ORIGINAL ARTICLE

New approaches to postprocessing of multi-model ensemble forecasts

Clair Barnes[1] | **Richard E. Chandler**[1] | **Christopher M. Brierley**[2]

[1]Department of Statistical Science, University College London, London WC1E 6BT, UK

[2]Department of Geography, Pearson Building, University College London, London WC1E 6BT, UK

Correspondence
Clair Barnes, Department of Statistical Science, University College London, London WC1E 6BT, UK
Email: clair.barnes.16@ucl.ac.uk

Funding information
EPSRC Grant: EP/N509577/1

Ensemble weather forecasts often under-represent uncertainty, leading to over-confidence in their predictions. Multi-model 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 are 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).

We propose a flexible multivariate Bayesian postprocessing framework, based on a directed acyclic graph representing the relationships between the ensembles and the observed weather. The posterior forecast is inferred from available ensemble forecasts and an estimate of the shared discrepancy, obtained from a collection of past forecast-observation pairs. We also propose a novel approach to selecting an appropriate training set for estimation of the required correction, using synoptic-scale analogues to obtain a regime-dependent estimate of the adjustment.

The proposed technique is applied to forecasts of surface temperature over the UK during the winter period from 2007-2013. Although the resulting parametric multivariate-normal probabilistic forecasts are marginally less sharp than those of the leading competitor, they capture the spatial

structure of the observations better than a correlation structure based on either the ensembles or climatology alone, and are robust to changes in the variables and spatial domain of the forecast, at a greatly reduced computational cost.

KEYWORDS
statistical postprocessing, multimodel ensemble, Bayesian inference, analogues, calibration, joint predictive distributions

1 | INTRODUCTION

Most weather forecasts are generated by numerical weather prediction (NWP) models, propagating a 'best guess' of the initial weather state through a model of the atmosphere in order to predict the future state. The uncertainty associated with the forecast is typically assessed from the dispersion of an ensemble obtained by running the model multiple times with perturbed initial conditions; if the forecast ensemble dispersion is small, then the uncertainty about the issued forecast is assumed to be small, and vice versa.

Despite many recent modelling improvements, numerical weather forecasts remain susceptible to errors from various sources, and it is generally accepted that the output of any ensemble prediction system (EPS) will require some form of postprocessing if it is to be useful (Wilks, 2011). Model biases are particularly prevalent when considering the surface weather quantities of most interest to many users (Atger, 2003), and although the uncertainty surrounding the initial conditions is sampled (at least partly) through the use of perturbed ensembles, further uncertainty arises from the choice of parameterisation schemes, boundary conditions, and processes at unresolved scales. Many NWP models now include schemes to partially account for this model uncertainty, for example by perturbing selected parameters - as in the UK Met Office's 'random parameter' scheme (Bowler et al., 2008; Baker et al., 2014) - or perturbing the effect of the parametrizations on certain variables - as in the ECMWF's 'stochastic perturbed parametrization tendencies' (Palmer et al., 2009) - to obtain a more representative ensemble spread that is physically consistent. However, many EPS forecasts are still found to be overconfident - that is, the ensemble spread tends to be smaller than the forecast error - with this underdispersiveness becoming worse at longer leadtimes (Weigel et al., 2008). 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.

A common approach to improving the calibration of the raw output from an EPS is to postprocess the forecast in some way, with the joint aims of bias correction and spread calibration (uncertainty quantification). Methods include Model Output Statistics (MOS) (Glahn and Lowry, 1972; Jewson et al., 2004; Gneiting et al., 2005), analogue ensembles (Hamill et al., 2006; Hamill and Whitaker, 2006), 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), ensemble best member dressing (Roulston and Smith, 2003), Kalman filtering (Delle Monache et al., 2011), and quantile regression methods (Taillardat et al., 2016; Bentzien and Friederichs, 2012).

Regardless of the numerical model and postprocessing method used, single-ensemble forecasts cannot fully account for uncertainty due to the choice of a particular model. A complementary approach, allowing further sampling of potential uncertainty, is to construct a multi-model ensemble (MME) combining the output of several EPSs. 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; Johnson and Swinbank, 2009; Weigel et al., 2009). More sophisticated approaches to MME postprocessing include ensemble BMA (Fraley et al., 2010) and ensemble MOS (Yuen et al., 2017), both of which apply a bias correction and variance adjustment to a weighted average of the individual EPS forecasts.

In this paper, we propose a novel Bayesian framework, developed around a directed acyclic graph representing the structure of the MME forecast system, to postprocess multivariate forecasts from several EPSs simultaneously. Performance of the new postprocessing technique is evaluated against an established MOS technique, and against uncorrected multi-ensemble forecasts.

Any statistical postprocessing of this kind requires a training dataset of past forecast-verification pairs (by which we mean a past forecast and its verifying observation or analysis), from which the necessary correction can be estimated. Typically, a moving-window approach has been used to select this training set (Gneiting et al., 2005); alternatively, previous years' reforecast data from the same operational model, date and synoptic time may be used to provide a training set of greater size (Hagedorn et al., 2008; Hamill, 2012), or analogues from an archive of prior forecasts may be selected on the basis of their similarity to the current prediction (Hamill et al., 2006; Delle Monache et al., 2013; Junk et al., 2015). We propose a new source of synoptic-scale analogues for construction of a training dataset, choosing candidate forecasts generated under weather regimes similar to that of the forecast of interest.

We begin with a motivating example in Section 2, with a discussion of two MME postprocessing methods in Section 3. Section 4 introduces the new postprocessing framework, and Section 5 suggests the process by which a training set can be obtained. Section 6 defines the metrics used to assess forecast performance, with results presented in Section 7. We conclude with a summary and discussion in Section 8.

2 | DATA

The MME system considered in this paper consists of component ensembles from the ECMWF, NCEP and UK Met Office, with 50, 20 and 23 perturbed members respectively - among the largest ensembles available during the study period. In addition, these three ensembles all apply different approaches to handling initial-condition uncertainty and model uncertainty, so should be exploring quite different areas of the state space of potential models. The ensemble forecasts were obtained from the TIGGE archive (Bougeault et al., 2010). We consider forecasts of 2m surface temperature at midnight during the winter period (December-January-February, excluding leap days) from December 2007 to February 2014, issued at 24h intervals up to 15 days ahead. October and November forecasts were also used as candidate training cases, but were not postprocessed. We use the term 'forecast instance' to refer to a forecast issued on a single day for a given leadtime and at a particular synoptic time; here, we have $7 \times 90 = 630$ forecast instances to postprocess at each leadtime. Forecasts were downloaded from TIGGE on a 1° latitude-longitude grid over the region from 50 to 60°N and 6°W to 2°E, covering the British Isles.

ERA-Interim reanalyses of 2m surface temperatures on the same grid are used to verify the postprocessed forecasts (Dee et al., 2011). It is possible that the ECMWF ensemble forecasts may perform better than the other two ensembles in this respect, since these forecasts are based on a similar model to the reanalysis. However, the relative performance of different combinations of contributing ensembles is outside of the scope of this paper.

The study area consists of forecasts at 13 'locations': alternating grid cells over the land mass of the UK (Figure 1). This choice was made to limit the size of the data set for ease of processing and interpretation, while including relatively heterogeneous climatologies.

3 | CURRENT METHODS IN MME POSTPROCESSING

We begin by considering two established postprocessing schemes by which a MME forecast might be corrected, in order to better understand the motivation for the new approach introduced in Section 4. These methods have been chosen because, like the proposed framework, they are parametric methods and produce multivariate-normal probabilistic forecasts; this allows for direct comparison of the skill of all three methods.

3.1 | Multi-model superensemble

The simplest method of combining multiple ensemble forecasts is to pool the ensemble members into a single superensemble, defining the predictive density as multivariate normal, with its mean and covariance matrix specified as the mean and covariance matrix of all members. Several studies have found pooling to be a straightforward way to improve the average performance of single-ensemble raw forecasts (Hagedorn et al., 2005; Doblas-Reyes et al., 2005; Johnson and Swinbank, 2009).

Figure 2 shows a bivariate example of a pooled superensemble forecast generated by this MME system, along with the verifying reanalysis. All three ensembles predict strong positive correlation between temperatures in the two locations; each ensemble's members are tightly clustered, indicating a high degree of confidence in the forecast, although the individual ensembles are fairly distinct. The ellipse contains 95% of the predictive density, and is much larger than the spread of any one ensemble. The pooled spread 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 the improvement to the various models exploring different regions of the phase space.

It has been suggested that one of the reasons for the superior performance of multi-model ensembles is that the errors from component ensembles cancel one another out, leading to a bias-corrected forecast (Hagedorn et al., 2005). This can only occur if the errors of the component ensembles (and their members) are independently distributed around the true value. However, many NWP models share grid resolutions, parameterisations, and even code, and so are likely to display errors of a similar type (such as similar wet/dry or cold/warm biases), as in Figure 2. A more sophisticated approach that can account for a potential common bias is clearly called for.

FIGURE 1 Locations in the study. Each cell is labelled with the name of the largest city within its boundaries.

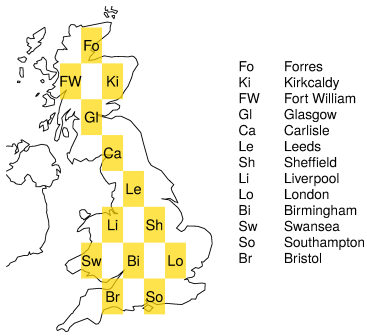
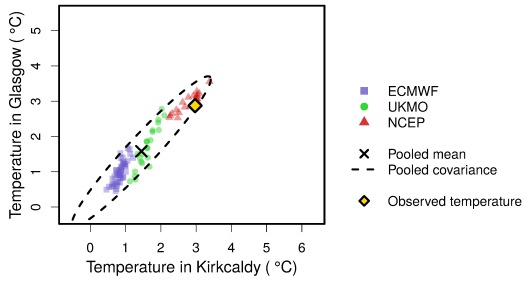


FIGURE 2 One-day-ahead ensemble predictions of temperatures in Kirkcaldy and Glasgow, issued on 19 January 2010.

The ellipse is a 95% prediction region calculated from a bivariate normal distribution with the same mean and covariance matrix as the pooled superensemble.



3.2 | Nonhomogeneous Gaussian Regression

Model Output Statistics (Glahn and Lowry, 1972; Wilks, 2011) is among the most commonly applied statistical postprocessing techniques. The exact form of MOS applied will vary according to the weather quantities to be calibrated; in the case of surface temperatures, where the ensemble forecasts generally have approximately Gaussian distributions, MOS often takes the form of a nonhomogeneous Gaussian regression (NGR) (Hagedorn et al., 2008, 2012; Junk et al., 2015). This means that, given an ensemble of forecasts $\{y_1, \dots, y_k\}$ of some weather quantity Y_0, with the forecasts having sample variance s^2, the NGR predictive distribution of Y_0 has the form

$$Y_0 \sim N(a + b_1 y_1 + \cdots + b_k y_k, c + ds^2) \tag{1}$$

The coefficients $a, b_1, \dots, b_k, c$ and d are estimated by least-squares regression and optimisation over a training set of forecast-observation pairs, with each training case consisting of forecasts from the same k ensemble members as the forecast to be postprocessed. The regression is described as nonhomogeneous because, just as the predictive mean depends on the predictors $y_1, \dots, y_k$, the predictive variance depends on the sample variance s^2 of the predictors. In order to ensure that the NGR variance is strictly positive, c and d are constrained to be greater than 0.

