Probabilistic Wind Vector Forecasting Using Ensembles and Bayesian Model Averaging

J. MCLEAN SLOUGHTER

Seattle University, Seattle, Washington

TILMANN GNEITING

Heidelberg University, Heidelberg, Germany

ADRIAN E. RAFTERY

University of Washington, Seattle, Washington

(Manuscript received 30 December 2011, in final form 26 May 2012)

ABSTRACT

Probabilistic forecasts of wind vectors are becoming critical as interest grows in wind as a clean and renewable source of energy, in addition to a wide range of other uses, from aviation to recreational boating. Unlike other common forecasting problems, which deal with univariate quantities, statistical approaches to wind vector forecasting must be based on bivariate distributions. The prevailing paradigm in weather forecasting is to issue deterministic forecasts based on numerical weather prediction models. Uncertainty can then be assessed through ensemble forecasts, where multiple estimates of the current state of the atmosphere are used to generate a collection of deterministic predictions. Ensemble forecasts are often uncalibrated, however, and Bayesian model averaging (BMA) is a statistical way of postprocessing these forecast ensembles to create calibrated predictive probability density functions (PDFs). It represents the predictive PDF as a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights reflect the forecasts' relative contributions to predictive skill over a training period. In this paper the authors extend the BMA methodology to use bivariate distributions, enabling them to provide probabilistic forecasts of wind vectors. The BMA method is applied to 48-h-ahead forecasts of wind vectors over the North American Pacific Northwest in 2003 using the University of Washington mesoscale ensemble and is shown to provide better-calibrated probabilistic forecasts than the raw ensemble, which are also sharper than probabilistic forecasts derived from climatology.

1. Introduction

While deterministic point forecasts have long been the standard in weather forecasting, there are many situations in which probabilistic information can be of value. In this paper, we consider the case of wind vectors. In many situations, simultaneously forecasting both the direction and speed of wind is of interest. For example boaters need to know not only how strong the winds might be, but also the direction in which they will be blowing. As another example, directional information can be key to predicting the movement of airborne pollutants.

In these situations, it can be valuable to have more than just a best guess at the speed and direction of the wind. Probabilistic forecasts can allow for the assessment of likely scenarios and can provide the probabilities of particular scenarios of interest. This information can be of particular value in situations calling for a cost–loss analysis, where probabilities of different outcomes need to be known in order to make an optimal decision. The optimal point forecast in these situations is often a quantile of the predictive distribution (Roulston et al. 2003; Pinson et al. 2007a; Gneiting 2011). Different situations can require different quantiles, and this needed flexibility can be provided by forecasts of a full predictive probability density function (PDF).

Corresponding author address: J. McLean Sloughter, Department of Mathematics, Seattle University, 901 12th Ave., P.O. Box 222000, Seattle, WA 98122.
E-mail: sloughtj@seattleu.edu

DOI: 10.1175/MWR-D-12-00002.1

Medium-range forecasts looking several days ahead are generally based on numerical weather prediction models, which can then be statistically postprocessed. To estimate the predictive distribution of a weather quantity, an ensemble forecast is often used. An ensemble forecast consists of a set of multiple forecasts of the same quantity, based on different estimates of the initial atmospheric conditions and/or different physical models (Palmer 2002; Gneiting and Raftery 2005). By using different estimates of initial conditions or different physical models, ensemble forecasts attempt to capture the uncertainty in forecasts. A statistical relationship between forecast errors and ensemble spread has been established for several ensemble systems (Buizza et al. 2005). However, it has also been shown that ensemble forecasts are typically uncalibrated, with a tendency for observed values to fall outside of the range of the ensemble too often (Grimit and Mass 2002; Buizza et al. 2005; Gneiting and Raftery 2005).

A number of methods have been proposed for statistically postprocessing ensemble forecasts of wind speed or wind power (Bremnes 2004; Nielsen et al. 2006; Pinson et al. 2007b; Møller et al. 2008; Pinson and Madsen 2009; Sloughter et al. 2010; Thorarinsdottir and Gneiting 2010). These methods only address the scalar wind speed or power quantity, however, and do not provide forecasts for the wind vector. Glahn and Lowry (1972) proposed a method for statistically postprocessing vector wind forecasts, but their method produces only deterministic forecasts. Bao et al. (2010) developed a method for statistically postprocessing wind direction forecasts, but this only gives a marginal distribution for wind direction and not a joint distribution for the wind vector.

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 component distributions, each a PDF centered on the individual bias-corrected forecasts, where the weights can be interpreted as 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 scalar weather quantities. Here we develop a BMA method for wind vectors by relating the component distribution for a given ensemble member to a bivariate normal distribution; the BMA PDF is then itself a mixture of the resulting distributions.

Section 2 describes the data used in this study. In section 3, we review the BMA technique and describe our extension of it to wind vectors. Then in section 4 we give results for 48-h-ahead forecasts of wind vectors over the North American Pacific Northwest in 2003 based on the eight-member University of Washington mesoscale ensemble (Grimit and Mass 2002; Eckel and Mass 2005). Throughout the paper we use illustrative examples drawn from these data, and we find that BMA is better calibrated than the raw ensemble and sharper than climatology for the period we consider. Finally, in section 5 we discuss alternative approaches and possible improvements to the method.

2. Data

This research considers 48-h-ahead forecasts of ''instantaneous'' wind vectors over the Pacific Northwest in the period 1 November 2002–31 December 2003, using the eight-member University of Washington mesoscale ensemble (Eckel and Mass 2005) at a 12-km resolution initialized at 0000 UTC, which is 1700 local time in summer, when daylight saving time operates, and 1600 local time otherwise. The dataset contains observations and forecasts at surface airway observation (SAO) stations, a network of automated weather stations located at airports throughout the United States and Canada. Observations are recorded at a temporal resolution of 1 min. An hourly instantaneous wind vector is a 2-min average from the period of 2 min before the hour to on the hour. Data were available for 343 days, and data for 83 days during this period were unavailable due to failures in the model runs for those days. In all, 38 091 station observations were used, an average of about 111 station observations (with a single observation per station) per day. Figure 1 shows observations of maximum wind speeds from a typical day. Data from 2003 were used for verification, and the November and December 2002 data were used only as initial training data for the model.

We use a sliding training period for the model. For each day, new model parameters are estimated based on data over the previous 30 days. Thus, the model is continually updated to reflect current trends, and all verification is conducted out of sample.

The forecasts were produced for observation locations by bilinear interpolation from forecasts generated on a 12-km grid, as is common practice in the meteorological community. The wind vector observations were subject to the quality control procedures described by Baars (2005).

