Using Bayesian Model Averaging to Calibrate Forecast Ensembles *

Adrian E. Raftery, Fadoua Balabdaoui, Tilmann Gneiting and Michael Polakowski
Department of Statistics, University of Washington, Seattle, Washington

Draft 2.0 (Adrian, Cliff, Eric): November 4, 2003

Abstract

Ensembles used for probabilistic weather forecasting often exhibit a spread-skill relationship, but they tend to be underdispersive. We propose a principled 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 PDF of any quantity of interest is a weighted average of PDFs centered around the individual (possibly bias-corrected) forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts, and reflect the models' skill over the training period. The BMA PDF can be represented as an unweighted ensemble of any desired size, by simulating from the BMA predictive distribution. 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 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-sill relationship but yet to be underdispersive.

The method was applied to short-range mesoscale forecasts of sea-level pressure in the Pacific Northwest in January–June, 2000 using the University of Washington MM5 ensemble. The predictive PDFs were much better calibrated than the raw ensemble, the BMA forecasts were sharp in that 90% BMA prediction intervals were 62% shorter on average than those produced by a crude form of climatology. As a byproduct, BMA yields a deterministic point forecast, and this had RMSE 11% lower than any of the ensemble members, and 6% lower than the ensemble mean.

*Corresponding author address: Adrian E. Raftery, Department of Statistics, University of Washington, Box 354320, Seattle, WA 98195-4320.

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 (Epstein 1969; Leith 1974). Ensembles based on global models have been found useful for medium-range probabilistic forecasting (Toth and Kalnay 1993; Molteni, Buizza, Palmer, and Petroliagis 1996; Houtekamer and Derome 1995; Hamill, Snyder, and Morss 2000). Typically the ensemble mean outperforms all or most of the individual ensemble members, and in some studies a spread-skill relationship 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.

Here we focus on short-range mesoscale forecasting. Several authors have studied the use of a synoptic ensemble, the 15-member NCEP Eta-RSM ensemble, for short-range forecasting (Hamill and Colucci 1997; Hamill and Colucci 1998; Stensrud, Brooks, Du, Tracton, and Rogers 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 there was no spread-skill relationship. The first short-range mesoscale ensemble forecasting experiment was SAMEX (Hou, Kalnay, and Droegemeier 2001). This found that the ensemble mean was more skillful than the individual forecasts, and that there was a significant spread-skill relationship, with a correlation on the order of 0.4. However, the ensemble was not well calibrated; see Figures 8 and 9 of Hou, Kalnay, and Droegemeier (2001).

Grimit and Mass (2002) described the University of Washington mesoscale SREF 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 MM5 model, in which initial conditions are taken from different operational centers. The UW ensemble was run at 36km and 12km grid spacing, while the NCEP SREF has been run at 48km. 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-skill correlation than other studies, ranging up to 0.6. Figure 1 is a scatterplot showing the spread-skill relationship for sea-level pressure for the UW ensemble for the same period as that on which Grimit and Mass (2002)'s report was based, namely January–June, 2000. The spread-skill correlation for daily average absolute errors, averaging spatially across the Pacific Northwest, was 0.42, which was highly statistically significant. However, the verification rank histogram for the same data, shown in Figure 2, shows the ensemble to be underdispersive and hence uncali-

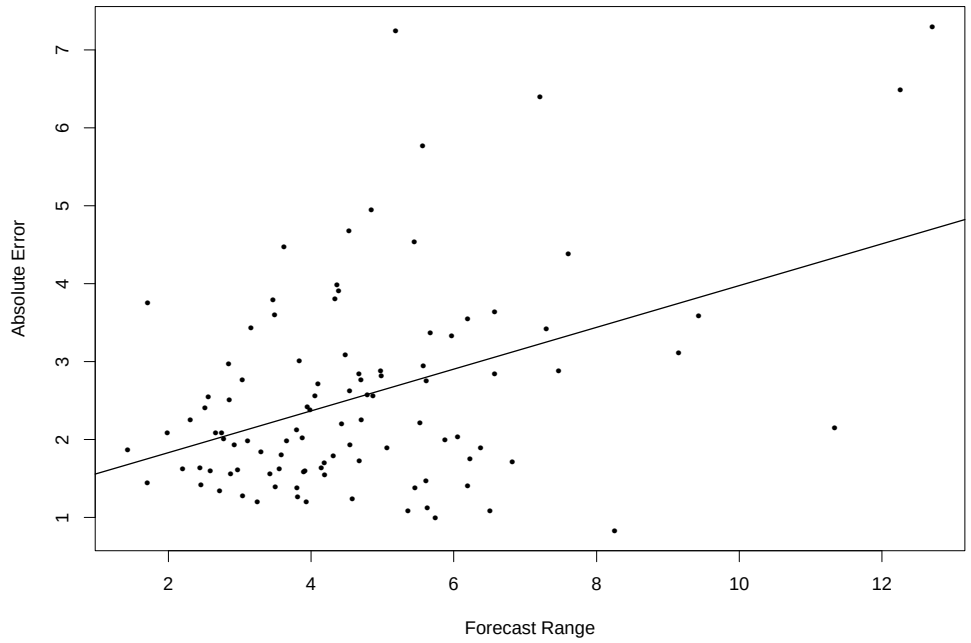

Figure 1: Spread-Skill Relationship for Daily Average Absolute Errors in the 48-hour Forecast of Sea-Level Pressure in the UW Ensemble, January-June, 2000. The vertical axis shows the daily average of the absolute errors of the ensemble mean forecast, and the horizontal axis shows the daily average of the difference between the highest and lowest forecasts in the ensemble. The solid line is the least-squares regression line. The correlation is 0.42, and is statistically significant at the 0.000001 level.

brated. In this case, the ensemble range based on five members would contain 4/6, or 67% of the observed values if the ensemble were calibrated, i.e. if the ensemble forecasts were a sample from the predictive PDF, whereas in fact it contained only 51% of them.

