The Minimum Spanning Tree Histogram as a Verification Tool for Multidimensional Ensemble Forecasts

D. S. WILKS

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 23 April 2003, in final form 21 October 2003)

ABSTRACT

The minimum spanning tree (MST) histogram is a multivariate extension of the ideas behind the conventional scalar rank histogram. It tabulates the frequencies, over n forecast occasions, of the rank of the MST length for each ensemble, within the group of such lengths that is obtained by substituting an observation for each of its ensemble members in turn. In raw form it is unable to distinguish ensemble bias from ensemble underdispersion, or to discern the contributions of forecast variables with small variance. The use of scaled and debiased MST histograms to diagnose attributes of ensemble forecasts is illustrated, both for synthetic Gaussian ensembles and for a small sample of actual ensemble forecasts. Also presented are adjustments to χ^2 critical values for evaluating rank uniformity, for both MST histograms and scalar rank histograms, given serial correlation in the forecasts.

1. Introduction

Ensemble forecasting is now well established as a technique that is relevant at a variety of spatial scales and lead times (e.g., Du et al. 1997; Eckel and Walters 1998; Hamill and Colucci 1997; Houtekamer et al. 1996; Molteni et al. 1996; Stensrud et al. 1999; Toth and Kalnay 1997). The aim in ensemble forecasting is to approximate the probability distribution reflecting the uncertain components of the forecast system (prominently, initial-state uncertainty) using an ensemble (i.e., a finite collection) of specific plausible initial conditions. If the initial ensemble consists of a random sample from the underlying probability distribution of initial-condition uncertainty, and each ensemble member is integrated forward in time according to a perfect dynamical model, the resulting ensemble of forecasts should represent a random sample from the probability distribution of future-state uncertainty, and the actual state to which the real atmosphere evolves should be yet another random sample from this distribution.

In practice the initial ensemble is not a random sample from the relevant distribution (for a variety of reasons, not least of which is that this distribution is unknown), and the forecast models are not perfect. Therefore, one aspect of interest in the verification of ensemble forecasts is the degree to which the observed (or analyzed) future atmospheric states appear to be plausible members of their forecast ensembles.

Corresponding author address: Daniel S. Wilks, Dept. of Earth and Atmospheric Sciences, 1113 Bradfield Hall, Cornell University, Ithaca, NY 14853.
E-mail: dsw5@cornell.edu

For one-dimensional (i.e., scalar, or univariate) forecasts, a popular graphical device for addressing this question is the rank histogram (Anderson 1996; Hamill and Colucci 1997; Harrison et al. 1995). To tabulate a rank histogram, the rank of the observation within the $n_{ens} + 1$ member collection defined by the union of the n_{ens}-member ensemble and the observation is determined. Equivalently [provided none of the ensemble members is exactly equal to the analysis; otherwise see Hamill and Colucci (1997)], one is added to the number of ensemble members exceeded in magnitude by the corresponding observation. If the premise is true that the observation and the ensemble members have been drawn from the same distribution, any of the $n_{ens} + 1$ ranks is an equally likely position for the observation on any particular forecast occasion. Collectively, over some number n forecast occasions, a histogram of these $n_{ens} + 1$ ranks—the rank histogram—will be uniform, or flat, within the limits of a finite sample. Particular deviations from the ideal situation of the observation and ensemble members being drawn from the same distribution are reflected in deviations of the rank histogram from uniformity: positive or negative ensemble biases produce overpopulation of the lowest or highest ranks, respectively; underdispersed ensembles produce U-shaped rank histograms; and overdispersed ensembles result in underpopulation of the extreme ranks, or mound-shaped rank histograms (Hamill 2001).

When the forecast is multidimensional, pertaining for example to several meteorological elements simultaneously at one location, or to forecasts of the same (or multiple) forecast elements at a collection of locations, the scalar rank histogram does not apply. However, a

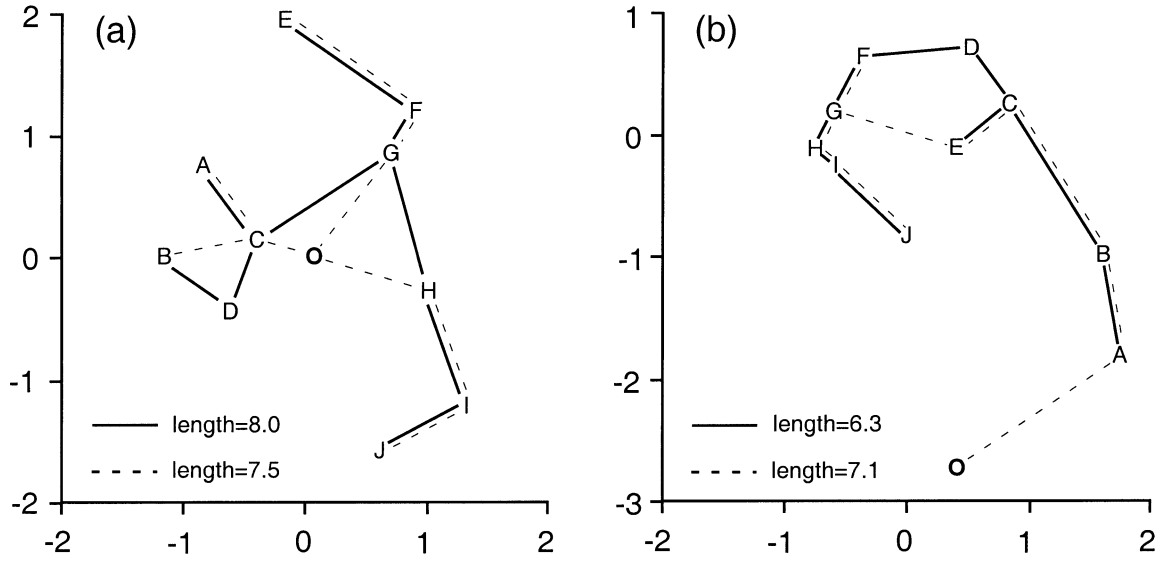

FIG. 1. Hypothetical example MSTs in $K = 2$ dimensions. The $n_{ens} = 10$ ensemble members are labeled A–J, and the corresponding observation is O. Solid lines indicate MSTs for the ensemble as forecast, and dashed lines indicate MSTs that result from the observation being substituted for ensemble member D. (a) A configuration that could result from an overdispersed ensemble, where the observation is interior to the point cloud of the ensemble. (b) A configuration that could result from an underdispersed ensemble and/or a substantial ensemble mean error.

conceptual extension of the basic approach from scalar to multidimensional (i.e., vector) forecasts has been suggested by Smith (2001), in terms of the lengths of minimum spanning trees (MSTs) (Ahuja et al. 1993). Consider a geometric space, each of whose K coordinate axes correspond to one of the forecast elements and/or locations. Let $x_{i,k}$ be the value of kth element of the ith ensemble member; $k = 1, \ldots, K$; $i = 1, \ldots, n_{ens}$. One can compute the (n_{ens}) $(n_{ens} - 1)/2$ pairwise Euclidean distances among the points in this space,

$$D_{i,j} = \left[\sum_{k=1}^{K} (x_{i,k} - x_{j,k})^2 \right]^{1/2}. \quad (1)$$

The MST for these n_{ens} points is the set of $n_{ens} - 1$ line segments connecting all of the points, such that the network contains no closed loops, and the sum of the lengths $D_{i,j}$ of these segments is minimized. The solid lines in Figs. 1a and 1b show MSTs in $K = 2$ dimensions for two ensembles with $n_{ens} = 10$, the members of which are labeled A–J.

