A PRACTICAL MODEL BLENDING TECHNIQUE BASED ON BAYESIAN MODEL AVERAGING

Bruce Veenhuis*

Meteorological Development Laboratory
Office of Science and Technology
National Weather Service
Silver Spring, Maryland

1. INTRODUCTION

The National Weather Service's Meteorological Development Laboratory (MDL) has developed a technique called Updatable Bayesian Model Averaging (UBMA) that combines multiple numerical weather prediction (NWP) forecasts into a single probabilistic consensus. Our technique is based on the Bayesian Model Averaging (BMA) method proposed by Raftery et al. (2005). Our technique, however, is easier to implement operationally, and despite its simplicity, creates accurate and reliable consensus forecasts.

Like the original BMA formulation, we dress each model with a probabilistic kernel and combine the kernels to create a calibrated probability distribution. The kernel shape depends on the climatological distribution of the element being postprocessed. For each model, the kernel location is determined by the bias corrected model forecast, while the kernel scale is controlled by the model's performance over a recent training period. More skillful models are given a more pronounced kernel and will ultimately have a stronger influence on the final forecast.

To implement BMA, we must estimate the kernel parameters for each model. Raftery et al. (2005) originally used the Expectation Maximization (EM; Dempster el al. 1977) algorithm. In operations however, the EM algorithm may be problematic for several reasons. Since the EM algorithm is iterative, the entire training set must be kept on hand which may not be feasible, especially when training to a high resolution gridded analysis. Likewise, Hamill (2007) demonstrated that the EM algorithm may overfit the training data if only a short training period is used. Overfitting causes unstable parameter estimates and poor performance when the algorithm is applied to independent data.

Instead of the EM, we use a decaying average algorithm that continuously updates the parameters based on recent performance. The specific details of our approach are explained below along with verification.

2. METHODOLOGY

As mentioned previously, we must estimate a set of model-specific kernel parameters. The details of the parameters depend on the chosen kernel shape. Here, we use a normal kernel, which is appropriate for elements such as temperature. The final BMA forecast probability density function (PDF) can then be expressed as

$$P(t) = \sum_{k=1}^{K} w_k N(t|f_k, \sigma). \qquad (1)$$

Here the term $N(t|f_k,\sigma)$ is the value of a normal density at point t with mean f_k and standard deviation σ. The term w_k is the weight for model k and K is the total number of models in the system. Thus, the probability that the temperature t takes any particular value is equal to the weighted sum of the probability from each model-specific kernel. We should note, if certain models are substantially more skillful, it may be advisable to estimate a unique standard deviation σ for each model. However, we have found estimating a single σ is adequate for the models we have tested.

To create probabilistic forecasts with (1) we must find a set of model-specific weights and an appropriate value for σ. For this work, we estimate a unique set of weights and standard deviations for each station, element, and time projection. In general, our computations closely follow Raftery et al. (2005); however, we use a decaying average algorithm to fit the parameters rather than the EM algorithm. Note that BMA assumes each model

**Corresponding author address:*
Bruce Veenhuis, Meteorological Development Laboratory, 1325 East-West Highway, Silver Spring, MD 20910; email: bruce.veenhuis@noaa.gov

forecast has been bias corrected prior to fitting the kernel parameters. Here, we apply a simple decaying average bias correction to each model which is similar to the algorithm used by Cui et al. (2012).

We initialize the system by equally weighting each model and set the standard deviation equal to standard error of the ensemble mean forecast over the preceding 30 days. Each day, a new observation arrives which we use to verify previously made forecasts and adjust the weights and standard deviation accordingly. To begin the update we compute

$$z_k^j = \frac{w_k^{j-1} N(x|f_k, \sigma^{j-1})}{\sum_{k=1}^{K} w_k^{j-1} N(x|f_k, \sigma^{j-1})}. \qquad (2)$$

Here z_k^j can be interpreted as an instantaneous estimate of the weight for model *k* at time *j* based on how well the model predicted the single observation *x*. The superscript *j* denotes the present time, while *j-1* indicates values from the last time we ran the algorithm. So, w^{j-1} and σ^{j-1} are the older parameter estimates we seek to update.