This behavior — an ensemble that yields a significant spread-skill relationship and useful predictions of forecast skill, and yet is uncalibrated — is not unique to the UW ensemble, as we have noted, and may seem contradictory. 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, model physics, and integration methods. Most ensembles capture only some of these uncertainties, and then probably only partially. In addition, even if these uncertainties are small, Lorenz (1963) has shown that forecast uncertainties generally become large at longer forecast lags because of the nonlinear

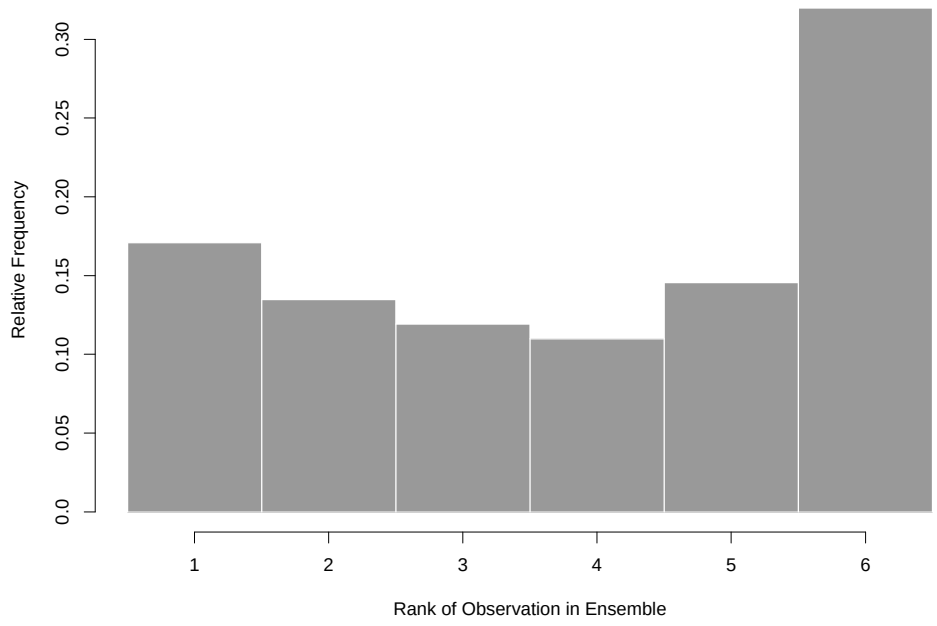

Figure 2: Verification Rank Histogram for the UW Ensemble 48-Forecasts of Sea-Level Pressure, January-June, 2000.

dynamics of the atmosphere. Thus it seems inevitable that ensembles based purely on perturbing initial and lateral boundary conditions, model physics and integration models will be underdispersive to some extent. Because they do capture some of the important sources of uncertainty, however, it is reasonable to expect a significant spread-skill relationship, even when the ensemble is uncalibrated. To obtain a calibrated forecast PDF, therefore, it seems necessary to carry out some form of statistical post-processing.

Our goal in this article is to propose an approach to 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 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, Raftery, Balabdaoui, and Westveld 2003).

To achieve this, we propose a principled statistical approach to postprocessing ensemble forecasts, based on Bayesian Model Averaging (BMA). This is a standard statistical 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. 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 the forecasts (possibly bias-corrected) 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 significant spread-skill relationship coexists with a lack of calibration.

In section 2 we describe BMA, show how the BMA model can be estimated using the EM algorithm, and give an example of BMA in action. In section 3 we give BMA results for the UW ensemble, and in section 4 we make some concluding remarks.

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, Madigan, Raftery, and Volinsky 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 $\sum_{k=1}^{K} p(M_k|y^T) = 1$, and they can thus be viewed as weights. The BMA PDF is a weighted average of the PDFs given 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 there is a "best" model, but we do not know what it is, and our uncertainty about the best model is quantified by BMA. Once again, we denote by y the quantity to be forecast. Each deterministic forecast, f_k, can be bias-corrected, yielding

a bias-corrected forecast $\tilde{f}_k$. The forecast f_k is then associated with a conditional PDF, $g_k(y|\tilde{f}_k)$, which can be interpreted as the conditional PDF of y conditional on $\tilde{f}_k$, *given that* f_k *is the best forecast in the ensemble*. The BMA predictive model is then

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

where w_k is the posterior probability of forecast k being the best one, and is based on forecast k's skill in the training period. The w_k's are probabilities and so they add up to 1, i.e. $\sum_{k=1}^{K} w_k = 1$. We will 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 $\tilde{f}_k$, so that $g_k(y|\tilde{f}_k)$ is a normal PDF with mean $\tilde{f}_k$ and an ensemble-member-specific standard deviation, σ_k. We denote this situation by

$$y|\tilde{f}_k \sim N(\tilde{f}_k, \sigma_k^2), \tag{3}$$

and we will describe how to estimate σ_k^2 in the next subsection. In that case, the BMA predictive mean is just the conditional expectation of y given the forecasts, namely

$$E[y|f_1, \ldots, f_K] = \sum_{k=1}^{K} w_k \tilde{f}_k. \tag{4}$$

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

b. Estimation by Maximum Likelihood and the EM Algorithm

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, w_k and σ_k^2, $k = 1, \ldots, K$, on the basis of training observational data. We denote the set of BMA model parameters to be estimated by θ. 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 horizon to be fixed; in practice we will estimate different models for each forecast horizon.