The NGR approach is readily applied to a multi-model context, as described in Gneiting et al. (2005) and implemented in the freely available R package `ensembleMOS` (Yuen et al., 2017). Given a collection of m ensemble forecasts of Y_0 at locations $l = 1, \dots, p$, with the ith ensemble having n_i members, we denote the jth forecast produced by the ith ensemble at location l as $y_{ij}(l)$. Here, upper case indicates a random variable, while lower case denotes realised values of those random variables. The n_i members of the ith ensemble are considered to be exchangeable, having identical statistical properties, and so the linear regression equation (1) is applied to the ensemble mean forecasts $\bar{y}_i(l) = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}(l)$, and $\bar{s}(l)$ is the standard deviation of the ensemble means $\bar{y}_i(l)$ (Gneiting et al., 2005).

NGR is a univariate postprocessing method, estimating a single set of regression coefficients; forecasts at several locations may be postprocessed independently, or the ensemble mean forecasts may be pooled across some or all regions, in which case forecasts at all of the pooled locations are assumed to obey a common regression relationship with their verifying observations (Fraley et al., 2007; Lerch and Baran, 2017). Pooling across locations gives a larger set of forecast-error pairs from which to estimate each set of coefficients, and so can produce more numerically stable estimates where the amount of available observations would otherwise be small. However, if the pooled sites are not sufficiently homogeneous, the resulting forecasts may retain substantial local biases and dispersion errors. Due to the inhomogeneity of the errors observed at each location in the temperature forecasts, and in order to ensure a fair comparison with the proposed alternatives, we choose the former option, estimating separate NGR parameters at each location from the same training set of forecast instances. Hence the predictive distribution at location l is

$$Y_0(l) \sim N(a(l) + b_1(l)\bar{y}_1(l) + \cdots + b_m(l)\bar{y}_m(l), c(l) + d(l)\bar{s}^2(l)) \qquad \forall\, l \in 1, \dots, p \tag{2}$$

The coefficients $a, b_1, \dots, b_m$ are initially estimated from an appropriately chosen set of training data by least-squares regression, then all coefficients are numerically optimised by minimising the Continuous Ranked Probability Score (CRPS, see Section 6.3), following Gneiting et al. (2005). The fitted coefficients provide information about the performance of the ensemble members over the training data. The $b_1, \dots, b_m$ are weights applied to the individual ensemble forecast means, according to their relative performances, with a a simple bias correction to the weighted mean forecast thus obtained. The variance component of (2) is intended to capture the fact that previous studies have observed a systematic relationship between the magnitude of forecast errors and the spread of the ensembles producing them

(Gneiting et al., 2005). An understanding of the strength of the spread-error relationship within the training set can be gained from d, with larger values of d indicating a stronger relationship; when $d = 0$, the spread and error are essentially independent, and the resulting distribution is reduced to a linear regression, inflated by c to replicate the uncertainty of the errors in the training data. As such, NGR offers an intuitive, appealing way to simultaneously correct a multi-ensemble forecast and assess the relative performance of the constituent ensembles. However, it should be noted that when - as is often the case - several ensembles are highly collinear, negligible weights may be assigned to all but the single most skilful forecast (Gneiting et al., 2005), so the weights should not be interpreted as direct measures of performance.

As described above, the NGR procedure produces postprocessed forecast distributions for each variable independently. In some applications however, it is important to consider all quantities simultaneously and hence to produce a postprocessed joint forecast distribution. Given that the marginal forecast distributions are normal, it is natural to specify a multivariate normal distribution for the joint forecasts. To do this, we follow the approach used by Feldmann et al. (2015), Berrocal et al. (2007) and others, and use a continuous copula function to provide the dependence structure between the NGR-postprocessed marginal forecasts and so obtain a multivariate predictive density

$$\mathbf{Y}_0 \sim MVN\left(\boldsymbol{\mu}_{ngr}, \mathbf{V}_{ngr}\mathbf{P}\mathbf{V}_{ngr}\right), \tag{3}$$

where $\boldsymbol{\mu}_{ngr}$ is the vector of NGR marginal predictive means, $\mathbf{V}_{ngr}$ is the diagonal matrix of NGR marginal predictive standard deviations, and $\mathbf{P}$ is a correlation matrix specified by the user.

In geospatial statistics, it is common to specify the elements ρ_{ij} of $\mathbf{P}$ through a parametric stationary, isotropic correlation function of the distance between location s_i and location s_j (Möller et al., 2013). However, attempts at fitting this type of function to the data set described in Section 2 produced unstable parameter estimates, suggesting that the assumptions of stationarity and isotropy do not hold in this case. Alternative recently proposed approaches construct a discrete empirical copula, based on the rank structure either of the forecast ensemble (Ensemble Copula Coupling, ECC - Schefzik et al. (2013)) or of a climatological sample (the Schaake Shuffle, Clark et al. (2004); Schefzik (2016)). These empirical copula approaches have been shown to be effective in obtaining jointly calibrated forecasts, but deliver a discrete forecast distribution based directly on the input samples, whereas the distribution in equation (3) is continuous. Since we require predictive densities, we adapt the approach of Schefzik (2016), and use the empirical correlation structure of the observations in the training set to estimate $\mathbf{P}$. A separate correlation matrix is estimated for each forecast instance, using the same set of training cases that are used to estimate the NGR parameters.

Fitting by CRPS minimisation means that NGR forecasts generally perform well when assessed via single scoring rules - particularly, of course, the CRPS. However, recent research suggests that optimisation by CRPS minimisation can lead to forecasts that are sharper (having lower variance) than competitors, but at the cost of calibration (Wilks, 2018): this contravenes the maxim of Gneiting et al. (2008) that sharpness should be improved only while respecting forecast calibration. Wilks (2018) proposes the introduction of a penalty function in the optimisation step to ensure that calibration, rather than sharpness, is maximised; however, this approach has not been widely adopted, so we have chosen to use the standard form of the algorithm provided by the `ensembleMOS` package in R (Yuen et al., 2017).

NGR methods also fail to exploit the full range of information provided by the available ensemble forecasts. By establishing the regression relationship only over the ensemble mean forecasts, and 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. In the next section, we present an intuitive and easily

applied framework designed to address these issues and produce a properly calibrated, bias-corrected multivariate multi-ensemble forecast, making use of all available information.

4 | A NEW APPROACH TO POSTPROCESSING OF MME FORECASTS

The proposed method applies a new approach to MME postprocessing, treating the ensemble forecasts as related elements of a single forecasting system. 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 method is developed from that presented in Chandler (2013) in the context of climate projections. A key difference, however, is that climate projections aim to make statements about the statistical properties of future weather, such as regional or global mean temperatures; in the current weather forecasting context, the aim is to forecast the actual weather quantities, rather than their statistical properties.

We begin by revisiting the representation of the available forecast data. Here, for each forecast instance we have a collection of ensemble forecasts from m models, made on a single day for a given leadtime and at a particular synoptic time; each forecast instance is postprocessed separately. For a given forecast instance, we have a vector of weather quantities of interest, $\mathbf{Y} = \{Y_1, \ldots, Y_p\}'$, with p the number of variables we wish to forecast. The value of $\mathbf{Y}$ that is eventually observed is denoted $\mathbf{Y}_0$. Again, we use upper case to denote random variables, and lower case to denote realisations of those random variables. In the case study presented in this paper, $\mathbf{Y}$ contains surface temperatures at each of the 13 grid cells shown in Figure 1, hence $p = 13$ here.

For each $i = 1, \ldots, m$, the ith ensemble provides a set of n_i forecasts of the vector $\mathbf{Y}$, with the jth forecast from the ith ensemble being labelled $\mathbf{Y}_{ij}$. Forecasts of surface temperature are commonly assumed to be reasonably well represented by multivariate normal distributions on the basis of plots such as Figure 2, in which the individual ensembles typically show a roughly elliptical scatter (Wilson et al., 1999; Wilks, 2002). The members of the ith ensemble are thus assumed to be drawn independently from multivariate normal distributions, conditional on the ensemble's population mean, $\boldsymbol{\mu}_i$:

$$\mathbf{Y}_{ij} \,|\, \boldsymbol{\mu}_i \sim MVN(\boldsymbol{\mu}_i, \mathbf{C}_i) \tag{4}$$

We now consider the fact that, while forecasts from a single ensemble are generally more similar to one another than they are to forecasts from other ensembles, ensembles may also be more similar to one another than they are to reality, for the reasons discussed in Section 3.1. The consequence of this is that if one ensemble predicts too low a temperature, it is likely that the other ensembles will display a similar tendency. To reflect this dependence, the individual ensemble mean forecasts $\boldsymbol{\mu}_i$ are themselves assumed to be dispersed around a mutual consensus, $\boldsymbol{\xi}$, according to the covariance matrix $\boldsymbol{\Sigma}$, and to be independent of one another only conditional on this consensus. This consensus $\boldsymbol{\xi}$ can be thought of as the centre of the population of possible ensembles; if we could sample an infinite number of ensembles from an infinite number of models for a particular forecast, the mean of the infinite sample of $\boldsymbol{\mu}_i$s would lie at $\boldsymbol{\xi}$, although each individual ensemble may be systematically offset from the consensus.

The consensus can be decomposed into the 'true' value, $\mathbf{Y}_0$, plus a shared discrepancy $\boldsymbol{\Delta}$. The distribution of $\boldsymbol{\Delta}$ for any forecast instance can be estimated using the mean and covariance of an appropriate training set of past

forecast-observation pairs, as discussed in Section 5. Thus we have

$$\boldsymbol{\mu}_i|\boldsymbol{\xi} \sim MVN(\boldsymbol{\xi}, \boldsymbol{\Sigma}) \qquad \text{where } \boldsymbol{\xi} = \mathbf{Y}_0 + \boldsymbol{\Delta} \tag{5}$$

$$\boldsymbol{\Delta} \sim MVN(\boldsymbol{\eta}, \boldsymbol{\Lambda}) \tag{6}$$

The structure of equations (4) to (6) is illustrated graphically in Figure 3. Here, the arrows 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 said to be 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 assumed to be independent, given $\boldsymbol{\mu}_i$. This means that, if $\boldsymbol{\mu}_i$ is known, then information about the value of $\mathbf{Y}_{i1}$ cannot tell us anything new about the value of $\mathbf{Y}_{i2}$. These assumed conditional independence relationships will be exploited in the subsequent derivations.

