Using Bayesian Model Averaging to Calibrate Forecast Ensembles

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski

Department of Statistics, University of Washington, Seattle, Washington

(Manuscript received 18 December 2003, in final form 29 September 2004)

ABSTRACT

Ensembles used for probabilistic weather forecasting often exhibit a spread-error correlation, but they tend to be underdispersive. This paper proposes a statistical method for postprocessing ensembles based on Bayesian model averaging (BMA), which is a standard method for combining predictive distributions from different sources. The BMA predictive probability density function (PDF) of any quantity of interest is a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts and reflect the models' relative contributions to predictive skill over the training period. The BMA weights can be used to assess the usefulness of ensemble members, and this can be used as a basis for selecting ensemble members; this can be useful given the cost of running large ensembles. The BMA PDF can be represented as an unweighted ensemble of any desired size, by simulating from the BMA predictive distribution.

The BMA predictive variance can be decomposed into two components, one corresponding to the between-forecast variability, and the second to the within-forecast variability. Predictive PDFs or intervals based solely on the ensemble spread incorporate the first component but not the second. Thus BMA provides a theoretical explanation of the tendency of ensembles to exhibit a spread-error correlation but yet be underdispersive.

The method was applied to 48-h forecasts of surface temperature in the Pacific Northwest in January–June 2000 using the University of Washington fifth-generation Pennsylvania State University–NCAR Mesoscale Model (MM5) ensemble. The predictive PDFs were much better calibrated than the raw ensemble, and the BMA forecasts were sharp in that 90% BMA prediction intervals were 66% shorter on average than those produced by sample climatology. As a by-product, BMA yields a deterministic point forecast, and this had root-mean-square errors 7% lower than the best of the ensemble members and 8% lower than the ensemble mean. Similar results were obtained for forecasts of sea level pressure. Simulation experiments show that BMA performs reasonably well when the underlying ensemble is calibrated, or even overdispersed.

1. Introduction

The dominant approach to probabilistic weather forecasting has been the use of ensembles in which a model is run several times with different initial conditions or model physics. This was proposed by Leith (1974) as a way of implementing the general framework presented by Epstein (1969). Ensembles based on global models have been found useful for medium-range probabilistic forecasting (Toth and Kalnay 1993; Molteni et al. 1996; Houtekamer and Derome 1995; Hamill et al. 2000). Typically the ensemble mean outperforms all or most of the individual ensemble members, and in some studies a spread-error correlation has been observed, in which the spread in the ensemble forecasts is correlated with the magnitude of the forecast error. Often, however, the ensemble is underdispersive and thus not calibrated. Both spread-error correlations and underdispersion have been observed in the National Centers for Environmental Prediction (NCEP) operational global ensemble (Toth et al. 2001; Eckel and Walters 1998), the Canadian Ensemble Prediction System (Pellerin et al. 2003), and the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (Buizza 1997; Buizza et al. 1999; Hersbach et al. 2000; Scherrer et al. 2004); for an overview see Buizza et al. (2005).

Here we focus on short-range mesoscale forecasting. Several authors have studied the use of a synoptic ensemble, the 15-member NCEP Eta–Regional Spectral Model (RSM) ensemble, for short-range forecasting (Hamill and Colucci 1997; Hamill and Colucci 1998; Stensrud et al. 1999). As was the case for medium-range forecasting, the ensemble mean was more skillful for short-range forecasting than the individual ensemble members, but the spread–skill relationship was weak. The first short-range mesoscale ensemble forecasting

Corresponding author address: Adrian E. Raftery, Department of Statistics, University of Washington, Box 354320, Seattle, WA 98195-4320.
E-mail: raftery@stat.washington.edu

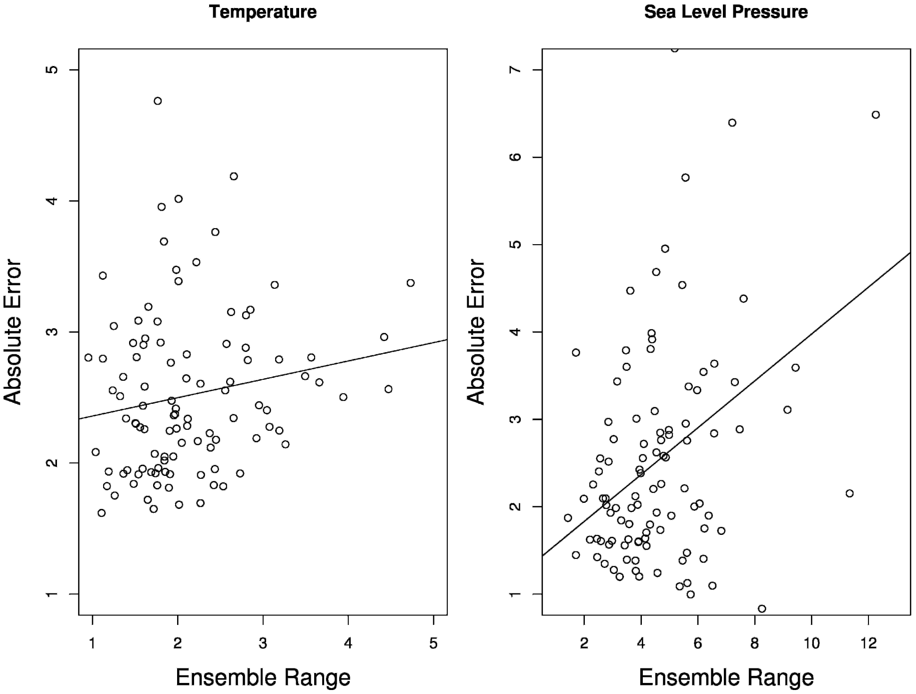

FIG. 1. Spread–skill relationship for daily average absolute errors in the 48-h forecast of (a) surface temperature and (b) sea level pressure in the UW ensemble, Jan–Jun 2000. The vertical axis shows the daily average of the absolute errors of the ensemble mean forecast, and the horizontal axis shows the daily average of the difference between the highest and lowest forecasts in the ensemble. The solid line is the least squares regression line. The correlation is 0.18 for temperature and 0.42 for sea level pressure.

experiment was the Storm and Mesoscale Ensemble Experiment (SAMEX; Hou et al. 2001). This found that the ensemble mean was more skillful than the individual forecasts, and that there was a significant spread-error correlation, with a correlation on the order of 0.4. However, the ensemble was not well calibrated; see Figs. 8 and 9 of Hou et al. (2001).

Grimit and Mass (2002) described the University of Washington mesoscale short-range ensemble system for the Pacific Northwest (hereafter referred to as the UW ensemble). This is a five-member multianalysis ensemble consisting of different runs of the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5), in which initial conditions are taken from different operational centers. The UW ensemble was run at 36- and 12-km grid spacing, while the NCEP Short-Range Ensemble Forecasting (SREF) has been run at 48 km. Like other authors, Grimit and Mass (2002) found the ensemble mean to be more skillful than the individual forecasts, and they reported a stronger spread-error correlation than other studies, ranging up to 0.6. Figure 1 is a scatterplot showing the spread-error relationships for surface temperature and sea level pressure for the UW ensemble for the same period as that on which Grimit and Mass' (2002) report was based, namely January–June 2000. The spread-error correlation for daily average absolute errors, averaging spatially across the Pacific Northwest, was 0.18 for temperature and 0.42 for sea level pressure; both correlations were positive and the latter was highly statistically significant. However, the verification rank histograms (Anderson 1996; Talagrand et al. 1997; Hamill 2001) for the same data, shown in Fig. 2, show the ensemble to be underdispersive and hence uncalibrated. In this case, the ensemble range based on five members would contain 4/6, or 66.7%, of the observed values if the ensemble were calibrated, that is, if the ensemble forecasts were a sample from the true predictive probability density function (PDF), whereas in fact it contained only 29% of them for temperature and 54% of them for sea level pressure.

This behavior—an ensemble that yields a positive spread-error correlation and hence useful predictions of forecast skill, and yet is uncalibrated—is not unique to the UW ensemble, as we have noted, and may seem contradictory at first sight. On reflection, though, it is not so surprising. There are several sources of uncertainty in numerical weather forecasts, including uncertainty about initial conditions, lateral boundary conditions, and model physics, as well as discretization and integration methods. Most ensembles capture only some of these uncertainties, and then probably only partially. Thus it seems inevitable that ensembles based

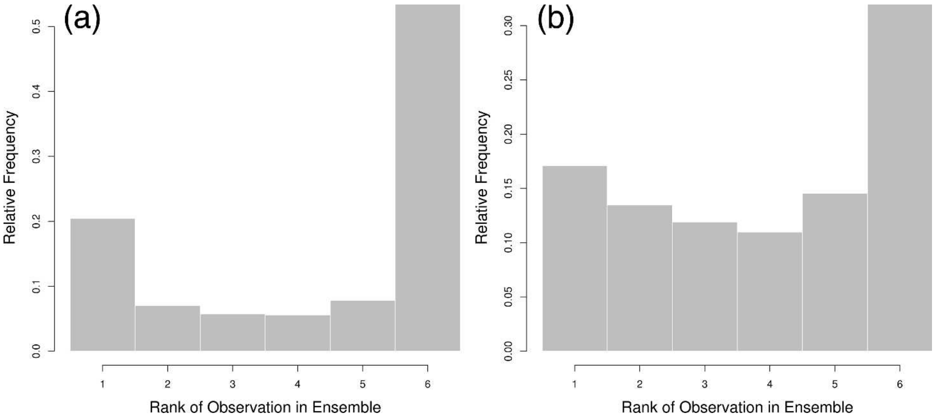

FIG. 2. Verification rank histograms for the UW ensemble 48-h forecasts of (a) surface temperature and (b) sea level pressure, Jan–Jun 2000.

purely on perturbing initial and lateral boundary conditions, model physics, and integration methods will be underdispersive to some extent. Because they do capture some of the important sources of uncertainty, however, it is reasonable to expect a positive spread-error correlation, even when the ensemble is uncalibrated. To obtain a calibrated forecast PDF, therefore, it seems necessary to carry out some form of statistical postprocessing, as suggested by Hamill and Colucci (1997, 1998).

Our goal in this article is to propose an approach for obtaining calibrated and sharp predictive PDFs of future weather quantities or events from the output of ensembles that may not be themselves calibrated. By calibrated we mean simply that intervals or events that we declare to have probability *P* contain the truth, or happen, a proportion *P* of the time on average in the long run. Sharpness is a function of the widths of prediction intervals. For example, a 90% prediction interval verifying at a given time and place is defined by a lower bound and an upper bound, such that the probability that the verifying observation lies between the two bounds is declared to be 90%. By sharp we mean that prediction intervals are narrower on average than those obtained from climatology. Clearly, the sharper the better. We adopt the principle that the goal of probabilistic forecasting is to maximize sharpness subject to calibration (Gneiting et al. 2003).