We estimate θ by maximum likelihood (Fisher 1922) from the training data. The likelihood function is defined as the probability of the training data given θ, viewed as a function

of θ. The maximum likelihood estimator is the value of θ that maximizes the likelihood function, i.e. the value of the parameter 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. The log-likelihood function for model (2) is

$$\ell(\theta) = \sum_{s,t} \log \left(\sum_{k=1}^{K} w_k g_k(y_{st} | \tilde{f}_{kst}) \right), \tag{5}$$

where the summation is over values of s and t that index observations in the training set. 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, or EM algorithm (Dempster, Laird, and Rubin 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 "missing data" such that, if we knew the missing data, the estimation problem would be straightforward. The missing data do not have to be actual data that are missing; instead, they are often latent or unobserved quantities, knowledge of which would simplify the estimation problem. The BMA model (2) is a finite mixture model (McLachlan and Peel 2000). Here we introduce "missing data" z_{kst} where $z_{kst} = 1$ if ensemble member k is the best forecast for verification place 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

$$\hat{z}_{kst}^{(j)} = \frac{g(y_{st} | \tilde{f}_{kst}, \sigma_k^{(j-1)})}{\sum_{i=1}^{K} g(y_{st} | \tilde{f}_{ist}, \sigma_i^{(j-1)})}, \tag{6}$$

where the superscript j refers to the jth iteration of the EM algorithm, and $g(y_{st} | \tilde{f}_{kst}, \sigma_k^{(j-1)})$ is a normal density with mean $\tilde{f}_{kst}$ and standard deviation $\sigma_k^{(j-1)}$ evaluated at y_{st}. The M

step then consists of estimating the w_k and the σ_k using as weights the current estimates of z_{kst}, i.e. $\hat{z}_{kst}^{(j)}$. Thus

$$\begin{aligned} w_k^{(j)} &= \frac{1}{n}\sum_{s,t}\hat{z}_{kst}^{(j)}, \\ \sigma_k^{2(j)} &= \frac{\sum_{s,t}\hat{z}_{kst}^{(j)}(y_{st}-\tilde{f}_{kst})^2}{\sum_{s,t}\hat{z}_{kst}^{(j)}}, \end{aligned}$$

where n is the number of observations (i.e. 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.

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

c. The BMA Predictive Variance Decomposition and the Spread-Skill Relationship

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

$$\mathrm{Var}(y_{st}|\tilde{f}_{1st},\ldots,\tilde{f}_{Kst}) = \sum_{k=1}^{K} w_k\left(\tilde{f}_{kst}-\sum_{i=1}^{K} w_i\tilde{f}_{ist}\right)^2 + \sum_{k=1}^{K} w_k\sigma_k^2 \qquad (7)$$

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

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

The first term represents the ensemble spread. Thus one would expect to see a spread-skill relationship, since the predictive variance includes the spread as a component. But it also implies that using the ensemble spread alone would underestimate uncertainty, because it ignores the second term on the right-hand side of (7) or (8). The second term would be zero only if the best forecast were always exact, which is not case.

Table 1: Forty-Eight Hour UW-MM5 Ensemble Forecasts of Sea-Level Pressure at Powell River, B.C. Initialized at 0000 UTC on February 23, 2000, Bias-Corrected Forecasts, BMA Weights, and Verifying Observation. The ensemble members are shown in increasing order of forecast value.

MM5 Initialization (Source)	ETA (NCEP)	NGM (NCEP)	NOGAPS (FNMOC)	AVN (NCEP)	GEM (CMC)
Forecast	991.5	993.6	995.7	996.8	1007.8
Bias-corrected forecast	992.5	994.2	1001.1	998.9	1009.5
BMA weight	.18	.27	.19	.03	.32
Observation			1009.7		

Thus BMA predicts a spread-skill relationship, but also predicts ensembles to 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.

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-hour forecast of sea-level pressure near Powell River, B.C., Canada, initialized at 0000 UTC on February 23, 2000 and verifying at 0000 UTC on February 25, 2000. As described below, a 40-day training period was used, in this case consisting of forecasts and observations in the 0000 UTC cycle from January 14 to February 23, 2000. A simple linear bias correction was used, with the coefficients estimated from the training period.

Table 1 shows the forecasts and the bias-corrected forecasts, the BMA weights, and the observation for the five members of the UW MM5 ensemble. There was strong disagreement among the ensemble members: four of the five forecasts were in close agreement with one another, while the forecast obtained with the CMC-GEM initialization differed from all the others by at least 11 mb. The verifying observation turned out to be outside the ensemble range, as happened for 49% of the cases in our dataset. The four forecasts that agreed closely were far from the verification, while the "outlying" CMC-GEM-MM5 forecast was close, and its bias-corrected version was very close indeed. The highest BMA weight was for the CMCGEM-MM5 forecast, which also turned out to be the best forecast in this case.