The term f_k is the bias-corrected forecast for model *k*. The function $N(x|f_k,\sigma^{j-1})$ is the height of a normal distribution with mean f_k and standard deviation σ^{j-1} at the observation *x*. This term is multiplied by the weight w^{j-1} which is our previous estimate for that model's weight. In essence, we are examining each bias-corrected model forecast and computing how much probability that model -- along with its kernel -- gave to the observation that actually occurred. A model that gave a high probability will be rewarded with a larger z_k^j value. The term in the denominator normalizes the values so the z_k^j values for all models sum to unity.

The instantaneous weights from (2) are quite noisy from day-to-day because even an inaccurate model may occasionally make the best forecast. Therefore we use a decaying average algorithm to smooth the instantaneous weight estimates as follows

$$w_k^j = (1-\alpha) w^{j-1} + \alpha z_k^j, \qquad (3)$$

where the latest stable estimate of the weight w_k^j is equal to the previous estimate multiplied by 1 minus the decaying weight α plus the instantaneous estimate z_k^j multipled by α. We have found an α value of 0.05 works well and indicates that ~95% of the training is determined by the previous 60 days.

Recall we must also find a value for the standard deviation in (1). Our strategy is similar to that for updating the weights however the specific computations are different. Each day, when a new observation arrives we compute

$$s^j = \left(\sum_{k=1}^{K} w_k^j (x - f_k)^2\right)^{\frac{1}{2}}, \qquad (4)$$

where w_k^j is our latest estimate of the weight for model *k* and *x* is the observation that verifies the bias-corrected forecast f_k. The term s^j can be viewed as an instantaneous estimate of the standard deviation parameter. Again, we use a decaying average algorithm to find a stable σ^j estimate

$$\sigma^j = (1-\beta)\sigma^{j-1} + \beta s^j, \qquad (5)$$

where we use the decaying weight β to combine our instantaneous estimate s^j with the previous estimate σ^{j-1}. A value of 0.05 for β has been found to work well.

We use the algorithm sketched above to maintain reasonable estimates of the weights and standard deviation in (1). In our operational implementation, when the latest model runs are available and ready for postprocessing, we create a UBMA probabilistic consensus forecast by dressing each model with the appropriate kernel and evaluating (1) to create the forecast PDF. To actually distribute the probabilistic information, we compute several points along the cumulative distribution function (CDF) by integrating the PDF. We also compute and distribute the weighted mean forecast and the standard deviation of the PDF.

3. VERIFICATION RESULTS

To test the UBMA algorithm, we created a consensus probabilistic MOS product by combining MDL's MOS guidance from the GFS, NAM, ECMWF, and EKDMOS (an ensemble-based product derived from the NAEFS; Glahn et al. 2009, Veenhuis 2013). We applied the technique to station-based forecasts at 335 stations distributed throughout the CONUS, Alaska, Hawaii, and Puerto Rico. We generated station-based UBMA forecasts for 2-m temperature for the 6-hr to 192-hr projections valid every 6 hours. The algorithm was initialized

1 October 2011, and forecasts evaluated over the period from 1 November 2011 to 31 March 2012. Decaying average bias correction has been shown to improve the accuracy of MOS forecast (Glahn 2014) and was applied to each MOS product prior to running the UBMA algorithm.

Figure 1 shows the mean absolute error (MAE) of the 2-m temperature forecasts for each individual MOS product, as well as the UBMA consensus mean. UBMA was always slightly more accurate than the best individual MOS product for all projections.

To judge the statistical reliability of the UBMA forecasts we created cumulative reliability diagrams (CRD; Fig. 2). CRDs are similar to the more familiar reliability diagrams; however, here the probabilities are cumulative from below. If the probabilistic forecasts are reliable, the points should fall along the diagonal reference line which indicates the predicted cumulative probability matched the observed frequency. At both projections the UBMA forecasts were well calibrated.

