## NOTES AND CORRESPONDENCE

# Interpretation of Rank Histograms for Verifying Ensemble Forecasts

THOMAS M. HAMILL

*National Center for Atmospheric Research,* Boulder, Colorado*

23 May 2000 and 9 August 2000

ABSTRACT

Rank histograms are a tool for evaluating ensemble forecasts. They are useful for determining the reliability of ensemble forecasts and for diagnosing errors in its mean and spread. Rank histograms are generated by repeatedly tallying the rank of the verification (usually an observation) relative to values from an ensemble sorted from lowest to highest. However, an uncritical use of the rank histogram can lead to misinterpretations of the qualities of that ensemble. For example, a flat rank histogram, usually taken as a sign of reliability, can still be generated from unreliable ensembles. Similarly, a U-shaped rank histogram, commonly understood as indicating a lack of variability in the ensemble, can also be a sign of conditional bias. It is also shown that flat rank histograms can be generated for some model variables if the variance of the ensemble is correctly specified, yet if covariances between model grid points are improperly specified, rank histograms for combinations of model variables may not be flat. Further, if imperfect observations are used for verification, the observational errors should be accounted for, otherwise the shape of the rank histogram may mislead the user about the characteristics of the ensemble. If a statistical hypothesis test is to be performed to determine whether the differences from uniformity of rank are statistically significant, then samples used to populate the rank histogram must be located far enough away from each other in time and space to be considered independent.

## 1. Introduction

The chaotic nature of the atmosphere (Lorenz 1963, 1969, 1982) ensures that errors will grow in any deterministic numerical weather forecast, eventually rendering that forecast no better than climatology. It is more appropriate, then, to envision the goal of numerical weather prediction as providing information on the relative likelihood of possible weather scenarios. A practical way of doing this is through ensemble forecasting (EF), whereby a set of numerical forecasts are generated from different initial conditions (Toth and Kalnay 1993, 1997; Molteni et al. 1996; Houtekamer et al. 1996; Houtekamer and Lefaivre 1997), different model physics or physical perturbations (e.g., Stensrud et al. 2000; Buizza et al. 1999), different models (Evans et al. 2000; Ziehmann 2000; Richardson 2000), and/or using differing fixed fields and constants (Houtekamer et al. 1996; Houtekamer and Lefaivre 1997). The ensemble is then typically used to generate a probabilistic forecast; for example, if 20 of 50 ensemble members forecast rain at a grid point, and if the ensemble is reliable (Wilks 1995), then the probability of rain may be estimated to be 40%.

How to produce probabilistic forecasts, how to use them, and how to evaluate them are questions still actively debated. Our focus here is strictly their evaluation. The problem is that conventional diagnostics for evaluating deterministic forecasts, measures such as ‘‘root-mean-square error,’’ are not useful with probabilistic forecasts. At a recent workshop on ensemble forecasting, a suite of useful verification techniques for ensemble forecasts was discussed and a subset suggested for common use (Hamill et al. 2000a). These techniques included probabilistic scoring measures such as the Brier score (Brier 1950; Murphy 1973; Wilks 1995), the ranked probability score (Epstein 1969; Murphy 1971), and their associated skill scores (Wilks 1995); reliability diagrams (Wilks 1995) plotted together with a distribution of the frequency of forecasts issued and a decomposition of the associated Brier score into reliability, resolution, and uncertainty terms (Murphy 1973); the relative operating characteristic, or ROC (Swets 1973; Mason 1982; Stanski et al. 1989); and the rank histogram, also known as the ‘‘Talagrand diagram.’’

The focus of this note is on one of these verification tools, the rank histogram. The rank histogram was developed contemporaneously and independently by An-

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

*Corresponding author address:* Dr. Thomas M. Hamill, NOAA–CIRES Climate Diagnostics Center, R/CDC1, 325 Broadway, Boulder, CO 80309-0216.
E-mail: hamill@cdc.noaa.gov

derson (1996), Hamill and Colucci (1996, 1997), and Talagrand (Harrison et al. 1995; Talagrand et al. 1997), though its inspiration goes back to long-established statistical ideas such as the *Q*–*Q* plot (e.g., Wilks 1995) and the probability integral transform (e.g., Casella and Berger 1990). The principle behind the rank histogram is quite simple. Ideally, one property that is desired from an EF is reliable probabilities; if ensemble relative frequency suggests *P* percent probability of occurrence, the event truly ought to have *P* probability of occurring. For this probability to be reliable, the set of ensemble member forecast values at a given point and the true state (the verification) ought to be able to be considered random samples from the same probability distribution. This reliability then implies in turn that if an *n*-member ensemble and the verification are pooled into a vector and sorted from lowest to highest, then the verification is equally likely to occur in each of the $n + 1$ possible ranks. If the rank of the verification is tallied and the process repeated over many independent sample points, a uniform histogram over the possible ranks should result.

The rank histogram permits a quick examination of some qualities of the ensemble. Consistent biases in the ensemble forecast will show up as a sloped rank histogram; a lack of variability in the ensemble will show up as a U-shaped, or concave, population of the ranks. Further, the rank histogram may be useful for more than just evaluating the forecast quality. Hamill and Colucci (1997, 1998) and Eckel and Walters (1999) also show how rank histograms provide information that may be used to recalibrate ensemble forecasts with systematic errors, thus achieving improved probabilistic forecasts.

While it is common for operational centers to produce probabilistic forecasts from their ensembles *as if* the ensembles were random samples from the same distribution as the truth, in fact many operational centers construct their ensembles under different assumptions. For example, the singular vector method used at the European Centre for Medium-Range Weather Forecasts (Molteni et al. 1996) generates initial perturbations that project strongly on the forecast modes where errors grow most quickly. This constitutes a sort of nonrandom sample, where the extremes of the forecast probability density function may be sampled more frequently than the center of the distribution. The interpretation of rank histograms under such different sampling strategies is not clear.