In a manner similar to the ordinary rank histogram for scalar ensemble forecasts, the MST histogram tabulates the rank of the MST length computed for the n_{ens} ensemble members only, within the $n_{ens} + 1$ element distribution consisting of the union of the ensemble-only MST length, with the n_{ens} MST lengths obtained by substituting the observation for each one of the ensemble members in turn. That is, one is added to the number of the n_{ens} MSTs in which the observation has been substituted for one of the ensemble members, whose lengths are exceeded by that for the MST of the ensemble as actually forecast. [Note that this convention is the reverse of that in Smith (2001) but is consistent with usual practice for rank histograms (e.g., Hamill 2001)]. If the ensemble and the subsequent analysis it is meant to predict have been drawn from the same (K dimensional) probability distribution, then the lengths of the MSTs obtained by substituting the observation for any of the ensemble members should be statistically indistinguishable from the length of the MST computed from the ensemble members only. Over a large number n of forecast occasions, the histogram of these ranks—the MST histogram—should be essentially uniform, or flat.

While the scalar rank histogram and the MST histogram are similar in concept, it should be noted that the MST histogram is not a mathematical generalization of the conventional rank histogram. In particular, the MST histogram does not reduce to the scalar rank histogram in the special case of $K = 1$ dimension. Indeed, in one dimension the MST length is trivially the range (maximum minus minimum) of the data.

The purpose of this paper is to outline some important considerations that bear on the use of the MST histogram and to catalog some typical behaviors under various deviations from perfect ensembles, which result in different types of nonuniform MST histograms. Section 2 details these considerations and typical behaviors in the context of synthetic data. Section 3 applies these to a particular small sample of actual ensemble forecasts. Section 4 considers the question of statistical significance for rank uniformity as a function of ensemble size, sample size, and nonindependence of the ensembles and provides corresponding results for scalar rank histograms. Section 5 provides conclusions.

2. The MST histogram

a. Raw MST histograms

As noted earlier, the solid lines in Fig. 1 indicate MSTs for two ensembles whose members are labeled A–J. The point representing the corresponding observation is labeled O in Figs. 1a and 1b, and the dashed lines show the MSTs that result when the observation is substituted for ensemble member D in each case. In Fig. 1a this substitution results in a shorter MST, with the sum of the lengths of the solid and dashed lined segments being 8.0 and 7.5, respectively. The rank of the solid-line MST depends also on the lengths of the other nine MSTs, resulting from each of the other nine ensemble members being replaced by the observation in turn. In Fig. 1a, the lengths of eight of these MSTs are shorter than 8.0, and they are also shorter than the one obtained by replacing point G by the observation, which is very slightly longer than 8.0. Therefore, the rank of the length of the solid MST in Fig. 1a is 10 out of 11. In Fig. 1b, the length 6.3 of the solid MST is shorter than all 10 of the MSTs obtained by replacing an ensemble member A–J by the observation point O, so its rank is 1 out of 11.

Each of these two hypothetical MST ranks pertains to one ensemble forecast and its corresponding observation or analysis. An MST histogram consists of the histogram, collectively over some large number n of such forecasts, of the frequencies of occurrence of the $n_{\text{ens}} + 1$ possible ranks. If, over this sample of n forecasts, the observations have been drawn from the same probability distributions as the respective ensemble members, this histogram will be essentially flat, or uniform. Figure 2 shows one such example, together with 29 others, exhibiting various deviations from this same-distribution condition. These are results for synthetic ensembles and observations, generated in each case from $n = 1000$ independent $K = 10$-dimensional multivariate normal distributions, with an ensemble size of $n_{\text{ens}} = 10$. The horizontal dimension is the ratio of the standard deviations (in all 10 dimensions simultaneously) of the ''truth'' distribution (from which the ''observation'' has been generated) to those of the ensemble distribution. Thus ratios of $\sigma_{\text{truth}}/\sigma_{\text{ensemble}}$ greater and less than 1 indicate underdispersed and overdispersed ensembles, respectively. The vertical dimension in Fig. 2 reflects ensemble bias, or systematic ensemble mean error, expressed in terms of Mahalanobis distance (e.g., Mardia et al. 1979):

$$D = [(\boldsymbol{\mu}_{\text{truth}} - \boldsymbol{\mu}_{\text{ens}})^{\text{T}}(\boldsymbol{\Sigma}_{\text{ens}})^{-1}(\boldsymbol{\mu}_{\text{truth}} - \boldsymbol{\mu}_{\text{ens}})]^{1/2}. \quad (2)$$

This is a nondimensionalized distance measure that is essentially the multivariate counterpart of the ''z score'' (or ''standardized anomaly;'' e.g., Wilks 1995), in which differences between the vector mean $\boldsymbol{\mu}_{\text{truth}}$ of the distribution from which the analysis is generated, and the mean $\boldsymbol{\mu}_{\text{ens}}$ of the distribution from which the ensemble is drawn, are scaled by ''dividing by'' ensemble standard deviations (multiplication by the inverse of the covariance matrix $[\boldsymbol{\Sigma}_{\text{ens}}]$ representing the ensemble dispersion, so that correlations among the K dimensions are also accounted for). The vertical scales in each of the panels of Fig. 2 have been varied for clarity of presentation, but in each case the horizontal dashed line indicates the level of the number of ''expected'' $[= n/(n_{\text{ens}} + 1) = 91]$ counts per bin rank under uniformity.

The top row in Fig. 2 shows behaviors of MST histograms for unbiased forecasts, that is, for cases where the (vector) means $\boldsymbol{\mu}_{\text{ens}}$ and $\boldsymbol{\mu}_{\text{truth}}$ of the distributions from which the ensemble and the observation are drawn are equal for each of the n forecast occasions (although not necessarily the same from occasion to occasion). Here the MST histogram for $\sigma_{\text{truth}}/\sigma_{\text{ensemble}} = 1$ exhibits uniformity, within typical sampling variability for this sample size. Unbiased but overdispersed ensembles (left panels of top row) exhibit overpopulation of the higher ranks, reflecting the preponderance cases in which the MST length for the ensemble alone is the largest or among the largest of the $n_{\text{ens}} + 1$ MSTs for a given forecast. This condition tends to occur for overdispersed ensembles because the observation is often interior to the scatter of the ensemble, as in Fig. 1a, allowing space in the middle of the ensemble to be bridged (e.g., between the groups A–D and E–J in Fig. 1a) through that point, while dropping the segments associated with the omitted point elsewhere in the tree. This condition is accentuated in higher dimensions, where it is increasingly unlikely for an ensemble member to occur near the ensemble mean, because its value in *all* K dimensions must be near the corresponding mean value simultaneously. Quantitatively, for multivariate normal data (although the qualitative result does not depend on the distribution), the square of the Mahalanobis distance D in Eq. (2) (but between individual data values and their mean) follows the χ^2 distribution, with degrees of freedom equal to the dimension K of the space. This is so because the transformation produces K-independent standard Gaussian random variables (Mardia et al. 1979), the sum of the squares of which is well known to follow the χ^2_K distribution. The result is that the most likely value of the distance D between a Gaussian ensemble member and its mean is greater than zero and increasing in K for $K \geq 3$.