The wind vector data we analyze come from recording stations with a startup speed of 3 kt, so that any wind vector whose magnitude is less than 3 kt is recorded as 0. Previous work on modeling wind speed found that models that address discretizations such as this are

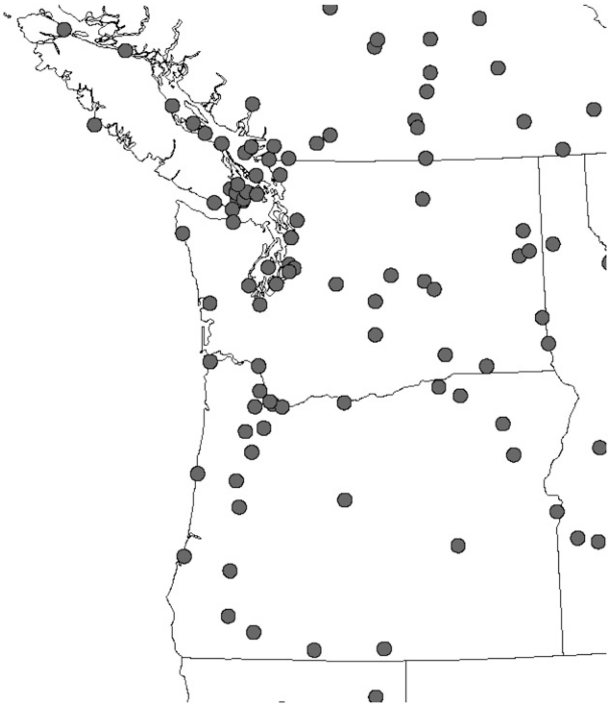

FIG. 1. Observation locations at SAO stations over the Pacific Northwest on 7 Aug 2003.

computationally more expensive, while not significantly improving forecasting results (Sloughter et al. 2010). As such, our model here does not attempt to model this discretization, instead treating these values as zero. One knot is equal to approximately 0.514 m s^{-1}, or 1.151 miles h^{-1}.

3. Methods

a. Bias correction

We aim to model the distribution of observed wind vectors conditional on the forecast wind vectors. To simplify the model, we assume that the mean of this conditional distribution is some function of the forecast vector, and that the covariance matrix does not depend upon the forecast. This assumption of homoscedasticity was assessed by examining the residuals (the error vectors after bias correction), and appears to be valid in our data.

Then our model, conditional upon a single forecast vector, can be expressed as

$$\mathbf{y} \mid \mathbf{f}_k \sim h(\mathbf{f}_k) + \mathrm{BV},$$

where BV is some bivariate distribution centered at 0. That is, if we think of $h(\mathbf{f}_k)$ as a bias-corrected forecast, then we can reframe the problem to consider modeling the bivariate distribution of the error vector rather than the observation vector. This allows us to model the mean and the covariance matrix separately.

For simplicity, we restrict ourselves here to considering bias-correction techniques that take the form of affine transformations. We look at two possible models, one modeling only an additive transformation:

$$\mathbf{Y} = \mathbf{a}_k + \mathbf{f}_k,$$

and the other a full affine transformation:

$$\mathbf{Y} = \mathbf{a}_k + \mathbf{B}_k \mathbf{f}_k.$$

Both methods are fit via least squares linear regression. We use a sliding training period for the model. For each day, new model parameters are estimated based on data over the previous 30 days. We evaluate performance here in terms of the bivariate root-mean-squared error, that is, the square root of the mean of the squared Euclidian norm of the error vectors. Table 1 shows how these two methods compare when applied to each of the eight ensemble members. We can see that the additive bias correction shows a lower bivariate root-mean-squared error than the raw forecasts, while the full affine bias correction shows further improvement over the additive bias correction. We therefore proceed with the full affine bias correction.

b. Bayesian model averaging

BMA (Leamer 1978; Madigan and Raftery 1994; 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 $\mathbf{f}_k$ is associated with a conditional PDF

TABLE 1. Bivariate root-mean-squared error (kt) for bias-correction methods applied to each ensemble member.

	GFS	GEM	ETA	GASP	JMA	NOGAPS	TCWB	UKMO
Raw forecasts	8.45	8.54	8.60	8.65	8.58	8.61	8.76	8.56
Additive bias correction	8.35	8.46	8.50	8.55	8.50	8.56	8.69	8.46
Affine bias correction	7.26	7.39	7.39	7.48	7.42	7.47	7.55	7.33

$p_k(\mathbf{y} \mid \mathbf{f}_k)$, which can be thought of as the PDF of the weather quantity $\mathbf{y}$, given $\mathbf{f}_k$, conditional on $\mathbf{f}_k$ being the best forecast in the ensemble. The BMA predictive PDF is then

$$p(\mathbf{y} \mid \mathbf{f}_1, \ldots, \mathbf{f}_K) = \sum_{k=1}^{K} w_k p_k(\mathbf{y} \mid \mathbf{f}_k), \qquad (1)$$

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_ks are probabilities and so they are nonnegative and add up to 1, that is, $\sum_{k=1}^{K} w_k = 1$. Here, K is the number of ensemble members.

c. Bivariate error model

For univariate weather quantities, BMA methods have been developed that fit the conditional PDFs using normal distributions (Raftery et al. 2005) or (possibly power transformed) gamma distributions (Sloughter et al. 2007).

For two-dimensional wind vectors, we consider bivariate normal distributions, centered at the bias-corrected forecasts. This is equivalent to modeling the error vector with a bivariate normal distribution with mean 0. Exploratory work showed that our error vectors had heavier tails than could be accounted for by the bivariate normal distribution. We address this by considering a transformation of the error vectors. Various transformations were assessed by quantile plots. We found that raising the magnitude of the error vector to the four-fifths power while preserving the angle produces values that can be modeled well by a bivariate normal distribution. It should be noted that there is no reason to assume that this particular transformation is universal, and similar exploratory work should be used in any new setting to determine what, if any, transformation would be appropriate.

We first consider the bias correction:

$$h_k(\mathbf{f}_k) = \mathbf{a}_k + \mathbf{B}_k \mathbf{f}_k .$$

We then find the angle θ_k and magnitude r_k of the error vector, $\mathbf{y} - h_k(\mathbf{f}_k)$, and produce a new transformed error vector:

$$\mathbf{e}_k(\mathbf{y}) = \begin{pmatrix} r_k^{4/5} \cos\theta_k \\ r_k^{4/5} \sin\theta_k \end{pmatrix}.$$

Thus our model is

$$\mathbf{e}_k(y) \sim \mathrm{BVN}(0, \boldsymbol{\Sigma}). \qquad (2)$$

Our final BMA model for the predictive PDF of the weather quantity $\mathbf{y}$, here the wind vector, is thus (1) with p_k the distribution of $\mathbf{y}$ implied by (2). We restrict the covariance matrix $\boldsymbol{\Sigma}$ to be equal 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. We found that it led to no degradation in predictive performance.

d. Parameter estimation

Parameter estimation is based on forecast–observation pairs from a training period, which we take here to be the N most recent available days preceding initialization. The training period is a sliding window, and the parameters are reestimated for each new initialization period.

