Bayesian Post-Processing of Multi-Model Ensemble Forecasts

Clair Barnes

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

May 1, 2022

I, Clair Barnes, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Ensemble weather forecasts often under-represent uncertainty, leading to overconfidence in their predictions. Multi-model ensemble (MME) forecasts combining several individual ensembles have been shown to display greater skill than single-ensemble forecasts in predicting temperatures, but tend to retain some bias in their joint predictions. Established postprocessing techniques may be able to correct bias and calibration issues in univariate forecasts, but are generally not designed to handle multivariate forecasts (of several variables or at several locations, say).

This thesis proposes a flexible multivariate Bayesian postprocessing framework, based on a directed acyclic graph representing the relationships between the ensembles and the weather quantity of interest. The posterior forecast is inferred from available ensemble forecasts and an estimate of the shared model error, obtained from a collection of past forecast-observation pairs.

Further contributions of the thesis address the problem of improving the estimate of this shared discrepancy, in order to obtain a more accurate and better calibrated posterior forecast. The first of these focuses on the selection of appropriate training cases from which to estimate the required correction, using analogues selected on the basis of a low-dimensional representation of the prevailing weather regime predicted by each ensemble. The second is motivated by reducing the uncertainty about the discrepancy, by combining two sources of information through Bayes linear updateing. The second-order exchangeability representation underpinning Bayes linear statistics is extended and used to derive a fully multivariate linear adjustment that is

able to approximate probabilistic Bayesian inference and is flexible enough to accommodate a judgement of non-zero excess marginal kurtosis.

The new methods are evaluated on their performance in postprocessing operational forecasts winter surface air temperatures over selected regions of the UK.

Impact Statement

This thesis considers the problem of how weather forecasts from multi-model ensemble (MME) prediction systems might be combined into a single probabilistic forecast in a principled way. The first contribution of the thesis is a Bayesian framework for postprocessing of such forecasts, developed from a graphical model of the structure of the MME forecasting system which represents the relationships between the available forecasts and the unobserved quantities of interest, and so allows inference to be carried out on the verifying observation. The framework is flexible enough to be of potential use in applications other than the mid-term weather forecasting considered here. The most obvious candidate application is in longer-term forecasting (such as at the subseasonal-to-seasonal scale), where the problem is less of prediction of the weather conditions at a particular time on a particular day, and more of forecasting statistical properties of the weather at a particular time scale. It is also likely that estimation of the required forecast adjustments using weather regime analogues, which are derived from pressure fields that may reflect longer-term atmospheric trends, may produce more skilful forecasts at these leadtimes than either moving window training cases, which may be several months removed from the forecast verification dates, or direct analogues, which are typically based on surface weather quantities that tend towards climatology at these time scales. However, the proposed method is also widely applicable in fields outside of meteorology and the environmental sciences: it may be of use in any situation where several competing probabilistic or ensemble forecasts are available, along with an archive of previous verifying observations from

which the required correction may be estimated, such as in financial modelling or astrostatistics.

Likewise, the extension to Bayes linear covariance adjustment proposed in this thesis may be adopted anywhere that Bayes linear methods are already used to understand and predict the behaviour of complex systems existing applications are as diverse as assessing medical risks, estimating crop yields, and parametrising galaxy formation, suggesting that, like the Bayesian MME postprocessing framework, the multivariate Bayes linear adjustment may be suitable for a wide range of applications.

Acknowledgements

First and foremost, I would like to thank my primary PhD supervisor, Richard Chandler: for agreeing to supervise my research proposal despite an already full schedule; for the many hours of insightful, illuminating and above all, enjoyable discussions over the past few years; and for his unfailing patience in the face of many tangents.

Thank you also to my secondary supervisor, Chris Brierley, for his invaluable feedback and constant encouragement.

Thank you to Wilfrid Kendall, my MSc supervisor at Warwick, for encouraging me to pursue a PhD in the first place.

And of course, thank you to Chris, for his constant support through my years of what must have seemed like never-ending studenthood.

This research was supported by the Engineering and Physical Sciences Research Council under grant EP/N509577/1.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Operational weather forecasts are generated by complex numerical models designed to replicate the key features of the atmospheric processes that affect the weather. Uncertainty about the forecast issued by each model is assumed to be reflected by the spread of an ensemble consisting of multiple model runs, each initialised with slightly different starting conditions. Each weather forecasting centre runs its own model, each with its own error characteristics; however, the models may also retain some common biases, due to shared model characteristics such as the model resolution, approaches to parametrisation of unresolved quantities, and so on. Combining several competing ensemble forecasts into a single prediction that incorporates all elements of the associated uncertainty is therefore not a straightforward task. Some established approaches to this problem may ignore the structure of the multi-model ensemble, treating the individual ensembles as if they were independent; others fail to take into account all of the sources of uncertainty in the forecasts.

This thesis will address the question of how the output from several models might be combined into a single coherent forecast, in which the contributions of all elements of the model are explicitly represented. It takes as its starting point a Bayesian framework for quantifying the uncertainty in MMEs of projections of future climate, introduced in Chandler (2013), in which the relationships

between the ensembles and the quantity they aim to predict are explicitly represented. The contributions of this thesis are motivated by the challenges of applying this methodology to the slightly different problem of correcting bias and calibration errors in MMEs of operational weather forecasts. The first of these challenges is to develop a framework similar to that proposed in Chandler (2013) to represent the relationships between the forecast ensembles and the verifying observation; later developments are motivated by questions around how estimation of the required corrections might be improved. The contributions of the thesis are outlined in the next section.

1.2 Thesis statement

The first contribution of this thesis is to adapt the Bayesian framework introduced in Chandler (2013), in order to postprocess weather forecasts produced by multi-model ensemble (MME) prediction systems. The relationship between the MME and the observed weather is represented by a directed acyclic graph, and the conditional relationships encoded by the graph are combined with an estimate of the forecast error to obtain a posterior forecast which explicitly incorporates the full forecast uncertainty. The proposed Bayesian approach is used to postprocess operational forecasts of winter surface temperatures over the UK.

Any statistical postprocessing method requires a training set consisting of prior forecast-observation pairs, from which corrections to the forecast of interest can be estimated. The second major contribution of the thesis is a novel approach to selection of such a training set, by identifying forecast instances that predict similar weather regimes to the forecast of interest. In the postprocessing of weather forecasts, training cases are most commonly selected from the period immediately preceding the date of issue of the forecast, implicitly assuming that forecast errors during the training period will persist. Less frequently, the forecast error is assumed to be dependent on characteristics

of the forecast itself, and the training set is constructed from forecasts similar to the instance to be postprocessed. The selected cases are known as analogues to the current forecast instance, and are usually selected on the basis of a distance metric applied over the variable or variables to be postprocessed. However, the quality of the analogues selected tends to deteriorate as the dimension of the selection space increases.

The method proposed here uses principal component analysis of pressure fields, which are known to be closely related to prevailing weather regimes, to obtain a low-dimensional representation of the synoptic conditions for each forecast. Analogue training cases are then selected based on their closeness to the current forecast in this low-dimensional space, rather than in the potentially higher-dimensional space of the variables to be postprocessed.

In many applications, the dominant source of uncertainty in the posterior forecasts may arise from uncertainty about the discrepancy between the forecasts and observations. An improved estimate of the expectation and variance of this discrepancy may be obtained by using Bayesian inference to update prior estimates of these quantities with more recent information. However, natural conjugate inference typically constrains the parametric form of the distribution of the variable of interest, which may not always be appropriate; more flexible models, on the other hand, can be computationally costly, particularly when the forecasts to be postprocessed are of high dimension. Bayes linear adjustment is proposed as a flexible alternative, allowing second-order approximation of non-Gaussian symmetric distributions, but remaining quick and cheap to compute.

The third contribution of this thesis is to extend the second-order exchangeability representation underpinning Bayes linear statistics, to accommodate cross-products of residuals from quantities that are themselves second-order exchangeable; and to use this representation to derive a novel multivariate adjustment for covariance matrices. The proposed covariance adjustment can

be specified using the same parametrisations as in the scalar case, and is shown to be able to closely approximate the posterior distribution obtained using the natural conjugate normal-inverse-Wishart prior, while also accommodating a judgement of non-zero excess marginal kurtosis.

In addition to these three contributions already listed, two new approaches to the verification of multivariate forecasts are suggested. The first of these is a modification of the band depth rank histogram, a tool originally proposed to diagnose issues in the joint calibration of forecasts issued in the form of ensembles of deterministic forecasts. The proposed modification removes the need to draw a synthetic ensemble from the predictive density when computing the band depth rank for probabilistic forecasts, thus avoiding introducing an additional source of uncertainty. The second innovation is a gridplot designed to visualise the joint distribution of a pair of histograms, each of which may individually capture some but not all issues with the dependence structure of the multivariate predictive density.

1.3 Thesis outline

Chapter 2 introduces the numerical weather prediction (NWP) methods used to produce operational weather forecasts, and reviews a number of commonly used postprocessing methods. The limitations of these methods with respect to the evaluation of multi-model ensemble (MME) forecasts are discussed. Section 2.2 introduces the first major contribution of the thesis, adapting the framework proposed by Chandler (2013) to postprocess multi-model ensemble weather forecasts. The second contribution of the thesis, a novel method for the selection of training cases to be used in postprocessing, is presented in Section 2.3.

The forecast verification methods required for evaluation of the postprocessed forecasts are reviewed in Chapter 3, where two new approaches – a semiparametric band depth rank and a method of jointly evaluating two measures of multivariate calibration – are suggested. In Chapter 4 the proposed

Bayesian framework is used to postprocess forecasts of nighttime surface temperatures over thirteen regions across the UK, using a multi-model ensemble combining predictions from three operational weather centres.

Chapter 5 begins by reviewing the Bayes linear framework described in Goldstein and Wooff (2007), before extending the second-order exchangeability representation and developing a fully multivariate adjustment of the covariance matrices and mean vectors of all variables jointly in Section 5.3. In Section 5.4 the Bayes linear adjustment is shown, under certain parametrisations, to produce an asymptotic approximation to conjugate Bayesian inference. In Chapter 6 the Bayes linear adjustment is used to refine the postprocessed forecasts previously considered in Chapter 4, and a detailed investigation of the effect of varying the parameters used in the adjustment is carried out. Finally, the findings of this thesis and areas of particular interest for future research are discussed in Chapter 7.

Chapter 2

Multi-model ensemble forecasting

This chapter presents two of the major contributions of the thesis: the development of a Bayesian framework for the statistical postprocessing of multi-model ensemble weather forecasts, based on a model originally proposed by Chandler (2013) to combine and correct climate projections; and a novel method for selecting appropriate training cases for use in estimating the required corrections.

The chapter begins with an overview of methods used in weather forecasting, and a review of some of the methods commonly used to postprocess raw forecast ensembles. The new postprocessing framework is introduced in Section 2.2, and the proposed method of selecting relevant training cases in Section 2.3. An application of the new methods to postprocessing operational forecasts of surface air temperatures will be presented in Chapter 4.

Much of the content in this chapter and Chapter 4 is expanded from work already published in Barnes et al. (2019).

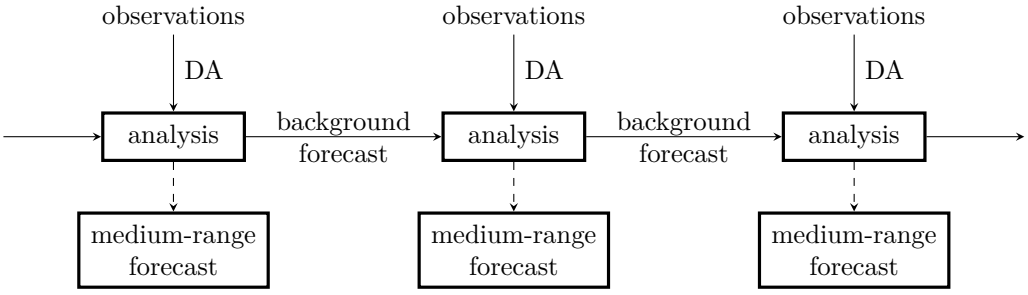
2.1 Current methods in weather forecasting

Operational weather forecasts looking more than a few hours ahead are generated by numerical weather prediction (NWP) systems: mathematical models which simulate the evolution of a large number – typically millions – of weather variables on a three-dimensional grid over the surface of the globe, representing

the physical processes of the atmosphere through a system of coupled partial differential equations.

In the simplest case a single sequence of forecasts is produced, in which each forecast is initialised using a 'best guess' approximation to the current atmospheric state known as the analysis (Bannister, 2017). Analyses are produced at regular time intervals – typically every 6-12 hours – by running a short-range 'background' forecast initialised using the previous analysis state, and adjusting this model output to more closely reflect observations made in the intervening hours. This ongoing process, known as data assimilation (DA), is represented in the schematic in Figure 2.1.

Figure 2.1: Schematic of the data assimilation (DA) cycle used to obtain initial conditions for medium-range forecasts (forecasts of weather up to 15 days ahead). Based on a similar representation in Garcia-Moya et al. (2016).



DA methods vary between weather centres: however, all are based on a representation of the relationship between the model state vector $\mathbf{x}$ and the observation vector $\mathbf{y}$ at time t (Carrassi et al., 2018). This relationship has the form

$$\mathbf{y}_t = H_t(\mathbf{x}_t) + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $H_t(\cdot)$ is a nonlinear function, known as the observation operator, that maps the model state $\mathbf{x}_t$ to observational space; and $\boldsymbol{\varepsilon}_t$ is an error term representing the discrepancy between the model state (after mapping into the same space as the observations) and the observations. This discrepancy accounts for the presence of observational errors, errors in the observation operator $H_t(\cdot)$, and errors arising from the model's representation of the physical processes

involved. The DA process is then formulated as a problem in Bayesian inference: given the new observations, what is the posterior distribution of the model state? Using Bayes' theorem, this can be expressed as

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}, \tag{2.2}$$

with the prior forecast distribution $p(\mathbf{x})$ provided by the background forecast, and $p(\mathbf{y})$ assumed to be constant; $p(\mathbf{y}|\mathbf{x})$ is determined by (2.1). In order to render this problem computationally tractable, all of the distributions involved are typically assumed to be Gaussian (Whitaker et al., 2008; Carrassi et al., 2018). Due to the huge number of variables involved, finding the full probability distribution of (2.2) remains computationally infeasible; instead, the DA process aims to find the posterior mean or mode.

The approaches used to solve (2.2) can generally be classified into two groups: variational methods, which minimise the difference between the model and the observations by minimising a cost function over all of the new observations simultaneously to find the mode of the distribution; and methods based on Kalman filtering and smoothing, which assimilate the observations in chronological order, 'nudging' the forecast trajectory with each new observation to find the posterior mean. A detailed discussion of these two families of methods is beyond the scope of this thesis: discussion of the underlying theory and recent developments can be found in, for example, Bannister (2017) and Carrassi et al. (2018). However, all of these methods share a need to estimate the discrepancy between the latest forecast and the observations, along with covariance matrices representing the uncertainty around the background forecast, and the observational uncertainty. Given estimates of these matrices, and the operator $H(\cdot)$, the model state $\mathbf{x}$ is optimised to find the best fit between the background forecast and the new observations. This optimal model state given the observations – the analysis – provides the initial conditions from which the next short-range background forecast will be produced, along with operational forecasts that will be run for longer periods before being issued. The model

is then again integrated forwards through time, and 'snapshots' of the model state at specified times after initialisation are stored and used to predict the corresponding weather states in the real world (Wilks, 2011).

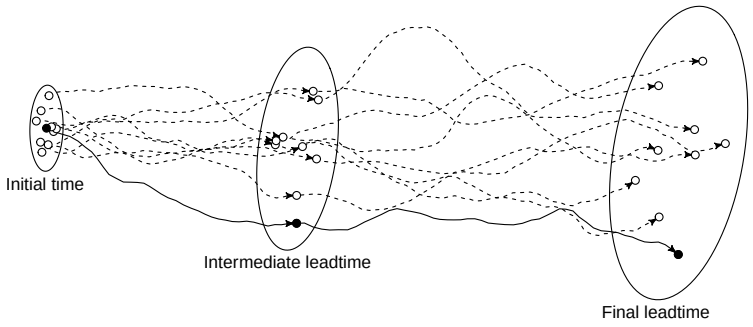
2.1.1 Ensemble forecasts

NWP models are deterministic, but highly nonlinear: small differences in the initial conditions can lead to very different model states relatively quickly (Kalnay, 2003). To understand the variability resulting from uncertainty about the 'true' initial state of the atmosphere, ensemble prediction systems (EPSs) are employed. An EPS is constructed by perturbing the initial analysis conditions that are propagated through the model, in order to create an ensemble that samples – at least, to some extent – the uncertainty about the true state of the atmosphere. Although, as with the data assimilation process, different forecasting centres design these perturbations in different ways, they are constructed in such a way that they are consistent with the uncertainties in the observations. The change in the spread of the ensemble as the trajectories of the individual members develop then reflects the changing uncertainty about the weather state over time. In recent years some operational forecasting centres have begun to use ensemble data assimilation methods or hybrid methods combining variational and ensemble elements to obtain an ensemble of initial conditions (Bonavita et al., 2012; Bowler et al., 2016; Carrassi et al., 2018); however, this makes little difference to the subsequent treatment of the forecasts, so the details of these methods are not considered here in detail.

An idealised two-dimensional schematic of the development of an ensemble forecast is presented in Figure 2.2. As the NWP simulation runs, each ensemble member follows its own trajectory through the model's state space, reflecting the ways in which the initial atmospheric state has been transformed by the dynamics of the model.

Epstein (1969) noted that, as in Figure 2.2, the behaviour of the ensemble mean differs from the behaviour of the ensemble member initialised with the

Figure 2.2: Schematic of the development of an ensemble forecast over time. The filled circle represents the ensemble member whose initial state represents the analysis, the 'best guess' at the current atmospheric conditions. Unfilled circles represent perturbations of the analysis state, forming an ensemble that approximates the probability distribution represented by the ellipse at each time step.

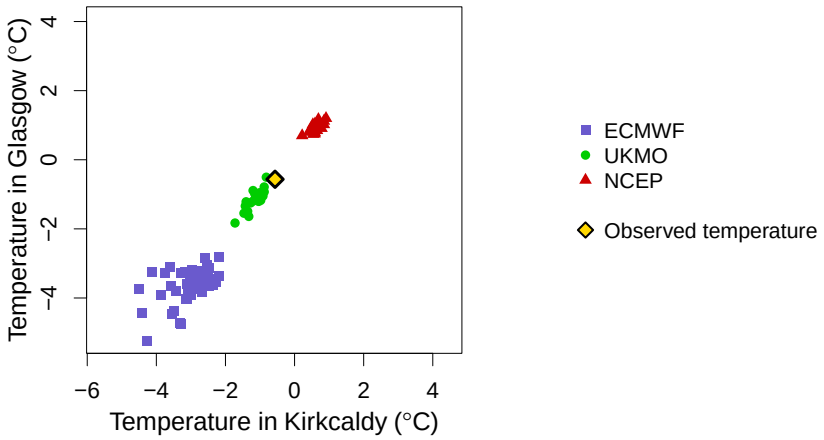


best estimate of initial conditions, and concluded that the best forecast provided by an ensemble is not given by the single member initialised with the best guess at the initial conditions. Intuitively, forecast users might have more confidence that the ensemble mean is close to the the eventual state of the atmosphere if the dispersion of the ensemble is small: if the ensemble members are all very different from each other, or if there is a bifurcation in the ensemble, the future state of the atmosphere is less certain (Wilks, 2011). However, quantifying this intuition with regards to an ensemble forecast is often not straightforward; furthermore, the problem of predicting the weather is sufficiently complex that this kind of simple intuition may be misleading. A common approach, particularly when the quantities being forecast are continuous, is to treat the ensemble members as if they were independent and identically distributed samples drawn from some underlying probability distribution with a known form, and to estimate the parameters of that distribution using summary statistics of the ensemble members (Wilks, 2011). In this way the salient features of the ensemble forecast can be expressed in an efficient, tractable and easily interpreted way, albeit at the cost of some detail. This approach relies on the implicit assumption that the realised value will be drawn from the estimated distribution.

2.1.2 Operational forecasts used in this thesis

An example of a collection of ensemble forecasts of surface temperature from three NWP models is shown in Figure 2.3. These ensemble forecasts were produced by three operational weather forecasting centres: the European Centre for Mid-term Weather Forecasting (ECMWF), National Centre for Environmental Prediction (NCEP) and UK Met Office (UKMO), which have 50, 20 and 23 perturbed members respectively. The methods described in this thesis will be evaluated in Chapters 4 and 6 using ensemble forecasts issued by these three centres of surface temperatures at midnight during the winter period (December-January-February, excluding leap days, giving 90 days of forecasts per year) during the seven years from December 2007 to February 2014. The forecasts were downloaded from the TIGGE archive (Bougeault et al., 2010), with these particular models selected for inclusion because they are among the largest of the ten ensembles for which data are consistently available during the study period. The study period was initially chosen as a pilot study, but could not be extended due to damage to the tapes storing the data at ECMWF (ECMWF, 2021e).

Figure 2.3: Two-day-ahead ensemble predictions of temperatures in Kirkcaldy and Glasgow, issued on 30 January 2010 by the European Centre for Medium-Range Weather Forecasting (ECMWF, 50 members), the UK Met Office (UKMO, 23 members), and the National Centers for Environmental Prediction (NCEP, 20 members).

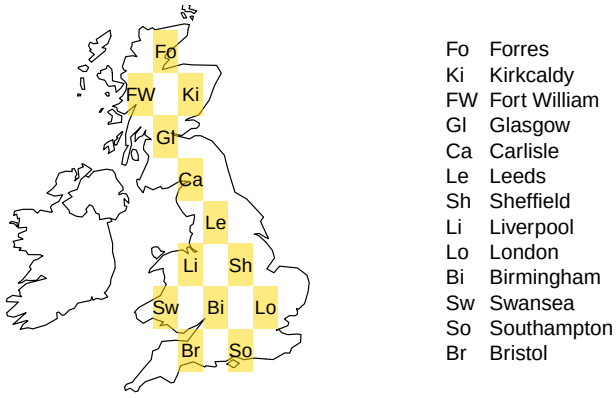


The forecasts used in the applications discussed in Chapters 4 and 6 are issued

at 24-hour intervals, at leadtimes of 0 to 15 days, with the 0-day-ahead forecasts corresponding to the initial conditions provided by the analysis. The term 'forecast instance' will be used to refer to forecasts issued on a single day for a given leadtime, and at a particular synoptic time: $7 \times 90 = 630$ forecast instances are available at each leadtime.

The forecasts were downloaded on a 1° latitude-longitude grid covering the British Isles, from 50 to 60°N and 6°W to 2°E. The study area consists of forecasts at 13 'locations': alternating grid cells over the land mass of Great Britain (Figure 2.4). This choice was made to limit the size of the data set for ease of processing and interpretation, while still including relatively heterogeneous climatologies. The observed value of the weather corresponding

Figure 2.4: Regions included in the study. Each cell is labelled with the name of the largest city within its boundaries for easier reference.



to a particular forecast instance is known as the verifying observation. However, observations in the real world are taken at precise geographical locations, and so cannot be compared directly to the gridded forecasts produced by NWP models. For this reason, weather forecasts are usually evaluated against gridded datasets known as reanalysis products. A meteorological analysis is a gridded dataset produced by assimilating historical observations of the real atmospheric state into a 'hindcast' – a forecast of historical weather – produced by a model similar to the operational NWP model. Throughout this thesis, forecasts will be evaluated against the ERA-Interim reanalysis of surface air temperatures at the same grid resolution as the ensemble forecasts (Dee et al., 2011). The term

'verifying observation' will be used for convenience, although strictly speaking, the forecasts are verified against reanalyses, rather than observations.

It should be noted that it is possible that the ECMWF ensemble forecasts may perform better than the other two ensembles in this respect, since these forecasts are produced by a similar model to the reanalysis. However, the relative performance of different combinations of contributing ensembles is outside of the scope of this thesis.

The example shown in Figure 2.3 is a two-day-ahead forecast of the temperature in two locations, showing the members of the three ensembles described above along with the verifying reanalysis. The temperature predictions are strongly correlated within each of the ensembles: each ensemble's members are tightly clustered, indicating a high degree of confidence in the forecast, although there is no overlap between the three ensembles, and only the UKMO ensemble is actually close to the verifying observation, with the NCEP ensemble predicting temperatures 1-2 degrees higher, and the ECMWF ensemble predicting temperatures 2-5 degrees lower. While this example was chosen as a particularly clear illustration of the specific issues discussed here, it is not difficult to find similar examples of forecasts in which some or all of the ensembles do not overlap. An immediate consequence of this is that the assumption that the verifying observation is sampled from the ensemble distribution has been shown to be incorrect – or, at the very least, to be possibly true of only one of the ensembles.

2.1.3 Multi-model ensemble forecasts

If uncertainty about the initial system state were the only source of uncertainty in the forecasts, a single ensemble of predictions could be treated as a sample from the probability distribution describing the atmospheric state, and used to make probabilistic predictions about the weather; if this were the case, the ensembles in Figure 2.3 would overlap. However, further uncertainty arises from modelling choices like boundary conditions and parametrisations, as well as processes at unresolved scales. In consequence, the ensemble spread tends

to underestimate the forecast error, with this underdispersiveness becoming worse at longer leadtimes (Weigel et al., 2008). Weather quantities that depend primarily on physical processes, such as pressure fields, are typically well represented in the models; however, the surface weather quantities that are of greatest interest to many users, such as surface temperatures, precipitation, and wind, are more sensitive to the precise formulation of the model and numerical schemes for solving the associated equations, and model biases are known to be particularly prevalent in these surface weather quantities, as can be seen in Figure 2.3.

Many NWP systems now include schemes to partially sample this model uncertainty, perturbing not only the initial conditions for each ensemble member but also the parameters used to represent unresolved atmospheric processes such as cloud formation and precipitation. These stochastic parametrisations aim to address errors arising from the choice of parametrisation algorithms (Baker et al., 2014; Palmer et al., 2009). However, no single forecast ensemble is able to fully represent the uncertainty due to errors and approximations in the model (Whitaker et al., 2008; Johnson and Swinbank, 2009; Carrassi et al., 2018).

One approach to understanding and incorporating this source of uncertainty is to construct a multi-model ensemble (MME) that is able to combine the output of several EPSs into a single forecast. Ideally, an MME should include models with different model physics and dynamics, in order to explore as much of the spectrum of model solutions as possible (Fritsch et al., 2000; Hagedorn et al., 2005; Sansom et al., 2021). Some of the key features of the three models introduced in Section 2.1.2 are summarised in Table 2.1 (ECMWF, 2021a). The details of the schemes used to generate the initial perturbations and to perturb the model physics are not discussed here; the purpose of the comparison is merely to illustrate that the three models chosen use different methods to perturb both the initial conditions and the model physics, and so might be expected to explore quite different regions of the parameter space.

Table 2.1: Key features of the three operational weather forecasting models in the MME considered in this thesis during the study period from 2007-2013: ensemble size, spatial and temporal resolution, and methods used to assimilate observations and to perturb initial conditions and physical parametrisations. In these models, data assimilation is carried out using either 4DVar (Rawlins et al., 2007) or an ensemble Kalman filter (Houtekamer and Mitchell, 1998, EnKF). The initial perturbations are generated by either singular vectors (Palmer, 1995, SV); EnKF; or an ensemble transform Kalman filter (Bowler and Mylne, 2009, ETKF). Model physics are perturbed by stochastic perturbed parameterisation tendencies (Lock et al., 2019, SPPT), stochastic kinetic energy backscatter (Tennant et al., 2011, SKEB), or a random parameter scheme (Bowler et al., 2008, RP).

Model	**No. of members**	**Horizontal resolution**	**Integration timestep**
ECMWF	50	16km*	20mins*
NCEP	23	25km	7.5mins
UKMO	20	21km	7.5mins

Model	**Data assimilation**	**Initial perturbations**	**Perturbed physics**
ECMWF	4DVar	SV	SPPT, SKEB
NCEP	EnKF	EnKF	SPPT, SKEB
UKMO	4DVar	Scaled ETKF	SKEB, RP

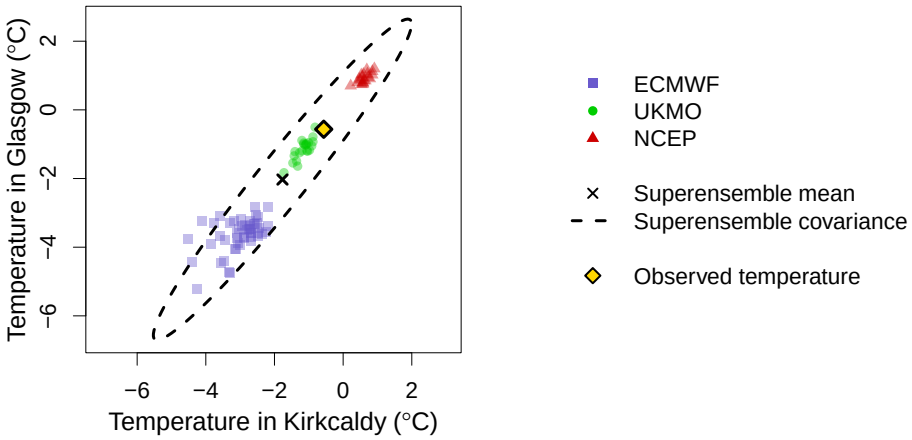
*After ten days the ECMWF model changes to 32km horizontal resolution with an integration timestep of 45 minutes.

2.1.3.1 The pooled MME 'superensemble' forecast

Several studies have shown that even a very simple MME forecast obtained by unweighted averaging of all available member forecasts can, in the long run, outperform even the best of its constituent models (Hagedorn et al., 2005; Doblas-Reyes et al., 2005; Matsueda et al., 2007; Weigel et al., 2008; Johnson and Swinbank, 2009; Weigel et al., 2009). Moreover, the spread of the pooled ensemble members has been shown to better reflect the true forecast uncertainty than that of any single component ensemble, with the improvements shown to be due to extra information in the additional ensembles, and not simply to increased ensemble size (Hagedorn et al., 2005): Johnson and Swinbank (2009) attribute this improvement to the various models exploring different regions of the phase space.

Figure 2.5 shows a bivariate example of this 'superensemble' forecast derived from the MME example shown in Figure 2.3, along with the verifying reanalysis. The individual ensembles show a roughly elliptical scatter; this is typical of short- and medium-range forecasts of surface temperatures, which are therefore commonly characterised by multivariate normal distributions (Wilson et al., 1999; Wilks, 2002, 2011). Similarly, the superensemble forecast is characterised by a multivariate normal distribution, using the mean vector and covariance matrix from the pooled ensemble members. The ellipse defines a region containing 95% of the predictive density, and is much larger than the spread of any single ensemble. The three ensembles in this MME are not balanced, having 50, 20 and 23 members; because of this, the superensemble mean is shifted toward the mean of the largest ensemble – which, in this instance, happens to be the furthest from the verifying observation. As a result, this superensemble predictive distribution places a lot of probability in the bottom-left corner of the plot, far from any ensemble members and from the temperature actually observed.

Figure 2.5: Superensemble derived from the MME forecast shown in Figure 2.3. The ellipse defines a 95% prediction region calculated from a bivariate normal distribution with the same mean and covariance matrix as the pooled superensemble.



The rationale behind this approach relies on an implicit assumption that the multi-model ensembles are all centred on the 'true' distribution of the verifying observation that they aim to forecast, so that any biases displayed by the

individual ensembles will cancel one another out (Hagedorn et al., 2005). If this were the case, then a MME constructed from a sufficiently large number of ensembles – or indeed a single sufficiently large ensemble – should be able to reduce the forecast uncertainty to almost zero. However, there is no guarantee that the individual ensembles will not share similar biases, perhaps due to the use of similar representations of physical processes, similar grid resolutions, or similar parametrisations of unresolved processes. As a result, there are limits to the improvements that can be achieved by simply adding more ensembles: this only produces an improved estimate of the MME consensus. Some form of bias correction is therefore necessary to account for the discrepancy between the MME consensus and the quantity (or quantities) that the MME intends to predict. Furthermore, if a forecast is to be useful to support planning and decision making, it is important not only to correct any biases in the deterministic forecast, but to accurately quantify the associated uncertainty (Wilks, 2011). Some form of postprocessing of the MME forecast is therefore required in order to correct for biases, and if necessary, to adjust the spread of the forecast to reflect the true uncertainty.

2.1.4 Forecast postprocessing

Forecast postprocessing is the process of converting the raw model output from the ensemble members into a form that can be used to issue probabilistic forecasts, while accounting for model biases and providing a defensible assessment of uncertainty. In statistical postprocessing, the necessary corrections are estimated using a training set of previous forecast-observation pairs, each consisting of a forecast issued by the same NWP model (or models) as the forecast of interest, along with its verifying observation; approaches to selecting an appropriate training set are considered in Section 2.3.

Many methods have been proposed for postprocessing the output from a single EPS (Vannitsem et al., 2018, §3). Commonly employed approaches include regression-based approaches (Glahn and Lowry, 1972; Jewson et al., 2004; Gneiting et al., 2005; Scheuerer and Hamill, 2015a; Hess, 2020), analogue

ensembles (Hamill and Whitaker, 2006; Hamill et al., 2006; Sperati et al., 2017; Hu et al., 2020), Bayesian Model Averaging (BMA) (Raftery et al., 2005), the adjustment of rank histograms to the desired uniformity (Hamill and Colucci, 1997; Eckel and Walters, 1998), 'dressing' forecasts with historical error statistics (Roulston and Smith, 2003), Kalman filtering (Delle Monache et al., 2011; Pelosi et al., 2017), quantile regression methods (Bentzien and Friederichs, 2012; Taillardat et al., 2016), and more recently, machine learning methods (Rasp and Lerch, 2018; Scher and Messori, 2018; Taillardat and Mestre, 2020). However, few of these methods have been extended to accommodate the postprocessing of MME forecasts. This is perhaps because MMEs are less commonly employed in weather prediction than in climate modelling, where – due to the relatively small numbers of runs available from each model – MMEs are often constructed in such a way that they can be treated as a single ensemble with exchangeable members (Rougier et al., 2013; Abramowitz et al., 2019; Sansom et al., 2021), rather than with the clustered structure seen in MMEs of weather forecasts.

Bayesian Model Averaging has been implemented for MME weather forecasts (Fraley et al., 2007, 2010), but the resulting predictive distributions are mixtures, which are not as straightforward to interpret as parametric predictive distributions, and so may be less appealing to end users; furthermore, the method rests on an implicit assumption that one of the models under consideration is in fact the model from which the verifying observation is drawn, which is a difficult claim to defend. Regression-based postprocessing – commonly known in the meteorological literature as Model Output Statistics, or MOS (Glahn and Lowry, 1972) – is a well-established approach commonly used in the postprocessing of operational forecasts (Wilks, 2011; Hess, 2020), and is easily adapted to accommodate MME forecasts, as will be shown in Section 2.1.4.1. For this reason, in this thesis MOS-based postprocessing will be used as a benchmark with which to compare the novel Bayesian method introduced in Section 2.2.

2.1.4.1 Nonhomogeneous Gaussian regression

Model Output Statistics (MOS; Glahn and Lowry, 1972) is the name given by weather forecasters to a family of postprocessing methods that use NWP model output as the covariates in a regression model. MOS postprocessing methods are well established in operational forecasting (Mylne et al., 2002; Glahn et al., 2009; Hess, 2020). The exact form of MOS applied will vary according to the weather quantities to be predicted; techniques based on different parametric forms have been developed to obtain probabilistic forecasts of wind speed (Thorarinsdottir and Gneiting, 2010) and precipitation (Scheuerer, 2014). In the case of the surface temperatures considered here, where the ensemble forecasts are generally assumed to have approximately Gaussian distributions, MOS typically takes the form of a nonhomogeneous Gaussian regression (NGR; Gneiting et al., 2005; Hagedorn et al., 2008, 2012; Junk et al., 2015). The NGR MME postprocessing model proposed by Gneiting et al. (2005) is described in some detail here, in part to motivate elements of the Bayesian framework in Section 2.2.

2.1.4.1.1 NGR for a univariate MME forecast

NGR was originally proposed as a technique to postprocess a single ensemble forecast of a univariate quantity, using all of the ensemble members as predictors. Because the members of a single ensemble are obtained by randomly perturbing the initial conditions, they are exchangeable: the joint probability distribution of the ensemble members does not change when the order of the ensemble members is permuted (De Finetti, 1992; Bröcker and Kantz, 2011). Jewson et al. (2004) and Gneiting et al. (2005) recommend that, in such cases, groups of exchangeable individuals should be replaced by their mean value. Therefore, given a multi-model ensemble forecast from p models, let $\bar{y}_1, \ldots, \bar{y}_p$ denote the ensemble means for the forecasts of some weather quantity Y, and let $\bar{s}^2$ denote

the sample variance of these means,

$$\bar{s}^2 = \frac{1}{p-1}\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2, \qquad \text{where } \bar{y} = \frac{1}{p}\sum_{i=1}^{p}\bar{y}_i. \tag{2.3}$$

The value of Y to be predicted is denoted Y_0. Here, upper case indicates a random variable, while lower case denotes realised values of those random variables.

In nonhomogeneous Gaussian regression – so called because the predictive variance is not fixed, but depends on the inter-ensemble variance $\bar{s}^2$ for each forecast instance – the predictive distribution is determined from the quantities $\bar{y}_1, \ldots, \bar{y}_p$ and $\bar{s}^2$, with

$$Y_0 \sim N\left(a + b_1\bar{y}_1 + \cdots + b_p\bar{y}_p, c + d\bar{s}^2\right). \tag{2.4}$$

The coefficients $a, b_1, \ldots, b_p, c, d$ are estimated either by maximising the likelihood (Jewson et al., 2004) or by minimising the continuous ranked probability score (CRPS; Gneiting et al., 2005) over a training set of previous forecast-observation pairs; the CRPS, which will be discussed in detail in Section 3.2.1, is a metric commonly used to evaluate the skill of probabilistic and ensemble weather forecasts. In order to ensure that the NGR variance is strictly positive, the estimators $\hat{c}$ and $\hat{d}$ are constrained to be greater than zero (Gneiting et al., 2005).

The fitted coefficients $\hat{b}_1, \ldots, \hat{b}_p$ can be interpreted as weights applied to each member of the MME depending on their relative skill in predicting the verifying observations in the training set, with $\hat{a}$ providing a simple bias correction to the weighted mean forecast thus obtained. When (as in Figure 2.5, for example) several of the ensembles are highly collinear, the estimates of $\hat{b}_1, \ldots, \hat{b}_p$ are likely to be unstable, and negligible weights may be assigned to all but the single most skilful ensemble member (Gneiting et al., 2005). The weights should therefore not be interpreted as direct measures of the relative skill of each ensemble.

The variance component of (2.4) is intended to capture the fact that a

systematic relationship is expected to exist between the magnitude of the forecast errors and the spread of the ensembles producing them (Gneiting et al., 2005). The strength of the spread-error relationship within the training data is reflected in $\hat{d}$, with larger values indicating a stronger relationship; when $\hat{d} = 0$, the spread and error are essentially independent, and the resulting distribution is reduced to a linear regression, with the ensemble variance inflated by $\hat{c}$ to replicate the variance of the errors in the training data.

2.1.4.1.2 NGR for multivariate forecasting

NGR is a univariate postprocessing method, estimating regression coefficients for a single variable. In weather forecasting, applications typically require joint forecasts of several variables simultaneously, whether of multiple weather quantities; of a single weather quantity at successive time steps; of a single weather quantity at several sites; or some combination of these. In principle, direct multivariate MOS methods are possible, but these require simultaneous estimation of large numbers of parameters even when the number of predictands is relatively low (Berrocal et al., 2007; Schuhen et al., 2012; Pinson, 2012; Wilks, 2015). Because of this, a more commonly used approach is to postprocess each of the m marginal forecasts independently, where m in this case denotes the number of locations; to concatenate the postprocessed forecast means into a single vector $\boldsymbol{\mu}_{ngr}$; and to combine this vector with a correlation matrix $\mathbf{R}$ defining the joint dependence structure, in order to produce a multivariate-normal joint predictive density

$$\mathbf{Y}_0 \sim MVN\Big(\boldsymbol{\mu}_{ngr}, \mathbf{V}_{ngr}^{1/2}\,\mathbf{R}\,\mathbf{V}_{ngr}^{1/2}\Big), \tag{2.5}$$

where $\mathbf{V}_{ngr}$ is the $m \times m$ diagonal matrix of NGR marginal predictive variances (Berrocal et al., 2007; Feldmann et al., 2015; Schefzik, 2016).

The method used to estimate $\mathbf{R}$ can be chosen depending on the context: time series or spatial correlation structures may be used if appropriate, but if the predictands are genuinely multivariate without a clear underlying temporal or

spatial structure then an unstructured correlation matrix can also be used. If the training datasets are large enough to support the estimation of an unstructured covariance matrix, this will usually be quicker and more straightforward than fitting a structured covariance matrix, a process that would typically involve either the fitting of multiple competing options to identify the best fit, or the manual inspection of sample correlation matrices to identify plausible candidates.

Equation (2.5) suggests that $\mathbf{R}$ should be the correlation matrix of the observed residuals after regressing out the effects of the raw forecasts. However, estimation of the dependence structure based on the forecast errors is not commonly used when postprocessing weather forecasts. Instead, $\mathbf{R}$ is typically estimated using one of two methods.

One is to use the correlation matrices of the raw forecasts that are undergoing postprocessing: this approach was originally proposed for postprocessing of forecasts issued in the form of discrete ensembles, and is known as Ensemble Copula Coupling (ECC; Schefzik et al., 2013; Schefzik, 2016). Alternatively, $\mathbf{R}$ may be estimated from a long series of historical observations; this method is usually known as the Schaake Shuffle (Clark et al., 2004). The rationale behind this choice is that, if the dependence structure is expected to depend on the prevailing atmospheric conditions, then an empirical copula based on the raw ensembles might be expected to yield better forecasts than one based on an unconditional climatology; however, if the raw ensembles fail to capture a realistic dependence structure, then estimation from historical observations may be expected to result in better joint calibration (Wilks, 2015). Several studies have found that forecast skill is improved when the dependence structure is estimated instead from forecasts that are similar, in some sense, to the current forecast (Junk et al., 2015; Schefzik, 2016; Lerch and Baran, 2017), essentially conditioning the climatological sample on the prevailing weather state of the forecasts. Such forecasts are known as analogues to the current forecast instance, and will be discussed further in Section 2.3. In the application

in Section 4.2 the correlation matrix **R** will be estimated using the sample correlation matrix of the verifying observations corresponding to the training cases used to estimate the coefficients of the marginal predictive distributions.

2.1.4.1.3 Limitations of the NGR approach

Nonhomogeneous Gaussian regression offers an intuitive approach to simultaneously correct a multi-ensemble forecast and assess the relative performance of the constituent ensembles. However, NGR methods fail to exploit the full range of information provided by the available ensemble forecasts. By working with only the ensemble mean forecasts, information about the spread (or confidence) within each ensemble is lost; and by potentially discarding whole ensembles in the weighted average, information from the full spread of the MME is lost – although this was found to be a key part of the success of superensemble forecasts in Hagedorn et al. (2005) and Weigel et al. (2009), where even less skilful ensembles were able to contribute to a well-calibrated combined forecast by increasing the forecast spread and exploring additional regions of the phase space. Furthermore, NGR fails to account for uncertainty in the coefficient estimates, which would usually be accounted for in a linear regression by issuing predictions of new observations in the form of a t-distribution with inflated variance.

The next section presents an intuitive and easily applied framework designed to incorporate all of the available information about the forecast of interest, and produce a properly calibrated, bias-corrected multivariate multi-ensemble forecast.

2.2 A new approach to MME postprocessing

This section describes a novel approach to the postprocessing of multivariate forecasts produced by multi-model ensemble prediction systems. A graphical representation of the relationships between the ensembles is used to derive an expression for the posterior distribution of the weather quantities of interest in a

Bayesian framework. Sources of uncertainty about each element of the forecast are explicitly quantified in a way that is easy to understand and interpret.

The section begins by introducing the conceptual MME framework, before developing a model for the postprocessing of quantities such as surface temperatures, which can be treated as having an approximately multivariate normal distribution.

2.2.1 Conceptual representation of the MME

The framework is developed from that proposed in Chandler (2013) to combine and correct climate projections. A key difference here is that, while climate projections aim to make statements about the statistical properties of future weather such as regional or global mean temperatures over a long period of time, the aim in the current weather forecasting context is to forecast the actual weather quantities themselves. The framework aims to accommodate the relationships not only between the ensemble members and the ensembles themselves, but also explicitly accounts for the fact that, due to shared elements of model design such as similar resolutions and parametrisations, the ensembles are likely to be more similar to each other than they are to the real world (Wilks, 2011).

For a single forecast instance, the aim of the postprocessing is to obtain a probabilistic prediction of the collection of weather quantities $\mathbf{Y} = \begin{bmatrix} Y_1 & \cdots & Y_m \end{bmatrix}$, say, where m denotes the number of weather quantities to be postprocessed simultaneously. Let $\mathbf{Y}_0$ denote the value of $\mathbf{Y}$ against which the prediction is to be verified.

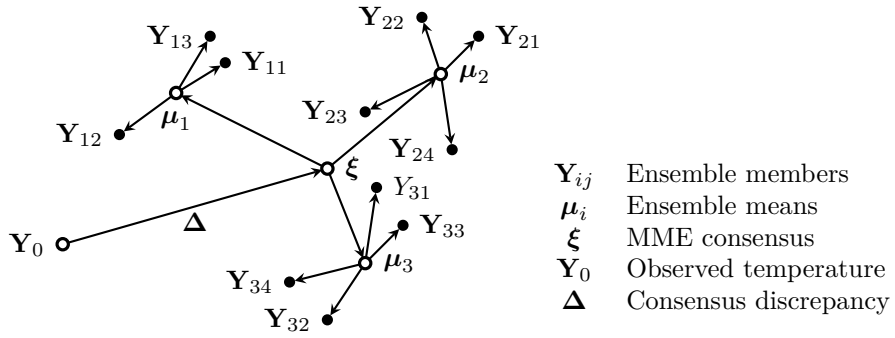
The MME consists of p ensembles, with the ith ensemble having n_i exchangeable members. The jth member of the ith ensemble is labelled $\mathbf{Y}_{ij}$. Reflecting the tendency (as seen in Figure 2.3) of the members of one ensemble to be more similar to each other than to the members of another ensemble, the members of the ith ensemble are centred on an ensemble-specific mean $\boldsymbol{\mu}_i$. These ensemble means are themselves centred on an unobserved MME 'consensus', $\boldsymbol{\xi}$, reflecting the fact that the ensembles may resemble each other

more closely than they resemble reality. This consensus can be thought of as the centre of the population of possible ensembles; if it were possible to sample an ensemble from each of an infinite collection of models for a particular forecast instance, the mean of the infinite sample of $\boldsymbol{\mu}_i$s would lie not at $\mathbf{Y}_0$ but at $\boldsymbol{\xi}$.

The discrepancy between the ensemble consensus and the value of $\mathbf{Y}_0$ actually observed is labelled $\boldsymbol{\Delta}$; this quantity explicitly accounts for the possibility that if, for example, one EPS predicts too low a temperature, the other EPSs will display a similar tendency.

The relationships between these quantities are represented in the schematic in Figure 2.6. The ensemble members $\{\mathbf{Y}_{ij}\}$ are represented by filled circles, indicating that these quantities are known at the time the forecast is issued. Quantities that are unknown, but either are of direct interest (the future value $\mathbf{Y}_0$) or are required to fully specify the MME model (the ensemble means $\{\boldsymbol{\mu}_i\}$ and consensus $\boldsymbol{\xi}$) are represented by hollow nodes.

Figure 2.6: Schematic representation of the relationships between the elements of the multi-model ensemble forecasting system. Quantities known at the time of issuing the forecast are shown as filled nodes, with unknown quantities represented by open nodes. The covariance matrices relating to each quantity are not shown.



The arrows linking the nodes encode conditional independence relationships: if there is no path between point A and point B without passing through point C, then A and B are conditionally independent, given C. Thus, since there is no path from $\mathbf{Y}_{i1}$ to $\mathbf{Y}_{i2}$ that does not pass through $\boldsymbol{\mu}_i$, $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ are independent, given $\boldsymbol{\mu}_i$. This assumption seems intuitively reasonable if the ini-

tial conditions of each ensemble member are sampled independently. Likewise, if the location of the 'consensus' node $\boldsymbol{\xi}$ is given, knowledge of the position of one ensemble mean $\boldsymbol{\mu}_1$ provides no additional information about the position of the ensemble mean $\boldsymbol{\mu}_2$. Without conditioning on the MME consensus $\boldsymbol{\xi}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are not independent, because knowing the mean forecast of one ensemble would provide some information about the likely value of the other. The conditional independences implied by this graph will be exploited in the subsequent derivations.

2.2.2 Postprocessing the MME

In the application described in Section 2.1.2, $\mathbf{Y}$ contains surface temperatures at each of the $m = 13$ grid cells shown in Figure 2.4, for which a multivariate-normal representation is appropriate. The members $\{\mathbf{Y}_{ij}\}$ of the ith ensemble are assumed to be exchangeable, and to be independent conditional on the ensemble's population mean $\boldsymbol{\mu}_i$, so that

$$\mathbf{Y}_{ij}|\boldsymbol{\mu}_i \sim MVN(\boldsymbol{\mu}_i, \mathbf{C}_i)\,, \tag{2.6}$$

where $\mathbf{C}_i$ denotes the ensemble covariance matrix. The individual ensemble mean forecasts $\boldsymbol{\mu}_i$ are themselves assumed to be dispersed around a mutual consensus, $\boldsymbol{\xi}$, and to be independent of one another only conditional on this consensus,

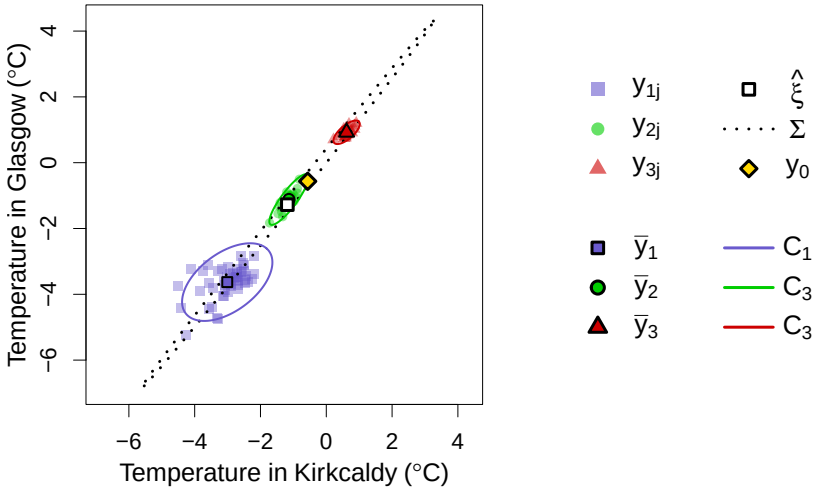
$$\boldsymbol{\mu}_i|\boldsymbol{\xi} \sim MVN(\boldsymbol{\xi}, \boldsymbol{\Sigma})\,. \tag{2.7}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix capturing the relationships between the ensemble means $\boldsymbol{\mu}_i$. The MME consensus $\boldsymbol{\xi}$ can be decomposed into the 'true' value, $\mathbf{Y}_0$, plus a shared discrepancy $\boldsymbol{\Delta}$ about which there is also some uncertainty, as

$$\boldsymbol{\xi} = \mathbf{Y}_0 + \boldsymbol{\Delta} \qquad \text{where } \boldsymbol{\Delta} \sim MVN(\boldsymbol{\eta}, \boldsymbol{\Lambda}) \tag{2.8}$$

Equations (2.6) to (2.8) correspond to the arrows in the schematic in Figure 2.6. Figure 2.7 represents the elements of the MME in Figure 2.3 in terms of the quantities defined in (2.6) and (2.7).

Figure 2.7: The MME forecast in Figure 2.3, represented in terms of the distributions described in (2.6)-(2.8). The coloured ellipses represent the individual covariance matrices C_1, C_2 and C_3 while the black dotted ellipse represents the covariance matrix $\mathbf{\Sigma}$.



2.2.2.1 Simplification of the MME structure

The structure in Figure 2.6 – and hence the derivation of the posterior – can be simplified by exploiting the fact that, having obtained $\mathbf{C}_i$, only the ensemble means $\overline{\mathbf{Y}}_i$ are required to represent all of the information contained in the ensemble. This means that, without loss of information, the individual members $\{\mathbf{Y}_{ij}\}$ of ensemble i can be replaced by the ensemble mean $\overline{\mathbf{Y}}_i$, and (2.6) can be replaced with

$$\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i \sim MVN\Big(\boldsymbol{\mu}_i, n_i^{-1}\mathbf{C}_i\Big)\,. \tag{2.9}$$

The proof of the equivalence of (2.6) and (2.9) is derived in Appendix A.2. This simplified form can in turn be combined with (2.7), and used to deduce that the sampled ensemble means $\left\{\overline{\mathbf{Y}}_i\right\}$ are independent of each other conditional

on the ensemble consensus $\boldsymbol{\xi}$, with

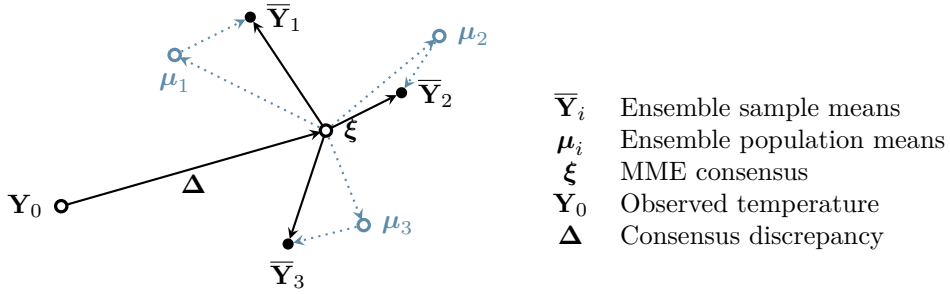
$$\overline{\mathbf{Y}}_i|\boldsymbol{\xi} \sim MVN\Big(\boldsymbol{\xi}, \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i\Big)\,. \tag{2.10}$$

To further simplify the notation, (2.10) can be written as

$$\overline{\mathbf{Y}}_i|\boldsymbol{\xi} \sim MVN(\boldsymbol{\xi}, \mathbf{D}_i) \qquad \text{where } \mathbf{D}_i = \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i. \tag{2.11}$$

The graph corresponding to this simplified structure is shown in Figure 2.8. All of the information contained in the $\{\mathbf{Y}_{ij}\}$ is now captured by the sample ensemble means $\left\{\overline{\mathbf{Y}}_i\right\}$, with associated uncertainty $n_i^{-1}\mathbf{C}_i$. The unobserved nodes corresponding to the ensemble population means $\{\boldsymbol{\mu}_i\}$ are no longer required for the inferential calculations, forming part of the unique path from $\boldsymbol{\xi}$ to each $\overline{\mathbf{Y}}_i$, and so can be removed from the graph structure.

Figure 2.8: Simplified schematic representation of the relationships between the elements of the multi-model ensemble forecasting system. Quantities known at the time of issuing the forecast are shown as filled nodes, with unknown quantities represented by open nodes. Dotted lines indicate redundant nodes that have been bypassed.



An important difference between this representation and NGR is that the latter approach uses only the sample variance of the ensemble means (the diagonal elements of $\boldsymbol{\Sigma}$), discarding the additional information on within-ensemble spread.

2.2.2.2 The posterior distribution of $\mathbf{Y}_0$

Having specified parametric forms for the quantities represented in Figure 2.8 and the conditional relationships between them, it is now possible to derive the distribution of $\mathbf{Y}_0$. As in Chandler (2013), this is most conveniently done in a Bayesian framework which also allows the incorporation of additional knowledge about $\mathbf{Y}_0$ via a prior distribution. To ensure a tractable form for the posterior distribution, a multivariate normal prior is used, with

$$\mathbf{Y}_0 \sim MVN(\boldsymbol{\alpha}, \boldsymbol{\Gamma}). \tag{2.12}$$

It can be shown, using arguments adapted from those in Chandler (2013), that the posterior distribution of $\mathbf{Y}_0$, conditioned on the set of all ensemble forecasts $\{\mathbf{Y}_{ij}\}$, is itself multivariate-normal, with

$$\mathbf{Y}_0 | \{\mathbf{Y}_{ij}\}, \boldsymbol{\eta}, \boldsymbol{\Lambda} \sim MVN(\boldsymbol{\tau}, \mathbf{S}) \tag{2.13}$$

with posterior precision

$$\mathbf{S}^{-1} = \boldsymbol{\Gamma}^{-1} + \left(\left(\sum_{i=1}^{p} \mathbf{D}_i^{-1} \right)^{-1} + \boldsymbol{\Lambda} \right)^{-1} \tag{2.14}$$

and posterior expectation

$$\boldsymbol{\tau} = \mathbf{S} \left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} + \left(\left(\sum_{i=1}^{p} \mathbf{D}_i^{-1} \right)^{-1} + \boldsymbol{\Lambda} \right)^{-1} \left\{ \left(\sum_{i=1}^{p} \mathbf{D}_i^{-1} \right)^{-1} \sum_{i=1}^{p} \mathbf{D}_i^{-1} \overline{\mathbf{y}}_i - \boldsymbol{\eta} \right\} \right], \tag{2.15}$$

where p is the number of ensembles in the MME. The derivation of the posterior expectation and covariance is given in Appendix A.3. In the expressions above, the expectation $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Lambda}$ of the discrepancy are considered to be known quantities; their estimation, and the estimation of the other required quantities, will be discussed subsequently.

2.2.2.3 Interpretation of the elements of the posterior distribution of $\mathbf{Y}_0$

Expressions (2.14) and (2.15) can be rewritten to clarify the contribution of each source of information to the posterior mean and covariance.

From (2.11), the term $\left(\sum_{i=1}^{p}\mathbf{D}_i^{-1}\right)^{-1}$ is the inverse of the total precision of the sample estimates of the $\left\{\overline{\mathbf{Y}}_i\right\}$, conditional on $\boldsymbol{\xi}$: this covariance matrix captures the uncertainty about the true position of the unobserved MME consensus $\boldsymbol{\xi}$, given the observed ensembles. The posterior form may be simplified by writing this quantity as

$$\boldsymbol{\Sigma}_{\boldsymbol{D}} = \left(\sum_{i=1}^{p}\mathbf{D}_i^{-1}\right)^{-1}, \tag{2.16}$$

say. The first term inside the braces {} in (2.15) is the weighted average of the sample ensemble means $\{\overline{\mathbf{y}}_i\}$, where the weights are determined by the relative precisions $\boldsymbol{\Sigma}_{\boldsymbol{D}}\mathbf{D}_i^{-1}$ of the sample means. This is the sample estimate of the MME consensus,

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{i=1}^{p}\mathbf{D}_i^{-1}\overline{\mathbf{y}}_i. \tag{2.17}$$

(2.14) and (2.15) can therefore be written in the simpler forms

$$\mathbf{S}^{-1} = \boldsymbol{\Gamma}^{-1} + \left(\boldsymbol{\Sigma}_{\boldsymbol{D}} + \boldsymbol{\Lambda}\right)^{-1}, \tag{2.18}$$

$$\boldsymbol{\tau} = \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} + \left(\boldsymbol{\Sigma}_{\boldsymbol{D}} + \boldsymbol{\Lambda}\right)^{-1}\left\{\hat{\boldsymbol{\xi}} - \boldsymbol{\eta}\right\}\right]. \tag{2.19}$$

It is now clear that the posterior precision matrix $\mathbf{S}^{-1}$ is the sum of the prior precision $\boldsymbol{\Gamma}^{-1}$ and the precision $(\boldsymbol{\Sigma}_{\boldsymbol{D}}+\boldsymbol{\Lambda})^{-1}$ of the estimate of the bias-corrected sample consensus, $\hat{\boldsymbol{\xi}}-\boldsymbol{\eta}$, which is represented by the term in braces {} in (2.19).

Likewise, the posterior mean vector $\boldsymbol{\tau}$ is a weighted sum of terms representing the prior expectation $\boldsymbol{\alpha}$ and the vector $\hat{\boldsymbol{\xi}}-\boldsymbol{\eta}$ inferred from the ensemble forecasts adjusted by the expected bias. The weights given to these two compo-

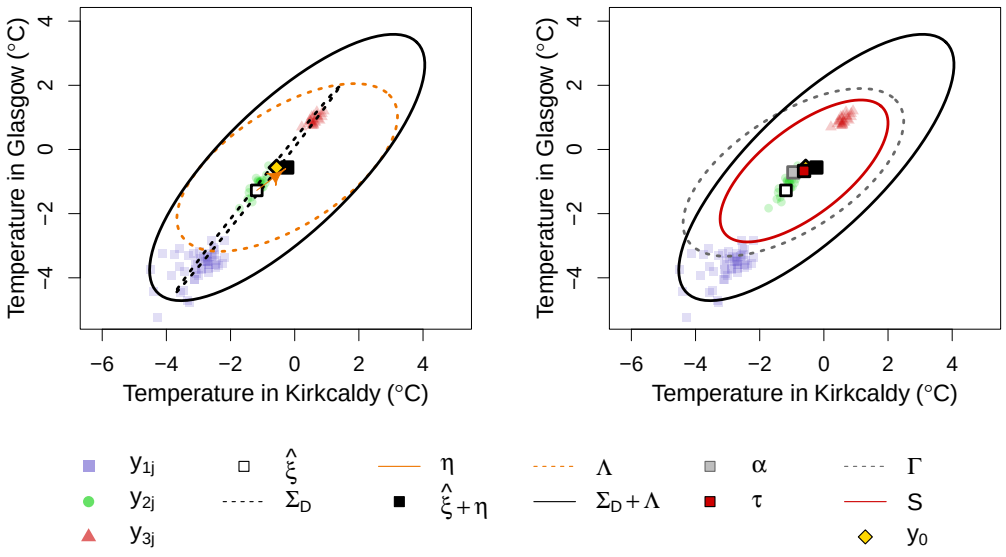
nents are determined by the covariance matrices of the prior distribution and the bias-corrected consensus, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma_D} + \boldsymbol{\Lambda}$ respectively.

The elements of the MME postprocessing framework are summarised in Table 2.2. Figure 2.9 shows the various components of the postprocessed MME forecast for the example shown in Figure 2.3, with the mean and covariance of the prior distribution estimated from a sample climatology using the observations from the week centred on the forecast verification date in the ten years prior to the forecast issue year, and $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ estimated using the sample mean and covariance matrix of the errors of the 2-day-ahead forecasts verified in the 25 days prior to this forecast being issued.

Figure 2.9: Elements of the posterior distribution of $\mathbf{Y}_0$ obtained from the MME forecast in Figure 2.3. Covariance matrices are represented by ellipses containing 95% of the respective distributions.

(a) The sample MME consensus $\hat{\boldsymbol{\xi}}$, adjusted by the bias correction $\boldsymbol{\eta}$ with uncertainty $\boldsymbol{\Lambda}$.

(b) The bias-corrected consensus $\hat{\boldsymbol{\xi}} - \boldsymbol{\eta}$ combined with the prior $\boldsymbol{\alpha}$ to obtain the posterior mean $\boldsymbol{\tau}$ and covariance matrix $\mathbf{S}$, using (2.18) and (2.19)



It is worth noting that even if the assumption of multivariate normality were not deemed reasonable, the same posterior form would be obtained by treating the problem as a form of Bayes linear analysis in which our prior expectations of the mean and variance of the temperature are adjusted by the forecasts and discrepancy: the posterior mean is then the optimum linear combination of the forecast information, and the posterior covariance matrix

Table 2.2: Summary of the elements of the proposed post-processing framework

$\mathbf{Y}_0$	Verifying observation: the quantity to be forecast
$\boldsymbol{\alpha}$	Prior expectation of $\mathbf{Y}_0$
$\boldsymbol{\Gamma}$	Prior covariance of $\mathbf{Y}_0$
$\mathbf{Y}_{ij}$	jth member of ith ensemble
n_i	Number of members in ith ensemble
$\boldsymbol{\mu}_i$	Expectation of ith ensemble
$\overline{\mathbf{Y}}_i$	Sample mean of ith ensemble
$\mathbf{C}_i$	Covariance matrix of members $\{\mathbf{Y}_{ij}\}$ of ith ensemble
$\mathbf{D}_i$	Covariance matrix of sample ensemble mean $\overline{\mathbf{Y}}_i$ given $\boldsymbol{\xi}$
$\hat{\boldsymbol{\xi}}$	Estimated MME forecast consensus
$\boldsymbol{\Sigma}$	Covariance matrix of ensemble means $\{\boldsymbol{\mu}_i\}$
$\boldsymbol{\Sigma_D}$	Uncertainty about MME consensus
$\boldsymbol{\Delta}$	Discrepancy between MME consensus $\boldsymbol{\xi}$ and verifying observation $\mathbf{Y}_0$
$\boldsymbol{\eta}$	Expected value of $\boldsymbol{\Delta}$
$\boldsymbol{\Lambda}$	Covariance of $\boldsymbol{\Delta}$
$\boldsymbol{\tau}$	Posterior expectation of $\mathbf{Y}_0$
$\mathbf{S}$	Posterior covariance of $\mathbf{Y}_0$

is a valid summary of the uncertainty in this optimum linear combination (Chandler, 2013).

2.2.2.4 Estimation of the required covariance matrices

When implementing the framework, each $\mathbf{C}_i$ may be estimated using the sample covariance matrix of the forecasts from the ith ensemble, and $\boldsymbol{\Sigma}$ using the sample covariance matrix of the ensemble means. This estimate of $\boldsymbol{\Sigma}$ will be singular when the number of ensembles p is smaller than the dimension m of the data, as occurs in the case study considered here and in later chapters. However, $\boldsymbol{\Sigma}$ appears in the posterior only as part of $\mathbf{D}_i = \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i$, which is invertible as long as $\mathbf{C}_i$ is: this is always the case when the ensemble size n_i used to estimate $\mathbf{C}_i$ is larger than the dimension m. This requirement is fulfilled by the MME considered here, but should be kept in mind when designing other applications of the method.

Strictly speaking, this estimate of $\boldsymbol{\Sigma}$ will be biased due to the use of the sample means $\overline{\mathbf{Y}}_i$ in place of the underlying means $\boldsymbol{\mu}_i$, tending to overestimate

the uncertainty by the average of the covariance matrices of the ensemble sample means, $\frac{1}{p}\sum_{i=1}^{p} n_i^{-1}\mathbf{C}_i$. If the ensembles are of moderate size then this bias will be very small unless the $\{\mathbf{C}_i\}$ are very large, and correcting it will have minimal impact on the postprocessed forecast variance. If the bias is not small, then the simplest approach to correcting it is to subtract it, to obtain the bias-corrected estimate

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \frac{1}{p}\sum_{i=1}^{p} n_i^{-1}\mathbf{C}_i. \tag{2.20}$$

However, $\widetilde{\boldsymbol{\Sigma}}$ is not guaranteed to be positive definite, and therefore may produce singular estimates of the $\{\mathbf{D}_i\}$. In an experiment not presented here, the Bayesian postprocessing reported in Section 4.2 was repeated with this simple bias correction; at all but the shortest leadtimes, an invalid posterior was obtained in around 4% of forecast instances (25-35 cases out of the 630 at each leadtime), and the forecast skill of the remaining cases was unchanged. For this reason, no bias correction of $\boldsymbol{\Sigma}$ is included in the results reported subsequently. An alternative approach would be to carry out the suggested bias correction and to use generalised inverses throughout the analysis to avoid the problem of non-invertibility; this approach has not been explored further due to time constraints.

The expectation and covariance matrix $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ of the discrepancy $\boldsymbol{\Delta}$ for a given forecast instance can be estimated directly using the mean and covariance matrix of an appropriate training set of past forecast-observation pairs. As with the estimate of $\boldsymbol{\Sigma}$, this approach will tend to overestimate the value of $\boldsymbol{\Lambda}$ because it is based on the sample forecast consensus $\hat{\boldsymbol{\xi}}$ rather than the population consensus $\boldsymbol{\xi}$; and, once again, a simple bias correction risks producing an estimate of $\boldsymbol{\Lambda}$ that is not positive definite, so no bias correction is included here. Approaches to selection of an appropriate training set from which to estimate $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ are considered in detail in Section 2.3. A more sophisticated method to estimate these quantities, using a linear approximation

to Bayesian inference, is introduced in Chapter 5.

It is also worth noting here the links between the Bayesian forecast postprocessing approach and the data assimilation (DA) process described in Section 2.1. The two problems are inverses of each other: while DA attempts to infer the model state from the observations, the postprocessing procedure infers the likely value of the observations from the model state. Both estimate a model discrepancy in order to do this; however, the DA process focuses on adjusting the initial conditions for the forecast, whereas postprocessing attempts to correct the subsequent trajectory of the forecasts, so no real-time observations are available from which to do so, and the error must be estimated from training data. If the method were in use operationally, it is possible that some information about the individual model errors and their associated covariances, and about the spread of each ensemble, could be obtained from the corresponding matrices used in the DA process; although the between-ensemble covariance matrix $\boldsymbol{\Sigma}$ would still need to be estimated for the MME as a whole.

2.2.3 Choosing a prior distribution for $\mathbf{Y}_0$

If no prior assumptions are to be made about the distribution of $\mathbf{Y}_0$, a non-informative prior can be used by setting $\boldsymbol{\Gamma}^{-1} = \mathbf{0}$. Such a prior will contribute nothing to the final posterior forecast covariance and expectation in (2.18) and (2.19), which will reduce to

$$\mathbf{S} = \boldsymbol{\Sigma}_{\boldsymbol{D}} + \boldsymbol{\Lambda}, \qquad \boldsymbol{\tau} = \hat{\boldsymbol{\xi}} - \boldsymbol{\eta}. \tag{2.21}$$

Two approaches to setting an informative prior are suggested below. However, an informative prior may be estimated from any subset of the available historical observations, with the proviso that the subset should be selected without reference to the forecast instance being postprocessed: the prior must not use any of the information used to construct the posterior.

2.2.3.1 Climatological prior

Perhaps the simplest informative prior is a climatological one, constructed from an archive of past observations for the time of year and time of day that the forecast will be verified (Vannitsem et al., 2018, §3.4). A typical approach to estimation of a climatological prior distribution for a forecast instance with verification calendar date d and year y, say, would be to select the verifying observations on days $d-3$ to $d+3$ and years $y-n$ to $y-1$. The sample mean and covariance matrix of this climatological sample are taken as estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ respectively. A climatological prior reflects a prior belief that the realised temperature will be close to the average for the time of year. However, this approach fails to account for the prevailing weather at the time the forecast is issued, or the conditions predicted by the ensembles themselves.

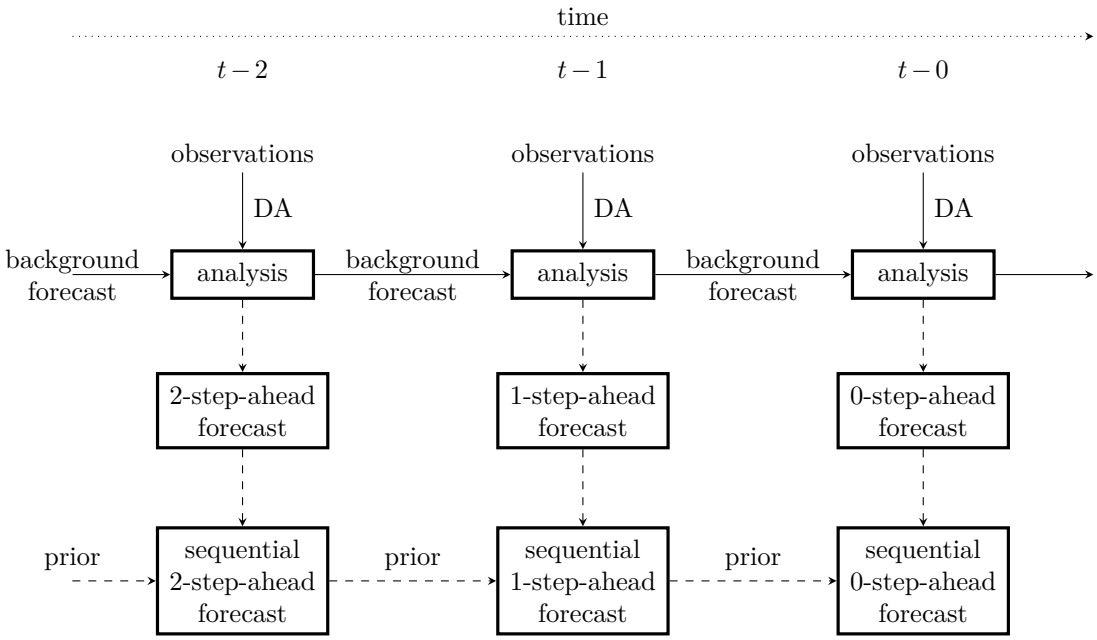
2.2.3.2 Sequential postprocessing

The Bayesian framework introduced here provides an opportunity for information from an earlier forecast of the weather at a particular time t to be used in the postprocessing of any new forecasts issued for the same verification time, in a process similar to the sequential data assimilation approach shown in Figure 2.1.

In the data set presented in Section 2.1.2 and used throughout this thesis, forecasts of the weather at time t are issued at one-day intervals, and are first issued 15 days in advance. For this initial forecast, no particular prior information is assumed, so these forecasts should be postprocessed using either a climatological or noninformative prior. The next day, a 14-day-ahead forecast is issued for the weather at time t; a natural choice for the prior distribution of the weather at time t is the 15-day-ahead posterior forecast distribution obtained one day previously. In this way, the forecasts for each verification date are postprocessed sequentially: for each new forecast issued for time t, the previous posterior forecast is used as the prior distribution, until the forecast verification date is reached. A schematic of the final stages of this

prior-to-posterior sequential approach is shown in Figure 2.10.

Figure 2.10: Schematic of the sequential forecasting process for forecasts of the weather at time t, and its relationship with the data assimilation (DA) process.



It is important to note that, for ensembles initialised using data assimilation as described in Section 2.1, the assimilation procedure can be regarded as partially updating the forecasts in the same way, at each time step using the posterior distribution of the previous background forecast (the analysis) as the prior for the analysis used to initialise the new background forecasts. This potentially complicates the postprocessing analysis; the effect of this will be explored in Section 4.1, when examining the postprocessing results obtained using different prior specifications.

2.3 Selection of a training set for statistical postprocessing

Statistical postprocessing methods like Model Output Statistics (Section 2.1.4.1) and the Bayesian framework introduced in Section 2.2 require a set of training cases consisting of past forecast-verification pairs, from which the necessary quantities can be estimated for each forecast instance. The more similar the

errors of the training case are to those of the current forecast instance, the better the estimate of the necessary adjustment will be, and the more skilful the postprocessed forecast will be. Forecast errors will depend to some extent on the time of day (also known as synoptic time), the time of year, or the state of the atmosphere at the time that the forecast is issued (Eckel and Mass, 2005; Greybush et al., 2008; Ferranti et al., 2015), so it is advisable to select training cases from forecasts verified at the same synoptic time and time of year as the forecast of interest.

2.3.1 Climatological training cases

As when setting a prior distribution for $\mathbf{Y}_0$, one possible approach to selecting training cases that are likely to have error characteristics similar to those of the current forecast instance is to choose previous forecasts that were verified at the same time of year – perhaps from the same day of the year, or a small window on either side (Hagedorn et al., 2008; Hamill, 2012) – to estimate the climatological forecast error. This may – depending on the length of the available archive of historical forecasts – offer a large set of training cases to choose from; however, climatology-based correction does not take into account the state of the atmosphere at the time that the current forecast is issued. Another issue is that operational weather forecasting models are updated fairly regularly: if this changes the typical biases of the models, then simply using historical forecast-observation pairs to estimate the correction risks introducing a new source of error.

2.3.2 'Moving window' training cases

An alternative is to use forecasts verified in the days prior to the forecast issue date to estimate the required correction (Gneiting et al., 2005). This 'moving window' approach limits the number of available training cases, but has the advantage of ensuring that, except for a handful of instances when a model change is implemented, the current forecast instance will have been generated by the same NWP model that was used to produce the forecasts in the training

set: however, it should also be noted that recent observation errors have already been incorporated into the individual forecasts, through the data assimilation process used to obtain the analysis with which the forecast is initialised (Section 2.1). Postprocessing using a 'moving window' training set will be successful if the forecast error characteristics persist until the current forecast is verified; however, as the forecast leadtime increases, forecasts within a moving window will become less relevant to the forecast discrepancy eventually observed, and choosing a long moving window risks including training cases that are not informative about the forecast instance of interest. Furthermore, the elements of a moving-window training set are autocorrelated: the information content of such a training set is therefore smaller than if the members of the training set were independent.

More pertinent information may be obtained by selecting a training set of forecasts that predict similar weather to that anticipated by the current forecast. The next section discusses the concept of selecting a training set based on some measure of similarity to the current forecast instance, and introduces the second major contribution of this thesis: a method for selecting a training set based on the similarity of the prevailing weather regime to that of the current forecast instance.

2.3.3 Analogues to the forecast of interest

Forecasts selected on the basis of some measure of similarity to the current forecast instance are known as analogues to the current forecast. The idea of postprocessing using analogues is by no means new (Toth, 1989; Van den Dool, 1989; Sievers et al., 2000), but the approach to identifying analogues presented here is new.

When using analogues as training cases, the size of the training set is no longer limited by temporal proximity to the forecast of interest, so potentially much larger training sets can be constructed. However, while surface temperatures and other variables such as surface wind speeds may be successfully postprocessed using relatively small training sets of 20-30 members (Hagedorn

et al., 2008; Delle Monache et al., 2013; Junk et al., 2015), postprocessing of precipitation requires a very long archive of forecasts in order to provide enough candidate examples of extreme events (Hamill and Whitaker, 2006), and the use of too short an archive can lead to selection of analogues of poor quality (Hu et al., 2020).

Using well-chosen analogues as training cases should ensure that the corresponding weather states are relevant to the forecast of interest – more so than by simply filtering on season and synoptic time to sample the climatology, as suggested in Section 2.3.1 – and so should produce a better estimate of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$.

2.3.3.1 Identifying analogues to the current forecast instance

Analogues to the current forecast are identified by first selecting a vector of summary statistics to characterise each forecast instance (perhaps the forecast values themselves, or some more compact summary of them, such as the ensemble means); and computing this vector of summary statistics for both the forecast instance of interest and for each potential analogue. Denoting the values for the forecast and a potential candidate by $\mathbf{F}$ and $\mathbf{C}$ respectively, analogues are selected on the basis of distance metric, typically

$$\|\mathbf{F}, \mathbf{C}\| = \sqrt{\sum_{i=1}^{m} \left(\frac{F(i) - C(i)}{\sigma_i} \right)^2} \tag{2.22}$$

where $F(i)$ and $C(i)$ are the ith elements of $\mathbf{F}$ and $\mathbf{C}$ respectively, each normalised by its climatological standard deviation σ_i (Delle Monache et al., 2013), computed over all potential candidates $\{\mathbf{C}\}$. Those candidates with the smallest values of $\|\mathbf{F}, \mathbf{C}\|$ are selected as analogues to $\mathbf{F}$, to be used as training case in postprocessing. Typically a fixed number of the closest analogues is selected (Delle Monache et al., 2013; Junk et al., 2015).

2.3.3.2 Analogue selection by weather regime

When the dimension m is not large, the vectors $\mathbf{F}$ and $\{\mathbf{C}\}$ typically contain elements corresponding to all of the forecast quantities of interest. However, when m is large, particularly if the forecasts have a large spatial domain, finding analogues that are close in all variables simultaneously may be difficult. As a result, the quality (or relevance) of the selected analogues is likely to fall as m increases, since the aggregated distance metric (2.22) cannot discriminate between (for example) candidates with a high proportion of moderate outliers, and candidates with only a single large outlier. In this situation, and in the context of forecasting simultaneously at multiple spatial locations, Hamill and Whitaker (2006) suggests selecting local analogues for each location, based on the corresponding subsets of $\mathbf{F}$ and $\{\mathbf{C}\}$. However, this approach requires a further postprocessing step to combine the independently postprocessed local forecasts into a single spatially continuous forecast. An alternative approach is therefore proposed here: to reduce the dimension of the vectors $\mathbf{F}$ and $\{\mathbf{C}\}$ for all locations and variables simultaneously, by defining them as low-dimensional summaries of the forecast pressure fields.

Dimension reduction techniques such as principal component analysis (PCA, often also known as Empirical Orthogonal Function (EOF) analysis in the climate and meteorological literature) have long been applied to pressure fields in order to characterise prevailing weather conditions (Jenkinson and Collison, 1977; Jones et al., 1993), and to obtain indices of large-scale synoptic structure (Wilks, 2011). Patterns of variation in mean sea level pressure fields, obtained using PCA, are used to classify forecasts into climatological weather regimes in the Met Office's operational weather forecasting (Neal et al., 2016; Richardson et al., 2020). Previous studies have found that incorporating regime-dependent errors derived from pressure fields can result in more skilful forecasts (Greybush et al., 2008; Allen et al., 2020). Pressure fields are at the core of the NWP models, and as such are almost always provided in the model output; furthermore, pressure fields are physical quantities that are directly

modelled in NWP models (unlike parametrised quantities like precipitation and surface temperatures, which suffer from greater model biases). As such, they are generally well predicted, and so form a robust basis in which to identify analogues.

2.3.3.2.1 Finding the modes of spatial variability

The first step in the proposed scheme is to carry out a spatial principal component analysis on a long archive of historical data to identify the principal modes of spatial variation in the pressure fields. Here, the entire archive of ERA-Interim mean sea level pressure (MSLP) reanalyses from 1979 to 2018 was used. Because the focus of the applications in this thesis is on winter temperatures observed at midnight, only reanalyses issued for midnight during the meteorological winter period (December, January and February) are included, giving a total of 3420 time points.

The efficiency of the PCA dimension reduction is such that the spatial domain need not be restricted to the forecast area of interest; in fact, since synoptic weather systems in the surrounding area are known to affect the weather over the UK (Scaife et al., 2014), MSLP fields for processing forecasts covering the UK and northern Europe should cover the entire North Atlantic European region and central Europe (35° to 70°N, 30°W to 20°E; shown in Figure 2.11) (Neal et al., 2016). Other studies have recommended this domain as being appropriate for purposes such as reconstructing surface temperatures from MSLP-derived regime classifications (Beck et al., 2016). Each MSLP field therefore contains 1836 values, arranged on a regular 1° latitude-longitude grid.

Following standard practice, the MSLP fields are first reweighted to account for differences in the areas covered by grid cells at each latitude θ, by multiplying the cell values by $\cos(\theta\pi/180)$ (North et al., 1982; Wilks, 2011). The climatological mean field is found by averaging all of the latitude-weighted MSLP fields over time; this mean field is then subtracted from the daily fields to obtain a daily MSLP anomaly $\mathbf{a}$. Spatial PCA is carried out on these daily

anomaly fields; the resulting eigenvectors (EOFs) represent the dominant modes of spatial variation, with the corresponding normalised eigenvalues indicating the proportion of the data's total variance explained by each eigenvector.

