MOS Uncertainty Estimates in an Ensemble Framework

BOB GLAHN, MATTHEW PEROUTKA, JERRY WIEDENFELD, JOHN WAGNER, GREG ZYLSTRA, AND BRYAN SCHUKNECHT

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

BRYAN JACKSON

NOAA/National Weather Service, Wakefield, Virginia

(Manuscript received 21 February 2008, in final form 22 May 2008)

ABSTRACT

It is being increasingly recognized that the uncertainty in weather forecasts should be quantified and furnished to users along with the single-value forecasts usually provided. Probabilistic forecasts of "events" have been made in special cases; for instance, probabilistic forecasts of the event defined as 0.01 in. or more of precipitation at a point over a specified time period [i.e., the probability of precipitation (PoP)] have been disseminated to the public by the Weather Bureau/National Weather Service since 1966. Within the past decade, ensembles of operational numerical weather prediction models have been produced and used to some degree to provide probabilistic estimates of events easily dealt with, such as the occurrence of specific amounts of precipitation. In most such applications, the number of ensembles restricts this "enumeration" method, and the ensembles are characteristically underdispersive. However, fewer attempts have been made to provide a probability density function (PDF) or cumulative distribution function (CDF) for a continuous variable. The Meteorological Development Laboratory (MDL) has used the error estimation capabilities of the linear regression framework and kernel density fitting applied to individual and aggregate ensemble members of the Global Ensemble Forecast System of the National Centers for Environmental Prediction to develop PDFs and CDFs. This paper describes the method and results for temperature, dewpoint, daytime maximum temperature, and nighttime minimum temperature. The method produces reliable forecasts with accuracy exceeding the raw ensembles. Points on the CDF for 1650 stations have been mapped to the National Digital Forecast Database 5-km grid and an example is provided.

1. Introduction

Weather forecasting[1] is not an exact science. Almost any weather forecast, whether it is for a dichotomous event like precipitation, or for a quasi-continuous variable like temperature, has an element of uncertainty associated with it. Whether the forecast is machine generated or humanly produced, the uncertainty may be hard to quantify. Nevertheless, it has been recognized within the meteorological community for many years that some expression or measure of uncertainty should accompany the forecast to better serve the user of the forecast.

This was well recognized by one of the earliest forecasters, Cleveland Abbe, who helped establish the U.S. Weather Service and was actually called "old probabilities." In 1965, the Weather Bureau made the probability of precipitation (PoP) product operational nationwide, by carefully defining the event as 0.01 in. or more of precipitation at a point over a 12-h period. This led the then Assistant Secretary of Commerce, Myron Tribus (Tribus 1970), to make the statement, "It was not too long ago that the major concession by the Weather Bureau to the existence of probability theory was the use of words such as 'likely,' 'probably,' or 'chance.' Fortunately, this policy has been abandoned. Today we have forecasts couched in the language of probability, which represents a distinct improvement

[1] No distinction is made here between weather and climate.

Corresponding author address: Bob Glahn, Meteorological Development Laboratory, W/OST2, Room 10214, SSMC2, 1325 East–West Highway, Silver Spring, MD 20910.
E-mail: Harry.Glahn@noaa.gov

DOI: 10.1175/2008MWR2569.1

over deterministic pronouncements."[2] Unfortunately, the progress of probability forecasting since that time has been excruciatingly slow.

There is currently a much renewed interest in probability forecasting. The American Meteorological Society (AMS) published a statement in 2002 that included, "Much of the informational content of meteorological data, models, techniques, and forecaster thought processes is not being conveyed to the users of weather forecasts. Making and disseminating forecasts in probabilistic terms would correct a major portion of the shortcoming" (AMS 2002). The National Research Council (NRC) undertook a study sponsored jointly by the National Weather Service (NWS) and the Office of Meteorological Research within the National Oceanic and Atmospheric Administration (NOAA) to suggest how we might make headway on this difficult problem. Their recent report (National Research Council 2006) makes several good suggestions that will provide guidance to the meteorological community.

While it might be a chicken and egg situation, much of the renewed interest in probability forecasting is likely because of the computer power now available to run ensembles of a (largely) deterministic model with slightly varying initial conditions (Toth and Kalnay 1997). The improvement in weather forecasting, for more than a few hours in advance, has come predominantly from better numerical models. Models and their output, including postprocessed products such as model output statistics (MOS), have driven the forecast enterprise. Forecasts for 5 days, 7 days, and even longer came about when the models showed some skill at those projections. Multiple results from a model immediately suggest, and provide the possibility of, probability forecasts. The desire to know and provide the uncertainty of numerical model output led to ensembles, a Monte Carlo approach suggested many years ago by Leith (1974) as an alternative to stochastic-dynamic prediction discussed by Epstein (1969) and Fleming (1971a,b), the latter being an elegant solution, but still impractical to implement with complex models on existing computers.

Numerical models characteristically do not provide forecasts of many of the weather variables needed by users, such as ceiling height, visibility, type of precipitation, cloud amount, cloud layer amount, and cloud layer height, nor do they provide probability output directly. Computing the relative frequency of an *event* from an ensemble is simple, but there is the question of skill and, particularly, reliability. Dealing with *continuous* variables, such as surface (i.e., 2 m) temperature, is even more challenging.

While a tremendous theoretical and implementation effort has been put into producing ensembles, considerably less effort has been put into developing methods to interpret and postprocess the ensemble output. Early emphasis was put on improving the model by studying upper-atmospheric variables such as 500-mb height (Atger 1999; Krishnamurti et al. 2003) rather than the weather for the man or woman on the street.

A number of techniques have appeared that use ensembles to make probabilistic forecasts of this so-called sensible weather. Hamill and Colucci (1997, 1998) and Eckel and Walters (1998) described a technique that used rank histograms to calibrate quantitative precipitation forecasts (QPFs) from ensembles. Krishnamurti et al. (2000) used a statistical combination of a multimodel ensemble; this so-called superensemble technique has been improved more recently with the addition of empirical orthogonal functions (Yun et al. 2005). "Ensemble dressing" techniques (Roulston and Smith 2003; Wang and Bishop 2005) address the error characteristics of one or more ensemble members. More recently, Bayesian model averaging (Raftery et al. 2005; Wilson et al. 2007), ensemble model output statistics (Gneiting et al. 2005), and analog techniques (Hamill et al. 2006) have all been applied to this problem. Recent evaluations of some of these techniques (and a few others) have been conducted using synthetic data (Wilks 2006a) and low-resolution Global Forecast System (GFS) reforecast datasets (Wilks and Hamill 2007).

The errors in numerical model forecasts are of two classes—inaccuracy of initial conditions and imperfect models. Numerical models start with a three-dimensional snapshot of the atmosphere characterized by values at grid points or possibly in spectral components. This snapshot is generated by assimilating many observations from a variety of sources, each source having certain, many times unknown, error characteristics. This data assimilation has become increasingly sophisticated and is a science in itself. Even so, it is not perfect; that is, the snapshot picture is produced with an imperfect lens. It is generally believed that if several different snapshots are used as initial conditions, each different, but reasonable from a synoptic and theoretical point of view, the evolution of the model forecasts from them will provide the needed basic uncertainty information. As Buizza et al. (2005) said, "the forecast probability density function [can be] approximated using a finite sample of forecast scenarios." No single best way to produce the multiple model initializations is

[2] Tribus was well known for his book *Rational Descriptions, Decisions, and Designs* in which he promoted Bayesian methods.

known, however (Descamps and Talagrand 2007). In any case, to date, the ensemble results in general do not cover the full range of possibilities of the verifying weather—the ensemble forecasts are underdispersive (e.g., Stensrud and Yussouf 2003; Gneiting et al. 2005).

It is not surprising that the ensembles are underdispersive, given that many of them do not build in the uncertainties inherent in the model itself or if they do, do so inadequately. But efforts still remain to make the ensemble results "dispersive enough" to cover the solution space, and to make the distribution of model solutions representative of the real, but unknowable, probability distribution. One wonders whether this is a reasonable expectation, given that a large source of error is dealt with inadequately, if at all. This situation is sometimes mitigated by making an ensemble composed of runs of more than one model, but not always will two or more models together produce a realistic and reliable probabilistic forecast.

The postprocessing of model data, MOS being the technique most used,[3] can produce quite unbiased forecasts. This is true for binary events as well as quasi-continuous variables, as was shown long ago for PoP (Glahn and Lowry 1969), and thunderstorm occurrence (Reap and Foster 1979). For the binary event, "unbiased" is equivalent to the forecasts being reliable—a basic desirable characteristic of a probability forecast.[4] Leith (1974) in his paper suggesting ensembles, states, ". . . any forecasting procedure can be made optimal in the least-square-error sense by the use of a final regression step." The Meteorological Development Laboratory (MDL) has been producing such probability forecasts for years (Carter et al. 1989), but there was not a strong pull for the information. Dealing in a probabilistic sense, a binary event is relatively straightforward for MOS.

To use regression, and other statistical models, to produce an objective estimate of a binary event, one has only to classify the event as occurring or nonoccurring, assign different values to those two conditions, and apply a model that tries to put those events into separate categories. Over the history of objective weather forecasting, (linear) regression is probably the most used. If the event is classified as a "1" and the nonevent as a "0," then the resulting value is an estimate of the probability of the event occurring. Least squares regression provides an estimate that minimizes the P score defined by Brier (1950). That is, the P score, or Brier score, which is more widely used and is ½ the P score for a dichotomous predictand, is a mean-square error score, which is exactly what regression minimizes. This specific application was dubbed Regression Estimation of Event Probabilities (REEP) by Miller (1964), who was a pioneer in objective weather forecasting. REEP has been used extensively by MDL and gives reliable results, even though the regression values can go outside the zero to unity range and have to be truncated to satisfy the definition of probability.

REEP can also be used to deal with a quasi-continuous variable by dividing it into several categories, either discrete or cumulative (see Glahn 1985, his Table 3), and performing regression, with the same predictors in each equation, on each category. When the discrete categories are mutually exclusive and exhaustive, the sum of the resulting probability estimates equals unity, as they should.[5] These estimates together form a, perhaps crude, cumulative distribution function (CDF), but is a viable means of dealing with predictands of a very nonnormal nature, such as ceiling height, visibility, and precipitation amount. However, when the predictand, such as temperature, is quasi normal, a more straightforward method is available.

This paper defines a capability of producing probability forecasts of quasi-normally distributed variables by regression methods, both with and without ensembles. The focus of this paper is on temperature at specific times (T); some results are also shown for dewpoint temperature (T_{dp}), daytime maximum (MaxT) temperature, and nighttime minimum (MinT) temperature.

2. Regression framework for probability forecasts

Relatively little has been done in producing a measure of uncertainty associated with a single value forecast of a continuous variable like temperature. Actually, the multiple regression framework provides for this, under the usual normality assumption. This capability has been used in two dimensions (e.g., see Wilks 2006b, p. 196; Neter and Wasserman 1974, p. 171; Glahn 2002)—one predictor and one predictand. The results of such an application can be plotted along with

[3] Actually, MOS by its very name and definition, is a synonym for model postprocessing except for purely mathematical or physically based algorithms devoid of statistics.

[4] We include here the concept of "reliable in the small" defined by Murphy and Dann (1985), which means that not only the overall relative frequency is (near) correct, but that the relative frequency of occurrence within small probability bands is also correct.

[5] The fact that any of the estimates could go outside the zero to the unity range may make recalibration necessary. Usually this has to be addressed, but is of minor importance for the final result.

underlying data. An error band at some level of probability, say 95%, can be put on the linear line such that for a new value of the predictor the probability of the verifying value being within the band is 95%. For a fairly normally distributed variable like temperature, the results should be quite good in this regard. (This is demonstrated later in Fig. 5.)

When the regression gives an estimate of Y as a function of X, the coefficients being computed from a sample of size n, the lines at probability level α can be placed according to the t distribution:

$$\hat{Y}_{h(\text{new})} \pm t(1 - \alpha/2; n - 2)s[\hat{Y}_{h(\text{new})}], \tag{1}$$

where $\hat{Y}_{h(\text{new})}$ represents a new value of Y found from a new value of X, $X_{h(\text{new})}$:

$$s^2[\hat{Y}_{h(\text{new})}] = \text{MSE}\left\{1 + \frac{1}{n} + \frac{[X_{h(\text{new})} - \overline{X}]^2}{\sum(X_i - \overline{X})^2}\right\}, \tag{2}$$

$$\text{MSE} = \frac{\sum(Y_i - \hat{Y})^2}{n - 2} \approx s^2(1 - r^2) \quad \text{for large } n, \tag{3}$$

where the MSE is the error (or residual or unexplained) sum of squares divided by the degrees of freedom, $n - 2$, r is the multiple correlation coefficient, s is the standard deviation of the predictand, and the summations are taken over the n sample cases. For samples of the size we usually deal with in meteorology, the normal distribution can be used instead of the t distribution, which is necessary when n is less than about 30.

