Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging

J. MCLEAN SLOUGHTER, ADRIAN E. RAFTERY, TILMANN GNEITING, AND CHRIS FRALEY

Department of Statistics, University of Washington, Seattle, Washington

(Manuscript received 23 February 2006, in final form 27 November 2006)

ABSTRACT

Bayesian model averaging (BMA) is a statistical way of postprocessing forecast ensembles to create predictive probability density functions (PDFs) for weather quantities. It represents the predictive PDF as a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights are posterior probabilities of the models generating the forecasts and reflect the forecasts' relative contributions to predictive skill over a training period. It was developed initially for quantities whose PDFs can be approximated by normal distributions, such as temperature and sea level pressure. BMA does not apply in its original form to precipitation, because the predictive PDF of precipitation is nonnormal in two major ways: it has a positive probability of being equal to zero, and it is skewed. In this study BMA is extended to probabilistic quantitative precipitation forecasting. The predictive PDF corresponding to one ensemble member is a mixture of a discrete component at zero and a gamma distribution. Unlike methods that predict the probability of exceeding a threshold, BMA gives a full probability distribution for future precipitation. The method was applied to daily 48-h forecasts of 24-h accumulated precipitation in the North American Pacific Northwest in 2003–04 using the University of Washington mesoscale ensemble. It yielded predictive distributions that were calibrated and sharp. It also gave probability of precipitation forecasts that were much better calibrated than those based on consensus voting of the ensemble members. It gave better estimates of the probability of high-precipitation events than logistic regression on the cube root of the ensemble mean.

1. Introduction

A number of existing methods generate probabilistic precipitation forecasts based on deterministic forecasts. Regression techniques such as model output statistics (MOS) can be used to generate probabilities of exceeding thresholds (Glahn and Lowry 1972; Klein and Glahn 1974; Bermowitz 1975; Charba 1998; Antolik 2000), or to generate quantiles of expected precipitation (Bremnes 2004; Friederichs and Hense 2007). Applequist et al. (2002) found that logistic regression can outperform standard regression, and Hamill et al. (2004) found that this can be further refined by using logistic regression on power-transformed forecasts.

These methods, however, do not yield a full predictive probability density function (PDF); rather, they give only probabilities for certain specific events. They also do not make use of all the information available in an ensemble forecast. Ensemble forecasts can give an indication of uncertainty, and a relationship between forecast errors and ensemble spread has been established for several ensemble systems (Buizza et al. 2005). Anderson (1996) suggested using the ensemble member forecasts to partition the real line into a series of bins, assuming each bin to be an equally likely range of possible outcomes, and probabilities uniformly distributed within the inner bins. Hamill and Colucci (1998) noted that this approach is not well calibrated, with far too many observations appearing at the extreme bins. They proposed an alternative method, fitting gamma distributions with parameters based on corrected ensembles or transformations of the ensemble mean. While they reported good results, it is not obvious how to obtain calibrated probability of precipitation (PoP) forecasts using this approach.

Bayesian model averaging (BMA) was introduced by Raftery et al. (2005) as a statistical postprocessing method for producing probabilistic forecasts from ensembles in the form of predictive PDFs. The BMA predictive PDF of any future weather quantity of interest is a weighted average of PDFs centered on the individual

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

DOI: 10.1175/MWR3441.1

bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts and reflect the forecasts' contributions to overall forecasting skill over a training period. The original development of BMA by Raftery et al. (2005) was for weather quantities whose predictive PDFs are approximately normal, such as temperature and sea level pressure.

BMA in the form described by Raftery et al. (2005) does not apply directly to precipitation. This is because the predictive distribution of precipitation is far from normal. It is nonnormal in two major ways: it has a positive probability of being equal to zero, and when it is not zero the predictive density is skewed. Here we extend BMA to precipitation by modeling the predictive distribution for a given ensemble member as a mixture of a point mass at zero and a gamma distribution; the BMA PDF is then itself a mixture of such distributions. In our experiments we show that BMA was calibrated and sharp for the period we considered. This indicates that BMA has the potential to provide both calibrated PoP forecasts, and calibrated and sharp probabilistic quantitative precipitation forecasts (PQPFs).

In section 2 we review the BMA technique and describe our extension of it to precipitation. Then in section 3 we give results for daily 48-h forecasts of 24-h accumulated precipitation over the North American Pacific Northwest in 2003–04 based on the nine-member University of Washington mesoscale ensemble (Grimit and Mass 2002; Eckel and Mass 2005), and associated verifying observations. Throughout the paper we use illustrative examples drawn from these data. Finally, in section 4 we discuss possible improvements to the method.

2. Methods

a. BMA

BMA (Leamer 1978; Kass and Raftery 1995; Hoeting et al. 1999) was originally developed as a way to combine inferences and predictions from multiple statistical models, and was applied to statistical linear regression and related models in social and health sciences. Raftery et al. (2005) extended BMA to ensembles of dynamical models and showed how it can be used as a statistical postprocessing method for forecast ensembles, yielding calibrated and sharp predictive PDFs of future weather quantities.

In BMA for ensemble forecasting, each ensemble member forecast f_k is associated with a conditional PDF $h_k(y|f_k)$, which can be thought of as the PDF of the weather quantity y given f_k, conditional on f_k being the best forecast in the ensemble. The BMA predictive PDF is then

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

where w_k is the posterior probability of forecast k being the best one, and is based on forecast k's relative performance in the training period. The w_k's are probabilities and so they are nonnegative and add up to 1, that is, $\Sigma_{k=1}^{K} w_k = 1$. Here K is the number of ensemble members.

b. Discrete–continuous model

For temperature and sea level pressure, the conditional PDF can be fit reasonably well using a normal distribution centered at a bias-corrected forecast, as shown by Raftery et al. (2005). For precipitation, however, the normal distribution is not appropriate. Figure 1 illustrates the distribution of precipitation accumulation among the verifying observations in our database of ensemble forecasts over the Pacific Northwest in 2003 and 2004, stratified by the accumulation amount predicted by the centroid member (Eckel and Mass 2005) of the forecast ensemble. These histograms show two important aspects of the distribution of precipitation. First, accumulated precipitation was zero in a large number of cases. Second, for the cases in which the accumulated precipitation was not zero, the distributions were highly skewed. The normal distribution does not fit data of this kind, and to extend BMA to precipitation we must develop a model for the conditional PDF $h_k(y|f_k)$ in (1) that takes account of these facts.