Even without ensemble bias, underdispersed ensembles ($\sigma_{\text{truth}}/\sigma_{\text{ensemble}} > 1$) characteristically exhibit overpopulation of the smallest ranks. The MST length for the ensemble members alone tends to be the smallest or among the smallest of the $n_{\text{ens}} + 1$ MST lengths because, for the remaining MSTs, the substantial distance between the observation and the ensemble is added to the MST length while a shorter segment within the ensemble is deleted (Fig. 1b). However, the observation is also usually well removed from the ensemble when there is a large ensemble mean error due to forecast bias. Thus, raw MST histograms for substantially

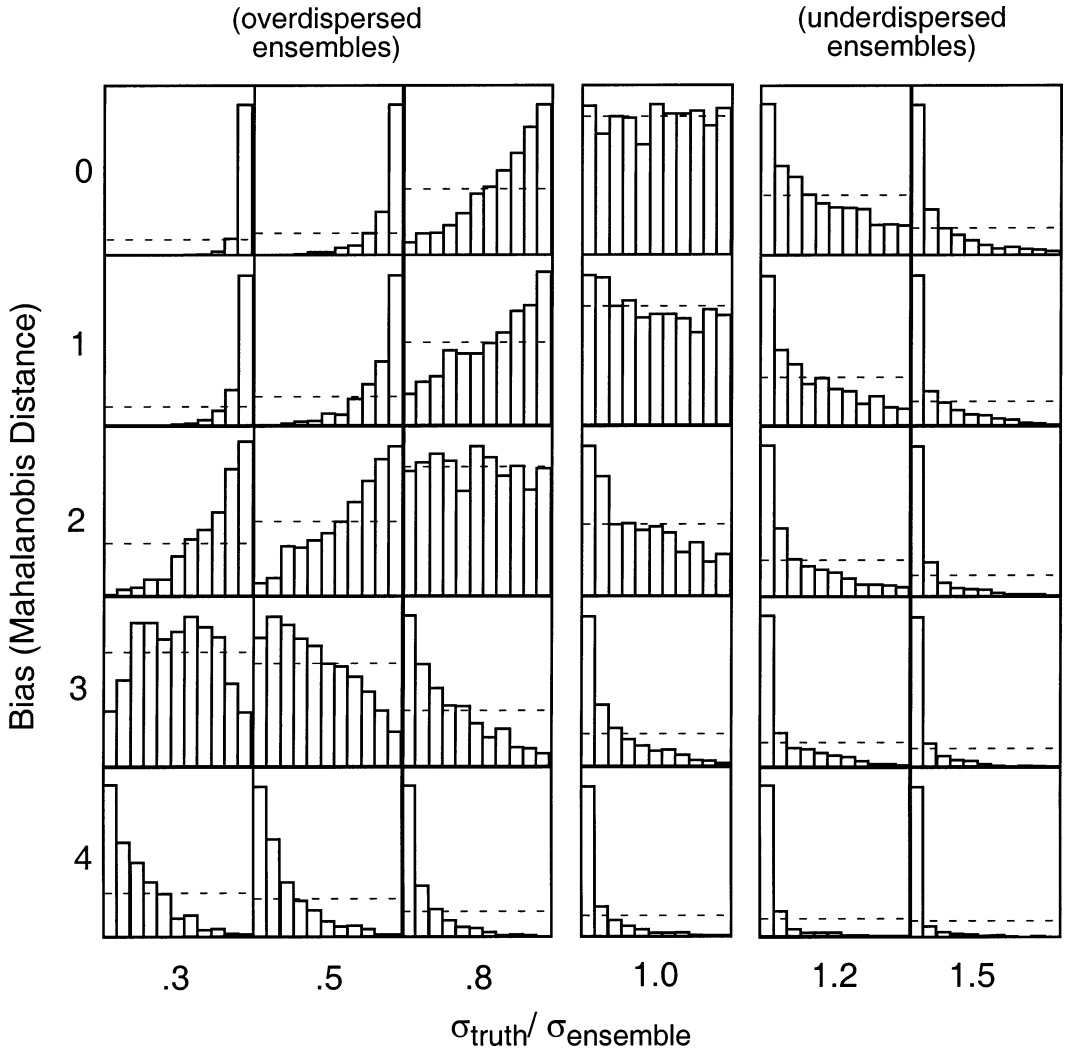

FIG. 2. Behaviors of MST histograms for $n_{ens} = 10$ in $K = 10$ dimensions, as functions of ensemble bias (vertical) and ensemble underdispersion (horizontal), from independent samples of size $n = 1000$. Vertical scales on each histogram have been varied for clarity of presentation, with the level of the expected number per bin under uniformity (1000/11 = 91) indicated in each case by the dashed line.

biased forecasts toward the bottom of Fig. 2 cannot be distinguished from MST histograms for underdispersed ensembles toward the right of Fig. 2. Similarly, the effects of ensemble bias and overdispersion can compensate to a degree, yielding MST histograms that are nearly uniform (e.g., bias = 2 and $\sigma_{truth}/\sigma_{ensemble} = 0.8$ in Fig. 2).

b. Scaled and bias-adjusted MST histograms

Figure 2 shows that raw MST histograms cannot distinguish between ensemble underdispersion and ensemble bias. Another problem may occur when there are different measurement scales or scales of variability on the different elements of the ensemble vector **x**. That is, if some of the K elements of **x** have variances that are very much smaller than the others, the MST will essentially ignore these dimensions because the corresponding terms in Eq. (1) will be small, so that the MST will essentially occupy only a subspace spanned by the high-variance elements.

These two problems can be addressed by computing MSTs using bias-corrected and scaled ensembles. First, the bias-corrected ensemble vector **x*** has elements

$$x_k^* = x_k - \left\langle \frac{1}{n_{ens}} \sum_{i=1}^{n_{ens}} x_k - o_k \right\rangle, \qquad k = 1, \ldots, K. \quad (3)$$

Here the x_k are the elements of a raw ensemble vector, and the angle brackets indicate the averages over all n ensembles and their corresponding observations, o_k.

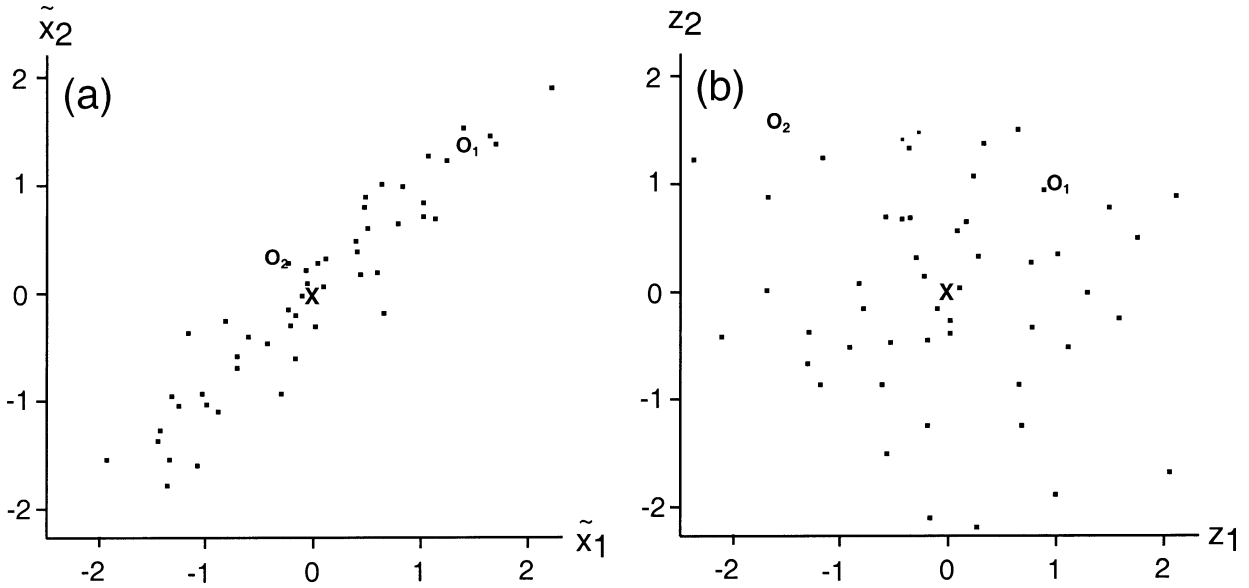

FIG. 3. Comparison of a hypothetical 50-member ensemble in $K = 2$ dimensions, as scaled by (a) dividing each dimension by the corresponding ensemble std dev [Eq. (5)] and (b) the Mahalanobis transformation [Eq. (6)]. Plots are centered at the ensemble mean (X) and show also two hypothetical observations O_1 and O_2 in relation to the ensemble.