As suggested by Jolliffe (2011), only the first q eigenvectors are retained, where q is the smallest number of eigenvectors required to capture at least 90% of the variance in the anomalies. In the dataset used here, the first six eigenvectors are retained; plots of the spatial patterns represented by these eigenvectors are shown in Figure 2.11. The retained modes of variation have fairly straightforward interpretations. The first is associated with the North Atlantic Oscillation, a large-scale pressure system known to be one of the key drivers of variability in the weather over the UK, particularly in winter (Scaife et al., 2014); the second captures situations in which a high- or low-pressure system is located to the south-west of the UK; the third reflects east-west pressure gradients over the UK, and so on. Higher-numbered modes display patterns of increasing complexity. The resulting $(L \times q)$ matrix of principal eigenvectors is denoted $\mathbf{E}$; the climatological modes that it represents need only be computed once from the reanalysis data, and updated whenever the climatological archive is updated.

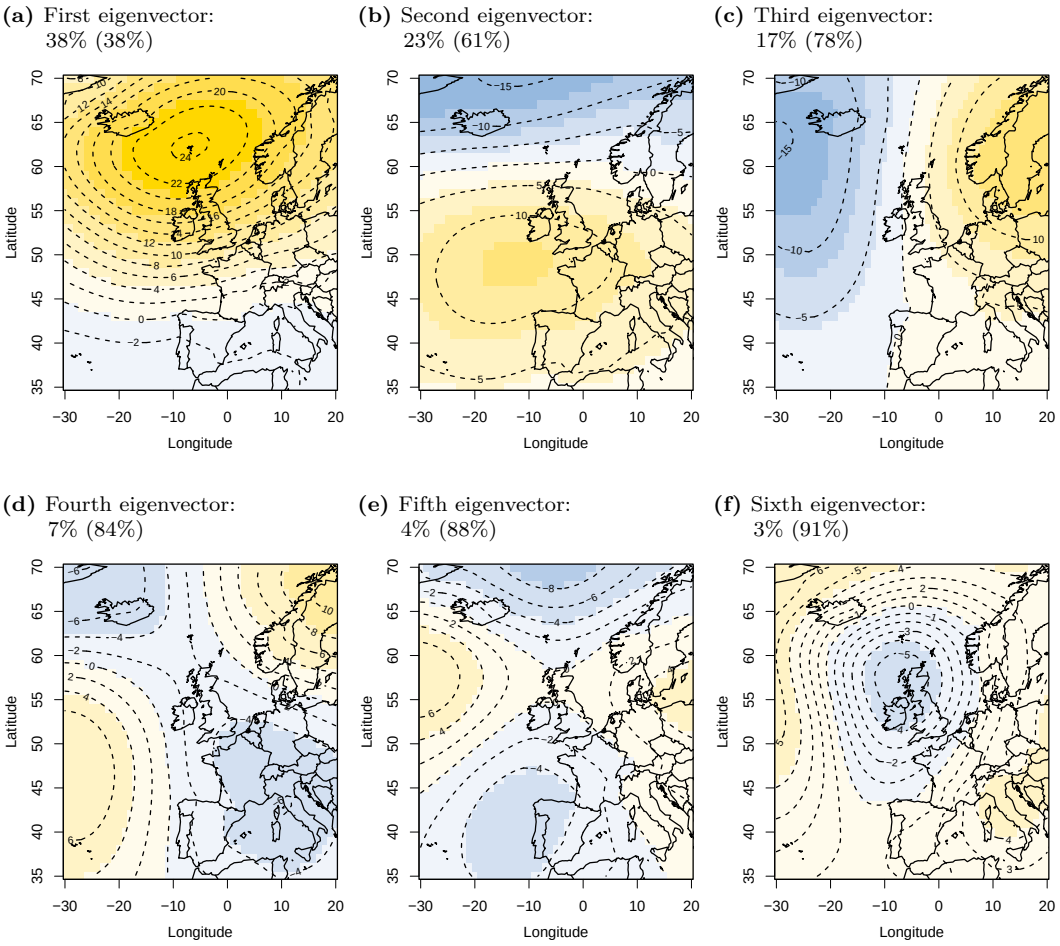
2.3.3.2.2 Selecting analogues

Vectors of principal component scores are now computed for each forecast instance by projecting the latitude-adjusted MSLP anomalies $\mathbf{a}_f$ of the ensemble mean MSLP field onto the eigenvector matrix $\mathbf{E}$, to obtain a q-length vector $\mathbf{u}_f$, defining the coordinates of the forecast in the basis defined by $\mathbf{E}$:

$$\mathbf{u}_f = \mathbf{E}' \mathbf{a}_f. \tag{2.23}$$

When a new forecast instance requires postprocessing, it is only necessary to obtain its anomaly field $\mathbf{a}_f$ and apply (2.23) to obtain the vector of principal component scores. Analogues to the instance of interest are selected in the

Figure 2.11: Spatial plots of the elements of the first six eigenvectors of the ERA-Interim winter archive of MSLP fields, with the percentage of variance explained by each eigenvector. Cumulative percentages of variance explained are given in parentheses.



q-dimensional principal component case using the distance metric

$$\left\|\mathbf{u}_f, \mathbf{u}_c\right\| = \sqrt{\sum_{i=1}^{q} \Big(\mathbf{u}_f(i) - \mathbf{u}_c(i)\Big)^2}, \tag{2.24}$$

where $\mathbf{u}_f$ and $\mathbf{u}_c$ are the vectors of scores for the forecast and candidate, respectively. Because they are not normalised to have unit variance, the elements of $\mathbf{u}_f$ and $\mathbf{u}_c$ will have magnitudes proportional to the amount of variance explained by the corresponding eigenvectors, and will contribute proportionally more to the total sum of squares, ensuring that the distance $\left\|\mathbf{u}_f, \mathbf{u}_c\right\|$ prioritises forecasts that are most similar in terms of the most important spatial patterns

of variation. After postprocessing, the principal component scores $\mathbf{u}_f$ can be added to the archive of potential candidate scores, to be searched when postprocessing the next forecast instance.

Where the forecast to be postprocessed is produced by an MME, principal component scores can be obtained separately for each ensemble by projecting the p ensemble mean fields onto $\mathbf{E}$; the joint state of the p-ensemble forecast is therefore represented by a vector of length $q \times p$, where q and p are both small, and analogues are selected on the basis of the Euclidean distance (2.24) calculated over this state space.

Unlike analogues identified in forecast variable space, the weather-regime candidate archive need not be recalculated if the forecast domain changes, either in terms of spatial extent or of the variables included: the analogues chosen will remain the same for any choice of surface weather variables, and for any locations for which the synoptic domain remains appropriate. This means, for example, that subsets of forecasts in north-western Europe could be postprocessed independently using the same weather regime analogues, and the resulting forecasts would be mutually consistent and coherent.

Training cases selected using the weather regime analogue method proposed in this section are used to postprocess forecasts of surface temperature in Section 4.3.

2.4 Summary and discussion

This chapter presents two of the major contributions of this thesis: in Section 2.2, a new framework for the postprocessing of multi-model ensemble forecasts; and in Section 2.3, a novel approach to the selection of training cases for use in statistical postprocessing of weather forecasts. An application of both methods to postprocessing the medium-range forecasts of surface air temperatures described in Section 2.1.2 is presented in the next chapter.

The MME postprocessing framework is derived from a graphical representation of the relationships between the quantity of interest – in the example

used here, the vector of temperatures $\mathbf{Y}_0$ – and the collections of individual forecasts issued by several EPSs. Unlike competing postprocessing methods, the Bayesian framework is able to accommodate information about the prior distribution of $\mathbf{Y}_0$, offering the possibility of producing sequentially postprocessed forecasts.

Although the framework was motivated by situations where the shared discrepancy is due to the approximations inherent in all NWP models, it could in principle be used to postprocess forecast ensembles where other potential sources of shared discrepancies are present, for example as a means of adding detail to low-resolution forecasts, with discrepancies estimated from a training set matching past low-resolution forecasts to higher-resolution verifying observations. In this case, the low resolution can itself be regarded as a source of shared discrepancies in the forecast ensembles due to unresolved processes, so that the same conceptual framework applies.

One further limitation of the framework in its present form is its scalability: as described here, the approach can only be applied to relatively small spatial domains, due to the difficulty of estimating the required covariance matrices from the relatively small forecast ensembles available. When the dimension m is greater than the number of members of any of the individual ensembles, the estimate of the corresponding covariance matrix $\mathbf{C}_i$ will be singular, and the posterior cannot be evaluated. In this setting, structured covariance models would be needed for the estimation of the required covariance matrices. Likewise, as discussed in Section 2.2.2.4, the between-ensemble covariance matrix $\mathbf{\Sigma}$ is estimated from only p points for each forecast instance, and so may be estimated imprecisely. In principle, this parameter uncertainty could itself be incorporated into the posterior distribution, although this is non-trivial and the computational complexity would increase dramatically. However, in the application presented in the next chapter, the contribution from $\mathbf{\Sigma}$ was found to be very small compared to both $\mathbf{\Sigma_D}$ and $\mathbf{\Lambda}$, so this source of uncertainty is not considered further here.

The second contribution introduced in this chapter is the proposed method for selecting analogues to the forecast of interest on the basis of their similarity in terms of the dominant patterns in the forecast mean sea level pressure (MSLP) fields. The proposed method allows analogues to be identified in a relatively low-dimensional space representing key aspects of the shape of the MSLP fields over a region known to affect the prevailing weather patterns over the UK. Analogues with similar patterns of MSLP fields may be expected to have similar prevailing weather regimes, and so to have similar forecast errors, to the forecast of interest; estimation of the discrepancy $\boldsymbol{\Delta}$ from such a training set may therefore be expected to produce a better estimate, and therefore more skilful postprocessed forecasts, than estimation using a moving-window or climatological approach. Choosing analogues on the basis of weather patterns over a large spatial domain has the added advantage that the training cases identified will be relevant for forecasts of all surface weather quantities throughout the domain. This is in contrast to the 'direct analogue' method described in Section 2.3.3.1, in which analogues are chosen according to their similarity to the current instance in terms of the variables to be postprocessed. Such an approach may require postprocessing of forecasts at each location using local analogues, in order to keep the dimension of the search space small enough to identify relevant instances.

One caveat when selecting analogues in any domain is that, ideally, a very large archive of forecasts from the same model configuration are required. For this reason, many operational centres now issue reforecasts (forecasts using the latest model initialised with historical observations) whenever the the NWP models undergo a major update, in order to provide just such an archive of relevant training cases for use in postprocessing. Constructing an archive of reforecasts for a multi-model ensemble, where models are updated on different dates, is particularly challenging, so the method may be more appropriate for use in the postprocessing of single-ensemble forecasts, where such a lengthy archive is more readily available.

Chapter 3

Review of forecast verification methods

Forecast verification is the process of evaluating how well weather forecasts are able to predict the quantity of interest. This chapter reviews the verification metrics that will be used to evaluate the forecast skill of probabilistic forecasts issued in the form of probability density functions, like those produced by the Bayesian framework and other postprocessing methods described in Chapter 2. Equivalent metrics for assessing ensemble forecasts are also presented; using these the skill of forecasts issued as an ensemble of point forecasts – such as the raw output from an ensemble prediction system or the forecasts in the simulation study in Section 6.3 – can be assessed and compared directly to that of the probabilistic forecasts.

Section 3.1 defines measures of forecast accuracy, bias and sharpness. More general scoring rules, which provide a scalar measure of the overall quality of probabilistic and ensemble forecasts, are considered in Section 3.2. Finally, in Section 3.3, tools to diagnose particular types of forecast error in the marginal and joint forecasts are discussed.

Two innovations are described in this chapter. First, in Section 3.3.2.2, the extension of the band depth rank histogram to more efficiently evaluate the joint calibration of forecasts issued in the form of a predictive density; and second, in Section 3.3.2.3, a graphical tool to facilitate the joint evaluation of

two types of calibration histogram.

3.1 Measures of bias, accuracy and sharpness

Forecast quality is typically summarised using one or more scalar verification measures. Among the simplest of these are measures describing particular attributes of forecast quality (Wilks, 2011; Jolliffe and Stephenson, 2012). Forecast bias and accuracy are both characteristics of a point forecast such as the mean of a forecast distribution, or the value obtained from a deterministic forecast system. The forecast bias is a measure of the average difference between a point forecast and the corresponding observation; for a collection of forecast distributions at times $t = 1, \ldots, n$ with means $\{\mu_t\}$ and verifying observations $\{y_t\}$, this is the mean error (ME),

$$ME = \frac{1}{n}\sum_{t=1}^{n}(\mu_t - y_t) = \overline{\mu} - \bar{y}. \tag{3.1}$$

Clearly, the smaller the bias, the better. Forecast bias is distinct from forecast accuracy, which quantifies the degree of correspondence between individual forecasts and their verifying observations. Accuracy is typically measured using either the mean absolute error (MAE) or the root mean squared error (RMSE).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\mu_t - y_t\right|, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\mu_t - y_t)^2}. \tag{3.2}$$

Both of these metrics are expressed in the same units as y, and can be interpreted as the typical magnitude of the forecast error; both will be zero if the forecasts are perfect, but the RMSE is more sensitive to large errors than the MAE. In evaluating the applications in Chapters 4 and 6, the MAE will be used to quantify the accuracy of the forecasts. The MAE is frequently used for the verification of operational forecasts (Wilks, 2011) and is typically presented alongside the Continuous Ranked Probability Score (Section 3.2.1), to which it generalises when forecasts are issued as predictive distributions, rather than point forecasts.

Forecast bias and accuracy are properties of point forecasts with respect to their verifying observations: for probabilistic forecasts, another useful quantity is the variance of the predictive distribution, which can be regarded as a measure of confidence in the forecast. In the context of forecast verification, the predictive variance is characterised in terms of the forecast sharpness; forecast distributions with a low variance are considered to be sharp, while forecast distributions with high variance lack sharpness. Marginal forecast sharpness is typically expressed via the forecast standard deviation, while the overall sharpness of a d-variate forecast with covariance matrix $\mathbf{\Sigma}$ is quantified using the determinant sharpness (Gneiting et al., 2008),

$$DS = |\mathbf{\Sigma}|^{1/2d}, \tag{3.3}$$

Sharpness is a property of the forecasts alone, measured without reference to the verifying observations; sharpness is a desirable quality in a forecast only to extent that the confidence of the forecast reflects the true uncertainty about its prediction. A forecast distribution that correctly characterises the uncertainty about the observed outcome is said to be well calibrated, or reliable: for a well-calibrated probabilistic forecast of a continuous variable, the verifying observation is indistinguishable from a random draw from the predictive distribution. Given two sets of well-calibrated forecasts, the sharper set should be preferred (Gneiting et al., 2008).

Unlike measures of accuracy and sharpness, calibration is a property that cannot be meaningfully evaluated for a single forecast instance, because only a single verifying observation is available for each forecast. Section 3.2 defines some more general scoring rules that measure the quality of probabilistic forecasts, taking into account both accuracy and sharpness; methods for diagnosing particular aspects of miscalibration of a collection of forecasts are considered in Section 3.3.

3.2 Measures of overall forecasting skill: proper scoring rules

A scoring rule can broadly be defined as any function of the forecast distribution and the observed weather quantity that provides a scalar measure of the overall quality – or skill – of the forecast. Typically, smaller values of the score indicate more skilful forecasts, and many common scores take non-negative values, so that a 'perfect' forecast achieves a score of zero. Scoring rules are sometimes also referred to as omnibus scoring rules, meaning that they measure the quality of several aspects of the forecast simultaneously (Gneiting et al., 2008).

A scoring rule is considered to be 'proper' if a forecast probability distribution has an optimal expected score when the verifying observation is, in fact, drawn from that probability distribution, and is considered to be 'strictly proper' if no other forecast distribution achieves this optimal expected score (Bröcker and Smith, 2007; Gneiting and Raftery, 2007). Propriety in a scoring rule is needed to ensure that honest forecasts are issued: a proper score rewards forecasters who issue predictions reflecting their true beliefs, while the use of an improper scoring rule may motivate the forecaster to issue a different forecast in an attempt to 'game' the system and obtain a better score (Murphy and Winkler, 1987). Some strictly proper scoring rules taking into account multiple aspects of forecast skill for probabilistic forecasts are now defined.

3.2.1 Continuous ranked probability score

The continuous ranked probability score (CRPS) generalises the MAE for probabilistic forecasts (Wilks, 2011). It is a measure of the difference between the predictive and empirical cumulative distribution functions (CDFs), expressed in the same units as the variable of interest (Hersbach, 2000). In the context of weather forecast verification, where there is only a single observation y, the

empirical CDF is $\mathbb{1}_{\{x \geq y\}}$ and, for a predictive distribution P with CDF $F(x)$,

$$CRPS(P, y) = \int_{-\infty}^{\infty} \left[F(x) - \mathbb{1}_{\{x \geq y\}} \right]^2 dx. \tag{3.4}$$

Equation (3.4) may be difficult to evaluate for forecasts of arbitrary form; however, for forecasts issued in the form of a normal distribution with mean μ and variance σ^2, the CRPS can be evaluated exactly as

$$CRPS(\mu, \sigma^2, y) = \sigma \left\{ \frac{y-\mu}{\sigma} \left[2\Phi\left(\frac{y-\mu}{\sigma} \right) - 1 \right] + 2\phi\left(\frac{y-\mu}{\sigma} \right) - \frac{1}{\sqrt{\pi}} \right\}, \tag{3.5}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are, respectively, the CDF and PDF of the standard normal distribution (Gneiting et al., 2005).

The CRPS can also be evaluated for an M-member ensemble forecast by replacing the parametric forecast CDF $F(x)$ in (3.4) with the empirical ensemble CDF defined by the piecewise constant function

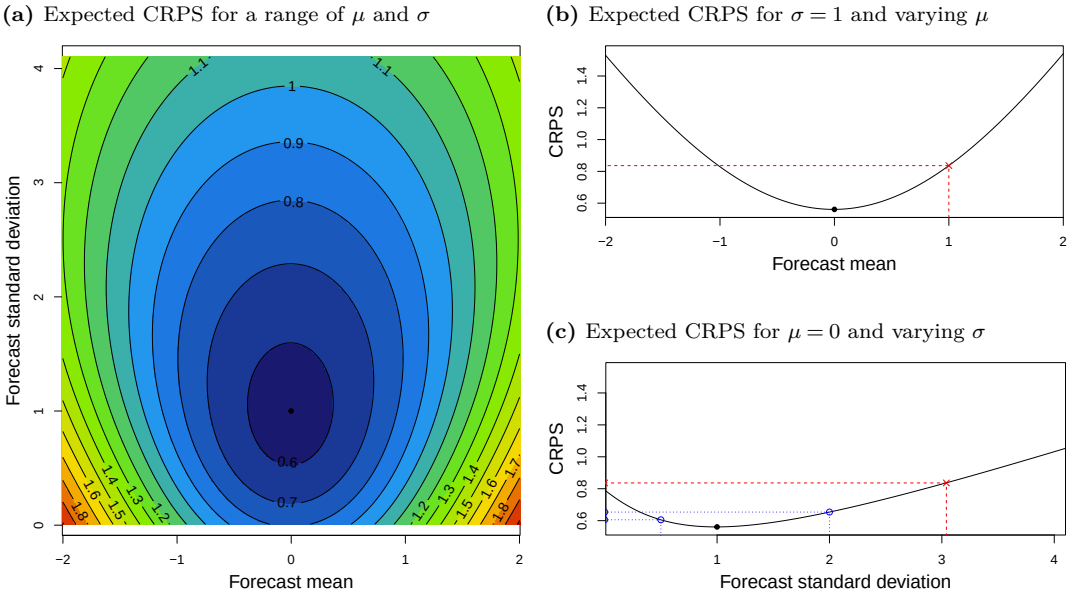
$$F_e(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{x \geq x_i\}}, \tag{3.6}$$

where x_i is the deterministic forecast issued by the ith ensemble member (Hersbach, 2000).

The CRPS is one of the most commonly used scoring rules in weather forecast verification, and is often cited as a measure of calibration, because it takes into account the whole predictive distribution. However, as Peirolo (2011) points out, the CRPS is known to strongly favour forecasts with a concentration of probability around the step function that jumps from 0 to 1 at the observation y: this characteristic is known as sensitivity to distance (Matheson and Winkler, 1976, §2). As a result, the CRPS tends to penalize lack of accuracy of the mean forecast vector more heavily than poor probabilistic calibration. The effect of this 'preference' can be seen in Figure 3.1, which shows the expected values of the CRPS when the observation is drawn from a standard normal distribution and the forecast distribution is normal, with mean

μ and standard deviation σ. The contours denoting pairs of forecast parameters with the same expected CRPS are elliptical, having a much shallower gradient as σ increases than as μ moves away from the correct value (3.1a): a forecast in which the standard deviation is correct, but the mean is misspecified by one unit, receives the same CRPS as a forecast in which the mean is correct, but the standard deviation is too high by a factor of three (Figures 3.1b and 3.1c). The expected CRPS is also not symmetric in σ, with underconfident forecasts being penalised more harshly than overconfident ones: a forecast with a predictive standard deviation that is double the correct value has a higher expected CRPS than a forecast with a predictive standard deviation that is half of the correct value (Figure 3.1c). Furthermore, Bröcker (2012) shows that ensemble forecasts calibrated by minimising the CRPS are not expected to produce the flat rank histograms usually expected from a well calibrated forecast (as discussed in Section 3.1), and Wilks (2018) demonstrates that the CRPS may reward sharper forecasts in preference to well calibrated ones; when comparing two imperfect forecasts, any claim that a lower CRPS score indicates a forecast with better calibration should therefore be treated with caution.

Figure 3.1: Expected value of the CRPS when forecasting the value of a standard normal random variable using a $N(\mu, \sigma^2)$ forecast distribution, estimated over 10000 synthetic 'observations' drawn from a standard normal distribution. The black dot indicates the parameters of the observation distribution.
In panels (b) and (c), the red lines indicate errors in the mean and standard deviation that receive the same CRPS. In panel (c), the blue lines indicate the CRPS for forecasts with a standard deviation of half and double the correct value.



3.2.1.1 The energy score

The CRPS is a univariate scoring rule; the multivariate generalisation of the CRPS is the energy score, defined as

$$ES(P, \mathbf{y}) = \mathbb{E}_P \|\mathbf{X} - \mathbf{y}\| - \frac{1}{2}\mathbb{E}_P \|\mathbf{X} - \mathbf{X}'\|, \tag{3.7}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbf{y}$ is the vector of verifying observations, P is the forecast distribution, and $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors with distribution P. Gneiting and Raftery (2007) showed that for univariate forecasts, the representation in (3.7) is equivalent to (3.4).

No closed form is available for the energy score, so following Gneiting et al. (2008), the energy score for a single forecast instance with realising observation $\mathbf{y}$ is evaluated over a random sample $\{\mathbf{X}_i : i = 1, \ldots, 10000\}$ of size $k = 10000$

from the multivariate forecast density P, using the Monte Carlo approximation

$$ES(P,\mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k}\|\mathbf{X}_i - \mathbf{y}\| - \frac{1}{2k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{X}_i - \mathbf{X}_j\|. \tag{3.8}$$

The energy score is often presented alongside the CRPS in forecast verification studies as a measure of the overall calibration of a multivariate forecasting method. However, Pinson and Tastu (2013) showed that, analogously to the CRPS, the energy score has extremely limited sensitivity to the covariance structure of the forecasts, and is strongly dominated by the forecast mean vector, particularly in higher dimensions. For this reason, the energy score should be viewed primarily as a measure of forecast accuracy, and more sensitive methods should be used to diagnose forecast calibration.

3.2.2 The logarithmic score

The CRPS and energy score may be best interpreted as a measure of whether a forecast is 'close enough' to the verifying observation (Peirolo, 2011), penalising errors in location more heavily than errors in spread. For a member of the public wanting to check whether to plan a picnic at the weekend, a forecasting method that consistently achieves a low CRPS is likely to be useful. Where forecasts are to be used in more precisely defined decision making, the opposite may be true: in applications that are sensitive to temperatures only beyond some critical threshold, a forecast that consistently places too little or too much probability on this threshold being reached is of little use, regardless of how close the mean forecast is to the truth.

One way of resolving this issue is by considering scoring rules that explicitly consider the probabilities assigned to the verifying observations. Any scoring rule that only takes into account the probability assigned to the verifying observation – that is, any scoring rule under which equal scores are given to forecasts that assign the same probability of occurrence to the verifying observation y – is known as a local scoring rule. Bernardo (1979) showed that any scoring rule for continuous variables that is smooth, proper, and local is

an affine function of the logarithmic or log score,

$$\log S(P, y) = -\log f(y) \tag{3.9}$$

where $f(y)$ is the predictive probability density function evaluated at the verifying observation. The unit of the log score depends on the base of the logarithm used and may be interpreted either in information-theoretic terms (Roulston and Smith, 2002; Peirolo, 2011) or in terms of betting returns (Hagedorn and Smith, 2009). The log score rewards forecasts that place a lot of probability density at the realised value, and heavily penalises forecasts for which the verifying observation falls in regions of low probability. For two forecasts $p(x)$ and $q(x)$, the difference in log score has a direct interpretation, indicating that the forecast p places $e^{\log p(y) - \log q(y)}$ times as much density at the observation y as the forecast q.

When P takes the form of a normal distribution with mean μ and variance σ^2, the logarithmic score is

$$\log S(\mu, \sigma^2, y) = \frac{\ln\left(2\pi\sigma^2\right)}{2} + \frac{(y-\mu)^2}{2\sigma^2}, \tag{3.10}$$

where the first term rewards sharpness in the predictive distribution, and the second term is proportional to the square of the standardised error of the deterministic forecast μ. The log score is not defined for raw ensemble forecasts, although it may be computed for a parametric distribution fitted to the ensemble (Siegert et al., 2019; Krüger et al., 2020).

When the predictive distribution P is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, (3.9) is a linear transformation of the Dawid-Sebastiani score (DSS; Dawid and Sebastiani, 1999):

$$DSS(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{y}) = \ln|\boldsymbol{\Sigma}| + (\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}). \tag{3.11}$$

Again, the first term rewards forecast sharpness, while the second is the squared

Mahalanobis distance of the verifying observation $\mathbf{y}$ with respect to the forecast distribution. The DSS can be used to measure the quality of any multivariate forecast distribution with finite first and second moments, for which it is a proper scoring rule; it is strictly proper for forecast distributions that, like the multivariate normal distribution, are fully characterised by their first and second moments (Gneiting and Raftery, 2007). Like the log score, the DSS may be computed for an ensemble forecast by estimating the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ from the ensemble forecast. However, unless the ensemble is substantially larger than the dimension of the data, sampling errors typically render this score unstable (Feldmann et al., 2015; Scheuerer and Hamill, 2015b); for this reason, the DSS is not reported in Chapters 4 and 6.

3.3 Diagnosing particular issues in forecast calibration

Scoring rules like those described in Section 3.2 give a single measure of the overall skill of a forecasting method, which can be used to identify which of several competing methods produces predictive distributions most like the generating distribution of the observations. However, it is also of interest to understand the particular strengths and weaknesses of each method, and scoring rules cannot provide much information about the nature of the forecast errors from which differences in skill arise. To understand the error characteristics of a given forecasting method, specific diagnostic tools are required.

3.3.1 Marginal calibration

As mentioned in Section 3.1, a forecasting system is considered to be well calibrated if the verifying observations are indistinguishable from random draws from the corresponding forecast distributions. To establish this, it is helpful to start by transforming the verifying observations in such a way that the transformed values should all be drawn from the same distribution. Specifically, suppose that a probabilistic forecast is issued with continuous CDF F_t. If

the observation y_t is indeed drawn from a distribution with CDF F_t, then the Probability Integral Transform (PIT)

$$PIT_t = F_t(y_t) \tag{3.12}$$

is distributed as $U(0,1)$ (Dawid, 1984; Diebold et al., 1998).

The PIT can be thought of as mapping each observation from its position within the corresponding forecast distribution onto the interval (0,1); the calibration of the collection of forecasts as a whole can thus be evaluated by considering any deviations of the PITs $\{F_t(y_t)\}$ from uniformity. The PIT is closely related to the coverage of a forecast: if PIT_t is in the interval (0.05, 0.95) then the verifying observation falls within the central 90% interval of the predictive distribution, and if a collection of forecasts is well calibrated, 90% of the corresponding observations should fall within this interval.

In weather forecasting, calibration is frequently assessed visually, by constructing histograms of the PITs for a collection of forecast instances (Gneiting et al., 2007; Jolliffe and Stephenson, 2012). These provide a useful diagnostic tool for understanding the marginal error characteristics of the forecasts, with different types of error resulting in different types of non-uniformity in the histograms. A ∪-shaped histogram indicates that the verifying observations fall too frequently in the tails of the predictive distribution, indicating that the forecasts are generally underdispersive, or overconfident. Conversely, a ∩-shaped histogram indicates overdispersion: the forecasts are under-confident, and the observation falls too often in the centre of the forecast distribution. Systematic bias in the forecasts will result in a skewed histogram, with the direction of the skew providing information about the direction of the bias: where the forecast mean is too high, the observation will fall towards the lower tail of the forecast distribution, and the leftmost bins of the histogram will be overpopulated, with the rightmost bins overpopulated when the forecast mean is too low.

A similar approach is used to evaluate the calibration of forecasts issued

in the form of ensembles: in that case, the rank of the verifying observation within the ordered ensemble members is used, rather than the PIT (Hamill, 2001). The interpretation of these verification rank histograms is exactly the same as that of PIT histograms.

3.3.1.1 Numerical summaries of histogram shapes

Histograms have long been used to visually assess the calibration of competing weather forecasting methods (Anderson, 1996; Hamill and Colucci, 1997). However, this is subjective, and quickly becomes impractical when assessing the marginal calibration of high-dimensional multivariate forecasts, or when comparing calibration between two or more collections of forecasts. It is therefore useful to be able to produce an objective numerical summary of the shape of the data used to construct the histogram. Deviations from uniformity may be quantified using Pearson's χ^2 statistic to compare the expected and observed number of PITs in each bin of the histogram (Wilks, 2017), but this does not provide any information about the nature of the miscalibration. Instead, the skewness and dispersion of the PITs may be used to characterise the histogram in an easily interpretable way.

Any non-central tendencies in the histogram are summarised using the sample skewness, with symmetric histograms obtaining a perfect score of 0. Positive skewness indicates that the forecasts are, on average, too high, and negative skewness that the forecasts are too low.

The extent of any under- or over-dispersion in the histogram data is quantified by comparing the variance of the PITs to that of the uniform distribution that would be obtained if the forecasts were perfectly calibrated, which has variance $1/12$ (Casella and Berger, 2002). A dispersion index for the PITs of n forecasts can therefore be defined as

$$disp(PIT) = \frac{12}{n-1}\sum_{t=1}^{n}\left(PIT_t - \overline{PIT}\right)^2, \quad \text{where } \overline{PIT} = \frac{1}{n}\sum_{t=1}^{n} PIT_t. \quad (3.13)$$

While the variance of the PITs could be used directly to compare the relative

dispersion characteristics of two competing forecasting methods, the dispersion index has the advantage of being equal to one when the forecasts are perfectly calibrated, while a symmetric, ∪-shaped histogram will have a dispersion index greater than one, and a ∩-shaped histogram will have a dispersion index of less than one.

3.3.2 Joint calibration of multivariate forecasts

A forecasting method that is well calibrated for one variable in a single location may be sufficient for some limited applications. However, many weather forecasting applications require multivariate forecasts, consisting of predictions of one or more weather quantities at multiple locations or time steps. Even if a forecasting method is marginally well calibrated, the dependencies between weather quantities, locations or times must also be well predicted if the forecast is to be useful for decision making in these situations.

3.3.2.1 Box Ordinate Transform histograms

Gneiting et al. (2008) proposed the use of the Box Ordinate Transform (BOT) histogram to evaluate the multivariate calibration of forecasts issued as a predictive density. When the predictive density is a d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with verifying observation $\mathbf{y}$, the BOT is defined as

$$u = 1 - \chi^2_d\left((\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{3.14}$$

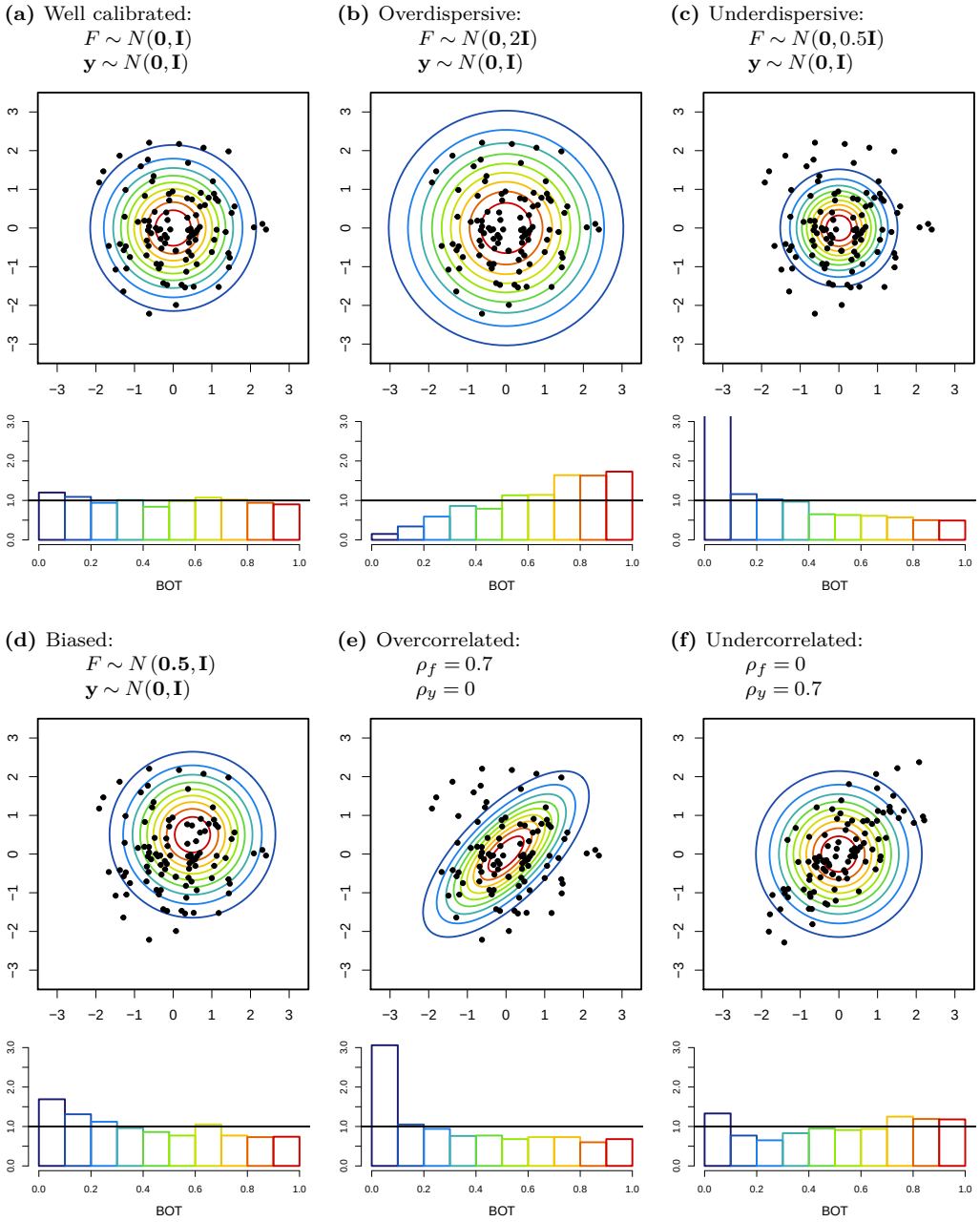
where $\chi^2_d(\cdot)$ denotes the CDF of a chi-squared distribution with d degrees of freedom, evaluated at the standardised observation (Box, 1980). If the predictive density F has a multivariate-normal distribution and $\mathbf{y}$ is drawn from F, then u will be uniformly distributed on the interval (0,1); a histogram of the BOTs for a collection of well-calibrated forecasts will therefore be uniform.

Interpretation of deviations from uniformity in the BOT histograms is less straightforward than interpretation of the PIT histograms. The BOT reflects

the centrality of the observation within the forecast distribution, with outlying observations being assigned low BOT values, and central observations assigned values close to one. Thus, a skewed histogram with too many high BOT values indicates an overdispersive forecast, with too many observations falling close to the centre of the distribution and too few in regions of low probability, as in Figure 3.2b. A preponderance of BOTs in the leftmost bins, indicating a high proportion of observations falling far from the centre of the forecast distribution, may indicate either that the forecasts are underdispersive, as in Figure 3.2c, that there is a bias in the forecast mean, as in Figure 3.2d, or that the forecasts are overcorrelated, as in Figure 3.2e. Undercorrelated forecasts produce ∪-shaped BOT histograms, as shown in Figure 3.2f.

It is clear from Figure 3.2 that, while the BOT histograms are particularly sensitive to underdispersiveness or overcorrelation in the forecast distributions, a BOT histogram in isolation cannot be used to distinguish between these types of miscalibration and biases in the forecasts: BOT histograms should therefore be interpreted in conjunction with the marginal PIT histograms or PIT summary statistics, which will reveal any significant marginal biases or dispersion issues that should be taken into account when diagnosing issues with joint calibration (Scheuerer and Hamill, 2015b).

Figure 3.2: Contours of predictive bivariate normal densities F and points representing 100 simulated 'observations' $\mathbf{y}$, with corresponding BOT histograms constructed from 1000 such observations. Contours indicate the values of the BOT at intervals of 0.1, corresponding to the bins used to construct the histograms.
Under- and over-correlated distributions have mean vector $\mathbf{0}$ and marginal variance 1, with forecast correlation ρ_f and observation correlation ρ_y.



3.3.2.2 Band Depth Rank histograms

A nonparametric alternative to the BOT histogram is the Band Depth Rank (BDR) histogram, proposed as a diagnostic tool for multivariate calibration by Thorarinsdottir et al. (2016). Like the BOT, the BDR is a measure of the centrality of an observation within a multi-dimensional forecast (López-Pintado and Romo, 2009). However, unlike the BOT, the BDR operates on forecasts represented as M-member ensemble forecasts, rather than on a predictive density having a known parametric form.

The band depth ranks are obtained as follows. Let $\mathcal{Z}$ denote the set of d-dimensional vectors $\{\mathbf{z}_1, ..., \mathbf{z}_{M+1}\} = \{\mathbf{y}, \mathbf{x}_1, ..., \mathbf{x}_M\}$, where $\mathbf{x}_1, ..., \mathbf{x}_M$ are the M members of an ensemble forecast and $\mathbf{y}$ is the verifying observation. The lth component z_l of each vector $\mathbf{z}$ is assigned a marginal rank among all of the lth components in $\mathcal{Z}$,

$$\text{rank}_{\mathcal{Z}}(z_l) = \sum_{i=1}^{M+1} \mathbb{1}_{\{z_{il} \leq z_l\}}, \tag{3.15}$$

and each vector $\mathbf{z}$ is assigned a prerank, based on the average of a quadratic function of the marginal ranks:

$$r(\mathbf{z}) = \frac{1}{d} \sum_{l=1}^{d} (M - \text{rank}_{\mathcal{Z}}(z_l)) (\text{rank}_{\mathcal{Z}}(z_l) - 1) + M - 1. \tag{3.16}$$

The preranks $r(\mathbf{z})$ measure the centrality of each element in $\mathcal{Z}$ with respect to the marginal axes, with more central elements attaining higher ranks, and extreme outlying elements attaining the lowest (Thorarinsdottir et al., 2016; Wilks, 2017). The band depth rank of the observation $\mathbf{y} = \mathbf{z}_1$ is the rank of $r(\mathbf{z}_1)$ in $\{r(\mathbf{z}_1), ..., r(\mathbf{z}_{M+1})\}$, with ties broken at random. For ease of interpretation, the ranks are normalised to lie between 0 and 1 by subtracting 1 and dividing by M before plotting. As with the BOT histograms, a histogram of BDRs from well-calibrated forecasts will be uniform (Thorarinsdottir et al., 2016). Band depth ranks can be computed using the `depthTools` package in R (Lopez-

Pintado and Torrente., 2021).

The BDR histogram was originally proposed as a means to evaluate the joint calibration of multivariate forecasts issued as ensembles; in order to compute the BDRs for forecasts issued as probability densities using the method outlined above, it is necessary to simulate an ensemble from the predictive distribution in order to obtain the marginal ranks $\text{rank}_{\mathcal{Z}}(z_l)$, introducing sampling uncertainty into the estimation of the BDR. However, as noted in Section 3.3.1, the verification rank of an observation within an ensemble is directly equivalent to the PIT of the observation with respect to a forecast density. When the forecast is issued as a predictive distribution, therefore, the marginal PITs can be computed directly, instead of using the verification ranks defined in (3.15). Thus, replacing $M - \text{rank}_{\mathcal{Z}}(z_l)$ with $1 - \text{PIT}(z_l)$ and $\text{rank}_{\mathcal{Z}}(z_l) - 1$ with $\text{PIT}(z_l)$ in (3.16), and omitting the term $M - 1$, which does not affect the ordering of the preranks, results in the exact parametric pre-rank function

$$r^*(\mathbf{z}) = \frac{1}{d} \sum_{l=1}^{d} \Big[1 - \text{PIT}(z_l)\Big] \text{PIT}(z_l). \tag{3.17}$$

To obtain a fully parametric equivalent to the band depth rank, it would be necessary to derive the distribution of $r^*(\mathbf{z})$ under the joint predictive density; this is outside the scope of this thesis, and is left as future work. Instead, a semiparametric band depth rank can be computed by drawing a synthetic ensemble from the multivariate predictive density and computing the exact prerank (3.17) for all ensemble members and for the observation; the semiparametric band depth rank is the rank of the prerank $r^*(\mathbf{y})$ of the observation within that simulated population of pre-ranks.

Figure 3.3 shows the contours of the BDR 'surface' for each of the synthetic bivariate normal forecasts examined in Figure 3.2, and the corresponding BDR histograms. The shape of the BDR contours remains fairly stable regardless of the form of the underlying forecast distribution, with the distance between the contours representing 10% intervals of the band depth rank changing to

reflect different changes in the forecast specification. Increasing or reducing the marginal variances stretches or compresses the entire surface, moving the contours further from or closer to the centre of the forecast distribution. Increasing the strength of the correlations decreases the gradient of the surface and pushes the contours apart, moving the most central bands towards the centre of the distribution and moving the outermost bands further away. The BDR, which is based on a function of the marginal depths in (3.15), is more sensitive to marginal misspecification than to errors in the forecast correlation structure: the difference between the BDR contours in Figures 3.3a and 3.3e, which show the effects of a substantial change in correlation, is much smaller than the difference between the contours in Figures 3.3a and 3.3b or 3.3c, which show the effects of a change in marginal variance.

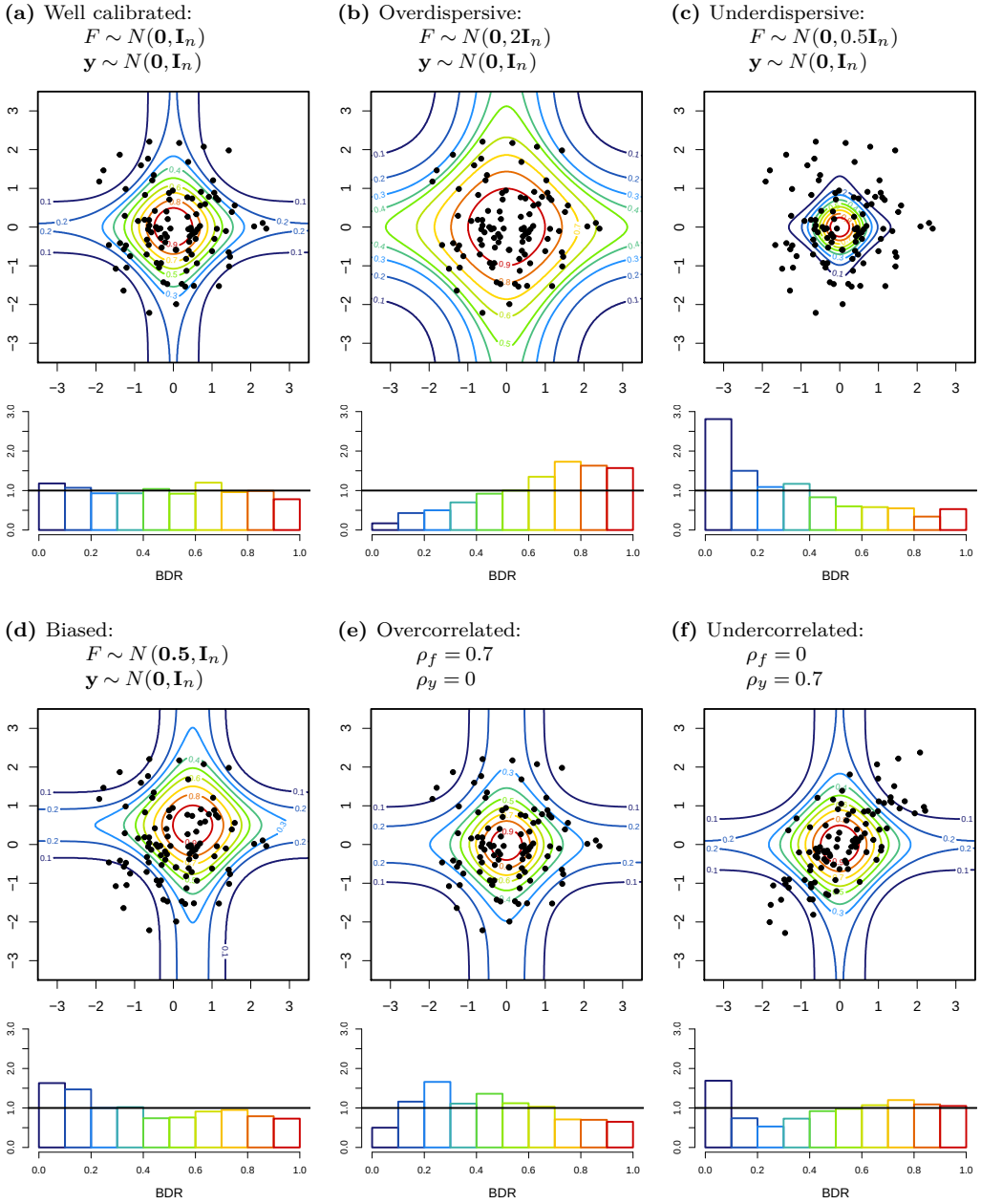
As with the BOT, a well-calibrated forecast will have a uniform BDR histogram, while an overdispersive forecast will have more values in the rightmost bins, and biased or underdispersive forecasts will produce in more values in the leftmost bins. When the forecast distribution is overcorrelated, as in Figure 3.3e, the band depth rank histogram will have a slight $\cap$-shape, while undercorrelated forecasts produce $\cup$-shaped histograms.

3.3.2.3 Joint interpretation of the BOT and BDR

The example histograms in Figures 3.2 and 3.3, each reflecting only a single type of specification error, display fairly clear patterns and are easy to interpret. However, this is not generally the case in real applications, where a collection of forecasts may display a combination of calibration errors, which can make interpretation of the resulting histograms difficult. In order to fully understand the nature of any joint miscalibration, it is generally advisable to consider multiple diagnostic methods together, as recommended by Wilks (2017).

Figure 3.4 shows an example of a situation where this is not only advisable, but necessary: both the BOT and BDR histograms are right-skewed, suggesting that the forecasts must be either underdispersive or biased.

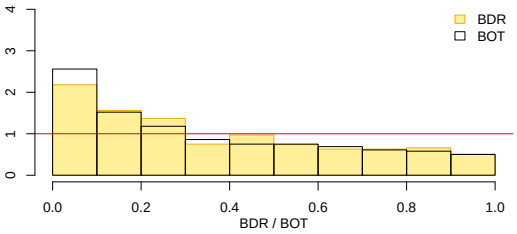
Figure 3.3: Contours of band depth ranks for predictive bivariate normal densities f and points representing 100 simulated 'observations' $\mathbf{y}$, with the corresponding BDR histograms constructed from 1000 such observations. Contours are drawn at intervals of 0.1, corresponding to the bins used to construct the histograms.
Under- and over-correlated distributions have mean vector $\mathbf{0}$ and marginal variance 1, with forecast correlation ρ_f and observation correlation ρ_y.



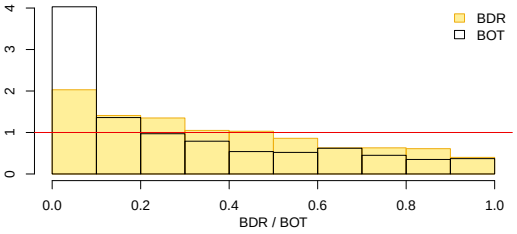
The forecasts are indeed marginally underdispersive – something that could be diagnosed by considering the PIT histograms or the PIT dispersion

Figure 3.4: An example of two different types of miscalibration, where the BOT and BDR histograms individually cannot clearly diagnose the underlying issue.
Observations are simulated from a 5-dimensional standard normal distribution.

(a) Forecast f_1 is underdispersive, with marginal variances 0.7



(b) Forecast f_2 is both underdispersive and overcorrelated. Its correlation matrix is a Toeplitz matrix with top row $\begin{bmatrix}1 & 0.4 & 0.16 & 0.064 & 0.0256\end{bmatrix}$, and all marginal variances are 0.7.



indices – but forecast f_2 is also over-correlated, something that is not obvious from the histograms alone.

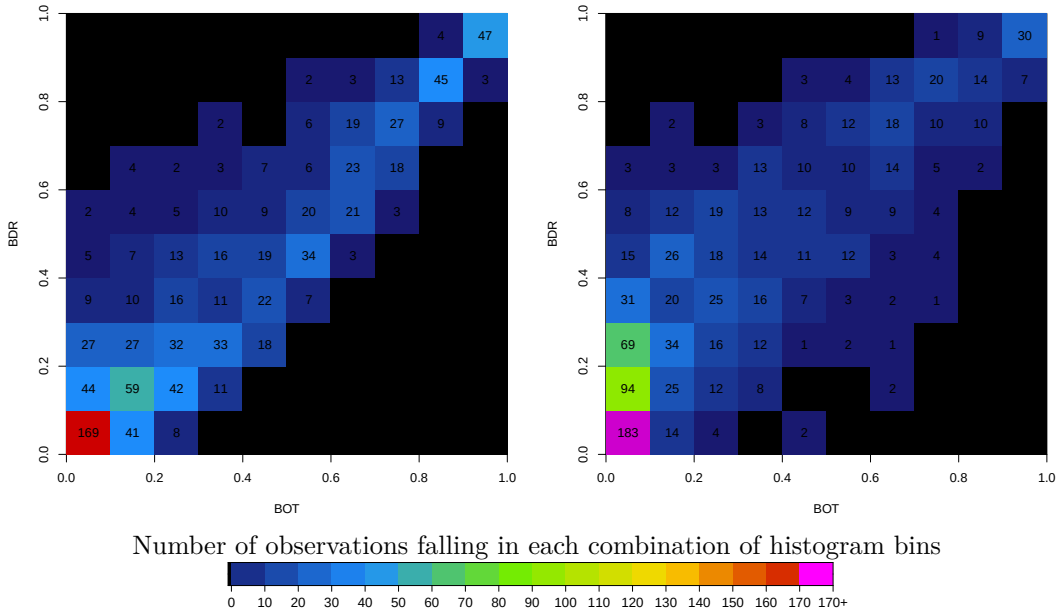
In cases like this, diagnosis of errors in the forecast dependence structure may be facilitated by considering the joint distribution of the two depth measures. Figure 3.5 displays this joint distribution in the form of a grid plot, with each cell showing the count of observations falling in the corresponding intervals of the BOT and BDR histograms. Unlike the histograms, the two grids show quite distinct patterns: in Figure 3.5a, most of the counts are clustered around the diagonal, whereas there is less agreement between the two depth measures in Figure 3.5b. Here, points in the leftmost column, corresponding to points in the leftmost bin of the BOT histogram, are spread among the four lowest cells, corresponding to the three leftmost bins of the BDR histogram. Experiments suggest that this pattern is typical of over-correlated forecasts, a calibration error that is often masked in the BOT and BDR histograms by marginal misspecifications.

Figure 3.6 shows gridplots for five-dimensional forecasts displaying the same types of miscalibration presented in Figures 3.2 and 3.3. Counts are more likely to fall in particular regions of the gridplot under different types of misspecification; the gridplots may therefore reveal patterns within the joint

Figure 3.5: Gridplots showing the joint distribution of the BOT and BDR for the misspecified forecasts f_1 and f_2, aggregated over the intervals used to construct the histograms in Figure 3.4.

(a) Forecast f_1 (underdispersive, correlations correct)

(b) Forecast f_2 (underdispersive, correlations too strong)



distribution of the BOT and BDR histograms that cannot be identified by examining the histograms alone. For a well calibrated forecast, roughly equal counts are observed in the neighbourhood of the diagonal, with slightly higher counts in the extreme top-right and bottom-left cells. When the forecast is overdispersive, counts tend to cluster in the top-right corner of the grid, and for underdispersive forecasts, in the bottom-left corner. A similar pattern is observed for biased forecasts; these two calibration errors are easily distinguished by considering the marginal PITs. When the correlation between the forecast variables is too strong, counts tend to move left from the diagonal, accumulating in the lower half of the leftmost column. When the forecasts are undercorrelated, counts tend to move down and to the right from the diagonal, and tend to cluster in the top-right and bottom-left corners of the grid, indicating that too many observations are falling either very close to or very far from the centre of the joint forecast distribution. While bias and dispersion errors are easily diagnosed from the marginal PIT histograms, consideration of the joint

distribution of the BOT and BDR histograms can therefore reveal issues in the dependence structure that would otherwise be concealed by marginal calibration issues.

Figure 3.6: BOT and BDR histograms, and gridplots of the joint distribution of the BOT and BDR, under the five-dimensional forecast f of 1000 simulated observations $\mathbf{y} \sim N(\mathbf{0}, \mathbf{T}(0.5))$, where $\mathbf{T}(\rho)$ is a Toeplitz matrix with top row $\begin{bmatrix} 1 & \rho & \rho^2 & \dots \end{bmatrix}$.

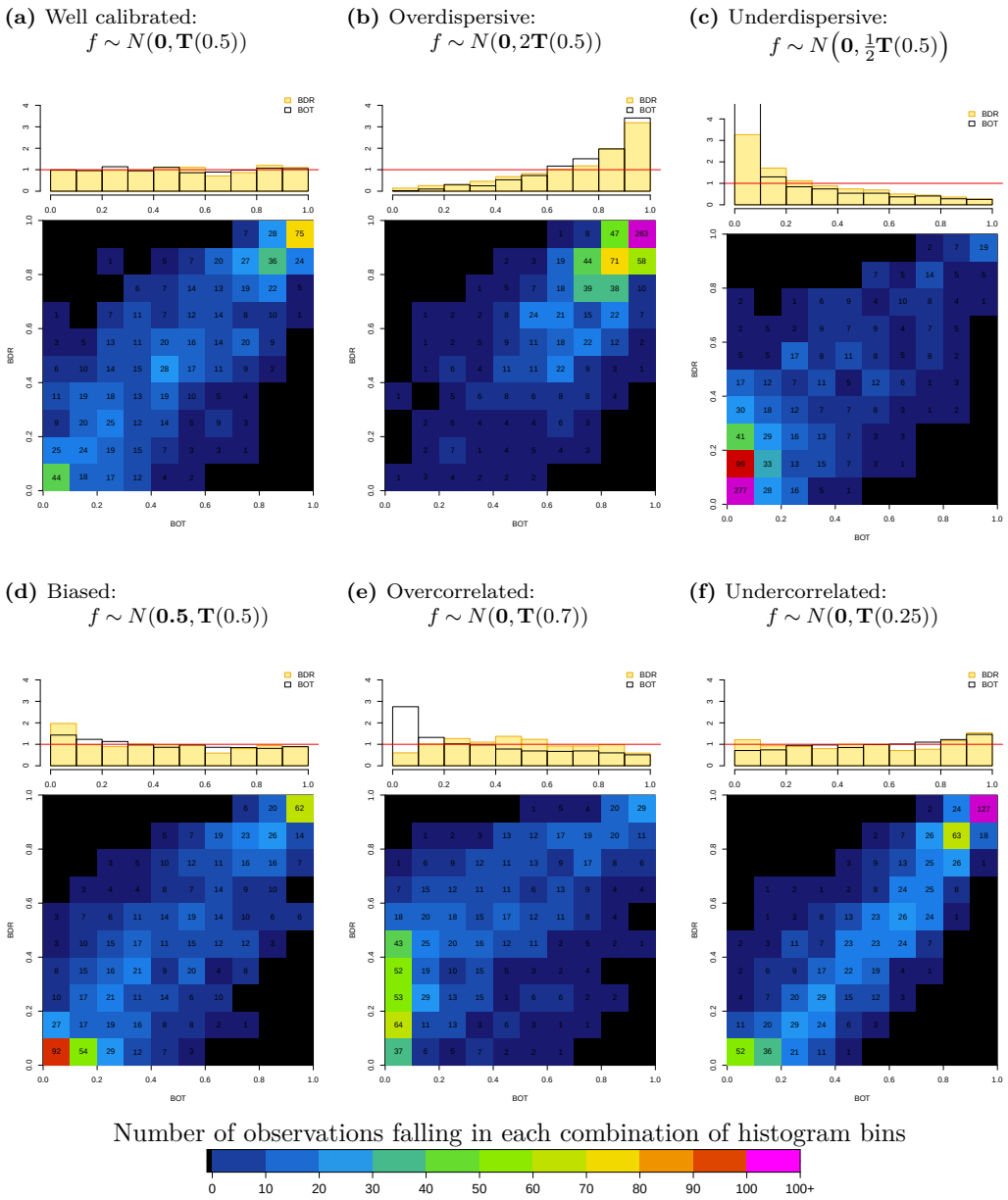
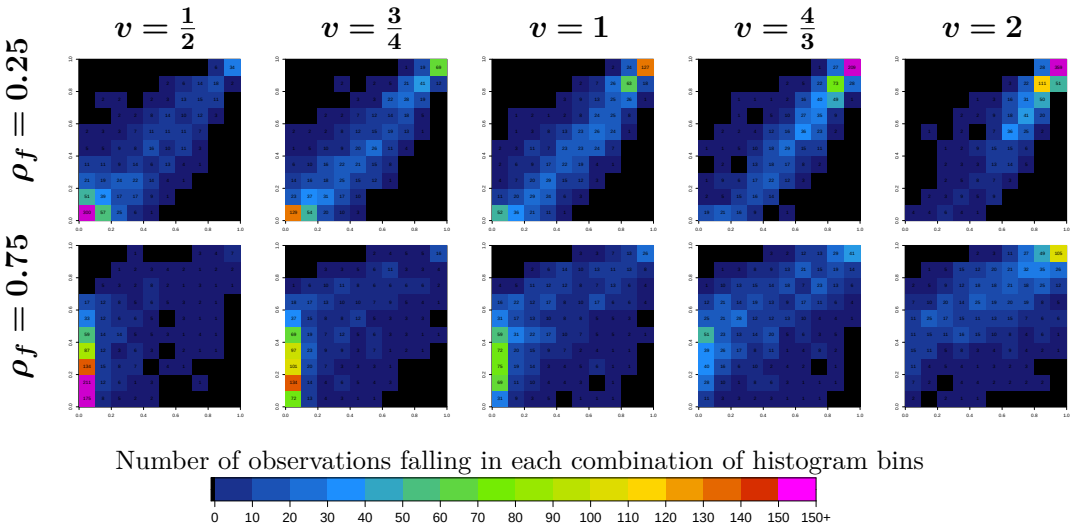


Figure 3.7 shows gridplots for forecasts where both the correlation strength and the marginal variance are misspecified, to varying degrees. In the top

row, the forecast correlations are too weak, while the bottom row suffers from too-strong forecast dependences. The marginal variance increases from left to right, being too low in the leftmost column and too high in the rightmost. In the top row, although it is possible to detect concentrations of values in the top-right and bottom-left of the grid when the variance is correctly specified, the pattern is not qualitatively different to that obtained when the forecasts are jointly well calibrated, as in Figure 3.6a; moreover, even this slight effect is masked when there is marginal over- or underdispersion. A stronger signal can be detected in the bottom row, where counts in the leftmost column are clustered a few cells up from the bottom of the grid. The peak in the leftmost column is shifted down towards the bottom corner when the forecast variance is too low, and up towards the middle of the BDR range when the forecast variance is too high; this clustering can be detected even under quite extreme misspecification of the marginal variance.

Figure 3.7: Gridplots showing the effect of compound misspecification, when forecasts have either variance, correlation, or both misspecified. The observations are drawn from $\mathbf{y} \sim N\left(\mathbf{0}, \mathbf{T}^{0.5}\right)$, where $\mathbf{T}^{\rho}$ is a Toeplitz matrix with top row $\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \end{bmatrix}$. Forecasts have the form $f \sim N(\mathbf{0}, v\mathbf{T}^{\rho_f})$.



The examples shown here are constructed from synthetic data in order to display the patterns characteristic of each type of error. In practice, it is unlikely

that a collection of forecasts will suffer from only one type of misspecification: any single forecast may suffer from compound miscalibration, while a collection of forecasts may display a mixture of miscalibration types. Considering the joint distribution of the BOT and BDR in this way can provide clearer patterns for diagnosis than a direct comparison of two histograms. In Chapters 4 and 6, BOT-BDR gridplots are used to diagnose miscalibration in the dependence structure that would otherwise be masked by marginal calibration issues. The simulations presented here suggest that the diagnostic patterns in the BOT-BDR grids are consistent, although further work is required to understand the expected joint distribution of the BOT and BDR under various types of forecast misspecification, and to consider whether plots of the joint distribution of other pairs of depth scores might be more informative about other types of miscalibration.

3.4 Summary

This chapter reviews verification methods appropriate for the evaluation of forecasts issued in the form of a predictive density or in the form of an ensemble of deterministic forecasts. The scoring rules and diagnostic tools described here will be used to quantify and compare the skill of forecasts postprocessed using competing methods in Chapters 4 and 6.

While this chapter is primarily a review of existing forecasting methods, it also introduces two potentially useful innovations. The first of these is a semiparametric equivalent to the band depth rank histograms proposed as a diagnostic tool to identify joint calibration issues by Thorarinsdottir et al. (2016). This PIT-based band depth rank exploits the fact that, when a forecast is issued in the form of a parametric density, the exact marginal band depths can be computed, removing the sampling variability introduced by using an ensemble approximation – although some sampling variability remains, due to the need to use simulation to obtain the final joint ranks. Further work is required to understand the joint distribution of the marginal PITs, and so to

remove this second source of uncertainty.

The second innovation introduced in this chapter is the gridplot used to visualise the joint distribution of the BOT and BDR. While the idea of using two or more depth histograms to diagnose joint calibration is by no means new (Wilks, 2017), the gridplots reveal certain patterns of behaviour in the two depth measures, which can provide useful information about the joint calibration of the forecasts beyond that available from the histograms alone. Again, further work is needed to understand the expected joint distribution of the BOT and BDR, which will vary not only with different types of forecast misspecification, but also with different parametric forms for the predictive distribution; and to understand whether plots of the joint distribution of other depth measures might provide more information about other types of dependence misspecification.

Chapter 4

Application of MME postprocessing to temperature forecasts

In this chapter, the Bayesian postprocessing framework proposed in Section 2.2 is used to produce posterior predictive distributions for winter surface temperatures over the UK. The raw forecasts are operational forecasts issued by the ECMWF, Met Office and NCEP weather centres, and downloaded from the TIGGE archive, as described in Section 2.1.2. The skill of the resulting forecasts is evaluated with respect to the ERA-Interim reanalysis dataset, using the measures of forecast skill and calibration described in Chapter 3. The chapter expands on the results presented in Barnes et al. (2019).

Section 4.1 examines the effect of the choice of prior distribution discussed in Section 2.2.3; this aspect of the problem is considered first so that the best-performing choice of prior can be used in subsequent analyses. Section 4.2 compares the skill of the Bayesian posterior forecasts to that of forecasts using a simple superensemble and the Nonhomogeneous Gaussian Regression method described in Section 2.1.4. The relative skill of forecasts postprocessed using training sets chosen using analogue methods rather than a moving window, as described in Section 2.3, is considered in Section 4.3.

Only forecasts verified during the meteorological winter months (December,

January and February) are postprocessed and evaluated, but training cases were also drawn from forecasts verified in October and November. All of the forecast postprocessing was carried out using 25 training cases per forecast instance. Training sets of 25-30 instances are commonly used in postprocessing short- to medium-range weather forecasts like the ones in this study (Raftery et al., 2005; Hagedorn et al., 2008, 2012; Junk et al., 2015; Zamo et al., 2021); a preliminary sensitivity analysis indicated that the exact size of the training set made no difference to the relative performance of any of the postprocessing methods discussed here. The sensitivity analysis is not presented in detail, in order to keep the focus on the relative skilfulness of the different postprocessing approaches given the same set of candidate training cases.

4.1 Effect of choice of prior

One of the potential advantages of the Bayesian framework over the nonhomogeneous Gaussian regression (NGR) described in Section 2.1.4.1 is the possibility to incorporate a prior distribution for the weather quantity of interest. In this section, the forecasts are postprocessed with several different choices of prior, and with $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ estimated from a 25-day moving window prior to the forecast issue date. The prior that produces the most skilful posterior forecasts will be retained in subsequent sections when comparing the performance of the Bayesian postprocessed forecasts with competing methods, and when comparing the relative skill of forecasts postprocessed with discrepancies estimated from training sets selected using different methods.

As a baseline, the noninformative prior obtained by setting $\boldsymbol{\Gamma} = \mathbf{0}$ was used. Next, the forecasts were postprocessed using a climatological prior constructed for each forecast instance by taking the sample mean and covariance matrix of the observations within a seven-day window centred on the verification date, from the ten years prior to the year in which the forecast was issued, as described in Section 2.2.3.1. The skill of these forecasts is compared to that of forecasts postprocessed using the sequential approach described in Section

2.2.3.2, with a noninformative prior used for the initial 15-day-ahead forecast.

The forecasts were also postprocessed using a twenty-year climatology, which resulted in very similar performance to those using a ten-year climatology; likewise, the results for forecasts postprocessed sequentially using a ten-year climatological prior for the first forecast issued were very similar to those using a noninformative prior for the first forecast. These variants are not discussed further.

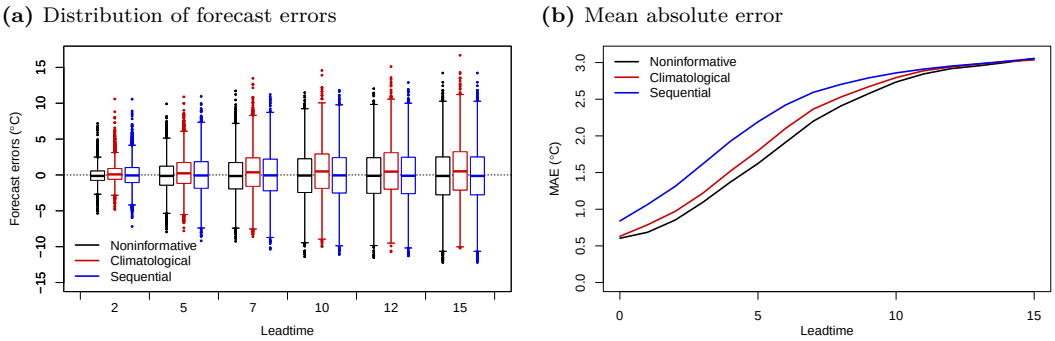
4.1.1 Forecast accuracy, sharpness and skill

Figure 4.1a shows the distribution of the Bayesian postprocessed forecast errors (the difference between the observation and the mean of the posterior forecast distribution) for all 630 forecast instances and all 13 study regions. At the shortest leadtimes, forecasts using a noninformative prior typically have slightly smaller errors than those using a climatological prior or sequential postprocessing, although at longer leadtimes, all three approaches produce a similar spread of errors. The forecasts using a climatological prior retain a slight warm bias increasing from 0.2°C at the shortest leadtimes to 0.5°C at the longest; those using a sequential or noninformative prior typically have a very small cold bias of less than 0.1°C at all leadtimes.

The mean absolute errors (MAE) for each choice of prior are shown in Figure 4.1b. At the shortest leadtimes, the sequentially postprocessed forecasts have higher MAE, reflecting the wider spread of the errors for those forecasts; the forecasts using a climatological prior have a slightly smaller spread and so achieve a somewhat lower MAE than those that are sequentially postprocessed, despite the residual bias already mentioned. Perhaps surprisingly, the most accurate forecasts at all leadtimes are those for which no informative prior is provided. Possible reasons for this are considered in Section 4.1.4.

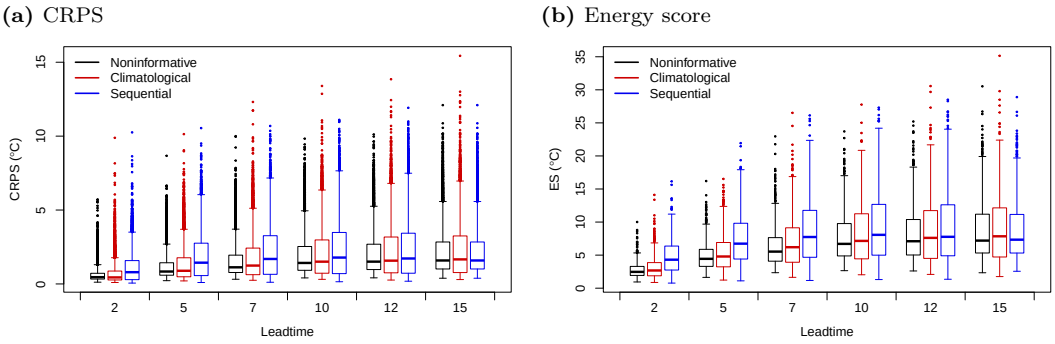
The MAEs of the three methods are closely reflected by the CRPS and its multivariate extension, the energy score (ES), shown in Figure 4.2a. At all but the longest leadtimes the sequentially postprocessed forecasts perform somewhat worse than those obtained using either of the other approaches. Similar median

Figure 4.1: Distribution of Bayesian postprocessed forecast errors and MAE over all 630 forecast instances for all 13 locations when forecasts are postprocessed with different prior distributions. Forecast errors are shown at leadtimes of 2, 5, 7, 10, 12 and 15 days.



scores are obtained when using either a climatological or noninformative prior, although the climatological prior produces a wider range of scores at each leadtime.

Figure 4.2: Distribution of continuous ranked probability score (CRPS) and its multivariate extension, the energy score (ES), at selected leadtimes over all 630 forecast instances for all 13 locations for forecasts postprocessed using different prior distributions.

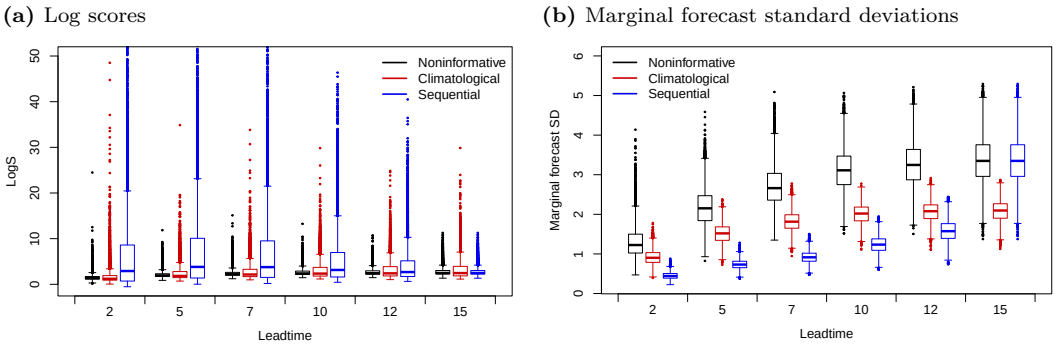


The difference in forecast skill is more striking when evaluated in terms of the logarithmic scoring rule, which rewards forecasts that place a high posterior density on the temperature actually observed, and which are therefore likely to be useful in threshold-based decision-making (Section 3.2.2). Figure 4.3a shows the log scores for all forecast instances, truncated at 50 so that the detail is visible; at all but the longest leadtime, the sequentially postprocessed forecasts typically receive substantially higher log scores than those postprocessed independently, indicating that the sequentially postprocessed forecasts

place low probability on the outcome actually observed. The median log scores of forecasts using a noninformative or climatological prior are again similar, although the log scores of forecasts postprocessed using climatological priors again have a wider spread than those using a noninformative prior, with the worst-performing instances receiving substantially higher scores.

The reason for this striking difference in log scores is clear when the forecast sharpness, shown in Figure 4.3b, is considered. The inclusion of any informative prior immediately reduces the variance of the marginal posterior forecasts from the variance that would be obtained with a zero-precision noninformative prior, as discussed in Section 2.2.3.1; the use of sequential postprocessing, which takes the previous posterior forecast as the new prior for each instance, propagates this variance reduction through the forecasts as the verification date approaches. As a result, the sequentially postprocessed forecasts are extremely sharp at the shortest leadtimes, but with no corresponding improvement in predictive accuracy to justify this increased confidence.

Figure 4.3: Distribution of logarithmic scores (LogS) and marginal forecast standard deviations at selected leadtimes over all 630 forecast instances for all 13 locations, for forecasts postprocessed with different prior distributions.



4.1.2 Marginal forecast calibration

Figure 4.4 shows the PIT histograms for the three methods in two grid cells, representing typical results for Scotland and southern England; histograms for the remaining regions, shown in Figure B.1 in Appendix B, are broadly similar in shape. The mean and range of the skewness and dispersion of the

PIT histograms at each leadtime are summarised in Figure 4.5.

The histograms for the sequentially postprocessed forecasts are extremely ∪-shaped at all but the longest leadtime (for which a noninformative prior was used, in the absence of an earlier posterior forecast for the same verification date); this reflects the low forecast standard deviations noted in Figure 4.3b. As a result, the sequentially-postprocessed forecasts have a very high dispersion index at all leadtimes (Figure 4.5b), although the histograms are fairly symmetric.

The forecasts using a climatological prior typically are also somewhat ∪-shaped – more so at longer than shorter leadtimes – with the leftmost bin typically containing more values than the rightmost, reflecting the residual bias noted in Figure 4.1a, and resulting in positive skewness in the PITs at all leadtimes (Figure 4.5a). The dispersion index is greater than 1 at all leadtimes, reflecting this underdispersiveness, although the dispersion index is lower than that of the sequentially postprocessed forecasts at all but the longest leadtimes (Figure 4.5b).

At the shortest leadtimes the histograms for the forecasts using a non-informative prior have a slight ∩ shape, indicating that they are slightly underconfident. The peak of the histograms tends to fall slightly to the right of the centre, reflecting the residual cold bias in the forecasts; however, this bias is small, with the verifying observations tending to land in the 50th-70th percentile of the predictive distribution, rather than in the tails. At longer leadtimes this ∩-shape is no longer evident, with histograms either slightly ∪-shaped – indicating overconfidence in the forecasts – or generally flat. At all but the shortest leadtimes, where the PITs have slight negative skewness, the forecasts using a noninformative prior are generally well calibrated, with close to zero skewness and dispersion indices close to one (Figure 4.5).

4.1.3 Joint forecast calibration

The univariate calibration issues just described are amplified in the depth rank histograms used to assess multivariate calibration. The sequentially postprocessed forecasts, which are marginally too sharp at all leadtimes, have

Figure 4.4: PIT histograms showing the marginal calibration of postprocessed forecasts of surface temperatures at selected locations in the north and south of the UK at a range of leadtimes, using different prior distributions for the observed temperature.

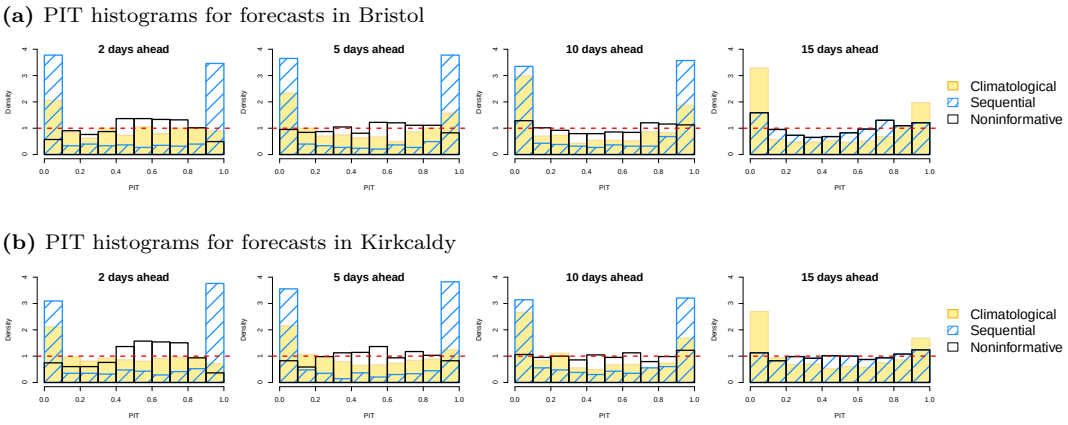
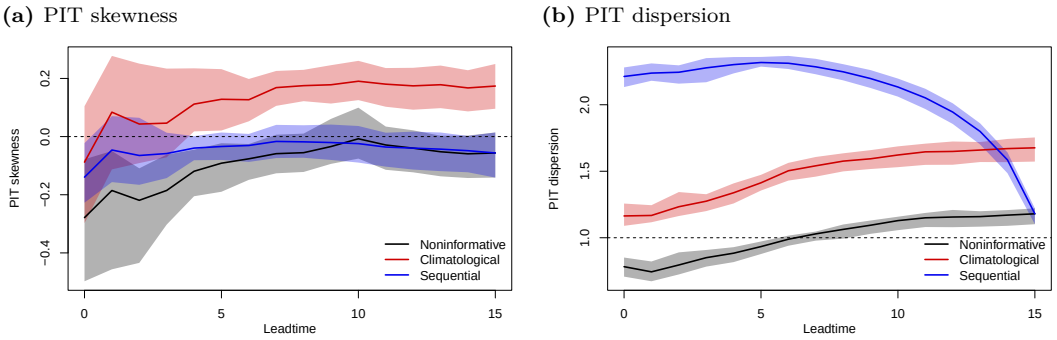


Figure 4.5: Characteristics of the PIT histograms at each leadtime. The lines indicate the mean value across the thirteen grid cells, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



a high proportion of values in the leftmost bin of both the Box ordinate transform (BOT) histograms (Figure 4.6) and the band depth rank (BDR) histograms (Figure 4.7). The histograms for the forecasts postprocessed using a climatological prior display a similar tendency, although to a lesser degree, reflecting the better marginal calibration of those forecasts.

The BOT histograms for the forecasts postprocessed using a noninformative prior also have a large proportion of values in the leftmost bin; however, a small spike of values in the rightmost bin can also be detected at all leadtimes, indicating a small group of observations falling too close to the centre of the distribution. At shorter leadtimes, the BDR histograms for these forecasts

are skewed in the opposite direction, with too many observations falling close to the centre of the joint predictive distribution; this reflects the slight bias visible in the ∩-shaped univariate PIT histograms. As the leadtime increases, the skew in the BDR histograms for these forecasts shifts, reflecting the slight underdispersiveness of the forecasts at longer leadtimes. The BDR histograms for the noninformative forecasts are the closest to uniform of the three at all leadtimes, indicating that those forecasts have better joint calibration.

The PIT and BDR histograms for the forecasts using a noninformative prior do not reveal any substantial lack of either marginal or joint calibration, while the BOT histograms indicate poor joint calibration; together, this suggests that some aspect of the multivariate dependence structure is misspecified. Gridplots of the joint distribution of the BOT and BDR for these forecasts are shown in Figure 4.8. At shorter leadtimes, there is a cluster of observations in the leftmost bin of the BOT histograms with BDRs between 0.3 and 0.4; as discussed in Section 3.3.2.3, this suggests that the dependences between the temperatures in different grid cells are typically overestimated by the corresponding forecasts. At the same time, a large number of forecast instances also fall in the top-right corner of the grid, indicating a group of forecasts for which the verifying observations fall very close to the centre of the forecast distribution. These two groups of instances are well separated in the gridplot, suggesting that the distribution of the forecast errors may in fact be a mixture of two different populations. Further work is required to determine whether there really are two sub-populations of forecast errors (and so, by implication, two sub-populations of forecast instances) within the set of winter temperature forecast-observation pairs. With increasing leadtime, as the forecasts switch from being generally overdispersive to generally underdispersive, the clusters become less distinct.

4.1.4 Summary

Forecasts postprocessed using a noninformative or climatological prior generally had similar marginal skill scores even though those using a climatological

Figure 4.6: Box Ordinate Transform (BOT) histograms showing the joint calibration of postprocessed forecasts of surface temperatures using different choices of prior.

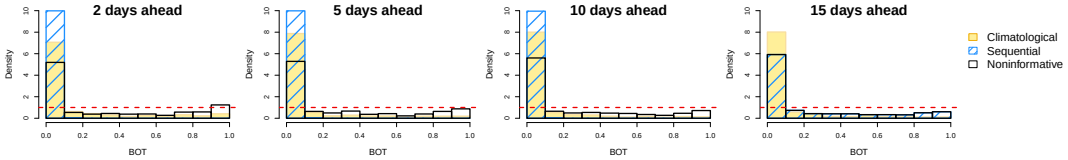


Figure 4.7: Band Depth Rank (BDR) histograms showing the joint calibration of postprocessed forecasts of surface temperatures using different choices of prior.