Figure 3 shows the BMA predictive PDF. This PDF (shown as the thick curve in the

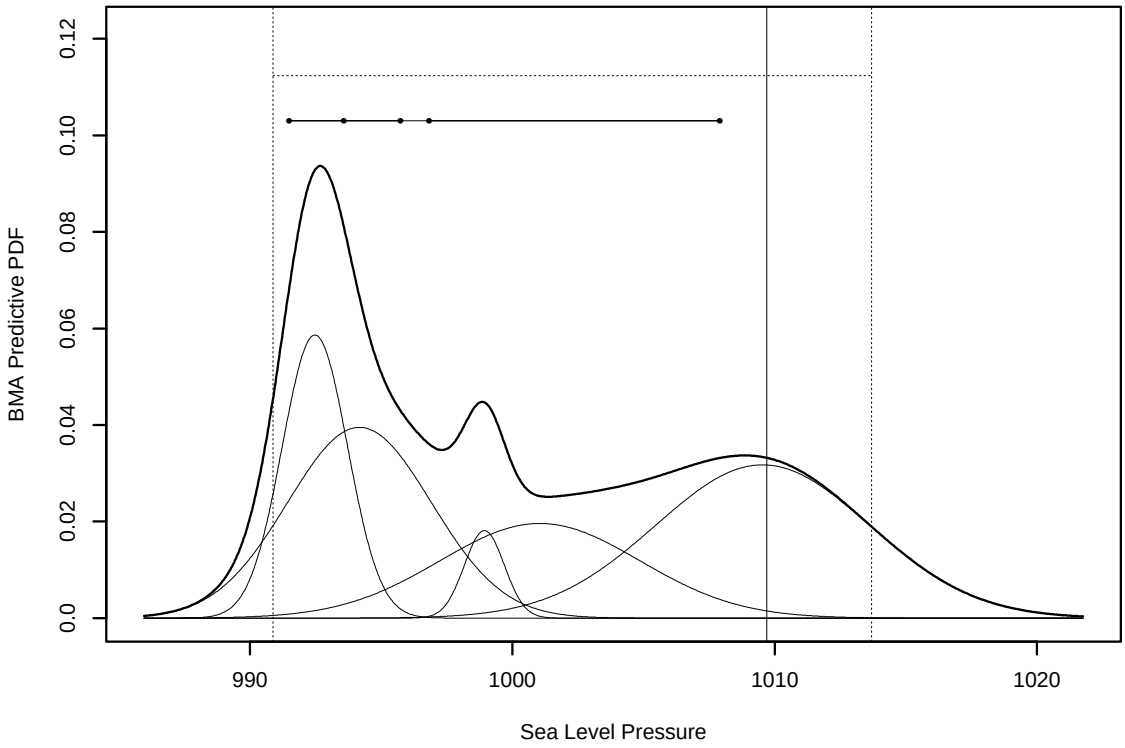

Figure 3: BMA Predictive PDF (thick curve) and its Five Components (thin curves) for the 48-Hour Sea-Level Pressure Forecast at Power River, B.C., Initialized at 0000 UTC on February 23, 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

figure), is a weighted sum of five normal PDFs (the components are the five thin lines). The distribution is multimodal, reflecting the disagreement between the forecasts. In particular, the broad right mode is centered around the bias-corrected CMC-GEM-MM5 forecast, which is far from the others. 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 size desired, simply by simulating from the predictive distribution (2). To simulate M values from the distribution (2), one can proceed as follows:
Repeat M times:

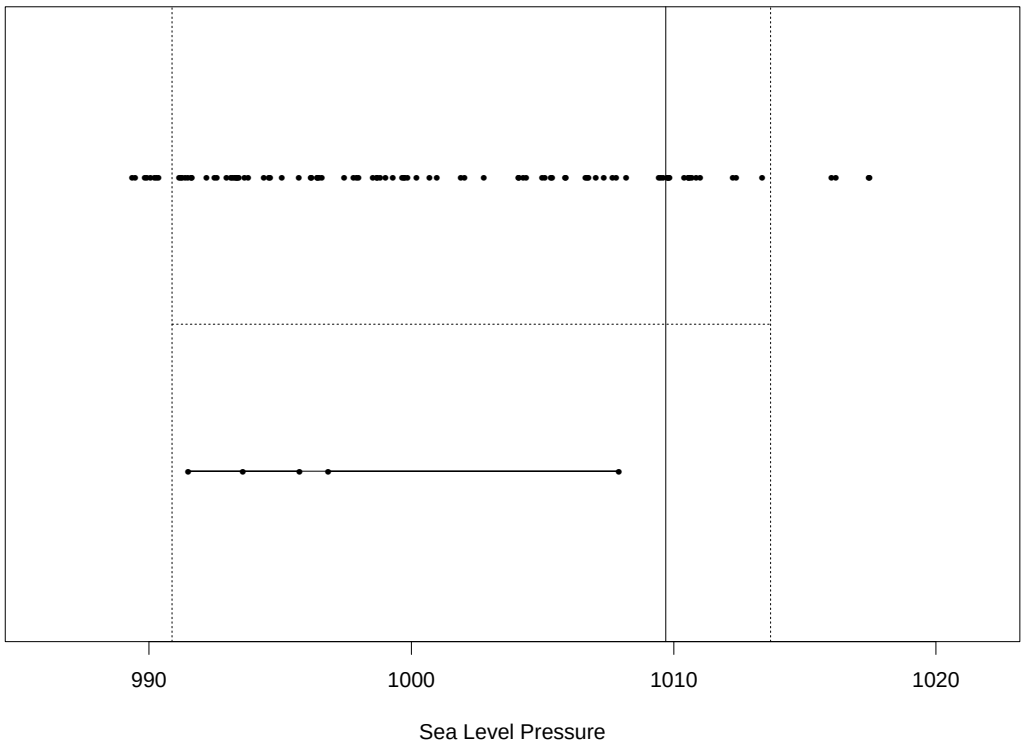

Figure 4: Ensemble of 100 Equally Likely Values from the BMA PDF (2) for the Powell River Sea-Level Pressure Forecast. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

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 $g_k(y|\tilde{f}_k)$. In the present case this will be a $N(\tilde{f}_k, \sigma_k^2)$ distribution.