To achieve this, we propose a statistical approach to postprocessing ensemble forecasts, based on Bayesian model averaging (BMA). This is a standard approach to inference in the presence of multiple competing statistical models and has been widely applied in the social and health sciences; here we extend it to forecasts from dynamical models. In BMA, the overall forecast PDF is a weighted average of forecast PDFs based on each of the individual forecasts; the weights are the estimated posterior model probabilities and reflect the models' forecast skill in the training period, relative to the other models. The weights can also provide a basis for selecting ensemble members: when they are small there is little to be lost by removing the corresponding ensemble member. This can be useful given the computational cost of running ensembles.

The BMA deterministic forecast is just a weighted average of linear functions of the (possibly bias-corrected) forecasts from the ensemble. The BMA forecast PDF can be written as an analytic expression, and it can also be represented as an equally weighted ensemble of any desired size, by simulating potential observations from the forecast PDF. The BMA forecast variance decomposes into two components, corresponding to between-model and within-model variance. The ensemble spread captures only the first component. This decomposition provides a theoretical explanation and quantification of the behavior observed in several ensembles, in which a positive spread-error correlation coexists with a lack of calibration.

In section 2 we describe BMA, show how the BMA model can be estimated, and give examples of BMA in action. In section 3 we give BMA results for the UW ensemble, in section 4 we give some results for simulated ensembles, and in section 5 we make some concluding remarks. While our experiments are with the UW ensemble—that is, a mesoscale, single-model, multianalysis ensemble system—the idea applies to other situations, including synoptic, perturbed observations, singular vector, and bred and multimodel ensembles, with small changes, as indicated below.

2. Bayesian model averaging

a. Basic ideas

Standard statistical analysis—such as, for example, regression analysis—typically proceeds conditionally on one assumed statistical model. Often this model has been selected from among several possible competing models for the data, and the data analyst is not sure that it is the best one. Other plausible models could give different answers to the scientific question at hand. This is a source of uncertainty in drawing conclusions, and the typical approach, that of conditioning on a single model deemed to be "best," ignores this source of uncertainty, thus underestimating uncertainty.

Bayesian model averaging (Leamer 1978; Kass and Raftery 1995; Hoeting et al. 1999) overcomes this problem by conditioning, not on a single "best" model, but on the entire ensemble of statistical models first considered. In the case of a quantity y to be forecast on the basis of training data y^T using K statistical models $M_1, \ldots, M_K$, the law of total probability tells us that the forecast PDF, $p(y)$, is given by

$$p(y) = \sum_{k=1}^{K} p(y|M_k)p(M_k|y^T), \tag{1}$$

where $p(y|M_k)$ is the forecast PDF based on model M_k alone, and $p(M_k|y^T)$ is the posterior probability of model M_k being correct given the training data, and reflects how well model M_k fits the training data. The posterior model probabilities add up to one, so that $\Sigma_{k=1}^{K} p(M_k|y^T) = 1$, and they can thus be viewed as weights. The BMA PDF is a weighted average of the conditional PDFs given each of the individual models, weighted by their posterior model probabilities. BMA possesses a range of theoretical optimality properties and has shown good performance in a variety of simulated and real data situations (Raftery and Zheng 2003).

We now extend BMA from statistical models to dynamical models. The basic idea is that for any given forecast ensemble there is a "best" model, or member, but we do not know what it is, and our uncertainty about the best member is quantified by BMA. Once again, we denote by y the quantity to be forecast. Each deterministic forecast can be bias corrected using any one of many possible bias-correction methods, yielding a bias-corrected forecast f_k. The forecast f_k is then associated with a conditional PDF, $g_k(y|f_k)$, which can be interpreted as the conditional PDF of y conditional on f_k, *given that f_k is the best forecast in the ensemble*. The BMA predictive model is then

$$p(y|f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k g_k(y|f_k), \tag{2}$$

where w_k is the posterior probability of forecast k being the best one and is based on forecast k's performance in the training period. The w_k's are probabilities and so they are nonnegative and add up to 1, that is, $\Sigma_{k=1}^{K} w_k = 1$. We describe how to estimate w_k in the next subsection.

When forecasting temperature and sea level pressure, it often seems reasonable to approximate the conditional PDF by a normal distribution centered at a linear function of the forecast, $a_k + b_k f_k$, so that $g_k(y|f_k)$ is a normal PDF with mean $a_k + b_k f_k$ and standard deviation σ. We denote this situation by

$$y|f_k \sim N(a_k + b_k f_k, \sigma^2), \tag{3}$$

and we describe how to estimate σ in the next subsection. In this case, the BMA predictive mean is just the conditional expectation of y given the forecasts, namely

$$E[y|f_1, \ldots, f_K] = \sum_{k=1}^{K} w_k(a_k + b_k f_k). \tag{4}$$

This can be viewed as a deterministic forecast in its own right and can be compared with the individual forecasts in the ensemble, or with the ensemble mean.

b. *Estimation by maximum likelihood, the EM algorithm, and minimum CRPS estimation*

For convenience, we restrict attention to the situation where the conditional PDFs are approximated by normal distributions. This seems to be reasonable for some variables, such as temperature and sea level pressure, but not for others, such as wind speed and precipitation; other distributions would be needed for the latter. The basic ideas carry across directly to other distributions also. We now consider how to estimate the model parameters, a_k, b_k, w_k, $k = 1, \ldots, K$, and σ^2, on the basis of a training dataset consisting of ensemble forecasts and verifying observations, where the forecasts have been interpolated to the observation sites. We denote space and time by subscripts s and t, so that f_{kst} denotes the kth forecast in the ensemble for place s and time t, and y_{st} denotes the corresponding verification. Here we will take the forecast lead time to be fixed; in practice we will estimate different models for each forecast lead time.

We first estimate a_k and b_k by simple linear regression of y_{st} on f_{kst} for the training data. If the forecasts have not yet been bias corrected, estimation of a_k and b_k can be viewed as a very simple bias-correction process, and it can also be considered as a very simple form of model output statistics (Glahn and Lowry 1972; Carter et al. 1989). Note that we retain the a_k and b_k in (3) even if the forecasts have been bias corrected.

We estimate w_k, $k = 1, \ldots, K$, and σ by maximum likelihood (Fisher 1922) from the training data. The likelihood function is defined as the probability of the training data given the parameters to be estimated, viewed as a function of the parameters. The maximum likelihood estimator is the value of the parameter vector that maximizes the likelihood function, that is, the value of the parameter vector under which the observed data were most likely to have been observed. The maximum likelihood estimator has many optimality properties (Casella and Berger 2001).

It is convenient to maximize the logarithm of the likelihood function (or log-likelihood function) rather than the likelihood function itself, for reasons of both algebraic simplicity and numerical stability; the same parameter value that maximizes one also maximizes the other. Assuming independence of forecast errors in space and time, the log-likelihood function for model (2) is

$$\ell(w_1, \ldots, w_k, \sigma^2) = \sum_{s,t} \log\left(\sum_{k=1}^{K} w_k g_k(y_{st}|f_{kst})\right), \quad (5)$$

where the summation is over values of s and t that index observations in the training set. The independence assumption is unlikely to hold, but estimates are unlikely to be very sensitive to this assumption, because we are estimating the conditional distribution for a scalar observation given forecasts, rather than for several observations simultaneously. This cannot be maximized analytically, and it is complex to maximize numerically using direct nonlinear maximization methods such as Newton–Raphson and its variants. Instead, we maximize it using the expectation-maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997).

The EM algorithm is a method for finding the maximum likelihood estimator when the problem can be recast in terms of unobserved quantities such that, if we knew what they were, the estimation problem would be straightforward. The BMA model (2) is a finite mixture model (McLachlan and Peel 2000). Here we introduce the unobserved quantities z_{kst}, where $z_{kst} = 1$ if ensemble member k is the best forecast for verification site s and time t, and $z_{kst} = 0$ otherwise. For each (s, t), only one of $\{z_{1st}, \ldots, z_{Kst}\}$ is equal to 1; the others are all zero.

The EM algorithm is iterative and alternates between two steps, the E (or expectation) step and the M (or maximization) step. It starts with an initial guess, $\theta^{(0)}$, for the parameter vector θ. In the E step, the z_{kst} are estimated given the current guess for the parameters; the estimates of the z_{kst} are not necessarily integers, even though the true values are 0 or 1. In the M step, θ is estimated given the current values of the z_{kst}.

For the normal BMA model given by (2) and (3), the E step is

$$\hat{z}_{kst}^{(j)} = \frac{w_k g(y_{st}|f_{kst}, \sigma^{(j-1)})}{\sum_{i=1}^{K} w_i g(y_{st}|f_{ist}, \sigma^{(j-1)})}, \quad (6)$$

where the superscript j refers to the jth iteration of the EM algorithm, and $g(y_{st}|f_{kst}, \sigma^{(j-1)})$ is a normal density with mean $a_k + b_k f_{kst}$ and standard deviation $\sigma^{(j-1)}$ evaluated at y_{st}. The M step then consists of estimating the w_k and σ using as weights the current estimates of z_{kst}, namely $\hat{z}_{kst}^{(j)}$. Thus

$$w_k^{(j)} = \frac{1}{n} \sum_{s,t} \hat{z}_{kst}^{(j)},$$

$$\sigma^{2(j)} = \frac{1}{n} \sum_{s,t} \sum_{k=1}^{K} \hat{z}_{kst}^{(j)} (y_{st} - f_{kst})^2,$$

where n is the number of observations in the training set [i.e., the number of distinct values of (s, t)].

The E and M steps are then iterated to convergence, which we defined as changes no greater than some small tolerances in any of the log likelihood, the parameter values, or the $\hat{z}_{kst}^{(j)}$ in one iteration. The log likelihood is guaranteed to increase at each EM iteration (Wu 1983), which implies that in general it converges to a local maximum of the likelihood. Convergence to a global maximum cannot be guaranteed, so the solution reached by the algorithm can be sensitive to the starting values. Starting values based on past experience usually give good solutions.

We finally refine our estimate of σ so that it optimizes the continuous ranked probability score (CRPS) for the training data. The CRPS is the integral of the Brier scores at all possible threshold values for the continuous predictand (Hersbach 2000). As such, it is an appropriate score when interest focuses on prediction intervals. We do this by searching numerically over a range of values of σ, centered at the maximum likelihood estimate, keeping the other parameters fixed.

In our implementation, the training set consists of a sliding window of forecasts and observations for the previous m days. We discuss the choice of m later.

c. The BMA predictive variance decomposition and the spread-error correlation

The BMA predictive variance of y_{st} given the ensemble of forecasts can be written as

$$\mathrm{Var}(y_{st}|f_{1st}, \ldots, f_{Kst}) = \sum_{k=1}^{K} w_k \left((a_k + b_k f_{kst}) - \sum_{i=1}^{K} w_i(a_i + b_i f_{ist})\right)^2 + \sigma^2 \quad (7)$$

(Raftery 1993). The right-hand side has two terms, the first of which summarizes between-forecast spread, and the second (equal to σ^2) measures the expected uncertainty conditional on one of the forecasts being best. We can summarize this verbally as

$$\text{Predictive Variance} = \text{Between-Forecast Variance} + \text{Within-Forecast Variance}. \quad (8)$$

The first term represents the ensemble spread. Thus one would expect to see a spread-error correlation, since the predictive variance includes the spread as a component. But it also implies that using the ensemble spread alone may underestimate uncertainty, because it ignores the second term on the right-hand side of (7) or (8).

