Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging

CHRIS FRALEY AND ADRIAN E. RAFTERY

University of Washington, Seattle, Washington

TILMANN GNEITING

Universitat Heidelberg, Heidelberg, Germany

(Manuscript received 28 April 2009, in final form 14 July 2009)

ABSTRACT

Bayesian model averaging (BMA) is a statistical postprocessing technique that generates calibrated and sharp predictive probability density functions (PDFs) from forecast ensembles. It represents the predictive PDF as a weighted average of PDFs centered on the bias-corrected ensemble members, where the weights reflect the relative skill of the individual members over a training period.

This work adapts the BMA approach to situations that arise frequently in practice; namely, when one or more of the member forecasts are exchangeable, and when there are missing ensemble members. Exchangeable members differ in random perturbations only, such as the members of bred ensembles, singular vector ensembles, or ensemble Kalman filter systems. Accounting for exchangeability simplifies the BMA approach, in that the BMA weights and the parameters of the component PDFs can be assumed to be equal within each exchangeable group. With these adaptations, BMA can be applied to postprocess multimodel ensembles of any composition.

In experiments with surface temperature and quantitative precipitation forecasts from the University of Washington mesoscale ensemble and ensemble Kalman filter systems over the Pacific Northwest, the proposed extensions yield good results. The BMA method is robust to exchangeability assumptions, and the BMA postprocessed combined ensemble shows better verification results than any of the individual, raw, or BMA postprocessed ensemble systems. These results suggest that statistically postprocessed multimodel ensembles can outperform individual ensemble systems, even in cases in which one of the constituent systems is superior to the others.

1. Introduction

Bayesian model averaging (BMA) was introduced by Raftery et al. (2005) as a statistical postprocessing method that generates calibrated and sharp predictive probability density functions (PDFs) from ensemble systems. The BMA predictive PDF of any future weather quantity of interest is a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights reflect the predictive skill of the member forecasts over a training period. The initial development of BMA was for weather quantities for which the forecast errors are approximately Gaussian, such as surface temperature and sea level pressure (Raftery et al. 2005; Wilson et al. 2007a). The approach was extended by Sloughter et al. (2007, 2009) to apply to skewed weather variables, such as quantitative precipitation and wind speed. For all variables considered, and for both mesoscale ensembles and synoptic ensembles, the BMA postprocessed PDFs outperformed the unprocessed ensemble forecast and were calibrated and sharp.

This work extends the BMA approach to accommodate situations that arise frequently in practice, namely, when one or more of the member forecasts are exchangeable and when there are missing ensemble members. We show how BMA can be adapted to handle these situations, and demonstrate good performance for the proposed extensions. There is considerable recent interest in the use of multimodel ensembles as a means of improving deterministic and probabilistic forecast skill (e.g., Hagedorn et al. 2005; Doblas-Reyes et al. 2005;

Corresponding author address: Chris Fraley, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
E-mail: fraley@stat.washington.edu

DOI: 10.1175/2009MWR3046.1

Park et al. 2008; Weigel et al. 2008). Our extensions allow for the application of the BMA technique to any type and configuration of multimodel ensemble. Contrary to earlier studies, our results show that statistically postprocessed multimodel ensembles are likely to outperform any of the raw or postprocessed constituent ensembles.

Exchangeable members lack individually distinguishable physical features. Examples include members of the bred or singular vector synoptic ensembles used by the National Centers for Environmental Prediction and the European Centre for Medium-Range Weather Forecasts (Toth and Kalnay 1993; Molteni et al. 1996), and the members of ensemble Kalman filter systems (Evensen 1994; Hamill 2005). Accounting for exchangeability facilitates postprocessing, in that the BMA weights and model parameters can be constrained to be equal within each exchangeable group. This simplifies the BMA approach, and speeds up the associated computations.

Missing ensemble members typically stem from disruptions in communications, and software or hardware failures, which can last for days or weeks at a time. The chance that any member forecast is missing increases with the ensemble size and ensemble diversity. Thus, with the current trend toward larger and multimodel ensembles, a considerable amount of potentially useful information would be ignored if instances with missing members were excluded from training sets.

The remainder of the paper is organized as follows. The basic BMA framework is reviewed in section 2, where we also show how to estimate the BMA parameters via the expectation-maximization (EM) algorithm. Section 3 shows how the BMA approach can be modified to properly account for exchangeability, and gives verification results for 24-h surface temperature forecasts from the University of Washington (UW) mesoscale ensemble (ME), ensemble Kalman filter (EnKF), and ME–EnKF combined systems over the Pacific Northwest (Grimit and Mass 2002; Eckel and Mass 2005; Dirren et al. 2007; Torn and Hakim 2008). Section 4 describes how missing ensemble members can be handled in estimation and prediction, and gives verification results with these methods for 48-h surface temperature and quantitative precipitation forecasts from the UW ME system. The paper ends with a discussion in section 5.

2. Ensemble postprocessing using Bayesian model averaging

In BMA for ensemble forecasting (Raftery et al. 2005) each ensemble member forecast, f_i, is associated with a component PDF, $g_i(y|f_i, \theta_i)$, where y represents the weather quantity of interest, and θ_i comprises the parameters of the ith component PDF. The BMA predictive PDF then is a mixture of the component PDFs:

$$p(y|f_1, \ldots, f_m; \theta_1, \ldots, \theta_m) = \sum_{i=1}^{m} w_i g_i(y|f_i, \theta_i), \qquad (1)$$

where the BMA weight w_i is based on ensemble member i's relative performance in the training period. The w_is are probabilities and so they are nonnegative and add up to 1, that is, $\sum_{i=1}^{m} w_i = 1$. Here m is the number of forecasts in the ensemble.

The component PDF $g_i(y|f_i)$ can be thought of as the conditional PDF of the weather quantity y given that ensemble member i provides the most skillful forecast. This heuristic interpretation is in line with how operational weather forecasters often work, by selecting one or a small number of ''best'' forecasts from a potentially large number available, based on recent predictive performance (Joslyn and Jones 2008). For weather variables such as temperature and sea level pressure, which have approximately Gaussian errors, the component PDFs can be taken to be normal distributions centered at bias-corrected ensemble member forecasts, as shown by Raftery et al. (2005). We refer to this case as the Gaussian mixture model. For quantitative precipitation, Sloughter et al. (2007) model the component PDFs using a mixture of a point mass at zero and a power-transformed gamma distribution. For wind speed, Sloughter et al. (2009) propose the use of gamma components.

Certain member-specific parameters of the BMA model are estimated individually and at an initial stage, prior to applying the EM algorithm that we describe here. For example, in the Gaussian mixture model, bias correction is done for each ensemble member individually by fitting a standard linear regression model to the training data (Raftery et al. 2005). For quantitative precipitation (Sloughter et al. 2007), a logistic regression model for the probability of precipitation is fit for each ensemble member. These initial procedures allow for straightforward adaptation when ensemble members are exchangeable and/or missing, in that training sets for exchangeable members are merged to estimate a single, common regression model for each exchangeable group, and instances with the given ensemble member missing are excluded from the training set. These adaptations are immediate and will not be further discussed here.

