Theory and Applications of the Minimum Spanning Tree Rank Histogram

DANIEL GOMBOS AND JAMES A. HANSEN*

Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts

JUN DU AND JEFF MCQUEEN

NOAA/NWS/NCEP, Environmental Modeling Center, Camp Springs, Maryland

(Manuscript received 17 February 2006, in final form 9 August 2006)

ABSTRACT

A minimum spanning tree (MST) rank histogram (RH) is a multidimensional ensemble reliability verification tool. The construction of debiased, decorrelated, and covariance-homogenized MST RHs is described. Experiments using Euclidean L_2, variance, and Mahalanobis norms imply that, unless the number of ensemble members is less than or equal to the number of dimensions being verified, the Mahalanobis norm transforms the problem into a space where ensemble imperfections are most readily identified. Short-Range Ensemble Forecast Mahalanobis-normed MST RHs for a cluster of northeastern U.S. cities show that forecasts of the temperature–humidity index are the most reliable of those considered, followed by mean sea level pressure, 2-m temperature, and 10-m wind speed forecasts. MST RHs of a Southwest city cluster illustrate that 2-m temperature forecasts are the most reliable weather component in this region, followed by mean sea level pressure, 10-m wind speed, and the temperature–humidity index. Forecast reliabilities of the Southwest city cluster are generally less reliable than those of the Northeast cluster.

1. Introduction

The sensitive dependence of atmospheric dynamics to initial conditions limits the utility of deterministic forecasts (Lorenz 1963). This sensitivity motivates an ensemble approach to forecasting that discretely approximates the time evolution of probability distribution functions (PDFs; Epstein 1969). The intrinsic probabilistic nature of ensemble forecasts necessitates changes in forecast verification. It is imperative to derive and implement systematic ensemble verification techniques in order to identify weaknesses of and expedite improvements of predictions and models.

The identification of ideal verification techniques requires an understanding of the nature of goodness in weather forecasting. Weather forecast goodness is typically defined in terms of a forecast's consistency, value, and quality (Murphy 1993), which is further subdivided into components that include sharpness, resolution, and reliability (Murphy 1993). Since no known verification measure satisfactorily addresses all aspects of goodness, it is necessary for a verification tool to address an individual aspect. This paper focuses on the assessment of ensemble reliability, which is defined as the correspondence between the mean of the observations associated with a particular forecast and that forecast, averaged over all forecasts. A perfectly reliable 30% chance precipitation forecast, for example, verifies exactly 30% of the time (Murphy 1993). It is important to reiterate that, although they are extremely important measures of forecast goodness, this paper is not concerned with the assessment of forecast sharpness and resolution.

Reliability can be measured by the degree to which the ensemble forecast members and truth are random samples from the same PDF. For scalar forecasts, this degree can be assessed by the shape of a rank histogram (RH), or Talagrand diagram (Anderson 1996; Talagrand et al. 1997). The scalar RH is simply a histogram of the N verification ranks over N independent forecast

* Current affiliation: Naval Research Laboratory, Monterey, California.

Corresponding author address: Daniel Gombos, Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139-4307.
E-mail: dgombos@mit.edu

DOI: 10.1175/MWR3362.1

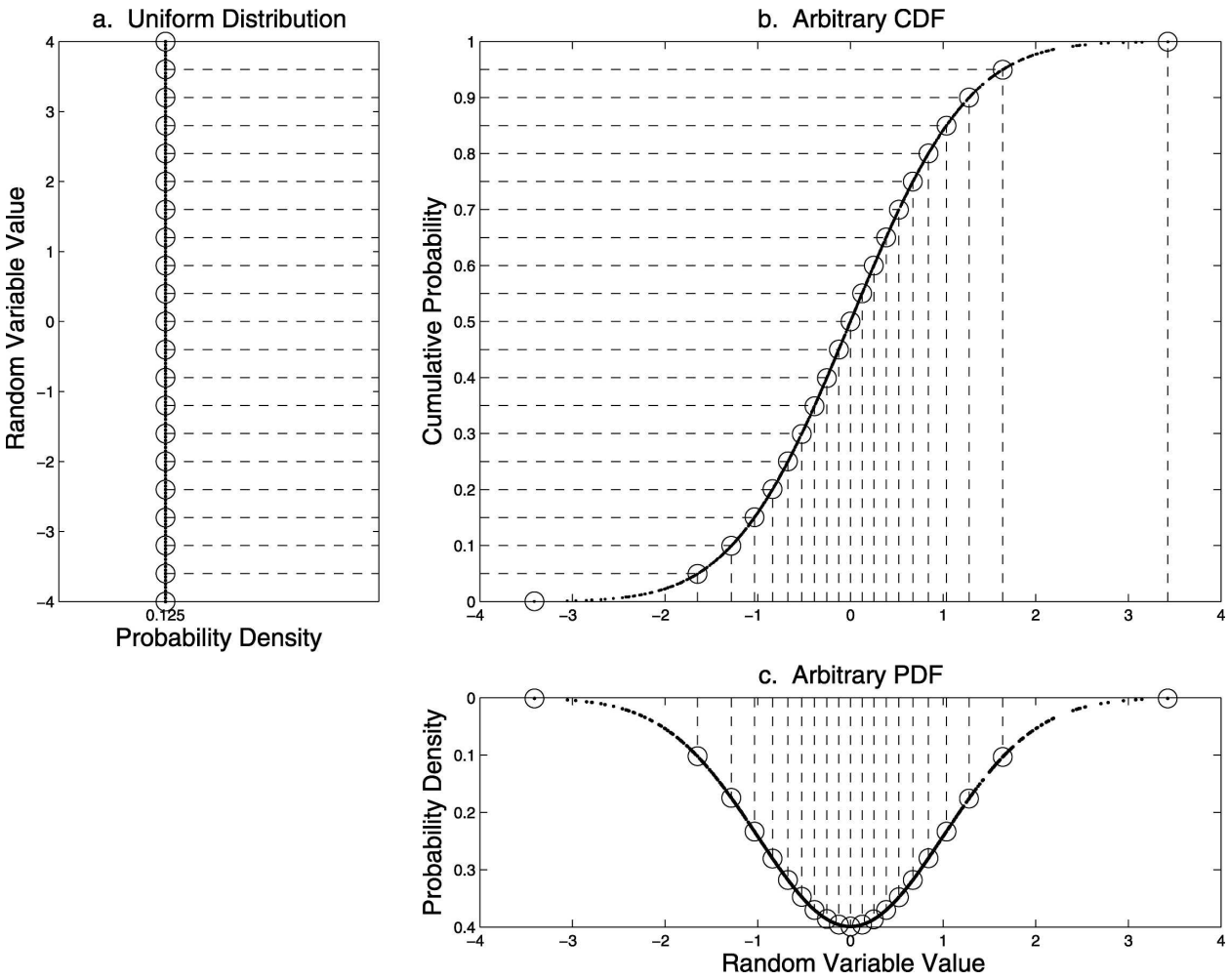

FIG. 1. An illustration of the CDF as a transfer function from a PDF to a uniform distribution. See text for details.

occasions. Each verification rank is defined as the rank of the verification entry in a forecast's $n_{ens} + 1$ member vector composing an individual forecast's n_{ens} ensemble entries and the corresponding verification entry, sorted in ascending order. Therefore, the histogram's shape depends on the population of the $n_{ens} + 1$ bins, as determined by the N ranks of the verification entries in the N vectors.

An equal representation of ranks, as indicated by a flat histogram, implies that the members of the ensemble forecast and the verification are random draws from the same PDF: they are statistically indistinguishable. This is easily conceptualized by thinking of the forecast cumulative distribution function (CDF) as a transfer function between the forecast PDF and a uniform distribution. Figure 1 shows a continuous schematic of this idea. The forecast PDF is shown in Fig. 1c (upside down for convenience), the associated CDF in Fig. 1b, and the uniform distribution that results from using the CDF to transform random draws from the PDF in Fig. 1a. The circles in Fig. 1c represent the boundaries between areas of equal probability; the integral of the PDF between each circle is the same. Note that the functional form of the PDF (which is arbitrary) results in unequal spacing between the circles. When these points are transformed by the CDF in Fig. 1b (dashed lines guide the eye), they result in the uniform distribution in Fig. 1a. In the construction of an RH, the circles in Fig. 1 are defined by the forecast ensemble, their rank ordering approximates the forecast CDF, and verification populates the bins of the transformed distribution.

Traditional RHs are used to assess one-dimensional forecasts. The atmosphere, however, is far from one-dimensional. Because of the covariance between dimensions, averaging univariate RHs to assess the multidimensional reliability can give misleading information (Smith and Hansen 2004). Therefore, in order to