That is, the angle-bracket term in Eq. (3) is a fixed quantity that is applied to each of the n_{ens} ensemble members in all of the n forecasts in a given verification sample. Note that some care is required in the implementation of Eq. (3), specifically that subsets of forecasts with nonhomogeneous bias characteristics (possibly, e.g., forecasts for winter versus summer seasons) are analyzed separately.

Define also the $(K \times K)$ ensemble variance–covariance matrix, including the observation, and calculated separately for each of the n forecast occasions, as

$$\mathbf{S}_{\text{ens}} = \frac{1}{n_{\text{ens}}}\left[(\mathbf{o} - \overline{\mathbf{x}}^*_{\text{ens}})(\mathbf{o} - \overline{\mathbf{x}}^*_{\text{ens}})^{\mathrm{T}} + \sum_{i=1}^{n_{\text{ens}}} (\mathbf{x}^*_i - \overline{\mathbf{x}}^*_{\text{ens}})(\mathbf{x}^*_i - \overline{\mathbf{x}}^*_{\text{ens}})^{\mathrm{T}}\right], \quad (4a)$$

where

$$\overline{\mathbf{x}}^*_{\text{ens}} = \frac{1}{n_{\text{ens}} + 1}\left(\mathbf{o} + \sum_{i=1}^{n_{\text{ens}}} \mathbf{x}^*_i\right) \quad (4b)$$

is the vector ensemble mean, including the observation $\mathbf{o}$ as the $n_{\text{ens}} + 1$st ensemble member. Inclusion of the observation in this way is necessary in order to obtain essentially flat MST histograms for ensembles of realistic size when it is truly the case that the observation and ensemble members are drawn from the same distribution.

One way to eliminate the effects of different measurement scales on the K elements of the forecast and observation vectors is to divide each by the corresponding standard deviation [square roots of the diagonal elements of Eq. (4a)],

$$\left.\begin{aligned} \tilde{x}_k &= x^*_k/(s^2_{\text{ens},k})^{1/2} \quad (5a)\\ \tilde{o}_k &= o_k/(s^2_{\text{ens},k})^{1/2} \quad (5b)\end{aligned}\right\} k = 1, \ldots, K.$$

However, this approach ignores the effects of correlation among the K elements of the forecast and observation vectors, which may be very substantial where forecasts for multiple locations are evaluated simultaneously.

An alternative scaling that also respects the correlation structure is the Mahalanobis transformation (e.g., Mardia et al. 1979),

$$\mathbf{z}_O = \mathbf{S}^{-1/2}_{\text{ens}}(\mathbf{o} - \overline{\mathbf{x}}^*_{\text{ens}}), \quad (6a)$$

$$\mathbf{z}_i = \mathbf{S}^{-1/2}_{\text{ens}}(\mathbf{x}^*_i - \overline{\mathbf{x}}^*_{\text{ens}}), \qquad i = 1, \ldots, n_{\text{ens}}. \quad (6b)$$

Here

$$\mathbf{S}^{-1/2}_{\text{ens}} = \mathbf{E}\boldsymbol{\Lambda}^{-1/2}\mathbf{E}^{\mathrm{T}}, \quad (7)$$

in which $\mathbf{E}$ is the matrix whose columns are the eigenvectors of $\mathbf{S}_{\text{ens}}$, and $\boldsymbol{\Lambda}^{-1/2}$ is the diagonal matrix whose elements are the reciprocals of the square roots of the corresponding eigenvalues. When $n_{\text{ens}} \leq K$, $\mathbf{S}_{\text{ens}}$ is not of full rank, and Eq. (7) is instead the generalized inverse (Mardia et al. 1979), or pseudoinverse (Stephenson 1997), in which $\mathbf{E}$ has n_{ens} columns corresponding to the nonzero eigenvalues, and the reduced matrix $\boldsymbol{\Lambda}^{-1/2}$ has dimension $(n_{\text{ens}} \times n_{\text{ens}})$.

Figure 3 illustrates the difference between the scalings in Eq. (5) (Fig. 3a) and Eq. (6) (Fig. 3b) for a hypothetical two-dimensional ensemble of size 50. In Fig. 3a the scaling has transformed both forecast variables to the same (unit) variance but has left the cor-

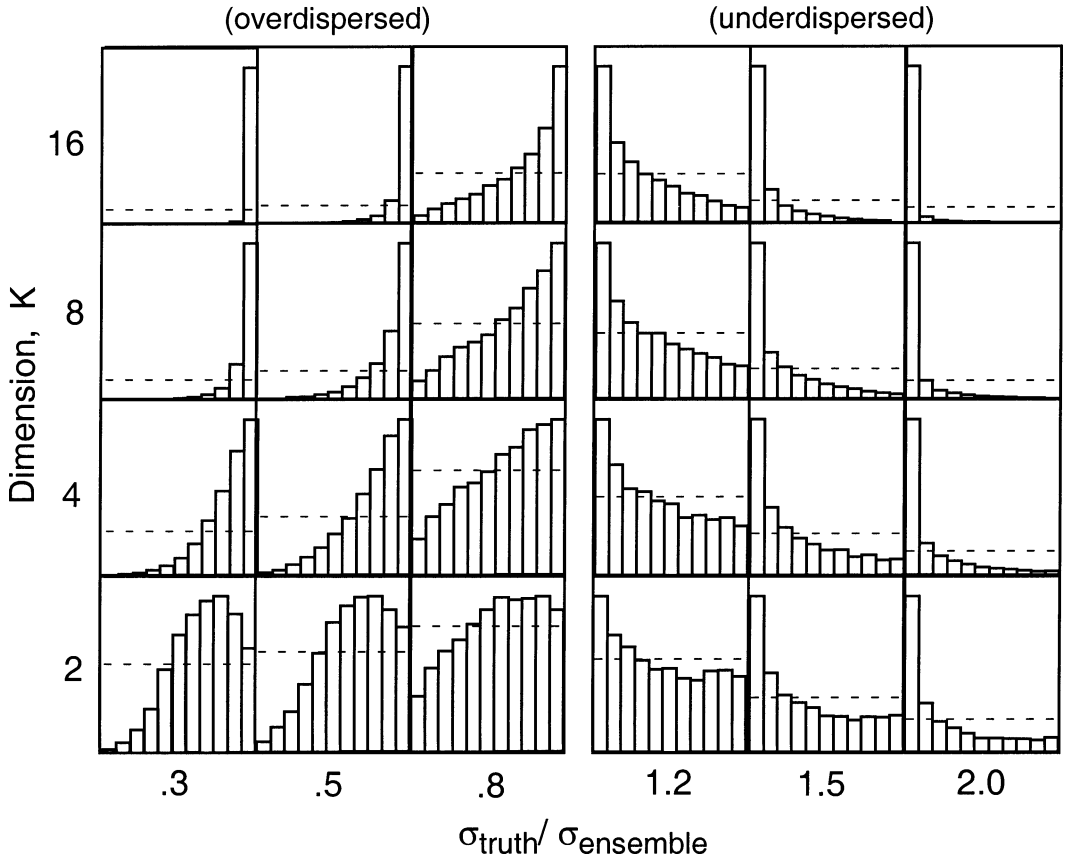

FIG. 4. Behaviors of scaled and debiased MST histograms for $n_{ens} = 10$, as functions of increasing dimensionality (vertical) and ensemble underdispersion (horizontal), from independent samples of size $n = 10\,000$. Vertical scales on each histogram have been varied for clarity of presentation, with the level of the expected number per bin under uniformity (10 000/11 = 909) indicated in each case by the dashed line.