The BMA weights, w_i, and the remaining parameters, θ_i, of the component PDFs are estimated by maximum likelihood (Wilks 2006) from training data. Typically, the training set comprises a temporally and/or spatially composited collection of past ensemble forecasts, $f_{1,s,t}, \ldots, f_{m,s,t}$, and the respective verifying observation, $y_{s,t}$, at location or station s and time t. The likelihood

function, ℓ, is then defined as the probability of the training data, viewed as a function of the w_is and θ_is:

$$\ell(w_1, \ldots, w_m; \theta_1, \ldots, \theta_m) = \prod_{s,t} \sum_{i=1}^{m} w_i g_i(y_{s,t}|f_{i,s,t}, \theta_i),$$

where the product extends over all instances (s, t) in the training set. The maximum likelihood estimates are those values of the w_is and θ_is that maximize the likelihood function, that is, the values under which the verifying observations were most likely to materialize.

The likelihood function typically cannot be maximized analytically, and so it is maximized using the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). The EM algorithm is iterative, and alternates between two steps, the expectation or E step, and the maximization or M step. For mixture models, it uses unobserved quantities $z_{i,s,t}$, which can be interpreted as the probability of ensemble member i being the most skillful forecast for location s at time t. The $z_{1,s,t}, \ldots, z_{m,s,t}$ are nonnegative and sum to 1 for each instance (s, t) in the training set.

In the E step, the $z_{i,s,t}$ are estimated given the current values of the BMA weights and component PDFs; specifically,

$$z_{i,s,t}^{(k+1)} = \frac{w_i^{(k)} g_i(y_{s,t}|f_{i,s,t}, \theta_i^{(k)})}{\sum_{p=1}^{m} w_p^{(k)} g_p(y_{s,t}|f_{p,s,t}, \theta_p^{(k)})}, \quad (2)$$

where the superscript "(k)" refers to the kth iteration of the EM algorithm, and thus $w_p^{(k)}$ and $\theta_p^{(k)}$ refer to the estimates at the kth iteration.

The M step then consists of maximizing the expected complete-data log likelihood as a function of the w_is and θ_is, where the expectation is conditional on the training data and the previous estimates. The expected complete-data log likelihood is the sum of two terms, one of which involves the w_is and not the θ_is, and the other of which involves the θ_is but not the w_is (Dempster et al. 1977). We refer to the second term as the partial expected complete-data log likelihood.

Maximizing the partial expected complete-data log likelihood as a function of the w_is yields updated estimates,

$$w_i^{(k+1)} = \frac{1}{n} \sum_{s,t} z_{i,s,t}^{(k+1)}, \quad (3)$$

of the BMA weights, where n is the total number of instances in the training set.

Updated estimates $\theta_1^{(k+1)}, \ldots, \theta_m^{(k+1)}$ of the component parameters are obtained by maximizing the partial expected complete-data log likelihood:

$$\sum_{s,t} \left[\sum_{i=1}^{m} z_{i,s,t}^{(k+1)} \log(g_i(y_{s,t}|f_{i,s,t}, \theta_i)) \right], \quad (4)$$

over $\theta_1, \ldots, \theta_m$, where the BMA weights are fixed at their current estimates. For the Gaussian mixture model, this optimization can be done analytically (Raftery et al. 2005). For the gamma mixtures that apply to quantitative precipitation and wind speed (Sloughter et al. 2007, 2009) numerical optimization is required. The E step and M step are then alternated iteratively to convergence.

Vrugt et al. (2008) compared the EM algorithm to a fully Bayesian, Markov chain Monte Carlo (MCMC)-based method for fitting the BMA parameters in the Gaussian mixture model. The MCMC approach has considerable flexibility and provides potentially useful information about the uncertainty associated with the estimated BMA weights and variances. However, although much more computationally intensive than EM estimation, the MCMC method gave similar predictive results.

The standard form of the EM algorithm assumes that none of the ensemble member forecasts, $f_{i,s,t}$, are missing, and that the ensemble members are individually distinguishable, so that distinct weights can be physically interpreted. In what follows, we extend BMA to cases in which the members can be grouped into subsets of exchangeable forecasts and/or some of the ensemble member forecasts are missing.

3. Accommodating exchangeable ensemble members

The methodology we have described so far has been for a situation where the ensemble members come from clearly distinguishable sources. We now show how to modify it to deal with situations in which some or all of the ensemble members are statistically indistinguishable, differing only in some random perturbations. Most of the synoptic ensembles that are currently in use are of this type (Buizza et al. 2005; Park et al. 2008). In these cases, members that are statistically indistinguishable should be treated as exchangeable, and thus should have equal BMA weight and equal BMA parameter values. The Gaussian mixture model for temperature already constrains the standard deviation parameters to be equal across all members (Raftery et al. 2005), so the only changes in this case are the added constraints that the BMA mean or bias parameters and the BMA weights be equal, as described by Wilson et al. (2007b). In the gamma models for quantitative precipitation (Sloughter et al. 2007) and wind speed (Sloughter et al. 2009), the variance parameters need to be constrained to be equal within exchangeable groups of ensemble members.

a. Estimation with exchangeable ensemble members

Consider an ensemble of size m, where the member forecasts can be divided into groups of exchangeable

members. We assume that there are I such groups, with the ith exchangeable group having $m_i \geq 1$ exchangeable members, so that $\sum_{i=1}^{I} m_i = m$. Let $f_{i,j}$ denote the jth member of the ith group. Then the BMA mixture distribution (1) can be rewritten to accommodate groups of exchangeable members:

$$p(y|\{f_{i,j}\}_{i=1,\ldots,I,j=1,\ldots,m_i}; \{\theta_i\}_{i=1,\ldots,I}) = \sum_{i=1}^{I}\sum_{j=1}^{m_i} w_i\, g_i(y|f_{i,j}, \theta_i), \quad (5)$$

where $w_i \geq 0$ and $\sum_{i=1}^{I} w_i = 1$. Note that BMA weights, probability functions and parameters are equal within each exchangeable group. For EM estimation, the E step in Eq. (2) can now be written as

$$z_{i,j,s,t}^{(k+1)} = \frac{w_i^{(k)} g_i(y_{s,t}|f_{i,j,s,t}, \theta_i^{(k)})}{\sum_{p=1}^{I}\sum_{q=1}^{m_p} w_p^{(k)} g_p(y_{s,t}|f_{p,q,s,t}, \theta_p^{(k)})}. \quad (6)$$

In the M step, the weight estimates are averaged over the ensemble members in the respective exchangeable group:

$$w_i^{(k+1)} = \frac{1}{n}\frac{1}{m_i}\sum_{s,t}\sum_{j=1}^{m_i} z_{i,j,s,t}^{(k+1)}, \quad (7)$$

and updated estimates, $\theta_1^{(k+1)}, \ldots, \theta_I^{(k+1)}$, of the component parameters, are obtained by maximizing the partial expected complete-data log likelihood, which takes the following form:

$$\sum_{s,t}\left[\sum_{i=1}^{I}\sum_{j=1}^{m_i} z_{i,j,s,t}^{(k+1)} \log(g_i(y_{s,t}|f_{i,j,s,t}, \theta_i))\right]. \quad (8)$$

b. Application to University of Washington ensemble systems

In the experiment below, we use both the UW ME (Grimit and Mass 2002; Eckel and Mass 2005) and the recently developed UW EnKF systems (Torn et al. 2006; Dirren et al. 2007; Torn and Hakim 2008) over the Pacific Northwest and adjacent areas of the Pacific Ocean.