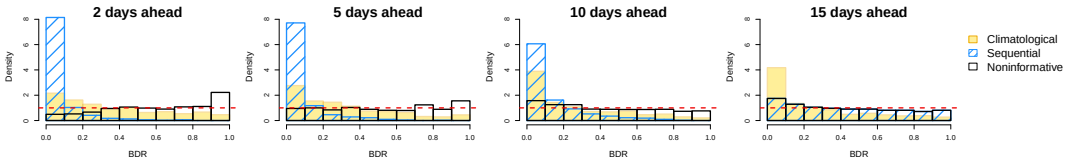
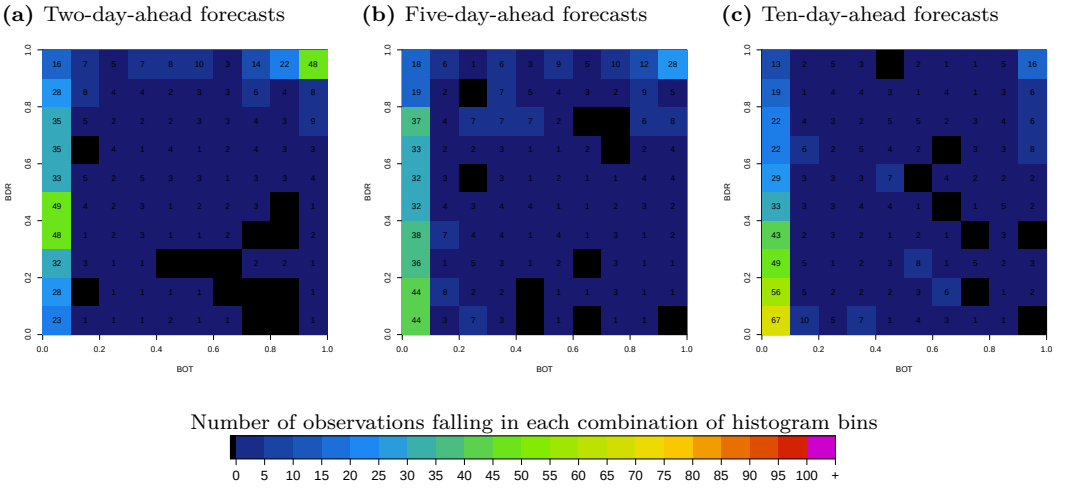


Figure 4.8: Gridplots summarising the joint distribution of the BOT and BDR histogram counts for forecasts postprocessed using a noninformative prior at selected leadtimes.



prior retained a slight residual bias. Those using a climatological prior had lower median, but higher mean CRPS and log scores than those using a noninformative prior. This may be because the forecasts using a climatological prior were sharper, improving the skill score for the majority of forecasts, but incurring greater penalties in the worst-performing cases. In spite of this, the forecasts using a noninformative prior had much better marginal and joint calibration.

It is possible that forecasts with an informative prior component are less

well calibrated than those with a non-informative prior (being sharper but less accurate) because the discrepancy-adjusted forecasts are the best source of information available when predicting the future weather. Adding further information in the form of an informative prior reduces the posterior variance but rarely brings a corresponding improvement in the accuracy of the mean forecast. An alternative perspective is that because the forecasts are based on dynamical data assimilation, they have already implicitly accounted for the 'prior' information, as described in Section 2.1.

At all leadtimes, the sequentially postprocessed forecasts were the least accurate and least well calibrated of the three methods, having the highest MAE and being far too sharp. This was an unexpected result: the noninformative prior forecasts are fairly well calibrated, so using the n-day-ahead noninformative posterior as the prior for the $n-1$-day-ahead forecasts is expected to produce more skilful forecasts. The poor performance of the sequentially postprocessed forecasts was initially suspected to be due to propagation of errors from poor long-leadtime forecasts through to shorter leadtimes; however, the same approach was tested starting from the five-day-ahead postprocessed forecasts, and also using only the single preceding posterior obtained with a noninformative prior at each leadtime, with similar results (not shown). This suggests that the issue lies with the postprocessing framework – or with the implementation of it used here – rather than with the quality of the prior distribution used.

One fundamental issue in the simple sequential postprocessing approach used here is that the framework implicitly assumes that the forecasts combined are independent predictions of the future weather, when in fact – as Figure 2.10 shows – much of the information in the raw n-day-ahead forecasts is already incorporated in the $n-1$-day-ahead forecasts through the data assimilation process, along with more recent observations of the atmosphere. This suggests that a possible improvement to the sequential postprocessing approach would be to explicitly account for correlation between the prior and ensemble forecasts,

and to adjust the posterior variance accordingly. However, although this may be able to mitigate the calibration issues discussed in Section 4.1.2, it is unlikely to resolve the lower accuracy of the sequential forecasts (Figure 4.1b).

4.2 Comparison of Bayesian postprocessing with other methods

In this section, the skill of the Bayesian posterior forecasts is evaluated against that of simple superensemble forecasts and forecasts postprocessed using nonhomogeneous regression (NGR). The Bayesian forecasts are postprocessed using a noninformative prior, which was found in Section 4.1 to produce the most well-calibrated forecasts.

The superensemble forecasts are obtained by taking the sample mean and covariance matrix of all ensemble members as described in Section 2.1.3.1, with no attempt made to correct for bias or dispersion errors; these provide a baseline against which the improvements due to the other methods can be compared. The marginal NGR forecasts are fitted by CRPS minimisation as described in Section 2.1.4.1; for each forecast instance, the sample correlation matrix of the verifying observations in the training dataset provides the dependence structure of the joint forecasts, as discussed in Section 2.1.4.1.2. Each forecast instance is postprocessed using the same 25 moving-window training cases for both the NGR and Bayesian forecasts.

4.2.1 Forecast accuracy, sharpness and skill

Figure 4.9 shows the biases of the postprocessed forecasts in each of the grid cells shown in Figure 2.4[1]. The superensemble forecasts have a pronounced cold bias, which tends to increase with latitude; this bias is almost completely removed for all grid cells by postprocessing with either NGR or the Bayesian method, although a very small cold bias does remain. However, the NGR-

[1]The jump in the biases between superensemble forecasts issued at 10- and 11-day leadtimes is due to a reduction in both the spatial and temporal resolution of the ECMWF model after 10 days of model time has elapsed (ECMWF, 2021b), as noted in Table 2.1.

postprocessed forecasts show a greater improvement in overall MAE (Figure 4.10a), particularly at longer leadtimes; as Figure 4.10b shows, the spread of the errors from the Bayesian-postprocessed forecasts is similar to that of the superensemble forecast, while the errors from the NGR forecasts have a narrower distribution. This can be ascribed to the fitting of the regressions by minimising the CRPS, a metric known to reward accuracy in the mean forecasts, as described in Section 3.2.1.

Figure 4.9: Regional biases for forecasts produced by each postprocessing method

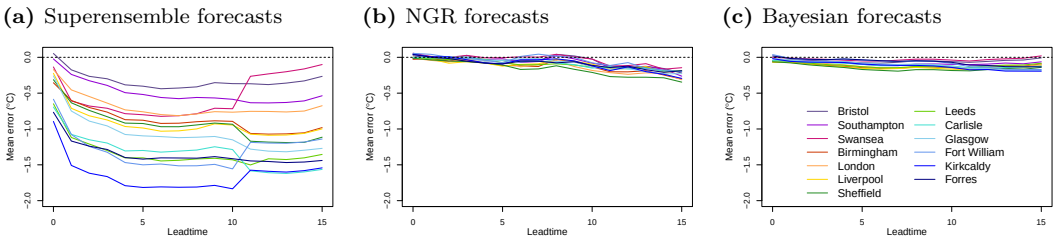
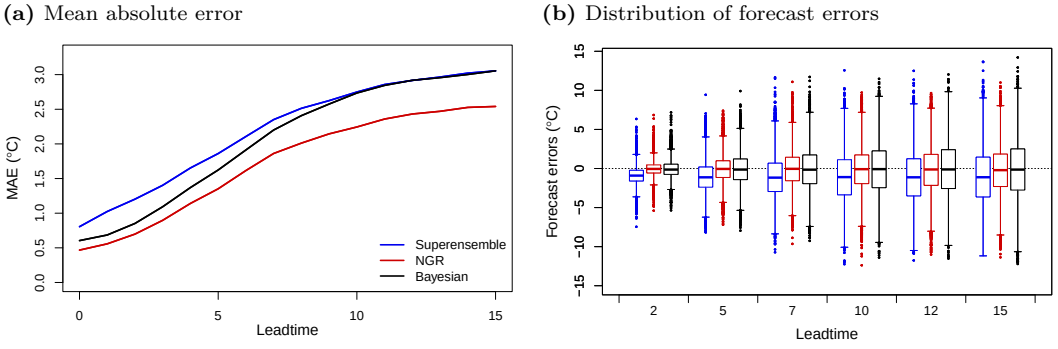


Figure 4.10: Mean absolute error (MAE) and distribution of forecast errors at selected leadtimes over all 630 forecast instances and all 13 locations, for forecasts produced by different postprocessing methods.



A similar pattern of performance can be seen in the CRPS and ES, shown in Figure 4.11. The entire distribution of the scores for the NGR forecasts is slightly closer to zero than that of either the superensemble or Bayesian forecasts, suggesting that the method is able to improve the skill of both the best- and worst-performing superensemble forecasts. At shorter leadtimes the Bayesian-postprocessed forecasts achieve broadly similar scores to the NGR forecasts, but at longer leadtimes, there is no clear improvement in CRPS or ES over the superensemble forecasts.

Figure 4.11: Distribution of continuous ranked probability score (CRPS) and its multivariate extension, the energy score (ES), at selected leadtimes over all 630 forecast instances for all 13 locations, for forecasts produced by different postprocessing methods.

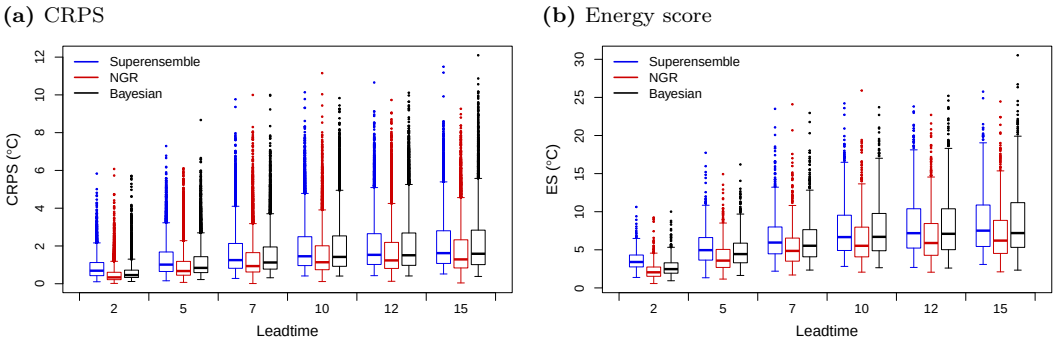
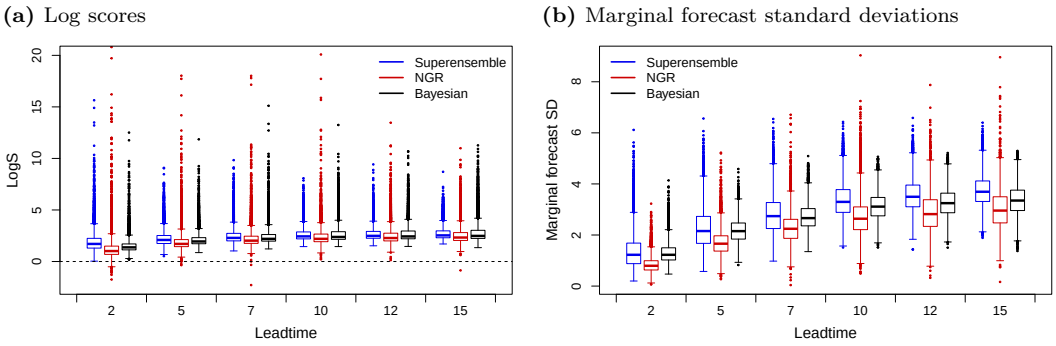


Figure 4.12a shows the distribution of logarithmic scores achieved by each postprocessing method; as noted in Section 3.2.2, this metric may be of more relevance than the CRPS in applications where the probability placed on the observed outcome is of primary interest. Again, the distribution of scores is generally fairly similar for the Bayesian and superensemble forecasts at all but the shortest leadtimes; while the NGR-postprocessed forecasts often achieve very low or even negative log scores, indicating a very high level of probability placed on the observed outcome, they also receive the highest log scores at most leadtimes. This is because, as Figure 4.12b shows, the NGR forecasts are generally rather sharper than the Bayesian posterior or superensemble forecasts at all leadtimes. The least confident Bayesian-postprocessed forecasts are somewhat sharper than the least confident superensemble forecasts at all leadtimes, but the difference in average sharpness is small, particularly at shorter leadtimes.

4.2.2 Marginal calibration

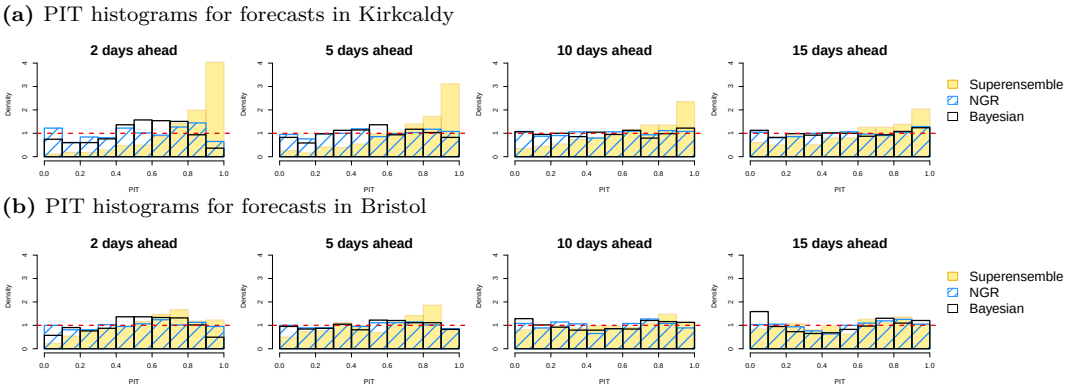
Inspection of the PIT histograms gives a more detailed understanding of the calibration of the postprocessed forecasts than the summary scores can. Figure 4.13 shows PIT histograms for forecasts in the cells containing Kirkcaldy and Bristol, which are typical of the histograms of other regions in, respectively, Scotland and northern England, and the south of the study area. PIT

Figure 4.12: Distribution of logarithmic scores (LogS) and marginal forecast standard deviations at selected leadtimes over all 630 forecast instances at all 13 locations, for forecasts produced by different postprocessing methods.



histograms for the other regions can be found in Figure B.2 in Appendix B.

Figure 4.13: PIT histograms showing the marginal calibration of forecasts of surface temperatures in selected regions, postprocessed using various methods. A 25-day moving window was used as a training set for the NGR and Bayesian postprocessors. The dashed line indicates the ideal uniform distribution.



The histograms for the uncorrected superensemble forecasts show significant negative skew in Scotland, reflecting the persistent substantial cold bias noted in Figure 4.9; the skewness in the histograms for Bristol is less pronounced, but still have a peak of values at around 0.7-0.9, reflecting the local bias in that region. As noted in Section 4.2.1, the NGR and Bayesian forecasts, which estimate a separate bias and calibration correction for each location, are able to remove this systematic regional bias almost completely, producing histograms that are much closer to uniformity than those of the superensemble forecasts. The slight residual bias manifests in a slight ∩-shape at around the

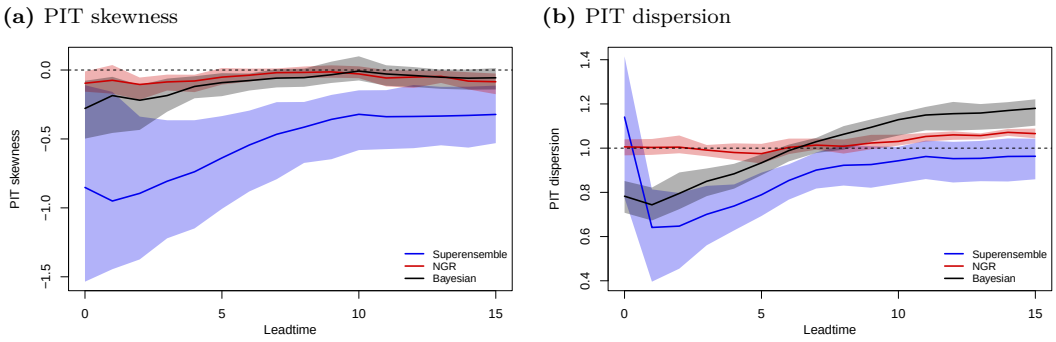
70th percentile of the transformed values, indicating that the bias is small with respect to the spread of the forecast errors, and is slightly more pronounced in the Bayesian histograms, reflecting the slightly higher accuracy of the NGR forecasts.

The histograms for the Bayesian and NGR forecasts are generally similar in shape, although at the shortest leadtimes the Bayesian forecasts have too few scores in the outermost bins, indicating that the forecasts are slightly overdispersive, or underconfident. The opposite is true of the forecasts at the longest leadtimes, when there are too many observations falling in the tails of the Bayesian forecasts, suggesting that they are somewhat underdispersive. Figure 4.14 summarises the histogram characteristics for each method at each leadtime: the shaded areas indicate the spread of values across all thirteen regional histograms for each method, with the lines indicating the mean values. The width of the blue bands reflects the widely differing skill of the superensemble forecasts; after postprocessing with either the NGR or Bayesian methods, the bands are much narrower, indicating that all regions have similar calibration. Figure 4.14b highlights the trend in dispersiveness of the Bayesian postprocessed forecasts: at shorter leadtimes the forecasts are somewhat overdispersive, while at longer leadtimes they are rather underdispersive. The NGR forecasts, on the other hand, have dispersion indices close to one at all leadtimes, indicating marginally well-calibrated forecasts.

4.2.3 Joint calibration

Joint calibration is assessed by considering the BDR and BOT histograms in Figures 4.15 and 4.16. The BDR histograms tend to reflect the overall marginal calibration, as discussed in Section 3.3.2.2. For the biased superensemble forecasts (shaded yellow in the histograms), this means that the BDR histograms have too many values in the leftmost bins, particularly at the shortest leadtimes. At the longest leadtimes, although the superensemble forecasts have regional biases of approximately the same magnitude as they did at leadtime 2, the BDR histograms are almost flat: this is because the overdispersiveness of the

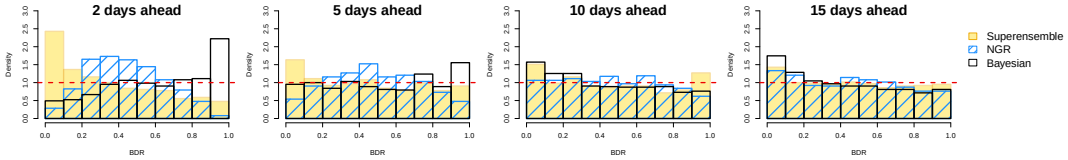
Figure 4.14: Characteristics of the PIT histograms for each postprocessing method at each leadtime. The lines indicate the mean value across the thirteen histograms, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



forecasts at these longer leadtimes is, to some extent, masking the bias.

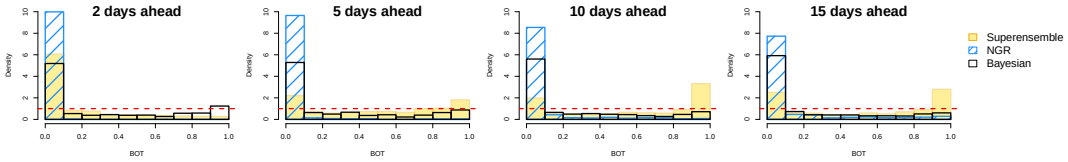
The shape of the BDR histograms for the less biased Bayesian forecasts (outlined in black) is dominated by the marginal dispersion characteristics, with the overdispersive forecasts at short leadtimes having too many values in the rightmost bins, and the underdispersive forecasts at longer leadtimes having too many values in the leftmost. For the NGR forecasts, which are marginally well calibrated, the BDR histograms are somewhat ∩-shaped at shorter leadtimes, indicating that the variables are too strongly correlated in the predictive distributions. At longer leadtimes this effect can barely be detected. However, the dependence structure of the NGR forecasts is determined by the correlations between the observations in the 25 days prior to the forecast issue date and is therefore independent of the forecast leadtime: if the forecast distributions at shorter leadtimes are over-correlated, the distributions at longer leadtimes must also be over-correlated. This suggests that the issue is masked at longer leadtimes by the slight underdispersiveness of the marginal forecasts, as seen in Figure 4.14b, highlighting the importance of considering multiple diagnostic histograms when diagnosing joint calibration. That the issue can be diagnosed at all from the BDR alone at the shortest leadtimes is partly due to the absence of marginal calibration issues, which would otherwise dominate the BDR; and partly due to the severity of the correlation misspecification.

Figure 4.15: Modified Bank Depth Rank (BDR) histograms showing the joint calibration of forecasts of surface temperatures using different postprocessing methods. The dashed line indicates the ideal uniform distribution.



This suspicion is supported by the BOT histograms in Figure 4.16; as noted in Section 3.3.2.1, the BOT histogram is known to be particularly sensitive to misspecifications of the predictive dependence structure and other distributional assumptions. The histograms for the NGR forecasts at all leadtimes are dominated by a large spike in the leftmost bin; in the absence of marginal calibration issues, this is indicative of too-strong correlations between variables in the predictive distributions. The Bayesian-postprocessed forecasts have a smaller spike in the leftmost bin at all leadtimes; as discussed in Section 4.1.3, there is some evidence that the correlations in the Bayesian predictive distributions are too strong, but to a lesser extent than those used to construct the multivariate NGR forecasts.

Figure 4.16: Box Ordinate Transform (BOT) histograms showing the joint calibration of forecasts of surface temperatures using different postprocessing methods. The dashed line indicates the ideal uniform distribution.

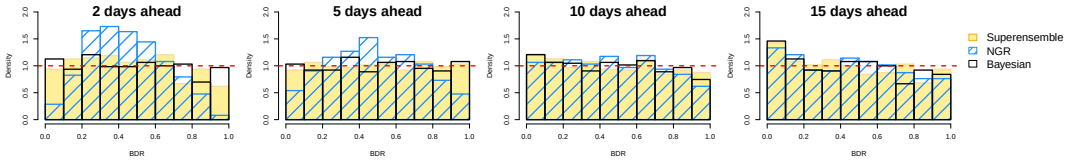


With the exception of the shortest leadtimes – where the histograms are skewed as a result of the marginal biases and overdispersiveness – the BOT histograms for the superensemble forecasts are ∪-shaped, with spikes in both the leftmost and rightmost bins. Given that these forecasts are already known to have substantial marginal biases and to be marginally overdispersive, it would be hard to justify concluding that this shape has arisen from underestimation of the correlations in the forecast distributions, rather than from the combined

effect of these two marginal issues.

It is, however, possible to separate the effect of the dependence structure from the effect of marginal calibration using (2.5), by combining the marginal NGR predictive distributions with correlation matrices $\mathbf{R}$ taken from the corresponding superensemble and Bayesian postprocessed forecasts to obtain hybrid forecast distributions for each forecast instance. BDR histograms for these hybrid forecasts – in which all three forecasts have identical marginal calibration, and differ only in terms of their joint calibration – are shown in Figure 4.17. The corresponding BOT histograms are, like those in Figure 4.16, too heavily skewed to be easily interpreted, and so are not shown. The BDR histograms produced by the hybrid forecasts using the superensemble and Bayesian dependence structure are almost flat at all but the longest leadtime, indicating improved joint calibration; at the shortest leadtimes, the superensemble hybrids produce slightly peaked histograms, but the difference in joint calibration between the superensemble and Bayesian hybrids is very small. At the longest leadtimes, where the marginal NGR forecasts are slightly underdispersive, the histogram is slightly skewed for all three sets of forecasts. The original NGR forecasts took their dependence structures from the verifying observations in the training set, as described in Section 2.1.4.1.2: this method of estimating correlations is known as the Schaake Shuffle (Clark et al., 2004). In contrast, the superensemble forecasts take their correlation structure from the MME forecast ensemble members, and so use a form of ensemble copula coupling, as described in Section 2.1.4.1.2; the dependence structure of the Bayesian postprocessed forecasts is derived partly from the correlations within the raw ensemble, through $\mathbf{\Sigma_D}$, and partly from the correlations between the forecast errors in the training set, through $\mathbf{\Lambda}$. Both of these methods of estimating the dependence structure improve on the joint calibration obtained by using the Schaake Shuffle, which produced forecasts that are jointly too sharp.

Figure 4.17: Modified Bank Depth Rank (BDR) histograms showing the joint calibration of hybrid forecasts of surface temperatures with identical marginal calibration, but using different postprocessing methods to estimate the dependence structure. The dashed line indicates the ideal uniform distribution.



4.2.4 Summary

The Bayesian postprocessing method is able to remove the regional biases at all leadtimes, and improves on the accuracy and forecast skill of the superensemble forecasts at leadtimes of up to nine days. However, at the longest leadtimes the Bayesian forecasts have similar accuracy, and a similar spread of errors, to the superensemble forecasts. This may be because the Bayesian bias correction simply shifts the consensus mean by $\boldsymbol{\eta}$, the mean discrepancy between the forecasts verified in the 25 days prior to the current instance being issued and their verifying observations. As the leadtime increases, the relevance of those 25 training cases to the present instance decreases, and so the accuracy of the resulting forecasts is lessened, although the overall bias is still well accounted for. Choosing a training set that is tailored to the forecast of interest may be expected to improve performance when using the Bayesian postprocessing method; two approaches to doing so are tested in the next section.

The NGR forecasts are both more accurate and sharper than the Bayesian postprocessed forecasts, despite using the same training data as the Bayesian correction, achieving a fairly constant improvement of 0.6°C in accuracy and around 0.3°C in CRPS over the superensemble forecasts at all leadtimes. This can partly be ascribed to the fact that the NGR forecasts were fitted by minimising the CRPS, a method known to maximise the accuracy of the forecasts as described in Section 3.2.1; but the fact that the NGR forecasts achieve a consistent level of improvement at even the longest leadtimes is likely due to the fact that the NGR approach is based on a model of the relationship

between the forecasts and the observed temperatures. Even if the average error in the training set is not representative of the error in the forecast of interest, the relationship between the forecasts and the observations is evidently stable enough to produce consistent improvements. Further work is required to understand how this relationship might be exploited within the Bayesian postprocessing framework.

One difference between the Bayesian and NGR postprocessing methods not reported in the results above is the difference in computational cost. NGR requires numerical optimisation and is therefore relatively slow: marginally postprocessing the full dataset of 630 forecasts of temperatures at 13 locations at each of 16 leadtimes using the `ensembleMOS` package in R took around 21 minutes, while jointly postprocessing the same forecasts using the Bayesian framework (including estimation of the discrepancy for each forecast instance) took around two minutes. The NGR predictive distribution is estimated independently at each location, so this computational cost will increase directly in proportion to the dimension of the data. While the time taken to carry out the Bayesian postprocessing also increases linearly with the dimension of the data, experiments indicate that the increase is of the order of 4-5 seconds per additional dimension, suggesting that the Bayesian method will scale much more readily to the problem of postprocessing higher-dimensional forecasts.

4.3 Effect of training set selection

In this section, the effect of the choice of training set on the performance of the Bayesian postprocessed forecasts is considered. Each forecast instance is postprocessed using a training set consisting of 25 forecast-observation pairs selected by one of several methods, and using a zero-precision noninformative prior. As a baseline, the moving-window (MW) training cases already discussed in Sections 4.1 and 4.2 are used; the skill of these forecasts is compared to that of training sets chosen according to the two analogue selection methods described in Section 2.3.3. Direct analogues (DA) are those that have the shortest

Euclidean distance to the forecast of interest in normalised temperature space, as described in Section 2.3.3.1, while weather regime (WR) analogues are the most similar to the forecast of interest in terms of the principal components of their MSLP fields, as described in Section 2.3.3.2.

To maximise the usefulness of the relatively short (7-year) available forecast archive, analogues are selected using a modified cross-validation approach (Wilks, 2011), rather than from past candidates in the strictly chronological sense. For each instance, candidates for the current year are excluded from the search, with the exception of the 25 days immediately preceding the forecast issue date; each method (including the moving-window approach) therefore has access to candidates drawn from 6 winters, plus 25 days immediately prior to the date on which the forecast was actually issued, ensuring parity between the three training sets.

Each selection method was found to produce quite different collections of forecast-observation pairs, with training sets having on average around 2-3 members in common with their counterparts.

4.3.1 Forecast accuracy, sharpness and skill

All three training sets produce forecasts with very similar error distributions, as Figure 4.18a shows; at the shortest leadtimes, the forecasts postprocessed using weather regime analogues have slightly higher MAE than those using moving window or direct analogue training sets, although the difference is very small. However, at longer leadtimes – after around 9 days – both sets of analogue-postprocessed forecasts are more accurate, on average, than those using recent forecasts to estimate the required correction. This trend is directly reflected in both the CRPS and energy scores, shown in Figure 4.19, and the log scores, shown in Figure 4.20a. All three sets of postprocessed forecasts have similar sharpness at shorter leadtimes (Figure 4.20b); at longer leadtimes, both sets of analogue-postprocessed forecasts are less sharp, on average, than those using a moving window to estimate $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$.

Figure 4.18: Distribution of forecast errors and MAE at selected leadtimes over all 630 forecast instances for all 13 locations when forecasts are postprocessed with different training sets.

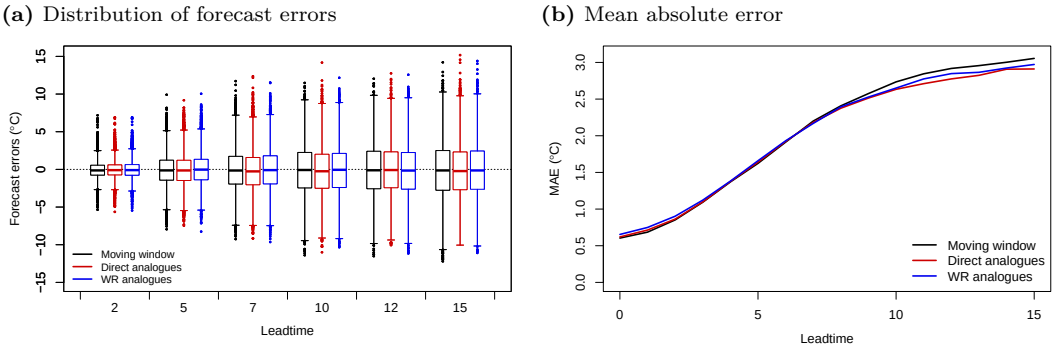


Figure 4.19: Average continuous ranked probability score (CRPS) and energy score (ES) over all 630 forecast instances at all 13 locations when forecasts are postprocessed with different training sets.

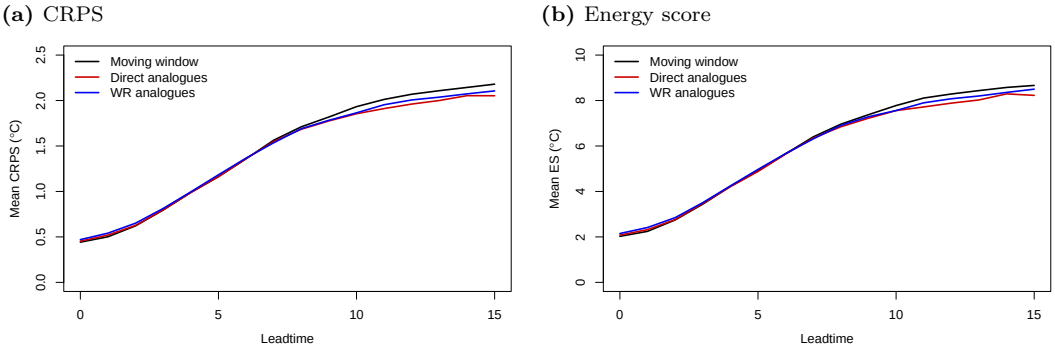
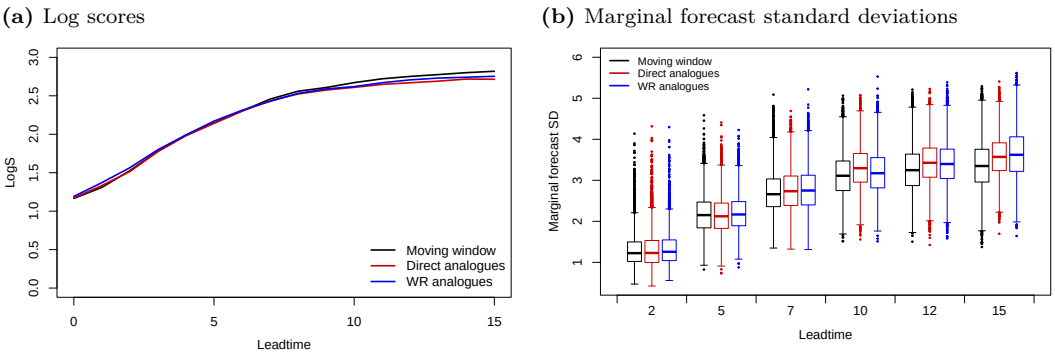


Figure 4.20: Mean of logarithmic scores (LogS) and distribution of marginal forecast standard deviations at selected leadtimes over all 630 forecast instances at all 13 locations, for forecasts postprocessed with different training sets.



4.3.2 Marginal calibration

PIT histograms for forecasts using the three training sets are shown for selected regions in Figure 4.21, and for the remaining study regions in Figure B.3 in

Appendix B. Given the similarity of the skill scores obtained by the three sets of postprocessed forecasts, it is unsurprising that the corresponding PIT histograms are, for the most part, also similar in shape; however, in regions in southern and central England, the histograms for forecasts postprocessed using moving-window training cases tend to have a small spike of values in the leftmost bin at the longest leadtimes, which does not appear in the histograms for the forecasts using analogues to estimate the discrepancy.

The skewness and dispersion characteristics of the histograms for each method are summarised in Figure 4.22. All three methods have slight negative skewness – reflecting a very small residual cold bias, as noted in Section 4.1.2 – but have close to zero skewness at all but the shortest leadtimes. The PITs of forecasts postprocessed using weather regime analogues are generally more symmetrically distributed than those of using a moving window at the shortest leadtimes and than those using direct analogues at leadtimes of five to ten days, but the difference is small. At leadtimes of up to six days, the three training sets all produce histograms with similar dispersion indices, with the predictive distributions tending to overdispersion. At longer leadtimes, the dispersion indices of forecasts postprocessed using a moving-window training set continue to increase with leadtime as the forecasts become more and more underdispersive: those using analogues are closer to one, indicating improved marginal calibration, with the forecasts using direct analogues achieving the best calibration overall.

4.3.3 Joint calibration

Like the PIT histograms, the BDR histograms for the three sets of forecasts are broadly similar in shape (Figure 4.23). At all leadtimes, the forecasts postprocessed using analogues to estimate the discrepancy are slightly closer to the ideal uniform shape than those using a moving window to estimate the discrepancy, reflecting the improved marginal calibration noted in Section 4.3.2: at shorter leadtimes, the forecasts using weather-regime analogues in particular have fewer values in the rightmost bin, reflecting the small improvement to

Figure 4.21: PIT histograms showing the marginal calibration of forecasts of surface temperatures in selected regions, postprocessed using various training sets to estimate the discrepancy. The dashed line indicates the ideal uniform distribution.

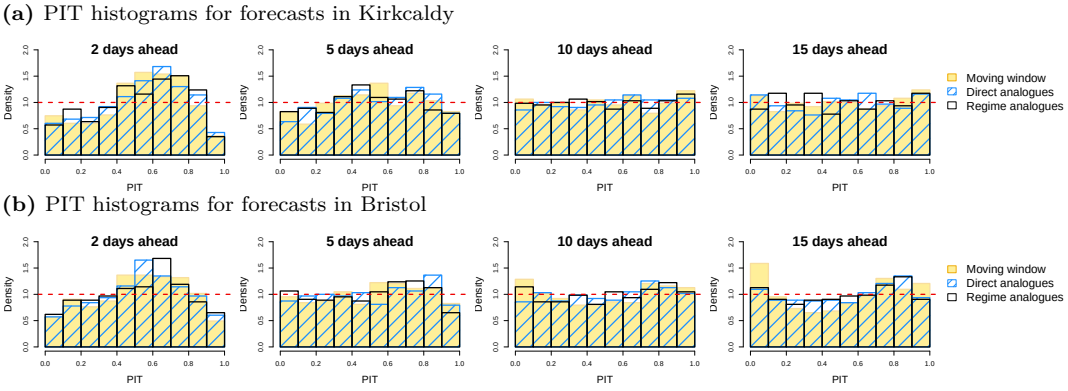
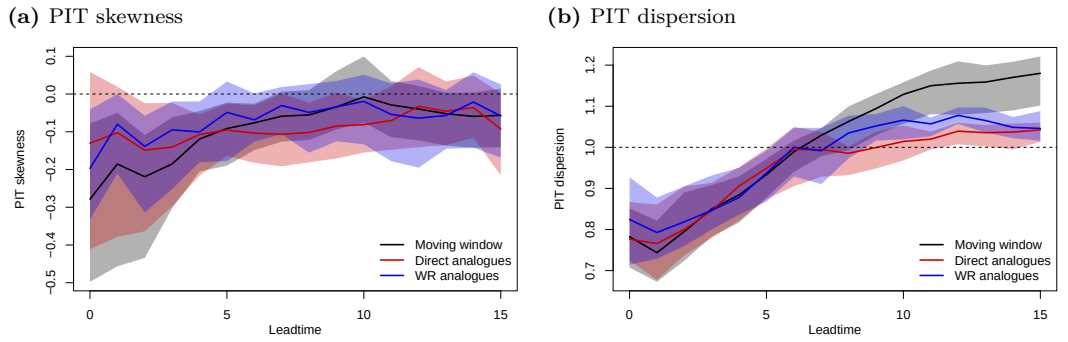
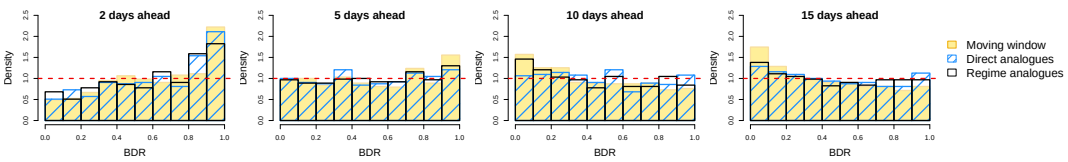


Figure 4.22: Characteristics of the PIT histograms at each leadtime for forecasts postprocessed using different training sets. The lines indicate the mean value across the thirteen histograms, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



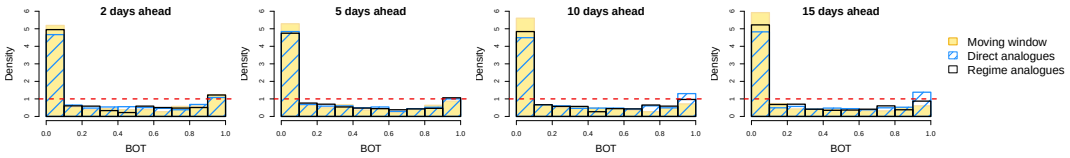
the bias correction; and at longer leadtimes, the histograms for both sets of analogue-corrected forecasts have fewer values in the leftmost bin at longer leadtimes than those of the moving-window-corrected forecasts, reflecting the improved marginal dispersion.

Figure 4.23: Modified Bank Depth Rank (BDR) histograms showing the joint calibration of forecasts of surface temperatures postprocessed using different training sets. The dashed line indicates the ideal uniform distribution.



The BOT histograms for forecasts postprocessed using all three training sets remain very skewed, with a large spike in the leftmost bin indicating a high proportion of the observations falling in regions of very low predicted probability (Figure 4.24); this spike is slightly smaller for both of the analogue-corrected sets of forecasts at longer leadtimes, with the direct analogues producing the least poorly calibrated forecasts according to this metric.

Figure 4.24: Box Ordinate Transform (BOT) histograms showing the joint calibration of forecasts of surface temperatures postprocessed using different training sets. The dashed line indicates the ideal uniform distribution.



In Figure 4.8 , it was noted that gridplots showing the joint distribution of the BOT and BDR of the forecasts using a noninformative prior and a discrepancy estimated using training cases from a moving window showed two clusters of points: one in the leftmost column, with BDRs between 0.3 and 0.5, indicating forecasts distributions specifying too-strong correlations between the variables; and a second in the top-right corner of the gridplots, indicating a group of very overdispersive forecasts. The same gridplots are presented alongside their equivalents for the analogue-corrected forecasts in Figures 4.25 and 4.26. The two clusters are still apparent in the gridplots for the analogue-postprocessed forecasts; for the direct analogues, there is still a peak between 0.3 and 0.5, but for the weather regime analogues, the values are more evenly distributed across all of the bins of the BDR histograms. This suggests that postprocessing using analogues chosen according to the similarity of the prevailing weather patterns results in forecasts with slightly lower correlations between the variables – and therefore slightly improved joint correlation – than using direct analogues.

Figure 4.25: Gridplots summarising the joint distribution of the BOT and BDR histogram counts for forecasts postprocessed using different training sets at leadtime 2.

(a) Moving window **(b)** Direct analogues **(c)** Weather regime analogues

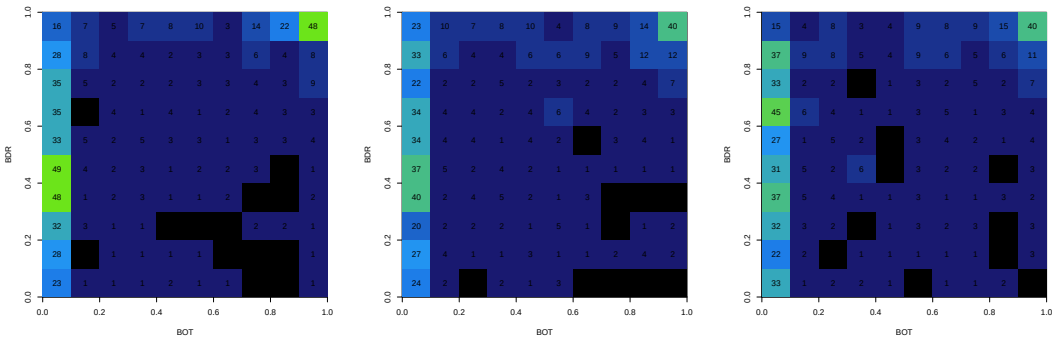
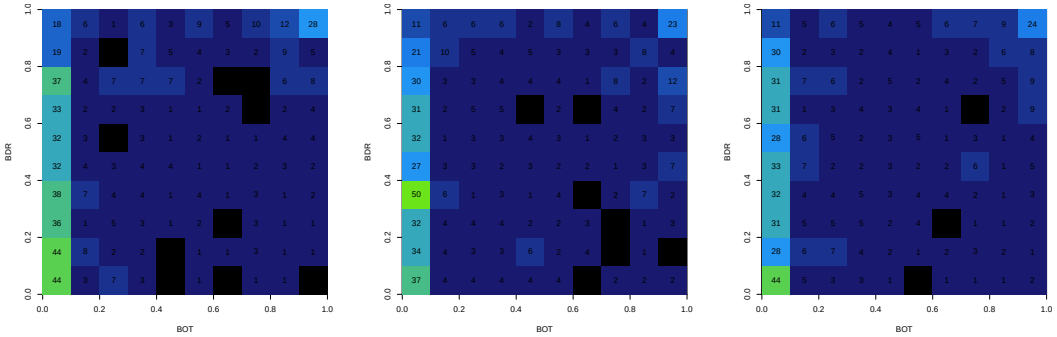


Figure 4.26: Gridplots summarising the joint distribution of the BOT and BDR histogram counts for forecasts postprocessed using different training sets at leadtime 5.

(a) Moving window **(b)** Direct analogues **(c)** Weather regime analogues



4.3.4 Summary

The method by which training cases for the estimation of the forecast discrepancy were selected has a relatively small effect on the forecast skill in this study. At the shortest leadtimes, marginal forecasts using all three methods achieved a similar level of skill under all of the metrics considered, while the analogue-postprocessed forecasts showed small improvements in joint calibration. For forecasts issued more than six days in advance, using either of the analogue methods to select training cases produced forecasts with better marginal calibration than the moving-window training cases, being both slightly more accurate and less underdispersive at long leadtimes. However, the difference in skill between the three methods remains small.

The lack of any substantial improvement in forecast skill when using analogues to estimate the discrepancy may be due to the use of an insufficiently long archive of candidate forecasts from which analogues could be selected. Due to constraints on data availability (ECMWF, 2021e), the dataset used for the application contains only seven years of data, of which only six were used as potential candidates for each forecast instance, as described in Section 4.3. Even with the inclusion of the 25 'moving window' candidates candidates, this offers only 565 candidates for each forecast instance; as noted by Hu et al. (2020), even the best analogues selected from such a short archive of candidates will tend to be mediocre. This problem is compounded by the fact that the forecasts are provided by a multi-model ensemble: candidate forecasts should be issued by the same model configuration as the forecast of interest, but the forecasts in the TIGGE archive used in this study are issued by operational NWP models, which underwent (sometimes substantial) updates during the study period. This will further reduce the effective size of the pool of suitable candidates for each forecast instance.

Model changes may have a greater impact on the quality of the weather regime analogues than on that of the direct analogues. This is because model changes are more likely to affect the way in which surface weather quantities like the temperature – which are known to be sensitive to the precise model formulation – than pressure fields, which form part of the core of the model and are governed by physical processes that are generally represented by well-established model physics. Thus, if a model undergoes a revision, the pressure fields used to select weather regime analogues may be largely unaffected: but the relationship between those pressure fields and the temperature predicted by the model may be changed. In this situation, a weather regime analogue produced by a different model version is likely to provide a poor estimate of the error in the forecast of interest, even if the pressure fields are almost identical. When using analogues chosen on the basis of similarity in terms of the predicted temperatures, on the other hand, the effective size of the archive

of suitable candidates will be reduced, as candidates produced by a different model version are likely to be filtered out as poor matches during the selection process. However, this approach will still identify relevant analogues to the current instance – albeit from a reduced pool of candidates – while selection by weather regime analogues may select training cases with entirely different error characteristics. Further work is planned to test whether postprocessing using weather regime analogues might be more successful when using an archive of true reforecasts, perhaps using the subseasonal-to-seasonal forecasts produced by the S2S project (Vitart et al., 2017).

Despite the lack of a long archive of reforecasts from which to draw analogues, both of the analogue selection methods tested here produced forecasts with skill comparable to or slightly better than the moving window training sets; furthermore, using weather regime analogues produced forecasts with skill comparable to that of forecasts postprocessed using analogues selected using the established method. The most appropriate approach to use is therefore likely to depend on the application. The moving-window training set is convenient to obtain, requiring no additional archive of candidate forecasts from which to select analogues; for the postprocessing of forecasts in a relatively small area, at leadtimes of less than a week, it most likely remains the best choice. However, if the forecast area is increased, a larger training set will be required for estimation of the necessary covariance matrices: simple increasing the length of the moving window used to select the training cases will eventually result in a reduction in forecast skill, due to the increasing remoteness of the training cases from the forecast of interest. Similarly, when postprocessing forecasts over a larger region, selection of direct analogues is likely to produce lower-quality training cases due to the corresponding increase in the dimension of the candidate vectors, as discussed in Section 2.3.3.2; this is not a problem when selecting analogues using the proposed weather regime method, because the dimension of the candidate search space remains small, regardless of the dimension of the forecast vectors themselves. Perhaps most usefully, if the

forecast domain changes slightly (within the bounds of the region to which the PCA was applied), there is no need to identify new analogues; forecasts postprocessed using weather regime analogues will not be changed substantially if they are recalculated as part of a different geographical domain, or with additional weather quantities incorporated in the forecast vector.

4.4 Summary of chapter

One of the potential benefits of the Bayesian postprocessing framework is the scope to include an informative prior. However, in Section 4.1, forecasts using a zero-precision noninformative prior were found to be more skilful than those using a climatological prior, or a sequential postprocessing approach. The sequential forecasts performed particularly poorly despite being initialised with a relatively well-calibrated forecast, suggesting that substantial modifications to the framework may be needed in order to successfully incorporate information from a previous forecast. The development of a framework to produce correctly calibrated sequences of forecasts is planned for future work.

In Section 4.2, the skill of forecasts postprocessed using the new Bayesian approach was compared to that of forecasts corrected using nonhomogeneous Gaussian regression (NGR). While the Bayesian method was able to outperform the raw 'superensemble' forecasts at shorter leadtimes, the NGR forecasts were both sharper and more accurate at all leadtimes, with better marginal calibration. A particular problem for the Bayesian forecasts was marginal calibration, with forecasts at shorter leadtimes being consistently overdispersive, and forecasts at longer leadtimes being consistently underdispersive. However, when combined with the marginal NGR forecasts, the correlation matrices produced by the Bayesian postprocessing method resulted in better joint calibration than matrices estimated using either the Schaake Shuffle or ensemble copula coupling methods described in Section 2.1.4.1.2. Moreover, the Bayesian postprocessed forecasts are computationally much cheaper than the NGR forecasts, taking around one-tenth of the time that NGR required to postprocess

the full dataset.

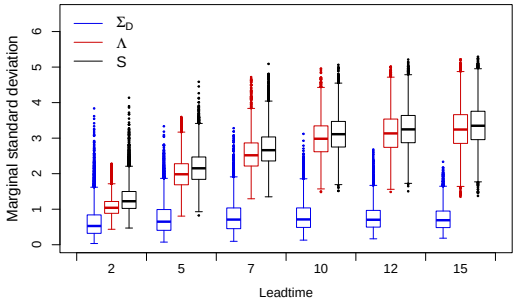
Finally, in Section 4.3, forecasts postprocessed using weather regime analogues, selected using the new approach proposed in Section 2.3.3.2, were found to have predictive skill comparable to or better than those using the established a moving-window or direct analogue approaches. Selection of training cases using the new approach may be particularly beneficial where a large number of locations or variables are to be postprocessed simultaneously, in which case the dimension of the candidate search space for direct analogues may become too large to reliably identify high-quality analogues. It should also be noted that forecasts postprocessed using analogues had skill comparable to those using a moving window approach, despite the fact that the archive of candidates consisted of past forecasts rather than reforecasts; the two analogue methods may be expected to perform better when an archive of true reforecasts, produced by the same model configuration as the forecast instances, is available. Further work is planned to investigate whether this is the case using forecasts and reforecasts from the S2S database (Vitart et al., 2017).

From (2.21), when a noninformative prior is used, the posterior variance reduces to the sum of $\boldsymbol{\Sigma_D}$, the uncertainty about the position of the unobserved MME consensus, and $\boldsymbol{\Lambda}$, the covariance matrix of the discrepancy. Figure 4.27 shows the contributions of these two quantities to the posterior variance $\mathbf{S}$ for forecasts postprocessed using the Bayesian framework with a noninformative prior and moving-window training cases.

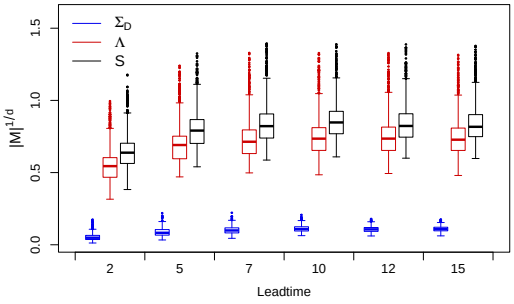
The dominant contribution to the posterior variance $\mathbf{S}$ is from $\boldsymbol{\Lambda}$, which increases steadily with leadtime, while the average sharpness of $\boldsymbol{\Sigma_D}$ remains constant at the longest leadtimes, as the ensemble forecasts all tend towards a common climatology. It is very likely that if a better estimate of the discrepancy uncertainty $\boldsymbol{\Lambda}$ can be obtained, the overall calibration of the Bayesian postprocessed forecasts will be improved, particularly at longer leadtimes; likewise, an improved estimate of $\boldsymbol{\eta}$ would result in higher forecast accuracy, and therefore overall skill. A method to improve the estimation of both $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$, using a

Figure 4.27: Distribution of all regional marginal standard deviations and determinant sharpnesses of $\mathbf{\Sigma_D}$, the uncertainty about the position of the unobserved MME consensus, $\mathbf{\Lambda}$, the covariance matrix of the discrepancy, and $\mathbf{S}$, the posterior covariance matrix, for forecasts postprocessed using the Bayesian framework, with a noninformative prior and discrepancy estimated using training cases from a moving window of 25 days.

(a) Marginal standard deviations



(b) Determinant sharpness



linear approximation to Bayesian inference, is proposed in the next chapter.

Chapter 5

Multivariate Bayes linear adjustment of the forecast discrepancy $\boldsymbol{\Delta}$

5.1 Motivation

The skill of the forecasts produced by the Bayesian framework depends on how well the discrepancy $\boldsymbol{\Delta}$ is estimated. The implementation described in Chapter 4 and Barnes et al. (2019) uses a simple moment-based approach, taking the sample mean of a set of training forecast errors as the estimate of the expected discrepancy $\boldsymbol{\eta}$ and the sample variance-covariance matrix of the same forecast errors as the estimate of the variance of the discrepancy, $\boldsymbol{\Lambda}$.

It seems likely that the estimate of the discrepancy might be improved – and hence the forecasts made more skilful – by applying a more sophisticated approach. In Section 4.3, corrections using errors from recently-verified forecasts tended to perform well at the shortest leadtimes; corrections using the errors from analogue forecasts, with local or synoptic weather regimes – and therefore errors – similar to those of the forecast to be postprocessed, tend to do better at longer leadtimes. This suggests that some advantage may be gained, at least at some leadtimes, by somehow combining the two different estimates of the forecast error.

Aside from this point, there is some evidence that a distribution with heavier tails may better represent the distribution of observed surface temperatures than the Gaussian model that is commonly used by default. Gneiting et al. (2005) remark that fitting the NGR parameters via minimisation of the CRPS is preferable to maximising the likelihood because the latter approach tends to favour overdispersive forecast PDFs, due to the greater sensitivity of the log score (LS) to outliers – although, as noted in Section 3.2.1, the CRPS tends to favour underdispersive forecasts, rather than well-calibrated ones. However, as Gebetsberger et al. (2018) point out, both the CRPS and LS should produce consistent estimators of the parameters – and therefore very similar regression coefficients – as long as the parametric form of the distribution is correctly specified. They argue that the very fact that NGR coefficients fitted using CRPS minimisation tend to differ substantially from those fitted using LS minimisation indicates that the Gaussian model is in fact not appropriate. In that paper, heavier-tailed parametric forms – both Student-t and logistic distributions – were generally found to produce similar parameter estimates when fitted by minimising either the LS or CRPS, while coefficients fitted using a Gaussian regression differed. Forecasts fitted using nonhomogeneous Student-t regression also had better calibration than those using nonhomogeneous Gaussian regression.

A natural way to combine two sources of information and obtain a predictive t distribution with slightly heaver tails than those of a normal distribution would be to perform Bayesian inference on the forecast error. The simplest approach would be to specify a normal-inverse-Wishart joint distribution for the expectation and covariance of $\boldsymbol{\Delta}$, and to infer the distribution of the discrepancy after observing a second sample of forecast errors which are also assumed to have a Gaussian distribution; this would result in a Student-t posterior density for the discrepancy mean vector $\boldsymbol{\eta}$ and an inverse-Wishart posterior density for the discrepancy covariance matrix $\boldsymbol{\Lambda}$, satisfying the requirement that the forecast discrepancy be assumed to have heavier tails than a normal distribution.

However, this leads to a new problem. The Bayesian framework described in Section 2 produces simple closed-form expressions for the postprocessed mean and covariance of the forecast because the ensemble forecasts, the multi-ensemble system as a whole, and the forecast discrepancy are all assumed to be approximately normally distributed. This means that all of the information that they provide can be captured by those two quantities; because the normal distribution is closed under summation, the mean and covariance of the postprocessed distribution can be calculated analytically. However, there is no simple way to compute the expectation and variance of the sum of a normal distribution and a t-distribution, so computationally intensive Monte Carlo techniques would be required to approximate the distribution of the posterior forecast. A further consideration is that, unless the sample size and the degrees of freedom of the Wishart prior used in the inference are very small, the degrees of freedom of the posterior t-distribution will be large enough that the density will closely approximate a normal distribution anyway, and the benefit of the heavy-tailedness will be minimal.

The approach proposed here makes use of Bayes linear statistics, which represents quantities of interest through their expectations and variances rather than full probabilistic specifications, and so produces adjusted expectations and variances that can be plugged directly into the Gaussian framework described in Section 2.2; the proposed method is also able to accommodate an assumption of heavier (or indeed lighter) tails in the observed forecast errors. The Bayes linear adjustment is carried out in two stages: first, the prior expectation and variance of the population covariance matrix are adjusted by the observed sample covariance matrix; this updated assessment of the variances is then used in the adjustment of the prior expectation and variance of the population mean by the observed sample mean.

Bayes linear adjustment of covariance matrices is typically presented in terms of geometric projection of matrices into a Hilbert space (Wilkinson, 1995; Goldstein and Wilkinson, 2001; Williamson et al., 2012); while this approach

allows a great deal of flexibility, it may greatly increase the complexity and cost of both specifying and computing the required quantities. Goldstein and Wooff (2007) suggest two simple heuristics as an alternative: either the marginal variances should first be adjusted independently, then combined with an estimate of the dependence structure in order to obtain a 'semi-adjusted' residual variance matrix; or the prior and observed covariance matrices should be weighted and averaged directly. This chapter proposes an alternative approach, extending the second-order exchangeability representation underpinning much of Bayes linear statistics to derive a fully multivariate adjustment of the covariance matrix, specification of which requires little more effort than these heuristic approaches.

Section 5.2 reviews the concepts of Bayes linear statistics, and existing approaches to Bayes linear adjustment of covariance matrices. The derivation of the multivariate adjustment in Section 5.3 depends on several special matrix operations and identities, which are reviewed in Appendix C.

The accuracy of the Bayes linear approximation to natural conjugate inference is examined in Section 5.4. Practical issues concerning implementation of the Bayes linear adjustment of the forecast discrepancy are discussed in Chapter 6, where an application to postprocessing forecasts of winter surface temperatures is presented.

5.2 Review of Bayes linear statistics

Unlike classical Bayesian statistics, Bayes linear statistics does not represent underlying populations in terms of probability distributions: instead, models are constructed from expectations of quantities that are judged to be exchangeable. Uncertainty about the values of quantities of interest is expressed in terms of mean vectors and variance-covariance matrices, which are specified according to the user's subjective beliefs and updated as additional data is observed.

Suppose that two quantities of interest, $\mathbf{X}$ and $\mathbf{Y}$, are to be observed in that order at two different time points, and that learning about $\mathbf{Y}$ is expected to

provide some information about $\mathbf{X}$. Bayes linear analysis begins by specifying prior beliefs about the expectation $\mathbb{E}[\mathbf{Y}]$ and variance $\mathbb{V}[\mathbf{Y}]$ of the random variable $\mathbf{Y}$; the expectation $\mathbb{E}[\mathbf{X}]$ and variance $\mathbb{V}[\mathbf{X}]$ of the random variable $\mathbf{X}$; and $\mathbb{C}[\mathbf{X}, \mathbf{Y}]$, the covariance between $\mathbf{X}$ and $\mathbf{Y}$. As in standard Bayesian inference, these prior beliefs should be specified using information that is independent of the observations that will be used to update them.

The relationship between $\mathbf{X}$ and $\mathbf{Y}$ can now be exploited to obtain an improved forecast for $\mathbf{X}$. In Bayes linear statistics this is done by constructing a linear estimate for each element of $\mathbf{X}$ from the elements of $\mathbf{Y}$, of the form $c_0 + \sum_k c_k Y_k$, and by choosing the coefficients $c_0^*, ..., c_k^*$ that minimise the prior expected squared error loss $\mathbb{E}\Big[\mathbb{E}[X_i] - (c_0 + \sum_k c_k Y_k)\Big]^2$. The estimator

$$\mathbb{E}_{\mathbf{Y}}[X_i] = c_0^* + \sum_k c_k^* Y_k \tag{5.1}$$

is known as the adjusted expectation for X_i given $\mathbf{Y}$; Goldstein and Wooff (2007, §1.4.5) show that the adjusted expectation of $\mathbf{X}$ given $\mathbf{Y}$ is

$$\mathbb{E}_{\mathbf{Y}}[\mathbf{X}] = \mathbb{E}[\mathbf{X}] + \mathbb{C}[\mathbf{X}, \mathbf{Y}]\, \mathbb{V}[\mathbf{Y}]^{-1} (\mathbf{Y} - \mathbb{E}[\mathbf{Y}]). \tag{5.2}$$

If $\mathbb{V}[\mathbf{Y}]^{-1}$ has been specified in such a way that it is singular, the Moore-Penrose generalised inverse should be used; following Goldstein and Wooff (2007, §1.4.11), no distinction is made here between the handling of full rank and singular variance matrices.

The residual variance not explained by the observed $\mathbf{Y}$ is referred to as the adjusted variance of $\mathbf{X}$ given $\mathbf{Y}$, which can be shown to be

$$\mathbb{V}_{\mathbf{Y}}[\mathbf{X}] = \mathbb{V}[\mathbf{X}] - \mathbb{C}[\mathbf{X}, \mathbf{Y}]\, \mathbb{V}[\mathbf{Y}]^{-1}\, \mathbb{C}[\mathbf{Y}, \mathbf{X}]. \tag{5.3}$$

Goldstein and Wooff (2007, §1.4.5) show that $\mathbb{V}_{\mathbf{Y}}[\mathbf{X}]$ is the mean squared error of the estimator $\mathbb{E}_{\mathbf{Y}}[\mathbf{X}]$, and can therefore be most usefully interpreted as a linear estimate of how much of the stated prior uncertainty about $\mathbf{X}$ remains

after the variability accounted for by $\mathbf{Y}$ is removed.

The adjusted expectation and variance of $\mathbf{X}$ are identical to the conditional mean and variance of a partition $\mathbf{Y}$ of a joint multivariate normal distribution for $\mathbf{Y}$ and $\mathbf{X}$, given a partition $\mathbf{Y}$ (Gelman et al. (2013), see also Krzanowski (2000), §7.2) – however, the Bayes linear framework can be applied and interpreted without reliance on any Gaussian probabilistic assumptions.

5.2.1 Second-order exchangeability

The intention in the proposed application is not to use observations of one quantity to predict another, as in the general approach outlined above; instead, observations of the multivariate forecast discrepancy will be used to adjust the prior expectation $\boldsymbol{\eta}$ and variance $\boldsymbol{\Lambda}$ of the forecast discrepancy, in order to obtain an improved estimate to be used in forecast postprocessing. Henceforth, $\mathbf{X}$ will be used to denote the quantity of interest in order to focus on the theoretical developments rather than the application.

Most statistical analysis proceeds from the assumption that a collection of observations $\{\mathbf{x}\}$ are exchangeable – meaning that the relationships between them can be expressed as a joint probability density $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ that is invariant to the ordering of the $\mathbf{x}_i$ (Gelman et al., 2013). Under the Bayes linear framework, full probability distributions are not specified for the underlying population, only expectations and variances; so only the slightly weaker assumption of second-order exchangeability is asserted. Following Goldstein and Wooff (2007, §6.4), a collection of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ is defined as second-order exchangeable if the first-and second-order belief specifications for the sequence of vectors are unaffected by any permutation of the order of the vectors. This means that all individuals share the same mean vector and variance matrix, and that the covariance matrix between any two different individuals is the same. This leads to the following representation theorem for second-order exchangeable random vectors.

5.2.1.1 Exchangeability representation for an infinite sequence of second-order exchangeable random vectors

If $\{\mathbf{X}_1, \mathbf{X}_2, ...\}$ is an infinite second-order exchangeable sequence of random vectors with common mean $\mathbb{E}[\mathbf{X}_i] = \boldsymbol{\mu}$ and common variance $\mathbb{V}[\mathbf{X}_i] = \boldsymbol{\Sigma}$ then $\{\mathbf{X}\}$ may be expressed in terms of the population mean vector $\mathcal{M}(\mathbf{X})$ and the infinite sequence of individual residual vectors $\{\mathcal{R}_1(\mathbf{X}), \mathcal{R}_2(\mathbf{X}), ...\}$, which satisfy the following properties.

1. For each individual i,

$$\mathbf{X}_i = \mathcal{M}(\mathbf{X}) + \mathcal{R}_i(\mathbf{X}). \tag{5.4}$$

2. The mean and variance for $\mathcal{M}(\mathbf{X})$ are

$$\mathbb{E}[\mathcal{M}(\mathbf{X})] = \boldsymbol{\mu}, \qquad \mathbb{V}[\mathcal{M}(\mathbf{X})] = \boldsymbol{\Gamma}. \tag{5.5}$$

3. The collection $\{\mathcal{R}_1(\mathbf{X}), \mathcal{R}_2(\mathbf{X}), ...\}$ is second-order exchangeable, with each individual i having

$$\mathbb{E}[\mathcal{R}_i(\mathbf{X})] = \mathbf{0}, \quad \mathbb{V}[\mathcal{R}_i(\mathbf{X})] = \boldsymbol{\Sigma} - \boldsymbol{\Gamma}, \quad \mathbb{C}[\mathcal{R}_i(\mathbf{X}), \mathcal{R}_j(\mathbf{X})] = \mathbf{0}. \tag{5.6}$$

4. Each $\mathcal{R}_i(\mathbf{X})$ is uncorrelated with $\mathcal{M}(\mathbf{X})$.

Strictly speaking, by analogy with De Finetti's theorem (De Finetti, 1992), this representation demands that the observations $\{\mathbf{x}\}$ can be plausibly thought of as a subset of an infinitely exchangeable sequence. However, following Rougier et al. (2013), the infinitely exchangeable representation is used here as a tractable representation of finite exchangeability.

5.2.2 Bayes linear sufficiency and belief separation

In a fully Bayesian analysis, posterior distributions are derived by constructing conditional probability densities. Two variables $\mathbf{A}$ and $\mathbf{B}$ are said to be

conditionally independent given $\mathbf{C}$ if the conditional distribution of $\mathbf{B}$ given both $\mathbf{A}$ and $\mathbf{C}$ is the same as the conditional distribution of $\mathbf{B}$ given $\mathbf{C}$ alone. Conditional independence plays an important role in Bayesian inference, in particular by guaranteeing that the posterior distribution depends on the sample data only via sufficient statistics (Dawid, 1979).

In the Bayes linear framework, beliefs are adjusted not by conditioning but by minimising the squared loss function. Thus, instead of relying on conditional independence, Goldstein and Wooff (2007, §5.15) introduce the concept of belief separation, a generalised conditional independence property. For three collections of random quantities $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, $\mathbf{C}$ is said to be Bayes linear sufficient for $\mathbf{A}$ for adjusting $\mathbf{B}$ if $\mathbb{E}_{\mathbf{C}\cup\mathbf{A}}[\mathbf{B}] = \mathbb{E}_{\mathbf{C}}[\mathbf{B}]$. Equivalently, $\mathbf{C}$ can be said to separate $\mathbf{A}$ and $\mathbf{B}$; this belief separation is written as

$$\lfloor \mathbf{A} \perp\!\!\!\perp \mathbf{B} \rfloor / \mathbf{C}. \tag{5.7}$$

This property simplifies the adjustment of expectations and variances: when $\lfloor \mathbf{A} \perp\!\!\!\perp \mathbf{B} \rfloor / \mathbf{C}$, it follows immediately that

$$\mathbb{E}_{\mathbf{C}\cup\mathbf{A}}[\mathbf{B}] = \mathbb{E}_{\mathbf{C}}[\mathbf{B}], \qquad \mathbb{V}_{\mathbf{C}\cup\mathbf{A}}[\mathbf{B}] = \mathbb{V}_{\mathbf{C}}[\mathbf{B}]. \tag{5.8}$$

In particular, if $\{\mathbf{C}\} \subseteq \{\mathbf{A}\}$ then

$$\mathbb{E}_{\mathbf{A}}[\mathbf{B}] = \mathbb{E}_{\mathbf{C}}[\mathbf{B}], \qquad \mathbb{V}_{\mathbf{A}}[\mathbf{B}] = \mathbb{V}_{\mathbf{C}}[\mathbf{B}].$$

These relationships will be exploited in the next section when adjusting the prior expectations and variances by a collection of observations.

5.2.3 Updating the expectation and variance of the population mean

Under the exchangeability representation described in Section 5.2.1.1, $\mathbf{X}_i = \mathcal{M}(\mathbf{X}) + \mathcal{R}_i(\mathbf{X})$, where the population mean $\mathcal{M}(\mathbf{X})$ and residuals $\mathcal{R}_i(\mathbf{X})$, $\mathcal{R}_j(\mathbf{X})$

are mutually uncorrelated. As a consequence,

$$\lfloor \mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j \rfloor / \mathcal{M}(\mathbf{X}) \qquad \text{for } i = 1, \ldots, n; j > n \tag{5.9}$$

and it is clear that $\mathcal{M}(\mathbf{X})$ induces belief separation between $\mathbf{X}_{1:n}$ and $\mathbf{X}_j$, where $j > n$. Furthermore, by writing

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i = \mathcal{M}(\mathbf{X}) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_i(\mathbf{X}), \tag{5.10}$$

(Goldstein and Wooff, 2007, §6.10) show that the sample mean vector $\overline{\mathbf{X}}$ is Bayes linear sufficient for $\mathbf{X}_{1:n}$ for adjusting $\mathcal{M}(\mathbf{X})$, and hence for adjusting beliefs over future observations $\mathbf{X}_j$ when $j > n$.

From the definition of $\mathbf{X}_i$ given in Section 5.2.1.1, the prior expectations, variances and covariances of $\mathbf{X}_i$ and $\mathbf{X}_j$ can be expressed in terms of the population mean and residuals as

$$\mathbb{E}[\mathbf{X}_i] = \mathbb{E}[\mathcal{M}(\mathbf{X}) + \mathcal{R}_i(\mathbf{X})] = \mathbb{E}[\mathcal{M}(\mathbf{X})] \tag{5.11}$$

$$\mathbb{V}[\mathbf{X}_i] = \mathbb{V}[\mathcal{M}(\mathbf{X}) + \mathcal{R}_i(\mathbf{X})] = \mathbb{V}[\mathcal{M}(\mathbf{X})] + \mathbb{V}[\mathcal{R}_i(\mathbf{X})] \tag{5.12}$$

$$\mathbb{C}[\mathbf{X}_i, \mathbf{X}_j] = \mathbb{C}[\mathcal{M}(\mathbf{X}) + \mathcal{R}_i(\mathbf{X}), \mathcal{M}(\mathbf{X}) + \mathcal{R}_j(\mathbf{X})] = \mathbb{V}[\mathcal{M}(\mathbf{X})], \tag{5.13}$$

whence it follows that

$$\mathbb{E}\left[\overline{\mathbf{X}}\right] = \mathcal{M}(\mathbf{X}), \quad \mathbb{V}\left[\overline{\mathbf{X}}\right] = \mathbb{V}[\mathcal{M}(\mathbf{X})] + \frac{1}{n}\mathbb{V}[\mathcal{R}_i(\mathbf{X})], \quad \mathbb{C}\left[\overline{\mathbf{X}}, \mathcal{M}(\mathbf{X})\right] = \mathbb{V}[\mathcal{M}(\mathbf{X})].$$

This determines all of the elements required to update the expectation and variance of the population mean $\mathcal{M}(\mathbf{X})$ and of future observations $\mathbf{X}_j$. Substituting these terms into the generic forms (5.2) and (5.3) gives the following form for the Bayes linear adjusted expectation and variance of the population

mean $\mathcal{M}(\mathbf{X})$ after updating by the sample mean $\overline{\mathbf{X}}$ of n observations:

$$\begin{aligned}\mathbb{E}_{\overline{\mathbf{X}}}[\mathcal{M}(\mathbf{X})] &= \mathbb{E}[\mathcal{M}(\mathbf{X})] + \mathbb{C}\left[\mathcal{M}(\mathbf{X}), \overline{\mathbf{X}}\right] \mathbb{V}\left[\overline{\mathbf{X}}\right]^{-1} \left(\overline{\mathbf{X}} - \mathbb{E}\left[\overline{\mathbf{X}}\right]\right) \\ &= \mathbb{E}[\mathcal{M}(\mathbf{X})] + \mathbb{V}[\mathcal{M}(\mathbf{X})] \left(\mathbb{V}[\mathcal{M}(\mathbf{X})] + \tfrac{1}{n}\mathbb{V}[\mathcal{R}_i(\mathbf{X})]\right)^{-1} \left(\overline{\mathbf{X}} - \mathbb{E}[\mathcal{M}(\mathbf{X})]\right)\end{aligned} \tag{5.14}$$

and

$$\begin{aligned}\mathbb{V}_{\overline{\mathbf{X}}}[\mathcal{M}(\mathbf{X})] &= \mathbb{V}[\mathcal{M}(\mathbf{X})] - \mathbb{C}\left[\mathcal{M}(\mathbf{X}), \overline{\mathbf{X}}\right] \mathbb{V}\left[\overline{\mathbf{X}}\right]^{-1} \mathbb{C}\left[\overline{\mathbf{X}}, \mathcal{M}(\mathbf{X})\right] \\ &= \mathbb{V}[\mathcal{M}(\mathbf{X})] - \mathbb{V}[\mathcal{M}(\mathbf{X})] \left(\mathbb{V}[\mathcal{M}(\mathbf{X})] + \tfrac{1}{n}\mathbb{V}[\mathcal{R}_i(\mathbf{X})]\right)^{-1} \mathbb{V}[\mathcal{M}(\mathbf{X})].\end{aligned} \tag{5.15}$$

The Woodbury matrix identity (Woodbury, 1950) states that

$$\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1} = (\mathbf{A} + \mathbf{U}\mathbf{C}\mathbf{V})^{-1} \tag{5.16}$$

and (5.15) has the form of the left-hand side, with $\mathbf{U} = \mathbf{V} = \mathbf{I}$, $\mathbf{A}^{-1} = \mathbb{V}[\mathcal{M}(\mathbf{X})]$, and $\mathbf{C}^{-1} = \frac{1}{n}\mathbb{V}[\mathcal{R}_i(\mathbf{X})]$. Hence

$$\mathbb{V}_{\overline{\mathbf{X}}}[\mathcal{M}(\mathbf{X})] = \left(\mathbb{V}[\mathcal{M}(\mathbf{X})]^{-1} + n\mathbb{V}[\mathcal{R}_i(\mathbf{X})]^{-1}\right)^{-1}, \tag{5.17}$$

and it is apparent that the adjusted variance of $\mathcal{M}(\mathbf{X})$ is simply the inverse of the sum of the precision of $\mathcal{M}(\mathbf{X})$ and the precision of n residuals.

5.2.4 Updating the expectation and variance of the population variance (scalar case)

Thus far, the prior expectations have only been adjusted by the observed sample mean; this is analogous to carrying out Bayesian inference with a variance assumed to be known a priori, which is not generally a realistic assumption. To determine how to adjust the prior variance by the observed sample variance, the exchangeability representation must be extended to include

higher-order exchangeability. The following exchangeability representation for scalar variances is taken from Goldstein and Wooff (2007, §6.4), and will be extended to the multivariate case in Section 5.3.

5.2.4.1 Exchangeability representation for an infinite sequence of second-order exchangeable squared residuals

Let $\{\mathcal{R}_1(X)^2, \mathcal{R}_2(X)^2, ...\}$ be the squared residuals from an infinite second-order exchangeable sequence of scalar random quantities $\{X_1, X_2, ...\}$ as defined in Section 5.2.1.1.

Suppose that the sequence $\{\mathcal{R}_1(X)^2, \mathcal{R}_2(X)^2, ...\}$ is also second-order exchangeable, with common mean $\mathbb{E}\Big[\mathcal{R}_i(X)^2\Big]$ and common variance $\mathbb{V}\Big[\mathcal{R}_i(X)^2\Big]$. Then the squared residuals $\{\mathcal{R}(X)^2\}$ may be expressed in terms of the population variance $\mathcal{M}(V)$ and the infinite sequence of individual residual-variance vectors $\{\mathcal{R}_1(V), \mathcal{R}_2(V), ...\}$, which satisfy the following properties.

1. For each individual i,

$$\mathcal{R}_i(X)^2 = \mathcal{M}(V) + \mathcal{R}_i(V). \tag{5.18}$$

2. The population variance $\mathcal{M}(V)$ has expectation and variance

$$\mathbb{E}[\mathcal{M}(V)] = V_R, \qquad \mathbb{V}[\mathcal{M}(V)] = V_M. \tag{5.19}$$

3. The collection $\{\mathcal{R}_1(V), \mathcal{R}_2(V), ...\}$ is second-order exchangeable, with each individual i having

$$\mathbb{E}[\mathcal{R}_i(V)] = 0, \quad \mathbb{V}[\mathcal{R}_i(V)] = V_{R(V)}, \quad \mathbb{C}[\mathcal{R}_i(V), \mathcal{R}_j(V)] = 0. \tag{5.20}$$

4. Each $\mathcal{R}_i(V)$ is uncorrelated with $\mathcal{M}(V)$.

5.2.4.2 Updating the population variance when the population mean is known

Suppose that the population mean $\mathcal{M}(X)$ is known; this means that the squared residuals $\mathcal{R}_i(X)^2 = (X_i - \mathcal{M}(X))^2$ are observable.

Just as the sample mean was Bayes linear sufficient for all of the X_i for adjusting $\mathcal{M}(X)$, the sample mean squared residual $\bar{X}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathcal{M}(X))^2$ is Bayes linear sufficient for all of the $\mathcal{R}_i(X)^2 = (X_i - \mathcal{M}(X))^2$ for adjusting $\mathcal{M}(V)$, with

$$\mathbb{E}\Big[\bar{X}^{(2)}\Big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\mathcal{M}(V) + \mathcal{R}_i(V)\Big)\right] = \mathbb{E}[\mathcal{M}(V)]\,, \tag{5.21}$$

$$\mathbb{V}\Big[\bar{X}^{(2)}\Big] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\mathcal{M}(V) + \mathcal{R}_i(V)\Big)\right] = \mathbb{V}[\mathcal{M}(V)] + \frac{1}{n}\mathbb{V}[\mathcal{R}_i(V)]\,, \tag{5.22}$$

$$\mathbb{C}\Big[\mathcal{M}(V), \bar{X}^{(2)}\Big] = \mathbb{C}\left[\mathcal{M}(V), \frac{1}{n}\sum_{i=1}^{n}\Big(\mathcal{M}(V) + \mathcal{R}_i(V)\Big)\right] = \mathbb{V}[\mathcal{M}(V)]\,. \tag{5.23}$$

The expectations and variances specified in Section 5.2.4.1 can therefore be combined with the update equations (5.2) and (5.3), giving the following equations for the adjustment of the variance of X by the sample squared residuals $\bar{X}^{(2)}$:

$$\begin{aligned}
\mathbb{E}_{\bar{X}^{(2)}}[\mathcal{M}(V)] &= \mathbb{E}[\mathcal{M}(V)] + \mathbb{C}\Big[\mathcal{M}(V), \bar{X}^{(2)}\Big]\mathbb{V}\Big[\bar{X}^{(2)}\Big]^{-1}\Big(\bar{X}^{(2)} - \mathbb{E}\Big[\bar{X}^{(2)}\Big]\Big) \\
&= \mathbb{E}[\mathcal{M}(V)] + \mathbb{V}[\mathcal{M}(V)]\Big(\mathbb{V}[\mathcal{M}(V)] + \tfrac{1}{n}\mathbb{V}[\mathcal{R}_i(V)]\Big)^{-1}\Big(\bar{X}^{(2)} - \mathbb{E}[\mathcal{M}(V)]\Big) \\
&= V_R + V_M\Big(V_M + \tfrac{1}{n}V_{R(V)}\Big)^{-1}\Big(\bar{X}^{(2)} - V_R\Big) \\
&= \frac{V_M\bar{X}^{(2)} + \frac{1}{n}V_{R(V)}V_R}{V_M + \frac{1}{n}V_{R(V)}},
\end{aligned} \tag{5.24}$$

and from (5.3),

$$\begin{aligned}\mathbb{V}_{\bar{X}^{(2)}}[\mathcal{M}(V)] &= \mathbb{V}[\mathcal{M}(V)] - \mathbb{C}\Big[\mathcal{M}(V), \bar{X}^{(2)}\Big]\mathbb{V}\Big[\bar{X}^{(2)}\Big]^{-1}\mathbb{C}\Big[\bar{X}^{(2)}, \mathcal{M}(V)\Big] \\ &= V_M - V_M\left(V_M + \tfrac{1}{n}V_{R(V)}\right)^{-1}V_M \\ &= \frac{\frac{1}{n}V_M V_{R(V)}}{V_M + \frac{1}{n}V_{R(V)}}. \end{aligned} \tag{5.25}$$

5.2.4.3 Updating the population variance when the population mean is not known

In the more realistic case when the population mean $\mathcal{M}(X)$ is not known, the 'true' squared residuals $\mathcal{R}_i(X)^2$ are not observable. Instead, the observable quantities are the sample squared residuals $(X_i - \bar{X})^2$, which are standardised to give the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

In order to define and exploit the relationship between the sample variance and the quantities to be adjusted, an expression for s^2 is required in terms of the quantities specified in Section 5.2.4.1. First, by expanding the sample sum of squares,

$$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \\ &= \sum_{i=1}^{n}\left([X_i - \mathcal{M}(X)] + \left[\mathcal{M}(X) - \bar{X}\right]\right)^2 \\ &= \sum_{i=1}^{n}\left(X_i - \mathcal{M}(X)\right)^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathcal{M}(X)\right)\right)^2. \end{aligned} \tag{5.26}$$

Now s^2 can be expressed in terms of the unobserved 'true' residuals $\mathcal{R}_i(X) =$

$\Big(X_i - \mathcal{M}(X)\Big)$ as

$$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^{n} \mathcal{R}_i(X)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \mathcal{R}_i(X)\right)^2 \\ &= \sum_{i=1}^{n} \mathcal{R}_i(X)^2 - \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{R}_i(X)^2 + \sum_{j\neq i} \mathcal{R}_i(X)\mathcal{R}_j(X)\right\} \\ &= \frac{n-1}{n}\sum_{i=1}^{n} \mathcal{R}_i(X)^2 - \frac{2}{n}\sum_{k<j} \mathcal{R}_j(X)\mathcal{R}_k(X) \qquad (5.27)\end{aligned}$$

and so

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_i(X)^2 - \frac{2}{n(n-1)}\sum_{k<j} \mathcal{R}_j(X)\mathcal{R}_k(X). \qquad (5.28)$$

This quantity can also be represented as $s^2 = \mathcal{M}(V) + T$, where T captures the portion of the variance not attributable to the population variance $\mathcal{M}(V)$. Recall from (5.18) that $\mathcal{M}(V) = [\mathcal{R}_i(X)]^2 - \mathcal{R}_i(V)$, and so

$$\begin{aligned}T &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_i(X)^2 - \frac{2}{n(n-1)}\sum_{k<j} \mathcal{R}_j(X)\mathcal{R}_k(X) - \frac{1}{n}\sum_{i=1}^{n}\Big([\mathcal{R}_i(X)]^2 - \mathcal{R}_i(V)\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_i(V) - \frac{2}{n(n-1)}\sum_{k<j} \mathcal{R}_j(X)\mathcal{R}_k(X). \qquad (5.29)\end{aligned}$$

Goldstein and Wooff (2007, §8.2) introduce the further assumption that products of pairs of residuals $\mathcal{R}_i(X)\mathcal{R}_j(X)$ satisfy certain fourth-order uncorrelated properties - analogous to defining these quantities as mutually independent in a classical treatment - such that

$$\mathbb{C}[\mathcal{M}(V), \mathcal{R}_i(X)\mathcal{R}_j(X)] = 0 \quad \text{for } i \neq j, \qquad (5.30)$$

$$\mathbb{C}[\mathcal{R}_i(V), \mathcal{R}_j(X)\mathcal{R}_j(X)] = 0 \quad \text{when } i \neq j \neq k, \qquad (5.31)$$

$$\mathbb{C}[\mathcal{R}_i(X)\mathcal{R}_j(X), \mathcal{R}_w(X)\mathcal{R}_u(X)] = 0 \quad \text{for } i > j, w > u, \text{ unless } i = w \text{ and } j = u. \qquad (5.32)$$

From Sections 5.2.1.1 and 5.2.4.1, $\mathbb{E}[\mathcal{R}_i(V)] = 0$ and $\mathbb{E}[\mathcal{R}_i(X)\mathcal{R}_j(X)] = 0$ for

$i \neq j$, so it can immediately be seen that $\mathbb{E}[T] = 0$.