Figure 4 shows a BMA ensemble of size $M = 100$ generated in this way. In this case, 88 of the 100 ensemble members lay within the exact 90% prediction interval.

3 Results

We now give results of the application of BMA to 48-hour sea-level pressure forecasts 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 temperature over the same period.

a. Length of Training Period

How many days should be used in the training period to estimate the BMA weights, variances 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 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 Gneiting principle that probabilistic forecasting methods should be designed to maximize sharpness subject to calibration, i.e. to make the prediction intervals as short as possible subject to their having the right coverage. We also tend towards making the training period as short as possible so as to be able to adapt as quickly as possible to the changing relative performance of ensemble members, lengthening it only if doing so seems to yield a clear advantage. Here we focus on 90% prediction intervals. We considered training periods of lengths 10, 20, ..., 100 calendar days. For comparability, the same verifications were used in evaluating all the training periods, so the verifications for the first 100 days were not used. For some days the data were missing (Grimit and Mass 2002), so that the effective number of days in the training data was smaller than the nominal number of calendar days in some instances.

Figure 5(a) shows the coverage of BMA 90% prediction intervals. It seems clear that 10 and 20-day training periods yield underdispersive PDFs, as do the 30-day periods, although less so. The 40-day training period gives accurate coverage, and longer periods give BMA PDFs that are slightly overdispersive. Figure 5(b) shows the average lengths of the 90% prediction intervals. These increase beyond 40 days, suggesting that there is little to be gained by increasing the length beyond 40 days. Thus a 40-day training period seems closest to maximizing sharpness subject to calibration: among those that are reasonably well calibrated (40 days and above), it has the shortest average intervals.

Figure 5(c) shows the RMSE of the BMA deterministic forecast given by (4). This decreases as the length of the training period increases from 10 to 40 days, and continues to decrease thereafter up to 80 days, but more slowly. Finally, Figure 5(d) 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. Like the RMSE, this decreases rapidly as the training period increases from 10 to 40 days, and more slowly as the training period is increased beyond 40 days.

To summarize these results, it seems that there are substantial gains in increasing the training period up to 40 days, and that beyond that there is little gain. We have therefore used 40 days here. It seems likely that different training periods would be best for other variables, forecast horizons, 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 calibration rank histogram for BMA is shown in Figure 6. To compute the calibration rank histogram we proceeded as follows. For each forecast initialization time at each station, 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 Figure 2. As can be seen, it was reasonably well calibrated, and clearly much more so than the ensemble itself.

Figure 7 shows the BMA weights for the five ensemble members over the evaluation period. The weights for ETA-MM5 and NGM-MM5 were consistently low throughout, while the weights for AVN-MM5 were consistently substantial. There seemed to be a tradeoff between CMC-GEM-MM5 and NOGAPS-MM5: in the first half of the evaluation period CMC-GEM-MM5 had high weight and NOGAPS-MM5 had low weight, while in the second half of the period the opposite was the case. The average weights of the five ensemble members over the evaluation period were as follows: AVN-MM5: 0.38; CMC-GEM-MM5: 0.16; ETA-MM5: 0.01; NGM-MM5: 0.01; NOGAPS-MM5: 0.44. This suggests that the ETA-MM5 and NGM-MM5 forecasts were not useful during this period, relative to the other three ensemble members.

Table 2 shows the coverage of various prediction intervals. We included the prediction interval from sample climatology, i.e. from the marginal distribution of our full dataset;