The version of the UW ME used here is an eight-member multianalysis ensemble based on the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model (MM5). The MM5 model is run with 36-km horizontal grid spacing over western North America and the eastern North Pacific Ocean, driven by initial and lateral boundary conditions from eight distinct global models. A summary description of the ensemble members is given in Table 1. Thus, the UW ME system consists of eight nonexchangeable members.

TABLE 1. Composition of the eight-member University of Washington Mesoscale Ensemble (UW ME; Eckel and Mass 2005), with member acronyms, and organizational and synoptic model sources for initial and lateral boundary conditions. The organizational sources are the National Centers for Environmental Prediction (NCEP), the Canadian Meteorological Centre (CMC), the Australian Bureau of Meteorology (ABM), the Japanese Meteorological Agency (JMA), the Fleet Numerical Meteorology and Oceanography Center (FNMOC), the Taiwan Central Weather Bureau (TCWB), and the Met Office (UKMO).

Member	Source	Driving synoptic model
GFS	NCEP	Global Forecast System
ETA	NCEP	Limited-Area Mesoscale Model
CMCG	CMC	Global-Environment Multiscale Model
GASP	ABM	Global Analysis and Prediction Model
JMA	JMA	Global Spectral Model
NGPS	FNMOC	Navy Operational Global Atmospheric Prediction System
TCWB	TCWB	Global Forecast System
UKMO	UKMO	Unified Model

The UW EnKF uses the Advanced Research (ARW) version of the Weather Research and Forecasting (WRF) model with 45-km horizontal grid spacing. Since 7 November 2007 the EnKF system has been run with 80 members; until then it used 90 members, which we restrict to the first 80 only, for consistency. Thus, the UW EnKF system comprises 80 exchangeable members, for which our BMA specification enforces equal weights and parameter values.

The UW ME–EnKF combined ensemble thus has 88 members, 80 of which are exchangeable. The resulting BMA specification requires the estimation of nine distinct weights and nine distinct sets of parameter values only.

We consider 24-h forecasts of surface (2 m) temperature in calendar year 2007. Forecasts were bilinearly interpolated from the model grid to observation stations. The UW EnKF data contains forecasts at 590 Automated Surface Observing System (ASOS) stations as well as fixed buoys, as shown in Fig. 1. While the UW ME data is more extensive, we restrict all verification results to the range of dates and sites common to both systems. The training period is a sliding window consisting of forecast and observation data from the most recent 30 days available. Thus, the first 30 days of data are used for training purposes only, leaving a total of 226 days and 106 333 unique forecast cases in the test period. To handle missing ensemble members, we use the renormalization method as described in section 4.

In probabilistic forecasting, the aim is to maximize the sharpness of the predictive distributions subject to calibration (Gneiting et al. 2007). Calibration of an

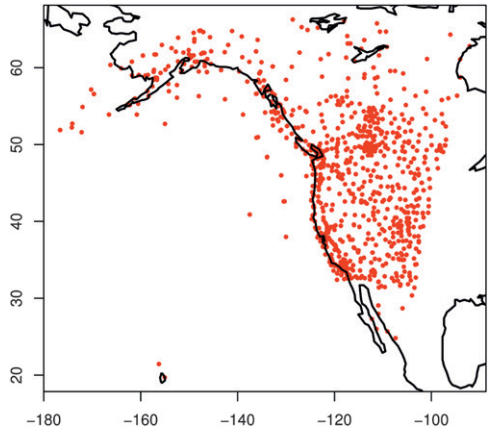

FIG. 1. Station locations common to UW ME and EnKF (there are 590 unique locations).

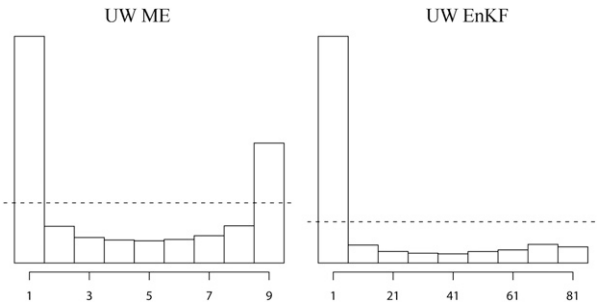

FIG. 2. Verification rank histograms for 24-h forecasts of surface temperature in 2007, using (left) the 8-member UW ME and (right) the 80-member UW EnKF systems.

ensemble system is assessed by means of the verification rank histogram (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Hamill 2001), while calibration of the BMA forecast distributions is evaluated by means of the probability integral transform (PIT) histogram (Raftery et al. 2005; Gneiting et al. 2007). The verification rank histogram plots the rank of each observed value relative to the ensemble member forecasts. The PIT is the value of the BMA cumulative distribution function when evaluated at the verifying observation, and is a continuous analog of the verification rank. In both cases, a more uniform histogram indicates better calibration.

Figure 2 shows the verification rank histograms for the UW ME and EnKF systems, which are considerably underdispersive. The verification rank histogram for the ME–EnKF combined ensemble is almost identical to that for the EnKF system and is therefore not shown. Figure 3 displays the PIT histograms for the BMA postprocessed UW ME, EnKF, and ME–EnKF combined ensembles, all of which show much improved calibration.

Table 2 gives summary measures of predictive performance for the raw and BMA postprocessed ensemble forecasts. As detailed in appendix A, the continuous ranked probability score (CRPS) and the mean absolute error (MAE) quantify probabilistic and deterministic forecast performance, respectively. Both quantities are negatively oriented, that is, the lower the better. The table shows that the unprocessed UW ME system outperforms the recently developed, maturing EnKF ensemble as well as the ME–EnKF combined system. However, this finding changes radically if we consider BMA postprocessed ensembles, in that the postprocessed ME–EnKF combined ensemble shows considerably better performance than either of the raw or BMA postprocessed individual ensemble systems. We believe that this is an important result, which demonstrates the potential for substantially improved forecasts from statistically postprocessed multimodel ensemble systems, even in cases in which one of the constituent systems is superior to the others.

Comparative results for BMA specifications with plausible exchangeable group assignments in a single-system situation can be found in Wilson et al. (2007a), Hamill (2007), and Wilson et al. (2007b).

c. *Results under misspecification*

Having seen results under correct exchangeability assumptions, we now compare to BMA specifications that ignore exchangeability. Figure 4 shows estimated

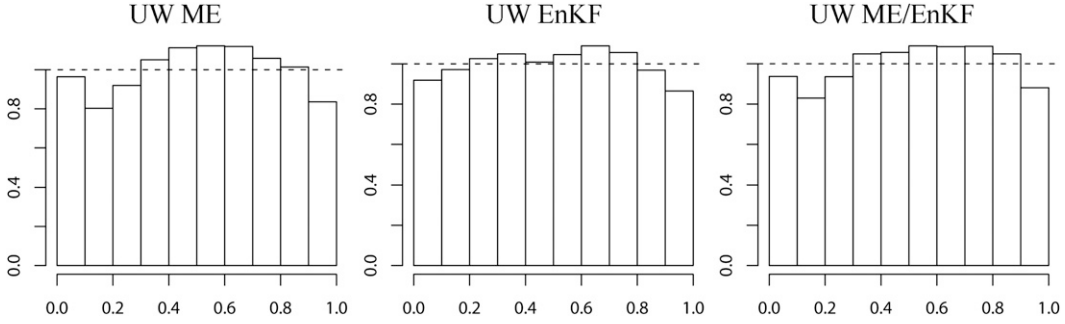

FIG. 3. PIT histograms for BMA postprocessed 24-h forecasts of surface temperature in 2007, based on the UW ME, EnKF, and ME–EnKF combined systems.