Since the rank histogram is a relatively new tool and collective experience with it is limited, some initial guidance is provided on its suggested use. We also explain some ways in which its uncritical use can lead to an inaccurate understanding of the characteristics of EFs. To this end, section 2 provides a general overview of the rank histogram and its link to other probabilistic verification tools. Section 3 describes some of the common problems in the interpretation of rank histograms. Section 4 describes the manner in which samples should be generated if one is to perform a formal hypothesis test of the uniformity of a rank histogram. Section 5 concludes.

## 2. Overview of the rank histogram

Suppose we are examining an ensemble of forecast values at a particular point. Assume we have a sorted *n*-member ensemble $\mathbf{X} = (x_1, \ldots, x_n)$ and the true state *V*. Because we have an imperfect knowledge of the true state, we describe it with a probability distribution. This distribution is calibrated, or ‘‘reliable,’’ if probabilities indicate the true likelihood of event occurrence. With a finite-sized ensemble, this will occur if the truth and the ensemble can be considered samples from the same probability distribution. If this is the case, then

$$E[P(V < x_i)] = \frac{i}{n + 1}. \qquad (1)$$

Here, $E(\cdot)$ denotes the expected value and *P* the probability. If we define fictional bounding ensemble members $x_0$ and $x_{n+1}$ such that $P(V < x_0) = 0$ and $P(V < x_{n+1}) = 1$, then (1) is equivalent to

$$E[P(x_{i-1} \le V < x_i)] = \frac{1}{n + 1}. \qquad (2)$$

Note that the expected value of the probability is the same for each of the $n + 1$ possible ranks relative to the sorted ensemble.

A rank histogram is found by repeatedly tallying the rank of the truth relative to an actual distribution of sorted ensemble forecasts (which may or may not be calibrated). Let $\mathbf{R} = (r_1, \ldots, r_{n+1})$ represent a rank histogram with $n + 1$ possible ranks. The population of a rank histogram element is determined from

$$\sum_{j=1}^{i} r_j = \overline{P(V < x_i)}, \qquad (3)$$

where $\overline{(\cdot)}$ denotes the average over a large sample of statistically independent points. Equation (3) is equivalent to

$$r_j = \overline{P(x_{j-1} \le V < x_j)}. \qquad (4)$$

In other words, the population of rank *j* is the fraction of times when the truth, when pooled with the sorted ensemble, is between sorted ensemble members $j - 1$ and *j*. Special rules are used for assigning ranks when many ensemble members have the exact same value as the verification, as may occur, for example, with no precipitation forecast and none observed; see Hamill and Colucci (1997, 1998). Note that the concept of the rank histogram is quite similar to that of the multicategory reliability diagram (MRCD; Hamill 1997). In fact, a diagram analogous to the MCRD can be generated from a rank histogram by plotting $\Sigma_{j=1}^{n} r_j$ (ordinate) versus $j/(n + 1)$ (abscissa).