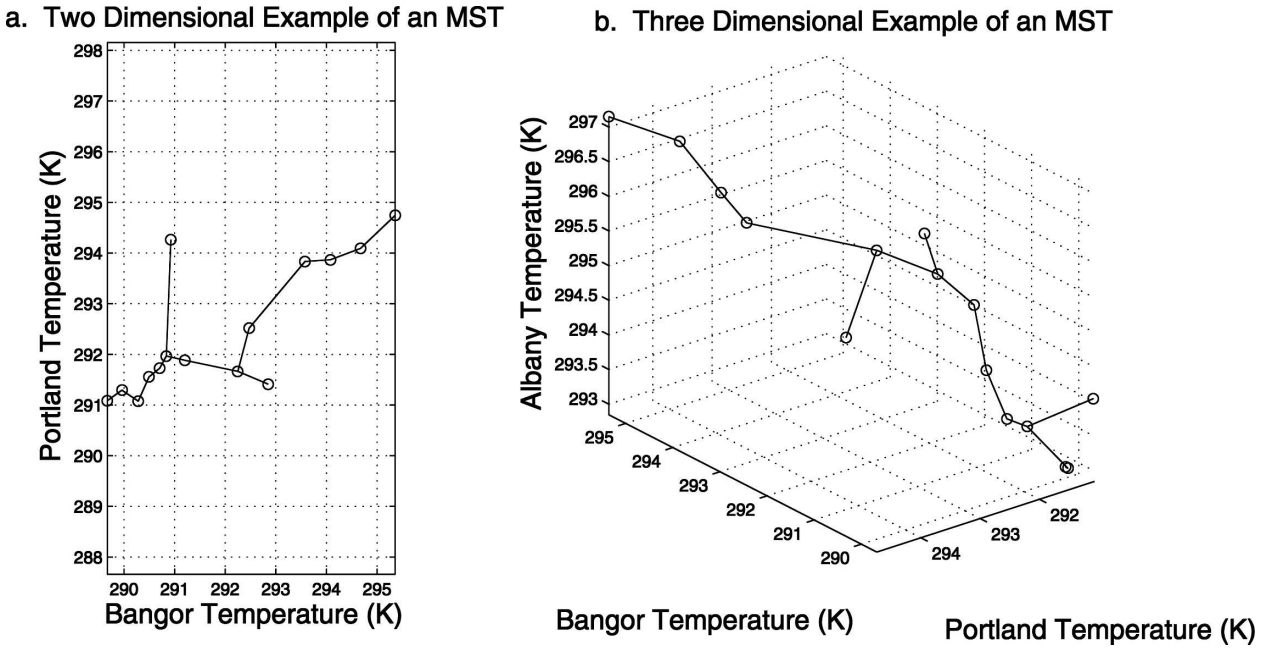

FIG. 2. Illustrations of (a) two- and (b) three-dimensional MSTs for a 24-h forecast. The dimensions are represented by the cities and the norm is 2-m temperature on 21 Aug 2004. Circles represent the $n_{\text{ens}} = 15$ points that could comprise either the ensemble only, or the union of $n_{\text{ens}} - 1$ ensemble members and the verification. The sum of the line segments represents the MST distance.

accurately assess the reliability of multidimensional fields, it is desirable to formulate a multidimensional extension of the RH that accounts for this covariance.

One such extension is the minimum spanning tree (MST) RH. Consider a K-dimensional space, where each dimension could correspond to one of K individual weather components, such as temperature or pressure, to the same component in one of K different locations, or to a combination of components and locations. Let each point, $x_{i,j,k}$, in this K-dimensional space correspond to the value of the kth element of the jth ensemble member on the ith forecast occasion, where $i = 1, \ldots, N$, $j = 1, \ldots, n_{\text{ens}}$, and $k = 1, \ldots, K$. The MST of this set of points is defined by the sum of the lengths (under a chosen norm) of the $n_{\text{ens}} - 1$ line segments that connect these points, subject to the restrictions that the resulting network has no closed loops and that the distance is minimized (Smith and Hansen 2004; Wilks 2004). Figure 2a shows an example of a two-dimensional MST with $n_{\text{ens}} = 15$ and 24-h lead time, where one dimension corresponds to the forecast temperature in Bangor, Maine, and the other to the forecast temperature in Portland, Maine, on 21 August 2004. Each circle represents $x_{i,j,k}$ and the sum of the $n_{\text{ens}} - 1$ line segments represents the MST. Figure 2b shows a three-dimensional example of an MST, where the third dimension corresponds to the forecast temperature in Albany, New York.

The calculation of each increment of an MST RH requires the computation of n_{ens} MST lengths. The first of these lengths is the MST distance of the n_{ens} ensemble points alone. The other n_{ens} lengths are the MST distances of the n_{ens} points consisting of the union of $n_{\text{ens}} - 1$ ensemble points and the verification. The verification replaces a different ensemble member for each of these n_{ens} lengths. If the ensemble members and the verification are random draws from the same PDF, the MST length of the ensemble-only points should be statistically indistinguishable from the n_{ens} MST lengths that include the verification. Analogous to a traditional scalar RH being a plot of the rank of the verification within the $n_{\text{ens}} + 1$ member vector over N one-dimensional forecasts, the MST RH is a plot of the rank of the ensemble-only MST length within the $n_{\text{ens}} + 1$ member MST length vector.

The degree to which the ensemble and verification points are statistically indistinguishable can be quantified using the Cramér–von Mises (CvM) goodness-of-fit test for a uniform distribution. The CvM test statistic, W^2, is given by

$$W^2 = N^{-1} \sum_{q=1}^{n_{\text{ens}}+1} Z_q^2 m_q, \tag{1}$$

where m_q is the probability of an observation landing in the qth bin, O_q and E_q are the observed and expected number of counts in the qth bin, respectively, and

$$Z_q = \sum_{r=1}^{q} (O_r - E_r). \quad (2)$$

Given the independence of each of the N forecast occasions, a histogram will be considered flat if this test statistic is less than the CvM critical value with n_{ens} degrees of freedom.[1] Note that the CvM statistic was chosen to assess flatness because, unlike the χ^2 statistic, it is sensitive to rank ordering and gives a more powerful goodness-of-fit assessment for small sample sizes (Elmore 2005). The CvM statistic is particularly sensitive to skewed histograms (Elmore 2005) and is therefore appropriate for the assessment of debiased MST RHs, which are characteristically right skewed for underdispersed ensembles and left skewed for overdispersed ensembles (Wilks 2004).

This paper addresses both theory and applications of the MST RH. Section 2 details MST distance norms and how the improper use of such norms causes misleading MST RH shapes. Section 3 describes the data used to construct the MSTs used in the application section of this paper. Section 4 is an analysis of separate MST RHs that were constructed by using a common city cluster, but different weather component norms. This section also compares the MST RHs from a southwestern U.S. city cluster and a northeastern U.S. city cluster. Section 5 presents the conclusions.

2. MST distance norms

A multidimensional ensemble reliability assessment determines the statistical similarities of the ensemble forecast distribution and the verification distribution. Because the MST RH determines this likeness using the ranks of MST distances, it is crucial to choose a norm for these distances that most accurately measures this statistical similarity.

The three choices of the norm considered in this paper are the Euclidean L_2, variance, and Mahalanobis norms. Each will be described below. Other than the circumstance using the Mahalanobis norm when $n_{\text{ens}} \leq K$ described below, in the limit of large numbers of realizations, the use of each of these norms in the construction of MST RHs will qualitatively yield the same determination of whether or not the two distributions are alike. However, as the number of realizations decreases, the choice of norm can potentially influence the CvM statistic's evaluation of population histogram flatness, motivating the use of the most sensitive and justifiable norm. The following describes how each of these norms can give misleading measurements about the degree of reliability of an ensemble forecast and outlines circumstances when certain norms should *not* be used.

a. L_2 norm

The familiar Euclidean L_2 norm is the most intuitive and straightforward norm to use when constructing MST RHs. However, because it does not homogenize the variances of the data, the L_2 norm yields a misleading MST RH when the standard deviation of the data in each of the K dimensions is not the same. Consider the $K = 8$, $n_{\text{ens}} = 15$, and $N = 140$ L_2 MST RH depicted in Fig. 3a. For this contrived example, each of the $K = 8$ dimensions represents the 2-m temperature in an individual city, and suppose that the true distribution is known. Assume that the true standard deviations of the temperatures in the first four cities are 5 K and that the forecasts for these cities are perfectly reliable, with standard deviations also equal to 5 K. Also assume that the true standard deviations in the other four cities are 1 K but that the forecasts for these cities are underdispersed, with standard deviations of only 0.1 K. Despite the underdispersion of half of the forecasts, the L_2 MST is relatively flat. Because it does not homogenize variances, the L_2 MST distances are dominated by the distances associated with the high standard deviation dimensions; the incorrect, but small, distances associated with the low variance cities are "lost in the noise." Of course, with a large enough number of samples, the L_2 MST RHs would correctly indicate that the ensembles are drawn from the incorrect $K = 8$ distribution.

b. Variance norm

The variance norm transforms each entry, $x^*_{i,j,k}$, of $\mathbf{X}^*_i$ into $x^{\text{var}}_{i,j,k}$ such that

$$x^{\text{var}}_{i,j,k} = x^*_{i,j,k}/\sigma_{i,k}, \quad (3)$$

where $\sigma_{\text{i,k}}$ is the standard deviation of the data in the kth dimension and $\mathbf{X}^*_i$ is an $(n_{\text{ens}} + 1) \times K$ matrix formed by the union of the verification vector, $\mathbf{o}_i$, and n_{ens} ensemble row vectors of length K, $\mathbf{x}^*_{i,j}$. (The star superscript indicates that the ensemble has been debiased. The bias transformation procedure will be explained in the following section.) The MST distance is then formed using the transformed $x^{\text{var}}_{i,j,k}$ entries.

A variance norm MST RH will equally weight the ensemble and verification dispersion differences in the unit directions of the cities. After each data point is divided by the standard deviation of its respective dimension, the data in each transformed dimension has unit variance. Therefore, from the previous example, a distance of 1 K in the low standard deviation temperature axis and a distance of 5 K in the high standard

[1] The CvM critical values can be found in Table 1 of Elmore (2005).

