Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over New England

DAVID J. STENSRUD

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

NUSRAT YUSSOUF*

Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma

(Manuscript received 27 December 2002, in final form 24 March 2003)

ABSTRACT

A multimodel short-range ensemble forecasting system created as part of a National Oceanic and Atmospheric Administration pilot program on temperature and air quality forecasting over New England during the summer of 2002 is evaluated. A simple 7-day running mean bias correction is applied individually to each of the 23 ensemble members. Various measures of accuracy are used to compare these bias-corrected ensemble predictions of 2-m temperature and dewpoint temperature with those available from the nested grid model (NGM) model output statistics (MOS). Results indicate that the bias-corrected ensemble mean prediction is as accurate as the NGM MOS for temperature predictions, and is more accurate than the NGM MOS for dewpoint temperature predictions, for the 48 days studied during the warm season. When the additional probabilistic information from the ensemble is examined, results indicate that the ensemble clearly provides value above that of NGM MOS for both variables, especially as the events become more unlikely. Results also indicate that the ensemble has some ability to predict forecast skill for temperature with a correlation between ensemble spread and the error of the ensemble mean of greater than 0.7 for some forecast periods. The use of a multimodel ensemble clearly helps to improve the spread–skill relationship.

1. Introduction

During the summer of 2002 the National Oceanic and Atmospheric Administration began a pilot program on temperature and air quality forecasting over New England. This project has three goals: 1) to quantify the improvements in the forecasting of temperature and air quality in the New England region that result from new and augmented observations and modeling, 2) to assess the benefits of the resulting better predictive capabilities to the energy sector as a primary customer, and 3) to provide a pathway to operational high-resolution temperature and air quality forecasting. As part of this pilot program, a short-range ensemble forecasting system was constructed to evaluate if an ensemble approach can provide improved 2-m temperature and dewpoint temperature predictions when compared against the statistical postprocessing available from present operational forecast models.

Statistical postprocessing of numerical weather prediction models has been provided to forecasters in the United States for over 30 years. This guidance is provided at present by the nested grid model (NGM) model output statistics (MOS), based upon a multiple linear regression approach (Glahn and Lowry 1972; Jacks et al. 1990). MOS is used, in part, to provide 2-m temperature and dewpoint temperature forecasts from 6 to 48 h for over 200 stations within the United States, and is an integral part of the guidance used to generate public forecasts. However, the regression approach used to create MOS requires a lengthy data archive period from an unchanged model, which makes it very difficult to apply this technique when model physical parameterizations and grid spacing are changing frequently [see discussions in Stensrud and Skindlov (1996) and Mao et al. (1999)].

This drawback to MOS has led to the exploration of other approaches to statistically postprocess model forecast data that do not require such long data archival periods. Homleid (1995) uses a Kalman filter technique to correct the systematic biases in short-term forecasts over Norway. A simple 7-day running mean bias calculation also is shown to improve upon raw model point forecasts (Stensrud and Skindlov 1996). Mao et al.

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

Corresponding author address: Dr. David J. Stensrud, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069. E-mail: David.Stensrud@noaa.gov

(1999) develop a multivariate linear regression approach that uses the past 2 to 4 weeks of data to produce refined 2-m temperature forecasts. And updateable MOS systems have been developed that do not require such an extensive data archive (Ross 1989; Wilson and Vallée 2002). Recent results with short-range ensembles suggest that this approach also is worth considering.

A number of studies show that the simple mean of an ensemble of short-range forecasts is as good as or better than a single forecast from a model with smaller grid spacing (Du et al. 1997; Hamill and Colucci 1997; Stensrud et al. 1999; Fritsch et al. 2000; Hou et al. 2001; Wandishin et al. 2001; Grimit and Mass 2002). Krishnamurti et al. (1999, 2001) further show the value of a bias-corrected ensemble for predicting precipitation, which could be extended to other variables. It appears that the main improvement from using an ensemble mean forecast is due to overlapping differences in the sign of the errors associated with individual traveling disturbances (Fritsch et al. 2000). However, the improved skill from the ensemble mean is only one measure of the value of an ensemble system.

Murphy and Winkler (1979) argue that forecasts cannot be used to their best advantage ''unless the uncertainty inherent in the forecasts is quantified and expressed in a concise and unambiguous manner.'' One advantage of ensemble approaches is that they provide information on the probabilities of various events, which may provide information on forecast uncertainty. While many early studies on short-range ensemble forecasting show little correlation between ensemble spread and forecast skill (Hamill and Colucci 1998; Stensrud et al. 1999; Hou et al. 2001), recent results by Grimit and Mass (2002) are more promising. It also may be that ensemble spread is a better predictor of forecast skill when it is either very small or very large in comparison with its climatological value (Houtekamer 1993; Whitaker and Loughe 1998). Nevertheless, Richardson (2000) clearly shows the potential economic value of imperfect ensembles for a wide range of end users, in which the additional probabilistic information from the ensemble provides an economic benefit to users that is equivalent to many years of effort in improving a deterministic model. Thus, we explore the benefits of an ensemble of short-range forecasts to the prediction of 2-m temperature and dewpoint temperature over New England by examining both the ensemble mean forecasts and the ensemble probabilities.

The ensemble system and methodology used is described in section 2. A comparison of the ensemble mean forecasts to NGM MOS is found in section 3. Section 4 illustrates the additional value that an ensemble can provide, followed by an examination of the ability of this multimodel ensemble to predict forecast skill. A final discussion follows in section 6.

2. Data and methodology

The models used in the pilot program are the National Centers for Environmental Prediction (NCEP) Eta model (Black 1994), the regional spectral model (RSM: Juang and Kanamitsu 1989), the Rapid Update Cycle model (RUC: Benjamin et al. 1994, 2001), and the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5: Dudhia 1993). A total of up to 23 forecasts are available each day, depending upon the availability of the various model runs in real time. Past studies indicate that multimodel ensembles are more skillful than single model ensembles (Richardson et al. 1996; Buizza et al. 1999; Harrison et al. 1999; Ziehmann 2000; Stensrud et al. 2000; Fritsch et al. 2000; Evans et al. 2000; Wandishin et al. 2001), providing hope that the results from this pilot program that uses several different models will be indicative of the potential for using ensembles to predict near-surface variables.

Five of these forecasts are from the 48-km Eta model using the breeding of growing modes technique to generate two positive and two negative bred perturbations, in addition to the control analysis, for the initial and boundary conditions (Toth and Kalnay 1993, 1997). Another five forecasts use a version of the Eta model that incorporates the Kain–Fritsch convective parameterization scheme (EtaKF: Kain et al. 2001) and fourth-order diffusion in a separate bred mode cycling. Five forecasts also are from the 48-km RSM,[1] again using a bred mode cycling to generate the initial and boundary conditions. Two forecasts are from the 22-km EtaKF model that is started from both the operational Eta model and the operational aviation run of the medium-range forecast (MRF) model initial conditions, but uses a smaller domain than the operational Eta model (Kain et al. 2001). One forecast is from the 20-km RUC, which uses its own optimal interpolation scheme (Benjamin 1989) to produce an initial condition and the Eta model forecast for boundary conditions. The final five forecasts are from MM5. Four of these forecasts use the Eta model initial and boundary conditions for the control run and a random coherent structure approach to generate another three initial conditions for the 32-km grid (see Stensrud et al. 2000). These four forecasts also mix the Kain–Fritsch (Kain and Fritsch 1990) and Betts–Miller–Janjić (Betts and Miller 1986; Janjić 1994) convective schemes, and the MRF (Hong and Pan 1996) and Blackadar (Zhang and Anthes 1982) planetary boundary layer schemes, to provide both initial condition and model physics variability. The final MM5 run uses 27-km grid spacing and is started from the RUC analysis with the Eta model forecast for boundary conditions.

Depending upon local resources, the start times from the model forecasts vary from 0000 to 1200 UTC, but all the model forecasts extend over the 48-h period beginning at 1200 UTC each day. (see Table 1 for further details). The experiment started on 15 July and ended

[1] The RSM forecasts do not include the 2-m dewpoint temperature as an output variable, so there are 18 ensemble members only for this variable.

TABLE 1. Descriptions of the models used to construct the ensemble, the organizations that provided the forecasts, the model grid spacing (km), the model forecast start times (UTC), the number of forecasts provided, and the perturbation strategy used: IC indicates the the initial conditions are perturbed, and PH indicates that the model physics are perturbed. The organizations providing forecasts are the National Centers for Environmental Prediction (NCEP), the National Severe Storms Laboratory (NSSL), and the Forecast Systems Laboratory (FSL).

Model	Organization	Grid spacing (km)	Start time (UTC) and forecast length (h)	Number of forecasts	Perturbation strategy (IC/PH)
Eta	NCEP	48	0900/51 h	5	IC
EtaKF	NCEP	48	0900/51 h	5	IC
RSM	NCEP	48	0900/51 h	5	IC
EtaKF	NSSL	22	0000/60 h	2	IC
MM5	NSSL	32	0000/60 h	4	IC and PH
RUC	FSL	20	1200/48 h	1	
MM5	FSL	27	1200/48 h	1	

on 31 August 2002, for a total of 48 forecasts. Model output is available every 3 h from 0 to 48 h and is bilinearly interpolated to a common 10-km Lambert conformal grid centered over the New England region (Fig. 1). The use of a common grid for all the model forecasts made the real-time dissemination of ensemble output much easier. Occasionally, one or more of the model forecasts is missing owing to computer problems, in which case the ensemble is created from the remaining available members.