Qualities of the calibration of the ensemble can be

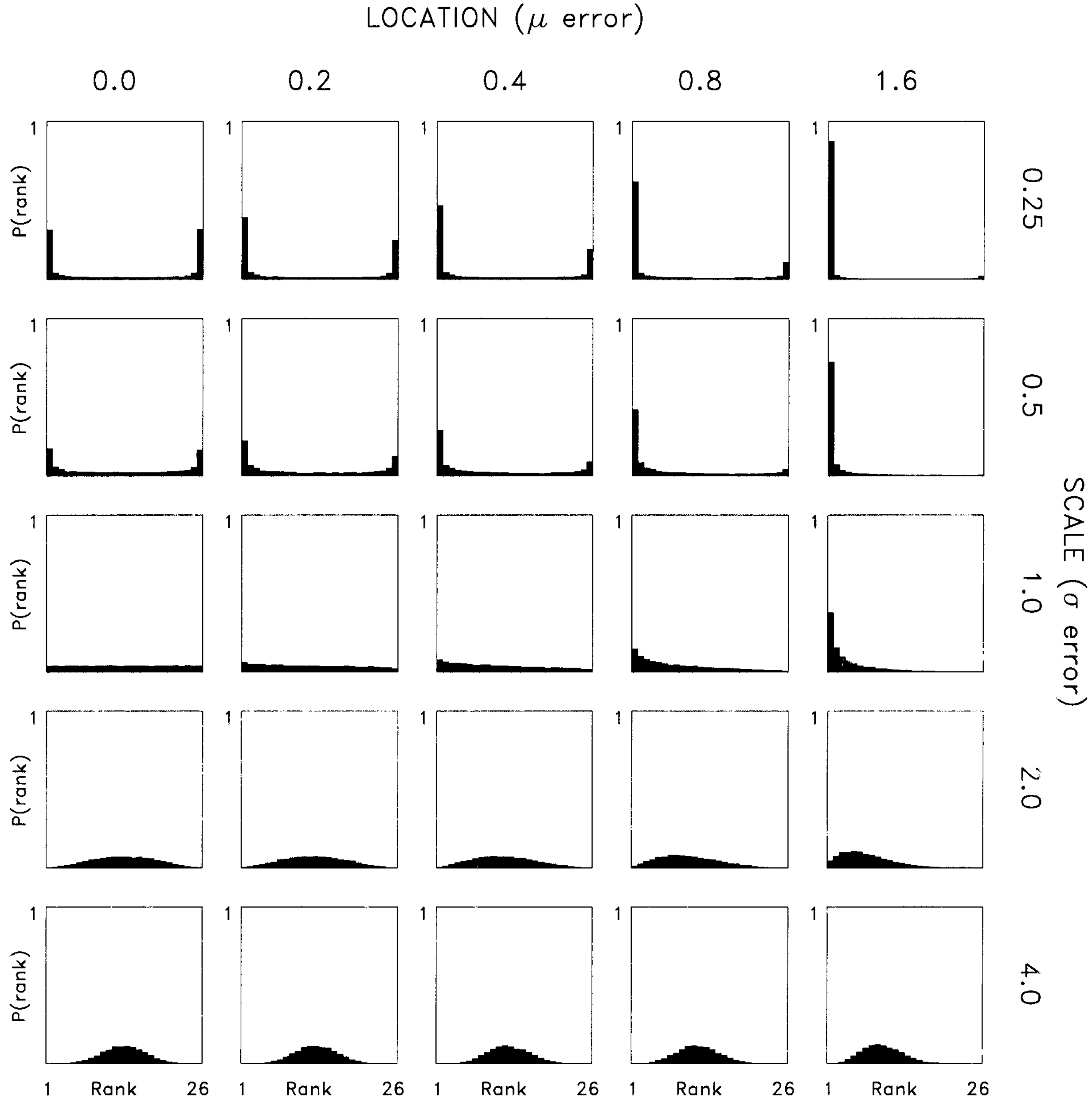

FIG. 1. Rank histograms where verification is sampled from a $N(0, 1)$ distribution and the ensemble ($n = 25$ members) is sampled from a $N(\mu, \sigma)$ distribution. The rank of the verification is tallied 10 000 times in each panel.

diagnosed from the shape of rank histograms. Suppose the ensemble forecast probability distributions errs in the location (wrongly forecasting the mean) of the distribution and/or in the scale (wrongly forecasting the standard deviation). Assume that the true state may be considered to be randomly drawn from a standard normal distribution $N(0, 1)$, that is, a distribution with mean 0.0 and standard deviation 1.0. Assume each member of a 25-member ensemble is randomly drawn from $N(\mu, \sigma)$. Figure 1 shows the shape of the resulting rank histogram for combinations of $\mu$ and $\sigma$ (only $+\mu$ errors are shown for brevity). When the ensemble samples are from a distribution with a lack of variability, a U-shaped rank histogram results. An excess of variability in the ensemble overpopulates the middle ranks. Bias (+/−) exessively populates the (left/right) extreme ranks.

The rank histogram, when correctly used and interpreted, measures the reliability of the ensemble (Talagrand et al. 1997; Hersbach 2000). There is another desirable property, namely, sharpness. In a calibrated forecast, sharpness is related to resolution, or the ability of the forecast to be sorted into subsamples where the verifying event is different (Wilks 1995). Given two well-calibrated forecast systems, any rational user would prefer the system that produces the more specific forecast. The rank histogram does not evaluate resolution, so it must be used in conjunction with other forecast tools such as the ROC, Brier scores, or ranked probability scores to generate a more complete picture of the quality of a probabilistic forecast. Hereafter, we will focus on how the rank histogram can be used (and misused) for evaluating reliability. Presumably, the user will also verify the ensemble with other techniques that measure resolution.

## 3. Problems interpreting rank histograms

### *a. Misdiagnosing ensemble characteristics from histogram shape*

We start by noting that a uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable. A rank histogram is populated with a set of sample points; if the ensemble at each sample point is reliable, then the resulting rank histogram should be uniform. However, a uniform rank histogram provides no guarantee that the ensemble is reliable at each point used to populate it. Figures 2a,b illustrates a scenario of how histogram flatness may be illusory. Assume that there is an ensemble that is forecasting the probability distribution incorrectly in one of three possible ways: the forecast distribution may have a negative bias, a positive bias, or excessive variability, depending on the sample point. This is simulated by assuming the verification is sampled from a $N(0, 1)$ distribution and the ensemble is sampled with equal likelihood from either a $N(-0.5, 1)$, a $N(0.5, 1)$, or a $N(0, 1.3)$ distribution. A relatively uniform rank histogram is achieved, though the ensemble was never sampled from the same probability distribution as the verification.

Suppose the sampling strategy for generating ensemble members is other than random. For example, singular vectors (Molteni et al. 1996) are designed to sample the most rapidly growing structures among the myriad of possible directions in the analysis error probability distribution. As a rough analogy to ensemble forecasts from singular vectors, suppose the tails of the distribution are sampled more frequently than its center. In such a case, the ensemble distribution can be quite different from the distribution from which the truth is sampled, yet a uniform rank histogram may result. Figures 3a–c illustrate such a nonrandom sampling process and the rank histogram that results. First, suppose the truth is a random sample from a $N(0, 1)$ distribution. However, the ensemble will be generated as a nonrandom sample from a $N(0, 0.7)$ distribution, where the tails of the distribution are more likely to be sampled from than the center. Such a nonrandom sample is simulated through the following process. First, generate a random number $X$ sampled from a $N(0, 0.7)$ distribution. Next, generate a random number $U$ sampled from a standard uniform distribution, so that its value is equally likely to take any value between 0.0 and 1.0. Using the function $I$ in Fig. 3b, accept the sample $X$ as an ensemble member only if $U < I(X)$; otherwise, start the process again [here, $I$ was generated from the normalized ratio of a $N(0, 1)$ to a $N(0, 0.7)$ distribution]. Consequently, when the ensemble is generated to simulate such a nonrandom sample, the resulting rank histogram can still be approximately uniform (Fig. 3c).

A related problem is that a rank histogram of any given shape may be generated in a variety of ways. Figure 4 shows that a U-shaped rank histogram, typi-