relation (=0.95) unaffected. According to this scaling, a hypothetical observation O_1 is at a distance of 2 (standard deviation units) from the ensemble mean (X), which is plotted at the origin for convenience. Observation O_2 is much closer (0.5 standard deviation units) to the ensemble mean although it is outside the main ensemble scatter, and thus further removed from the ensemble mean according to the ensemble dispersion. In Fig. 3b both forecast variables have also been scaled to unit variance, but in addition the scaling in Eq. (6) reflects nearness of points in terms of the ensemble scatter itself, so that the distance [i.e., the Mahalanobis distance; Eq. (2)] from the ensemble mean to O_1 is 1.4, while the distance to O_2 is 2.2. That is, the Mahalanobis scaling emphasizes distances that are perpendicular to the main directions of scatter in the ensemble, reflecting the fact that points separated in such directions are less alike than points at an equal Euclidean distance apart in directions of the main ensemble scatter. Relative to Fig. 3a, the Mahalanobis scaling has in effect stretched the ensemble in the direction between the upper-left-hand and lower-right-hand corners of Fig. 3b. The result is that the two scaled variables z_1 and z_2 are uncorrelated and more correctly reflect (in terms of distances within the transformed space) the fact that O_1 is inside but at the edge of the ensemble while O_2 is near but outside.

Tabulation of MST histograms using the Mahalanobis transformation [Eq. (6)] is recommended in order to judge MST lengths in a way that is consistent with the shape of the ensemble scatter. The rank of the MST length for the scaled and debiased ensemble $\mathbf{z}_i$, $i = 1, \ldots, n_{ens}$, is then determined with respect to the MSTs obtained by substituting $\mathbf{z}_0$ in turn for each of the $\mathbf{z}_i$, and tabulating the MST histogram collectively for all n forecast occasions. In order not to lose the bias information, which will often be an important aspect of the forecast verification exercise, the K biases that are subtracted (angle-bracket term) in Eq. (3) need to be tabulated and presented with the MST histogram.

Figures 4 and 5 show characteristic shapes of the MST histograms derived from bias-corrected and scaled [according to Eq. (6)] ensembles, for ensemble sizes of 10 and 54, respectively. Again, these are results for synthetic, Gaussian ensembles and observations and are pre-

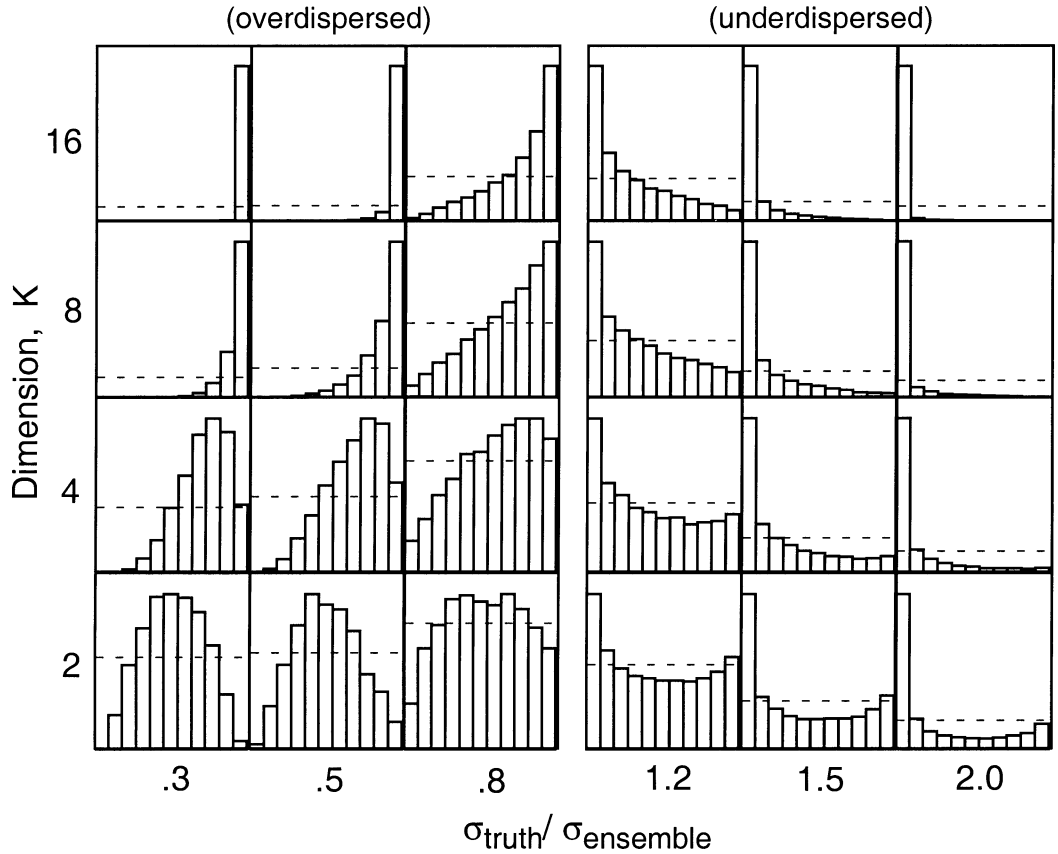

FIG. 5. As in Fig. 4, but for $n_{ens} = 54$, with each of the 11 bars indicating counts in five consecutive MST histogram bins for clarity of presentation.

sented as functions of ensemble underdispersion (horizontal) and the dimension K. Results for $\sigma_{truth}/\sigma_{ensemble} = 1$ have been omitted since these result in uniform MST histograms regardless of the ensemble size or dimension. For the larger dimensions K, the results are relatively insensitive to ensemble size, and the MST histograms are similar to the no-bias cases (top row) in Fig. 2. As the dimension increases, the MST histogram is increasingly sensitive to dispersion errors.

Overdispersed ensembles typically contain the observation as an interior point in a K-dimensional ''shell'' (because the probability of an ensemble member very near the ensemble mean is small in high-dimensional spaces) through which the MST can traverse a distance that would need to be bridged in any case. The result is that the MST excluding the observation is the longest or among the longest, leading to overpopulation of the high ranks. The members of underdispersed ensembles are typically farther from the observation than from each other, so the MST excluding the observation tends to be the shortest or among the shortest, leading to the characteristic overpopulation of the smaller ranks. The effects of ensemble size are more noticeable for smaller-dimension K, particularly for the overdispersed ensembles. Here there is a tendency for hump-shaped MST histograms rather than overpopulated high ranks, since in lower dimensions the ensemble tends to be more of a filled ball rather than a hollow shell, so the MST excluding the observation is often not extraordinarily long or short (Fig. 1a is thus somewhat atypical of $K = 2$-dimensional MSTs but has been chosen to illustrate the higher-dimensional behavior). This effect extends to higher dimensions for larger ensemble size, for example, $K = 4$ and $n_{ens} = 54$ in Fig. 5.