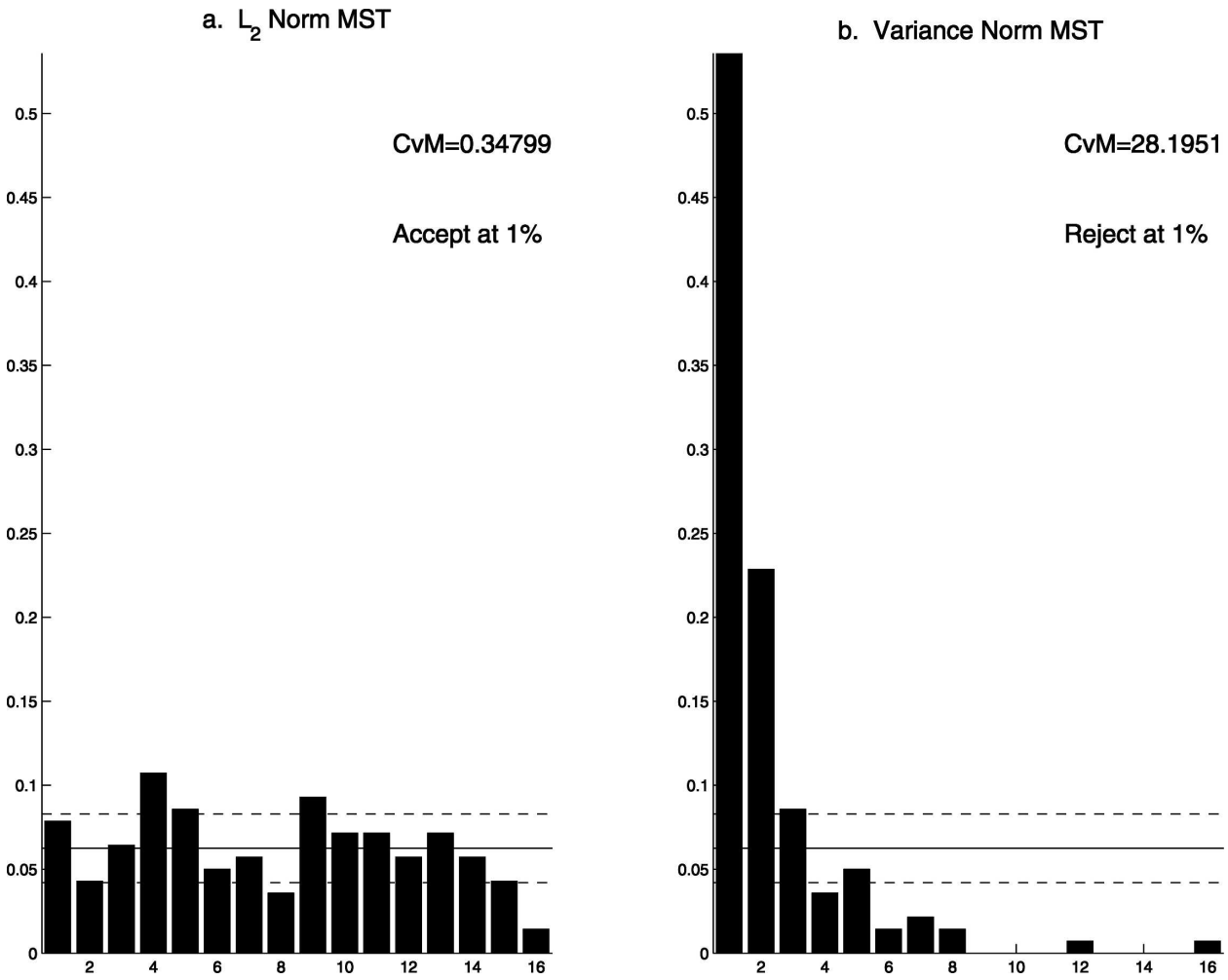

FIG. 3. (a) The L_2 MST RH is nearly flat, even though four of the eight ensembles are highly underdispersed. However, since it homogenizes variances and thereby equally weights all dimensions, (b) the variance norm MST RH is underdispersed. The solid line represents the expected number of counts in each bin, p, given a perfectly flat histogram. Dotted lines represent a one standard deviation bound of this expectation, $(1/\sqrt{N})\sqrt{p(1-p)}$ (Smith and Hansen 2004).

deviation temperature axis will be weighted equally under the variance norm, thereby enabling the MST distance to equitably account for each dimension in the reliability assessment.

Figure 3b shows the variance norm MST of the same $K = 8$, $n_{\text{ens}} = 15$, and $N = 140$ temperature data as used in the L_2 example above. As portrayed by its right skewed shape, the variance-norm MST RH properly shows the underdispersed relationship between the ensemble and verification. By homogenizing the variances of all dimensions, the variance norm equally weights all dimensions when computing MST distances and therefore is able to capture the extreme underdispersion of the low standard deviation elements.

Although it averts the problems presented in the previous example, the variance norm is not ideal when the ensemble dimensions covary. Consider a two-dimensional linear cluster of highly correlated ensemble points and two hypothetical verification points portrayed in Fig. 4a. The x dimension has a standard deviation of 0.1 and the y dimension has a standard deviation of 10. Verification point B has a short Euclidean distance but a large statistical distance from the mean of the cluster of points measured in terms of standard deviation units of the associated two-dimensional PDF. Verification point A has a relatively large Euclidean distance but a short statistical distance from the mean of the cluster of points compared to point B. Because it is farther than A from the mean in terms of standard deviation units, point B is significantly less statistically similar to the ensemble than is point A.

Figures 4a,b show how these ensemble and verifica-

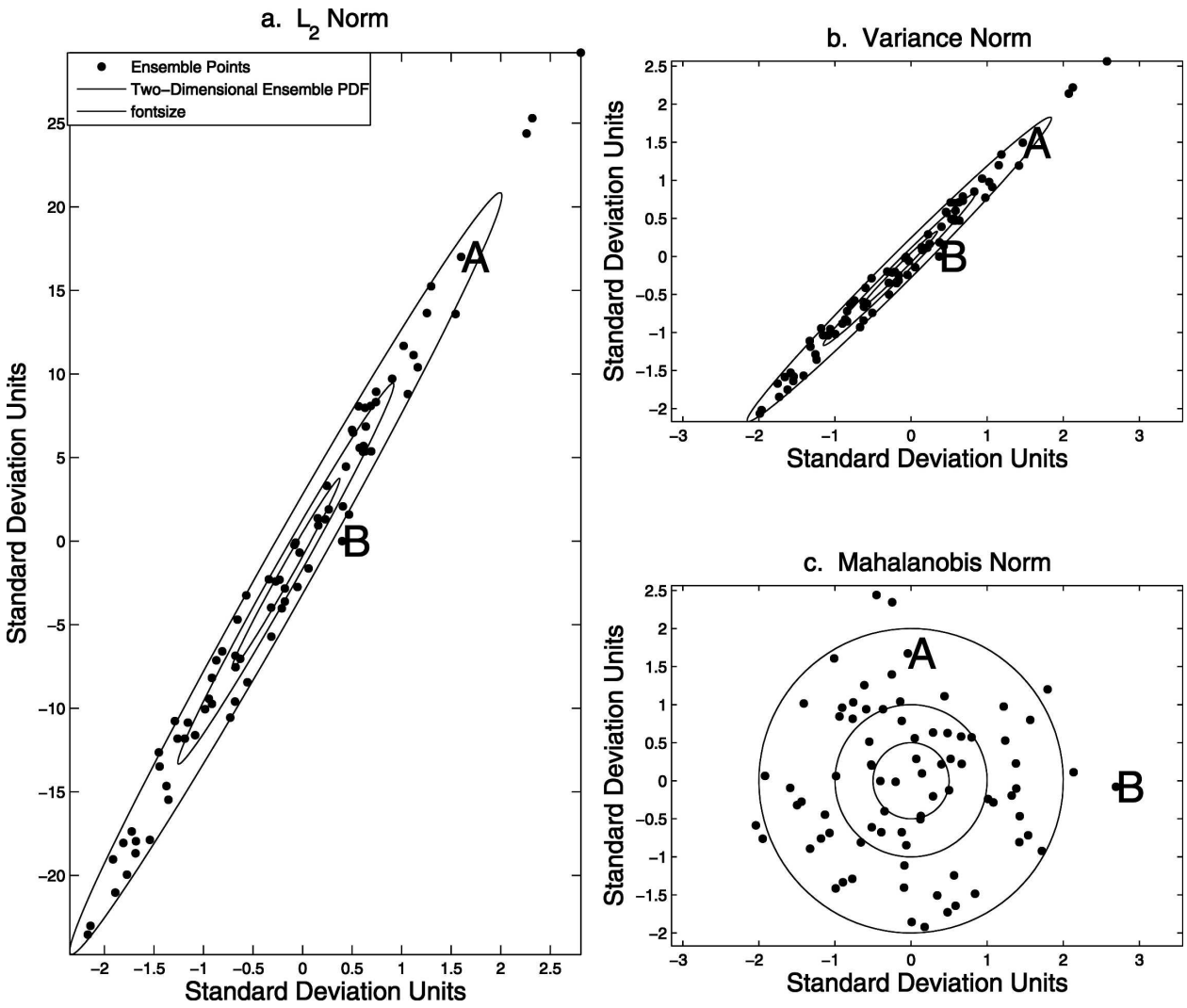

FIG. 4. A comparison of the behavior of ensemble and verification points under the L_2, variance, and Mahalanobis norms. See text for details.

tion points behave under the L_2 and variance norms, respectively. Since it has undergone no transformations of variance or covariance, the L_2 norm simply reflects the case described above. The collection of points in the variance norm space portrays a similar structure as the L_2 norm, with the notable difference that the x and y dimensions have been homogenized. However, since it has not accounted for the covariance of the data, the variance norm improperly implies that point A is less similar to the ensemble than is point B, as seen by its greater Euclidean distance (in variance-normed space) from the ensemble PDF. Therefore, a variance-normed MST RH systematically using points similar to verification point A for all forecast occasions will be significantly less flat than one using points similar to point B, even though each B point is less likely to be a random draw from the same distribution that forms the respective ensemble.

c. *Mahalanobis norm*

The Mahalanobis transformation is a conversion of the verification vector and ensemble vectors to the multivariate counterparts of the z score:

$$\mathbf{z}_{i_o} = \mathbf{C}_i^{-1/2}(\mathbf{o}_i - \bar{\mathbf{x}}_i^*) \quad \text{and} \tag{4}$$

$$\mathbf{z}_{i,j} = \mathbf{C}_i^{-1/2}(\mathbf{x}_{i,j}^* - \bar{\mathbf{x}}_i^*) \text{ (Wilks 2004).} \tag{5}$$

Here $\bar{\mathbf{x}}_i^*$ is the length K vector whose entries are the averages of the columns of $\mathbf{X}_i^*$ (as defined above), $\mathbf{C}_i$ is the covariance matrix of $\mathbf{X}_i^*$, and

$$\mathbf{C}_i^{-1/2} = \mathbf{E}_i\mathbf{D}_i^{-1/2}\mathbf{E}_i^{\mathrm{T}}, \tag{6}$$

where the columns of $\mathbf{E}_i$ are the eigenvectors of $\mathbf{C}_i$ and the entries of the diagonal matrix $\mathbf{D}_i$ are the corresponding eigenvalues of $\mathbf{C}_i$. Mahalanobis-normed MSTs are computed in the same way as are L_2-normed MSTs, except that $\mathbf{z}_{i_o}$ is substituted, in turn, for one of the $\mathbf{z}_{i,j}$, instead of the L_2 verification vector being substituted, in turn, for one of the L_2 ensemble vectors (Wilks 2004; Mardia et al. 1979). Note that this forecast error covariance norm, $\mathbf{C}^{-1/2}$, performs the same function as the analysis error covariance norm used to transform nonisotropic initial uncertainty into isotropic initial uncertainty in singular vector computations. In each case, this operation simply defines the mean of the multivariate distribution to be zero and the covariance to be the identity matrix.