Our model for $h_k(y|f_k)$ is in two parts. The first part specifies PoP as a function of the forecast f_k. We follow Hamill et al. (2004) in using logistic regression with a power transformation of the forecast as a predictor variable. Hamill et al. (2004) recommended using the fourth root of the forecast as a predictor variable, but we found that using the cube root was adequate. All else being equal, it seems desirable to use a predictor variable that is as close to the original forecast as possible, and the cube root is closer than the fourth root, and so is preferable if its performance is adequate. We found that this model did not provide the best possible predictions when the forecast was equal to zero, and so we included a second predictor δ_k, equal to 1 if $f_k = 0$ and equal to 0 otherwise. Our logistic regression model then is

$$\begin{aligned} \text{logit}P(y = 0|f_k) &\equiv \log \frac{P(y = 0|f_k)}{P(y > 0|f_k)} \\ &= a_{0k} + a_{1k} f_k^{1/3} + a_{2k}\delta_k. \end{aligned} \tag{2}$$

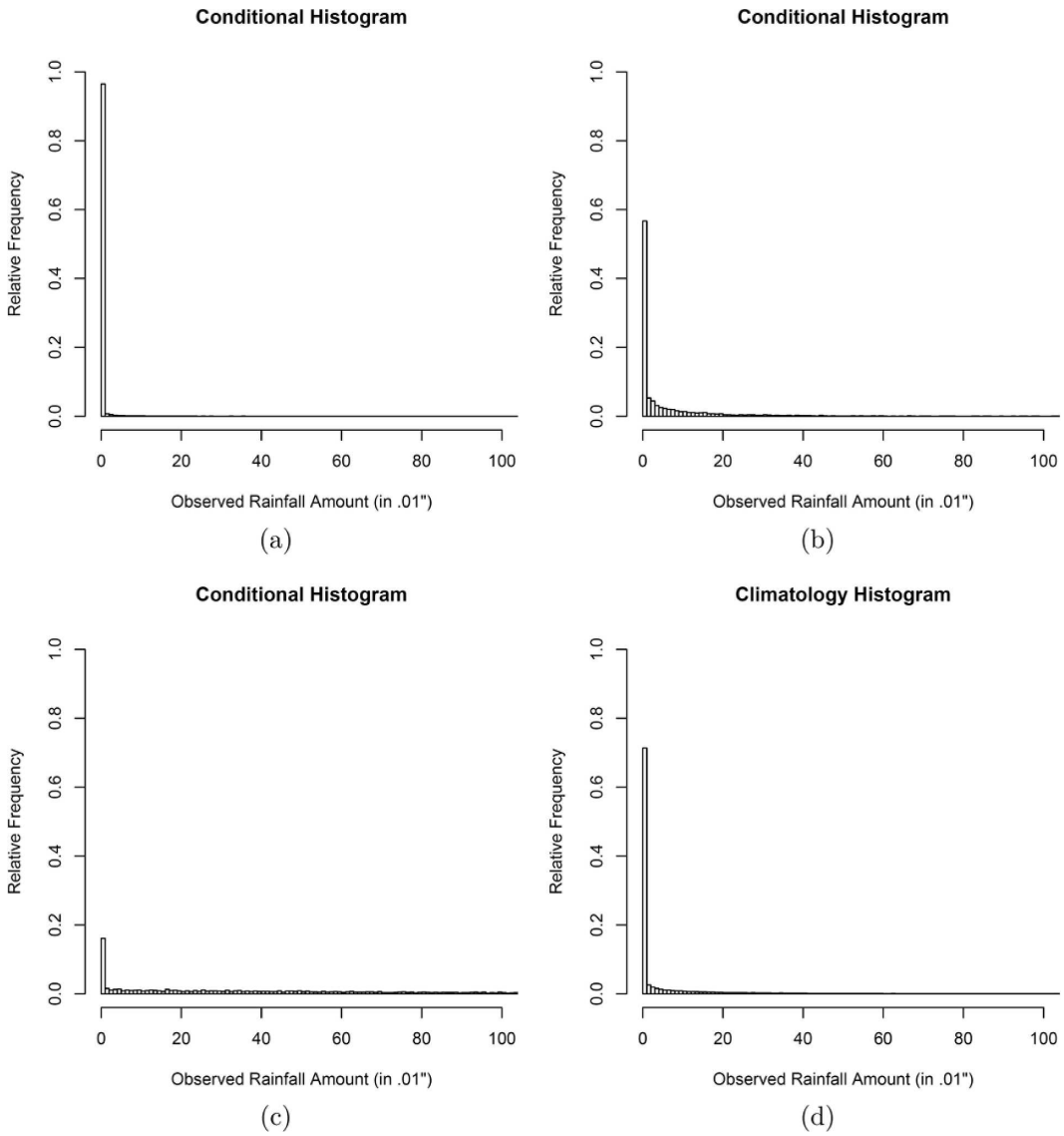

FIG. 1. Histograms of observed precipitation accumulation for cases in which the centroid member forecast of precipitation was (a) zero, (b) between 6.4 and 9.6 hundredths of an inch, and (c) greater than 0.0594 in., and (d) all cases.

The probability $P(y > 0|f_k)$ is the probability of nonzero precipitation given the forecast f_k, if f_k is the best ensemble member forecast for that day.

The second part of our model specifies the PDF of the amount of precipitation given that it is not zero. Previous authors have fit gamma distributions to precipitation amounts (Coe and Stern 1982; Stern and Coe 1984; Wilks 1990; Hamill and Colucci 1998; Wilson et al. 1999) as they can fit skewed data and are quite flexible, and we also took the gamma distribution as our starting point. The gamma distribution with shape parameter α and scale parameter β has the PDF

$$g(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta)$$

for $y > 0$, and $g(y) = 0$ for $y \leq 0$. The mean of this distribution is $\mu = \alpha\beta$, and its variance is $\sigma^2 = \alpha\beta^2$. We found that fitting gamma distributions to the raw observed accumulation amounts did not give an especially good fit. We found the same issues with high values being fit poorly that Hamill and Colucci (1998) reported. In light of this, rather than fitting the gamma distribution to the observed precipitation amounts themselves, we fit the gamma distribution to powers of

the observed values. We found that the best fit was achieved when the gamma distribution was fit to the cube root of the observed precipitation amounts.