We have also applied UBMA to the Short Range Ensemble Forecast (SREF) system. Specifically we used UBMA to calibrate the spread of upper air temperature and layer thickness forecasts. These products give a reliable estimate of the forecast uncertainty associated with thermal profile of the atmosphere and help forecasters determine precipitation type.

The SREF contains 21 members but only 3 separate model cores. Members with the same core should have the same skill and bias characteristics, therefore we only estimated 3 distinct weights (one for each core) and a single standard deviation. We used the North American Mesoscale (NAM) Data Assimilation (NDAS) gridded analysis as our proxy for truth. We interpolated both the SREF forecasts and the NDAS to stations and applied a simple decaying average bias correction prior to UBMA. We initialized the algorithm 15 November 2012 and evaluated the performance over the period 15 December 2012 to 10 February 2013.

For brevity, we only present results for 850-hpa temperature from the 0900 UTC cycle of the SREF. Figure 3 shows the continuous ranked probability score (CRPS) which is a negatively oriented score that verifies the predicted CDF. The UBMA CRPS was lower than that obtained by forming a CDF from the raw members. Probability integral transform (PIT) histograms comparing the raw members to UBMA at the 48-hr projection are shown in Figure 4. The raw members were underdispersed while the UBMA forecasts were much more reliable.

We produce SREF-based UBMA guidance in real-time on an experimental basis. The UBMA products and additional documentation are available online at:
http://www.mdl.nws.noaa.gov/~BMA-SREF/BMAindex.php

4. CONCLUSIONS

MDL has developed a technique called Updateable Bayesian Model Averaging (UBMA) which creates statistically reliable probabilistic consensus forecasts from multiple inputs. UBMA is similar to the BMA technique proposed by Raftery et al. (2005) but uses a decaying average algorithm to update the statistical model parameters rather than an iterative algorithm. UBMA is easier to implement operationally because less data must be retained for training. For example, to calibrate an operational global model we must only store previous model forecasts for 16 days (assuming the model is integrated to 384-hr) but can effectively train to a much longer sample size by choosing a small decaying weight.

Despite the algorithm's simplicity, we have shown UBMA can improve forecast accuracy and reliability. The UBMA consensus MOS forecasts were more accurate than any individual MOS product. Likewise, UBMA-calibrated SREF forecasts were more reliable than the raw ensemble.

A limitation of UBMA and BMA in general is that the technique can only increase the ensemble spread which may be problematic if the raw ensemble is overdispersed. We have also only used UBMA to calibrate normally distributed elements such as temperature. Future work should explore if UBMA or a similar technique can be applied to non-normal elements such as wind speed or precipitation amount.

5. ACKNOWLDEGMENTS

The author wishes to acknowledge John Wagner, Michelle Cohen, and Mark Oberfield who helped implement and maintain the SREF UBMA products.

6. REFERENCES

Cui B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396-410.

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.

Glahn, H. R., M. R. Peroutka, J. Wiedenfeld, J. L., Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246-268.

Glahn, H. R., 2014: Determining the optimal decay factor for bias-correcting MOS temperature and dewpoint forecasts. Preprints*, 22nd Conference on Probability and Statistics in the Atmospheric Sciences,* Atlanta, GA, Amer. Meteor. Soc., 6.3

Hamill, T. M., 2007: Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging". *Mon. Wea. Rev.*, **135**, 4226-4230.

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.

Veenhuis, B. A., 2013: Spread calibration of ensemble MOS forecasts. *Mon. Wea. Rev.*, **141**, 2467-2482.