The Mahalanobis transformation homogenizes the variances and decorrelates the points that form the MST by operating on the ensemble and verification points with the covariance matrix, thereby eliminating the problems associated with the L_2 and variance norms. This operation effectively alters the Euclidean distances (in Mahalanobis-normed space) so that they properly reflect the statistical "closeness" of points from the mean (Wilks 2004; Mardia et al. 1979). Therefore, as depicted in Fig. 4c, the Mahalanobis transformation decreases the Euclidean distance (in Mahalanobis-normed space) of point A from the mean of the cluster and increases the Euclidean distance (in Mahalanobis-normed space) of point B from the mean of the cluster.

Although it effectively accounts for covariance information, the Mahalanobis norm gives misleading results when $R - 1 \leq K$, where R is the number of samples used to compute the covariance. (In the case of the Mahalanobis-normed MST RH, R was previously defined to be $n_{\text{ens}} + 1$, the number of rows of $\mathbf{X}_i^*$.) This circumstance results in a symmetric configuration of the Mahalanobis-normed points in which every pair of points in the transformed space is separated by a distance of *exactly* $\sqrt{2(R-1)}$. These points form a perfect R hedron, analogous to a two-dimensional equilateral triangle or three-dimensional tetrahedron. Figure 5 shows nine examples of a $K = 2$ and $R = 3$ set of random points with zero mean and unit variance. The times signs and circles represent these same points under the L_2 and Mahalanobis norm, respectively. The line segments connecting the circles are shown to indicate the shapes of the triangles formed by the three points; they are not the MST, but they do indicate the problem encountered by the MST RH. For $R - 1 \leq K$, *all* MST distances are *exactly* the same, rendering the Mahalanobis-normed MST RH useless. Note that the implication of $R - 1 \leq K$ is that the ensemble is spanning a rank deficient space, and the Mahalanobis norm always chooses R hedrons as the most efficient way to isotropically span that space.

Figure 6 presents six Mahalanobis-normed MST RHs in which the verification and the ensembles are random draws from the same distribution. The only difference between the panels is the number of dimensions used to calculate the MST distances. Given that the ensembles and the verification are random draws from the same distribution, the MST RHs should be flat. However, when $n_{\text{ens}} \leq K$ (Figs. 6d,e,f), the Mahalanobis-normed MST RHs appear to indicate an underdispersed ensemble. In reality, all MST distances are exactly the same (to within machine precision) and the verifying MST distance satisfies the condition for populating the leading bin. The fact that bins other than the leading bin are populated is an artifact of round-off error creating differences in MST distances at the level of machine precision. Note that when $n_{\text{ens}} \leq K$, it is necessary to calculate $\mathbf{C}_i^{-1/2}$ using truncations of $\mathbf{E}_i$ and $\mathbf{D}_i$. The $\mathbf{E}_i$ comprises the n_{ens} columns corresponding to the nonzero eigenvalues of $\mathbf{C}_i$ and the entries of the diagonal $n_{\text{ens}} \times n_{\text{ens}}$ matrix $\mathbf{D}_i$ are the nonzero eigenvalues of $\mathbf{C}_i$ (Wilks 2004).

Researchers familiar with the ensemble-based data assimilation literature will likely be concerned about spurious correlations due to sampling errors. While sampling errors will certainly exist for small ensemble sizes, since each increment to a Mahalanobis-normed MST RH uses a common covariance matrix, each of the $n_{\text{ens}} + 1$ MST distances are subject to the same errors. The sampling errors may increase the number of forecast occasions needed to discern that a Mahalanobis-normed MST RH is nonflat, but they will not make flat histograms appear nonflat.

3. Description of data

The following section applies the MST rank histogram to assess multidimensional ensemble reliability using the National Centers for Environmental Prediction (NCEP) Short-Range Ensemble Forecast (SREF) datasets. SREF consists of 15 ensemble members, 5 of which come from the Eta model with a Betts–Miller–Janjic (BMJ) convective scheme, 5 of which come from the same Eta model but with a Kain–Fritch convective scheme, and 5 of which come from the Regional Spectral Model (RSM) with a simplified Arakawa–Schubert (SAS) convective scheme. All 15 SREF ensembles are perturbed in their initial conditions. For a description of this system, the reader is referred to Du et al. (2003).

Two separate SREF datasets are used to construct the MST rank histograms in this work. Aiming to im-

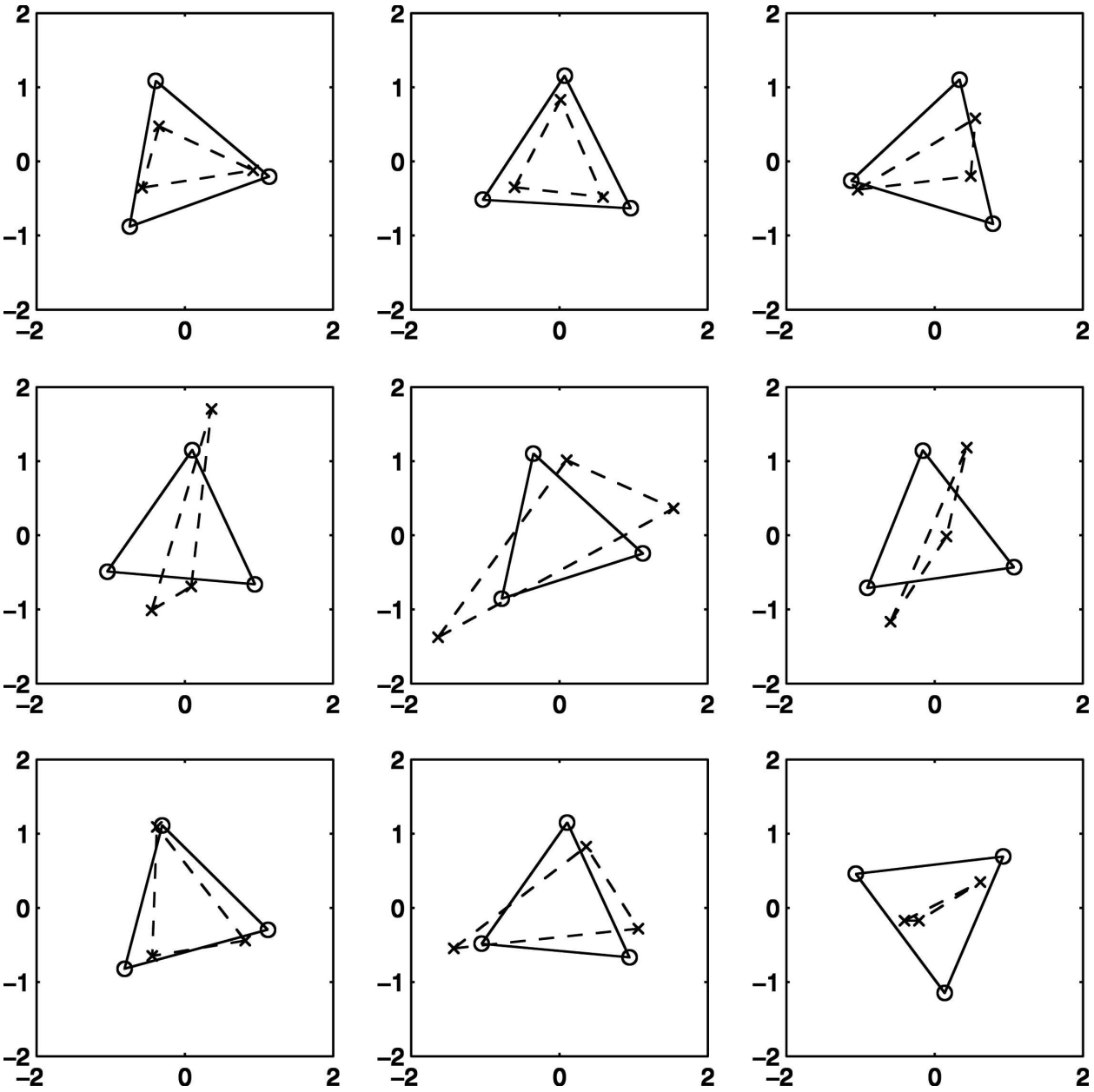

FIG. 5. The three times signs represent three randomly chosen points with zero mean and unit variance. Line segments connect the times signs in order to show the shape of the triangle formed by the three points. These lines do not represent the MST. The three circles indicate these same points under the Mahalanobis norm. Notice that in each of the nine randomly chosen examples, the Mahalanobis-normed points form an equilateral triangle with sides of length 2. When $R - 1 \leq K$, the Mahalanobis-normed points form an R hedron with equally spaced points separated by a distance of exactly $\sqrt{2(R-1)}$.