It remains to specify how the parameters of the gamma distribution depend on the forecast. We found that the mean of the fitted gamma distribution of the cube root of precipitation was approximately linear as a function of the cube root of the forecasted accumulation. We also found that the variance of the fitted gamma distribution was approximately linear as a function of the forecast.

Putting these components together, we get the following model for the conditional PDF of precipitation accumulation, given that forecast f_k is best:

$$h_k(y|f_k) = P(y = 0|f_k)I[y = 0] + P(y > 0|f_k)g_k(y|f_k)I[y > 0],$$

where y is the cube root of the precipitation accumulation, the general indicator function $I[]$ is unity if the condition in brackets holds and zero otherwise, and $P(y = 0|f_k)$ is specified by (2). The conditional PDF $g_k(y|f_k)$ of the cube root precipitation amount y given that it is positive is a gamma distribution with PDF

$$g_k(y|f_k) = \frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} y^{\alpha_k - 1} \exp(-y/\beta_k).$$

The parameters of the gamma distribution depend on the original forecast f_k through the relationships

$$\mu_k = b_{0k} + b_{1k} f_k^{1/3}$$

and

$$\sigma_k^2 = c_{0k} + c_{1k} f_k, \quad (3)$$

where $\mu_k = \alpha_k\beta_k$ is the mean of the distribution, and $\sigma_k^2 = \alpha_k\beta_k^2$ is its variance.

c. BMA for discrete–continuous models

For the variances, we observed that the parameters c_{0k} and c_{1k} in (3) did not vary much from one model to another, and so we restricted the variance parameters to be constant across all ensemble members. This simplifies the model by reducing the number of parameters to be estimated, makes parameter estimation computationally easier, and reduces the risk of overfitting. It is analogous to the assumption of equal variances in Raftery et al. (2005).

Our final BMA model (1) for the predictive PDF of the weather quantity, y—here the cube root of precipitation accumulation—is thus

$$p(y|f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k[P(y = 0|f_k)I[y = 0] + P(y > 0|f_k)g_k(y|f_k)I[y > 0]], \quad (4)$$

where w_k is the posterior probability of ensemble member k being best, f_k is the original forecast from this member,

$$\text{logit}P(y = 0|f_k) = a_{0k} + a_{1k} f_k^{1/3} + a_{2k}\delta_k,$$

where δ_k is equal to 1 if $f_k = 0$ and equal to 0 otherwise, and

$$g_k(y|f_k) = \frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} y^{\alpha_k - 1} \exp(-y/\beta_k).$$

The parameters $\alpha_k = \mu_k^2/\sigma_k^2$ and $\beta_k = \sigma_k^2/\mu_k$ of the gamma distribution depend on f_k through the relationships

$$\mu_k = b_{0k} + b_{1k} f_k^{1/3}$$

and

$$\sigma_k^2 = c_0 + c_1 f_k,$$

which specify the mean and the variance of the distribution, respectively. While (4) is stated in terms of the cube root of the precipitation amount, it is easy to express the resulting probability statements in terms of the original amounts.

d. Parameter estimation

Parameter estimation is based on data from a training period, which we take here to be the N days of forecast and verifying observation data preceding initialization, following Raftery et al. (2005). The training period is a sliding window, and the parameters are reestimated for each new initialization period. The required data consist of forecast–observation pairs from a collection of observation sites for each of the ensemble members.

To assess the length of the training period, we computed the average continuous ranked probability score (CRPS) for the probabilistic forecasts and the mean absolute errors (MAE) of the resulting deterministic forecasts, for each of a set of possible training period lengths, $N = 15, 20, \ldots, 50$. The results are shown in Fig. 2. It is clear that making the training period longer leads to improved forecasts up to 30 days, but that increasing it beyond that does not yield any further improvement. As a result, we used a training period of length $N = 30$ days.

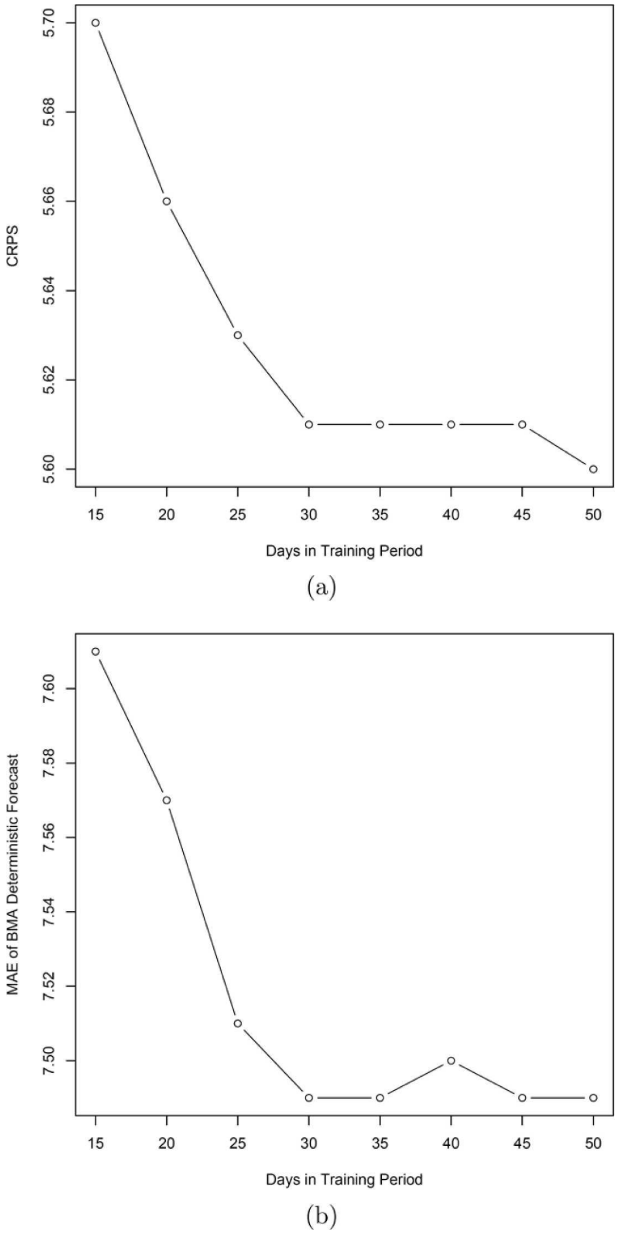

FIG. 2. Comparison of training period lengths: (a) CRPS of BMA forecasts and (b) MAE of BMA deterministic forecasts.