To compare the model results with temperature observations, the point on the common 10-km grid that is closest to the observation site is selected. Similarity theory (Stull 1988) is used to interpolate from the lowest model level of each model to the 2-m level for temperatures and dewpoint temperatures in a manner that is consistent with the model physical parameterization schemes prior to the bilinear horizontal interpolation. However, it is clear that postprocessing of the raw model data increases the accuracy of the forecasts (Dallavalle 1996). Thus, a technique is needed to reduce the model 2-m temperature and dewpoint temperature biases. Since one of the goals of the pilot program is to provide an approach that could be quickly transferred to operations, a simple 7-day running mean bias correction is applied to each model at each of 395 National Weather Service (NWS) observing station locations within the common grid domain and for each 3-h forecast interval[2] (Fig. 2).

Values of the bias correction vary from model to model and station to station and depend upon the forecast time, but typically are within the range of $+5$ K to -5 K. As an example, for both the Eta and MM5 forecasts the bias corrections calculated for Pittsburgh (PIT), Pennsylvania, are between ± 1.5 K at the 9-h forecast time for all days and vary slowly over time as expected from using a 7-day mean calculation. In contrast, the bias corrections for Albany (ALB), New York, vary over a larger range, from $+4$ K to -2 K, at the 9-h forecast time. The ALB bias corrections from MM5 remain above zero for all days except one, ranging from 0 to 4 K, whereas the ALB bias corrections from Eta range from -1.5 to 4 K. Bias corrections during the nighttime hours typically show smaller variations in time than those during the daytime. The bias corrections are needed to account for differences between the model and actual terrain heights, errors in the depth and intensity of the surface superadiabatic layer, the amount of entrainment from the boundary layer top, the ratio of sensible to latent heat fluxes, and net radiation. The bias-corrected forecasts are used to calculate the ensemble mean, ensemble spread, and the raw ensemble probabilities for each observation location. The ensemble forecast start time is defined as 1200 UTC, and the 1200 UTC NGM MOS data are used as a benchmark to evaluate the usefulness of the ensemble data.

The NGM MOS guidance for 2-m temperature and dewpoint temperature is available from 6 to 48 h at 3-h intervals for the same days as the ensemble forecasts. However, the NGM MOS is only available for 108 NWS observing sites within the common grid domain. Thus, the comparisons between the ensemble mean and the NGM MOS forecasts only occur at the NGM MOS locations. Calculations of bias, mean absolute error (MAE), and root-mean-square error (rmse; see Wilks 1995 for definitions) are conducted for each observation location and averaged over all locations for each forecast time. Owing to missing data, typically over 3000 observations are used to compute the verification statistics at each forecast time.

It should be recognized that the training data and verification data from the NGM MOS are mutually exclusive (i.e., from different years), whereas they are not mutually exclusive for the bias calculations using the simple 7-day running mean bias calculation approach. This might cause one to wonder about the validity of comparing the resulting biases from these two approaches. However, it is not clear that one wants to discard prior information when trying to improve a forecast. In essence, if we have prior knowledge about the bias for today's forecast based upon information from yesterday, then it seems reasonable that we would want to use this knowledge to our advantage. Thus, we proceed with making comparisons of the resulting biases produced by these two approaches since the practical

[2] For the first 7 days of the pilot program, the subsequent 7 days are used to calculate the bias corrections for each ensemble member.

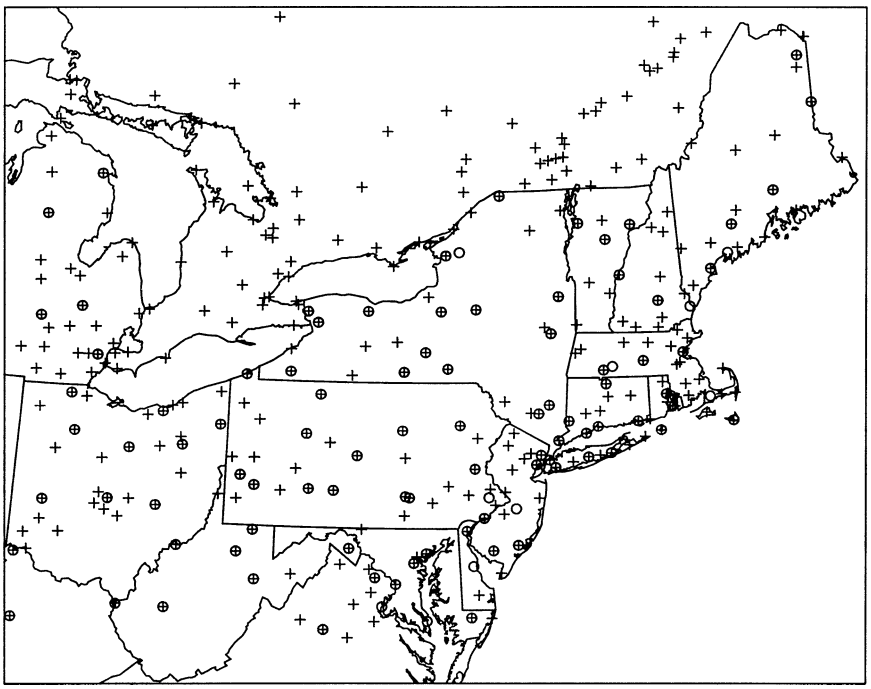

FIG. 1. Common domain to which all model forecasts are bilinearly interpolated to a 10-km grid spacing. A plus sign indicates the location of a NWS surface observing station, while a circle indicates the location for which a MOS forecast is made. Note that roughly 20%–25% of the stations likely are close enough to large water bodies to be influenced by sea breezes.

end result we seek is an improved forecast postprocessing system.

3. Comparisons with NGM MOS

Results indicate that the average bias-corrected ensemble (BCE) forecasts of 2-m temperature and dewpoint temperature have smaller bias, MAE, and rmse values than those produced by NGM MOS for the 48 days of the pilot program (Table 2). While these differences are significant at the 95% level using a Student's *t* test (Wilks 1995), this significance level likely is due to the large dataset size of over 55 000 observations. If a Wilcoxon signed-rank test (Wilks 1995) is used on the entire paired dataset (i.e., using all stations and forecast hours), then the differences in temperature predictions are not significant. Thus, these results simply indicate that the BCE temperature forecasts are as accurate as the NGM MOS. However, the differences in the dewpoint temperature predictions are significant at the 95% level for bias, MAE, and rmse when using the Wilcoxon signed-rank test on the entire dataset, indicating that the BCE is more accurate than the NGM MOS for this variable.

If the errors are examined as a function of forecast time (Fig. 3), then we see that the MAE and rmse for both the MOS and BCE predictions increase during the daytime heating hours (6–12 and 30–36 h). This likely is due to errors in both the net radiation and the evolution of the planetary boundary layer. The BCE dewpoint temperature predictions are consistently better than the NGM MOS for all forecast times and verification measures, except for late in the forecast period

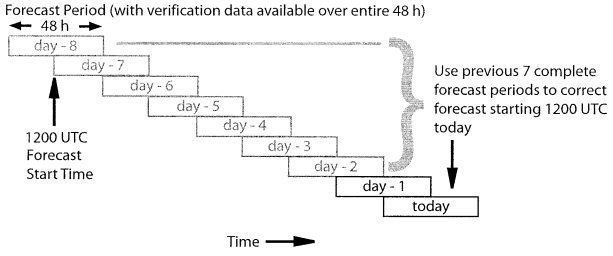

FIG. 2. Schematic illustrating the 7-day running mean bias correction applied to each of the individual forecast model output fields of 2-m temperature and dewpoint temperature. These bias corrections are calculated at 3-h intervals for the 395 NWS observing sites located within the common domain depicted in Fig. 1.

TABLE 2. Values of bias, MAE, and rmse from the NGM MOS and the bias-corrected ensemble (BCE) averaged over all observation locations, all times, and all 48 forecast days during the pilot program.

