Evaluation of Eta–RSM Ensemble Probabilistic Precipitation Forecasts

THOMAS M. HAMILL* AND STEPHEN J. COLUCCI

Department of Soil, Crop, and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 8 October 1996, in final form 3 June 1997)

ABSTRACT

The accuracy of short-range probabilistic forecasts of quantitative precipitation (PQPF) from the experimental Eta–Regional Spectral Model ensemble is compared with the accuracy of forecasts from the Nested Grid Model's model output statistics (MOS) over a set of 13 case days from September 1995 through January 1996. Ensembles adjusted to compensate for deficiencies noted in prior forecasts were found to be more skillful than MOS for all precipitation categories except the basic probability of measurable precipitation. Gamma distributions fit to the corrected ensemble probability distributions provided an additional small improvement.

Interestingly, despite the favorable comparison with MOS forecasts, this ensemble configuration showed no ability to ''forecast the forecast skill'' of precipitation—that is, the ensemble was not able to forecast the variable specificity of the ensemble probability distribution from day-to-day and location-to-location. Probability forecasts from gamma distributions developed as a function of the ensemble mean alone were as skillful at PQPF as forecasts from distributions whose specificity varied with the spread of the ensemble. Since forecasters desire information on forecast uncertainty from the ensemble, these results suggest that future ensemble configurations should be checked carefully for their presumed ability to forecast uncertainty.

1. Introduction

Researchers are now exploring short-range ensemble forecasting (SREF) as a possible alternative way of using available computational power for producing numerical weather forecasts. As computational power increases, higher and higher resolution forecasts of the weather become possible. SREF represents an alternative approach, allocating the available computer time to multiple, reduced-resolution integrations.

Although the ensemble methodology is used operationally for medium-range forecasts (Tracton and Kalnay 1993; Toth and Kalnay 1993; Molteni et al. 1996), the practice to date for short-range forecasts was to allocate the available computer resources to a single, high-resolution forecast. It was presumed the atmosphere behaved pseudodeterministically for short-range forecasts; hence, the effects of sensitive dependence on initial condition, or ''chaos'' (Lorenz 1963) and the concomitant loss of forecast skill should dominate only after several days. Though the benefits of higher-resolution forecasts are many, surface features and precipitation display significant spatial variability at short wavelengths and behave chaotically even within the first few hours or days of the forecast (Lorenz 1969; Brooks et al. 1992). Hence, alternatives to single-integration forecasts are being considered. A primary candidate is ensemble forecasting (Leith 1974), whereby a varied set of initial conditions are generated, all consistent with the observations and their errors. Separate deterministic forecasts are integrated from each initial condition. Potentially, an ensemble can have the appealing characteristics of better defining the most likely weather outcome and more accurately assessing probabilities of rare, damaging events. The drawback is the computational necessity of using reduced resolution for the multiple ensemble member forecasts.

Ensemble forecast methodologies are now being considered for use operationally with shorter-range forecasts (0–2 days). This is a new approach, and there are yet many questions. As a first attempt to answer some of these questions, the National Centers for Environmental Prediction has provided a test set of short-range ensemble forecasts generated with the Eta Model (Black 1994; Rogers et al. 1996) and the Regional Spectral Model (RSM; Juang and Kanamitsu 1994). In this dataset, 10 ensemble forecast members were generated using the Eta Model and a mix of perturbation methodologies. Five initial conditions are interpolated from various in-house objective analyses, and five others, a control and four bred initial conditions, are interpolated from the Medium Range Forecast (MRF) ensemble (Toth and Kalnay 1993). Similarly, five ensemble forecast members are generated with the RSM, also using

*Current affiliation: National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author address: Thomas M. Hamill, NCAR/RAP, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: hamill@ucar.edu

TABLE 1. Root-mean-square magnitudes of perturbations for each individual ensemble member domain averaged and averaged over each case. Perturbation is calculated with reference to the ensemble mean excluding that member.

Model/source of IC	500-mb heights (m)	850-mb temperatures (K)
Eta/Bred P1	11.0	0.72
Eta/Bred P2	10.9	0.69
Eta/Opnl	8.3	0.85
Eta/AVN	3.1	0.41
Eta/Control	7.2	1.01
Eta/EDAS	9.9	1.43
Eta/3DVAR	8.4	0.66
Eta/NGM	11.6	1.26
Eta/Bred N1	11.0	0.74
Eta/Bred N2	10.9	0.68
RSM/Control	3.4	0.70
RSM/N1	11.0	0.94
RSM/N2	10.9	0.89
RSM/P1	11.1	0.91
RSM/P2	11.0	0.89

the same MRF control and bred initial conditions. Table 1 provides information on typical perturbation magnitudes for each ensemble member. These magnitudes were determined from an average over 13 case days from September 1995 through January 1996. The magnitudes are measured as a domain average root-mean-square difference of the member forecast relative to the average of all other ensemble members, excluding the member of interest. As shown, the perturbations vary substantially in magnitude. Despite this, the root-mean-square error of the resulting precipitation forecasts for each individual ensemble member were quite similar (Hamill and Colucci 1997, hereafter HC97), indicating that the member forecasts for precipitation could be considered interchangeable. The ensemble forecasts were also found to be underdispersive, with the member forecasts typically resembling each other more closely than the forecasts resembled the verification data. Despite this, HC97 determined that the precipitation forecasts could be postprocessed rather simply to correct for their undervariability, yielding an adjusted ensemble with more desirable statistical characteristics.

Future research may correct or ameliorate the deficiencies noted in this ensemble. In the interim, we demonstrate the existing ensemble configuration may still prove beneficial to the practicing weather forecaster. This paper first reviews the method of HC97 for generating reliable statistical forecasts from an imperfect ensemble (section 2). Other candidate methods for postprocessing the ensemble precipitation forecasts are also described. Next, we quantify the accuracy of PQPFs generated from this prototype Eta–RSM ensemble (section 3). The performance of the ensemble will be examined before and after a correction based on previous model forecasts, as well as before and after the fitting of several plausible gamma distributions. The PQPFs are also compared to the most viable current alternative, forecasts from model output statistics, or MOS (Carter et al. 1989; Dallavalle et al. 1992). Additionally, this paper will address whether the ensemble really can ‘‘forecast precipitation forecast skill’’—that is, forecast the uncertainty of precipitation forecasts (section 4). Section 5 provides conclusions.

In this study, 13 case days were used: 5 September 1995, 18 September 1995, 25 September 1995, 2 October 1995, 23 October 1995, 8 November 1995, 13 November 1995, 20 November 1995, 27 November 1995, 18 December 1995, and 26 December 1995, 23 January 1996, and 31 January 1996, all with forecasts started from 1200 UTC. Verification was limited to MOS sites in the conterminous United States with available forecasts and precipitation data. Over the 13 case days, there are approximately 4000 points with valid forecasts and verifications, or approximately 300 sites on each day. Twelve-hourly precipitation totals valid at the various MOS sites were used as verification.

Since MOS forecasts are prepared using English units of inches for precipitation, this convention will be used throughout this paper. Conversions to millimeters will be supplied where essential (1.0 in. = 25.4 mm).

2. Methodologies for generating probabilistic forecasts

Computer-generated forecasts are never perfect; they inevitably contain a mix of errors due to insufficient model physics, inadequate resolution, and incorrect initial conditions. For the evaluation of an ensemble, a reference is needed. Here the standard of comparison will be a hypothetical ‘‘perfect model’’ ensemble where all errors are attributable to errors in the initial condition. Further, in this perfect-model ensemble, members forecasts are assumed to have independent and identically distributed (iid) errors, and the verification is considered a plausible member of the ensemble, differing from the actual forecasts only by choice of initial condition. Under these assumptions, the value of the verification observation when pooled with N ensemble forecasts and sorted from lowest to highest is equally likely to occur in each of the $N + 1$ possible ranks. Counting the rank of the verification over many independent samples, an approximately uniform distribution is expected across the possible ranks. If the rank distribution was nonuniform, this indicates that the assumptions were not being met; the model was not perfect, or the selection of initial conditions was inappropriate, or both.

