# Spread Calibration of Ensemble MOS Forecasts

BRUCE A. VEENHUIS

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

(Manuscript received 10 July 2012, in final form 18 October 2012)

ABSTRACT

Ensemble forecasting systems often contain systematic biases and spread deficiencies that can be corrected by statistical postprocessing. This study presents an improvement to an ensemble statistical postprocessing technique, called ensemble kernel density model output statistics (EKDMOS). EKDMOS uses model output statistics (MOS) equations and spread–skill relationships to generate calibrated probabilistic forecasts. The MOS equations are multiple linear regression equations developed by relating observations to ensemble mean-based predictors. The spread–skill relationships are one-term linear regression equations that predict the expected accuracy of the ensemble mean given the ensemble spread. To generate an EKDMOS forecast, the MOS equations are applied to each ensemble member. Kernel density fitting is used to create a probability density function (PDF) from the ensemble MOS forecasts. The PDF spread is adjusted to match the spread predicted by the spread–skill relationship, producing a calibrated forecast. The improved EKDMOS technique was used to produce probabilistic 2-m temperature forecasts from the North American Ensemble Forecast System (NAEFS) over the period 1 October 2007–31 March 2010. The results were compared with an earlier spread adjustment technique, as well as forecasts generated by rank sorting the bias-corrected ensemble members. Compared to the other techniques, the new EKDMOS forecasts were more reliable, had a better calibrated spread–error relationship, and showed increased day-to-day spread variability.

## 1. Introduction

Ensemble forecasting systems benefit operational forecasters by predicting a range of possible forecast outcomes. An ensemble contains multiple members, each a separate run of a numerical weather prediction (NWP) model. With a well-constructed ensemble, the ensemble mean is on average more accurate than the individual members (Leith 1974). Likewise, the spread of the ensemble may correlate with the expected accuracy of the ensemble mean (Kalnay and Dalcher 1987; Whitaker and Loughe 1998). The ensemble members sample the various sources of error that degrade NWP forecasts. To quantify the error in the underlying analysis, the ensemble members are initialized with perturbed initial conditions. Over the years, a range of perturbation techniques have been proposed including the breeding method (Toth and Kalnay 1997), singular vectors (Palmer et al. 1998), and the ensemble transform Kalman filter (Wang and Bishop 2003). The numerical model itself is also a source of error because it must parameterize subgrid-scale processes and only approximates the true atmosphere. Thus, to quantify the numerical model uncertainty, the ensemble members may use the same parameterization schemes with varied constants or have stochastic parameterizations (Houtekamer et al. 1996; Buizza et al. 1999).

Many operational meteorological centers run ensembles, including the National Centers for Environmental Prediction's (NCEP) Global Ensemble Forecast System (GEFS; Toth et al. 2001), the Short-Range Ensemble Forecast System (SREF; McQueen et al. 2005), the Canadian Meteorological Centre's (CMC) ensemble (Charron et al. 2010), the European Centre for Medium-Range Weather Forecasts' (ECMWF) ensemble (Buizza 1997), and the U.S. Navy's Fleet Numerical Meteorology and Oceanography Center's (FNMOC) ensemble (Peng et al. 2004).

Unfortunately, there are often systematic errors in the near-surface weather elements predicted directly by

*Corresponding author address:* Bruce Veenhuis, Meteorological Development Laboratory, Office of Science and Technology, NOAA/National Weather Service, Room 10400, 1325 East-West Hwy., Silver Spring, MD 20910.
E-mail: bruce.veenhuis@noaa.gov

DOI: 10.1175/MWR-D-12-00191.1

ensembles. In addition, the spread of the ensemble members is often too small and underestimates the true forecast error. To correct these deficiencies, numerous statistical postprocessing techniques have been suggested, including ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005), Bayesian Model Averaging (BMA; Raftery et al. 2005; Wilson et al. 2007), Nonhomogeneous Gaussian Regression (NGR; Gneiting et al. 2005; Wilks 2006; Wilks and Hamill 2007), ensemble regression (Unger et al. 2009), variance inflation techniques (Johnson and Bowler 2009), and shift-and-stretch calibration (Eckel et al. 2012). The Meteorological Development Laboratory (MDL) of the National Weather Service (NWS) has used a technique called model output statistics (MOS; Glahn and Lowry 1972) for decades to statistically postprocess numerical models. Typically, MOS has been applied to output from a single numerical model. More recently, MDL developed a MOS-based technique called ensemble kernel density MOS (EKDMOS; Glahn et al. 2009a) which can generate calibrated probabilistic forecasts from ensembles.

Calibrating the ensemble spread is important for ensemble postprocessing. The original EKDMOS methodology explained by Glahn et al. (2009a) used an empirical spread-adjustment factor to calibrate the final predicted spread. Overall, the technique produced statistically reliable results; however, the spread calibration lacked station specificity as the same spread adjustment factor was applied to all stations. In addition, the method did not attempt to leverage a possible spread–skill relationship between the ensemble spread and the expected accuracy of the ensemble mean to conditionally calibrate the spread. Many studies have found that ensembles do contain a spread–skill relationship (Kalnay and Dalcher 1987; Buizza et al. 2005; Whitaker and Loughe 1998; Barker 1991). That is, the spread of the ensemble positively correlates with the expected error of the ensemble mean. Some postprocessing techniques explicitly account for a spread–skill relationship. For example, NGR models the predicted error variance as a linear function of the ensemble member variance. Grimit and Mass (2007) proposed a technique that modeled the spread–skill relationship in a probabilistic sense. They used a stochastic model to simulate ensemble forecasts and binned the forecasts by ensemble spread. For each bin, they computed the corresponding ensemble mean standard error and showed there was a linear relationship between the ensemble member standard deviation and the ensemble mean standard error. They suggested a linear function could be fit to the data and used to calibrate future forecast. Other studies have applied similar spread-dependent calibration techniques to a variety of weather elements including 2-m temperature (Eckel et al. 2012) and upper-level winds (Kolczynski et al. 2009).

Building upon the work of earlier studies, we have improved the EKDMOS method by including a spread–skill relationship within the multiple linear regression framework. In keeping with the Glahn et al. (2009a) methodology, we develop MOS equations with ensemble mean-based predictors. In addition, we develop spread–skill relationships that conditionally calibrate the predicted spread. We model the spread–skill relationship with a simple one-term linear regression equation that uses the ensemble spread to predict the expected accuracy of the ensemble mean. We have found we can improve the linearity of the relationship between spread and the expected error by applying variance stabilizing transforms (see Box et al. 2005, p. 320) to our predictor and predictand. Our technique resembles NGR in that we model the predictive error variance as a linear function of the ensemble spread. However, unlike NGR, we fit the statistical model analytically rather than iteratively. Our method also avoids the Grimit and Mass (2007) approach's requirement to bin and sort data.

The remainder of this paper is organized as follows: in section 2 we discuss the numerical models and observational datasets used to perform this work. In section 3 we give an overview of EKDMOS and explain our improved methodology. In section 4 we present verification results for independent cross-validated data. Finally, section 5 finishes with a discussion and conclusions.

## 2. Data

We used numerical model forecasts from the North American Ensemble Forecast System (NAEFS; Candille 2009). NAEFS is the suite of 42 ensemble-member forecasts that combines the CMC's Global Environment Multiscale Model (GEM) and NCEP's GEFS. The operational centers distribute a matched set of model output fields from their respective ensembles on the same $1^\circ \times 1^\circ$ latitude–longitude grid. The CMC ensemble has a control run and 20 members that use varied physical parameterizations. Table 1 of Charron et al. (2010) lists additional details regarding the operational configuration of the CMC ensemble. The GEFS also has a control and 20 perturbation members, but unlike the CMC ensemble, the ensemble members use an identical set of physical parameterizations (Zhu et al. 2007). MDL maintains an archive of operational NAEFS forecasts from July 2007 to the present.

We used station-based observations from an archive maintained by MDL. We applied the EKDMOS technique

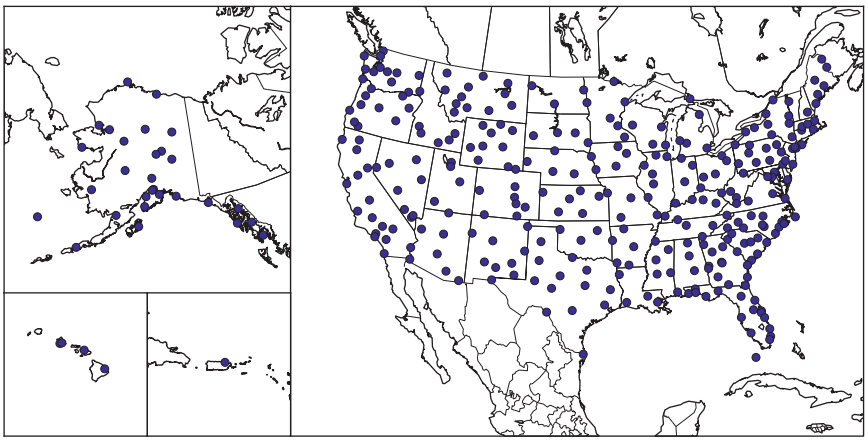

FIG. 1. Map of the 335 stations used for verification.

to 2303 stations distributed throughout the conterminous United States (CONUS), Canada, Alaska, Hawaii, and Puerto Rico. For testing, we generated 2-m temperature forecasts, which we verified at a subset of 335 stations that MDL has judged to have reliable observations. Figure 1 shows a map of these stations.