The parameters a_{0k}, a_{1k}, and a_{2k} are member specific, and they are determined separately for each ensemble member, using the forecasts from that ensemble member only and the associated verifying observations. They are estimated by logistic regression with precipitation/no precipitation as the dependent variable, and $f_k^{1/3}$ and δ_k as the two predictor variables.

The parameters b_{0k} and b_{1k} are also member specific, and they are determined by linear regression with the nonzero precipitation observations as cases, the cube root of the amount of precipitation as the dependent variable, and the cube root of the forecasted accumulation amount as the predictor variable.

We estimate w_k, $k = 1, \ldots, K$; c_0; and c_1 by the maximum likelihood technique (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.

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 the BMA model (4) is

$$\ell(w_1, \ldots, w_K; c_0; c_1) = \sum_{s,t} \log p(y_{st} | f_{1st}, \ldots, f_{Kst}), \tag{5}$$

where the summation is over values of s and t that index observations in the training set by space and time, and $p(y_{st}|f_{1st}, \ldots, f_{Kst})$ is given by (4), with subscripts s and t added to y and f_k. This cannot be maximized analytically, and instead we maximize it numerically 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 (4) 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 expectation (E) step, and the maximization (M) step. It starts with an initial guess for the parameters. 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, the parameters are reestimated given the current values of the z_{kst}.

TABLE 1. Raw ensemble, logistic regression PoP, and BMA forecasts for two example stations. The quantitative precipitation forecasts and observations are given in hundredths of an inch. Descriptions of the University of Washington ensemble can be found in Eckel and Mass (2005). CENT is the ensemble centroid, AVN is the Global Forecast System from the National Centers for Environmental Prediction (NCEP), CMCG is the Global Environmental Multi-scale from the Canadian Meteorological Centre, ETA is the limited-area mesoscale model from NCEP, GASP is the Global Analysis and Prediction Model from the Australian Bureau of Meteorology, JMA is the Global Spectral Model from the Japan Meteorological Agency, NGPS is the Navy Operational Global Atmospheric Prediction System from the Fleet Numerical Meteorological and Oceanographic Center, TCWB is the Global Forecast System from the Taiwan Central Weather Bureau, and UKMO is the Unified Model from the Met Office.

Ensemble member	CENT	AVN	CMCG	ETA	GASP	JMA	NGPS	TCWB	UKMO
Station KCLM on 19 May 2003									
BMA weight	0.00	0.39	0.00	0.30	0.19	0.13	0.00	0.00	0.00
Member PoP	0.19	0.16	0.21	0.18	0.17	0.21	0.22	0.23	0.19
BMA PoP	0.17								
Member forecast	0	0	0	0	0	0	0	0	0
BMA forecast	0								
BMA upper bound	3								
Observation	2								
Station KPWT on 26 Jan 2003									
BMA weight	0.00	0.30	0.23	0.23	0.00	0.00	0.03	0.16	0.05
Member PoP	0.46	0.64	0.74	0.59	0.44	0.44	0.70	0.46	0.72
BMA PoP	0.63								
Member forecast	2	8	10	4	1	1	11	1	12
BMA forecast	3								
BMA upper bound	32								
Observation	26								

For the BMA model (4), the E step is

$$\hat{z}_{kst}^{(j+1)} = \frac{w_k^{(j)} p^{(j)}(y_{st}|f_{kst})}{\sum_{l=1}^{K} w_l^{(j)} p^{(j)}(y_{st}|f_{lst})},$$

where the superscript j refers to the jth iteration of the EM algorithm, and thus $w_k^{(j)}$ refers to the estimate of w_k at the jth iteration, and $p^{(j)}(y_{st}|f_{kst})$ is $p(y_{st}|f_{kst})$ as defined in (4), using the estimates of c_0 and c_1 from the jth iteration. The M step then consists of estimating the w_k, c_0, and c_1 using as weights the current estimates of z_{kst}, namely, $\hat{z}_{kst}^{(j+1)}$. Thus

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

where n is the number of cases in the training set, that is, the number of distinct values of (s, t). There are no analytic solutions for the maximum likelihood estimates of the parameters c_0 and c_1, and so they must be estimated numerically by optimizing (5) using the current estimates of the w_k parameters.

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. Choosing the starting value for day $t + 1$ to be equal to the converged estimate for day t usually leads to a good solution.

e. Examples

To illustrate how the method works, we show two examples. Our first example is on 19 May 2003, at station KCLM in Port Angeles, Washington. Table 1 shows the raw ensemble forecasts, the logistic regression PoP results, the BMA results, and the verifying observation. The probability of exceeding a given amount is given in Fig. 3 by the proportion of the area under the upper curve (BMA PDF) to the right of it, multiplied by 1 − PoP, that is, by the height of the thick vertical line at zero.

All nine ensemble members predicted no rain, but it actually did rain. The BMA predictive PDF is shown in Fig. 3a; the observation was below the BMA 90th percentile upper bound, which is shown as a dashed vertical line, and so was within the BMA 90% prediction interval.