Temperature	Bias (K)	MAE (K)	Rmse (K)
NGM MOS	−0.25	1.49	1.98
BCE	0.01	1.48	1.93
	Dewpoint temperature		
NGM MOS	0.36	1.58	2.29
BCE	0.10	1.47	1.98

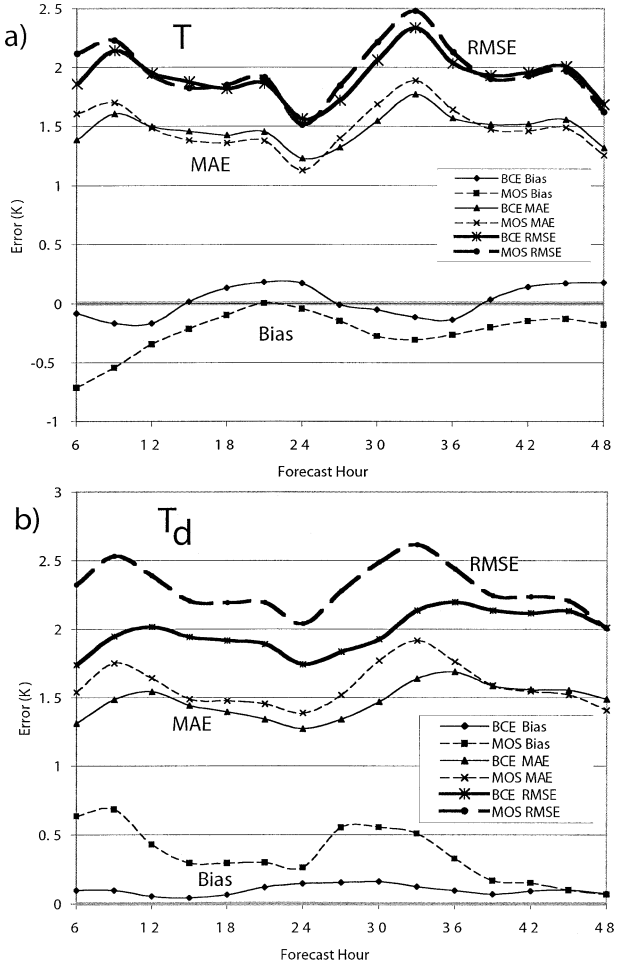

FIG. 3. Values of mean bias, MAE, and rmse plotted as a function of forecast hour for (a) 2-m temperature and (b) 2-m dewpoint temperature calculated from over 100 NGM MOS sites within the common domain. Solid lines indicate BCE error curves, while dashed lines indicate the NGM MOS error curves. Further details are shown in the legend.

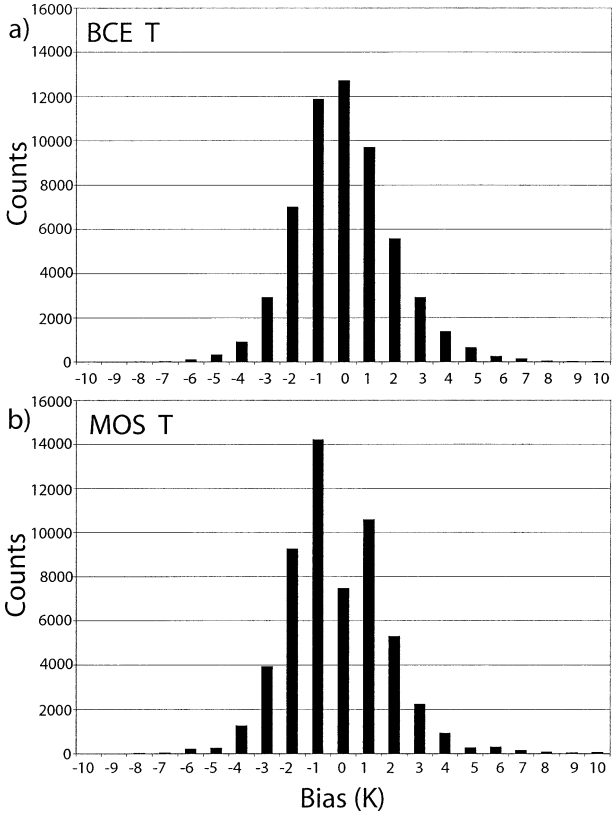

FIG. 4. Histograms of the 2-m temperature errors from (a) the mean BCE and (b) the NGM MOS for all forecast times from 6 to 48 h calculated at the NGM MOS sites. Bias (forecast − observed) values are binned into 1-K intervals and counted.

(Fig. 3b). If the errors are binned into 1-K intervals, then we find that the BCE forecast errors are more normally distributed around zero than the MOS forecast errors and that the BCE forecast errors have fewer large errors (magnitude ≥ 6 K) when compared with the MOS (Figs. 4 and 5).

The errors at individual observation locations show similar behaviors to the means over the common grid (Figs. 6 and 7). The temperature rmse mimics the diurnal heating cycle at both Boston (BOS), Massachusetts, and Portland (PWM), Maine. We also see that the BCE bias values are less than 0.5 K for all forecast times, whereas the MOS bias values exceed 0.5 K at some times. In addition, the BCE temperature rmse is lower than the MOS value for most times after 9 h, and the BCE dewpoint temperature rmse is always lower than or equal to the MOS value. These results merely reinforce the conclusions formed by examining results over the entire New England region in that the BCE forecasts are competitive with or better than the NGM MOS forecasts. However, the ensemble has far greater value than suggested by just an examination of the ensemble mean.

4. Ensemble probabilities

One of the obvious ways to evaluate the behavior of an ensemble is to examine a rank histogram. If the verifying observation is considered a member of the ensemble, then one would expect that the verifying observation falls with equal probability anywhere in the envelope of the ensemble forecasts when ranked from smallest to largest. However, results from counting the rank of the verifying observation for 2-m temperature and dewpoint temperature at all 395 NWS stations within the common domain indicates that the ensemble system is underdispersive (Fig. 8) since the rank histogram takes on a characteristic U shape (Hamill 2001). This means that the verifying observation often falls outside the envelope of forecasts generated by the ensemble members. The ensemble is more underdispersive for 2-m dewpoint temperature, where the ensemble often produces warmer dewpoint temperatures than are observed

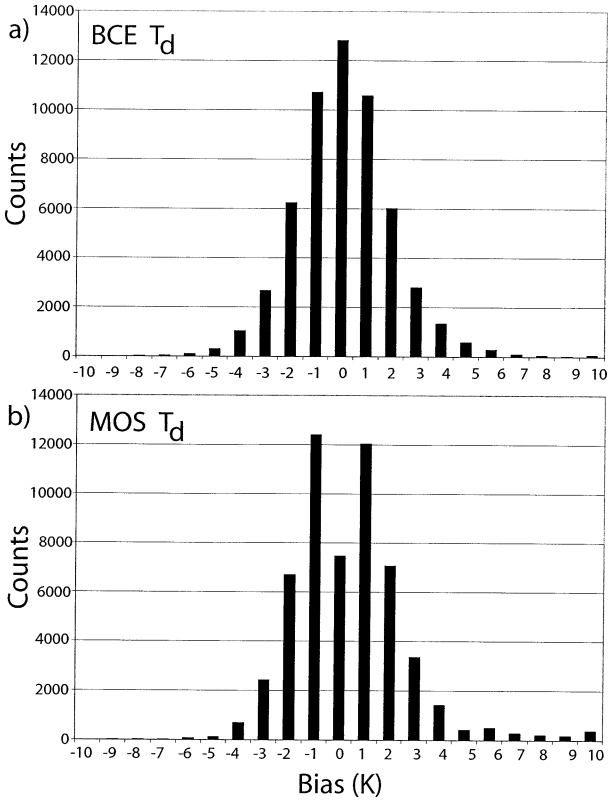

FIG. 5. As in Fig. 4 but for 2-m dewpoint temperature errors.

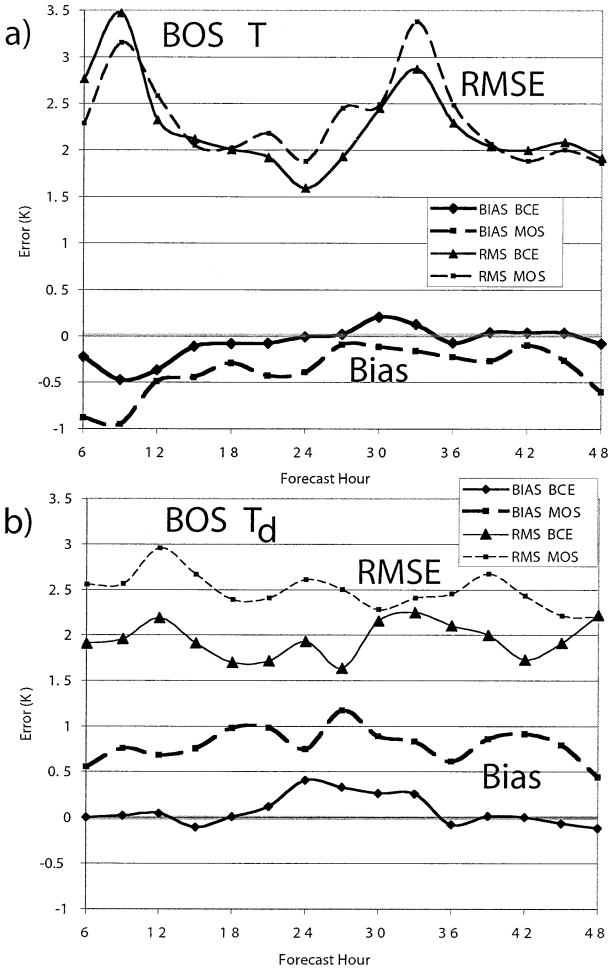

FIG. 6. Values of mean bias and rmse plotted as a function of forecast hour for both (a) 2-m temperature and (b) 2-m dewpoint temperature for Boston, MA. Solid lines indicate BCE error curves, while dashed lines indicate the NGM MOS error curves.

as indicated by the large percentage of occurrences in the first rank.