## 3. EKDMOS method

For convenience, we present here a brief overview of the EKDMOS technique that is described in full by Glahn et al. (2009a). Separate sets of MOS equations were created for the GEFS and CMC with ensemble mean-based predictors. MOS uses forward screening multiple linear regression to relate observations (i.e., predictands) to model-based predictors. Typically, the regression selects model fields closely related to the predictand. For example, for hourly 2-m temperature, the most common predictors were model forecasts of 2-m temperature and geopotential thickness. Harmonics, such as the cosine and sine of the day of the year, were also offered as predictors, and became important in the later projections. MOS equations were developed for each projection, station, cycle, and element. We applied the MOS equations to each ensemble member to generate 42 ensemble MOS forecasts. We constructed a probability density function (PDF) from the 42 ensemble MOS forecasts with kernel density fitting (Wilks 2011). Following Glahn et al. (2009a) we used Gaussian kernels with a standard deviation equal to the standard error estimate predicted by the MOS equation. At this point, the spread of the PDF from kernel density fitting is uncalibrated and will be adjusted later.

As mentioned previously, we developed MOS equations with ensemble mean-based predictors and applied those equations to the ensemble members. Unger et al. (2009) argues that this is theoretically sound if the ensemble member forecasts represent equally likely outcomes. For each forecast case, one ensemble member will be ''best''; however, it is impossible to identify the best member beforehand. Unger et al. (2009) demonstrates that the expected coefficients for the best member regression equation are identical to those obtained by developing a regression equation with the ensemble means as predictors. Since the GEFS is composed of random perturbations of the same model, the assumption of equally likely members seems plausible. However, as the CMC members contain different model physics and parameterizations, greater caution must be exercised. We tested developing 2-m temperature MOS equations for each CMC ensemble member versus developing CMC ensemble mean-based equations. Examining three years of independent cross-validated results, we found that forecasts generated by applying the mean-based equations to each member were more accurate, as measured by mean absolute error (MAE), at all projections. Therefore, for both GEFS and CMC, we develop MOS equations with the ensemble means and apply those equations to the members.

### a. Original Spread-Adjustment technique

For clarity, we refer to the original second moment calibration technique, described by Glahn et al. (2009a), as Spread-Adjustment. To review, the Spread-Adjustment technique computes a spread-adjustment factor $x$ given by

$$x = \frac{3(\sigma_{\min} + \sigma_{\max}) + \mathrm{SF}(F_{\max} - F_{\min})}{3(\sigma_{\min} + \sigma_{\max}) + (F_{\max} - F_{\min})}, \qquad (1)$$

where $F_{\min}$ and $F_{\max}$ are the smallest and largest ensemble member MOS forecasts, respectively; $\sigma_{\min}$ and

$\sigma_{\max}$ are the associated standard errors predicted by the MOS equations, respectively; and SF is an empirical spread-adjustment factor. The standard deviation of the final PDF will differ from the original by a factor of $(1 - x)$. Glahn et al. (2009a) found that the kernel density fitting technique produced overdispersive PDFs. They performed testing with dependent EKDMOS GEFS forecasts and found that a spread-adjustment factor of SF = 0.5 optimized reliability by reducing the spread. Our own testing with dependent EKDMOS NAEFS forecasts suggested a spread-adjustment factor of SF = 0.4 was optimal. The smaller spread-adjustment factor indicates the NAEFS required more spread reduction following the kernel density fitting compared to the GEFS. For this study, we generated a set of EKDMOS NAEFS forecasts with a spread adjustment factor of SF = 0.4, which we refer to as Spread-Adjustment.

### b. EKDMOS Spread-Skill calibration

Rather than use the spread-adjustment factor from (1) we instead develop station-specific spread–skill relationships to calibrate the final forecast PDF. Following the general formulation for multiple linear regression, MOS fits the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \qquad (2)$$

where for each forecast case $i$, the predictand $y_i$ is related to a set of $k$ predictors $x_{1i} \ldots x_{ki}$ via the coefficients $\beta_0 \ldots \beta_k$, with residual error $\varepsilon_i$. In matrix form, (2) may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (3)$$

where $\mathbf{Y}$ is the vector of predictand values and $\mathbf{X}$ is the design matrix containing the predictors. The least squares estimate of the regression coefficients vector $\boldsymbol{\beta}$ can be obtained via

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \qquad (4)$$

The vector of fitted values $\hat{\mathbf{Y}}$ is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \qquad (5)$$

and the vector of residual error values $\mathbf{e}$ by

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}. \qquad (6)$$

The error variance estimate of the regression residuals is given by

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k - 1}, \qquad (7)$$

where $n$ is the sample size. If we assume the regression residuals are normally distributed, the standard error (s.e.) of a future response is given by

$$\text{s. e.}\,(\hat{y} \mid \mathbf{x}^* + \varepsilon) = \hat{\sigma}\sqrt{1 + \mathbf{x}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^*}. \qquad (8)$$

Here, $\hat{\sigma}$ is the regression residual standard error from (7) and $\mathbf{x}^*$ is the vector of current predictor values. Equation (8) could be used to calibrate the forecast PDF; however, in practice, the predicted standard error from (8) varies little day-to-day because MOS typically uses large development samples containing several years of data. Likewise, the spread predicted by (8) is not a function of the ensemble spread.

We should note that if the residual errors are not normally distributed, the MOS equation still provides the least squares estimate of a future response given the current predictor values, but the predicted standard error from (8) may not yield statistically reliable forecasts. Our experience with 2-m temperature suggests assuming normally distributed error is reasonable at most stations.

To include a spread–skill relationship, we perform a second regression step and model the regression residual error in (6) as a function of the ensemble spread. As mentioned previously, we apply the MOS equations to each ensemble member to obtain 42 ensemble MOS forecasts. For each case $i$, we calculate the standard deviation of the 42 ensemble MOS forecasts $s_i$ as

$$s_i = \sqrt{\frac{\sum_{k=1}^{K} (f_{ki} - \overline{f}_i)^2}{K - 1}}, \qquad (9)$$

where $f_{ki}$ is the MOS forecast for member $k$, $\overline{f}_i$ is the mean of the ensemble MOS forecasts, and $K$ is the number of ensemble members. We have found that if we apply variable transformations to $s_i$ and the residual errors $\varepsilon_i$ we can model the spread–skill relationship with linear regression. Specifically, we compute the square root of the absolute error of the ensemble mean MOS forecast $\sqrt{|\varepsilon_i|}$ and the square root of the standard deviation of the ensemble MOS forecasts $\sqrt{|s_i|}$. We specify a linear relationship between $\sqrt{|\varepsilon_i|}$ and $\sqrt{|s_i|}$ according to

$$\sqrt{|\varepsilon_i|} = \alpha_0 + \alpha_1\sqrt{s_i}. \qquad (10)$$

The term $\sqrt{|\varepsilon_i|}$ is simply an alternative measure of the ensemble mean MOS forecast accuracy for case $i$. To estimate the parameters $\alpha_0$ and $\alpha_1$, we define $\mathbf{S}$ to be the vector of transformed ensemble MOS standard deviations:

$$\mathbf{S} = \begin{pmatrix} \sqrt{s_1} \\ \vdots \\ \sqrt{s_n} \end{pmatrix}, \tag{11}$$

$\mathbf{E}$ to be the vector of transformed ensemble mean MOS forecast errors:

$$\mathbf{E} = \begin{pmatrix} \sqrt{|\varepsilon_1|} \\ \vdots \\ \sqrt{|\varepsilon_n|} \end{pmatrix}, \tag{12}$$

and $\boldsymbol{\alpha}$ to be the vector of coefficients to estimate

$$\boldsymbol{\alpha} = (\alpha_0 \quad \alpha_1). \tag{13}$$

Here, $n$ is the number of forecast cases. We use linear regression to estimate $\boldsymbol{\alpha}$ according to

$$\hat{\boldsymbol{\alpha}} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}\mathbf{E}. \tag{14}$$

The relationship defined by the parameters $\hat{\boldsymbol{\alpha}}$ is called a spread–skill relationship because it relates the spread of the ensemble MOS to the expected accuracy of the ensemble mean MOS forecast. The spread–skill relationship will predict the expected value of the square root of the absolute error, Expected$[\sqrt{|\varepsilon|}]$; however, to use the spread–skill relationship within the regression framework we wish to know the expected standard error $\sigma$. We can find an approximate back transformation to convert from Expected$[\sqrt{|\varepsilon|}]$ to $\sigma$ if we assume the errors $\varepsilon$ are normally distributed with mean 0 and variance $\sigma^2$. We use the general formula for expected value to write

$$\text{Expected}[\sqrt{|\varepsilon|}] = \int_{-\infty}^{+\infty} \frac{\sqrt{|\varepsilon|}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) d\varepsilon. \tag{15}$$

The integral on the right-hand side of (15) cannot be solved analytically; however, we can substitute trial values for $\sigma$ and numerically evaluate the integral to obtain values of Expected$[\sqrt{|\varepsilon|}]$. This allows us to build a lookup table to perform the back transformation. The lookup table is general and can be used for any element with normally distributed errors.