To complete the specification, the variance of T is required. From Section 5.2.4.1, we have $\mathbb{V}[\mathcal{R}_i(V)] = V_{R(V)}$, and (5.31) tells us that $\mathbb{C}[\mathcal{R}_i(V), \mathcal{R}_j(X)\mathcal{R}_k(X)] = 0$ when $i \neq j \neq k$, as is the case here. For $\mathbb{V}[\mathcal{R}_j(X)\mathcal{R}_k(X)]$,

$$\begin{aligned}\mathbb{V}[\mathcal{R}_j(X)\mathcal{R}_k(X)] = \mathbb{C}\Big[\mathcal{R}_j(X)^2, \mathcal{R}_k(X)^2\Big] + \\ \Big(\mathbb{V}[\mathcal{R}_j(X)] + \mathbb{E}[\mathcal{R}_j(X)]^2\Big)\Big(\mathbb{V}[\mathcal{R}_k(X)] + \mathbb{E}[\mathcal{R}_k(X)]^2\Big) - \\ \left(\mathbb{C}[\mathcal{R}_j(X), \mathcal{R}_k(X)] + \mathbb{E}[\mathcal{R}_j(X)]\,\mathbb{E}[\mathcal{R}_k(X)]\right)^2. \quad (5.33)\end{aligned}$$

Note that $\mathbb{C}\Big[\mathcal{R}_j(X)^2, \mathcal{R}_k(X)^2\Big] = V_M$, and so

$$\mathbb{V}[\mathcal{R}_j(X)\mathcal{R}_k(X)] = V_M + (V_R + 0)(V_R + 0) - 0^2 = V_M + V_R^2. \quad (5.34)$$

The variance of T is therefore

$$\mathbb{V}[T] = \frac{1}{n} V_{R(V)} + \frac{2\left(V_M + V_R^2\right)}{n(n-1)} = V_T, \text{ say.} \quad (5.35)$$

And, since T is independent of $\mathcal{M}(V)$, it is also true that $\mathbb{C}[\mathcal{M}(V), T] = 0$.

The expectation and variance of the sample variance s^2 can therefore be expressed in terms of quantities about which belief specifications can be made under the exchangeability representation in Section 5.2.4.1, with

$$\mathbb{E}\Big[s^2\Big] = \mathbb{E}[\mathcal{M}(V) + T] = V_R, \quad (5.36)$$

$$\mathbb{V}\Big[s^2\Big] = \mathbb{V}[\mathcal{M}(V) + T] = V_M + V_T, \quad (5.37)$$

$$\mathbb{C}\Big[s^2, \mathcal{M}(V)\Big] = \mathbb{C}[\mathcal{M}(V) + T, \mathcal{M}(V)] = V_M. \quad (5.38)$$

5.2.4.3.1 Adjusted expectation and variance of $\mathcal{M}(V)$

All of the quantities required to carry out Bayes linear adjustment of the expectation and variance of $\mathcal{M}(V)$ by the sample variance s^2 have now been

determined. Using (5.2),

$$\begin{aligned}\mathbb{E}_{s^2}[\mathcal{M}(V)] &= \mathbb{E}[\mathcal{M}(V)] + \mathbb{C}\Big[\mathcal{M}(V), s^2\Big]\,\mathbb{V}\Big[s^2\Big]^{-1}\Big(s^2 - \mathbb{E}\Big[s^2\Big]\Big) \\ &= V_R + V_M\,(V_M + V_T)^{-1}\Big(s^2 - V_R\Big) \\ &= \frac{V_M s^2 + V_T V_R}{V_M + V_T}, \end{aligned} \tag{5.39}$$

and

$$\begin{aligned}\mathbb{V}_{s^2}[\mathcal{M}(V)] &= \mathbb{V}[\mathcal{M}(V)] - \mathbb{C}\Big[\mathcal{M}(V), s^2\Big]\,\mathbb{V}\Big[s^2\Big]^{-1}\,\mathbb{C}\Big[\mathcal{M}(V), s^2\Big] \\ &= V_M - V_M\,(V_M + V_T)^{-1}\,V_M \\ &= \frac{V_M V_T}{V_M + V_T}. \end{aligned} \tag{5.40}$$

5.2.4.3.2 Variance-modified Bayes linear assessments

The adjusted expectation (5.39) of $\mathcal{M}(V)$ is treated as an updated estimate of the residual variance of X_i. $\mathbb{V}[\mathcal{R}_i(X)]$ can therefore now be replaced by $\mathbb{E}_{s^2}[\mathcal{M}(V)]$ in equations (5.14) and (5.15) to obtain the variance-modified adjusted expectation and variance,

$$\mathbb{E}_{\mathbf{x}}[\mathcal{M}(X)] = \mathbb{E}[\mathcal{M}(X)] + \mathbb{V}[\mathcal{M}(X)]\left(\mathbb{V}[\mathcal{M}(X)] + \tfrac{1}{n}\mathbb{E}_{s^2}[\mathcal{M}(V)]\right)^{-1}\left(\bar{X} - \mathbb{E}[\mathcal{M}(X)]\right), \tag{5.41}$$

and

$$\begin{aligned}\mathbb{V}_{\mathbf{x}}[\mathcal{M}(X)] &= \mathbb{V}[\mathcal{M}(X)] - \mathbb{V}[\mathcal{M}(X)]\left(\mathbb{V}[\mathcal{M}(X)] + \tfrac{1}{n}\mathbb{E}_{s^2}[\mathcal{M}(V)]\right)^{-1}\mathbb{V}[\mathcal{M}(X)] \\ &= \left(\mathbb{V}[\mathcal{M}(X)]^{-1} + n\mathbb{E}_{s^2}[\mathcal{M}(V)]^{-1}\right)^{-1}. \end{aligned} \tag{5.42}$$

The notation $\mathbb{E}_{\mathbf{x}}[\mathcal{M}(X)]$ reflects the fact that the adjustment of $\mathcal{M}(X)$ incorporates both the sample mean and sample variance of the data $\mathbf{x}$.

5.2.5 Priors for higher-order quantities

The representation derived in Section 5.2.4.3 enables the creation of update equations for the expectation $\mathbb{E}[\mathcal{M}(V)]$ and variance $\mathbb{V}[\mathcal{M}(V)]$ of the squared residuals. The next step in carrying out the adjustment is to specify priors on each of the required components.

Under the representation given in Section 5.2.4.1, $\mathbb{V}[X] = \mathcal{M}(V) + \mathcal{R}_i(V)$, where $\mathbb{E}[\mathcal{R}_i(V)] = 0$, and so $\mathbb{E}[\mathcal{M}(V)] = \mathbb{V}[X] = V_R$. Specifications are still required for two quantities: for $V_{R(V)}$, which expresses judgements about the shape of X, in the sense of whether X is expected to have particularly heavy or light tails, and represents the fourth-order moments of X; and also for V_M, the variance of the population variance, which reflects confidence in the prior V_R.

5.2.5.1 Expressing kurtosis through $V_{R(V)}$

In specifying $V_{R(V)}$, Goldstein and Wooff (2007, §8.3) propose that the population variance $\mathcal{M}(V)$ be treated as a scale parameter acting on the standardised residuals $Z_i = (X_i - \mathcal{M}(X)) / \sqrt{\mathcal{M}(V)}$, so that

$$\mathcal{R}_i(X) = \sqrt{\mathcal{M}(V)} Z_i \tag{5.43}$$

where

$$\mathbb{E}[Z_i] = 0, \qquad \mathbb{V}[Z_i] = 1, \qquad \mathbb{C}[Z_i, Z_j] = 0, \qquad \mathbb{C}[\mathcal{M}(V), Z_i] = 0. \tag{5.44}$$

It follows from equation (5.18) that

$$\mathcal{R}_i(V) = \mathcal{M}(V)(Z_i^2 - 1). \tag{5.45}$$

From (5.20), $\mathbb{E}[\mathcal{R}_i(V)] = 0$; therefore $\mathbb{V}[\mathcal{R}_i(V)] = \mathbb{E}\Big[\mathcal{R}_i(V)^2\Big] - \mathbb{E}[\mathcal{R}_i(V)]^2 = \mathbb{E}\Big[\mathcal{R}_i(V)^2\Big]$, and so

$$\begin{aligned} V_{R(V)} = \mathbb{V}[\mathcal{R}_i(V)] &= \mathbb{E}\Big[\mathcal{M}(V)^2(Z_i^2-1)^2\Big] \\ &= \mathbb{E}\Big[\mathcal{M}(V)^2\Big]\Big\{\mathbb{V}\Big[Z_i^2-1\Big] + \mathbb{E}\Big[Z_i^2-1\Big]^2\Big\} \\ &= \Big(\mathbb{V}[\mathcal{M}(V)] + \mathbb{E}[\mathcal{M}(V)]^2\Big)\,\mathbb{V}\Big[Z_i^2\Big] \\ &= \Big(V_M + V_R^2\Big)\,\mathbb{V}\Big[Z_i^2\Big]. \end{aligned} \tag{5.46}$$

Thus $V_{R(V)}$ can be determined by quantities specified elsewhere, and the variance $\mathbb{V}\Big[Z_i^2\Big]$ of the standardised squared residuals.

5.2.5.1.1 $\mathbb{V}\Big[Z_i^2\Big]$ and excess kurtosis

An appropriate choice of values for $\mathbb{V}\Big[Z_i^2\Big]$ can be determined by appealing to distribution theory. Following Kendall and Stuart (1969), let $\mathbb{K}\text{ur}\,[X]$ denote the excess kurtosis of X_i with respect to a normal distribution, defined in terms of the fourth and second moments of X_i as

$$\mathbb{K}\text{ur}\,[X] = \frac{\mu_4}{\mu_2^2} - 3 = \mathbb{E}\left[\left(\frac{X_i - \mathcal{M}(X)}{\sqrt{\mathcal{M}(V)}}\right)^4\right] - 3 = \mathbb{E}\Big[Z_i^4\Big] - 3, \tag{5.47}$$

where μ_r denotes the rth central moment of X_i. Note that Goldstein and Wooff (2007, §8.3) use the raw kurtosis $\mathbb{K}\text{ur}\,[X] + 3$ in their derivation: the two formulations are interchangeable, but parametrising in terms of the excess kurtosis will result in a more direct interpretation of the effect of the adjustment in later sections.

This fourth-order moment of Z_i can be expressed in terms of the standardised residuals Z_i as

$$\mathbb{E}\Big[Z_i^4\Big] = \mathbb{E}\Big[\Big(Z_i^2\Big)^2\Big] = \mathbb{V}\Big[Z_i^2\Big] + \mathbb{E}\Big[Z_i^2\Big]^2. \tag{5.48}$$

By construction, $\mathbb{E}\Big[Z_i^2\Big]$ is equivalent to the second standardised moment μ_2/σ^2,

which is always one, and so

$$\mathbb{K}\text{ur}[X] = \left(\mathbb{V}\left[Z_i^2\right] + 1\right) - 3 = \mathbb{V}\left[Z_i^2\right] - 2. \tag{5.49}$$

$\mathbb{V}\left[Z_i^2\right]$ can now be specified according to the prior judgement about the shape of the distribution. Setting $\mathbb{V}\left[Z_i^2\right] = 2$ implies that $\mathbb{K}\text{ur}[X] = 0$, hence that the residuals are approximately Gaussian; larger values of $\mathbb{V}\left[Z_i^2\right]$ imply an increasingly heavier-tailed distribution, while negative values imply a distribution with lighter tails than a normal distribution.

A natural approach to eliciting an appropriate value for $\mathbb{V}\left[Z_i^2\right]$ to indicate heavier-than-normal tails is to proceed as though the residuals follow a Student-t distribution, as suggested by Goldstein and Wooff (2007, §8.3). Let T_ν have a t distribution with $\nu > 4$ degrees of freedom. Then $\mathbb{V}[T_\nu] = \nu/(\nu-2)$; to standardise this to have unit variance, define the standardised residuals as

$$Z_i = \sqrt{\frac{\nu-2}{\nu}} T_\nu \tag{5.50}$$

so that the squared residuals are F distributed as

$$Z_i^2 \sim \frac{\nu-2}{\nu} F(1,\nu) \qquad \text{with} \qquad \mathbb{V}\left[Z_i^2\right] = \frac{2(\nu-1)}{\nu-4}, \tag{5.51}$$

and, using (5.49), the corresponding kurtosis is

$$\mathbb{V}\left[Z_i^2\right] - 2 = \frac{2(\nu-1) - 2(\nu-4)}{\nu-4} = \frac{6}{\nu-4}. \tag{5.52}$$

For lighter-tailed distributions, Goldstein and Wooff (2007, §8.3) note that the excess kurtosis of any regular unimodal symmetric distribution cannot be less than -1.2 (Stuart et al., 1994); from (5.49), therefore, a choice of $\mathbb{V}\left[Z_i^2\right] > 0.8$ will be appropriate.

With $V_{R(V)} = \left(V_M + V_R^2\right)\mathbb{V}\left[Z_i^2\right]$ defined in this way, with $\mathbb{V}\left[Z_i^2\right] = \mathbb{K}\text{ur}[X] +$

2, (5.35) can be written as

$$\begin{aligned} V_T &= \frac{\left(V_M + V_R^2\right)}{n}\left\{\mathbb{V}\left[Z_i^2\right] + \frac{2}{n-1}\right\} \\ &= \frac{\left(V_M + V_R^2\right)}{n}\left\{\mathbb{K}\text{ur}\left[X\right] + \frac{2n}{n-1}\right\}. \end{aligned} \tag{5.53}$$

Increasing $\mathbb{V}\left[Z_i^2\right]$ to represent heavier tails thus increases V_T; as a result, the prior variance V_R receives slightly more weight relative to the sample variance s^2 in (5.39).

5.2.5.2 Expressing confidence through V_M

The remaining term to be specified is $V_M = \mathbb{V}[\mathcal{M}(V)]$, the prior uncertainty about the population variance. Small values indicate a relatively high level of confidence in the assessment of $\mathbb{E}[\mathcal{M}(V)]$, with larger values indicating that it would not be surprising to find substantial changes to the prior variance after observing a large sample.

Rather than attempt to specify V_M directly - which would involve somehow quantifying and articulating the user's uncertainty about the variance - Goldstein and Wooff (2007, §8.3) propose setting $V_M = cV_R^2$, and instead choosing some value of $c > 0$ that reflects the user's beliefs about the value of the prior information.

The scaling parameter c has no interpretable units, and so it is not immediately obvious how to specify a sensible value for c to reflect prior beliefs about the relationship between V_M and V_R. Goldstein and Wooff (2007, §8.3) suggest that the exact value for c be determined by the size n of the observed sample from which s^2 is estimated, the kurtosis parameter $\mathbb{V}\left[Z_i^2\right]$, and a parameter m denoting the notional size of sample from which V_R might be considered to have been estimated, with these four quantities related by

$$c = \frac{\kappa n}{m(n-1) - \kappa n}, \qquad \text{where} \qquad \kappa = \frac{1}{n}\left\{(n-1)\mathbb{V}\left[Z_i^2\right] + 2\right\} \tag{5.54}$$

is a term introduced by Goldstein and Wooff (2007, §8.3) for convenience. Setting $V_M = cV_R^2$ in (5.53), this definition of c leads to

$$V_T = \frac{c+1}{n}\left\{\mathbb{V}\left[Z_i^2\right] + \frac{2}{n-1}\right\}V_R^2 = (c+1)\frac{\kappa}{n-1}V_R^2, \tag{5.55}$$

while from (5.46),

$$V_{R(V)} = (c+1)\mathbb{V}\left[Z_i^2\right]V_R^2. \tag{5.56}$$

Substituting (5.55) and (5.56) into (5.39), the adjusted expectation of $\mathcal{M}(V)$ is

$$\begin{aligned}
\mathbb{E}_{s^2}[\mathcal{M}(V)] &= \frac{cV_R^2 s^2 + (c+1)\frac{\kappa}{n-1}V_R^2 V_R}{cV_R^2 + (c+1)\frac{\kappa}{n-1}V_R^2} \\
&= \frac{cs^2 + (c+1)\frac{\kappa}{n-1}V_R}{c + (c+1)\frac{\kappa}{n-1}} \\
&= \frac{\kappa n s^2 + m(n-1)\frac{\kappa}{n-1}V_R}{\kappa n + m(n-1)\frac{\kappa}{n-1}} = \frac{ns^2 + mV_R}{n+m}.
\end{aligned} \tag{5.57}$$

Similarly, from (5.40), the adjusted variance of $\mathcal{M}(V)$ is

$$\begin{aligned}
\mathbb{V}_{s^2}[\mathcal{M}(V)] &= \frac{cV_R^2(c+1)\frac{\kappa}{n-1}V_R^2}{cV_R^2 + (c+1)\frac{\kappa}{n-1}V_R^2} \\
&= \frac{c(c+1)\frac{\kappa}{n-1}V_R^2}{c + (c+1)\frac{\kappa}{n-1}} \\
&= \frac{\frac{\kappa n \kappa m}{(m(n-1)-\kappa n)^2}}{\frac{\kappa n + \kappa m}{m(n-1)-\kappa n}}V_R^2 \\
&= \frac{\kappa n m}{(m(n-1)-\kappa n)(n+m)}V_R^2.
\end{aligned} \tag{5.58}$$

Specifying c in this way fixes the relationship between the prior and observed sample sizes in such a way that, once both m and n are specified, changes to the kurtosis parameter $\mathbb{V}\left[Z_i^2\right]$ have no effect on the adjusted expectation of the population variance, but are effectively absorbed by c. Given that c itself has no physical interpretation, it is hard to justify the effort of carefully specifying

$\mathbb{V}\left[Z_i^2\right]$ when using this approach, and it seems likely that many users would simply dispense with this step altogether as a result.

An alternative approach to the specification of c is proposed in the next section, which separates the effect of adjusting the notional sample size from that of adjusting the kurtosis, and in so doing allows direct interpretation of the effect of changes to the kurtosis on the weights assigned to the prior and observed sample variances.

5.2.5.2.1 Specification of the relationship between V_M and V_R

In order to separate the effect of sample size from that of the kurtosis parameter, c should be defined in such a way that the relationship between V_M and V_R is affected only by the sample size. This can be achieved by borrowing from a commonly-used parametric form for the prior on the population variance $\mathcal{M}(V)$. Suppose that $\mathcal{M}(V)$ is assumed to behave as if it were the variance of a normally distributed random variable, say X, thus fixing the excess kurtosis $\mathbb{K}\text{ur}[X]$ at zero. The natural conjugate choice for the implied prior distribution of $\mathcal{M}(V)$ would then be the scaled inverse-Chi-squared distribution with ν degrees of freedom. Under this parametrisation,

$$\mathbb{E}[\mathcal{M}(V)] = \frac{\nu\tau^2}{\nu-2}, \qquad \mathbb{V}[\mathcal{M}(V)] = \frac{2\left(\nu\tau^2\right)^2}{(\nu-2)^2(\nu-4)} \tag{5.59}$$

where τ^2 denotes a scaling parameter; hence, with $c = V_M/V_R^2$,

$$c = \frac{\mathbb{V}[\mathcal{M}(V)]}{\mathbb{E}[\mathcal{M}(V)]^2} = \frac{2}{\nu-4}. \tag{5.60}$$

The user remains free to specify V_R however they wish; however, the relationship between V_M and V_R is now fixed in terms of the degrees of freedom ν. Under this specification,

$$V_M = \frac{2}{\nu-4}V_R^2, \qquad V_T = \frac{\nu-2}{(\nu-4)n}\left\{\mathbb{V}\left[Z_i^2\right] + \frac{2}{n-1}\right\}V_R^2. \tag{5.61}$$

The adjusted expectation of $\mathcal{M}(V)$ is now

$$\begin{aligned}\mathbb{E}_{s^2}[\mathcal{M}(V)] &= \frac{\frac{2}{\nu-4}V_R^2 s^2 + \frac{\nu-2}{(\nu-4)n}\left\{\mathbb{V}\left[Z_i^2\right] + \frac{2}{n-1}\right\}V_R^2 V_R}{\frac{2}{\nu-4}V_R^2 + \frac{\nu-2}{(\nu-4)n}\left\{\mathbb{V}\left[Z_i^2\right] + \frac{2}{n-1}\right\}V_R^2} \\ &= \frac{2(n-1)s^2 + \frac{\nu-2}{n}\left\{(n-1)\mathbb{V}\left[Z_i^2\right] + 2\right\}V_R}{2(n-1) + \frac{\nu-2}{n}\left\{(n-1)\mathbb{V}\left[Z_i^2\right] + 2\right\}}. \end{aligned} \tag{5.62}$$

Note that this can be written in terms of κ, the quantity used in (5.54) by Goldstein and Wooff (2007, §8.3) to capture the effect of the kurtosis on the adjustment, as

$$\mathbb{E}_{s^2}[\mathcal{M}(V)] = \frac{2(n-1)s^2 + \kappa(\nu-2)V_R}{2(n-1) + \kappa(\nu-2)}. \tag{5.63}$$

Setting $\mathbb{V}\left[Z_i^2\right] = 2$, reflecting a belief that there is no excess kurtosis in the residuals, gives $\kappa = 2$ from (5.54); so when there is no excess kurtosis, the weights on the observed and prior sample variances are governed directly by the observed and notional sample sizes, as in (5.57). Any increase in the kurtosis, expressed via $\mathbb{V}\left[Z_i^2\right] > 2$, will lead to a greater weight on the prior, while reducing the kurtosis by setting $\mathbb{V}\left[Z_i^2\right] < 2$ will increase the weight on the observations.

It is also informative to express the above in terms of $\mathbb{K}\text{ur}\,[X] = \mathbb{V}\left[Z_i^2\right] - 2$, the excess kurtosis with respect to a normal distribution. Substituting this into (5.61),

$$V_T = \frac{\nu-2}{(\nu-4)n}\left\{\mathbb{K}\text{ur}\,[X] + \frac{2n}{n-1}\right\}V_R^2, \tag{5.64}$$

from which

$$\mathbb{E}_{s^2}[\mathcal{M}(V)] = \frac{(n-1)s^2 + (\nu-2)\left\{\frac{(n-1)\mathbb{K}\text{ur}[X]}{2n} + 1\right\}V_R}{(n-1) + (\nu-2)\left\{\frac{(n-1)\mathbb{K}\text{ur}[X]}{2n} + 1\right\}}. \tag{5.65}$$

Expressed in this form, it is easy to see that if $\mathbb{K}\text{ur}\,[X] = 0$ then the weights

on s^2 and V_R are proportional to $n-1$ and $(\nu-1)-1$ respectively, as before. Furthermore, the effect of changing the kurtosis can immediately be quantified in terms of the cost in terms of the number of additional samples required to balance the effect on the weights: an increase of one unit of kurtosis has the effect of adding $[(\nu-2)(n-1)/2n]$ to the notional prior sample size.

5.2.6 Bayes linear adjustment of covariance matrices

While the exchangeability representation used in Section 5.2.3 permits adjustments to the whole mean vector simultaneously, the scalar framework for variance adjustments reviewed above only permits adjustments to one variable at a time. Goldstein and Wooff (2007, §8.11-8.13) suggest that this method might be extended to a multivariate setting by either applying the weighting scheme used in (5.57) to the prior and observed variance-covariance matrices; or by carrying out scalar adjustments to the variances of interest, and combining those adjusted marginal variances with a weighted sum of prior and observed correlation matrices. This heuristic approach may provide an adequate adjustment in cases where the variances are of interest only insofar as they are informative about the mean vector; however, in a weather forecasting context, well-calibrated variances and covariances are necessary in order to issue skilful multivariate probabilistic forecasts, so it is worth considering a more rigorous approach. Goldstein and Wooff (2007) also provide a more formal alternative for the multivariate setting, based on geometrical considerations.

Recall from Section 5.2.4.3 that $s^2 = \mathcal{M}(V) + T$, where $\mathcal{M}(V)$ is the population variance and T describes the portion of the sample variance attributed to individual variability within each sample. The multivariate analogues of these quantities are the sample covariance matrix $\mathbf{S}$, population covariance matrix $\mathcal{M}(\mathbf{V})$, and excess variation matrix $\mathbf{T}$. These quantities are discussed in more detail in Section 5.3; for now, a key observation is that by direct analogy with the scalar case we can write $\mathbf{S} = \mathcal{M}(\mathbf{V}) + \mathbf{T}$ with $\mathbb{E}[\mathbf{T}] = \mathbf{0}$, and with all elements of $\mathcal{M}(\mathbf{V})$ uncorrelated with those of $\mathbf{T}$.

The starting point for the geometric approach of Goldstein and Wooff (2007,

§8.11) is rooted in the work of Goldstein (1981), who provided a geometrical interpretation for Bayes linear adjustment by considering the various uncertain quantities as elements of a vector space of random variables, with inner products defined by covariances. In the present context, the extension to covariance matrices considers the space ($\mathcal{H}$, say) spanned by $\mathbf{S}$ and $\mathcal{M}(\mathbf{V})$ with inner product defined by

$$\langle \mathbf{P}, \mathbf{Q} \rangle = \mathrm{tr}\Big((\mathbf{P} - \mathbb{E}[\mathbf{P}]) (\mathbf{Q} - \mathbb{E}[\mathbf{Q}]) \Big) \qquad (\mathbf{P}, \mathbf{Q} \in \mathcal{H}) \tag{5.66}$$

Strictly speaking, this is not a true inner product: it satisfies the properties $\langle \mathbf{P}, \mathbf{Q} \rangle = \langle \mathbf{Q}, \mathbf{P} \rangle$, $\langle a\mathbf{P} + b\mathbf{Q}, \mathbf{R} \rangle = a \langle \mathbf{P}, \mathbf{R} \rangle + b \langle \mathbf{Q}, \mathbf{R} \rangle$ and $\langle \mathbf{P}, \mathbf{P} \rangle \geq 0$, but the subtraction of expectations means that there may be more than one element $\mathbf{P}$ such that $\langle \mathbf{P}, \mathbf{P} \rangle = 0$. As noted in Goldstein (1981), this can be resolved by considering equivalence classes: this is not necessary for the level of detail considered here, however, and the issue does not affect the subsequent development.

With the inner product defined by (5.66), the norm of $\mathbf{P} \in \mathcal{H}$ is

$$\|\mathbf{P}\| = \sqrt{\langle \mathbf{P}, \mathbf{P} \rangle} = \sqrt{\sum_{ij} \mathbb{V}[P_{ij}]} \tag{5.67}$$

where P_{ij} is the (i, j)th element of $\mathbf{P}$.

In the scalar-valued case, Goldstein (1981) demonstrates that the Bayes linear update of prior judgements based on data can be regarded as a projection of the random variable of interest onto the subspace spanned by the data. In the present context, the analogue of this is to project the random matrix $\mathcal{M}(\mathbf{V}) - \mathbb{E}[\mathcal{M}(\mathbf{V})]$ into the subspace spanned by $\mathbf{S} - \mathbb{E}[\mathbf{S}]$. This projection is $\alpha (\mathbf{S} - \mathbb{E}[\mathbf{S}])$, where

$$\alpha = \frac{\langle \mathcal{M}(\mathbf{V}), \mathbf{S} \rangle}{\langle \mathbf{S}, \mathbf{S} \rangle} = \frac{\langle \mathcal{M}(\mathbf{V}), \mathbf{S} \rangle}{\|\mathbf{S}\|^2}. \tag{5.68}$$

Rearranging, this suggests an adjustment of $\mathcal{M}(\mathbf{V})$ by $\mathbf{S}$ as

$$\mathbb{E}[\mathcal{M}(\mathbf{V})] + \alpha\left(\mathbf{S} - \mathbb{E}[\mathbf{S}]\right). \tag{5.69}$$

Now, as noted above, we have $\mathbf{S} = \mathcal{M}(\mathbf{V}) + \mathbf{T}$, with $\mathbb{E}[\mathbf{T}] = \mathbf{0}$. Therefore $\mathbf{S} - \mathbb{E}[\mathbf{S}] = \mathbf{S} - \mathbb{E}[\mathcal{M}(\mathbf{V})]$, which, on substitution into (5.69), yields

$$\mathbb{E}[\mathcal{M}(\mathbf{V})] + \alpha\left(\mathbf{S} - \mathbb{E}[\mathbf{S}]\right) = \alpha\mathbf{S} + (1-\alpha)\mathbb{E}[\mathcal{M}(\mathbf{V})]. \tag{5.70}$$

This is a weighted average of the sample covariance matrix and prior expectation of $\mathcal{M}(\mathbf{V})$, with the weights determined by (5.68). An explicit expression for these weights is obtained by noting that, as will be shown in Section 5.3.3.2, all elements of $\mathcal{M}(\mathbf{V})$ are uncorrelated with those of $\mathbf{T}$, and so $\langle \mathcal{M}(\mathbf{V}), \mathbf{T} \rangle = 0$. Hence

$$\langle \mathcal{M}(\mathbf{V}), \mathbf{S} \rangle = \langle \mathcal{M}(\mathbf{V}), \mathcal{M}(\mathbf{V}) + \mathbf{T} \rangle = \langle \mathcal{M}(\mathbf{V}), \mathcal{M}(\mathbf{V}) \rangle = \|\mathcal{M}(\mathbf{V})\|^2 \tag{5.71}$$

and

$$\langle \mathbf{S}, \mathbf{S} \rangle = \langle \mathcal{M}(\mathbf{V}) + \mathbf{T}, \mathcal{M}(\mathbf{V}) + \mathbf{T} \rangle = \|\mathcal{M}(\mathbf{V})\|^2 + \|\mathbf{T}\|^2. \tag{5.72}$$

Therefore, from (5.67),

$$\alpha = \frac{\sum_{ij} \mathbb{V}[\mathcal{M}(V)_{ij}]}{\sum_{ij} \mathbb{V}[\mathcal{M}(V)_{ij}] + \sum_{ij} \mathbb{V}[T_{ij}]}, \tag{5.73}$$

where $\mathbb{V}[\mathcal{M}(V)_{ij}]$ and $\mathbb{V}[T_{ij}]$ are the scalar variances and covariances associated with the (i,j)th elements of, respectively, $\mathcal{M}(\mathbf{V})$ and $\mathbf{T}$.

Extensions of this geometrical approach have been developed by Wilkinson and Goldstein (1995), who aim to provide more flexible updating schemes by decomposing the covariance matrices into component parts using orthogonal basis representations, and updating each component separately. This allows, for example, separate updating schemes for the diagonal and off-diagonal elements.

However, the additional flexibility offered by this approach comes at the cost of having to specify prior expectations and covariances for all of the relevant components individually: this is not trivial, although examples are given by Wilkinson and Goldstein (1996), Wilkinson (1997) and Williamson et al. (2012).

In the next section an alternative to these existing approaches to the adjustment of covariance matrices is proposed. The method is derived from a second-order exchangeability representation for vectors of cross-products of residuals from quantities that are themselves second-order exchangeable: this representation is directly analogous to the scalar representation already discussed and requires specification of the same parameters that are used in the scalar adjustment, greatly simplifying elicitation.

5.3 Multivariate Bayes linear adjustment of the population variance

In this section, the exchangeability representation reviewed in Section 5.2.4.1 is extended to accommodate second-order exchangeability between cross products of vectors of residuals from quantities that are themselves second-order exchangeable. This representation provides the basis for the development of the multivariate Bayes linear covariance matrix adjustment, which will be shown not only to be a generalisation of the scalar variance adjustment, but also to closely approximate the posterior distribution that would be obtained by probabilistic Bayesian inference using the natural conjugate normal-inverse-Wishart prior distribution.

The development makes extensive use of various matrix identities and operations, which are reviewed in Appendix C. In particular, when dealing with variances of the elements of a matrix $\mathbf{A}$, say, the standard approach is to transform the matrix into a vector. This makes it possible to define the matrix of variances and covariances between the elements of $\mathbf{A}$, which does not exist when $\mathbf{A}$ is in matrix form. This operation is carried out using the vec operator: given any $m \times n$ matrix $\mathbf{A}$, $\text{vec}\big(\mathbf{A}\big)$ denotes the $mn \times 1$ vector obtained by

stacking the columns of $\mathbf{A}$ one underneath the other (Schott, 2016).

5.3.1 A multivariate exchangeability representation

Let $\{\mathcal{R}_1(\mathbf{X})\mathcal{R}_1(\mathbf{X})', \mathcal{R}_2(\mathbf{X})\mathcal{R}_2(\mathbf{X})', ...\}$ be the residual cross-product matrices from an infinite second-order exchangeable sequence of random quantities $\{\mathbf{X}_1, \mathbf{X}_2, ...\}$, where each $\mathbf{X}_i$ is an $m \times 1$ vector as defined in Section 5.2.1.1. Denote by $\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big)$ the $m^2 \times 1$ vector obtained by stacking the columns of $\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'$ one underneath the other.

Suppose that the sequence $\{\text{vec}\Big(\mathcal{R}_1(\mathbf{X})\mathcal{R}_1(\mathbf{X})'\Big), \text{vec}\Big(\mathcal{R}_2(\mathbf{X})\mathcal{R}_2(\mathbf{X})'\Big), ...\}$ is also second-order exchangeable, with common $m^2 \times 1$ mean vector $\mathbb{E}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big)\Big] = \text{vec}\Big(\mathbb{V}[\mathcal{R}_i(\mathbf{X})]\Big)$ and common $m^2 \times m^2$ variance-covariance matrix $\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big)\Big]$.

Then $\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big)$ may be expressed in terms of the $m^2 \times 1$ vector of population variances $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$, and the infinite sequence of individual $m^2 \times 1$ residual vectors $\Big\{\text{vec}\Big(\mathcal{R}_1(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_2(\mathbf{V})\Big), \ldots\Big\}$, which satisfy the following properties.

1. For each individual i,

$$\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big) = \text{vec}\Big(\mathcal{M}(\mathbf{V})\Big) + \text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big), \tag{5.74}$$

   hence

$$\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' = \mathcal{M}(\mathbf{V}) + \mathcal{R}_i(\mathbf{V}). \tag{5.75}$$

2. The population variance $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ has expectation and variance

$$\mathbb{E}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \text{vec}\Big(\mathbf{V}_R\Big), \qquad \mathbb{V}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbf{V}_M \tag{5.76}$$

   where $\text{vec}\Big(\mathbf{V}_R\Big)$ is an $m^2 \times 1$ vector and $\mathbf{V}_M$ is an $m^2 \times m^2$ matrix.

3. The collection $\{\text{vec}\Big(\mathcal{R}_1(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_2(\mathbf{V})\Big), ...\}$ is also second-order ex-

changeable, with each individual i having

$$\mathbb{E}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] = \mathbf{0}, \tag{5.77}$$

$$\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] = \mathbf{V}_{R(V)}, \tag{5.78}$$

$$\mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_j(\mathbf{V})\Big)\Big] = \mathbf{0} \tag{5.79}$$

where each $\mathcal{R}_i(\mathbf{V})$ is an $m \times m$ matrix, and $\mathbf{V}_{R(V)}$ and $\mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_j(\mathbf{V})\Big)\Big]$ are both $m^2 \times m^2$ matrices.

4. All elements of each $\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)$ are uncorrelated with those of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$.

The additional fourth-order properties introduced in (5.30)-(5.32) in the univariate setting also hold between all elements of the vectors of residuals for different individuals, so that

$$\mathbb{C}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] = \mathbf{0} \quad \text{for } i \neq j, \tag{5.80}$$

$$\mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_j(\mathbf{X})\mathcal{R}_k(\mathbf{X})'\Big)\Big] = \mathbf{0} \quad \text{for } i \neq j, j \neq k, i \neq k \tag{5.81}$$

$$\mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big), \text{vec}\Big(\mathcal{R}_w(\mathbf{X})\mathcal{R}_u(\mathbf{X})'\Big)\Big] = \mathbf{0} \quad \begin{array}{l}\text{for } i \neq j, w \neq u, \\ \text{unless } i = w \text{ and } j = u,\end{array} \tag{5.82}$$

where all of these are $m^2 \times m^2$ matrices.

As in the univariate case, $\mathbf{V}_R$, $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ are quantities that will be specified by the user to reflect their prior beliefs about the dispersion and shape of $\mathbf{X}$. $\mathbf{V}_R$ is assumed to be a valid covariance matrix; that is, it is assumed to be a symmetric positive definite matrix of full rank. Specification of $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ will be considered subsequently.

5.3.2 Adjusting the population variance matrix when the population mean vector is known

Suppose that the population mean $\mathcal{M}(\mathbf{X})$ is known, so that the residual cross-product matrices $\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' = [\mathbf{X}_i - \mathcal{M}(\mathbf{X})][\mathbf{X}_i - \mathcal{M}(\mathbf{X})]'$ are observable. As in the scalar case presented in Section 5.2.4, the vectorised sample mean of the residual cross-product matrices, $\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)$, is Bayes linear sufficient for all of the $\text{vec}\left(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\right)$ for adjusting $\text{vec}\left(\mathcal{M}(\mathbf{V})\right)$, with

$$\mathbb{E}\left[\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right] = \mathbb{E}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right] = \text{vec}\left(\mathbf{V}_R\right), \tag{5.83}$$

$$\begin{aligned}\mathbb{V}\left[\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right] &= \mathbb{V}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right] + \frac{\mathbb{V}\left[\text{vec}\left(\mathbf{V}_{R(V)}\right)\right]}{n} \\ &= \mathbf{V}_M + \frac{\mathbf{V}_{R(V)}}{n}, \end{aligned} \tag{5.84}$$

$$\mathbb{C}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right), \text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right] = \mathbb{V}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right] = \mathbf{V}_M. \tag{5.85}$$

This specification can be used in the update equations (5.2) and (5.3) to obtain expressions for the adjustment of the vector of variances of $\mathbf{X}$ by the sample mean of the residual cross-product matrices, $\overline{\mathbf{X}}^{(2)}$,with one small adjustment: because $\overline{\mathbf{X}}^{(2)}$ is a symmetric matrix, $\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)$ will contain duplicated elements, so $\mathbb{V}\left[\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right]$ will be singular, and the matrix inverses $\mathbf{A}^{-1}$ in (5.2) and (5.3) must be replaced by the generalised inverses $\mathbf{A}^{\dagger}$.

$$\begin{aligned}&\mathbb{E}_{\overline{\mathbf{X}}^{(2)}}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right] = \mathbb{E}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right] + \\ &\quad \mathbb{C}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right), \text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right] \mathbb{V}\left[\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right]^{\dagger} \left(\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right) - \mathbb{E}\left[\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right)\right]\right) \\ &= \text{vec}\left(\mathbf{V}_R\right) + \mathbf{V}_M \left(\mathbf{V}_M + \frac{1}{n}\mathbf{V}_{R(V)}\right)^{\dagger} \left(\text{vec}\left(\overline{\mathbf{X}}^{(2)}\right) - \text{vec}\left(\mathbf{V}_R\right)\right), \end{aligned} \tag{5.86}$$

$$\mathbb{V}_{\overline{\mathbf{X}}^{(2)}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] - \mathbb{C}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big), \mathrm{vec}\Big(\overline{\mathbf{X}}^{(2)}\Big)\Big]\mathbb{V}\Big[\mathrm{vec}\Big(\overline{\mathbf{X}}^{(2)}\Big)\Big]^{\dagger}\mathbb{C}\Big[\mathrm{vec}\Big(\overline{\mathbf{X}}^{(2)}\Big), \mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$$
$$= \mathbf{V}_M - \mathbf{V}_M\left(\mathbf{V}_M + \frac{1}{n}\mathbf{V}_{R(V)}\right)^{\dagger}\mathbf{V}_M, \tag{5.87}$$

where $\mathrm{vec}\Big(\mathbf{V}_R\Big)$ and $\mathrm{vec}\Big(\overline{\mathbf{X}}^{(2)}\Big)$ are $m^2 \times 1$ column vectors, and $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ are $m^2 \times m^2$ matrices.

5.3.3 Adjusting the population variance matrix when the population mean vector is unknown

As in the scalar case reviewed in Section 5.2.4.3, the formulation above must be adapted to account for the more realistic case when the population mean $\mathcal{M}(\mathbf{X})$ is not known. The residuals $\mathcal{R}_i(\mathbf{X})$ are now not observable; instead, the observable quantities are the sample residual cross-product matrices

$$\Big(\mathbf{X}_i - \overline{\mathbf{X}}\Big)\Big(\mathbf{X}_i - \overline{\mathbf{X}}\Big)' = \Big(\mathcal{R}_i(\mathbf{X}) - \bar{\mathbf{R}}\Big)\Big(\mathcal{R}_i(\mathbf{X}) - \bar{\mathbf{R}}\Big)', \tag{5.88}$$

where $\bar{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})$. Instead of adjusting $\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ by $\mathrm{vec}\Big(\overline{\mathbf{X}}^{(2)}\Big)$, the adjustment will be by the observed residual cross-product matrices, which are standardised in the usual way to obtain the vectorised sample variance matrix,

$$\mathrm{vec}\Big(\mathbf{S}\Big) = \mathrm{vec}\left(\frac{1}{n-1}\sum_{i=1}^{n}\Big(\mathbf{X}_i - \overline{\mathbf{X}}\Big)\Big(\mathbf{X}_i - \overline{\mathbf{X}}\Big)'\right). \tag{5.89}$$

5.3.3.1 Representation of the sample variance matrix S

In order to determine how the observed sample variance can be used in the Bayes linear adjustment, $\mathbf{S}$ must first be expressed in terms of the quantities described in Section 5.3.1. The first step in this process is to expand the sample sum of squares, via a multivariate generalisation of the derivation in Section

5.2.4.3:

$$\begin{aligned}(n-1)\mathbf{S} &= \sum_{i=1}^{n}(\mathbf{X}_i-\overline{\mathbf{X}})(\mathbf{X}_i-\overline{\mathbf{X}})' \\ &= \sum_{i=1}^{n}\Big([\mathbf{X}_i-\mathcal{M}(\mathbf{X})]+\Big[\mathcal{M}(\mathbf{X})-\overline{\mathbf{X}}\Big]\Big)\Big([\mathbf{X}_i-\mathcal{M}(\mathbf{X})]+\Big[\mathcal{M}(\mathbf{X})-\overline{\mathbf{X}}\Big]\Big)' \\ &= \sum_{i=1}^{n}\Big(\mathbf{X}_i-\mathcal{M}(\mathbf{X})\Big)\Big(\mathbf{X}_i-\mathcal{M}(\mathbf{X})\Big)' - \\ &\qquad n\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i-\mathcal{M}(\mathbf{X})\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i-\mathcal{M}(\mathbf{X})\right)'. \end{aligned} \tag{5.90}$$

Now writing $\mathcal{R}_i(\mathbf{X}) = \mathbf{X}_i - \mathcal{M}(\mathbf{X})$,

$$\begin{aligned}(n-1)\mathbf{S} &= \sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \frac{1}{n}\left(\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\right)\left(\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\right)' \\ &= \sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' + \sum_{j\neq i}\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\right\} \\ &= \frac{n-1}{n}\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})' \end{aligned} \tag{5.91}$$

and so

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})', \tag{5.92}$$

where $\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'$ denotes the cross product of the residuals from any two individuals.

By analogy with the scalar case, it is helpful to rewrite this expression in terms of the population variance $\mathcal{M}(\mathbf{V})$: thus $\mathbf{S} = \mathcal{M}(\mathbf{V}) + \mathbf{T}$, where $\mathbf{T}$ represents the additional variation due to sampling. From the second-order multivariate exchangeability representation (5.75),

$$\mathcal{M}(\mathbf{V}) = \mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{R}_i(\mathbf{V})$$

for each i, whence $\mathcal{M}(\mathbf{V})$ is also equal to $\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{R}_i(\mathbf{V})\right)$, so

$$\begin{aligned}\mathbf{T} &= \mathbf{S} - \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{R}_i(\mathbf{V})\Big\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_i(\mathbf{V}) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'. \end{aligned} \tag{5.93}$$

For the subsequent development, the expectation and variance of $\mathbf{S}$ are required: however, because $\mathbf{S}$ is a matrix, $\mathbb{V}[\mathbf{S}]$ is not defined. Instead, the expectation and variance of $\mathrm{vec}\Big(\mathbf{S}\Big)$ will be used.

5.3.3.2 Expectation and variance of vec (S)

First, vectorising $\mathbf{S} = \mathcal{M}(\mathbf{V}) + \mathbf{T}$ and taking expectations,

$$\mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{S}\Big)\Big] = \mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V}) + \mathbf{T}\Big)\Big] = \mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] + \mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{T}\Big)\Big]. \tag{5.94}$$

From (5.76) in the exchangeability representation, $\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathrm{vec}\Big(\mathbf{V}_R\Big)$, where $\mathbf{V}_R$ is specified by the user. From (5.93), the second term is

$$\begin{aligned}\mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{T}\Big)\Big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big]. \end{aligned} \tag{5.95}$$

Also from the exchangeability representation, $\mathbb{E}[\mathcal{R}_i(\mathbf{V})] = \mathbf{0}$ (5.77), and from (5.6), $\mathbb{E}[\mathcal{R}_i(\mathbf{X})] = \mathbf{0}$ and $\mathbb{C}[\mathcal{R}_i(\mathbf{X}), \mathcal{R}_j(\mathbf{X})'] = \mathbf{0}$, so

$$\mathbb{E}\Big[\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big] = \mathbb{E}[\mathcal{R}_i(\mathbf{X})]\,\mathbb{E}[\mathcal{R}_j(\mathbf{X})]' + \mathbb{C}[\mathcal{R}_i(\mathbf{X}), \mathcal{R}_j(\mathbf{X})] = \mathbf{0}, \tag{5.96}$$

and $\mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{T}\Big)\Big] = \mathbf{0}$. Hence

$$\mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{S}\Big)\Big] = \mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathrm{vec}\Big(\mathbf{V}_R\Big). \tag{5.97}$$

Turning now to the variance of $\text{vec}\big(\mathbf{S}\big)$: from (5.76), $\mathbb{V}\big[\text{vec}\big(\mathcal{M}(\mathbf{V})\big)\big] = \mathbf{V}_M$. For now, let $\mathbb{V}\big[\text{vec}\big(\mathbf{T}\big)\big] = \mathbf{V}_T$; this expression will be expanded in Section 5.3.4. From property 4 in the exchangeability representation (Section 5.3.1) and the fourth-order uncorrelated property (5.80), the elements of $\mathcal{M}(\mathbf{V})$ are uncorrelated with those of $\mathcal{R}_i(\mathbf{V})$ and $\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'$, so that $\mathbb{C}\big[\text{vec}\big(\mathcal{M}(\mathbf{V})\big), \text{vec}\big(\mathbf{T}\big)\big] = \mathbf{0}$. The expectation and variance of $\text{vec}\big(\mathbf{S}\big)$ can therefore now be expressed in terms of specifiable quantities, with

$$\mathbb{E}\big[\text{vec}\big(\mathbf{S}\big)\big] = \text{vec}\big(\mathbf{V}_R\big), \tag{5.98}$$

$$\mathbb{V}\big[\text{vec}\big(\mathbf{S}\big)\big] = \mathbf{V}_M + \mathbf{V}_T, \tag{5.99}$$

$$\mathbb{C}\big[\text{vec}\big(\mathbf{S}\big), \text{vec}\big(\mathcal{M}(\mathbf{V})\big)\big] = \mathbf{V}_M. \tag{5.100}$$

5.3.3.3 Adjusted expectation and variance of vec $(\mathcal{M}(\mathbf{V}))$

The expectation and variance of $\text{vec}\big(\mathbf{S}\big)$ defined above are required in order to carry out Bayes linear adjustment of the prior covariance specifications by the observed residual cross products. Using (5.100) in the update equations (5.2) and (5.3), the adjusted expectation of $\text{vec}\big(\mathcal{M}(\mathbf{V})\big)$ given the vectorised observed sample variance matrix $\text{vec}\big(\mathbf{S}\big)$ is therefore

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\big[\text{vec}\big(\mathcal{M}(\mathbf{V})\big)\big] &= \mathbb{E}\big[\text{vec}\big(\mathcal{M}(\mathbf{V})\big)\big] + \\
&\qquad \mathbb{C}\big[\text{vec}\big(\mathcal{M}(\mathbf{V})\big), \text{vec}\big(\mathbf{S}\big)\big]\, \mathbb{V}\big[\text{vec}\big(\mathbf{S}\big)\big]^{\dagger} \Big(\text{vec}\big(\mathbf{S}\big) - \mathbb{E}\big[\text{vec}\big(\mathbf{S}\big)\big]\Big) \\
&= \text{vec}\big(\mathbf{V}_R\big) + \mathbf{V}_M \left(\mathbf{V}_M + \mathbf{V}_T\right)^{\dagger} \Big(\text{vec}\big(\mathbf{S}\big) - \text{vec}\big(\mathbf{V}_R\big)\Big) \\
&= \Big(\mathbf{I}_{m^2} - \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\Big)\, \text{vec}\big(\mathbf{V}_R\big) + \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\text{vec}\big(\mathbf{S}\big) \\
&= \mathbf{V}_T(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\text{vec}\big(\mathbf{V}_R\big) + \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\text{vec}\big(\mathbf{S}\big),
\end{aligned} \tag{5.101}$$

with corresponding adjusted uncertainty

$$\begin{aligned}\mathbb{V}_{\mathbf{S}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] &= \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] - \\ &\qquad \mathbb{C}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big), \mathrm{vec}\Big(\mathbf{S}\Big)\Big]\, \mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{S}\Big)\Big]^{\dagger}\, \mathbb{C}\Big[\mathrm{vec}\Big(\mathbf{S}\Big), \mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] \\ &= \mathbf{V}_M - \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\mathbf{V}_M. \end{aligned} \tag{5.102}$$

Again, the inverse of $\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{S}\Big)\Big]$ does not exist due to the duplication of elements in $\mathrm{vec}\Big(\mathbf{S}\Big)$, so the generalised inverse is used. The matrices $\mathbf{V}_T$ and $\mathbf{V}_M$ will be defined in the coming sections.

5.3.3.4 Variance-modified Bayes linear adjustments

Having carried out the multivariate adjustment of $\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ and $\mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ by the observed sample variance $\mathbf{S}$, the Bayes linear adjusted expectation and variance of $\mathcal{M}(\mathbf{X})$ are modified by using $\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]$ as an updated estimate of the residual variance of $\mathbf{X}_i$. $\mathbb{V}[\mathcal{R}_i(\mathbf{X})]$ is replaced by $\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]$ in (5.14) and (5.15), to give the variance-modified adjusted expectation and variance, after adjustment by the observed sample mean and covariance matrix:

$$\mathbb{E}_{\mathbf{X}}[\mathcal{M}(\mathbf{X})] = \mathbb{E}[\mathcal{M}(\mathbf{X})] + \mathbb{V}[\mathcal{M}(\mathbf{X})]\left(\mathbb{V}[\mathcal{M}(\mathbf{X})] + \frac{1}{n}\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]\right)^{-1}\left(\overline{\mathbf{X}} - \mathbb{E}[\mathcal{M}(\mathbf{X})]\right) \tag{5.103}$$

and

$$\begin{aligned}\mathbb{V}_{\mathbf{X}}[\mathcal{M}(\mathbf{X})] &= \mathbb{V}[\mathcal{M}(\mathbf{X})] - \mathbb{V}[\mathcal{M}(\mathbf{X})]\left(\mathbb{V}[\mathcal{M}(\mathbf{X})] + \frac{1}{n}\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]\right)^{-1}\mathbb{V}[\mathcal{M}(\mathbf{X})] \\ &= \left(\mathbb{V}[\mathcal{M}(\mathbf{X})]^{-1} + n\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]^{-1}\right)^{-1}. \end{aligned} \tag{5.104}$$

We now turn to expanding $\mathbf{V}_T = \mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{S}\Big)\Big] - \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$, the term introduced in Section 5.3.3.2 to account for the variance of the additional uncertainty in the sample covariance matrix, in terms of specifiable quantities.

5.3.4 Variance of the residual sampling variance $\mathbf{vec}(\mathbf{T}) = \mathbf{vec}\,(\mathbf{S} - \mathcal{M}(\mathbf{V}))$

In Section 5.3.3, expressions for the adjusted expectation and variance of the vectorised sample variance $\mathrm{vec}\big(\mathbf{S}\big)$ were derived in terms of the matrices $\mathbf{V}_R = \mathbb{E}[\mathcal{M}(\mathbf{V})]$, $\mathbf{V}_M = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$, the observed sample covariance matrix $\mathbf{S}$ and the matrix $\mathbf{V}_T$, which represents the sampling uncertainty of $\mathbf{S}$. The problem of expressing $\mathbf{V}_T = \mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{S} - \mathcal{M}(\mathbf{V})\Big)\Big]$ in terms of the quantities introduced in Section 5.3.1 is now considered. From (5.93),

$$\begin{aligned}\mathbf{V}_T &= \mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{T}\Big)\Big] \\ &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right]. \qquad (5.105)\end{aligned}$$

First, note that the double summation in the second term can be expanded into

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) &= \sum_{i=1}^{n}\left\{\sum_{j<i}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) + \sum_{j>i}\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right\} \\ &= \sum_{i=1}^{n}\sum_{j>i}\Big\{\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) + \mathrm{vec}\Big(\mathcal{R}_j(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big)\Big\}. \qquad (5.106)\end{aligned}$$

The following steps require the use of special matrices known as commutation matrices (see Appendix C.4), defined as

$$\mathbf{K}_{m,m} = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_j\mathbf{e}_i', \qquad (5.107)$$

where $\mathbf{e}_i\mathbf{e}_j$ is an $m \times m$ matrix with only one nonzero element, a one in the (i,j)th position. This matrix has the useful property that, for any $m \times m$ matrix

$\mathbf{A}$, $\text{vec}\Big(\mathbf{A}\Big) = \mathbf{K}_{m,m}\text{vec}\Big(\mathbf{A}'\Big)$, which allows us to write (5.106) as

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j\neq i}\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) &= \sum_{i=1}^{n}\sum_{j>i}\Big\{\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) + \mathbf{K}_{m,m}\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big\} \\ &= \sum_{i=1}^{n}\sum_{j>i}\left(\mathbf{I}_{m^2} + \mathbf{K}_{m,m}\right)\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big) \\ &= \sum_{i=1}^{n}\sum_{j>i} 2\mathbf{N}_m\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big), \qquad (5.108)\end{aligned}$$

where $\mathbf{N}_m = \frac{1}{2}(\mathbf{I}_{m^2} + \mathbf{K}_{m,m})$ as defined in (C.25). The original expression for $\mathbf{V}_T = \mathbb{V}\Big[\text{vec}\Big(\mathbf{T}\Big)\Big]$ in (5.105) can therefore be written as

$$\mathbf{V}_T = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}\mathbf{N}_m\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right]. \qquad (5.109)$$

This change of summation is analogous to that carried out in (5.27) in the scalar case.

Next, from (5.74) in the exchangeability representation we know that

$$\begin{aligned}&\mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] = \\ &\qquad \mathbb{C}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big) - \text{vec}\Big(\mathcal{M}(\mathbf{V})\Big), \text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big].\end{aligned}$$

In conjunction with the fourth-order uncorrelated properties listed in (5.80)-(5.82), this shows that the covariance between the two terms on the right-hand side of (5.109) is zero, hence

$$\mathbf{V}_T = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] + \frac{4}{n^2(n-1)^2}\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j>i}\mathbf{N}_m\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right]. \qquad (5.110)$$

Also from (5.82), cross-product residual matrices from different (i, j) pairs

of individuals are uncorrelated; thus

$$\mathbb{V}\left[\sum_{i=1}^{n}\sum_{j>i}\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\right] = \sum_{i=1}^{n}\sum_{j>i}\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big],$$

so that

$$\begin{aligned}\mathbf{V}_T &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] + \frac{4}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j>i}\mathbb{V}\Big[\mathbf{N}_m\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] \\ &= \frac{1}{n}\mathbf{V}_{R(V)} + \frac{2}{n(n-1)}\mathbf{N}_m\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big]\mathbf{N}_m, \qquad (5.111)\end{aligned}$$

where $\mathbf{V}_{R(V)} = \mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big]$ as defined in Section 5.3.1. Specification of $\mathbf{V}_{R(V)}$ will be considered in Section 5.3.5.3.

5.3.4.1 Variance of vec$(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})')$

The aim of this section is to express $\mathbf{V}_T$ – and hence ultimately the adjusted expectation and variance of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ derived in Section 5.3.3.3 – solely in terms of the specifiable elements described in Section 5.3.1. In order to achieve this, an expression for $\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big]$, the variance of the cross product of residuals from two different individuals in (5.111), is still required, and will now be derived. This derivation again makes use of matrix identities reviewed in Appendix C.

Using the identity $\mathbb{V}[\mathbf{Y}] = \mathbb{E}\Big[\mathbf{Y}\mathbf{Y}'\Big] - \mathbb{E}[\mathbf{Y}]\,\mathbb{E}[\mathbf{Y}]'$ gives

$$\begin{aligned}\mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] = \mathbb{E}\Big[\Big(\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big)\Big(\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big)'\Big] - \\ \mathbb{E}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big]\,\mathbb{E}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big]'.\end{aligned}$$

From (5.96), $\mathbb{E}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] = \mathbf{0}$ for $i \neq j$.

Hence, using the identities $\text{vec}\Big(\mathbf{x}\mathbf{y}'\Big) = \mathbf{y}\otimes\mathbf{x}$, $(\mathbf{A}\otimes\mathbf{B})' = \mathbf{A}'\otimes\mathbf{B}'$ and $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{A}\mathbf{C}\otimes\mathbf{B}\mathbf{D}$ for vectors $\mathbf{x}$ and $\mathbf{y}$ and matrices $\mathbf{A}$,$\mathbf{B}$,$\mathbf{C}$ and $\mathbf{D}$

– respectively, equations (C.2), (C.4) and (C.8) in Appendix C – we have

$$\begin{aligned}\mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] &= \mathbb{E}\Big[\Big(\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big)\Big(\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big)'\Big] \\ &= \mathbb{E}\Big[\Big(\mathcal{R}_j(\mathbf{X})\otimes\mathcal{R}_i(\mathbf{X})\Big)\Big(\mathcal{R}_j(\mathbf{X})\otimes\mathcal{R}_i(\mathbf{X})\Big)'\Big] \\ &= \mathbb{E}\Big[\mathcal{R}_j(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\otimes\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big]. \qquad (5.112)\end{aligned}$$

Each term in the Kronecker product now describes the residual cross-product matrix for a single individual. This means that (5.112) can be decomposed using the exchangeability representation (5.74) into

$$\begin{aligned}\mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] &= \mathbb{E}\Big[\Big\{\mathcal{M}(\mathbf{V})+\mathcal{R}_j(\mathbf{V})\Big\}\otimes\Big\{\mathcal{M}(\mathbf{V})+\mathcal{R}_i(\mathbf{V})\Big\}\Big] \\ &= \mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]+\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{R}_i(\mathbf{V})]+ \\ &\qquad \mathbb{E}[\mathcal{R}_j(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]+\mathbb{E}[\mathcal{R}_j(\mathbf{V})\otimes\mathcal{R}_i(\mathbf{V})]. \qquad (5.113)\end{aligned}$$

Again from the exchangeability representation in Section 5.3.1, $\mathbb{E}[\mathcal{R}_i(\mathbf{V})] = \mathbb{E}[\mathcal{R}_j(\mathbf{V})] = \mathbf{0}$ and $\mathbb{C}[\mathcal{R}_i(\mathbf{V}),\mathcal{M}(\mathbf{V})] = \mathbb{C}[\mathcal{R}_j(\mathbf{V}),\mathcal{M}(\mathbf{V})] = \mathbb{C}[\mathcal{R}_i(\mathbf{V}),\mathcal{R}_j(\mathbf{V})] = \mathbf{0}$, so that the second, third and fourth terms are all zero, and

$$\mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_j(\mathbf{X})'\Big)\Big] = \mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]. \qquad (5.114)$$

Hence the expression for $\mathbf{V}_T$ derived in (5.111) can be written as

$$\mathbf{V}_T = \frac{1}{n}\mathbf{V}_{R(V)} + \frac{2}{n(n-1)}\mathbf{N}_m\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]\,\mathbf{N}_m. \qquad (5.115)$$

From (C.29) we have $\mathbf{N}_m\,(\mathbf{B}\otimes\mathbf{B})\,\mathbf{N}_m = \mathbf{N}_m\,(\mathbf{B}\otimes\mathbf{B})$, so this expression can be further simplified to

$$= \frac{1}{n}\mathbf{V}_{R(V)} + \frac{2}{n(n-1)}\mathbf{N}_m\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]. \qquad (5.116)$$

The expectation $\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]$ is not easily expressed directly in terms of the quantities $\mathbf{V}_R$, $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ about which belief statements can be

made. However, using the identities given in Appendices C.3 and C.4, it can be shown that the elements of $\mathcal{M}(\mathbf{V}) \otimes \mathcal{M}(\mathbf{V})$ are a permutation of those of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\,\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)'$, the expectation of which is easily expressed in terms of these quantities as

$$\begin{aligned}\mathbb{E}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\,\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)'\Big] &= \mathbb{V}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] + \mathbb{E}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]\,\mathbb{E}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]' \\ &= \mathbf{V}_M + \text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'. \end{aligned} \tag{5.117}$$

Let $\mathbf{V}_M{}^*$ denote the unique permutation of the elements of $\mathbf{V}_M$ – the details of which are derived in Appendix D.1 – such that

$$\mathbb{E}[\mathcal{M}(\mathbf{V}) \otimes \mathcal{M}(\mathbf{V})] = \mathbf{V}_M{}^* + \mathbf{V}_R \otimes \mathbf{V}_R. \tag{5.118}$$

Equation (5.116) can now be expressed in terms of the specifiable quantities $\mathbf{V}_R$, $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ as

$$\mathbf{V}_T = \frac{1}{n}\mathbf{V}_{R(V)} + \frac{2}{n(n-1)}\mathbf{N}_m\left(\mathbf{V}_M{}^* + \mathbf{V}_R \otimes \mathbf{V}_R\right). \tag{5.119}$$

No further simplification of this expression is possible at present. However, recall from Section 5.3.3.3 that the expectation and variance of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ adjusted by $\mathbf{S}$ are, respectively,

$$\mathbb{E}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbf{V}_T(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\text{vec}\Big(\mathbf{V}_R\Big) + \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\text{vec}\Big(\mathbf{S}\Big) \tag{5.120}$$

and

$$\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbf{V}_M - \mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\mathbf{V}_M. \tag{5.121}$$

With $\mathbf{V}_T$ defined as in (5.119) the adjusted expectation and variance of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ can now be written in terms of the observed sample covariance $\mathbf{S}$ and quantities $\mathbf{V}_R$, $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$, where $\mathbf{V}_R = \mathbb{E}[\mathcal{M}(\mathbf{V})]$ is a valid covariance

matrix specified by the user; $\mathbf{V}_M$ is the prior variance of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$; and $\mathbf{V}_{R(V)} = \mathbb{V}\Big[\text{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{M}(\mathbf{V})\Big)\Big]$ is the variance of the residual variance for any one individual. Appropriate specifications for $\mathbf{V}_{R(V)}$ and $\mathbf{V}_M$ are considered in the next section.

5.3.5 Specifying higher-order priors

In the univariate case described in Section 5.2.5.2.1, specification of the fourth-order priors is relatively straightforward: the marginal kurtosis is controlled through the parameter $\mathbb{V}\Big[Z_i^2\Big]$ in $V_{R(V)} = \mathbb{V}\Big[Z_i^2\Big]\,(V_M + V_R^2)$, with confidence in the prior variance reflected in $V_M = \frac{2}{\nu-4}V_R^2$ through the degrees of freedom ν. This approach is now generalised to the multivariate setting, as part of a framework that suggests a tractable form for $\mathbb{V}\Big[\text{vec}\Big(\mathbf{z}_i\mathbf{z}_i'\Big)\Big]$, the multivariate extension of $\mathbb{V}\Big[Z_i^2\Big]$, and also generalises the relationship between $\mathbf{V}_M$ and $\mathbf{V}_R$, while retaining the same intuitive parameterisations used in the univariate setting.

5.3.5.1 A convenient paradigm: elliptical distributions

In the univariate setting, following Goldstein and Wooff (2007, §8.3), a value of $\mathbb{V}\Big[Z_i^2\Big]$ was chosen to reflect the kurtosis of known parametric distributions: in Section 5.2.5.1.1, the normal and t distributions were suggested as possible templates for elicitation of a specific value. Both of these distributions and their multivariate (vector-variate) and matrix-variate forms are members of the broader class of elliptical distributions.

The random variable $\mathbf{X}$ is a member of the class of multivariate elliptical distributions if its characteristic function – the Fourier transform of its probability density function – has the form $\phi(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}}\,\psi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$ (Kelker, 1970; Muirhead, 2009). The class of multivariate elliptical distributions is parameterised by the location $\boldsymbol{\mu}$, spread matrix $\boldsymbol{\Sigma}$ and function ψ; a variable $\mathbf{X}$ having a distribution of this form is denoted by $\mathbf{X} \sim E\,(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \psi)$.

Informally, multivariate elliptical distributions are a class of symmetric distributions whose contours of equal density have the same elliptical shape as

those of the multivariate normal distribution, and whose marginal distributions all have the same functional form. They have a number of convenient properties (Bentler, 1983; Muirhead, 2009):

1. Provided the first moment exists, $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$.

2. Provided the second moment exists, $\mathbb{V}[\mathbf{X}] = \alpha\boldsymbol{\Sigma}$, where

$$\alpha = -2\psi'(\mathbf{0}) \qquad \psi'(\mathbf{0}) = \left.\frac{d\psi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})}{d\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}\right|_{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}=\mathbf{0}} \tag{5.122}$$

3. The marginal distributions of $\mathbf{X}$ all have zero skewness.

4. The marginal distributions of $\mathbf{X}$ all have the same excess kurtosis, which can be expressed in terms of a kurtosis parameter κ as

$$3\kappa = \mathbb{K}\mathrm{ur}[\mathbf{X}] = \frac{\left(\psi''(\mathbf{0}) - \psi'(\mathbf{0})^2\right)}{\psi'(\mathbf{0})^2}, \quad \psi''(\mathbf{0}) = \left.\frac{d^2\psi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})}{d\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}^2}\right|_{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}=\mathbf{0}} \tag{5.123}$$

   where $\mathbb{K}\mathrm{ur}[\mathbf{X}]$ denotes the excess kurtosis with respect to a normal distribution, as defined in Section 5.2.5.1.1. Note that κ here is unrelated to the quantity defined in (5.54).

5. All fourth-order central moments are determined by κ and the second-order moments, with

$$\mathbb{E}[(X_p - \mu_p)(X_q - \mu_q)(X_r - \mu_r)(X_s - \mu_s)] = \alpha^2(\kappa + 1)\left(\sigma_{pq}\sigma_{rs} + \sigma_{pr}\sigma_{qs} + \sigma_{ps}\sigma_{qr}\right), \tag{5.124}$$

   where X_p is the pth element of $\mathbf{X}$ and σ_{pq} is the (p,q)th element of $\boldsymbol{\Sigma}$.

From the properties above, it is clear that the first four moments of any multivariate elliptical distribution are fully determined by $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the scalar parameters α and κ. This is extremely useful in the Bayes linear framework, because a fourth-order approximation of any elliptical distribution can be made by appropriate choices of α and κ. The fourth-order moments can be written

in matrix form as

$$\begin{aligned}\mathbb{E}\Big[\mathbf{Y}\mathbf{Y}'\otimes\mathbf{Y}\mathbf{Y}'\Big] &= \alpha^2(\kappa+1)\Big\{\mathrm{vec}\Big(\mathbf{\Sigma}\Big)\,\mathrm{vec}\Big(\mathbf{\Sigma}\Big)' + 2\mathbf{N}_m\left(\mathbf{\Sigma}\otimes\mathbf{\Sigma}\right)\Big\} \\ &= (\kappa+1)\Big\{\mathrm{vec}\Big(\mathbb{V}[\mathbf{Y}]\Big)\,\mathrm{vec}\Big(\mathbb{V}[\mathbf{Y}]\Big)' + 2\mathbf{N}_m\left(\mathbb{V}[\mathbf{Y}]\otimes\mathbb{V}[\mathbf{Y}]\right)\Big\}\end{aligned} \tag{5.125}$$

and the variance of $\mathrm{vec}\Big(\mathbf{Y}\mathbf{Y}'\Big)$ as

$$\begin{aligned}\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Y}\mathbf{Y}'\Big)\Big] &= \alpha^2\Big\{\kappa\mathrm{vec}\Big(\mathbf{\Sigma}\Big)\,\mathrm{vec}\Big(\mathbf{\Sigma}\Big)' + 2(\kappa+1)\mathbf{N}_m\left(\mathbf{\Sigma}\otimes\mathbf{\Sigma}\right)\Big\} \\ &= \kappa\mathrm{vec}\Big(\mathbb{V}[\mathbf{Y}]\Big)\,\mathrm{vec}\Big(\mathbb{V}[\mathbf{Y}]\Big)' + 2(\kappa+1)\mathbf{N}_m\left(\mathbb{V}[\mathbf{Y}]\otimes\mathbb{V}[\mathbf{Y}]\right).\end{aligned} \tag{5.126}$$

Derivations of these matrix forms are given in Appendix D.2.

These results will now be used to extend the second-order exchangeability representation used in Section 5.2.5.2.1 to specify the scalar quantities V_M and $V_{R(V)}$ to their multivariate analogues, $\mathbf{V}_M = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ and $\mathbf{V}_{R(V)} = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{M}(\mathbf{V})\Big)\Big]$.

5.3.5.2 Representing confidence: $\mathbf{V}_M$

In the univariate case presented in Section 5.2.5.2.1, uncertainty about the population variance was specified by treating V_R and V_M as the expectation and variance of a scaled inverse Chi-squared random variable; this implies that $V_M = cV_R^2$, with $c = 2/(\nu-4)$ determined by the degrees of freedom ν used in estimating V_R.

The same approach to specifying the relationship between $\mathbf{V}_R$ and $\mathbf{V}_M$ is used here, with $\mathcal{M}(\mathbf{V})$ treated as though it were a sample covariance matrix estimated from β observations of a vector random variable $\mathbf{Y}_i \overset{iid}{\sim} E(\mathbf{0}, \mathbf{\Psi}, \psi)$, where $\mathbf{\Psi}$ is a symmetric positive semi-definite scaling matrix. Then

$$\mathbf{V}_M = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbb{V}\left[\mathrm{vec}\left(\frac{1}{\beta}\sum_{i=1}^{\beta}\mathbf{Y}_i\mathbf{Y}_i'\right)\right] = \frac{1}{\beta}\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Y}_i\mathbf{Y}_i'\Big)\Big], \tag{5.127}$$

and obtaining $\mathbb{V}\Big[\text{vec}\Big(\mathbf{Y}_i\mathbf{Y}_i'\Big)\Big]$ from (5.126),

$$\begin{aligned}\mathbf{V}_M &= \frac{\alpha^2}{\beta}\Big\{\kappa\text{vec}\Big(\mathbf{\Psi}\Big)\,\text{vec}\Big(\mathbf{\Psi}\Big)' + 2(\kappa+1)\mathbf{N}_m\,(\mathbf{\Psi}\otimes\mathbf{\Psi})\Big\} \\ &= \frac{1}{\beta}\Big\{\gamma\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2(\gamma+1)\mathbf{N}_m\,(\mathbf{V}_R\otimes\mathbf{V}_R)\Big\}, \qquad (5.128)\end{aligned}$$

say, where $\mathbf{V}_R = \alpha\mathbf{\Psi} = \mathbb{V}[\mathbf{Y}_i]$. The parameter γ will be used to denote the excess marginal kurtosis in the particular case where the specification of $\mathbf{V}_M$ is being considered, to avoid confusion with the user-specified kurtosis parameter κ that will be introduced shortly.

Having expressed $\mathbf{V}_M$ in this form, it is now possible to express the permuted matrix $\mathbf{V}_M{}^*$ introduced in (5.3.4.1) similarly in terms of vec and Kronecker products of $\mathbf{V}_R$:

$$\mathbf{V}_M{}^* = \frac{1}{\beta}\Big\{\gamma\,(\mathbf{V}_R\otimes\mathbf{V}_R) + (\gamma+1)\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + (\gamma+1)\mathbf{K}_{m,m}\,(\mathbf{V}_R\otimes\mathbf{V}_R)\Big\}. \qquad (5.129)$$

The full derivation of this form is given in Appendix D.1.2.

(5.128) could be used to specify $\mathbf{V}_M$ to emulate the variance of the covariance matrix of a sample drawn from any multivariate elliptical distribution, by expressing prior beliefs about the shape of the population variance in terms of the marginal kurtosis parameter γ and a notional sample size β. However, as in the scalar specification in Section 5.2.5.2.1, when specifying $\mathbf{V}_M$ it is convenient to separate the effect of sample size from that of the marginal kurtosis, which will be specified through $\mathbf{V}_{R(V)}$ in the next section.

This separation is achieved by treating $\mathcal{M}(\mathbf{V})$ as though it were the covariance matrix of a multivariate normal distribution, and adopting the form of the natural conjugate prior. The relationship between $\mathbf{V}_R$ and $\mathbf{V}_M$, respectively the prior expectation and variance-covariance matrix of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$, is then specified as if the random variance matrix $\mathcal{M}(\mathbf{V})$ were believed to have an inverse-Wishart distribution with ν degrees of freedom and scatter matrix $\mathbf{\Psi}$

(Mardia et al., 1979), so that

$$\mathbf{V}_R = \mathbb{E}[\mathcal{M}(\mathbf{V})] = \frac{1}{\nu - m - 1}\boldsymbol{\Psi} \tag{5.130}$$

and the covariance between the (i,j)th and (k,l)th elements of $\mathcal{M}(\mathbf{V})$ is

$$\mathbb{C}[\mathcal{M}(V_{ij}), \mathcal{M}(V_{kl})] = \frac{2\psi_{ij}\psi_{kl} + (\nu - m - 1)\left(\psi_{ik}\psi_{jl} + \psi_{il}\psi_{jk}\right)}{(\nu - m)(\nu - m - 1)^2(\nu - m - 3)}. \tag{5.131}$$

It is fairly straightforward, following the steps used in the derivation of the fourth-order moments of $\mathbf{Y}$ in Appendix D.2 to obtain the matrix of fourth-order moments, to arrange these covariances into matrix form, and so to show that

$$\mathbf{V}_M = \mathbb{V}[\mathrm{vec}(\mathcal{M}(\mathbf{V}))] = \frac{2\mathrm{vec}(\boldsymbol{\Psi})\,\mathrm{vec}(\boldsymbol{\Psi})' + 2(\nu - m - 1)\mathbf{N}_m(\boldsymbol{\Psi} \otimes \boldsymbol{\Psi})}{(\nu - m)(\nu - m - 1)^2(\nu - m - 3)} \tag{5.132}$$

$$= \frac{2\mathrm{vec}(\mathbf{V}_R)\,\mathrm{vec}(\mathbf{V}_R)' + 2(\nu - m - 1)\mathbf{N}_m(\mathbf{V}_R \otimes \mathbf{V}_R)}{(\nu - m)(\nu - m - 3)}. \tag{5.133}$$

This has the same form as (5.128), with $\beta = \nu - m$, $\gamma = 2/(\nu - m - 3)$, and (implicitly) $\alpha = 1/(\nu - m - 1)$; but requires that the user specify only a single scalar parameter, ν, with $\nu - m$ reflecting the notional equivalent sample size β from which $\mathcal{M}(\mathbf{V})$ was considered to be estimated in (5.128). Applying the same parametrisation to (5.129) gives the slightly less elegant form

$$\mathbf{V}_M{}^* = \frac{2(\mathbf{V}_R \otimes \mathbf{V}_R) + (\nu - m - 1)\mathrm{vec}(\mathbf{V}_R)\,\mathrm{vec}(\mathbf{V}_R)' + (\nu - m - 1)\mathbf{K}_{m,m}(\mathbf{V}_R \otimes \mathbf{V}_R)}{(\nu - m)(\nu - m - 3)}. \tag{5.134}$$

Interpretation of ν using this specification is relatively simple. Larger values of ν correspond to greater confidence in the prior estimate of the expected value of $\mathcal{M}(\mathbf{V})$, resulting in smaller values for all elements of $\mathbf{V}_M$. A more precise interpretation, which can be used to support elicitation, will be given in Section 5.3.6.1.

5.3.5.3 Representing marginal kurtosis in the residuals: $\mathbf{V}_{R(V)}$

The higher-order moments of elliptical distributions derived in Section 5.3.5.1 are also instrumental in specifying prior beliefs about $\mathbf{V}_{R(V)} = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big]$, the term controlling the shape of the residuals $\mathcal{R}_i(\mathbf{X})$. Recall that in Section 5.2.5, following the approach used by Goldstein and Wooff (2007, §8.3), this quantity was specified as $\mathbb{V}[\mathcal{R}_i(V)] = \Big(V_M + V_R^2\Big)\mathbb{V}\Big[Z_i^2\Big]$, where $\mathbb{V}\Big[Z_i^2\Big]$ represents the marginal excess kurtosis. This expression was derived by treating the residuals $\mathcal{R}_i(X)$ as the product of the population variance $\mathcal{M}(V)$ with standardised residuals Z_i. This approach is now extended to the multivariate case.

Let the $m \times 1$ vector of standardised residuals for the ith observation be $\mathbf{Z}_i = \mathcal{M}(\mathbf{V})^{-1/2}(\mathbf{X}_i - \mathcal{M}(\mathbf{X}))$, where $\mathbf{X}_i \overset{iid}{\sim} E(\mathbf{0}, \mathbf{\Psi}, \psi)$ and $\mathcal{M}(\mathbf{V})^{1/2}$ is the symmetric square root of $\mathcal{M}(\mathbf{V})$, so that

$$\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}, \qquad \mathbb{V}[\mathbf{Z}_i] = \mathbf{I}_m, \qquad \mathbb{C}[\mathbf{Z}_i, \mathbf{Z}_j] = \mathbf{0}, \tag{5.135}$$

where $\mathbf{Z}_i$ is independent of the value of $\mathcal{M}(\mathbf{V})$; hence $\mathcal{R}_i(\mathbf{X}) = \mathcal{M}(\mathbf{V})^{1/2}\,\mathbf{Z}_i$.