Rules must be specified for assigning the rank. Matters are simple when the verification is different from all ensemble members. For example, a verification precipitation forecast of 0.08 in. when pooled with five ensemble forecasts of 0.0, 0.01. 0.03, 0.07, and 0.09 in. is assigned rank 5 of 6. For situations where the verification exactly equals some of the forecast members, such as precipitation forecasts of zero and a verification of zero, a supplemental rule for rank assignment is needed. For these cases, the number (M) of members tied with the verification are counted. A total $M + 1$ uniform random deviates (Press et al. 1992) are generated for the M members and one verification, and the rank of the verification's deviate in the pool of $M + 1$ deviates is determined. All ensemble members with a lower rank have an insignificantly small number (0.0001 in.) subtracted from their values; similarly, all ensemble members with higher rank have the tiny number added. This randomly assigned the rank among the ties without substantially affecting later calculations.

HC97 provides extensive detail on the characteristics of rank distributions from the Eta–RSM ensemble, and Anderson (1996) reviews their usefulness in low-order and climate model ensembles. In general, the Eta–RSM distributions were found to be highly nonuniform, with a greater percentage at the extreme ranks than at the intermediate ranks. This indicates insufficient variability within the ensemble and that the perfect-model assumptions were not met. No results are yet available to indicate whether the insufficient variability was due to model errors, the selection of initial conditions, or both.

Given a rank distribution preferentially populated at the extreme ranks, it is inappropriate to use the relative frequency from the unmodified ensemble to make probabilistic forecasts. For example, just because one-fifth of the ensemble members are above a precipitation threshold, the probability of the event being above the threshold is not necessarily one-fifth. However, if the shape of the rank distribution generated from past model forecasts is representative of the distribution that can be expected for new forecast sample points, then it can be used in conjunction with the member forecasts to assess probabilities. For example, Fig. 1 shows a *hypothetical* rank distribution for precipitation forecasts. Here the rank distribution indicates that the verification is higher than the highest ensemble forecast on average 10% of the time. Hence, subsequent ensemble forecasts can be sorted, and the highest ensemble member can be used to define the event threshold at which the verification is expected to be greater 10% of the time. Similarly, the verification is likely to be higher than the second highest ensemble member 17% of the time, the sum of the top two ranks. Continuing in this manner, points in the probability distribution can be estimated. Unfortunately, in such a case there is no specific information on the distribution of probabilities above the 90th percentile, and the probability of extreme events such as heavy rainfall are of great interest. Hence, an alternative method will be necessary to assign probabilities in the tails.

A method for calibrating an ensemble forecast using rank histogram information is now described. This overall methodology will hereafter be referred to as the "corrected ensemble" forecast, and is also discussed in HC97. Suppose there is a sorted ensemble precipitation forecast **X** with N members, a verifying observation V, and a corresponding representative verification rank histogram distribution **R** with $N + 1$ ranks representing the past probability of the verification location compared to the ensemble. Then probabilities of forecast events can be assigned using (1):

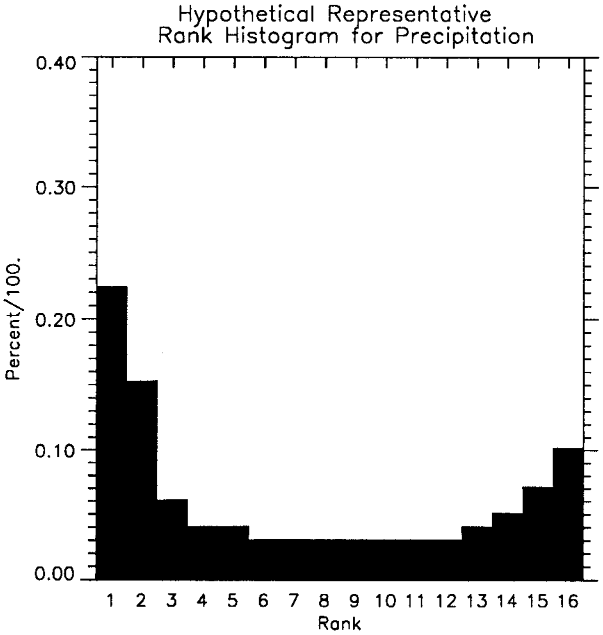

FIG. 1. Hypothetical rank distribution for precipitation forecast with 15 members.

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

or equivalently, above the first rank

$$p(X_{i-1} \le V < X_i) = R_i. \qquad (2)$$

The following additional assumptions were also made. First, the rank histogram probability is uniformly distributed between the lowest ensemble member and zero. For a threshold T less than the lowest ensemble forecast X_i,

$$p(0 \le V < T) = \left(\frac{T}{X_1}\right) R_1, \qquad 0 < T < X_1. \qquad (3)$$

For example, if the lowest ensemble member forecast were 0.03 in., the threshold 0.01 in., and the probability of the verification occurring below the lowest ensemble member 15%, the probability of 0.0–0.01 in. is set to 5%. Similarly, it is assumed that a given rank's probability is equally distributed between ensemble members:

$$p(X_i \le V < T) = \left(\frac{T - X_i}{X_{i+1} - X_1}\right) R_{i+1}, \qquad X_i < T \le X_{i+1} \qquad (4)$$

and

$$p(T \le V < X_{i+1}) = \left(\frac{X_{i+1} - T}{X_{i+1} - X_i}\right) R_{i+1}, \qquad X_i < T \le X_{i+1}. \qquad (5)$$

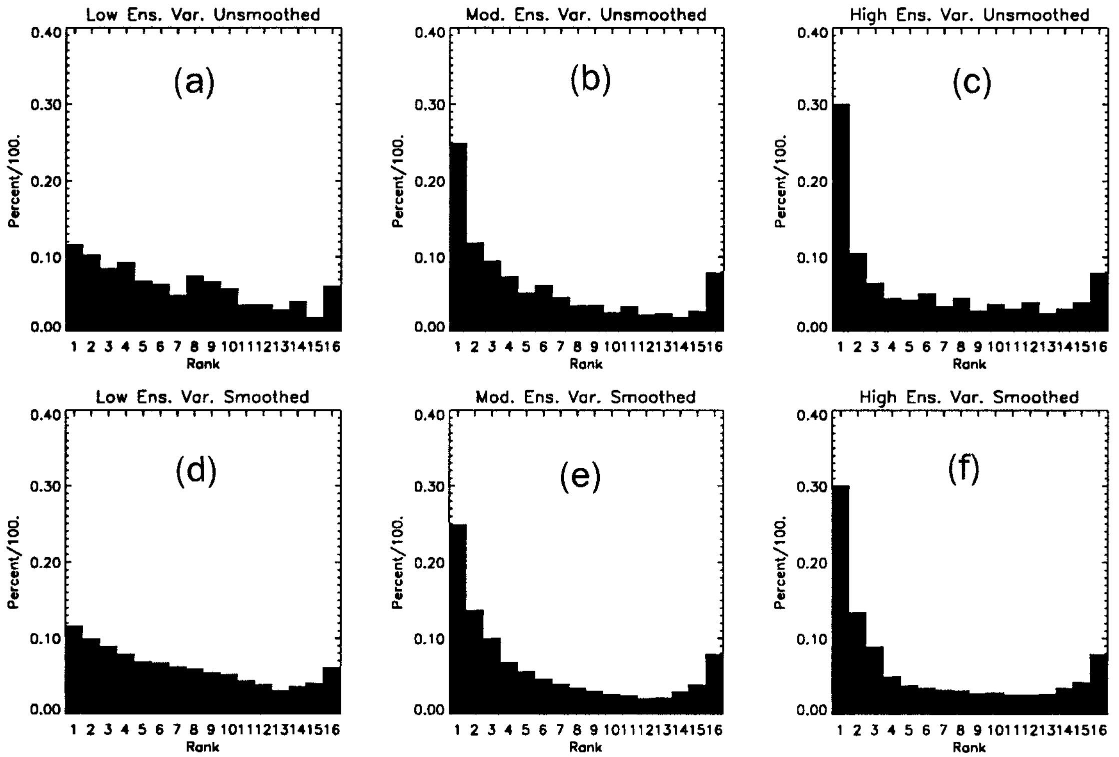

FIG. 2. Cross-validated rank histograms for 23 January 1996 as a function of 24–36-h precipitation forecast ensemble variability at MOS sites, before and after application of smoother. (a) Low ensemble variability (EV) histogram before smoothing. (b) Moderate EV before smoothing. (c) High EV before smoothing. (d) Low EV after smoothing. (e) Moderate EV after smoothing. (f) High EV after smoothing.