Using the subscript SS to denote Spread-Skill we can write

$$\sqrt{|\varepsilon_{\text{SS}}|} = \hat{\alpha}_0 + \hat{\alpha}_1\sqrt{s_{\text{Mem}}}, \tag{16}$$

and

$$\hat{\sigma}_{\text{SS}} = G(\sqrt{|\varepsilon_{\text{SS}}|}), \tag{17}$$

where $G$ is a function, in this case a lookup table, that performs the back transformation.

Substituting $\hat{\sigma}_{\text{SS}}$ for $\hat{\sigma}$ in (8) yields

$$\text{s.e.}\,(\hat{y}\,|\,\mathbf{x}^* + \varepsilon) = \hat{\sigma}_{\text{SS}}\sqrt{1 + \mathbf{x}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^*}. \tag{18}$$

Now, the ensemble spread influences the predictive standard error via the spread–skill relationship. When generating a forecast, we adjust the spread of the final PDF so that its spread equals the standard error estimate from (18). We refer to these forecasts as Spread-Skill.

In practice, we enact several quality control criteria to ensure that reasonable spread–skill relationships are developed. First, we require the slope parameter $\hat{\alpha}_1$ in (16) to be positive. Second, we use an $F$ test to test the hypothesis that $\hat{\alpha}_1$ is significantly different from 0 and require the resulting $p$ value to be less than 0.25. If the spread–skill relationship fails either criteria we reject it and use the standard error estimate from (8).

### c. *Justification for variable transformations*

The reason for applying the square root transformation above is, perhaps, not immediately clear. For simplicity, we wish to fit the spread–skill relationship with standard linear regression; however, to be optimal, the regression theory assumes the errors are normally distributed with constant variance. One might fit a regression line between the ensemble member standard deviation and the ensemble mean absolute error, but clearly, this will violate the regression assumptions. In such situations, a common strategy is to apply variable transformations, such as the square root, to stabilize the variance and improve the normality of errors (see Box et al. 2005, p. 320).

In Fig. 2, we demonstrate pragmatically why the square root transformation works well for ensemble forecasts. To begin, we assume the errors of the ensemble mean $\mathbf{e}$ are normally distributed with mean 0 and variance $\sigma$ that is proportional to ensemble spread, $\mathbf{e} \sim N(0, \sigma)$. A few hypothetical distributions of $\mathbf{e}$ for different values of $\sigma$ are shown in Fig. 2a. As the ensemble standard deviation increases along the abscissa,

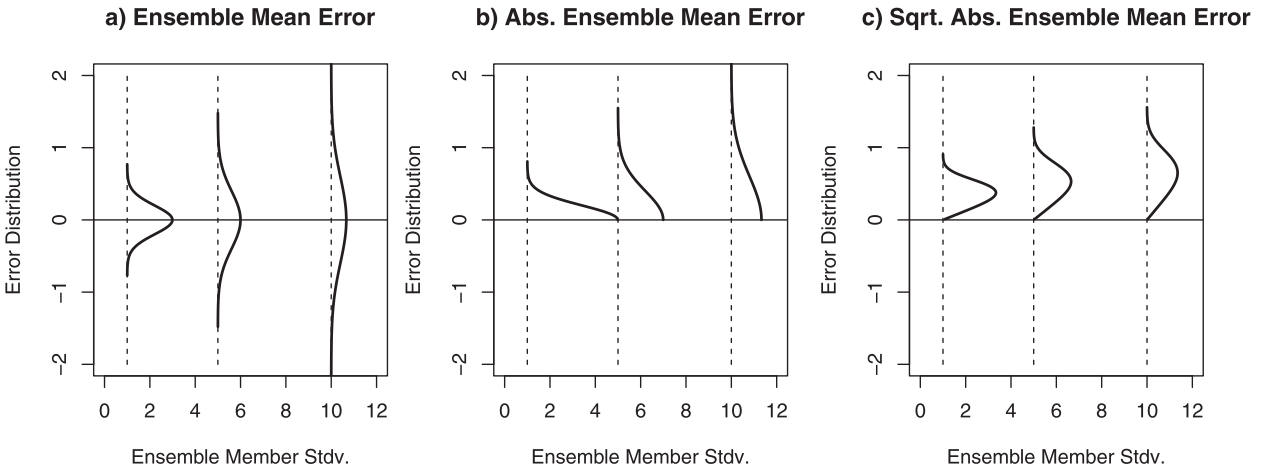


FIG. 2. Examples illustrating how the square root transformation improves the normality and stabilizes the variance of the ensemble mean error distribution. (a) The distribution of the original ensemble mean error, (b) the absolute ensemble mean error, and (c) the square root of the absolute ensemble mean error are shown.

the variance of the errors **e** also increases. Taking the absolute value of **e** transforms the data to the half-normal distributions shown in Fig. 2b. At this point, a regression line could be fit between the ensemble standard deviation and the ensemble mean absolute error because the expected absolute error is proportional to the original error variance. However, the basic regression assumptions of normally distributed error and constant variance are not valid. If instead, we compute the square root of the absolute errors (i.e., compute $\sqrt{|\mathbf{e}|}$), the data become more normally distributed with less heteroscedasticity (Fig. 2c). The mean of the new distribution is still proportional to the original error variance, thus a regression fit will produce a spread–skill relationship. In practice, the ensemble member standard deviation may also be transformed by taking the square root. We have found that doing so improves the regression reduction of variance (not shown).

### *d. Example EKDMOS Spread-Skill forecast*

To further explain the EKDMOS Spread-Skill method, we provide an example application with the 102-h daytime maximum temperature forecast at the Baltimore/Washington International Thurgood Marshall Airport. First, MOS equations were developed with ensemble mean-based predictors. Separate equations were developed for the GEFS and CMC ensembles. Table 1 summarizes the predictors and corresponding coefficients chosen for these particular equations. The stepwise regression selected more predictors for the CMC equation than for the GEFS equation. The most common predictors were model forecasts of near-surface temperature and zonal wind.

Next, we created a spread–skill relationship following the procedure outlined in section 3b. Working with our dependent dataset, we applied the MOS equations listed in Table 1 to the ensemble members to generate 42 ensemble MOS forecasts. For each forecast case, we calculated the ensemble mean MOS forecast, which was simply the equally weighted average of the 42 ensemble MOS forecasts. We also calculated the standard deviation of the ensemble MOS forecasts to quantify the ensemble spread. Figure 3a compares the absolute error of the ensemble mean MOS forecasts to the standard deviation of the ensemble MOS forecasts. For each forecast case, the standard deviation of the ensemble MOS forecasts is plotted along the abscissa while the absolute error of the ensemble mean MOS forecast is plotted along the ordinate. As the spread of the ensemble MOS forecasts increased, larger errors of the ensemble mean MOS forecast tended to occur, indicating there was a spread–skill relationship.

TABLE 1. Summary of predictors and coefficients for the EKDMOS 102-h daytime maximum temperature MOS equations at the Baltimore/Washington International Thurgood Marshall Airport.

| Model | Predictor | Coef |
|---|---|---|
| CMC | Intercept | −2.09 |
| | 2-m temperature, 90 h | 1.08 |
| | 2-m temperature, 102 h | −0.48 |
| | 925-hPa temperature, 96 h | 0.42 |
| | 1000-hPa *U* wind, 90 h | 0.23 |
| GEFS | Intercept | −2.55 |
| | 2-m temperature, 90 h | 1.02 |
| | 10-m *U* wind, 90 h | 0.28 |