TABLE 1. Forty-eight-hour UW-MM5 ensemble forecasts of surface temperature at Packwood, WA, initialized at 0000 UTC on 12 Jun 2000, bias-corrected forecasts, BMA weights, and verifying observation. The kth bias-corrected forecast is equal to $a_k + b_k f_k$, where f_k is the kth forecast. Initial conditions (ICs) and lateral boundary conditions (LBCs) were obtained from AVN, the NGM Regional Data Assimilation System, and the ETA Data Assimilation System, all run by NCEP; the GEM analysis run by the CMC; and the NOGAPS analysis run by FNMOC. See Grimit and Mass (2002) for details.

MM5 initialization (source)	AVN (NCEP)	ETA (NCEP)	NGM (NCEP)	GEM (CMC)	NOGAPS (FNMOC)
Forecast	284.5	290.6	291.7	290.0	283.9
Bias-corrected forecast	285.2	291.2	292.4	290.8	285.5
BMA weight	0.38	0.27	0.03	0.24	0.08
Observation			292.6		

Thus BMA predicts a spread-error correlation, but it also accounts for the possibility that ensembles may be underdispersive. This is exactly what we observed in the UW ensemble, and it is also the case in other ensembles. BMA provides a theoretical framework for understanding these apparently contradictory phenomena and suggests ways to remedy them.

d. Example of BMA predictive PDF

To illustrate the operation of BMA, we first describe the prediction of just one quantity at one place and time; later we will give aggregate performance results. We consider the 48-h forecast of temperature at Packwood, Washington, initialized at 0000 UTC on 12 June 2000 and verifying at 0000 UTC on 14 June 2000. As described below, a 25-day training period was used, in this case consisting of forecasts and observations in the 0000 UTC cycle from 16 April to 9 June 2000. No bias correction was applied, apart from the estimation of the a_k and the b_k in the model, which can be viewed as a simple linear bias correction.

Table 1 shows the forecasts, the bias-corrected forecasts, the BMA weights for the five members of the UW MM5 ensemble, and the observation. There was strong disagreement among the ensemble members: two of them (AVN-MM5 and NOGAPS-MM5) were around 284 K, while the other three (ETA-MM5, NGM-MM5, and GEM-MM5) were around 291 K. This difference of 7 K is quite large. The verifying observation turned out to be outside the ensemble range, as happened for 71% of the cases in our dataset. The veri-

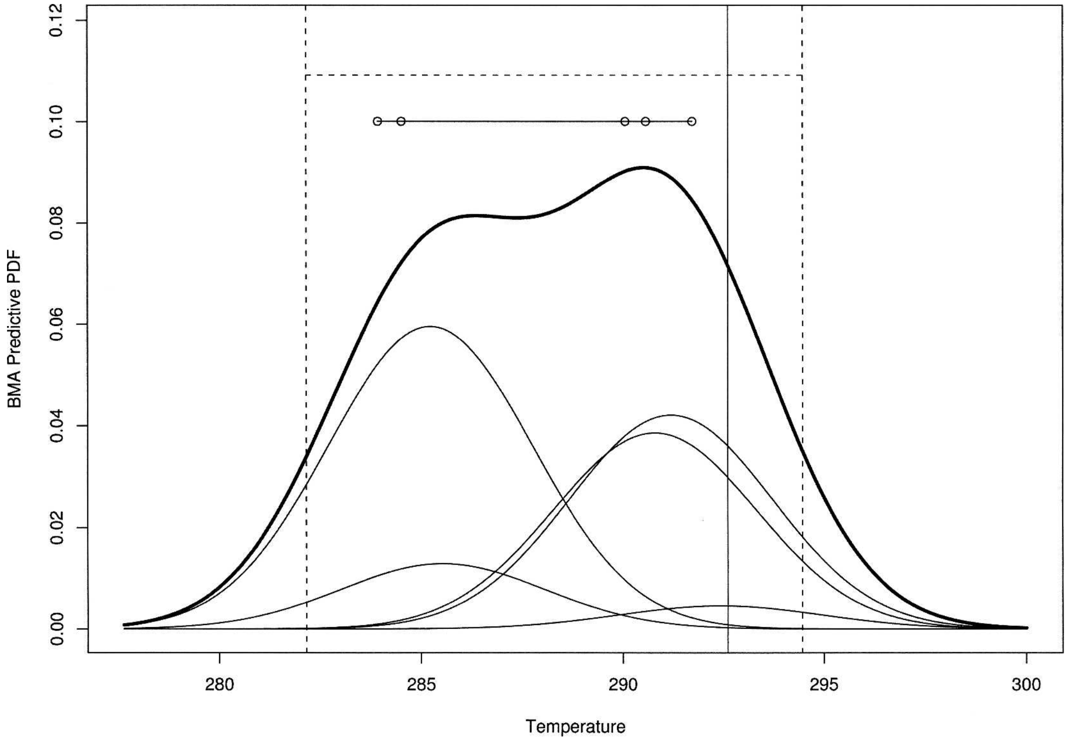

FIG. 3. BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-h surface temperature forecast at Packwood, WA, initialized at 0000 UTC on 12 Jun 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

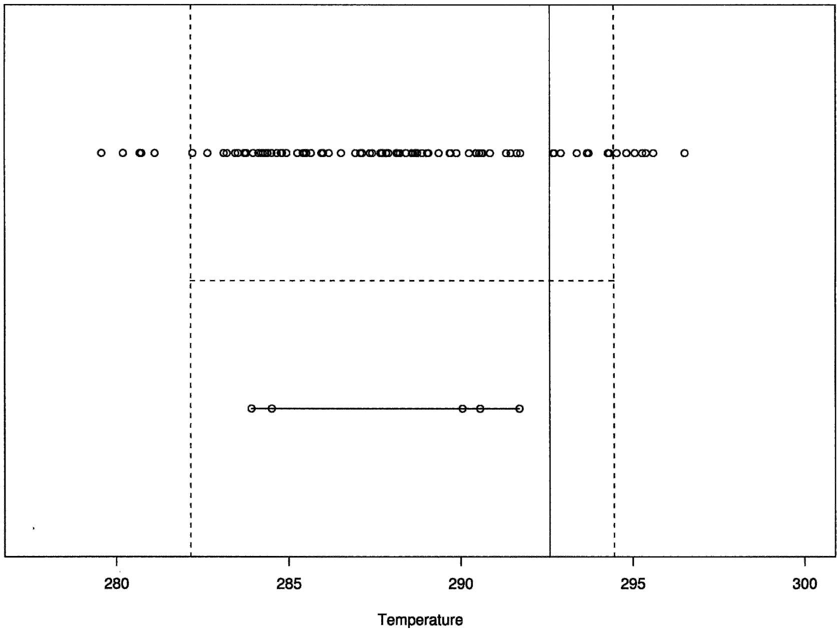

FIG. 4. Ensemble of 100 equally likely values from the BMA PDF (2) for the Packwood surface temperature forecast. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

fying observation was well forecast by the three higher forecasts (especially after the linear bias correction) and was far from the two lower forecasts.

Figure 3 shows the BMA predictive PDF. This PDF (shown as the thick curve in the figure) is a weighted sum of five normal PDFs (the components are the five thin lines). The distribution is bimodal, reflecting the fact that there are two groups of forecasts that disagree with one another. The right mode is centered around the cluster of three higher forecasts (after the linear bias correction), while the left mode is centered around the cluster of two lower forecasts. The observation fell within the 90% BMA prediction interval, even though it was outside the ensemble range.

The BMA PDF can also be represented as an unweighted ensemble of any desired size, simply by simulating from the predictive distribution (2). To simulate M values from the distribution (2), one can proceed as follows:

Repeat M times:

1) Generate a value of k from the numbers $\{1, \ldots, K\}$ with probabilities $\{w_1, \ldots, w_K\}$.
2) Generate a value of y from the PDF $g_k(y|f_k)$. In the present case this will be a $N(a_k + b_k f_k, \sigma^2)$ distribution.

Figure 4 shows a BMA ensemble of size $M = 100$ generated in this way. In this case, 87 of the 100 ensemble members lay within the exact 90% prediction interval. This differs slightly from the expected number of 90, but the difference is well within the range of what would be observed by chance.

The weights, w_k, reflect the ensemble members' overall performance over the training period, relative to the other members. Their rank order tends to be similar to that of the forecast root-mean-square errors (RMSEs), but this is not a direct relationship; they also reflect the correlations between the forecasts. Table 2 shows the RMSEs of both the raw and bias-corrected forecasts over the training period for the Packwood, Washington, forecast that we have been looking at, together with the BMA weights. With one exception, the rank order of the weights is the same as that of the rmses (reversed). The weights vary more than the RMSEs , however. This reflects the fact that the forecasts are highly correlated,

TABLE 2. RMSEs, bias-corrected RMSEs, and BMA weights for the forecasts over the 25-day training period preceding 12 Jun 2000.

MM5 initialization (source)	AVN (NCEP)	ETA (NCEP)	NGM (NCEP)	GEM (CMC)	NOGAPS (FNMOC)
RMSE	3.16	3.20	3.28	3.42	3.74
Bias-corrected RMSE	3.11	3.17	3.22	3.36	3.49
BMA weight	0.38	0.27	0.03	0.24	0.08

TABLE 3. Correlations between the surface temperature forecasts over the 25-day training period preceding 12 Jun 2000. The three NCEP-based forecasts (AVN-MM5, ETA-MM5, and NGM-MM5) are grouped together. They are more highly correlated with one another than with the forecasts from the other forecasting organizations (GEM from CMC, and NOGAPS from FNMOC).

	AVN	ETA	NGM	GEM	NOGAPS
AVN-MM5	1.00	0.95	0.95	0.90	0.92
ETA-MM5	0.95	1.00	0.98	0.91	0.91
NGM-MM5	0.95	0.98	1.00	0.91	0.91
GEM-MM5	0.90	0.91	0.91	1.00	0.89
NOGAPS-MM5	0.92	.91	0.91	0.89	1.00

as shown in Table 3. The AVN-MM5 forecast had the lowest RMSE and the highest BMA weight, at 0.38. The second ranked forecast, ETA-MM5, had an RMSE that was not much worse than AVN-MM5, but a much lower BMA weight, at 0.27. This reflects the fact that once one knows the AVN-MM5 forecast, the additional information provided by the ETA-MM5 forecast is much less than it would be if the two forecasts were uncorrelated.