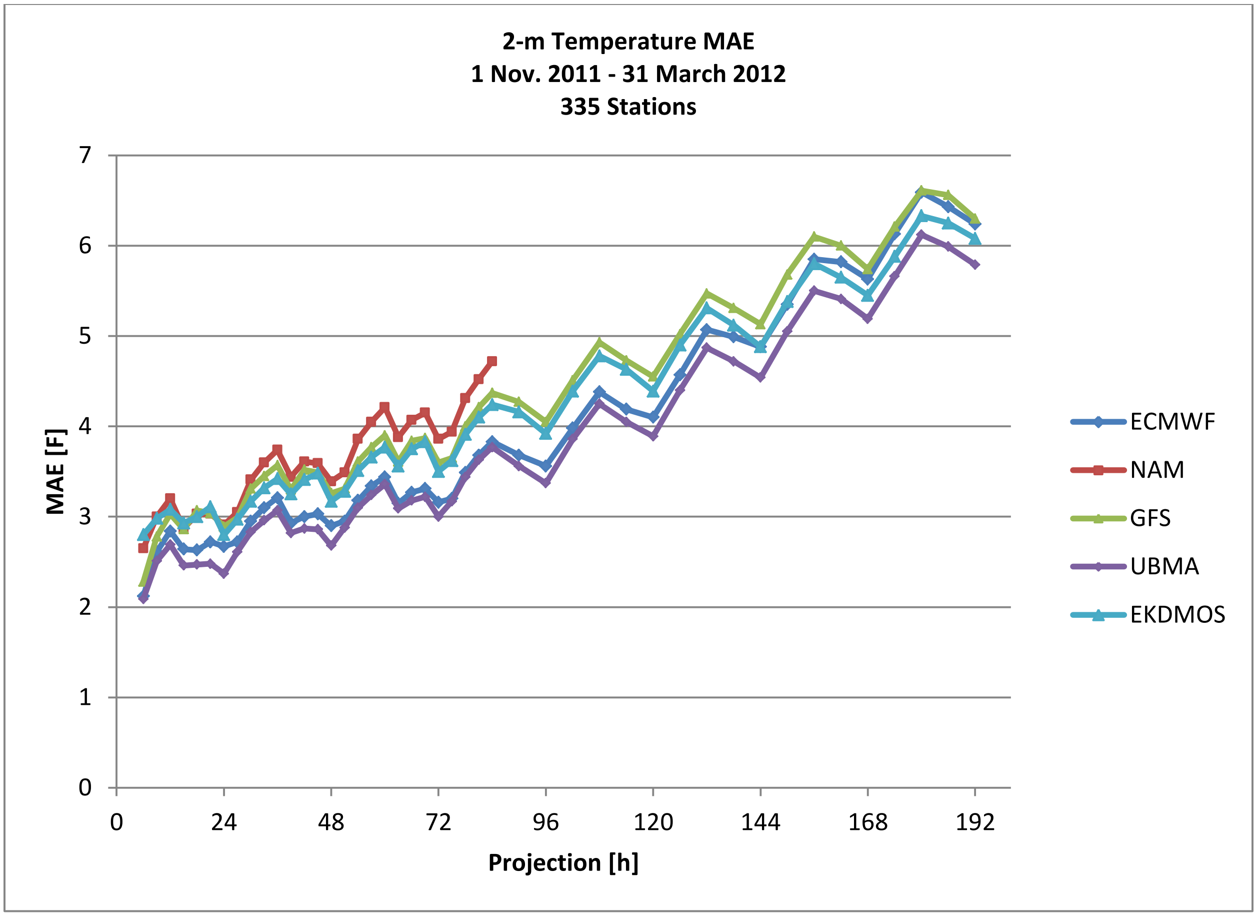

Fig. 1. Mean absolute error (MAE) for the individual MOS products and for the UBMA consensus.