TABLE 2. Verification results for 24-h forecasts of surface temperature in 2007 using the raw and BMA postprocessed UW ME, EnKF, and ME–EnKF combined systems. The mean CRPS applies to the probabilistic forecast and the MAE applies to the deterministic forecast given by the median of the predictive distribution, in units of °C. For the postprocessed forecasts of the combined ensemble, results are given for the case in which the BMA specification accounts for exchangeability of the EnKF members, as well as for the case in which the EnKF members are assumed nonexchangeable. Note that the exchangeability assumption has no effect on the predictive performance.

Ensemble forecast	CRPS	MAE
ME	1.96	2.31
EnKF	2.84	3.32
ME–EnKF combined	2.64	3.25
BMA postprocessed forecast	**CRPS**	**MAE**
ME	1.55	2.15
EnKF	1.76	2.49
ME–EnKF exchangeable	1.48	2.09
ME–EnKF nonexchangeable	1.48	2.09

BMA weights for 9 of the 80 exchangeable EnKF members in the ME–EnKF combined ensemble, using a BMA specification in which they are mistakenly treated as if they were individually distinguishable. Since the members are not considered exchangeable, the weights (and also the BMA parameter values) vary across the 80 EnKF members.

The panels display the evolution of these over time, with each estimate based on the trailing 30-day training period at hand. There is considerable noise that can, and ought to be, avoided by invoking the exchangeability assumption. However, the sum of the BMA weights for the 80 EnKF members is nearly the same, regardless of whether or not the BMA specification accounts for exchangeability, and the predictive performance is not affected by the exchangeability assumption (Table 2). Figure 5 shows that near equality holds for the BMA weights of the nonexchangeable ME members as well.

Furthermore, the CRPS and MAE scores under the misspecified BMA model are essentially unchanged from those in Table 2, where the BMA specification accounts for exchangeability. However, the computational effort for the (erroneous) BMA specification with 88 nonexchangeable members is considerably greater than that for the (correct) specification with 8 nonexchangeable ME and 80 exchangeable EnKF members.

4. Accommodating missing ensemble members

In this section we show how to adapt the standard BMA implementation in order to take missing ensemble members into account. Missing ensemble members stem from disruptions in communications, and software or hardware failures, which can last for days or weeks at a time. The chance that any member forecast is missing

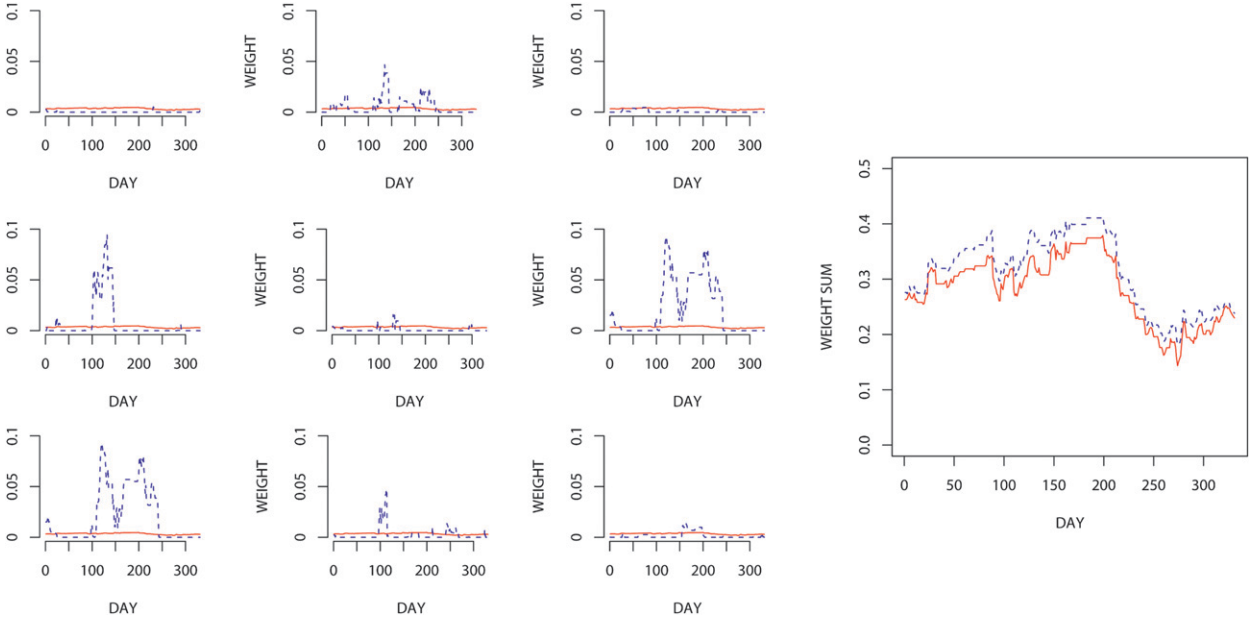

FIG. 4. (left) BMA weights for the first nine of the 80 EnKF members and (right) sum of the BMA weights for the EnKF members in the UW ME–EnKF combined system in 2007. The weights are shown under a BMA specification in which the EnKF members are not considered exchangeable (broken blue line) and under the correct BMA specification in which the EnKF members are treated as exchangeable (solid red line). Under the exchangeability assumption, the BMA weights are the same across the 80 EnKF members. However, the sum of the BMA weights for the EnKF members is nearly the same regardless of whether or not exchangeability is assumed.

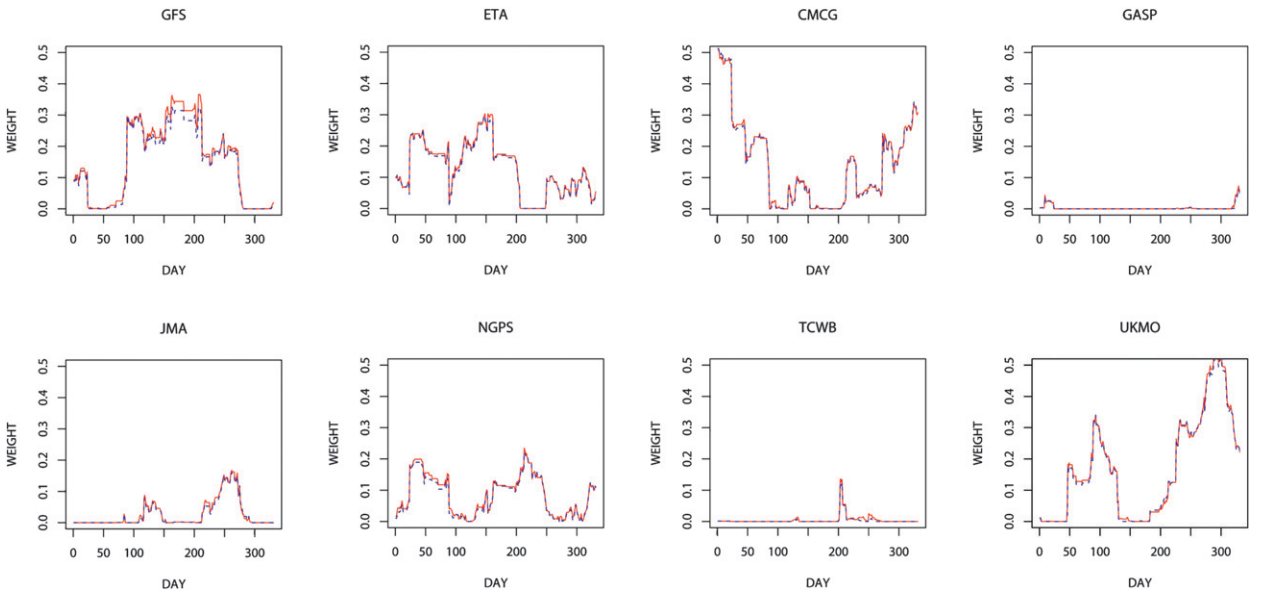

FIG. 5. BMA weights for the eight ME members in the UW ME–EnKF combined system in 2007. The weights are shown under a BMA specification in which the EnKF members are not considered exchangeable (broken blue line) and under the correct BMA specification in which the EnKF members are treated as exchangeable (solid red line).