Given the nonuniform rank histograms, it is unclear whether or not the ensemble probabilities have forecast value. This is determined by examining the reliability, or conditional bias, of the ensemble probability forecasts for 2-m temperatures exceeding a specified value. Results indicate that the raw ensemble probabilities underpredict the frequency of occurrence of temperatures greater than or equal to 303 K (~30°C or 86°F) for probabilities less than 70% (Fig. 9a). This underprediction is less when the temperature threshold is decreased to 296 K (~23°C or 73°F) (Fig. 9b). However, following Hamill and Colucci (1998), it may be that the information from a rank histogram can be used to calibrate the ensemble probabilities (see appendix). These adjusted probabilities, calibrated only using the rank histogram information from the previous 7 complete forecast days, show an improvement in the reliability of the ensemble forecasts for temperatures exceeding 303 K and 296 K (Fig. 9). Calculations of the Brier score (Brier 1950), basically a mean-square error of the probability forecasts (Wilks 1995), from the raw and calibrated ensemble probabilities are completed for each case using all 17 forecast output times. The scores are nearly identical for both temperature thresholds. These differences are not significant when evaluated using a Wilcoxon signed-rank test on the 48 daily paired values of the Brier score.

Similar results are found for the probabilities of 2-m dewpoint temperatures exceeding a specified value (Fig. 10). The adjusted probabilities, again calibrated only using the rank histogram information from the previous 7 days, are an improvement over the raw values with better values of reliability and similar values of resolution calculated from the decomposed Brier score (Sanders 1963; Wilks 1995). Brier scores are calculated from the forecast probabilities for each case day. The differences in the Brier scores are significant at the 95% level for both dewpoint temperature thresholds when evaluated using a Wilcoxon signed-rank test on the 48 daily paired values of Brier score.

A useful measure for evaluating the ability of probabilistic forecasts to discriminate dichotomous events is the relative operating characteristic (ROC: Swets 1973; Mason 1982). The ROC uses information from a 2×2 contingency table to determine the hit rate (HR), the

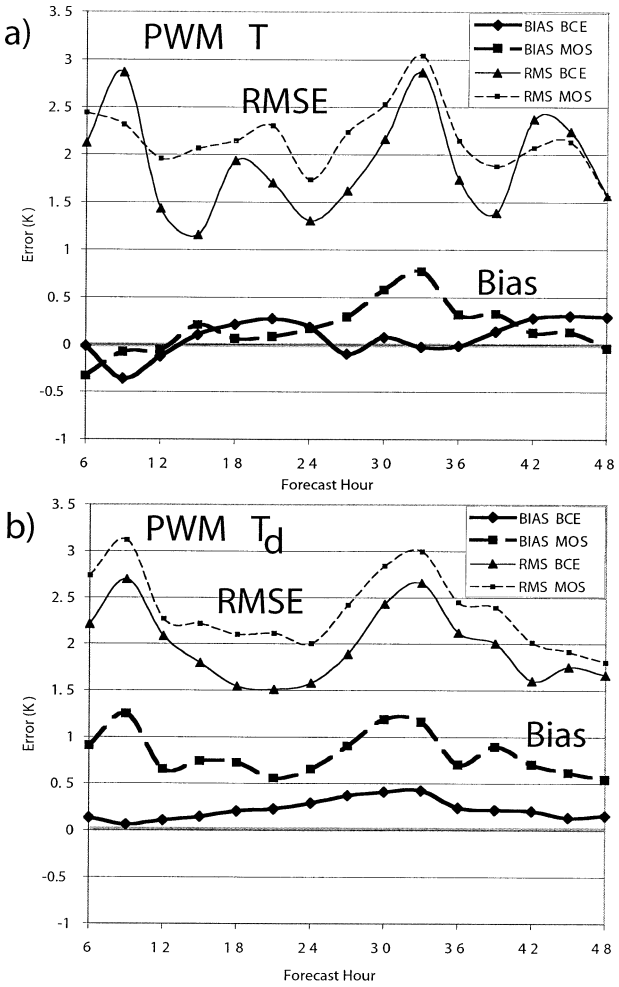

FIG. 7. As in Fig. 6 but for Portland, ME.

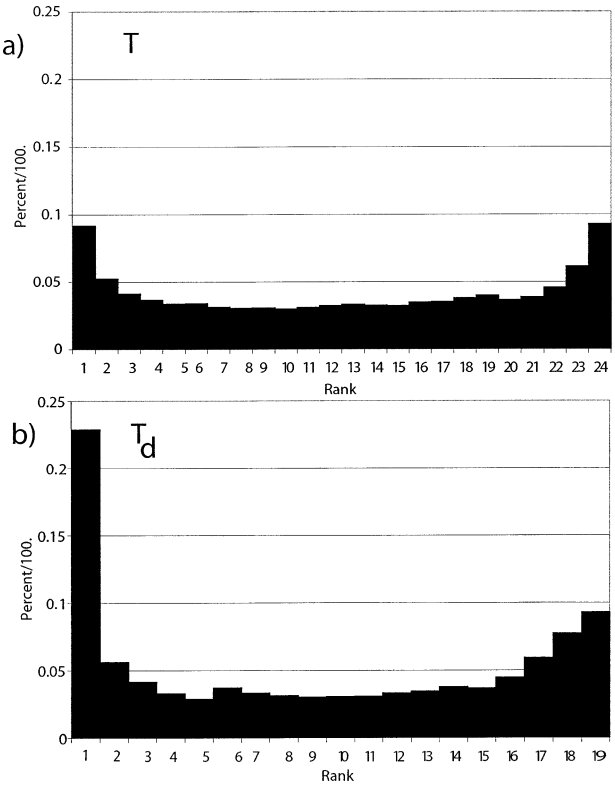

FIG. 8. Rank histogram from the bias-corrected ensemble forecasts of (a) 2-m temperature and (b) 2-m dewpoint temperature for all forecast times at all 395 NWS observing sites. Recall that only 18 ensemble members provide a 2-m dewpoint temperature forecast, instead of the full 23 members that predict 2-m temperature.

proportion of events that are correctly forecast, and the false alarm rate (FAR), the proportion of nonevents that are incorrectly forecast as events. For probabilistic forecasts, the HR and FAR are calculated for different probability values and plotted as a curve on a FAR versus HR diagram. An unskillful forecast has HR = FAR since it is unable to discriminate between events and nonevents, yielding a diagonal line from the (FAR, HR) point (0, 0) to (1, 1). In contrast, a perfect forecast yields a FAR = 0 and HR = 1. Any skillful forecast produces a ROC curve that lies to the upper left of the diagonal line, and the closer the curve to the upper-left corner (0, 1) the more accurate the forecast. When comparing with deterministic forecasts, which produce only a single point on the ROC curve, one examines whether the deterministic forecast (FAR, HR) point falls above or below the ensemble ROC curve. If it falls below the ensemble ROC curve, then the deterministic system is less accurate; it is more accurate if the point falls above the ensemble ROC curve.

The quality of a forecast system often is summarized in terms of the area (A) under the ROC curve (Mason 1982). Skillful forecast systems have $A > 0.5$, with $A = 1.0$ being a perfect forecast system. Mason (1982) presents two methods for calculating A. The second method, which is independent of the number of decision criteria selected, is chosen. Here a line is fitted to the data when transformed into normal deviate space, yielding a continuous ROC curve under which the area A_z is calculated.

Results from the ensemble using various threshold values for temperature indicates that the ensemble is reasonably accurate when compared with the NGM MOS (Fig. 11). For the warmer temperature thresholds, the ensemble may be more accurate than NGM MOS as the MOS (FAR, HR) point is inside (below) the ensemble ROC curve. However, for lower temperatures the NGM MOS (FAR, HR) point is outside (above) the ensemble ROC curve, indicating that the NGM MOS may be more accurate. Calculations of the 95% confidence intervals (see Wandishin et al. 2001) indicate that none of these differences are actually significant. Note that the area A_z under the ensemble ROC curve increases as the temperature threshold both increases and decreases to less frequently observed values, highlighting the value of the ensemble for more unlikely events.