From (5.74) in the exchangeability representation in Section 5.3.1, $\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})'\Big) = \mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big) + \mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)$, so

$$\begin{aligned}
\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big) &= \mathrm{vec}\Big(\mathcal{R}_i(\mathbf{X})\mathcal{R}_i(\mathbf{X})' - \mathcal{M}(\mathbf{V})\Big) \\
&= \mathrm{vec}\Big(\Big[\mathcal{M}(\mathbf{V})^{1/2}\,\mathbf{Z}_i\Big]\Big[\mathcal{M}(\mathbf{V})^{1/2}\,\mathbf{Z}_i\Big]' - \mathcal{M}(\mathbf{V})\Big) \\
&= \mathrm{vec}\Big(\mathcal{M}(\mathbf{V})^{1/2}\,\mathbf{Z}_i\mathbf{Z}_i'\mathcal{M}(\mathbf{V})^{1/2\prime} - \mathcal{M}(\mathbf{V})\Big) \\
&= \mathrm{vec}\Big(\mathcal{M}(\mathbf{V})^{1/2}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big)\mathcal{M}(\mathbf{V})^{1/2\prime}\Big).
\end{aligned} \tag{5.136}$$

Also from Section 5.3.1, $\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] = \mathbf{0}$, so $\mathbf{V}_{R(V)} = \mathbb{V}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\Big] = \mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)\,\mathrm{vec}\Big(\mathcal{R}_i(\mathbf{V})\Big)'\Big]$, and

$$\mathbf{V}_{R(V)} = \mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})^{1/2}\,(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m)\,\mathcal{M}(\mathbf{V})^{1/2\prime}\Big)\,\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})^{1/2}\,(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m)\,\mathcal{M}(\mathbf{V})^{1/2\prime}\Big)'\Big]. \tag{5.137}$$

Each of the vectors in this expectation has the form $\text{vec}\Big(\mathbf{ABC}\Big)$: using the identity $\text{vec}\Big(\mathbf{ABC}\Big) = \Big(\mathbf{C}' \otimes \mathbf{A}\Big)\,\text{vec}\Big(\mathbf{B}\Big)$ given in (C.9), these can be written as

$$\text{vec}\Big(\mathcal{M}(\mathbf{V})^{1/2}\left(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\right)\mathcal{M}(\mathbf{V})^{1/2\prime}\Big) = \Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)\,\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big),$$

whence

$$\begin{aligned}\mathbf{V}_{R(V)} &= \mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)\,\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)\Big\{\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)\,\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)\Big\}'\Big] \\ &= \mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)\,\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)\,\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)'\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)'\Big].\end{aligned} \tag{5.138}$$

Because $\mathbf{Z}_i$ is independent of the value of $\mathcal{M}(\mathbf{V})$, by the law of iterated expectation (5.138) is equivalent to

$$\mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)\,\mathbb{E}\Big[\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)\,\text{vec}\big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\big)'\Big]\Big(\mathcal{M}(\mathbf{V})^{1/2} \otimes \mathcal{M}(\mathbf{V})^{1/2}\Big)'\Big]. \tag{5.139}$$

Expanding the inner expectation over $\mathbf{Z}_i\mathbf{Z}_i'$,

$$\begin{aligned}\mathbb{E}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big)\,\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big)'\Big] = \\ \mathbb{E}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\,\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)'\Big] - \text{vec}\Big(\mathbb{E}\Big[\mathbf{Z}_i\mathbf{Z}_i'\Big]\Big)\,\text{vec}\Big(\mathbf{I}_m\Big)' - \\ \text{vec}\Big(\mathbf{I}_m\Big)\,\text{vec}\Big(\mathbb{E}\Big[\mathbf{Z}_i\mathbf{Z}_i'\Big]\Big)' + \text{vec}\Big(\mathbf{I}_m\Big)\,\text{vec}\Big(\mathbf{I}_m\Big)'.\end{aligned} \tag{5.140}$$

Recall that by definition, $\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}$ and $\mathbb{V}[\mathbf{Z}_i] = \mathbf{I}_m$: hence $\mathbb{E}\Big[\mathbf{Z}_i\mathbf{Z}_i'\Big] = \mathbb{V}[\mathbf{Z}_i] + \mathbb{E}[\mathbf{Z}_i]\,\mathbb{E}[\mathbf{Z}_i]' = \mathbf{I}_m$. The first term in (5.140) is therefore

$$\begin{aligned}\mathbb{E}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\,\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)'\Big] &= \mathbb{V}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big] + \mathbb{E}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big]\,\mathbb{E}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big]' \\ &= \mathbb{V}\Big[\text{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big] + \text{vec}\Big(\mathbf{I}_m\Big)\,\text{vec}\Big(\mathbf{I}_m\Big)',\end{aligned} \tag{5.141}$$

while

$$\mathrm{vec}\Big(\mathbb{E}\Big[\mathbf{Z}_i\mathbf{Z}_i'\Big]\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)' = \mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbb{E}\Big[\mathbf{Z}_i\mathbf{Z}_i'\Big]\Big)' = \mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)', \tag{5.142}$$

and therefore

$$\mathbb{E}\Big[\mathrm{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i' - \mathbf{I}_m\Big)'\Big] = \mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big]. \tag{5.143}$$

Substituting this into the expression for $\mathbf{V}_{R(V)}$ in (5.138) gives

$$\mathbf{V}_{R(V)} = \mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)\,\mathbb{V}[\mathrm{vec}(\mathbf{Z}_i\mathbf{Z}_i')]\,\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)'\Big]. \tag{5.144}$$

Recall that in (5.126) a general expression was given for $\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Y}\mathbf{Y}'\Big)\Big]$ for elliptically distributed $\mathbf{Y}$ in terms of $\mathbb{V}[\mathbf{Y}]$. Here, we require $\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big]$, where $\mathbb{V}[\mathbf{Z}_i] = \mathbf{I}_m$. Hence

$$\begin{aligned}\mathbb{V}\Big[\mathrm{vec}\Big(\mathbf{Z}_i\mathbf{Z}_i'\Big)\Big] &= \kappa\mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)' + 2(\kappa+1)\mathbf{N}_m\,(\mathbf{I}_m\otimes\mathbf{I}_m)\\ &= \kappa\mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)' + 2(\kappa+1)\mathbf{N}_m.\end{aligned} \tag{5.145}$$

All of the required elements are now available to define $\mathbf{V}_{R(V)}$, the term controlling the shape of the residuals $\mathcal{R}_i(\mathbf{X})$. Combining (5.144) and (5.145),

$$\begin{aligned}\mathbf{V}_{R(V)} = \mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)\Big(&\kappa\mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)' +\\ &2(\kappa+1)\mathbf{N}_m\Big)\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)'\Big].\end{aligned} \tag{5.146}$$

For the contribution from the first component in the central term of (5.146), the identity $(\mathbf{C}'\otimes\mathbf{A})\mathrm{vec}\Big(\mathbf{I}_m\Big) = \mathrm{vec}\Big(\mathbf{AC}\Big)$ from (C.9) is again required:

$$\begin{aligned}&\kappa\mathbb{E}\Big[\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)\,\mathrm{vec}\Big(\mathbf{I}_m\Big)'\Big(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\Big)'\Big]\\ &\qquad = \kappa\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\,\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)'\Big].\end{aligned} \tag{5.147}$$

Because $\mathcal{M}(\mathbf{V})^{1/2}$ is symmetric, the contribution from the second component in the central term has the form $(\mathbf{B}\otimes\mathbf{B})\mathbf{N}_m(\mathbf{B}\otimes\mathbf{B})$; from (C.30), an expression of this form is equal to $\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A})$ where $\mathbf{A}=\mathbf{B}\mathbf{B}'$, so

$$\begin{aligned}&2(\kappa+1)\mathbb{E}\Big[\left(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\right)\mathbf{N}_m\left(\mathcal{M}(\mathbf{V})^{1/2}\otimes\mathcal{M}(\mathbf{V})^{1/2}\right)'\Big]\\&\quad=2(\kappa+1)\mathbf{N}_m\mathbb{E}\Big[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})\Big].\end{aligned}\tag{5.148}$$

Combining (5.147) and (5.148), the required expectation is

$$\mathbf{V}_{R(V)}=\kappa\mathbb{E}\Big[\mathrm{vec}(\mathcal{M}(\mathbf{V}))\,\mathrm{vec}(\mathcal{M}(\mathbf{V}))'\Big]+2(\kappa+1)\mathbf{N}_m\mathbb{E}\big[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})\big].\tag{5.149}$$

Expressions for both $\mathbb{E}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\,\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)'\Big]$ and $\mathbb{E}\Big[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})\Big]$ were given in terms of $\mathbf{V}_R$, $\mathbf{V}_M$ and the permuted matrix $\mathbf{V}_M{}^*$ in equations (5.117) and (5.118) in Section 5.3.4.1. Substituting those expressions into (5.149) gives

$$\mathbf{V}_{R(V)}=\kappa\left(\mathbf{V}_M+\mathrm{vec}(\mathbf{V}_R)\,\mathrm{vec}(\mathbf{V}_R)'\right)+2(\kappa+1)\mathbf{N}_m\left(\mathbf{V}_M{}^*+(\mathbf{V}_R\otimes\mathbf{V}_R)\right).\tag{5.150}$$

5.3.5.4 Expressing $\mathbf{V}_{R(V)}$ in terms of $\mathbf{V}_R$, ν and κ

Recall from (5.133) and (5.134) in Section 5.3.5.2 that $\mathbf{V}_M$ and $\mathbf{V}_M{}^*$ can be expressed in terms of $\mathbf{V}_R$, the expectation of $\mathcal{M}(\mathbf{V})$, and the confidence parameter ν: hence it is now possible to write $\mathbf{V}_{R(V)}$, the dispersion of the individual residual cross-product matrices, in terms of $\mathbf{V}_R$ and ν, plus the marginal kurtosis parameter κ and dimension m, as

$$\begin{aligned}\mathbf{V}_{R(V)}=\kappa\frac{\nu-m-1}{\nu-m-3}\Big\{&\mathrm{vec}(\mathbf{V}_R)\,\mathrm{vec}(\mathbf{V}_R)'+2\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)\Big\}+\\&\frac{\nu-m-1}{\nu-m}\left\{\frac{2}{\nu-m-3}\mathrm{vec}(\mathbf{V}_R)\,\mathrm{vec}(\mathbf{V}_R)'+\frac{\nu-m-1}{\nu-m-3}2\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)\right\}.\end{aligned}\tag{5.151}$$

By gathering all terms involving κ, it is possible to gain some intuition about the expected behaviour of $\mathbf{V}_{R(V)}$ for different values of κ:

$$\mathbf{V}_{R(V)} = \kappa \frac{\nu-m-1}{\nu-m-3}\left\{\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2\mathbf{N}_m\left(\mathbf{V}_R \otimes \mathbf{V}_R\right)\right\} + (\nu-m-1)\mathbf{V}_M. \tag{5.152}$$

Recall that setting $\kappa = 0$ indicates a judgement that the marginal distributions are assumed to have no excess kurtosis with respect to a normal distribution. When $\kappa = 0$ the first term in (5.152) is zero, and $\mathbf{V}_{R(V)}$, the dispersion of the individual residual cross-product matrices, is proportional to $\mathbf{V}_M$, the prior dispersion of the population variance. Setting $\kappa < 0$ will reduce $\mathbf{V}_{R(V)}$ from this baseline, while setting $\kappa > 0$ will increase it.

5.3.5.5 Relationship between scalar & multivariate parameters

The multivariate terms $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ are exact multivariate extensions of the scalar terms V_M and $V_{R(V)}$, although they are parameterised differently. In the scalar case in Section 5.2.5.2,

$$V_M = cV_R^2, \qquad \text{and} \qquad V_{R(V)} = (c+1)\mathbb{V}\Big[Z_i^2\Big]\,V_R^2,$$

where $c = 2/(\nu-4)$; the multivariate equivalents given in (5.133) and (5.152) are

$$\mathbf{V}_M = \frac{1}{\nu-m}\left\{\frac{2}{\nu-m-3}\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + \frac{2(\nu-m-1)}{\nu-m-3}\mathbf{N}_m\left(\mathbf{V}_R \otimes \mathbf{V}_R\right)\right\},$$

$$\mathbf{V}_{R(V)} = \frac{\nu-m-1}{\nu-m-3}\kappa\left\{\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2\mathbf{N}_m\left(\mathbf{V}_R \otimes \mathbf{V}_R\right)\right\} + (\nu-m-1)\mathbf{V}_M.$$

Suppose now that $m = 1$, so that $\mathbf{V}_M$ and $\mathbf{V}_R$ are scalar variances, denoted by V_R and V_M respectively, and $2\mathbf{N}_m = 2$. Then

$$\begin{aligned} V_M &= \frac{1}{\nu-1}\left\{\frac{2}{\nu-4}V_R^2 + \frac{2(\nu-2)}{\nu-4}V_R^2\right\} \\ &= \frac{1}{\nu-1}\left\{\frac{2\nu-2}{\nu-4}\right\}V_R^2 \\ &= \frac{2}{\nu-4}V_R^2, \end{aligned} \tag{5.153}$$

where $2/(\nu-4)$ is the kurtosis parameter $\gamma = 2/(\nu-m-3)$ that was used to fix the inverse-Wishart relationship between $\mathbf{V}_M$ and $\mathbf{V}_R$ in Section 5.3.5.2; hence γ is the multivariate analogue to the scalar parameter c, with $c = \gamma$ exactly when $m = 1$.

Substituting V_R for $\mathbf{V}_R$ in $\mathbf{V}_{R(V)}$ along with (5.153),

$$\mathbf{V}_{R(V)} = \frac{\nu-2}{\nu-4}\kappa\left\{V_R^2 + 2V_R^2\right\} + (\nu-2)\frac{2}{\nu-4}V_R^2 \quad = \frac{\nu-2}{\nu-4}(3\kappa+2)V_R^2. \tag{5.154}$$

Again, $(\nu-2)/(\nu-4) \equiv c+1$, so the remaining term must be

$$\mathbb{V}\left[Z_i^2\right] = 3\kappa + 2 = \mathbb{K}\mathrm{ur}\left[X_i\right] + 2, \tag{5.155}$$

and it is clear that, although the derivation and final form for $\mathbf{V}_M$ and $\mathbf{V}_{R(V)}$ are very different in the multivariate case, the multivariate parameters γ (determined via ν) and 3κ are in fact direct analogues of the scalar parameters c (determined, again, by ν) and $\mathbb{V}\left[Z_i^2\right]$.

5.3.5.6 Expressing $\mathbf{V}_T$ in terms of $\mathbf{V}_R$, ν and κ

It is now also possible to write the expression derived for $\mathbf{V}_T$ in (5.119) in terms of $\mathbf{V}_R$, the confidence parameter ν, kurtosis parameter κ, dimension m and

observed sample size n as

$$\mathbf{V}_T = \frac{\kappa(\nu-m-1)}{n(\nu-m-3)}\left\{\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)\right\} + \frac{\nu-m-1}{n-1}\mathbf{V}_M. \tag{5.156}$$

Recall that $\mathbf{V}_T$ is the variance of $\text{vec}\Big(\mathbf{T}\Big) = \text{vec}\Big(\mathbf{S}-\mathcal{M}(\mathbf{V})\Big)$, and captures the sampling variance of $\mathbf{S}$. When $\kappa = 0$, $\mathbf{V}_T$ is again proportional to $\mathbf{V}_M$, the prior variance of the population variance $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$: for positive κ, $\mathbf{V}_T$ will increase, and for negative κ, it will decrease.

From (5.101) and (5.102) the adjusted expectation and variance of $\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$ are, respectively,

$$\mathbb{E}_{\mathbf{S}}\left[\text{vec}(\mathcal{M}(\mathbf{V}))\right] = \mathbf{V}_T\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\text{vec}(\mathbf{V}_R) + \mathbf{V}_M\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\text{vec}(\mathbf{S}) \tag{5.157}$$

and

$$\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \mathbf{V}_M - \mathbf{V}_M\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\mathbf{V}_M. \tag{5.158}$$

In the case where $\kappa = 0$ and $\mathbf{V}_T \propto \mathbf{V}_M$, the adjusted expectation of $\mathcal{M}(\mathbf{V})$ simplifies to a scalar-weighted sum of $\mathbf{V}_R$ and $\mathbf{S}$, while the adjusted variance is proportional to $\mathbf{V}_M$. It will now be shown that, in fact, this is true of $\mathbb{E}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ for any value of κ; and that $\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ has the same form as $\mathbf{V}_M$ for any value of κ.

5.3.6 The adjusted expectation and variance of $\text{vec}\left(\mathcal{M}(\mathbf{V})\right)$, revisited

Recall from (5.133) and (5.156) that $\mathbf{V}_M$ and $\mathbf{V}_T$ are, respectively,

$$\mathbf{V}_M = \frac{2\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2(\nu-m-1)\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)}{(\nu-m)(\nu-m-3)}$$

and

$$\mathbf{V}_T = \frac{\kappa(\nu-m-1)}{n(\nu-m-3)}\left\{\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)\right\} + \frac{\nu-m-1}{n-1}\mathbf{V}_M.$$

Both of these expressions – and therefore also $\mathbf{V}_M+\mathbf{V}_T$ – can be rearranged into the form

$$\mathbf{V}_M = d_1\mathbf{N}_m\left\{\mathbf{V}_R\otimes\mathbf{V}_R + d_2\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'\right\} \tag{5.159}$$

$$\mathbf{V}_T = f_1\mathbf{N}_m\left\{\mathbf{V}_R\otimes\mathbf{V}_R + f_2\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'\right\} \tag{5.160}$$

$$\mathbf{V}_M+\mathbf{V}_T = c_1\mathbf{N}_m\left\{\mathbf{V}_R\otimes\mathbf{V}_R + c_2\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'\right\}, \tag{5.161}$$

where

$$d_1 = \frac{2(\nu-m-1)}{(\nu-m)(\nu-m-3)}, \tag{5.162}$$

$$d_2 = \frac{1}{\nu-m-1}, \tag{5.163}$$

$$f_1 = \frac{2(\nu-m-1)\left[\kappa(n-1)(\nu-m)+n(\nu-m-1)\right]}{n(n-1)(\nu-m)(\nu-m-3)}, \tag{5.164}$$

$$f_2 = \frac{\kappa(n-1)(\nu-m)+2n}{2\left[\kappa(n-1)(\nu-m)+n(\nu-m-1)\right]}, \tag{5.165}$$

$$c_1 = d_1+f_1, \tag{5.166}$$

$$c_2 = \frac{d_1d_2+f_1f_2}{d_1+f_1}. \tag{5.167}$$

The generalised inverse of an expression of this form is derived in Appendix C.5.2: in particular, the generalised inverse of (5.161) is

$$(\mathbf{V}_M+\mathbf{V}_T)^\dagger = \frac{1}{c_1}\mathbf{N}_m\left\{\mathbf{V}_R^{-1}\otimes\mathbf{V}_R^{-1} - \frac{c_2}{1+c_2m}\text{vec}\Big(\mathbf{V}_R^{-1}\Big)\,\text{vec}\Big(\mathbf{V}_R^{-1}\Big)'\right\}. \tag{5.168}$$

This generalised inverse can now be used to obtain explicit terms for the matrices $\mathbf{V}_M(\mathbf{V}_M+\mathbf{V}_T)^\dagger$ and $\mathbf{V}_T(\mathbf{V}_M+\mathbf{V}_T)^\dagger$ in (5.157) and (5.158). First,

$$\begin{aligned}\mathbf{V}_M(\mathbf{V}_M+\mathbf{V}_T)^\dagger = \frac{d_1}{c_1}\mathbf{N}_m\mathbf{N}_m\Big\{ &\Big(\mathbf{V}_R\otimes\mathbf{V}_R\Big)\Big(\mathbf{V}_R^{-1}\otimes\mathbf{V}_R^{-1}\Big)+\\ &d_2\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R\Big)'\Big(\mathbf{V}_R^{-1}\otimes\mathbf{V}_R^{-1}\Big)-\\ &\frac{c_2}{1+c_2m}\Big(\mathbf{V}_R\otimes\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'-\\ &\frac{c_2d_2}{1+c_2m}\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R\Big)'\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\Big\}.\end{aligned} \tag{5.169}$$

As shown in Section C.5, $\mathbf{N}_m$ is idempotent, so $\mathbf{N}_m\mathbf{N}_m=\mathbf{N}_m$; and from (C.27), we have $\mathbf{N}_m\mathrm{vec}\Big(\mathbf{V}_R\Big)=\mathrm{vec}\Big(\mathbf{V}_R\Big)$. Using identities (C.8)-(C.10),

$$\Big(\mathbf{V}_R\otimes\mathbf{V}_R\Big)\Big(\mathbf{V}_R^{-1}\otimes\mathbf{V}_R^{-1}\Big)=\mathbf{I}_{m^2}, \quad \text{(from C.8)}$$

$$\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R\Big)'\Big(\mathbf{V}_R^{-1}\otimes\mathbf{V}_R^{-1}\Big)=\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)', \quad \text{(from C.9)}$$

$$\Big(\mathbf{V}_R\otimes\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'=\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)', \quad \text{(from C.9)}$$

$$\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R\Big)'\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'=m\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)', \quad \text{(from C.10)}$$

and so

$$\mathbf{V}_M(\mathbf{V}_M+\mathbf{V}_T)^\dagger=\frac{d_1}{c_1}\left\{\mathbf{N}_m+\frac{d_2-c_2}{1+c_2m}\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\right\}. \tag{5.170}$$

Similarly,

$$\mathbf{V}_T(\mathbf{V}_M+\mathbf{V}_T)^\dagger=\frac{f_1}{c_1}\left\{\mathbf{N}_m+\frac{f_2-c_2}{1+c_2m}\mathrm{vec}\Big(\mathbf{V}_R\Big)\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\right\}. \tag{5.171}$$

5.3.6.1 The adjusted expectation $\mathbb{E}_{\mathbf{S}}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right]$ as a scalar-weighted sum

Using (5.171) and (C.10), the first term in (5.157) can now be written as

$$\begin{aligned}\mathbf{V}_T\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\text{vec}\left(\mathbf{V}_R\right) &= \frac{f_1}{c_1}\left\{\mathbf{N}_m+\frac{f_2-c_2}{1+c_2m}\text{vec}\left(\mathbf{V}_R\right)\text{vec}\left(\mathbf{V}_R^{-1}\right)'\right\}\text{vec}\left(\mathbf{V}_R\right)\\ &= \frac{f_1}{c_1}\text{vec}\left(\mathbf{V}_R\right)+\frac{f_1}{c_1}\left(\frac{f_2-c_2}{1+c_2m}\right)\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{V}_R\right)\text{vec}\left(\mathbf{V}_R\right)\\ &= \frac{f_1}{c_1}\text{vec}\left(\mathbf{V}_R\right)+\frac{f_1}{c_1}\left(\frac{f_2-c_2}{1+c_2m}\right)m\text{vec}\left(\mathbf{V}_R\right). \quad (5.172)\end{aligned}$$

Similarly, the second term is

$$\begin{aligned}\mathbf{V}_M\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\text{vec}\left(\mathbf{S}\right) &= \frac{d_1}{c_1}\left\{\mathbf{N}_m+\frac{d_2-c_2}{1+c_2m}\text{vec}\left(\mathbf{V}_R\right)\text{vec}\left(\mathbf{V}_R^{-1}\right)'\right\}\text{vec}\left(\mathbf{S}\right)\\ &= \frac{d_1}{c_1}\text{vec}\left(\mathbf{S}\right)+\frac{d_1}{c_1}\left(\frac{d_2-c_2}{1+c_2m}\right)\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\text{vec}\left(\mathbf{V}_R\right).\end{aligned} \quad (5.173)$$

Thus the adjusted expectation of $\mathcal{M}(\mathbf{V})$ (5.157) can be expressed as a scalar-weighted sum of $\text{vec}\left(\mathbf{V}_R\right)$ and $\text{vec}\left(\mathbf{S}\right)$ as

$$\mathbb{E}_{\mathbf{S}}\left[\text{vec}(\mathcal{M}(\mathbf{V}))\right]=\frac{d_1}{c_1}\text{vec}(\mathbf{S})+\left\{\frac{f_1}{c_1}+\frac{f_1\left(f_2-c_2\right)m+d_1\left(d_2-c_2\right)\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)}{c_1\left(1+c_2m\right)}\right\}\text{vec}(\mathbf{V}_R). \quad (5.174)$$

From (5.162)-(5.167), $d_1(d_2-c_2)=-f_1(f_2-c_2)$, so this can be further simplified to

$$\mathbb{E}_{\mathbf{S}}\left[\text{vec}\left(\mathcal{M}(\mathbf{V})\right)\right]=\frac{d_1}{c_1}\text{vec}\left(\mathbf{S}\right)+\left\{\frac{f_1}{c_1}+\frac{f_1\left(f_2-c_2\right)\left[m-\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]}{c_1\left(1+c_2m\right)}\right\}\text{vec}\left(\mathbf{V}_R\right). \quad (5.175)$$

This means that it is no longer necessary to vectorise the elements of $\mathbf{V}_R$ and $\mathbf{S}$: instead, the adjusted variance-covariance matrix can be obtained directly

using

$$\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})] = \frac{d_1}{c_1}\mathbf{S} + \left\{ \frac{f_1}{c_1} + \frac{f_1\,(f_2 - c_2)\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]}{c_1\,(1 + c_2 m)} \right\} \mathbf{V}_R. \tag{5.176}$$

By replacing c_1, c_2, d_1, d_2, f_1 and f_2 with (5.162)-(5.167), the adjusted expectation can be obtained in terms of the parameters used to specify the adjustment: then the adjusted expectation of $\mathcal{M}(\mathbf{V})$ is

$$\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})] = w_1\mathbf{S} + \left\{ (1 - w_1) + w_2\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right] \right\} \mathbf{V}_R, \tag{5.177}$$

where

$$w_1 = \frac{n(n-1)}{\tilde{\kappa} + n\tilde{\nu}} \quad \text{and} \quad w_2 = \frac{n(n-1)(\nu - m - 3)\tilde{\kappa}}{\left[\tilde{\kappa} + n\tilde{\nu}\right]\left[(m+2)(\nu - m - 1)\tilde{\kappa} + 2(\nu - 1)n\tilde{\nu}\right]}, \tag{5.178}$$

with $\tilde{\kappa} = \kappa(n-1)(\nu - m)$ and $\tilde{\nu} = (\nu - m - 1) + (n-1)$. The adjusted expectation therefore consists of a weighted average of $\mathbf{S}$ and $\mathbf{V}_R$, with an additional contribution from $\mathbf{V}_R$ when κ is non-zero. As noted in Section 5.3.5.4, when $\kappa = 0$ the adjusted expectation of the population variance reduces to

$$\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})] = \frac{n-1}{\tilde{\nu}}\mathbf{S} + \frac{\nu - m - 1}{\tilde{\nu}}\mathbf{V}_R \tag{5.179}$$

and the effect of changing the notional prior sample size $\nu - m$ becomes clear: if $\nu - m$ is equal to the observed sample size n, then the prior and observed covariance matrices are given equal weight in the adjusted expectation. Increasing ν to reflect greater confidence in the prior beliefs expressed in $\mathbf{V}_R$ means that $\mathbf{V}_R$ is given greater weight in the adjusted expectation, while setting $\nu - m < n$ means that $\mathbf{S}$ is given greater weight.

Selection of appropriate non-zero values for κ will be considered in Section 5.3.7.

5.3.6.2 Simplifying the adjusted variance $\mathbb{V}_{\mathbf{S}}\Big[\mathbf{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$

A similar approach can be used to expand (5.158), the adjusted variance of $\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)$, without the need to explicitly compute the generalised inverse. First, note that

$$\begin{aligned}\mathbb{V}_{\mathbf{S}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] &= \mathbf{V}_M - \mathbf{V}_M\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\mathbf{V}_M \\ &= \mathbf{V}_T\left(\mathbf{V}_M+\mathbf{V}_T\right)^{\dagger}\mathbf{V}_M, \end{aligned} \tag{5.180}$$

so that, using using (5.171),

$$\mathbb{V}_{\mathbf{S}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \frac{f_1}{c_1}\left\{\mathbf{N}_m + \frac{f_2-c_2}{1+c_2 m}\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\right\}\mathbf{V}_M. \tag{5.181}$$

Now, using the fact that $\mathbf{N}_m\mathrm{vec}\Big(\mathbf{V}_R\Big) = \mathrm{vec}\Big(\mathbf{V}_R\Big)$ from (C.27) and the idempotence of $\mathbf{N}_m$, along with the definition of $\mathbf{V}_M$ from (5.133), we see that

$$\begin{aligned}\mathbf{N}_m\mathbf{V}_M &= \mathbf{N}_m\left\{\frac{2\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)' + 2(\nu-m-1)\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)}{(\nu-m)(\nu-m-3)}\right\} \\ &= \frac{2\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)' + 2(\nu-m-1)\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right)}{(\nu-m)(\nu-m-3)} \\ &= \mathbf{V}_M. \end{aligned} \tag{5.182}$$

To expand $\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\mathbf{V}_M$, we again make use of identities (C.10) and (C.9), along with (C.27), to see that

$$\begin{aligned}\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\,\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)' &= \mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{V}_R\right)\mathrm{vec}\Big(\mathbf{V}_R\Big)' \\ &= m\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)' \end{aligned} \tag{5.183}$$

and

$$\begin{aligned}\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R^{-1}\Big)'\mathbf{N}_m\left(\mathbf{V}_R\otimes\mathbf{V}_R\right) &= \mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\mathbf{V}_R^{-1}\mathbf{V}_R\Big)' \\ &= \mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)'. \end{aligned} \tag{5.184}$$

Hence

$$\begin{aligned}\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R^{-1}\Big)'\,\mathbf{V}_M &= \frac{2m\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)' + 2(\nu-m-1)\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'}{(\nu-m)(\nu-m-3)} \\ &= \frac{2(\nu-1)\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'}{(\nu-m)(\nu-m-3)}, \end{aligned} \tag{5.185}$$

and (5.181) can be written as

$$\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \frac{f_1}{c_1}\mathbf{V}_M + \frac{f_1}{c_1}\frac{f_2-c_2}{1+c_2m}\frac{2(\nu-1)\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'}{(\nu-m)(\nu-m-3)}. \tag{5.186}$$

Comparing this to (5.176), the adjusted expectation of $\mathcal{M}(\mathbf{V})$, we see that the scalar coefficients involving c_1, c_2, f_1 and f_2 are the same, and so the adjusted variance of $\mathcal{M}(\mathbf{V})$ can be written as

$$\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = (1-w_1)\mathbf{V}_M + \frac{2w_2(\nu-1)\text{vec}\Big(\mathbf{V}_R\Big)\,\text{vec}\Big(\mathbf{V}_R\Big)'}{(\nu-m)(\nu-m-3)}, \tag{5.187}$$

where

$$w_1 = \frac{n(n-1)}{\tilde{\kappa}+n\tilde{\nu}} \quad \text{and} \quad w_2 = \frac{n(n-1)(\nu-m-3)\tilde{\kappa}}{\Big[\tilde{\kappa}+n\tilde{\nu}\Big]\Big[(m+2)(\nu-m-1)\tilde{\kappa}+2(\nu-1)n\tilde{\nu}\Big]},$$

with $\tilde{\kappa} = \kappa(n-1)(\nu-m)$ and $\tilde{\nu} = (\nu-m-1)+(n-1)$, as in (5.178). As already noted, $w_2 = 0$ when $\kappa = 0$, implying that the second term captures fourth-order variability due to excess kurtosis; in this case, the adjusted variance can be simplified significantly to

$$\mathbb{V}_{\mathbf{S}}\Big[\text{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \frac{\nu-m-1}{\nu-m-1+n-1}\mathbf{V}_M, \tag{5.188}$$

and the adjusted variance of $\mathcal{M}(\mathbf{V})$ is simply the prior variance scaled by a factor determined by the prior and observed sample sizes.

5.3.7 Accommodating non-zero kurtosis in the adjusted expectation of $\mathcal{M}(\mathbf{V})$

A critical part of the Bayes linear framework is its ability to incorporate a judgement of non-zero marginal kurtosis in the adjustment of the population variance, through the scalar parameter κ. Recall from (5.177)-(5.178) that the adjusted expectation of $\mathcal{M}(\mathbf{V})$ is

$$\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})] = w_1\mathbf{S} + \Big\{(1-w_1) + w_2\Big[m - \mathrm{tr}\Big(\mathbf{V}_R^{-1}\mathbf{S}\Big)\Big]\Big\}\mathbf{V}_R, \tag{5.189}$$

where

$$w_1 = \frac{n(n-1)}{\tilde{\kappa}+n\tilde{\nu}} \quad \text{and} \quad w_2 = \frac{n(n-1)(\nu-m-3)\tilde{\kappa}}{\big[\tilde{\kappa}+n\tilde{\nu}\big]\big[(m+2)(\nu-m-1)\tilde{\kappa}+2(\nu-1)n\tilde{\nu}\big]}, \tag{5.190}$$

with $\tilde{\kappa} = \kappa(n-1)(\nu-m)$ and $\tilde{\nu} = (\nu-m-1)+(n-1)$. Changing the value of κ modifies both of the scaling factors w_1 and w_2, and so alters the relative contributions of the prior $\mathbf{V}_R$ and the observed sample covariance matrix vSn to $\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]$, as well as adjusting the value of $\mathbb{V}_{\mathbf{S}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big]$ given by (5.177).

In this section a range of plausible values for κ are identified, and the implications of particular choices of the parameter are considered.

5.3.7.1 A plausible range for κ

The role of κ is essentially to allow for potentially surprising values of $\mathbf{X}$ in the observed sample: setting $\kappa > 0$ can be interpreted as a judgement that there is likely to be a high proportion of outliers in the data, while setting $\kappa < 0$ reflects a belief that the observations are unlikely to fall very far from the mean or very close to it. Elicitation of κ may be simplified by considering the expected shape of the residuals in terms of known parametric distributions, and choosing the appropriate value of κ accordingly. Table 5.1 shows a range of plausible candidate distributions, together with their excess marginal kurtoses $\mathbb{K}\mathrm{ur}[\mathbf{X}]$ and corresponding κ values; as described in Section 5.3.5.1, $\kappa = \mathbb{K}\mathrm{ur}[\mathbf{X}]/3$.

The range of plausible values of κ is fairly small. It has been shown that

any regular unimodal symmetric distribution has excess kurtosis no smaller than -1.2 (Stuart et al., 1994), suggesting that an operational lower bound for κ should be around -0.4. For positive κ, setting $\kappa = 2$ implies a judgement that the tails of $\mathbf{X}$ are as heavy as those of a Student-t distribution with five degrees of freedom. As κ increases beyond 2, the tails of $\mathbf{X}$ may be considered to resemble those of a Student-t distribution with degrees of freedom approaching arbitrarily close to four; κ is undefined for a t distribution with four or fewer degrees of freedom.

Table 5.1: Selected values of κ and the equivalent marginal kurtosis specification, expressed in terms of representative parametric distributions.

$\mathbf{X} \sim$	U	N	t_{10}	t_7	t_6	t_5
$\mathbb{K}\mathrm{ur}[\mathbf{X}]$	-1.2	0	1	2	3	6
κ	-0.4	0	$\frac{1}{3}$	$\frac{2}{3}$	1	2

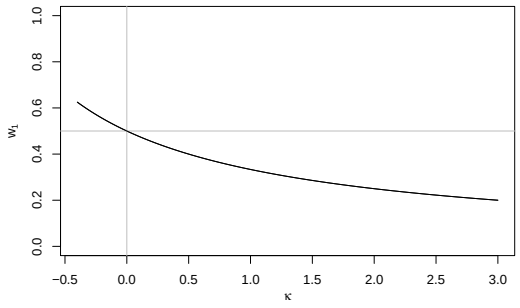
Increasing the value of κ decreases the scaling factor w_1 applied to $\mathbf{S}$ in (5.189), as Figure 5.1a shows, with a corresponding increase in the contribution from $\mathbf{V}_R$ in the second term. Changes in κ also affect the adjusted expectation of $\mathcal{M}(\mathbf{V})$ through w_2 in the third term in (5.189), which is non-zero only for $\kappa \neq 0$.

Figure 5.1b shows the scaling factor w_2 as a function of κ when $n = 25$, $m = 13$, and $\nu = 13 + 25$; these are, respectively, the observed sample size, dimension and confidence parameter (notional sample size + dimension) used in the application in Chapter 6, but the shape of the function will be similar for any choices of these values. For positive κ, w_2 initially increases fairly rapidly to its maximum, after which it begins to decay slowly. The exact value that maximimises w_2 can be found by differentiating the expression given in (5.190) with respect to κ; for the chosen parameters, the maximum of w_2 is reached at $\kappa \approx 0.9$. The change in w_2 around this maximum is very small; for example, with the parameters used here, setting $0.24 \leq \kappa \leq 3.46$ results in $0.01 \leq w_2 \leq 0.015$. w_2 is negative for $\kappa < 0$, and decreases rapidly as κ decreases; Figure 5.1b is truncated for clarity at $\kappa = -0.2$, but for the minimum

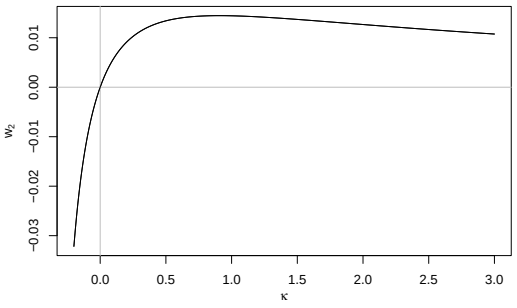
permissible value of $\kappa = -0.4$, w_2 is close to -1.4.

Figure 5.1: Scaling factors w_1 and w_2 given by (5.190) as a function of κ. These values are calculated with $n = 25$, $m = 13$, and $\nu = 13 + 25$, the dimensions and sample sizes used in the application in Chapter 6.

(a) Relative contribution from $\mathbf{S}$: w_1



(b) Contribution from the third term: w_2 (truncated at $\kappa = -0.2$)



5.3.7.2 Interpreting the trace $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$

Depending on the value of κ chosen to reflect the user's prior beliefs about the marginal kurtosis of $\mathbf{X}$, w_2 may be very small, as Figure 5.1b shows. However, the third term in (5.189) may still be large, depending on the magnitude of $\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]$. This term reflects the dispersion of the observed residuals compared with the prior expected dispersion described by $\mathbf{V}_R$.[1]

Recalling that $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$, the trace can be expressed as

$$\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}_R^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}), \tag{5.191}$$

which is the squared Mahalanobis distance of the n observations from the sample mean with respect to $\mathbf{V}_R$, averaged over the $n-1$ available degrees of freedom. When $\mathbf{X}$ is normally distributed with covariance matrix $\mathbf{V}_R$, the Mahalanobis distance has a chi-squared distribution with m degrees of freedom (Hardin and Rocke, 2005) and expectation m: the expected value of

[1] In fact, as shown in Section 5.3.6.1, the trace arises from $\mathbf{V}_M(\mathbf{V}_M + \mathbf{V}_T)^{\dagger}\mathrm{vec}(\mathbf{S})$ in (5.173), where $\mathbf{V}_M$ is the prior variance of the population variance and $\mathbf{V}_T$ reflects sampling variation: it may therefore more accurately be interpreted as reflecting the magnitude of $\mathbf{S}$ with respect to the prior variance of $\mathcal{M}(\mathbf{V})$. However, since both $\mathbf{V}_M$ and $\mathbf{V}_T$ are ultimately specified in terms of $\mathbf{V}_R$, the direct interpretation given here is preferred.

$\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$, if the observations $\mathbf{x}$ are drawn from a normal distribution with mean $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{V}_R$, is therefore $nm/(n-1)$. When n is large, this expected value is approximately m, and the expectation of the third term $\left[m - \text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]\mathbf{V}_R$ is approximately zero: therefore, if the prior is consistent with the data the third term in (5.189) may be relatively unimportant.

The Mahalanobis distance is often used to identify outliers in a data set, and $\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$ can be interpreted similarly as a summary of the dispersion of the observed residuals $\mathbf{x}_i - \overline{\mathbf{x}}$ with respect to the prior expected covariance matrix $\mathbf{V}_R$. If the observations are very dispersed or contain a high proportion of outliers, the trace will be large; small values of the trace (less than the expected value, which is approximately m) indicate that the sample observations are generally less dispersed than they would be if they were drawn from a multivariate-normal distribution with covariance matrix $\mathbf{V}_R$. The trace can therefore be viewed as a proxy for the kurtosis in the observed sample, measured with respect to the prior expectation $\mathbf{V}_R$ of the population variance.

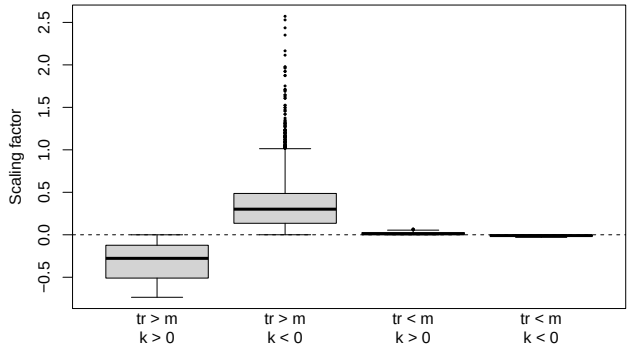
5.3.7.3 Effect of the third term in the adjusted expectation of $\mathcal{M}(\mathbf{V})$

The prior specification of κ has specific implications in the event of differences between the prior dispersion $\mathbf{V}_R$ and observed dispersion $\mathbf{S}$. Setting $\kappa > 0$ reflects a judgement that extreme values in the observations should be interpreted as outliers, rather than indicating a highly dispersed distribution. For $\kappa > 0$, w_2 is also positive, and typically very small, as shown in Figure 5.1b. When the observed trace is larger than m, the term in square braces will be negative, and may be quite large; the effect of the third term will therefore be to reduce the adjusted expectation of the variance from the weighted sum $w_1\mathbf{S} + (1-w_1)\mathbf{V}_R$ that would be obtained if $\kappa = 0$. Heuristically, the adjusted expectation of $\mathcal{M}(\mathbf{V})$ is reduced to reflect the judgement that extreme values in the observations are outliers, rather than evidence of high variability per se. The converse interpretation holds when $\kappa < 0$.

If the observed trace is smaller than m - indicating that $\mathbf{S}$ is smaller than

$\mathbf{V}_R$ - there is no evidence that there are any outliers in the observations. In this case, the contribution from the third term will typically be very small for both positive and negative κ. Figure 5.2 shows the scaling factors $w_2\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]$ obtained in the application presented in Section 6.2.4 in each of these scenarios.

Figure 5.2: Distribution of the scaling factor $w_2\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right]$ over all leadtimes for positive and negative κ, for the application considered in Chapter 6. These values are computed using $\kappa = \kappa_{lt}$ with nearest-κ replacement, as described in Section 6.2.4.



5.3.7.4 A constraint on the value of κ

The left-hand side of (5.189) is an expectation taken over a population of covariance matrices, and so must be positive semidefinite. Assuming that $\mathbf{V}_R$ is a valid covariance matrix, this is guaranteed to be the case when $\kappa = 0$, because the adjusted expectation is reduced to a nonnegative linear combination of $\mathbf{V}_R$ and $\mathbf{S}$. If $\kappa > 0$ and $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$, or if $\kappa < 0$ and $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) < m$, the adjusted expectation is no longer guaranteed to produce a valid covariance matrix. This occurs when the observed value of $\mathbf{S}$ disqualifies the joint prior choices for κ and $\mathbf{V}_R$, and may be regarded as a diagnostic warning of a conflict between the observed data and the prior specifications.

It is possible to identify a range of values of κ for which the right-hand side of (5.189) is guaranteed to be positive semidefinite, by finding those values of κ for which the term in braces $\{\}$ in that equation is nonnegative. This

involves solution of the quadratic inequality $a\tilde{\kappa}^2 + b\tilde{\kappa} + c \geq 0$, where

$$
\begin{aligned}
a &= (m+2)(\nu - m - 1), \\
b &= n(n-1)(\nu - m - 3)\left[m - \mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)\right] + n(\nu - m - 1)^2(m+2) + 2(\nu - 1)n\tilde{\nu}, \\
c &= 2n^2(\nu - m - 1)(\nu - 1)\tilde{\nu}.
\end{aligned}
$$

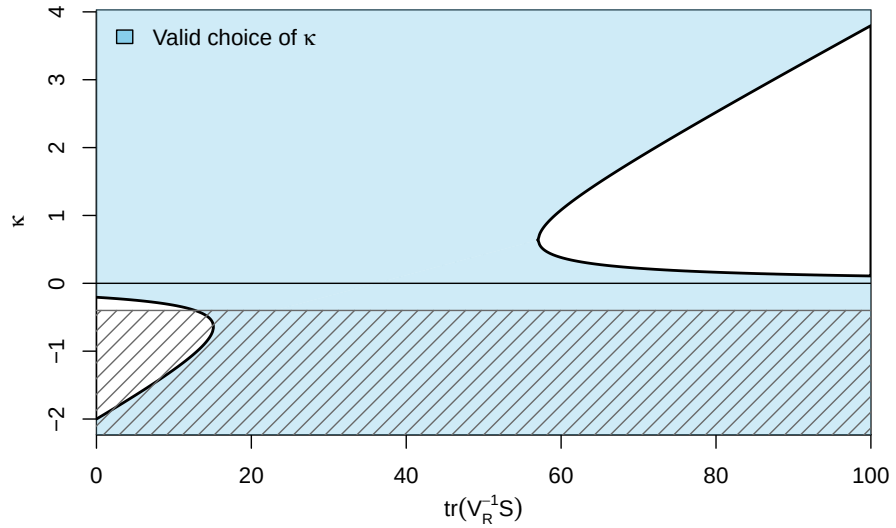
Since m is the dimension of $\mathbf{X}$, and $\nu - m$ is the notional sample size from which $\mathcal{M}(\mathbf{V})$ is considered to be estimated (see Section refsec:VM), both $\nu - m$ and m are positive; thus a is strictly positive, and the quadratic expression $a\tilde{\kappa}^2 + b\tilde{\kappa} + c$ has a unique minimum at $-b/2a$. The right-hand side of (5.189) is therefore guaranteed to be positive semidefinite for $\tilde{\kappa}$ in the set

$$
\left\{ \tilde{\kappa} \leq \frac{-b}{2a} - \sqrt{\frac{b^2 - 4ac}{4a^2}} \cup \tilde{\kappa} \geq \frac{-b}{2a} + \sqrt{\frac{b^2 - 4ac}{4a^2}} \right\}. \tag{5.192}
$$

The only coefficient actually affected by the data is b, through $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$. An example of valid values of κ for a range of values of this trace is shown in Figure 5.3. When $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \leq m$, only large negative values of κ are excluded; when $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$ is slightly larger than m, any value of κ is deemed compatible with the data. However, when $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$, implying that the observations are much more dispersed than was expected a priori, κ must be either close to zero or very large: recall from Table 5.1, for example, that $\kappa = 2$ suggests excess marginal kurtosis of 6, equivalent to that of a Student-t distribution with 5 degrees of freedom. The discontinuity in the range of valid values of κ may be interpreted as reflecting two competing explanations when the sample covariance matrix $\mathbf{S}$ is observed to be much larger than the prior expectation $\mathbf{V}_R$, leading to $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$. Setting κ close to zero implies a judgement that the observed sample will not contain many outliers: under this assumption, if $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$, the implication is that the sample is much more dispersed than was expected a priori. Choosing a relatively high value of κ implies the opposite judgement, that we would not be surprised to see a very high proportion of

outliers in the observed sample, and consequently that we would not be very surprised to observe $\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$.

Figure 5.3: Regions of valid κ values given by (5.192) as a function of $\text{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right)$ when $n = 25$, $m = 13$, and $\nu = 13 + 25$, as in the application in Chapter 6. The blue shaded area is the region for which (5.189) is guaranteed to produce a valid covariance matrix, while the hatched area is the region $\kappa < -0.4$ excluded by the minimum possible κ for a unimodal symmetric distribution, as described in Section 5.3.7.1.



Because the invalid range depends on a scalar summary statistic - the trace - of a matrix product, the bounds on 'permissible' values of κ are not exact limits in the sense that, depending on the particular instances of $\mathbf{V}_R$ and $\mathbf{S}$, some 'forbidden' values of κ close to the threshold may still result in a valid covariance matrix. However, selecting values of κ based on (5.192) will guarantee that the right-hand side of (5.189) is a nonnegative linear combination of $\mathbf{V}_R$ and $\mathbf{S}$, and so will produce a valid covariance matrix. In this sense, the limits in (5.192) define a sufficient condition for the right-hand side of (5.189) to be a valid covariance matrix; further work is required to identify a range of $\tilde{\kappa}$ that defines a necessary condition.

Although the prior covariance $\mathbf{V}_R$ and marginal excess kurtosis parameter κ are set before any data is observed, this provides a useful check that the specified $\mathbf{V}_R$ and κ are compatible with the sample actually obtained. Where the resulting adjusted matrix is found not to be a valid covariance matrix, the

bounds given by (5.192) may also be used to suggest alternative values for κ that are guaranteed to produce a valid covariance matrix. In particular, choosing the closest valid value of κ to the original proposal provides a guaranteed positive definite matrix that is in some sense close to the invalid adjusted expectation. Practical issues around implementing this substitution are considered in Section 6.2.4.

5.3.8 Summary and discussion

This section extends the second-order exchangeability representation used in Bayes linear adjustment – reviewed in Section 5.2 – to accommodate multivariate adjustment of covariance matrices in a principled but relatively straightforward framework. While the derivations presented require extensive matrix manipulation, it is not necessary for a user to understand these operations before applying the method: they need only specify the matrix $\mathbf{V}_R$ of prior variances and covariances, the notional sample size $\nu - m$ reflecting how confident they are in their estimate of $\mathbf{V}_R$, and the marginal kurtosis parameter κ, indicating whether the data are believed to have a greater or lesser propensity to outliers than a normal distribution. The parameters ν and κ can be obtained directly from the parameters used in the scalar case, with $c = 2/(\nu - m - 3)$ and $\mathbb{V}\left[Z_i^2\right] = 3\kappa + 2$.

Goldstein and Wooff (2007, §8.13) remark that specifying a full set of variances and covariances between all elements of $\mathbf{S}$ and all elements of $\mathcal{M}(\mathbf{V})$ requires a more detailed level of specification than may be reasonable, a concern echoed by Wilkinson and Goldstein (1995), who note that the effort required for both specification and computation may increase significantly for large covariance matrices if an enriched projection space is to be used, as described in Section 5.2.6. The heuristic solution proposed by Goldstein and Wooff (2007) is either to weight the prior and observed covariance matrices by their relative sample sizes and average them, or to perform independent adjustments of the scalar quantities, and to combine these marginal adjusted expectations and variances with a correlation matrix obtained by averaging a prior correlation

matrix and the sample correlation matrix, with the resulting matrix referred to as the semi-adjusted residual variance matrix. However, having specified all of the quantities required to obtain this semi-adjusted residual variance matrix, it is no more difficult to instead combine the prior marginal expectations and variances with the prior correlation matrix to give $\mathbf{V}_R$, and proceed with the multivariate adjustment as presented above; all of the required covariances are then fully specified in terms of this prior variance matrix and the parameters κ and ν.

Furthermore, considering the full matrix of variances and covariances between all variables forces the user to consider an appropriate value for each element: simply carrying out univariate variance adjustments implies a judgement that the variances are independent, which is not generally likely to be the case - particularly when the adjustment is of highly dependent variables such as spatial data. It is still perfectly possible to encode an assumption of independence between some or indeed all variables, by setting the relevant covariances to zero - but by doing so in the multivariate framework, the judgement is considered and made explicit, rather than assuming independence by default.

5.4 Bayes linear adjustment as an approximation to Bayesian inference

The original motivation for using a Bayes linear adjustment was to develop a framework to combine information from two sources in an approximation to Bayesian inference, while retaining a tractable form that is easily assimilated into the Bayesian postprocessing framework of Chapter 2. This approximation can only be checked directly when a closed-form solution to the probabilistic Bayesian analysis is available; this is most easily achieved in the conjugate setting, when the quantity of interest is assumed to be Gaussian. It will be shown that in this case, Bayes linear adjustment produces asymptotic approximations to the posterior expectation and variance of the population

mean that would be obtained using a natural conjugate joint prior.

5.4.1 Conjugate Bayesian inference for the parameters of a multivariate normal distribution

When the data $\mathbf{x}$ are assumed to have a multivariate normal distribution, the natural conjugate form for simultaneously expressing prior beliefs about the distributions of the population mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ is a joint prior distribution with the same normal-inverse-Wishart form as the joint likelihood (O'Hagan and Forster, 2004). Under this specification, $\boldsymbol{\mu}$ is assumed to be normally distributed conditional on $\boldsymbol{\Sigma}$, with prior expectation and variance

$$\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{\Sigma}] = \boldsymbol{\mu}_0, \qquad \mathbb{V}[\boldsymbol{\mu}|\boldsymbol{\Sigma}] = \frac{1}{\beta}\boldsymbol{\Sigma}, \tag{5.193}$$

say, where β is a scaling parameter defining the relationship between the variance of $\boldsymbol{\mu}$ and the population variance $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is assumed to have an inverse-Wishart distribution, with prior expectation and variance (Mardia et al., 1979)

$$\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\mathbf{SS}_0}{\nu_0 - m - 1} \tag{5.194}$$

and

$$\mathbb{V}\Big[\mathrm{vec}\Big(\boldsymbol{\Sigma}\Big)\Big] = \frac{2\mathrm{vec}\Big(\mathbf{SS}_0\Big)\,\mathrm{vec}\Big(\mathbf{SS}_0\Big)' + 2(\nu_0 - m - 1)\mathbf{N}_m\,(\mathbf{SS}_0 \otimes \mathbf{SS}_0)}{(\nu_0 - m)(\nu_0 - m - 1)^2(\nu_0 - m - 3)}, \tag{5.195}$$

where $\mathbf{SS}_0$ is a scale matrix and ν_0 denotes the prior degrees of freedom.

The joint posterior density of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, conditional on the observed data $\mathbf{x}$, is also a normal-inverse-Wishart density. The posterior distribution of $\boldsymbol{\Sigma}$ is inverse-Wishart (O'Hagan and Forster, 2004), with posterior expectation

$$\mathbb{E}[\boldsymbol{\Sigma}|\mathbf{x}] = \frac{\mathbf{SS}_0 + \mathbf{SS}_x + \frac{\beta n}{\beta + n}\,(\boldsymbol{\mu}_0 - \overline{\mathbf{x}})\,(\boldsymbol{\mu}_0 - \overline{\mathbf{x}})'}{\nu_0 + n - m - 1} = \frac{\mathbf{SS}_n}{\nu_n - m - 1}, \tag{5.196}$$

say, where n is the size of the observed sample, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the observed

sample mean, and $\mathbf{SS}_x = \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ denotes the observed sum of squares; $\mathbf{SS}_n$ denotes the posterior sum of squares, and ν_n the posterior degrees of freedom. The posterior variance is then

$$\mathbb{V}\Big[\mathrm{vec}\Big(\boldsymbol{\Sigma}\Big)\,|\mathbf{x}\Big] = \frac{2\mathrm{vec}\Big(\mathbf{SS}_n\Big)\,\mathrm{vec}\Big(\mathbf{SS}_n\Big)' + 2(\nu_n - m - 1)\mathbf{N}_m\,(\mathbf{SS}_n \otimes \mathbf{SS}_n)}{(\nu_n - m)(\nu_n - m - 1)^2(\nu_n - m - 3)}. \tag{5.197}$$

The marginal posterior distribution of $\boldsymbol{\mu}$, conditional on the observed data, is obtained by integrating $\boldsymbol{\Sigma}$ out of the joint posterior distribution. The resulting distribution is a multivariate Student-t distribution with ν_n degrees of freedom, with expectation and variance

$$\mathbb{E}[\boldsymbol{\mu}|\mathbf{X}] = \frac{\beta\boldsymbol{\mu}_0 + n\overline{\mathbf{x}}}{\beta + n} = \boldsymbol{\mu}_n, \tag{5.198}$$

and

$$\mathbb{V}[\boldsymbol{\mu}|\mathbf{X}] = \frac{\mathbf{SS}_n}{(\beta + n)\,(\nu_0 + n - m - 1)} = \frac{\mathbf{SS}_n}{(\kappa_n)\,(\nu_n - m - 1)}. \tag{5.199}$$

5.4.2 Bayes linear adjustment as an approximation to inference using the natural conjugate prior

The multivariate Bayes linear adjustment described in Section 5.3 is now used to update the prior specifications used in Section 5.4.1 with the observed sample mean and covariance matrix, and the resulting expectations and variances compared to those obtained using a fully probabilistic prior specification.

The population variance $\mathcal{M}(\mathbf{V})$ that appears in the Bayes linear exchangeability representation is directly analogous to the quantity $\boldsymbol{\Sigma}$ in the Bayesian inference carried out above, with prior expectation $\mathbf{V}_R = \mathbb{E}[\boldsymbol{\Sigma}]$ and prior variance $\mathbf{V}_M = \mathbb{V}[\boldsymbol{\Sigma}]$ as set out in (5.194) and (5.195). Because the data are assumed to be Gaussian, the kurtosis parameter κ is 0, so that the prior beliefs are updated using equations (5.179) and (5.188) respectively. First using (5.179),

the adjusted expectation of the population variance is

$$\begin{aligned}\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})] &= \frac{n-1}{\nu_0-m-1+n-1}\mathbf{S}+\frac{\nu_0-m-1}{\nu_0-m-1+n-1}\mathbf{V}_R \\ &= \frac{(\mathbf{SS}_x+\mathbf{SS}_0)}{\nu_0+n-1-m-1} \\ &= \frac{\mathbf{SS}_n^*}{\nu_n^*-m-1}, \end{aligned} \tag{5.200}$$

say, where $\mathbf{S}$ is the observed sample covariance matrix. This has the same form as the expectation $\mathbb{E}[\boldsymbol{\Sigma}|\mathbf{X}]$ obtained in (5.196), but has one less degree of freedom in ν_n^*, and lacks the term in the numerator of (5.196) reflecting the discrepancy between the prior expectation and observed value of the mean. However, as n increases, $\mathbf{SS}_n$ and $\mathbf{SS}_n^*$ will both be dominated by the sample sum of squares $\mathbf{SS}_x$, and so the Bayes linear adjusted expectation of the population variance will asymptotically approach that obtained from Bayesian inference with a natural conjugate prior.

However, this asymptotic equivalence no longer holds when considering the Bayes linear adjusted variance of the population variance. From (5.188), when $\kappa = 0$,

$$\mathbb{V}_{\mathbf{S}}\Big[\mathrm{vec}\Big(\mathcal{M}(\mathbf{V})\Big)\Big] = \frac{\nu_0-m-1}{\nu_0-m-1+n-1}\mathbf{V}_M. \tag{5.201}$$

The adjusted variance of $\mathcal{M}(\mathbf{V})$ is simply a scaled version of the prior variance, with no contribution from the data, and so may be quite different to the Bayesian posterior variance of $\boldsymbol{\Sigma}$ given in (5.197).

The Bayes linear population mean $\mathcal{M}(\mathbf{X})$ is directly equivalent to $\boldsymbol{\mu}$ in the Bayesian inference above. Under the joint conjugate prior specification, the prior distribution of $\boldsymbol{\mu}$ was specified conditional on the population variance $\boldsymbol{\Sigma}$; an analogous Bayes linear prior belief specification gives the prior expectation and variance of $\mathcal{M}(\mathbf{X})$ in terms of the updated beliefs about $\mathcal{M}(\mathbf{V})$ captured

in (5.200), with

$$\mathbb{E}[\mathcal{M}(\mathbf{X})] = \boldsymbol{\mu}_0 \tag{5.202}$$

and

$$\mathbb{V}[\mathcal{M}(\mathbf{X})] = \frac{1}{\beta}\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]\,. \tag{5.203}$$

Then from (5.103),

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathcal{M}(\mathbf{X})] &= \mathbb{E}[\mathcal{M}(\mathbf{X})] + \mathbb{V}[\mathcal{M}(\mathbf{X})]\left(\mathbb{V}[\mathcal{M}(\mathbf{X})] - \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{n}\right)^{-1}(\overline{\mathbf{x}} - \mathbb{E}[\mathcal{M}(\mathbf{X})]) \\
&= \boldsymbol{\mu}_0 + \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{\beta}\left(\left[\frac{1}{\beta} + \frac{1}{n}\right]\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]\right)^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0) \\
&= \boldsymbol{\mu}_0 + \frac{n\,(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)}{\beta + n} \\
&= \boldsymbol{\mu}_n.
\end{aligned} \tag{5.204}$$

From (5.104), the adjusted variance of $\mathcal{M}(\mathbf{X})$ is

$$\begin{aligned}
\mathbb{V}_{\mathbf{X}}[\mathcal{M}(\mathbf{X})] &= \mathbb{V}[\mathcal{M}(\mathbf{X})] - \mathbb{V}[\mathcal{M}(\mathbf{X})]\left(\mathbb{V}[\mathcal{M}(\mathbf{X})] + \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{n}\right)^{-1}\mathbb{V}[\mathcal{M}(\mathbf{X})] \\
&= \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{\beta} - \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{\beta}\left(\left[\frac{1}{\beta} + \frac{1}{n}\right]\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]\right)^{-1}\frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{\beta} \\
&= \frac{\mathbb{E}_{\mathbf{S}}[\mathcal{M}(\mathbf{V})]}{\beta + n} \\
&= \frac{1}{\beta + n}\frac{\mathbf{SS}^*_n}{\nu^*_n - m - 1}.
\end{aligned} \tag{5.205}$$

Thus the Bayes linear adjusted expectation of the population mean is identical to that obtained by Bayesian inference, and the Bayes linear adjusted variance asymptotically approaches the posterior variance $\mathbb{V}[\boldsymbol{\mu}|\mathbf{x}]$, as discussed in Section 5.4.2.

5.5 Summary of chapter

This chapter presents one of the major contributions of the thesis, a multivariate extension of the variance adjustment for second-order exchangeable quantities laid out in Goldstein and Wooff (2007). This extension was motivated by the need to find a more flexible alternative to conjugate inference that would accommodate a judgement that the quantity of interest may not be normally distributed, without the use of computationally costly simulation to find the full posterior distribution.

The chapter begins by reviewing the Bayes linear statistical methodology presented by Goldstein and Wooff (2007). In Section 5.2.5.2.1 an adjustment is made to the parametrisation of higher-order quantities through the kurtosis parameter $\mathbb{V}\left[Z_i^2\right]$ and the notional sample size m, which have a fixed relationship in the original framework: by allowing these two quantities to act independently of one another, greater flexibility and interpretability of the adjustment are obtained.

Section 5.3 derives the multivariate extension to the Bayes linear adjustment. Although Goldstein and Wooff (2007) argue that specifying all of the variances and covariances required to carry out a full multivariate adjustment may be an unreasonably complex task – particularly when working with the geometric representation described in Section 5.2.6 (Wilkinson, 1997) – the multivariate adjustment presented here requires specification only of the prior covariance matrix $\mathbf{V}_R$ and the scalar parameters κ and ν, which can be obtained from the parameters used in the scalar adjustment. This approach is, arguably, simpler than the copula-based multivariate adjustment suggested as an alternative by Goldstein and Wooff (2007), particularly where $\mathbf{V}_R$ can be obtained empirically, because the dependence structure does not need to be estimated separately.

Where the assumption of common marginal kurtosis across all variables is reasonable, the multivariate adjustment is very efficient, reducing to a scalar-weighted sum of the prior and observed covariance matrices. The motivating

application for this work required Bayes linear adjustment of the covariance matrix of a single variable at multiple locations; as will be seen in Chapter 6, the assumption of common (or, at least, similar) kurtosis across all regions is indeed a reasonable one, and the question of how to handle different marginal kurtoses is not addressed here. It is possible that the framework could be extended to use a kurtosis matrix, $\mathbf{K}$ say, rather than the scalar parameter; investigation of this possibility is left as further work. Instead, if common kurtosis can only be assumed to apply for a subset of the variables, multivariate Bayes linear adjustment should be carried out for each subset that can be assumed to be approximately homogeneous, and may be used to construct the 'semi-adjusted residual variance matrix' proposed by Goldstein and Wooff (2007, §8.13) to facilitate adjustment of covariance matrices, or to determine appropriate submatrices with which to enhance the projection space (Wilkinson and Goldstein, 1995; Williamson et al., 2012).

In Section 5.4 it was shown that, with an appropriate choice of parametrisation, Bayes linear adjustment can produce a second-order approximation to the posterior distribution of $\boldsymbol{\mu}$, and a first-order approximation to the posterior distribution of $\boldsymbol{\Sigma}$, obtained using fully probabilistic inference with the natural conjugate normal-inverse-Wishart prior.

More practical questions concerning the implementation of the multivariate Bayes linear adjustment are addressed in the next chapter, where the method is used to obtain an updated estimate of the forecast discrepancy used in the postprocessing of temperature forecasts in Chapter 4.

Chapter 6

Bayes linear adjustment of UK temperature forecasts

In Chapter 4, the forecasts of surface temperature described in Section 2.1.2 were postprocessed using simple empirical estimates of the expectation and covariance matrix of the forecast discrepancy. This chapter evaluates the benefits gained by instead using Bayes linear adjustment to determine the discrepancy correction required to postprocess the forecasts. Prior beliefs about $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ are updated with later information using the multivariate second-order Bayes linear adjustment described in Chapter 5, and these updated beliefs are used to postprocess weather forecasts of surface temperatures.

In the postprocessing framework described in Chapter 2, all available information about the forecast discrepancy $\boldsymbol{\Delta}$ is assumed to be captured by the mean discrepancy vector $\boldsymbol{\eta}$ and discrepancy covariance matrix $\boldsymbol{\Lambda}$; these quantities are equivalent to the population mean $\mathcal{M}(\mathbf{X})$ and population variance $\mathcal{M}(\mathbf{V})$ introduced in the Bayes linear framework in Sections 5.2.1.1 and 5.3. For clarity, this chapter will use $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ to refer to the population mean and covariance of the forecast discrepancy $\boldsymbol{\Delta}$. Observed forecast discrepancies are denoted by $\boldsymbol{\delta}$ rather than $\mathbf{X}$, with $\bar{\boldsymbol{\delta}}$ and $\boldsymbol{S}_{\boldsymbol{\delta}}$ denoting the sample mean and covariance, respectively, of the forecast errors in a training dataset; these terms are summarised for easy reference in Tables 6.1 and 6.2.

Section 6.1 describes the general form of the Bayes linear update used to

Table 6.1: Table of notation used for general Bayes linear adjustment in Chapter 5 with corresponding terms used for adjustment of forecast discrepancy in this chapter.

	General adjustment	**Discrepancy adjustment**
Forecast consensus discrepancy	$\mathbf{X}$	$\boldsymbol{\Delta}$
Observed forecast errors	$\mathbf{x}$	$\boldsymbol{\delta}$
Expected value of consensus discrepancy	$\mathcal{M}(\mathbf{X})$	$\mathbb{E}[\boldsymbol{\Delta}] = \boldsymbol{\eta}$
Covariance of consensus discrepancy	$\mathcal{M}(\mathbf{V})$	$\mathbb{V}[\boldsymbol{\Delta}] = \boldsymbol{\Lambda}$
Kurtosis parameter	$\kappa = \mathbb{K}\text{ur}[\mathbf{X}]/3$	$\kappa = \mathbb{K}\text{ur}[\boldsymbol{\Delta}]/3$

Table 6.2: Prior and adjusted expectations and covariance matrices of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$, and scalar quantities used in the adjustment.

	Expectation $\boldsymbol{\eta}$	**Covariance $\boldsymbol{\Lambda}$**
Prior expectation	$\mathbb{E}[\boldsymbol{\eta}]$	$\mathbb{E}[\boldsymbol{\Lambda}] = \mathbf{V}_R$
Prior covariance	$\mathbb{V}[\boldsymbol{\eta}]$	$\mathbb{V}[\boldsymbol{\Lambda}] = \mathbf{V}_M$
Sample estimator	$\bar{\boldsymbol{\delta}}$	$\boldsymbol{S}_{\boldsymbol{\delta}}$
Adjusted expectation	$\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$	$\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$
Adjusted covariance	$\mathbb{V}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$	$\mathbb{V}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$
Dimension	m	m
Confidence parameter	ν	ν
Size of observed sample	n	n
Size of sample used to set prior $\mathbb{E}[\boldsymbol{\eta}]$	z	z

obtain adjusted expectations and covariances of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ throughout this chapter. In Section 6.2, Bayes linear adjusted expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ are plugged directly into the Bayesian postprocessing model described in Section 2.2, and the skill of the resulting forecasts is evaluated against forecasts postprocessed using only the 'prior' or 'observed' estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$. Each of the parameters in the Bayes linear adjustment is varied in turn in order to investigate the sensitivity of forecast skill to various choices of prior specification. In Section 6.3, uncertainty about $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ is incorporated into the postprocessed forecasts through simulation, in order to evaluate the effect on forecast skill.

6.1 Bayes linear adjustment of the forecast discrepancy $\boldsymbol{\Delta}$

Throughout this chapter, Bayes linear adjusted expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ are used in place of direct empirical estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ in the Bayesian postprocessing framework described in Section 2.2. Sections 6.2.1 – 6.2.4 explore the effect

of varying the user-specified parameters, while Section 6.3 includes additional parameter uncertainty: all of these experiments use the same approach to specifying the prior expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$, and to estimating the sample means $\bar{\boldsymbol{\delta}}$ and sample covariances $\boldsymbol{S_\delta}$ used in the adjustment, as outlined below.

As in classical Bayesian statistics, the priors should be specified before observing any data – that is, before the new forecasts that are to be postprocessed have been issued. The priors should correspond as closely as possible to the distribution of discrepancies that might be expected at the time the forecast are issued; one way to achieve this is to use a moving window (MW) training set to estimate the prior expectation and variance of the forecast errors, as described in Section 2.3.2, since these data are available before the new forecasts have been generated. Throughout this chapter, the prior expectations $\mathbb{E}[\boldsymbol{\eta}]$ and $\mathbb{E}[\boldsymbol{\Lambda}]$ for each forecast instance are the sample mean and covariance of the forecast errors in 25 MW training cases.

After the forecasts are issued, they can be used to identify similar forecasts from a historical archive, as described in Section 2.3.3.1; a second training set drawn from the closest of these analogues to the forecast instance of interest will be treated as the observed sample. In this way, information about persistent errors from recent forecasts is combined with information about flow-dependent errors from the analogues. The observed sample mean $\bar{\boldsymbol{\delta}}$ and sample covariance $\boldsymbol{S_\delta}$ used to adjust the prior beliefs for each forecast instance are estimated from 25 analogues, obtained using the method described in Section 2.3.3.1[1]: similar results were obtained when using analogues selected on the basis of the prevailing weather regime, as discussed in Section 2.3.3.2.

The prior quantities derived from the MW training cases are updated by the sample mean and covariance matrix of the AN training cases to obtain the necessary Bayes linear adjusted quantities, as described in Section 5.3. From

[1]The selection of analogues in this chapter differs from the implementation in Chapter 4, in that analogues here are selected from all years excluding the year in which the forecast was issued: this is to avoid the possibility of any training cases appearing in both the MW and AN sets for any given forecast instance, and so being used to determine both prior and observed expectations and variances.

(5.177) and (5.178), and using the notation of this chapter as defined in Table 6.2, the adjusted expectation of $\boldsymbol{\Lambda}$ given $\boldsymbol{S_\delta}$ is

$$\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = w_1 \boldsymbol{S_\delta} + \left\{(1-w_1) + w_2\left[m - \operatorname{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)\right]\right\}\mathbb{E}[\boldsymbol{\Lambda}] \tag{6.1}$$

where

$$w_1 = \frac{n(n-1)}{\tilde{\kappa}+n\tilde{\nu}} \quad \text{and} \quad w_2 = \frac{n(n-1)(\nu-m-3)\tilde{\kappa}}{[\tilde{\kappa}+n\tilde{\nu}]\,[(m+2)(\nu-m-1)\tilde{\kappa}+2(\nu-1)n\tilde{\nu}]}, \tag{6.2}$$

with $\tilde{\kappa} = \kappa(n-1)(\nu-m)$ and $\tilde{\nu} = \nu-m-1+n-1$. As noted in Section 5.3.6.1, when $\kappa = 0$, this can be simplified to

$$\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = \frac{n-1}{\tilde{\nu}}\boldsymbol{S_\delta} + \frac{\nu-m-1}{\tilde{\nu}}\mathbb{E}[\boldsymbol{\Lambda}]. \tag{6.3}$$

Similarly, from (5.103) and (5.104),

$$\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}] = \mathbb{E}[\boldsymbol{\eta}] + \mathbb{V}[\boldsymbol{\eta}]\left(\mathbb{V}[\boldsymbol{\eta}] + \frac{1}{n}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\right)^{-1}\left(\bar{\boldsymbol{\delta}} - \mathbb{E}[\boldsymbol{\eta}]\right) \tag{6.4}$$

and

$$\mathbb{V}_{\boldsymbol{\delta}}[\boldsymbol{\eta}] = \left(\mathbb{V}[\boldsymbol{\Lambda}]^{-1} + n\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]^{-1}\right)^{-1}. \tag{6.5}$$

The effect of varying the prior specifications of $\mathbb{V}[\boldsymbol{\eta}]$, ν and κ is considered in the next section.

6.2 Postprocessing with 'plug-in' Bayes linear adjusted estimate of $\boldsymbol{\Delta}$

Recall from (2.13) in Section 2.2 that the posterior distribution of the weather quantity $\mathbf{Y}_0$ is

$$\mathbf{Y}_0 | \{\mathbf{Y}_{ij}\}, \boldsymbol{\Delta} \sim MVN(\boldsymbol{\tau}, \mathbf{S}) \tag{6.6}$$

where from (2.18) and (2.19),

$$\mathbf{S}^{-1} = \boldsymbol{\Gamma}^{-1} + (\boldsymbol{\Sigma}_{\boldsymbol{D}} + \boldsymbol{\Lambda})^{-1}, \tag{6.7}$$

$$\boldsymbol{\tau} = \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} + (\boldsymbol{\Sigma}_{\boldsymbol{D}} + \boldsymbol{\Lambda})^{-1}\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\eta}\right)\right], \tag{6.8}$$

where $\hat{\boldsymbol{\xi}}$ is the sample ensemble forecast consensus, with associated uncertainty $\boldsymbol{\Sigma}_{\boldsymbol{D}}$, and $\boldsymbol{\alpha}$ is the prior expectation of $\mathbf{Y}_0$, with associated covariance matrix $\boldsymbol{\Gamma}$. This distribution is conditional on the observed ensemble members $\{\mathbf{Y}_{ij}\}$ via the multi-model ensemble consensus $\boldsymbol{\xi}$ and associated covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{D}}$, as described in Section 2.2.1; and the expectation $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Lambda}$ of the forecast discrepancy $\boldsymbol{\Delta}$ are treated as known quantities.

In this section, $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ will be replaced in (6.7) and (6.8) with their Bayes linear adjusted expectations $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ and $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ to obtain the conditional posterior distribution of $\mathbf{Y}_0$, parametrised by

$$\mathbf{S}^{-1} = \boldsymbol{\Gamma}^{-1} + \left(\boldsymbol{\Sigma}_{\boldsymbol{D}} + \mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\right)^{-1}, \tag{6.9}$$

$$\boldsymbol{\tau} = \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} + \left(\boldsymbol{\Sigma}_{\boldsymbol{D}} + \mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\right)^{-1}\left(\hat{\boldsymbol{\xi}} - \mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]\right)\right]. \tag{6.10}$$

This approach essentially treats $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ and $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ as if there is no uncertainty about their values, and so allows a direct comparison with the postprocessed forecasts obtained in Chapter 4 using plug-in estimates. The effect of incorporating additional sources of uncertainty in the posterior distribution is investigated in Section 6.3.

In all instances, the posterior forecast is obtained using a noninformative prior with precision $\boldsymbol{\Gamma}^{-1} = \mathbf{0}$ for the temperature $\mathbf{Y}_0$, as suggested in Section 4.1. The forecast skill of the postprocessed forecasts is compared using the metrics for density forecasts described in Chapter 3; forecasts are judged to have greater skill if they are more accurate and, subject to good probabilistic calibration, sharper.

Throughout this chapter, the skill of forecasts postprocessed using Bayes linear adjusted expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ in (6.9) and (6.10) is compared to that

of the MW-postprocessed and AN-postprocessed forecasts already discussed in detail in Section 4.3. In the notation of this chapter, the MW-postprocessed forecasts use $\mathbb{E}[\boldsymbol{\eta}]$ and $\mathbb{E}[\boldsymbol{\Lambda}]$ in place of $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ and $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ in (6.9) and (6.10), while the AN-postprocessed forecasts use $\bar{\boldsymbol{\delta}}$ and $\boldsymbol{S_\delta}$. This enables a direct comparison of the skill of forecasts corrected using the two sources of information with the skill of forecasts using only one source.

6.2.1 Bayes linear adjustment when $\kappa = 0$

In Section 5.4 it was shown that Bayes linear adjustment can be used to approximate the joint posterior distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ that would be obtained using a fully probabilistic Bayesian inference with a normal-inverse Wishart prior, by setting $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$, where z denotes the notional sample size used to determine the prior expectation of $\boldsymbol{\eta}$, with $\kappa = 0$ and $\nu = z + m$. This specification provides a useful baseline against which to measure the effect of alternative specifications in later sections.

Recall from Section 5.3.5.2 that ν controls the relationship between $\mathbf{V}_R = \mathbb{E}[\boldsymbol{\Lambda}]$ and $\mathbf{V}_M = \mathbb{V}[\boldsymbol{\Lambda}]$, and so reflects the user's degree of confidence in their prior beliefs about $\boldsymbol{\eta}$: where, as here, $\mathbb{E}[\boldsymbol{\Lambda}]$ is estimated empirically from a set of z samples, the natural choice is to set $\nu = z + m$, so that the notional sample size reflects the actual sample size used. Such a judgement is appropriate if both the MW and AN training sets are believed to be sampling independently from the same population of forecast errors as the current instance; the effect of reducing z to indicate reduced confidence in the prior will be considered in Section 6.2.2.

As well as the the MW- and AN-postprocessed forecasts, the skill of forecasts postprocessed using the Bayes linear adjusted expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ is compared to that of forecasts postprocessed using the sample mean and covariance of the larger sample obtained by simply 'pooling' the MW and AN training cases into a single set, in order to understand the benefits of using Bayes linear adjustment to estimate the discrepancy, rather than this simpler approach.

6.2.1.1 Bayes linear adjusted estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$

Under the specification used here, Bayes linear adjustment produces the same estimate of $\boldsymbol{\eta}$ as would be obtained by simply taking the sample mean of the $z = 25$ MW forecast errors and the $n = 25$ AN forecast errors: with $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$, the adjusted expectation of the population mean discrepancy $\boldsymbol{\eta}$ given by (6.4) is

$$\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}] = \frac{z}{z+n}\mathbb{E}[\boldsymbol{\eta}] + \frac{n}{z+n}\bar{\boldsymbol{\delta}} \tag{6.11}$$

and where, as here, $z = n$, this is the unweighted average of the MW and AN-adjusted forecast errors. When $\kappa = 0$ and with $\nu = z + m$, the Bayes linear adjusted expectation of $\boldsymbol{\Lambda}$ given by (6.3) is

$$\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = \frac{1}{z+n-2}\left\{(n-1)\boldsymbol{S_\delta} + (z-1)\mathbb{E}[\boldsymbol{\Lambda}]\right\}, \tag{6.12}$$

which is the standard estimate of the common variance of two samples.

It is worth highlighting the difference between postprocessing using the Bayes linear expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$, and using empirical estimates from the pooled MW and AN training cases, which may seem like an appealing alternative. In particular, the variance $\hat{\boldsymbol{\Lambda}}_{pooled}$ of the pooled training sets is divided by one less degree of freedom, so if the prior $\mathbb{E}[\boldsymbol{\eta}]$ is sufficiently close to $\bar{\boldsymbol{\delta}}$, the estimate of $\boldsymbol{\Lambda}$ obtained from the pooled training sets will be slightly sharper than that obtained by the baseline Bayes linear adjustment, although the improvement is bounded at $\frac{z+n-1}{z+n-2}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$. However, a large discrepancy between $\mathbb{E}[\boldsymbol{\eta}]$ and $\bar{\boldsymbol{\delta}}$ will increase $\hat{\boldsymbol{\Lambda}}_{pooled}$, in which case the pooled variance will typically be somewhat larger than $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$.

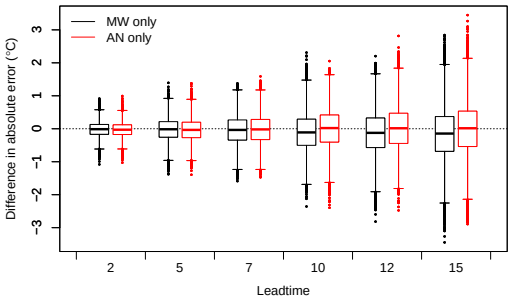
6.2.1.2 Forecast accuracy and sharpness

Because the pooled and Bayes linear adjusted estimates of $\boldsymbol{\eta}$ are the same under this specification, postprocessing with either results in identical deterministic

(mean) forecasts. Figure 6.1 shows the distribution of the differences in absolute error between the forecasts postprocessed using the Bayes linear adjusted and MW/AN-only estimates, for all forecasts in selected regions of the study. A similar pattern appears in all regions, with the majority of forecast errors differing by less than 1°C even at the longest leadtimes. The distribution of differences is almost symmetric about zero; this occurs because the MW-adjusted forecasts are more accurate (having smaller absolute errors) than the AN-adjusted forecast in about 50% of all forecast instances across all leadtimes and all regions. Averaging the two sources of information therefore results in roughly equal numbers of forecast instances with better and worse accuracy, and an overall change in MAE close to zero. Although the improvement is small, the Bayes linear adjusted forecasts achieve the lowest MAE at all leadtimes, with MAE around 0.1-0.2°C lower than the MW-adjusted forecasts at longer leadtimes (Figure 6.1b).

Figure 6.1: Distributions (over 630 forecast instances) at selected leadtimes of differences in absolute marginal errors in selected regions when using Bayes linear adjusted expectation of $\boldsymbol{\eta}$ to postprocess forecasts in place of the MW or AN estimates; and overall mean absolute marginal error for each method. Negative differences indicate that the Bayes linear adjusted forecasts were more accurate, having smaller absolute errors.

(a) Distribution of differences in absolute error



(b) MAE for each method

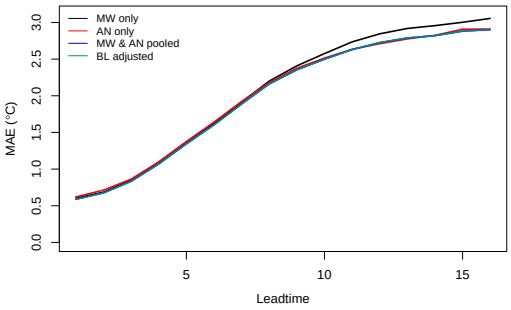
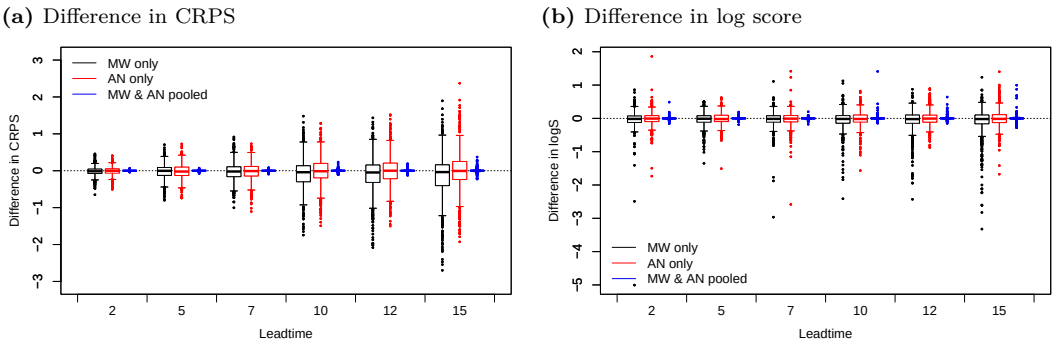


Figure 6.2 shows the difference in log scores and CRPS for each estimation method compared to the Bayes linear adjusted forecasts. The average log scores are almost the same for all methods, although relatively large changes in skill are seen for some individual forecasts. The distribution of differences between the AN-adjusted and Bayes linear adjusted forecasts is generally symmetric about

zero, while the MW-adjusted forecasts are more likely to have a higher log score, and the pooled-adjustment forecasts often have a slightly lower logarithmic score than the Bayes linear adjusted forecasts, particularly at longer leadtimes.

Figure 6.2: Distributions (over 630 forecast instances for each location) at selected leadtimes of differences in CRPS and logS when using Bayes linear adjusted $\boldsymbol{\eta}$ to postprocess forecasts in place of the MW, AN or pooled estimate. Negative changes indicate that the Bayes linear adjusted forecasts were more skilful, having lower CRPS/logS.

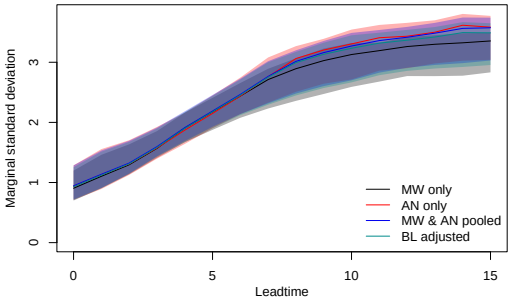


As noted in Section 4.3, the MW-adjusted forecasts are sharper than the AN-adjusted forecasts in roughly half of all instances at lead times of a week or less, rising to 60% of all instances at longer leadtimes. As a result, forecast sharpness is insensitive to the choice of $\boldsymbol{\Lambda}$ at shorter leadtimes, while at eight days or longer, the Bayes linear adjusted foreasts tend to be somewhat sharper than the AN-adjusted forecasts, but less sharp than the MW-adjusted forecasts (Figure 6.3a).

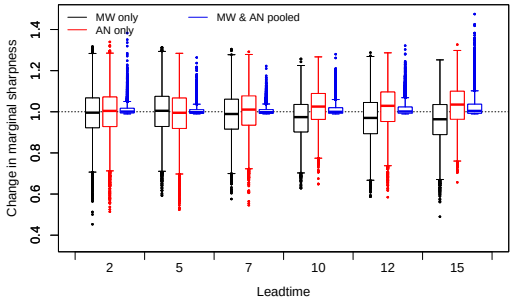
Pooling the two training sets to estimate the discrepancy produces a forecast of comparable sharpness to Bayes linear adjustment in roughly half of all of the forecast instances at all leadtimes and in all regions (Figure 6.3b). In the remaining instances, differences between the means of the AN and MW training sets mean that the Bayes linear adjusted expectations of $\boldsymbol{\Lambda}$ are typically somewhat sharper, as discussed in Section 6.2.1.1; as a result, the Bayes linear adjusted forecasts are slightly marginally sharper on average. Overall, as Figure 6.3a shows, forecast sharpness varies more between regions than between estimation methods, and the difference in average forecast sharpness is very small at all but the longest leadtimes.

Figure 6.3: Spread of average marginal sharpness in each region, where the lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values; and distribution of changes in marginal sharpness of postprocessed forecasts across all regions at selected leadtimes, expressed as the ratio of each forecast standard deviation to that of the Bayes linear adjusted forecast. Values greater than one indicate that the Bayes linear adjusted forecasts were sharper than the competitor.

(a) Average marginal sharpness in each region



(b) Distribution of changes in marginal sharpness



6.2.1.3 Marginal forecast calibration

Figure 6.4 shows examples of PIT histograms for forecasts at two locations at selected leadtimes, with the PIT skewness and dispersion across all regions shown in Figure 6.5. Histograms for the remaining regions can be found in Figure B.4 in Appendix B.