increases with ensemble size and ensemble diversity, and a considerable amount of potentially useful information would be ignored if instances with missing members were excluded from training sets.

For example, Park et al. (2008) describe the pattern and extent of missing data in The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) system, which is substantial, with some of the constituent ensembles missing for weeks or months at a time. Tables 3 and 4 illustrate the extent of missing members in the eight member UW ME system, which we now consider in its nested 12-km grid version over the Pacific Northwest. For 48-h forecasts of surface (2 m) temperature, there was a total of 557 109 instances in 2006, of which 90 130 or 16.2% had missing member forecasts, in numbers ranging from 1 to 3. In 2007, there was a total of 922 267 instances, of which 58 580 or 6.4% had missing ensemble members, in numbers ranging from 1 to 5. For 48-h forecasts of daily precipitation accumulation, there is much less observational data available, but the extent and pattern of missing data is similar. Figure 6 shows the 12-km UW ME domain along with the unique station locations in the verification database in calendar years 2006 and 2007.

a. Estimation with missing ensemble members

Our goal now is to estimate a full BMA model with terms for all ensemble members, while allowing for instances (s, t) in the training data that lack one or more ensemble member forecasts. We assume initially that the ensemble members are nonexchangeable, that is, we return to the scenario of section 2. We now propose an approach that enables the handling of all configurations of missing data that might conceivably arise in practice.

To achieve this, we modify the EM algorithm as follows. When there are no missing ensemble members, the E step in Eq. (2) can be written as

$$z_{i,s,t}^{(k+1)} = \frac{w_i^{(k)} g_i(y_{s,t}|f_{i,s,t}, \theta_i^{(k)})}{\sum_{p=1}^{m} w_p^{(k)} g_p(y_{s,t}|f_{p,s,t}, \theta_p^{(k)}) \Big/ \sum_{p=1}^{m} w_p^{(k)}}.$$

TABLE 3. Extent of missing members in the 8-member UW ME system for 48-h forecasts of surface temperature over the Pacific Northwest in 2006 and 2007.

2006	Instances	Percent
Tot	557 109	100.0
Complete	466 979	83.8
1 Missing member	76 496	13.7
2 Missing members	11 667	2.1
3 Missing members	1967	0.4
2007	**Instances**	**Percent**
Tot	922 267	100.0
Complete	863 687	93.6
1 Missing member	40 369	4.4
2 Missing members	7324	0.8
3 Missing members	6607	0.7
4 Missing members	2276	0.2
5 Missing members	2004	0.2

TABLE 4. Extent of missing members in the 8-member UW ME system for 48-h forecasts of daily precipitation accumulation over the Pacific Northwest in 2006 and 2007.

2006	Instances	Percent
Tot	117 573	100.0
Complete	97 559	83.0
1 Missing member	17 002	14.4
2 Missing members	2667	2.3
3 Missing members	345	0.3

2007	Instances	Percent
Tot	128 447	100.0
Complete	120 262	93.6
1 Missing member	6037	4.7
2 Missing members	837	0.7
3 Missing members	838	0.7
4 Missing members	303	0.2
5 Missing members	170	0.1

Define

$$\mathcal{A}_{s,t} = \{i | \text{ensemble member } i \text{ available at location } s \text{ and time } t\},$$

which is a subset of, or equal to, the full index set, $\{1, \ldots, m\}$. If there are missing members at location s and time t, the E step is modified, in that

$$z_{i,s,t}^{(k+1)} = \frac{w_i^{(k)} g_i(y_{s,t} | f_{i,s,t}, \theta_i^{(k)})}{\sum_{p \in \mathcal{A}_{s,t}} w_p^{(k)} g_p(y_{s,t} | f_{p,s,t}, \theta_p^{(k)}) \Big/ \sum_{q \in \mathcal{A}_{s,t}} w_p^{(k)}} \tag{9}$$

if i belongs to $\mathcal{A}_{s,t}$, and $z_{i,s,t}^{(k+1)} = 0$ otherwise. Note that the denominator is normalized to account for the missing ensemble members. Accordingly, the M step in Eq. (3) is modified:

$$w_i^{(k+1)} = \frac{\sum_{s,t} z_{i,s,t}^{(k+1)}}{\sum_{s,t} \sum_{p=1}^{m} z_{p,s,t}^{(k+1)}}. \tag{10}$$

Updated estimates $\theta_1^{(k+1)}, \ldots, \theta_m^{(k+1)}$ of the component parameters are obtained by maximizing a renormalized partial expected complete-data log likelihood:

$$\sum_{s,t} \left[\frac{\sum_{p \in \mathcal{A}_{s,t}} z_{p,s,t}^{(k+1)} \log(g_p(y_{s,t} | f_{p,s,t}, \theta_p))}{\sum_{q \in \mathcal{A}_{s,t}} z_{q,s,t}^{(k+1)}} \right], \tag{11}$$

as a function of $\theta_1, \ldots, \theta_m$.

It is straightforward to combine these formulas with those of section 3, to extend them to situations in which there are both missing and exchangeable members, and we do so in appendix B.

b. *Forecasting with missing ensemble members*

We have described above how to adapt the EM estimation algorithm to account for missing member forecasts in the training set. The resulting full BMA model in (1) includes terms for all ensemble members. In forecasting, however, the full BMA model cannot be used when one or more of the members are missing.

We have tested a number of different approaches to forecasting with missing ensemble members, including the following:

- Ensemble mean: Missing members are replaced by the mean of the nonmissing member forecasts. (We also experimented with the median, with nearly identical results.)