Similar results are seen in the ROC curves from the

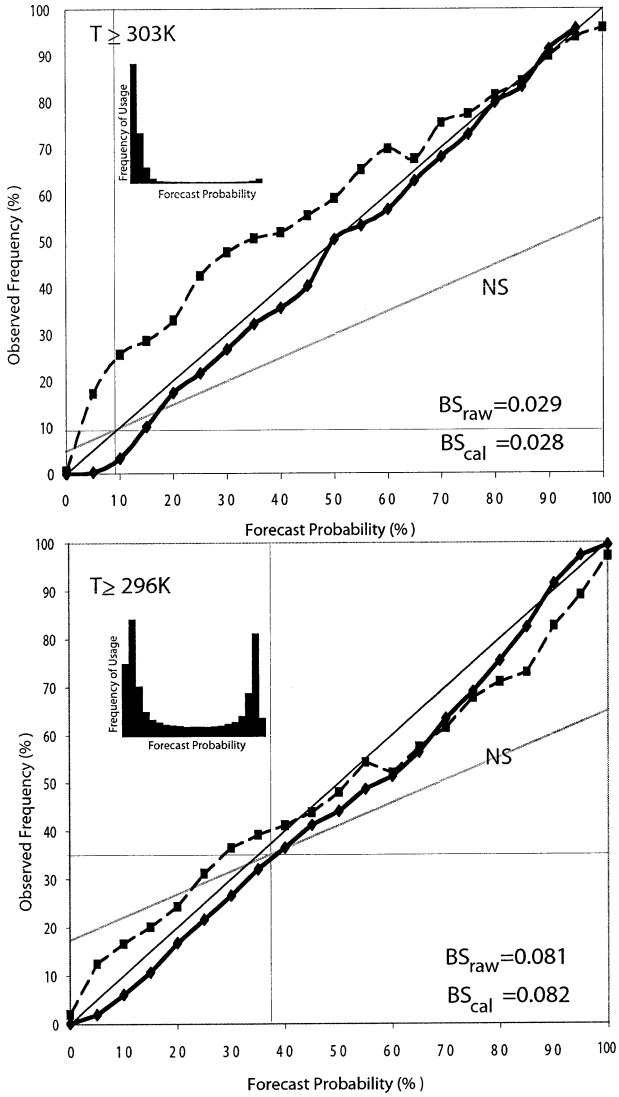

FIG. 9. Attribute diagrams for the bias-corrected ensemble forecasts of 2-m temperatures equal to or exceeding (a) 303 and (b) 296 K. Inset histogram indicates the frequency of usage of each 5%-inteval forecast probability category for the uncalibrated ensemble. Dashed line is from the uncalibrated (raw) ensemble probabilities, while the solid line is from the corrected ensemble probabilities. Horizontal gray line indicates the frequency of the event in the observed dataset, and the diagonal gray line is the no skill (NS) line. Brier scores for the calibrated and uncalibrated ensembles indicated.

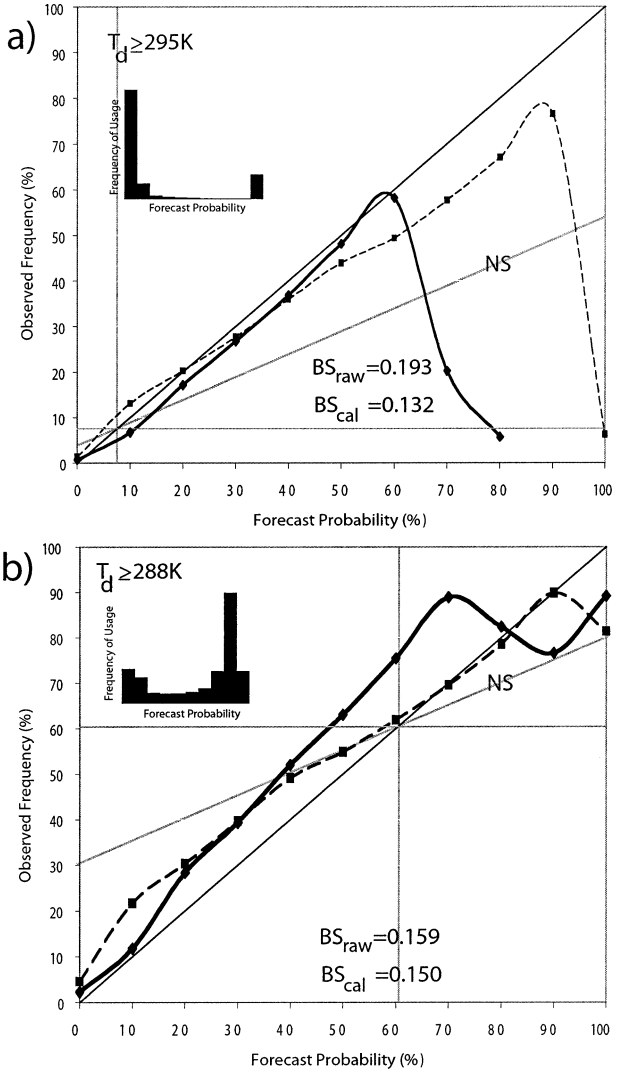

FIG. 10. As in Fig. 9 but for the bias-corrected ensemble forecasts of 2-m dewpoint temperatures equal to or exceeding (a) 295 and (b) 288 K.

2-m dewpoint temperature forecasts (Fig. 12). For the lowest dewpoint temperature threshold of 283 K (~10°C or 50°F), the point calculated from the NGM MOS forecasts is outside (above) the ensemble ROC curve and this difference is significant at the 95% confidence level. However, as the dewpoint temperature threshold increases, two behaviors are seen: 1) the ensemble ROC curves move to the upper left, indicating that the ensemble has greater forecast quality, and 2) the point calculated from the NGM MOS forecasts falls inside (below) the ensemble ROC curves. The difference between the ensemble and NGM MOS values is significant at the 95% confidence interval for the dewpoint temperature threshold of 293 K. Thus, the ensemble again adds value above that of the NGM MOS when the events become more unlikely.

While we have seen that the bias-corrected ensemble is very competitive with the NGM MOS forecasts of 2-m temperature and dewpoint temperature using a variety of measures, these results tell us very little about the economic value of the forecasts. As discussed by Murphy (1994), it is impossible to know the economic value of forecasts unless you know that a particular user took action based upon a forecast, have detailed knowledge of the decision-making processes of this particular user, and know the skill of the forecasts. However, it is possible to evaluate the potential economic value of forecasts using various decision-making models (Murphy

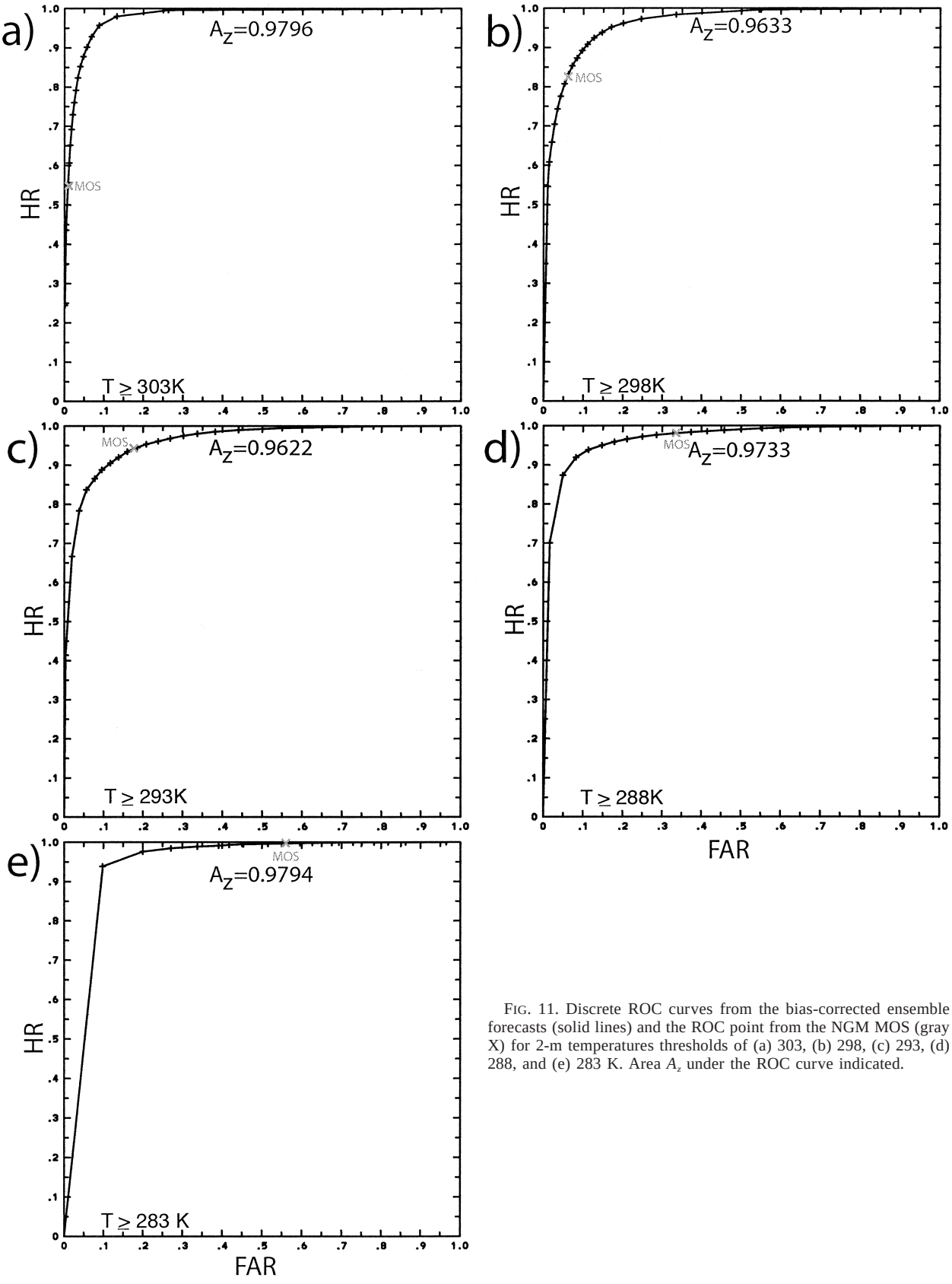

FIG. 11. Discrete ROC curves from the bias-corrected ensemble forecasts (solid lines) and the ROC point from the NGM MOS (gray X) for 2-m temperatures thresholds of (a) 303, (b) 298, (c) 293, (d) 288, and (e) 283 K. Area A_z under the ROC curve indicated.