However, assumption of uniformity of probability beyond the highest ensemble forecast X_N is certainly inappropriate. For example, given the highest ensemble forecast is 0.75 in., the probability of 1–2-in. precipitation should typically be greater than the probability of 2–3 in. Hence we assume that the probability beyond the highest ensemble member has the *shape* of a Gumbel distribution (Wilks 1995) fit to the ensemble data by the method of moments. The Gumbel distribution is the distribution of choice for assigning probabilities to extreme events. Given the cumulative distribution function F of the fitted Gumbel distribution, the forecast probability that the verification will occur above X_N and below the next threshold is

$$P(X_N \leq V < T) = \frac{F(T) - F(X_N)}{1.0 - F(X_N)} R_{N+1}. \qquad (6)$$

Similarly, the probability that the verification will be between any two thresholds $T_2 > T_1 > X_N$ is defined as

$$P(T_1 \leq V < T_2) = \frac{F(T_2) - F(T_1)}{1.0 - F(X_N)} R_{N+1}. \qquad (7)$$

For a practical example of how to use (1)–(6), see the appendix.

Using these equations, at each MOS site, the ensemble data and rank histograms were used to generate probabilities for each MOS precipitation category. The MOS categories here are $0 \leq V < 0.01$ in., $0.01 \leq V < 0.10$, $0.10 \leq V < 0.25$, $0.25 \leq V < 0.5$, $0.5 \leq V < 1.0$, $1.0 \leq V < 2.0$, and $2.0 \leq V$ (0.01, 0.10, 0.25, 0.50, 1.00, and 2.00 in. equals 0.2, 2.5, 6.4, 12.7, 25.4, and 50.8 mm, respectively). The rank histograms were generated using 12-h observed precipitation totals at the MOS sites as verification and the technique of cross-validation, whereby all sample points from all case days except the forecast day of interest are used to generate rank histograms (note that this a much shorter training dataset than is used with MOS). The shape of the rank histogram changed significantly with ensemble variability, or "spread," defined as the standard deviation of the ensemble about its mean. Hence, a different rank histogram was used for low, moderate, and high ensemble variability forecasts. A low ensemble variability (EV) was defined as EV $<$ 0.03 in.; moderate, $0.03 \leq$ EV $<$ 0.12 in., and high, $0.12 \leq$ EV. Further, the rank histograms were smoothed with a running line smoother (Hastie and Tibshirani 1990) to smooth out the variations in the rank histograms due to small sample size. Sample unsmoothed and smoothed rank histograms at low, moderate, and high ensemble variability are shown in Fig 2. Further stratification by geographical area and/

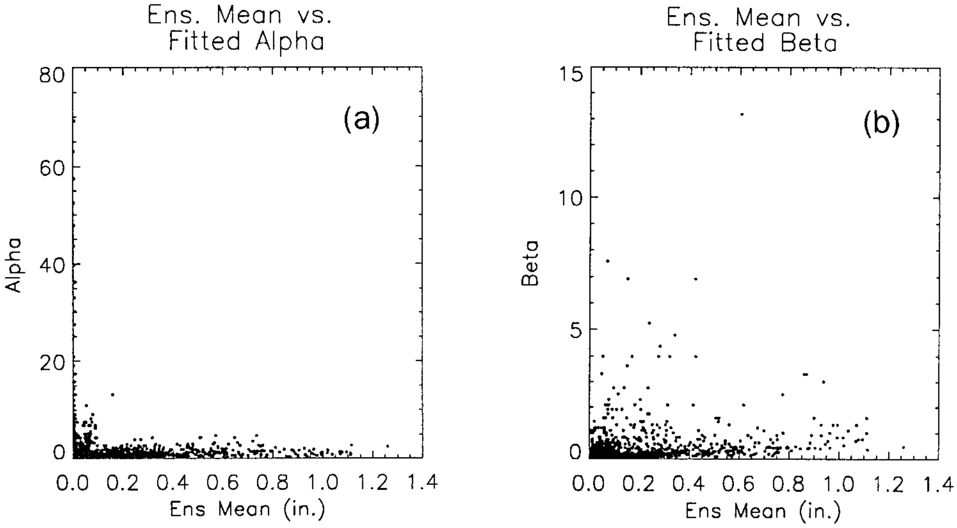

FIG. 3. Scatterplots of fitted gamma parameters α and β as a function of the ensemble mean for 5 September 1995: (a) α vs mean and (b) β vs mean.

or climate regime may prove beneficial in future studies when larger training datasets are available.

To compare the corrected ensembles with an uncorrected ensemble, the same equations (1)–(6) were used, but a uniform rank distribution was used—that is, $R_1 = R_2 = \cdot\cdot\cdot = R_{N+1} = 1/(N + 1)$. This will be referred to as the ‘‘uncorrected’’ forecast. This method yields similar results to setting probabilities by relative frequency.

Two gamma distributions were also generated from each ensemble forecast point, and the probabilities were evaluated for the MOS categories. Gamma distributions were chosen for their ability to take on a variety of shapes based on the distribution of the input data; they are used frequently to fit distributions to precipitation climatologies (e.g., Wilks 1995). For the first of the two fitted gamma distributions, the shape of the distribution was designed to vary with the spread of the ensemble, as do the corrected ensemble forecasts; when the spread of this ensemble is small, the probability distribution is rather sharp, and vice versa. Hence, this first gamma distribution was selected, which best fit the corrected ensemble forecast. However, it was difficult to accurately fit gamma distributions to the corrected forecasts partitioned to the coarsely binned MOS precipitation categories, so a temporary alternative corrected forecast was generated using a larger number of categories ($0 \leq V < 0.01$ in., $0.01 \leq V < 0.03$, $0.03 \leq V < 0.06$, $0.06 \leq V < 0.10$, $0.10 \leq V < 0.20$, $0.20 \leq V < 0.35$, $0.35 \leq V < 0.50$, $0.50 \leq V < 0.75$, $0.75 \leq V < 1.0$, $1.0 \leq V < 1.5$, $1.5 \leq V < 2.0$, $2.0 \leq V < 3.0$, $3.0 \leq V < 4.0$, and $V > 4.0$ in.). The same rank histograms and methodology that were used to generate the original corrected ensemble were used here. Next, a set of gamma distributions was generated by varying the parameters α and β through a range of values spanning the range of distributions realistic to precipitation forecasts. For each distribution the probabilities were computed for each of the previously listed categories. The particular (α, β) combination that most closely fit the alternative corrected forecast was selected. Probabilities were then computed for the MOS categories. This method of distribution fitting was developed because more common methods of distribution fitting proved inadequate for forecasts including many zero precipitation events (Wilks 1990, 1995).

A second set of gamma distributions was also developed. These gamma distributions were a function of only the ensemble mean. If forecasts from gamma distributions that vary with the ensemble spread are more skillful than these more generic gamma distributions, this then indicates some ability of the ensemble to forecast the forecast skill. First, for each MOS location on each case day, the value of the ensemble mean and the previously fitted (α, β) combination described above were archived. Next, through cross-validation, a scatterplot of the fitted α and β versus the ensemble mean were generated using data from all other case days. Representative plots are shown in Figs. 3a and 3b. As shown, statistical relationships are obscured by the strong nonnormality of the data. Hence, power transformations (Wilks 1995) were applied to α, β, and the ensemble mean. Figure 4a and 4b plot transformed ln(α) and ln(β) against a transformed ensemble mean, with the transformed ensemble mean $\overline{X}'$ defined by

$$\overline{X}' = \frac{(\overline{X} + 0.01)^{-0.3} - 1.0}{-0.3}. \tag{8}$$

As shown, the resulting distribution is much more normally distributed and easier to interpret. Next, a running line smoother (Hastie and Tibshirani 1990) was applied to determine the optimal transformed ln(α) and ln(β) as