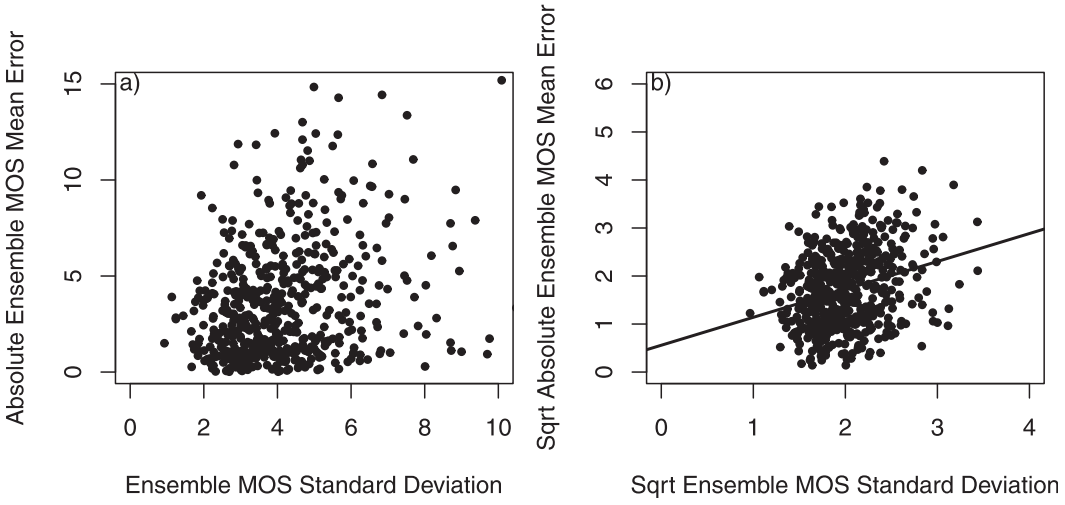


FIG. 3. Scatterplots demonstrating the spread–skill relationship for the 102-h daytime maximum temperature forecast at the Baltimore/Washington International Thurgood Marshall Airport. (a) The standard deviation of the ensemble MOS forecasts versus the absolute error of the ensemble mean MOS forecast is shown. (b) The transformed data are shown.

To fit the spread–skill relationship we applied variable transformations to the data. Specifically, for each case, we computed the square root of the absolute ensemble mean MOS forecast error and the square root of the standard deviation of the ensemble MOS forecasts. As shown in Fig. 3b, the transformations reduced the heteroscedasticity and normalized the error distribution. The black line in Fig. 3b is a regression fit to the transformed data. Although the scatter about the regression line is large, the $p$ value for the regression slope parameter is highly statistically significant ($\ll 0.001$). A check of the regression residuals plotted against the explanatory variable (Fig. 4a) does not show any systematic pattern that would cause alarm. In Fig. 4b, a normal $Q$–$Q$ plot compares the distribution of the standardized regression residual with those of a theoretical normal distribution. Each point on a $Q$–$Q$ plot is the quantile of one distribution plotted against the quantile of a comparison distribution. If both sets of data follow the same distribution then the points will form a straight, diagonal line. Examining Fig. 4b, we see that this was generally the case, suggesting the regression residuals were at least approximately normally distributed.

To demonstrate the generation of an actual forecast we use the 102-h daytime maximum temperature forecast created from the 0000 UTC 1 December 2009 run of the NAEFS. The MOS equations were applied to each

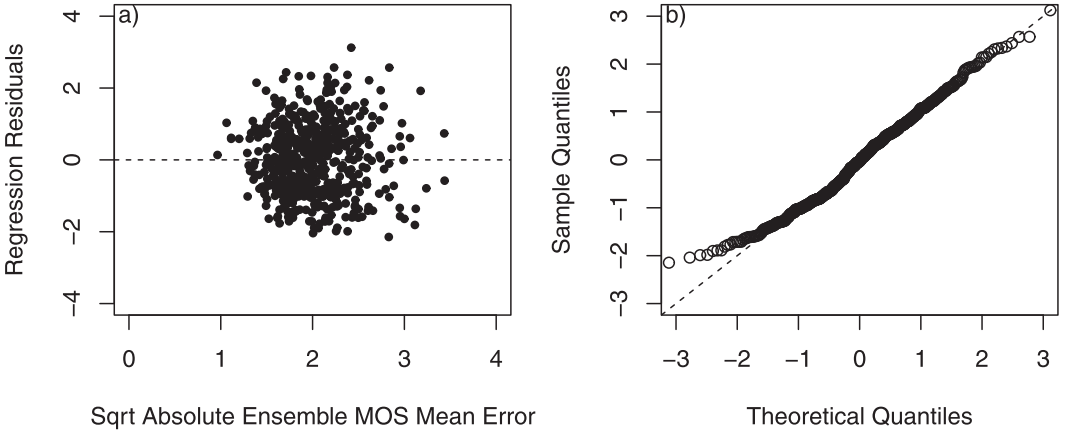


FIG. 4. The standardized regression residuals after fitting (a) the spread–skill relationship and (b) a normal $Q$–$Q$ plot of the standardized regression residuals.

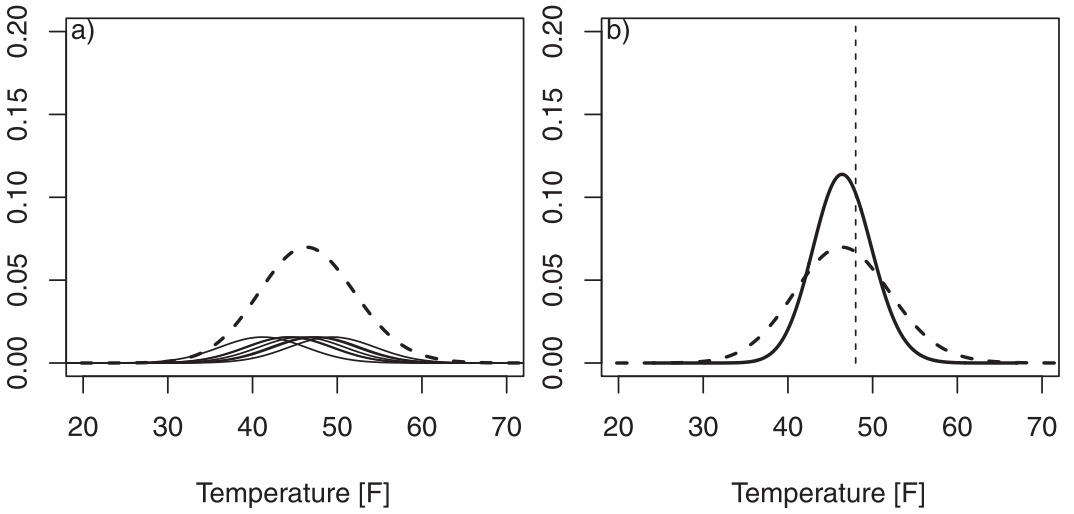


FIG. 5. Example EKDMOS forecast for the 102-h daytime maximum temperature at the Baltimore/Washington International Thurgood Marshall Airport from the 0000 UTC 1 Dec 2009 run of the NAEFS. (a) The solid curves are the individual member kernels and the dashed curve is the kernel density estimated PDF. (b) The dashed curve is the original PDF and the solid curve is the Spread-Skill adjusted PDF with the verifying observation indicated by the vertical dashed line.

member to generate 42 ensemble MOS forecasts. The standard deviation of those forecasts was calculated and used to evaluate the spread–skill relationship and corresponding back transformation. As shown in Fig. 5a, a normal kernel was fit to each ensemble MOS forecast. For clarity we only plot a subset of the 42 kernels. Kernel density estimation was used to sum the area under the individual kernels and obtain the PDF shown by the dashed curve in Fig. 5a. The PDF spread was adjusted to match the standard error estimate predicted by the spread–skill relationship. In Fig. 5b, the dashed curve is the original PDF from kernel density fitting while the solid curve is the PDF after spread calibration. For this case, the final PDF was sharper than the original PDF. The vertical black line indicates the verifying observation of 48°F (~9°C).

### e. *Baseline for spread calibration evaluation*

EKDMOS probabilistic forecasts should be better calibrated than the raw ensemble. To verify this, we computed a set of probabilistic forecasts whose spread was dependent on the original ensemble spread. A cumulative distribution function can be constructed from the raw ensemble by rank sorting the ensemble members. However, near-surface weather forecasts from ensembles will often be poorly calibrated simply due to systematic model biases. To provide a more competitive candidate for evaluation, we first performed some limited postprocessing to correct the systematic bias before computing the CDF. Here we used a modification of an approach suggested by Hamill (2007). For each station-based forecast, we found the direct model output (DMO) ensemble mean and the ensemble mean MOS forecast. We computed the difference between the DMO and ensemble mean MOS forecast and added the difference to each DMO ensemble member. This centered the DMO ensemble members around the EKDMOS mean but preserved their original scatter. A CDF was computed from the relocated DMO ensemble members using the Tukey plotting position estimator (see Wilks 2011, p. 41),

$$\Pr(F \le f) = \frac{\mathrm{Rank}(f) - 1/3}{(K+1) + 1/3}, \tag{19}$$

which determines the probability that the verifying observation $F$ will be less than or equal to ensemble member forecast $f$, where $\mathrm{Rank}(f)$ is the rank of the ensemble member forecast within the ensemble. These forecasts are hereafter referred to as the bias correction rank sorted (BC-Sorted) forecasts.