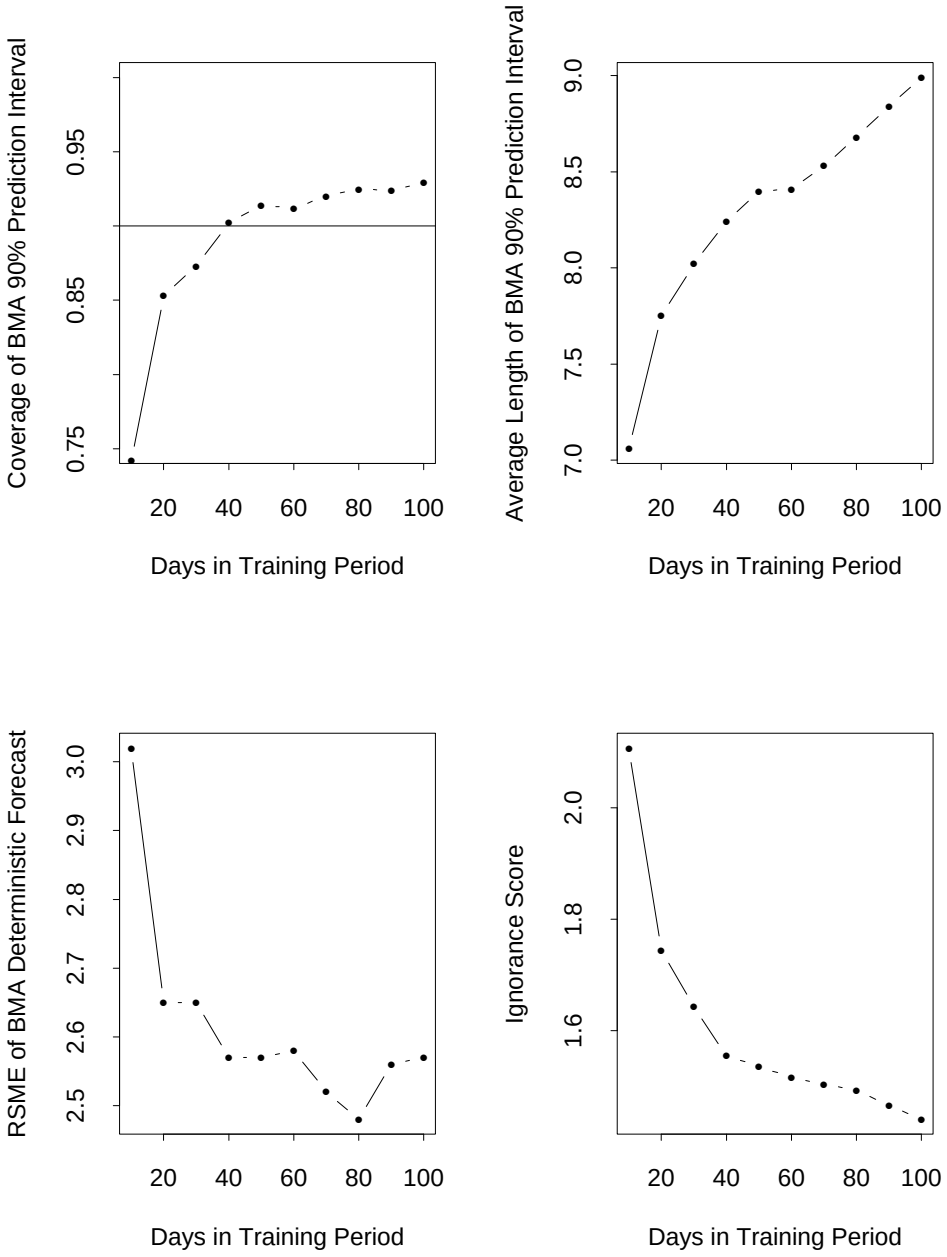

Figure 5: Comparison of Training Period Lengths for Sea-Level Pressure: (a) Coverage of 90% prediction intervals. (b) Average width of 90% prediction intervals. (c) RMSE of BMA deterministic forecasts. (d) Ignorance score.

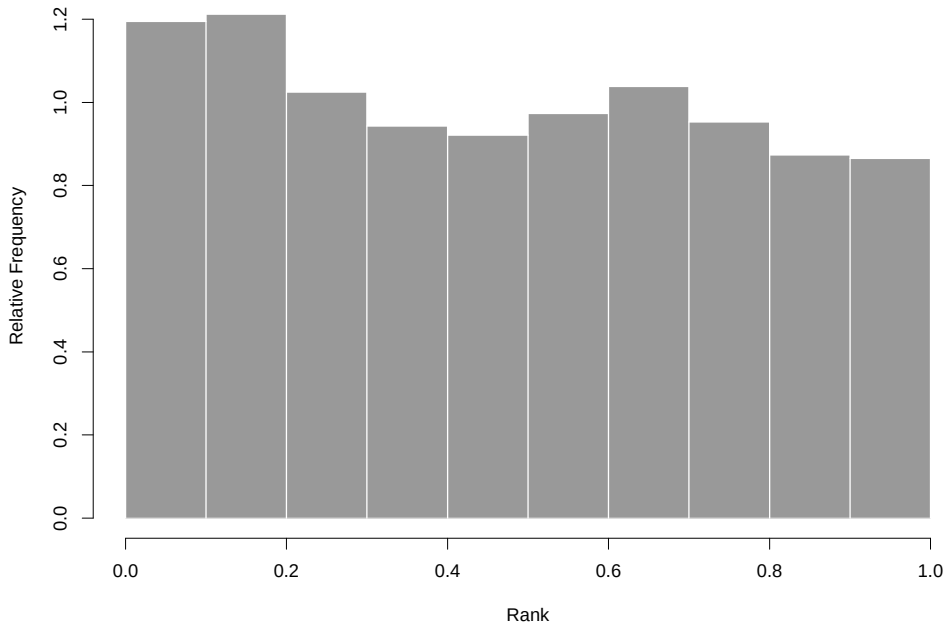

Figure 6: Calibration Rank Histogram for Bayesian Model Averaging for Sea-Level Pressure