The theory for placing prediction bands for a *multiple* regression solution is contained in various texts, Montgomery and Peck (1982) being especially good. To illustrate with three predictors, let the predictand values be arranged in the vector (subscripts on matrices and vectors indicate dimension):

$${}_n\mathbf{Y}_1 = \left|\begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \cdots \\ y_n \end{matrix}\right|, \tag{4}$$

the three-predictor matrix:

$${}_n\mathbf{X}_4 = \left|\begin{matrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{matrix}\right|, \tag{5}$$

and the coefficient vector:

$${}_4\mathbf{A}_1 = \left|\begin{matrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{matrix}\right|. \tag{6}$$

Then the regression equation that will produce estimates for the n points is written

$${}_n\hat{\mathbf{Y}}_1 = {}_n\mathbf{X}_4\mathbf{A}_1 = {}_n\mathbf{Y}_1 + {}_n\mathbf{e}_1, \tag{7}$$

${}_n\mathbf{e}_1$ being the errors of the estimates.

The coefficient vector is found by

$${}_4\mathbf{A}_1 = ({}_4\mathbf{X}'_n\mathbf{X}_4)^{-1}{}_4\mathbf{X}'_n\mathbf{Y}_1, \tag{8}$$

where ${}_4\mathbf{X}'_n$ is the transpose of ${}_n\mathbf{X}_4$ and -1 denotes the inverse.

A new value of Y_h is given for the predictor values X_{ih} by

$$\hat{Y}_h = {}_1\mathbf{X}_4\mathbf{A}_1, \tag{9}$$

with

$${}_1\mathbf{X}_4 = |1 \quad x_{1h} \quad x_{2h} \quad x_{3h}|. \tag{10}$$

The error of the new prediction value is

$$s_h^2[\hat{Y}_{h(\text{new})}] = \pm\{\hat{\sigma}^2[1 + {}_1\mathbf{X}_4({}_4\mathbf{X}'_n\mathbf{X}_4)^{-1}{}_4\mathbf{X}'_1]\}^{1/2}, \tag{11}$$

where

$$\hat{\sigma}^2 \approx s^2(1 - R^2) \tag{12}$$

for large n, and R is the multiple correlation afforded by the equation. Finally,

$$\hat{Y}_{h(\text{new})} \pm \{s^2(1 - R^2)[1 + {}_1\mathbf{X}_4({}_4\mathbf{X}'_n\mathbf{X}_4)^{-1}{}_4\mathbf{X}'_1]\}^{1/2} \tag{13}$$

can be used to put error bounds around the new value of Y.

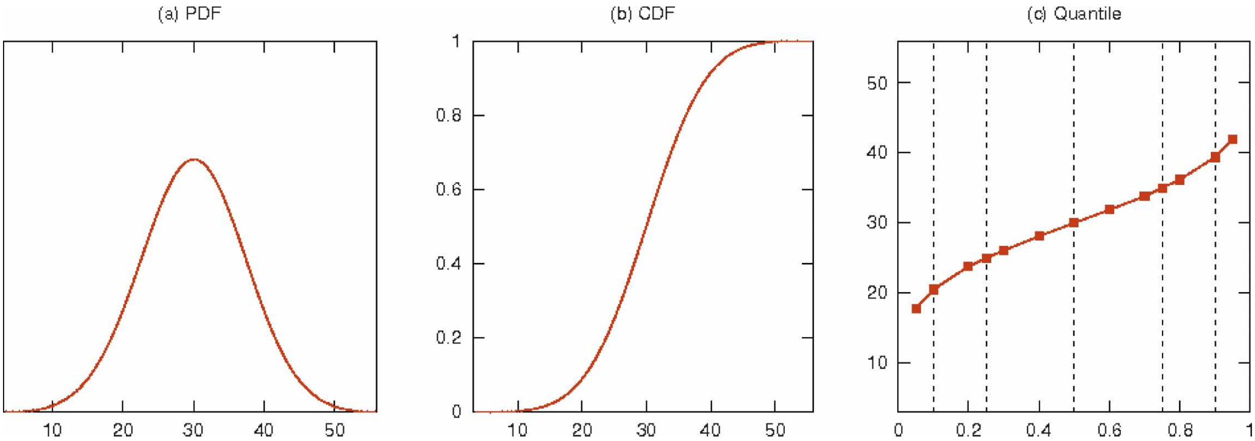

FIG. 1. An example of an (a) PDF, (b) CDF, and (c) a quantile function (shown for the 120-h forecast issued at 0000 UTC 21 Jan 2007 for Kansas City, MO); *T* (°F) represented on the *x* axes of the PDF and CDF and the *y* axis.

3. Data available

The National Centers for Environmental Prediction (NCEP) saved a sample of their Global Ensemble Forecast System (GEFS) from May 2004 until the present. This archive was made for the purpose of model improvement and not postprocessing into operational products; therefore, it is not optimum for the latter purpose. The data were retrieved and put into a format conducive to processing by MDL software. Consistent with other MOS techniques, the data were divided into 6-month cool seasons; two seasons were used for development (October 2004–March 2005 and October 2005–March 2006) and the third season (October 2006–March 2007) was used for validation. Our study was limited to the 0000 UTC forecast cycle, since it was the most complete dataset. It is noted that several changes were made to the ensemble system during the sample period. Changes were made to the model, the model resolution, and the resolution at which the data were archived. Notably, the method of establishing initial conditions was changed between the developmental and independent datasets. Therefore, some disagreement between samples would be expected, even discounting any possible climatic change over that period. The number of ensemble members was increased from 11 to 15 starting at 1200 UTC 30 May 2006, and increased again to 21 starting at 1200 UTC 27 March 2007. To remain consistent with the dependent dataset, only the first 11 members were used as the independent dataset.

MDL maintains an archive of observations at regularly reporting sites within the United States; these observations furnished the predictand data. The NWS forecasts MaxT and MinT, but these variables are not directly observed. We calculated them from the hourly observations and the 6- and 12-h reported maximum and minimum values. A set of 1650 stations was chosen for development and testing, generally matching those stations used in recent MOS developments. Stations were distributed throughout the conterminous United States (CONUS), Alaska, Hawaii, and the U.S. territories. In all discussions, the development (or dependent) data sample consists of the two cool seasons and 1650 stations. The independent sample consists of one cool season for the same stations.

4. Presentation of probability forecasts

A probabilistic forecast of a dichotomous event, as defined in section 1, is easily stated or communicated. It is just a single value, like 20%. However, when dealing with a continuous variable, a CDF, or a few values from it, is needed to communicate the forecast to the user community.

An example of a CDF (Fig. 1b), the probability density function (PDF) from which it was derived (Fig. 1a), and an associated quantile plot (Fig. 1c) are shown. These are presented in terms of temperature. Here, the CDF represents the probability that a particular value of temperature will not be exceeded. The full CDF may not be known, but rather specific values on it. These can be represented on a quantile plot. In Fig. 1c, one can see that the following 13 probability values have been used to delineate the probability distribution: 0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.90, and 0.95. These points are marked with squares in the figure. Dashed lines have been added at 0.10, 0.25, 0.50, 0.75, and 0.90 to enhance readability. The selection of these 13 values is important since it has an impact on the evaluation techniques described be-

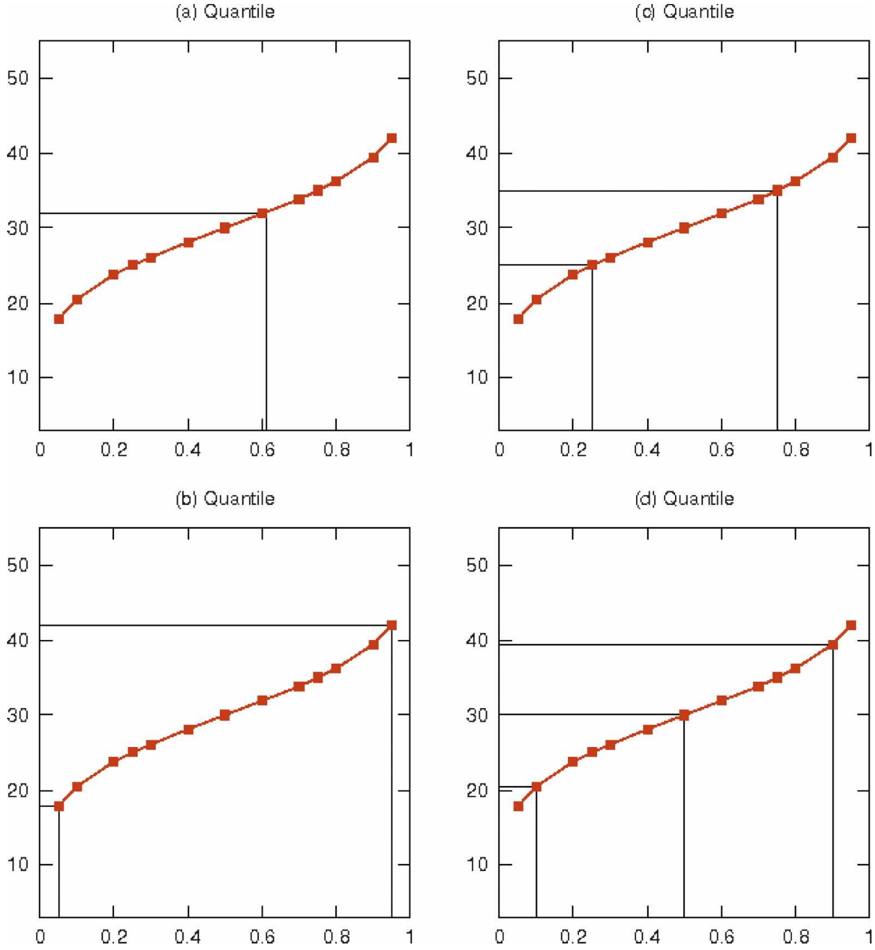

FIG. 2. Examples of quantile plots (shown for the 120-h forecast issued at 0000 UTC 21 Jan 2007 for Kansas City, MO): (a) how to calculate the probability of the temperature being at or below freezing, (b) how to compute the 90% credible interval, (c) how to calculate the 50% credible interval, and (d) how to compute the cold and warm tails.

low. The nine evenly spaced probability values (0.10, 0.20, etc.) were chosen to provide a basic outline of the distribution. The two quartile boundaries (0.25 and 0.75) were added to this set so they would not have to be interpolated. Finally, two values were added (0.05 and 0.95) to give additional definition to the tails of the distributions. These last two values may be useful to users as upper and lower bounds for any numerical processing they perform on the distributions.

The 13 points shown in Fig. 1c can be manipulated to yield a considerable amount of information about the forecast. For simplicity, we adopt the subscript notation such that T_p is defined to be the nonexceedance temperature for a given p, such that

$$\Pr(T \leq T_p) = \frac{p}{100}. \tag{14}$$

Figure 2a illustrates how one can interpolate T_{60} and T_{70} to estimate that the probability of a temperature below freezing is 0.61. Figure 2b shows that the 90% confidence interval can be determined to be 24.2°F (42.0°F–17.8°F). Figure 2c shows that the 50% confidence interval is 10.0°F (35.0°F–25.0°F). We find it useful and intuitive to define the terms "warm tail" and "cold tail" as ($T_{90} - T_{50}$) and ($T_{50} - T_{10}$), respectively. Figure 2d illustrates these terms. Note that the warm tail and cold tail are nearly equal, suggesting a symmetric forecast PDF.

5. Methods of evaluation

Whether a probability forecast has been made in a Bayesian framework or from a frequentist point of view, it is universally accepted that the forecast should be unbiased. That is, when the forecast of a well defined event is 20%, and that same forecast is made numerous times, then the relative frequency (RF) of the event should approach 20% as the number of forecasts increases. So, reliability is a primary component of our evaluation. Reliability is easily calculated for an event, as defined in section 1, but the evaluation of a CDF is more challenging. We have chosen three methods to represent reliability. One is the probability integral transform (PIT) histogram (Gneiting et al. 2005). The PIT is essentially the value of the CDF at the value that is observed (Czado et al. 2007). A histogram can be generated from the PITs, given a sample large enough to support the number of probability bins in which the observations are counted. The PIT histogram should be uniform, and the calculated RF for each bin should be unity. Visually, a PIT histogram shares many characteristics with the rank histogram (also known as the "Talagrand diagram"). Hamill (2001) provides useful information about the interpretation of rank histograms, much of which can be applied directly to PIT histograms. The shape of both histograms gives a visual way to identify biases, and under and overdispersion.