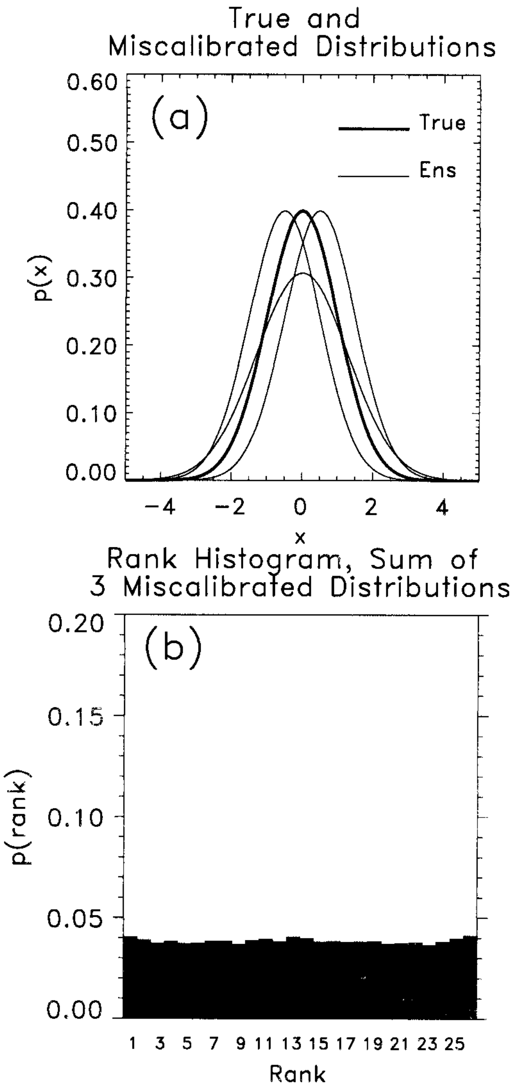

FIG. 2. Illustration of shape of rank histograms when ensemble members are selected from one of three different probability distributions other than the truth. (a) Probability distributions from which truth and ensemble are selected; samples are generated from each of three distributions with either a bias or excess variability; $N(-0.5, 1)$, $N(0.5, 1)$, and $N(0, 1.3)$. (b) Rank histogram corresponding to (a). The rank of the verification is tallied 30 000 times (10 000 with respect to each of the three forecast distributions).

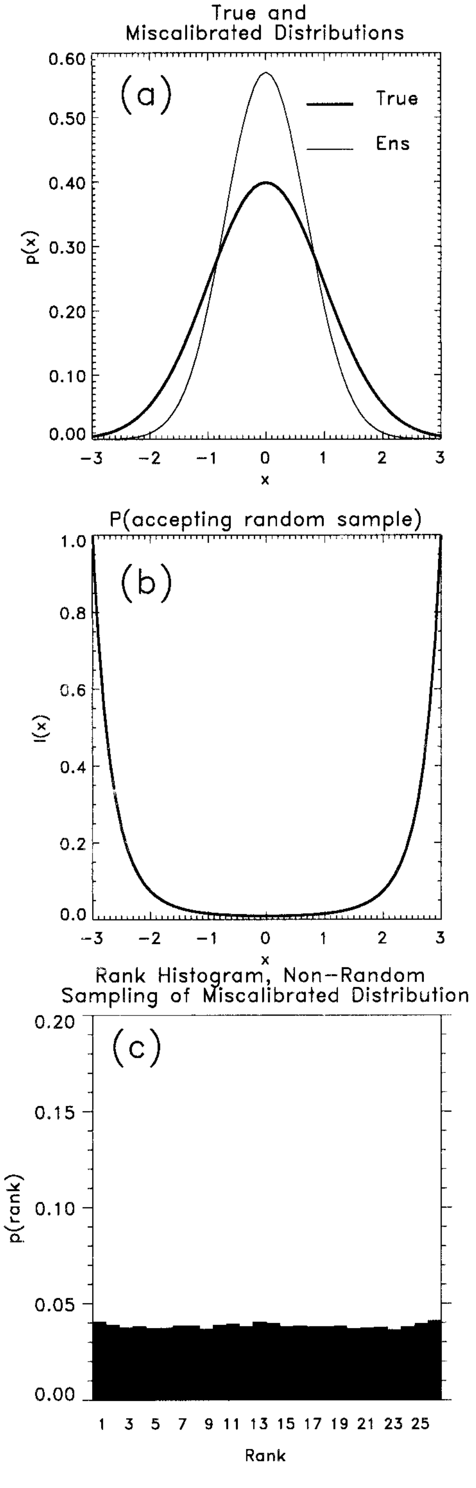

FIG. 3. (a) Probability distribution from which truth is sampled, $N(0, 1)$, and ensemble $N(0, 0.7)$. (b) Probability of accepting ensemble random sample; the random sample $X$ is accepted if a uniform random number $U$ is generated such that $U < I(X)$. (c) Rank histogram created under nonrandom sampling of miscalibrated distribution in (a) using nonrandom sampling function in (b). Verification rank tallied 10 000 times.

cally thought of as indicating undervariability in the ensemble, could also indicate that the ensemble is sampling a population with some combination of conditional biases. Sampling half of the time from an ensemble with a negative bias and half of the time from an ensemble with a positive bias can generate a U-shaped rank histogram indistinguishable from one created as a result of undervariability. In general, if the model developer has reason to believe that the ensemble may perform differently under different weather regimes (i.e., different synoptic situations, regions, seasons, etc.), then it may be worth generating rank histograms separately for each regime. If the shape differs from one regime to the next, this probably indicates conditional biases in the ensemble. Hamill and Colucci (1997, 1998) show, for example, that rank histograms from an Eta/Regional Spectral Model ensemble are very differently shaped when precipitation spread is small (generally, corresponding to low precipitation amounts) compared to when spread is large (higher amounts). Spread here refers to the standard deviation of the ensemble about its mean.

These scenarios illustrate that reliability alone is not a good metric of forecast quality, and reliability apparently can be achieved even if samples from ensemble forecasts and the verification are not drawn from the same distribution. When, then, is reliability as diagnosed from rank histograms indicative of proper random sampling and when is it not? The results of Gilmour and Smith (1997) and Smith (1999) suggest that reliability may be illusory unless it is possible to find a model state that ‘‘shadows,’’ or follows, the evolution of the real atmosphere within an error tolerance consistent with magnitude of analysis uncertainty. If a shadowing trajectory can be found, it can be attributed to be sampled from the same distribution as the truth. Conversely, if no model state can be found with this property, then the ensemble is sampling some other probability distribution than the one the truth is drawn from, and hence any noted reliability from a rank histogram may be considered illusory. Because finding a shadowing model forecast trajectory is difficult for large dynamical systems like current weather prediction models, this idea is just beginning to be explored with operational forecasts.