For computational purposes, the data structure in Figure 3a can be simplified without loss of information by exploiting the fact that the sample mean and covariance matrix are sufficient statistics for the parameters of a multivariate normal distribution - they capture all of the available information provided by the sample about the population mean and covariance (Cox and Hinkley, 1974, p173). This enables us to replace the individual members $\mathbf{Y}_{ij}$ of ensemble i by the ensemble mean $\overline{\mathbf{Y}}_i$, and to replace equation (4) with

$$\overline{\mathbf{Y}}_i|\boldsymbol{\mu}_i \sim MVN(\boldsymbol{\mu}_i, n^{-1}\mathbf{C}_i) \tag{7}$$

without loss of information (see supplementary paper, Section S1.2, for details of this equivalence). This in turn can be combined with equation (5), 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

$$\overline{\mathbf{Y}}_i|\boldsymbol{\xi} \sim MVN\left(\boldsymbol{\xi}, \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i\right). \tag{8}$$

This leads to a simplified graphical structure, shown in Figure 3b. $\mathbf{C}_i$ is estimated using the sample covariance matrix of forecasts from the ith ensemble, and $\boldsymbol{\Sigma}$ using the sample covariance matrix of the ensemble means. 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 $\boldsymbol{\mu}_i$: however, when (as here) each ensemble has many members, this bias will be small. A further issue is that the elements of $\boldsymbol{\Sigma}$ will be estimated imprecisely if m, the number of ensembles, is small. Again, this is the case here (m=3): this should be borne in mind below when considering the performance of the method in practice. For more discussion of the relevant estimation issues in a closely related problem, see Chandler (2013).

The role of the individual members of each ensemble is to provide an estimate of $\mathbf{C}_i$: having obtained this, equation (8) implies that only the ensemble means are needed. An important difference between this and other approaches such as NGR is that the latter approach uses only the sample variance of the ensemble means (the diagonal elements of $\boldsymbol{\Sigma}$), thereby losing the additional information on ensemble spread.

To simplify the notation, let $\mathbf{D}_i = \boldsymbol{\Sigma} + n_i^{-1}\mathbf{C}_i$, and so

$$\overline{\mathbf{Y}}_i|\boldsymbol{\xi} \sim MVN\left(\boldsymbol{\xi}, \mathbf{D}_i\right) \qquad \text{where } \boldsymbol{\xi} = \mathbf{Y}_0 + \boldsymbol{\Delta}. \tag{9}$$

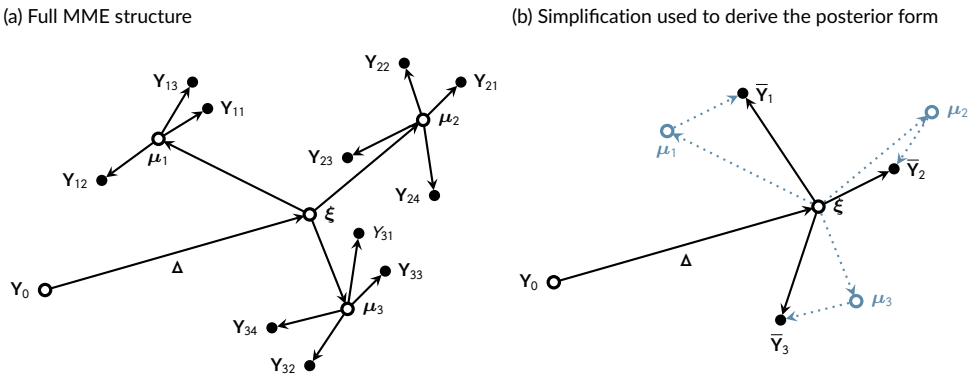
In the framework set out above, the aim of the postprocessing is to use the forecast ensembles $\left\{\mathbf{Y}_{ij}\right\}$ to make statements about the value of $\mathbf{Y}_0$ that will be realised. As in Chandler (2013), this is most conveniently done in a Bayesian framework which also allows the incorporation of additional knowledge about $\mathbf{Y}_0$ via a prior distribution. We use a

multivariate normal prior here for convenience:

$$\mathbf{Y}_0 \sim MVN(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \tag{10}$$

If we have no particular prior assumptions about the distribution of $\mathbf{Y}_0$, we can set a non-informative prior with $\boldsymbol{\Gamma}^{-1} = \mathbf{0}$ (meaning that the variance, $\boldsymbol{\Gamma}$, is infinite); regardless of the value of $\boldsymbol{\alpha}$ specified, such a prior will contribute nothing to the final posterior forecast. In working with the inverse $\boldsymbol{\Gamma}^{-1}$, we follow standard practice in Bayesian analyses (Bernardo and Smith, 2001), where the inverse of any covariance matrix is usually referred to as a precision matrix.

FIGURE 3 Schematic diagram of relationships between elements of the multi-ensemble system
Quantities known at the time of forecasting are shown as filled nodes, with unknown quantities represented by open nodes.



It can be shown, using arguments adapted from those in Chandler (2013) and presented in detail in the online supplement to this paper, that the posterior distribution of $\mathbf{Y}_0$, conditioned on the ensemble forecasts, is itself multivariate-normal, with

$$\mathbf{Y}_0 | \mathbf{Y}_{ij}, \boldsymbol{\Delta} \sim MVN(\boldsymbol{\tau}, \mathbf{S}) \tag{11}$$

$$\mathbf{S}^{-1} = \boldsymbol{\Gamma}^{-1} + (\Sigma_{\mathbf{D}} + \boldsymbol{\Lambda})^{-1}, \tag{12}$$

$$\boldsymbol{\tau} = \mathbf{S}\left[\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} + (\Sigma_{\mathbf{D}} + \boldsymbol{\Lambda})^{-1}\left\{\Sigma_{\mathbf{D}} \sum_{i=1}^{m} \mathbf{D}_i^{-1}\overline{\mathbf{y}}_i - \boldsymbol{\eta}\right\}\right], \tag{13}$$

where $\Sigma_{\mathbf{D}} = \left(\sum_{i=1}^{m} \mathbf{D}_i^{-1}\right)^{-1}$, a covariance matrix representing the uncertainty about the 'true' position of the ensemble consensus $\boldsymbol{\xi}$.

The posterior precision matrix $\mathbf{S}^{-1}$ is the sum of the prior precision $\boldsymbol{\Gamma}^{-1}$ and the precision $(\Sigma_{\mathbf{D}}+\boldsymbol{\Lambda})^{-1}$ of the estimate of the discrepancy-corrected consensus, $\boldsymbol{\xi} - \boldsymbol{\eta}$, which is represented by the term in braces $\{\}$ in equation (13). The precision of $\boldsymbol{\xi} - \boldsymbol{\eta}$ is the inverse of the sum of the uncertainty $\boldsymbol{\Lambda}$ about the estimated discrepancy, and the uncertainty $\Sigma_{\mathbf{D}}$ about the ensemble consensus.

The posterior mean vector $\boldsymbol{\tau}$ is a weighted sum of terms representing the prior estimate $\boldsymbol{\alpha}$ and the mean vector $\boldsymbol{\xi} - \boldsymbol{\eta}$, inferred from the ensemble forecasts adjusted by the expected discrepancy, as described in equation (6). The weights given to these two components are determined by the covariance matrices of the prior distribution, $\boldsymbol{\Gamma}$, and

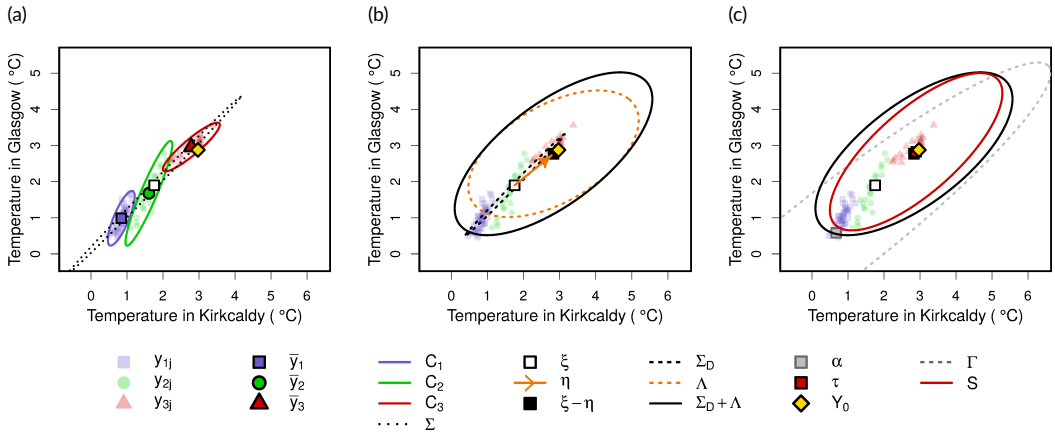
FIGURE 4 Example of postprocessing a collection of ensemble forecasts using the Bayesian framework.

Ellipses represent regions containing 95% of the multivariate-normal density with the specified mean and covariance matrix.

(a) Ensemble members $\mathbf{Y}_{ij}$ from three ensembles, with ensemble sample means $\overline{\mathbf{Y}}_i$ and covariances $\mathbf{C}_i$.

(b) The ensemble consensus $\boldsymbol{\xi}$ and associated uncertainty $\Sigma_{\mathbf{D}}$, adjusted by the estimated discrepancy correction $\boldsymbol{\eta}$ and uncertainty $\boldsymbol{\Lambda}$.

(c) The discrepancy-adjusted consensus $\boldsymbol{\xi} - \boldsymbol{\eta}$ is combined with the prior estimate $\boldsymbol{\alpha}$, weighted by their respective covariances $\Sigma_{\mathbf{D}} + \boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$, to obtain the posterior mean $\boldsymbol{\tau}$ and covariance $\mathbf{S}$.



of the discrepancy-adjusted consensus, $\Sigma_{\mathbf{D}} + \boldsymbol{\Lambda}$, respectively. Figure 4 shows the various components involved in postprocessing a multi-model ensemble forecast, using the ensembles introduced in Figure 2 as a case study and with the mean and covariance of the prior distribution estimated from a sample climatology, using the observations from the week centred on the forecast issue date in the ten years prior to the forecast issue year. The postprocessed forecast variance reflects both the relationship between the temperatures in the two regions and the spread of the errors in the training set.

Experiments not presented in this paper have shown that postprocessed forecasts with an informative prior component are, in general, less skilful than those with a non-informative prior having $\boldsymbol{\Gamma}^{-1} = \mathbf{0}$, being sharper but with less accurate mean forecasts. We conjecture that this is because the discrepancy-adjusted forecasts are, to the best of our knowledge, the best source of information available to us 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, shifting the posterior mean away from the forecasts and towards the prior. An alternative perspective is that because the forecasts are based on dynamical data assimilation, they have already implicitly accounted for the 'prior' information. For this reason, a non-informative prior is used in the case study described in Section 7, which means that the posterior mean vector and covariance matrix in (12) and (13) can be simplified to

$$\mathbf{S}^{-1} = (\boldsymbol{\Lambda} + \Sigma_{\mathbf{D}})^{-1} \qquad \boldsymbol{\tau} = \Sigma_{\mathbf{D}} \sum_{i=1}^{m} \mathbf{D}_i^{-1} \overline{\mathbf{y}}_i - \boldsymbol{\eta}. \tag{14}$$

Finally, we note that even if we did not consider the multivariate-normal assumption to hold, 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 is a valid summary of

the uncertainty in this optimum linear combination (Goldstein and Wooff, 2007).