## 4. Verification results

To evaluate our EKDMOS Spread-Skill technique, we generated 2-m temperature forecasts using NAEFS data covering the period 1 October 2007–31 March 2010. We stratified our sample into warm (1 April–30 September) and cool (1 October–31 March) seasons. We performed cross validation (Wilks 2011) whereby we

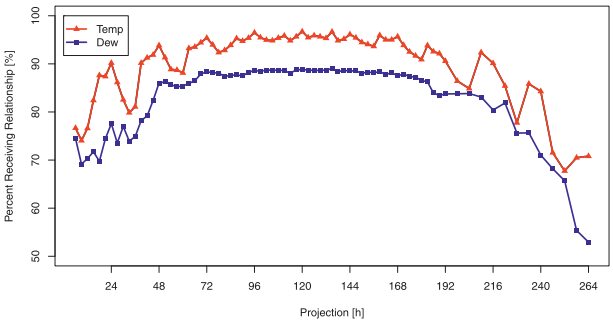


FIG. 6. Percentage of stations receiving a spread–skill relationship by projection. Results for 2-m temperature (Temp) and 2-m dewpoint (Dew).

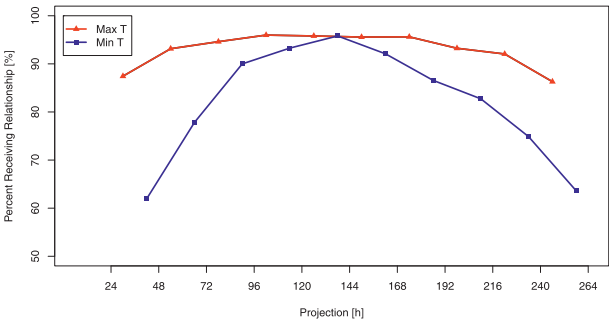


FIG. 7. As in Fig. 6, but for daytime maximum temperature (''Max T'') and nighttime minimum temperature (''Min T'').

developed MOS equations and spread–skill relationships with two years of data, withholding the third year for independent verification. The results presented are for cross-validated cool season forecasts. Unless otherwise stated, all verification scores were computed at a subset of 335 stations distributed uniformly throughout the CONUS, Alaska, Hawaii, and Puerto Rico and judged by MDL to have reliable data (Fig. 1).

### *a. Spread–skill relationships*

We investigated how successful we were at developing spread–skill relationships. We attempted to develop spread–skill relationships at a total of 2303 stations. Figures 6 and 7 present the percentage of stations, by projection, that passed the quality control criteria mentioned in section 3b and received a spread–skill relationship. The values ranged between 50%–95% depending on the element and projection. Coastal and marine stations along the western CONUS most commonly failed to receive a spread–skill relationship. At these sites, the cold ocean sea surface temperature likely suppresses the day-to-day temperature variability and weakens the spread–skill relationship. For hourly 2-m temperature and dewpoint there was a diurnal cycle, with the percentage peaking during the daytime hours. The percentage of stations that received a spread–skill relationship was less for nighttime minimum temperature compared to daytime maximum temperature (Fig. 7), suggesting the ensemble spread was a poorer predictor of accuracy at the nocturnal hours. For all four weather elements, the percentages peaked at the mid-range forecast projections between 74 and 192 h. At the earlier projections the spread–skill relationships may be weak because the ensemble spread is mostly due to the random perturbations used to initialize the ensemble. At the later projections, after 216 h, the percentage may decrease because, when model skill is low the MOS equations reduce spread variability by hedging toward the development sample mean.

### *b. Spread-Skill 2-m temperature forecasts*

We compared EKDMOS Spread-Skill 2-m temperature forecasts with those generated by the Spread-Adjustment and BC-Sorted methods. Figure 8 shows probability integral transform (PIT) histograms for the 48-, 120-, and 168-h forecast projections. PIT histograms measure ensemble consistency and are similar to rank histograms (see Wilks 2011, p. 371). Here, we define consistency to mean that for each forecast case, the verifying observation and ensemble MOS forecasts are drawn from the same probability distribution. A PIT histogram has multiple bins that correspond to discrete percentile ranges on the forecasted CDF. If the ensemble is consistent, then the verifying observation should be equally likely to fall within any bin. Thus, after sampling many cases, the PIT histogram should be uniformly flat with bin heights near unity.

To assess statistical significance we constructed boxplot PITs with a bootstrap technique following Marzban et al. (2011). We individually verified the 18 months of our independent dataset and randomly sampled the results with replacement 10 000 times. The boxplot whiskers in Figs. 8a–i are the 95% confidence intervals for the bin means. Examining Figs. 8a–c, we see that the BC-Sorted forecasts were underdispersive at all forecast projections. The Spread-Adjustment (Figs. 8d–f) and Spread-Skill (Figs. 8g–i) PIT histograms were much flatter, indicating better consistency. Spread-Skill was more consistent than Spread-Adjustment because the 95% confidence intervals more often straddled unity.

Cumulative reliability diagrams (CRD) also diagnose ensemble consistency. CRDs are similar to reliability diagrams; however, the CRD probabilities are cumulative from below (see Wilks 2011, p. 334). Relative frequency to the left of the dashed reference line indicates underforecasting while relative frequency to the right implies overforecasting. We computed 95% confidence intervals for each point on our CRDs with

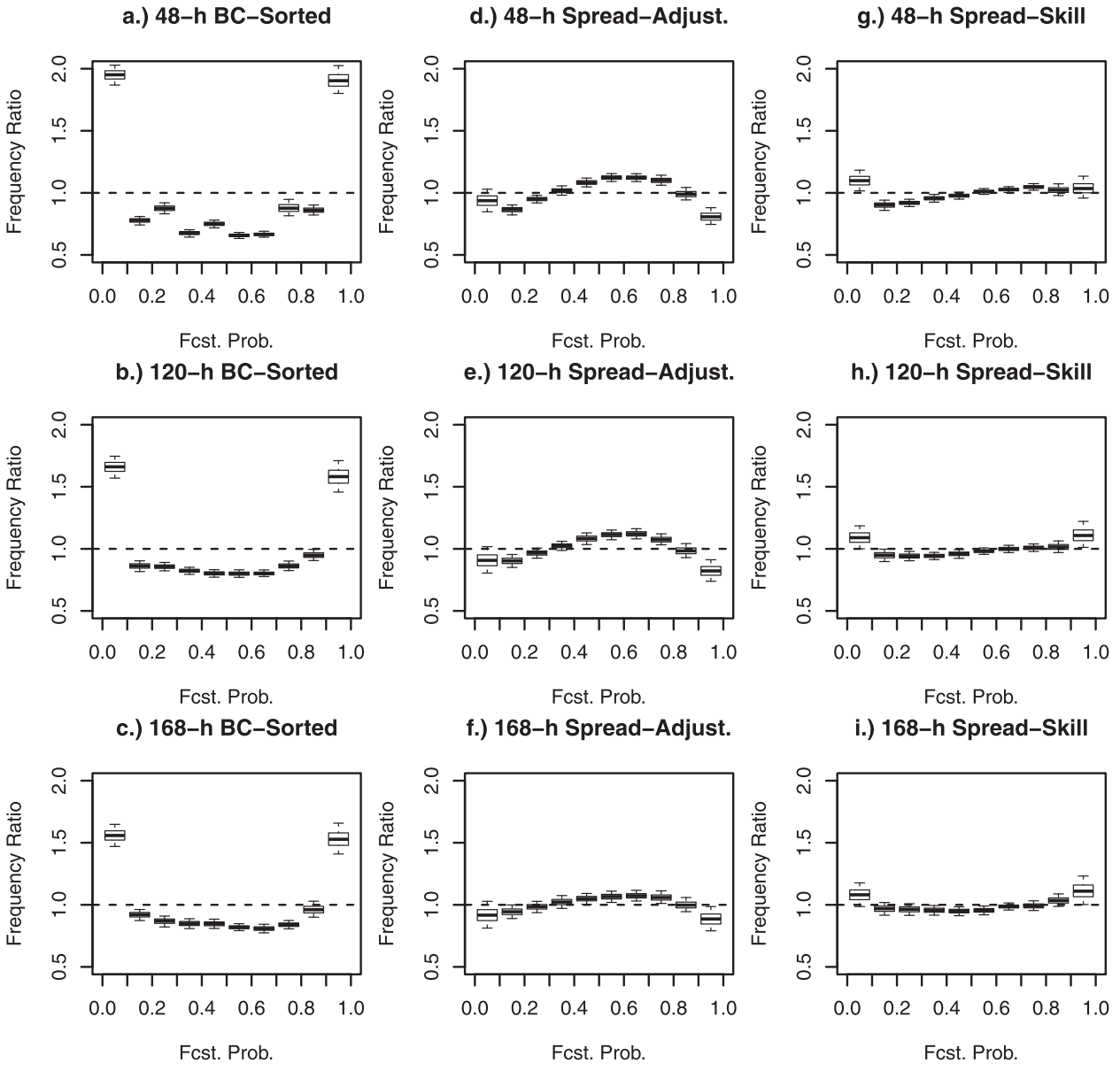


FIG. 8. Boxplot probability integral transform (PIT) histograms comparing (a)–(c) BC-Sorted, (d)–(f) Spread-Adjustment, and (g)–(i) Spread-Skill forecasts. The whiskers are the 95% confidence interval for the bin height.

the bootstrap technique described above. The CRDs confirmed that the BC-Sorted forecasts were unreliable; the individual points often differed significantly from the diagonal reference line (Figs. 9a–c). The BC-Sorted curves did however cross the diagonal reference line near the 0.5 percentile point, indicating the bias correction successfully relocated the center of the PDFs. Therefore, underdispersion most likely caused the poor reliability. For each projection, the Spread-Adjustment (Figs. 9d–f) and Spread-Skill (Figs. 9g–i) cumulative frequency curves were close to the diagonal reference line.