The one exception was the NGM-MM5 model, which had the third best RMSE, but the lowest BMA weight, at 0.03. One can understand why this occurred by look-

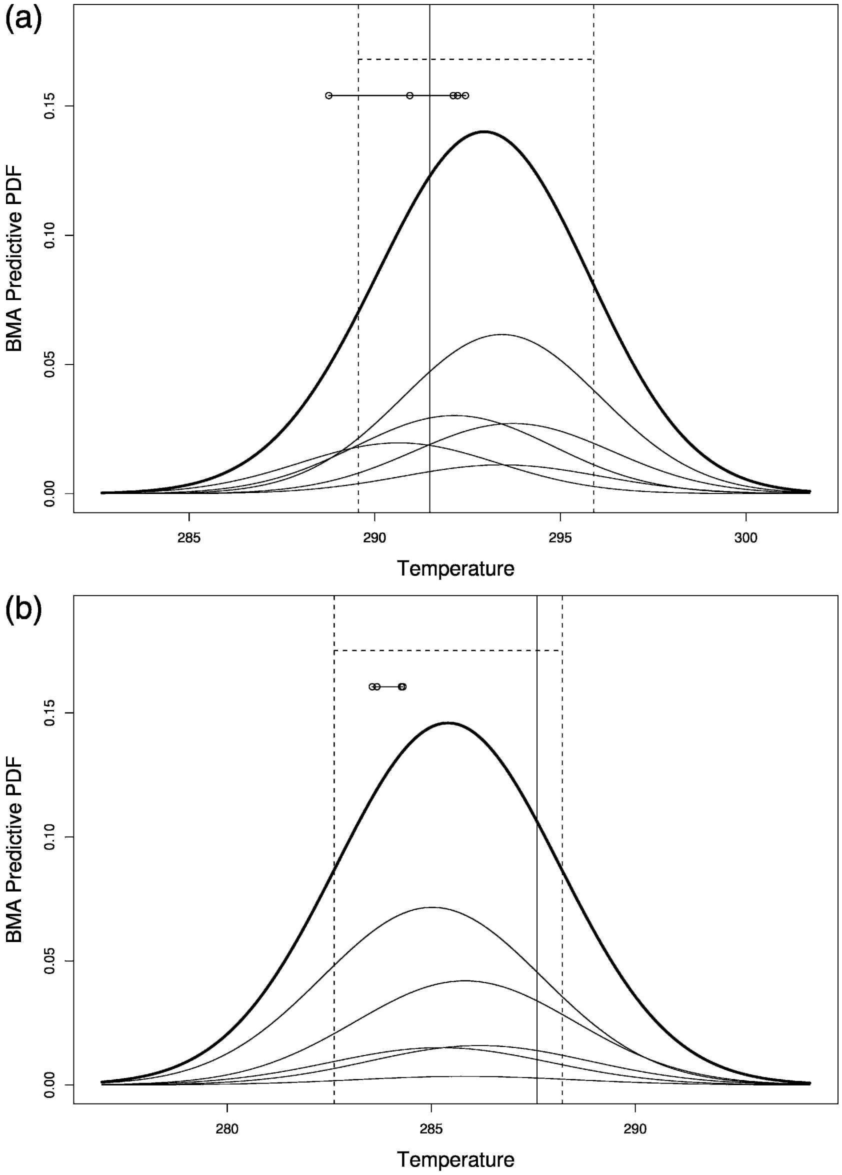

FIG. 5. BMA predictive PDFs for (a) an averagely dispersed ensemble and (b) an underdispersed ensemble. Both are for temperature at Packwood on different days. The same symbols are used as in Fig. 3.

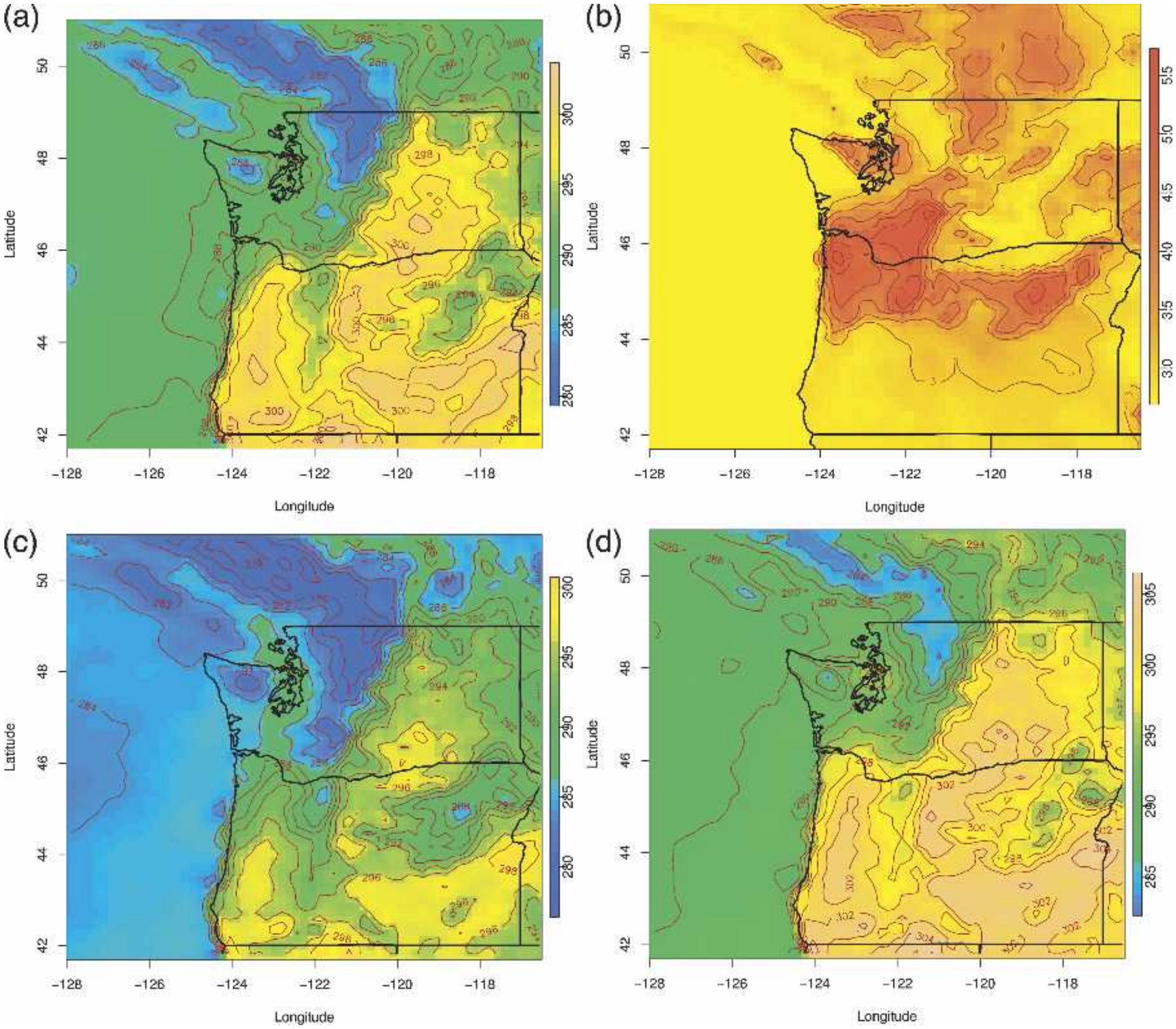

FIG. 6. BMA probabilistic 48-h forecast of surface temperature in the Pacific Northwest, initialized at 0000 UTC on 12 Jun 2000: (a) BMA deterministic forecast, (b) BMA margin of error, defined as half the width of the 90% prediction interval, (c) lower, and (d) upper bound of the 90% prediction interval.

ing again at the correlations between the forecasts in Table 3. The first three forecasts in the table, AVN-MM5, ETA-MM5, and NGM-MM5, are based on initializations produced by the same organization, NCEP. The remaining two forecasts are based on initializations from other organizations: GEM-MM5 from the Canadian Meteorological Centre (CMC) and NOGAPS-MM5 from the U.S. Navy [Fleet Numerical Meteorology and Oceanography Center (FNMOC)]. The correlations are in line with this: the three NCEP forecasts are very highly correlated, with correlations of 0.95–0.98, while the other two forecasts have correlations with one another and with the NCEP forecasts that are lower, even though still high. Thus once one knows the AVN-MM5 and ETA-MM5 forecasts, the NGM-MM5 forecast contributes very little additional information, because it is very highly correlated with the AVN-MM5 and ETA-MM5 forecasts, and of lower quality. On the other hand, the GEM-MM5 and NOGAPS-MM5 forecasts contribute more additional information because they are less correlated with the others, although they have worse RMSEs than the NGM-MM5 forecast.

Figure 5 shows the BMA predictive PDFs for two other days at Packwood. The first, in Fig. 5a, shows an ensemble with an average amount of dispersion, while the second, in Fig. 5b, shows an ensemble with a smaller than average amount of dispersion. Both are unimodal, as indeed were the majority of BMA PDFs in our dataset. Figure 5b illustrates the way in which BMA can yield reasonable intervals and PDFs, even when the ensemble is highly concentrated.

Figure 6 shows the BMA probabilistic 48-h forecast of temperature initialized at 0000 UTC on 12 June 2000

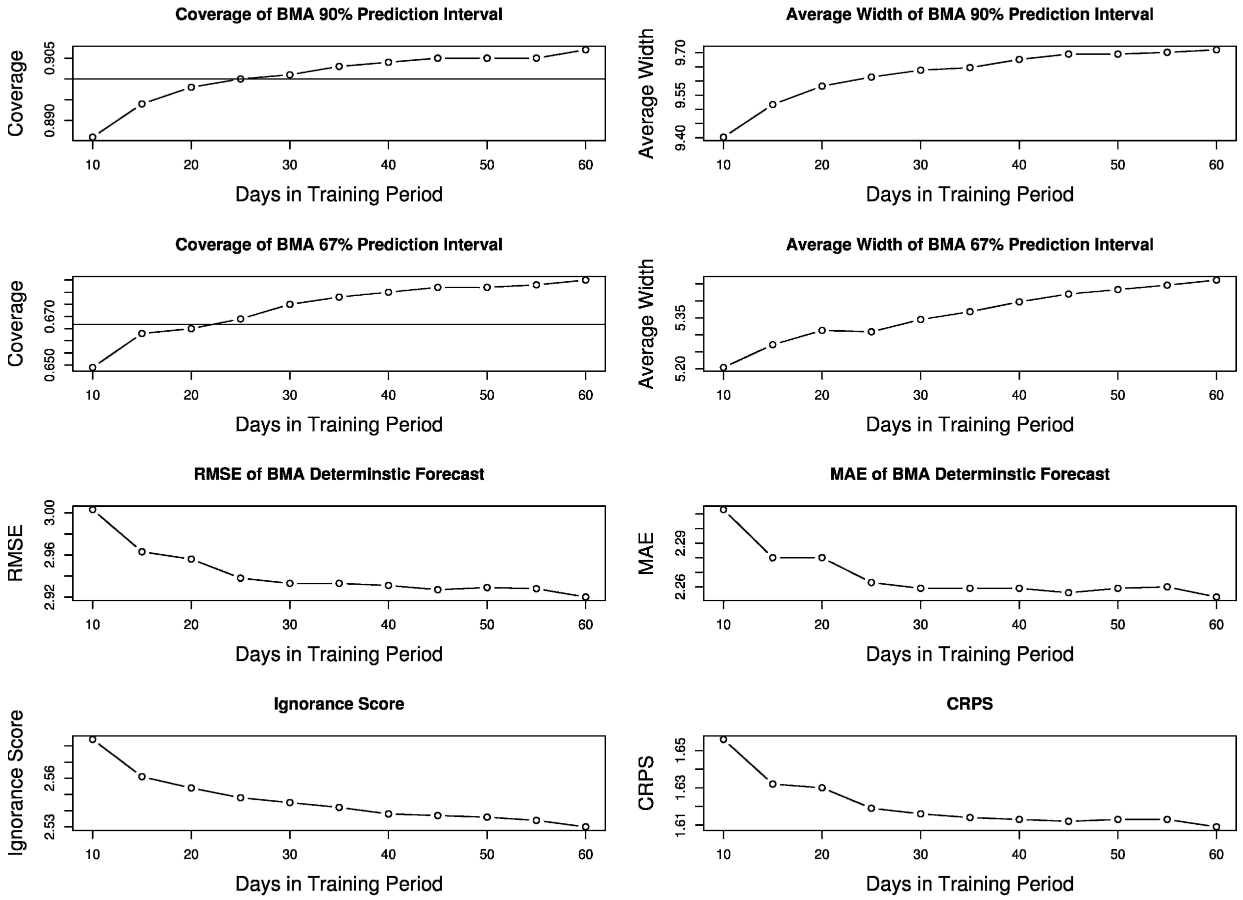

FIG. 7. Comparison of training period lengths for surface temperature: (a) coverage of 90% prediction intervals; (b) average width of 90% prediction intervals; (c) coverage of 66.7% prediction intervals; (d) average width of 66.7% prediction intervals; (e) RMSE of BMA deterministic forecasts; (f) MAE; (g) ignorance score; and (h) CRPS.