3. Example

In this section the foregoing ideas are applied to a small sample of ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) (Molteni et al. 1996). These are $n_{ens} = 51$-member ensembles initialized at 0000 UTC during the winter months of January and February 1997 and December 1997 through February 1998 and compared to the subsequent ECMWF analysis as the ''observation.'' Forecasts at 180-h lead time for 2-m air temperature, 10-m wind speed, and fractional cloud cover are considered, as interpolated to five lo-

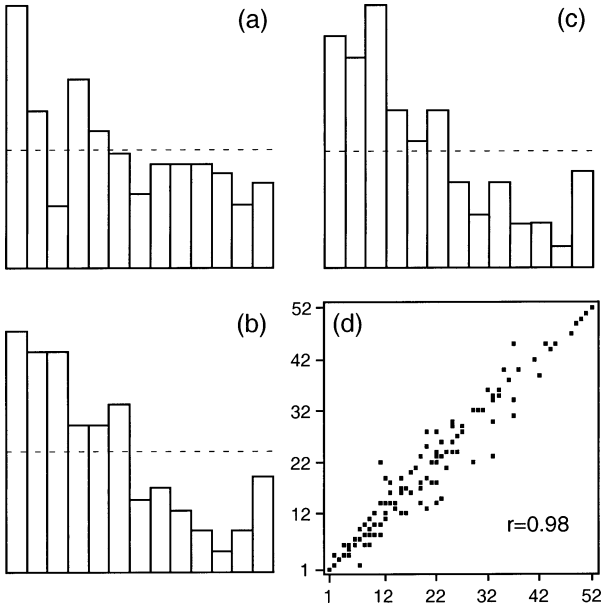

FIG. 6. (a) MST histogram for ECMWF EPS forecasts of temperature (°C), wind speed (m s^{-1}), and cloud fraction (%), at Birmingham, Bristol, Leeds, London, and Manchester, United Kingdom (i.e., considering 15-dimensional forecast vectors) at 180-h lead time for the 149 forecasts initialized during Jan and Feb 1997 and Dec 1997–Feb 1998. (b) Results for the same data, except with cloud fractions expressed as %/100, and (c) results omitting the cloud forecasts (10-dimensional forecasts). The ensemble size is 51, each of the 13 bars indicates counts in four consecutive MST histogram bins for clarity of presentation, and the expected number of counts (11.5) under uniformity is indicated by the dashed lines. (d) Scatterplot of MST ranks corresponding to (b) (vertical) and (c) (horizontal) and their correlation over the 149 cases, illustrating domination of the MST lengths by variables with larger scales of variation.

cations in the United Kingdom: Birmingham, Bristol, Heathrow (London), Leeds, and Manchester. Since there are forecasts for three weather elements at five locations for each of the $n = 149$ forecast occasions, the dimension K of the forecast vector $\mathbf{x}$ is 15.

Figure 6 shows raw MST rank histograms for these forecasts, with (a) indicating results when the cloud cover is expressed as percent, (b) showing the same results but with cloud cover expressed as a decimal fraction (percent/100), and (c) showing results for the reduced ($K = 10$) forecasts that include only temperature and wind speed at the five locations. Because of the wide disparity in measurement scales, the ensemble scatter in the five cloud cover dimensions dominates the MSTs summarized in Fig. 6a, whereas expressing cloud cover as decimal fractions (Fig. 6b) results in their being essentially ignored, so that these MSTs are nearly confined to the 10-dimensional subspace spanned by the five temperature and five wind speed variables. This result is confirmed by Fig. 6c, which shows the MST histogram for the $K = 10$-dimensional forecasts of the temperatures and wind speeds only. Figure 6c is very similar to Fig. 6b, with both exhibiting more extreme overpopulation of the smaller ranks than Fig. 6a. Figure 6d compares the MST ranks for these $n = 149$ cases, with ranks from Fig. 6c on the horizontal and ranks from Fig. 6b on the vertical. Here the correlation is 0.98, while the corresponding correlations between the points in Fig. 6a and the other two MST histograms are about 0.25.

TABLE 1. Ensemble biases [angle-bracket term in Eq. (3)] over the $n = 149$ forecasts.

Location (United Kingdom)	Temperature (°C)	Wind (m s^{-1})	Cloud cover (%)
Birmingham	−1.15	−0.73	−9.1
Bristol	−0.94	−0.77	−9.9
Heathrow	−0.60	−0.19	−7.8
Leeds	−0.95	−0.58	−6.4
Manchester	−1.11	−0.38	−9.6

In order to remove the effects of different measurement scales, and to separate the effects of possible bias and dispersion errors, the same ensemble forecasts were scaled and bias adjusted as described in section 2b. Table 1 shows the 15 bias corrections [angle-bracketed term in Eq. (3)]. These are all negative, indicating underforecasting of all three elements (too cool, calm, and clear, on average) at all five locations, although the absolute magnitudes are generally modest. Figure 7 shows the MST histograms for these forecasts (corresponding to Fig. 6a) when scaled (a) according to the Mahalanobis transformation [Eq. (6)] and (b) by dividing by corresponding ensemble standard deviations only [Eq. (5)]. Both Figs. 7a and 7b indicate that the ensembles are underdispersed, with the Mahalanobis scaling in Fig. 7a

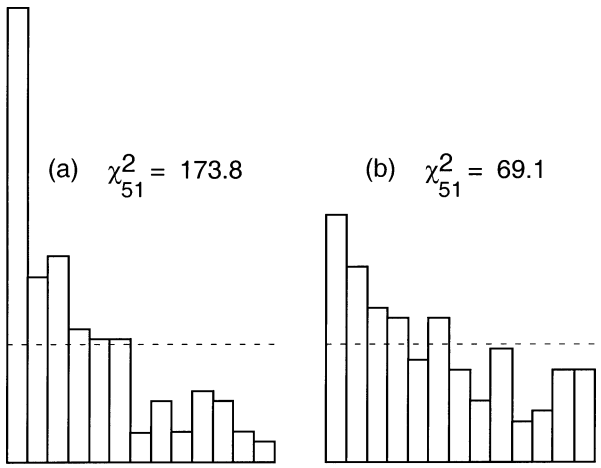

FIG. 7. MST histograms for the 15-dimensional forecasts, as in Fig. 6a, after removal of biases, and standardization to common scales according to (a) the Mahalanobis transformation [Eq. (6)] and (b) division of each ensemble vector element by its ensemble sd dev only [Eq. (5)]. Each of the 13 bars indicates counts in four consecutive MST histogram bins for clarity of presentation, and the expected number of counts (11.5) under uniformity is indicated by the dashed lines.

reflecting also the effects of the correlations among the forecast elements on the distances between ensemble members. These correlations are substantial, with average correlations among the five sites of 0.988, 0.876, and 0.935 for the temperature, wind, and cloud cover forecasts, respectively.

4. Chi-square tests for histogram uniformity given autocorrelated forecasts

It is conventional and appropriate to test for uniformity of the scalar rank histogram using the χ^2 statistic [e.g., Wilks 1995; Eq. (5.18)],

$$\chi^2_{n_{\text{ens}}} = \sum_{i=1}^{n_{\text{ens}}+1} \frac{[m_i - n/(n_{\text{ens}} + 1)]^2}{n/(n_{\text{ens}} + 1)}, \qquad (8)$$

where m_i is the number of counts in the ith bin, and $n/(n_{\text{ens}} + 1)$ is the expected number of counts in each bin under rank uniformity. Under the null hypothesis that a given rank histogram was drawn from a process in which assignment to any of the $n_{\text{ens}} + 1$ bins is independent and equally likely, this test statistic follows the χ^2 distribution with n_{ens} degrees of freedom (as indicated by the subscript on the left-hand side). If the statistic in Eq. (8) is larger than the appropriate critical value of this distribution, the null hypothesis is rejected. The same concepts and test statistic are appropriate in the case of the MST histogram.