5 | CALIBRATION USING ANALOGUES TO THE CURRENT FORECAST

All statistical postprocessing techniques require an appropriate training set of past forecast-observation pairs, from which a correction to the current forecast can be derived; 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. Training cases are often taken from the w days immediately preceding the forecast date, using a 'moving window' approach (Gneiting et al., 2005). The use of such a training set implicitly assumes that the biases of recent forecasts will persist for the current forecast. This may be true to some extent, but since forecast biases are also known to be flow-dependent and to vary with the dominant weather pattern (Eckel and Mass, 2005; Greybush et al., 2008; Ferranti et al., 2015), it does not necessarily provide the most appropriate basis for estimation of future forecast biases; the relevance of a moving window training set is likely to reduce with both increasing forecast leadtime and increasing size of the training set (and corresponding earlier start date). More pertinent information may be obtained by selecting a training set of forecasts that predict similar weather to that anticipated by the current forecast.

5.1 | Selection of analogues to the current forecast

We begin by identifying instances in the training set that are, in some sense, similar to the instance to be postprocessed. Given an archive of 'candidate' forecasts, we can define a distance metric to identify those candidates that are closest to the current forecast instance. The candidates found to be most similar to the forecast of interest are referred to as analogues; their errors provide a sample of forecast errors that is expected to be representative of those of the present forecast instance.

Following the metric proposed in Delle Monache et al. (2011), we calculate the Euclidean distance $\|\mathbf{F}, \mathbf{C}\|$ from each p-dimensional candidate vector $\mathbf{C}$ to the current forecast vector $\mathbf{F}$. Each variable l is first normalised by dividing by its standard deviation σ_l over all candidates. Those candidates with the smallest values of

$$\|\mathbf{F}, \mathbf{C}\| = \sqrt{\sum_{l=1}^{p} \left(\frac{F(l) - C(l)}{\sigma_l} \right)^2} \tag{15}$$

are selected as analogues to $\mathbf{F}$ (Delle Monache et al., 2013; Junk et al., 2015).

The vectors $\mathbf{F}$ and $\mathbf{C}$ are typically chosen to contain the specific forecast quantities of interest, with candidates selected from the same season and at the same synoptic time and leadtime as the forecast to be postprocessed. In the multi-ensemble framework presented here, $\mathbf{F}$ and $\mathbf{C}$ are vectors of length $m \times p$ containing the forecast ensemble mean temperatures at all locations, with $m = 3$ ensembles and $p = 13$ locations giving a 39-dimensional candidate search space. We refer to analogues selected in this way as direct analogues (DA).

5.2 | Analogue selection by weather regime

A potential difficulty with the direct analogue approach is that as the dimension p increases, the quality of the selected analogues is likely to fall, since the aggregated distance cannot discriminate between (for example) candidates with several moderate outliers, and candidates with a single large outlier. This issue is likely to become particularly acute when forecasts at a large number of spatial locations are to be postprocessed. To address this problem, it will be

helpful to reduce the dimensions of the candidate vectors before calculating the distances in (15). Here, we propose to identify analogues by applying dimension reduction techniques to the forecast pressure fields, on the basis that these provide a good characterisation of the physical state of the atmosphere.

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 categorise prevailing weather conditions (Jenkinson and Collison, 1977; Jones et al., 1993), and to obtain indices of large-scale synoptic structure (Wilks, 2011). Since pressure fields are physical quantities, and not parameterised by the NWP models, they tend to be fairly well forecast, and so provide a robust basis on which to identify analogues. The predictands in our case study are surface air temperatures, which are known to be particularly affected by large-scale circulation patterns in mean sea level pressure (MSLP) fields (Della-Marta et al., 2007). We therefore choose to apply PCA to MSLP fields, and search for analogues among the resulting lower-dimensional candidates.

The efficiency of the principal component reduction is such that we need not constrain ourselves only to searching for analogous weather patterns within the relatively small forecast area shown in Figure 1. Since synoptic weather conditions in the surrounding regions also affect the local weather (Neal et al., 2016), the MSLP fields used in this study cover the North Atlantic European region and central Europe (35° to 70°N, 30°W to 20°E; shown in Figure 5); other studies have found this to be the optimal domain size for reconstructing surface temperatures from MSLP-derived regime classifications (Beck et al., 2016).

We begin by identifying the principal modes of climatological variation in the region of interest. For this, a long archive of data is necessary; in this study, the entire available archive of ERA-Interim MSLP reanalysis data was used, giving 38 winters from 1979 to 2016: a total of $T = 3420$ time points. Each field contains $L = 1836$ MSLP values, arranged on a 1° latitude-longitude grid. Following standard practice (e.g. North et al. (1982); Wilks (2011)), the MSLP fields are adjusted so that each point on the regular grid is weighted by the area it represents, by multiplying the value at each latitude θ by $\sqrt{\cos(\theta)}$. For each forecast instance, we convert each pressure field into a daily anomaly field by subtracting its mean, and spatial principal components analysis is performed on these daily anomalies. The resulting eigenvectors (EOFs) represent the dominant modes of spatial variation in the climatological record, with the corresponding normalised eigenvalues indicating the proportion of the data's total variance explained by each eigenvector.

Following Jolliffe (2011), we retain only the first q eigenvectors, where q is the smallest number of eigenvectors needed to capture at least 90% of the variance in the raw data. In the present study, the first six eigenvectors are retained; plots of the spatial patterns represented by the first four of these eigenvectors are shown in Figure 5. The retained modes have fairly straightforward interpretations: the first is associated with pressure systems centred to the north-west of Scotland, the second and third are indices of the strength of north-south and east-west pressure gradients over the UK, and so on. Higher-numbered modes display patterns of increasing complexity.

Having obtained the $(L \times q)$ matrix of principal eigenvectors, denoted $\mathbf{E}$, we project each forecast's latitude-adjusted MSLP anomaly field $\tilde{\mathbf{a}}_f$ onto the eigenvectors to obtain a q-vector $\mathbf{u}$ of principal component scores, which form the coordinates of the points in the basis defined by $\mathbf{E}$:

$$\mathbf{u} = \mathbf{E}^T \tilde{\mathbf{a}}_f \tag{16}$$

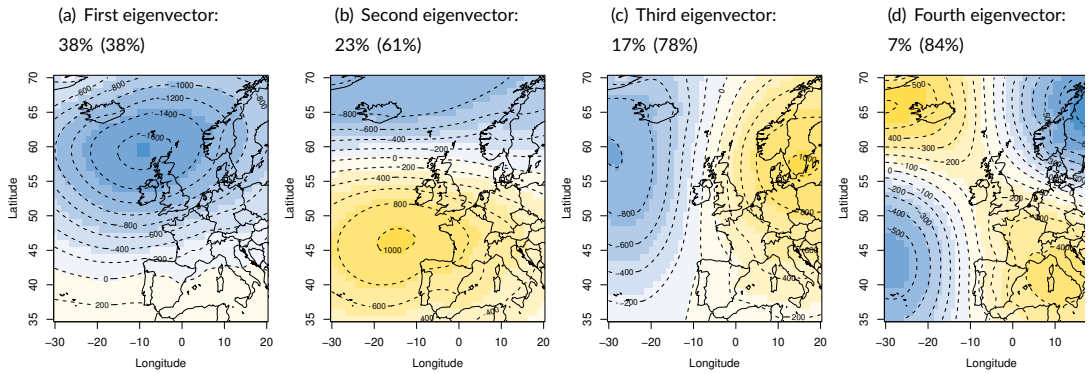
The climatological modes represented by $\mathbf{E}$ need only be obtained once, from the reanalysis data. When a new forecast instance requires postprocessing, we need only obtain its anomaly field $\tilde{\mathbf{a}}_f$ and apply (16) to obtain the principal component scores. Analogues to the new instance are then selected in the q-dimensional principal component space, as in Section 5.1. The new principal component scores are then added to the archive of potential candidate scores, to

be searched when postprocessing the next forecast instance.

In the particular case of a multi-ensemble forecast, we first obtain the mean latitude-adjusted MSLP anomaly field of each of the m ensembles. Principal component scores are obtained separately for each ensemble by projecting the m mean fields onto $\mathbf{E}$; the joint state of the m-ensemble forecast is thus represented by $q \times m$ variables, where q and m are both small, and analogues are selected on the basis of the Euclidean distance (15) calculated over this state space. We refer to analogues selected in this principal component space as weather regime analogues (WR).

Unlike direct analogues in the forecast variable space, the WR candidate archive need not be recalculated if the forecast domain changes slightly. Not only is the principal-component representation a very compact and efficient way to store and represent the search space, but - perhaps more importantly - the analogues chosen will remain the same for any choice of forecast variables or locations for which the synoptic domain remains appropriate. This means, for example, that any subset of forecasts in western Europe could be postprocessed independently using the same training cases obtained in the WR search space, and could be expected to produce mutually consistent and coherent forecasts.

FIGURE 5 Spatial plots of the elements of the first four 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.



6 | FORECAST VERIFICATION METHODS

6.1 | Calibration

A successfully postprocessed forecast should not only reduce biases in the mean forecast, but should also be well calibrated; that is, it should correctly represent the uncertainty in the forecast. A forecast is well calibrated if the verifying observation is indistinguishable from a random draw from the predictive distribution. Alternatively framed, a forecast is considered to be well calibrated if it gives accurate probabilistic forecasts; that is, does it rain on 10% of the days when the forecast says that there is a 10% chance of rain?

We employ several verification methods to evaluate the accuracy and calibration of both the marginal (single-location) and joint (regional) forecasts. Following the principle of Gneiting et al. (2007), we aim to maximise the sharpness of the forecasts, subject to calibration.

6.1.1 | PIT histograms

Calibration of a forecast distribution for a single variable - in this case, a temperature forecast for a single location - can be assessed through the probability integral transform (PIT). This is the value attained by a predictive cumulative distribution function at its verifying observation (Gneiting et al., 2007; Jolliffe and Stephenson, 2012). If a postprocessing method produces well calibrated forecasts for a particular location at a given leadtime, a histogram of the PITs of all 630 forecast instances (90 days for each of the 7 available years) will be uniform. ∩-shaped histograms indicate overdispersion: the forecasts are under-confident, and the observation falls too often in the centre of the forecast distribution. A ∪-shape indicates underdispersion: an over-confident forecast, with the observation falling too often in the tails of the forecast. Systematic bias in the forecasts will result in a skewed or triangular histogram.