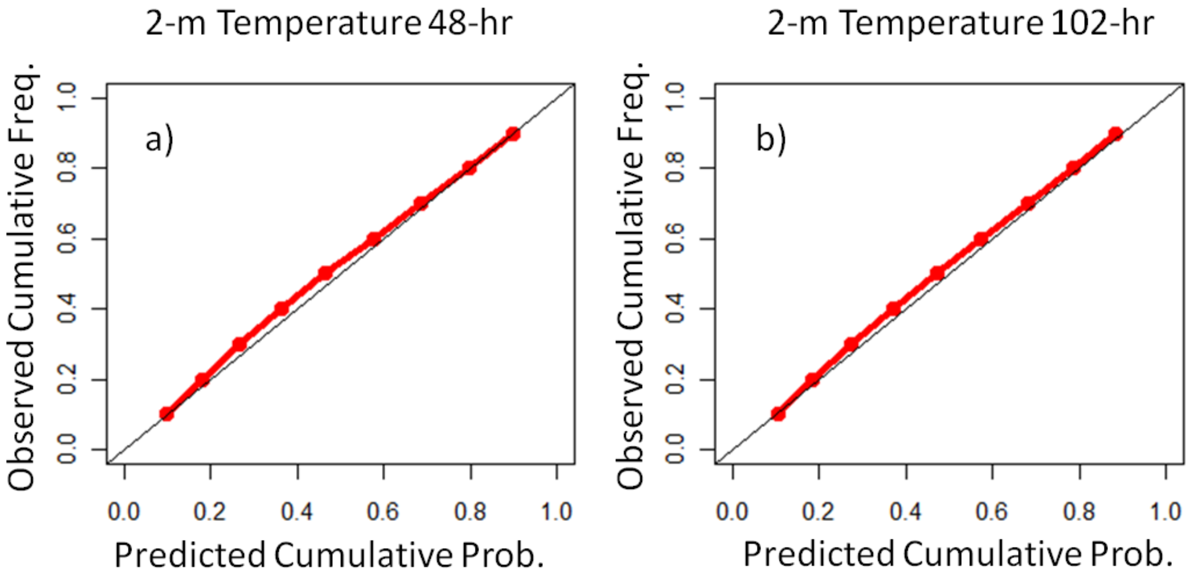

Fig. 2. Cumulative reliability diagrams (CRD) for the 48-hr (a) and 102-hr (b) 2-m UBMA temperature forecasts.