Our second example is on 26 January 2003, at station KPWT in Bremerton, Washington. Again, Table 1 shows the raw ensemble forecasts, logistic regression

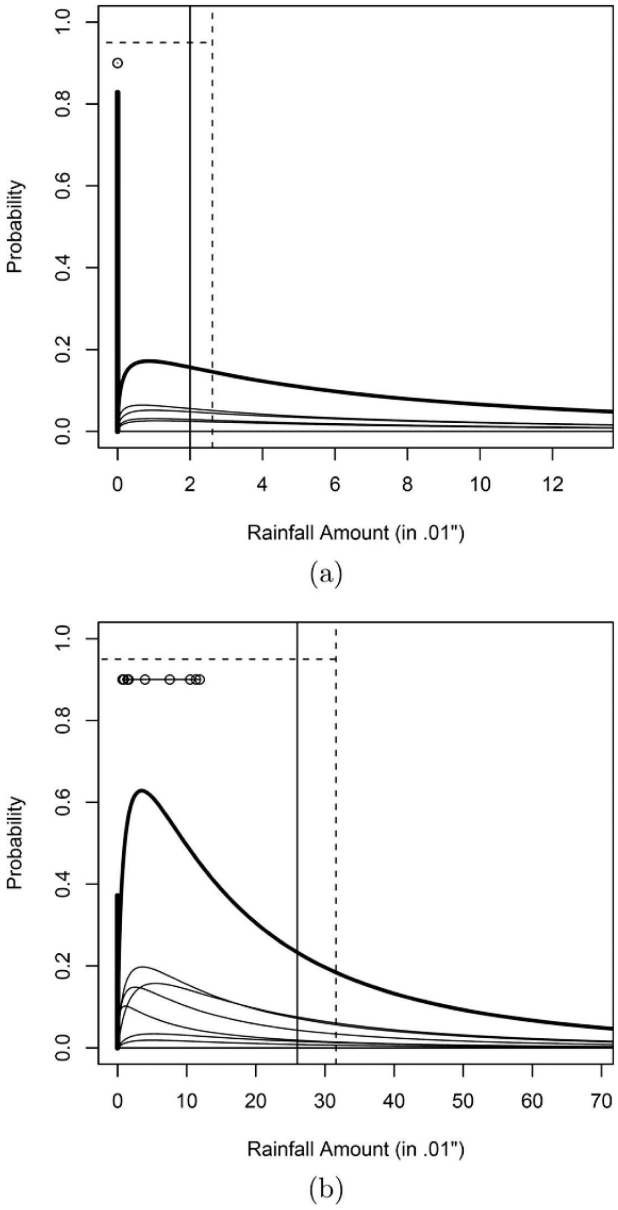

FIG. 3. BMA-fitted PDFs for (a) station KCLM on 19 May 2003 and (b) station KPWT on 26 Jan 2003. The thick vertical line at zero represents the BMA estimate of the probability of no precipitation, and the upper solid curve is the BMA PDF of the precipitation amount given that it is nonzero. The lower curves are the components of the BMA PDF, namely, the weighted contributions from the ensemble members. The dashed vertical line represents the 90th percentile upper bound of the BMA PDF; the dashed horizontal line is the respective prediction interval; the dots represent the ensemble member forecasts; and the solid vertical line represents the verifying observation.

PoP results, BMA results, and the observed value. The BMA deterministic forecast, that is, the median of the BMA predictive PDF, was about 0.03 in. The BMA predictive PDF itself is shown in Fig. 3b. The observation is far outside the ensemble range, but it is contained within the BMA upper bound.

Spatial displays of the BMA PoP forecast, and of the BMA deterministic forecast and 90th percentile upper bound for the precipitation amount, are shown in Figs. 4 and 5 for these two dates. Spatial displays of the PoP seem potentially useful in communicating probability forecasts to the general public, and might assist in the use and interpretation of the forecasts (Gigerenzer et al. 2005). Figure 6 shows a graphical comparison between the probabilistic precipitation forecasts and the verifying observations for all stations on 26 January 2003.

3. Results

BMA was applied to 48-h forecasts of 24-h precipitation accumulation in the Pacific Northwest for the 0000 UTC cycle over the 2-yr period of 1 January 2003 through 31 December 2004, using the nine-member University of Washington mesoscale ensemble (Eckel and Mass 2005). Data were available for 560 days, and data for 171 days during these 2 yr were unavailable. In all, 109 996 station observations were used, an average of about 196 per day. The forecasts were produced for observation locations by bilinear interpolation from the forecast grids. The observations were subject to the quality control procedures described by Baars (2005).

We begin with a discussion of the PoP forecasts. Figure 7 shows the reliability diagram (Wilks 2006, section 7.4.4). As can be seen, BMA produced well-calibrated results, while a consensus vote from the raw ensemble produced severely uncalibrated results. Table 2 shows that the Brier score (Wilks 2006, p. 284) for the BMA PoP forecasts was better than that for either the raw ensemble or logistic regression based on the cube root of the ensemble mean.

In assessing probabilistic forecasts of quantitative precipitation, we follow Gneiting et al. (2005) and aim to maximize the sharpness of the predictive PDFs, subject to calibration. Calibration refers to the statistical consistency between the forecast PDFs and the observations, and in the context of precipitation forecasts it was discussed by Krzysztofowicz and Sigrest (1999). To assess calibration, we consider Fig. 8, which shows the verification rank histogram for the ensemble forecasts and the probability integral transform (PIT) histogram for the BMA forecast distributions. The verification rank histogram illustrates the lack of calibration in the raw ensemble, similar to results reported by Hamill and Colucci (1998), Eckel and Walters (1998), and Mullen and Buizza (2001) for other ensembles. The PIT histogram is a continuous analog of the verification rank

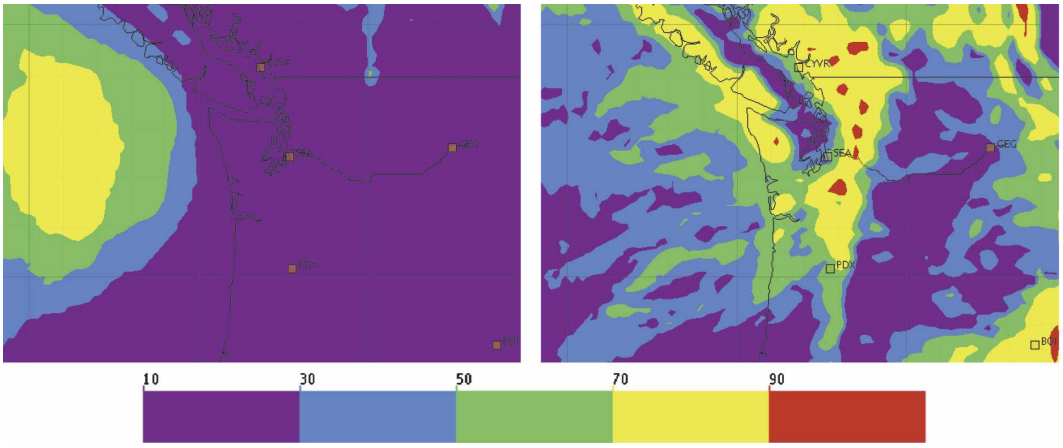

FIG. 4. BMA PoP forecast for (left) 19 May 2003 and (right) 26 Jan 2003.

histogram (Gneiting et al. 2005), and it shows that the BMA forecast distributions were considerably better calibrated than the raw ensemble.