As mentioned above, bias-correction parameters were fit using linear regression. These parameters are member specific and are thus estimated separately for each ensemble member using the observed wind vector as the dependent variable and the forecasted wind vector $\mathbf{f}_k$ as the independent variable.

We estimate the remaining parameters, $w_1, \ldots, w_K$, and $\boldsymbol{\Sigma}$, by maximum likelihood from the training data. Assuming independence of forecast errors in space and time, the log-likelihood function for the BMA model is

$$\ell(w_1, \ldots, w_K; \boldsymbol{\Sigma}) = \sum_{s,t} \log p(\mathbf{y}_{st} \mid \mathbf{f}_{1st}, \ldots, \mathbf{f}_{Kst}),$$

where the sum extends over all station locations s and times t in the training data.

The log-likelihood function 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 iterative, alternating between two steps, the expectation (E) step, and the maximization (M) step. It uses the unobserved quantities z_{kst}, which are latent variables equal to 1 if observation $\mathbf{y}_{st}$ comes from the kth mixture component and to 0 otherwise.

In the E step, the z_{kst} are estimated given the current estimate of the parameters. Specifically,

$$\hat{z}_{kst}^{(j+1)} = \frac{w_k^{(j)} g_k^{(j)}[\mathbf{e}_k(\mathbf{y}_{st})]}{\sum_{l=1}^{K} w_l^{(j)} g_l^{(j)}[\mathbf{e}_l(\mathbf{y}_{st})]},$$

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. The quantity $g_k^{(j)}[\mathbf{e}_k(\mathbf{y}_{st})]$ is defined as the probability density function for $\mathbf{e}_k(\mathbf{y}_{st})$ using the

estimate of $\mathbf{\Sigma}$ from the jth iteration. Note that, although the z_{kst} are equal to either 0 or 1, the $\hat{z}_{kst}$ are real numbers between 0 and 1. The $\hat{z}_{1st}, \ldots, \hat{z}_{Kst}$ are nonnegative and sum to 1 for each (s, t).

The M step then consists of maximizing the expected log-likelihood as a function of $w_1, \ldots, w_K$, and $\mathbf{\Sigma}$, where the expectation is taken over the distribution of z_{kst} given the data and the previous estimates. This is the same as maximizing the log-likelihood given the z_{kst} as well as $w_1, \ldots, w_K$, and $\mathbf{\Sigma}$, evaluated at $z_{kst} = \hat{z}_{kst}^{(j+1)}$. 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). Also,

$$\mathbf{\Sigma}^{(j+1)} = \frac{\sum_{l=1}^{K} S_l^{(j+1)}}{n},$$

where

$$S_l^{(j+1)} = \sum_{s,t} \hat{z}_{lst}^{(j+1)} \mathbf{e}_l(\mathbf{y}_{st}) \mathbf{e}_l(\mathbf{y}_{st})^{\mathrm{T}}.$$

These updated parameter estimates result in an updated likelihood function, which is then used in the subsequent iteration. The E and M steps are iterated to convergence, which we define as a change no greater than some small tolerance in the log-likelihood 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 values based on equal weights and the marginal variance usually led to a good solution in our experience.

e. Forecasting

Once the parameters of the model have been fit, it remains to use the model to generate probabilistic forecasts. As the interest may be in either the forecast of the full wind vector or of some derived quantity such as wind speed or direction, we simulate a large number of forecasts from the distribution. We can then use the empirical distribution of these simulated forecasts, or of any derived quantity from these forecasts, as our forecast distribution. This essentially creates a new, larger, and better-calibrated ensemble of vector wind forecasts.

4. Results

We begin by looking at aggregate results over the entire Pacific Northwest domain for the full 2003 calendar year, with the data available from late 2002 used only as training data, to allow us to create forecasts starting in January. We then examine using the vector forecasts to create forecasts for marginal wind speed and wind direction. Finally, we look at some more specific examples of results for individual locations and/or times.

a. Vector forecast results

In assessing probabilistic forecasts of wind vectors, we aim to maximize the sharpness of the predictive PDFs subject to calibration (Gneiting et al. 2007). Calibration refers to the statistical consistency between the forecast PDFs and the observations. For ensembles of univariate quantities, calibration is often assessed using a verification rank histogram. For multivariate quantities as we have here, the analog is to use the multivariate rank histogram (Gneiting et al. 2008). If the observation and the ensemble members come from the same distribution, then the observed and forecasted values are exchangeable which should result in a flat multivariate rank histogram.

Figure 2 shows these results for our data. The multivariate rank histogram illustrates the lack of calibration in the raw ensemble, which is underdispersed, similarly to results reported by Eckel and Walters (1998), Hamill and Colucci (1998), and Mullen and Buizza (2001) for other ensembles. From the multivariate rank histogram for the BMA forecast distribution, we see that our forecast is substantially better calibrated than the raw ensemble. It has been noted that in some situations, rank histograms are insufficient to accurately assess calibration (Hamill 2001; Gneiting et al. 2007; Marzban et al. 2011; Kolczynski et al. 2011). As such, we consider other results that also examine calibration.

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

$$\mathrm{es}(P, \mathbf{x}) = E_P \|\mathbf{X} - \mathbf{x}\| - \frac{1}{2} E_P \|\mathbf{X} - \mathbf{X}'\|,$$

where P is the predictive distribution, $\mathbf{x}$ is the observed wind vector, and $\mathbf{X}$ and $\mathbf{X}'$ are independent random variables with distribution P (Gneiting et al. 2008). The

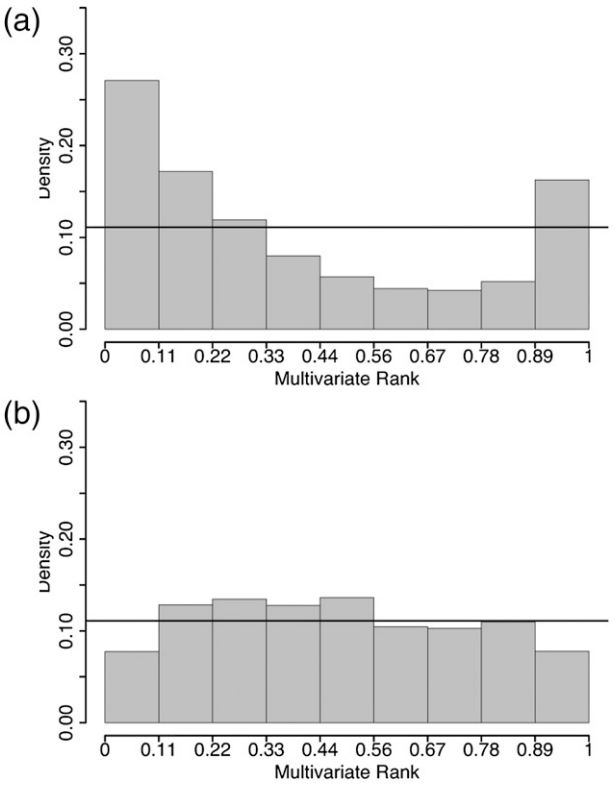

FIG. 2. Calibration checks for probabilistic forecasts of vector wind over the Pacific Northwest in 2003. Multivariate rank histogram for (a) the raw ensemble and (b) the BMA forecast distribution.

energy score is negatively oriented (i.e., the smaller the better).