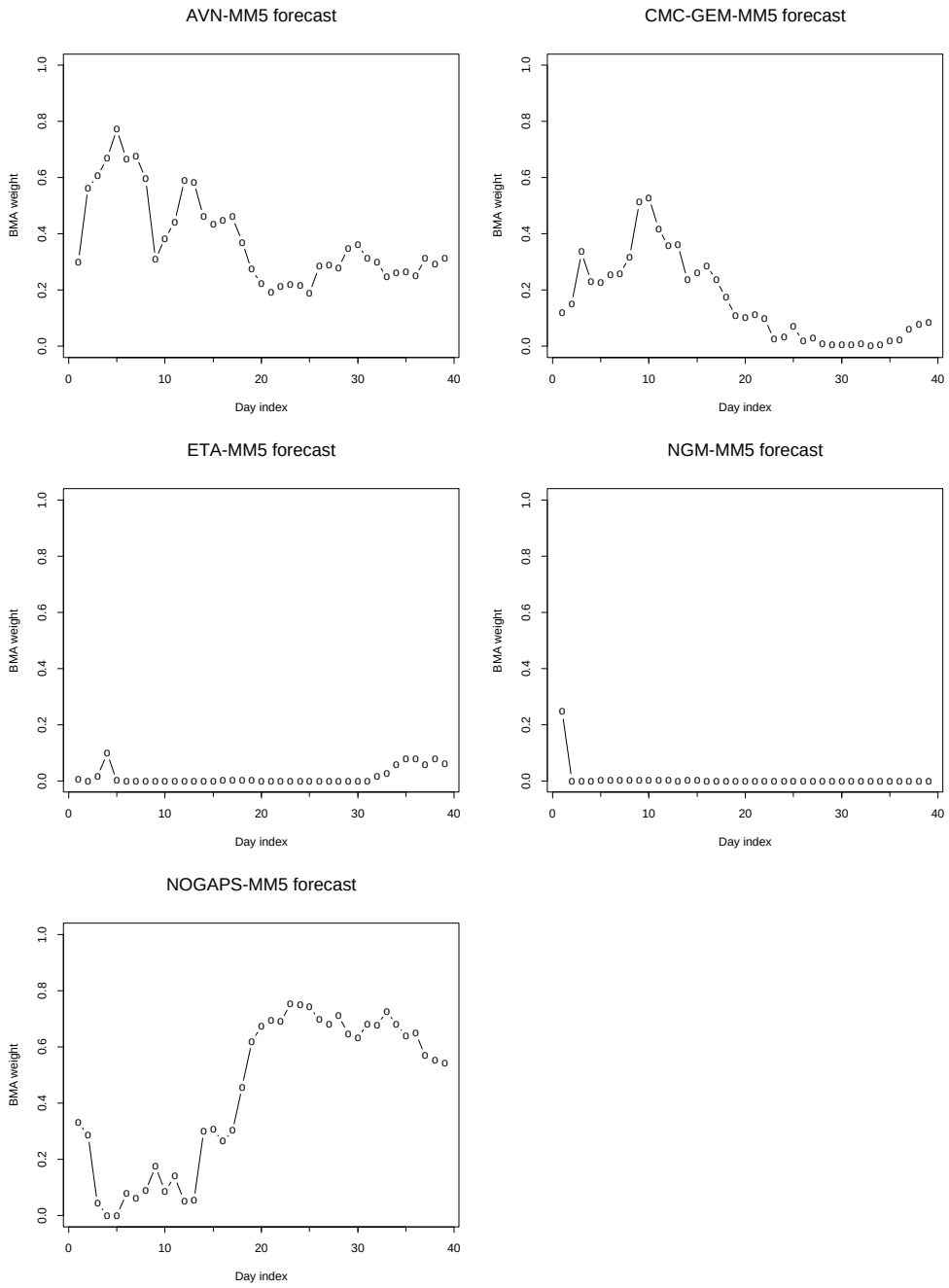

Figure 7: Bayesian Model Averaging Weights for the Five Models Over the Evaluation Period for Sea-Level Pressure

Table 2: 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	65.6	90.2

Table 3: 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	4.8	8.2

this interval is the same for each day and is useful as a baseline for methods that use the numerical weather predictions. There were no strong seasonal effects in our data and the distribution of sea-level pressure was roughly stationary in time, and spatial effects were a small part of overall variability, making it more reasonable to use sample climatology as a baseline. 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 close to having the right coverage.

Table 3 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 not much wider on average than the ensemble range, and it is calibrated. The climatological and BMA 90% intervals were both approximately calibrated, but the BMA intervals were over 60% narrower on average. STOPPED

Table 4 shows the RMSEs of the various deterministic forecasts considered over the evaluation period. The numerical weather prediction forecasts performed much better than the approximate climatological forecast (with all forecasts equal to the sample mean), and among these the AVN-MM5 forecast was best on average. The ensemble mean performed better than any of the individual ensemble members, with RMSE 6% lower than that for the best ensemble member. This agrees with published results for several ensembles, including the UW ensemble on which our work is based (Grimit and Mass 2002). The BMA deterministic

Table 4: RMSEs of Deterministic Forecasts for Sea-Level Pressure

Forecast	RMSE
Sample climatology	5.70
AVN-MM5	2.90
CMC-GEM-MM5	3.00
ETA-MM5	3.25
NGM-MM5	3.40
NOGAPS-MM5	3.21
Ensemble mean	2.72
BMA	2.57

forecast given by (4) in turn performed better than the ensemble mean, with RMSE 6% lower than that of the ensemble mean, and was the best forecast considered for these data.

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. Often 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. 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, two of the ensemble members, ETA-MM5 and NGM-MM5 had very low weights on average, both averaging 0.01. One might then consider eliminating these two members, and using instead a reduced three-member ensemble.