One complication in the application of Eq. (8) to assessing rank uniformity, either for scalar rank histograms or for MST histograms, is that the tabulated critical values from the χ^2 distribution pertain to independent sequences of ensembles. This condition implies that sequences of forecasts must exhibit no serial correlation, which of course is often not the case. For example, the daily sequences of 180-h lead time temperature and wind forecasts described in section 3 exhibit lag $-$ 1 autocorrelations of approximately 0.5 and 0.4, respectively (the cloud cover forecasts are essentially uncorrelated).

Tabulated critical values from the χ^2 distribution can be adjusted to reflect the effects of serial correlation on the sampling variability of MST histograms, using the values provided as functions of the lag $-$ 1 autocorrelation ϕ, in Table 2. These have been computed using the simple stochastic model of ensemble behavior described in the appendix, in which the observation is statistically indistinguishable from the ensemble members by construction, and which reflect the Mahalanobis scaling of Eq. (6) through simulation of uncorrelated ensemble members [the submatrices on the diagonal of Eq. (A4), shown later in the appendix, are themselves diagonal]. The resulting adjustments are insensitive to the dimensionality K ($K \geq 2$) of the ensembles and depend on the ensemble size only through the degrees-of-freedom parameter of the χ^2 distribution, which in this setting is equal to the ensemble size. While a conventional rule of thumb states that there should be sufficient data to have at least five counts in each bin on average (in the present setting, $n/n_{\text{ens}} \geq 5$), the testing approach and the adjustments in Table 2 were found to be valid for $n/n_{\text{ens}} \geq 2$ or less.

TABLE 2. Additive corrections to tabulated χ^2 critical values to test uniformity of MST histograms as functions of lag $-$ 1 autocorrelation ϕ. Corrections for $\phi < 0.4$ are negligible.

	Test level, α			
ϕ	0.10	0.05	0.01	0.001
0.4	0.4	0.5	0.6	1.1
0.5	0.6	0.9	1.3	2.2
0.6	1.3	1.6	2.4	4.4
0.7	2.6	3.4	5.0	8.8
0.8	5.4	7.1	11.9	22.6
0.9	15.6	21.0	37.2	68.6

The χ^2 values [Eq. (8)] for the example scaled and bias-corrected MST histograms presented in section 3 are included in Fig. 7. Even though the histograms drawn in this figure have been collected into only 13 bins, the χ^2 values were computed using the 51 + 1 bin counts separately, as indicated by the subscripts in Fig. 7. The critical levels of χ^2_{51} at the $\alpha = 0.10$, 0.05, 0.01, and 0.001 levels are, respectively, 64.2, 68.7, 77.4, and 88.0. Accordingly, assuming zero serial correlation in the ensemble data, uniformity of the MST histogram in Fig. 6a would be rejected at the 0.1% level (173.8 $>$ 88.0). Because of the serial correlation in these forecasts, the correct critical levels are larger than the χ^2 quantiles (the ''effective sample size'' is smaller than that for an equal number of independent data), although only slightly so unless the serial correlation is quite strong. A complication here is that the lag $-$ 1 autocorrelations for the temperature, wind, and cloud cover variables are different, but the relative insensitivity of the adjustments in Table 2 to all but the largest values of ϕ minimize the problem, and in any case the example in Fig. 7a leads to an obvious negative conclusion on rank uniformity. The adjustment in Table 2 for $\phi = 0.5$ and $\alpha = 0.001$ is 2.2, so the adjusted critical level 88.0 + 2.2 $\ll$ 173.8, leading to easy rejection of the null hypothesis. In cases where adjustments appropriate to different values of ϕ for different forecast variables might be needed, they could be generated using these unequal values in each of the matrices on the diagonal of Eq. (A3).

Because the effects of the large correlations among the forecast elements have not been accounted for in Fig. 7b, quantitative interpretation of the χ^2 value for that MST histogram is not straightforward. It would be possible to evaluate adjusted χ^2 values for particular cases through simulations using Eq. (A1), in which the diagonal submatrices in Eq. (A4) reflected the observed correlations (see appendix).

Finally, Table 3 contains additive adjustments to tabulated χ^2 critical values, appropriate to evaluating uni-

TABLE 3. Additive corrections to tabulated χ^2 critical values to test uniformity of conventional (scalar) rank histograms as functions of lag − 1 autocorrelation ϕ.

	Test level, α			
ϕ	0.10	0.05	0.01	0.001
0.1	0.3	0.3	0.6	1.1
0.2	0.8	0.9	1.4	2.4
0.3	1.5	1.8	2.8	4.6
0.4	2.6	3.1	4.9	8.3
0.5	4.1	5.1	8.4	14.6
0.6	6.6	8.6	14.3	25.3
0.7	11.2	14.8	25.2	44.3
0.8	20.9	28.1	48.6	85.1
0.9	50.5	69.0	121.7	214.2

formity of scalar rank histograms. These were tabulated from simulations with the simple stochastic model described in the appendix, with $K = 1$ so that the submatrices in Eqs. (A3) and (A4) reduce to scalars. Again, dependence on the ensemble size is subsumed in the χ^2 critical values through its degrees-of-freedom parameter, and the results are valid for $n/n_{ens} \geq 2$, at least. Comparison of Tables 2 and 3 shows that the adjustments appropriate to scalar rank histograms are much more sensitive to serial correlation than are the values for MST histograms in Table 2.

5. Conclusions

This paper has examined the MST histogram, a conceptual extension (and not a mathematical generalization) for multidimensional ensemble forecasts of the conventional rank histogram for scalar forecasts. While not a complete verification tool, in the sense that it does not portray the joint distribution of forecasts and observations (Murphy and Winkler 1987), it does provide diagnostic information that may be useful in interpreting and improving ensemble forecasts. Notably, however, the MST histogram does not provide information on the resolution of the forecasts. That is, other things being equal, forecasts with smaller ensemble dispersion (provided it is appropriate to the forecast accuracy) yield more refined probabilities (and thus will be better forecasts to the extent that those refined probabilities are well calibrated, or reliable), but this attribute is not reflected in the MST histogram. This deficiency is also a characteristic of the conventional scalar rank histogram (e.g., Hamill 2001).

The MST histogram presents frequencies of ranks of lengths of ensemble MSTs, relative to the group of such lengths derived by substituting the observation in turn for each of its ensemble members. This convention is consistent with usual practice for scalar rank histograms but is opposite to the original proposal for the MST histogram made by Smith (2001), which results in histograms that are flipped horizontally relative to those described here. In raw form, the MST histogram cannot distinguish ensemble bias from ensemble underdispersion and will downweight or ignore forecast dimensions with small ensemble variability. This paper has advocated computing the MST histograms using forecasts that have been debiased ex post facto and scaled according to the Mahalanobis transformation [Eq. (6)], to eliminate the effects of different ensemble spreads in different dimensions and to account for the effects of correlations within the ensemble on effective distances between ensemble members. The bias information should be retained and reported with the MST histograms.

The behavior of MST histograms has been explored for synthetic Gaussian data, as a function of ensemble over- or underdispersion, ensemble size n_{ens}, and data dimension K; but this catalog of behaviors is not exhaustive. As noted by Hamill (2001) in the context of scalar rank histograms, qualitative deviations from these synthetic results may occur for real forecasts, for example, when ensemble properties are not homogeneous within a particular sample of n forecasts.