### *b. Sampling properly in multiple dimensions*

To this point it has been assumed that rank histograms were to be generated by sampling independent points, and it was noted that if the ensemble were reliable at each sample point, the rank histogram should be uniform. Such an analysis, however, neglects the possibility that ensemble forecast fields (many variables at many grid points) should also be reliable in a more highly dimensional subspace as well. To illustrate this, let us assume that an ensemble at two adjacent grid points correctly specifies the variance at each grid point but incorrectly specifies the covariance between the two. This might happen, for example, if one were to naively

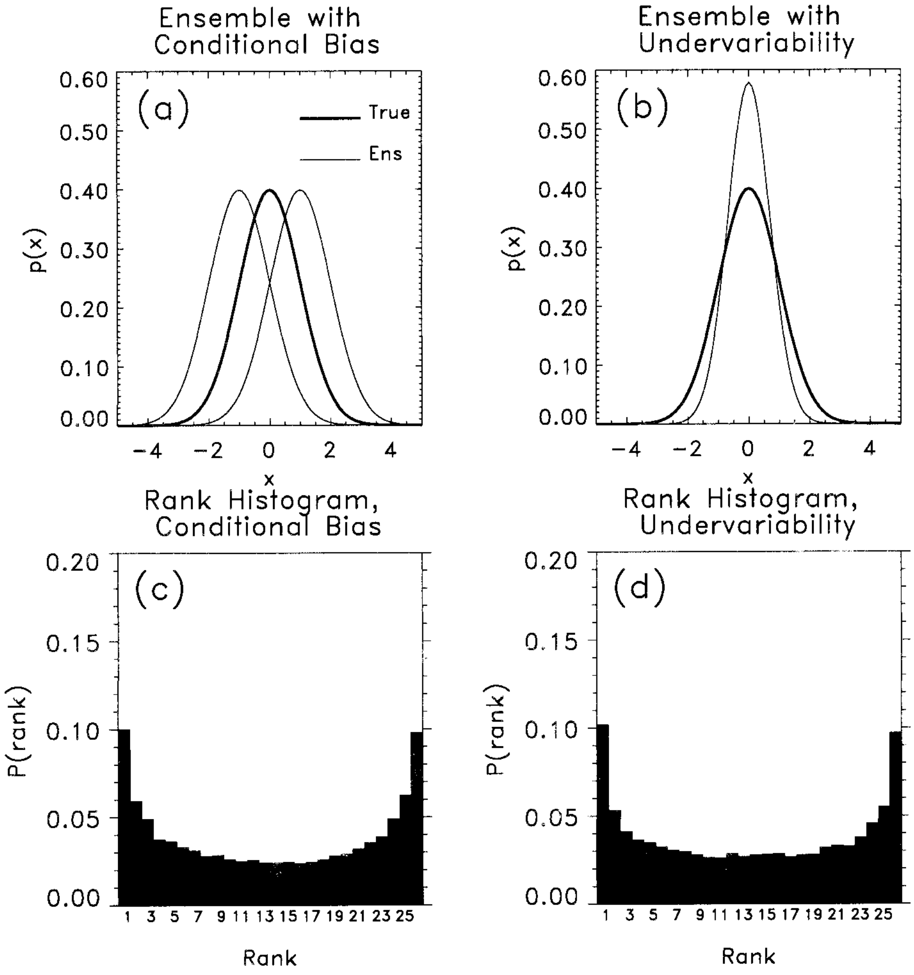

FIG. 4. (a) As in Fig. 2a, but where ensemble is selected with equally likely probability from one of the two biased distributions, a $N(-1, 1)$ or $N(1, 1)$ distribution, with the verification tallied 10 000 times for each distribution. (b) As in (a), but where ensemble forecasts are selected from a probability distribution with a lack of variability, $N(0, 0.69)$. (c) Rank histogram corresponding to (a). (d) Rank histogram corresponding to (b). Verification rank tallied 20 000 times.

generate perturbations by adding white noise at every grid point consistent with the analysis uncertainty, but neglecting the correlations of errors between grid points. If variances are correctly specified but covariances are not, a rank histogram formed from some combination of the values of two grid points is not necessarily uniform (Fig. 5). Practically, thus, it is wise to check the rank histogram's uniformity for not only fields such as geopotential height, but also variables that are related to its spatial derivatives, such as winds and vorticity. A practical example of how rank histograms can appear to be quite different for geopotential, winds, and vorticity is shown in Hamill et al. (2000b; Fig. 8).

Ideas for extending the rank histogram to multiple dimensions are just beginning to be explored. See Smith (1999) for details on another possible way of examining the reliability in a phase space of very many directions, through use of a diagnostic called the ''minimal spanning tree.''

### *c. Errors in observations*

To this point it has been assumed that verification samples are error free; consequently, if the ensemble forecast and verification are sampled from the same distribution, the rank histogram should appear flat. In practice, imperfect observations will be used for verification. Let us assume that observations are unbiased but are contaminated by noise, be it from instrument error, representativeness error, or both. In this case, the effects of including observational errors on the shape of rank histograms should be considered (see also Anderson 1996). The problem caused by observational errors is illustrated in Fig. 6. Here, it is assumed that a 25-member ensemble is sampled from a $N(0, 1)$ distribution, and the verification is created from a random sample from a $N(0, \sigma)$ distribution added to a random sample from a $N(0, 1)$ distribution, where $\sigma$ represents the standard deviation of the observational errors. This models the