A point forecast can be created from a forecast distribution by finding the spatial median of the predictive distribution. The spatial median is the unique value that minimizes the multivariate mean absolute error (Milasevic and Ducharme 1987), that is, the mean magnitude of the error vectors. It is the multivariate analog of the median, reducing to the median in the one-dimensional case.

Table 2 shows energy score and multivariate mean absolute error values for climatology, that is, the marginal distribution of observed wind speed across space and time for the dataset, the raw ensemble forecast, and the BMA forecasts, all in units of knots. We see that BMA outperforms both climatology and the raw ensemble in both of these measures.

b. Wind speed forecast results

We can also consider the performance of the marginal forecast distribution for wind speed obtained from this method. In the univariate case, to assess calibration, we consider Fig. 3, which shows the verification rank histogram for the raw ensemble forecast and probability integral transform (PIT) histograms for the BMA

TABLE 2. Mean energy score (ES) and multivariate mean absolute error (MMAE) for probabilistic forecasts of wind vectors over the Pacific Northwest in 2003 (kt). The MMAE refers to the point forecast given by the spatial median of the respective forecast distribution.

Forecast	ES	MMAE
Climatology	5.303	7.187
Ensemble	5.421	6.793
BMA	4.323	6.037

forecast distributions. In both cases, a more uniform histogram indicates better calibration. The verification rank histogram plots the rank of each observed wind speed relative to the eight ensemble member forecasts. If the observation and the ensemble members come from the same distribution, then the observed and forecasted values are exchangeable so that all possible ranks are equally likely. The PIT is the value that the predictive cumulative distribution function attains at the observation and is a continuous analog of the verification rank.

We again see the lack of calibration in the raw ensemble, which is underdispersed. From the PIT histograms for the BMA forecast distribution, we see substantially improved calibration.

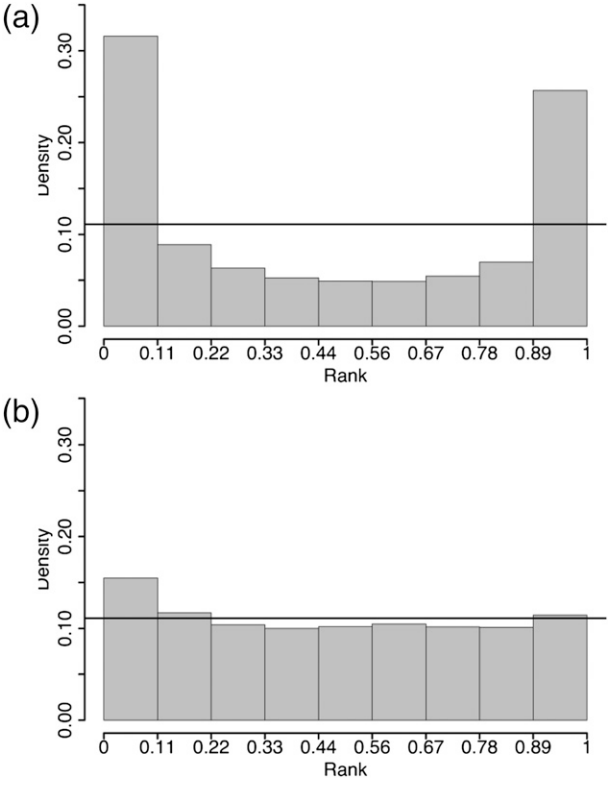

FIG. 3. As in Fig. 2, but for wind speed. Verification rank histogram for (a) the raw ensemble and (b) PIT histogram for the BMA forecast distribution.

TABLE 3. Mean CRPS and MAE, and coverage and average width of 77.8% central prediction intervals for probabilistic forecasts of wind speed over the Pacific Northwest in 2003. Coverage is in percent; all other values in kt. The MAE refers to the point forecast given by the median of the respective forecast distribution.

Forecast	CRPS	MAE	Coverage	Width
Climatology	2.856	3.982	77.8	12.99
Ensemble	3.064	3.713	42.7	5.38
BMA	2.591	3.468	69.4	9.16

If the eight-member raw ensemble were calibrated, there would be a 1/9 probability of the wind speed observation falling below the ensemble range, and a 1/9 probability of it falling above the ensemble range. As such, to allow direct comparisons to the raw ensemble, we will consider 7/9 or 77.8% central prediction intervals from the BMA PDF. Table 3 shows the empirical coverage of 77.8% prediction intervals, and the results echo what we see in the verification rank and PIT histograms. The raw ensemble was highly uncalibrated. The BMA intervals were much better calibrated. The table also shows the average width of the prediction intervals, which characterizes the sharpness of the forecast distributions. While the raw ensemble provides a narrower interval, this comes at the cost of much poorer calibration.

The continuous ranked probability score (CRPS) is the scalar equivalent of the energy score. A generalization of the mean absolute error (MAE), it can be directly compared to the latter. It is a proper scoring rule and is defined as

$$\begin{aligned}\mathrm{crps}(P, x) &= \int_{-\infty}^{\infty} [P(y) - I(y \geq x)]^2 dy \\ &= E_P|X - x| - \frac{1}{2} E_P|X - X'|,\end{aligned}$$

where P is the predictive distribution, here taking the form of a cumulative distribution function; x is the observed wind speed; and X and X' are independent random variables with distribution P (Grimit et al. 2006; Wilks 2006; Gneiting and Raftery 2007). Both CRPS and MAE are negatively oriented (i.e., the smaller the better).