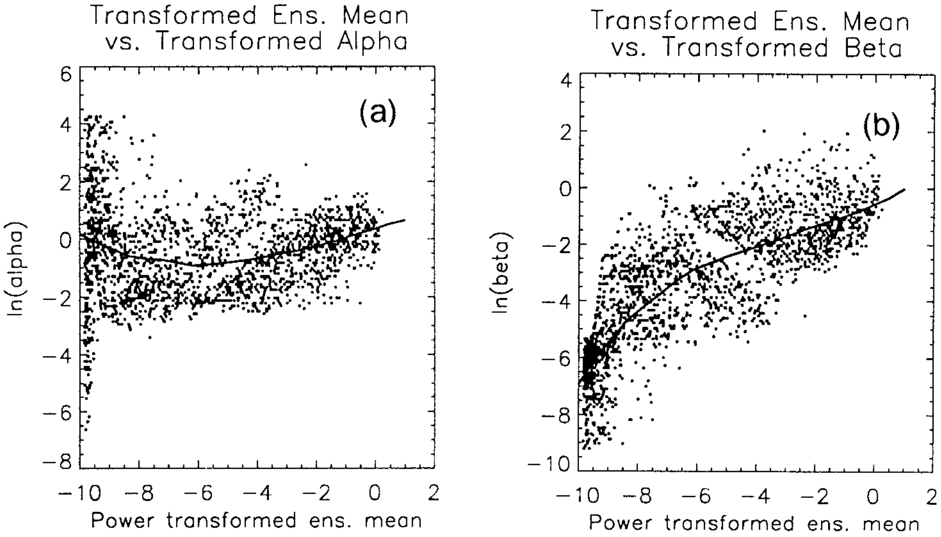

FIG. 4. Scatterplots of fitted (and power transformed) gamma parameters α and β as a function of the power-transformed ensemble mean: (a) α vs mean and (b) β vs mean. Fitted regression line is overplotted on each.

a function of $\overline{X}'$. The running line smoother used a neighborhood of 2.5 and a Gaussian kernel with a standard deviation of 0.7. The optimally fitted values were overplotted in Figs. 4a and 4b. Next, inverse transforms were applied to the regression relationship to predict α and β simply as a function of $\overline{X}$. Some resulting probability density functions are illustrated in Figs. 5a–d. Generally, the parameter estimates nicely meet the constraint that the expected value $E(\overline{X}) = \alpha\beta$ at low precipitation thresholds, but not as well at higher thresholds. Other methods, such as including variational constraints on the product $\alpha\beta$ while selecting the parameters were not tried.

3. Comparison against MOS forecasts

Despite the theoretical appeal of forecasting precipitation amount probabilistically, it is rarely done. Automated probabilistic precipitation forecasts are generated by the NGM MOS system (Carter et al. 1989; Dallavalle et al. 1992). Unlike perfect prog approaches (Wilks 1995), MOS can compensate for systematic errors in the forecast model. The notable disadvantages of the MOS technique are that many training case days are necessary to sample adequately the range of potential weather regimes, and the model physics or resolution should not be changed once the predictive equations have been developed. This retards the rapid development and implementation of model improvements. The Eta Model has since replaced the NGM as the primary development model at NCEP, but because of frequent improvements to the Eta Model, no MOS forecasts have been developed for it. Hence, the NGM MOS still provides the most sophisticated automated statistical guidance for precipitation routinely available in the United States.

MOS forecasts are disseminated to the field in the National Weather Service's FOUS14 bulletin. This bulletin gives unconditional probabilities of measurable precipitation in 12-h increments as well as a ''best'' precipitation category, but the full information of probabilities for each precipitation is not transmitted regularly as part of this bulletin. However, such probabilities are generated in house by the MOS developers at the Techniques Development Lab (TDL) and were obtained for comparison against the ensemble. For this comparison, quantitative precipitation probabilities were obtained for the MOS 12–24-, 24–36-, and 36–48-h forecasts for the mutually exclusive and collectively exhaustive categories $0 \leq V < 0.01$ in., $0.01 \leq V < 0.10$, $0.10 \leq V < 0.25$, $0.25 \leq V < 0.5$, $0.5 \leq V < 1.0$, $1.0 \leq V < 2.0$, and $2.0 \leq V$.

The overall accuracy of the probability distribution generated from each forecast is evaluated by the ranked probability skill score (Wilks 1995), or ''RPSS.'' This is based on the ranked probability score (Epstein 1969; Murphy 1971; Daan 1985), which compares the cumulative distribution vector derived from the verification to the cumulative distribution vector derived from the forecast. Here, the RPSS measures the fractional improvement in ranked probability score over MOS. Higher scores are better, with 1.0 indicating a perfect forecast and 0.0 indicating the skill of the MOS forecast. Forecasts are also evaluated here using the Brier skill score, or ''BSS'' (Brier 1950; Wilks 1995) for various precipitation thresholds. Again, scores are computed against the reference MOS forecast, and higher scores are better.

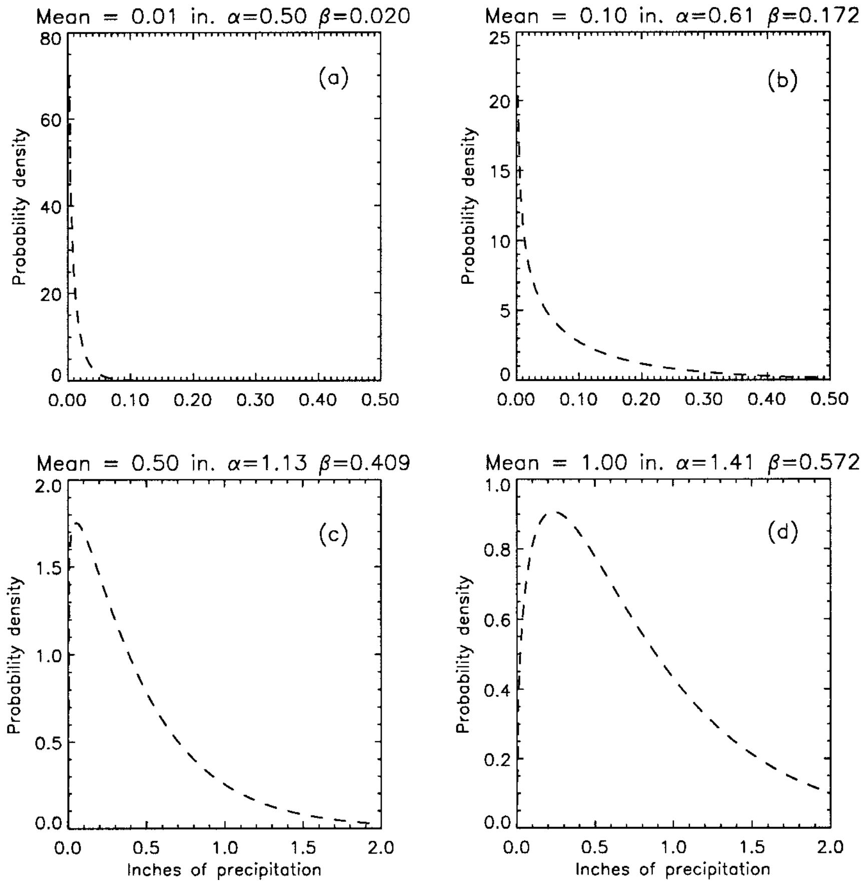

FIG. 5. Representative probability density functions fitted to various ensemble means. (a) Distribution for 0.01-in. ensemble mean, (b) 0.10 in., (c) 0.50 in., and (d) 1.0 in.

Tables 2–4 summarize the RPSSs. Table 2 shows RPSSs for all forecasts combined; Table 3 shows RPSSs for the subset of sample points where the verification was greater than 0.25 in. Finally, Table 4 shows RPSSs for the subset of points where the ensemble mean was greater than 0.25 in. As shown, for the sample as a whole, the MOS is the best performer at 24–36 and 36–48 h, but the fitted gamma distribution was the best performer at 24 h. However, as indicated in Tables 3 and 4, ensemble-based methods consistently outperform MOS in the subsets with higher precipitation events. However, the scores are more variable between these subsets and are based on a smaller sample, and should thus be regarded as less trustworthy. Nonetheless, these results are quite encouraging, especially the competitive performance of the ensembles for higher precipitation amounts and considering the small training dataset used to establish the rank histograms.

Similar performance was seen in the BSS, as shown in Table 5. For the lowest precipitation thresholds, MOS generally outperforms the ensemble forecasts, and vice versa for all higher precipitation thresholds. Interestingly, the gamma distribution fit to the ensemble mean was frequently competitive with the other ensemble-based forecasts, an indication that the presumed ability of the ensemble to forecast the precipitation forecast skill should be questioned.

Another important characteristic of probabilistic forecasts is their reliability, or calibration, which measures the relationship between the forecast probability and the relative frequency of event occurrence at a given probability. Reliability diagrams for $p > 0.10$ in. at 12–24 h are shown in Fig. 6; the decomposition of the Brier score into reliability, resolution, and uncertainty (Murphy 1973) are also indicated in this figure. The forecasts appear reasonably well calibrated, except for the uncorrected ensemble forecasts, which show a tendency to overforecast the likelihood of precipitation over 0.10 in. Similar conclusions were drawn from the interpretation of reliability diagrams for other thresholds and forecast intervals (not shown).