We also computed the continuous ranked probability score (CRPS; Matheson and Winkler 1976; Unger 1985; Hersbach 2000) which is the continuous analog of the ranked probability score (Epstein 1969). CRPS is a negatively oriented score that evaluates the resolution and reliability of the forecasted CDF. Figure 10a shows the 2-m temperature CRPS for BC-Sorted, Spread-Skill, and Spread-Adjustment. For all methods, the CRPS values were quite similar. To test statistical significance we used a bootstrap technique. We computed the CRPS for each of the 18 months in our independent sample. Using each monthly result, we computed the difference

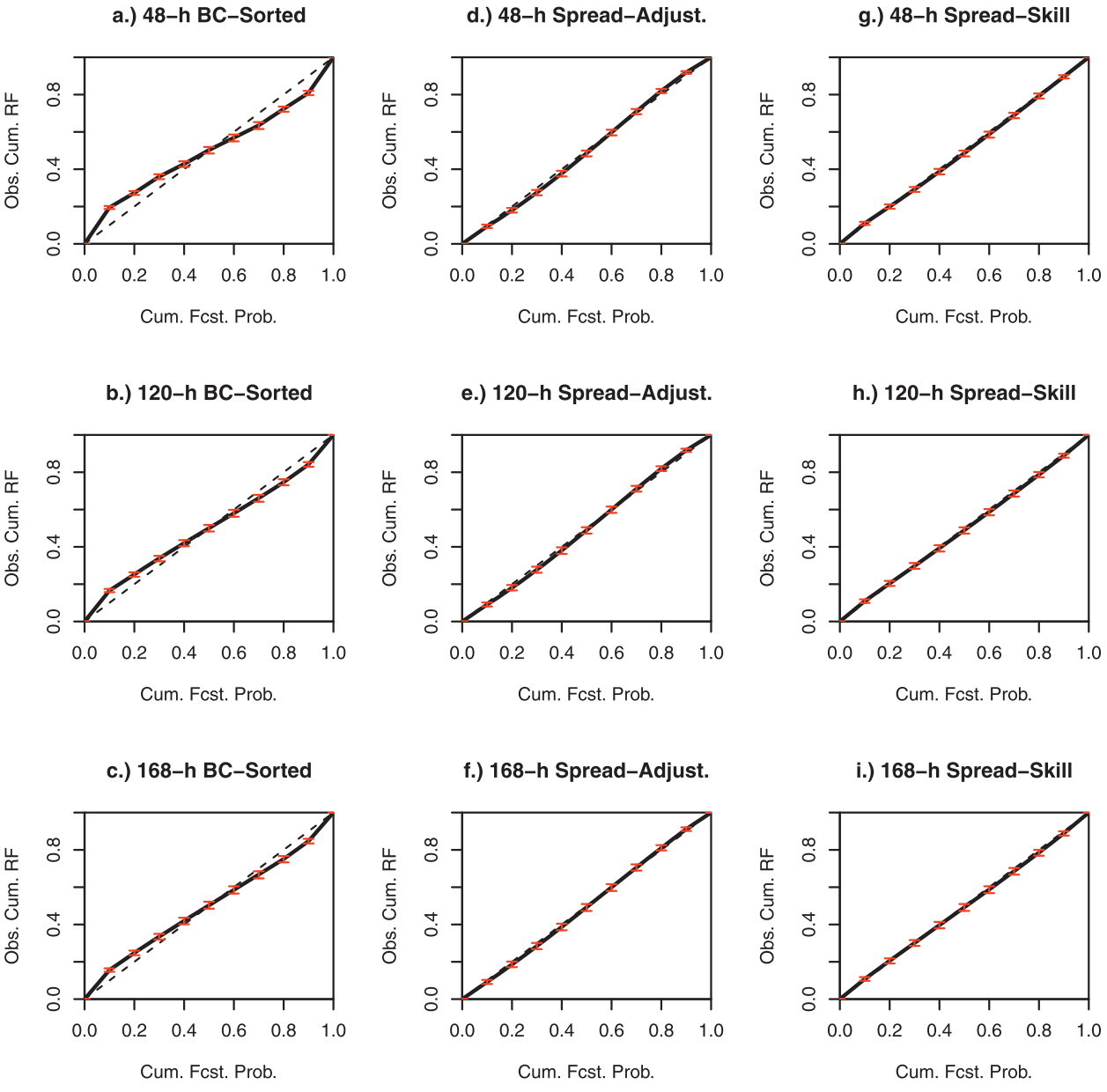


FIG. 9. Cumulative reliability diagrams (CRD) comparing (a)–(c) BC-Sorted, (d)–(f) Spread-Adjustment, and (g)–(i) Spread-Skill forecasts. The 95% confidence interval for each point is shown in red.

between Spread-Skill and Spread-Adjustment, and the difference between Spread-Skill and BC-Sorted. The difference was computed as the comparison method minus Spread-Skill, thus, since CPRS is negatively oriented, positive (negative) values imply Spread-Skill was better (worse). We sampled the 18 monthly differences 10 000 times with replacement and computed the 95% confidence interval for the difference in means. At most projections, the difference between Spread-Skill and Spread-Adjustment was not statistically significant (Fig. 10b). In contrast, Spread-Skill was significantly better than BC-Sorted at all projections (Fig. 10c).

### *c. Spread-error verification diagrams*

To further test the calibration, we grouped 2-m temperature forecasts with similar predicted spread and computed the corresponding standard error of ensemble mean MOS forecast. If the forecast spread was calibrated, then on average the standard deviation of the predicted PDF should match the standard error of the

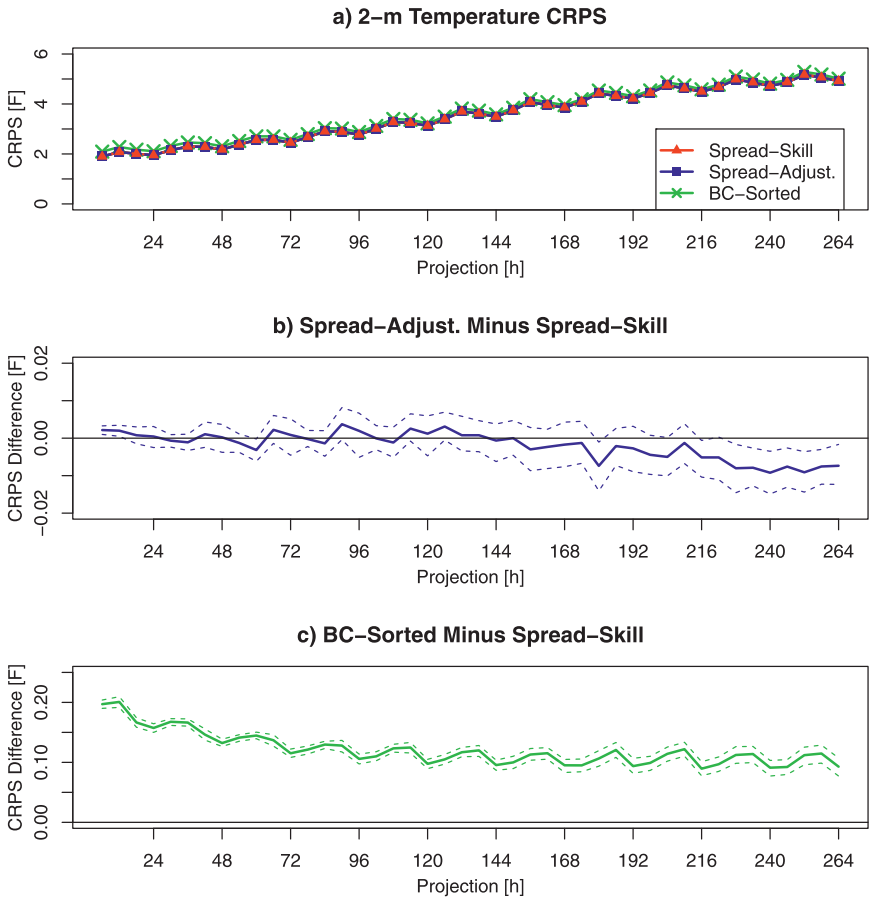


FIG. 10. (a) Continuous ranked probability score (CRPS) for Spread-Skill, Spread-Adjustment, and BC-Sorted 2-m temperature forecasts. Testing the difference in means between (b) Spread-Adjustment and Spread-Skill and (c) BC-Sorted and Spread-Skill forecasts. The dashed lines are 95% confidence intervals.

ensemble mean MOS forecast (Van den Dool 1989; Grimit and Mass 2007; Eckel et al. 2012). Such spread-error verification diagrams are shown in Fig. 11 for the 120-h, 2-m temperature forecast. We present both an overall spread-error verification diagram constructed using the 335 stations, as well as three single-station diagrams. For each figure, the standard deviation of the forecasted PDF is plotted along the abscissa while the standard error of the verified ensemble mean MOS forecast is plotted along the ordinate. The data points were formed by grouping cases with similar predicted spread into equal case count bins. For the overall diagram (Fig. 11a) there were approximately 43 000 forecast cases per bin while for the single-station diagrams (Figs. 11b–d) there were approximately 130. For each bin, we computed the 95% confidence interval for the ensemble mean MOS forecast standard error by comparing the estimated value with the critical values of a chi-squared distribution.