Table 3 shows CRPS and MAE values for climatology, the raw ensemble forecast, and the BMA forecasts, all in units of knots. A point forecast can be created from the BMA forecast distribution by finding the median of the predictive distribution, and the MAE refers to this forecast. Similarly, we show the MAE for the median of the eight-member forecast ensemble, with the results for BMA being by far the best. The results for the CRPS were similar, in that BMA outperformed the raw ensemble and climatology.

c. Wind direction forecast results

We can then also consider the performance of the marginal forecast distribution of wind direction obtained through our method. The verification rank histogram requires an ability to rank the observation relative to the ensemble forecast. For a circular variable, there is no absolute starting or ending point, and so we must in some fashion choose a starting point in order to be able to define a rank. While it would be possible to choose an arbitrary direction, say 0°, as the starting point, the histogram would no longer have the easy interpretability that it has in other situations. To retain the ability to interpret over- and underdispersion from these histograms, we wish to define a starting point that captures a sense of the ''inside'' and ''outside'' of the ensemble range.

We propose doing so by first finding the directional mean (Fisher 1993). This seems a reasonable candidate to be considered inside the ensemble range. Because we are effectively testing an assumption of exchangeability with the observation and the ensemble, this directional mean is calculated using both the observation and the ensemble. We then take a point 180° opposite the directional mean, which should be outside the ensemble range, and consider that to be our starting point. We then count counterclockwise to obtain the rank of the observation relative to the ensemble. Thus observations ''below'' the majority of the ensemble member forecasts will have a low rank, and observations ''above'' the majority of the ensemble member forecasts will have a high rank.

Figure 4 shows these histograms for both the raw ensemble and a simulated BMA ensemble. As in the earlier instances we see a more uniform histogram for the simulated BMA ensemble, indicating better calibration of the BMA forecast than that of the raw ensemble.

The directional CRPS is an analog of the CRPS that can be applied to circular variables (Grimit et al. 2006). It too is a proper scoring rule and is defined as

$$\mathrm{dcrps}(P, \theta) = E_P \alpha(\Theta, \theta) - \frac{1}{2} E_P \alpha(\Theta, \Theta'),$$

where $\alpha(\theta, \theta')$ denotes the angular distance between any two directions θ and θ' on the circle $[-\pi, \pi)$ and where Θ and Θ' are independent random variables with common circular probability distribution P.

Point forecasts can be obtained by calculating the directional median (Fisher 1993), and they can be assessed by considering the mean directional error. Table 4 compares these results from BMA to the raw ensemble, and we again see improved performance from the BMA forecast.

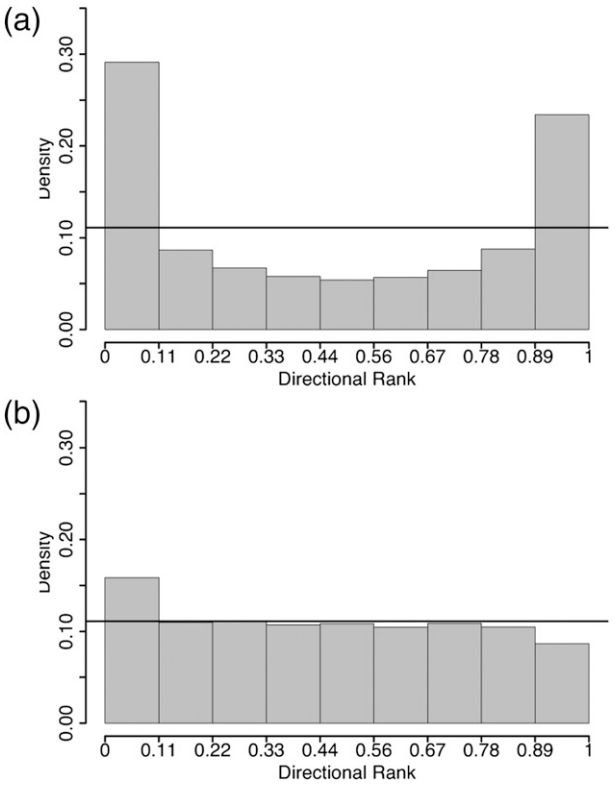

FIG. 4. Calibration checks for probabilistic forecasts of wind direction over the Pacific Northwest in 2003. Directional verification rank histogram for (a) the raw ensemble and (b) the BMA forecast distribution.

d. Example

To illustrate the BMA forecast distributions for wind vectors, we show an example on 4 February 2003, at Omak, Washington. Table 5 gives the parameters of the BMA density at this location for this day. Figure 5 shows a contour plot of the BMA forecast density, as well as the ensemble member forecasts and the observed wind vector. We can see that the observation is outside of the range of the raw ensemble, as well as being slightly outside of the range of the bias-corrected ensemble, but well within the BMA density.

We can examine some of the details of the parameters in this instance. Figure 5 has the BMA weights labeled next to each of the ensemble member forecasts. As one would expect, the density is largely centered around the forecasts with higher weights.

We then examine the bias-correction parameters. There is first an additive bias-correction vector. In this case, we see that the first component, corresponding to east–west direction, is very close to zero in all cases, while the second component is consistently negative, suggesting a small northerly bias to the forecasts. We then decompose the transformation matrix $\mathbf{B}_k$ in terms of skew, squeeze, rotation, and scaling, for interpretability. For a matrix of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we first define

$$f = \frac{1}{a^2 + c^2}.$$

The skew can then be calculated as

$$n = (ab + cd)f.$$

The squeeze can be calculated as

$$t = \frac{1}{\sqrt{f(ad - bc)}}.$$

The scaling can be calculated as

$$r = \sqrt{(ad - bc)}.$$

The rotation can then be calculated as

$$w = \arccos\frac{a}{rt}.$$

We see a negative skew—this indicates that the y axis is being transformed from a vertical line to a line with a negative slope. The squeeze values indicate a change in the relative dimensions along the x and y axes—in this case, they are close to one, suggesting very little change in the relative dimensions. The rotation is an angular rotation of the forecast vector; here we see a small but relatively consistent counterclockwise rotation. Finally we see that the scaling is less than one, indicating that the bias correction is decreasing the magnitude of the forecast vectors.

We additionally compare the BMA forecasts at Omak over the 2003 calendar year to the raw ensemble forecast. Table 6 shows energy scores and multivariate mean absolute errors for climatology (here a station-specific climatology considering only the marginal distribution of observed values at this station over the entire data period), the raw ensemble, and BMA at this location.

TABLE 4. Mean directional continuous ranked probability score (DCRPS) and mean directional error (MDE) for probabilistic forecasts of wind direction over the Pacific Northwest in 2003 (°).

Forecast	DCRPS	MDE
Climatology	30.24	45.42
Ensemble	36.76	46.86
BMA	28.82	43.29

TABLE 5. BMA forecast parameters for 4 Feb 2003.