For the verification rank histogram, there were incidences where the observed value was zero (no precipitation), and one or more forecasts were also zero. To obtain a rank in these situations, a ranking was randomly chosen between zero and the number of forecasts that equaled zero. To calculate the values for the PIT histogram, each BMA cumulative distribution

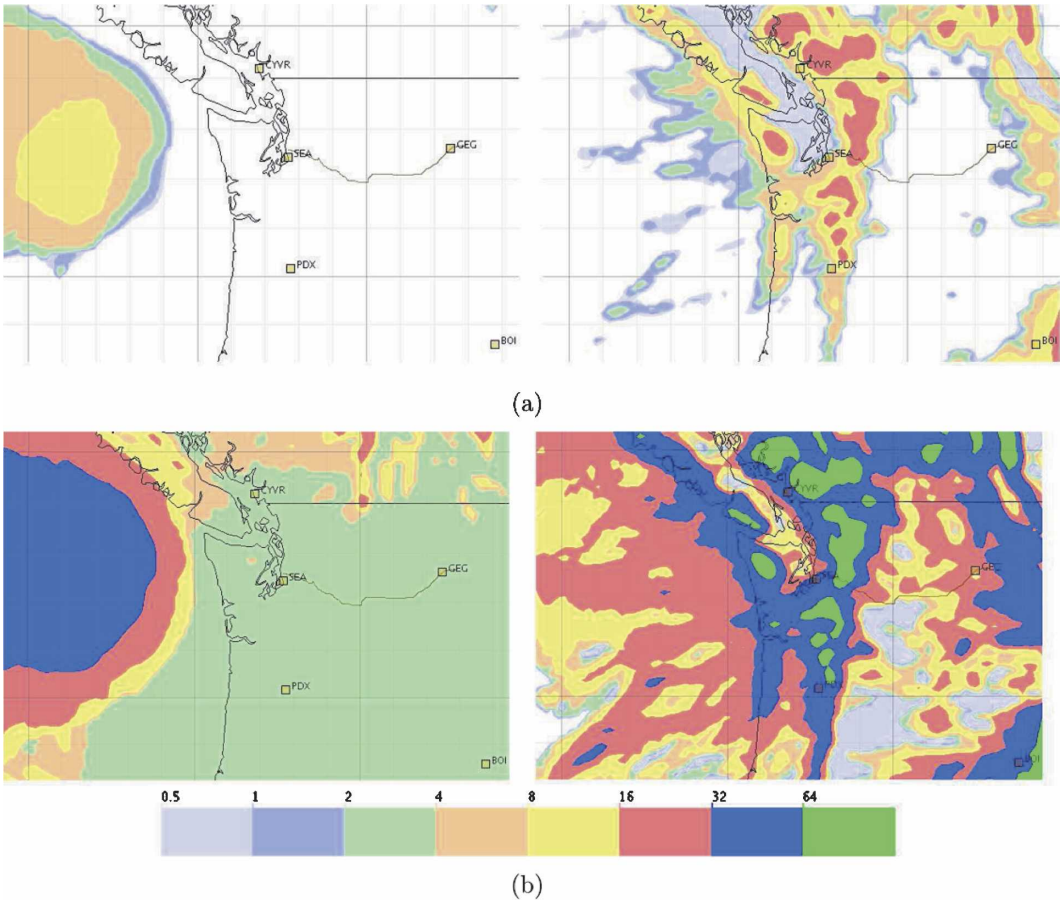

FIG. 5. (a) BMA deterministic forecast and (b) BMA 90th percentile upper bound forecast for (left) 19 May 2003 and (right) 26 Jan 2003, in hundredths of an inch.

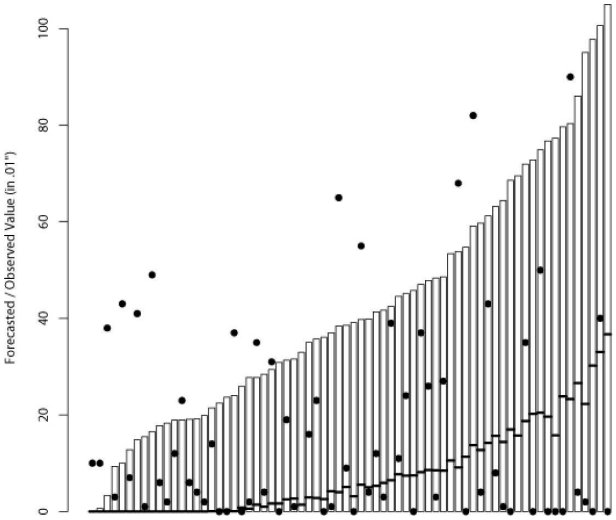

FIG. 6. BMA forecasts and upper bounds (bars) against observed values (dots) for all locations on 26 Jan 2003.

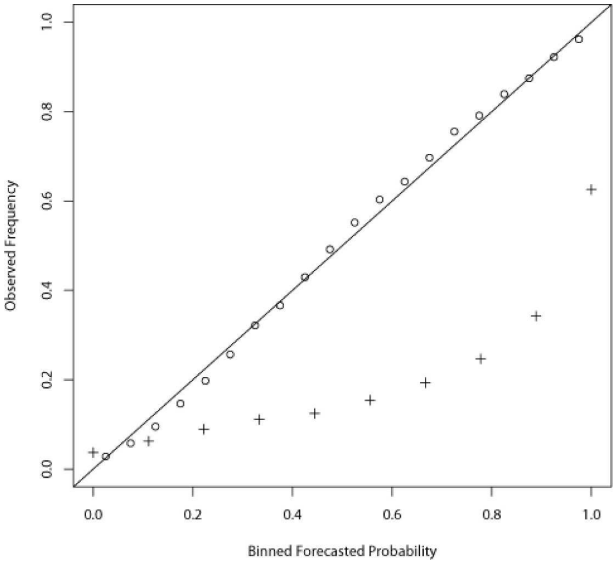

FIG. 7. Reliability diagram of binned PoP forecast vs observed relative frequency of precipitation, for consensus voting of the raw ensemble (crosses) and BMA (circles).

function was evaluated at its corresponding observation. In the case of an observation of zero, a value was randomly drawn between zero and the probability of no precipitation.

