Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds

Tilmann Gneiting[1], Larissa I. Stanberry[1], Eric P. Grimit[2], Leonhard Held[3] and Nicholas A. Johnson[4]

Technical Report no. 537

Department of Statistics, University of Washington

June 2008

Abstract We discuss methods for the evaluation of probabilistic predictions of vector-valued quantities, that can take the form of a discrete forecast ensemble or a density forecast. In particular, we propose a multivariate version of the univariate verification rank histogram or Talagrand diagram that can be used to check the calibration of ensemble forecasts. In the case of density forecasts, Box's density ordinate transform provides an attractive alternative. The multivariate energy score generalizes the continuous ranked probability score. It addresses both calibration and sharpness, and can be used to compare deterministic forecasts, ensemble forecasts and density forecasts, using a single loss function that is proper. An application to the University of Washington mesoscale ensemble points at strengths and deficiencies of probabilistic short-range forecasts of surface wind vectors over the North American Pacific Northwest.

Keywords Calibration · Density forecast · Ensemble postprocessing · Exchangeability · Forecast verification · Probability integral transform · Proper scoring rule · Sharpness · Rank histogram

1 Introduction

One of the major purposes of statistical analysis is to make forecasts for the future, and to provide suitable measures of the uncertainty associated with them. Consequently, forecasts ought to be issued in a probabilistic format, taking the form of probability distributions over future quantities or events (Dawid 1984). Stigler (1975) gives a lucid account of the 19th century transition from point estimation to distribution estimation. Today, we may be witnessing what future generations might

[1] Department of Statistics, University of Washington, Seattle, Washington, USA

[2] 3Tier Environmental Forecast Group, Seattle, Washington, USA

[3] Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik, Universität Zürich, Switzerland

[4] Department of Statistics, Stanford University, Stanford, California, USA

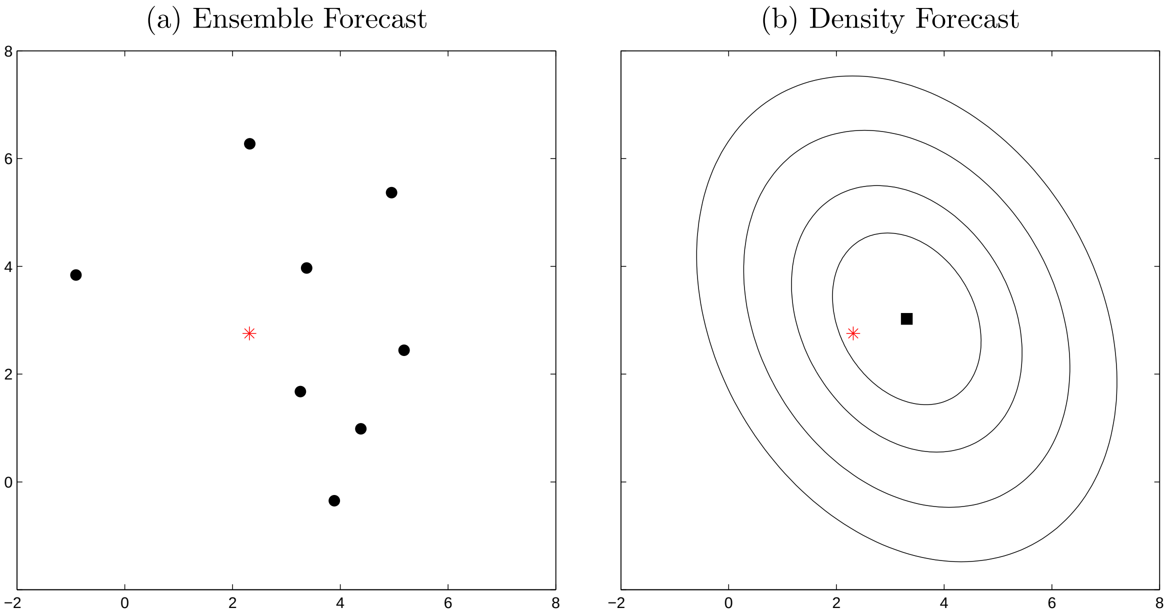

Figure 1: Probabilistic forecasts of the surface wind vector at Olympia Airport, Washington valid October 22, 2003 at 4 pm local time, at a prediction horizon of 24 hours. (a) University of Washington mesoscale ensemble (UW). The eight ensemble member forecasts are shown as black dots, and the verifying wind vector is indicated by the red star. (b) Bivariate normal density forecast fitted to the ensemble values. The black square shows the ensemble mean forecast. The nested regions within the contour lines have probability content 25%, 50%, 75% and 90%, respectively.