Table 5 compares the results for the five-member and the reduced three-member ensemble over the evaluation period. They are almost indistinguishable, and the ignorance score actually improves slightly when the two less useful members are removed. This would suggest that these members can be removed with little cost in terms of performance, and the operational gain could be considerable. Before making such a decision, however, it would be necessary to study the BMA weights over a longer period and for all the variables and forecast horizons of interest. Ensemble members that contribute little to forecasting one

Table 5: Comparison of BMA PDFs from the Five-Member Ensemble and the Reduced Three-Member Ensemble for Sea-Level Pressure

	90% Prediction Interval			
	Coverage	Average Length	RMSE	Ignorance Score
5-member ensemble	90.2%	8.25	2.57	1.554
3-member ensemble	90.5%	8.27	2.57	1.548

Note: The ignorance score is the mean negative log predictive probability of the observed values.

variable might be useful for others.

d. Results for Temperature

4 Discussion

We have proposed a new method for statistical postprocessing of ensemble output to produce calibrated and sharp predictive PDFs. It is based on Bayesian model averaging, a statistically principled way of combining forecasts from different models and analyses, and provides a theoretical explanation of the empirical phenomenon of ensembles exhibiting a spread-skill relationship while still being underdispersive. In our experiment, the BMA PDFs were much better calibrated than the ensemble itself, and produced prediction intervals that were much sharper than those produced by approximate 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 homogenous 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 ECMWF ensemble, 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 accomodated 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 conditional densities $g_k(y|\tilde{f}_k)$ in the BMA model (2) can reasonably be taken to be normal densities. This works well for sea-level pressure, and in experiments not reported here it also worked well for temperature 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 itself could be extended to these situations by using a non-normal conditional distribution $g_k(y|\tilde{f}_k)$ 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 improve the bias correction method. In our experiment we used a very simple linear bias correction method. 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 Eckel and Mass (2003) and Gel, Raftery, and Gneiting (2003b). 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 and Mitchell 1998; Houtekamer and Mitchell 2001; Gel, Raftery, and Gneiting 2003a). 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 a statistical model, and in not accounting for the spread-skill relationship. The method was applied to the ETA-RSM short-range ensemble, which did not exhibit a spread-skill relationship.

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. In contrast, BMA takes account of this uncertainty through the use of the mixture likelihood, and it is estimated explicitly by the $\hat{z}_{kst}$ that are produced by the EM algorithm.

BMA is designed to produce probabilistic forecasts, but as a byproduct it also produces a deterministic forecast, and this outperformed all the ensemble members as well as the ensemble mean in our experiment. 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, Kishtawal, LaRow, Bachiochi, Zhange, Williford, Gadgil, and Surendan 1999). 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 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.

Acknowledgements: The authors are grateful to Mark Albright, Eric Grimit and Clifford Mass for helpful discussions and useful comments, and for providing data. 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

Casella, G. and R. L. Berger (2001). *Statistical Inference* (2nd ed.). Brooks Cole.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–39.

Eckel, F. and C. F. Mass (2003). Towards an effective short-range ensemble forecast system. In *Proceedings of the Workshop on Ensemble Forecasting, Val-Morin, Québec.* Available at www.cdc.noaa.gov/~hamill/ef_workshop_2003_schedule.html.

Eckel, F. and M. Walters (1998). Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting 13*, 131–1147.

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

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A 222*, 309–368.

Gel, Y., A. E. Raftery, and T. Gneiting (2003a). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method. Technical Report 427, Department of Statistics, University of Washington. Available at

www.stat.washington.edu/tech.reports.

Gel, Y., A. E. Raftery, and T. Gneiting (2003b). Combining local and global grid-based bias removal for mesoscale numerical weather prediction models. Unpublished manuscript, Department of Statistics, University of Washington.

Gneiting, T., A. E. Raftery, F. Balabdaoui, and A. Westveld (2003). Verifying probabilistic forecasts: Calibration and sharpness. In *Proceedings of the Workshop on Ensemble Forecasting, Val-Morin, Québec*. Available at www.cdc.noaa.gov/~hamill/ef_workshop_2003_schedule.html.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B 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. *Weather and Forecasting 17*, 192–205.

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

Hamill, T. M. and S. J. Colucci (1998). Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review 126*, 711–724.

Hamill, T. M., C. Snyder, and R. E. Morss (2000). A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Monthly Weather Review 128*, 1835–1851.

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science 14*, 382–401. A corrected version with typos corrected is available 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. *Monthly Weather Review 129*, 73–91.

Houtekamer, P. L. and J. Derome (1995). Methods for ensemble prediction. *Monthly Weather Review 123*, 2181–2196.

Houtekamer, P. L. and H. L. Mitchell (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review 126*, 796–811.

Houtekamer, P. L. and H. L. Mitchell (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review 128*, 123–137.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. Bachiochi, Z. Zhange, 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*. New York: Wiley.

Leith, C. E. (1974). Theoretical skill of Monte-Carlo forecasts. *Monthly Weather Review 102*, 409–418.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science 20*, 131–140.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. New York: Wiley.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis (1996). The ECMWF ensemble system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society 122*, 73–119.

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

Raftery, A. E. and Y. Zheng (2003). Long-run performance of Bayesian model averaging. *Journal of the American Statistical Association 98*, to appear.

Roulston, M. S. and L. A. Smith (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review 130*, 1653–1660.

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

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. *Journal of Climate 15*, 537–544.

Stensrud, D. J., H. E. Brooks, J. Du, S. Tracton, and E. Rogers (1999). Using ensembles for short-range forecasting. *Monthly Weather Review 127*, 433–446.

Stern, R. D. and R. Coe (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society, Series A 147*, 1–34.

Toth, Z. and E. Kalnay (1993). Ensemble forecasting at the NMC: The generation of perturbations. *Bulletin of the American Meteorological Society 74*, 2317–2330.

van den Dool, H. M. and L. Rukhovets (1994). On the weights for an ensemble-averaged 6-10-day forecast. *Weather and Forecasting 3*, 457–465.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. San Diego: Academic Press.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*, 95–103.