Another way for visually identifying departures from reliability is to plot the cumulative observed relative frequency against the cumulative probability; this gives a cumulative reliability diagram (CRD). Reliability on a CRD can be evaluated in a similar fashion as on a reliability diagram (see Wilks 2006b, p. 287): RFs to the right of the dashed diagonal reference line (which indicates perfect reliability) show overforecasting (forecast cumulative probabilities were higher than the associated cumulative frequencies), while RFs to the left show underforecasting. Note, however, that this is *cumulative*—a summation from below the bins that are represented in a PIT histogram and that would be represented on a reliability diagram.

The discrete nature of the surface temperature and dewpoint observations gave rise to a complication in the generation of PIT histograms and CRDs. Both temperature and dewpoint are generally reported in units of tenths of a degree Celsius. Our data have been converted to Fahrenheit and rounded to whole degrees. For those cases where a forecast distribution exhibits a small variance, the verifying observation can satisfy the criteria for more than one bin in the histogram. To manage this situation and the biases to which it can lead, we generate a random number from the continuous uniform distribution $U(-0.5, +0.5)$ and add it to the observed value before computing its PIT. This technique eliminates integers from the subsequent comparisons. This procedure is analogous to the one described by Hamill and Colucci (1998) for dealing with "ties" when generating rank histograms.

While the PIT and CRD give a visual check on the reliability, it is useful to summarize the reliability information in a single value. We compute a negatively oriented measure we call the squared bias (SB) in RF, which is the squared difference between the RF and unity, weighted by the width of the probability bin, summed over the entire range of probabilities on a PIT histogram.

The other measure we have used to evaluate the forecasts is the continuous ranked probability score (CRPS). This measure of accuracy (also negatively oriented) is well known and will not be explained in detail here (e.g., see Matheson and Winkler 1976; Unger 1985; Hersbach 2000). Suffice to say, it is a squared measure of the difference between the CDF and the verifying observation and degenerates to the mean absolute error (MAE) in the case of single value (nonprobabilistic) forecasts. CRPS is measured with the units of the weather element, which in this work are degrees Fahrenheit.

6. Baseline for evaluation

The 11 forecasts from the GEFS, referred to hereafter as the raw ensembles (RawEns), can be rank ordered, and as such provide a crude CDF. This can be done by assigning a probability to each ensemble member using a plotting position estimator attributed to Weibull (see Wilks 2006b, p. 41),

$$\Pr(T \le T_i) = \frac{i}{n+1}, \qquad (15)$$

where i is the rank of the ensemble member and n is the number of ensemble members. For 11 ensemble members, this yields 11 points on the CDF, all evenly spaced on the probability axis. The points derived from this estimator were interpolated to find nonexceedance temperatures for the list of probabilities described in section 4, above. Tails for this CDF were inferred by assuming they were half as wide as their neighboring segments.

Figures 3a–c show, for the 48-, 120-, and 168-h projections from model run time, the PIT histograms for our 1650 stations combined (see section 3) for the RawEns. The time projections were chosen to represent forecast days 2, 5, and 7. The "U" shape of these histograms implies that the ensembles are underdispersed while the higher bar on the right side of each implies a cold bias in the forecasts (meaning too many observations fell outside the ensemble extremes, especially on the warm side). Figures 3g–i show the associated CRDs. While the CRDs do not show the bias and underdispersion as readily as the PIT histograms do,

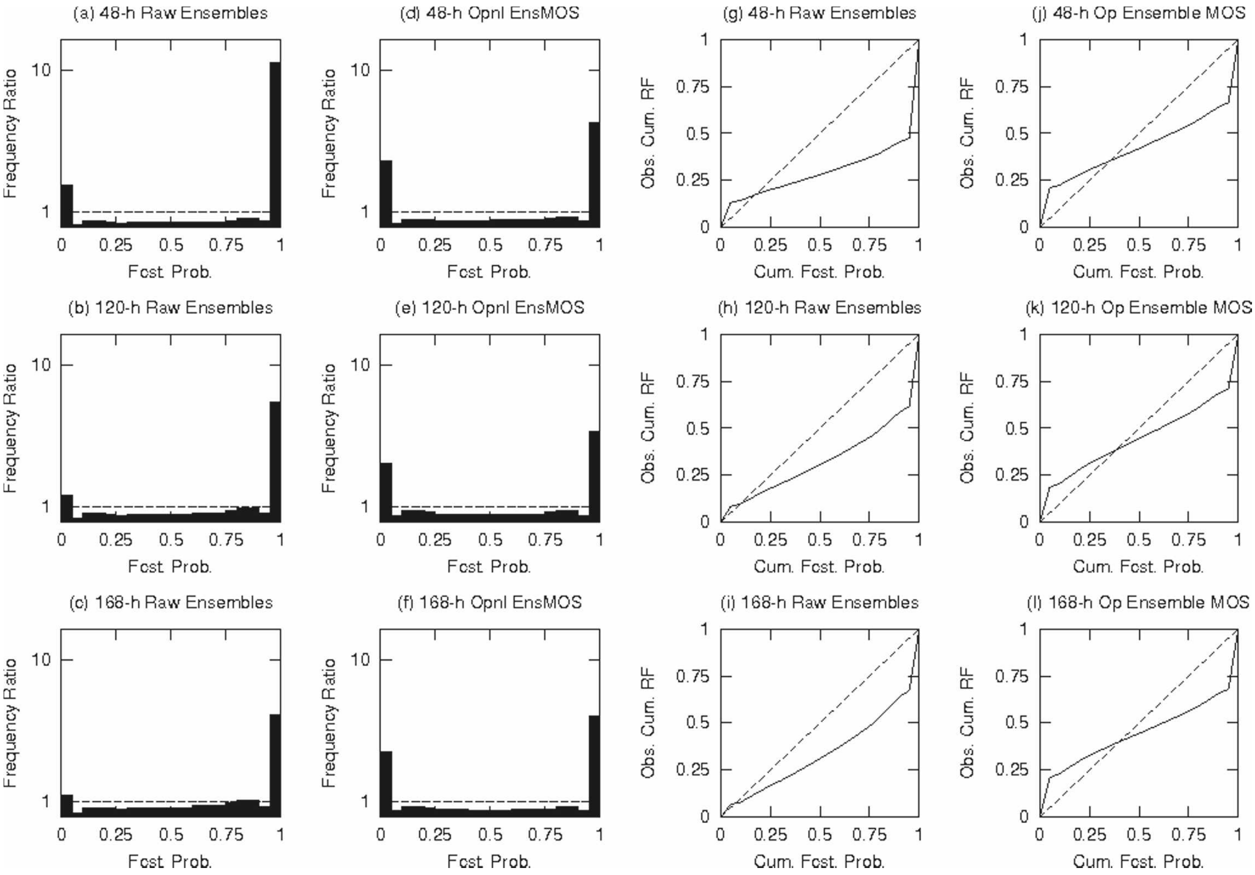

FIG. 3. PIT histograms and CRDs for the raw ensembles and operational ensemble MOS at 48-, 120-, and 168-h projections for two cool seasons covering the dependent sample. The CRDs show the observed cumulative relative frequency (Obs. Cum. RF) as a function of the cumulative forecast probability (Cum. Fcst. Prob.).

the unreliability of the distributions is easier to quantify in the CRDs. It is easy to find regions of the distribution where RF deviates from the ideal by more than 0.15. At some points the deviation exceeds 0.45.

MOS single station equations for temperature based on NCEP's GFS (formerly the Global Spectral Model; Erickson 1996) have been applied to NCEP's raw ensembles and made available to forecasters as operational guidance forecasts for a number of years. As with the RawEns, these can be rank ordered, and evaluated with the techniques described above. Figures 3d–f show the 48-, 120-, and 168-h PIT histograms, respectively, for the ensemble MOS (EnsMOS) and Figs. 3j–l show the CRDs. Like the raw ensembles, the EnsMOS forecasts are underdispersed. The bars on both ends of the histograms are more level, however, when compared to the raw ensembles, showing less bias, as one would expect from a MOS forecast. The CRDs also show bias improvement over RawEns, but are far from reliable. The EnsMOS equations were developed on a single run of the GFS. The application of these equations to individual ensemble runs mimic the developmental model. Because the GFS is underdispersive, the EnsMOS will be also. This characteristic has been noted previously (see Wilks 2006a), and is borne out here. The skill of the EnsMOS equations, developed on an older version of the GFS, is expected to be less than for equations developed on newer data, provided an adequate sample is used. While the exact behavior of "old" equations on "new" data cannot be known for sure, the results here conform to expectations.

Figure 4a compares SB at each 6-h time projection for RawEns and EnsMOS. Note the strong diurnal variation in the RawEns, which is largely removed by the EnsMOS. The SB score for RawEns lowers with increasing time projection. This indicates that the underdispersion in the ensemble output is worse at earlier projections than at later projections, as also indicated in the PIT histograms in Figs. 3a–c. The EnsMOS SB scores drift upward somewhat in the later projections. Since the MOS forecasts trend toward climatology in the later projections, the CDFs generated from them

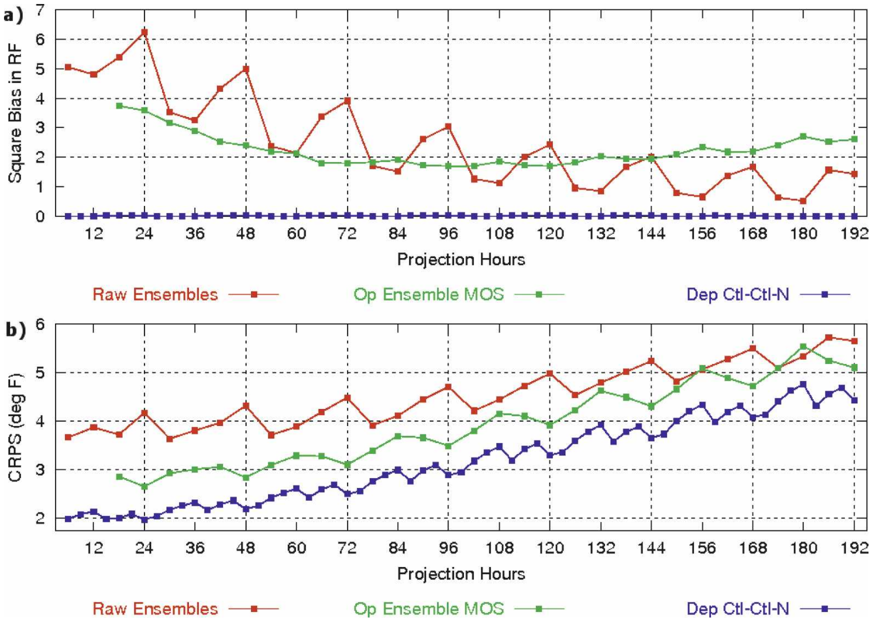

FIG. 4. (a) SB in RF and (b) CRPS for the raw ensembles, operational ensemble MOS, and the Ctl-Ctl-N technique for the dependent sample.

exhibit less variance than the RawEns. This result is unfortunate, but not unexpected, because the MOS ensemble spread narrows around the mean climatological value at long projections, resulting in smaller variance when larger variance is needed.

Figure 4b compares CRPS for RawEns and EnsMOS. For most projections, EnsMOS handily improves over RawEns although a few exceptions can be seen at later time projections. The diurnal variation noted in SB for RawEns can also be seen in CRPS, and for EnsMOS as well, although the variation is out of phase with RawEns. (Previous verification of MOS forecasts has shown that the times of day corresponding to these 24-, 48-, . . . 192-h forecasts have lower MAEs than other times of the day. Figure 4 shows that this is a more difficult time of day for the model forecasts.)

For comparison, both charts in Fig. 4 include scores for a technique named Ctl-Ctl-N, which is described below.