refer to as the transition from point prediction to distributional prediction (Gneiting 2008). Since the early 1990s distributional prediction or probabilistic forecasting has become routine in a wealth of applications, including but not limited to weather and climate (Palmer 2002; National Research Council 2006) and economics and finance (Timmermann 2000; Granger 2006). In the statistical literature, advances in Markov chain Monte Carlo methodology (see, for example, Besag, Green, Higdon and Mengersen 1995) have led to explosive growth in the use of predictive distributions, mostly in the form of Monte Carlo samples from posterior predictive distributions of quantities of interest.

With the proliferation of probabilistic forecasting, principled statistical techniques for the evaluation of distributional predictions have been sought. However, the extant literature focuses on predictions of a binary or real-valued, continuous variable (Jolliffe and Stephenson 2003; Pepe 2003; Clements 2005), and there is very little work that applies to more general types of quantities or events, such as multivariate continuous data in $\mathbb{R}^d$. This is the problem that we are addressing here. We focus on the situation in which the dimension is small, say $d = 2$, as in the case of wind vectors. When the dimension of the forecast vector is large, attention typically focuses on low-dimensional functionals, for which the methods discussed here continue to apply.

Figure 1 illustrates the two types of probabilistic forecasts that are commonly employed. Panel (a) shows an ensemble forecast for the surface wind vector at Olympia Airport, Washington valid October 22, 2002 at 4 pm local time. The forecast ensemble has $m = 8$ members and is generated by a regional weather forecasting ensemble that is run in real time at the University of Washington

(Eckel and Mass 2005). The predictive distribution is discrete and assigns mass $1/m$ to each of the ensemble members. Ensemble forecasts have been the preferred format for probabilistic weather and climate predictions (Palmer 2002; Gneiting and Raftery 2005), which typically comprise on the order of 5 to 100 members. Panel (b) illustrates a continuous predictive distribution that takes the form of a density forecast. Density forecasts have been particularly popular in economic and financial applications (Timmermann 2000; Granger 2006), and they are seeing steadily increasing usage in the meteorological community as well.

Arguably, the distinction between ensemble forecasts and density forecasts is artificial, in that we can sample from a predictive density to obtain a forecast ensemble. Conversely, a forecast ensemble can be replaced by a density estimate, particularly, but not exclusively, in cases in which the ensemble size is large. For example, a Markov chain Monte Carlo sample from a posterior predictive distribution can be represented by a kernel density estimate.

Gneiting, Balabdaoui and Raftery (2007) contend that the goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and is a joint property of the predictive distributions and the vector-valued events that materialize. Sharpness refers to the concentration of the predictive distributions, and is a property of the forecasts only: The sharper the distributional forecasts, the less the uncertainty, and the sharper, the better, subject to calibration.

The remainder of the paper is organized as follows. Section 2 turns to calibration checks. In the case of ensemble forecasts, we propose a multivariate version of the univariate verification rank histogram or Talagrand diagram. This can be interpreted within the general framework of tests for exchangeability. In the case of density forecasts, Box's density ordinate transform provides an attractive alternative. Section 3 proposes sharpness measures, and Section 4 reviews the use of proper scoring rules. Proper scoring rules provide omnibus performance measures that address calibration and sharpness simultaneously, by assigning a numerical penalty based on the predictive distribution and the event or value that materializes. The energy score generalizes the univariate continuous ranked probability score and allows for a direct comparison of point forecasts, ensemble forecasts and density forecasts. Section 5 applies these tools in an assessment of probabilistic predictions of surface winds vectors over the North American Pacific Northwest, based on the University of Washington mesoscale ensemble. The paper closes with a discussion in Section 6.