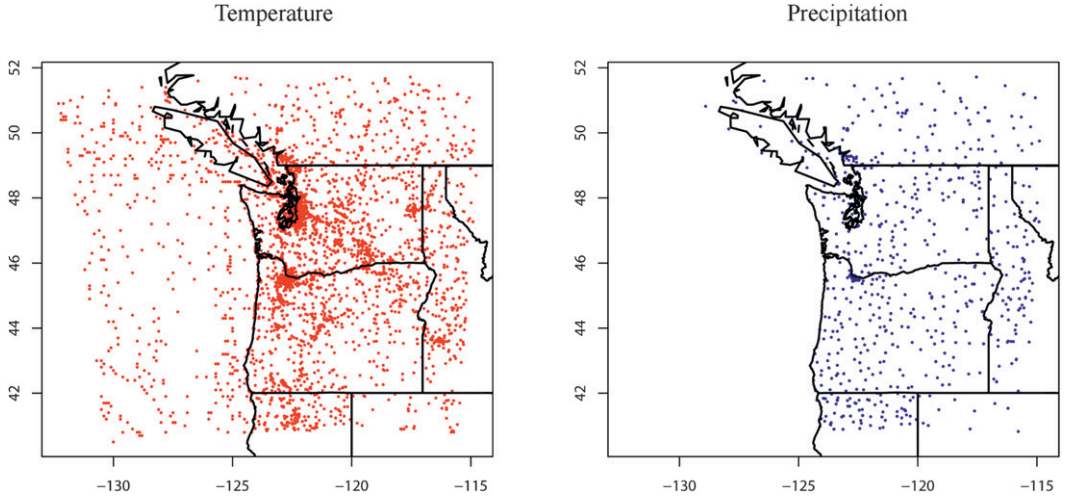

FIG. 6. Nested 12-km UW ME domain over the Pacific Northwest with station locations for (left) temperature (4709 locations) and (right) precipitation (1016 locations).

TABLE 5. Verification results for 48-h forecasts of surface temperature over the Pacific Northwest in 2006 and 2007, using the raw and BMA postprocessed UW ME system, with missing member handling as described in the text. The mean CRPS applies to the probabilistic forecast and the MAE applies to the deterministic forecast given by the median of the predictive distribution (in units of °C).

2006	CRPS	MAE
UW ME	2.08	2.43
BMA ensemble mean	1.73	2.40
BMA conditional mean	1.73	2.40
BMA renormalized	1.73	2.40
BMA restrict/drop	1.72	2.40
BMA restrict/keep	1.72	2.40
2007	**CRPS**	**MAE**
UW ME	2.08	2.44
BMA ensemble mean	1.70	2.36
BMA conditional mean	1.70	2.36
BMA renormalized	1.70	2.36
BMA restrict/drop	1.68	2.34
BMA restrict/keep	1.70	2.36

TABLE 6. Verification results for 48-h forecasts of daily precipitation accumulation over the Pacific Northwest in 2006 and 2007, using the raw and BMA postprocessed UW ME system, with missing member handling as described in the text. The mean CRPS applies to the probabilistic forecast, and the MAE applies to the deterministic forecast given by the median of the predictive distribution (in units of mm).

2006	CRPS	MAE
UW ME	1.74	2.15
BMA ensemble mean	1.46	1.98
BMA conditional mean	1.46	1.98
BMA renormalized	1.44	1.98
BMA restrict/drop	1.45	1.99
BMA restrict/keep	1.44	1.98
2007	**CRPS**	**MAE**
UW ME	1.75	2.14
BMA ensemble mean	1.36	1.83
BMA conditional mean	1.36	1.83
BMA renormalized	1.35	1.83
BMA restrict/drop	1.34	1.83
BMA restrict/keep	1.35	1.83

- Conditional mean: Missing members are replaced using a standard procedure for missing data handling that is known as single imputation (e.g., Schafer 1997), as implemented in the R package norm. For details see appendix C.
- Renormalized: The model is reduced to the terms for the nonmissing forecasts, with the BMA weights renormalized to sum to 1. A small quantity (we use 0.0001) is added to each weight before renormalization, to account for cases in which all nonmissing members have small BMA weights.

Numerical weather forecasts are generally generated on grids, from which they are bilinearly interpolated to observation sites. Thus, on any given day, any specific member will either be available on the entire model grid, or not at all. In such cases, the BMA model can be estimated with the training data restricted to the members available on the forecast day. Any remaining instances with missing forecasts in the training data can still be retained, using the missing member modeling with renormalization as described above (restrict–keep), or else be eliminated from the training data (restrict–drop). McCandless et al. (2009) refer to this latter approach as case deletion.

c. *Application to the University of Washington mesoscale ensemble*

To compare the methods described above, we apply them to 48-h forecasts of surface temperature and daily precipitation accumulation with the UW ME system in calendar years 2006 and 2007. As noted before, the model domain and the extent of missing ensemble members are described in Fig. 6 and Tables 3 and 4, respectively.

Tables 5 and 6 show verification results in terms of the mean CRPS and MAE. For both temperature and precipitation accumulation, the BMA forecasts outperform the raw ensemble, but there is little difference in forecasting performance between the various alternatives for missing member handling. Our overall recommendation thus is the use of the renormalization approach. It is possible that a different conclusion would be reached if there are higher proportions of missing members, or in situations such as the European Poor Man's Ensemble Prediction System (PEPS; Heizenreder et al. 2005), in which the ensemble members have overlapping but distinct model domains.

5. Discussion

We have shown that the BMA approach to the statistical postprocessing of forecast ensembles can be extended to accommodate situations in which member forecasts may be exchangeable and/or missing. In the case of exchangeable members, as in most bred, singular vector or ensemble Kalman filter systems, these modifications result in a physically principled BMA specification, simplify the approach, and facilitate the associated computations. An implementation is available within the R package ensemble BMA (Fraley et al. 2007), both for the Gaussian mixture model for temperature and pressure and for the gamma models that apply to quantitative precipitation and wind speed.

With these extensions, the BMA postprocessing approach applies to any type and configuration of multimodel ensemble system. Recently, Park et al. (2008) concluded that a single ensemble system that has markedly better performance than its competitors, performs as well or better than a multimodel combined system that gives equal weight to all members. However, the authors acknowledge that constraining the weights to be equal may be the reason for this outcome. Our results support this explanation, by showing that a BMA postprocessed combined ensemble system, in which the weights are allowed to vary among the nonexchangeable members, can perform considerably better than any of the constituent ensembles by itself. This is a strongly encouraging conclusion, which suggests BMA postprocessing for other types of multimodel ensembles, including but not limited to the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER; Hagedorn et al. 2005; Doblas-Reyes et al. 2005) and TIGGE (Park et al. 2008; Bougeault et al. 2009, manuscript submitted to *Bull. Amer. Meteor. Soc.*) systems.

Acknowledgments. We are indebted to Cliff Mass, Greg Hakim, Jeff Baars, Brian Ancell, and McLean Sloughter for sharing their insights and providing data. This research was sponsored by the National Science Foundation under Joint Ensemble Forecasting System (JEFS) Subaward S06-47225 with the University Corporation for Atmospheric Research (UCAR), as well as Grants ATM-0724721 and DMS-0706745.

APPENDIX A

Verification Scores

Scoring rules provide summary measures of predictive performance that address calibration and sharpness simultaneously. A particularly attractive scoring rule for probabilistic forecasts of a scalar variable is the continuous ranked probability score, which is defined as

$$\text{crps}(P, x) = \int_{-\infty}^{\infty} (P(y) - \mathbb{I}\{y \geq x\})^2 \, dy,$$

where P is the forecast distribution, here taking the form of a cumulative distribution function, $\mathbb{I}$ is an indicator function, equal to 1 if the argument is true and equal to 0 otherwise, and x is the observed weather quantity (Hersbach 2000; Wilks 2006). An alternative form is

$$\text{crps}(P, x) = E|X - x| - \frac{1}{2} E|X - X'|,$$

where X and X' are independent random variables with distribution P (Grimit et al. 2006; Gneiting and Raftery 2007), and E denotes the expectation operator. The CRPS is proper and generalizes the absolute error, to which it reduces in the case of a deterministic forecast. Both the CRPS and the absolute error are reported in the same units as the forecast variable, and both are negatively oriented, that is, the lower the better.