6.1.2 | Modified band depth rank histograms

A multivariate analogue to the PIT histogram, allowing evaluation of the calibration of all variables simultaneously, is the modified band depth rank (BDR) histogram, proposed by Thorarinsdottir et al. (2016). The band depth is a measure of the centrality of an observation within a multi-dimensional forecast (López-Pintado and Romo, 2009); in the modified case used here, each probabilistic forecast is represented by a synthetic ensemble of p-dimensional forecasts. The method is computationally slow, limiting the size of ensemble that can practically be used; reflecting the size of the largest available ensemble in the source data, we have used $M = 50$ ensemble members, with each member forecast randomly generated from the multivariate Gaussian predictive density. The verifying observation is then ranked within the synthetic ensemble according to its modified band depth.

Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_{M+1}\} = \{\mathbf{Y}_0, \mathbf{F}_1, ..., \mathbf{F}_M\}$ denote a vector of length $p \times (M + 1)$, consisting of the vector observations $\mathbf{Y}_0$ and the synthetic forecast vectors $\mathbf{F}_1, ..., \mathbf{F}_M$. The observation and forecast vectors all have dimension p, corresponding to the number of grid cells considered simultaneously. We first calculate the prerank of each vector $\mathbf{x}$ in $\mathcal{X}$:

$$r(\mathbf{x}) = \frac{1}{p} \sum_{l=1}^{p} (M + 1 - \text{rank}_{\mathcal{X}}(x_l))(\text{rank}_{\mathcal{X}}(x_l) - 1) + M \tag{17}$$

where $\text{rank}_{\mathcal{X}}(x_l) = \sum_{i=1}^{M+1} \mathbb{1}\{x_{il} \leq x_l\}$ is the component-wise rank of the lth element of the vector $\mathbf{x}$ within $\{\mathcal{X}\}$. The preranks measure the centrality of each element in $\mathcal{X}$, with more central elements attaining higher ranks, and extreme outlying elements attaining the lowest (Thorarinsdottir et al., 2016). The band depth rank of the observation, $\mathbf{y}_0 = \mathbf{x}_1$, is the rank of $r(\mathbf{x}_1)$ in $\{r(\mathbf{x}_1), ..., r(\mathbf{x}_{M+1})\}$, 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 PIT, a calibrated forecast will produce a uniform histogram; however, non-uniform BDR histograms have a somewhat different interpretation. Modified BDR reflects a centre-outward ordering of the rank of the observation within the predictive distribution: the closer the BDR is to one, the more central (deeper) the observation is within the multivariate forecast (Thorarinsdottir et al., 2016). Here, a ∩-shaped histogram indicates an over-correlated joint forecast, while ∪-shaped histograms indicate insufficient correlation in the predictive distributions. A skew histogram with too many high ranks indicates an overdispersive forecast, with the observation falling close to the centre of the forecast distribution more often than it should, and too many low ranks indicates that the observation often falls far from the centre of the predictive distribution, suggesting that the forecast is either underdispersive or systematically biased.

6.1.3 | Summaries of histogram shapes

Histograms are most commonly used descriptively, as a visual diagnostic tool; however, when comparing multiple marginal distributions across several postprocessing methods, it is useful to be able to produce an objective numerical summary of the shape of the data used to construct a histogram.

Any non-central tendencies in the histogram are summarised using the sample skewness, with symmetric histograms obtaining a perfect score of 0. The extent of under- or over-dispersion in the histogram data is quantified by comparing the variance of the values to that of the ideal uniform distribution. For PITs, this is the continuous uniform distribution on the interval $[0, 1]$, which has variance $1/12$ (Casella and Berger, 2002). For the unnormalised band depth ranks R, this would be the discrete uniform distribution on the interval $[1, M + 1]$, which has variance $\frac{M(M+2)}{12}$ (Casella and Berger, 2002); for the normalised ranks $Z = \frac{R-1}{M}$, the calibrated discrete uniform distribution would therefore have variance $\frac{M(M+2)}{12}/M^2 = \frac{M+2}{12M}$. Thus, we define a dispersion index for the PITs and BDRs respectively as

$$disp_{PIT} = 12\,\mathrm{Var}(PIT) \qquad disp_{BDR} = 12\frac{M}{M+2}\,\mathrm{Var}(BDR) \tag{18}$$

A symmetric, $\cup$-shaped histogram has higher variance than a uniform histogram, and will have a dispersion index greater than 1, while a $\cap$-shaped histogram will have a dispersion index less than 1. Any skew in the histogram will reduce the value of the dispersion index slightly. The interpretation of the shape statistics is slightly different for PIT and BDR histograms, as described above.

6.1.4 | Bootstrapped confidence intervals

The significance of any departures from uniformity in the histograms is assessed by a bootstrap procedure, following Efron and Tibshirani (1994) and Hamill (1999). Taking the source data for a single histogram (630 values), we take 10000 bootstrap samples of size 630 with replacement, and calculate the required summary statistic for each of these 10000 bootstrap samples. A significant departure from uniformity is suggested if the 2.5% and 97.5% percentiles of the resulting bootstrap distribution do not bracket the theoretical value for a uniform distribution. When testing for skewness, the theoretical value under the assumption of uniformity is 0, while the theoretical value of the dispersion index is 1.

6.2 | Sharpness

Subject to calibration, we aim to maximise the sharpness of the forecast distribution (Gneiting et al., 2008); sharper, more confident forecasts have a smaller spread than less confident forecasts. The sharpness of the univariate marginal forecasts is measured by the standard deviation of the marginal predictive distribution, with sharpness of the joint forecasts measured by the multivariate equivalent, the determinant sharpness, given by

$$\mathrm{DS} = (\det \mathbf{A})^{1/(2p)} \tag{19}$$

where $\mathbf{A}$ is the postprocessed forecast covariance matrix and p is the dimension of $\mathbf{A}$ (Gneiting et al., 2008). Given two equally well calibrated forecasts, the sharper (ie. with the lower value of DS) is more desirable; however, a sharper forecast that is not well calibrated is overconfident in its prediction, and so of less use in decision making.

6.3 | Accuracy

The accuracy of the marginal deterministic (mean) forecasts produced by each postprocessing method is assessed using the mean absolute error (MAE), which gives the average magnitude of the expected errors for each postprocessing method.

The Brier score (Brier, 1950; Jolliffe and Stephenson, 2012) is used to evaluate success in forecasting binary events, such as the event that the temperature falls below a certain threshold. By integrating the Brier score over all possible thresholds, we obtain the analogue for continuous probabilistic forecasts: the Continuous Ranked Probability Score or CRPS (Hersbach, 2000), a negatively-oriented scoring rule with lower scores indicating better overall forecast performance. The CRPS is generally highly correlated with the MAE, but rewards sharpness and calibration of the predictive density as well as accuracy; in fact, it has been shown that it may reward sharper forecasts in preference to well calibrated ones (Wilks, 2018), so care should be taken when interpreting the CRPS. The CRPS can be extended to the multivariate case in the form of the energy score (Gneiting et al., 2008), defined as

$$ES(F, \mathbf{x}) = \mathbb{E}_F \|\mathbf{X} - \mathbf{x}\| - \frac{1}{2}\mathbb{E}_F \|\mathbf{X} - \mathbf{X}'\|, \tag{20}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbf{x}$ is the verifying observation, F is the forecast distribution, and $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors with distribution F. When $d = 1$, the energy score reduces to the CRPS, which in turn reduces to the MAE when evaluating point forecasts.

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 $\mathbf{X}$ of size $k = 10000$ from the multivariate predictive density F, using the computationally efficient Monte Carlo approximation

$$ES(F, \mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{X}_i - \mathbf{y}\| - \frac{1}{2k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{X}_i - \mathbf{X}_j\|. \tag{21}$$

We use the implementation provided by the R package `scoringRules` (Jordan et al., 2018).

The energy score has been shown by Pinson and Tastu (2013) to have very limited sensitivity to the covariance structure of the forecasts, and to be strongly dominated by the forecast mean vector, particularly in higher dimensions. Our experience has shown that the energy score is closely related to the mean of the MAE across all locations, and it is for this reason that we treat the energy score primarily as a measure of forecast accuracy, and rely on histogram methods to diagnose forecast calibration.

All scores are aggregated over the 630 forecast instances for each synoptic time, leadtime and location. The significance of differences in scores between postprocessing methods is again assessed using a bootstrapped confidence interval. 10000 bootstrap samples of size 630 are drawn with replacement from the 630 instances and used to recalculate the required score difference 10000 times. If the central 95% interval of the resulting bootstrap distribution does not contain zero, the difference between the two scores is said to be significant.

7 | RESULTS

In this section we compare the performance of different ensemble postprocessing techniques, when applied to the forecasts of UK temperatures introduced in Section 2.

We begin by comparing the performance of the various postprocessing techniques, with training datasets ob-

tained using the standard 'moving window' (MW) approach (Section 5). An uninformative prior is used in calculating the Bayesian posterior forecast, as described in Section 4. We then move on to comparing the relative performance of MW-calibrated forecasts with those corrected using direct analogues in temperature space (DA), and with a training set drawn from weather regime analogues (WR), using the methods of Section 5.

All postprocessing was carried out using 25 training cases. A sensitivity analysis (not presented here) indicates that the size of the training set is not a key factor in the performance of any of the postprocessing methods or training sets, with most scores seen to change by less than 0.1 at short to medium leadtimes, and up to 0.3 at leadtimes of greater than ten days; the choice of training set size made no qualitative difference to the conclusions drawn below.

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 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.

7.1 | Bayesian postprocessing vs NGR

7.1.1 | Marginal forecasts

We first evaluate the performance of the postprocessed marginal forecasts. Table 1 shows the mean and range of the MAE and CRPS over all 13 locations. Both postprocessing methods show a significant improvement over the raw superensemble forecast at leadtimes up to six days ahead; the largest improvements are made at higher latitudes, where the raw forecasts tend to be less skilful due to systematic regional biases.

TABLE 1 Mean *(min, max)* MAE and CRPS (in °C), and ES over all locations, at selected leadtimes, for each postprocessing method. A 25-day moving window was used as a training set for the NGR and Bayesian postprocessors.

	MAE			CRPS; ES		
Leadtime	**Superensemble**	**Bayesian**	**NGR**	**Superensemble**	**Bayesian**	**NGR**
2 days	1.2 *(0.9, 1.9)*	0.9 *(0.7, 1.0)*	0.7 *(0.6, 0.8)*	0.9 *(0.6, 1.3)*; 3.6	0.6 *(0.5, 0.7)*; 2.7	0.5 *(0.4, 0.6)*; 2.3
5 days	1.9 *(1.7, 2.4)*	1.6 *(1.4, 1.8)*	1.4 *(1.2, 1.5)*	1.3 *(1.2, 1.7)*; 5.4	1.2 *(1.0, 1.3)*; 4.8	1.0 *(0.8, 1.1)*; 4.2
10 days	2.7 *(2.4, 3.0)*	2.7 *(2.2, 2.9)*	2.3 *(1.9, 2.5)*	1.9 *(1.7, 2.2)*; 7.7	1.9 *(1.6, 2.0)*; 7.7	1.6 *(1.4, 1.7)*; 6.6
15 days	3.1 *(2.9, 3.3)*	3.1 *(2.5, 3.2)*	2.7 *(2.2, 2.9)*	2.1 *(2.0, 2.3)*; 8.5	2.2 *(1.8, 2.3)*; 8.6	1.9 *(1.5, 2.0)*; 7.6

NGR postprocessed forecasts generally have lower MAE and CRPS than their Bayesian posterior counterparts, with the MAE generally around 0.2-0.4°C lower. The lower CRPS achieved by the NGR forecasts is due partly to this improved accuracy, but also largely to the greater sharpness of those forecasts (Table 2). The NGR forecasts are in fact slightly too sharp at the shortest and longest leadtimes, with the 90% predictive interval found to contain the verifying observation in around 88% of all instances at all leadtimes. The Bayesian forecasts achieve around 94% coverage at the shortest leadtimes, dropping to around 83% at the longest, suggesting that the forecasts are initially overdispersive, covering too large an area, and later become overconfident.