2 Assessing calibration

As noted above, the goal of probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration (Gneiting, Balabdaoui and Raftery 2007). We now consider calibration checks, using the case of a univariate predictand and an ensemble forecast of size m as an initial example. The ordered ensemble values partition the real line into $m + 1$ bins.[5] A minimal assumption on a calibrated ensemble is that the verifying observation be equally likely to fall into any of the bins. Note that this is a much weaker assumption than exchangeability between the ensemble members and the observation, which can be thought of as a strict and possibly unattainable, ideal notion of calibration.

[5] For simplicity, we assume that the m ensemble member values and the verifying observation are pairwise distinct. Ties can be handled easily, using tools discussed by Czado, Gneiting and Held (2007).

The respective diagnostic tool for calibration checks is the Talagrand diagram or rank histogram, proposed in the geophysical literature by Anderson (1996), Hamill and Colucci (1997) and Talagrand, Vautard and Strauss (1997) and extensively used since. Similar devices have been applied in the statistical literature to assess predictive distributions that take the form of Markov chain Monte Carlo samples; see, for example, Shephard (1994, p. 129). Given an ensemble forecast of size m, we find the verification rank, that is, the bin occupied by the verifying observation, which is a number between 1 and $m + 1$. We repeat over a sizable number of individual forecast cases and aggregate ranks. The histogram of these ranks is called a Talagrand diagram or verification rank histogram. Calibration then is assessed diagnostically, by checking for deviations from uniformity. U-shaped histograms indicate underdispersed ensembles with ensemble ranges that are too narrow, hump or inverse U-shaped histograms point at overdispersion. Skewed histograms occur when central tendencies are biased. To quantify the deviation from uniformity in a Talagrand diagram, we use the discrepancy or reliability index

$$\Delta = \sum_{j=1}^{m+1} \left| f_j - \frac{1}{m+1} \right|, \tag{1}$$

where f_j is the observed relative frequency of rank j (Delle Monache, Hacker, Zhou, Deng and Stull 2006; Berrocal, Raftery and Gneiting 2007). Formal tests for uniformity can be developed; they require care in interpretation and have been used in economic as well as meteorological applications (Hamill 2001; Clements 2005).

In the subsequent Section 2.1, we introduce an analogue of the Talagrand diagram that applies to ensemble forecasts of a vector-valued quantity in $\mathbb{R}^d$. We call this tool the multivariate rank histogram. Then in Section 2.2, we compare the multivariate rank histogram to the minimum spanning tree histogram, a related tool that has been proposed and used before (Smith and Hansen 2004; Wilks 2004; Gombos, Hansen, Du and McQueen 2007). Section 2.3 takes a unifying view within the general framework of tests for exchangeability. Section 2.4 turns to calibration checks for density forecasts. We discuss generalizations of the probability integral transform (PIT) histogram (Diebold, Gunther and Tay 1998; Gneiting, Balabdaoui and Raftery 2007), which is the continuous analogue of the Talagrand diagram, and advocate the use of Box's (1980) density ordinate transform. Section 2.5 hints at the marginal calibration diagram, a simple but powerful diagnostic tool for unmasking dispersion errors and forecast biases.

2.1 The multivariate rank histogram

We consider ensemble forecasts for a vector-valued quantity that takes values in $\mathbb{R}^d$. Given vectors $\boldsymbol{x} = (x_1, \ldots, x_d)' \in \mathbb{R}^d$ and $\boldsymbol{y} = (y_1, \ldots, y_d)' \in \mathbb{R}^d$, we write

$$\boldsymbol{x} \preceq \boldsymbol{y} \qquad \text{if and only if} \qquad x_l \leq y_l \quad \text{for} \quad l = 1, \ldots, d. \tag{2}$$

In other words, $\boldsymbol{x} \preceq \boldsymbol{y}$ if and only if $\boldsymbol{x}$ lies in the orthant or hypercube to the left and below $\boldsymbol{y}$.

Let m be the number of ensemble members. To construct a multivariate rank histogram, we compute multivariate ranks as follows, repeat over individual forecast cases, and plot the resulting rank histogram. Here and in the remainder of the article, we use the symbol $\mathbb{I}$ to denote an indicator function.