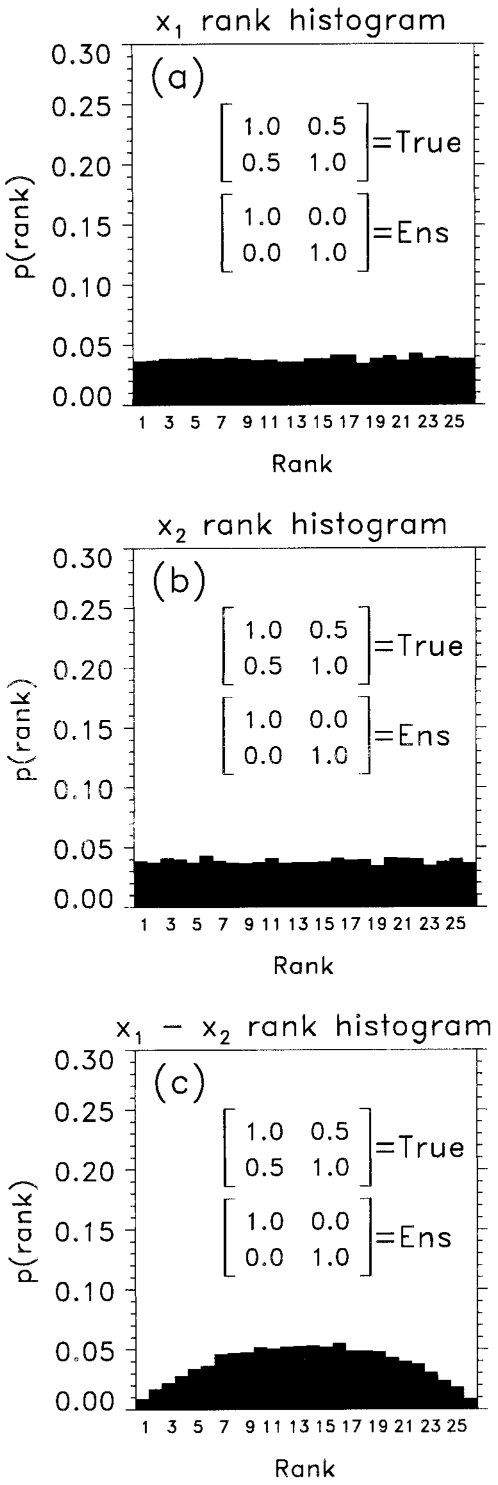

FIG. 5. Illustration of how rank histograms for two grid points with correlated errors are affected if the correlation is misspecified in the ensemble. Assume 25-member ensemble is sampled at two grid points from population with variance of 1.0 but 0.0 covariance between the grid points. True distribution has 0.5 covariance between grid points. Rank of verification is tallied 10 000 times. (a) Rank histogram of variable at first grid point, $x_1$. (b) Rank histogram of variable and second grid point, $x_2$. (c) Rank histogram of $x_1 - x_2$.

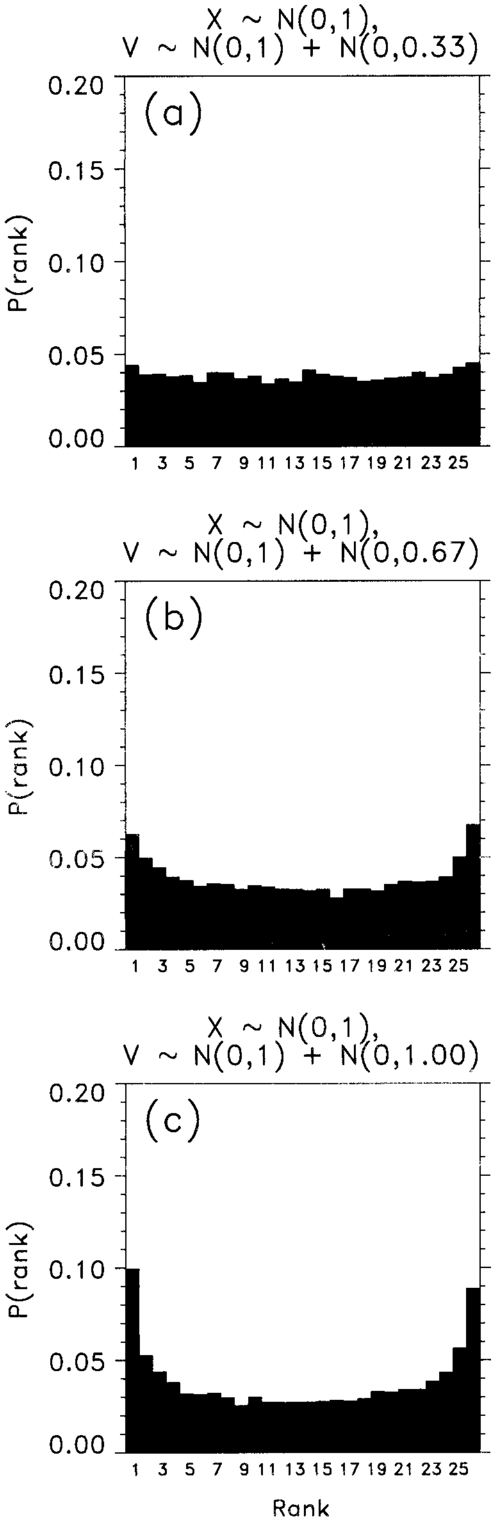

FIG. 6. Rank histograms in presence of observational error. Ensemble sampled from $N(0, 1)$ distribution, observation from $N(0, 1) + N(0, \sigma)$ distribution. Verification rank tallied 10 000 times. (a) $\sigma = 0.33$, (b) $\sigma = 0.67$, (c) $\sigma = 1.00$.

situation where the ensemble and the true state are sampled from the same distribution, but where imperfect observations are used for verification rather than the truth. For $\sigma \ll 1$, there is little deviation from uniformity of rank, but as $\sigma$ increases, the extreme ranks become more highly populated. Obviously, the situation is yet more complicated if the observations are biased as well.

Since overpopulation of the extreme ranks of a histogram are commonly observed in operational ensemble forecasts, it may be worth determining how much of the population of these ranks is due to observational errors. The results of Fig. 6 suggest that if the observational errors are a significant fraction of the spread in the ensemble, then rank histograms should *not* be generated by ranking the observation relative to the sorted ensemble. Rather, the rank histogram should be generated by *ranking the observation relative to an sorted ensemble with random observational noise added to each member* (see also Anderson 1996). In this manner, if the ensemble is reliable, the adjusted ensemble and the observations are presumed to both be sampled from the same probability distribution. In principle, adding these random errors for at least some common observational types such as raobs should not be very difficult, since these error statistics have been estimated for data assimilation purposes.