4. Forecasting the forecast skill

Ensemble forecasts are expected to provide information on the variable uncertainty of the precipitation forecast, that is, the extent to which forecast probabilities are to be dispersed across the MOS categories.

TABLE 2. Ranked probability skill scores for MOS and four ensemble forecast methodologies averaged over all sample points. The asterisk indicates highest RPSS value.

Time of forecast	Sample size	MOS	Uncorrected ensemble	Corrected ensemble	Gamma on corrected	Gamma on ens. mean
12–24 h	3901	0.0	−0.140	−0.001	0.004*	−0.023
24–36 h	4014	0.0*	−0.194	−0.033	−0.027	−0.024
36–48 h	3868	0.0*	−0.198	−0.033	−0.032	−0.018

TABLE 3. As in Table 1 but for the subset of points where the ensemble mean forecast was greater than 0.25 in.

Time of forecast	Sample size	MOS	Uncorrected ensemble	Corrected ensemble	Gamma on corrected	Gamma on ens. mean
12–24 h	296	0.0	−0.069	0.092	0.099*	0.081
24–36 h	334	0.0	−0.016	0.158	0.167	0.176*
36–48 h	307	0.0	−0.156	0.030	0.027	0.048*

TABLE 4. As in Table 1 but for the subset of points where the verification was greater than 0.25 in.

Time of forecast	Sample size	MOS	Uncorrected ensemble	Corrected ensemble	Gamma on corrected	Gamma on ens. mean
12–24 h	233	0.0	0.282*	0.152	0.157	0.210
24–36 h	272	0.0	0.341*	0.162	0.173	0.246
36–48 h	203	0.0	0.293*	0.016	0.028	0.109

However, as shown previously, forecasts generated from gamma distributions, which are only a function of the ensemble mean, were generally competitive with other ensemble forecast methods, especially at higher precipitation thresholds. We now attempt to determine more specifically whether this particular ensemble has the ability to forecast the forecast skill.

To examine this, consider first the decomposition of squared error of the ensemble X at a particular point and time into bias and variability components (e.g., Brankovic et al. 1990):

$$\overline{(X_f - V)^2} = (\overline{X}_f - V)^2 + \overline{(X_f - \overline{X}_f)^2}. \qquad (9)$$

Here the subscript f represents an individual ensemble member forecast, the overbar represents an average over all ensemble members, and V is the verification. The first term on the right-hand side is the square of the bias of the ensemble, representing how far the ensemble mean is from the verification. The square root of this term will hereafter be called the ‘‘absolute bias.’’ The second terms represents the spread, or variability of the ensemble; its square root will be denoted as the ‘‘ensemble variability.’’ Generally, if indeed the verification can be considered a member of the ensemble, as is assumed with a perfect ensemble, then when the

TABLE 5. Brier skill scores for MOS and four ensemble methodologies for various thresholds and forecast intervals. The asterisk indicates highest Brier skill score for this threshold/forecast interval.

Threshold	Time of forecast	MOS	Uncorrected ensemble	Corrected ensemble	Gamma on corrected	Gamma on ens. mean
0.01 in.	12–24 h	0.000*	−0.199	−0.032	−0.042	−0.100
	24–36 h	0.000*	−0.340	−0.162	−0.153	0.167
	36–48 h	0.000*	−0.252	−0.104	−0.101	−0.078
0.10 in.	12–24 h	0.000	−0.115	0.024	0.035*	0.013
	24–36 h	0.000	−0.174	0.014	0.014	0.020*
	36–48 h	0.000	−0.193	0.018	0.013	0.025*
0.25 in.	12–24 h	0.000	−0.086	0.029	0.045*	0.034
	24–36 h	0.000	−0.051	0.075	0.082	0.086*
	36–48 h	0.000	−0.197	0.008	0.010*	0.010*
0.50 in.	12–24 h	0.000	−0.166	0.009	0.017	0.025*
	24–36 h	0.000	−0.118	0.043	0.065	0.104*
	36–48 h	0.000	−0.049	0.071*	0.068	0.063
1.00 in.	12–24 h	0.000	0.009	0.012	0.033	0.034*
	24–36 h	0.000	0.021	0.066	0.055	0.079*
	36–48 h	0.000	0.002	0.040	0.054*	0.033

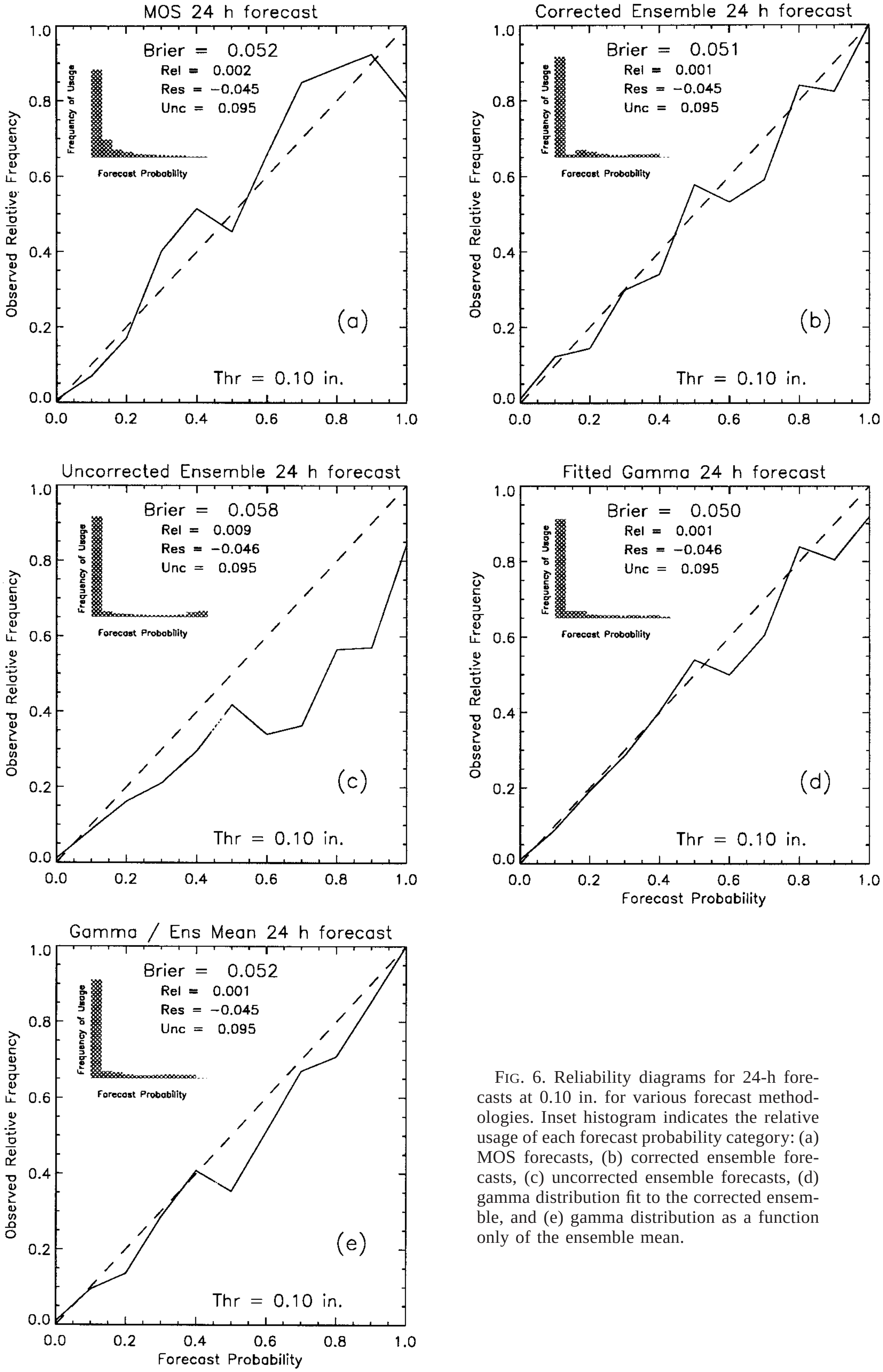

FIG. 6. Reliability diagrams for 24-h forecasts at 0.10 in. for various forecast methodologies. Inset histogram indicates the relative usage of each forecast probability category: (a) MOS forecasts, (b) corrected ensemble forecasts, (c) uncorrected ensemble forecasts, (d) gamma distribution fit to the corrected ensemble, and (e) gamma distribution as a function only of the ensemble mean.