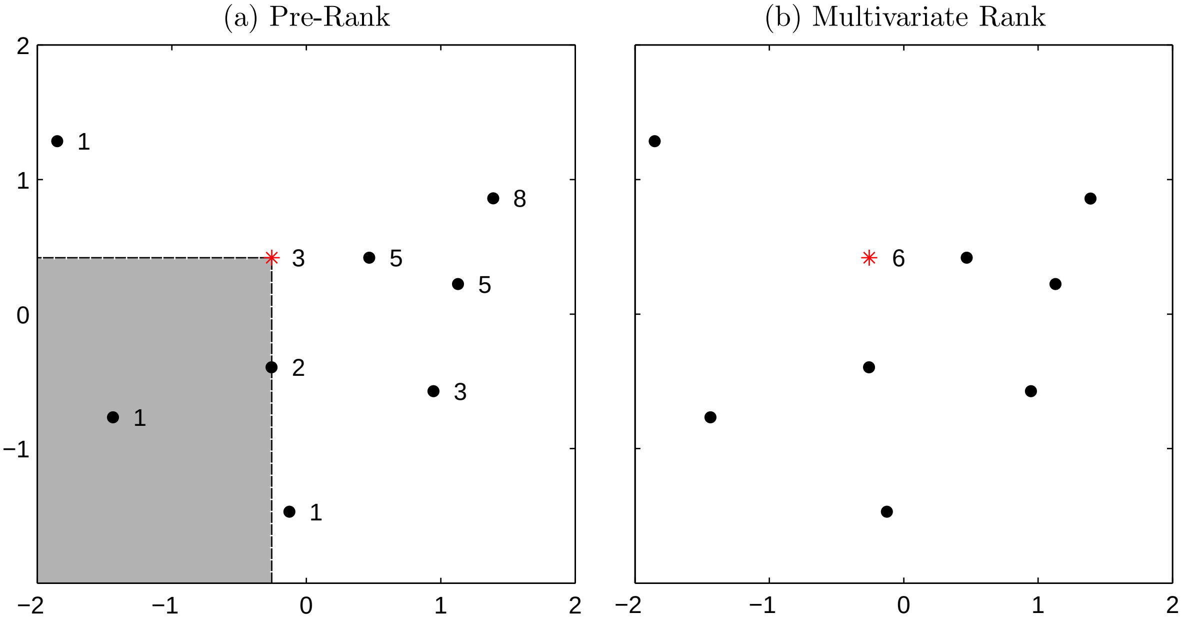

Figure 2: Computation of the multivariate rank for the ensemble forecast in Figure 1. (a) Ensemble member forecasts and verifying observation in standardized coordinates with associated pre-ranks. The observation pre-rank is 3, because three of the nine standardized vectors lie to its lower left, including the standardized observation itself. (b) From panel (a), four vectors have pre-rank ≤ 2, and two vectors have pre-rank 3, that is, $s^< = 4$ and $s^+ = 2$. Hence, the multivariate rank is randomized over the set $\{5, 6\}$. The specific choice here is $r = 6$.

Given an ensemble forecast $\{\boldsymbol{x}_j \in \mathbb{R}^d : j = 1, \ldots, m\}$ and the respective verifying observation $\boldsymbol{x}_0 \in \mathbb{R}^d$:

(a) Standardize (optional, but frequently useful): Apply a principal component transform to the pooled set $\{\boldsymbol{x}_j : j = 0, 1, \ldots, m\}$, to obtain a standardized observation $\boldsymbol{x}_0^*$ and standardized ensemble member forecasts, $\boldsymbol{x}_j^*$, $j = 1, \ldots, m$.

(b) Assign pre-ranks: For $j = 0, 1, \ldots, m$, find the pre-rank,

$$\rho_j = \sum_{k=0}^{m} \mathbb{I}(\boldsymbol{x}_k^* \preceq \boldsymbol{x}_j^*),$$

of $\boldsymbol{x}_j^*$ among the union of the (possibly standardized) observation and the (possibly standardized) ensemble member forecasts. Each pre-rank is an integer between 1 and $m + 1$.