The PIT histograms tend to have slight negative skewness at all leadtimes, suggesting that the forecasts retain a slight cold bias regardless of the method used to estimate the discrepancy, and with the magnitude of the bias varying somewhat between regions. The bias is small compared to the spread of the forecasts; even at the shortest leadtimes, when the skewness is most pronounced, observations are more likely to fall in the 50th-80th percentile of the forecast distributions than in the extreme tail. All estimation methods produce PIT histograms with dispersion indices well below one at these leadtimes, reflecting a degree of overdispersiveness in the marginal forecast distributions, with the pooled and Bayes linear adjusted forecast having the lowest dispersion indices.

At leadtimes of greater than a week the MW-corrected forecasts have PIT dispersion indices somewhat greater than one, with too many observations falling in regions of low forecast probability; the AN, pooled, and Bayes linear

adjusted forecasts, which tend to have higher variances at these leadtimes, all have similar dispersion indices that are closer to one, indicating better marginal calibration.

Figure 6.4: PIT histograms showing the marginal calibration of postprocessed forecasts of surface temperatures at selected locations in the north and south of the UK at a range of leadtimes, for forecasts postprocessed using either direct, Bayes linear adjusted or pooled estimates of the discrepancy **Δ**. The dashed line indicates the ideal uniform distribution.

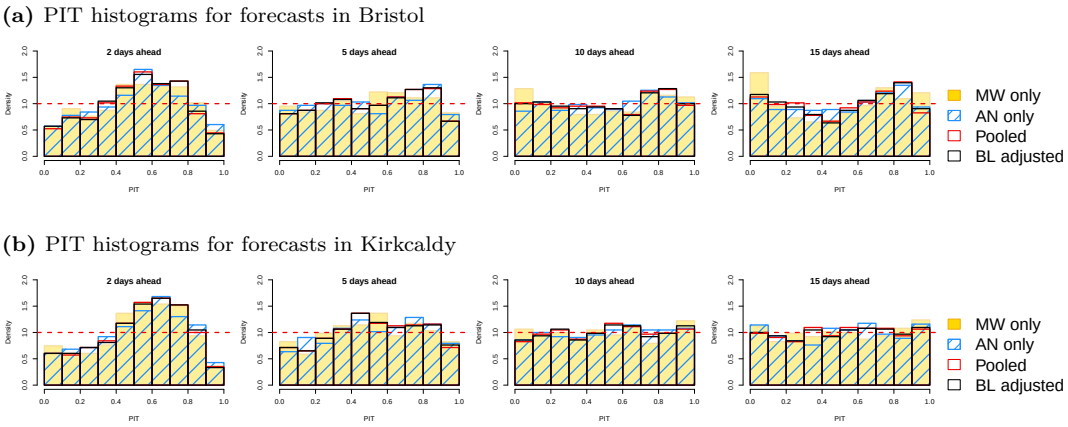
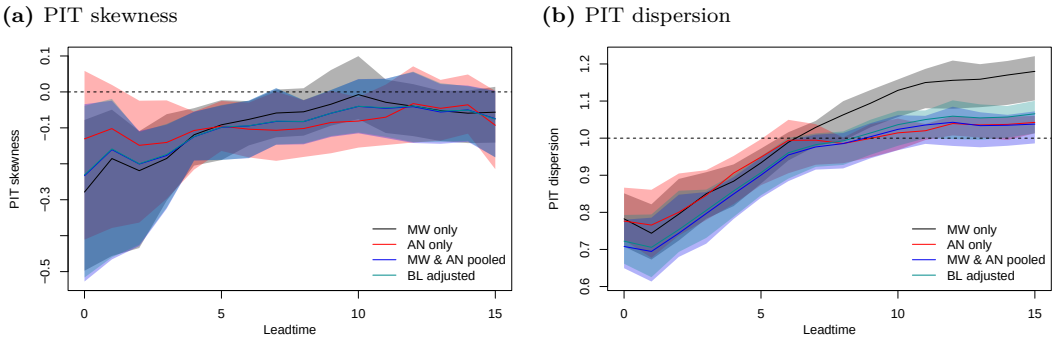


Figure 6.5: Characteristics of the PIT histograms at each leadtime. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



6.2.1.4 Joint forecast calibration

The marginal calibration performance shown in the PIT histograms is reflected and amplified in the Band Depth Rank (BDR) histograms in Figure 6.6. At the shortest leadtimes, the rightmost bins of the histograms are heavily overpopulated for all forecasting methods, indicating that the forecasts are

jointly overdispersive; the pooled and Bayes linear forecasts, being marginally less sharp than the MW and AN-adjusted forecasts, suffer more from this issue. As the leadtime increases, more observations begin to fall too often in the leftmost bins, indicating that the forecasts are too narrow. This issue particularly affects the MW-adjusted forecasts at longer leadtimes, with the histograms for AN-adjusted forecasts being closer to uniform at the longest leadtimes than either the Bayes linear adjusted or pooled forecasts.

Figure 6.6: Band Depth Rank (BDR) histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes.

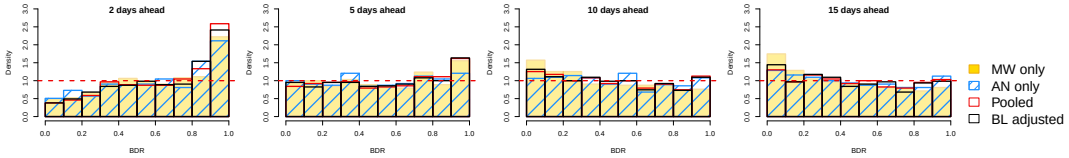


Figure 6.7 shows histograms of the Box Ordinate Transforms of the observations under each set of forecast distributions. The histograms for the MW and AN-adjusted forecasts at all leadtimes are dominated by spikes in the leftmost bins, indicating a high number of observations falling far from the centre of the forecast distribution, indicating that the forecast distributions are both overdispersive and over-correlated.

The BOT histograms for the Bayes linear adjusted and pooled forecasts at the same leadtimes have a more symmetric $\cup$ shape, with a high proportion of the observations falling in either the centre or the extreme tails of the distribution. However, this pattern does not indicate that the forecasts are consistently underestimating the correlations between variables (as, for example, in Figure 3.2f in Section 3.3.2.3). As in the results presented in Section 4.1.3, consideration of the joint distribution of the binned BDRs and BOTs in Figures 6.8c and 6.9c reveals a cluster of observations falling into the leftmost bin of the BOT histograms with BDRs between 0.2 and 0.6, indicating that the dependences between variables are actually typically overestimated by the corresponding forecasts. At the same time, a large number of forecast instances also fall in the top-right corner of the grids for all postprocessing methods,

indicating a group of forecasts for which the verifying observations fall very close to the centre of the forecast distribution.

These two sets of instances remain well separated in the gridplot, suggesting that the distribution of the forecast errors may in fact be a mixture of two different populations. However, the clusters of counts indicating forecasts with too-strong correlations are less pronounced for the Bayes linear adjusted forecasts than for the corresponding MW and AN forecasts, indicating a reduction in the number of overcorrelated forecasts compared to both the MW-adjusted and AN-adjusted forecasts at all leadtimes.

Figure 6.7: Box Ordinate Transform (BOT) histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes. BOTs for pooled-dispersion forecasts are not shown, but typically have a very similar distribution to those of forecasts with Bayes linear adjusted discrepancies.

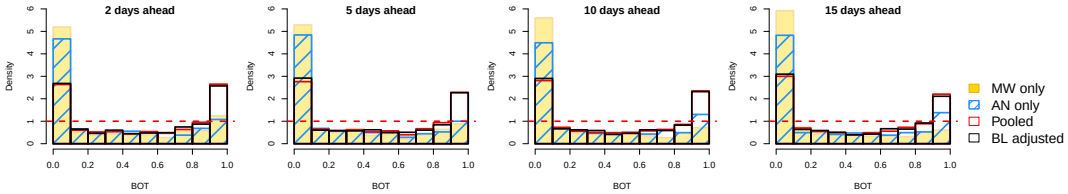


Figure 6.8: Gridplots summarising the joint distribution of the BOT and BDR histogram counts at leadtime 2. Plots are not shown for the pooled forecasts, but closely resemble those of the Bayes linear (BL) adjusted forecasts.

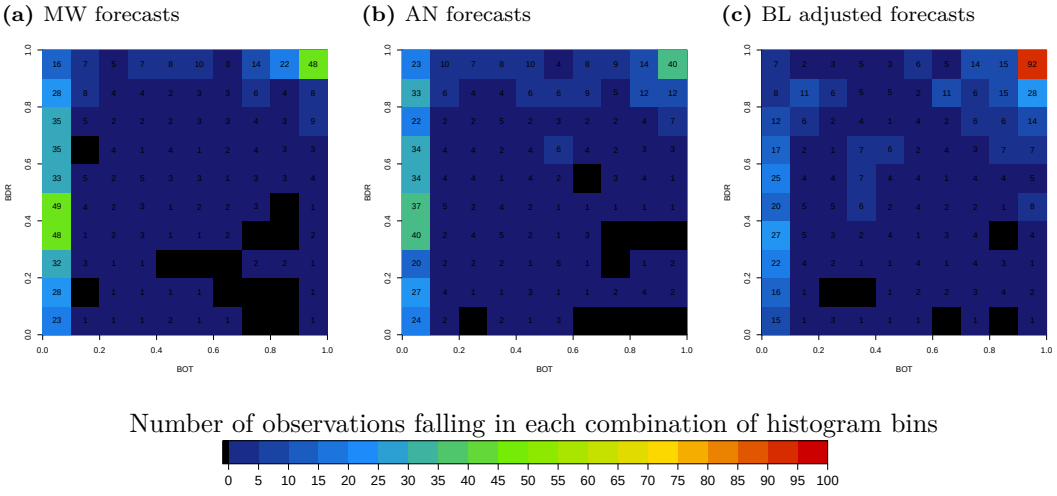
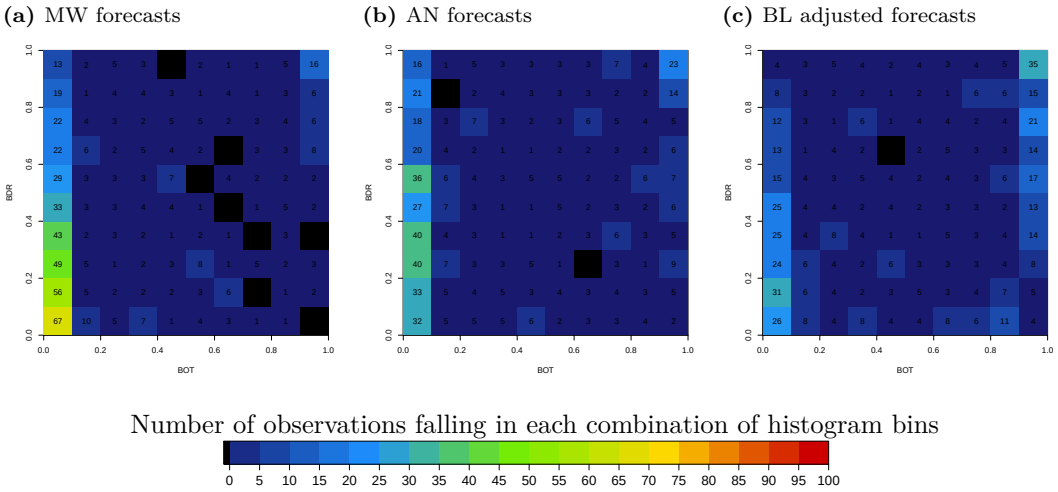


Figure 6.9: Gridplots summarising the joint distribution of the BOT and BDR histogram counts at leadtime 10. Plots are not shown for the pooled forecasts, but closely resemble those of the Bayes linear (BL) adjusted forecasts.



6.2.1.5 Summary

When using the baseline specifications applied here, the Bayes linear adjusted estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ are simple averages of the 'prior' MW estimates and the 'observed' AN estimates. Using either of these sources of information in isolation to postprocess the forecasts results in similar levels of overall forecast accuracy, and as a result, forecasts postprocessed using all four estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ have similar forecast accuracy on average, although the MW-adjusted forecasts have slightly higher MAE at longer leadtimes.

Likewise, at shorter leadtimes all four methods produce forecasts of broadly similar sharpness, being generally overdispersive; the MW- and AN-adjusted forecasts, each being slightly sharper than the Bayes linear adjusted forecasts on average, tend to have better marginal calibration at these leadtimes. At longer leadtimes, the Bayes linear adjusted forecasts are somewhat sharper than the AN-adjusted and pooled-adjustment forecasts, but less sharp than the MW-adjusted forecasts. The BOT-BDR grids in Figures 6.8 and 6.9 indicate that the Bayes linear adjusted forecasts better represent the dependences between variables.

Simply averaging the two available sources of information – whether by

Bayes linear adjustment or simply pooling – has only a small effect on forecast skill in this case, largely because the difference in skill between forecasts adjusted by either source in isolation is small. In particular, it cannot correct biases or calibration errors that are shared by both the MW and AN training sets, such as the tendency to issue forecasts that are overdispersive at shorter leadtimes and underdispersive at longer leadtimes.

6.2.2 Alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$

The 'baseline' forecasts in Section 6.2.1 used $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ with $z = n = 25$, which means that equal weight is given to the two sources of information when adjusting the expectation of $\boldsymbol{\eta}$, which is therefore the mean of the prior $\mathbb{E}[\boldsymbol{\eta}]$ and the observed sample mean $\bar{\boldsymbol{\delta}}$. In this section, the skill of the baseline forecasts is first compared to that of forecasts in which less confidence is placed in the prior estimate of $\boldsymbol{\eta}$; this judgement is reflected by reducing z, the notional sample size used to set the priors. Setting z to a smaller value reflects reduced confidence in the prior expectation $\mathbb{E}[\boldsymbol{\eta}]$, corresponding to a judgement that the MW training set may not be sampling from the same population of errors as the current forecast instance, or that – due to autocorrelation between the forecast errors on consecutive days – the MW training set may have a lower information content than would an independent sample from that population. This reduces the weight on the prior, so forecasts postprocessed using $\boldsymbol{\eta}$ estimated with a smaller notional sample size z will tend to more closely resemble the AN 'observed' forecasts. Results are presented here for $z = \{19, 13, 7\}$ to illustrate the effect of reducing the notional sample size by a range of values.

An alternative approach is to set $\mathbb{V}[\boldsymbol{\eta}]$ based only on the information provided by the 'prior' MW training set, and to treat $\boldsymbol{\eta}$ as if it were a sample mean estimated from z independent observations; then the sampling distribution of $\boldsymbol{\eta}$ has variance $z^{-1}\boldsymbol{\Lambda}$. Although $\boldsymbol{\Lambda}$ itself is unknown, $\mathbb{V}[\boldsymbol{\eta}]$ may be considered to be proportional to the prior expectation of $\boldsymbol{\Lambda}$. Under this specification – and indeed any specification where $\mathbb{V}[\boldsymbol{\eta}]$ is not proportional to $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ – the weights placed on $\mathbb{E}[\boldsymbol{\eta}]$ and $\bar{\boldsymbol{\delta}}$ in the adjusted expectation are no longer scalar,

as can be seen by writing (6.4) as

$$\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}] = \frac{1}{n}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\left(\mathbb{V}[\boldsymbol{\eta}] + \frac{1}{n}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\right)^{-1}\mathbb{E}[\boldsymbol{\eta}] + \mathbb{V}[\boldsymbol{\eta}]\left(\mathbb{V}[\boldsymbol{\eta}] + \frac{1}{n}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]\right)^{-1}\bar{\boldsymbol{\delta}}. \quad (6.13)$$

Generally, if $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ is larger than $\mathbb{V}[\boldsymbol{\eta}]$ then more weight will be placed on the prior $\mathbb{E}[\boldsymbol{\eta}]$ in the adjusted expectation; otherwise, the sample mean discrepancy $\bar{\boldsymbol{\delta}}$ will dominate.

Note that, in the particular case where $\kappa = 0$, $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$, and $z = n$, the adjusted expectation of $\boldsymbol{\Lambda}$ given by (6.3) simplifies to $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = (\mathbb{E}[\boldsymbol{\Lambda}] + \boldsymbol{S_\delta})/2$, and (6.13) becomes

$$\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}] = \left(\mathbb{E}[\boldsymbol{\Lambda}] + \boldsymbol{S_\delta}\right)\left(3\mathbb{E}[\boldsymbol{\Lambda}] + \boldsymbol{S_\delta}\right)^{-1}\mathbb{E}[\boldsymbol{\eta}] + 2\mathbb{E}[\boldsymbol{\Lambda}]\left(3\mathbb{E}[\boldsymbol{\Lambda}] + \boldsymbol{S_\delta}\right)^{-1}\bar{\boldsymbol{\delta}}. \quad (6.14)$$

Under this specification, if $\boldsymbol{S_\delta}$ is larger than $\mathbb{E}[\boldsymbol{\Lambda}]$, more weight will be placed on the prior expectation of $\boldsymbol{\eta}$ than on the observations; conversely, more weight is placed on the observed sample mean $\bar{\boldsymbol{\delta}}$ if $\boldsymbol{S_\delta}$ is smaller than $\mathbb{E}[\boldsymbol{\Lambda}]$. Setting $\mathbb{V}[\boldsymbol{\eta}] \propto \mathbb{E}[\boldsymbol{\Lambda}]$ and expressing equal confidence in both the 'prior' MW and 'observed' AN training sets therefore places more weight on whichever source of information about $\boldsymbol{\eta}$ has the greater precision, which may be a desirable property, particularly if forecast accuracy is expected to be strongly positively correlated with forecast precision. Results are also presented here for this particular case.

6.2.2.1 Bayes linear adjusted estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$

Changing the specification of $\mathbb{V}[\boldsymbol{\eta}]$ has no effect on the adjusted expectation of $\boldsymbol{\Lambda}$ in (6.1), but changes the weights assigned to the prior $\mathbb{E}[\boldsymbol{\eta}]$ and observed sample mean $\bar{\boldsymbol{\delta}}$ in the adjusted expectation of $\boldsymbol{\eta}$ (6.4). All changes in forecast skill in this section therefore occur only through changes in the forecast accuracy, and forecast sharpness is not discussed.

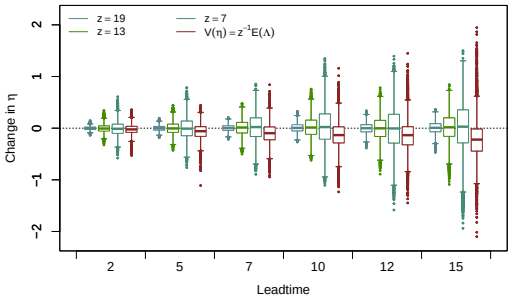
The distribution of changes in $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ under each of the alternative specifications for $\mathbb{V}[\boldsymbol{\eta}]$ is shown in Figure 6.10a. The first three variants use

$\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$, with z reduced by an additional six degrees of freedom in each variant. The distribution of changes is almost symmetric about zero in all cases, so that the average change remains very close to zero.

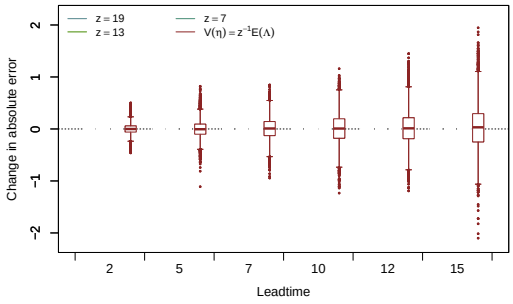
The fourth variant in the plots uses $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$, and so scales the contribution from the prior and observed sample means according to their respective sample precisions. This specification typically produces a wider spread of changes particularly at the longest leadtimes, with the majority of instances having slightly cooler $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ than the corresponding baseline forecasts. This occurs because, particularly at longer leadtimes, the 'prior' MW estimates of $\boldsymbol{\Lambda}$ typically have lower marginal variances, and so more weight is placed on the MW estimates of $\boldsymbol{\eta}$, which are typically slightly cooler than their 'observed' AN counterparts, as discussed in Section 4.3. The result is that forecasts postprocessed with this specification tend to have a larger negative mean discrepancy $\boldsymbol{\eta}$ than either the baseline Bayes linear adjusted forecasts or the MW or AN-adjusted forecasts alone, so are likely to predict slightly warmer temperatures.

Figure 6.10: Distribution of changes from baseline $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ when using alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$, and corresponding change in absolute forecast error. Results are shown for all regions at selected leadtimes .

(a) Change in $\boldsymbol{\eta}$ used to postprocess forecasts



(b) Change in absolute forecast error

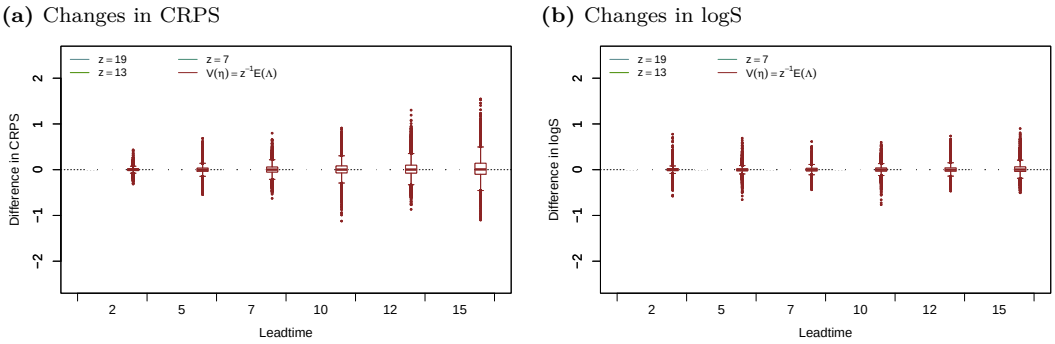


6.2.2.2 Forecast accuracy & skill

Figure 6.10b shows the change in absolute error for forecasts postprocessed using Bayes linear adjustments with alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$. For forecasts using $\mathbb{V}[\boldsymbol{\eta}]$ proportional to $\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$, which place more weight on the

'observed' AN training data in $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ as the notional sample size z is reduced, the distribution of changes in absolute error closely resembles the distribution of changes in $\boldsymbol{\eta}$. For the forecasts using $\mathbb{V}[\boldsymbol{\eta}]$ proportional to $\mathbb{E}[\boldsymbol{\Lambda}]$ instead – represented by the rightmost bars at each leadtime in Figures 6.10a and 6.10b – the MAE changes by somewhat less than the estimate of $\boldsymbol{\eta}$ itself. These changes in MAE are reflected in the changes in the CRPS and logS, shown in Figure 6.11.

Figure 6.11: Distributions (over 630 forecast instances and thirteen locations) of differences in CRPS and logS at selected leadtimes when using alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$ in place of the baseline $z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$. Negative changes indicate that the forecasts with alternative $\mathbb{V}[\boldsymbol{\eta}]$ were more skilful than the baseline forecasts, having lower CRPS/logS.



6.2.2.3 Marginal forecast calibration

Figure 6.12 shows the PIT histograms produced using $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$, where $z = 25$, overlaid with those obtained using $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$ with the lowest value of $z = 7$ and the highest of $= 25$; all values of z intermediate between the two produced very similar histograms, and so are not shown. The skewness and dispersion of the PIT histograms are summarised in Figure 6.13. There is almost no change in PIT skewness even for the smallest choice of z tested, while histograms for forecasts using $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$ have slightly lower negative skewness at longer leadtimes, suggesting that the slightly larger negative discrepancies have somewhat offset the residual cold bias noted in Section 6.2.1.3. PIT dispersion indices, which are more sensitive to changes in the forecast variance than the forecast mean, are almost unchanged at all leadtimes.

Figure 6.12: PIT histograms showing the marginal calibration of postprocessed forecasts of surface temperatures at selected locations in the north and south of the UK at a range of leadtimes.

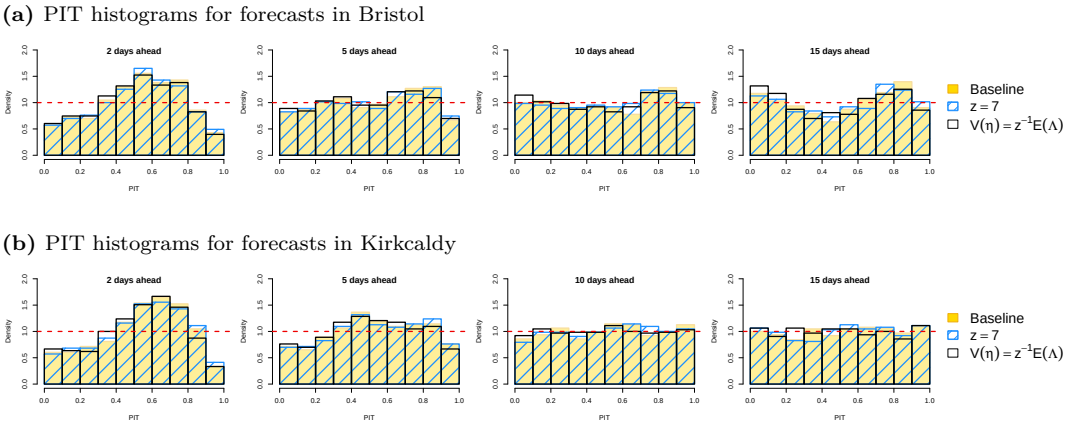
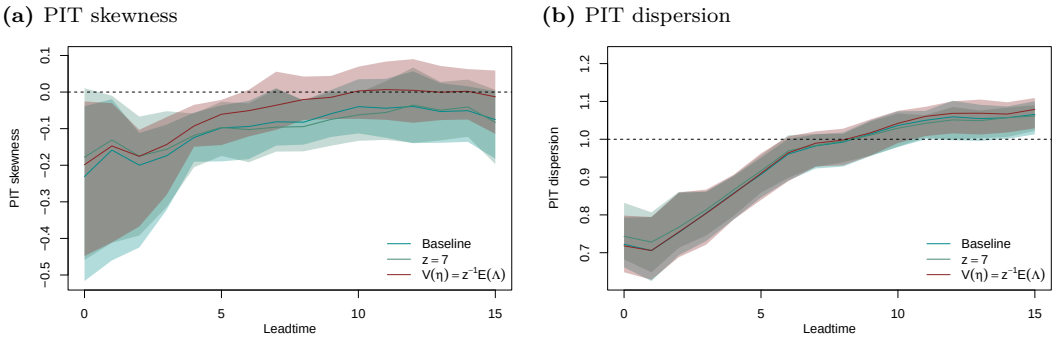


Figure 6.13: Characteristics of the PIT histograms at each leadtime for forecasts using alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



Because changing the specification of $\mathbb{V}[\boldsymbol{\eta}]$ makes no difference to the adjusted expectation of $\boldsymbol{\Lambda}$, there will be no changes to the BOT and BDR histograms that cannot be explained in terms of the marginal forecasts already considered. Results for joint calibration are therefore not presented here.

6.2.2.4 Summary

In the current study, changing the prior variance of the population mean $\mathbb{V}[\boldsymbol{\eta}]$ has a minimal effect on the overall skill of the forecasts, although the mean vectors of individual forecasts may change by up to 2°C at longer leadtimes when z is much smaller than n, or when $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$. If one method of

estimating the mean discrepancy $\boldsymbol{\eta}$ were known to consistently outperform the other in terms of forecast accuracy, there would be a greater justification for adjusting the value of z to take advantage of the fact. Likewise, if there is expected to be a strong relationship between forecast precision & forecast accuracy in both the prior and observed sources of information, it would be advisable to use $\mathbb{V}[\boldsymbol{\eta}] \propto \mathbb{E}[\boldsymbol{\Lambda}]$, so that the sharper forecasts place more weight on the more accurate source of information. However, the two sets of forecasts used in this case are generally sufficiently similar, and insufficiently well calibrated, that changing the specification of $\mathbb{V}[\boldsymbol{\eta}]$ has almost no effect on the average forecast skill.

6.2.3 Sensitivity to choice of ν

When deriving the Bayes linear update in Section 5.3.5.2, the relationship between $\mathbb{E}[\boldsymbol{\Lambda}] = \mathbf{V}_R$ and $\mathbb{V}[\boldsymbol{\Lambda}] = \mathbf{V}_M$ was specified by treating the population variance $\boldsymbol{\Lambda} = \mathcal{M}(\mathbf{V})$ as if it were a random covariance matrix with an inverse-Wishart distribution with ν degrees of freedom and scatter matrix $\boldsymbol{\Psi}$, so that

$$\mathbb{E}[\boldsymbol{\Lambda}] = \frac{1}{\nu - m - 1}\boldsymbol{\Psi} \tag{6.15}$$

and

$$\begin{aligned}\mathbb{V}\Big[\mathrm{vec}\Big(\boldsymbol{\Lambda}\Big)\Big] &= \frac{2\mathrm{vec}\Big(\boldsymbol{\Psi}\Big)\,\mathrm{vec}\Big(\boldsymbol{\Psi}\Big)' + 2(\nu-m-1)\mathbf{N}_m\,(\boldsymbol{\Psi}\otimes\boldsymbol{\Psi})}{(\nu-m)(\nu-m-1)^2(\nu-m-3)}. \\ &= \frac{2\mathrm{vec}\Big(\mathbb{E}[\boldsymbol{\Lambda}]\Big)\,\mathrm{vec}\Big(\mathbb{E}[\boldsymbol{\Lambda}]\Big)' + 2(\nu-m-1)\mathbf{N}_m\,(\mathbb{E}[\boldsymbol{\Lambda}]\otimes\mathbb{E}[\boldsymbol{\Lambda}])}{(\nu-m)(\nu-m-3)}\end{aligned} \tag{6.16}$$

In this way, confidence in the prior expectation of $\boldsymbol{\Lambda}$ can be expressed through the scalar parameter ν, with $\nu - m$ reflecting the notional equivalent sample size z used to determine $\mathbb{E}[\boldsymbol{\Lambda}]$. While higher values of ν could, in theory, be specified to increase the weight on the prior, it would be hard to justify placing greater confidence in the prior expectation than the actual sample size would suggest. Decreasing ν to reflect a lack of confidence that the prior MW

training set is sampled independently from the same population as the current forecast instance is a more defensible choice, although the degrees of freedom must be greater than $m+1$ in order that the expectation of $\boldsymbol{\Lambda}$ be defined; similarly, the covariance matrix of an inverse-Wishart random variable is not defined when $\nu \leq m+3$. A reasonable range of values for ν is therefore to set $m+4 \leq \nu \leq z+m$.

In this section, Bayes linear adjustments are carried out using $\nu = z+m$; $\nu = z$, reflecting moderate confidence in the prior; and $\nu = m+4$, reflecting extremely low confidence in the prior. Reducing ν in this way reduces the weight assigned to the prior $\mathbb{E}[\boldsymbol{\Lambda}]$ in the adjusted expectation of $\boldsymbol{\Lambda}$; as a result, the Bayes linear adjusted expectation will more closely resemble the estimate of $\boldsymbol{\Lambda}$ derived from the AN training set as ν decreases. All other quantities are defined as in Section 6.2.1.

The confidence parameter ν only plays a role in adjusting the expected value of $\boldsymbol{\Lambda}$: the adjusted expectation of $\boldsymbol{\eta}$ is the same for all values of ν. Consequently, the accuracy of the deterministic postprocessed forecasts with varying ν is not discussed here.

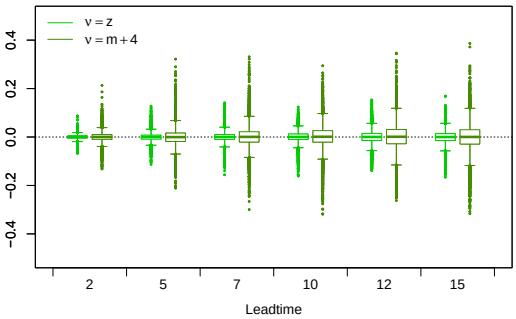
6.2.3.1 Forecast skill, sharpness & calibration

The differences in CRPS and log score resulting from reducing ν from the baseline of $z+m$ are shown in Figure 6.14: both sets of changes are of small magnitude and broadly symmetric about zero, with the average forecast skill almost unchanged from the baseline case.

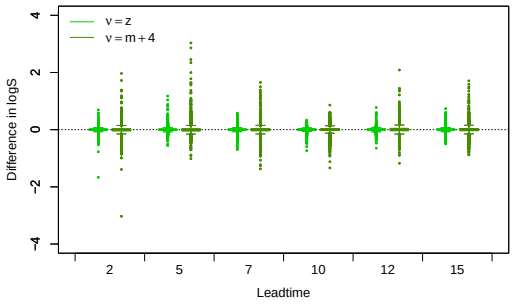
Decreasing the prior degrees of freedom to reflect reduced confidence in the prior does not necessarily result in a higher adjusted variance: instead, more weight is given to the 'observed' sample variance $\boldsymbol{S_\delta}$ obtained from the AN-adjusted forecasts, so reducing the prior degrees of freedom will result in a lower combined variance only where $\boldsymbol{S_\delta}$ is sharper than the 'prior' $\mathbb{E}[\boldsymbol{\Lambda}]$. At leadtimes greater than a week, the MW-adjusted forecasts used to set $\mathbb{E}[\boldsymbol{\Lambda}]$ are typically significantly sharper than the AN-adjusted forecasts used to adjust

Figure 6.14: Distribution of differences in CRPS and logS when ν is changed from the baseline of $\nu = z + m$. Differences are shown for all regions at selected leadtimes. Negative changes indicate that the forecasts with smaller ν have lower scores, indicating greater skill.

(a) Difference in CRPS



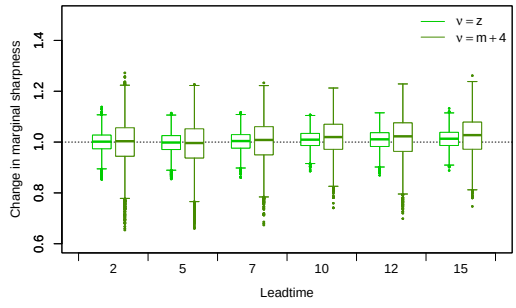
(b) Difference in log score



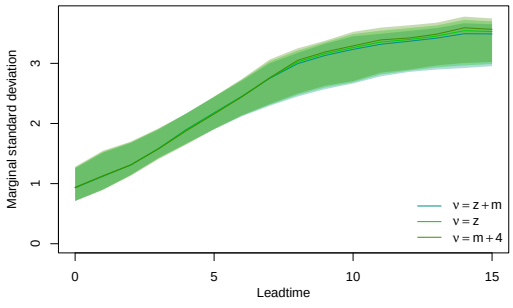
the variance, and so reducing ν tends to result in a slight increase, on average, in forecast variance at these leadtimes, as shown in Figure 6.15a. However, the effect is very small, with only around a 2% increase in variance when ν is reduced from $z + m$ to $m + 4$ (Figure 6.15b).

Figure 6.15: Distribution of changes in marginal sharpness of postprocessed forecasts when ν is reduced from the baseline of $\nu = z + m$, across all regions at selected leadtimes, expressed as the ratio of each forecast standard deviation to that of the Bayes linear adjusted forecast, where values greater than one indicate that the Bayes linear adjusted forecasts were sharper than the alternative; and spread of average marginal sharpness in each region for each choice of ν, where the lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values.

(a) Distribution of changes in marginal sharpness



(b) Average marginal sharpness in each region



Because the changes in forecast variance due to reducing ν are generally very small, the effect on the marginal calibration is negligible, with PIT histograms for all values of ν having almost the same PIT skewness and dispersion even at the longest leadtimes (Figure 6.16). Similarly, the BDR

histograms shown in Figure 6.17 are almost identical for all values of ν.

Figure 6.16: Characteristics of the PIT histograms at each leadtime for forecasts using alternative specifications of ν. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.

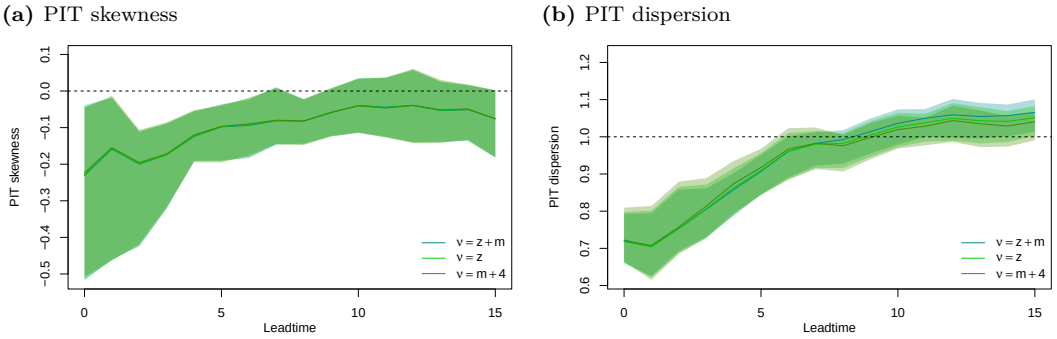
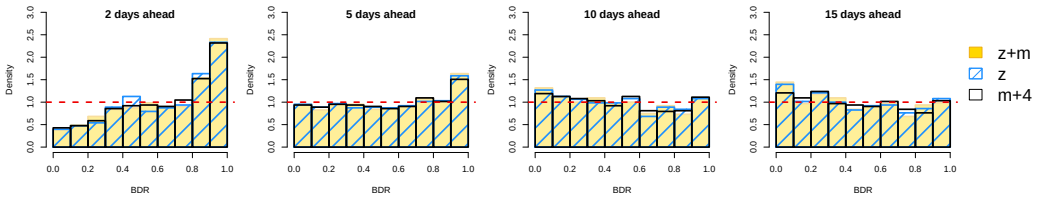
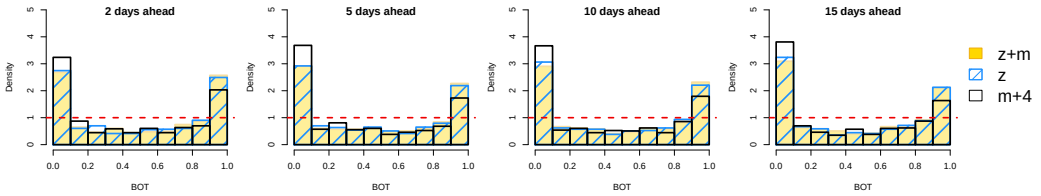


Figure 6.17: BDR histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes when $\boldsymbol{\Lambda}$ is estimated using a range of values of ν.



Interestingly, a somewhat larger change can be observed in the BOT histograms in Figure 6.18; while forecasts using $\nu = z + m$ and $\nu = z$ produce almost identical BOT histograms at all leadtimes, histograms of forecasts with $\nu = m + 4$ display a small but distinct shift, with lower counts in the rightmost bin – indicating fewer observations in the centre of the joint distribution – and higher counts in the leftmost bin, indicating more observations falling in regions of very low forecast probability. The change is small, but suggests that the dependence structure of forecasts with lower ν and correspondingly greater contribution from the 'observed' AN-adjusted forecasts is too strongly specified. This supports the conclusion in Section 6.2.1.4 that the correlation structure is better represented when the two sources of information are combined.

Figure 6.18: BOT histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes when $\boldsymbol{\Lambda}$ is estimated using a range of values of ν.



6.2.3.2 Summary

The mean forecast vectors are unaffected by changes in ν, and although the change in weighting between MW and AN-adjusted forecasts may have a large impact on the variance of individual forecasts, the average change in forecast calibration over all instances is very small. In an operational setting it may be useful to vary ν across leadtimes based on the earlier performance of the prior and observed datasets; however, the change in forecast skill is so small in this data set that such an approach is unlikely to yield any substantial benefits, so is not investigated further here.

6.2.4 Effect of specifying non-zero κ

This section investigates the effect of increasing or decreasing the kurtosis parameter $\kappa = \mathbb{K}\text{ur}[\mathbf{X}]/3$ to reflect a prior belief that the forecast errors $\boldsymbol{\delta}$ are drawn not from a normal distribution but from a heavier- or lighter-tailed distribution – meaning that any sample of forecast errors will tend to contain a higher or lower proportion of outliers than would a sample drawn from a normal distribution. Bayes linear adjustment of $\boldsymbol{\Lambda}$ is initially carried out using a fixed value of κ at all leadtimes. Results are then presented for the skill of forecasts using more realistic approaches, either by using the sample kurtosis of the training sets used to determine the other prior quantities, or by using the sample kurtosis over the full archive at each leadtime.

With the exception of κ, all adjustments in this section use the same specifications that were used in Section 6.2.1, with $\mathbb{V}[\boldsymbol{\eta}] \propto \mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$, $\nu = z + m$, and $z = 25$. The adjusted expectation of $\boldsymbol{\eta}$ is therefore the same as (6.11),

and $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ is unaffected by the choice of κ. Forecast accuracy is therefore not reported here, and any changes in forecast skill can be directly attributed to the change in κ alone.

It is worth noting that changing the kurtosis parameter κ modifies the adjusted expectation of the covariance matrix $\boldsymbol{\Lambda}$, but does not change the shape of the Bayes linear adjusted distribution as it would in a fully probabilistic Bayesian analysis: all quantities are still assumed to be fully specified by their expectations and covariance matrices.

6.2.4.1 Effect of κ on adjusted expectation of $\boldsymbol{\Lambda}$

From (6.1) and (6.2), the adjusted expectation of $\boldsymbol{\Lambda}$ can be written as

$$\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = w_1 \boldsymbol{S_\delta} + \left\{(1-w_1) + w_2\left[m - \operatorname{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)\right]\right\}\mathbb{E}[\boldsymbol{\Lambda}] \tag{6.17}$$

where

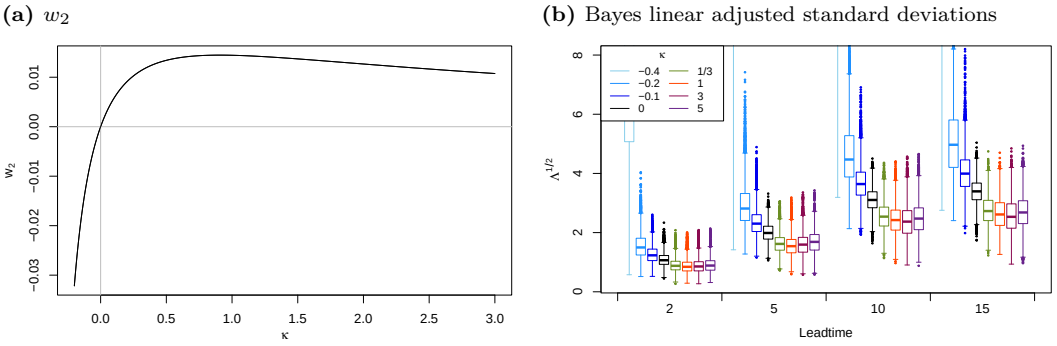
$$w_1 = \frac{n(n-1)}{\tilde{\kappa}+n\tilde{\nu}} \quad \text{and} \quad w_2 = \frac{n(n-1)(\nu-m-3)\tilde{\kappa}}{[\tilde{\kappa}+n\tilde{\nu}]\,[(m+2)(\nu-m-1)\tilde{\kappa}+2(\nu-1)n\tilde{\nu}]}, \tag{6.18}$$

where $\tilde{\kappa} = \kappa(n-1)(\nu-m)$ and $\tilde{\nu} = \nu-m-1+n-1$. Using the baseline specifications of $z=n=25$, $m=13$, and $\nu = z+m = 38$, w_2 varies with κ as shown in Figure 6.19a, with positive κ producing positive values of w_2, and negative choices of κ producing negative w_2. In the dataset used in this study, $\operatorname{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)$ was typically larger than m; when this is the case, positive values of κ will result in a negative contribution from $w_2\left[m - \operatorname{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)\right]$ and so in lower adjusted variances than would using $\kappa = 0$, while negative values of κ will produce forecast distributions with higher variances than the baseline. w_2 is maximised by setting κ to approximately 0.9, as described in Section 5.3.7.1: Bayes linear adjustments using $\kappa \approx 0.9$ are therefore expected to produce the sharpest possible forecasts, although the difference in sharpness is likely to be small for values of κ moderately close to this value unless

$\left[m - \mathrm{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)\right]\mathbb{E}[\boldsymbol{\Lambda}]$ is very large.

This expected pattern of behaviour is clearly reflected in Figure 6.19b, which shows the sharpness of the Bayes linear adjusted forecast discrepancies for a single region under a range of arbitrarily chosen values of κ. The same pattern is found in all regions and at all leadtimes, with the difference between the baseline and κ-adjusted estimates generally being larger at longer leadtimes, when $\mathbb{E}[\boldsymbol{\Lambda}]$ is typically larger.

Figure 6.19: Values of w_2 in the third term in (6.17) with $n = 25$, $m = 13$, and $\nu = 38$, and the distributions (over all 630 forecast instances and all thirteen regions, at selected leadtimes) of the resulting Bayes linear adjusted marginal standard deviations for several choices of κ. The Y-axis of (b) is truncated to show the detail.



In Section 6.2.1.3 it was observed that forecasts using $\kappa = 0$ are typically under-confident at shorter leadtimes and over-confident at longer leadtimes, suggesting that using a fixed value of κ in the Bayes linear adjustment at all leadtimes will result in significantly worse forecast calibration at either longer or shorter leadtimes, and therefore worse forecast skill overall. Better results may therefore be expected if the kurtosis is allowed to vary with the forecast leadtime, and perhaps with the forecast instance.

6.2.4.2 More realistic estimates of κ

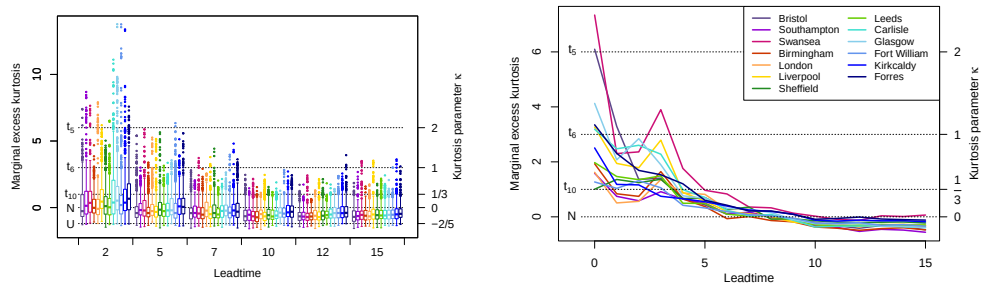
The value of κ may be set separately for each forecast instance by evaluating the kurtosis of the corresponding training set directly; however, estimates of kurtosis from such a small sample are likely to be very variable (Fisher, 1930). Figure 6.20a shows distributions of sample kurtoses for all available

25-member MW training sets within the forecast archive described in Section 2.1.2; particularly at the shortest leadtimes, a wide spread of values is obtained, both within and between regions. Higher sample kurtosis values tend to be observed at shorter leadtimes, with most training sets at longer leadtimes having slightly negative kurtosis. For operational purposes, the distributions of sample kurtoses at each leadtime are fairly similar across the regions.

Figure 6.20: Marginal kurtosis of forecast errors at each leadtime and region

(a) Distribution of sample kurtosis of each 25-member MW training set in each region for selected leadtimes

(b) Sample kurtosis of all 630 forecast errors at each leadtime and region



Where a large archive of previous training cases is available, a more stable estimate of the common kurtosis parameter κ at each leadtime may be obtained by computing the kurtosis over all available training cases for each region, and averaging over the m regions. Figure 6.20b shows the marginal kurtosis of the 630 forecast errors at each leadtime and each region in the study archive. The plot suggests that it would be appropriate to set $\kappa > 0$ for forecasts at leadtimes of up to roughly a week, with errors in most regions having marginal excess kurtosis between 1 and 3 at up to four days ahead, and of less than 1 at four to five days ahead; at longer leadtimes the excess kurtosis is again negative.

Here, two practical approaches to the estimation of κ are investigated. In the first, values of κ are estimated independently for each forecast instance as one-third of the excess kurtosis $\mathbb{K}\mathrm{ur}[\boldsymbol{\delta}]$ of the discrepancies in the 'prior' MW training set, with the parameter estimated in this way denoted by κ_{ts} (recall from Section 5.3.5 that $\kappa = \mathbb{K}\mathrm{ur}[\mathbf{X}]/3$, where $\mathbb{K}\mathrm{ur}[\mathbf{X}]$ is the excess kurtosis of the quantity of interest $\mathbf{X}$). The second approach adjusts all forecast instances at a given leadtime using the same value of κ, denoted κ_{lt}, where κ_{lt} is one-third

of the excess kurtosis of the 630 training cases in the archive at that leadtime. Note that the subscripts 'ts' and 'lt' are not indices, but merely indicate which of the methods was used to estimate κ for each set of forecasts.

6.2.4.3 Replacing invalid κ

The Bayes linear adjusted expectation obtained using (6.17) is not guaranteed to be positive-semidefinite when $\kappa \neq 0$, as discussed in Section 5.3.7.4. A potential solution to this problem is to replace the proposed value of κ with one that will result in a valid covariance matrix in (5.177). The range of values of κ for which $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ is guaranteed to be positive-semidefinite was found in Section 5.3.7.4 to be

$$\left\{\tilde{\kappa} \leq \frac{-b}{2a} - \sqrt{\frac{b^2-4ac}{4a^2}} \cup \tilde{\kappa} \geq \frac{-b}{2a} + \sqrt{\frac{b^2-4ac}{4a^2}}\right\}, \tag{6.19}$$

where

$$a = (m+2)(\nu-m-1),$$
$$b = n(n-1)(\nu-m-3)\left[m - \operatorname{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)\right] + n(\nu-m-1)^2(m+2) + 2(\nu-1)n\tilde{\nu},$$
$$c = 2n^2(\nu-m-1)(\nu-1)\tilde{\nu}.$$

Where the proposed value of κ does not fall within this set, a replacement must be chosen. Reasonable candidates include the closest valid κ to the proposed value, or the value of κ that produces the most similar value w_2 to that given by the proposed κ.

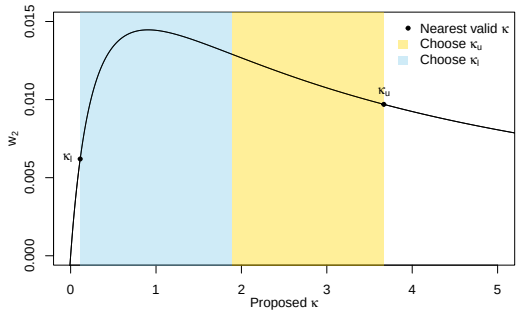
Figure 6.21 shows the replacement values of κ corresponding to each of these methods of replacement: κ_l is the largest valid value of κ such that $\tilde{\kappa} \leq \frac{-b}{2a} - \sqrt{\frac{b^2-4ac}{4a^2}}$, and κ_u the smallest value of κ such that $\tilde{\kappa} \geq \frac{-b}{2a} + \sqrt{\frac{b^2-4ac}{4a^2}}$. Where the proposed value of κ falls within the blue shaded region, κ_l will be chosen as the replacement value; where the proposed value falls within the yellow region, the replacement value will be κ_u. Where the closest value of w_2 is used to choose the replacement, there is a range of invalid choices

of κ where $w_2(\kappa_l) < w_2(\kappa) < w_2(\kappa_u)$, denoted by the orange region in Figure 6.21b, for which there is always a replacement value greater than κ_u that matches $w_2(\kappa)$ exactly: this potentially much larger replacement value of κ would always be chosen under this strategy, suggesting that it may result in physically implausible choices for κ.

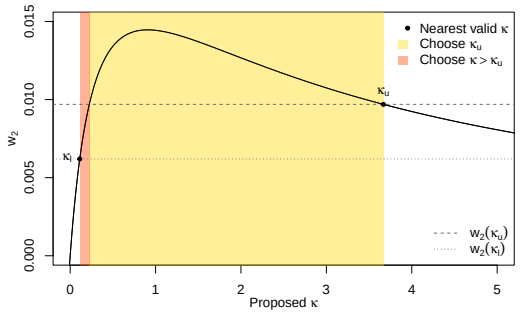
This assertion is supported by the empirical evidence from this study: in the dataset used here, 991 proposed values of κ_{ts} (out of $630 \times 16 = 10080$ forecast instances) resulted in invalid $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$; in 984 of these cases the closest valid κ to these proposals was κ_l, while κ_u was selected in only seven instances, suggesting that the proposed values of κ were generally quite close to zero. When the replacement was chosen by matching the closest value of w_2 κ, κ_u was selected in 748 of the 991 cases. In the 243 remaining cases where the proposed κ was very close to κ_l, the replacement values were larger than κ_u, and 143 of these were greater than two, the largest of the plausible values of κ suggested in Table 5.1. This supports the assertion made above that the closest-w_2 approach is likely to result in physically implausible choices for κ.

Figure 6.21: Example of the likely values of κ chosen by each replacement method. κ_l and κ_u denote the closest valid values of κ to proposals in the shaded region.

(a) Closest value of κ to proposal



(b) Closest value of w_2 to proposal



Replacing the proposed κ with either κ_l or κ_u will set the term in braces $\{\}$ in (6.17) to zero, removing any contribution from $\mathbb{E}[\boldsymbol{\Lambda}]$ to the adjusted expectation of $\boldsymbol{\Lambda}$. In this case, $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = w_1 \boldsymbol{S_\delta}$, where from (6.18), w_1 is also

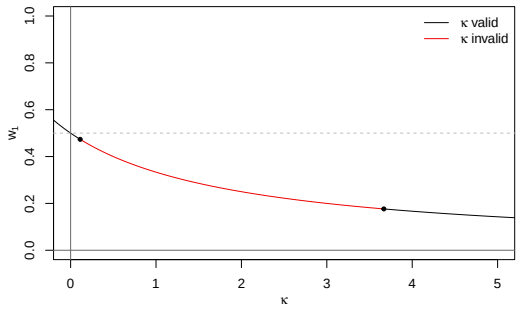
determined by κ, with

$$w_1 = \frac{n(n-1)}{\kappa(n-1)(\nu-m)+n(\nu-m-1+n-1)}. \tag{6.20}$$

This quantity, illustrated in Figure 6.22a, has the value 0.5 when $\kappa = 0$ and decays with increasing κ: if κ_l is chosen as the replacement, w_1 will be close to 0.5, while if κ_l is chosen as the replacement, w_1 will be somewhat smaller. When – under the nearest-w_2 strategy – $\kappa > \kappa_u$ is used as a replacement for the proposal, w_1 will be smaller still, although $\mathbb{E}[\boldsymbol{\Lambda}]$ will also make a non-zero contribution to $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$. Either approach to choosing a replacement value of κ will therefore result in a relatively small adjusted variance matrix.

Figure 6.22: Scaling factors assigned to $\boldsymbol{S_\delta}$ and $\mathbb{E}[\boldsymbol{\Lambda}]$ in (6.17) for a range of κ when $\nu - m = n$; in this example, $\text{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right) = 98$, but the shape of the curve will be the same regardless of the value of $\text{tr}\left(\mathbb{E}[\boldsymbol{\Lambda}]^{-1}\boldsymbol{S_\delta}\right)$.

(a) Scaling applied to $\boldsymbol{S_\delta}$



(b) Scaling applied to $\mathbb{E}[\boldsymbol{\Lambda}]$

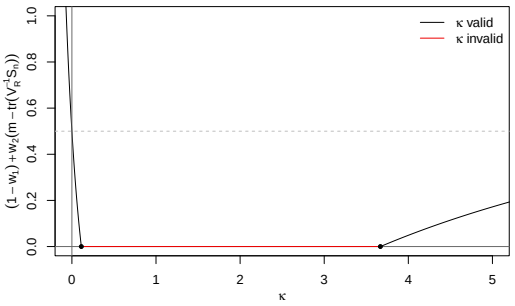
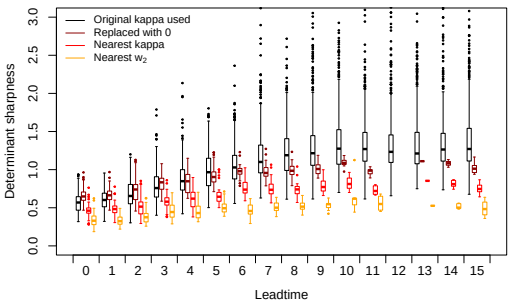


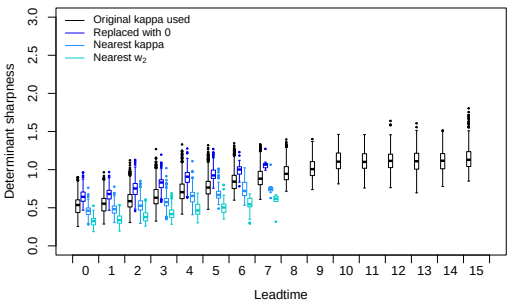
Figure 6.23 illustrates this, showing the distribution of the determinant sharpness (DS) of the postprocessed forecasts under various specifications of κ. The left-hand boxplot for each leadtime shows the distribution of DS for that subset of forecasts for which the original value of κ produced a valid covariance matrix, indicating the range of sharpness that might be considered reasonable under the original specifications. The remaining boxplots at each leadtime show the distribution of DS for those forecast instances for which the original choice of κ was replaced: either with $\kappa = 0$, with the nearest valid value of κ, or with the value of κ giving the closest value of w_2 to that obtained using the original choice of κ.

Figure 6.23: Distribution of forecast determinant sharpness (DS) at each leadtime when invalid κ is replaced with an alternative.
Black boxplots show the distribution of DS for the subset of forecasts where no replacement was necessary; coloured boxplots show the distribution of DS for the subset of forecasts where κ was replaced with an alternative. Fewer observations appear in the coloured boxplots at longer leadtimes because the proposed κ is less frequently invalid at these leadtimes.

(a) Original κ is κ_{ts}.
Plot is truncated at DS=3 for clarity.



(b) Original κ is κ_{lt}



Regardless of whether κ_{ts} or κ_{lt} was originally proposed, setting $\kappa = 0$ is a conservative choice of replacement: for every forecast instance, replacing the proposed value of κ with zero resulted in forecast distributions that were less sharp than if a non-zero replacement was used. The DS of forecasts postprocessed using the closest value of κ tend to fall in the lower tail of the distribution of DS of the forecasts for which the original κ was valid, but still fairly close to the centre of the distribution, suggesting that this choice most closely replicates the desired behaviour under the original specifications. When κ is replaced with the value giving the closest valid w_2, the DS of the adjusted forecasts is typically lower still, frequently falling outside the range of DS for the 'uncorrected' forecasts, and indicating that this approach often produces very over-confident forecasts. Because of this tendency, coupled with the fact that choosing the nearest-w_2 κ generally produces a value of κ that is very far from that originally proposed, and which may be extremely large, choosing the nearest-w_2 replacement for the proposed κ is not recommended.

In what follows, for any instance where the adjusted expectation $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ was not positive definite, the adjustment was carried out again using the closest valid κ, and that estimate of $\boldsymbol{\Lambda}$ used instead, ensuring that valid postprocessed

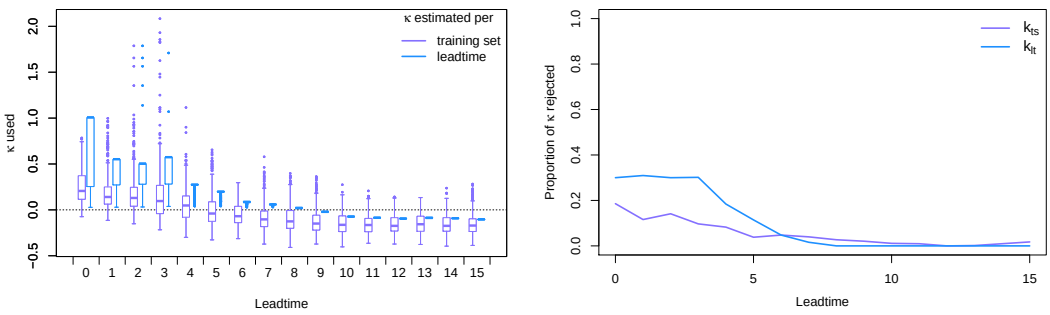
covariance matrices were obtained for all forecast instances. This approach to replacing κ ensures the greatest possible consistency with the original proposed values, and as shown in Figure 6.23, produces more plausible adjusted expectations of the covariance matrix.

Figure 6.24a shows the range of values of κ used to carry out the Bayes linear adjustments. The values of κ_{ts} estimated for each training set are generally lower than those of κ_{lt} estimated across all training cases at each leadtime, although a few forecast instances received substantially higher estimates. At the shortest leadtimes, κ_{lt} is close to the sharpness-maximising value of $\kappa = 0.9$; as a result, κ_{lt} was replaced with the nearest valid κ in around 30% of all forecast instances (Figure 6.24b), while the lower κ_{ts} estimates were more likely to produce a valid variance-covariance matrix. At longer leadtimes both methods usually produced an estimate of κ close enough to zero that the resulting adjusted expectation of $\mathbf{\Lambda}$ was positive-semidefinite, with κ_{lt} always producing valid covariance matrices at leadtimes of eight or more days. Note also that forecasts postprocessed using the proposed κ_{ts} were much more likely to produce very large adjusted expectations, and therefore to produce forecasts with large determinant sharpness in Figure 6.23, while those using κ_{lt} produced less variable adjusted expectations

Figure 6.24: Distribution of values of κ used to produce valid adjusted expectations of $\mathbf{\Lambda}$ at each leadtime. For κ_{lt} a single value of κ was originally proposed at each leadtime; the boxes and whiskers therefore indicate the spread of replacement values.

(a) Values of κ used in Bayes linear adjustment.

(b) Proportion of forecast instances in which $\kappa \neq 0$ resulted in an invalid estimate of $\mathbf{\Lambda}$, and was replaced with the nearest valid κ.



6.2.4.4 Forecast sharpness and skill

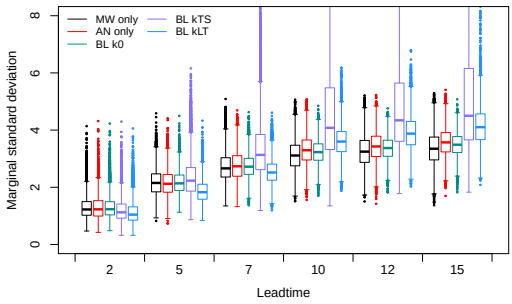
Figure 6.25a shows the spread of the marginal standard deviations of the postprocessed forecasts, with the mean sharpness and spread across all regions in Figure 6.25b. At the very shortest leadtimes, forecasts using non-zero kappa are typically sharper than not only the baseline forecasts using $\kappa = 0$, but also the forecasts postprocessed using $\boldsymbol{\Lambda}$ estimated from the MW or AN training sets alone; in Section 6.2.1 the baseline forecasts and those adjusted using only the MW or AN training sets were found to be insufficiently sharp, on average, at these leadtimes, so this change might be expected to result in consistently better-calibrated forecasts. At leadtimes of five days ahead, the median value of κ_{ts} is less than zero, and as a result, the forecasts using κ_{ts} tend to be less sharp than those using κ_{lt}. The κ_{lt} forecasts, on the other hand, remain sharper on average than the baseline forecasts or those using MW or AN estimates of $\boldsymbol{\Lambda}$ alone.

At longer leadtimes, where κ_{ts} and κ_{lt} are generally both less than zero, both estimates of κ generally result in less sharp forecasts than the baseline. These are the leadtimes at which the AN-adjusted, baseline and particularly the MW-adjusted forecasts tend to be over-confident, so increasing the variance to some extent may be expected to result in improved calibration. However, due to the extremely low values of κ estimated from the training data the variance of κ_{ts} forecasts is often much higher than that of the κ_{lt} or baseline forecasts, which is likely to result in a different form of miscalibration.

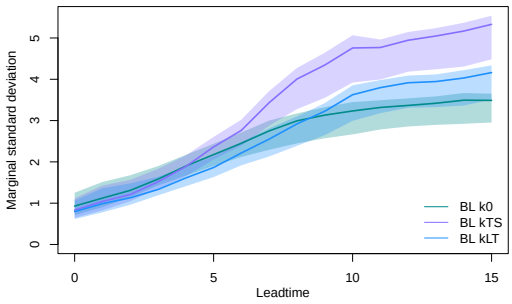
Figure 6.26 shows the difference in CRPS and logarithmic score when $\kappa = 0$ is replaced with either κ_{ts} or κ_{lt} in the Bayes linear adjustment. The κ_{lt}-adjusted forecasts achieve similar CRPS to the baseline forecasts at all leadtimes. However, while the κ_{ts}-adjusted forecasts receive similar scores at the shortest leadtimes, their performance becomes increasingly variable with increasing leadtime, with some instances receiving substantially worse CRPS under this specification than with $\kappa = 0$; the effect is large enough that the mean CRPS for the κ_{ts} forecasts is 0.2°C higher at longer leadtimes, suggesting

Figure 6.25: Distribution of marginal standard deviations of all 630 postprocessed forecasts in all regions at selected leadtimes, and spread of regional mean standard deviation for selected discrepancies: the lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values.
The model acronyms in the legend denote forecasts corrected using a discrepancy estimated directly from the MW or AN training sets (MW only, AN only); and forecasts corrected using a Bayes linear adjusted discrepancy with κ set to zero (BL κ0), estimated per MW training set for each instance (BL κ_{ts}), and estimated per leadtime (BL κ_{lt}). In instances where the proposed κ_{ts} or κ_{lt} were found to be invalid, the nearest valid κ was used.

(a) Marginal standard deviations (truncated to show detail)



(b) Average marginal sharpness in each region



that the κ_{ts} forecasts should be considered to be particularly poorly calibrated.

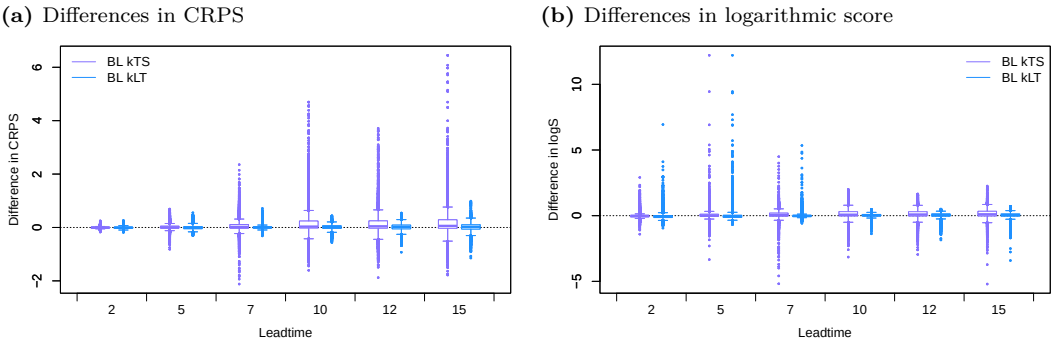
The logarithmic scores display a slightly different pattern: at the shortest leadtimes, both κ_{ts} and κ_{lt}-adjusted forecasts are more likely to be less skilful than the baseline forecasts when evaluated by this metric, while at longer leadtimes, they are likely to be more skilful. The mean logarithmic score for the κ_{ts}-adjusted forecasts is approximately 6% higher than that of the baseline forecasts at the longest leadtimes, while the κ_{lt}-adjusted forecasts have a similar mean logarithmic score to the baseline forecasts at all leadtimes.

6.2.4.5 Marginal forecast calibration

The effect of the changing forecast sharpness can be clearly seen in the PIT histograms in Figure 6.27; histograms for regions not shown in the main text can be found in Figure B.5 in Appendix B.

Figure 6.28 summarises the characteristics of the PIT histograms for all regions; the PIT skewness is broadly similar for all three methods at all

Figure 6.26: Distribution of differences in CRPS and logS when realistic values of κ are used in Bayes linear adjustment instead of the baseline $\kappa = 0$. Results are presented for all regions. Negative differences mean that the forecasts with non-zero κ achieved a better score.



leadtimes because all three sets of forecasts have the same mean vectors. At shorter leadtimes, observations are slightly more likely to fall in the outer regions of the slightly sharper kurtosis-adjusted forecasts, with the κ_{lt} forecasts having dispersion indices slightly closer to one than either the baseline or κ_{ts} forecasts. For five-day-ahead forecasts, the κ_{ts} forecasts are, on average, less sharp than the baseline forecasts, and the resulting PIT dispersion index is slightly lower than that of the baseline forecasts. κ_{lt} forecasts at the same leadtime are generally sharper than the baseline, and tend to produce well-calibrated PIT histograms with dispersion indices close to one. Furthermore, comparing the PIT dispersion indices for the κ_{lt} forecasts to those of the MW- and AN-only forecasts in Figure 6.28b indicates that the κ_{lt} forecasts have substantially better marginal calibration than forecasts using either single source of information at shorter leadtimes.

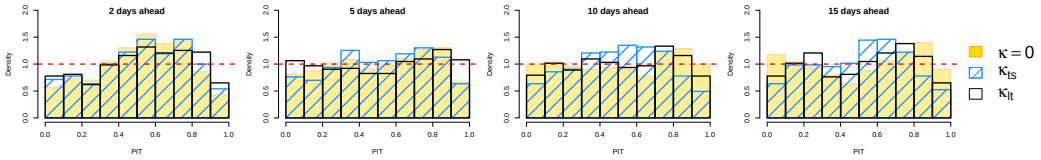
At longer leadtimes, when the excess kurtosis was typically estimated to be negative, the PIT histograms for the κ_{ts} forecasts are visibly peaked, suggesting that the resulting forecasts are typically overdispersive. The κ_{lt} forecasts also suffer from this underconfidence, but to a lesser degree than the κ_{ts} forecast.

6.2.4.6 Joint forecast calibration

The lack of marginal calibration of the κ_{ts} forecasts is reflected in the BDR histograms presented in Figure 6.29. At the shortest leadtimes, the difference in

Figure 6.27: PIT histograms showing the marginal calibration of forecasts postprocessed using Bayes linear adjusted discrepancies with different choices of κ at selected locations in the north and south of the UK, at a range of leadtimes. The dashed line indicates the ideal uniform distribution.

(a) PIT histograms for forecasts in Bristol



(b) PIT histograms for forecasts in Kirkcaldy

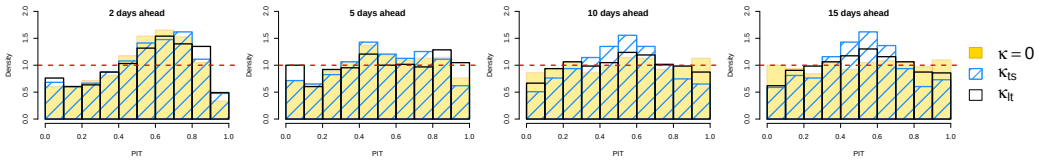
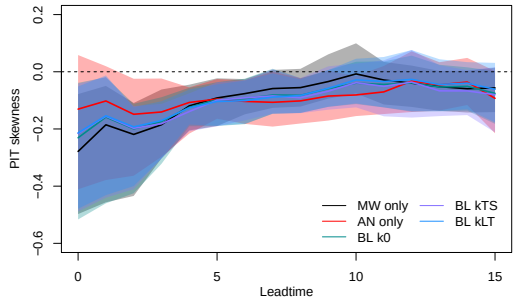
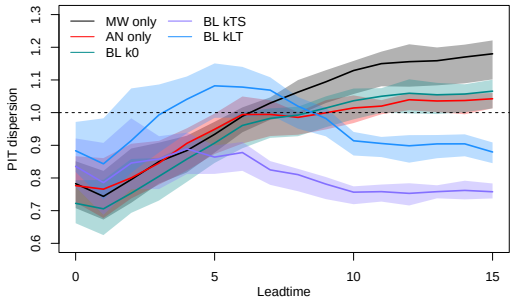


Figure 6.28: Characteristics of the PIT histograms at each leadtime for Bayes linear adjusted forecasts using realistic choices for κ. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.

(a) PIT skewness



(b) PIT dispersion



marginal calibration between the three sets of forecasts is relatively small, with all three having too many observations falling in the centre of the marginally overdispersive forecast distributions; however, this tendency persists at all leadtimes for the κ_{ts} forecasts, reflecting the tendency to marginal overdispersiveness already noted. Fewer observations fall in the rightmost bins of the BDR histograms for the κ_{lt} forecasts at the shortest leadtimes, but by leadtime five, there is a small spike in the leftmost bin, indicating that the forecasts are jointly slightly underdispersive at this leadtime, with a high proportion of observations falling in regions of extremely low forecast probability. At longer

leadtimes, where the baseline forecasts are themselves slightly underdispersive, the κ_{lt} forecasts are again slightly overdispersive, reflecting the pattern of marginal calibration.

All three methods produce highly ∪-shaped BOT histograms (Figure 6.30), indicating that the κ-adjusted forecasts still overstate the dependence between the temperatures in each region.

Figure 6.29: Band Depth Rank (BDR) histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes.

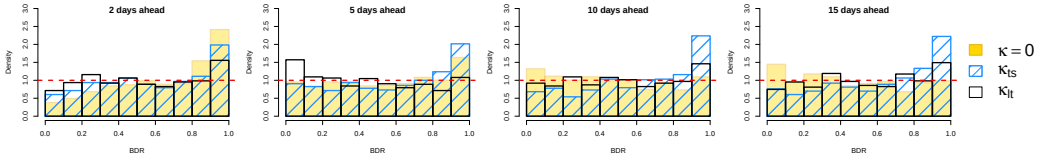
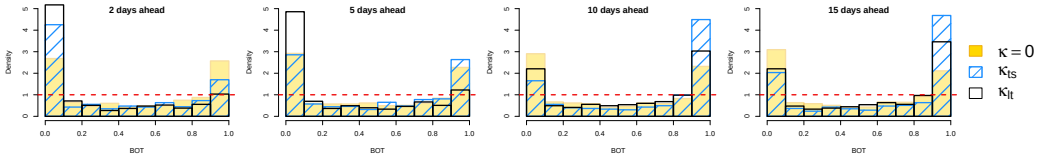


Figure 6.30: Box Ordinate Transform (BOT) histograms showing the joint calibration of postprocessed forecasts of surface temperatures at a range of leadtimes.



6.2.4.7 Summary

Adjusting the kurtosis parameter κ has a larger effect on the forecast variance – and so on the forecast skill – than adjusting any of the other prior specifications. In particular, setting $\kappa > 0$ at shorter leadtimes, when the forecast errors have been shown to have positive marginal excess kurtosis, tends to improve forecast skill not only against the baseline forecasts with $\kappa = 0$, but also against either the MW or AN-adjusted forecasts alone. However, setting $\kappa < 0$ at longer leadtimes does not generally improve on the forecast skill obtained by the baseline forecasts with $\kappa = 0$.

Estimating the kurtosis parameter independently for each forecast instance generally produces lower estimates of κ than estimating across all instances for a single leadtime. In particular, at the longest leadtimes – when forecasts

tend to be very close to the observed climatology – this method of estimating κ frequently produced large negative values, which massively inflates the variance of the postprocessed forecasts. This suggests that a sample size of 25 is too small to adequately estimate the true rate of outliers in the underlying population; more reliable estimates of κ were obtained when all available data were used.

6.2.5 Forecast skill of Bayes linear adjusted forecasts compared to NGR-postprocessed forecasts

In Section 4.2 forecasts were postprocessed using the Bayesian framework laid out in Section 2.2, with $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ estimated empirically from a single training set. The skill of those forecasts was compared to that achieved by forecasts postprocessed using nonhomogeneous Gaussian regression (NGR). The NGR-postprocessed forecasts were generally found to be more accurate (having lower MAE) and sharper, which results in better probabilistic calibration at shorter leadtimes, but may produce too-sharp forecasts at longer leadtimes. In this section, the skill of NGR forecasts trained using the MW training sets is compared to that of forecasts postprocessed using the Bayesian framework with $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ obtained by Bayes linear adjustment using the baseline specifications, first with $\kappa = 0$ and then with $\kappa = \kappa_{lt}$.

6.2.5.1 Forecast accuracy

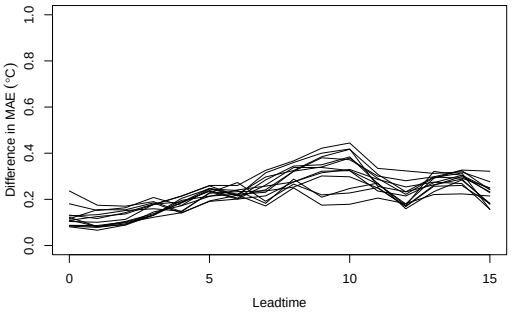
The distribution of the difference in mean absolute error (MAE) between the NGR and Bayes linear adjusted forecasts at each leadtime is shown in Figure 6.31; results are only shown for the baseline Bayes linear adjusted forecasts, because the mean forecast vector is the same regardless of the value of κ used in the adjustment.

Forecasts postprocessed using NGR have slightly lower MAE than forecasts using the Bayesian framework with Bayes linear adjusted $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$; however, as Figure 6.31 shows, the average improvement is very small compared to the spread of forecast errors at all leadtimes, which is very similar for both postprocessing methods. In fact, the Bayes linear adjusted forecasts are more

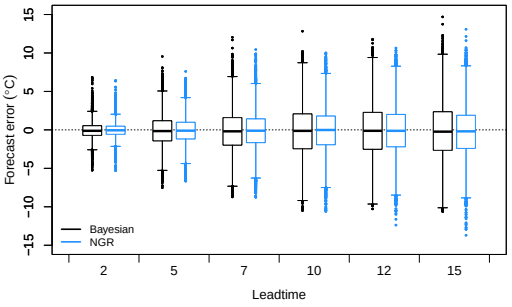
accurate than the NGR forecasts in around 40% of all cases at all leadtimes and in all regions, with colder temperatures remaining particularly poorly predicted by both methods.

Figure 6.31: Forecast errors for Bayes linear adjusted and NGR-postprocessed forecasts.

(a) Difference in mean absolute error in each of the 13 study regions. Positive differences indicate that the Bayes linear adjusted forecasts are less accurate.



(b) Distribution of forecast errors.



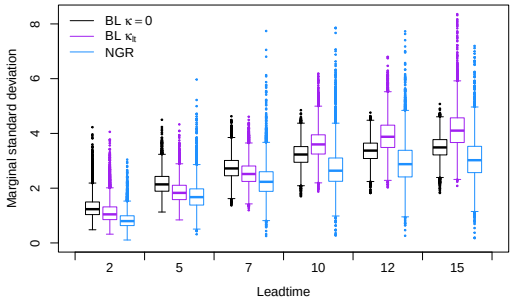
6.2.5.2 Forecast sharpness and calibration

Figure 6.32b shows the distribution of the marginal standard deviations of the postprocessed forecasts in Bristol for all three postprocessing approaches. At all leadtimes, NGR postprocessing is able to produce extremely sharp forecasts; at longer leadtimes, the NGR forecasts tend to span a wider range of variances than the Bayes linear adjusted forecasts, particularly those using the baseline specifications, suggesting that they may be better able to capture the full spread of variability in the forecasts. On average, the NGR forecasts are substantially sharper than the Bayes linear adjusted forecasts at all leadtimes, even when the kurtosis adjustment is applied at shorter leadtimes.

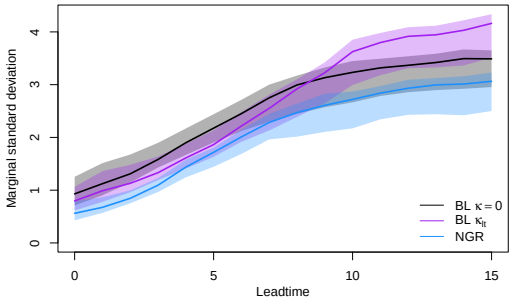
The NGR forecasts have slightly lower CRPS and logS overall than the Bayesian postprocessed forecasts, with mean CRPS around 0.1-0.2°C lower and mean logS between 0.05 and 0.25 lower, corresponding to between 5% and 30% more density placed at the verifying observation by the NGR forecasts. However, as Figure 6.33 shows, this improvement is not consistent across all forecast instances. In particular, while the median difference in log score is close to zero,

Figure 6.32: Distribution of marginal standard deviations of all 630 postprocessed forecasts in all regions at selected leadtimes, and spread of regional mean standard deviation, for forecasts postprocessed using NGR, Bayesian postprocessing with Bayes linear adjusted discrepancy with zero marginal kurtosis in the forecast errors, and Bayesian postprocessing with Bayes linear adjusted discrepancy with marginal kurtosis estimated separately for each leadtime. In instances where the proposed κ_{lt} was found to be invalid, the nearest valid κ was used. The lines in (b) indicate the mean value across the thirteen regions, while the shaded area shows the range of values.

(a) Marginal standard deviations



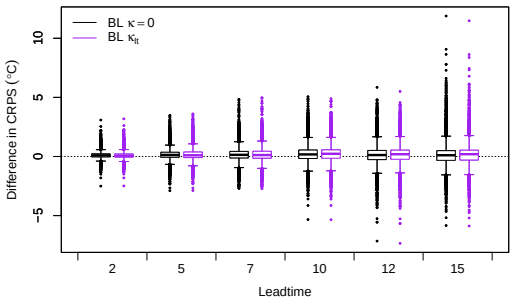
(b) Average marginal sharpness in each region



the distribution of logS differences in Figure 6.33b is quite asymmetric; the NGR forecasts sometimes receive much higher log scores than their Bayes linear adjusted counterparts (represented by large negative differences), indicating NGR forecasts that assign very low probability to the outcome that actually occurred.

Figure 6.33: Distributions of differences in CRPS and log score between NGR and Bayes linear adjusted forecasts. Positive scores mean that the Bayesian postprocessed forecasts were less skilful.

(a) Difference in CRPS



(b) Difference in logS

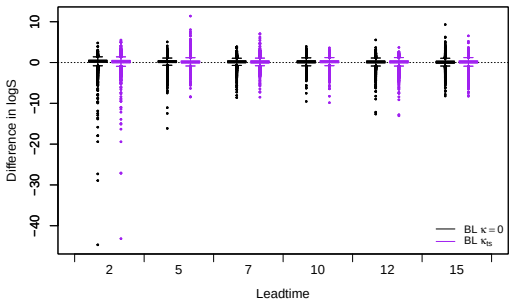


Figure 6.34 shows PIT histograms for forecasts in regions in the north and south of the UK for all three methods. At the shortest leadtimes, histograms for the NGR-postprocessed forecasts are more uniform than those using the

Bayesian framework to carry out postprocessing, with fewer observations falling between the 40th and 80th percentiles of the forecast distributions, and more falling in the extreme tails, indicating better forecast calibration. At longer leadtimes the NGR forecasts are slightly underdispersive, with PIT dispersion indices greater than one (Figure 6.13); the Bayes linear adjusted forecasts using $\kappa = \kappa_{lt}$ are generally overdispersive at these leadtimes, while the Bayes linear adjusted forecasts using $\kappa = 0$ – while still slightly underdispersive – tend to achieve better probabilistic calibration, with dispersion indices closer to one.

Figure 6.34: PIT histograms showing the marginal calibration of postprocessed forecasts of surface temperatures at selected locations in the north and south of the UK at a range of leadtimes.