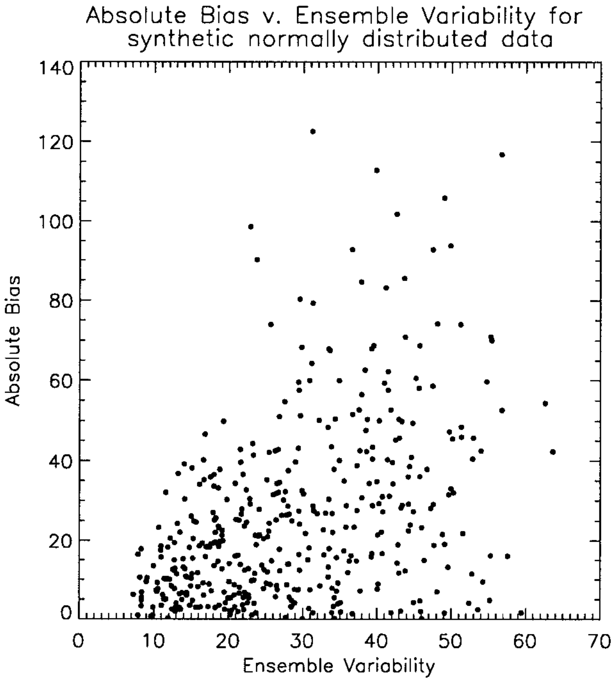

FIG. 7. Scatterplot of synthetic ensemble dataset's absolute bias vs ensemble variability.

ensemble is more dispersed and the ensemble variability is larger, then the expected value of the absolute bias should be larger as well. To illustrate this, a synthetic group of normally distributed data was created. A total of 10 sets of random normal data were created for each variance between 10 and 50 in increments of 1, yielding a total of 410 sets. For each individual set, 16 random, normally distributed samples were created. One sample was arbitrarily denoted the verification *V*, and the remaining 15 were denoted the ensemble. From this, an absolute bias and ensemble variability was calculated and plotted in Fig. 7. As shown, though there is much scatter, as the ensemble variability increases, there is a tendency for the absolute bias to increase as well. Hence, a real ensemble ought to show some ability to forecast its forecast skill based on the ensemble variability. However, in evaluating the ability to forecast the forecast skill, the variability of high-precipitation events should not be compared to variability of low-precipitation events. If done this way, the ''spread–skill'' relationship is contaminated by the ensemble mean, since low-precipitation events usually have lower variability than high-precipitation events. A more meaningful analysis must distinguish the ability for two forecasts with the same ensemble mean but differing ensemble variabilities to differently forecast the forecast skill.

To examine whether the ensemble can forecast the forecast skill in the same manner that would be expected of a perfect-model ensemble, we constructed such a perfect-model ensemble for comparison (see also Buizza 1997). As shown in HC97, the error characteristics of each individual member's precipitation forecasts were very similar. Hence, for the perfect-model ensemble, a new synthetic verification was constructed at each sample point by randomly using one of the 15 forecasts, leaving 14 remaining forecasts in the ensemble. Statistics such as ensemble mean and variability were calculated from the remaining 14 members. For the real ensemble, the verification data was obtained from precipitation analyses derived from the River Forecast Center precipitation database as in HC97, and the sampling locations for the ensemble were also done as in HC97.

Scatterplots of the ensemble variability plotted as a function of the ensemble mean are shown in Figs. 8a and 8b for both the real and perfect-model data. As

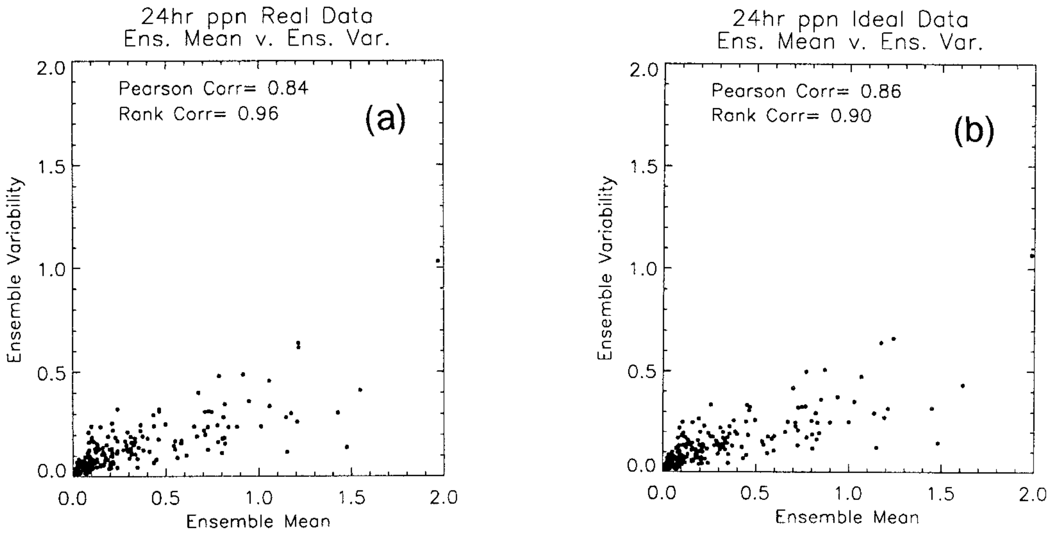

FIG. 8. Scatterplots of 24-h precipitation forecast ensemble variability as a function of the ensemble mean for (a) real ensemble data and (b) perfect-model ensemble.

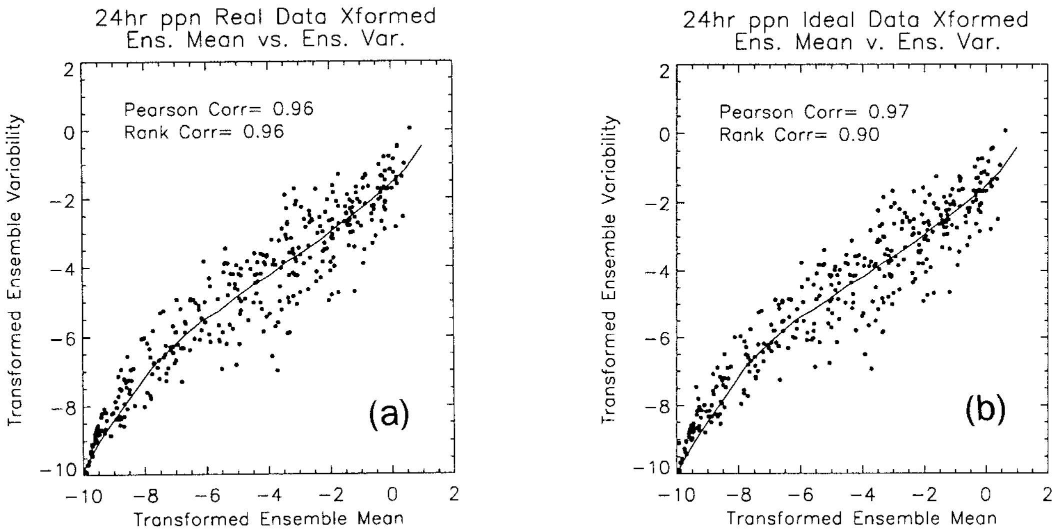

FIG. 9. Scatterplots of power-transformed ensemble variability as a function of the transformed ensemble mean: (a) real ensemble and (b) perfect-model ensemble. Regression line dividing above and below average subsets overplotted.