prove ensemble diversity and forecast spread, SREF physics diversity was modified by increasing the number of convective schemes (from three to six) and cloud microphysics parameterizations (Du et al. 2004; McQueen et al. 2005). This upgrade, which occurred on 17 August 2004, also included an increase in the model resolution; the 10 ETA members have 60 levels and a 32-km horizontal resolution and the 5 RSM members have 28 levels and a 40-km horizontal resolution. The first dataset, which will be referred to as SREF1, comprises these upgraded ensemble forecasts from 18 August 2004 to 13 May 2005. The second set, which will be referred to as SREF2, is an older ensemble prediction system with slightly less ensemble diversity and decreased model resolution; the 10 ETA members have 45 levels and a 48-km resolution and the 5 RSM mem-

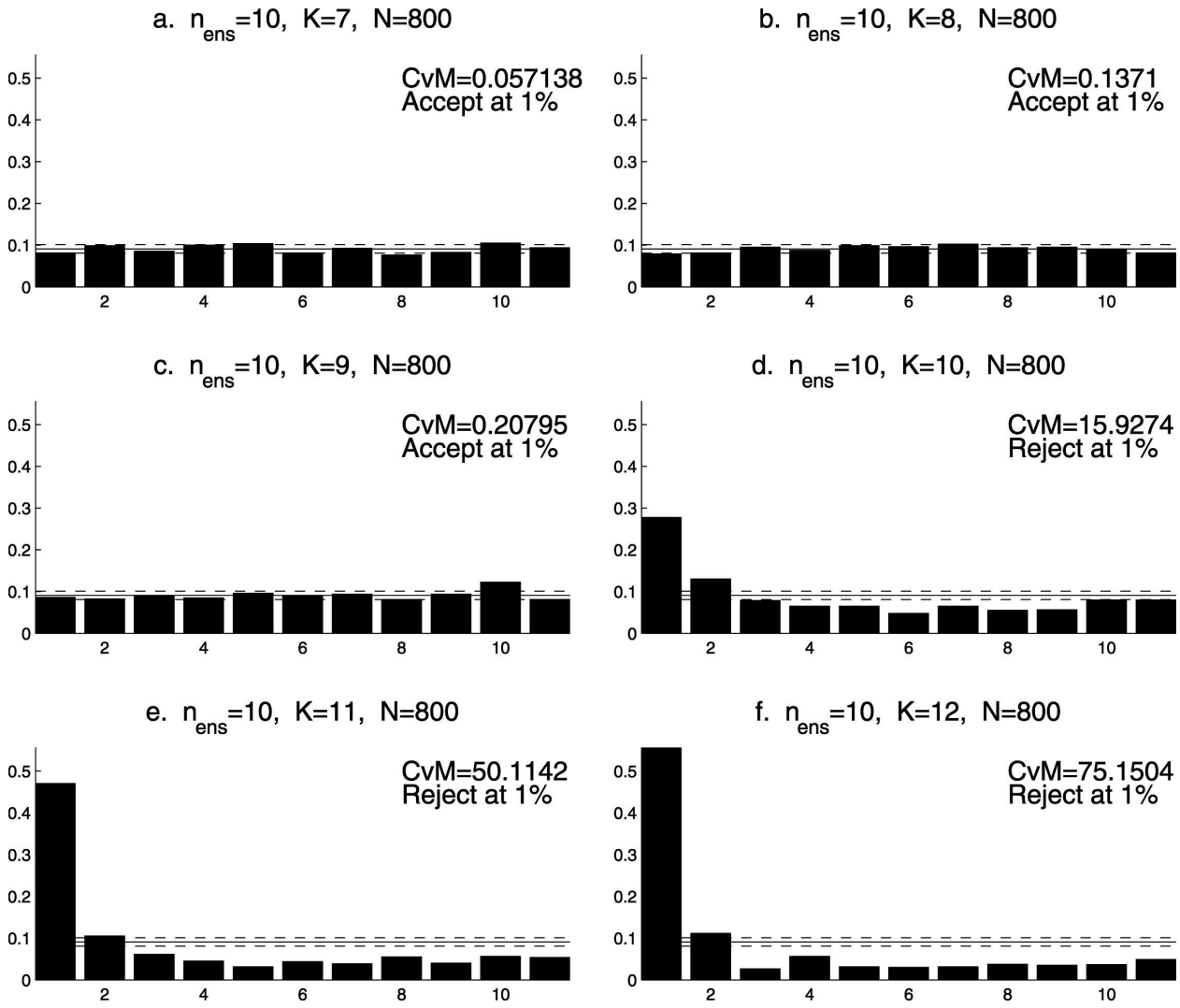

FIG. 6. Mahalanobis-normed MST RHs with $n_{ens} = 10$ and varying dimensions. Since both the ensemble and the verification are random draws from a random Gaussian distribution with zero mean and unit variance, all RHs should be flat. However, when $n_{ens} \leq K$, the Mahalanobis-normed MST RHs spuriously indicate an underdispersed ensemble. The solid line represents the expected number of counts in each bin, p, given a perfectly flat histogram. Dotted lines represent a one standard deviation bound of this expectation, $(1/\sqrt{N})\sqrt{p(1-p)}$ (Smith and Hansen 2004).

bers have 28 levels and a 48-km horizontal resolution. This set uses data from 3 June to 17 August 2004.

The SREF forecasts were verified using two separate verification datasets. The first verification was obtained by randomly selecting between the two Eta analysis controls and the one RSM analysis control. These controls were not averaged because averaging significantly reduced the variance of the verification. The second verification consisted of station observations obtained from the National Climate Data Center that undergo extensive automated quality control. Ensemble forecast values were linearly interpolated to the station locations.

4. Analysis of the multidimensional reliability of weather components

This section uses MST RHs to compare the multivariate reliabilities of forecast components for various cities clusters, forecast components, and lead times. For all of the following examples, $K = 7$ (7 different cities) and $n_{ens} = 15$ (the 10 Eta forecasts and the 5 RSM forecasts). Because of unlike variances of the data in the different dimensions, significant covariance between dimensions, and $n_{ens} > K$, the Mahalanobis norm has been used to calculate MST distances. Note that for the cases considered, L_2-normed MST RHs give quali-

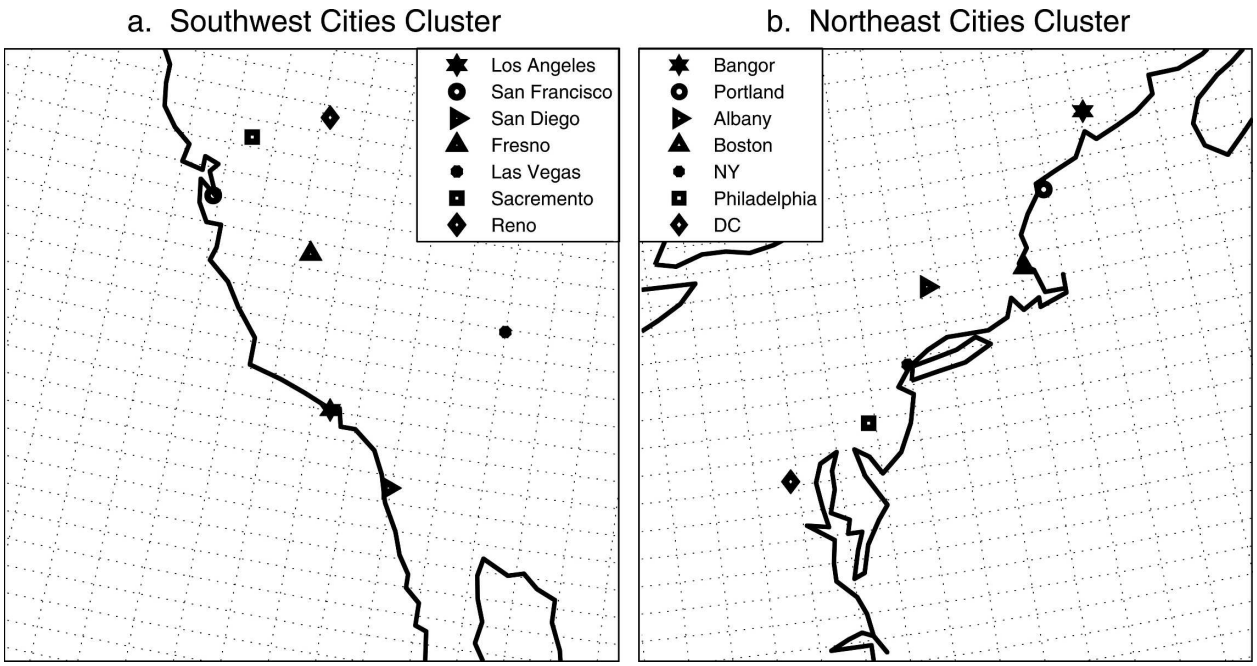

FIG. 7. Locations of the $K = 7$ city (a) Southwest cluster the (b) Northeast cluster.

tatively similar results to the Mahalanobis-normed MST RHs, but the Mahalanobis-normed MST RHs are significantly less flat.

Mahalanobis-normed MST RHs were separately computed for two clusters of $K = 7$ cities. The first cluster comprises the northeastern U.S. cities shown in Fig. 7a: Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Washington, D.C.; Albany, New York; Portland and Bangor; the second cluster comprises the southwest U.S. cities shown in Fig. 7b: San Diego, Los Angeles, San Francisco, Sacramento, and Fresno, California; Reno, and Las Vegas, Nevada.[2] Each MST RH uses all seven cities to separately assess the ensemble forecast reliability of one of four different weather components: mean sea level pressure (P_{MSL}), 2-m temperature ($T_{2\mathrm{m}}$), 10-m wind speed ($u_{10\mathrm{m}}$), and the temperature–humidity index (THI). SREF1 was used to compute the P_{MSL}, $T_{2\mathrm{m}}$, and $u_{10\mathrm{m}}$ MST RHs and SREF2 was used to compute the THI MST RHs. SREF1 was not used to compute the THI MST RHs because necessary dewpoint temperature information was not available as part of the SREF1 dataset. Also note that the analysis THI MST RHs were verified using the SREF2 analyses.