## 4. Hypothesis testing for uniformity of rank

Rank histograms will naturally appear somewhat irregular if populated with a relatively small sample; they look progressively smoother with more and more samples. If samples used to populate the rank histogram are independent, then a $\chi^2$ hypothesis test (Wilks 1995; Anderson 1996; Hamill and Colucci 1997) may be performed to determine whether or not the distribution for a given sample size is significantly different from uniform. However, the statistical hypothesis test may produce misleading results if samples have correlated errors, as may happen if two adjacent grid points are both used as samples. Ideally, the model user thus ought to have information on the spatial and temporal correlation of errors for the model being used and the variable being examined, and the sample points ought to be spaced far enough apart from each other so as to be reasonably independent.

The potential problem of correlated samples is demonstrated using the quasigeostrophic (QG) channel model used for perfect-model ensemble simulations in Hamill et al. (2000b), Hamill and Snyder (2000), and Morss et al. (2000, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*). It is a midlatitude, beta-plane, gridpoint channel model that is periodic in $x$ (east–west), has impermeable walls on the north–south boundaries, and has rigid lids at the top and bottom. There is no terrain, nor are there surface variations such as land and water. Pseudo–potential vorticity (PV) is conserved except for Ekman pumping at the surface, $\nabla^4$ horizontal diffusion, and forcing by relaxation to a zonal mean state. The domain is 16 000 × 8000 × 9 km; there are 129 grid points east–west and 65 north–south, and eight model forecast levels, with additional staggered top and bottom levels at which potential temperature $\theta$ is specified. The grid spacing is 125 km.

It can be shown that there is a rather simple distribution for the expected difference in the verification ranks between two samples with uncorrelated errors. Define $D$ as the difference in ranks between the two samples in an $n$-member ensemble. The expected value for the probability $P(D)$ of the difference is

$$P(D) = \begin{cases} \dfrac{n + 1 - |D|}{(n + 1)^2}, & \text{if } |D| < (n + 1) \\ 0, & \text{otherwise.} \end{cases} \qquad (5)$$

This equation can be verified in the following manner. Generate an $(n + 1) \times (n + 1)$ array, and populate each array element with the row number minus the column number. The row number represents possible ranks of the first sample, the column number the ranks of the second, and the value assigned to a particular element is the difference in ranks between the two samples. Count the fraction of array elements with a particular $D$ and (5) will result by inspection.

We now explore whether different sampling strategies bring us closer to the desired distribution achieved by independent samples. To do so, we use the QG model and the perturbed observation ensemble analysis data described in Hamill et al. (2000b), shown in that paper to have approximately uniform rank histograms (in a perfect-model context). Using this data, the rank of the truth is determined relative to the sorted ensemble for model level 4 (~500 hPa) geopotential, $u$-wind component, and pseudo–PV at every model grid point and every analysis time. We then examined the distribution of the difference in ranks for various spatial and temporal lags using points in the center ½ of the channel. Figures 7a–c show the distribution of differences in ranks for various spatial lags plotted over the top of the distribution that is expected from (5) assuming uncorrelated data. Geopotential heights are somewhat correlated even at a spatial lag of 20 grid points (samples 2500 km apart) whereas PV, with much more small-scale structure, is nearly uncorrelated by lag 6 (750 km). Similarly, an analysis was done of the difference in ranks for various temporal lags (Figs. 8a–c). For collocated samples with a 1-day lag between samples, there is a slight correlation for geopotential, wind, and PV; for 2- and 3-day lags, wind and PV samples are effectively uncorrelated. These results may well differ from those that would be obtained with a primitive equation model. Nonetheless, they do illustrate that if the aim of generating rank histograms is for formal hypothesis testing of uniformity, then samples must often be located far apart in space and in time, and that the difference

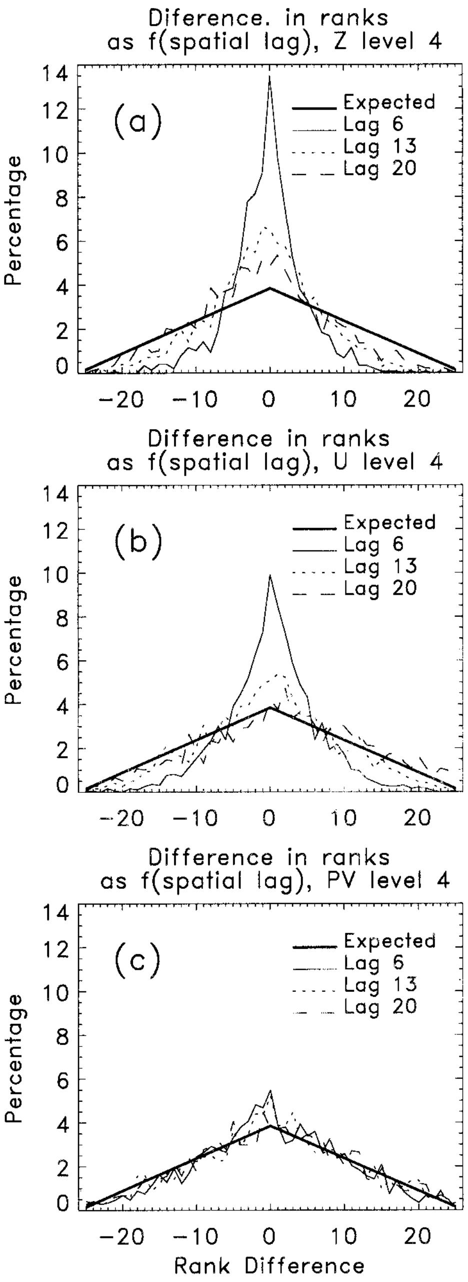

FIG. 7. Frequency of difference in ranks for samples with spatial lags of 6, 13, and 20 grid points in QG model, plotted over distribution expected if samples are uncorrelated. (a) Differences for model level 4 geopotential, (b) differences for *U* wind component, and (c) differences for PV.