7. Regression equations based on control ensemble run

The development of regression equations based on the single "control" model run, and evaluated on the control run is the most basic test of the probabilistic regression framework. The predictands were temperature and dewpoint, developed simultaneously at 3-h intervals from 6 to 192 h, and MaxT and MinT, developed independently out to 390 and 378 h, respectively. Simultaneous development is a term used to denote selecting predictors in a way that ensures the same predictors are chosen for both elements, enhancing the meteorological consistency of the forecasts (National Weather Service 1985). The potential predictors were taken from the control member of the GEFS and used in a forward selection screening procedure, which selects predictors for the regression equation from a pool of variables based on their additional reduction of variance of the predictand (Lubin and Summerfield 1951); no observations were used as predictors. Table 1 shows the most frequently chosen predictors for each element. We have found that the GEFS does not have much temporal bias, so the predictors were valid at the same time as the predictand for the temperature and dewpoint, and covered a range of projections for the max and min. Because the model data were saved at 6-h resolution, we used a quadratic time interpolation to get projections at 3-h intervals. The name Ctl-Ctl-N was given to this technique to summarize how the equations were developed (on the control or "Ctl" member only), how the equations were applied (to the Ctl member only), and what type of distribution was applied (a normal distribution, abbreviated "N"). Equations (9) and (13), above, give the framework used to determine the two parameters of the normal distribution.

Figure 4 indicates the Ctl-Ctl-N method had very small square bias on the dependent sample and its

TABLE 1. Predictors many times selected by the screening procedure for the development of equations.

Element	Predictors
T, T_{dp}	2-m surface T, 2-m surface T_{dp}, geoclimatic predictors, RH (layer), low- and midlevel lapse rates, thickness fields
Maximum T	2-m surface T at 6- and 12-h intervals, geoclimatic predictors, 850-mb RH at 12-h intervals, 850-mb u and v components of wind at 12-h intervals, 850-mb equivalent potential temperature at 12-h intervals
Minimum T	2-m T at 6- and 12-h intervals, geoclimatic predictors, 850-mb RH at 6-h intervals, 2-m T_{dp} at 6-h intervals, 850-mb equivalent potential temperature at 12-h intervals

CRPS was considerably lower than for RawEns or EnsMOS.

Figure 5 shows the regression for a 24-h prediction of temperature at Milwaukee, Wisconsin, based on a single predictor, the GFS 2-m temperature. Prediction intervals are shown for 50% and 95%. The regression error estimation is seen to fit rather well. While not noticeable, the prediction intervals are not bounded by straight lines, but rather by hyperbolae [see Eq. (2)]. Thus, our predictions are less certain as we move away from the predictor mean.

Figures 6a–c show the PIT histograms for our station set for the 48-, 120-, and 168-h projections, respectively, for dependent Ctl-Ctl-N data. When compared to the PIT histograms for the raw ensembles and of ensemble MOS (Fig. 3), these histograms appear relatively flat, showing an increase in reliability. Please note the scales for Figs. 3 and 6 are different by about a factor of 8. The associated CRDs are not shown, but they plainly show the increase in reliability with deviations from the diagonal generally less than 0.03. These results indicate that the normality assumption in the regression framework holds quite well for temperature.

8. Development based on mean of ensemble forecasts

The same procedure as described in section 7 was used to develop equations based on the means of the 11 ensemble variables (mean equations). The equations were then applied to those means to make forecasts. This technique was given the name "Mn-Mn-N." Figures 6d–f show the PIT histograms for dependent Mn-Mn-N data at the same projections as those from the control run from the previous section. The Mn-Mn-N technique does not seem to improve the reliability of the forecasts over Ctl-Ctl-N; however, below we will show that forecasts made by Mn-Mn-N are more accurate at the longer projections, but slightly less accurate at shorter projections.

9. Mean equations applied to each ensemble member

The results above show that the regression framework gives quite well-behaved distributions, but they are symmetric and unimodal. To attain the full benefit of ensembles and to achieve nonnormal distributions, the individual members must be used. All members have been used in developing the mean equations, so we now apply the mean equations to each member individually. This gives 11 forecasts, F_i, each with an error estimate, σ_i. We combine these using kernel density fitting [(or estimation (KDE); Wilks 2006b]. We use a normal kernel in concert with the regression framework, and the 11 kernel functions are generated using F_i and σ_i. This, then, provides a CDF that can be nonsymmetric and multimodal. We named this technique Mn-Ens-KDE, indicating that the forecast equations were developed using the ensemble mean, the forecast equations were applied to each ensemble member individually, and KDE was used to combine the various forecasts into a single distribution.

Figures 6g–i show the PIT histograms obtained from the Mn-Ens-KDE technique. The Mn-Ens-KDE forecasts are much more reliable than RawEns or EnsMOS (see Fig. 3), but they are not as reliable as the Mn-Mn-N or Ctl-Ctl-N forecasts. In particular, the Mn-Ens-KDE forecasts are overdispersed. Figure 7a compares SB for Ctl-Ctl-N, Mn-Mn-N, and Mn-Ens-KDE techniques. Note that the scales for Figs. 4a and 7a are quite

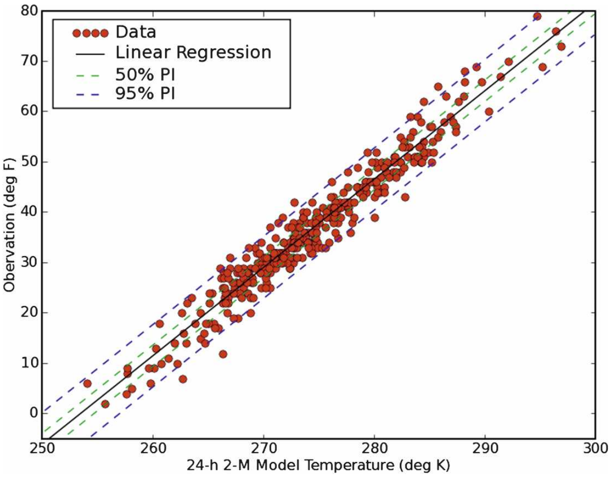

FIG. 5. Regression equation predicting T based on the 2-m model T, plotted data, and the 50% and 95% prediction intervals for Milwaukee, WI, for the dependent sample.

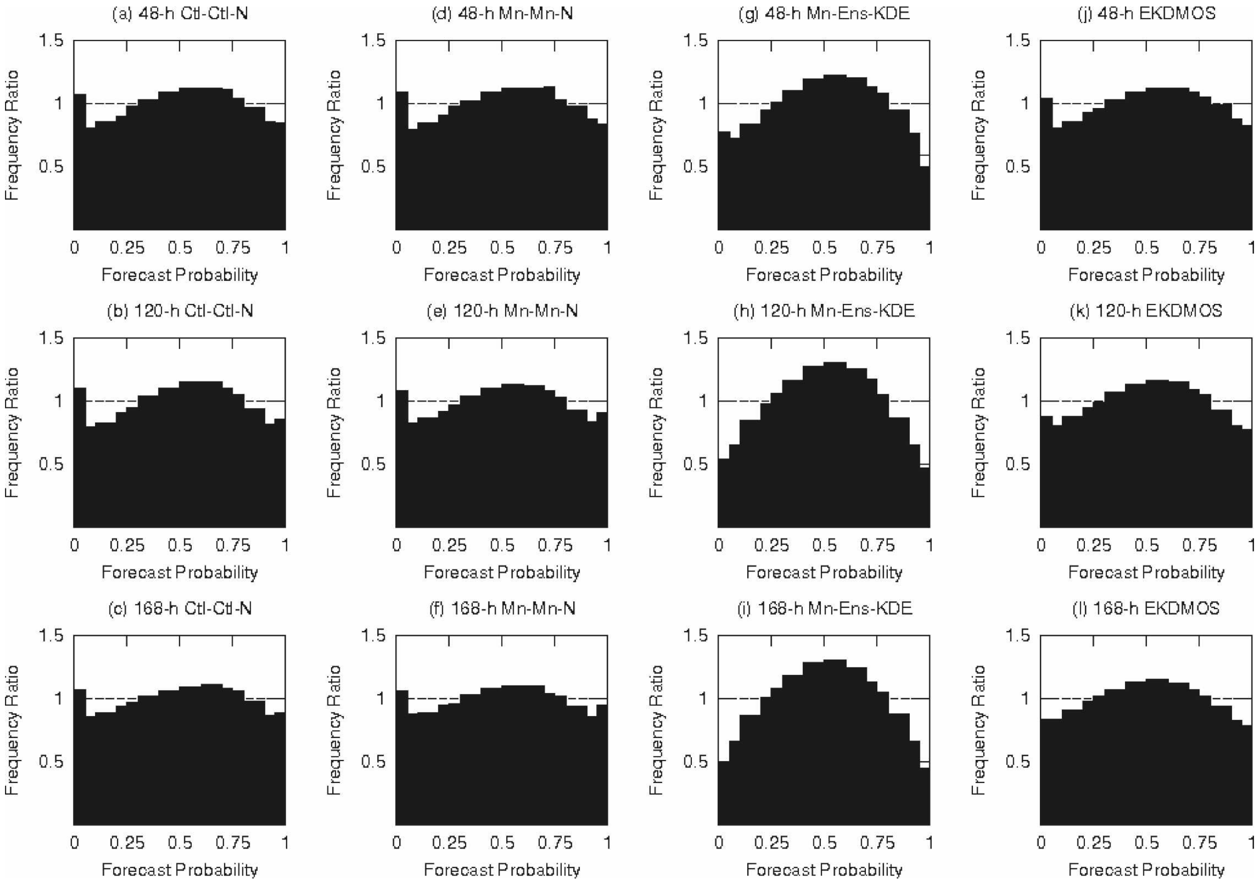

FIG. 6. PIT histograms for the Ctl-Ctl-N, Mn-Mn-N, Mn-Ens-KDE, and EKDMOS techniques for the dependent sample.

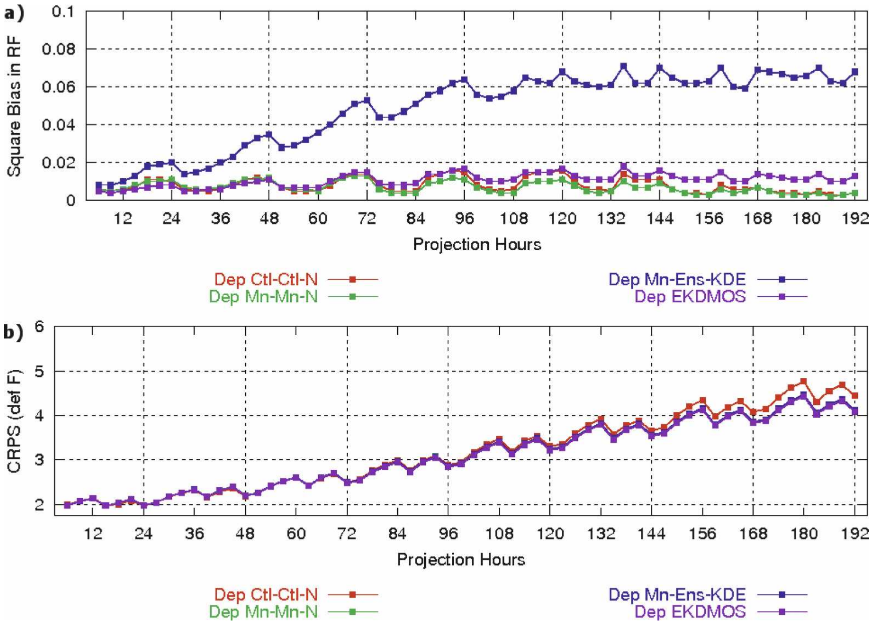

FIG. 7. (a) SB in RF and (b) CRPS for the Ctl-Ctl-N, Mn-Mn-N, Mn-Ens-KDE, and EKDMOS techniques for the dependent sample. In (b), the Mn-Mn-N plot is masked by the Mn-Ens-N plot.