Table 3 shows the empirical coverage of lower 50% and 90% prediction intervals, and the results echo what we see in the histograms. Sample climatology was perfectly calibrated, as expected, while the raw ensemble was substantially uncalibrated. The BMA intervals were close to being calibrated. The table also shows the average width of the prediction intervals, which characterizes the sharpness of the forecast distributions. The BMA PDFs produced narrower intervals than the raw ensemble forecasts for both intervals considered, and narrower intervals than climatology for 90% intervals.

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 CRPS, which generalizes the MAE, and can be directly compared to the latter (Gneiting et al. 2005; Wilks 2006, section 7.5.1). Table 2 shows MAE and CRPS values for sample climatology, raw ensemble forecasts, and BMA forecasts, all in units of hundredths of an inch. A deterministic forecast can be created from the BMA

TABLE 2. MAE, CRPS, and Brier skill score (BSS) relative to sample climatology for probabilistic precipitation forecasts. The thresholds, MAE, and CRPS values are given in hundredths of an inch, and the MAE refers to the deterministic forecast given by the median of the respective forecast distribution.

Score	Threshold	Sample climatology	Ensemble forecast	BMA forecast	Logistic regression
MAE		8.7	9.4	7.5	
CRPS		7.8	7.6	5.6	
BSS	0		−0.18	0.38	0.37
BSS	5		0.00	0.36	0.37
BSS	10		−0.02	0.34	0.35
BSS	25		−0.02	0.31	0.33
BSS	50		−0.02	0.26	0.30
BSS	100		0.05	0.21	0.25
BSS	150		0.05	0.17	0.19
BSS	200		0.11	0.14	0.12
BSS	250		0.10	0.11	0.03
BSS	300		0.09	0.09	0.05
BSS	350		0.10	0.08	0.00
BSS	400		0.05	0.07	−0.02

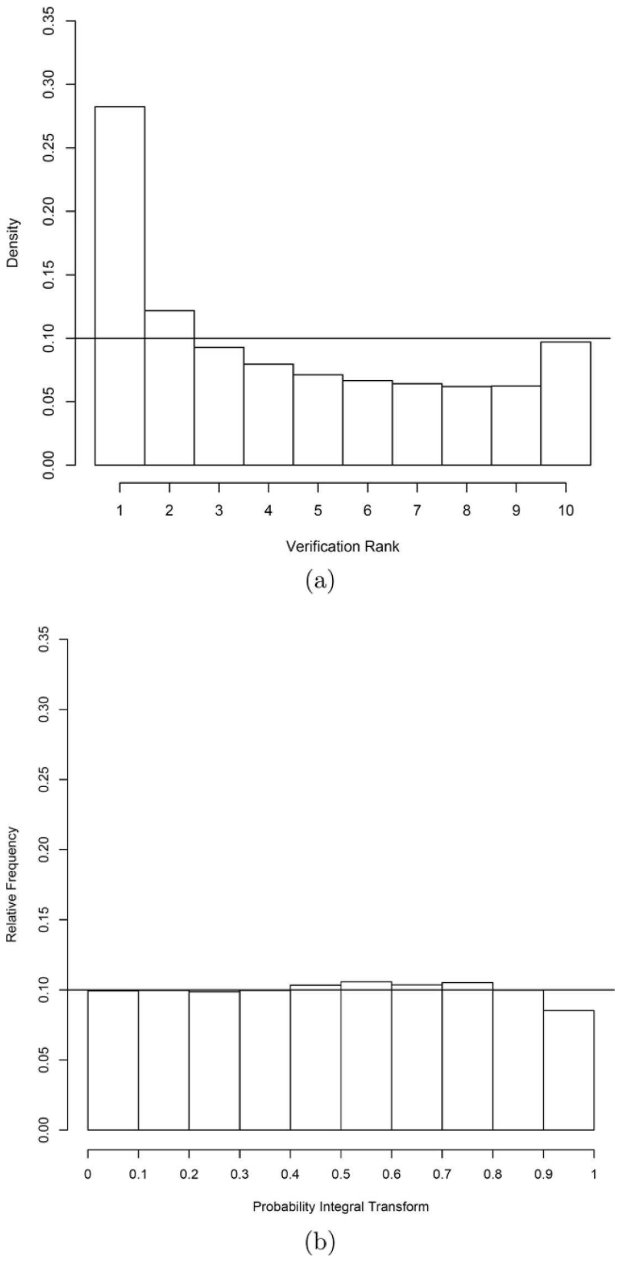

FIG. 8. (a) Verification rank histogram for raw ensemble forecasts and (b) PIT histogram for BMA forecast distributions of precipitation accumulation.

forecast by finding the median of the predictive PDF, and the MAE refers to this forecast. Similarly, we show the MAE for the median of the sample climatology and the median of the nine-member forecast ensemble, with the results for BMA being by far the best. We also computed MAE values for deterministic forecasts based on the respective means; these were much higher than the MAE values for the median forecasts, as is generally true when the predictive PDFs are highly skewed. The results for the CRPS were similar, in that the BMA forecast substantially outperformed the others.

TABLE 3. Coverage and average width of lower 50% and 90% prediction intervals for precipitation accumulation, in percentages and hundredths of an inch, respectively.

	Coverage		Width	
Interval	50%	90%	50%	90%
Sample climatology	50.0	90.0	0.0	24.0
Ensemble forecast	68.5	92.9	11.8	24.2
BMA forecast	50.7	91.1	3.2	22.8

Table 2 shows Brier skill scores (Wilks 2006, p. 285) relative to sample climatology at various thresholds. These are shown for three probabilistic forecasting methods: ensemble consensus voting (taking PoP to be equal to the proportion of ensemble members that predict precipitation), BMA, and logistic regression based on the cube root of the ensemble mean.