Inspection of selected PIT histograms (Figure 6) gives a more detailed understanding of the calibration of the postprocessed forecasts than the summary scores can. The histograms shown for Kirkcaldy forecasts are typical of

TABLE 2 Mean *(min, max)* marginal and joint sharpness and coverage over all locations at selected leadtimes, for each postprocessing method. A 25-day moving window was used as a training set for the NGR and Bayesian postprocessors. Coverages are proportions of occasions for which the verifying observation fell between the 5th and 95th percentile of the corresponding marginal forecast distribution.

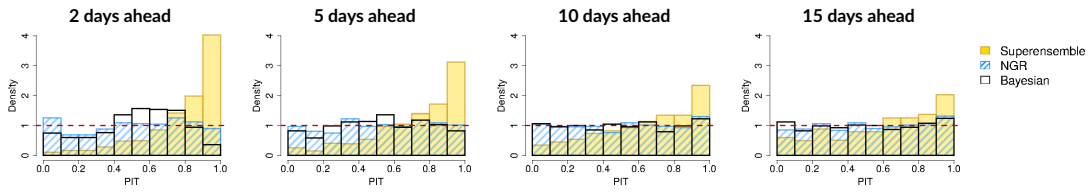
	Sharpness; determinant sharpness			Marginal coverage		
Leadtime	**Superensemble**	**Bayesian**	**NGR**	**Superensemble**	**Bayesian**	**NGR**
2 days	1.4 *(1.2, 1.8)*; 0.7	1.3 *(1.1, 1.6)*; 0.7	0.8 *(0.8, 1.0)*; 0.2	0.81 *(0.69, 0.94)*	0.94 *(0.91, 0.95)*	0.88 *(0.86, 0.90)*
5 days	2.3 *(2.0, 2.7)*; 1.0	2.2 *(1.9, 2.4)*; 0.8	1.7 *(1.4, 1.8)*; 0.4	0.88 *(0.81, 0.95)*	0.91 *(0.89, 0.92)*	0.90 *(0.87, 0.92)*
10 days	3.4 *(2.8, 3.8)*; 1.2	3.1 *(2.6, 3.3)*; 0.9	2.7 *(2.2, 2.9)*; 0.6	0.89 *(0.83, 0.92)*	0.86 *(0.84, 0.88)*	0.89 *(0.85, 0.90)*
15 days	3.7 *(3.1, 4.0)*; 1.1	3.4 *(2.8, 3.5)*; 0.9	3.1 *(2.5, 3.2)*; 0.7	0.89 *(0.84, 0.92)*	0.83 *(0.82, 0.86)*	0.87 *(0.86, 0.88)*

those obtained for predictions in Scotland and northern England, while those for Bristol are broadly representative of those obtained for forecasts in the south of the study area.

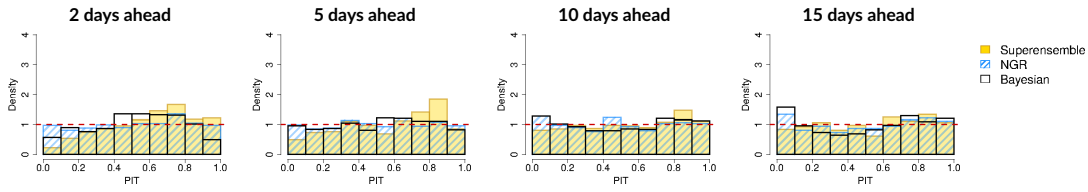
The raw superensemble PIT histograms show significant negative skew in Scotland, with a spike of values in the right-hand bin revealing that the observed temperature was very often in the extreme upper tail of the forecasts, indicating a persistent, systematic cold bias. Table 3 shows the range of skewness seen in the PIT histograms at all 13 locations; all of the superensemble histograms have some degree of negative skew at all leadtimes, indicating that the bias is common to forecasts throughout the UK. However, the bias is not constant at all locations, being larger in more northerly regions and at shorter leadtimes, with forecasts in Scotland being, on average, 1 – 1.5°C too low at the shortest leadtimes, and those in southern England between 0.4 and 1°C too low.

FIGURE 6 PIT histograms showing the marginal calibration of postprocessed forecasts of surface temperatures at selected locations in the north and south of the UK at various leadtimes. 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.

(a) Forecasts of temperatures in Kirkcaldy



(b) Forecasts of temperatures in Bristol



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. At the shortest leadtimes the NGR and Bayesian forecasts still

display a very slight residual bias, particularly (as in Figure 6a) in the northernmost regions where the raw ensemble bias is the strongest. The residual bias manifests in the PIT histograms as a slight bulge at around the 70th percentile of the transformed values, rather than as a tail of high 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 shapes of the Bayesian and NGR histograms at all locations (including those not shown here) are very similar, indicating that consistent improvements are achieved with both postprocessing methods.

The PIT histograms of the Bayesian postprocessed forecasts are slightly humped at the shortest leadtimes, having dispersion indices lower than 1, reflecting the fact that the forecasts are generally slightly underconfident in their predictions. The forecast overdispersion lessens with increasing leadtime, with longer-leadtime forecasts becoming slightly underdispersive. The NGR forecasts are well calibrated at all leadtimes, while the low dispersion indices of the superensemble PITs is due to the systematic bias, with a large concentration of values in the rightmost bin of the histogram.

TABLE 3 Mean *(min, max)* skewness and dispersion of PITs at all locations at selected leadtimes, for each postprocessing method. A 25-day moving window was used as a training set for the NGR and Bayesian postprocessors.

	PIT skewness			PIT dispersion		
Leadtime	**Superensemble**	**Bayesian**	**NGR**	**Superensemble**	**Bayesian**	**NGR**
2 days	-0.9 *(-1.4, -0.3)*	-0.2 *(-0.4, -0.1)*	-0.1 *(-0.2, 0.0)*	0.6 *(0.5, 0.8)*	0.8 *(0.7, 0.9)*	1.0 *(1.0, 1.1)*
5 days	-0.6 *(-1.0, -0.3)*	-0.1 *(-0.2, 0.0)*	-0.1 *(-0.1, 0.0)*	0.8 *(0.7, 0.9)*	0.9 *(0.9, 1.0)*	1.0 *(0.9, 1.1)*
10 days	-0.3 *(-0.6, -0.1)*	0.0 *(-0.1, 0.1)*	0.0 *(0.0, 0.1)*	0.9 *(0.8, 1.0)*	1.1 *(1.1, 1.2)*	1.1 *(1.0, 1.1)*
15 days	-0.3 *(-0.5, -0.1)*	-0.1 *(-0.1, 0.0)*	-0.1 *(-0.1, 0.0)*	1.0 *(0.9, 1.0)*	1.2 *(1.1, 1.2)*	1.1 *(1.0, 1.1)*

7.1.2 | Joint forecasts

The results for the energy score follow a similar pattern to those for the MAE (Table 1), with the NGR postprocessed forecasts having improved accuracy over the superensemble forecasts at all leadtimes, while the Bayesian forecasts were jointly more accurate only at shorter leadtimes. The NGR predictive densities are also jointly much sharper than either the Bayesian or superensemble forecasts, with a much lower determinant sharpness. This is partly due to the sharper marginal forecasts, but the NGR correlation matrices also generally specify stronger correlations between the forecast errors than either the superensemble or Bayesian postprocessed forecasts, further increasing the joint sharpness.

The modified band depth rank histograms in Figure 7a show the joint calibration of the postprocessed forecasts. The histograms for the superensemble and Bayesian joint forecasts are dominated by the marginal effects already discussed in Section 7.1.1, with the effects particularly obvious at the shortest leadtimes. Here, the peak at the right-hand side of the distribution represents a high proportion of the observations falling close to the centre of the Bayesian posterior predictive distributions, indicating that the forecasts are jointly overdispersive; although the effect in any single marginal forecast is quite small, the cumulative effect is magnified in the joint forecast. As in the marginal forecasts, the effect reduces and reverses with increasing leadtime, with a preponderance of points in the leftmost bin indicating too many points falling too far from the centre of the distribution, and reflecting the underdispersiveness of the marginal forecasts. At these longer leadtimes, apart from the spikes in the leftmost or rightmost bins, the Bayesian BDR histograms are fairly uniform, and give no indication of any misspecification in the correlation structure of the

forecasts.

The raw superensemble's systematic marginal cold bias is reflected in heavy over-population of the lowest ranks of the histograms, as the observation frequently falls far from the centre of the forecast distribution. As the leadtime increases, the superensemble histograms become close to uniform, but care must be taken to remember the marginal calibration issues when interpreting this. The PIT skewness indicates that, even at 15 days' leadtime, there is still some residual systematic cold bias in the forecasts, which we would expect to result in a peak of values in the leftmost bins of the histogram. At the same time, the superensemble forecasts are less sharp than the Bayesian posterior forecasts (Table 2), which we have already seen to be overdispersive at these leadtimes, and which would manifest in a peak of values in the rightmost bin of the histogram. It seems reasonable to conclude, then, that the too-large spread of the joint forecasts is, to some extent, counteracting the bias at all leadtimes, and that the joint superensemble forecasts are in fact both biased and overdispersive, although there is no reason to suspect that the correlations between the variables are misspecified.

The NGR histograms display a different problem: the $\cap$-shaped histograms, particularly at shorter leadtimes, indicate that the correlations specified by the predictive distributions are too high, with too many observations falling in the 'shoulders' of the distribution, rather than in the centre and the fringes of the joint distribution. This is because, being based purely on a sample of climatology as described in Section 3.2, the NGR correlation structure is unable to adapt fully to the specific features of each individual forecast, while the Bayesian correlations are based on a mixture of the forecasts and previous forecast errors. Bootstrapped confidence intervals indicate that, while the histograms look quite symmetric, there is a small but persistent positive skewness at all leadtimes, indicating that the joint forecasts are also slightly too sharp, with too many observations falling far from the centre of the predictive distributions. Thus, although the marginal distributions appear to be well calibrated, their cumulative effect is to produce a slightly underdispersive forecast.