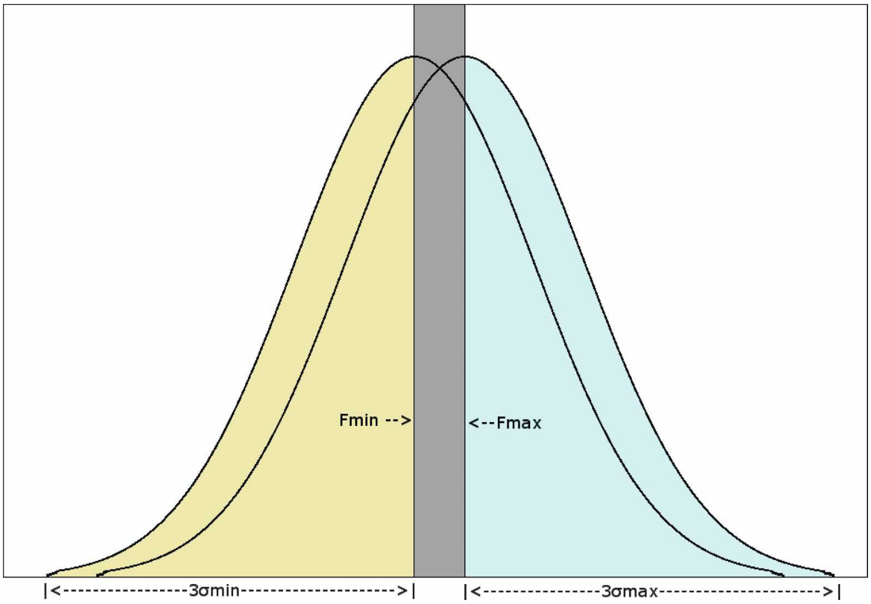
Fmin -->
<--Fmax
|<----------------3σmin--------------------->|
|<---------------3σmax---------------------->|

FIG. 8. Example of EKDMOS spread adjustment. The dispersion is adjusted by changing the total width of the (combined) distribution by a factor sf that is a function of the width of the gray area.

different, the difference being almost two orders of magnitude. While all these ensemble MOS results have a low square bias in comparison to the raw ensembles, using the 11 members gives higher (worse) bias than Ctl-Ctl-N or Mn-Mn-N. This was expected because when one ensemble member has about the correct spread, adding the spread of the ensemble members is bound to cause overdispersion.

Figure 7b compares CRPS for the same techniques and time projections as Fig. 7a. At this scale, there is little discernable difference in their accuracy as measured by the CRPS. However, both Mn-Mn-N and Mn-Ens-KDE improved on Ctl-Ctl-N beyond about 108 h, which indicates the value of ensembles. Also, a closer look on an expanded scale shows Ctl-Ctl-N to be slightly better at projections less than 72 h.

10. Adjustment of dispersion

Since the spread of the Mn-Mn-KDE distribution is too large, we decreased the spread by a factor based on the spread between the most different ensemble members. The adjusted distribution has a spread that is smaller than the original distribution by a factor of $(1 - x)$, based on Eq. (16):

$$x = \frac{3(\sigma_{\min} + \sigma_{\max}) + \mathrm{sf}(F_{\max} - F_{\min})}{3(\sigma_{\min} + \sigma_{\max}) + (F_{\max} - F_{\min})}, \qquad (16)$$

where $F_{\min}$ and $F_{\max}$ are the smallest and largest ensemble forecasts, respectively; $\sigma_{\min}$ and $\sigma_{\max}$ are the associated standard deviations of their kernels; and sf is a factor that can be used to tune the process. This is illustrated in Fig. 8. When the ensembles are tightly packed, very little adjustment is made; when they are widely dispersed, the adjustment is more pronounced. At one extreme, if the value and spread of each member were the same (or if there were only one member), no adjustment would be made, and the spread would be the same as for a single member.

We found through testing on dependent data that a spread factor (sf) of 0.5 produced distributions that were only slightly overdispersed. Once the width of the PDF is adjusted, the height of the curve is then normalized so that the area under the curve is unity. This technique—Mn-Ens-KDE with an adjustment to dispersion—was given the name "EKDMOS" for "Ensemble-Kernel Density-MOS."

Figures 6j–l show the PIT histograms obtained from the EKDMOS technique. These plots show that much of the overdispersion that was present in the Mn-Ens-KDE technique has been corrected. Figure 7a shows the SB for the EKDMOS technique to be on scale with the single member techniques. Figure 7b shows that accuracy as defined by the CRPS was not noticeably changed by making the adjustment; however, viewed in greater detail, the EKDMOS was slightly superior beyond about 72 h.

Figure 9 compares PIT histograms and CRDs generated with the EKDMOS technique on both dependent and independent data. The slight overdispersion that is

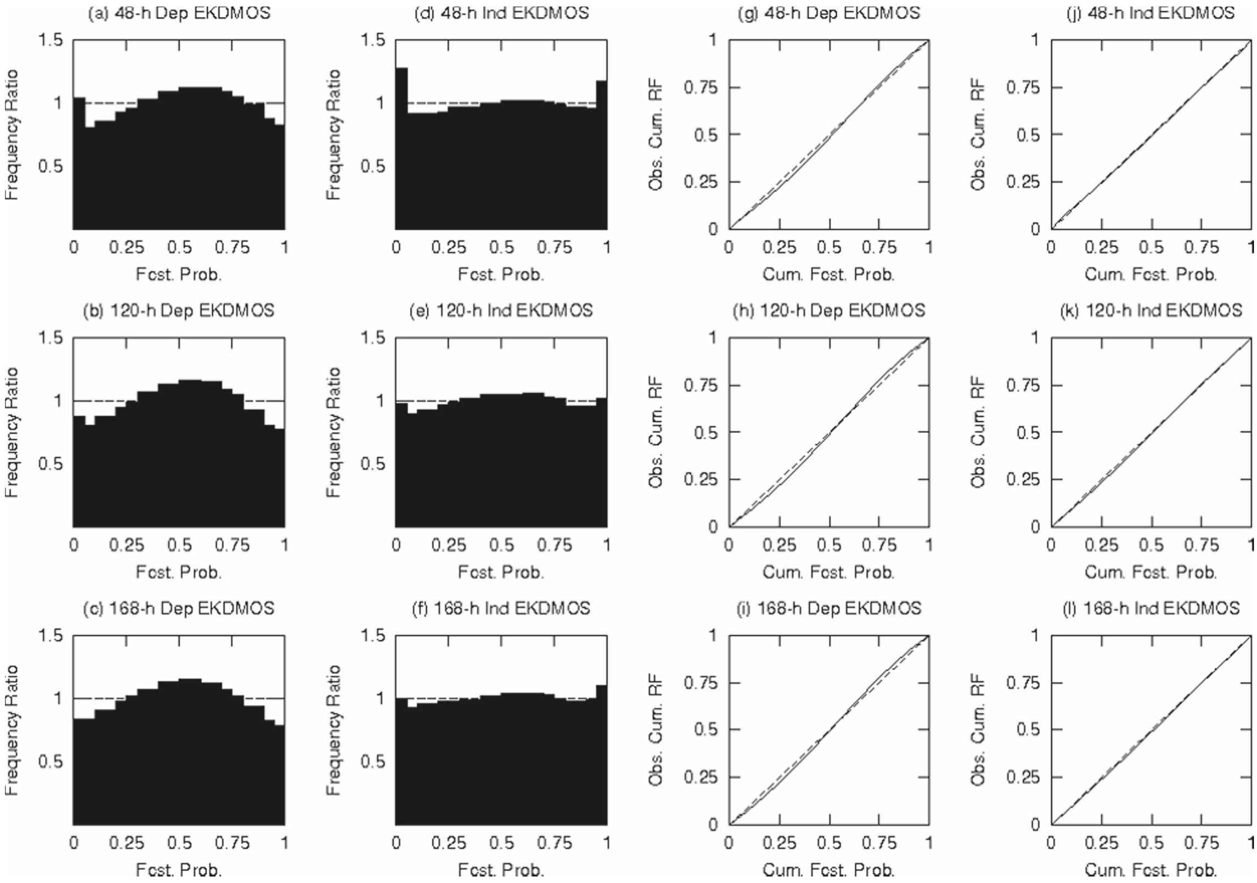

FIG. 9. EKDMOS PIT histograms and CRDs for both dependent and independent samples. The CRDs show the observed cumulative relative frequency (Obs. Cum. RF) as a function of the cumulative forecast probability (Cum. Fcst. Prob.).

present in the dependent data PIT histograms (Figs. 9a–c) has largely disappeared in the independent data (Figs. 9d–f), especially at the later projections. The maximum deviations present in the CRDs have also decreased from 0.025 in the dependent data (Figs. 9g–i) to 0.01 in the independent data (Figs. 9j–l).

Figure 10a compares the SB for EKDMOS for dependent and independent data. The fluctuations in the independent data in the early projections can be explained by the slight overdispersion at some projections mentioned above. The CRPS plots in Fig. 10b show that the accuracy of independent data is within 0.2°–0.4°F of the dependent data, and the independent data results are better than the raw ensemble shown in Fig. 4 by 1.0°–1.2°F. Although not shown here, SB and CRPS were computed for all four techniques (Ctl-Ctl-N, Mn-Mn-N, Mn-Ens-KDE, and EKDMOS) for the independent data sample. The application to independent data had little impact on the relative performance of the techniques. However, expanded scales for Fig. 7 indicate the Ctl-Ctl-N method is slightly superior to other methods, including EKDMOS, at most early projections, being about equal at 72 h, and being less accurate at long projections. As stated previously, the Ctl-Ctl-N method used only the high-resolution control member, while the other techniques used all members, all but the control member being at a lower resolution; this bears on the question of trade-off between resolution and a larger number of lower-resolution ensembles. Most differences between CRPS scores of EKDMOS and other techniques were highly significant as judged by a paired *t* test.

One can wonder why the bias characteristics were better on the independent data than the dependent data (see Fig. 9); was it fortuitous, or was it because we purposely set the sf value so that there was still some overdispersion on the dependent data? We thought that because the regression would likely not fit the independent data as well as the dependent data, the dispersion would, on average, increase.

Although the EKDMOS produced nonsymmetric distributions and we were able to control the bias, it showed only slight, but statistically significant, improvement over the simpler technique Mn-Mn-N in terms of

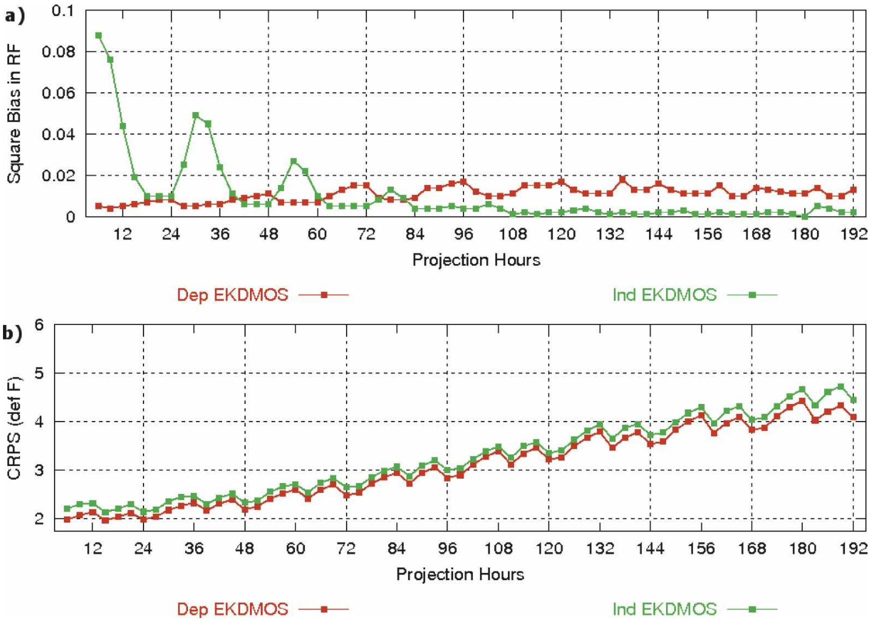

FIG. 10. (a) SB in RF and (b) CRPS for EKDMOS for both dependent and independent samples.

the CRPS. Likely, the majority of the distributions are quite close to normal and overshadow the few that depart significantly; the CRPS is not sensitive enough to register much improvement. It also appears the CRPS is much more sensitive to the mean of the distribution than to its dispersion. This does lead to the question, though, as to whether the nonsymmetric distributions, as derived from the ensemble individual members, really were an improvement. Perhaps all the information is contained in the mean and standard deviation from the regression, and the distribution of the members is not furnishing additional information. This deserves further study.

As a final comparison, we examine the results of EDKMOS temperature forecasts using more conventional measures widely used to evaluate single-valued forecasts. The simplest way to extract a single-valued forecast from a probability distribution is to use the nonexceedance value at the 0.50 level (the median of the distribution, T_{50}). Figure 11 shows the bias (Fig. 11a) and mean absolute error (Fig. 11b) values computed for T_{50} for the two baseline techniques and EKDMOS verified over the independent test sample. Results for the operational GFS MOS T forecast are included for comparison. Figure 11a shows that the raw ensemble forecast exhibits a considerable cold bias that varies diurnally. All three of the MOS-based techniques manage to correct much of this bias with EKDMOS being best in this regard. Figure 11b shows the same general result as Fig. 4b with the MOS-based techniques generally improving on the error performance of the raw ensembles. As in Fig. 7b, EKDMOS performs better than the other MOS-based techniques especially after day 5. It is interesting to note that, for this sample at least, the EKDMOS forecasts perform better than the operational GFS MOS; results for T_{dp} are not presented here, but are similar. Also not shown, the Ctl-Ctl-N method gave slightly lower MAEs for temperature than EKDMOS (between 0.05° and 0.1°F) at short projections, being about the same at 72 h, and larger by up to 0.5°F at 192 h. While these differences are small under 72 h, it is interesting that the single control member produced better results than all combined and the differences are highly significant as judged by the paired t test. Of course, this is with our methods; other methods could produce other results.