	GFS	GEM	ETA	GASP	JMA	NOGAPS	TCWB	UKMO
Weight	0.33	0.11	0.01	0.00	0.04	0.17	0.07	0.27
Additive bias	0.059	−0.074	−0.053	0.024	−0.048	0.049	−0.061	0.013
Correction	−0.323	−0.174	−0.185	−0.269	−0.208	−0.240	−0.140	−0.343
Skew	−0.321	−0.300	−0.265	−0.291	−0.223	−0.326	−0.268	−0.225
Squeeze	0.944	0.932	0.939	0.935	0.914	0.945	0.980	0.935
Rotation	7.24	7.26	7.59	12.57	11.17	6.00	8.69	12.81
Scaling	0.448	0.450	0.422	0.409	0.429	0.434	0.401	0.423

The BMA forecast showed substantially better scores than the raw ensemble. Furthermore, the multivariate rank histograms in Fig. 6 show that the BMA forecast distribution was substantially better calibrated than the raw ensemble.

e. Weights

There was substantial variability in the weights over time. Figure 7 shows the weights over time. On average, the most highly weighted ensemble member was the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS; with an average weight of 0.32), followed by the Met Office (UKMO; average weight of 0.23), then the ETA Model (0.10), Canadian Meteorological Centre (CMC) Global Environmental Multiscale Model (GEM) (0.08), Japan Meteorological Agency (JMA; 0.08), Taiwan Central Weather Bureau Operational Model (TCWB) (0.07), Navy Operational Global Atmospheric Prediction System (NOGAPS) (0.06), and Australian Bureau of Meteorology (BoM) Global Analysis and Prediction model (GASP) (0.06).

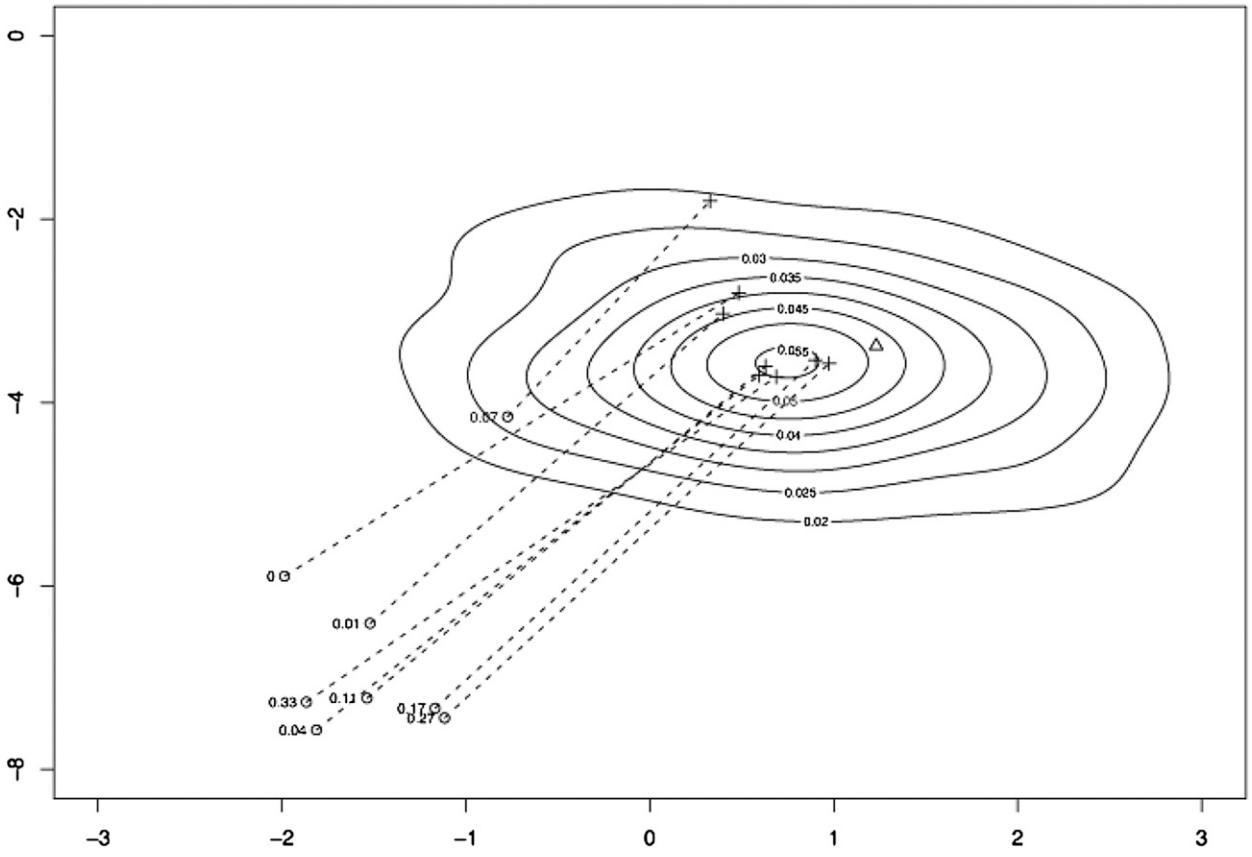

FIG. 5. BMA forecast density for 4 Feb 2003, at Omak, WA. Contours represent the BMA density, the small circles are the raw ensemble forecasts, the crosses are the bias-corrected ensemble forecasts, and the triangle is the observed wind vector. The dashed lines connect the raw ensemble forecasts to the corresponding bias-corrected forecasts.

TABLE 6. Mean ES and MMAE for probabilistic forecasts of wind vectors at Omak, WA, in 2003 (kt). The MMAE refers to the point forecast given by the spatial median of the respective forecast distribution.

Forecast	ES	MMAE
Climatology	5.341	7.202
Ensemble	5.335	6.773
BMA	4.256	5.974

f. Comparison to independent BMA model

To further demonstrate the value of this new multivariate BMA model, we compare the results to what could have been achieved using existing methodologies. We construct independent BMA models for the U and V components of the wind vector, each using a mixture of normal distributions, similar to the models developed by Raftery et al. (2005). By simulating from these independently modeled U and V components, we create an empirical BMA distribution for wind vectors that does not use a multivariate modeling approach. We show here that such an approach, while improving over the raw ensemble, does not provide the same quality of forecasts that the multivariate BMA model provides.

As discussed above, Table 2 shows energy score and multivariate mean absolute error values for climatology, the raw ensemble forecast, and the multivariate BMA forecasts, all in units of knots. For the independent BMA model, the mean energy score was 4.430 and the multivariate mean absolute error was 6.209. We see that the multivariate BMA outperforms both the raw ensemble and the independent BMA in both of these measures. Figure 2, as previously discussed, shows the multivariate rank histograms for the raw ensemble and the multivariate BMA forecasts. Figure 8 shows that the independent BMA forecast, while better calibrated than the raw ensemble, is not as well calibrated as the multivariate BMA forecast.