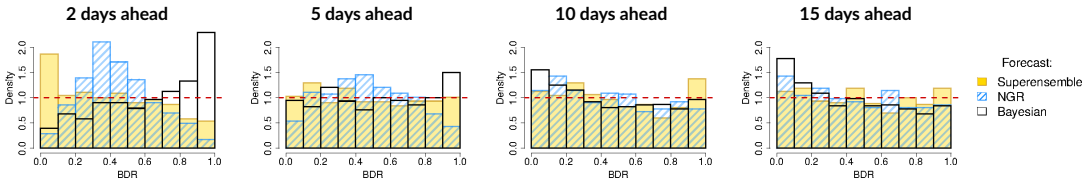
A further investigation was carried out to investigate how well the correlation structure of the observations is captured by each postprocessing method. To fully separate the effect of marginal calibration from joint calibration, each set of NGR marginal forecasts was combined with the correlation matrices of the superensemble and Bayesian posterior forecasts, and the joint calibration assessed again. Figure 7b shows the BDR histograms obtained from these hybrid forecasts. At shorter leadtimes, forecasts taking their dependence structure only from either the superensemble or the training set are rather over-correlated, producing peaked BDR histograms; forecasts using the Bayesian posterior correlation matrix as their copula are close to uniform, and show no significant dispersion issues. At longer leadtimes, all three copula methods produce similar BDR histograms, with the only significant departure from uniformity being the skewness resulting from a preponderance of observations falling in the tails of the distribution, due to the underdispersiveness of the NGR marginal distributions mentioned above.

7.2 | Effect of training set selection

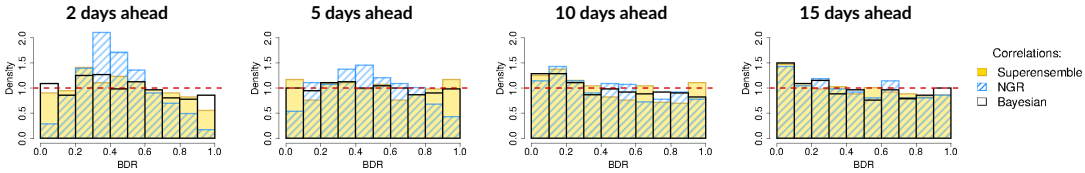
We now consider the effect that the choice of training set has on forecast performance. Results are presented for Bayesian postprocessed forecasts, with the discrepancy estimated using training sets obtained by a moving window (MW), direct analogues (DA), and by weather regime analogues (WR), as described in Section 5. All training sets consisted of 25 members. Forecasts postprocessed by applying NGR to the same training sets showed a similar pattern of results, which are not reported here. Each selection method was found to produce quite different collections of forecast-observation pairs, with training sets having on average around 10% of their members (2-3 instances) in common with their counterparts.

FIGURE 7 Modified band depth rank (BDR) histograms showing the joint calibration of postprocessed forecasts of surface temperatures across all grid cells at various leadtimes. 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.

(a) Postprocessed forecasts



(b) NGR marginals combined with superensemble and Bayesian posterior correlation functions



7.2.1 | Marginal forecasts

All three training sets obtain almost identical MAE, CRPS, and ES at all leadtimes (Table 4), with MAE differences between the three forecasts of extremely small magnitude (generally only around 0.1°C), and only found to be statistically significant at leadtimes greater than 8 or 9 days. Conversely, the CRPS for the DA and WR training sets is slightly (but, again, significantly) higher at these longer leadtimes; this is because the analogue-trained forecasts are less sharp than those using MW training sets at those leadtimes (Table 5). This decreased sharpness leads to improved coverage at leadtimes greater than 10 days, with nominal 90% predictive intervals for both DA and WR-trained sets achieving an average of 88% coverage at this range. The PIT histograms in Figure 8 reflect this improved coverage,

TABLE 4 Mean *(min, max)* MAE and CRPS (in °C), and ES over all locations, at selected leadtimes, for Bayesian posterior forecasts using each training set.

	MAE			CRPS; ES		
Leadtime	**MW**	**DA**	**WR**	**MW**	**DA**	**WR**
2 days	0.9 *(0.7, 1.0)*	0.8 *(0.7, 1.0)*	0.9 *(0.8, 1.1)*	0.6 *(0.5, 0.7)*; 2.7	0.6 *(0.5, 0.7)*; 2.7	0.7 *(0.6, 0.8)*; 2.8
5 days	1.6 *(1.4, 1.8)*	1.6 *(1.4, 1.8)*	1.7 *(1.5, 1.8)*	1.2 *(1.0, 1.3)*; 4.8	1.1 *(1.0, 1.3)*; 4.8	1.2 *(1.1, 1.3)*; 4.9
10 days	2.7 *(2.2, 2.9)*	2.7 *(2.1, 2.8)*	2.6 *(2.1, 2.8)*	1.9 *(1.6, 2.0)*; 7.7	1.9 *(1.5, 2.0)*; 7.5	1.9 *(1.5, 2.0)*; 7.5
15 days	3.1 *(2.5, 3.2)*	2.9 *(2.4, 3.1)*	3.0 *(2.6, 3.2)*	2.2 *(1.8, 2.3)*; 8.6	2.1 *(1.7, 2.2)*; 8.3	2.1 *(1.9, 2.2)*; 8.4

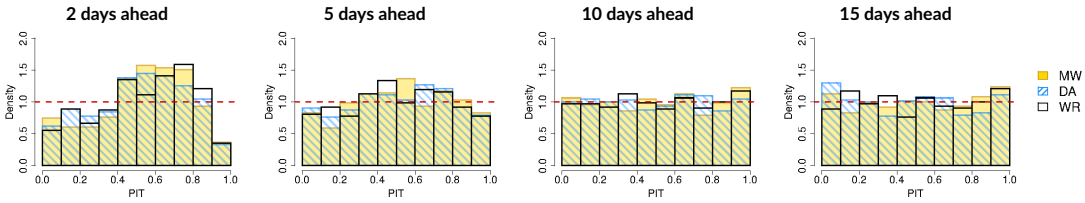
with the DA and WR histograms being closer to uniform - and having dispersion indices closer to 1 - after around 7 days (Table 6); MW overdispersiveness continues to increase with leadtime, but this does not occur with the DA and WR forecasts. Furthermore, at the shortest leadtimes, the DA and WR PIT histograms are both less skewed and closer to uniformity than the MW histograms, suggesting that replacing a training set that assumes persistence in the forecast errors with a training set that assumes flow dependence of the errors may be able to improve both the short-term overdispersiveness and longer-term underdispersiveness of the forecasts.

TABLE 5 Mean *(min, max)* marginal and joint sharpness and coverage over all locations at selected leadtimes, for Bayesian posterior forecasts using each training set. Coverages are proportions of occasions for which the verifying observation fell between the 5th and 95th percentile of the corresponding forecast distribution.

	Sharpness; determinant sharpness			Marginal coverage		
Leadtime	**MW**	**DA**	**WR**	**MW**	**DA**	**WR**
2 days	1.3 *(1.1, 1.6)*; 0.7	1.2 *(1.1, 1.7)*; 0.6	1.3 *(1.2, 1.7)*; 0.7	0.94 *(0.91, 0.95)*	0.93 *(0.91, 0.95)*	0.94 *(0.91, 0.96)*
5 days	2.2 *(1.9, 2.4)*; 0.8	2.0 *(1.8, 2.4)*; 0.8	2.2 *(1.9, 2.5)*; 0.8	0.91 *(0.89, 0.92)*	0.90 *(0.88, 0.92)*	0.91 *(0.90, 0.93)*
10 days	3.1 *(2.6, 3.3)*; 0.9	3.2 *(2.7, 3.4)*; 0.9	3.2 *(2.6, 3.5)*; 0.9	0.86 *(0.84, 0.88)*	0.88 *(0.87, 0.90)*	0.88 *(0.87, 0.90)*
15 days	3.4 *(2.8, 3.5)*; 0.9	3.5 *(3.0, 3.7)*; 0.9	3.7 *(3.1, 3.8)*; 0.9	0.83 *(0.82, 0.86)*	0.88 *(0.86, 0.90)*	0.89 *(0.87, 0.90)*

FIGURE 8 PIT histograms showing the marginal calibration of Bayesian postprocessed forecasts of surface temperatures in the north and south of the UK at various leadtimes, comparing the performance of training sets selected using a moving window, direct analogue or principal-component analogue approach. The dashed line indicates the ideal uniform distribution.

(a) Forecasts of temperatures in Kirkcaldy



(b) Forecasts of temperatures in Bristol

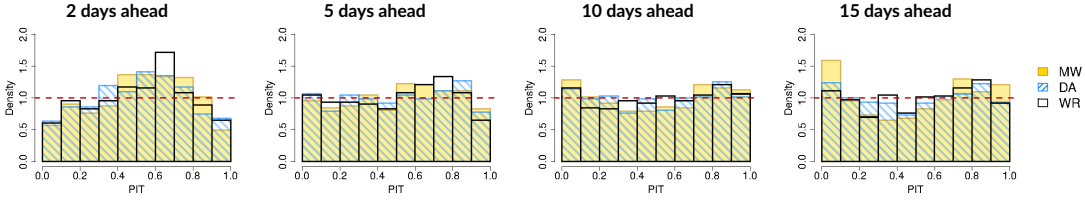


TABLE 6 Mean *(min, max)* skewness and dispersion of PIT histograms at all locations at selected leadtimes, for Bayesian posterior forecasts using each training set.

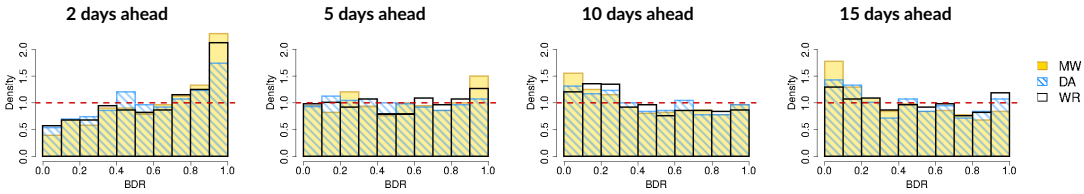
	PIT skewness			PIT dispersion		
Leadtime	**MW**	**DA**	**WR**	**MW**	**DA**	**WR**
2 days	-0.2 *(-0.4, -0.1)*	-0.1 *(-0.3, 0.0)*	-0.1 *(-0.3, -0.1)*	0.8 *(0.7, 0.9)*	0.8 *(0.7, 0.9)*	0.8 *(0.8, 0.9)*
5 days	-0.1 *(-0.2, 0.0)*	-0.1 *(-0.2, 0.0)*	0.0 *(-0.2, 0.0)*	0.9 *(0.9, 1.0)*	1.0 *(0.9, 1.0)*	0.9 *(0.9, 1.0)*
10 days	0.0 *(-0.1, 0.1)*	0.0 *(-0.1, 0.1)*	0.0 *(-0.1, 0.1)*	1.1 *(1.1, 1.2)*	1.1 *(1.0, 1.1)*	1.1 *(1.0, 1.1)*
15 days	-0.1 *(-0.1, 0.0)*	0.0 *(-0.1, 0.1)*	-0.1 *(-0.2, 0.0)*	1.2 *(1.1, 1.2)*	1.1 *(1.0, 1.1)*	1.0 *(1.0, 1.1)*

7.2.2 | Joint forecasts

As with their marginal score equivalents, the three training sets produce forecasts with very similar energy scores (Table 4), with the DA-trained forecasts jointly slightly sharper than either the MW or WR forecasts at the shortest

leadtimes, and the MW-trained forecasts slightly sharper at longer leadtimes (Table 5). Although the change in these summary scores is small, their impact is clear in the BDR histograms in Figure 9, with the DA histograms in particular having a reduced number of values in the rightmost bin at shorter leadtimes. As the leadtime increases and the MW forecasts become jointly more underdispersive, the DA and WR histograms are again more symmetric, supporting the conclusion in Section 7.2.1 that the flow-dependent training sets produce forecasts with better dispersion characteristics.