The choice of P_{MSL}, $T_{2\mathrm{m}}$, and $u_{10\mathrm{m}}$ was motivated by their obvious importance in typical weather forecasts. The THI, as defined by

$$\mathrm{THI}(^\circ\mathrm{F}) = 0.55 \times T_{2\mathrm{m}}(^\circ\mathrm{F}) + 0.2 \times T_{d_{2\mathrm{m}}}(^\circ\mathrm{F}) + 17.5 \text{ (Glockman 2000)}, \quad (7)$$

where $T_{d_{2\mathrm{m}}}$ is the 2-m dewpoint temperature, was chosen because of its importance in energy markets. This index is an indicator of the sultriness due to the combined effects of temperature and humidity. Therefore, the accurate prediction of the THI is crucial for markets

[2] The northeastern and southwestern city clusters were chosen because they are the most populous regions in the United States. The reader should be aware that the SREF reliabilities depicted by the MST RHs in this paper may be significantly different than the SREF reliabilities for other regions.

TABLE 1. Averaged 7-day running biases for the Northeast cluster cities from Fig. 8.

	Bangor	Portland	Albany	Boston	New York	Philadelphia	Washington
P_{MSL} (mb)	−0.63	−0.39	−0.51	−0.27	−0.50	−0.63	−0.76
$T_{2\mathrm{m}}$ (°C)	−0.08	−0.41	−0.37	−0.81	−0.77	−0.26	−0.19
$u_{10\mathrm{m}}$ (m s^{-1})	2.46	3.30	2.46	3.32	2.53	2.49	2.80
THI (°F)	0.64	−0.59	−0.50	−1.03	−0.62	−0.30	−0.33

TABLE 2. Same as in Table 1, except that the observations are used as the verification.

	Bangor	Portland	Albany	Boston	New York	Philadelphia	Washington
P_{MSL} (mb)	−0.59	−0.52	−0.81	−0.33	−0.17	−0.60	−0.64
$T_{2\mathrm{m}}$ (°C)	−0.46	0.54	−1.30	−2.13	−2.93	−1.86	−2.20
$u_{10\mathrm{m}}$ (m s^{-1})	−1.95	−0.92	−1.67	−4.52	−4.64	−3.83	−2.19
THI (°F)	−0.82	−0.33	−1.95	−2.18	−3.38	−1.61	−2.61

that are sensitive to supply and demand fluctuations induced by air conditioner energy usage. Because energy companies are particularly interested in regional forecasts, a multivariate reliability assessment of the THI is especially important. As it is an indicator of sultriness, the THI RHs were computed using only summer data.

Following Stensrud and Skindlov (1996), 7-day running mean biases have been removed from all MST RHs in this section. Define $x^*_{i,j,k}$ to be

$$x^*_{i,j,k} = x_{i,j,k} - \left[\frac{1}{7n_{\mathrm{ens}}} \sum_{m=1}^{7} \sum_{j=1}^{n_{\mathrm{ens}}} (x_{i-m,j,k} - o_{i-m,k}) \right], \tag{8}$$

where $o_{i,k}$ is an individual verification data point in the kth dimension. This transformation simply subtracts the average bias of each dimension, for the 7-day period prior to the ith day, from each ensemble data point of the corresponding dimension on the ith day. All MST RHs in the application sections of this paper have been computed using the debiased $x^*_{i,j,k}$ points. The biases reported in Tables 1–4 are the averages of these 7-day running mean biases for each city and weather component.

Figures 8–11 show Mahalanobis-normed MST histograms for 24-h forecasts valid at 0900 UTC, the associated CvM test statistic, and a histogram flatness assessment at the 1% significance level. The verification for each increment in Figs. 8 and 10 is a random selection of one of the three SREF control analyses for the corresponding day; the verification for each increment in Figs. 9 and 11 is the actual observation for the corresponding day, with forecast values interpolated to the location of the observing station. Figures 8 and 9 are for the Northeast cluster of cities and Figs. 10 and 11 are for the Southwest cluster. Note that the number of counts in each bin has been divided by $N = 81$ for the P_{MSL}, $T_{2\mathrm{m}}$, and $u_{2\mathrm{m}}$ histograms and by $N = 21$ for the THI histogram to yield relative frequency histograms. The solid line represents the expected number of counts in each bin, p, given a perfectly flat histogram. To give an indication of the effects of the small sample size on the flatness, dotted lines representing a one standard deviation bound of this expectation, $(1/\sqrt{N})\sqrt{p(1 - p)}$, have also been included (Smith and Hansen 2004). Also note that, because the proper interpretation of an MST RH requires that each increment be statistically independent of others, the MST RHs are constructed using data from every third day, the lag at which bin population autocorrelations were found to be relatively negligible.

Regardless of the verification type, city cluster location, or weather component, multidimensional SREF forecasts are underdispersed, as indicated by the right-skewed MST RHs in Figs. 8–11. Despite recent attempts by NCEP to increase ensemble diversity, short-range ensembles members lack sufficient differences to capture the PDF of the verification. Further initial condition, physics, and/or parameterization diversifications are needed. Although all RHs are right skewed, the degree of underdispersion depends on the choice of the verification, city cluster location, and weather component. Figures 8 and 9 indicate that forecasts for the THI are the most reliable (or more accurately, the least unreliable) for 24-h lead times in the Northeast cluster, followed by P_{MSL}, $T_{2\mathrm{m}}$, and $u_{2\mathrm{m}}$. This bodes well for those that rely on THI forecasts in the energy markets. Note, however, that the small sample sizes for these and all MST RHs in this section limits the significance of the differences between the CvM statistics. Although it is clear that the forecasts for these weather components

TABLE 3. Same as in Table 1, but for the Southwest cluster.

	Los Angeles	San Francisco	San Diego	Fresno	Las Vegas	Sacramento	Reno
P_{MSL} (mb)	0.06	0.08	0.01	0.11	0.61	0.19	0.82
$T_{2\mathrm{m}}$ (°C)	−0.63	−0.15	−0.79	−0.25	−0.95	0.45	−0.69
$u_{10\mathrm{m}}$ (m s^{-1})	1.13	1.31	1.78	1.73	1.53	1.45	1.76
THI (°F)	−4.03	3.85	−0.08	9.63	5.25	1.49	3.80

TABLE 4. Same as in Table 2, but for the Southwest cluster.

	Los Angeles	San Francisco	San Diego	Fresno	Las Vegas	Sacramento	Reno
P_{MSL} (mb)	0.49	0.39	−0.28	0.04	1.58	0.53	2.20
T_{2m} (°C)	−2.00	−1.41	−2.91	−2.17	−4.43	−0.87	−3.13
u_{10m} (m s^{-1})	−3.08	−4.18	−0.70	−2.00	−3.38	−2.32	−1.08
THI (°F)	−6.24	0.02	−5.39	−4.47	−6.78	2.98	3.91

are underdispersed, the relative reliability may change with increased sample sizes.

Differences between Figs. 8 and 9 can be attributed to differences between and limitations of the two choices of verification. Using observations as the verification introduces representativeness errors that reflect the fact that the observations resolve scales that the model does not; a simple interpolation of low-resolution forecast fields to an observation station location is a particularly crude form of downscaling. Geographic areas that tend to generate steep gradients in the forecast component are particularly prone to such representativeness errors. A more fair comparison would be to compare station observations with forecast

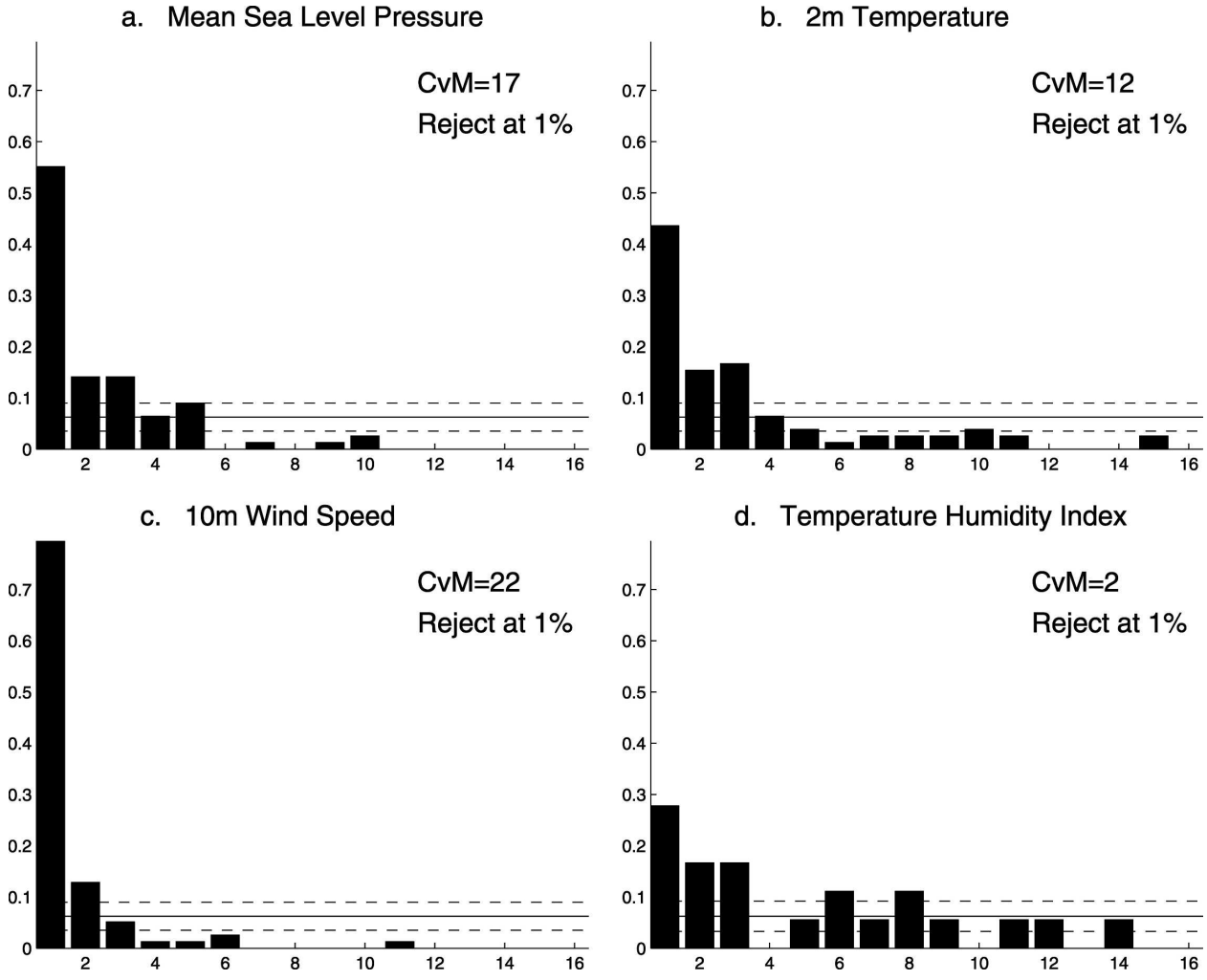

FIG. 8. Mahalanobis-normed and debiased MST RHs for the Northeast cluster for 24-h lead time valid at 0900 UTC. The verification for each increment is taken as a random selection of one of the three SREF control analyses. The dotted lines represent a one standard deviation bound on this expectation. CvM statistics are also included, as well as an assessment of flatness at the 1% significance level. A rejection implies that the histogram is not flat, whereas an acceptance indicates that the histogram is flat.