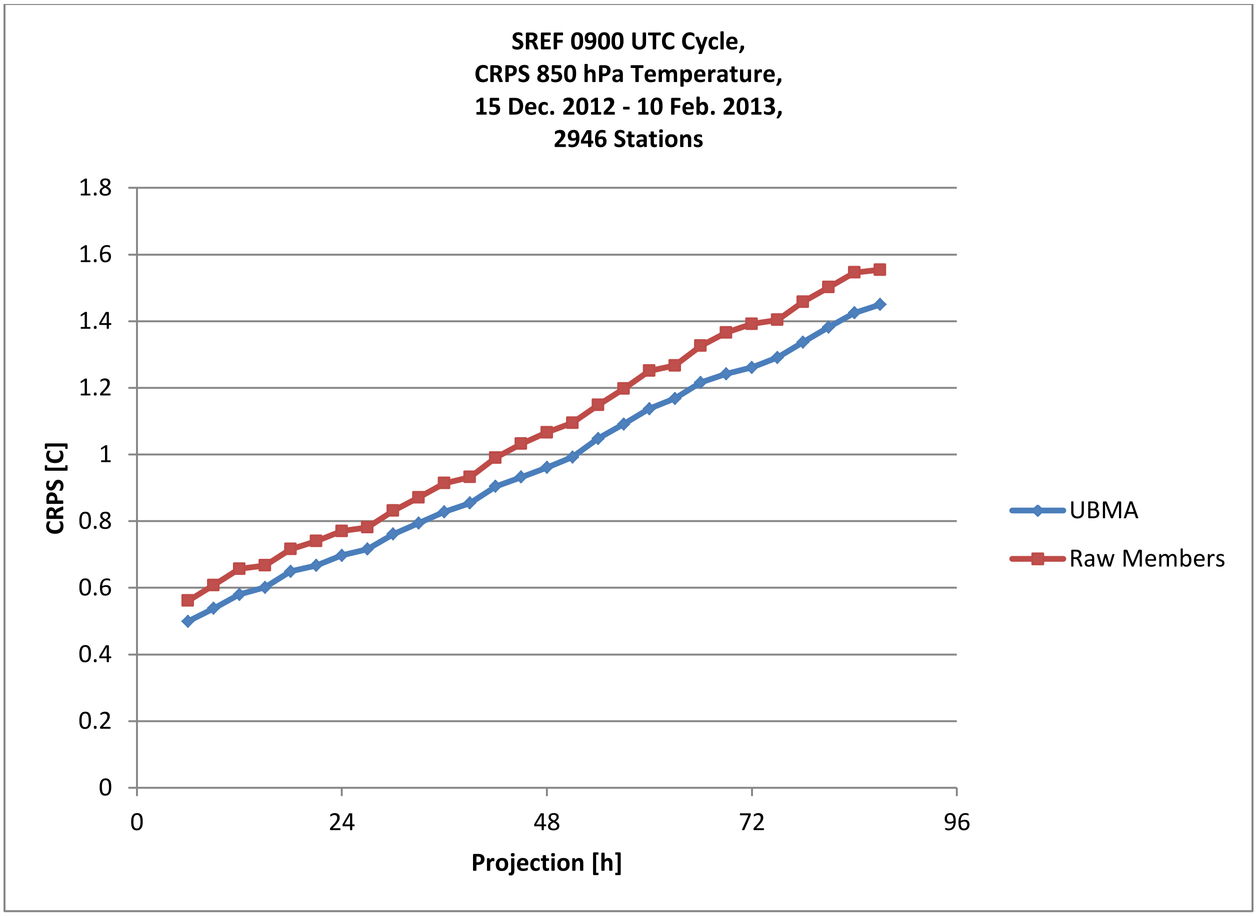

Fig. 3. Continuous ranked probability score (CRPS) comparing the skill of the raw members (red) to the UBMA forecasts (blue).

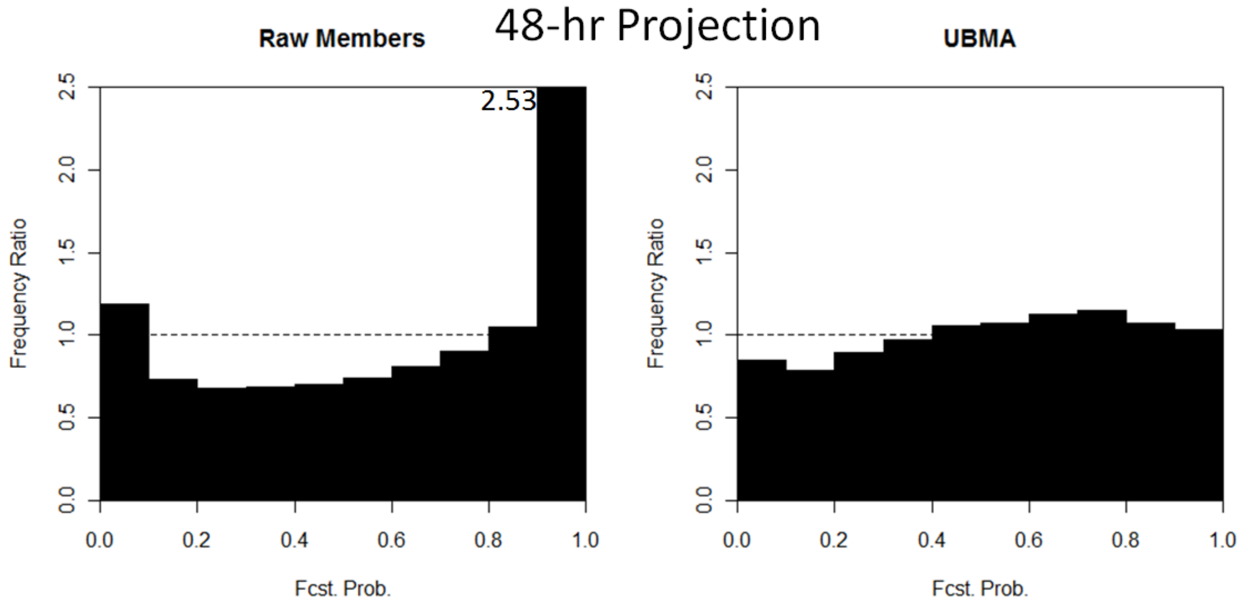

Fig. 4. Probability integral transform (PIT) histograms comparing the reliability of the raw members (a) to that of the UBMA forecasts (b).