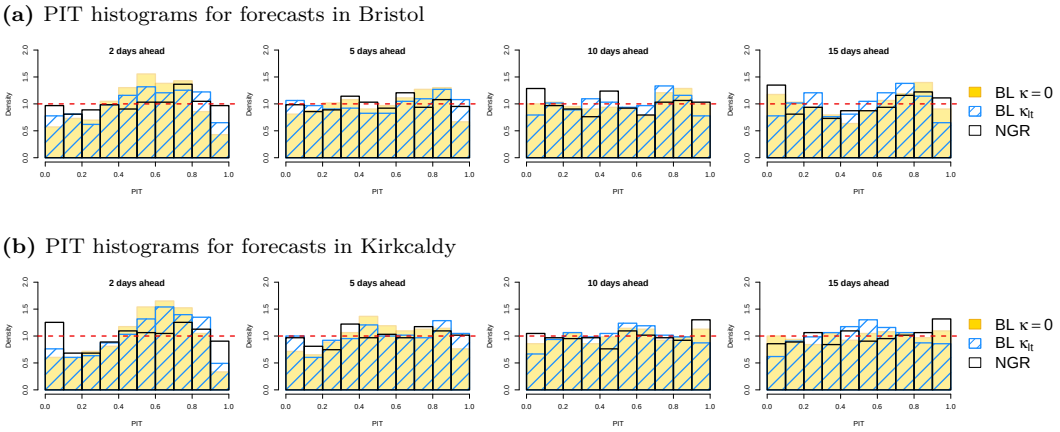
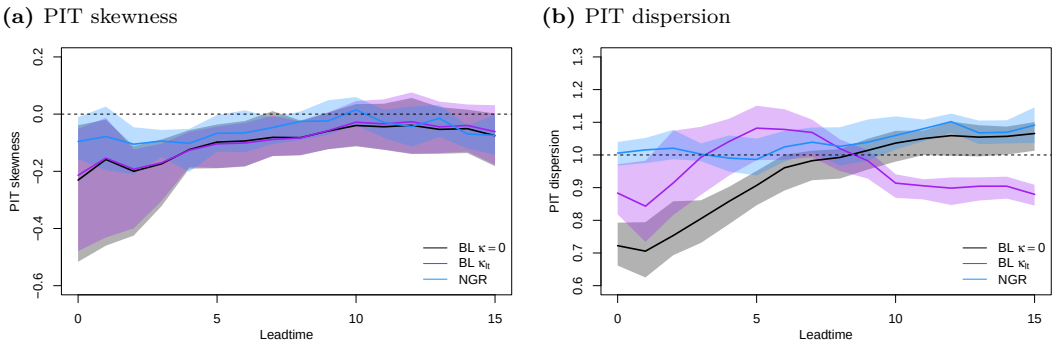


Figure 6.35: Characteristics of the PIT histograms at each leadtime for forecasts using alternative specifications of $\mathbb{V}[\boldsymbol{\eta}]$. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal values of zero for skewness and one for dispersion.



The joint calibration of the NGR-postprocessed forecasts is determined

by the choice of dependence structure used to combine the marginal forecasts into a multivariate distribution. The effect of this choice has already been considered in Section 4.2.3, so joint calibration is not discussed further here.

6.2.5.3 Summary

As in Section 4.2, the NGR-postprocessed forecasts are both sharper and more accurate than those obtained using the Bayesian postprocessing framework regardless of the approach used to estimate the discrepancy $\boldsymbol{\Delta}$, although the improvement in average accuracy is small compared to the magnitude of the errors that remain. As a result of these improvements, the NGR forecasts have better marginal probabilistic calibration than the Bayes linear adjusted forecasts at shorter leadtimes, reflected in slightly improved CRPS skill and more uniform PIT histograms. At longer leadtimes, while the NGR forecasts remain slightly more accurate on average, the baseline Bayes linear adjusted forecasts achieve comparable or slightly better marginal calibration overall. This suggests that, if forecast accuracy could be improved, the Bayes linear adjusted forecasts could be competitive with the NGR forecasts at longer leadtimes, although in order to compete at shorter leadtimes the forecasts would need to be sharper even than those incorporating an assumption of positive kurtosis.

6.3 Postprocessing with full assessment of uncertainty

In Section 6.2, the postprocessed forecasts were the posterior distributions of the weather quantity $\mathbf{Y}_0$ conditional on the observed ensemble members $\{\mathbf{Y}_{ij}\}$ and the expectation $\boldsymbol{\eta}$ and covariance $\boldsymbol{\Lambda}$ of the forecast discrepancy. However, by simply 'plugging in' the values of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$, this approach fails to account for any uncertainty about their true values. The posterior distribution of $\mathbf{Y}_0$ conditioned only on the observed ensemble members $\{\mathbf{Y}_{ij}\}$ would be obtained

by integrating the joint posterior distribution over $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$:

$$\pi\left(\mathbf{Y}_0 | \{\mathbf{Y}_{ij}\}\right) = \int \pi\left(\mathbf{Y}_0 | \{\mathbf{Y}_{ij}\}, \boldsymbol{\eta}, \boldsymbol{\Lambda}\right) \pi\left(\boldsymbol{\eta}, \boldsymbol{\Lambda}\right) d\boldsymbol{\eta}\, d\boldsymbol{\Lambda} \tag{6.21}$$

where $\pi(\cdot)$ denotes a probability density function, and $\pi\left(\boldsymbol{\eta}, \boldsymbol{\Lambda}\right)$ denotes the joint distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ after Bayes linear adjustment. Obtaining a closed form for this posterior distribution is not trivial, so simulations are used in this section to provide an approximation to the posterior distribution, in order to evaluate the impact on forecast skill of omitting this source of uncertainty in the original forecasts.

Due to the computing time needed to sample even a moderately-sized ensemble from the full posterior distribution of $\mathbf{Y}_0 | \{\mathbf{Y}_{ij}\}$, results are only presented here for forecasts postprocessed using the baseline adjustment considered in Section 6.2.1, using $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$, $\kappa = 0$, and $\nu = z + m$; and for the same adjustment with $\kappa = \kappa_{lt}$ presented in Section 6.2.4. This choice of prior for $\mathbb{V}[\boldsymbol{\eta}]$ results in an adjusted variance of $\boldsymbol{\eta}$ that depends on the adjusted expectation of $\boldsymbol{\Lambda}$, effectively replicating the effect of a joint distribution for $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$.

6.3.1 Simulation of uncertainty about $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$

Although the Bayes linear approach does not provide a joint posterior for $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ from which to sample the discrepancy $\boldsymbol{\Delta}$, it does provide for each forecast instance the adjusted expectation and variance of the mean discrepancy $\boldsymbol{\eta}$, along with the adjusted expectation of the covariance $\boldsymbol{\Lambda}$. A natural approach to simulating an approximation to the joint posterior distribution of $\boldsymbol{\Delta}$ is therefore to draw samples from a normal-inverse-Wishart distribution with parameters determined by these quantities and the scalar parameters ν, z and n used in the adjustment of $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}$.

Simulation is carried out for each forecast instance by first sampling an instance $\boldsymbol{\Lambda}_i$, say, of $\boldsymbol{\Lambda}$ from the inverse-Wishart distribution with expectation $\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ and $(\nu + n - 1)$ degrees of freedom. This parametrisation ensures that,

in the baseline forecasts, the scale matrix $\boldsymbol{\Psi}$ is

$$\boldsymbol{\Psi} = (\nu - m - 1 + n - 1)\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}] = (n-1)\boldsymbol{S_\delta} + (\nu - m - 1)\mathbb{E}[\boldsymbol{\Lambda}]\,, \qquad (6.22)$$

consistent with the natural conjugate specification that the baseline Bayes linear adjustment was shown to approximate in (5.200) in Section 5.4.2.

An instance $\boldsymbol{\eta}_i$ of $\boldsymbol{\eta}$ is then sampled from the normal distribution with mean vector $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\eta}]$ and covariance matrix $\boldsymbol{\Lambda}_i/(z+n)$; these realisations are then used in (6.8) to obtain the corresponding mean vector $\boldsymbol{\tau}_i$ and covariance matrix $\mathbf{S}_i$ of the forecast distribution. Finally, a single realisation $\mathbf{Y}_0^{(i)}$ is sampled from this conditional distribution. This process is repeated 100 times for each forecast instance, providing a 100-member sample from the posterior distribution (6.21). The verification tools described in Chapter 3 as suitable for the evaluation of ensemble forecasts are used to obtain measures of forecast accuracy and calibration that can be compared directly to those already presented in Sections 6.2.1 and 6.2.4 for the normally distributed 'plug-in' forecasts.

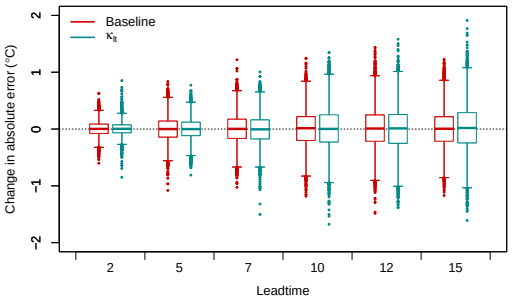
6.3.1.1 Forecast accuracy and sharpness

Figure 6.36 shows the distribution of differences in absolute error and CRPS between the simulated forecasts and their 'plug-in' equivalents. At all leadtimes the change in MAE due to simulation is very close to zero, although changes of up to 1°C for individual forecasts are fairly common. The change in MAE is reflected almost exactly in the distribution of the CRPS. The log score is not computed for these nonparametric ensemble forecasts, because this would require some assumptions regarding the underlying parametric form of the distribution, as discussed in Section 3.2.2.

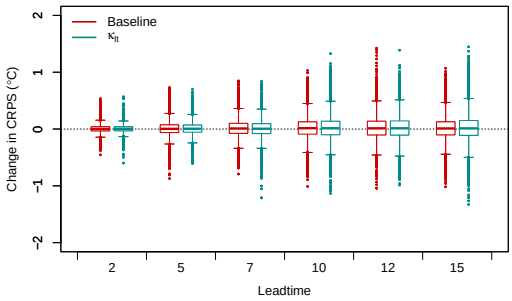
The change in forecast sharpness resulting from the additional source of uncertainty is shown in Figure 6.37, expressed as the ratio of the standard deviations of the simulated forecasts to the standard deviations of their 'plug-in' equivalents. The distribution of the proportional changes in forecast sharpness is constant at all leadtimes, and forecasts with simulated uncertainty are as

Figure 6.36: Distribution of changes in forecast accuracy when $\mathbf{Y}_0$ is simulated, compared to forecast accuracy of corresponding 'plug-in' estimate. Changes are shown for all regions.
Negative values indicate greater skill (lower MAE/CRPSS) from the forecasts with simulated uncertainty.

(a) Change in MAE



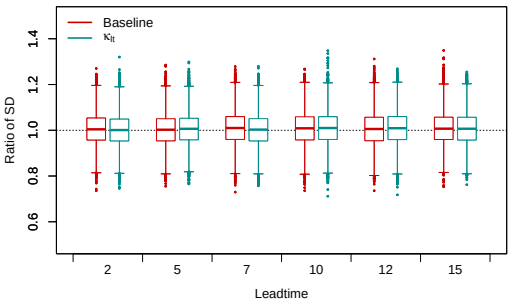
(b) Change in CRPS



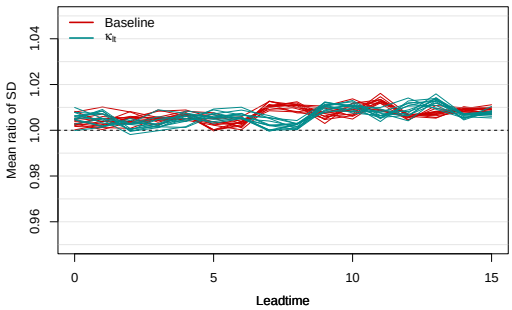
likely to be sharper than their plug-in counterparts as they are less sharp. This suggests that the random variations introduced by simulating from $\mathbf{\Lambda}$ are responsible for the majority of the change in sharpness for individual forecasts, masking the smaller increase that might be expected as a result of sampling the uncertainty about $\boldsymbol{\eta}$. Figure 6.37b shows that such an increase does indeed occur, with the marginal standard deviations of simulated forecasts being around 1% larger, on average, than those of forecast distributions not incorporating this additional source of uncertainty.

Figure 6.37: Distribution of changes in forecast sharpness when $\mathbf{Y}_0$ is simulated, compared to forecast accuracy of corresponding 'plug-in' estimate.
Values less than one indicate that the forecasts with simulated uncertainty are sharper (having a lower standard deviation) than the forecasts using 'plug-in' adjustment.

(a) Ratio of marginal standard deviations for simulated forecasts to corresponding 'plug-in' forecasts, for all regions.



(b) Mean ratio of simulated forecast standard deviations to 'plug-in' standard deviations for each region.



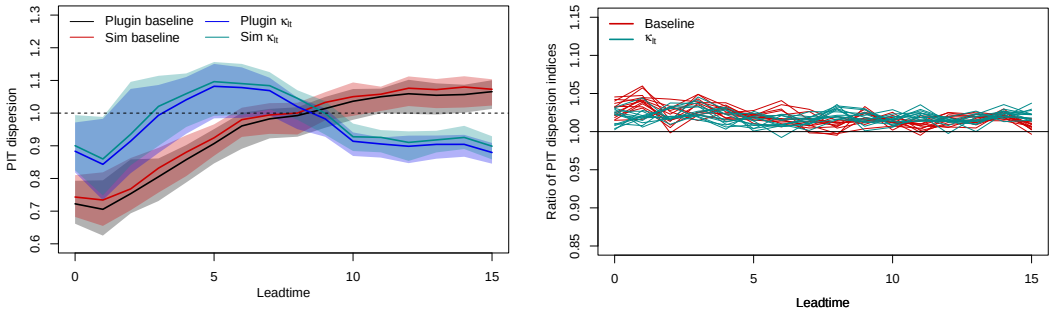
6.3.1.2 Forecast calibration

Figure 6.38 shows the changes in marginal calibration resulting from the additional uncertainty about $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$. The PITs of the plug-in forecasts are directly analogous to the verification ranks (VR) of the simulated ensemble forecasts. Because the changes to the forecast error are essentially symmetric, the skewness of the PIT/VR (not shown) is almost unchanged by the addition of simulated uncertainty, while the PIT/VR dispersion indices are inflated by around 2% on average, reflecting the mean increase in marginal standard deviation.

Figure 6.38: Dispersion of the verification rank (VR) and PIT histograms for forecasts with and without simulated parameter uncertainty, and with zero and non-zero kurtosis included in the Bayes linear adjustment of $\boldsymbol{\Delta}$; and ratio of PIT/VR dispersion with and without simulated uncertainty for each choice of κ.

(a) PIT/VR dispersion. The lines indicate the mean value across the thirteen regions, while the shaded area shows the range of values. The dashed horizontal lines indicate the ideal value of one.

(b) Ratio of PIT/VR dispersion of forecasts with simulated uncertainty to 'plug-in' forecasts for each choice of κ.



This blanket increase in marginal variance means that the observations will generally fall slightly closer to the centre of the joint forecast distributions, and manifests as a slight decrease in the skewness of the BDR histograms at all leadtimes (Figure 6.39). However, the BOT histograms in Figure 6.40 show that the number of observations falling in regions of low probability have increased. This is because the BOT is based on an assumption that the joint distribution is close to multivariate normal (3.14); the simulated forecasts are expected to have heavier tails than those of a normal distribution, and as such, a higher proportion of points should be expected to be classified as outliers by the BOT

for the simulated forecasts than for the plug-in forecasts. The BOT-BDR grids in Figure 6.41 reflect this tendency, but are otherwise similar to those of the plug-in forecasts, suggesting that the calibration of the dependence structure is unaffected by the simulation.

Figure 6.39: Band Depth Rank (BDR) histograms showing the joint calibration of postprocessed 'baseline' forecasts of surface temperatures at selected leadtimes, with and without simulated uncertainty. A similar pattern of changes is observed in the histograms for forecasts with $\kappa = \kappa_{lt}$.

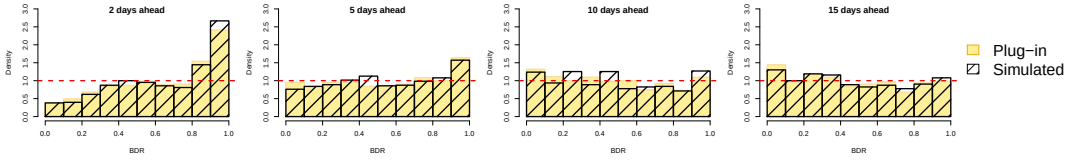


Figure 6.40: Box Ordinate Transform (BOT) histograms showing the joint calibration of postprocessed forecasts of surface temperatures at selected leadtimes, with and without simulated uncertainty. A similar pattern of changes is observed in the histograms for forecasts with $\kappa = \kappa_{lt}$.

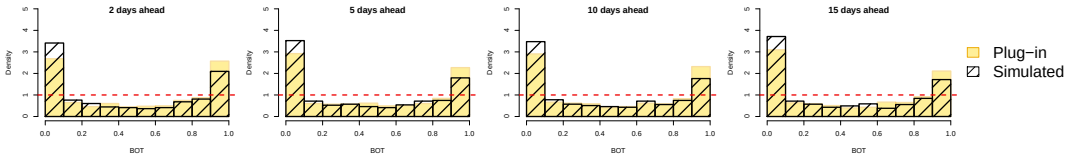
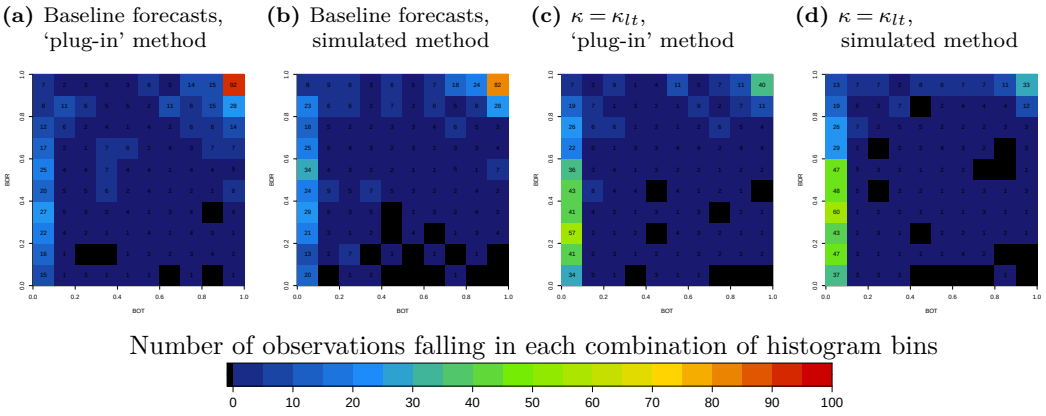


Figure 6.41: Gridplots summarising the joint distribution of the BOT and BDR histogram counts at leadtime 2.



6.3.1.3 Summary

Including a fuller representation of the uncertainty about the true values of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ results in slightly wider forecast intervals than when the mean and

variance of the discrepancy are treated as known quantities, although the effect on calibration is small in this application. Failing to incorporate this source of uncertainty is unlikely to qualitatively alter any conclusions drawn or decisions made, while the cost of simulating the additional values of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ may be rather large.

However, it is possible that simulating the estimation uncertainty will have a larger effect on forecast calibration in other settings. The effect of simulating – or failing to simulate – the additional uncertainty can be approximately quantified by using the law of total variance to decompose the uncertainty about the forecast discrepancy $\boldsymbol{\Delta}$ as

$$\mathbb{V}[\boldsymbol{\Delta}] = \mathbb{E}\Big[\mathbb{V}[\boldsymbol{\Delta}|\boldsymbol{\delta}]\Big] + \mathbb{V}\Big[\mathbb{E}[\boldsymbol{\Delta}|\boldsymbol{\delta}]\Big] \qquad = \mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}] + \mathbb{V}_{\mathbf{X}}[\boldsymbol{\eta}] \tag{6.23}$$

where $\boldsymbol{\delta}$ is the set of observed forecast errors. The adjusted expectation of $\boldsymbol{\Lambda}$ is already incorporated in the 'plug-in' postprocessed forecasts; however, the uncertainty about $\boldsymbol{\eta}$ is omitted. The ratio of the missing variance $\mathbb{V}_{\mathbf{X}}[\boldsymbol{\eta}]$ to the adjusted expectation $\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$ is approximately the amount by which the plug-in variance will, on average, be inflated when simulation is used to capture this missing source of variability, although this approximation still fails to take into account the fact that the simulated forecasts will also have heavier tails than their multivariate-normal 'plug-in' equivalents. In the examples presented here, where the prior is $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$, the adjusted variance of $\boldsymbol{\eta}$ is $\mathbb{V}_{\mathbf{X}}[\boldsymbol{\eta}] = (z+n)^{-1}\mathbb{E}_{\boldsymbol{S}_{\boldsymbol{\delta}}}[\boldsymbol{\Lambda}]$ using (5.104), with $z = n = 25$; so the additional uncertainty due to incorporating $\mathbb{V}_{\mathbf{X}}[\boldsymbol{\eta}]$ into the postprocessing is expected to be around 2%, as seen in Figure 6.37b; this change in the marginal variances corresponds to a change in the marginal standard deviations of around 1%. In any situation where the ratio is small, as here, the user may judge that the additional uncertainty is unlikely to have a meaningful impact in the final analysis.

6.4 Summary & discussion

The parameters used in Bayes linear adjustment can be tailored to reflect the user's level of confidence in each aspect of their prior judgements. The distribution of weights assigned to the prior and observed mean when adjusting the expected value of $\boldsymbol{\eta}$ is controlled by the prior variance $\mathbb{V}[\boldsymbol{\eta}]$; setting $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}_{\boldsymbol{S_\delta}}[\boldsymbol{\Lambda}]$ and decreasing z will shift weight directly from the prior to the observations when adjusting the expected value of $\boldsymbol{\eta}$, while setting $\mathbb{V}[\boldsymbol{\eta}] = z^{-1}\mathbb{E}[\boldsymbol{\Lambda}]$ allows the weighting to depend on the relative precisions of the prior and sample estimates of $\boldsymbol{\Lambda}$. Similarly, reducing the notional prior sample size ν will shift weight from the prior to sample variance when adjusting the expectation of $\boldsymbol{\Lambda}$. By changing these two specifications, the user can assign different levels of confidence to the expectation and variance of the two sources of information: so, if one training set was known to result in very accurate but over-confident forecasts, and the other in less accurate but less confident forecasts, a high value of z and low value of ν could be used, to allow the adjusted forecast to be dominated by the mean vector of the first and the variance-covariance matrix of the second. This flexibility is supplemented by the kurtosis parameter κ, which provides a mechanism to account for particularly high or low expected numbers of extreme values in the data, scaling the covariance matrix to accommodate the assumed frequency of contamination by or absence of outliers.

In the temperature data analysed here, the MW-adjusted forecasts used to specify the prior expectation and variance of the forecast error $\boldsymbol{\Delta}$ are more accurate than the 'observed' AN-adjusted forecasts in roughly half of all cases, and sharper in roughly half of all cases at leadtimes up to seven days, and around 60% of all cases thereafter. Both methods have fairly similar forecast skill, although the MW-adjusted forecasts are consistently less well calibrated at longer leadtimes as a result of the increased sharpness. Because of this, Bayes linear adjustment typically results in roughly equal numbers of forecasts with improved and degraded forecast skill under most metrics, with the net

effect being a very small change in overall skill. However, improvements in average forecast skill are still possible: in particular, combining the two sources of forecast information results in an improved estimate of the forecast spread at longer leadtimes, when either single source would underestimate the true uncertainty. Changes to $\mathbb{V}[\boldsymbol{\eta}]$ and ν, which reflect confidence in the prior expectations of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ respectively, resulted in very minor changes to forecast skill.

The largest change in forecast skill was obtained by setting $\kappa \neq 0$, in order to accommodate the marginal sample kurtosis at each leadtime. At leadtimes of eight days or shorter, the kurtosis was estimated to be greater than zero, with the effect that the Bayes linear adjusted expectation of the discrepancy variance $\boldsymbol{\Lambda}$ was sharper than when kurtosis was not accounted for, significantly improving forecast calibration. At longer leadtimes, using negative kurtosis – thereby generally increasing the adjusted variance – produced slightly overdispersive forecasts, which were less well calibrated. These results lend support to the findings of Denholm-Price (2003) and Gebetsberger et al. (2018) that better-calibrated forecasts might be obtained by using a postprocessing method that is able to accommodate heavier tails in the forecast distribution at shorter leadtimes.

Forecasts obtained using the Bayesian postprocessing framework have slightly worse accuracy than those obtained by the NGR postprocessing method described in Section 2.1.4.1, which use only a single source of training data (in this case, the MW training set used to specify the priors for the Bayes linear adjustment). The NGR forecasts have good marginal calibration at all leadtimes, although the Bayes linear adjusted forecasts using $\kappa = 0$ have slightly more uniform PIT histograms at the longest leadtimes, indicating slightly improved marginal calibration despite the larger forecast errors. This indicates that better calibration may be achieved at longer leadtimes by combining the two sources of information; however, further work is required to improve the accuracy of the forecasts.

Chapter 7

Discussion

7.1 Summary

The work in this thesis was motivated by the problem of how weather forecasts from multi-model ensemble (MME) prediction systems might be combined into a single forecast in a principled way. Established approaches to the problem tend to ignore the relationships between the component ensembles, treating them as if they were mutually independent and failing to account for the fact that some or all models may, in fact, share common biases and calibration errors.

The first contribution of this thesis, presented in Section 2.2, is the modification of the framework proposed by Chandler (2013), which combines climate projections from multiple models into a single predictive density, to accommodate the slightly different problem of weather forecasting: this requires a model that issues predictions of the actual weather quantities, rather than of the statistical properties of the future weather. The mathematical representation of the problem is developed from a graphical model of the structure of the MME weather forecasting system, which represents the relationships between the available ensemble forecasts and the unobserved quantities of interest, and so allows inference to be carried out on the verifying observation. Unlike the competing postprocessing methods reviewed in Section 2.1.4, the Bayesian framework is able to incorporate the user's prior beliefs about the quantity

of interest, offering the possibility of generating sequentially postprocessed forecasts, although the 'posterior-to-prior' sequential postprocessing approach tested in Section 4.1 was unsuccessful: plans for further work in this area are discussed in Section 7.2.2. Several alternative approaches to specifying an informative prior, and to selecting an appropriate training set for the estimation of the necessary correction, were also considered.

A comparison of Bayesian postprocessed forecasts with the raw 'superensemble' forecasts and forecasts postprocessed using nonhomogeneous Gaussian regression (NGR) showed that, while the Bayesian method was able to outperform the superensemble forecasts at shorter leadtimes, the NGR forecasts were both sharper and more accurate overall, although this improved skill comes at considerable computational cost due to the numerical optimisation required to estimate the NGR parameters. A particular problem for the Bayesian forecasts is marginal calibration, with forecasts at shorter leadtimes being consistently overdispersive, and forecasts at longer leadtimes being consistently underdispersive. This may be partially attributable to the method of estimating $\mathbf{\Lambda}$, the variance of the forecast discrepancy, which contributes most of the uncertainty in the posterior covariance matrix $\mathbf{S}$, and which increases with leadtime. This suggests that improving the estimates of $\boldsymbol{\eta}$ and $\mathbf{\Lambda}$ may be the most straightforward way to improve the skill of forecasts postprocessed using the Bayesian method.

The second contribution of the thesis concerns a new approach to selecting an appropriate training set for the estimation of the correction. Many studies simply use a 'moving window' of training cases to correct the forecast of interest; however, especially at longer leadtimes, this risks including training cases that are not particularly relevant to the current forecast instance. The use of analogues to the current instance – forecasts that are, in some sense, similar to the current forecast – is expected to produce more relevant training cases, and so to produce a better estimate of the forecast discrepancy. The proposed method uses principal component analysis (PCA) as a dimension reduction

technique to obtain low-dimensional summaries of the forecast mean sea level pressure fields, which are known to be well predicted by the numerical weather prediction models, and are commonly used to characterise prevailing weather conditions and to classify weather forecasts into climatological weather regimes.

The skill of forecasts postprocessed using analogues selected on the basis of these low-dimensional summaries was found to be comparable to or better than that of forecasts postprocessed using analogues that were similar in terms of their predicted temperatures, and that of forecasts postprocessed using a moving-window approach. Greater improvements may be achievable when postprocessing higher-dimensional forecasts, or when a longer archive of high-quality candidate forecasts is available.

The third and final contribution of the thesis is a coherent multivariate framework for Bayes linear adjustment, proposed in Section 5.3. This extends the second-order exchangeability assumptions used in the Bayes linear variance adjustment described in Goldstein and Wooff (2007) to handle Bayes linear adjustment of the both the expectations and covariance matrices of multivariate random quantities. The resulting method is parametrised by a prior covariance matrix $\mathbf{V}_R$ and scalar parameters ν and κ reflecting, respectively, the user's confidence in their estimate of $\mathbf{V}_R$ and the degree to which the data are believed to have a greater or lesser propensity to outliers than a normal distribution. These parameters are shown to be directly analogous to the parameters used in the scalar case, and the multivariate adjustment is shown to closely approximate the posterior distribution that would be obtained using Bayesian inference with a conjugate normal-inverse Wishart prior and normally distributed data.

Chapter 6 presents a detailed investigation of the skill of forecasts again postprocessed using the Bayesian postprocessing method, but now with $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ obtained using Bayes linear adjustment, rather than estimated directly from the training data. The impact on forecast skill of changing the parameter specifications was generally small – due, in part, to the similar skill levels of the two sets of forecasts being combined – with the largest change in forecast

skill obtained by varying the scalar parameter κ to reflect an assumption of non-zero excess marginal kurtosis. In particular, forecasts postprocessed using the Bayesian method with Bayes linear adjusted estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ using realistic estimates of the marginal excess kurtosis for each leadtime were found to be better calibrated than NGR forecasts at longer leadtimes. This suggests that this method may be competitive with NGR postprocessing over slightly longer-term forecasts, perhaps in the medium to subseasonal range from 15 days' leadtime onwards.

7.2 Future work

A number of areas of further interest have been highlighted throughout this thesis. These are summarised below.

7.2.1 Improved verification metrics for parametric probabilistic forecasts

The review of verification methods in Chapter 3 introduced two potentially useful innovations. The first of these is a semiparametric equivalent to the band depth rank, which is used to construct histograms that can be used to diagnose various types of miscalibration in a collection of joint forecast distributions. The original approach, proposed by Thorarinsdottir et al. (2016), was designed to evaluate the joint calibration of ensemble forecasts, and depends on the ordering of a quantity known as the 'prerank' (3.16), which depends in turn on the marginal ranks of the verifying observation and the ensemble members.

Where the forecast is issued as a predictive density with a known parametric form, the marginal ranks can be replaced by the PITs when computing the prerank function; however, the ordering of the prerank of the observation within the ensemble is still estimated by drawing a synthetic ensemble from the multivariate predictive density. In order to obtain a fully parametric equivalent to the band depth rank, it will be necessary to derive the distribution of the parametric pre-rank function (3.17).

The second innovation suggested in this chapter is the use of grid plots of

the joint distribution of the BOT and BDR in order to more precisely diagnose calibration issues in the joint forecast distribution. A simple extension to this suggestion would be to consider whether plots of the joint distribution of other pairs of depth scores (for example, the minimum spanning tree depth or average marginal rank, both of which are used by Wilks (2017) to evaluate joint calibration) might be more informative about other types of calibration misspecification.

The grid plots showing different types of misspecification presented in Section 3.3.2.3 are based on synthetic data with various error characteristics; in order to fully understand how to interpret the patterns that might arise in these plots, it will be necessary to derive the expected joint distribution of the BOT and BDR for various types of forecast misspecification. For forecasts issued as a multivariate normal predictive density, this will likely be much simplified if the fully parametric band depth rank function is first derived.

7.2.2 Sequential forecast postprocessing

In Section 4.1, the sequentially postprocessed forecasts were the least accurate and least well calibrated of the methods compared, having the highest MAE and being far too sharp. This was an unexpected result: the noninformative prior forecasts are reasonably well calibrated, so using the n-day-ahead noninformative posterior as the prior for the $(n-1)$-day-ahead forecasts was expected to produce more skilful forecasts. Experiments with sequential postprocessing starting from a shorter leadtime – or only using one sequential step – produced similar results, suggesting that the problem is most likely to do with the sequential postprocessing model, rather than with the quality of the prior distribution.

One potential improvement to the sequential postprocessing approach tested here would be to explicitly account for correlation between successive forecasts, rather than treating the prior and MME forecasts as if they were independent. This may be able to counter the increasing sharpness seen in the sequentially postprocessed forecasts with decreasing leadtime, and so improve

calibration.

7.2.3 Testing weather regime analogues on a longer dataset

One of the contributions of this thesis was the method of selecting analogues to the forecast of interest on the basis of principal component scores reflecting the dominant modes of spatial variability in the forecast mean sea level pressure fields, which are known to be closely related to the prevailing weather conditions. In Section 4.3, forecasts postprocessed using analogue training sets were found to perform as well as or better than those postprocessed using recent training cases, although the improvements in forecast skill were relatively modest.

It is known that a critical requirement of any analogue selection method is the availability of a sufficiently long archive of candidate forecasts to provide a large enough number of high-quality analogues (Hu et al., 2020). These candidate forecasts should be produced by the same model – or configuration of models – that issued the forecast of interest; ideally, the archive would consist of reforecasts, issued by the current operational model(s) and initialised using historical data. The analogues in the application presented in Section 4.3 were drawn from a relatively small archive of 565 candidates, consisting of six full winters of ninety days each (ie. omitting from the full seven-year dataset the winter in which the forecast of interest was issued) plus the 25-day moving window immediately preceding the forecast issue date. Not only does the short length of the archive mean that there may be a dearth of high-quality analogues – particularly when predicting extreme weather events – but, because the forecasts were issued operationally, it is known that the model configuration changed several times throughout the study period (ECMWF, 2021b,d,c).

The results presented in Section 4.3 demonstrate that forecasts post-processed using the proposed weather regime analogues are able to achieve skill comparable to those postprocessed using analogues selected by the usual method, and that both analogue methods were able to improve on the skill of forecasts using moving-window training cases, despite this disadvantage.

However, it is likely that the analogue-postprocessed forecasts will be more successful when analogues are selected from an archive of reforecasts produced by the same model configuration as the forecast of interest. No such archive is available for the TIGGE forecasts, but an archive of weather forecasts including reforecasts at subseasonal to seasonal time scales is available through the S2S project (Vitart et al., 2017). This dataset provides forecasts at leadtimes of up to 65 days, along with up to twenty years of reforecasts accompanying each forecast instance, from eleven forecasting centres: so it should provide an ideal test case to investigate the potential improvements in forecast skill. One caveat is that many of the models are only run two or three times each week, with different forecasting centres initialising the model on different days; as a result, constructing a multi-model ensemble from the S2S dataset is not straightforward. Nonetheless, the archive of reforecasts offers an opportunity to compare the skill of forecasts postprocessed using weather regime analogues, traditional analogues and a moving window training set, albeit only for a single-model ensemble.

7.2.4 Improving the treatment of nonzero kurtosis in Bayes linear adjustment

A key element of the multivariate Bayes linear variance adjustment described in Section 5.3 is the facility to incorporate a judgement of non-zero marginal excess kurtosis. However, care must be taken when specifying the kurtosis parameter κ to ensure that the resulting adjusted covariance matrix is indeed a valid covariance matrix. The positive semidefiniteness of the adjusted expectation of $\mathcal{M}(\mathbf{V})$ given in (5.177) depends not only on κ but on the prior and observed covariance matrices, $\mathbf{V}_R$ and $\mathbf{S}$: as noted in Section 5.3.7.4, if $\kappa > 0$ and $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) \gg m$, or if $\kappa < 0$ and $\mathrm{tr}\left(\mathbf{V}_R^{-1}\mathbf{S}\right) < m$, the adjusted expectation of $\mathcal{M}(\mathbf{V})$ is no longer guaranteed to be a positive semidefinite matrix. A range of values of κ has been identified for which the adjusted expectation of $\mathcal{M}(\mathbf{V})$ is guaranteed to be positive semidefinite; however, the bounds of this range are fuzzy, in the sense that, depending on the particular matrices $\mathbf{V}_R$ and $\mathbf{S}$, some

'invalid' values of κ that are close to the threshold may still result in a valid covariance matrix. Further work is required to define the exact range of κ that will produce a valid adjusted covariance matrix.

7.3 Conclusions

This thesis considered the problem of improving forecasts of surface temperatures over the UK: temperature forecasts were chosen because they are fairly well represented by Gaussian distributions, meaning that analytical expressions could be found for all of the required quantities; and the geographical region was limited to a relatively small number of grid cells in order to ensure that all of the required covariance matrices could be estimated empirically. In principle, there is no reason that this approach could not be extended to postprocessing over a much larger spatial domain, although this would require a different approach to estimation of the required covariance matrices: the size of the matrices that can be estimated using the empirical approach suggested here is limited not only by the size of the training set, but by the size of the forecast ensembles. With increasing spatial domain it is likely that estimating the forecast discrepancy using weather regime analogues, rather than a moving window or direct analogues, will produce more skilful forecasts.

Extending the Bayesian postprocessing framework to variables other than surface temperatures may prove more complicated. The framework could be used without significant modifications to the current implementation for any variable that could reasonably be considered to have a multivariate-normal distribution, perhaps after a suitable transformation. It would also be relatively straightforward to implement the method using other parametric forms, although the parameters of the resulting posterior distributions would need to be found using computational methods. A similar approach has been applied to the forecasting of fish stocks (Spence et al., 2018), based on the climate modelling framework originally proposed by Chandler (2013); however, this increased computation time may render this approach impractical if postpro-

cessed forecasts need to be issued very rapidly after the raw forecasts are issued. Furthermore, it is not immediately clear how the framework should be adapted to handle the forecasting of precipitation, a variable often represented as comprising two components: a binary variable indicating the presence or absence of any precipitation, and a continuous variable representing the amount of precipitation to be expected. This remains a potential limitation of the method, albeit one shared by most postprocessing techniques.

Despite this limitation, the Bayesian MME postprocessing framework proposed here is flexible enough to be of potential use in applications other than the mid-term weather forecasting presented here. As already noted, the approach may be useful in longer-term forecasting, for example at the subseasonal-to-seasonal scale, where the focus is less on prediction of the weather conditions on a particular day and more on statistical properties of the weather over a short period of time. It is also likely that estimation of the required forecast adjustments using weather regime analogues, which are derived from pressure fields that may reflect longer-term atmospheric trends, may produce more skilful forecasts at these leadtimes than either moving window training cases, which may be several months removed from the forecast verification dates, or direct analogues, which are typically based on surface weather quantities that tend towards climatology at these time scales. However, the postprocessing method is also widely applicable in fields outside of meteorology and the environmental sciences: it may be of use in any situation where several competing probabilistic or ensemble forecasts are available, along with an archive of previous verifying observations from which the required correction may be estimated, such as in financial modelling or astrostatistics.

Likewise, the multivariate Bayes linear adjustment proposed here may be adopted anywhere that Bayes linear methods are already used to understand and predict the behaviour of complex systems. The suggested joint adjustment requires no input from the user beyond that already required for the semi-

adjusted variance assessments proposed by Goldstein and Wooff (2007), so could be implemented immediately anywhere that the standard Bayes linear method could be: existing applications are as diverse as assessing medical risks (Gosling et al., 2013), estimating crop yields (Makowski, 2017), and parametrising galaxy formation (Bower et al., 2010), suggesting that, like the Bayesian MME postprocessing framework, the multivariate Bayes linear adjustment may be suitable for a wide range of applications.

Appendix A

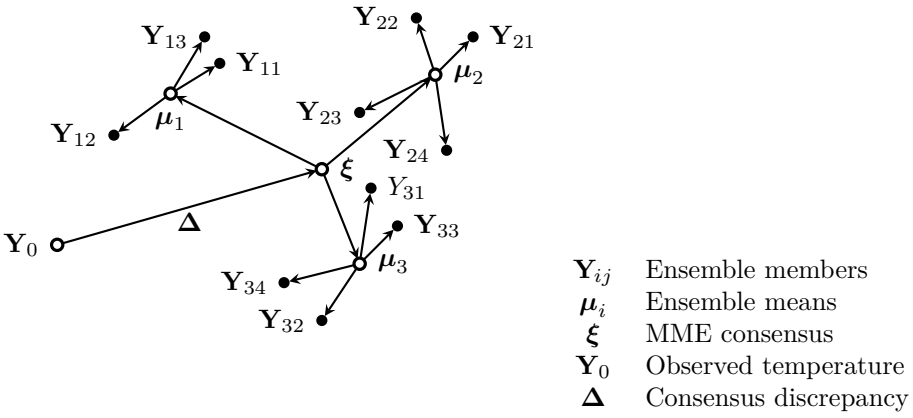
Derivation of the posterior density of $\mathbf{Y}_0$

A.1 The generic posterior

The derivation of the posterior density given here follows similar arguments to those in the supplement to Chandler (2013), and is expanded from the supplement to Barnes et al. (2019). The derivation is based on the graphical representation of the MME structure presented in Figure 2.6 and reproduced in Figure A.1. This directed acyclic graph represents the system of ensemble forecasts described in Section 2.2.1; as noted in the main thesis, the arrows linking the various nodes can be taken to represent conditional probability distributions, with each 'child' node being modelled by a distribution conditioned on its 'parent'. The structure encodes assumptions of conditional independence; that is to say, if there is no path connecting nodes A and B that does not pass through C en route, then A and B are conditionally independent given C. So, for example, there is no path from $\mathbf{Y}_{11}$ to $\mathbf{Y}_{12}$ that does not pass through $\boldsymbol{\mu}_1$, which means that those forecasts are independent of one another given the value of $\boldsymbol{\mu}_1$. In other words, if $\boldsymbol{\mu}_1$ is known then $\mathbf{Y}_{11}$ cannot give us any new information about the position of $\mathbf{Y}_{12}$. It is possible to derive a Bayesian posterior density to infer the true value of the weather quantity of interest, $\mathbf{Y}_0$ from the available information, using the conditional independence structure

implied by Figure A.1 as a framework.

Figure A.1: Schematic representation of the relationships between the elements of the multi-model ensemble forecasting system, originally introduced in Figure 2.6. Quantities known at the time of issuing the forecast are shown as filled nodes, with unknown quantities represented by open nodes.



The necessary calculations are simplified by shifting the distribution of $\mathbf{Y}_0$ by the expected value of the discrepancy $\boldsymbol{\Delta}$, and performing inference over $\widetilde{\mathbf{Y}}_0 = \mathbf{Y}_0 + \boldsymbol{\eta}$, with $\widetilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta} - \boldsymbol{\eta}$; the posterior forecast can then be recovered by subtracting the added discrepancy from the posterior mean of $\mathbf{Y}_0$. This device ensures that the framework and derivation are conceptually identical to those derived in the supplement to Chandler (2013).

Generically, the posterior density forecast is the product of a prior distribution describing beliefs about $\mathbf{Y}_0$ before any forecasts are made, and a likelihood function describing the evidence obtained from the forecasts. Let $\pi(\cdot)$ denote any probability density function (pdf), and $\pi(a|b,c)$ the density of a conditional upon the values of b and c. Let lower-case $\overline{\mathbf{y}}_i$ and $\mathbf{y}_0$ indicate realisations of the corresponding random variables $\overline{\mathbf{Y}}_i = n_i^{-1} \sum_j \mathbf{Y}_{ij}$ and $\mathbf{Y}_0$. All covariance matrices specified as part of the MME framework are assumed to be nonsingular.

Using Bayes' Theorem, the posterior distribution of $\mathbf{Y}_0$ conditional on the

ensemble members $\{\mathbf{Y}_{ij}\}$ is

$$\pi\left(\widetilde{\mathbf{y}}_0|\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}\right) = \frac{\pi(\widetilde{\mathbf{y}}_0)\,\pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}|\widetilde{\mathbf{y}}_0\right)}{\pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}\right)}. \tag{A.1}$$

Any terms not containing the quantity of interest, $\widetilde{\mathbf{y}}_0$, can be absorbed into an arbitrary normalising constant, the sole function of which is to ensure that the posterior density integrates to one. The posterior distribution can therefore be written as

$$\pi\left(\widetilde{\mathbf{y}}_0|\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}\right) \propto \pi(\widetilde{\mathbf{y}}_0)\,\pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}|\widetilde{\mathbf{y}}_0\right). \tag{A.2}$$

The second term on the right-hand side of this expression can be written as an integral,

$$\pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}|\widetilde{\mathbf{y}}_0\right) = \int \pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p},\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_p|\widetilde{\mathbf{y}}_0\right) d\boldsymbol{\mu}_1,\dots,d\boldsymbol{\mu}_p \tag{A.3}$$

which, under the conditional assumptions encoded by Figure A.1, can be factorised as

$$\begin{aligned} &\int \pi\left(\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}|\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_p\right)\pi\left(\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_p|\widetilde{\mathbf{y}}_0\right) d\boldsymbol{\mu}_1,\dots,d\boldsymbol{\mu}_p \\ &= \int \prod_{i=1}^{p}\prod_{j=1}^{n_p} \pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right)\pi\left(\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_p|\widetilde{\mathbf{y}}_0\right) d\boldsymbol{\mu}_1,\dots,d\boldsymbol{\mu}_p. \end{aligned} \tag{A.4}$$

The posterior (A.2) can therefore be expressed as

$$\pi\left(\widetilde{\mathbf{y}}_0|\mathbf{y}_{11},\dots,\mathbf{y}_{pn_p}\right) \propto \pi(\widetilde{\mathbf{y}}_0) \int \prod_{i=1}^{p}\prod_{j=1}^{n_i} \pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right)\pi\left(\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_p|\widetilde{\mathbf{y}}_0\right) d\boldsymbol{\mu}_1,\dots,d\boldsymbol{\mu}_p. \tag{A.5}$$

A.2 A simplified Gaussian posterior

In (2.2.2), all elements of the MME are described by multivariate normal distributions. In particular, from (2.6), each ensemble member $\mathbf{Y}_{ij}$ is assumed

to be normally distributed conditional on the ensemble population mean $\boldsymbol{\mu}_i$, with expectation $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{C}_i$, so that

$$\pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_{ij}-\boldsymbol{\mu}_i\right)' \mathbf{C}_i^{-1}\left(\mathbf{y}_{ij}-\boldsymbol{\mu}_i\right)\right\} \tag{A.6}$$

and

$$\prod_{j=1}^{n_i} \pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right) \propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij}-\boldsymbol{\mu}_i\right)' \mathbf{C}_i^{-1}\left(\mathbf{y}_{ij}-\boldsymbol{\mu}_i\right)\right\}. \tag{A.7}$$

By adding and subtracting the ensemble sample mean $\overline{\mathbf{y}}_i = n_i^{-1}\sum_j \mathbf{y}_{ij}$, and noting that $\sum_j\left(\mathbf{y}_{ij}-\overline{\mathbf{y}}_i\right)=\mathbf{0}$, the exponent of (A.7) can be expanded as

$$-\frac{1}{2}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij}-\overline{\mathbf{y}}_i\right)' \mathbf{C}_i^{-1}\left(\mathbf{y}_{ij}-\overline{\mathbf{y}}_i\right) - \frac{n_i}{2}\left(\overline{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)' \mathbf{C}_i^{-1}\left(\overline{\mathbf{y}}_i-\boldsymbol{\mu}_i\right). \tag{A.8}$$

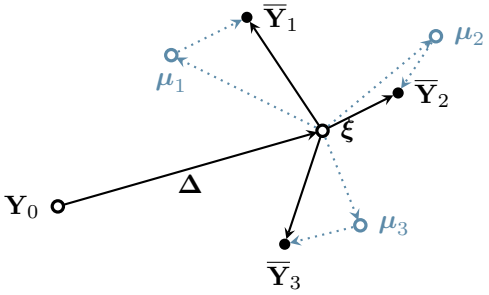
Again, any factors not involving $\boldsymbol{\mu}_i$, on which the distribution is conditioned, can be subsumed into the normalising constant, leaving

$$\prod_{j=1}^{n_i} \pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right) \propto \exp\left\{-\frac{n_i}{2}\left(\overline{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)' \mathbf{C}_i^{-1}\left(\overline{\mathbf{y}}_i-\boldsymbol{\mu}_i\right)\right\}. \tag{A.9}$$

This is the kernel of the multivariate normal density of $\overline{\mathbf{y}}_i|\boldsymbol{\mu}_i$, which has mean $\boldsymbol{\mu}_i$ and covariance matrix $n_i^{-1}\mathbf{C}_i$. This is directly equivalent to $\prod_{j=1}^{n_i}\pi\left(\mathbf{y}_{ij}|\boldsymbol{\mu}_i\right)$ because the sample mean $\overline{\mathbf{y}}_i$ is a sufficient statistic for the population mean in a multivariate normal distribution (Cox and Hinkley, 1974). The posterior (A.5) can therefore be simplified, without loss of information, to a representation based on the schematic in Figure 2.8 and reproduced in Figure A.2, which uses the ensemble means rather than the ensemble members:

$$\pi\left(\widetilde{\mathbf{y}}_0|\mathbf{y}_{11},\ldots,\mathbf{y}_{pn_p}\right) \propto \pi(\widetilde{\mathbf{y}}_0)\int\prod_{i=1}^{p}\pi(\overline{\mathbf{y}}_i|\boldsymbol{\mu}_i)\,\pi\left(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_p|\widetilde{\mathbf{y}}_0\right)d\boldsymbol{\mu}_1,\ldots,d\boldsymbol{\mu}_p. \tag{A.10}$$

Figure A.2: Simplified schematic representation of the relationships between the elements of the multi-model ensemble forecasting system, originally introduced in Figure 2.8. Quantities known at the time of issuing the forecast are shown as filled nodes, with unknown quantities represented by open nodes. Dotted lines indicate redundant nodes which have been bypassed.



$\overline{\mathbf{Y}}_i$	Ensemble sample means
$\boldsymbol{\mu}_i$	Ensemble population means
$\boldsymbol{\xi}$	MME consensus
$\mathbf{Y}_0$	Observed temperature
$\boldsymbol{\Delta}$	Consensus discrepancy

A.3 Obtaining the posterior mean & covariance of $\mathbf{Y}_0$

The value of the integral in (A.10) can be found without carrying out the integration explicitly. In Sections 2.2.2.1 and 2.2.2.2, the remaining model components were specified as in Table A.1, with $\widetilde{\mathbf{Y}}_0$ and $\widetilde{\boldsymbol{\Delta}}$ obtained from $\mathbf{Y}_0$ and $\boldsymbol{\Delta}$ by adding and subtracting $\boldsymbol{\eta}$ as described above. All of the components

Table A.1: Multivariate-normal distributions of the components of the reduced MME framework in Figure 2.8

$$\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i \sim MVN\Big(\boldsymbol{\mu}_i, n_i^{-1}\mathbf{C}_i\Big) \tag{A.11}$$

$$\boldsymbol{\mu}_i|\boldsymbol{\xi} \sim MVN(\boldsymbol{\xi}, \boldsymbol{\Sigma}) \qquad \text{where } \boldsymbol{\xi} = \widetilde{\mathbf{Y}}_0 + \widetilde{\boldsymbol{\Delta}} \tag{A.12}$$

$$\widetilde{\boldsymbol{\Delta}} \sim MVN(\mathbf{0}, \boldsymbol{\Lambda}) \tag{A.13}$$

$$\widetilde{\mathbf{Y}}_0 \sim MVN(\boldsymbol{\alpha}+\boldsymbol{\eta}, \boldsymbol{\Gamma}) \tag{A.14}$$

are multivariate normal quantities, so the joint density $\pi(\{\overline{\mathbf{y}}_i\}|\widetilde{\mathbf{y}}_0)$ represented by the integral in (A.10) will also be multivariate normal, and can be fully specified by its mean and covariance.

From (A.11), because $\overline{\mathbf{Y}}_i$ is independent of $\widetilde{\mathbf{Y}}_0$ and $\widetilde{\boldsymbol{\Delta}}$ given $\boldsymbol{\mu}_i$, $\mathbb{E}\left[\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i, \widetilde{\mathbf{Y}}_0\right] = \mathbb{E}\left[\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i\right] = \boldsymbol{\mu}_i$; and from (A.12) and (A.13), $\mathbb{E}\left[\boldsymbol{\mu}_i|\widetilde{\mathbf{Y}}_0\right] = \widetilde{\mathbf{Y}}_0$. Thus, by the law of iterated expectation,

$$\mathbb{E}\left[\overline{\mathbf{Y}}_i|\widetilde{\mathbf{Y}}_0\right] = \mathbb{E}\left[\mathbb{E}\left[\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i, \widetilde{\mathbf{Y}}_0\right]\right] = \widetilde{\mathbf{Y}}_0. \tag{A.15}$$

From the graph in Figure 2.8, the ith ensemble mean can be written as

$$\overline{\mathbf{Y}}_i = \left[\widetilde{\mathbf{Y}}_0 + \widetilde{\boldsymbol{\Delta}}\right] + \left[\boldsymbol{\mu}_i - \left(\widetilde{\mathbf{Y}}_0 + \widetilde{\boldsymbol{\Delta}}\right)\right] + \left[\overline{\mathbf{Y}}_i - \boldsymbol{\mu}_i\right], \tag{A.16}$$

where each of the terms in square brackets is independent of the others. Thus for any two ensembles i and j,

$$\mathbb{C}\left[\overline{\mathbf{Y}}_i, \overline{\mathbf{Y}}_j|\widetilde{\mathbf{Y}}_0\right] = \begin{cases} \boldsymbol{\Lambda} + \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i & \text{for } i = j \\ \boldsymbol{\Lambda} & \text{for } i \neq j \end{cases} \tag{A.17}$$

The joint forecast distribution of the MME system as a whole, conditional on $\widetilde{\mathbf{Y}}_0 + \widetilde{\boldsymbol{\Delta}}$, is obtained by concatenating all of the ensemble forecasts into a single $m \times p$ vector $\hat{\boldsymbol{\Xi}} = \begin{bmatrix} \overline{\mathbf{y}}_1 & \dots & \overline{\mathbf{y}}_p \end{bmatrix}$. For notational convenience, let $\mathbf{D}_i = \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i$; then from (A.15) and (A.17), $\hat{\boldsymbol{\Xi}}$ has a multivariate normal distribution, with expectation $\boldsymbol{\Xi} = \begin{bmatrix} \widetilde{\mathbf{y}}_0 & \dots & \widetilde{\mathbf{y}}_0 \end{bmatrix}$ and covariance matrix

$$\begin{bmatrix} \mathbf{D}_1 + \boldsymbol{\Lambda} & \boldsymbol{\Lambda} & \dots & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda} & \mathbf{D}_2 + \boldsymbol{\Lambda} & \dots & \boldsymbol{\Lambda} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda} & \boldsymbol{\Lambda} & \dots & \mathbf{D}_p + \boldsymbol{\Lambda} \end{bmatrix} = \mathbf{K}, \text{ say.} \tag{A.18}$$

(A.4) is therefore proportional to

$$\exp\left\{-\frac{1}{2}\left(\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}\right)' \mathbf{K}^{-1}\left(\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}\right)\right\}. \tag{A.19}$$

A.3.1 The likelihood of the MME forecasts

To find an explicit expression for the matrix inverse $\mathbf{K}^{-1}$, note that $\mathbf{K}$ can be decomposed as $\mathbf{K} = \mathbf{D} + \mathbf{LV}$, where $\mathbf{D}$ is a block-diagonal matrix with $\mathbf{D}_i$ in the ith block, $\mathbf{L}$ is the $mp \times m$ matrix $\begin{bmatrix} \mathbf{\Lambda} & \mathbf{\Lambda} & \dots & \mathbf{\Lambda} \end{bmatrix}'$, and $\mathbf{V}$ is the $m \times mp$ matrix $\begin{bmatrix} \mathbf{I}_m & \mathbf{I}_m & \dots & \mathbf{I}_m \end{bmatrix}$, where p denotes the number of ensembles in the MME, and m the dimension of $\mathbf{Y}_0$. Using this decomposition of $\mathbf{K}$, the inverse can be computed using the Woodbury matrix identity (Woodbury, 1950; Press et al., 1996), which states that for conformable matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{C}$ and $\mathbf{V}$,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}. \tag{A.20}$$

Setting $\mathbf{A} = \mathbf{D}$, $\mathbf{U} = \mathbf{L}$, $\mathbf{C} = \mathbf{I}_m$, and $\mathbf{V} = \mathbf{V}$ as defined above, the precision matrix $\mathbf{K}^{-1}$ is

$$(\mathbf{D} + \mathbf{LV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{L}\left(\mathbf{I}_m + \mathbf{VD}^{-1}\mathbf{L}\right)^{-1}\mathbf{VD}^{-1}. \tag{A.21}$$

Because $\mathbf{D}$ is block diagonal, its inverse is also block diagonal, with the ith non-zero block given by $\mathbf{D}_i^{-1}$. The matrix product $\mathbf{VD}^{-1}\mathbf{L}$ is therefore

$$\mathbf{VD}^{-1}\mathbf{L} = \begin{bmatrix} \mathbf{I}_m & \mathbf{I}_m & \dots & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{D}_1^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{-1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{D}_p^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{\Lambda} \\ \dots \\ \mathbf{\Lambda} \end{bmatrix} = \sum_{k=1}^{p} \mathbf{D}_k^{-1}\mathbf{\Lambda}. \tag{A.22}$$

Now (A.21) becomes

$$(\mathbf{D}+\mathbf{LV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\begin{bmatrix}\boldsymbol{\Lambda}\\ \boldsymbol{\Lambda}\\ \dots\\ \boldsymbol{\Lambda}\end{bmatrix}\left(\mathbf{I}_m + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\boldsymbol{\Lambda}\right)^{-1}\begin{bmatrix}\mathbf{I}_m & \mathbf{I}_m & \dots & \mathbf{I}_m\end{bmatrix}\mathbf{D}^{-1}. \tag{A.23}$$

Note that

$$\left(\mathbf{I}_m + \sum_{k=1}^{n}\mathbf{D}_k^{-1}\boldsymbol{\Lambda}\right)^{-1} = \left[\left(\boldsymbol{\Lambda}^{-1} + \sum_{k=1}^{n}\mathbf{D}_k^{-1}\right)\boldsymbol{\Lambda}\right]^{-1} = \boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Lambda}^{-1} + \sum_{k=1}^{n}\mathbf{D}_k^{-1}\right)^{-1} \tag{A.24}$$

and so

$$\boldsymbol{\Lambda}\left(\mathbf{I}_m + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\boldsymbol{\Lambda}\right)^{-1} = \left(\boldsymbol{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}, \tag{A.25}$$

whence

$$\mathbf{K}^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\left\{\begin{bmatrix}\mathbf{I}_m\\ \mathbf{I}_m\\ \dots\\ \mathbf{I}_m\end{bmatrix}\left(\boldsymbol{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\begin{bmatrix}\mathbf{I}_m & \mathbf{I}_m & \dots & \mathbf{I}_m\end{bmatrix}\right\}\mathbf{D}^{-1}. \tag{A.26}$$

The matrix product enclosed in braces $\{\}$ is a $mp \times mp$ block matrix, with each block containing (A.25). Because $\mathbf{D}^{-1}$ is a block diagonal matrix, the (i,j)th block of $\mathbf{K}^{-1}$ can therefore be written as

$$\mathbf{K}_{ij}^{-1} = \mathbf{D}_{ij}^{-1} - \sum_{q=1}^{p}\sum_{r=1}^{p}\left[\mathbf{D}_{iq}^{-1}\left(\boldsymbol{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\mathbf{D}_{rj}^{-1}\right], \tag{A.27}$$

where (with a slight abuse of notation in the case where $i \neq j$) $\mathbf{D}_{ij}^{-1} = \mathbf{D}_i^{-1}$ when $i = j$ and $\mathbf{0}$ otherwise; $\mathbf{D}_{iq}^{-1} = \mathbf{D}_i^{-1}$ when $q = i$ and $\mathbf{0}$ otherwise; and $\mathbf{D}_{rj}^{-1} = \mathbf{D}_j^{-1}$ when $r = j$ and $\mathbf{0}$ otherwise. The term in square brackets will therefore be zero unless $q = i$ and $r = j$; hence

$$\mathbf{K}_{ij}^{-1} = \mathbb{1}_{\{i=j\}}\mathbf{D}_i^{-1} - \mathbf{D}_i^{-1}\left(\mathbf{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\mathbf{D}_j^{-1}, \tag{A.28}$$

where $\mathbb{1}_{\{i=j\}} = 1$ if $i = j$, and 0 otherwise.

The matrix product in the likelihood function (A.19) can now be expressed as a summation over the blocks of $\mathbf{K}^{-1}$,

$$\begin{aligned}\left(\hat{\mathbf{\Xi}} - \mathbf{\Xi}\right)' \mathbf{K}^{-1} \left(\hat{\mathbf{\Xi}} - \mathbf{\Xi}\right) &= \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right)' \mathbf{K}_{ij}^{-1}\left(\overline{\mathbf{y}}_j - \widetilde{\mathbf{y}}_0\right) \\ &= \sum_{i=1}^{p}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right)' \mathbf{D}_i^{-1}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right) - \\ &\qquad \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right)' \mathbf{D}_i^{-1}\left(\mathbf{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\mathbf{D}_j^{-1}\left(\overline{\mathbf{y}}_j - \widetilde{\mathbf{y}}_0\right). \end{aligned} \tag{A.29}$$

Finally, because the matrix product is distributive with respect to matrix addition, this can be rearranged into

$$\begin{aligned}&\sum_{i=1}^{p}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right)' \mathbf{D}_i^{-1}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right) - \\ &\qquad \sum_{i=1}^{p}\left(\overline{\mathbf{y}}_i - \widetilde{\mathbf{y}}_0\right)' \mathbf{D}_i^{-1}\left(\mathbf{\Lambda}^{-1} + \sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\left(\overline{\mathbf{y}}_j - \widetilde{\mathbf{y}}_0\right).\end{aligned} \tag{A.30}$$

This expansion of the exponent of the likelihood function (A.4) can now be used to write an expression proportional to the posterior density of $\widetilde{\mathbf{Y}}_0$, in order to identify the corresponding posterior expectation and variance.

A.3.2 The posterior expectation and covariance of $\mathbf{Y}_0$

Let $\boldsymbol{\tau}$ denote the posterior expectation of $\mathbf{Y}_0$, and $\mathbf{S}$ the posterior covariance; the posterior density of $\widetilde{\mathbf{Y}}_0$ can therefore be written as

$$\pi\left(\widetilde{\mathbf{y}}_0|\mathbf{y}_{11},\cdots,\mathbf{y}_{pn_p}\right) \propto \exp\left\{-\frac{1}{2}\left(\widetilde{\mathbf{y}}_0-(\boldsymbol{\tau}+\boldsymbol{\eta})\right)'\mathbf{S}^{-1}\left(\widetilde{\mathbf{y}}_0-(\boldsymbol{\tau}+\boldsymbol{\eta})\right)\right\}. \quad \text{(A.31)}$$

It has already been shown that the likelihood (A.4) is proportional to (A.19), and hence to the exponential of (A.30). The posterior density is therefore proportional to

$$\exp\left\{-\frac{1}{2}\left[\left(\widetilde{\mathbf{y}}_0-(\boldsymbol{\alpha}+\boldsymbol{\eta})\right)'\boldsymbol{\Gamma}^{-1}\left(\widetilde{\mathbf{y}}_0-(\boldsymbol{\alpha}+\boldsymbol{\eta})\right)+\right.\right.$$
$$\sum_{i=1}^{p}\left(\overline{\mathbf{y}}_i-\widetilde{\mathbf{y}}_0\right)'\mathbf{D}_i{}^{-1}\left(\overline{\mathbf{y}}_i-\widetilde{\mathbf{y}}_0\right)-$$
$$\left.\left.\sum_{i=1}^{p}\left(\overline{\mathbf{y}}_i-\widetilde{\mathbf{y}}_0\right)'\mathbf{D}_i^{-1}\left(\boldsymbol{\Lambda}^{-1}+\sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\left(\overline{\mathbf{y}}_j-\widetilde{\mathbf{y}}_0\right)\right]\right\}. \quad \text{(A.32)}$$

The posterior covariance matrix $\mathbf{S}$ is obtained by equating the quadratic terms in $\widetilde{\mathbf{y}}_0$ between (A.31) and (A.32):

$$\mathbf{S}^{-1}=\boldsymbol{\Gamma}^{-1}+\sum_{i=1}^{p}\mathbf{D}_i^{-1}-\sum_{i=1}^{p}\mathbf{D}_i^{-1}\left(\boldsymbol{\Lambda}^{-1}+\sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}. \quad \text{(A.33)}$$

Since $\sum_{i=1}^{p}\mathbf{D}_i^{-1}=\sum_{j=1}^{p}\mathbf{D}_j^{-1}=\sum_{k=1}^{p}\mathbf{D}_k^{-1}$, the Woodbury formula (A.20) can again be applied, this time with $\mathbf{A}^{-1}=\sum_{i=1}^{p}\mathbf{D}_i^{-1}$, $\mathbf{C}^{-1}=\boldsymbol{\Lambda}^{-1}$, and $\mathbf{U}=\mathbf{V}=\mathbf{I}$, to obtain

$$\mathbf{S}^{-1}=\boldsymbol{\Gamma}^{-1}+\left[\boldsymbol{\Lambda}+\left(\sum_{i=1}^{p}\mathbf{D}_i^{-1}\right)^{-1}\right]^{-1}. \quad \text{(A.34)}$$

Because $\mathbf{Y}_0$ is simply $\widetilde{\mathbf{Y}}_0$ shifted by a known constant $\boldsymbol{\eta}$, this is also the posterior covariance of $\mathbf{Y}_0$. For notational compactness, let $\boldsymbol{\Sigma_D}=\left(\sum_{i=1}^{p}\mathbf{D}_i^{-1}\right)^{-1}$, a quantity describing the inverse of the total precision of the sampled $\left\{\overline{\mathbf{Y}}_i\right\}$ and

representing the uncertainty about the position of the consensus $\boldsymbol{\xi}$, given the observed ensemble means $\overline{\mathbf{y}}_i$. The posterior precision matrix $\mathbf{S}^{-1}$ is therefore the sum of the prior precision $\boldsymbol{\Gamma}^{-1}$ and the precision of the discrepancy-adjusted consensus $\boldsymbol{\xi}-\boldsymbol{\eta}$:

$$\mathbf{S}^{-1}=\boldsymbol{\Gamma}^{-1}+\left[\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}}\right]^{-1} \tag{A.35}$$

Similarly, the posterior expectation of $\widetilde{\mathbf{Y}}_0$ is obtained by equating the coefficients of $\tilde{\mathbf{y}}_0$ in (A.31) and (A.32):

$$\begin{aligned}\mathbf{S}^{-1}(\boldsymbol{\tau}+\boldsymbol{\eta}) &= \boldsymbol{\Gamma}^{-1}(\boldsymbol{\alpha}+\boldsymbol{\eta})+\sum_{i=1}^{p}\mathbf{D}_i^{-1}\overline{\mathbf{y}}_i-\sum_{i=1}^{p}\mathbf{D}_i^{-1}\left(\boldsymbol{\Lambda}^{-1}+\sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j \\ &= \boldsymbol{\Gamma}^{-1}(\boldsymbol{\alpha}+\boldsymbol{\eta})+\left\{\mathbf{I}-\sum_{i=1}^{p}\mathbf{D}_i^{-1}\left(\boldsymbol{\Lambda}^{-1}+\sum_{k=1}^{p}\mathbf{D}_k^{-1}\right)^{-1}\right\}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j.\end{aligned} \tag{A.36}$$

The term inside the braces {} is another occurrence of the Woodbury formula (A.20), this time with $\mathbf{A}=\mathbf{V}=\mathbf{I}$, $\mathbf{C}=\boldsymbol{\Lambda}$, and $\mathbf{U}=\sum_{i=1}^{p}\mathbf{D}_i^{-1}=\sum_{k=1}^{p}\mathbf{D}_k^{-1}$; hence

$$\mathbf{S}^{-1}\left(\boldsymbol{\tau}+\boldsymbol{\eta}\right)=\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\alpha}+\boldsymbol{\eta}\right)+\left(\mathbf{I}+\sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j, \tag{A.37}$$

and the posterior expectation of $\widetilde{\mathbf{Y}}_0$ is

$$\boldsymbol{\tau}+\boldsymbol{\eta}=\mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\alpha}+\boldsymbol{\eta}\right)+\left(\mathbf{I}+\sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j\right]. \tag{A.38}$$

Removing the translation by $\boldsymbol{\eta}$, the posterior expectation of $\mathbf{Y}_0$ is therefore

$$\boldsymbol{\tau}=\mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\alpha}+\boldsymbol{\eta}\right)+\left(\mathbf{I}+\sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j\right]-\boldsymbol{\eta}. \tag{A.39}$$

A.3.2.1 A more interpretable form for the posterior expectation

Interpretation of the contribution of each element of the framework to the posterior expectation in (A.39) is not straightforward. However, it is possible to rearrange this expression into a weighted sum of the contributions from the prior and likelihood.

First, note that from (A.35), $\mathbf{S} = \left(\boldsymbol{\Gamma}^{-1} + [\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}]^{-1}\right)^{-1}$ has the form $\left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1}$, with $\mathbf{A} = \boldsymbol{\Gamma}$ and $\mathbf{B} = \boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}$. This expression can be expanded as

$$\begin{aligned}\left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1} &= \left(\mathbf{B}^{-1}\mathbf{A}\mathbf{A}^{-1} + \mathbf{B}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)^{-1} \\ &= \left(\mathbf{B}^{-1}\left[\mathbf{A} + \mathbf{B}\right]\mathbf{A}^{-1}\right)^{-1} \\ &= \mathbf{A}\left(\mathbf{A} + \mathbf{B}\right)^{-1}\mathbf{B}, \end{aligned} \tag{A.40}$$

whence

$$\mathbf{S} = \boldsymbol{\Gamma}\left(\boldsymbol{\Gamma} + \boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)^{-1}\left(\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right) \tag{A.41}$$

and

$$\mathbf{S}^{-1} = \left(\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)^{-1}\left(\boldsymbol{\Gamma} + \boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)\boldsymbol{\Gamma}^{-1}. \tag{A.42}$$

Bringing $\boldsymbol{\eta}$ inside the square brackets in (A.39) and using this expansion of $\mathbf{S}^{-1}$,

$$\begin{aligned}\boldsymbol{\tau} &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\alpha} + \boldsymbol{\eta}\right) + \left(\mathbf{I} + \sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j - \mathbf{S}^{-1}\boldsymbol{\eta}\right] \\ &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\alpha} + \boldsymbol{\eta}\right) + \left(\mathbf{I} + \sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j - \left(\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)^{-1}\left(\boldsymbol{\Gamma} + \boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right]. \end{aligned} \tag{A.43}$$

The term $\left(\mathbf{I} + \sum_{i=1}^{p}\mathbf{D}_i^{-1}\boldsymbol{\Lambda}\right)^{-1}$ can be written as $\left(\mathbf{I} + \boldsymbol{\Sigma}_{\boldsymbol{D}}^{-1}\boldsymbol{\Lambda}\right)^{-1} = \left(\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{\boldsymbol{D}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{D}}$,

so that

$$\begin{aligned}\boldsymbol{\tau} &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}(\boldsymbol{\alpha}+\boldsymbol{\eta})+(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j-(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}(\boldsymbol{\Gamma}+\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right] \\ &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}(\boldsymbol{\alpha}+\boldsymbol{\eta})+(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}\left\{\boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j-(\boldsymbol{\Gamma}+\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right\}\right]\end{aligned} \tag{A.44}$$

Now by expanding and gathering terms involving $\boldsymbol{\Gamma}^{-1}$,

$$\begin{aligned}\boldsymbol{\tau} &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha}+\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}+(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}\left\{\boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j-\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}-(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right\}\right] \\ &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha}+\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}+(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}\left\{\boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j-\boldsymbol{\eta}\right\}-\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right] \\ &= \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha}+(\boldsymbol{\Lambda}+\boldsymbol{\Sigma}_{\boldsymbol{D}})^{-1}\left\{\boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j-\boldsymbol{\eta}\right\}\right].\end{aligned} \tag{A.45}$$

The term summing the $\left\{\overline{\mathbf{y}}_j\right\}$ can be considered as an estimate of the MME consensus,

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\Sigma}_{\boldsymbol{D}}\sum_{j=1}^{p}\mathbf{D}_j^{-1}\overline{\mathbf{y}}_j. \tag{A.46}$$

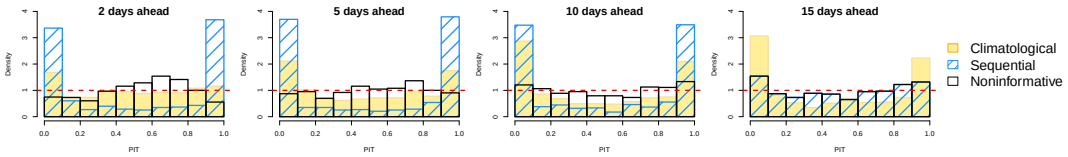
This term is a weighted average of the ensemble means, with weights proportional to the precision of each estimated $\overline{\mathbf{y}}_j$; $\boldsymbol{\Sigma}_{\boldsymbol{D}} = \left(\sum_{i=1}^{p}\mathbf{D}_i^{-1}\right)^{-1}$ is the inverse of the total precision of the sampled $\overline{\mathbf{y}}_i$ over all ensembles. The posterior mean $\boldsymbol{\tau}$ is thus a weighted sum of contributions from the prior distribution with expectation $\boldsymbol{\alpha}$ and covariance matrix $\boldsymbol{\Gamma}$, and the bias-corrected MME consensus $\hat{\boldsymbol{\xi}}-\boldsymbol{\eta}$, with combined covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{D}}+\boldsymbol{\Lambda}$.

Appendix B

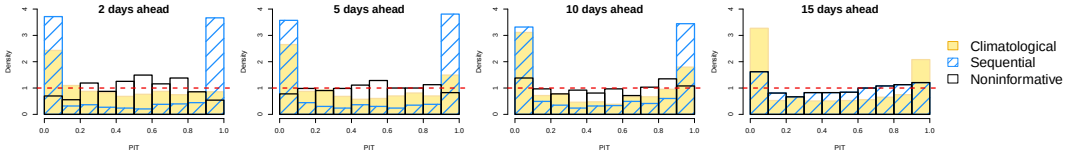
PIT histograms for regions not shown in the main text

Figure B.1: PIT histograms accompanying those in Figure 4.4 showing the marginal calibration of forecasts of surface temperatures at a range of leadtimes, postprocessed using the Bayesian method with different prior distributions for the observed temperature. The dashed line indicates the ideal uniform distribution.

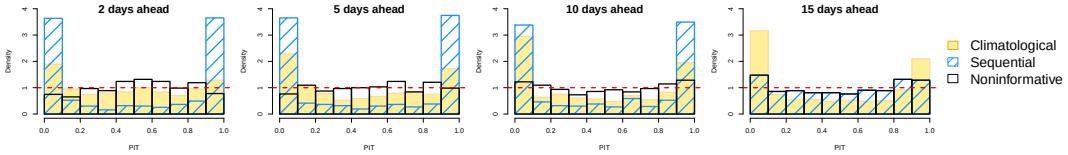
(a) PIT histograms for forecasts in Southampton



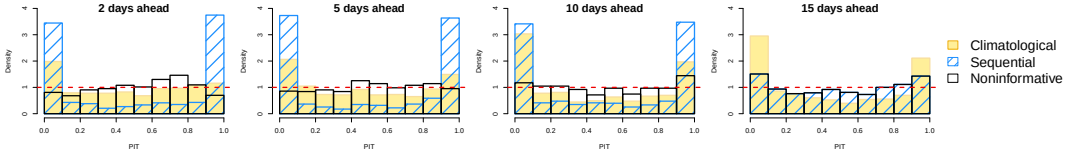
(b) PIT histograms for forecasts in Swansea



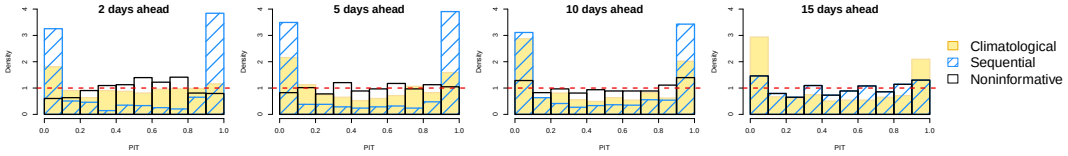
(c) PIT histograms for forecasts in Birmingham



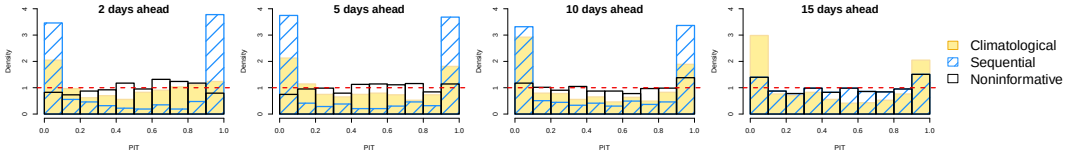
(d) PIT histograms for forecasts in London



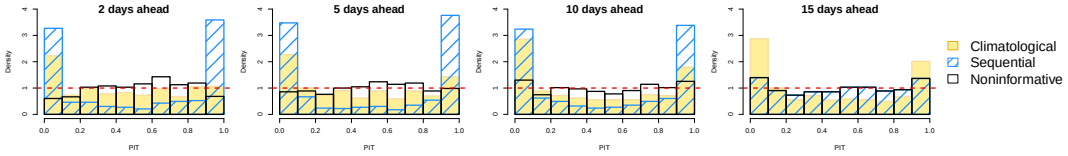
(e) PIT histograms for forecasts in Liverpool



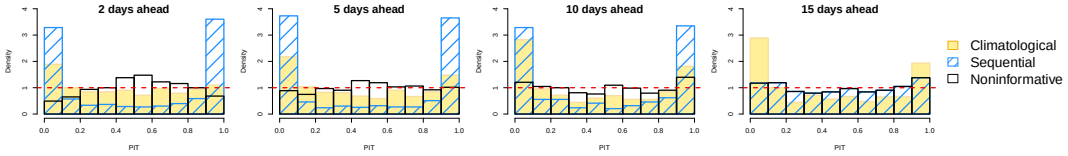
(f) PIT histograms for forecasts in Sheffield



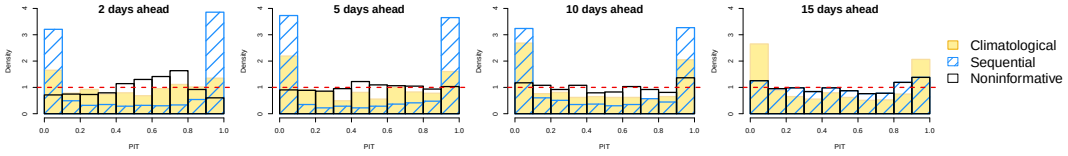
(g) PIT histograms for forecasts in Leeds



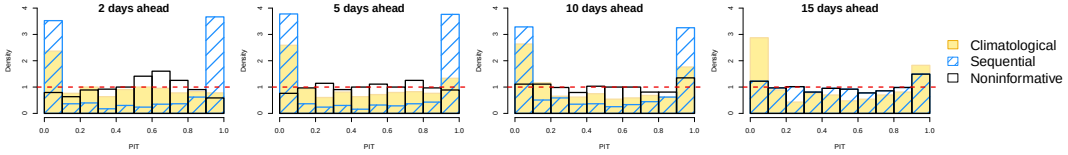
(h) PIT histograms for forecasts in Carlisle



(i) PIT histograms for forecasts in Glasgow



(j) PIT histograms for forecasts in Fort William



(k) PIT histograms for forecasts in Forres

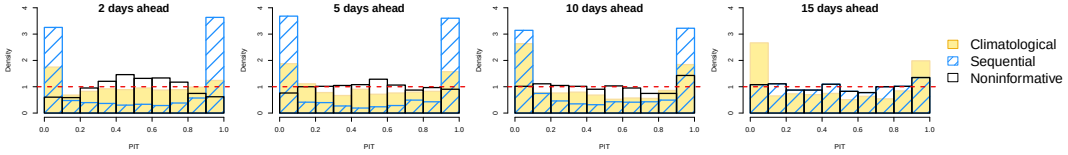
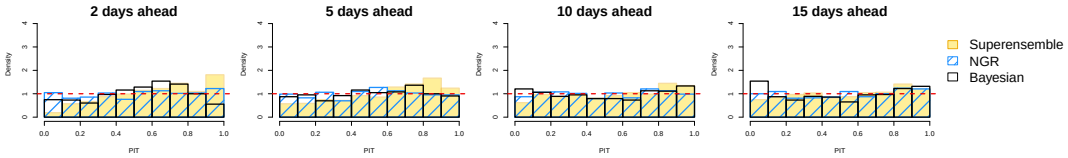
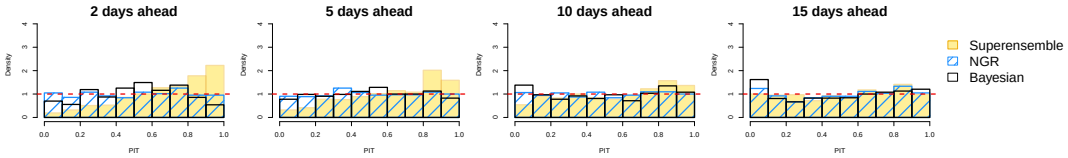


Figure B.2: PIT histograms accompanying those in Figure 4.13 showing the marginal calibration of postprocessed forecasts of surface temperatures at a range of leadtimes, using different postprocessing methods. The dashed line indicates the ideal uniform distribution.