In Figs. 11a–d, Spread-Skill forecasts are shown in red, Spread-Adjustment forecasts are shown in blue, and BC-Sorted forecasts are shown in green. In each figure, if there is a spread–skill relationship, the points will spread along the abscissa and form a positively sloped curve. If the spread–skill relationship is weak, the points will either tightly cluster along the abscissa or the slope of the curve may be close to zero. In all cases, the points should fall along the dashed diagonal reference line if the forecast spread is calibrated.

Examining Fig. 11a, we observe that all three techniques produced a spread–skill relationship. The BC-Sorted spread–skill relationship was underdispersive and poorly calibrated; the dashed diagonal reference line is outside the 95% confidence interval for each point. In contrast, Spread-Skill and Spread-Adjustment were better calibrated. In agreement with the PIT diagrams, the Spread-Adjustment points fell to the right of the reference line suggesting overdispersion while the

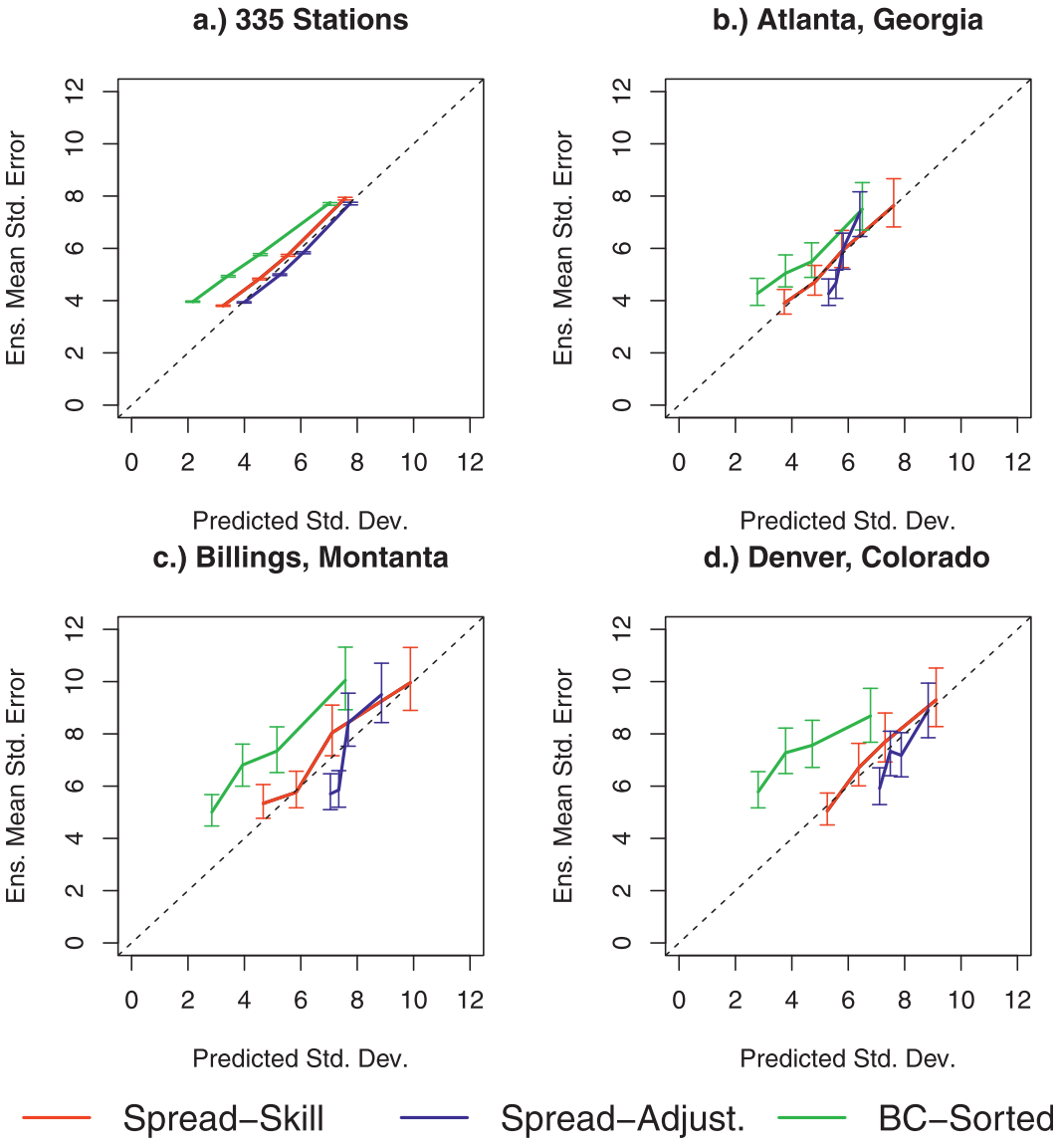


FIG. 11. Spread-error verification diagrams for (a) 335 stations and (b)–(d) select individual stations. Spread-Skill is plotted in red, Spread-Adjustment in blue, and BC-Sorted in green. The error bars are the 95% confidence intervals for each point estimate.

Spread-Skill points were to the left, suggesting underdispersion.

Since Fig. 11a presents station-pooled results it does not guarantee the spread–skill relationships were well calibrated at individual stations. Therefore, we also constructed station-specific diagrams, three of which we show in Figs. 11b–d. Typically, we found that the BC-Sorted performed worst while Spread-Skill was best. In addition, we found Spread-Skill produced greater day-to-day spread variability, compared to Spread-Adjustment, which is apparent because the points were more widely distributed along the abscissa.

To quantify the reliability and strength of the spread–skill relationship at individual stations, we created two verification scores. We computed the first, which we call the spread-error reliability score, by finding the squared difference between the ensemble mean MOS forecast standard error and the predicted standard deviation for each bin. We summed the results for all bins and then took the square root. The score was computed individually for each station and then averaged over all stations. The score is negatively oriented and measures how closely the predicted spread matched the observed standard error.

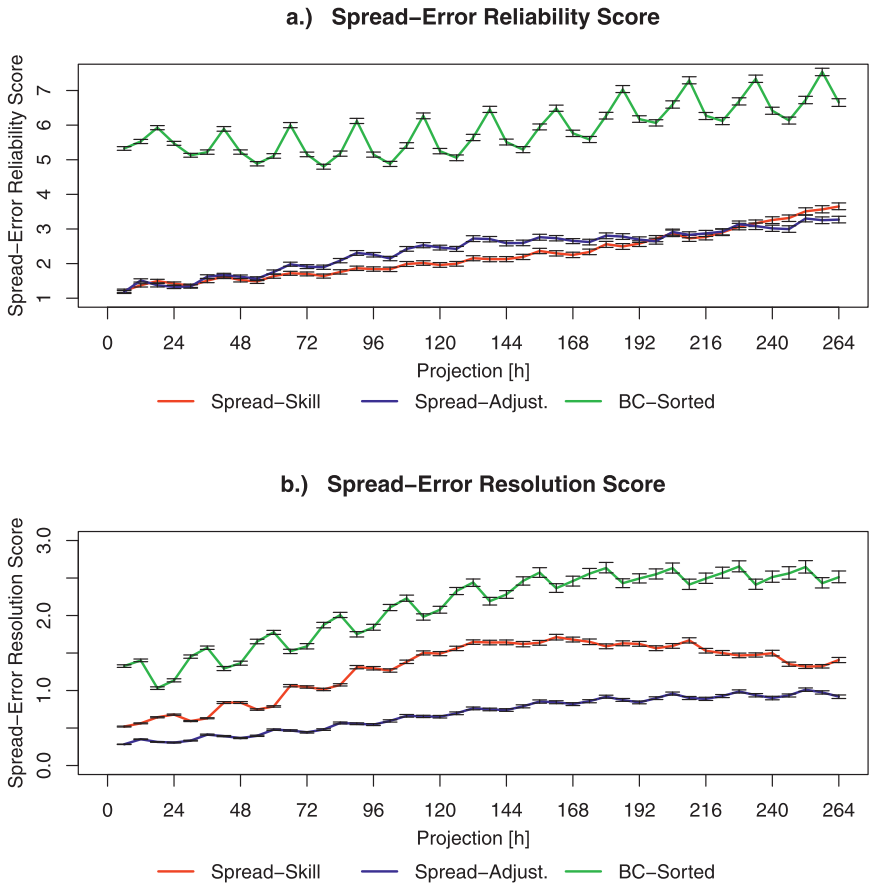


FIG. 12. (a) Spread-error reliability score and (b) spread-error resolution score. Spread-Skill is plotted in red, Spread-Adjustment in blue, and BC-Sorted in green.