for the entire Pacific Northwest. Figure 6a shows the deterministic BMA forecast. Figure 6b shows the margin of error of the 90% prediction interval, defined as half the width of the interval. Roughly speaking, the prediction interval is approximately equal to the deterministic forecast plus or minus the margin of error, and the margin of error plot indicates where the uncertainty is greatest. Figures 6c and 6d display the lower and upper bounds of the 90% prediction intervals. These four plots show one way to summarize the probabilistic forecast of an entire field visually.

3. Results

We now give results of the application of BMA to 48-h forecasts of 2-m temperature in the Pacific Northwest for the 0000 UTC cycle in January–June 2000, using the UW-MM5 ensemble described by Grimit and Mass (2002). We first describe how we chose the length of the training period, then we give the main results, and finally we outline how the results could be used to select the members of a possibly reduced ensemble. We also give summary results for sea level pressure over the same period.

a. Length of training period

How many days should be used in the sliding-window training period to estimate the BMA weights, variance, and bias-correction coefficients? There is a trade-off here, and no automatic way of making it. Both weather patterns and model specification change over time, and there is an advantage to using a short training period so as to be able to adapt rapidly to such changes. In particular, the relative performance of the models changes. On the other hand, the longer the training period, the better the BMA parameters are estimated.

In making our choice, we were guided by the principle that probabilistic forecasting methods should be designed to maximize sharpness subject to calibration, that is, to make the prediction intervals as short as possible subject to their having the right coverage (Gneiting et al. 2003). We also tend toward making the training period as short as possible so as to be able to adapt as quickly as possible to the changing relative perfor-

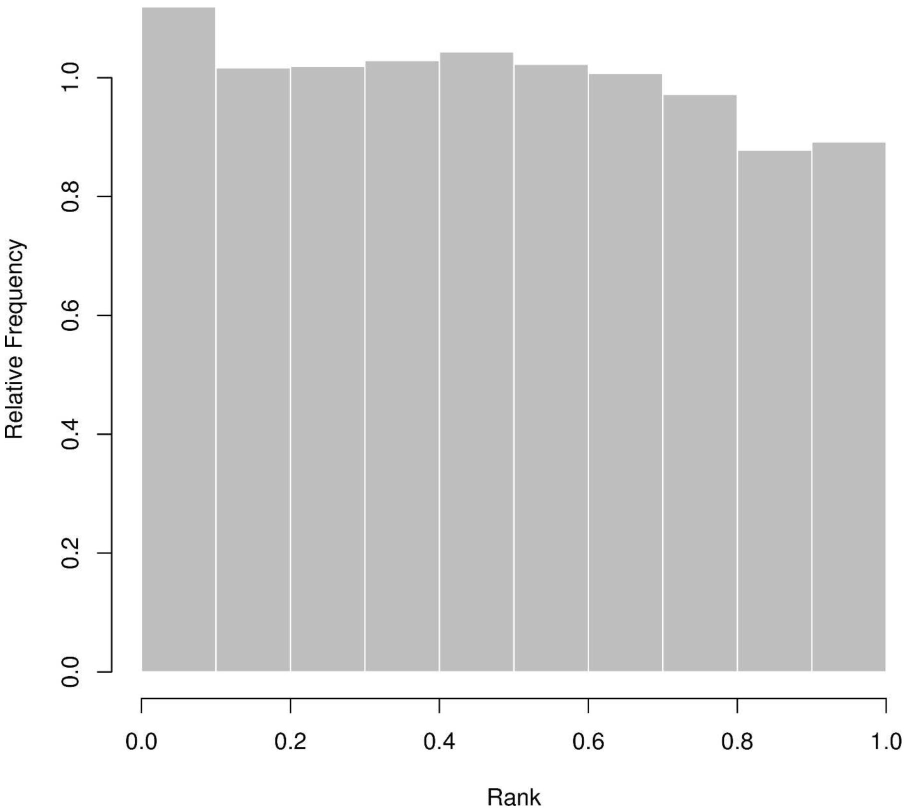

FIG. 8. PIT histogram for BMA for surface temperature.

mance of ensemble members, lengthening it only if doing so seems to confer a clear advantage.

Here we focus on 66.7% and 90% prediction intervals. We considered training periods of lengths 10, 15, 20, . . . , 60 calendar days. For comparability, the same verifications were used in evaluating all the training periods, and the verifications for the first 63 days were not used for evaluation. For some days the data were missing (Grimit and Mass 2002), so that the number of calendar days spanned by the training dataset was typically larger than the actual number of days of training data used.

Figure 7a shows the coverage of BMA 90% prediction intervals. The coverage increases with the number of training days, hitting the correct 90.0% at 25 days, and increasing beyond that. Figure 7b shows the average width of BMA 90% prediction intervals. This increases with the number of training days, indicating that shorter training periods yield sharper forecasts. Figures 7c and 7d show the same quantities for the 66.7% intervals, with similar conclusions.

Figures 7e and 7f show the RMSE and the mean absolute error (MAE) of BMA deterministic forecasts corresponding to different lengths of the training period. These decrease substantially as the number of training days increases, up to 25 days, and change little as the number of days is increased beyond 25. Figure 7g shows the ignorance score for BMA. This is the average of the negative of the natural logarithms of the BMA PDFs evaluated at the observations. It was proposed by Good (1952), and its use in the present context was suggested by Roulston and Smith (2002). Smaller scores are preferred. This decreases sharply at first, and then more slowly. Figure 7h shows the CRPS. This also decreases with the number of training days, sharply from 10 to 25 days, and flattens out after that.

To summarize these results, it seems that there are substantial gains in increasing the training period up to 25 days, and that beyond that there is little gain. We have therefore used 25 days here. It seems likely that different training periods would be best for other variables, forecast cycles, forecast lead times, time periods, and regions. Further research on how best to choose the length of the training period is needed, and a good automatic way of doing this would be useful.

b. *Results*

We now give results for BMA, using the same evaluation dataset as was used to compare the different training periods. The probability integral transform (PIT) histogram for BMA is shown in Fig. 8. This is a continuous analog of the verification rank histogram. To compute the PIT histogram we proceeded as follows. For each forecast initialization time at each sta-

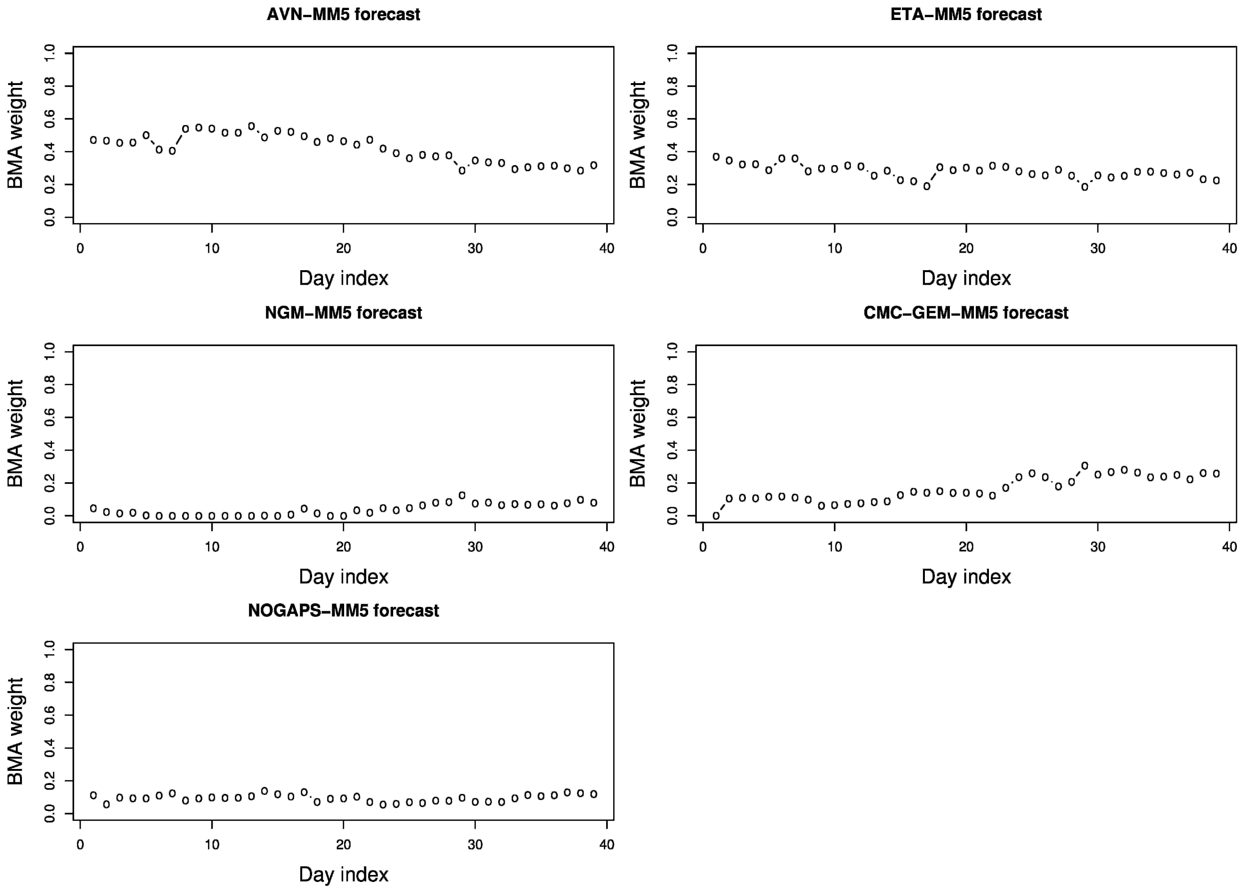

FIG. 9. BMA weights for the five models over the evaluation period for surface temperature.

tion, we computed the BMA cumulative distribution function (CDF), and we found its value at the verifying observation. We then formed the histogram of these BMA CDF values. This should be uniform if the predictive PDF is calibrated; it can be compared directly with the verification rank histogram of the underlying ensemble in Fig. 2a. As can be seen, it was reasonably well calibrated, and clearly much more so than the ensemble itself.