In this paper, we report the mean CRPS and the MAE, which average individual scores over forecast cases. From any probabilistic forecast we can form a deterministic forecast, by taking the median of the respective predictive distribution, and the MAE reported here refers to this forecast.

APPENDIX B

Estimation with Exchangeable and Missing Ensemble Members

When there are groups of exchangeable members as well as missing member forecasts, let

$$\mathcal{A}_{i,s,t} = \{j | \text{member } j \text{ of group } i \text{ available at location } s \text{ and time } t\}.$$

The E step in Eq. (6) is then adapted as follows:

$$z_{i,j,s,t}^{(k+1)} = \frac{w_i^{(k)} g_i(y_{s,t} | f_{i,j,s,t}, \theta_i^{(k)})}{\sum_{p=1}^{I} \sum_{r \in \mathcal{A}_{p,s,t}} w_p^{(k)} g_p(y_{s,t} | f_{p,r,s,t}, \theta_p^{(k)}) \Big/ \sum_{q=1}^{I} \sum_{r \in \mathcal{A}_{q,s,t}} w_q^{(k)}} \tag{B1}$$

if j belongs to $\mathcal{A}_{i,s,t}$, and $z_{i,j,s,t}^{(k+1)} = 0$ otherwise. The M step update in Eq. (7) for the BMA weight estimates is generalized as

$$w_i^{(k+1)} = \frac{1}{m_i} \frac{\sum_{s,t} \sum_{j=1}^{m_i} z_{i,j,s,t}^{(k+1)}}{\sum_{s,t} \sum_{p=1}^{I} \sum_{r=1}^{m_p} z_{p,r,s,t}^{(k+1)}}, \tag{B2}$$

and the M step update of θ maximizes

$$\sum_{s,t} \left[\frac{\sum_{i=1}^{I} \sum_{j \in \mathcal{A}_{i,s,t}} z_{i,j,s,t}^{(k+1)} \log(g_i(y_{s,t} | f_{i,j,s,t}, \theta_i))}{\sum_{p=1}^{I} \sum_{r \in \mathcal{A}_{p,s,t}} z_{p,r,s,t}^{(k+1)}} \right]. \tag{B3}$$

TABLE B1. The M step variance updates for the Gaussian mixture model.

Missing	Exchangeable	M step update for $(\sigma^2)^{(k+1)}$
No	No	$\dfrac{\sum_{s,t}\sum_{i=1}^{m} z_{i,s,t}^{(k+1)}(y_{s,t}-f_{i,s,t})^2}{n}$
No	Yes	$\dfrac{\sum_{s,t}\sum_{i=1}^{I}\sum_{j=1}^{m_i} z_{i,j,s,t}^{(k+1)}(y_{s,t}-f_{i,j,s,t})^2}{n}$
Yes	No	$\dfrac{\sum_{s,t}\sum_{i=1}^{m} z_{i,s,t}^{(k+1)}(y_{s,t}-f_{i,s,t})^2}{\sum_{s,t}\sum_{i\in\mathcal{A}_{s,t}} z_{i,s,t}^{(k+1)}}$
Yes	Yes	$\dfrac{\sum_{s,t}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_{i,s,t}} z_{i,j,s,t}^{(k+1)}(y_{s,t}-f_{i,j,s,t})^2}{\sum_{s,t}\sum_{i=1}^{I}\sum_{j\in\mathcal{A}_{i,s,t}} z_{i,j,s,t}^{(k+1)}}$

For example, the Gaussian mixture model for temperature (Raftery et al. 2005) has a single variance parameter, σ^2, that is common to all members and needs to be estimated in each M step. Table B1 summarizes the respective updates, both with and without missing members, and with and without exchangeability assumptions.

APPENDIX C

Imputing Missing Ensemble Members: Conditional Mean Approach

In the conditional mean approach to the imputation of missing ensemble members we use a standard procedure for missing data handling that is known as single imputation (e.g., Schafer 1997) and implemented in the R package norm.

Suppose, for now, that the ensemble has m nonexchangeable members. Let $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ be the standard maximum likelihood estimates for the mean vector and the covariance matrix of the ensemble members, based on training data. The conditional mean approach imputes the conditional mean for any missing members in the ensemble forecast, based on an assumption of multivariate normality and the above estimates, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

This is now explained in more detail, using an illustrative example. Let $\mathbf{f} \in \mathbb{R}^m$ denote the ensemble forecast, with $q \geq 1$ members missing. Let

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_{\mathrm{A}} \\ \mathbf{f}_{\mathrm{M}} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathrm{A}} \\ \boldsymbol{\mu}_{\mathrm{M}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathrm{AA}} & \boldsymbol{\Sigma}_{\mathrm{AM}} \\ \boldsymbol{\Sigma}_{\mathrm{MA}} & \boldsymbol{\Sigma}_{\mathrm{MM}} \end{pmatrix}$$

be partitions of $\mathbf{f}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ corresponding to the nonmissing or available (A), and the missing (M) members, respectively. Then the conditional distribution of $\mathbf{f}_{\mathrm{M}}$ given $\mathbf{f}_{\mathrm{A}}$ is $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where

$$\boldsymbol{\mu}_0 = \boldsymbol{\mu}_{\mathrm{M}} + \boldsymbol{\Sigma}_{\mathrm{MA}}\boldsymbol{\Sigma}_{\mathrm{AA}}^{-1}(\mathbf{f}_{\mathrm{A}} - \boldsymbol{\mu}_{\mathrm{A}}) \in \mathbb{R}^q$$

and

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\mathrm{MM}} - \boldsymbol{\Sigma}_{\mathrm{MA}}\boldsymbol{\Sigma}_{\mathrm{AA}}^{-1}\boldsymbol{\Sigma}_{\mathrm{AM}} \in \mathbb{R}^{q\times q}.$$

For example, in the case of 48-h UW ME forecasts of surface temperature, five member forecasts are missing for initialization date 4 July 2007, that is, $m = 8$ and $q = 5$. To facilitate the presentation, we do not distinguish row and column vectors, leaving the (obvious) identifications and transpositions to the reader. Using a 30-day sliding training period and the R package norm (Schafer 1997), the ensemble mean vector is

$$\boldsymbol{\mu} = \begin{pmatrix} \texttt{GFS} & \texttt{CMCG} & \texttt{ETA} & \texttt{GASP} & \texttt{JMA} & \texttt{NGPS} & \texttt{TCWB} & \texttt{UKMO} \\ 18.82 & 18.95 & 18.38 & 18.26 & 18.03 & 18.77 & 17.98 & 18.67 \end{pmatrix},$$

and the ensemble covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} & \texttt{GFS} & \texttt{CMCG} & \texttt{ETA} & \texttt{GASP} & \texttt{JMA} & \texttt{NGPS} & \texttt{TCWB} & \texttt{UKMO} \\ \texttt{GFS} & 31.49 & 30.41 & 30.73 & 30.31 & 30.07 & 30.95 & 31.04 & 31.18 \\ \texttt{CMCG} & 30.41 & 33.06 & 31.00 & 30.96 & 30.84 & 31.90 & 31.47 & 31.69 \\ \texttt{ETA} & 30.73 & 31.00 & 32.96 & 30.64 & 30.47 & 31.42 & 31.58 & 31.49 \\ \texttt{GASP} & 30.31 & 30.96 & 30.64 & 32.91 & 30.69 & 31.92 & 31.57 & 31.68 \\ \texttt{JMA} & 30.07 & 30.84 & 30.47 & 30.69 & 31.97 & 31.12 & 31.19 & 31.57 \\ \texttt{NGPS} & 30.95 & 31.90 & 31.42 & 31.92 & 31.12 & 33.83 & 32.20 & 32.38 \\ \texttt{TCWB} & 31.04 & 31.47 & 31.58 & 31.57 & 31.19 & 32.20 & 33.66 & 32.14 \\ \texttt{UKMO} & 31.18 & 31.69 & 31.49 & 31.68 & 31.57 & 32.38 & 32.14 & 33.97 \end{pmatrix}.$$