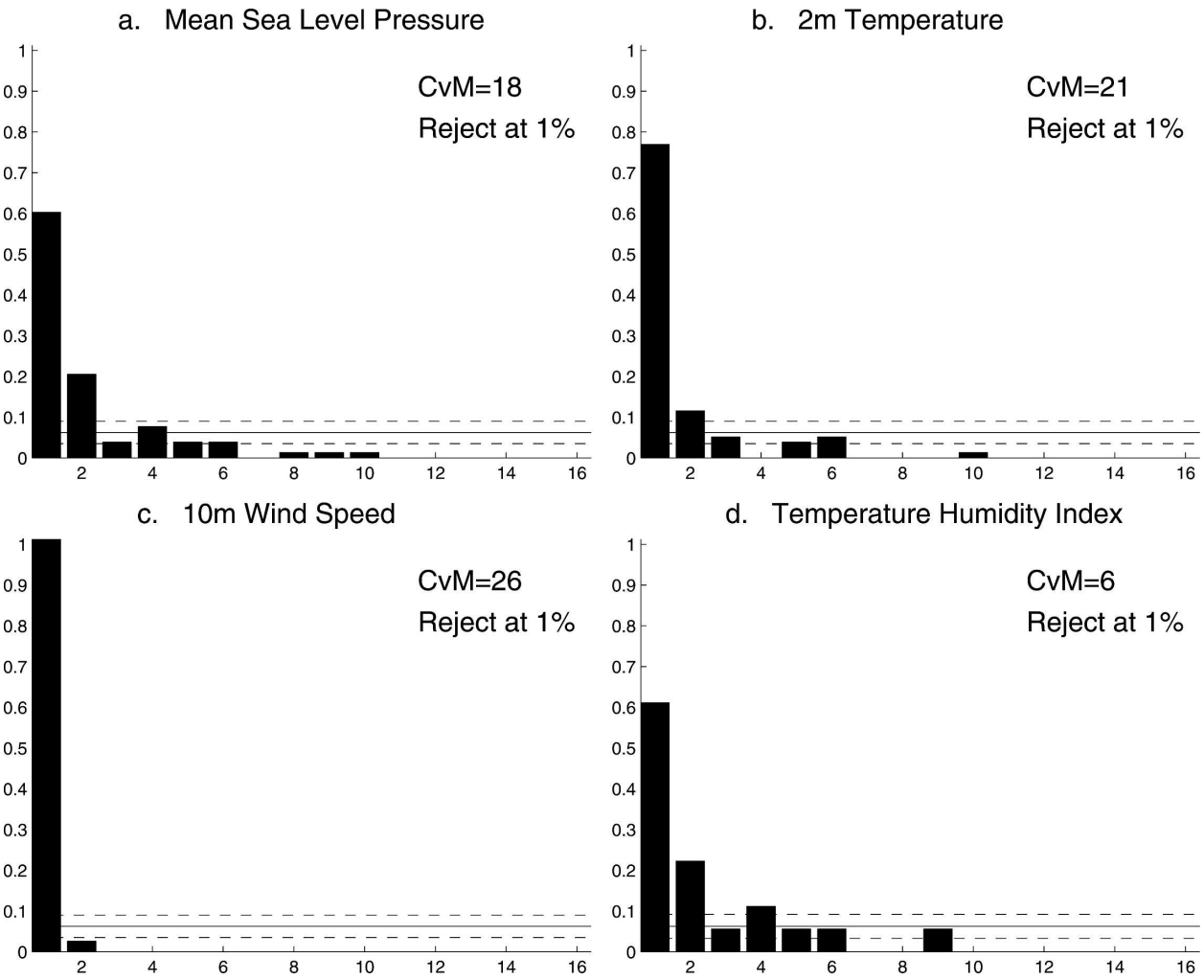

FIG. 9. Same as in Fig. 8, except that the observations are used as the verification.

values that have been mapped from model space into observation space via model output statistics (MOS) or some other form of calibration. Such a comparison is beyond the scope of this work.

Although they mitigate representativeness errors, RHs using the analysis as the verification are subject to forecast dependence errors. Because the control analysis is a weighted combination of a short-term forecast and observations, the analysis and forecast are incestuously dependent, especially at short lead times. For example, by construction, the ensemble control analysis is not an outlier of the ensemble forecast at analysis time; all other ensemble members are perturbations around this control analysis. Therefore, a sufficient lead time is required to ensure that the ensemble can evolve such that the verifying analysis is different from the median of the forecast ensemble (Saetra et al. 2004). Additionally, because analyses are in model space, not observation space, one expects model forecasts to be in some sense "closer" to analyses than to observations, which lie in a completely different space.

Because of the incestuous relationship between ensemble forecasts and the analysis at short lead times, the 24-h Northeast cluster analysis RHs in Fig. 8 are susceptible to forecast dependence errors. However, the observation RHs in Fig. 9 of weather components that can support relatively steep spatial gradients, such as T_{2m}, THI, and particularly u_{2m}, are highly prone to representativeness errors. Therefore, it is difficult to determine which figures' histograms measure the ensemble reliability most accurately. It is the view of the authors that verification in observation space is preferred. Note, however, that the relative similarities of the two P_{MSL} histograms (Figs. 8a and 9a), which are not prone to high representativeness errors, may indicate that representativeness errors have a larger impact

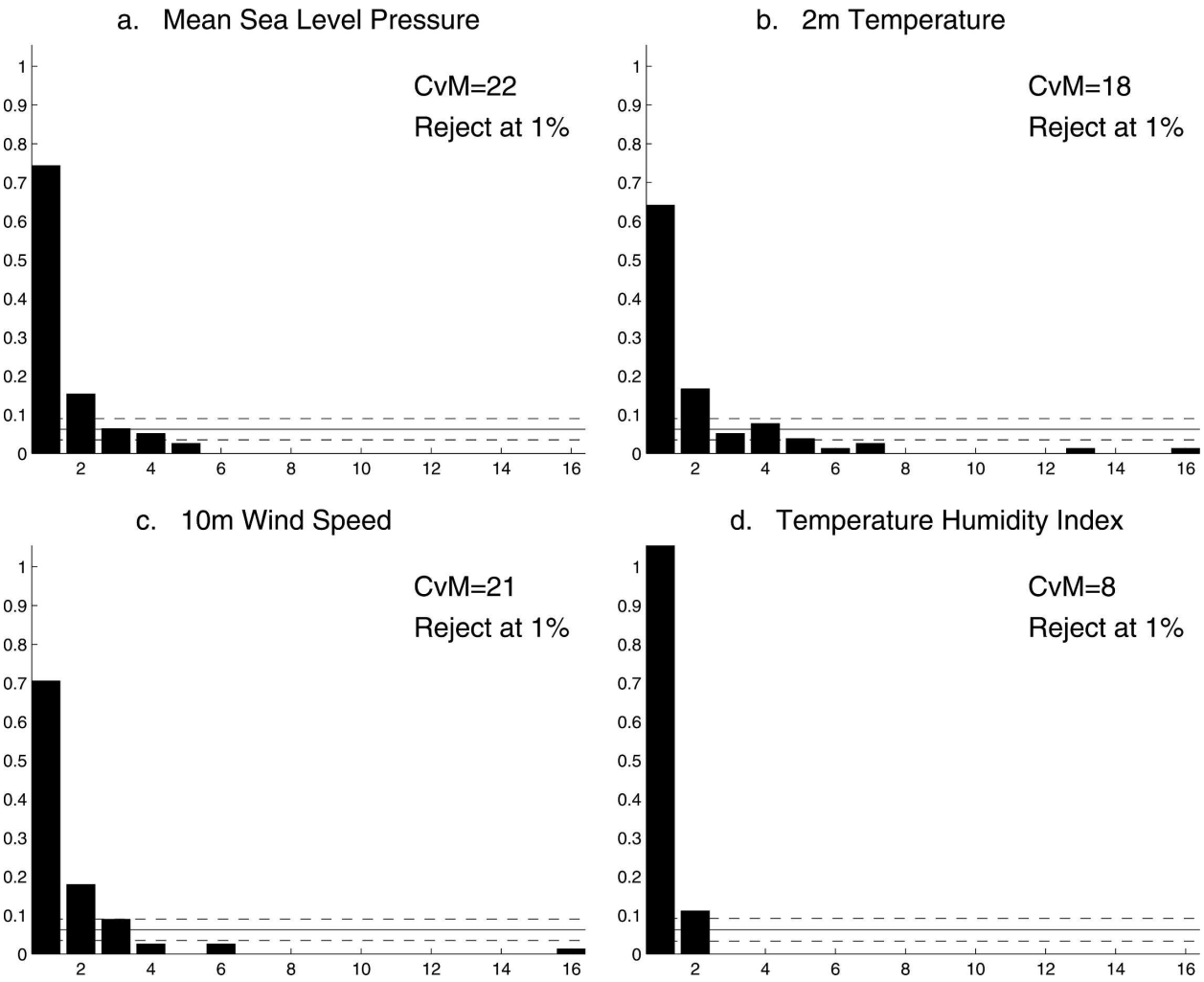

FIG. 10. Same as in Fig. 8, but for the Southwest cluster.

than dependence errors, even at 24-h lead times. Again, the authors reiterate that the preferred method of forecast assessment is to project forecast ensemble members into observation space using some form of calibration and to verify them using station observations.