FIGURE 9 Modified band depth rank (BDR) histograms showing the spatial calibration of Bayesian postprocessed forecasts of surface temperatures across all grid cells at various leadtimes. The dashed line indicates the ideal uniform distribution.



Bootstrapped confidence intervals indicate that the dispersion indices for BDR histograms for all three training sets are generally close to 1; all three were found to have significant departures from uniformity only at the longest leadtimes, and those departures were very close to the threshold of the 5% significance level, suggesting that any correlation misspecification in any of the Bayesian postprocessed forecasts is very slight.

8 | SUMMARY AND DISCUSSION

We have proposed a novel postprocessing method for multi-model ensemble forecasts, based on an understanding of the relationships between the component ensembles and the true forecast. The specification presented here may be applied in its current form to any variables for which the forecasts may be assumed to have an approximately multivariate-normal distribution; even where this is not the case, the posterior mean gives the optimum linear combination of the available forecast information, with the posterior covariance matrix summarising the uncertainty about the mean (Goldstein and Wooff, 2007). The postprocessed forecasts are significantly less biased and better calibrated, both marginally and jointly, than those of the raw superensemble, and the spatial dependencies of the Bayesian postprocessed forecasts better capture the correlation structure of the observations than a copula based on either the ensembles or a sampled climatology alone, particularly at leadtimes of up to one week.

The NGR postprocessed marginal forecasts are sharper than those produced by the Bayesian framework in its present form, having consistently lower MAE and CRPS; however, this improvement comes at a high computational cost due to the numerical optimisation required. Postprocessing of all available forecast instances using the `ensembleMOS` package in R (Yuen et al., 2017) was found to take around 30 times as long as producing the Bayesian posterior forecasts, even over this relatively small spatial domain, and with longer postprocessing times recorded as larger training sets were tested. Thus, if computational cost is an issue, we would always recommend using the Bayesian framework in preference to the NGR approach, unless very high accuracy is critical.

The implementation of the Bayesian framework presented here is a fairly naive one, in that all forecast instances are postprocessed independently of one another, and the distribution of the consensus discrepancy $\boldsymbol{\Delta}$ is simply estimated from the means and covariances of the forecast errors in the training data. However, part of the appeal of

the Bayesian approach is its potential flexibility. The full Bayesian framework specified in equations (11)-(12) includes a prior estimate of the distribution of the vector of 'true' weather quantities, which is omitted in the single-instance implementation used here. However through this prior, the Bayesian approach lends itself naturally to sequential postprocessing of a sequence of forecasts, with one day's t-day-ahead posterior forecast providing the next day's $(t-1)$-day-ahead prior, for example.

Similarly, a more sophisticated approach to estimation of $\boldsymbol{\eta}$ and $\boldsymbol{\Lambda}$ might be expected to produce further improvements in forecast skill. Instead of estimating the parameters directly from historical forecast errors as we have done here, the distribution of $\boldsymbol{\Delta}$ could be obtained using Bayesian inference over the training data. This, again, offers scope for inclusion of an informative prior, perhaps reflecting expert judgement of the expected correlation structure. Inference of the distribution of $\boldsymbol{\Delta}$ is of particular interest, because the resulting posterior distribution would no longer be multivariate normal, but would instead describe a multivariate t distribution; this may better reflect the distribution of the observations and so produce better-calibrated forecasts, particularly at longer leadtimes, where the Bayesian posterior forecasts currently tend to be slightly underdispersive.

One limitation of the proposed framework is its scalability: as presented here, the approach can only be applied to relatively small spatial domains, due to the difficulty of estimating the necessary covariance matrices from the small ensembles available. When the dimension p is greater than the number of members of any of the member ensembles, the estimate of the corresponding covariance matrix $\mathbf{C}_i$ will be singular, and the posterior cannot be evaluated. Further work is planned to investigate alternative methods of estimating the spatial covariance matrices to allow the method to be applied to larger regions. Likewise, $\boldsymbol{\Sigma}$ is estimated from only m 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: research into this possibility is ongoing.

The approach could also be extended to postprocess low-resolution forecasts with $\boldsymbol{\Delta}$ estimated from a training set matching past forecasts to higher-resolution observations, providing scope for applications to forecast downscaling. In this case, the low resolution can itself be regarded as a source of shared discrepancies in the forecast ensembles, so that the same framework applies.

In addition to the new postprocessing framework, we suggest a new method for selection of a training set of synoptic-scale analogues. Forecasts postprocessed using analogues selected in MSLP principal component space (WR) and analogues selected in variable space (DA) were found to have joint and marginal calibration comparable to or better than forecasts postprocessed using a moving-window (MW) training set, suggesting that choosing a training set based on an assumption of flow-dependence, rather than persistence, of forecast errors produces better-calibrated forecasts. Since there is no evidence that one single method of training set selection can be said to be universally 'best', selection of the most appropriate approach is likely to depend on the application.

The MW training set is convenient to obtain, requiring no additional archive of candidate forecasts; for the postprocessing of forecasts in a relatively small area, at leadtimes of less than a week, it remains a reasonable choice. However, if the forecast area is increased, a larger training set will be required for estimation of the necessary covariance matrices; simply increasing the size of an MW training set will eventually result in a reduction in the quality of the postprocessed forecasts, due to the increasing remoteness of the training cases from the forecast under consideration. Similarly, when postprocessing larger regions, selection of direct analogues will become impractical due to the corresponding increase in the size of the analogue search space; a great advantage of the WR method is that the dimension of the candidate search space remains relatively small. Perhaps most usefully, if the forecast region changes slightly (within the bounds of the region to which PCA was applied), new analogues do not need to be identified - thus, forecasts postprocessed using the Bayesian framework with a synoptic-scale WR analogue training set can be

expected to remain stable even if they are recalculated as part of a different forecast domain. In addition, while the DA-postprocessed forecasts performed slightly better over the collection of surface temperature forecasts examined in the case study, a WR training set may produce more skilful forecasts when several different weather quantities are to be postprocessed together.

Acknowledgements

The authors would like to thank Helen Dacre and Andrew Charlton-Perez at the University of Reading for sharing their thoughts and expertise during the development of this work, and two anonymous reviewers whose perceptive comments led to substantial improvements to the final paper. This project was supported by the Engineering and Physical Sciences Research Council under grant EP/N509577/1.

Supporting information

A full derivation of the framework described in Section 4 is provided in an electronic supplement to this paper.

references

Atger, F. (2003) Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, **131**.

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.

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**, 2692–2709.

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**, 988–1002.

Bernardo, J. and Smith, A. (2001) Bayesian theory.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007) Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, **135**, 1386–1402.

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**, 1059–1072.

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: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **134**, 703–722.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78**, 1–3.

Casella, G. and Berger, R. L. (2002) *Statistical inference*, vol. 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*, 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**, 243–262.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical statistics*. Chapman & Hall, London.

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**, 553–597.

Della-Marta, P. M., Luterbacher, J., von Weissenfluh, H., Xoplaki, E., Brunet, M. and Wanner, H. (2007) Summer heat waves over western Europe 1880–2003, their relationship to large-scale forcings and predictability. *Climate Dynamics*, **29**, 251–275.

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**, 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**, 3554–3570.

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**, 234–252.

Eckel, F. A. and Mass, C. F. (2005) Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, **20**, 328–350.

Eckel, F. A. and Walters, M. K. (1998) Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting*, **13**, 1132–1147.

Efron, B. and Tibshirani, R. J. (1994) *An introduction to the bootstrap*. CRC press.

Feldmann, K., Scheuerer, M. and Thorarinsdottir, T. L. (2015) Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, **143**, 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**, 916–924.

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**, 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. *Tech. rep.*, DTIC Document.

Glahn, H. R. and Lowry, D. A. (1972) The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11**, 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**, 243–268.

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**, 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**, 211–235.

Goldstein, M. and Wooff, D. (2007) *Bayes linear statistics: Theory and methods*, vol. 716. John Wiley & Sons.

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**, 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**, 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**, 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**, 2608–2619.

Hamill, T. M. (1999) Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14**, 155–167.

— (2012) Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, **140**, 2232–2252.

Hamill, T. M. and Colucci, S. J. (1997) Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312–1327.

Hamill, T. M. and Whitaker, J. S. (2006) Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, **134**, 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**, 33.

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–570.

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**, 96–102.

Johnson, C. and Swinbank, R. (2009) Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, **135**, 777–794.

Jolliffe, I. T. (2011) Principal component analysis. In *International encyclopedia of statistical science*, 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**, 655–663.

Jordan, A., Krueger, F. and Lerch, S. (2018) Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*. Forthcoming.

Junk, C., Delle Monache, L. and Alessandrini, S. (2015) Analog-based ensemble model output statistics. *Monthly Weather Review*, **143**, 2909–2917.

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**, 29–51.

López-Pintado, S. and Romo, J. (2009) On the concept of depth for functional data. *Journal of the American Statistical Association*, **104**, 718–734.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, **139**, 982–991.

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**, 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**, 699–706.

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.

Pinson, P. and Tastu, J. (2013) Discrimination ability of the energy score. *Tech. rep.*, Technical University of Denmark.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Roulston, M. and Smith, L. A. (2003) Combining dynamical and statistical ensembles. *Tellus A*, **55**, 16–30.

Schefzik, R. (2016) A similarity-based implementation of the Schaake Shuffle. *Monthly Weather Review*, **144**, 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**, 616–640.

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**, 2375–2393.

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**, 105–122.

Weigel, A., Liniger, M. and Appenzeller, C. (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, **134**, 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**, 1460–1479.

Wilks, D. S. (2002) Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, **128**, 2821–2836.

— (2011) *Statistical methods in the atmospheric sciences*. Academic Press, third edn.

— (2018) Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, **144**, 76–84.

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**, 956–970.

Yuen, R., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M. and Thorarinsdottir, T. (2017) *ensembleMOS: Ensemble Model Output Statistics*. URL: `https://CRAN.R-project.org/package=ensembleMOS`. R package version 0.8.