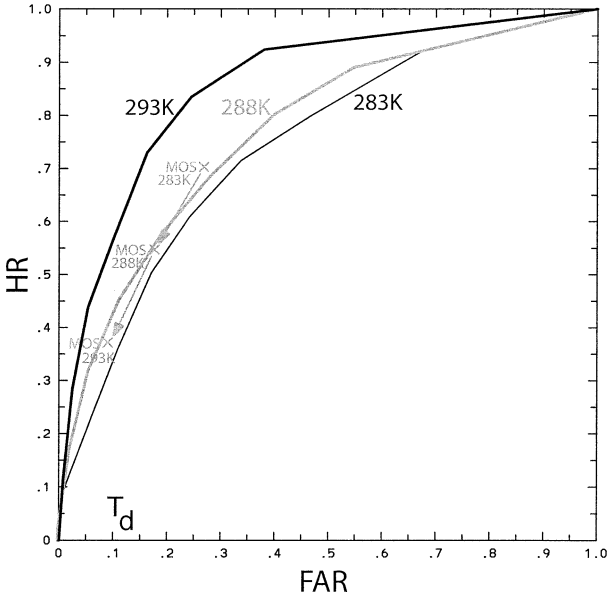

FIG. 12. Discrete ROC curves from the bias-corrected ensemble forecast (solid lines) and the ROC points from the NGM MOS (gray X) for 2-m dewpoint temperatures for thresholds of 293, 288, and 283 K. For 283 K, the NGM MOS point is initially above the ensemble ROC curve, but it moves below the ensemble ROC curve for the 288- and 293-K thresholds as indicated by the arrows.

1977; Katz and Murphy 1997). Of these, the cost–loss ratio model is the simplest and has a long tradition of study (Murphy 1977).

The cost–loss model assumes the existence of a decision maker who must choose to either take action or do nothing when presented with guidance from an imperfect weather forecasting system that a given weather event will, or will not, occur. A cost C is incurred when action is taken to protect from this weather event, irrespective of the outcome. However, if the event occurs and no protective action is taken, then a loss L is incurred. With respect to the energy sector, C could be the cost of turning on gaspowered turbines when energy demands are expected to exceed the available power supply of a given power company, owing to warmer than normal forecast temperatures. The loss L is incurred when power demand exceeds that available from this power company and the additional power needed to meet demand must be purchased from other power suppliers. The goal of the decision maker is to minimize the costs of weather events over a large number of cases. Dempsey et al. (1998) discuss in greater detail many of the weather sensitivities in the energy industry.

Richardson (2000) defines the value V, relative to climatology, of a forecast system as the reduction in expense incurred in proportion to that which would be achieved by a perfect forecast system. Thus, $V = 1$ implies a perfect forecast system and $V = 0$ implies a forecast system that is no better than climatology. For all $V > 0$ the forecast system has value and the user benefits from these forecasts. Richardson finds that

$$V = \frac{\min(\alpha, \overline{o}) - \mathrm{FAR}\alpha(1 - \overline{o}) + \mathrm{HR}\overline{o}(1 - \alpha) - \overline{o}}{\min(\alpha, \overline{o}) - \overline{o}\alpha}, \quad (1)$$

where $\alpha = C/L_1$, the cost of taking action divided by the potential loss protected by the action (L_1), $\overline{o}$ is the observed relative frequency of the event, and FAR and HR are as defined previously. The maximum value of V always occurs when $\alpha = \overline{o}$. For each calibrated ensemble forecast probability, the values of FAR and HR are determined. Then α is varied from 0 to 1 and the curve for V is calculated as a function of α and plotted. The same is done for the NGM MOS forecasts, but the deterministic system only yields a single V curve. Results indicate that for three different temperature thresholds, the BCE provides greater value to users than the deterministic NGM MOS forecasts (Fig. 13). Note that the threshold value criterion is different in Fig. 13c, where temperatures are specified to be $\leq$283 K in order to determine the likelihood of cooler temperatures. The ensemble clearly provides value to a larger number of users than NGM MOS since the values of α for which value can be derived ($V > 0$) has a larger range for the ensemble than for NGM MOS.

The NGM MOS data also is postprocessed into a probabilistic form and compared with the BCE. This is accomplished by calculating the rmse of the NGM MOS predictions over the entire domain at each forecast output time. Similar to the BCE, a 7-day running mean rmse is determined for each forecast hour and used to develop the forecast probabilities for the present day's forecast. We assume that the NGM MOS errors are normally distributed with a standard deviation equal to the calculated rmse, and this allows us to determine probability forecasts from our knowledge of the characteristics of the normal distribution. Results indicate that the probabilistic NGM MOS has greater value than the deterministic NGM MOS (Fig. 13), but the BCE still provides greater value for most users.

The value of the dewpoint temperature forecasts is less than those of the temperature forecasts for both the ensemble and the NGM MOS. For a 293-K ($\sim$20°C or 68°F) dewpoint temperature threshold, the BCE provides more value to more users than NGM MOS (Fig. 14). However, for a 283-K dewpoint temperature threshold, the NGM MOS provides the largest value to some users (as expected from the ROC curve calculations), while the BCE provides the largest value to other users owing to the larger range of users that benefit from the ensemble. As seen previously, the probabilistic NGM MOS provides greater value than the deterministic NGM MOS, but the BCE provides the most value overall.

5. Forecasting forecast skill

Since many of the early short-range ensemble studies find little correlation between the skill of the forecasts

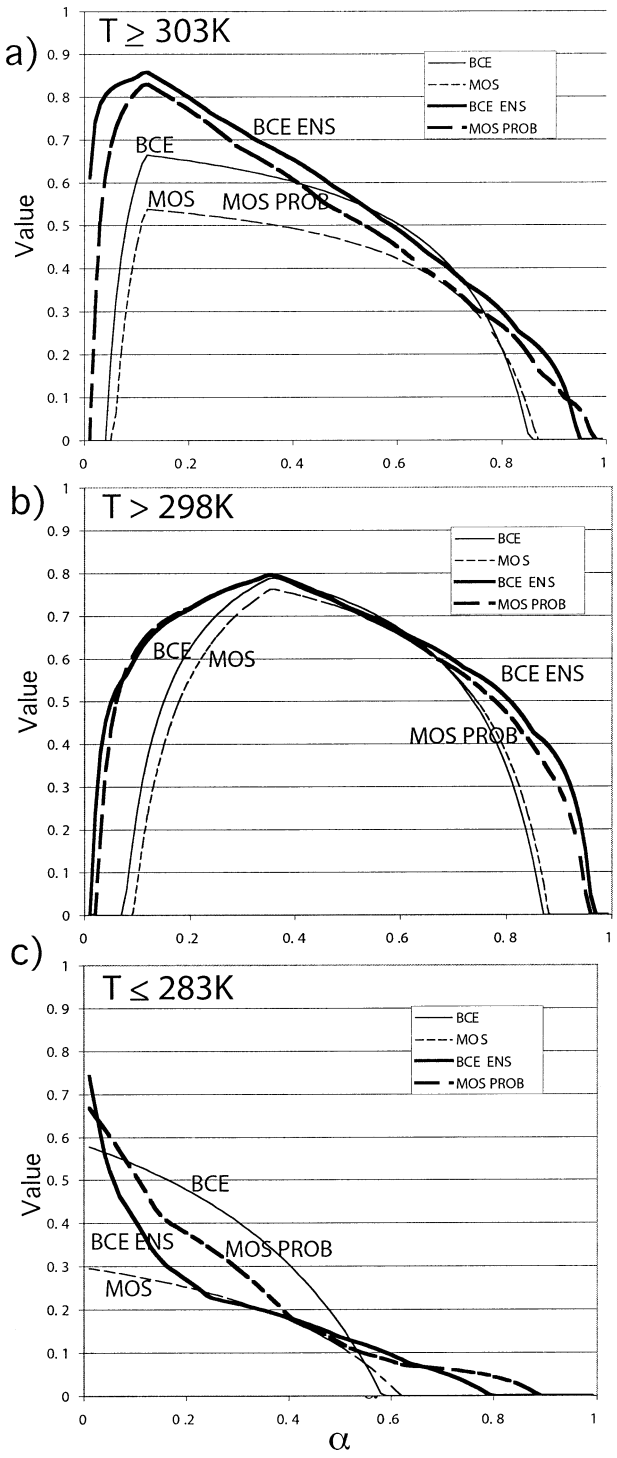

FIG. 13. Relative value V of the bias-corrected ensemble (solid thick line) and probabilistic NGM MOS (dashed thick line) forecasts of 2-m temperature for thresholds of (a) ≥303, (b) ≥298, and (c) ≤283 K plotted as a function of α. The relative value is calculated for specified probabilities, but only the upper envelope of these curves is shown as in Richardson (2000). The curves for the BCE mean (solid thin line) and deterministic NGM MOS (dashed thin line) are indicated.