The ensemble forecast had poor skill at lower thresholds, but better skill at higher thresholds. Logistic regression had good skill at lower thresholds, but not much skill at higher thresholds (2.5 in. or above). BMA had good skill at both higher and lower thresholds. Its skill at higher thresholds indicates that BMA may be useful for identifying a risk of extreme precipitation events.

4. Discussion

We have shown how to apply BMA to precipitation forecasts. This provides a statistical postprocessing method for ensembles that yields a full predictive distribution for quantitative precipitation. The predictive distribution has two components: the probability of zero precipitation, and the PDF for the precipitation accumulation given that it is greater than zero. It thus provides both PoP and PQPF in a unified form. In our experiments with the University of Washington ensemble, the BMA forecast PDFs were better calibrated and sharper than the raw ensemble, which was uncalibrated. The BMA median forecast had lower MAE than the ensemble median, and the BMA forecast PDFs had substantially lower CRPS than the raw ensemble.

BMA probabilistic forecasts of precipitation above a given threshold had good Brier skill scores across the full range of thresholds of interest. In comparison, the ensemble forecasts had very poor skill at lower thresholds, and power-transformed logistic regression based on the ensemble mean, as suggested by Hamill et al. (2004), had less skill at higher thresholds.

Our implementation has been for the situation where the ensemble members come from clearly distinguishable sources. In other cases, such as the current synoptic National Centers for Environmental Prediction and European Centre for Medium-Range Weather Forecasts 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 (1) would be constrained to be equal. The EM algorithm can still be used, with a small modification, as pointed out by Raftery et al. (2005, p. 1170).

BMA offers the added advantage, by giving a full predictive PDF, of being able to give probabilities of exceeding arbitrary precipitation amounts, rather than having to create a new logistic regression model for each threshold of interest. This may explain why it performs well for forecasts at high thresholds where the amount of training data is small. This suggests that BMA may be useful both for routine precipitation forecasting and for forecasting the risk of extreme precipitation events.

Various improvements to the method may be possible. The BMA parameters were estimated using data on observations from the entire Pacific Northwest, and a more local approach; for example, partitioning the region into climatologically homogeneous subregions, or fitting BMA locally for each location using only observations within a given radius, might perform better. This latter possibility was suggested by E. Grimit and C. Mass (2004, personal communication). Our method of estimation assumes independence of forecast errors in space and time. This is unlikely to hold, but it is also unlikely to have much effect on the results, because we are focusing here on the predictive distribution of a single scalar quantity. A calibrated probabilistic forecasting method for temperature and sea level pressure that does take account of spatial dependence was proposed by Gel et al. (2004), and it would be interesting to extend this to precipitation. Herr and Krzysztofowicz (2005) proposed a generic bivariate probability model for rainfall accumulation in space and gave a critique of the simulation technique of Seo et al. (2000), which generates multiple realizations of downscaled precipitation fields from PQPF.

Hamill et al. (2004) recommended the use of reforecasts from past years computed on the same basis as the current ensemble forecasts. If such reforecasts were available, it seems possible that expanding the training period to include days from the same season in previous years could improve performance. The University of Washington ensemble is frequently updated, however, and as a result prior forecasts were not available to us.

Our experiments were carried out for 24-h precipitation accumulation, and other experiments not reported here suggest that the method also performs well for other accumulation durations, such as 3, 6, or 12 h. This would have to be verified for the particular forecasting task at hand.

We used a moving window training dataset of 30 days. Our choice of training period is specific to our dataset and region, the Pacific Northwest, where it rains relatively frequently. The best training period could be different for other regions, and this could be assessed in a similar way. Nevertheless, our training period of 30 days is similar to training periods for temperature reported elsewhere. Good results for temperature have been obtained with a 25-day training period in the Pacific Northwest (Raftery et al. 2005), and with a 40-day training period for stations across Canada (Wilson et al. 2007). Although this should be assessed empirically for the region to which BMA is to be applied, it does seem reasonable to expect that a 30-day training period will often give good results.

Probabilistic forecasts using BMA based on the University of Washington mesoscale ensemble prediction system are currently being produced in real time for temperature and precipitation. They are available online (at http://bma.apl.washington.edu and http://probcast.washington.edu). These Web sites provide median forecasts, upper bound forecasts, and forecasts of exceeding thresholds for precipitation accumulation. We apply the BMA technique directly on the model grid, and the Web site provides the ability to look at the BMA PDF for any grid cell in the forecast domain.

Acknowledgments. The authors are grateful to Clifford Mass, Mark Albright, Jeff Baars, and Eric Grimit 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 acknowledge helpful comments by the editor and two reviewers. 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.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.,* **239,** 306–337.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **17,** 783–799.

Baars, J., cited 2005: Observations QC summary page. [Available online at http://www.atmos.washington.edu/~qcreport/.]

Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. *Mon. Wea. Rev.,* **103,** 149–153.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.,* **132,** 338–347.

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.

Charba, J. P., 1998: The LAMP QPF products. Part I: Model development. *Wea. Forecasting,* **13,** 934–962.

Coe, R., and R. D. Stern, 1982: Fitting models to daily rainfall data. *J. Appl. Meteor.,* **21,** 1024–1031.

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. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

——, and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **A222,** 309–368.

Friederichs, P., and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.,* **135,** 2365–2378.

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.

Gigerenzer, G., R. Hertwig, E. van den Broeck, B. Fasolo, and K. V. Katsikoupolos, 2005: "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Anal.,* **25,** 623–629.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

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

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

Herr, H. D., and R. Krzysztofowicz, 2005: Generic probability distribution of rainfall in space: The bivariate model. *J. Hydrol.,* **306,** 234–263.

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.

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

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.,* **55,** 1217–1227.

Krzysztofowicz, R., and A. A. Sigrest, 1999: Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting,* **14,** 427–442.

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

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **129,** 638–663.

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.

Seo, D. J., S. Perica, E. Welles, and J. C. Schaake, 2000: Simulation of precipitation fields from probabilistic quantitative precipitation forecast. *J. Hydrol.,* **239,** 203–229.

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

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

——, 2006: *Statistical Methods in the Atmospheric Sciences.* 2d ed. Academic Press, 627 pp.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verifying weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.,* **127,** 956–970.

——, S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

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