Figure 9 shows the BMA weights for the five ensemble members over the evaluation period. These varied relatively little. The NGM-MM5 forecast had low weights throughout, suggesting that it is not useful relative to the other four ensemble members.

Table 4 shows the coverage of various prediction intervals. We included the prediction interval from sample climatology, that is, from the marginal distribution of our full dataset; this interval is the same for each day and is useful as a baseline for methods that use the numerical weather predictions. The climatological forecast is of course well calibrated, but at the expense of producing very wide intervals, as we will see. The ensemble range is underdispersive, as we have already seen. The BMA intervals are very close to having the right coverage.

Table 5 shows the average widths of the prediction intervals considered. The ensemble range was much narrower on average than the climatological 66.7% interval, but the price of this was that the ensemble range was far from being a calibrated interval and was underdispersive. The BMA 66.7% interval was wider on average than the ensemble range, but still much less so

TABLE 4. Coverage of prediction intervals for surface temperature (%).

Interval	66.7% interval	90% interval
Sample climatology	66.7	90.0
Ensemble range	28.7	—
BMA	66.9	90.0

TABLE 5. Average width of prediction intervals for surface temperature.

Interval	66.7% interval	90% interval
Sample climatology	17.2	28.3
Ensemble range	2.5	—
BMA	5.3	9.6

TABLE 6. RMSEs and MAEs of deterministic forecasts for surface temperature. The bias-corrected forecast for the kth ensemble member at place s and time t is equal to $a_k + b_k f_{kst}$.

Forecast	RMSE	MAE
Sample climatology	9.58	7.69
Raw forecasts		
AVN-MM5	3.15	2.45
ETA-MM5	3.23	2.52
NGM-MM5	3.28	2.56
CMC-GEM-MM5	3.40	2.64
NOGAPS-MM5	3.76	2.96
Bias-corrected forecasts		
Bias-corrected AVN-MM5	3.01	2.32
Bias-corrected ETA-MM5	3.11	2.40
Bias-corrected NGM-MM5	3.14	2.43
Bias-corrected CMC-GEM-MM5	3.25	2.49
Bias-corrected NOGAPS-MM5	3.25	2.50
Ensemble forecasts		
Ensemble mean	3.18	2.49
BMA	2.94	2.26

than sample climatology, and it is calibrated, unlike the ensemble range. The climatological and BMA 90% intervals were both approximately calibrated, but the BMA intervals were 69% narrower on average.

Table 6 shows the RMSEs and MAEs of the various deterministic forecasts considered over the evaluation period. The numerical weather prediction forecasts performed much better than the forecast from sample climatology (with all forecasts equal to the sample mean), and among these the AVN-MM5 forecast was best on average. The bias-corrected forecasts, $a_k + b_k f_{kst}$, have lower RMSEs and MAEs than the raw forecasts, by about 5%. The ensemble mean is as good as any of the raw forecasts, but not as good as the best bias-corrected forecasts. The BMA deterministic forecast given by (4) performed better than any of the other 12 forecasts considered, in terms of both RMSE and MAE. In terms of MAE, it was better than sample climatology by 71%, the best single forecast (AVN-MM5) by 8%, the best single bias-corrected forecast by 3%, and the ensemble mean by 9%. The results were similar in terms of RMSE.

c. *Selecting ensemble members: Results for a reduced ensemble*

Ensemble forecasts are very demanding in terms of computational time, and so it is important that the members of the ensemble be carefully selected. The number of ensemble runs that can be done by an organization is limited, and large ensembles make demands on computer and personnel resources that could be used for other purposes.

Our approach provides a way of selecting ensemble members in situations where the individual ensemble members come from different, identifiable sources. The BMA weights provide a measure of the relative usefulness of the ensemble members, and so it would seem reasonable to consider eliminating ensemble members that consistently have low weights. Over our evaluation period, the NGM-MM5 forecasts had low weights on average, averaging under 0.04. One might then consider eliminating this member, and using instead a reduced four-member ensemble.

Table 7 compares the results for the five-member and the reduced four-member ensemble over the evaluation period. They are almost indistinguishable, and the ignorance score actually improves slightly when the least useful member is removed. This would suggest that this member can be removed with little cost in terms of performance, and the operational gain could be considerable. Indeed, the NGM-MM5 model was removed from the UW ensemble shortly after the end of our test period, in August 2000. Before making such a decision in general, however, it would be necessary to study the BMA weights over a longer period and for all the variables and forecast lead times of interest. Ensemble members that contribute little to forecasting one variable might be useful for others.

d. *Results for sea level pressure*

We now briefly summarize the results for 48-h forecasts of sea level pressure for the same region and time period. These are qualitatively similar to those for 2-m temperature. The results for choice of training period are similar to those for temperature, and again point to a 25-day training period as being best. The unit used is the millibar (mb).

Table 8 shows the coverage of the various prediction intervals. The ensemble range is underdispersive, but less so than for temperature, while BMA is well calibrated. Table 9 shows the average widths of the prediction intervals for sea level pressure. The ensemble range is narrow but, as we have seen, this is at the cost of not being well calibrated. The BMA intervals are

TABLE 7. Comparison of BMA probabilistic forecasts from the five-member ensemble and the reduced four-member ensemble for surface temperature.

	90% prediction interval					
	Coverage	Average width	RMSE	MAE	Ignorance score	CRPS
Five-member ensemble	90.0%	9.6	2.94	2.26	2.55	1.62
Four-member ensemble	90.0%	9.6	2.93	2.26	2.55	1.62

TABLE 8. Coverage of prediction intervals for sea level pressure (%).

Interval	66.7% interval	90% interval
Sample climatology	66.7	90.0
Ensemble range	53.9	—
BMA	65.4	90.4

63% narrower than the intervals from sample climatology. Sample climatology seems like a more useful baseline for sea level pressure than for temperature, because sea level pressure is a synoptic variable with relatively little sensitivity to topography and a weak seasonal effect, if any. So the good performance of BMA relative to sample climatology, achieving considerable sharpness while remaining calibrated, is striking.

Table 10 shows the RMSEs and MAEs of the various deterministic forecasts. Once again, the BMA deterministic forecast outperforms the other 12 forecasts considered in terms of both RMSE and MAE. In terms of MAE, it is 56% better than sample climatology, 7% better than the best single forecast (AVN-MM5), 2% better than the best bias-corrected forecast, and 3% better than the ensemble mean.

4. Experiments with simulated ensembles

We now report the results of several experiments with simulated ensembles. Some of these relate to calibrated ensembles. Most ensembles in current use appear to be underdispersive, so the current primary need is for methods that work well with underdispersive ensembles. However, as ensembles become better, it will become critical that postprocessing methods work well with calibrated ensembles as well as underdispersive ones.

Our experiments use the same data structure as the temperature dataset we have already considered. The observations and forecasts in the temperature dataset are replaced by simulated values. BMA is implemented as before, using 25 training days to estimate the model.

a. Experiment 1: A calibrated ensemble with varying means and variances

Our first experiment simulated a calibrated ensemble in which the true predictive PDF for each day and station is normally distributed, with its own mean and variance. The means were themselves simulated from a distribution, which was chosen to reflect the temperature observations and forecasts in the ensemble we have analyzed; the same was true for the variances. For each day and station, six observations were drawn from the normal distribution for that day and station; five of these were the ensemble forecast members, and one was the verifying observation. Thus the ensemble was calibrated by design.

TABLE 9. Average width of prediction intervals for sea level pressure.

Interval	66.7% interval	90% interval
Sample climatology	13.2	21.8
Ensemble range	3.9	—
BMA	4.9	8.3

TABLE 10. RMSEs of deterministic forecasts for sea level pressure.

Forecast	RMSE	MAE
Sample climatology	5.70	4.61
Raw forecasts		
AVN-MM5	2.90	2.20
ETA-MM5	3.25	2.50
NGM-MM5	3.40	2.70
CMC-GEM-MM5	3.00	2.35
NOGAPS-MM5	3.21	2.50
Bias-corrected forecasts		
Bias-corrected AVN-MM5	2.66	2.09
Bias-corrected ETA-MM5	3.14	2.40
Bias-corrected NGM-MM5	3.20	2.47
Bias-corrected CMC-GEM-MM5	2.92	2.29
Bias-corrected NOGAPS-MM5	2.67	2.11
Ensemble forecasts		
Ensemble mean	2.73	2.11
BMA	2.59	2.05

The simulation was implemented as follows. The mean, μ_{st}, for station s on day t was simulated from a normal distribution with mean $\overline{\mu}$ and variance σ_{μ}^2, a situation we denote by $\mu_{st} \sim N(\overline{\mu}, \sigma_{\mu}^2)$. The variance, υ_{st}, for station s on day t, was simulated from a chi-square distribution with 12 degrees of freedom, multiplied by $\overline{\upsilon}/12$, a situation we denote by $\upsilon_{st} \sim \overline{\upsilon}\chi_{12}^2/12$. We chose $\overline{\mu} = 286$, $\sigma_{\mu}^2 = 66.8$, and $\overline{\upsilon} = 6.0$. This ensured that the mean and variance of the observations and the forecasts, and the average RMSE of the forecasts, were close to those in the data. With this setup, the correlation between the observation and any single forecast is 0.92, and this is also equal to the correlation between any two ensemble members. The setups for all six experiments, including this one, are summarized in Table 11.

The results are shown in Table 12. The RMSE for BMA is considerably smaller than for sample climatology, or for any single forecast. The 66.7% prediction interval is slightly overdispersed, but it is no wider on average than the ensemble range, which also has nominal coverage 66.7%. The 90% BMA prediction interval has coverage 91.7% and is much narrower on average than the climatological interval: 9.4 compared to 28.1, or 67% shorter. The ensemble does not provide a 90% prediction interval directly, of course. The BMA PIT histogram is shown in Fig. 10a.

Figure 11 gives an example of how BMA achieves this result with a calibrated ensemble. The ensemble members are essentially equally weighted, and the BMA PDF is in this case close to being a normal dis-

TABLE 11. Summary of experiments with simulated ensembles. Shown are the mean and variance of the true predictive PDF and the mean and variance of the forecast ensemble. All observations and forecast ensemble members were simulated from normal distributions. Note that μ_{st} and υ_{st} were simulated independently for each day t and station s, as follows: $\mu_{st} \sim N(\overline{\mu}, \sigma_\mu^2)$, $\upsilon_{st} \sim \overline{\upsilon}\chi_{12}^2/12$, where $\overline{\mu} = 286$, $\sigma_\mu^2 = 66.8$, and $\overline{\upsilon} = 6.0$.