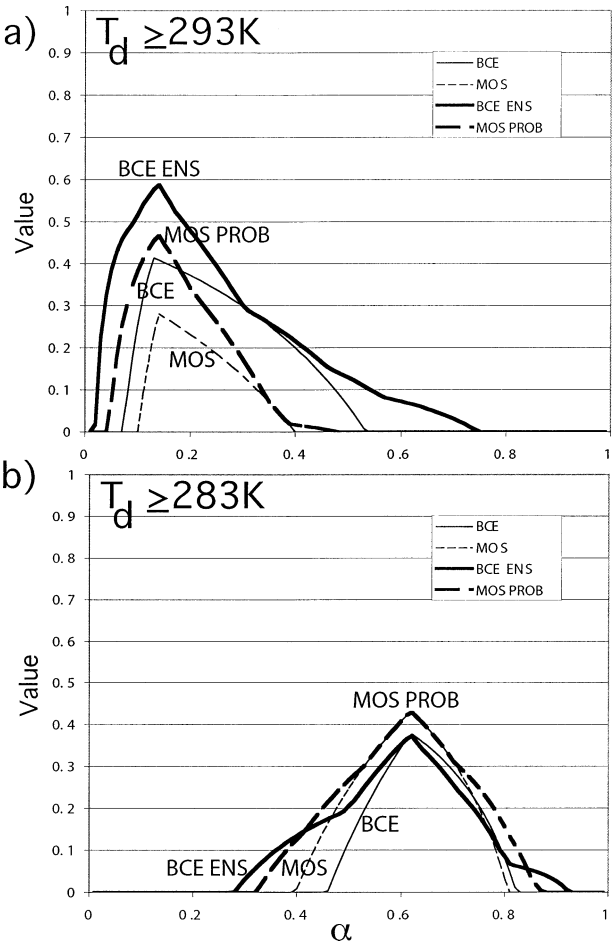

FIG. 14. As in Fig. 13 but for the 2-m dewpoint temperature thresholds of (a) 293 and (b) 283 K.

and ensemble spread (Hamill and Colucci 1998; Stensrud et al. 1999; Hou et al. 2001), it is important to examine this question with the somewhat larger ensemble used in the pilot program that also includes a greater variation in model diversity than previously studied. The approach of Grimit and Mass (2002) is followed in which both the ensemble mean error and the ensemble spread are averaged over the domain at each forecast time period for each forecast day. Both the ensemble spread and error are calculated only at the 395 NWS observing station locations. The resulting time series of ensemble mean error and ensemble mean spread are correlated using the Pearson's correlation coefficient (Wilks 1995). Results indicate that there is a relatively strong correlation between ensemble mean error and mean spread for 2-m temperatures during the daytime hours (Fig. 15), when differences in boundary layer structure are most evident. The correlations are above 0.7 for two forecast times, and above 0.6 for another five forecast times, indicating that between 36% and 50% of the variance is explained.

These correlations are higher than those reported for

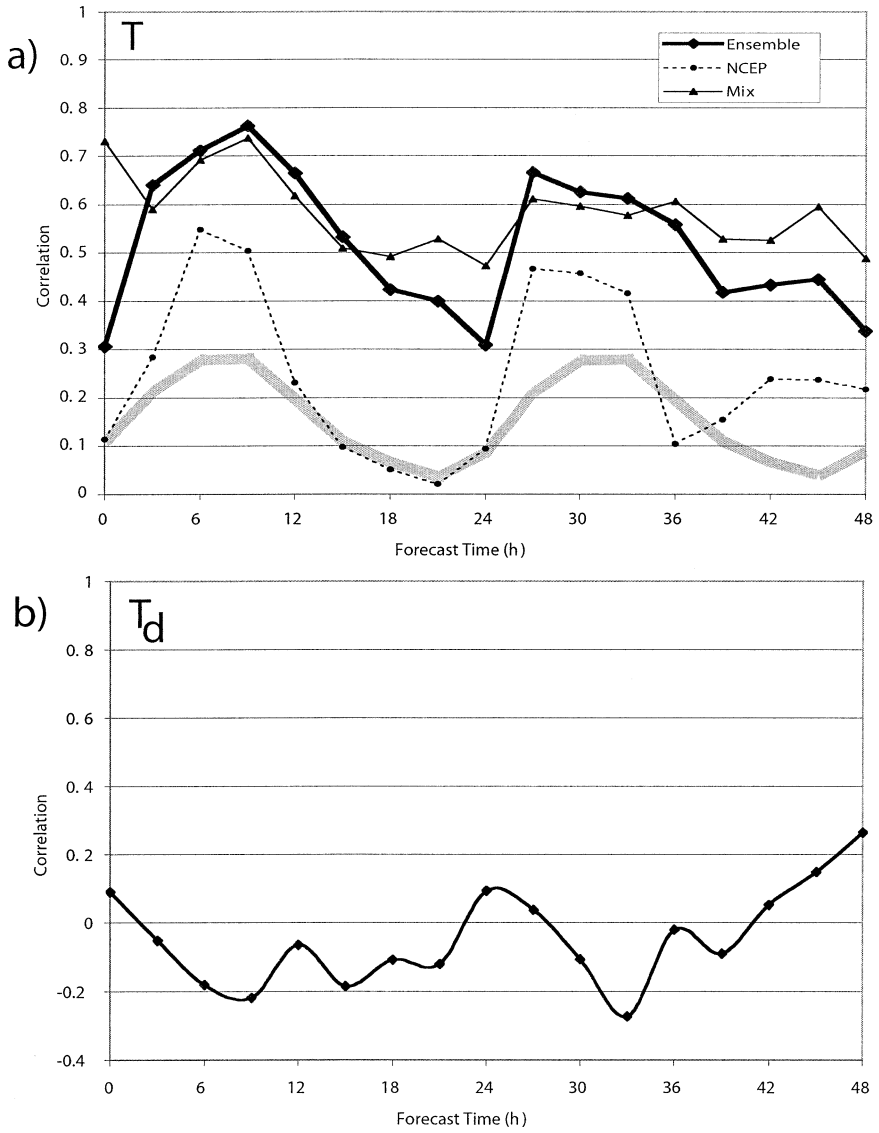

FIG. 15. Correlations between the ensemble mean spread and ensemble mean error as a function of forecast time (h) for (a) 2-m temperature, and (b) 2-m dewpoint temperature. Thick gray line in (a) represents the mean diurnal cycle of the observations. The full ensemble (thick black line) consists of all 23 members, the NCEP ensemble consists of only the 15 Eta, EtaKF, and RSM members, and the Mix ensemble consists of 7 members with one member from each forecast model and from each forecast group (NCEP, NSSL, FSL).

wind direction by Grimit and Mass (2002) and likely are due to the use of multiple modeling systems. This hypothesis is confirmed by an examination of the spread/skill relationship when a 7-member ensemble is created using one forecast from each of the modeling systems. The result is a slight decrease in the maximum correlation at 9 h, but an increase in the correlations during the nighttime hours (Fig. 15). In contrast, the correlation from a 15-member ensemble of model forecasts generated at NCEP using the breeding of growing mode technique shows a smaller spread–skill correlation. Thus, the use of a multimodel ensemble system appears to produce large benefits when trying to relate ensemble spread with the uncertainty of the resulting forecast. Results also show very little correlation between the spread in the dewpoint temperature and the skill of the resulting dewpoint temperature forecast (Fig. 15b). Looking at various subsets of the ensemble forecasts for dewpoint temperature does not improve this result.

6. Discussion

Results from a pilot program exploring the use of short-range ensembles for improved 2-m temperature and dewpoint temperature over New England have been

shown. A simple bias-corrected ensemble mean, where the past 7 complete days of forecasts and observations are used to bias correct both the 2-m temperature and dewpoint temperature predictions for each individual model at each forecast output time, produces slightly smaller bias, MAE, and rmse values than the NGM MOS over the 48 days of study during the warm season. Thus, this simple ensemble approach is competitive with, or better than, NGM MOS, and it has the additional benefit in that it does not require a long data archive to produce good results. While seven frontal passages occurred during the study period, this approach needs to be tested during the cold season when weather regime transitions are more frequent and stronger. It also may be that selecting longer time periods over which the bias correction is calculated may produce more accurate forecasts.

The ensemble probability data further show that the additional information provided by the ensemble can be quite valuable to many end users of weather forecast data when used in a simple cost–loss model. Results indicate that the ensemble adds the most value above that provided by NGM MOS for the more unlikely events. Other postprocessing methods beyond the simple bias correction used here, such as neural networks, may provide even better results. Finally, the ensemble shows some ability to predict forecast skill for 2-m temperature, with a correlation between the ensemble spread and the error of the ensemble mean exceeding 0.7 for some forecast periods. The use of a multimodel ensemble is one reason why this correlation occurs at this relatively high level.