As can be seen from Figs. 10 and 11, the reliabilities of weather components in the Southwest cluster at 24-h lead times also depend on the type of verification. THI reliability is consistently poor, whereas the poorness of the T_{2m}, u_{2m}, and P_{MSL} reliabilities differs. Because of the greater topographic changes in the Southwest than in the Northeast, we speculate that representativeness errors are more influential in the Southwest cluster RHs than in the Northeast cluster RHs. Topographic channeling effects and valley inversion layers induce high mesoscale wind speed and temperature variability. Because mesoscale P_{MSL} gradients are primarily thermally driven in this region (Zhong et al. 2004), even P_{MSL} histograms are prone to representativeness errors. Therefore, to a greater extent than for the Northeast cluster RHs, the Southwest analysis RHs are likely to be more accurate than the Southwest observation RHs.

Comparing Fig. 10 with Figs. 8 and 9, forecast reliability is generally worse in the Southwest than in the Northeast. This is especially true for the THI, which is particularly foreboding considering the heavy air conditioner usage in this region. Note, however, that the observation RHs of these two clusters are extremely similar, other than that of the THI.

5. Conclusions

The MST RH is an effective multidimensional ensemble reliability assessment tool. After eliminating biases, spatial and temporal correlations, and variance

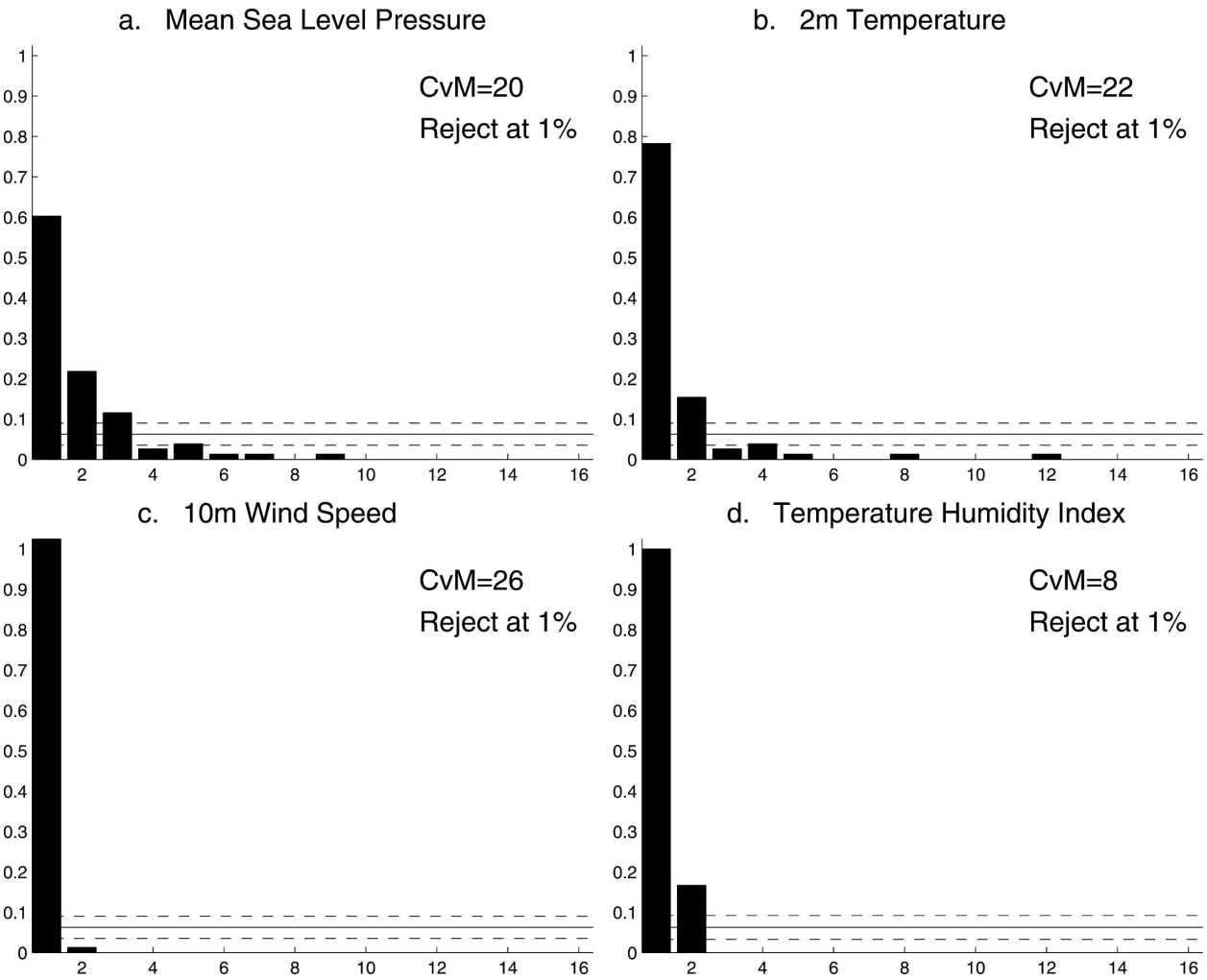

FIG. 11. Same as in Fig. 9, but for the Southwest cluster.

inconsistencies among the K dimensions, the shape of an MST RH can be used to diagnose the relationship between the distribution of the ensemble and of the verification. This information can ultimately help improve forecast reliability through the modification of the ensemble prediction system.

The Mahalanobis norm transforms the forecast data in the most meaningful and interpretable way when the number of ensemble members is greater than the number of forecast locations and/or weather components; this paper advocates the use of the Mahalanobis norm under this circumstance. However, given the misleading results when $n_{\text{ens}} \leq K$, this paper suggests that the variance norm be used when $n_{\text{ens}} \leq K$ and the variances in all dimensions are not identical. The L_2 norm should only be used when the covariance matrix is a scalar multiple of the identity matrix.

Although results are somewhat obscured by verification errors, the analysis of Mahalanobis-normed MST RHs has revealed several important characteristics of the SREF ensemble forecast system. For the components and city clusters analyzed, the right-skewed RHs imply that SREF ensembles are underdispersed at a 24-h lead time. For the Northeast cluster, THI forecasts are the least underdispersed, followed by P_{MSL}, and $u_{10\text{m}}$ forecasts. Depending on the type of verification used, the most reliable weather component forecasts in the Southwest cluster are for $T_{2\text{m}}$, followed by P_{MSL}, $u_{10\text{m}}$, and the THI. Reliability in the Northeast cluster is generally greater than Southwest cluster reliability, especially when the analysis is used as the verification.

It is important to note that absolute uniformity of a reliable RH requires that initial ensemble distributions are correct, and that ensembles be evolved under a perfect forecast model. Since no models of the atmosphere are perfect, RH interpreters must realize that

both model error and initial distribution error will impact the histograms (Smith and Hansen 2004), and that it is not clear how to disentangle these two types of inadequacies.

This paper has presented some preliminary applications of the MST RH. Subsequent studies of the detailed effects of imperfect model scenarios, variable sample sizes, ensemble sizes, dimension sizes, and norm definitions, among others, are needed. Given the multidimensionality of the atmosphere and the need to jointly assess the reliability of these dimensions, the authors feel that the MST RH will evolve into a standard ensemble reliability assessment tool that is available in the toolbox of all ensemble forecasting practitioners.

Acknowledgments. The authors thank Leonard Smith of the University of Oxford and Daniel Wilks of Cornell University whose previous works on the MST RH inspired the writing of this paper. The authors also thank W. Gregory Lawson of the California Institute of Technology and Jonathan Moskaitis of the Massachusetts Institute of Technology for their helpful comments concerning this work. The authors are grateful for the funding provided by NSF Grant ATM-0216866 and the Climate Modeling Initiative.

REFERENCES

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

Du, J., G. DiMego, M. S. Tracton, and B. Zhou, 2003: NCEP Short-Range Ensemble Forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. Research activities in atmospheric and oceanic modeling, Rep. 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/Tech. Doc. 1161, 5.09–5.10.

——, and Coauthors, 2004: The NOAA/NWS/NCEP Short-Range Ensemble Forecast (SREF) system: Evaluation of an initial condition vs. multi-model physics ensemble approach. Preprints, *16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 21.3.

Elmore, K. L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting,* **20,** 789–795.

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

Glickman, T., Ed., 2000: *Glossary of Meteorology*. 2d ed. Amer. Meteor. Soc., 855 pp.

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

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

McQueen, J., J. Du, B. Zhou, G. Manikin, B. Ferrier, H. Chuang, G. DiMego, and Z. Toth, 2005: Recent upgrades to the NCEP Short Range Ensemble Forecasting System (SREF) and future plans. Preprints, *17th Conf. on Numerical Weather Prediction/21st Conf. on Weather Analysis and Forecasting,* Washington, DC, Amer. Meteor. Soc., CD-ROM, 11.2.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

Saetra, O., H. Hersbach, J. R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.,* **132,** 1487–1501.

Smith, L. A., and J. A. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.,* **132,** 1522–1528.

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

Talagrand, O., B. Strauss, and R. Vautard, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25.

Wilks, D. S., 2004: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.,* **132,** 1329–1340.

Zhong, S., C. D. Whiteman, and X. Bian, 2004: Diurnal evolution of three-dimensional wind and temperature structure in California's Central Valley. *J. Appl. Meteor.,* **43,** 1679–1699.