Adjustments to χ^2 values for evaluation of uniformity of the MST histograms to accommodate serial correlation in forecast data have also been presented. These adjustments are generally modest, except for the largest magnitudes of serial dependence. The values in Table 2 pertain to ensembles that have been scaled according to Eq. (6) and are not appropriate to MST histograms in which the effects of ensemble correlation on proximity of ensemble members has not been accounted for. Corresponding χ^2 adjustments for assessing uniformity of scalar rank histograms have also been presented.

Verification approaches and other interpretation methods for ensemble forecasts are only just developing. In addition to the scalar rank histogram, alternative ensemble verification methods that recently have been suggested include Bayesian probabilities of the observation given the ensemble distribution (Wilson et al. 1999), scalar performance measures based on economic value (Richardson 2000; Wilks 2001), bounding boxes (Smith 2001), multidimensional scaling (Stephenson and Doblas-Reyes 2000), and time evolution of the ensemble eigenvalues and eigenvectors, and of the ensemble entropy (Stephenson and Doblas-Reyes 2000). Given the intrinsically high dimensionality (Murphy 1991) of ensemble forecast verification, it seems possible that a unified approach to ensemble verification that intelligibly expresses the full joint distribution of forecasts and observations may not be achieved. The MST histogram may develop as one of a number of useful and important diagnostics for ensemble forecasts.

Acknowledgments. I thank ECMWF for supplying the EPS forecast data. The comments of Tom Hamill and two anonymous reviewers have improved the presen-

tation of the paper. This work was supported by NSF under Grant ATM-0221542.

APPENDIX

A Multivariate Autoregressive Model for Ensemble Forecast Behavior

Critical values for the χ^2 statistic assessing MST and ordinary rank histogram uniformity for nonindependent (serially dependent) forecast ensembles were obtained by simulation using the standard [e.g., Wilks 1995; Eq. (8.51)] first-order vector autoregression

$$\mathbf{x}(t) = \mathbf{\Phi}\mathbf{x}(t-1) + \mathbf{B}\,\mathbf{e}(t). \tag{A1}$$

While other forms for the underlying stochastic model would affect the results reported in Tables 2 and 3, although possibly only to a small degree, first-order autoregressions have been chosen because they are often very reasonable models for daily weather data (e.g., Wilks 1995). Here the forecast vector $\mathbf{x}$ simultaneously encompasses the vector observation $\mathbf{x}_0$ and all n_{ens} K-dimensional forecasts $\mathbf{x}_k$, as

$$\underset{K(n_{\text{ens}}+1)\times 1}{\mathbf{x}} = [\mathbf{x}_0^{\mathrm{T}} | \mathbf{x}_1^{\mathrm{T}} | \mathbf{x}_2^{\mathrm{T}} | \cdots | \mathbf{x}_{n_{\text{ens}}}^{\mathrm{T}}]^{\mathrm{T}}. \tag{A2}$$

That is, $\mathbf{x}$ is partitioned into $n_{\text{ens}} + 1$ K-dimensional subvectors, the first of which corresponds to the observation vector. The matrix $\mathbf{\Phi}$ is block diagonal, according to

$$\mathbf{\Phi} = \begin{bmatrix} \phi\mathbf{I} & 0 & 0 & \cdots & 0 \\ 0 & \phi\mathbf{I} & 0 & \cdots & 0 \\ 0 & 0 & \phi\mathbf{I} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \phi\mathbf{I} \end{bmatrix}, \tag{A3}$$

where each submatrix is $(K \times K)$, corresponding to the partition of $\mathbf{x}$ in Eq. (A2). Here all autoregressive coefficients ϕ (equal to the lag -1 autocorrelation for the respective scalar time series), $0 \le \phi < 1$, for the K forecast elements are equal, and the submatrix $\phi\,\mathbf{I}$ for the observation vector in the upper-left-hand corner is equal to those for each of the ensemble members. These assumptions are consistent with the purpose of section 4, in which the observation must be drawn from the same distribution as the ensemble by construction, although they could be relaxed in other applications.

The matrix $\mathbf{B}$ in Eq. (A1) depends also on the matrix of simultaneous (i.e., unlagged) variances and covariances, specified here as

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \mathbf{I} & r\mathbf{I} & r\mathbf{I} & \cdots & r\mathbf{I} \\ r\mathbf{I} & \mathbf{I} & r\mathbf{I} & \cdots & r\mathbf{I} \\ r\mathbf{I} & r\mathbf{I} & \mathbf{I} & \cdots & r\mathbf{I} \\ \vdots & \vdots & \vdots & & \vdots \\ r\mathbf{I} & r\mathbf{I} & r\mathbf{I} & \cdots & \mathbf{I} \end{bmatrix}. \tag{A4}$$

Here the diagonal submatrices are all the identity $\mathbf{I}$, indicating unit variance for, and no correlation among, all forecast elements within each ensemble member. The uncorrelatedness of the ensemble members is consistent with the Mahalanobis scaling in Eq. (6) that has been recommended for MST histogram calculation. In other applications these conditions could be relaxed, allowing in particular that different forecast elements have different variances and nonzero correlation. The parameter r, $0 \le r < 1$, in Eq. (A4) controls the ensemble dispersion, and so the lengths of the MSTs, but is immaterial with respect to the ranks of the MST lengths. It has been set to the value 0.9 in the simulations that produced Tables 2 and 3. When the off-diagonal submatrices in Eq. (A4) are not themselves diagonal, an ellipsoidal region, rather than a (hyper-) spherical region, of the K-dimensional forecast space is occupied by the simulated ensembles.

Using Eqs. (A3) and (A4), $\mathbf{B}$ can be any matrix satisfying

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{\Sigma}_0 - \mathbf{\Phi}\mathbf{\Sigma}_0\mathbf{\Phi}^{\mathrm{T}} \tag{A5}$$

This equation is arrived at by postmultiplying Eq. (A1) by $\mathbf{x}_t^{\mathrm{T}}$ and taking expectations to yield the simultaneous covariance matrix $\mathbf{\Sigma}_0 = \mathbf{\Phi}\mathbf{\Sigma}_1^{\mathrm{T}} + \mathbf{B}\mathbf{B}^{\mathrm{T}}$, similarly multiplying Eq. A1 by $\mathbf{x}_{t-1}^{\mathrm{T}}$ and taking expectations to yield the lag -1 autocovariance matrix $\mathbf{\Sigma}_1 = \mathbf{\Phi}\mathbf{\Sigma}_0$, and combining the two equations (e.g., Bras and Rodriguez-Iturbe 1985). A consistent solution for $\mathbf{B}$ can be obtained using the Cholesky factorization of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ (e.g., Atkinson 1978; Bras and Rodriguez-Iturbe 1985), or through its eigenvalues and eigenvectors [i.e., the inverse of Eq. (7)]. Equation (A1) can then be used for stochastic simulation by subsituting, at each time step t, a vector of independent standard normal variates for $\mathbf{e}(t)$.

REFERENCES

Ahuja, R., T. Magnanti, and J. Orlin, 1993: *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 846 pp.

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

Atkinson, K. E., 1978: *An Introduction to Numerical Analysis.* Wiley, 587 pp.

Bras, R. L., and I. Rodriguez-Iturbe, 1985: *Random Functions and Hydrology.* Addison-Wesley, 559 pp.

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

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

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.

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 Met Office Library, London Road, Bracknell, Berkshire RG12 2SZ, United Kingdom.]

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

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Academic Press, 518 pp.

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., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

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

Smith, L. A., 2001: Disentangling uncertainty and error: On the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* A. E. Mees, Ed., Birkhauer Press, 31–64.

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

Stephenson, D. B., 1997: Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in predictions. *Tellus,* **49A,** 513–527.

——, and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus,* **52A,** 300–322.

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

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8,** 209–219.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.,* **127,** 956–970.