11. Results for dewpoint and maximum and minimum temperature

The EKDMOS technique was also applied to dewpoint, MaxT, and MinT. All results shown here for these elements are for the independent sample. Figure 12 compares the PIT histograms for these three elements. While some bias in the distributions appears to exist in the early projections (Figs. 12a,d,g), the corresponding CRDs (not shown) show that this technique is still quite reliable for these elements, with a maximum deviation of less than 0.05. The SB scores were generally below 0.02 at all time projections.

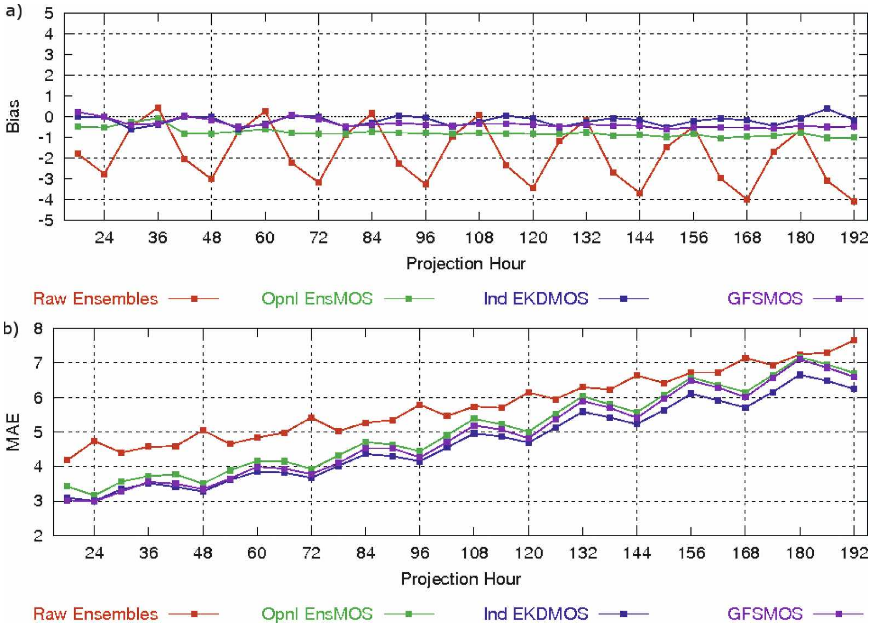

FIG. 11. (a) Bias and (b) MAE computed for T_{50} for the raw ensembles, operational ensemble MOS, and EDKMOS techniques for the independent sample. Results for GFS MOS for the same dates are plotted for reference.

Figure 13 shows the CRPS for T_{dp}, MinT, and MaxT at each available time projection. In general, these scores are consistent with those seen for T. The results for these three elements were obtained with *exactly* the same process as with temperature; the sf determined on dependent data for temperature was used unchanged as 0.5. This shows the technique to be quite robust for these quasi-normally distributed variables.

12. Sample PDFs

We present a case study to demonstrate these techniques operating in a “real world” meteorological setting (i.e., an interesting temperature contrast across southern Alaska). PDFs are displayed to aid in the visualization of the distributions. We find PDFs to be more useful for visualization and CDFs more useful when seeking quantitative answers to specific questions.

Figure 14a shows NCEP’s surface analysis for North America valid at 0000 UTC 28 November 2006, and Fig. 14b shows the locations of Juneau and Homer, Alaska. Note that the two stations were influenced by two different weather regimes. A cold dome of high pressure centered over the Yukon kept Juneau unseasonably cold. Southerly flow ahead of an occluded low pressure system brought maritime air from the Gulf of Alaska to Homer.

Figure 15 shows a time series of eight EKDMOS PDFs all forecasting the 0000 UTC temperature at Homer on 28 November 2006 along with the verifying observation. A climatological normal is not available for this particular weather element (temperature at a given time). Since 0000 UTC occurs during the midafternoon, we use the climatological normal MaxT of 33°F for Homer for this date to represent the normal temperature for this time. The earliest forecast (192 h) shows a mode that is remarkably close to the verifying observation of 36°F; however, the mean is colder because of the heavy (cold) tail of the distribution on the left side. The mode of the next forecast (168 h) is not so close. The six subsequent PDFs seem to quickly converge on the verifying observation. Note that the first three PDFs show a visible skew toward colder temperatures. In two of these cases the skew is oriented toward the climatological normal value. It is interesting to see the progressive decrease in forecast variance as lead time decreases; statistically, this is expected.

Figure 16 shows a similar time series of PDFs forecasting the 0000 UTC temperature at Juneau for the same date. The climatological normal MaxT for this date at Juneau is 35°F. Obviously, the forecasts for Juneau do not converge as well as they did for Homer, especially at 24 h. Note that the 192-h forecast has a mode that is close to the climatological normal and is skewed toward colder temperatures. The next two PDFs (168 and 144 h) are closer to the verifying obser-

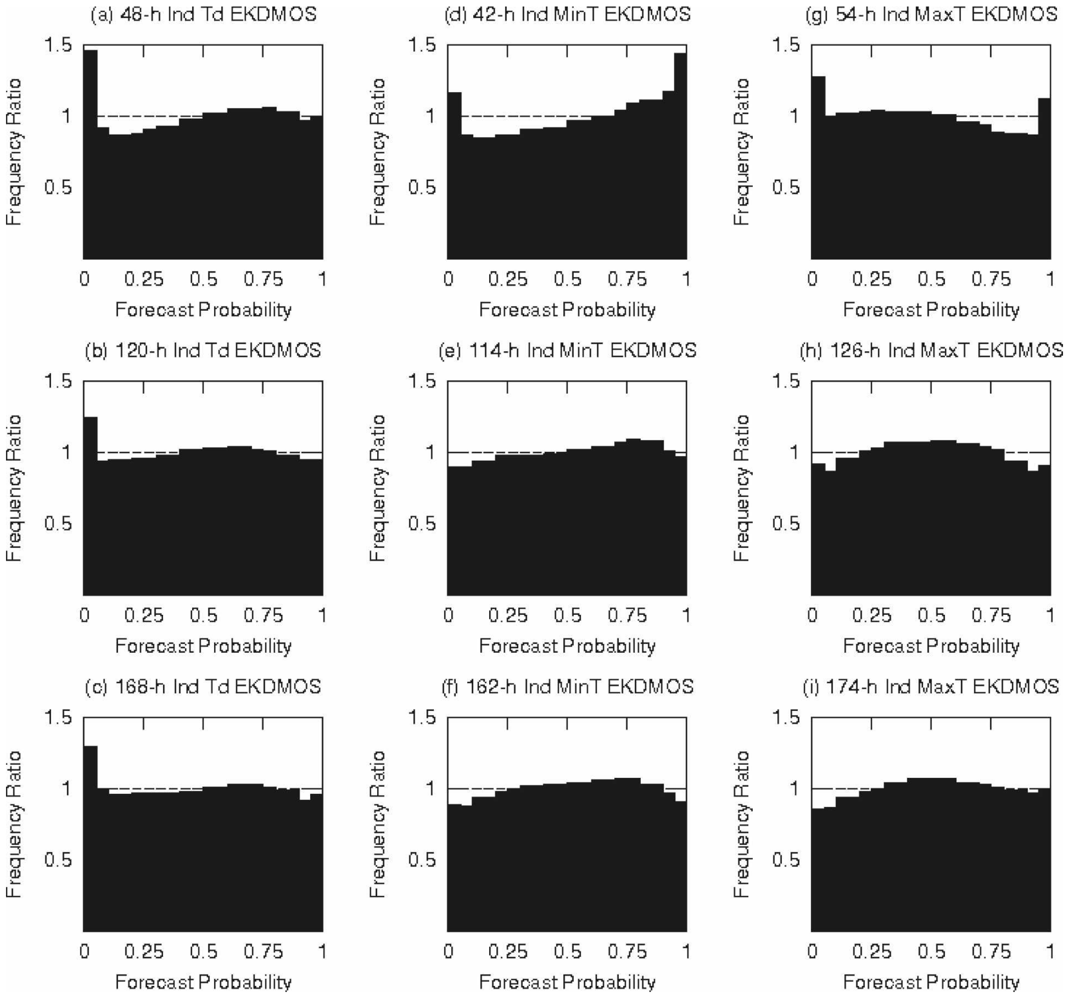

FIG. 12. PIT histogram for EKDMOS dewpoint, MinT, and MaxT for the independent sample.

vation, but are skewed in the direction of climatology. The rest of the PDFs show decreasing variance as lead time decreases, but the 24-h mean and mode forecast was in error by about 8°F.

Visual examination of the PDFs for these two cases show that probability distributions created with EKDMOS exhibit characteristics that are consistent with their respective lead times, climatologies, and meteorological scenarios.

13. Gridded forecasts

The forecasts presented above have all been valid for stations. Both the private and public sectors of the weather enterprise have seen a growing demand for digital forecasts in gridded form (see Glahn and Ruth 2003). To support this demand, we are producing analyses of our station-based EKDMOS using the gridded MOS approach (Glahn et al. 2008). Figure 17 shows an image generated from one of these gridded forecasts. It is the median temperature forecast generated from the 0000 UTC run of the GEFS on 5 February 2008 and for 1500 UTC 5 February 2008 for the CONUS. One can readily locate a frontal boundary across the midwestern CONUS as well as temperature associated with the Rocky Mountains and Appalachian Mountains.

We expect much of the value of EKDMOS will come from better-informed economic decisions. Given reliable probabilistic forecasts, weather can be incorporated into decision models such as the cost–loss model introduced by Thompson (1952). Figure 18 illustrates a case where weather impacted the local economy—a freezing event over northern Florida on the morning of 3 January 2008. That morning a story in *The St. Petersburg Times* read "Floridians race to prepare plants and people for what is predicted to be an icy, but short, blast" (Wesner 2008). The day before, the financial Web site Bloomberg.com reported "Orange-juice fu-

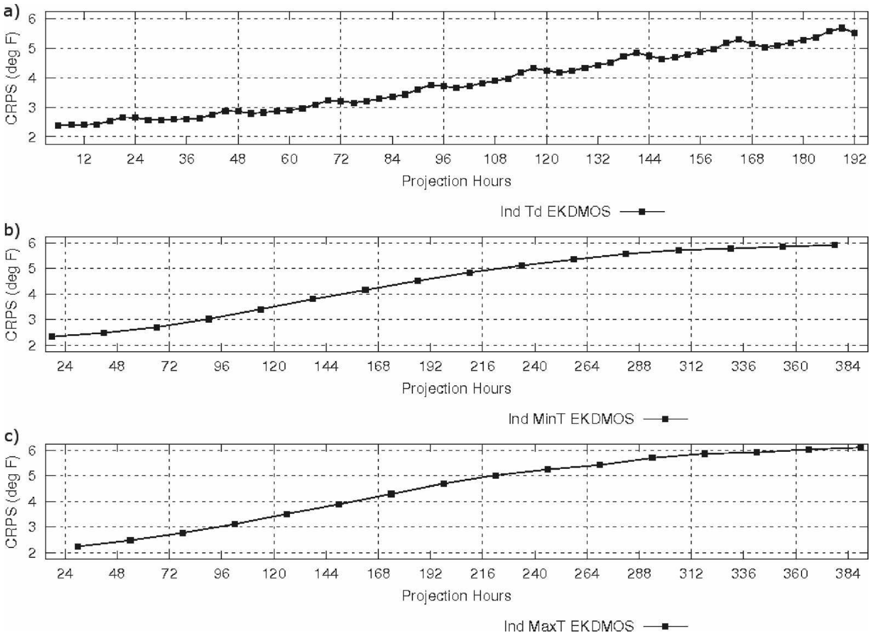

FIG. 13. CRPS for EKDMOS dewpoint, MinT, and MaxT for the independent sample. Note that the projection scales are different for dewpoint than for MinT and MaxT.

tures for March delivery rose $0.04 or 2.8%, to $1.488 a pound" (Day 2008). The MinT forecast for that morning was big news in Florida.