The second score measures day-to-day spread variability. At each station, we created a histogram of the predicted spread from all the cases in our sample and computed the histogram interquartile range. We averaged the interquartile ranges for all stations and dubbed this metric the spread-error resolution score. The score is positively oriented; larger values indicate greater day-to-day spread variability, and perhaps, a greater ability to distinguish between low and high forecast confidence.

We used a bootstrap technique to build confidence intervals for both scores. We randomly sampled from the total set of possible verification dates with replacement 10 000 times. For each iteration, we computed the spread-error reliability and spread-error resolution scores as outlined above with the random date list. We sorted the results to find the 95% confidence interval.

In terms of the spread-error reliability score, Spread-Skill and Spread-Adjustment forecasts were more reliable than BC-Sorted forecasts (Fig. 12a). Here we overlay the 95% confidence intervals in black. The Spread-Skill and Spread-Adjustment scores were not significantly different between 6 and 72 h. Spread-Skill was better between 72 and 192 h; however, after 192 h Spread-Adjustment was better. Both Spread-Skill and Spread-Adjustment were significantly better than BC-Sorted at all projections.

Examining the spread-error resolution score (Fig. 12b), we found that BC-Sorted had the greatest spread variability, but as we showed previously, the spread was uncalibrated. The Spread-Skill spread variability was around twice that of Spread-Adjustment, and as we demonstrated, the spread was well calibrated. For all methods, the 95% confidence intervals indicated the differences were statistically significant.

## 5. Discussion

This study described recent modifications to the EKDMOS statistical postprocessing technique. We explained how we improved the technique by including a spread–skill relationship within the multiple linear regression framework. Our new EKDMOS statistical

model was fit in two stages. First, we generated ensemble mean-based MOS equations with forward screening multiple linear regression. Second, to produce flow-dependent spread variability, we created spread–skill relationships that model the predictive standard error as a function of the ensemble member standard deviation. We used square root variable transformations to improve the linearity between ensemble mean accuracy and ensemble spread, and fit the spread–skill relationship with a one-term linear regression equation. When applied to 3 years of operational NAEFS forecasts, the technique produced statistically consistent probabilistic forecasts.

The EKDMOS methodology presented here resembles other postprocessing methods, such as NGR, because we explicitly include a spread–skill relationship in the statistical model. In the case of NGR, the statistical model is fit with an iterative technique that minimizes the CRPS. In contrast, the EKDMOS statistical model is fit via two closed-form regression steps. Since the EKDMOS spread–skill relationship is a one-term regression equation, the relationship could be derived from accumulated information and incorporated into an updateable MOS system such as the one proposed by Wilson and Vallée (2002). Thus, we could train the statistical model with large samples but only have to retain a small dataset.

We did not attempt to unequally weight the ensemble members when fitting the MOS equations. Rather, we developed separate MOS equations for each ensemble system and later joined the equally weighted MOS forecasts with kernel density fitting. From an operational perspective, this approach is desirable in case an ensemble member is missing for a given day, or if one of the operational models undergoes a major change that requires redeveloping the MOS equations. However, weighting could be readily incorporated into the technique. For example, Bayesian model averaging (Raftery et al. 2005) or mean absolute error (MAE)-based (Woodcock and Engel 2005) weights could be applied to the ensemble MOS forecasts prior to kernel density fitting.

The PIT histograms, CRDs, and CRPS results suggested that there was little difference in calibration between Spread-Skill and Spread-Adjustment. In contrast, the station-specific spread-error verification diagrams and accompanying spread-error reliability and spread-error resolution scores demonstrated that Spread-Skill was in fact superior. The Spread-Skill forecasts were comparable to the original Spread-Adjustment technique in terms of calibration but produced a much greater range in day-to-day predicted spread. Thus, the Spread-Skill forecasts better discriminate between low and high forecast-confidence situations.

The verification results we presented were for 2-m temperature cool season forecasts. We have also successfully used the EKDMOS technique to postprocess dewpoint, daytime maximum temperature, and nighttime minimum temperature in both the warm and cool seasons. Preliminary work suggests the technique may also be used for probabilistic apparent temperature forecasts and other reasonably normally distributed elements.

NAEFS-based EKDMOS forecasts have been operationally implemented on the NCEP Central Computing System (CCS) beginning with the 1200 UTC 29 May 2012 run of the NAEFS. The current implementation uses a slightly different method to fit the spread–skill relationship described by Veenhuis and Wagner (2012). A future update to EKDMOS will use the methodology from this study. Currently, EKDMOS produces forecasts at 2303 stations from the 0000 and 1200 UTC run of the NAEFS. The station-based forecasts are analyzed to 2.5-km grids covering the CONUS and Alaska with the BCDG technique (Glahn et al. 2009b). Station-based and gridded GRIB2 forecasts for 2-m temperature, dewpoint, daytime maximum, and nighttime minimum temperature are available online (http://www.mdl.nws.noaa.gov/~naefs_ekdmos/). Future implementations of EKDMOS will include probabilistic guidance for additional elements.

*Acknowledgments.* The author is grateful to Bob Glahn, Matt Peroutka, John Wagner, and two anonymous reviewers for useful suggestions. The author acknowledges Tony Eckel who suggested including a spread-skill-based calibration technique within EKDMOS. This work was made possible by the Statistical Modeling Branch of the Meteorological Development Laboratory, which maintains an archive of quality controlled station observations. The operational ensemble forecasts furnished by NCEP and the CMC were also invaluable for this work.

## REFERENCES

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733–742.

Box, E. P., J. S. Hunter, and W. G. Hunter, 2005: *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd ed. Wiley-Interscience, 664 pp.

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

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

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

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.,* **138,** 1877–1901.

Eckel, F. A., M. S. Allen, and M. C. Sittel, 2012: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting,* **27,** 50–69.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

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

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

——, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009b: The gridding of MOS. *Wea. Forecasting,* **24,** 520–529.

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

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.,* **135,** 203–221.

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.

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

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchei, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.,* **137,** 1717–1720.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, and A. Deng, 2009: Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *J. Appl. Meteor. Climatol.,* **48,** 2001–2012.

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

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.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22,** 1087–1096.

McQueen J., J. Du, B. Zhou, G. Manikin, B. Ferrier, H.-Y. Chuang, G. DiMego, and Z. Toth, 2005: Recent upgrades to the NCEP Short Range Ensemble Forecasting System (SREF) and future plans. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 11.2. [Available online at http://ams.confex.com/ams/pdfpapers/94665.pdf.]

Palmer, T. N., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics and adaptive observations. *J. Atmos. Sci.,* **55,** 633–653.

Peng, M. S., J. A. Ridout, and T. F. Hogan, 2004: Recent modifications of the Emanuel convective scheme in the Navy Operational Global Atmospheric Prediction System. *Mon. Wea. Rev.,* **132,** 1254–1268.

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., and L. A. Smith, 2003: Combining dynamic and statistical ensembles. *Tellus,* **55A,** 16–30.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

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

Unger, D. A., 1985: A method to estimate the Continuous Ranked Probability Score. Preprints, *Ninth Conf. on Probability and Statistics in the Atmospheric Sciences,* Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

——, H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.,* **137,** 2365–2379.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.,* **117,** 2230–2247.

Veenhuis, B. A., and J. L. Wagner, 2012: Second moment calibration and comparative verification of ensemble MOS forecasts. Preprints, *21st Conf. on Probability and Statistics in the Atmospheric Sciences,* New Orleans, LA, Amer. Meteor. Soc., 2.3. [Available online at https://ams.confex.com/ams/92Annual/flvgateway.cgi/id/19514?recordingid=19514.]

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

——, and ——, 2005: Improvements of ensemble reliability using a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.,* **126,** 3292–3302.

Wilks, D. S., 2006: Comparison of ensemble MOS methods in the Lorenz 96 setting. *Meteor. Appl.,* **13,** 243–256.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Academic Press, 740 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Wilson, L. J., and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206–222.

——, S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1365–1385.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting,* **20,** 101–111.

Zhu Y., R. Wobus, M. Wei, B. Cui, and Z. Toth, 2007: March 2007 NAEFS upgrade. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]