No.	True mean	True variance	Ensemble mean	Ensemble variance	Description of ensemble
1	μ_{st}	υ_{st}	μ_{st}	υ_{st}	Calibrated, with varying mean and variance
2	0	1	0	1	Calibrated, uncorrelated
3	0	1	0	0.25	Underdispersive, uncorrelated
4	0	1	0	4	Overdispersive, uncorrelated
5	μ_{st}	υ_{st}	μ_{st}	0.25 υ_{st}	Underdispersive, with varying mean and variance
6	μ_{st}	υ_{st}	μ_{st}	4 υ_{st}	Overdispersive, with varying mean and variance

tribution itself. The parameters b_k in Eq. (3) are about 0.90, so that the forecasts are slightly shrunk toward the mean, largely avoiding overdispersion.

b. *Experiments 2, 3, and 4: Uncorrelated experiments*

We now consider three experiments in which the true predictive PDF was $N(0, 1)$. In the first of these, experiment 2, the ensemble members were also drawn from $N(0, 1)$, so this was a calibrated ensemble. However, the forecasts and the observations were uncorrelated in this case, and so the ensemble was uninformative. While one would hope that this is a rare situation, it can be viewed as a limiting case of a forecast with little skill. This does arise; for example, in the Pacific Northwest, wind speed forecasts have relatively little skill because of the Pacific data void and the mountainous terrain. It seems desirable that a statistical postprocessing method would accurately convey uncertainty in this kind of situation.

In experiment 3, the ensemble members were drawn from $N(0, 0.25)$, so this was an underdispersive uninformative ensemble, while in experiment 4 they were drawn from $N(0, 4)$, so this was an overdispersive uninformative ensemble.

The results are shown in Table 12. For experiment 2, the calibrated ensemble, BMA was well calibrated, and actually had 66.7% prediction intervals that were shorter on average than the ensemble range. In this case, the parameters b_k in Eq. (3) are close to zero, and BMA essentially cuts back automatically to climatology, which is the best one can do in this situation.

For experiments 3 and 4, the under- and overdispersive ensembles, the ensemble range was very poorly calibrated, as one would expect, but BMA remained well calibrated, as can be seen from the coverages in Table 12 and the PIT histograms in Fig. 10.

c. *Experiments 5 and 6: Under- and overdispersive ensembles with varying means and variances*

Experiment 5 was a modified version of experiment 1, modified so that the ensemble is underdispersive. Experiment 6 was also a modified version of experiment 1, this time modified to make it overdispersive.

As can be seen from Table 12, the ensemble range was very poorly calibrated in both cases, not surprisingly. The coverage of the ensemble range for experiment 5 was similar to that observed in the actual data we analyzed; thus this is perhaps the most relevant experiment to the actual ensemble forecasting data we have been looking at. For experiment 5, BMA is well calibrated and sharp. For experiment 6, BMA is much better calibrated than the ensemble range, and much sharper than climatology, but it is still slighly overdispersed. This seems hard to avoid when the ensemble itself is highly overdispersive, but this is rarely observed in practice.

Figure 11b gives an example of a BMA PDF for the underdispersive ensemble. Essentially, BMA spreads out the ensemble range. Figure 11c gives an example of a BMA PDF for the overdispersive ensemble. Here the PDF is slightly multimodal, reflecting the fact that when the ensemble is very dispersed, the forecasts are to some extent in conflict with one another.

TABLE 12. Results of experiments with simulated ensembles.

	Experiment					
	1	2	3	4	5	6
RMSE						
Climatology	8.5	1.0	1.0	1.0	8.5	8.5
Single forecast	3.5	1.4	1.1	2.2	2.5	10.1
BMA	2.6	1.0	1.0	1.0	2.5	3.7
66.7% interval: Coverage (%)						
Climatology	66.7	66.7	66.7	66.7	66.7	66.7
Ensemble range	66.7	66.7	41.5	84.6	22.8	91.3
BMA	71.7	66.9	66.4	66.5	68.4	72.0
66.7% interval: Width						
Climatology	16.5	1.9	1.9	1.9	16.5	16.5
Ensemble range	5.6	2.3	1.2	4.7	1.4	22.3
BMA	5.6	1.9	1.9	1.9	4.8	8.2
90% interval: Coverage (%)						
Climatology	90.0	90.0	90.0	90.0	90.0	90.0
BMA	91.7	90.1	90.0	90.1	90.4	91.9
90% interval: Width						
Climatology	28.1	3.3	3.3	3.3	28.1	28.1

5. Discussion

We have proposed a new method for statistical postprocessing of ensemble output to produce calibrated

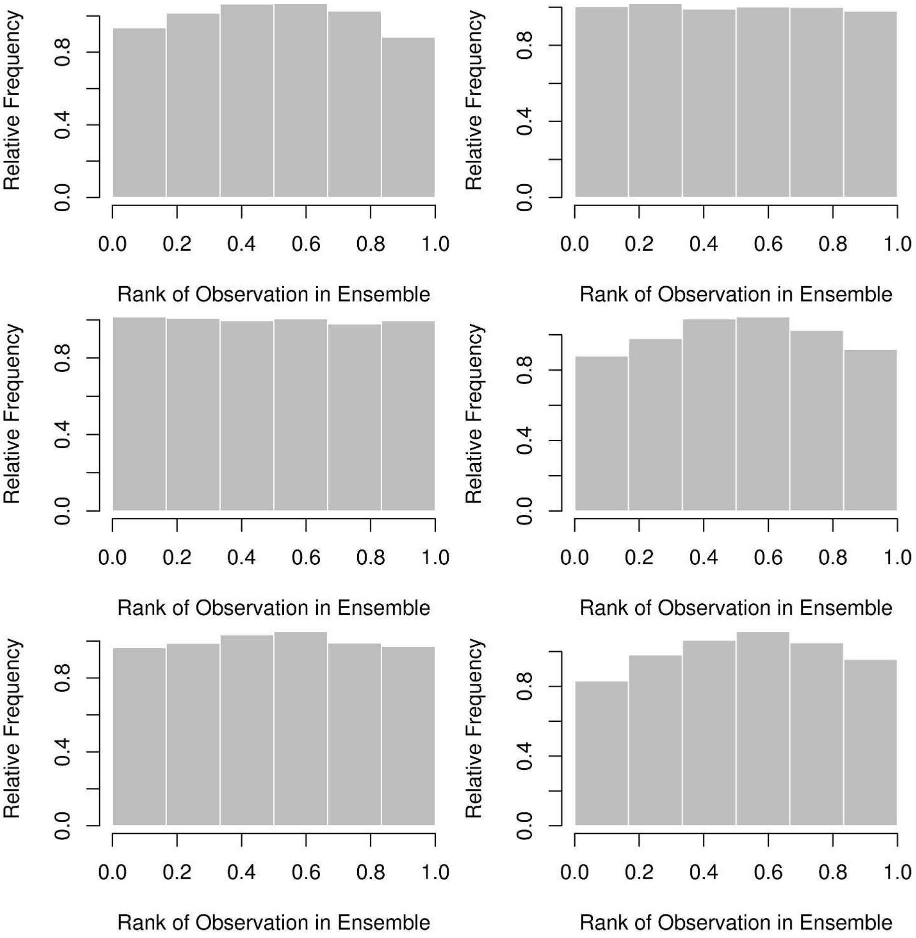

FIG. 10. PIT histograms for BMA for the six simulated ensemble experiments, ordered by row.

and sharp predictive PDFs. It is based on Bayesian model averaging, a statistical method for combining forecasts from different models and analyses, and provides a theoretical explanation of the empirical phenomenon of ensembles exhibiting a spread-error correlation while still being underdispersive. In our case study, the BMA PDFs were much better calibrated than the ensemble itself and produced prediction intervals that were much sharper than those produced by sample climatology. In addition, the BMA deterministic forecast had a lower RMSE than any of the individual ensemble members, and also than the ensemble mean, although the latter was also better than any of the ensemble members.

Our approach uses observations and forecasts to estimate the BMA model for a spatial region and is thus applicable to the production of probabilistic forecasts on a grid. In our experiment we applied it to the UW-MM5 ensemble's 12-km domain, the Pacific Northwest, and it would seem desirable that the model be estimated separately for different spatial regions. Clearly such regions should be fairly homogeneous with respect to the variable being forecast, but precisely how to determine them needs further research. We have used observations to estimate the model, but it would be possible to do so also using an analysis, and this may be preferable in regions where there are few observational assets.

Our experiments here have been with short-range mesoscale probabilistic forecasting from multianalysis ensembles, but it would seem feasible also to apply the idea to other situations, including medium-range and synoptic forecasting, and to perturbed observations, singular vector, bred, and poor man's ensembles. Our implementation has been for the situation where the ensemble members come from clearly distinguishable sources. In other cases, such as the current synoptic NCEP and ECMWF ensembles, it may be more appropriate to consider some or all of the ensemble members as being from the same source, and hence to treat them equally. This can be accommodated within our approach with a small change in the model: for ensemble members viewed as equivalent, the BMA weights w_k in (2) would be constrained to be equal. The EM algorithm can still be used, with a small modification.

In our experiments we have assumed that the condi-

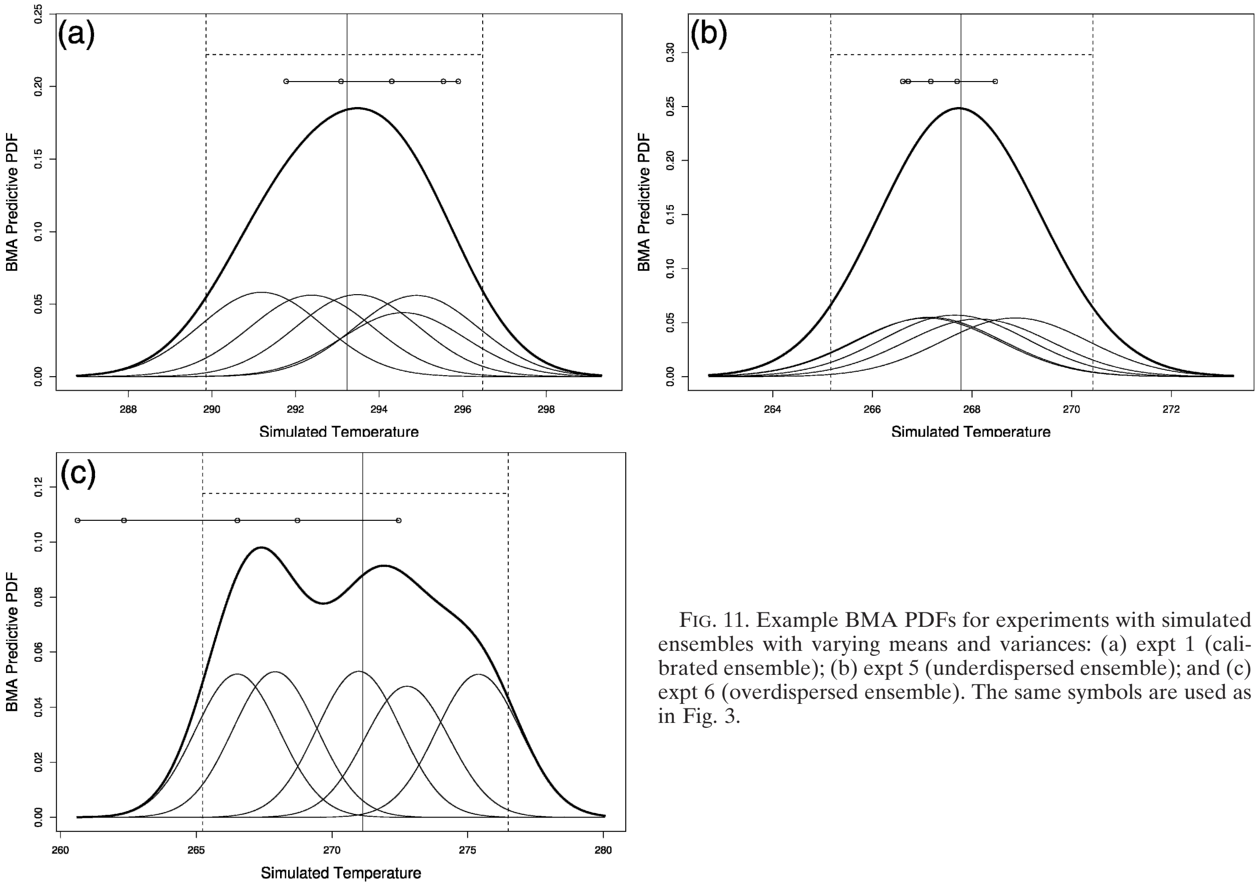

FIG. 11. Example BMA PDFs for experiments with simulated ensembles with varying means and variances: (a) expt 1 (calibrated ensemble); (b) expt 5 (underdispersed ensemble); and (c) expt 6 (overdispersed ensemble). The same symbols are used as in Fig. 3.