Figure 18 is a series of gridded probabilistic EKDMOS MinT forecasts for Florida, all valid 3 January 2008. Figure 18j shows the verifying analysis. There is no simple, obvious way to display gridded forecasts that are probability distributions; Fig. 18 explores one option. The analyses in Figs. 18d–f show the median of the EKDMOS forecasts (MinT_{50}). Figures 18a–c show the cold tail ($\mathrm{MinT}_{50} - \mathrm{MinT}_{10}$), and Figs. 18g–i show the warm tail ($\mathrm{MinT}_{90} - \mathrm{MinT}_{50}$) of the forecast probability distribution. Lead times decrease from the top to bottom in this figure and are valid for days 7 (Figs. 18a,d,g), 5 (Figs. 18b,e,8h), and 2 (Figs. 18c,f,i). All the forecasts indicate a dangerous freezing event for northern Florida and generally verify well. The size of the tails decreases with time, indicating an overall decrease in the dispersion of the distributions. The median temperature forecasts decrease with time as well, but this is more subtle. One can discern a warm skew in the day 7 forecasts (warm tail larger than cold tail; a trend toward climatology, perhaps) and a cold skew in the day 2 forecasts (indicating the possibility of colder temperatures).

14. Conclusions and summary

Raw ensembles have proven to be notoriously underdispersed. A method has been developed to postprocess ensemble data that yields forecasts that have very good bias characteristics for weather elements that have a quasi-normal distribution. The method was developed on two cool seasons of temperature data for 1650 stations and tested on one season. Even the most basic regression application (the Ctl-Ctl-N) shows dramatic improvement in bias and CRPS (see Fig. 4).

The full EKDMOS method encompasses four significant elements:

1) screening multiple regression applied to the means of the variables forecast by the ensembles,
2) estimation of the error variance directly from the regression,
3) application of the equations to individual ensemble members and combining the results with the kernel density fitting in which a Gausian kernel with the standard deviation produced by the regression is used, and
4) a spread adjustment based on the spread of the ensemble members.

This technique seems to be quite robust, as the method was developed on temperature data and then tested without change on independent data for dewpoint, maximum temperature, and minimum temperature. It also surmounted the several changes in the model, its running and archival resolutions, and initial conditions made over the course of the 3-yr period.

These results have been presented in terms of PIT

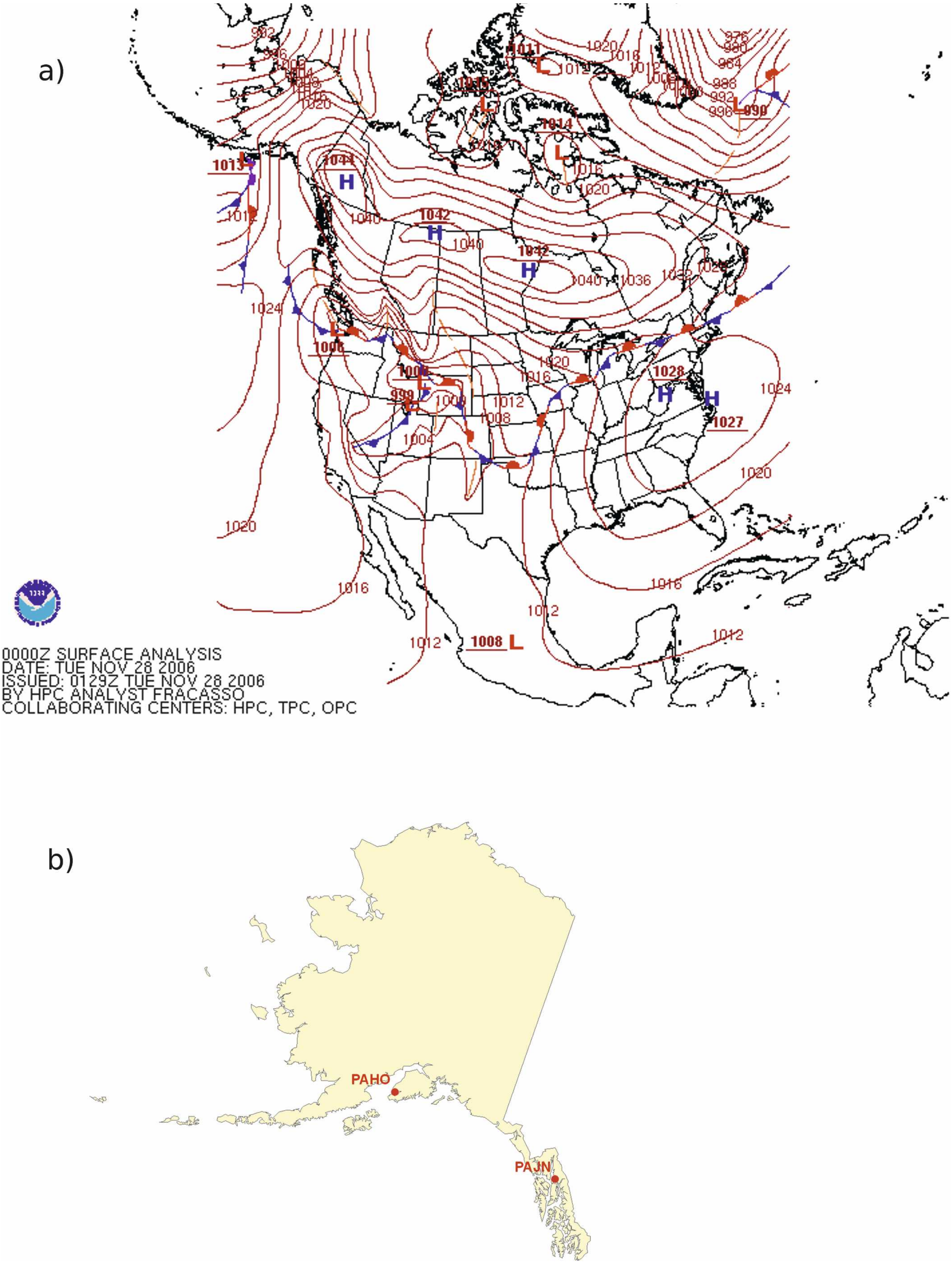

FIG. 14. (a) Surface analysis for North America for 0000 UTC 28 Nov 2006. (b) Locations of Homer, AK (PAHO), and Juneau, AK (PAJN), are displayed.

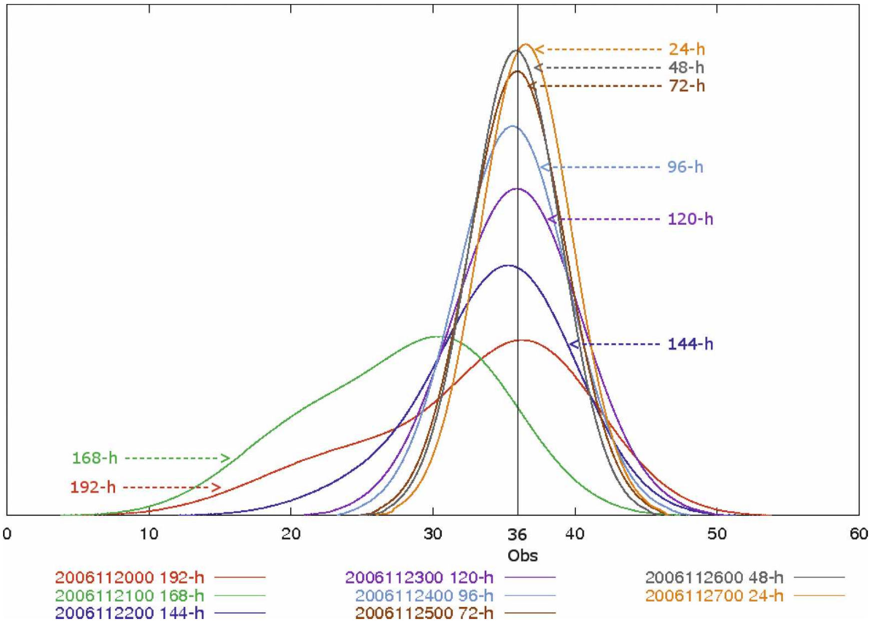

FIG. 15. PDFs for Homer, AK (PAHO). All of the PDFs above verify at 0000 UTC 28 Nov 2006. The temperature observed at this time was 36°F, as indicated by the vertical black line.

histograms, a square bias measure (SB), and a cumulative reliability measure (CRD), which indicate quite good reliability; the accuracy is judged by the continuous ranked probability score (CRPS).

Points on the CDF have been mapped to the National Digital Forecast Database (NDFD) grid, and we intend to place enough of these thresholds into the National Digital Guidance Database (NDGD), an adjunct to the NDFD, so that users can reconstruct the CDFs at individual points of their choice. This will enable users

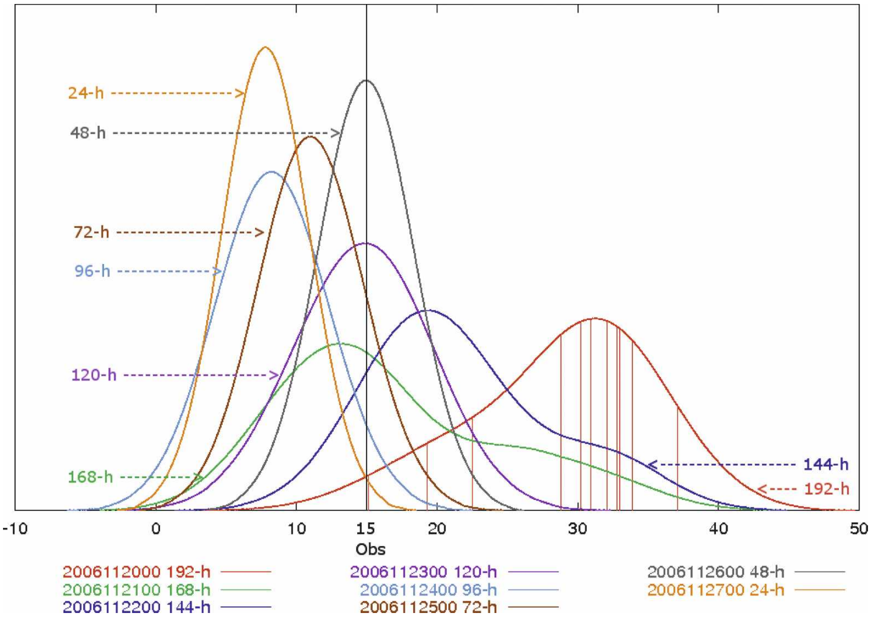

FIG. 16. PDFs for Juneau, AK (PAJN). All of the PDFs above verify at 0000 UTC 28 Nov 2006. The *T* observed at this time was 15°F, as indicated by the vertical black line. The positions of the 11 ensemble members used to create the 192-h forecast are the red vertical lines shown.