The UW ME ensemble forecast at the Seattle–Tacoma Airport ASOS station is

$$\mathbf{f} = \begin{pmatrix} \text{GFS} & \text{CMCG} & \text{ETA} & \text{GASP} & \text{JMA} & \text{NGPS} & \text{TCWB} & \text{UKMO} \\ 25.75 & \text{NA} & 27.57 & \text{NA} & \text{NA} & \text{NA} & 25.90 & \text{NA} \end{pmatrix},$$

where NA indicates a missing value. We replace the $q = 5$ missing member by the respective conditional mean, μ_0, namely

$$\begin{aligned} \boldsymbol{\mu}_0 &= \boldsymbol{\mu}_{\text{M}} + \boldsymbol{\Sigma}_{\text{MA}}\boldsymbol{\Sigma}_{\text{AA}}^{-1}(\mathbf{f}_{\text{A}} - \boldsymbol{\mu}_{\text{A}}) \\ &= \begin{pmatrix} \text{CMCG} & \text{GASP} & \text{JMA} & \text{NGPS} & \text{UKMO} \\ 18.95 & 18.26 & 18.03 & 18.77 & 18.67 \end{pmatrix} \\ &\quad + \begin{pmatrix} & \text{GFS} & \text{ETA} & \text{TCWB} \\ \text{CMCG} & 30.41 & 31.00 & 31.47 \\ \text{GASP} & 30.31 & 30.64 & 31.57 \\ \text{JMA} & 30.07 & 30.47 & 31.19 \\ \text{NGPS} & 30.95 & 31.42 & 32.20 \\ \text{UKMO} & 31.18 & 31.49 & 32.14 \end{pmatrix} \begin{pmatrix} & \text{GFS} & \text{ETA} & \text{TCWB} \\ \text{GFS} & 31.49 & 30.73 & 31.04 \\ \text{ETA} & 30.73 & 32.96 & 31.58 \\ \text{TCWB} & 31.04 & 31.58 & 33.66 \end{pmatrix}^{-1} \begin{pmatrix} 25.75 - 18.82 \\ 27.57 - 18.38 \\ 25.90 - 17.98 \end{pmatrix} \\ &= \begin{pmatrix} \text{CMCG} & \text{GASP} & \text{JMA} & \text{NGPS} & \text{UKMO} \\ 26.73 & 25.18 & 25.59 & 26.59 & 26.42 \end{pmatrix}. \end{aligned}$$

For weather parameters such as temperature, it is also possible to perform the imputation based on the deviations from the ensemble mean forecast, rather than the ensemble forecasts themselves. In experiments, we found no appreciable difference between these two imputation alternatives on UW ME temperature forecasts for 2006 and 2007.

The method as described above assumes nonexchangeable members, but it can easily be extended to account for exchangeability. This is done by constraining the elements of the mean vector, $\boldsymbol{\mu}$, and the diagonal elements in the covariance matrix, $\boldsymbol{\Sigma}$, to be equal for exchangeable members, and requiring that the cross-covariance terms depend on the group assignments only. As an illustration, consider a four-member ensemble in which the second and third members are exchangeable. Then we use

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_2, \mu_3)$$

with the second and third entry constrained to be equal, and the covariance matrix, $\boldsymbol{\Sigma}$, is constrained to be of the following form:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \alpha_{12} & \alpha_{12} & \alpha_{12} \\ \alpha_{12} & \sigma_{22}^2 & \alpha_{22} & \alpha_{22} \\ \alpha_{12} & \alpha_{22} & \sigma_{22}^2 & \alpha_{22} \\ \alpha_{12} & \alpha_{22} & \alpha_{22} & \sigma_{33}^2 \end{pmatrix},$$

where α_{12} and α_{22} are cross-covariance terms.

REFERENCES

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

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. Ser. B,* **39,** 1–38.

Dirren, S., R. D. Torn, and G. J. Hakim, 2007: A data assimilation case study using a limited-area ensemble Kalman filter. *Mon. Wea. Rev.,* **135,** 1455–1473.

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

Eckel, F. A., and C. F. Mass, 2005: Effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Evensen, G., 1994: Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99,** 10 143–10 162.

Fraley, C., A. E. Raftery, T. Gneiting, and J. M. Sloughter, 2007: EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging. Tech. Rep. 516, Department of Statistics, University of Washington, 17 pp.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B,* **69,** 243–268.

Grimit, E. P., and C. F. Mass, 2002: Initial results for a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

——, T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.,* **132,** 3209–3220.

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

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, 2005: Ensemble-based atmospheric data assimilation: A tutorial. *Predictability of Weather and Climate,* T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 124–156.

——, 2007: Comments on ''Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging.'' *Mon. Wea. Rev.,* **135,** 4226–4230.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

Heizenreder, D., S. Trepte, and M. Denhard, 2005: SRNWP-PEPS: A regional multi-model ensemble in Europe. *Euro. Forecaster,* **11,** 29–35.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble predictions. *Wea. Forecasting,* **15,** 559–570.

Joslyn, S., and D. W. Jones, 2008: Strategies in naturalistic decision-making: A cognitive task analysis of naval weather forecasting. *Naturalistic Decision Making and Macrocognition,* J. M. Schraagen et al., Eds., Ashgate Publishing, 183–202.

McCandless, T. C., S. E. Haupt, and G. Young, 2009: Replacing missing data for ensemble systems. Preprints, *Seventh Conf. on Artificial Intelligence and Its Applications to the Environmental Sciences,* Phoenix, AZ, Amer. Meteor. Soc., 1.2. [Available online at http://ams.confex.com/ams/89annual/techprogram/paper_150305.htm.]

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions*. Wiley, 274 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Park, Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.,* **134,** 2029–2050.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Schafer, J. L., 1997: *Analysis of Incomplete Multivariate Data by Simulation*. Chapman and Hall, 430 pp.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

——, T. Gneiting, and A. E. Raftery, 2009: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.,* in press.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability,* Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.

Torn, R. D., and G. J. Hakim, 2008: Performance characteristics of a pseudo-operational ensemble Kalman filter. *Mon. Wea. Rev.,* **136,** 3947–3963.

——, ——, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.,* **134,** 2490–2502.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at the NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Vrugt, J. A., C. G. H. Diks, and M. P. Clark, 2008: Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environ. Fluid Mech.,* **134,** 1–17.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of forecasts? *Quart. J. Roy. Meteor. Soc.,* **134,** 241–260.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 648 pp.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007a: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

——, ——, ——, and ——, 2007b: Reply. *Mon. Wea. Rev.,* **135,** 4231–4236.