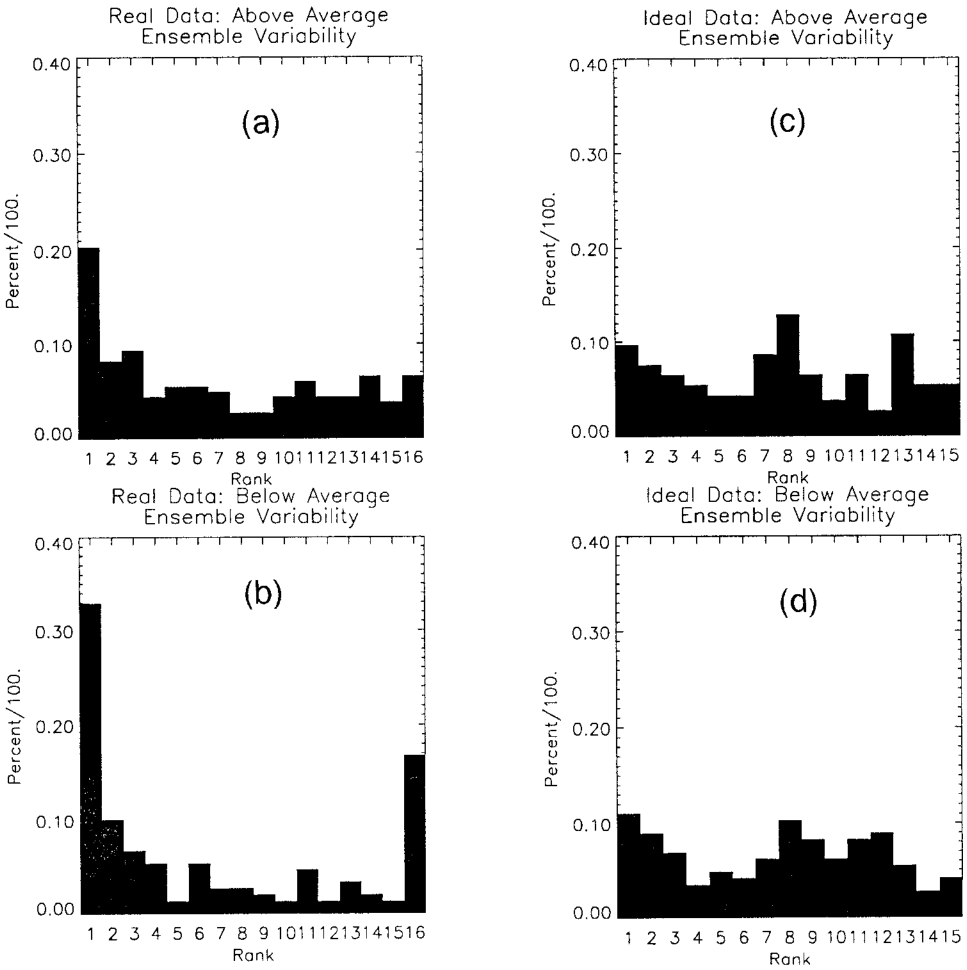

FIG. 10. The 0–24-h total precipitation forecast rank histograms for low- and high-variability subsets: (a) high-variability subset for real data, (b) low variability subset for real data, (c) high variability subset for perfect-model data, and (d) low variability subset for perfect-model data.

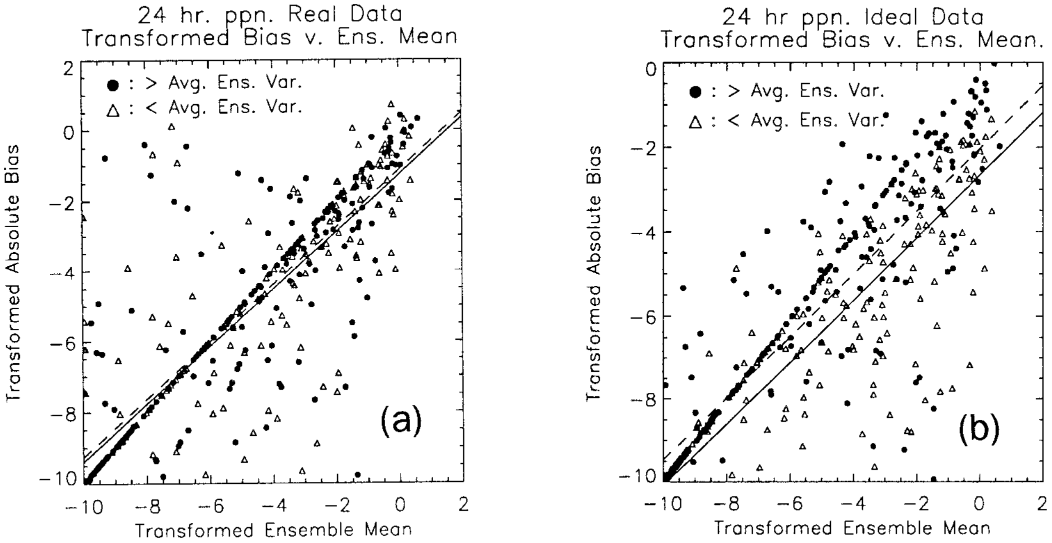

FIG. 11. Scatterplots of power-transformed absolute bias as a function of the transformed ensemble mean for each subset, with the above average variability subset represented with dots and the below average subset with triangle: (a) real ensemble and (b) perfect-model ensemble. Regression lines for above average (dashed) and below average (solid) ensemble variability subsets are overplotted.

shown, the data is highly nonnormally distributed, so both the ordinate and abscissa were power transformed as in (8) and replotted in Figs. 9a and 9b. Each dataset was also divided into two halves, one with above average ensemble variability for a given ensemble mean, and the other with below average variability. The dividing line for the two is overplotted in Fig. 9; this nonparametric regression line was generated with a running line smoother with a neighborhood of 2.0 and a Gaussian kernel with a standard deviation of 0.2.

Figures 10a–d plot rank histograms for low- and high-variability subsets of both real and perfect-model data. For the real data in Figs. 10a and 10b, the extreme ranks are much more highly populated for the low variability subset, indicating that when the ensemble variability is lower than average, then the ensemble is typically underdispersive to a greater extent than for higher than average ensemble variability. Conversely, for the perfect-model data in Figs. 10c and 10d, the histograms are relatively uniform both for above and below average subsets.

We now attempt to quantify whether the real ensemble data shows a statistically significant ability to forecast the forecast skill. Plots of the absolute bias as a function of the ensemble mean [after power transformations to each using (8) are plotted in Figs. 11a and 11b]. For the perfect-model data, the subset with above average ensemble variability appears to have higher absolute bias than the subset with below average ensemble variability. However, there appears to be much more overlap in the distributions with the real data in Fig. 11a. To assess the statistical significance of this difference, a regression equation was fit to the data of the form

$$AB_t = b_0 + b_1\overline{X}' + b_2I. \qquad (10)$$

Here AB_t is the predicted transformed absolute bias, $\overline{X}'$ is the transformed ensemble mean, and I is an indicator variable ($I = 1$ for above average $\overline{X}'$, $I = 0$ for below average). Use of a regression equation with this form permits one regression equation to describe the entire dataset. The most important coefficient from the regression is b_2, which measures the magnitude of the discrimination of absolute bias between the high- and low-variability subsets. The regression lines for $I = 0$ and $I = 1$ are overplotted in Fig. 11. After regression, the coefficient b_2 is equal to 0.729 for the perfect-model data and 0.134 for the real data. To quantify the statistical significance of the magnitude of b_2, a resampling permutation test (Wilks 1995) was performed. The regression analysis was redone, and the magnitude of b_2 noted. This was repeated a total of 1000 times, yielding a sampling distribution of b_2 that would be expected under the null hypothesis of no difference in absolute bias between subsets. For the perfect-model data, the original b_2 was higher than all 1000 resampled b_2's, indicating a statistically significant ability to discriminate between higher and lower than average error. For the real data, however, 230 of the 1000 resampled b_2's were higher than the original b_2, indicating that there is little evidence to conclude the real ensemble data can ''forecast the forecast skill,'' even crudely. Apparently, the ensemble does not adequately discriminate between above and below average variability subsets.

5. Conclusions

This paper tested the skill of various probabilistic precipitation forecasts generated from a prototype short-range ensemble against the current operational standard, MOS. Corrected ensemble forecasts and gamma distributions fit to these forecasts were competitive with MOS forecasts, and even slightly outperformed MOS for thresholds higher than 0.01 in. This result was based on 13 case days over primarily the fall and winter seasons, so the apparent positive benefit of the ensemble must be regarded as preliminary until tested with a larger number of case days over all seasons. Nonetheless, this ensemble was competitive with MOS despite a small training sample, indicating that it may be possible to generate probabilistic forecasts from ensembles that are as skillful as MOS yet do not require a long training dataset. This would permit more rapid implementation of model improvements, since the forecast system would not have to be frozen for many years so the forecasts behave similarly to the training data.