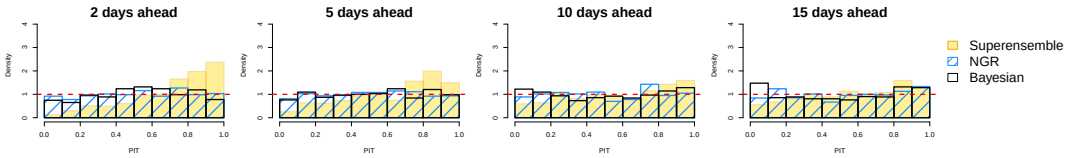
(a) PIT histograms for forecasts in Southampton



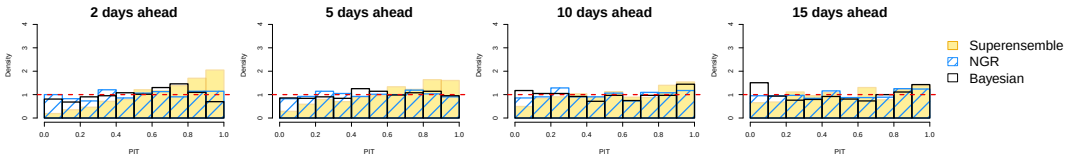
(b) PIT histograms for forecasts in Swansea



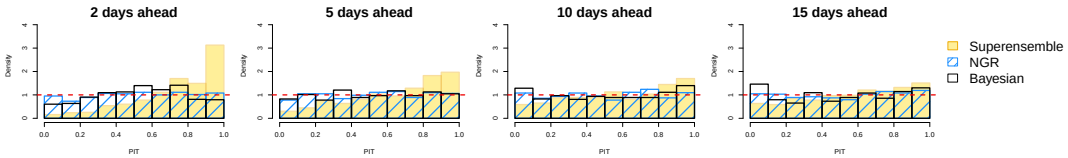
(c) PIT histograms for forecasts in Birmingham



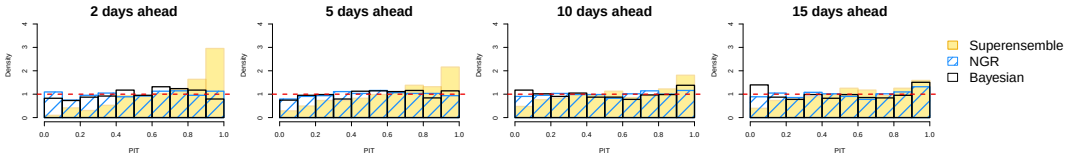
(d) PIT histograms for forecasts in London



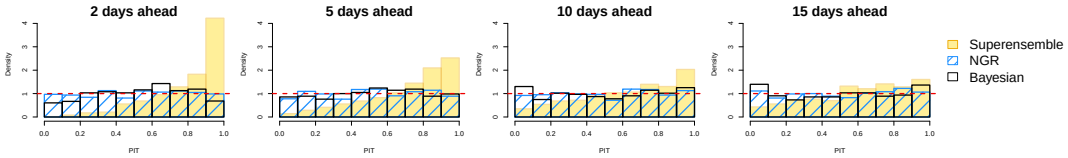
(e) PIT histograms for forecasts in Liverpool



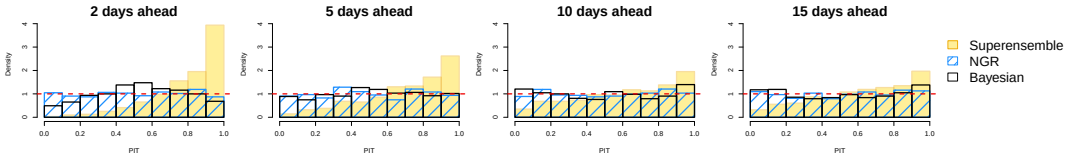
(f) PIT histograms for forecasts in Sheffield



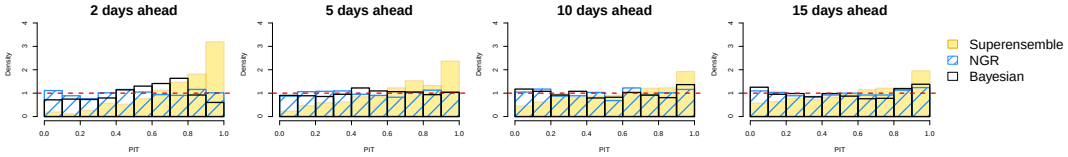
(g) PIT histograms for forecasts in Leeds



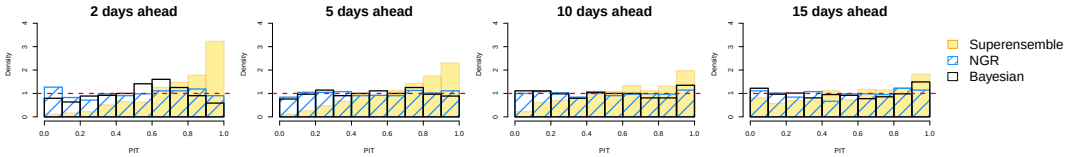
(h) PIT histograms for forecasts in Carlisle



(i) PIT histograms for forecasts in Glasgow



(j) PIT histograms for forecasts in Fort William



(k) PIT histograms for forecasts in Forres

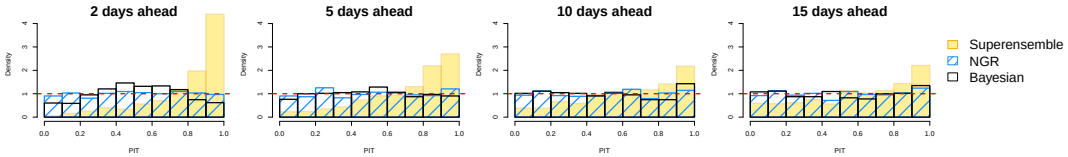
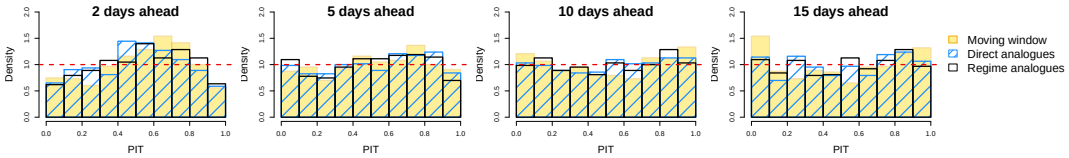
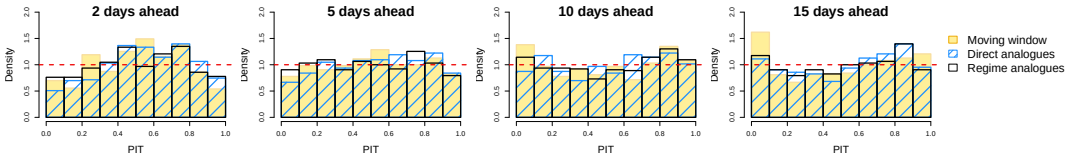


Figure B.3: PIT histograms accompanying those in Figure 4.21 showing the marginal calibration of forecasts of forecasts of surface temperatures at a range of leadtimes, postprocessed using the Bayesian method with different training sets. The dashed line indicates the ideal uniform distribution.

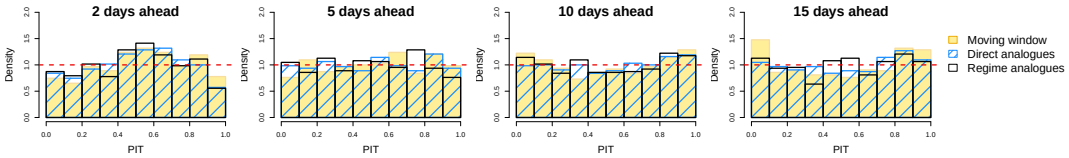
(a) PIT histograms for forecasts in Southampton



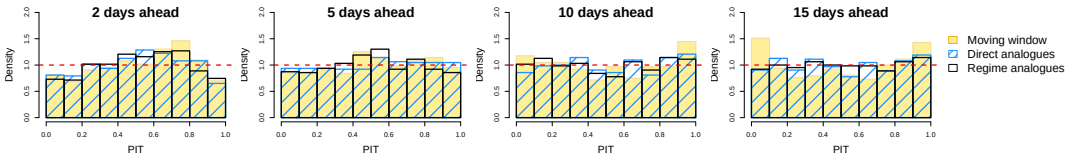
(b) PIT histograms for forecasts in Swansea



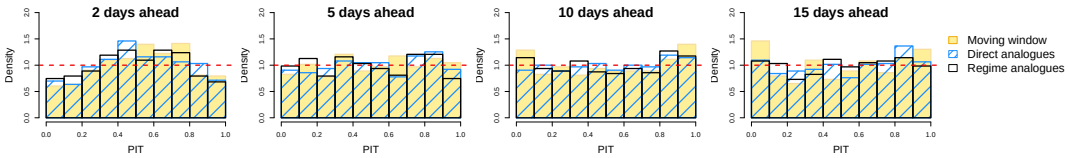
(c) PIT histograms for forecasts in Birmingham



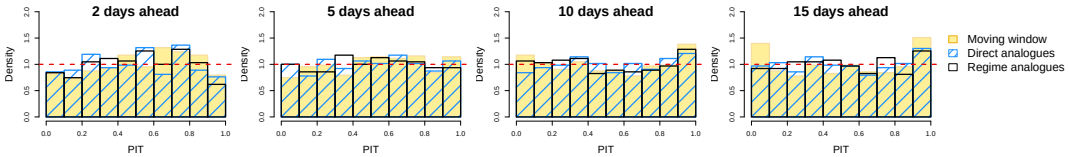
(d) PIT histograms for forecasts in London



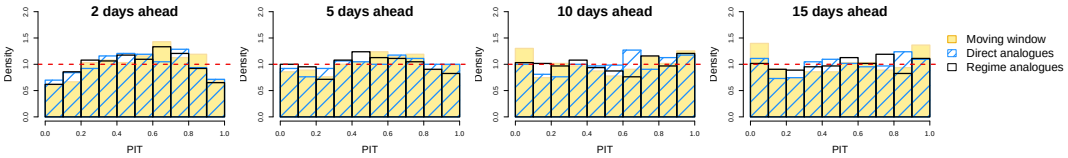
(e) PIT histograms for forecasts in Liverpool



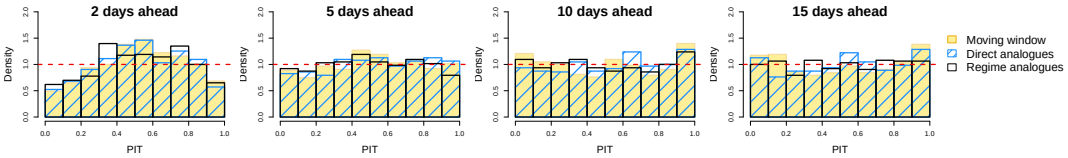
(f) PIT histograms for forecasts in Sheffield



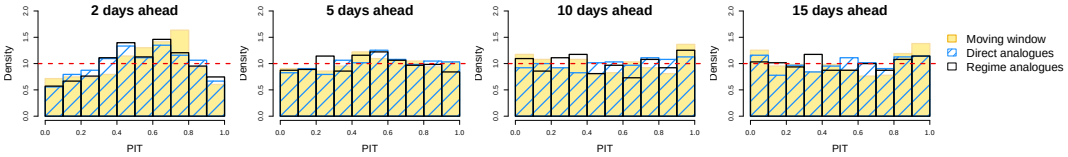
(g) PIT histograms for forecasts in Leeds



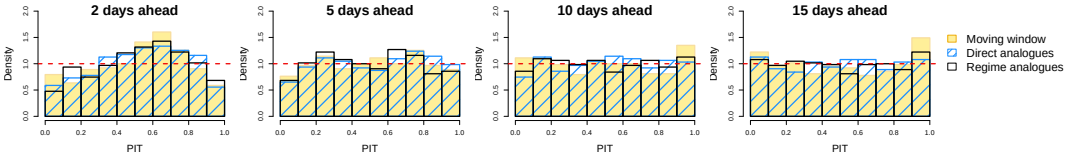
(h) PIT histograms for forecasts in Carlisle



(i) PIT histograms for forecasts in Glasgow



(j) PIT histograms for forecasts in Fort William



(k) PIT histograms for forecasts in Forres

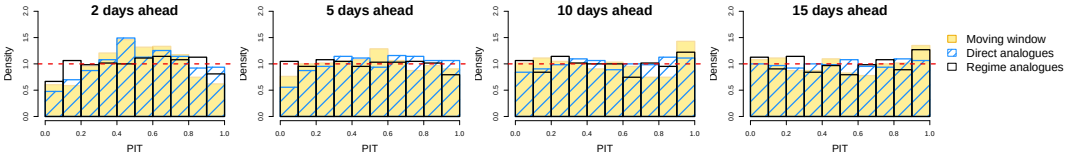
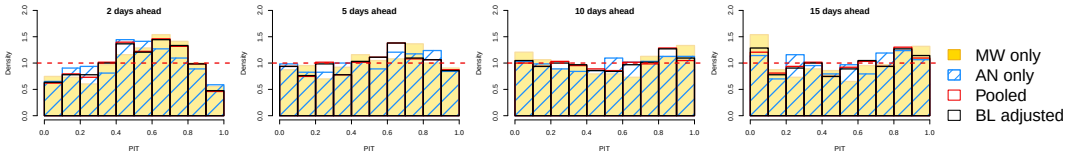
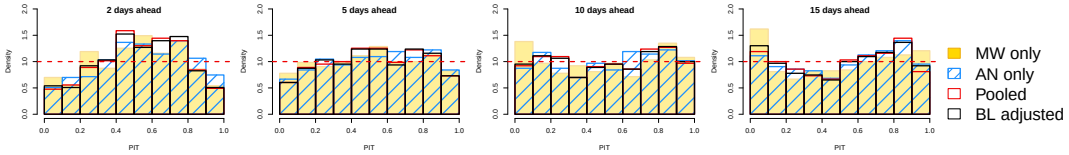


Figure B.4: PIT histograms accompanying those in Figure 6.4 showing the marginal calibration of postprocessed forecasts of surface temperatures at a range of leadtimes, for forecasts postprocessed using either direct, Bayes linear adjusted or pooled estimates of the expectation and variance of the discrepancy $\boldsymbol{\Delta}$. The dashed line indicates the ideal uniform distribution.

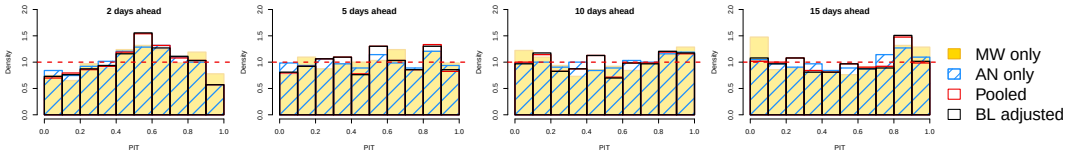
(a) PIT histograms for forecasts in Southampton



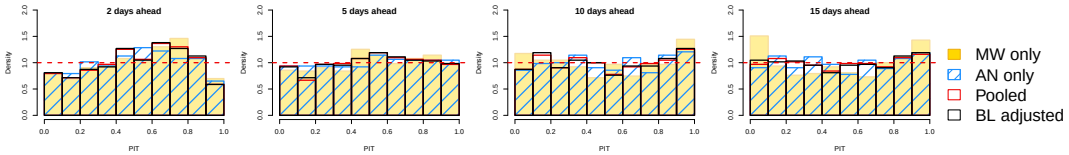
(b) PIT histograms for forecasts in Swansea



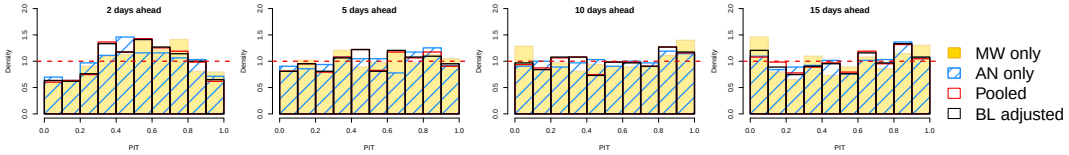
(c) PIT histograms for forecasts in Birmingham



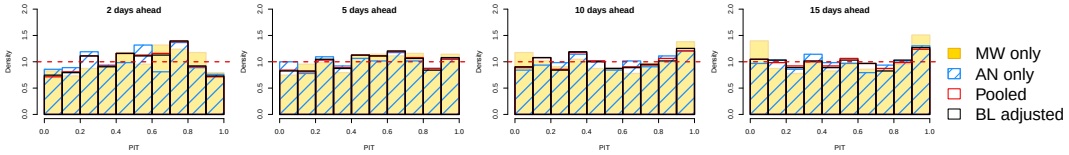
(d) PIT histograms for forecasts in London



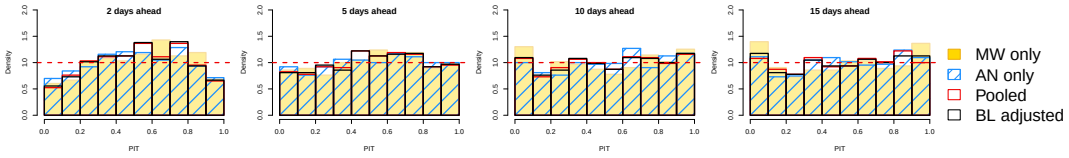
(e) PIT histograms for forecasts in Liverpool



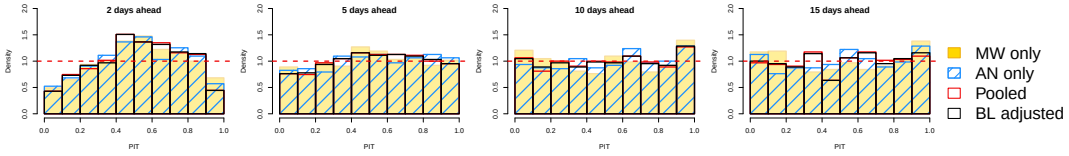
(f) PIT histograms for forecasts in Sheffield



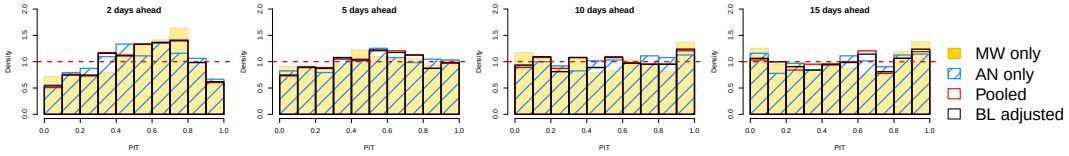
(g) PIT histograms for forecasts in Leeds



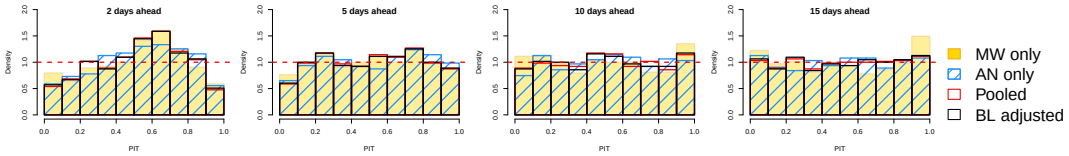
(h) PIT histograms for forecasts in Carlisle



(i) PIT histograms for forecasts in Glasgow



(j) PIT histograms for forecasts in Fort William



(k) PIT histograms for forecasts in Forres

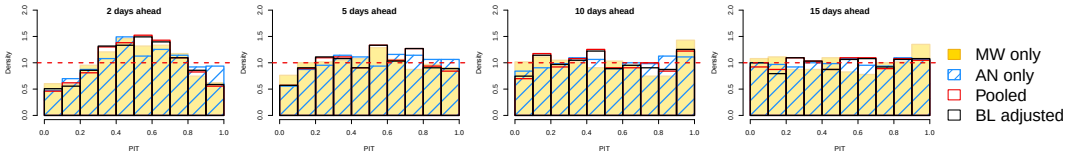
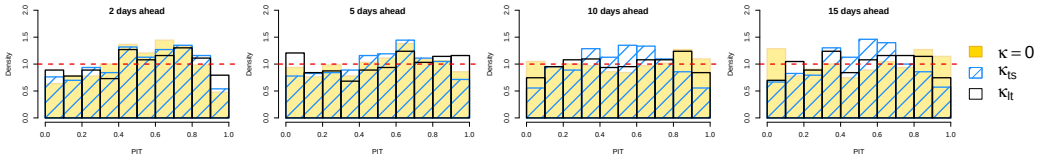
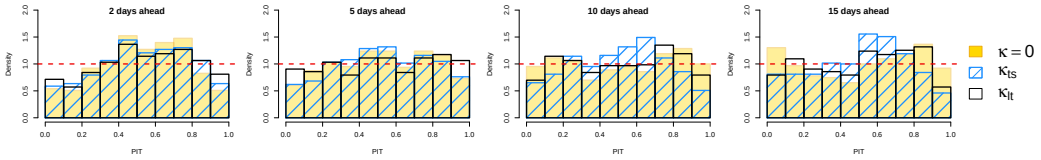


Figure B.5: PIT histograms accompanying those in Figure 6.27 showing the marginal calibration of forecasts postprocessed using Bayes linear adjusted discrepancies with different choices of κ at a range of leadtimes. The dashed line indicates the ideal uniform distribution.

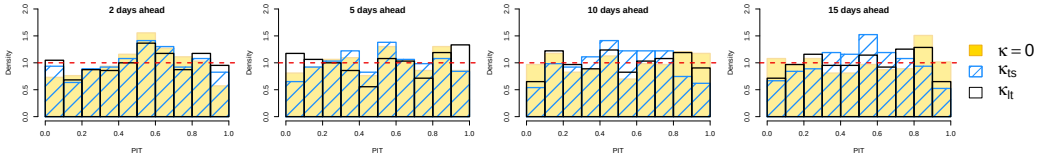
(a) PIT histograms for forecasts in Southampton



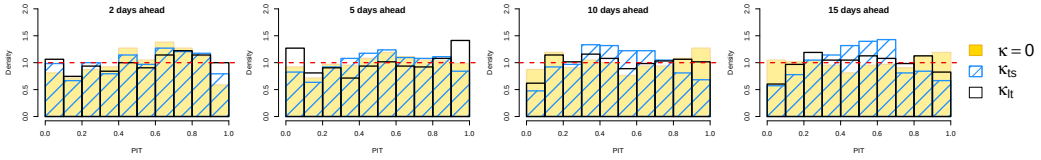
(b) PIT histograms for forecasts in Swansea



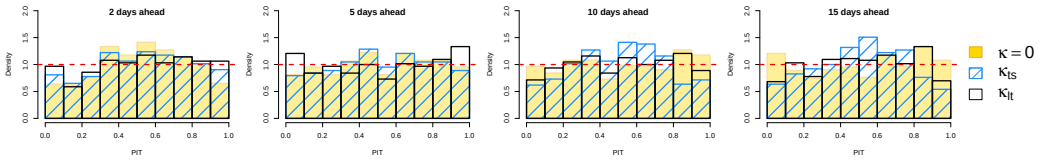
(c) PIT histograms for forecasts in Birmingham



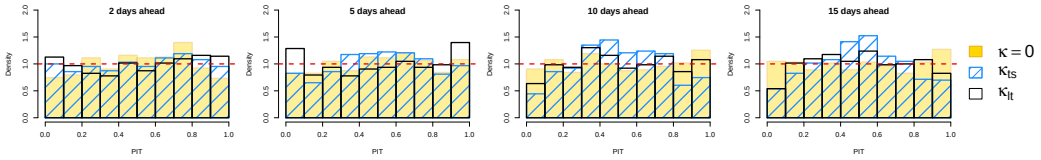
(d) PIT histograms for forecasts in London



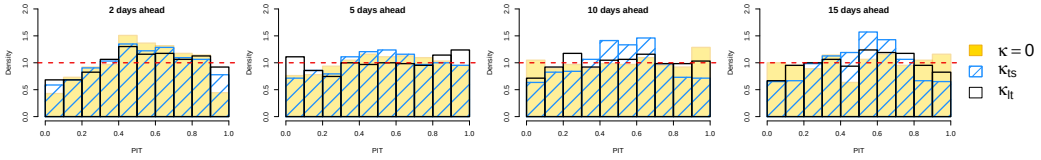
(e) PIT histograms for forecasts in Liverpool



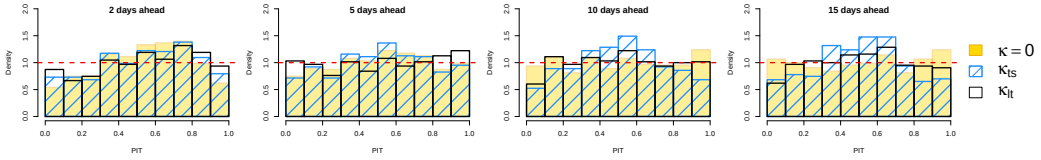
(f) PIT histograms for forecasts in Sheffield



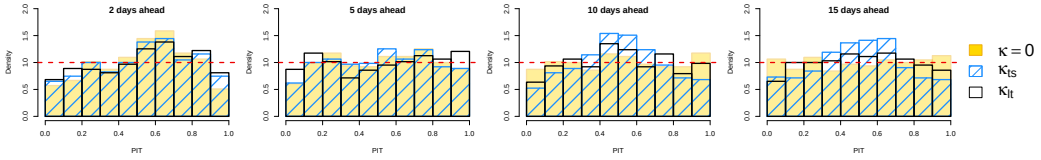
(g) PIT histograms for forecasts in Carlisle



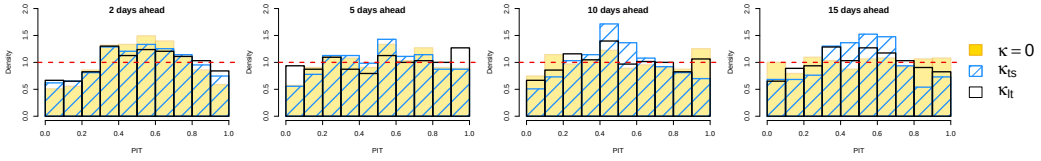
(h) PIT histograms for forecasts in Glasgow



(i) PIT histograms for forecasts in Fort William



(j) PIT histograms for forecasts in Forres



Appendix C

Review of some key concepts in matrix algebra

C.1 The vec operator

When dealing with variances of the elements of a matrix $\mathbf{A}$, say, the standard approach is to transform the matrix into a vector. This makes it possible to define the matrix of variances and covariances between the elements of $\mathbf{A}$, which does not exist when $\mathbf{A}$ is in matrix form. This operation is carried out using the vec operator: given any $m \times n$ matrix $\mathbf{A}$, $\text{vec}\Big(\mathbf{A}\Big)$ denotes the $mn \times 1$ vector obtained by stacking the columns of $\mathbf{A}$ one underneath the other (Schott, 2016). Products of vectorised matrices are conveniently expressed in terms of their Kronecker products.

C.2 The Kronecker product

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is a $p \times q$ matrix, then the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{C.1}$$

The derivations in Section 5.3 make use of the following identities presented in Schott (2016):

1. For two vectors $\mathbf{x}$ and $\mathbf{y}$ of any length,

$$\text{vec}\left(\mathbf{x}\mathbf{y}'\right) = \mathbf{y} \otimes \mathbf{x}, \tag{C.2}$$

$$\mathbf{x}\mathbf{y}' = \mathbf{x} \otimes \mathbf{y}' = \mathbf{y}' \otimes \mathbf{x} \tag{C.3}$$

2. For matrices $\mathbf{A}$ and $\mathbf{B}$ of any size,

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}' \tag{C.4}$$

3. For square matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} \otimes \mathbf{B}$ is nonsingular then

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}. \tag{C.5}$$

   If $\mathbf{A} \otimes \mathbf{B}$ is singular, then for any generalised inverses $\mathbf{A}^{\dagger}$ and $\mathbf{B}^{\dagger}$ of $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})^{\dagger} = \mathbf{A}^{\dagger} \otimes \mathbf{B}^{\dagger}. \tag{C.6}$$

4. For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, where $\mathbf{B}$ and $\mathbf{C}$ are of the same size,

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{C}) \tag{C.7}$$

5. For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ of sizes $m \times h$, $p \times k$, $h \times n$ and $k \times q$ respectively,

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \tag{C.8}$$

6. For matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ of sizes $m \times n$, $n \times p$, and $p \times q$,

$$\text{vec}\Big(\mathbf{ABC}\Big) = \Big(\mathbf{C}' \otimes \mathbf{A}\Big)\,\text{vec}\Big(\mathbf{B}\Big) \tag{C.9}$$

7. For $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\text{vec}\Big(\mathbf{A}\Big)'\,\text{vec}\Big(\mathbf{B}\Big) = \text{tr}\left(\mathbf{A}'\mathbf{B}\right). \tag{C.10}$$

C.3 The position matrix $\mathbf{e}_i\mathbf{e}'_j$

Let $\mathbf{e}_{i,m}$ be the ith column of $\mathbf{I}_m$ and $\mathbf{e}_{j,n}$ be the jth column of $\mathbf{I}_n$. Then $\mathbf{e}_{i,m}\mathbf{e}'_{j,n}$ is an $m \times n$ matrix that has its only nonzero element, a one, in the (i,j)th position. Any matrix can be expressed as a sum of its elements multiplied by these position matrices: for example,

$$\begin{aligned}\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} &= a_{11}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + a_{12}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + a_{21}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + a_{22}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}\mathbf{e}_{i,m}\mathbf{e}'_{j,n}.\end{aligned}$$

In general, for any $m \times n$ matrix $\mathbf{A}$ with (i,j)th element a_{ij},

$$\mathbf{A} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\mathbf{e}_{i,m}\mathbf{e}'_{j,n} \tag{C.11}$$

$\mathbf{A}$ can equivalently be expressed as a sum of cross products of its columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ with the corresponding columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of the $n \times n$ identity matrix (Magnus and Neudecker, 1979):

$$\mathbf{A} = \sum_{j=1}^{n} \mathbf{a}_j\mathbf{e}'_{j,n}, \qquad \text{where} \qquad \mathbf{a}_j = \sum_{i=1}^{m} a_{ij}\mathbf{e}_{i,m} \tag{C.12}$$

To simplify notation, the dimensions of $\mathbf{e}_i$ and $\mathbf{e}_j$ will henceforth be omitted, but are implied by the limits of the respective summations.

Equations (C.11) and (C.12) lead to some useful identities for the Kronecker square, $\mathbf{A} \otimes \mathbf{A}$, and for the cross product of a vectorised matrix with its transpose, $\text{vec}\Big(\mathbf{A}\Big)\,\text{vec}\Big(\mathbf{A}\Big)'$.

C.3.1 Decomposition of the Kronecker square $\mathbf{A} \otimes \mathbf{A}$

Decomposing $\mathbf{A}$ into its elements using (C.11), the Kronecker square can be expressed as a sum of Kronecker products as

$$\mathbf{A} \otimes \mathbf{A} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{A}. \tag{C.13}$$

Using (C.12), (C.8), (C.4) and (C.2), the Kronecker square can also be decomposed into a sum of vectorised matrices, with

$$\begin{aligned}\mathbf{A} \otimes \mathbf{A} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{a}_j\mathbf{e}_j' \otimes \mathbf{a}_i\mathbf{e}_i' \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} (\mathbf{a}_j \otimes \mathbf{a}_i)\left(\mathbf{e}_j' \otimes \mathbf{e}_i'\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} \text{vec}\Big(\mathbf{a}_i\mathbf{a}_j'\Big)\,\text{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)'. \end{aligned} \tag{C.14}$$

C.3.2 Decomposition of the cross product $\mathbf{vec(A)vec(A)'}$

Using (C.11), the cross product of a vectorised matrix $\text{vec}\Big(\mathbf{A}\Big)$ with its transpose can be expressed as

$$\text{vec}\Big(\mathbf{A}\Big)\,\text{vec}\Big(\mathbf{A}\Big)' = \sum_{i=1}^{m}\sum_{j=1}^{n} \text{vec}\Big(\mathbf{A}\Big)\,\text{vec}\Big(a_{ij}\mathbf{e}_i\mathbf{e}_j'\Big)'. \tag{C.15}$$

This can also be expressed as a sum of Kronecker products of column products and position matrices by using (C.12), (C.2) and (C.8).

$$\begin{aligned}\operatorname{vec}\Big(\mathbf{A}\Big)\operatorname{vec}\Big(\mathbf{A}\Big)' &= \sum_{i=1}^{m}\sum_{j=1}^{n}\operatorname{vec}\Big(\mathbf{a}_i\mathbf{e}_i'\Big)\operatorname{vec}\Big(\mathbf{a}_j\mathbf{e}_j'\Big)' \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{e}_i\otimes\mathbf{a}_i\right)\Big(\mathbf{e}_j'\otimes\mathbf{a}_j'\Big) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{e}_i\mathbf{e}_j'\otimes\mathbf{a}_i\mathbf{a}_j'. \end{aligned} \tag{C.16}$$

C.4 The commutation matrix $\mathbf{K}_{m,n}$

Magnus and Neudecker (1979) used a sum of Kronecker products of $\mathbf{e}_i\mathbf{e}_j'$ and its transpose to define the $mn\times mn$ matrix

$$\mathbf{K}_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{e}_i\mathbf{e}_j'\otimes\mathbf{e}_j\mathbf{e}_i'. \tag{C.17}$$

Hence, for example,

$$\mathbf{K}_{2,2} = \begin{bmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{bmatrix}, \qquad \mathbf{K}_{3,3} = \begin{bmatrix}1&0&0&0&0&0&0&0&0\\0&0&0&1&0&0&0&0&0\\0&0&0&0&0&0&1&0&0\\0&1&0&0&0&0&0&0&0\\0&0&0&0&1&0&0&0&0\\0&0&0&0&0&0&0&1&0\\0&0&1&0&0&0&0&0&0\\0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&0&1\end{bmatrix}.$$

The matrix $\mathbf{K}_{m,n}$ is known as a commutation matrix because it provides the factors that allow a Kronecker product of vectors to commute. For vectors $\mathbf{x}$ and $\mathbf{y}$ of length m and n respectively, and $m\times n$ matrix $\mathbf{A}$,

$$\mathbf{K}_{m,n}(\mathbf{x}\otimes\mathbf{y}) = \mathbf{y}\otimes\mathbf{x}, \qquad \mathbf{K}_{m,n}\operatorname{vec}\Big(\mathbf{A}\Big) = \operatorname{vec}\Big(\mathbf{A}'\Big). \tag{C.18}$$

Furthermore, for a matrix $\mathbf{A}$ with m rows, vectors $\mathbf{x}$ of length n and vector $\mathbf{y}$ of arbitrary length,

$$\mathbf{y}' \otimes \mathbf{A} \otimes \mathbf{x} = \mathbf{K}_{m,n} \left(\mathbf{x}\mathbf{y}' \otimes \mathbf{A} \right). \tag{C.19}$$

The commutation matrix satisfies the further properties

$$\mathbf{K}'_{m,n} = \mathbf{K}_{n,m}, \tag{C.20}$$

$$\mathbf{K}^{-1}_{m,n} = \mathbf{K}_{n,m}, \tag{C.21}$$

$$\mathbf{K}_{m,n}\mathbf{K}_{n,m} = \mathbf{I}_{mn} \tag{C.22}$$

C.4.1 Decomposition of the commuted Kronecker square of a symmetric matrix

Several steps in Section 5.3 require permutation of the blocks or columns of an $m \times m$ symmetric matrix $\mathbf{A}$ into an alternative arrangement. In the particular case where $\mathbf{A}$ is symmetric, (C.12) can be used to write

$$\mathbf{K}_{m,m} (\mathbf{A} \otimes \mathbf{A}) = \mathbf{K}_{m,m} \left(\mathbf{A}' \otimes \mathbf{A} \right) = \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{K}_{m,m} \left(\mathbf{a}_j \mathbf{e}'_j \otimes \mathbf{e}_i \mathbf{a}'_i \right).$$

Then, using (C.19) and (C.3),

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{K}_{m,m} \left(\mathbf{a}_j \mathbf{e}'_j \otimes \mathbf{e}_i \mathbf{a}'_i \right) &= \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{e}_i \otimes \mathbf{a}_j \mathbf{e}'_j \otimes \mathbf{a}'_i \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{e}_i \otimes \mathbf{e}'_j \otimes \mathbf{a}_j \otimes \mathbf{a}'_i \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{e}_i \mathbf{e}'_j \otimes \mathbf{a}_j \mathbf{a}'_i. \end{aligned} \tag{C.23}$$

This can also be expressed as a sum of vector cross products by applying (C.18) to (C.14):

$$\begin{aligned}\mathbf{K}_{m,m}(\mathbf{A}\otimes\mathbf{A}) &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{K}_{m,m}\mathrm{vec}\Big(\mathbf{a}_i\mathbf{a}_j'\Big)\,\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)' \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{vec}\Big(\mathbf{a}_j\mathbf{a}_i'\Big)\,\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)'. \end{aligned} \tag{C.24}$$

C.5 The N matrix

The square commutation matrix $\mathbf{K}_{m,m}$ appears in several matrix moment formulas through terms of the form

$$\mathbf{N}_m = \frac{1}{2}\left(\mathbf{I}_{m^2} + \mathbf{K}_{m,m}\right). \tag{C.25}$$

Thus, for example,

$$\mathbf{N}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{N}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using (C.18) it follows that

$$\mathbf{N}_m\mathrm{vec}\Big(\mathbf{A}\Big) = \frac{1}{2}\mathrm{vec}\Big(\mathbf{A}+\mathbf{A}'\Big), \tag{C.26}$$

and so when $\mathbf{A}$ is symmetric,

$$\mathbf{N}_m\mathrm{vec}\Big(\mathbf{A}\Big) = \mathrm{vec}\Big(\mathbf{A}\Big), \tag{C.27}$$

$\mathbf{N}_m$ is singular, and so the inverse $\mathbf{N}_m^{-1}$ does not exist. However, since both

$\mathbf{I}_{m^2}$ and $\mathbf{K}_{m,m}$ are symmetric and real-valued, $\mathbf{N}_m$ also shares these properties; furthermore, it is straightforward to show that $\mathbf{N}_m$ is idempotent using (C.25) and (C.22), since

$$\begin{aligned}\mathbf{N}_m\mathbf{N}_m &= \frac{1}{4}\left(\mathbf{I}_{m^2}+\mathbf{K}_{m,m}\right)\left(\mathbf{I}_{m^2}+\mathbf{K}_{m,m}\right)\\ &= \frac{1}{4}\left(\mathbf{I}_{m^2}\mathbf{I}_{m^2}+\mathbf{K}_{m,m}\mathbf{I}_{m^2}+\mathbf{I}_{m^2}\mathbf{K}_{m,m}+\mathbf{K}_{m,m}\mathbf{K}_{m,m}\right)\\ &= \frac{1}{4}\left(2\mathbf{I}_{m^2}+2\mathbf{K}_{m,m}\right)\\ &= \mathbf{N}_m. \end{aligned} \tag{C.28}$$

Being idempotent, symmetric and real-valued, $\mathbf{N}_m$ satisfies the four Moore-Penrose criteria (Penrose, 1955), thus the generalised inverse of $\mathbf{N}_m$ is $\mathbf{N}_m^{\dagger} = \mathbf{N}_m$.

The $\mathbf{N}$ matrix enables a number of useful manipulations of Kronecker products. From Schott (2016, Theorem 8.31), for $\mathbf{A} = \mathbf{B}\mathbf{B}'$ where $\mathbf{A}$ and $\mathbf{B}$ are both $m \times m$ matrices,

$$\mathbf{N}_m\left(\mathbf{B}\otimes\mathbf{B}\right)\mathbf{N}_m = \left(\mathbf{B}\otimes\mathbf{B}\right)\mathbf{N}_m = \mathbf{N}_m\left(\mathbf{B}\otimes\mathbf{B}\right), \tag{C.29}$$

$$\left(\mathbf{B}\otimes\mathbf{B}\right)\mathbf{N}_m\left(\mathbf{B}'\otimes\mathbf{B}'\right) = \mathbf{N}_m\left(\mathbf{A}\otimes\mathbf{A}\right). \tag{C.30}$$

C.5.1 Positive semi-definiteness of linear combinations of vec and Kronecker products of symmetric matrices

Section 5.3 will make extensive use of linear combinations of terms of the form $\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}\right)'$ and $\mathbf{N}_m\left(\mathbf{A}\otimes\mathbf{A}\right)$, where $\mathbf{A}$ is a symmetric $m \times m$ positive semidefinite matrix, to define the $m^2 \times m^2$ covariance matrices used in multivariate Bayes linear variance adjustment. In order for the matrices so defined to be valid covariance matrices, they must be both symmetric and positive semidefinite.

For any matrix $\mathbf{A}$, $\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}\right)'$ is symmetric positive semidefinite. To

see this, notice that each column is a multiple of $\mathrm{vec}\Big(\mathbf{A}\Big)$, and so $\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'$ has rank one and has only one non-zero eigenvalue: when $\mathbf{A}$ is an $m \times m$ matrix, this is equal to $\mathrm{tr}\left(\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\right) = \sum_{i=1}^{m} a_{ii}^2$.

From Schott (2016, Theorem 8.5),the m^2 eigenvalues of $\mathbf{A} \otimes \mathbf{A}$ are $\{\lambda_i \lambda_j : i = 1, \ldots, m; j = 1, \ldots, m\}$, where λ_i denotes the ith eigenvalue of $\mathbf{A}$; hence, if $\mathbf{A}$ is positive semidefinite, $\mathbf{A} \otimes \mathbf{A}$ must be also. Furthermore, since $\mathbf{N}_m$ is idempotent its eigenvalues are all either zero or one (Horn and Johnson, 1985); thus $\mathbf{N}_m$ is also positive semidefinite. From (C.29), $\mathbf{N}_m\,(\mathbf{A} \otimes \mathbf{A}) = (\mathbf{A} \otimes \mathbf{A})\,\mathbf{N}_m$; thus $\mathbf{N}_m\,(\mathbf{A} \otimes \mathbf{A})$ must also be symmetric and positive semidefinite. Any matrix constructed from a nonnegative linear combination of $\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'$ and $\mathbf{N}_m\,(\mathbf{A} \otimes \mathbf{A})$ must therefore also be both symmetric and positive semidefinite, if $\mathbf{A}$ is positive semidefinite.

Such matrices are rank deficient, and so the inverse does not exist. However, due to the special structure of these matrices, it can be shown that when $\mathbf{A}$ is nonsingular, the generalised inverse can be written in terms of $\mathbf{A}^{-1}$, and has the general form

$$\begin{aligned}\Big[\mathbf{N}_m\,(\mathbf{A} \otimes \mathbf{A}) + \alpha \mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\Big]^{\dagger} = \\ \mathbf{N}_m \left\{\mathbf{A}^{-1} \otimes \mathbf{A}^{-1} - \frac{\alpha}{1 + \alpha m}\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'\right\}. \qquad \text{(C.31)}\end{aligned}$$

The following proof of this statement is adapted from Magnus (1988), Theorem 4.16.

C.5.2 Derivation of the generalised inverse of $\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A})+\alpha\mathbf{vec}\big(\mathbf{A}\big)\,\mathbf{vec}\big(\mathbf{A}\big)'$

From (C.6) and (C.8), $(\mathbf{A}\otimes\mathbf{A})^\dagger(\mathbf{A}\otimes\mathbf{A})=\mathbf{I}_{m^2}$, so that

$$\begin{aligned}
&\Big[\mathbf{N}_m\,(\mathbf{A}\otimes\mathbf{A})+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\Big]\\
&\quad=\Big[\mathbf{N}_m\,(\mathbf{A}\otimes\mathbf{A})+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\Big]\,(\mathbf{A}\otimes\mathbf{A})^\dagger\,(\mathbf{A}\otimes\mathbf{A})\,,\\
&\quad=\Big[\mathbf{N}_m\,(\mathbf{A}\otimes\mathbf{A})\,(\mathbf{A}\otimes\mathbf{A})^\dagger+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\,(\mathbf{A}\otimes\mathbf{A})^\dagger\Big]\,(\mathbf{A}\otimes\mathbf{A})\,;\\
&\quad=\Big[\mathbf{N}_m+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}\Big)'\,\Big(\mathbf{A}^\dagger\otimes\mathbf{A}^\dagger\Big)\Big]\,(\mathbf{A}\otimes\mathbf{A})\,. \qquad \text{(C.32)}
\end{aligned}$$

Because $\mathbf{A}$ is invertible, $\mathbf{A}^\dagger=\mathbf{A}^{-1}$, and from (C.9), $\mathrm{vec}\Big(\mathbf{A}\Big)'\Big(\mathbf{A}^{-1}\otimes\mathbf{A}^{-1}\Big)=\mathrm{vec}\Big(\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\Big)'=\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'$, so (C.32) is

$$=\Big[\mathbf{N}_m+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'\Big]\,(\mathbf{A}\otimes\mathbf{A})\,. \qquad \text{(C.33)}$$

Finally, using (C.29) and (C.27), this is equivalent to

$$=\Big[\mathbf{I}_{m^2}+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'\Big]\,\mathbf{N}_m\,(\mathbf{A}\otimes\mathbf{A})\,. \qquad \text{(C.34)}$$

The required generalised inverse is therefore now

$$\Big\{\Big[\mathbf{I}_{m^2}+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'\Big]\,\mathbf{N}_m\,(\mathbf{A}\otimes\mathbf{A})\Big\}^\dagger, \qquad \text{(C.35)}$$

which has the form $(\mathbf{PQR})^\dagger$, with $\mathbf{P}=\mathbf{I}_{m^2}+\alpha\mathrm{vec}\Big(\mathbf{A}\Big)\,\mathrm{vec}\Big(\mathbf{A}^{-1}\Big)'$, $\mathbf{Q}=\mathbf{N}_m$, and $\mathbf{R}=(\mathbf{A}\otimes\mathbf{A})$. It will now be shown that, in this case, $(\mathbf{PQR})^\dagger=\mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger$.

First, it is necessary to obtain $\mathbf{P}^\dagger$, $\mathbf{Q}^\dagger$ and $\mathbf{R}^\dagger$. Note that $\mathbf{P}$ has the form $\mathbf{B}+\mathbf{uv}'$. According to the Sherman-Morrison formula (Bartlett, 1951), a matrix of this form is invertible if $1+\mathbf{v}'\mathbf{B}^{-1}\mathbf{u}\neq 0$, in which case

$$\Big(\mathbf{B}+\mathbf{uv}'\Big)^{-1}=\mathbf{B}^{-1}-\frac{\mathbf{B}^{-1}\mathbf{uv}'\mathbf{B}^{-1}}{1+\mathbf{v}'\mathbf{B}^{-1}\mathbf{u}}. \qquad \text{(C.36)}$$

Here, $\mathbf{B} = \mathbf{I}_{m^2}$, and from (C.10), $1 + \mathbf{v}'\mathbf{B}^{-1}\mathbf{u} = 1 + \text{vec}\left(\mathbf{A}^{-1}\right)' \text{vec}\left(\mathbf{A}\right) = 1 + \text{tr}\left(\mathbf{A}^{-1}\mathbf{A}\right) = 1 + m$, so $\mathbf{P}$ is invertible, with

$$\mathbf{P}^{\dagger} = \mathbf{P}^{-1} = \mathbf{I}_{m^2} - \frac{\alpha \text{vec}\left(\mathbf{A}\right) \text{vec}\left(\mathbf{A}^{-1}\right)'}{1 + \alpha m}. \tag{C.37}$$

From Section C.5, $\mathbf{Q}^{\dagger} = \mathbf{N}_m^{\dagger} = \mathbf{N}_m$, and using (C.6), $\mathbf{R}^{\dagger} = (\mathbf{A} \otimes \mathbf{A})^{\dagger} = (\mathbf{A}^{\dagger} \otimes \mathbf{A}^{\dagger}) = (\mathbf{A}^{-1} \otimes \mathbf{A}^{-1})$.

C.5.2.1 Satisfying the Penrose conditions

Having obtained the generalised inverses of $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$, it is now necessary to show that the generalised inverse of $\mathbf{PQR}$ also exists. The matrix $\mathbf{M}^{\dagger} = (\mathbf{PQR})^{\dagger}$ is a pseudoinverse (and therefore a generalised inverse) of the matrix $\mathbf{M} = \mathbf{PQR}$) if it satisfies the four Penrose conditions (Rao, 1972):

1. $\mathbf{M}\mathbf{M}^{\dagger}\mathbf{M} = \mathbf{M}$,
2. $\mathbf{M}^{\dagger}\mathbf{M}\mathbf{M}^{\dagger} = \mathbf{M}^{\dagger}$,
3. $\left(\mathbf{M}\mathbf{M}^{\dagger}\right)^{*} = \mathbf{M}\mathbf{M}^{\dagger}$,
4. $\left(\mathbf{M}^{\dagger}\mathbf{M}\right)^{*} = \mathbf{M}^{\dagger}\mathbf{M}$,

where $\mathbf{M}^{*}$ denotes the conjugate transpose of $\mathbf{M}$; here, $\mathbf{M} = \mathbf{PQR}$, where all of the matrices concerned are real-valued, so $\mathbf{M}^{*}$ is the transpose $\mathbf{M}'$.

First, consider condition 4. Recall that $\mathbf{P}$ is invertible, so that $\mathbf{P}\mathbf{P}^{-1} = \mathbf{I}_{m^2}$. Similarly, as noted earlier in this proof, $(\mathbf{A} \otimes \mathbf{A})^{\dagger}(\mathbf{A} \otimes \mathbf{A}) = \mathbf{I}_{m^2}$, so that $\mathbf{R}\mathbf{R}^{\dagger} = \mathbf{I}_{m^2}$. $\mathbf{N}_m$ is symmetric and idempotent (Section C.5), so that $\mathbf{Q}\mathbf{Q}^{\dagger} = \mathbf{N}_m\mathbf{N}_m = \mathbf{N}_m$.

Then, using (C.29) and (C.8),

$$\begin{aligned}\mathbf{M}^\dagger\mathbf{M} = \mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger\mathbf{P}\mathbf{Q}\mathbf{R} &= \left(\mathbf{A}^{-1}\otimes\mathbf{A}^{-1}\right)\mathbf{N}_m^\dagger\mathbf{P}^{-1}\mathbf{P}\mathbf{N}_m\left(\mathbf{A}\otimes\mathbf{A}\right)\\ &= \left(\mathbf{A}\otimes\mathbf{A}\right)^\dagger\mathbf{N}_m^\dagger\mathbf{N}_m\left(\mathbf{A}\otimes\mathbf{A}\right)\\ &= \left(\mathbf{A}\otimes\mathbf{A}\right)^\dagger\left(\mathbf{A}\otimes\mathbf{A}\right)\mathbf{N}_m\\ &= \mathbf{N}_m. \end{aligned} \tag{C.38}$$

$\mathbf{N}_m$ is symmetric, so $\mathbf{M}^\dagger\mathbf{M} = \left(\mathbf{M}^\dagger\mathbf{M}\right)'$, and the fourth condition is satisfied.

Similarly, for the third condition,

$$\begin{aligned}\mathbf{M}\mathbf{M}^\dagger = \mathbf{P}\mathbf{Q}\mathbf{R}\mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger &= \mathbf{P}\mathbf{N}_m\left(\mathbf{A}\otimes\mathbf{A}\right)\left(\mathbf{A}\otimes\mathbf{A}\right)^\dagger\mathbf{N}_m^\dagger\mathbf{P}^{-1}\\ &= \mathbf{P}\mathbf{N}_m\mathbf{P}^{-1}. \end{aligned} \tag{C.39}$$

Using (C.27), which states that for symmetric $\mathbf{A}$, $\mathbf{N}_m\text{vec}\left(\mathbf{A}\right) = \text{vec}\left(\mathbf{A}\right)$ and $\text{vec}\left(\mathbf{A}\right)' = \text{vec}\left(\mathbf{A}\right)\mathbf{N}_m$, it is possible to write

$$\begin{aligned}\mathbf{P}\mathbf{N}_m &= \left[\mathbf{I}_{m^2} + \alpha\text{vec}\left(\mathbf{A}\right)\text{vec}\left(\mathbf{A}^{-1}\right)'\right]\mathbf{N}_m\\ &= \left[\mathbf{N}_m + \mathbf{N}_m\alpha\text{vec}\left(\mathbf{A}\right)\text{vec}\left(\mathbf{A}^{-1}\right)'\mathbf{N}_m\right]\\ &= \mathbf{N}_m\left[\mathbf{I}_{m^2} + \alpha\text{vec}\left(\mathbf{A}\right)\text{vec}\left(\mathbf{A}^{-1}\right)'\right] = \mathbf{N}_m\mathbf{P}, \end{aligned} \tag{C.40}$$

so that

$$\mathbf{M}\mathbf{M}^\dagger = \mathbf{N}_m\mathbf{P}\mathbf{P}^{-1} = \mathbf{N}_m. \tag{C.41}$$

Again, because $\mathbf{N}_m$ is symmetric, $\mathbf{M}\mathbf{M}^\dagger = \left(\mathbf{M}\mathbf{M}^\dagger\right)'$, and the third condition is satisfied.

Now using (C.41) in condition 2, and this time using (C.27) to write

$\mathbf{N}_m\mathbf{P}^{-1}\mathbf{N}_m = \mathbf{N}_m\mathbf{P}^{-1}$,

$$\begin{aligned}\mathbf{M}^\dagger\left(\mathbf{M}\mathbf{M}^\dagger\right) = \mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger\mathbf{N}_m &= (\mathbf{A}\otimes\mathbf{A})^\dagger\mathbf{N}_m^\dagger\mathbf{P}^{-1}\mathbf{N}_m \\ &= (\mathbf{A}\otimes\mathbf{A})^\dagger\mathbf{N}_m^\dagger\mathbf{P}^{-1} \qquad = \mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger \\ &\qquad\qquad\qquad\qquad\quad = \mathbf{M}^\dagger. \qquad \text{(C.42)}\end{aligned}$$

Finally, applying the same approach to condition 1 gives

$$\begin{aligned}\left(\mathbf{M}\mathbf{M}^\dagger\right)\mathbf{M} = \mathbf{N}_m\mathbf{P}\mathbf{Q}\mathbf{R} = \mathbf{N}_m\mathbf{P}\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A}) \qquad &= \mathbf{P}\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A}) \\ &= \mathbf{M}. \qquad \text{(C.43)}\end{aligned}$$

All four Penrose conditions are therefore satisfied, proving that the generalised inverse of $\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A}) + \alpha\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}\right)'$ is

$$\mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{P}^\dagger = (\mathbf{A}\otimes\mathbf{A})^{-1}\mathbf{N}_m\left[\mathbf{I}_{m^2} + \alpha\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}^{-1}\right)'\right]^{-1}. \qquad \text{(C.44)}$$

Using (C.37), and (C.29), this is

$$\begin{aligned}&(\mathbf{A}\otimes\mathbf{A})^{-1}\mathbf{N}_m\left[\mathbf{I}_{m^2} - \frac{\alpha\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}^{-1}\right)'}{1+\alpha m}\right] \\ &\quad = \mathbf{N}_m\left[(\mathbf{A}\otimes\mathbf{A})^{-1} - \frac{\alpha(\mathbf{A}\otimes\mathbf{A})^{-1}\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}^{-1}\right)'}{1+\alpha m}\right]. \qquad \text{(C.45)}\end{aligned}$$

Finally, using (C.5) and (C.9), $(\mathbf{A}\otimes\mathbf{A})^{-1}\mathrm{vec}\left(\mathbf{A}\right) = \left(\mathbf{A}^{-1}\otimes\mathbf{A}^{-1}\right)\mathrm{vec}\left(\mathbf{A}\right) = \mathrm{vec}\left(\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\right) = \mathrm{vec}\left(\mathbf{A}^{-1}\right)$, and the generalised inverse can be written as

$$\begin{aligned}&\left[\mathbf{N}_m(\mathbf{A}\otimes\mathbf{A}) + \alpha\mathrm{vec}\left(\mathbf{A}\right)\mathrm{vec}\left(\mathbf{A}\right)'\right]^\dagger = \\ &\qquad \mathbf{N}_m\left\{\mathbf{A}^{-1}\otimes\mathbf{A}^{-1} - \frac{\alpha\mathrm{vec}\left(\mathbf{A}^{-1}\right)\mathrm{vec}\left(\mathbf{A}^{-1}\right)'}{1+\alpha m}\right\}, \qquad \text{(C.46)}\end{aligned}$$

as required.

Appendix D

Derivations relating to multivariate Bayes linear variance adjustment

D.1 Permuting $\mathbf{V}_M$ into $\mathbf{V}_M{}^*$

In Section 5.3.4.1 the permuted variance $\mathbf{V}_M{}^*$ was introduced, in order to express the expectation $\mathbb{E}[\mathcal{M}(\mathbf{V}) \otimes \mathcal{M}(\mathbf{V})]$ in terms of specifiable quantities. The exact form of the permutation is derived here. Recall that the unpermuted variance is $\mathbf{V}_M = \mathbb{V}[\mathcal{M}(\mathbf{V})]$, and that $\mathbf{V}_R = \mathbb{E}[\mathcal{M}(\mathbf{V})]$.

D.1.1 Permutation of submatrices of $\mathbf{V}_M$

Let $\mathbf{v}_i$ denote the ith column of $\mathcal{M}(\mathbf{V})$ and let $\mathbf{e}_i$ denote the ith column of $\mathbf{I}_m$, as defined in Section C.3. Then using (C.16),

$$\begin{aligned}\mathbb{E}\Big[\mathrm{vec}\big(\mathcal{M}(\mathbf{V})\big)\,\mathrm{vec}\big(\mathcal{M}(\mathbf{V})\big)'\Big] &= \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{v}_i\mathbf{v}_i'\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbb{E}\Big[\mathbf{v}_i\mathbf{v}_i'\Big], \qquad \text{(D.1)}\end{aligned}$$

while using (C.14),

$$\begin{aligned}\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})] &= \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\operatorname{vec}\left(\mathbf{v}_i\mathbf{v}_i'\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)'\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\operatorname{vec}\left(\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_i'\right]\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)'. \end{aligned} \tag{D.2}$$

Now let $\mathbf{M}_{ij}$ denote the (i,j)th $m\times m$ block of $\mathbf{V}_M$, so that

$$\mathbf{V}_M = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{e}_i\mathbf{e}_j'\otimes\mathbf{M}_{ij}, \tag{D.3}$$

and let $\mathbf{r}_i$ denote the ith column of $\mathbf{V}_R$, so that the (i,j)th block of $\mathbb{E}\left[\operatorname{vec}\left(\mathcal{M}(\mathbf{V})\right)\operatorname{vec}\left(\mathcal{M}(\mathbf{V})\right)'\right]$ can be written as

$$\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_i'\right] = \mathbf{M}_{ij} + \mathbf{r}_i\mathbf{r}_j'. \tag{D.4}$$

Substituting (D.4) into (D.2) and again using (C.14), $\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})]$ can be written as

$$\begin{aligned}\mathbb{E}[\mathcal{M}(\mathbf{V})\otimes\mathcal{M}(\mathbf{V})] &= \sum_{i=1}^{m}\sum_{j=1}^{m}\operatorname{vec}\left(\mathbf{M}_{ij}+\mathbf{r}_i\mathbf{r}_j'\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)' \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\left\{\operatorname{vec}\left(\mathbf{M}_{ij}\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)' + \operatorname{vec}\left(\mathbf{r}_i\mathbf{r}_j'\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)'\right\} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\operatorname{vec}\left(\mathbf{M}_{ij}\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)' + \mathbf{V}_R\otimes\mathbf{V}_R \\ &= \mathbf{V}_M{}^* + (\mathbf{V}_R\otimes\mathbf{V}_R), \end{aligned} \tag{D.5}$$

where

$$\mathbf{V}_M{}^* = \sum_{i=1}^{m}\sum_{j=1}^{m}\operatorname{vec}\left(\mathbf{M}_{ij}\right)\operatorname{vec}\left(\mathbf{e}_i\mathbf{e}_j'\right)'. \tag{D.6}$$

D.1.2 Expressing $\mathbf{V}_M{}^*$ in terms of $\mathbf{V}_R$

With $\mathbf{V}_M$ expressed in terms of $\mathbf{V}_R$ as in Section 5.3.5.2, it is possible to derive an explicit form for $\mathbf{M}_{ij}$, the (i,j)th $m \times m$ block of $\mathbf{V}_M$, and so to derive an explicit form for $\mathbf{V}_M{}^*$, the permutation of $\mathbf{V}_M$ required to complete the specification of $\mathbf{V}_T$ in (5.119). The derivation is carried out using the more compact general elliptical notation used in (5.128).

Using (C.16)-(C.23), (5.128) can be written as

$$\begin{aligned}\mathbf{V}_M &= \frac{1}{\beta}\Big\{\gamma \mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big)' + (\gamma+1)(\mathbf{V}_R \otimes \mathbf{V}_R) + (\gamma+1)\mathbf{K}_{m,m}(\mathbf{V}_R \otimes \mathbf{V}_R)\Big\} \\ &= \frac{1}{\beta}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big\{\gamma\left(\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{r}_i\mathbf{r}_j'\right) + (\gamma+1)\left(r_{ij}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{V}_R\right) + (\gamma+1)\left(\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{r}_j\mathbf{r}_i'\right)\Big\} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{e}_i\mathbf{e}_j' \otimes \frac{1}{\beta}\Big\{\gamma\mathbf{r}_i\mathbf{r}_j' + (\gamma+1)r_{ij}\mathbf{V}_R + (\gamma+1)\mathbf{r}_j\mathbf{r}_i'\Big\}, \end{aligned} \tag{D.7}$$

where r_{ij} denotes the (i,j)th element of $\mathbf{V}_R$ and $\mathbf{r}_i$ denotes the ith column of $\mathbf{V}_R$. This has the form $\mathbf{V}_M = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{M}_{ij}$ that was obtained in (D.3), so it follows that

$$\mathbf{M}_{ij} = \frac{1}{\beta}\Big\{\gamma\mathbf{r}_i\mathbf{r}_j' + (\gamma+1)r_{ij}\mathbf{V}_R + (\gamma+1)\mathbf{r}_j\mathbf{r}_i'\Big\}. \tag{D.8}$$

As noted in Section 5.3.1, $\mathbf{V}_R$ is assumed to be a valid covariance matrix and therefore of full rank; therefore each column $\mathbf{r}_i$ is unique, and so each block $\mathbf{M}_{ij}$ has a unique value for any given $\mathbf{V}_M$. Now substituting these blocks into the permuted form (D.6) and using (C.15)-(C.24), $\mathbf{V}_M{}^*$ can be expressed in terms of specifiable quantities as

$$\begin{aligned}\mathbf{V}_M{}^* &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{vec}\left(\frac{1}{\beta}\Big\{\gamma\mathbf{r}_i\mathbf{r}_j' + (\gamma+1)r_{ij}\mathbf{V}_R + (\gamma+1)\mathbf{r}_j\mathbf{r}_i'\Big\}\right)\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)' \\ &= \frac{1}{\beta}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big\{\gamma\mathrm{vec}\Big(\mathbf{r}_i\mathbf{r}_j'\Big)\,\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)' + (\gamma+1)\mathrm{vec}\Big(r_{ij}\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)' + \\ &\qquad (\gamma+1)\mathrm{vec}\Big(\mathbf{r}_j\mathbf{r}_i'\Big)\,\mathrm{vec}\Big(\mathbf{e}_i\mathbf{e}_j'\Big)'\Big\}\end{aligned}$$

$$= \frac{1}{\beta}\left\{\gamma\left(\mathbf{V}_R \otimes \mathbf{V}_R\right) + (\gamma+1)\mathrm{vec}\Big(\mathbf{V}_R\Big)\,\mathrm{vec}\Big(\mathbf{V}_R\Big) + (\gamma+1)\mathbf{K}_{m,m}\left(\mathbf{V}_R \otimes \mathbf{V}_R\right)\right\}. \tag{D.9}$$

D.2 Fourth-order moments of a multivariate elliptical distribution

The fourth-order central moments (5.124) of an $m \times 1$ random variable $\mathbf{X} \sim E(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \psi)$ can be expressed in matrix form using the definition of the Kronecker product given in (C.1). The covariances here are no longer between individuals i and j, say, so the subscript i is omitted here; instead, p and q refer to the elements of $\mathbf{X}$. Also let $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$, so that $\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' \otimes (\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})']$ can be expressed more compactly as $\mathbb{E}\Big[\mathbf{Y}\mathbf{Y}' \otimes \mathbf{Y}\mathbf{Y}'\Big]$.

From (C.12), $\mathbf{Y} = \sum_{p=1}^{m} Y_p \mathbf{e}_p$. Then by gathering the scalar elements over which the expectation is taken,

$$\begin{aligned}\mathbb{E}\Big[\mathbf{Y}\mathbf{Y}' \otimes \mathbf{Y}\mathbf{Y}'\Big] &= \mathbb{E}\left[\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}(Y_p\mathbf{e}_p)(Y_q\mathbf{e}_q)' \otimes (Y_r\mathbf{e}_r)(Y_s\mathbf{e}_s)'\right] \\ &= \sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathbf{e}_p\mathbf{e}_q' \otimes \mathbf{e}_r\mathbf{e}_s'\,\mathbb{E}[Y_pY_qY_rY_s], \end{aligned} \tag{D.10}$$

and, taking $\mathbb{E}[Y_pY_qY_rY_s]$ from (5.124),

$$\mathbb{E}\Big[\mathbf{Y}\mathbf{Y}' \otimes \mathbf{Y}\mathbf{Y}'\Big] = \alpha^2(\kappa+1)\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathbf{e}_p\mathbf{e}_q' \otimes \mathbf{e}_r\mathbf{e}_s'\left[\sigma_{pq}\sigma_{rs} + \sigma_{pr}\sigma_{qs} + \sigma_{ps}\sigma_{qr}\right]. \tag{D.11}$$

This summation can be carried out separately for each element in the square brackets.

First, using (C.11),

$$\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\mathbf{e}_r\mathbf{e}_s'\,\sigma_{pq}\sigma_{rs}=\sum_{p=1}^{m}\sum_{q=1}^{m}\sigma_{pq}\mathbf{e}_p\mathbf{e}_q'\otimes\sum_{r=1}^{m}\sum_{s=1}^{m}\sigma_{rs}\mathbf{e}_r\mathbf{e}_s'$$
$$=\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}. \tag{D.12}$$

For the second term, (C.12) is used to to sum over r and s, then using (C.16),

$$\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\mathbf{e}_r\mathbf{e}_s'\sigma_{pr}\sigma_{qs}=\sum_{p=1}^{m}\sum_{q=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\sum_{r=1}^{m}\sum_{s=1}^{m}\sigma_{pr}\mathbf{e}_r\sigma_{qs}\mathbf{e}_s'$$
$$=\sum_{p=1}^{m}\sum_{q=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\boldsymbol{\sigma}_p\boldsymbol{\sigma}_q'$$
$$=\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)', \tag{D.13}$$

where $\boldsymbol{\sigma}_p$ denotes the pth column of $\boldsymbol{\Sigma}$.

Finally, (C.12) is used again to rearrange the third term, this time as

$$\sum_{p=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\mathbf{e}_r\mathbf{e}_s'\sigma_{ps}\sigma_{qr}=\sum_{p=1}^{m}\sum_{q=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\sum_{r=1}^{m}\sum_{s=1}^{m}\sigma_{qr}\mathbf{e}_r\sigma_{ps}\mathbf{e}_s'$$
$$=\sum_{p=1}^{m}\sum_{q=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\boldsymbol{\sigma}_q\boldsymbol{\sigma}_p'. \tag{D.14}$$

Then, using (C.23),

$$\sum_{p=1}^{m}\sum_{q=1}^{m}\mathbf{e}_p\mathbf{e}_q'\otimes\boldsymbol{\sigma}_q\boldsymbol{\sigma}_p'=\mathbf{K}_{m,m}\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right). \tag{D.15}$$

Combining (D.12), (D.13) and (D.15), the matrix of fourth-order central moments can be expressed as

$$\mathbb{E}\Big[\mathbf{Y}\mathbf{Y}'\otimes\mathbf{Y}\mathbf{Y}'\Big]=\alpha^2(\kappa+1)\Big\{\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)'+\left(\mathbf{I}_{m^2}+\mathbf{K}_{m,m}\right)\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\Big\}$$
$$=\alpha^2(\kappa+1)\Big\{\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)\operatorname{vec}\Big(\boldsymbol{\Sigma}\Big)'+2\mathbf{N}_m\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\Big\}. \tag{D.16}$$

D.2.1 Variance of the second-order moments of a multivariate elliptical distribution

$\mathbb{E}\left[\mathbf{Y}\mathbf{Y}'\otimes\mathbf{Y}\mathbf{Y}'\right]$ can be used to obtain $\mathbb{V}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\right]$ because, using (C.8) and (C.2),

$$\mathbb{E}\left[\mathbf{Y}\mathbf{Y}'\otimes\mathbf{Y}\mathbf{Y}'\right]=\mathbb{E}\left[\left(\mathbf{Y}\otimes\mathbf{Y}\right)\left(\mathbf{Y}\otimes\mathbf{Y}\right)'\right]=\mathbb{E}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)'\right], \quad \text{(D.17)}$$

and so, since $\mathbb{E}[\mathbf{Y}]=\mathbf{0}$,

$$\begin{aligned}\mathbb{V}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\right]&=\mathbb{E}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)'\right]-\mathbb{E}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\right]\mathbb{E}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\right]'\\&=\mathbb{E}\left[\mathbf{Y}\mathbf{Y}'\otimes\mathbf{Y}\mathbf{Y}'\right]-\text{vec}\left(\mathbb{V}[\mathbf{Y}]\right)\text{vec}\left(\mathbb{V}[\mathbf{Y}]\right)'. \quad \text{(D.18)}\end{aligned}$$

From property 2 in Section 5.3.5.1, $\mathbb{V}[\mathbf{Y}]=\alpha\boldsymbol{\Sigma}$, so

$$\begin{aligned}\mathbb{V}\left[\text{vec}\left(\mathbf{Y}\mathbf{Y}'\right)\right]&=\alpha^2(\kappa+1)\left\{\text{vec}\left(\boldsymbol{\Sigma}\right)\text{vec}\left(\boldsymbol{\Sigma}\right)'+2\mathbf{N}_m\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\right\}-\text{vec}\left(\alpha\boldsymbol{\Sigma}\right)\text{vec}\left(\alpha\boldsymbol{\Sigma}\right)'\\&=\alpha^2\left\{\kappa\text{vec}\left(\boldsymbol{\Sigma}\right)\text{vec}\left(\boldsymbol{\Sigma}\right)'+2(\kappa+1)\mathbf{N}_m\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\right\}\\&=\kappa\text{vec}\left(\mathbb{V}[\mathbf{Y}]\right)\text{vec}\left(\mathbb{V}[\mathbf{Y}]\right)'+2(\kappa+1)\mathbf{N}_m\left(\mathbb{V}[\mathbf{Y}]\otimes\mathbb{V}[\mathbf{Y}]\right). \quad \text{(D.19)}\end{aligned}$$

Bibliography

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, 10(1):91–105.

Allen, S., Ferro, C., and Kwasniok, F. (2020). Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 146(731):2576–2596.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530.

Baker, L., Rudd, A., Migliorini, S., and Bannister, R. (2014). Representation of model error in a convective-scale ensemble prediction system. *Nonlinear Processes in Geophysics*, 21:19–39.

Bannister, R. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):607–633.

Barnes, C., Brierley, C. M., and Chandler, R. E. (2019). New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(725):3479–3498.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111.

Beck, C., Philipp, A., and Streicher, F. (2016). The effect of domain size on the relationship between circulation type classifications and surface climate. *International Journal of Climatology*, 36(7):2692–2709.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika*, 48(4):493–517.

Bentzien, S. and Friederichs, P. (2012). Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting*, 27(4):988–1002.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690.

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.

Bonavita, M., Isaksen, L., and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. *Quarterly journal of the royal meteorological society*, 138(667):1540–1559.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., et al. (2010). The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072.

Bower, R. G., Goldstein, M., and Vernon, I. (2010). Galaxy formation: a Bayesian uncertainty analysis. *Bayesian analysis*, 5(4):619–669.

Bowler, N., Clayton, A., Jardak, M., Lee, E., Jermey, P., Lorenc, A., Piccolo, C., Pring, S., Wlasak, M., Barker, D., et al. (2016). A 4d-ensemble-variational system for data assimilation and ensemble initialization. In *EGU General Assembly Conference Abstracts*, pages EPSC2016–17140.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E. (2008). The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134(632):703–722.

Bowler, N. E. and Mylne, K. R. (2009). Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(640):757–766.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society*, 143(4):383–430.

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*, 18:1–5.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: the importance of being proper. *Weather and Forecasting*, 22(2):382–388.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chandler, R. E. (2013). Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371:20120388.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. (2004). The Schaake shuffle: a method for reconstructing space-time variability in

forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman & Hall, London.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27(1):65–81.

De Finetti, B. (1992). Foresight: its logical laws, its subjective sources. In *Breakthroughs in Statistics*, pages 134–174. Springer.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10):3498–3516.

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., and Stull, R. (2011). Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139(11):3554–3570.

Denholm-Price, J. (2003). Can an ensemble give anything more than Gaussian probabilities? *Nonlinear Processes in Geophysics*, 10(6):469–475.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

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

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 20(3):328–350.

Eckel, F. A. and Walters, M. K. (1998). Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting*, 13(4):1132–1147.

ECMWF (2021a). `https://confluence.ecmwf.int/display/TIGGE/Model+upgrades`. *ECMWF wiki (accessed December 2021)*.

ECMWF (2021b). `https://confluence.ecmwf.int/display/TIGGE/Model+upgrades#Modelupgrades-ECMWF`. *ECMWF wiki (accessed December 2021)*.

ECMWF (2021c). `https://confluence.ecmwf.int/display/TIGGE/Model+upgrades#Modelupgrades-UKMO`. *ECMWF wiki (accessed December 2021)*.

ECMWF (2021d). `https://confluence.ecmwf.int/display/TIGGE/Model+upgrades#NCEP`. *ECMWF wiki (accessed December 2021)*.

ECMWF (2021e). `https://confluence.ecmwf.int/display/UDOC/MARS+data+unavailability+in+ECMWF+tape+library?type=dataset&val=tigge`. *ECMWF wiki (accessed December 2021)*.

Epstein, E. S. (1969). The role of initial uncertainties in prediction. *Journal of Applied Meteorology and Climatology*, 8(2):190–198.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, 143(3):955–971.

Ferranti, L., Corti, S., and Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141(688):916–924.

Fisher, R. A. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 130(812):16–28.

Fraley, C., Raftery, A. E., and Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138(1):190–202.

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. Technical report, DTIC Document.

Fritsch, J., Hilliker, J., Ross, J., and Vislocky, R. (2000). Model consensus. *Weather and Forecasting*, 15(5):571–582.

Garcia-Moya, J. A., Casado, J., Marco, I., Fernández-Peruchena, C. M., and Gastón, M. (2016). *Deterministic and probabilistic weather forecasting*. PhD thesis, AEMET.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Glahn, B., Gilbert, K., Cosgrove, R., Ruth, D. P., and Sheets, K. (2009). The gridding of MOS. *Weather and Forecasting*, 24(2):520–529.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11(8):1203–1211.

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(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211–235.

Goldstein, M. (1981). Revising previsions: a geometric interpretation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):105–120.

Goldstein, M. and Wilkinson, D. J. (2001). Restricted prior inference for complex uncertainty structures. *Annals of Mathematics and Artificial Intelligence*, 32(1):315–334.

Goldstein, M. and Wooff, D. (2007). *Bayes Linear Statistics: theory and methods*, volume 716. John Wiley & Sons.

Gosling, J. P., Hart, A., Owen, H., Davies, M., Li, J., and MacKay, C. (2013). A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. *Bayesian Analysis*, 8(1):169–186.

Greybush, S. J., Haupt, S. E., and Young, G. S. (2008). The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Weather and Forecasting*, 23(6):1146–1161.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated

ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1814–1827.

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

Hagedorn, R., Hamill, T. M., and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part I: Two-meter temperatures. *Monthly Weather Review*, 136(7):2608–2619.

Hagedorn, R. and Smith, L. A. (2009). Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, 16(2):143–155.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560.

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, 140(7):2232–2252.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134(11):3209–3229.

Hamill, T. M., Whitaker, J. S., and Mullen, S. L. (2006). Reforecasts: an important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, 87(1):33.

Hardin, J. and Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Hess, R. (2020). Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst. *Nonlinear Processes in Geophysics*, 27(4):473–487.

Horn, R. A. and Johnson, C. R. (1985). *Matrix analysis*. Cambridge university press.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811.

Hu, W., Cervone, G., Clemente-Harding, L., and Calovi, M. (2020). Parallel analog ensemble - the power of weather analogs. In *Proceedings of the 2020 Improving Scientific Software Conference*, pages 1–14.

Jenkinson, A. and Collison, F. (1977). An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memorandum*, 62:18.

Jewson, S., Brix, A., and Ziehmann, C. (2004). A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts. *Atmospheric Science Letters*, 5(5):96–102.

Johnson, C. and Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, 135(640):777–794.

Jolliffe, I. T. (2011). Principal Component Analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Jones, P., Hulme, M., and Briffa, K. (1993). A comparison of Lamb circulation types with an objective classification scheme. *International Journal of Climatology*, 13(6):655–663.

Junk, C., Delle Monache, L., and Alessandrini, S. (2015). Analog-based ensemble model output statistics. *Monthly Weather Review*, 143(7):2909–2917.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability.* Cambridge university press.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, 32(4):419–430.

Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics. Vol. 1: Distribution theory.* Charles Griffin and Company Ltd, 42 Drury Lane, London, 3rd edition.

Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2020). Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review*, 89(2):274–301.

Krzanowski, W. (2000). *Principles of multivariate analysis*, volume 23. OUP Oxford.

Lerch, S. and Baran, S. (2017). Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1):29–51.

Lock, S.-J., Lang, S. T., Leutbecher, M., Hogan, R. J., and Vitart, F. (2019). Treatment of model uncertainty from radiation by the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and associated revisions in the ECMWF ensembles. *Quarterly Journal of the Royal Meteorological Society*, 145:75–89.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Lopez-Pintado, S. and Torrente., A. (2021). *`depthTools`: Depth Tools package*. R package version 0.7.

Magnus, J. R. (1988). Linear structures. *Griffin's Statistical Monographs and Courses*, (42).

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *The Annals of Statistics*, 7(2):381–394.

Makowski, D. (2017). A simple Bayesian method for adjusting ensemble of crop model outputs to yield observations. *European journal of agronomy*, 88:76–83.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press: London.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

Matsueda, M., Kyouda, M., HL, T., and Tsuyuki, T. (2007). Daily forecast skill of multi-center grand ensemble. *SOLA*, 3:29–32.

Muirhead, R. J. (2009). *Aspects of Multivariate Statistical Theory*, volume 197. John Wiley & Sons.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

Mylne, K., Woolcock, C., Denholm-Price, J., and Darvell, R. (2002). Operational calibrated probability forecasts from the ECMWF ensemble prediction system: implementation and verification. In *Preprints of the Symposium on Observations, Data Asimmilation and Probabilistic Prediction*, pages 113–118.

Neal, R., Fereday, D., Crocker, R., and Comer, R. E. (2016). A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, 23(3):389–400.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review*, 110(7):699–706.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*. Arnold, second edition.

Palmer, T. (1995). The singular-vector structure of the atmospheric general circulation. *Journal of the Atmospheric Sciences*, 52:1434–1456.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Technical Memoranda*, 598:1–42.

Peirolo, R. (2011). Information gain as a score for probabilistic forecasts. *Meteorological Applications*, 18(1):9–17.

Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S., and Chirico, G. B. (2017). Adaptive Kalman filtering for postprocessing ensemble numerical weather predictions. *Monthly Weather Review*, 145(12):4837–4854.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413. Cambridge University Press.

Pinson, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(666):1273–1284.

Pinson, P. and Tastu, J. (2013). Discrimination ability of the energy score. Technical report, Kgs. Lyngby: Technical University of Denmark (DTU).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1996). *Numerical Recipes in Fortran*, volume 2. Cambridge University Press.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Rao, C. R. (1972). Generalized inverse of a matrix and its applications. In *Vol. 1 Theory of Statistics*, pages 601–620. University of California Press.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900.

Rawlins, F., Ballard, S., Bovis, K., Clayton, A., Li, D., Inverarity, G., Lorenc, A., and Payne, T. (2007). The Met Office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 133(623):347–362.

Richardson, D., Neal, R., Dankers, R., Mylne, K., Cowling, R., Clements, H., and Millard, J. (2020). Linking weather patterns to regional extreme precipitation for highlighting potential flood events in medium-to long-range forecasts. *Meteorological Applications*, 27(4):e1931.

Rougier, J., Goldstein, M., and House, L. (2013). Second-order exchangeability analysis for multimodel ensembles. *Journal of the American Statistical Association*, 108(503):852–863.

Roulston, M. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, 55(1):16–30.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660.

Sansom, P. G., Stephenson, D. B., and Bracegirdle, T. J. (2021). On constraining projections of future climate using observations and simulations from multiple climate models. *Journal of the American Statistical Association*, 116(534):546–557.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K.,

Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41(7):2514–2519. 2014GL059637.

Schefzik, R. (2016). A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, 144(5):1909–1921.

Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., et al. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640.

Scher, S. and Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2830–2841.

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 Hamill, T. M. (2015a). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596.

Scheuerer, M. and Hamill, T. M. (2015b). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334.

Schott, J. R. (2016). *Matrix analysis for statistics*. John Wiley & Sons.

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., Ferro, C. A., Stephenson, D. B., and Leutbecher, M. (2019). The ensemble-adjusted ignorance score for forecasts issued as normal distributions. *Quarterly Journal of the Royal Meteorological Society*, 145:129–139.

Sievers, O., Fraedrich, K., and Raible, C. C. (2000). Self-adapting analog ensemble predictions of tropical cyclone tracks. *Weather and Forecasting*, 15(5):623–629.

Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., Serpetti, N., Speirs, D. C., Thorpe, R. B., and Blackwell, P. G. (2018). A general framework for combining ecosystem models. *Fish and Fisheries*, 19(6):1031–1042.

Sperati, S., Alessandrini, S., and Delle Monache, L. (2017). Gridded probabilistic weather forecasts with an analog ensemble. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2874–2885.

Stuart, A., Arnold, S., Ord, J. K., O'Hagan, A., and Forster, J. (1994). *Kendall's advanced theory of statistics*. Wiley.

Taillardat, M. and Mestre, O. (2020). From research to applications–examples of operational ensemble post-processing in france using machine learning. *Nonlinear Processes in Geophysics*, 27(2):329–347.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393.

Tennant, W. J., Shutts, G. J., Arribas, A., and Thompson, S. A. (2011). Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Monthly Weather Review*, 139(4):1190–1206.

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.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. (2016). Assessing the

calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25(1):105–122.

Toth, Z. (1989). Long-range weather forecasting using an analog approach. *Journal of Climate*, 2(6):594–607.

Van den Dool, H. (1989). A new look at weather forecasting through analogues. *Monthly Weather Review*, 117(10):2230–2247.

Vannitsem, S., Wilks, D. S., and Messner, J. (2018). *Statistical postprocessing of ensemble forecasts*. Elsevier.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., et al. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, 98(1):163–173.

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

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, 137(4):1460–1479.

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., and Toth, Z. (2008). Ensemble data assimilation with the NCEP global forecast system. *Monthly Weather Review*, 136(2):463–482.

Wilkinson, D. J. (1995). *Bayes linear covariance matrix adjustment*. PhD thesis, Durham University, Durham.

Wilkinson, D. J. (1997). Bayes linear variance adjustment for locally linear DLMs. *Journal of Forecasting*, 16(5):329–342.

Wilkinson, D. J. and Goldstein, M. (1995). Bayes linear adjustment for variance matrices. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 5*, pages 791–799. Oxford University Press.

Wilkinson, D. J. and Goldstein, M. (1996). Bayes linear covariance matrix adjustment for multivariate dynamic linear models. *Bayesian analysis e-print* *http://xxx.lanl.gov/abs/bayes-an/9506002*.

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2821–2836.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. Academic press.

Wilks, D. S. (2015). Multivariate ensemble model output statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, 141(688):945–952.

Wilks, D. S. (2017). On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143(702):164–172.

Wilks, D. S. (2018). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 144(710):76–84.

Williamson, D., Goldstein, M., and Blaker, A. (2012). Fast linked analyses for scenario-based hierarchies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5):665–691.

Wilson, L. J., Burrows, W. R., and Lanzinger, A. (1999). A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, 127(6):956–970.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report*, 42(106):336.

Zamo, M., Bel, L., and Mestre, O. (2021). Sequential aggregation of probabilistic forecasts – application to wind speed ensemble forecasts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):202–225.