5. Discussion

We have shown how to apply BMA to wind vector forecasts. This provides a statistical postprocessing method for ensembles of numerical weather predictions that yields a full predictive distribution for wind vectors.

In our experiments with the University of Washington mesoscale ensemble, the BMA forecast PDFs were better calibrated than the raw ensemble, which was underdispersive. The BMA spatial median forecast had lower multivariate mean absolute error than the ensemble spatial median, and the BMA forecast PDFs had a substantially lower energy score than the raw ensemble or climatology. We also saw that the marginal

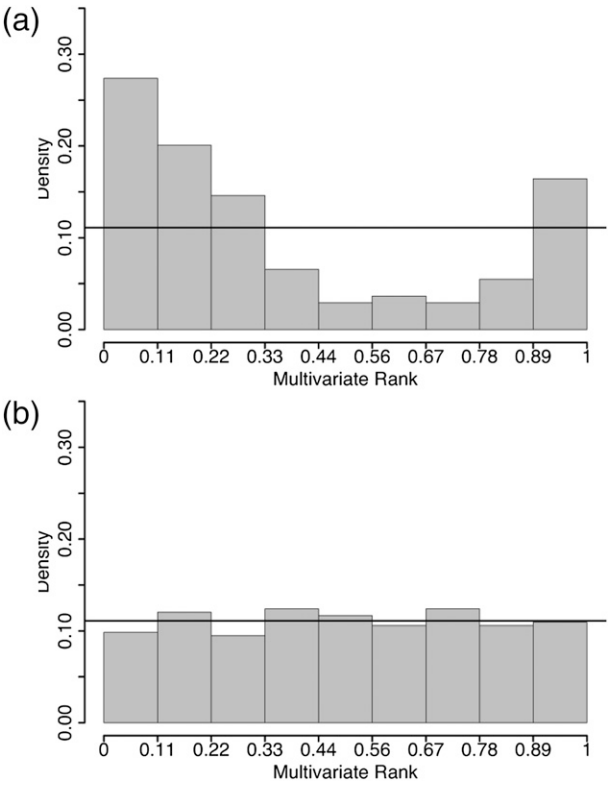

FIG. 6. As in Fig. 2, but at Omak, WA, in 2003.

probabilistic forecasts for both wind speed and wind direction that can be obtained from the BMA forecast outperformed the raw ensemble forecasts.

While our implementation has been for a situation where the ensemble members come from clearly distinguishable sources, it can easily be modified to deal with situations in which ensemble members come from the same model, differing only in some random perturbations. Examples include the global ensembles that are currently used by the National Centers for Environmental Prediction and the European Centre for Medium-Range Weather Forecasts (Buizza et al. 2005). In these cases, members coming from the same source should be treated as exchangeable, and thus should have equal weight and equal BMA parameter values across members. This would be analogous to the method described by Raftery et al. (2005, p. 1170), Wilson et al. (2007), and Fraley et al. (2010).

Our method produces forecasts at individual locations, which is the focus of many, and possibly most applications. As a result we have not had to model spatial correlation between observed values, although these are definitely present. In applications that involve forecasting at more than one location simultaneously, it would be vital to take account of spatial correlation. These include

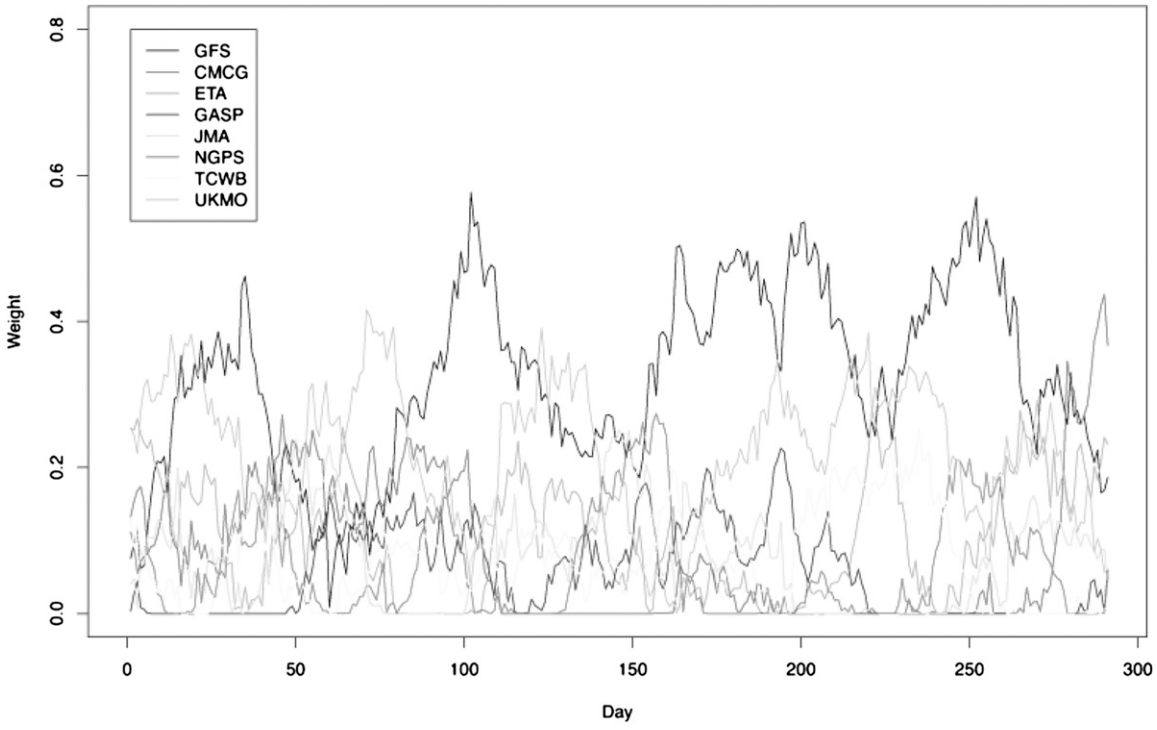

FIG. 7. BMA ensemble member weights across time.

forecasting maximum wind speeds over an area or trajectory, for example for shipping or boating, and forecasting the total energy from several wind farms in a region. Methods for probabilistic weather forecasting at multiple locations simultaneously have been developed for temperature (Gel et al. 2004; Berrocal et al. 2007), for precipitation (Berrocal et al. 2008), and for temperature and precipitation simultaneously (Berrocal et al. 2010). These methods could potentially be extended to variables such as wind speed and wind vectors. Schuhen et al. (2012) discuss the ensemble copula coupling approach to forecasting wind vectors, which can model spatial correlation in forecasts.