Interestingly, though the probabilistic precipitation forecasts were competitive with MOS, the important information content can be extracted by judiciously using the ensemble mean. The presumed ability of the ensemble to accurately forecast the precipitation forecast skill using the dispersion of the ensemble could not be demonstrated. This was noted first in the similar skill of probability forecasts generated from gamma distributions whose specificity either varied or did not vary with the spread of the ensemble. This was further demonstrated through a comparison of the real ensemble data and a perfect-model ensemble dataset, each divided into subsets with above and below average variability. Whereas the perfect-model data showed the ability to discriminate between higher and lower than average precipitation forecast error, the real ensemble data showed no such ability.

There are a number of interesting issues raised by this research. First, though other authors (e.g., Molteni et al. 1996) have shown some ability of medium-range ensembles to forecast the midtropospheric forecast skill on the large scale, users should not assume the skill of surface weather effects at specific locations can also be forecast from day to day until rigorous testing confirms this. Second, since a forecast of the skill is often desired, research is needed into designing an ensemble forecast system that will indeed be able to forecast the forecast skill better, whether through a different perturbation methodology, changes to the model physics, or changes in the postprocessing strategy. Third, there is interest in using the spread of short-range ensemble weather forecasts to find areas where adaptive observations would produce the most improvement to the analysis and subsequent forecast (e.g., Emanuel et al. 1996). Areas with greater than normal spread would be preferentially targeted. The success of such a strategy is predicated on the operational short-range ensemble forecast possessing a temporally continuous spread–skill relationship *at specific points*, not for the domain as a whole. We suggest the methodology demonstrated here can be used to test the extent to which such a spread–skill relationship exists. The work here with precipitation forecasts suggests that the relationship may not be as strong as presumed.

Last, we suggest that ensemble model development should include testing to determine an optimal ensemble size and resolution. Coarser resolution forecasts are generally less accurate than finer-resolution forecasts, so an increase in the size of the ensemble typically comes at the expense of somewhat lessened accuracy of each member forecast. The decrease in error through ensemble averaging is largest when the ensemble size is small; increasing the size of the ensemble from 1 to 10 members lowers the error substantially. Further increasing from 10 to 100 members does little to improve the ensemble mean, even if all are computed at the same resolution (Leith 1974; Du et al. 1997). This suggests if the size is based only on the accuracy of ensemble mean, a moderately sized ensemble is likely to yield the lowest error. If probabilistic assessments are important, additional forecast members may prove useful for assessing the probabilities of rare events. However, the results with the Eta Model experiments above suggest that this particular model configuration was not able to forecast precipitation forecast skill, and competitive probabilistic forecasts could be generated strictly from the ensemble mean. Hence, we suggest a positive spread–skill relationship and the usefulness of additional members should first be demonstrated before increasing the ensemble size beyond that which produces the lowest ensemble mean error. We plan to explore this issue quantitatively in future research.

Acknowledgments. This work was supported under NSF Grant ATM-9508645 and represents a portion of the first author's (TMH's) Ph.D. dissertation research. The authors thank Eric Rogers, Henry Juang, and Zoltan Toth at NCEP for the production of the ensemble and Danny Mitchell at the National Severe Storms Lab for its archival. Paul Dallavalle and Jack Settelmaier at TDL provided the MOS data, and Bob Vislocky at Penn State is thanked for his constructive criticism and for supplying the verification data. Roberto Buizza and two anonymous reviewers are thanked for constructive criticism, which substantially improved the quality of this manuscript.

APPENDIX

Sample Calculation of ''Corrected'' Forecast Probability Distribution Using an Ensemble and Rank Histogram

Assume a sorted vector of ensemble forecasts $\mathbf{X}$ at a given point and time, a corresponding rank histogram $\mathbf{R}$, and a verification V.

Probabilities are to be set for the MOS categories $0 \leq V < 0.01$ in., $0.01 \leq V < 0.10$, $0.10 \leq V < 0.25$, $0.25 \leq V < 0.5$, $0.5 \leq V < 1.0$, $1.0 \leq V < 2.0$, and $2.0 \leq V$.

Assume the precipitation forecast (in inches) is as follows:

$$\mathbf{X} = [X_1, \ldots, X_{15}]$$
$$= [0, 0, 0, 0, 0, 0, 0.02, 0.04, 0.05, 0.07,$$
$$0.10, 0.11, 0.23, 0.26, 0.35].$$

First, calculate an ensemble mean and an ensemble variability, the standard deviation of the ensemble about its mean. Here, the ensemble mean is 0.082 in. and the ensemble variability is 0.111 in. According to the criteria from section 2, this indicates ''moderate'' ensemble variability, and hence the rank histogram illustrated in Fig. 2e is used. Assume thus that

$$\mathbf{R} = [R_1, \ldots, R_{16}]$$
$$= [0.25, 0.13, 0.09, 0.07, 0.05, 0.05, 0.04, 0.04,$$
$$0.03, 0.03, 0.03, 0.02, 0.02, 0.03, 0.05, 0.07].$$

Work upward through the precipitation categories, starting with the first category, $p(0.0 \leq V < 0.01)$. There are six precipitation forecasts of zero, and one forecast of 0.02, above the first threshold of 0.01 in. Hence, using (2) and (4), ranks 1–6 and a fraction of rank 7 are summed. Hence, $p(0.0 \leq V < 0.01) = 0.25 + 0.13 + 0.09 + 0.07 + 0.05 + 0.05 + 0.04[(0.01 - 0.00)/(0.02 - 0.00)] = 0.66$.

The probability for the next category, $p(0.01 \leq V < 0.10)$, is now calculated. The ensemble members of interest in calculating this probability are X_7 to X_{11}. Hence, the remaining part of rank 7 is summed with ranks 8–11: using (5) and (2), $p(0.01 \leq V < 0.10) = 0.04[(0.02 - 0.01)/(0.02 - 0.00)] + 0.04 + 0.03 + 0.03 + 0.03 = 0.15$.

Similarly, $p(0.10 \leq V < 0.25)$ is calculated using (2) and (4). Ranks 12 and 13 are added to a fraction of rank 14: $p(0.10 \leq V < 0.25) = 0.02 + 0.02 + 0.03[(0.25 - 0.23)/(0.26 - 0.23)] = 0.06$.

The remaining precipitation forecasts are X_{14} and X_{15}, 0.26 and 0.35 in., respectively. The largest remaining issue is how to allocate the last rank's probability among the existing categories. Using the method of moments (Wilks 1995), the estimated Gumbel parameters $\hat{\xi}$ and $\hat{\beta}$ are 0.030 and 0.0898, respectively. The cumulative distribution functions for the Gumbel distribution $F(0.35 \text{ in.})$ and $F(0.50 \text{ in.})$ are 0.9721, and 0.9946, respectively. Hence, using (2) and (6), $p(0.35 \leq V < 0.50) = 0.05 + 0.07(0.9946 - 0.9721)/(1.0 - 0.9721)$, which is approximately 0.11.

Finally, (7) is used to calculate $p(0.50 \leq V < 1.00)$; $F(1.0) = 0.99998$. Hence, $p(0.50 \leq V < 1.00) = 0.07(0.99998 - 0.9946)/(1.0 - 0.9721)$, which is approximately 0.01.

The probability above 1.0 in. and 2.0 in. is negligible.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1529.

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

Brankovic, C., T. N. Palmer, F. Molteni, and S. Tibaldi, 1990: Extended-range predictions with EMCWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

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

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

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

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.

Daan, H., 1985: Sensitivity of the verification scores to classification of the predictand. *Mon. Wea. Rev.,* **113,** 1384–1392.

Dallavalle, J. P., J. S. Jensenius Jr., and S. A. Gilbert, 1992: NGM-based MOS guidance—The FOUS 14/FWC message. Technical Procedures Bulletin 408, NOAA/National Weather Service, Washington, DC, 9 pp. [Available from NOAA/NWS, Services Development Branch, 1325 East-West Highway, Room 13466, Silver Spring, MD 20910.]

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

Emanuel, K. A., E. N. Lorenz, and R. E. Morss, 1996: Adaptive observations. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 67–69.

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

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

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models.* Chapman and Hall, 335 pp.

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

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

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–140.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran.* 2d ed. Cambridge University Press, 963 pp.

Rogers, E., T. L. Black, D. G. Deaven, and G. J. DiMego, 1996: Changes to operational ''early'' Eta analysis forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting,* **11,** 391–413.

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

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

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