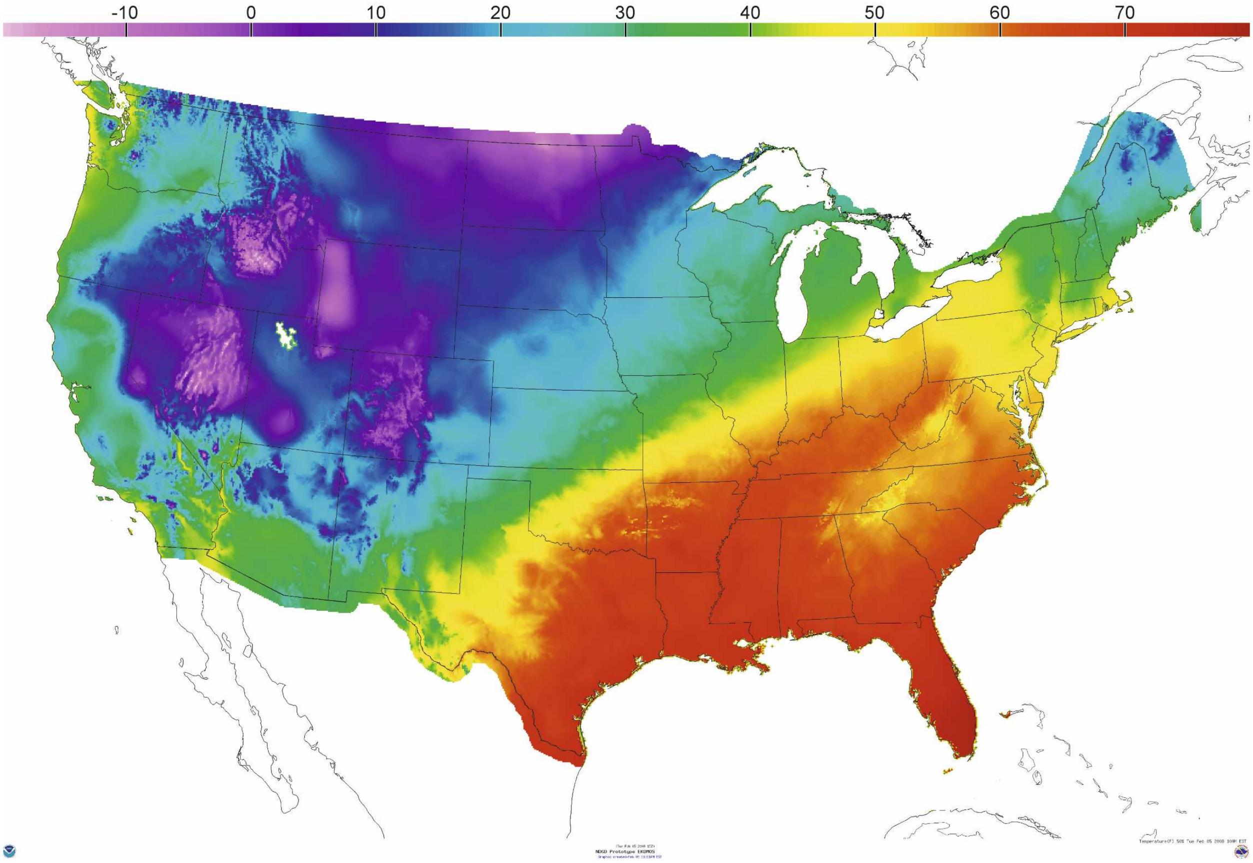
-10
0
10
20
30
40
50
60
70
(Tue Feb 05 2008 15Z)
NDGD Prototype EKDMOS
Graphic created-Feb 05 11:11AM EST
Temperature(F) 50% Tue Feb 05 2008 10AM EST

FIG. 17. Forecast median *T* for land areas of the CONUS at 1500 UTC 5 Feb 2008, based on the 0000 UTC run of GEFS on the same date. Note the frontal boundary across the Midwest and the temperature structure over the Rocky Mountains and Appalachian Mountains. Forecasts for the Great Salt Lake and the Great Lakes are not available in this prototype.

to take advantage of the powerful ensemble–post-processing system to make better decisions than they can make with raw, underdispersed ensembles. This will respond directly to the NRC report (National Research Council 2006) subtitled "Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts," and the AMS statement "Enhancing Weather Information with Probability Forecasts" (AMS 2002).

An interesting and unexpected conclusion was that the Ctl-Ctl-N method (only elements 1 and 2 above) gave statistically significant, although small, smaller errors than the use of all members for all projections less than 3 days than EKDMOS. The control member was run at higher resolution than the other members for those projections. The situation reversed for longer-range forecasts. While one might conclude that this poor performance of the ensembles at short projections, relative to the control run, was due only to the pronounced underdispersion at short projections compared to longer ones, the same result was found for the single value Ctl-Ctl-N median (and mean) compared to the EKDMOS median (and mean). That is, one could have obtained a better single-value MOS temperature forecast from the control run than our combination of the ensembles.

15. Future work

The results presented here are based on the 0000 UTC forecast cycle of the GEFS and the so called "cool season." Similar results can be expected for the warm season and 0600, 1200, and 1800 UTC GEFS forecast cycles, but the work must be done and the results tested. We have limited ourselves to 11 ensemble members because that was the number available during the development period. One might study the impact of developing and implementing EKDMOS guidance on a variable number of ensemble members. There is also the question of how to balance the number of ensemble members against their spatial and temporal resolution.

The method presented here is applicable to quasi-

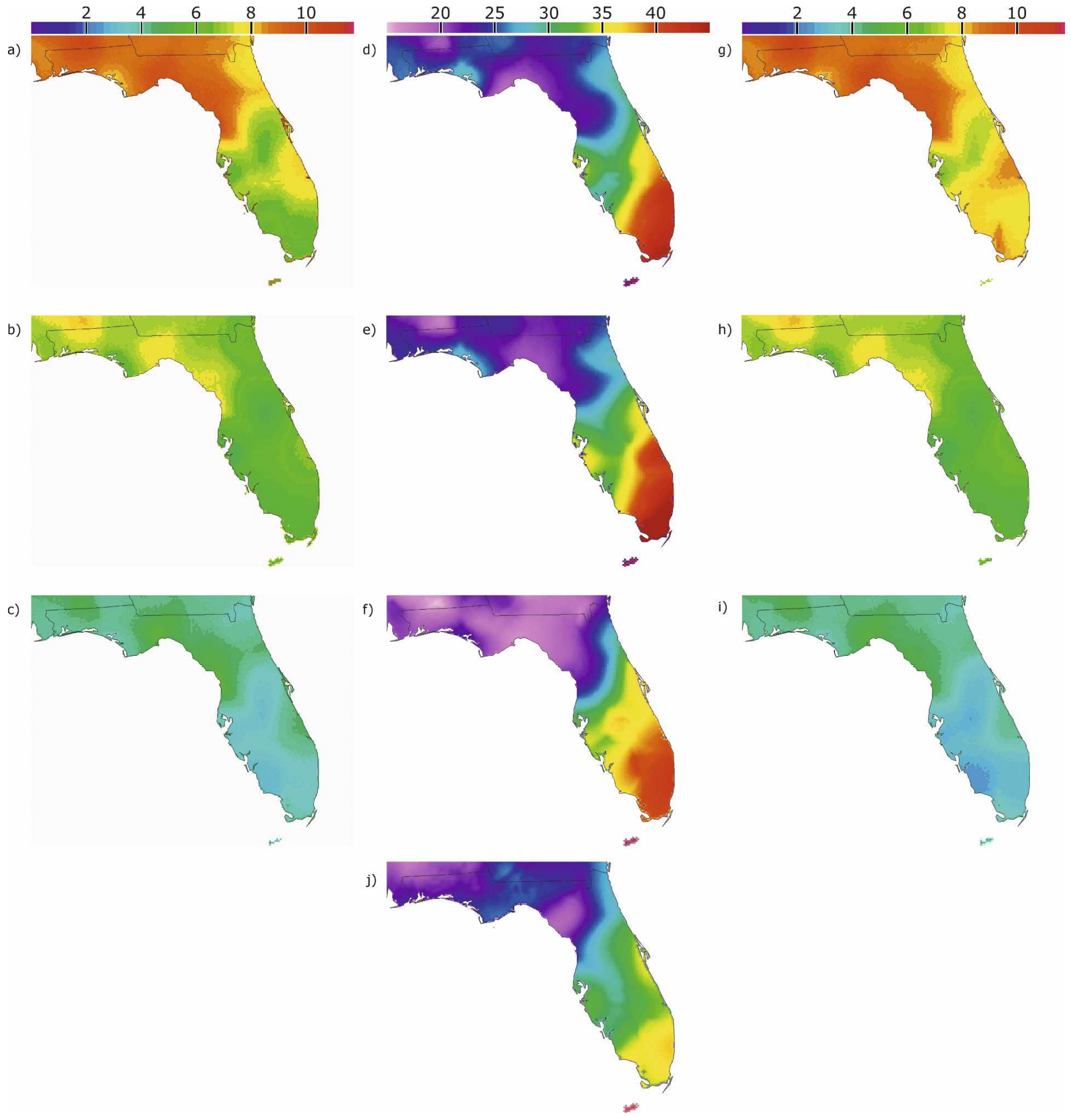

FIG. 18. Forecast median minimum *T*s for Florida on 3 Jan 2008 with cold and warm tails. The top row shows the day 7 forecast (center) with the width of the (left) cool tail and (right) warm tail. The second and third rows show the forecasts for days 5 and 2, respectively. (j) The verifying minimum *T* is shown.

normally distributed variables. When the variable has a high nonnormal distribution, and especially when a part of that distribution is far more important than another part (e.g., definitive ceiling heights under 500 ft are very important, but the difference between 3500 and 4000 ft is much less important), then sufficient and relevant points on the CDF can be determined by defining and forecasting a series of events by thresholding.

Acknowledgments. We acknowledge the work of David Rudack in initial retrievals of NCEP ensemble data. The Statistical Modeling Branch of the Meteorological Development Laboratory has made this work possible by the archival and quality control of the hourly observations. We also acknowledge use of NCEP's archive of the model data. We especially thank a reviewer for suggesting a significance test, which led to a more detailed analysis of the scores.

REFERENCES

AMS, 2002: AMS statement: Enhancing weather information with probability forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 450–452.

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.,* **127,** 1941–1953.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

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.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* **4,** 401–412.

Czado, C., T. Gneiting, and L. Held, 2007: Predictive model assessment for count data. University of Washington Department of Statistics, Tech. Rep. 518, 19 pp.

Day, R., cited 2008: Orange juice gains most in 2 months on freeze threat in Florida. [Available online at http://www.bloomberg.com/apps/news?pid=newsarchive&sid=a7PBBc8oP6xE.]

Descamps, L., and O. Talagrand, 2007: On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Wea. Rev.,* **135,** 3260–3272.

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

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

Erickson, M. C., 1996: Medium-range prediction of PoP and Max/Min in the era of ensemble model output. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J35–J38.

Fleming, R. J., 1971a: On stochastic dynamic prediction. Part I: The energetics of uncertainty and the question of closure. *Mon. Wea. Rev.,* **99,** 851–872.

——, 1971b: On stochastic dynamic prediction. Part II: Predictability and utility. *Mon. Wea. Rev.,* **99,** 927–938.

Glahn, H. R., 1985: Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 289–335.

——, 2002: A methodology for evaluating and estimating performance metrics. MDL Office Note 02-1, NOAA/National Weather Service, 18 pp.

——, and D. A. Lowry, 1969: An operational method for objectively forecasting probability of precipitation. ESSA Tech. Memo. WBTM TDL 27, Environmental Science Services Administration, 24 pp.

——, and D. P. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.,* **84,** 195–201.

——, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2008: Gridded MOS guidance in the National Digital Guidance Database. Preprints, *19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 11.3. [Available online at http://ams.confex.com/ams/pdfpapers/132526.pdf.]

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.

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

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

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

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

——, and Coauthors, 2003: Improved skill for the anomaly correlation of geopotential heights at 500 hPa. *Mon. Wea. Rev.,* **131,** 1082–1102.

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

Lubin, A., and A. Summerfield, 1951: A square root method of selecting a minimum set of variables in multiple regression: I. The method. *Psychometrika,* **16,** 271–284.

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

Miller, R. G., 1964: Regression estimation of event probabilities. The Travelers Research Center, Inc., Hartford, CT, U.S. Weather Bureau Contract Cwb-10704, Tech. Rep. 4, 153 pp.

Montgomery, D. C., and E. A. Peck, 1982: *Introduction to Linear Regression Analysis*. John Wiley & Sons, 504 pp.

Murphy, A. H., and H. Dann, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

National Research Council, 2006: *Completing the Forecast*. The National Academies Press, 112 pp.

National Weather Service, 1985: Automated daytime maximum, nighttime minimum, 3 hourly surface temperature, and 3-hourly surface dew-point guidance. NWS Tech. Procedures Bull. 356, NOAA/NWS, Silver Spring, MD, 14 pp.

Neter, J., and W. Wasserman, 1974: *Applied Linear Statistical Models*. Richard D. Irwin, Inc., 842 pp.

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.

Reap, R. M., and D. S. Foster, 1979: On producing categorical forecasts from operational probability forecasts of thunderstorms and severe local storms. Preprints, *11th Conf. of Severe Local Storms,* Omaha, NE, Amer. Meteor. Soc., 619–624.

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

Stensrud, D. J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510–2524.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.,* **33,** 223–226.

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

Tribus, M., 1970: Uncertainty and the weather. *Bull. Amer. Meteor. Soc.,* **51,** 4–10.

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.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986.

Wesner, J., 2008: Before freeze comes frenzy. *St. Petersburg Times,* 3 January 2008, 124.

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

——, 2006b: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Elsevier/Academic Press, 627 pp.

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

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

Yun, W. T., L. Stefanova, A. K. Mitra, T. S. V. Vijaya Kumar, W. Dewar, and T. N. Krishnamurti, 2005: A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus,* **57A,** 280–289.