tional densities $g_k(y|f_k)$ in the BMA model (2) can reasonably be taken to be normal densities. This works well for surface temperature, and in the experiments that we summarized above it also worked well for sea level pressure data. However, this may not apply so directly to wind speed and precipitation data, because they tend to have a positive probability of being equal to or very close to zero, and because their distribution tends to be skewed. The BMA approach could be extended to these situations by using a nonnormal conditional distribution $g_k(y|f_k)$ in (2). It has been common to model wind speed using a Weibull distribution and precipitation using a gamma distribution, and it may be necessary to augment these with a component representing a positive probability of being equal to zero. This can be done within the framework of generalized linear models (McCullagh and Nelder 1989), and one example of how to model precipitation in this way was given by Stern and Coe (1984).

One way to improve the performance of this method is to bias correct the forecasts before applying BMA. The linear regression of observations on forecasts is incorporated in our implementation of the BMA method and can be viewed as a very simple bias correction, but it is possible to do much better. Model output statistics (MOS) is the dominant approach to bias correction and may give improved results (Wilks 1995). Approaches based on spatial and temporal neighborhoods have also been proposed, for example, by Eckel and Mass (2003). Note that to be useful in our context, bias-correction methods need to be applicable to grid-based forecasts and not just to forecasts at observation sites. It is clear from (2) that the MOS and neighborhood bias-correction methods mentioned can be combined with BMA to produce probabilistic forecasts.

Our method produces a predictive PDF for one location, but it does not reproduce the spatial correlation properties of error fields. Various ways of creating ensembles of entire fields that reproduce the spatial correlation of the error field have been proposed for the situation where just one numerical weather prediction model and initialization is used (Houtekamer 1993; Houtekamer and Mitchell 1998, 2001; Gel et al. 2004). Such methods could be combined with the present proposal to produce multimodel and/or multianalysis ensembles that reproduce spatial correlation of error fields by creating ensembles of fields corresponding to each ensemble member and simulating a number of fields from each of these ensembles that is proportional to the corresponding BMA weight.

Hamill and Colucci (1997, 1998) proposed a statistical postprocessing method based on directly adjusting the probabilities in the rank histogram; this method was applied by Eckel and Walters (1998). This worked well, but differs from the present approach in not being based on an explicit probabilistic model. Wilks (2002) proposed smoothing forecast ensembles by fitting mixtures of Gaussian distributions to ensemble data. This is related to the BMA approach, but does not account for ensemble underdispersion.

A different approach to postprocessing an ensemble called "best member dressing" has been proposed by Roulston and Smith (2003). This consists of identifying the best member of an ensemble for each element of a historic record, finding the error in that ensemble member forecast, finding the empirical distribution of such errors, and then "dressing" each forecast in the ensemble with the empirical error distribution found in this way. Viewed in this way, BMA could also be viewed as a way of dressing an ensemble of forecasts. The approaches differ in some ways, however. The method of Roulston and Smith (2003) is designed for the situation where all the ensemble members can be treated equally, as exchangeable, and, for example, it treats the ECMWF control forecast in the same way as the other 50 members of the ECMWF ensemble. In contrast, BMA is applicable to the situation where the ensemble members come from different, identifiable sources, but is also applicable to the exchangeable situation, as we have noted. For example, BMA would allow different treatment of the control and other ECMWF ensemble members in a straightforward way.

Best member dressing is based on the assumption that the best member can be identified with high probability, and as such does not take uncertainty about the identification of the best member into account. Usually, however, there is considerable uncertainty about which is the best member. The best member dressing method attempts to overcome this problem by using a large number of variables when identifying the best member. But this is based on the assumption that all the variables used have a common best member. This is an issue with the best member dressing approach, whereas BMA does not make this assumption.

In contrast, BMA takes account of this uncertainty through the use of the mixture likelihood (5), and it is estimated explicitly by the $\hat{z}_{kst}$ that are produced by the EM algorithm, given by Eq. (6). The quantity $\hat{z}_{kst}$ can be interpreted as the posterior probability that forecast k was the best member of the ensemble at place s on day t.

BMA is designed to produce probabilistic forecasts, but as a by-product it also produces a deterministic forecast, and this outperformed all the ensemble members as well as the ensemble mean in our experiments. It has also been proposed that forecasts be combined using multiple linear regression to produce a single deterministic forecast or "superensemble" (Van den Dool and Rukhovets 1994; Krishnamurti et al. 1999; Kharin and Zwiers 2002). It seems likely that BMA and regression would give similar forecasts. However, one difference is that the weights in BMA are constrained to be positive, whereas those in regression are not; see, for example, Tables 2, 4, 5, and 6 in Van den Dool and Rukhovets (1994). Negative weights seem hard to interpret in this context; they imply that, all else being equal, temperature (for example) is predicted to be higher when the forecast with the negative weight is lower. Stefanova and Krishnamurti (2002) have proposed a way of using superensembles to estimate the probability of a dichotomous event. This does not appear to apply to estimating the PDFs of continuous weather quantities, and the problem of interpreting negative coefficients continues to apply in this case.

We have used a training period consisting of the past 25 days of data. This works well for the UW ensemble because it allows the method to adapt quickly not just to changes in the distribution of forecast errors and the relative performance of the models, but also to changes in the ensemble itself. The UW ensemble itself has experienced significant changes, in membership or in other ways, several times per year on average. In addition, ensemble members may also change because of improvements in data coverage, data assimilation methods, resolution, model physics, integration methods, and computing power. Thus we have chosen a training period in the recent past.

However, it would also be possible to use data from previous years while restricting the training data to the same season, as suggested by Hamill et al. (2004). This would have the advantage of providing more data. But it would have the disadvantage of using data from an ensemble that might be quite different from the one in current use, unless one takes the ambitious route of ensemble reforecasting (Hamill et al. 2004), which is extremely computationally intensive. There is a tradeoff here, and how best to make it is an empirical question.

Probabilistic temperature forecasts using BMA and the (now eight-member) UW ensemble are now being produced on a regular basis and are available online at http://isis.apl.washington.edu/bma/index.jsp. This provides maps of the BMA deterministic forecast, the upper and lower bounds for any desired probability levels, and the margin of error, as in Fig. 6. It also provides a map of the probability of temperature being greater than or less than any specified threshold. In addition, one can click on any of the maps and obtain a picture of the BMA PDF at the grid point clicked on, with its component densities, similar to Fig. 3.

Acknowledgments. The authors are grateful to Clifford Mass, Mark Albright, Eric Grimit, and Peter Houtekamer for helpful discussions and useful comments, and for providing data. They are also grateful to Patrick Tewson for implementing the UW Ensemble

BMA Web site. They thank the three reviewers for very helpful comments that greatly improved the quality of the article. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

REFERENCES

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

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

——, 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.

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

Casella, G., and R. L. Berger, 2001: *Statistical Inference.* 2d ed. Brooks Cole, 660 pp.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, **39B,** 1–39.

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

——, and C. F. Mass, 2003: Towards an effective short-range ensemble forecast system. *Proc. Workshop on Ensemble Forecasting,* Val-Morin, QC, Canada. [Available online at http://www.cdc.noaa.gov/people/tom.hamill/ef_workshop_2003.html.]

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **222A,** 309–368.

Gel, Y., A. E. Raftery, and T. Gneiting, 2004: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion). *J. Amer. Stat. Assoc.,* **99,** 575–590.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1202–1211.

Gneiting, T., A. E. Raftery, F. Balabdaoui, and A. Westveld, 2003: Verifying probabilistic forecasts: Calibration and sharpness. *Proc. Workshop on Ensemble Forecasting,* Val-Morin, QC, Canada. [Available online at http://www.cdc.noaa.gov/people/tom.hamill/ef_workshop_2003.html.]

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **14,** 107–114.

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

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

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

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

——, J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

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

——, R. Mureau, J. D. Opsteegh, and J. Barkmeijer, 2000: A short-range to early-medium-range ensemble prediction system for the European area. *Mon. Wea. Rev.,* **128,** 3501–3519.

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.,* **14,** 382–401. [A corrected version is available online at www.stat.washington.edu/www/research/online/hoeting1999.pdf.]

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecast. *Mon. Wea. Rev.,* **129,** 73–91.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834–1846.

——, and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

——, and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.,* **129,** 123–137.

Kass, R. E., and A. E. Raftery, 1995: Bayes factors. *J. Amer. Stat. Assoc.,* **90,** 773–795.

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

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendan, 1999: Improved weather and seasonal climate forecasts from multimodel superensembles. *Science,* **258,** 1548–1550.

Leamer, E. E., 1978: *Specification Searches.* Wiley, 370 pp.

Leith, C. E., 1974: Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models.* 2d ed. Chapman and Hall, 511 pp.

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

——, and D. Peel, 2000: *Finite Mixture Models.* Wiley, 419 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.

Pellerin, G., L. Lefaivre, P. L. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes Geophys.,* **10,** 463–468.

Raftery, A. E., 1993: Bayesian model selection in structural equation models. *Testing Structural Equation Models,* K. A. Bollen and J. S. Long, Eds., Sage, 163–180.

——, and Y. Zheng, 2003: Long-run performance of Bayesian model averaging. *J. Amer. Stat. Assoc.,* **98,** 931–938.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

——, and ——, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Wea. Forecasting,* **19,** 552–565.

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

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

Stern, R. D., and R. Coe, 1984: A model fitting analysis of daily rainfall data (with discussion). *J. Roy. Stat. Soc.,* **147A,** 1–34.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

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

——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting,* **16,** 463–477.

Van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10-day forecast. *Wea. Forecasting,* **9,** 457–465.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.,* **128,** 2821–2836.

Wu, C. F. J., 1983: On the convergence properties of the EM algorithm. *Ann. Stat.,* **11,** 95–103.