There are many reasons why ensembles should be seriously considered as a routine and important component of any operational short-range forecasting system. Mass et al. (2002) indicate that only very minor improvements, if any, occur with present numerical models as the grid spacing drops below 12 km when verified using standard measures of forecast skill. Brooks et al. (1992) and Stensrud et al. (1999) show that model forecasts can be very sensitive, even in the short-range, to slight changes in the large-scale pattern. Numerous studies have shown that the ensemble mean forecast is more accurate than forecasts from a single model with smaller grid spacing (Du et al. 1997; Hamill and Colucci 1997; Stensrud et al. 1999; Fritsch et al. 2000; Hou et al. 2001; Wandishin et al. 2001; Grimit and Mass 2002). Krishnamurti et al. (1999, 2001) and the present study indicate that simple postprocessing of ensemble forecasts can provide even more useful information. While it is clear that there are many aspects of ensemble forecasting that we do not presently understand, including how best to construct initial condition perturbations for short-range forecasts, the number of studies showing the value of an ensemble approach is growing. Since uncertainty needs to be an explicit part of weather forecasting (Murphy 1977), ensembles appear to be a reasonable and important step towards constructing a fully probabilistic short-range forecasting system.

Acknowledgments. The authors are thankful to Jun Du, Steve Tracton, Michael Baldwin, Jack Kain, Stan Benjamin, Tracy Lorraine Smith, Steven Peckham, and Georg Grell for providing us with the output from the various forecast models used in this experiment. We further appreciate the local computer support provided by Doug Kennedy, Steven Fletcher, and Brett Morrow. Discussions with Harold Brooks and Matt Wandishin were insightful and we are grateful for their help. Constructive and helpful reviews from three anonymous reviewers are greatly appreciated. Partial funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227.

APPENDIX

Calibrating Ensemble Probability Forecasts

The technique used to calibrate the raw ensemble probability forecasts using a rank histogram largely follows that of Hamill and Colucci (1998). Assume that there exists a sorted ensemble temperature forecast $\mathbf{X}$ with N members, and a verifying observation V, and a corresponding representative verification rank histogram distribution $\mathbf{R}$ with $N + 1$ ranks representing the past probability of the verification location compared to the ensemble forecasts over the past 7 complete forecast days (as shown in Fig. 2). The probabilities of forecast events are given by

$$p(V < X_i) = \sum_{j=1}^{i} R_j. \tag{A1}$$

It is assumed that a given rank's probability is linearly distributed between the ensemble member forecasts such that for a given threshold T

$$p(X_i < T) = \sum_{j=1}^{i} R_j + \frac{(T - X_i)}{(X_{i+1} - X_i)} R_{i+1}, \tag{A2}$$

for $X_i < T \leq X_{i+1}$. The difference between this approach and that of Hamill and Colucci (1998) is that the Gumbel distribution (Wilks 1995) is used to calculate the forecast probabilities of events that are both *above* the highest ensemble forecast X_N and *below* the lowest ensemble forecast X_1. Following Hamill and Colucci (1998) the forecast probability that the verification will occur above X_N is

$$p(X_N < T) = \frac{F(T) - F(X_N)}{1.0 - F(X_N)} R_{N+1}, \tag{A3}$$

where F is the cumulative distribution function of the fitted maximum Gumbel distribution. The Gumbel parameters are estimated using the method of moments (Wilks 1995). Thus, we have

$$F = \exp\left[-\exp\left(-\frac{(x - \xi)}{\beta}\right)\right] \quad \text{(A4)}$$

with

$$\beta = \frac{s\sqrt{6}}{\pi} \quad \text{and} \quad \xi = \overline{x} - \gamma\beta, \quad \text{(A5)}$$

where s is the sample standard deviation and γ is Euler's constant (0.577 21). The forecast probability that the verification will occur below X_1 is

$$p(T < X_1) = \frac{G(T)}{G(X_1)}R_1, \quad \text{(A6)}$$

where G is the cumulative distribution function of the fitted minimum Gumbel distribution defined as

$$G = 1.0 - \exp\left[-\exp\left(\frac{x - \xi}{\beta}\right)\right]. \quad \text{(A7)}$$

The verification rank histograms are computed only when all 23 forecasts are available. Calibrating the ensemble probabilities when all 23 forecasts are available follows the steps indicated above in (A1) to (A7) and is straightforward.

However, for the occasional times when today's ensemble has a missing forecast member and a forecast probability is desired, the prior rank histogram must be collapsed. This is because $M + 1$ ranks are needed to calibrate the forecast, where M is the number of forecasts available. Thus, if we only have 21 forecasts, we need a verification rank histogram with 22 ranks. In this case the verification rank histogram with 24 ranks is modified before being used to calibrate the ensemble probabilities. If one ensemble member forecast is missing, then the rank R_i with the lowest probability is removed. This rank's probability is added to the remaining ranks proportional to the relative probabilities in these ranks, resulting in a verification rank histogram with one less rank. Up to three ranks can be removed sequentially in this manner before the forecast data are rejected as being incomplete. Once the number of ranks equals the number of forecasts plus one ($M + 1$), then the procedure outlined above in (A1) to (A7) can be followed with $N = M$.

REFERENCES

Benjamin, S. G., 1989: An isentropic meso-α scale analysis system and its sensitivity to aircraft and surface observations. *Mon. Wea. Rev.*, **117,** 1586–1605.

——, K. J. Brundage, P. A. Miller, T. L. Smith, G. A. Grell, D. Kim, J. M. Brown, and T. W. Schlatter, 1994: The Rapid Update Cycle at NMC. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 566–568.

——, and Coauthors, 2001: The 20-km version of the RUC. Preprints, *14th Conf. on Numerical Weather Prediction,* Fort Lauderdale, FL, Amer. Meteor. Soc., J75–J79.

Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.*, **112,** 693–709.

Black, T. L., 1994: The new NMC mesoscale eta model: Description and forecast examples. *Wea. Forecasting*, **9,** 265–278.

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

Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7,** 120–132.

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

Dallavalle, J. P., 1996: A perspective on the use of model output statistics in objective weather forecasting. Preprints, *15th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 479–482.

Dempsey, C. L., K. W. Howard, R. A. Maddox, and D. H. Phillips, 1998: Developing advanced weather technologies for the power industry. *Bull. Amer. Meteor. Soc.*, **79,** 1019–1035.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125,** 2427–2459.

Dudhia, J., 1993: A nonhydrostatic version of the Penn State–NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121,** 1493–1513.

Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128,** 3104–3127.

Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting*, **15,** 571–582.

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

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17,** 192–205.

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.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.*, **125,** 2487–2515.

Homleid, M., 1995: Diurnal corrections of short-term temperature forecasts using the Kalman filter. *Wea. Forecasting*, **10,** 689–707.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124,** 2322–2339.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Wea. Rev.*, **129,** 73–91.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.*, **121,** 1834–1846.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, **5,** 128–138.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122,** 927–945.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.*, **122,** 3–26.

Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47,** 2784–2802.

——, M. E. Baldwin, P. Janish, and S. J. Weiss, 2001: Utilizing the

Eta Model with two different convective parameterizations to predict convective initiation and evolution at the SPC. Preprints, *Ninth Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 91–95.

Katz, R. W., and A. H. Murphy, 1997: Forecast value: Prototype decision-making models. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 183–217.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

——, and Coauthors, 2001: Real-time multianalysis–multimodel superensemble forecasts of precipitation using TRMM and SSM/I products. *Mon. Wea. Rev.,* **129,** 2861–2883.

Mao, Q., R. T. McNider, S. F. Mueller, and H.-M. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting,* **14,** 190–202.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce beter forecasts? *Bull. Amer. Meteor. Soc.,* **83,** 407–430.

Murphy, A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.

——, 1994: Assessing the economic value of weather forecasts: An overview of method, results and issues. *Meteor. Appl.,* **1,** 69–73.

——, and R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12–19.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

——, M. S. J. Harrison, K. B. Robertson, and A. P. Woodcock, 1996: Joint medium-range ensembles using UKMO, ECMWF, and NCEP ensemble systems. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J26–J28.

Ross, G. H., 1989: Model output statistics—An updateable scheme. Preprints, *11th Conf. on Probability and Statistics in Atmospheric Sciences,* Monterey, CA, Amer. Meteor. Soc., 93–97.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.,* **2,** 191–201.

Stensrud, D. J., and J. A. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting,* **11,** 103–110.

——, H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

——, J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensembles of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

Stull, R. L., 1988: *An Introduction to Boundary Layer Meteorology.* Kluwer Academic, 666 pp.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

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

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.,* **129,** 729–747.

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., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

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

Zhang, D.-L., and R. A. Anthes, 1982: A high-resolution model of the planetary boundary layer—Sensitivity tests and comparisons with SESAME-79 data. *J. Appl. Meteor.,* **21,** 1594–1609.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multimodel ensemble consisting of a few operational models. *Tellus,* **52A,** 280–299.