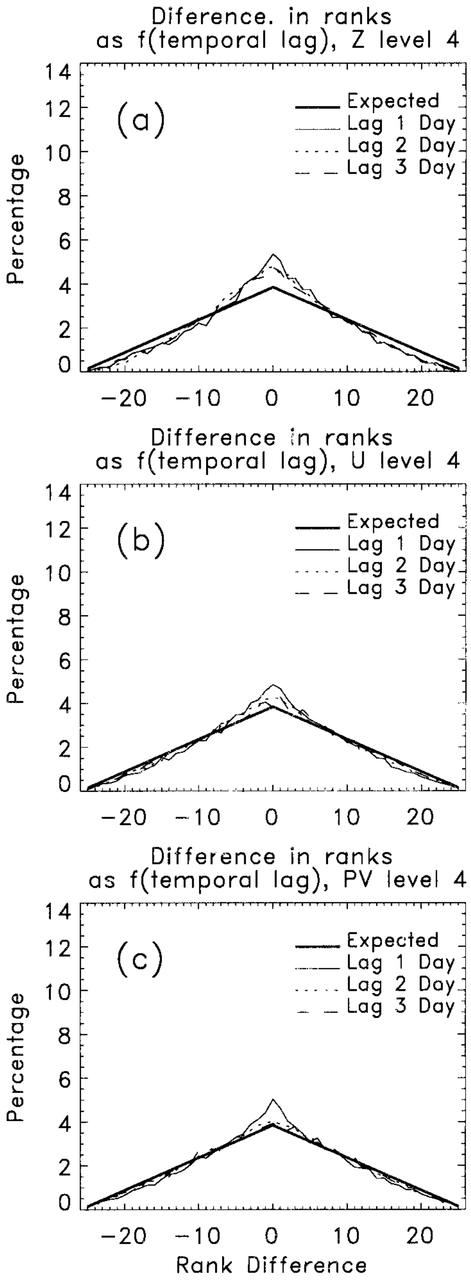

FIG. 8. As in Fig. 7, but for frequency of difference in ranks for samples with temporal lags of 1, 2, and 3 days.

in distance and time may depend on the variable in question and the model being used.

## 5. Conclusions

A rank histogram is a tool for evaluating the reliability of ensemble forecasts. Errors in the mean and spread of the ensemble can be diagnosed with the rank histogram.

Uncritical use of rank histograms, however, can lead to misinterpretations of the qualities of that ensemble. Some potential problems to be cognizant of include the following.

- A flat rank histogram does not necessarily indicate reliability of the ensemble. A flat rank histogram can still be generated from ensembles with different conditional biases, or by nonrandom sampling of a different probability distribution than that from which the truth is drawn.

- Flat rank histograms may also indicate that the ensemble is correctly specifying the variance at a grid point, but covariances may still be misspecified. This can be checked somewhat by generating rank histograms for differences between values at different grid points or by using other diagnostics such as the minimal spanning tree.
- A U-shaped rank histogram, commonly understood as indicating a lack of variability in the ensemble, can also be a sign of conditional biases. If possible, rank histograms for subpopulations should be generated to determine if the shape varies from one to the next; this can provide some perspective on whether the U shape indicates conditional biases or undervariability.
- Imperfect observations are commonly used for the verification value in generating the rank histogram. Observational errors, if not accounted for, may affect the shape of the rank histogram; the larger the error, the more U-shaped the rank histogram will appear, even if the ensemble is reliable. If observational error characteristics are known, this can be dealt with by adding random noise to each ensemble member, consistent with the observational error statistics.
- If a statistical hypothesis test is to be performed to test for uniformity, then samples used to populate the rank histogram must be located far enough away from each other in time and space to be considered independent.

*Acknowledgments.* Tony Eckel (U.S. Air Force and University of Washington), Isla Gilmour (NCAR/ASP), and three anonymous reviewers are thanked for their careful reviews of this manuscript.

## REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

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

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

Casella, G., and R. L. Berger, 1990: *Statistical Inference.* Duxbury Press, 650 pp.

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

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

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

Gilmour, I., and L. A. Smith, 1997: Enlightenment in shadows. *Applied Nonlinear Dynamics and Stochastic Systems near the Millenium,* J. B. Kadtke and A. Bulsara, Eds., AIP, 335–340.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting,* **12,** 736–741.

——, and S. J. Colucci, 1996: Random and systematic error in NMC's short-range Eta ensembles. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 51–56.

——, and ——, 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.

——, and C. Snyder, 2000: A hybrid ensemble Kalman filter/3-dimensional variational analysis scheme. *Mon. Wea. Rev.,* **128,** 2905–2919.

——, S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000a: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor Soc.,* **81,** 2653–2664.

——, C. Snyder, and R. E. Morss, 2000b: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

Harrison, M. S. J., D. S. Richardson, K. Robertson, and A. Woodcock, 1995: Medium-range ensembles using both the ECMWF T63 and unified models—An initial report. UKMO Tech. Rep. 153, 25 pp. [Available from U. K. Met Office Library, London Road, Bracknell, Berkshire RG12 2SZ, United Kingdom.]

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

Houtekamer, P. L., and L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.,* **125,** 2416–2426.

——, ——, and J. Derome, 1996: The RPN ensemble prediction system. *Proceedings, ECMWF Seminar on Predictability.* Vol. II. ECMWF, 121–146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

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

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus,* **34,** 505–513.

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

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.

Richardson, D. S., 2000: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.,* in press.

Smith, L. A., 1999: Disentangling uncertainty and error: On the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* Alistair E. Mees, Ed., Birkhauer Press, 31–64.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Research Rep. 89-5, Environment Canada, 114 pp. [Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4 Canada.]

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

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–999.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability,* ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

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.,* **12,** 3297–3319.

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

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