Spatial information could also improve forecasts by taking into consideration the differing statistical relationships between forecasts and observations at different locations. Our method estimates a single set of parameters across the entire domain. Nott et al. (2001) noted that localized statistical postprocessing can address issues of locally varying biases in numerical weather forecasts. Localized versions of BMA for temperature, based on modeling the spatial variation in the bias of the forecasts or on taking sets of forecasts and observations within a carefully selected neighborhood, have shown substantial improvements over the global version (Kleiber et al. 2011), and it seems plausible that similar improvements would be seen for a localized version of BMA for wind vectors.

We have shown how to extend BMA to model multivariate quantities whose errors have an ellipsoidal distribution. While we have only investigated this for the particular case of wind vectors, it seems likely that this family of power-transformed multivariate normal distributions could provide the flexibility to model a variety of multivariate quantities whose error distributions are ellipsoidal. Additionally, this method could be reduced to the univariate case, where power-transformed normal distributions could potentially be used to model quantities whose error distributions are symmetric but nonnormal.

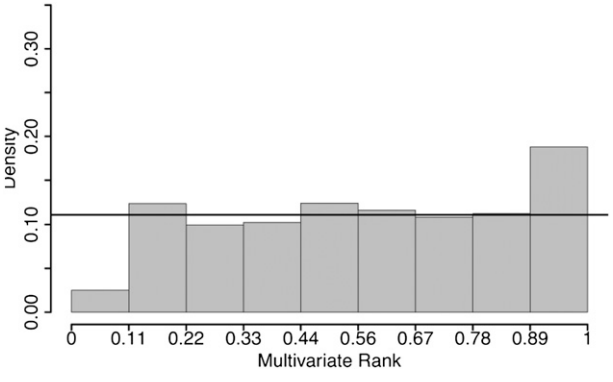

FIG. 8. Calibration checks for probabilistic forecasts of vector wind over the Pacific Northwest in 2003 for the independent BMA forecasts.

One general open modeling question is how to model multivariate quantities in which the error distribution is not ellipsoidal. This issue may arise in other wind vector situations, or in situations where a joint distribution of two or more skewed weather quantities is of interest (such as modeling a joint distribution for wind speed and pressure, or temperature and precipitation). Other work has investigated modeling wind vectors using a bivariate skew-*t* distribution (Hering and Genton 2008), and a BMA method could be developed using these as the component distributions, or possibly by fitting each component as a mixture of multiple multivariate normal distributions. Pinson (2012) has proposed an alternative method that, unlike what we do here, does not fit probability distributions based on the ensembles. Instead, he carries out a two-dimensional translation and dilation of the set of ensemble forecasts based on an estimated bivariate Gaussian model, to yield an ensemble of the same nature as the original one, but that is calibrated. Schuhen et al. (2012) propose an approach to forecasting wind vectors via ensemble model output statistics.

Acknowledgments. The authors are grateful to Jeff Baars, Chris Fraley, Eric Grimit, and Clifford F. 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, by the National Science Foundation under Awards ATM-0724721 and DMS-0706745, by the Joint Ensemble Forecasting System (JEFS) under Subcontract S06-47225 from the University Corporation for Atmospheric Research (UCAR), and by the Center for Statistics and the Social Sciences at the University of Washington.

REFERENCES

Baars, J., cited 2005: Observations QC summary page. [Available online at http://www.atmos.washington.edu/mm5rt/qc_obs/qc_obs_stats.html.]

Bao, L., T. Gneiting, E. P. Grimit, P. Guttorp, and A. E. Raftery, 2010: Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Wea. Rev.,* **138,** 1811–1821.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.,* **135,** 1386–1402.

——, ——, and ——, 2008: Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.,* **2,** 1170–1193.

——, ——, and ——, 2010: Probabilistic weather forecasting for winter road maintenance. *J. Amer. Stat. Assoc.,* **105,** 522–537.

Bremnes, J. B., 2004: Probabilistic wind power forecasts using local quantile regression. *Wind Energy,* **7,** 47–54.

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

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

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

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

Fisher, N. I., 1993: *Statistical Analysis of Circular Data.* Cambridge University Press, 295 pp.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.,* **138,** 190–202.

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

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

Gneiting, T., 2011: Quantiles as optimal point forecasts. *Int. J. Forecasting,* **27,** 197–207.

——, and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science,* **310,** 248–249.

——, and ——, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

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

——, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion). *TEST,* **17,** 211–264.

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.

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

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

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

Hering, A. S., and M. G. Genton, 2008: Powering up with space-time wind forecasting. *J. Amer. Stat. Assoc.,* **105,** 92–104.

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–417. [Available online at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.]

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.,* **139,** 2630–2649.

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, N. S. Altman, and A. Deng, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.,* **139,** 3954–3963.

Leamer, E. E., 1978: *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* John Wiley & Sons, 370 pp.

Madigan, D., and A. E. Raftery, 1994: Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Stat. Assoc.,* **89,** 1335–1346.

Marzban, C., R. Wang, F. Kong, and S. Leyton, 2011: On the effect of correlations on rank histograms: Reliability of temperature and wind speed forecasts from finescale ensemble reforecasts. *Mon. Wea. Rev.,* **139,** 295–310.

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions.* Wiley-Interscience, 304 pp.

Milasevic, P., and G. R. Ducharme, 1987: Uniqueness of the spatial median. *Ann. Stat.,* **15,** 1332–1333.

Møller, J. K., H. A. Nielsen, and H. Madsen, 2008: Time-adaptive quantile regression. *Comput. Stat. Data Anal.,* **52,** 1292–1303.

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.

Nielsen, H. A., H. Madsen, and T. S. Nielsen, 2006: Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy,* **9,** 95–108.

Nott, D. J., W. T. M. Dunsmuir, R. Kohn, and F. Woodcock, 2001: Statistical correction of a deterministic numerical weather prediction model. *J. Amer. Stat. Assoc.,* **96,** 794–804.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128,** 747–774.

Pinson, P., 2012: Adaptive calibration of (*u, v*)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **138,** 1273–1284, doi: 10.1002/qj.1873.

——, and H. Madsen, 2009: Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy,* **12,** 137–155.

——, C. Chevallier, and G. N. Kariniotakis, 2007a: Trading wind generation with short-term probabilistic forecasts of wind power. *IEEE Trans. Power Syst.,* **22,** 1148–1156.

——, H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis, 2007b: Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy,* **10,** 497–516.

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.

Roulston, M. S., D. T. Kaplan, J. Hardenberg, and L. A. Smith, 2003: Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy,* **28,** 585–602.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.,* **140,** 3204–3219.

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

——, T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.,* **105,** 25–35.

Thorarinsdottir, T., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics using heteroskedastic censored regression. *J. Roy. Stat. Soc.,* **173A,** 371–388.

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

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Reply. *Mon. Wea. Rev.,* **135,** 4231–4236.

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