(c) Find the multivariate rank: The multivariate rank, r, is the rank of the observation pre-rank, with ties resolved at random. Specifically, if

$$s^< = \sum_{j=0}^{m} \mathbb{I}(\rho_j < \rho_0) \qquad \text{and} \qquad s^= = \sum_{j=0}^{m} \mathbb{I}(\rho_j = \rho_0),$$

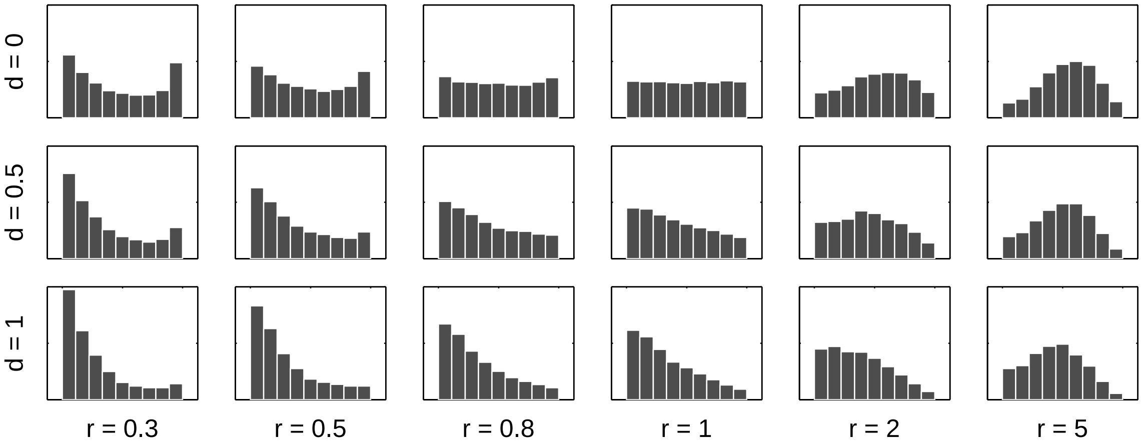

Figure 3: Simulation study for the multivariate rank histogram. The observation is bivariate standard normal. Ensemble forecasts of size $m = 8$ are sampled from a bivariate normal distribution with mean vector $\boldsymbol{\mu}_{\text{ens}} = (d, 0)'$ and covariance matrix $\boldsymbol{\Sigma}_{\text{ens}} = \text{diag}(r, r)$. The rows correspond to distance $d = \|\boldsymbol{\mu}_{\text{ens}}\| = 0$, 0.5 and 1 between the distribution centers, and the columns to variance ratio $r = \sigma^2_{\text{ens}}/\sigma^2_{\text{obs}} = 0.3$, 0.5, 0.8, 1, 2 and 5.

the multivariate rank, r, is chosen from a discrete uniform distribution on the set $\{s^< + 1, \ldots, s^< + s^=\}$. It is an integer between 1 and $m + 1$.

This construction is illustrated in Figure 2. Briefly, we decorrelate and standardize using principal component coordinates. Based on the semi-order (2), we find pre-ranks both for the observation and the ensemble member forecasts. The multivariate rank is the possibly randomized rank of the observation pre-rank, when pooled with the pre-ranks of the ensemble member forecasts. The randomization in part (c) can be avoided, if desired, using a construction proposed by Czado, Gneiting and Held (2007). It is straightforward to see that the multivariate rank is uniform if the ensemble members and the verifying observation are exchangeable. Furthermore, in dimension $d = 1$ the multivariate rank reduces to the familiar univariate verification rank.

The multivariate rank histogram is a plot of the empirical frequency of the multivariate ranks. Underdispersed ensembles lead to U-shaped, overdispersed ensembles to hump-shaped, and biased ensembles to skewed rank histograms, as in the case of the Talagrand diagram (Hamill 2001; Gneiting, Balabdaoui and Raftery 2007). This is illustrated in the simulation study in Figure 3, which uses ensembles of size $m = 8$. For each panel, 10,000 ensemble forecasts were sampled.

2.2 The minimum spanning tree (MST) rank histogram

We now discuss the minimum spanning tree (MST) rank histogram, which is another diagnostic tool for checking the calibration of multivariate ensemble forecasts. It was proposed by Smith (2001) and has been studied by Smith and Hansen (2004), Wilks (2004) and Gombos et al. (2007).

Its construction is similar to that of the multivariate rank histogram, with steps (b) and (c) replaced by (b') and (c'), as follows.

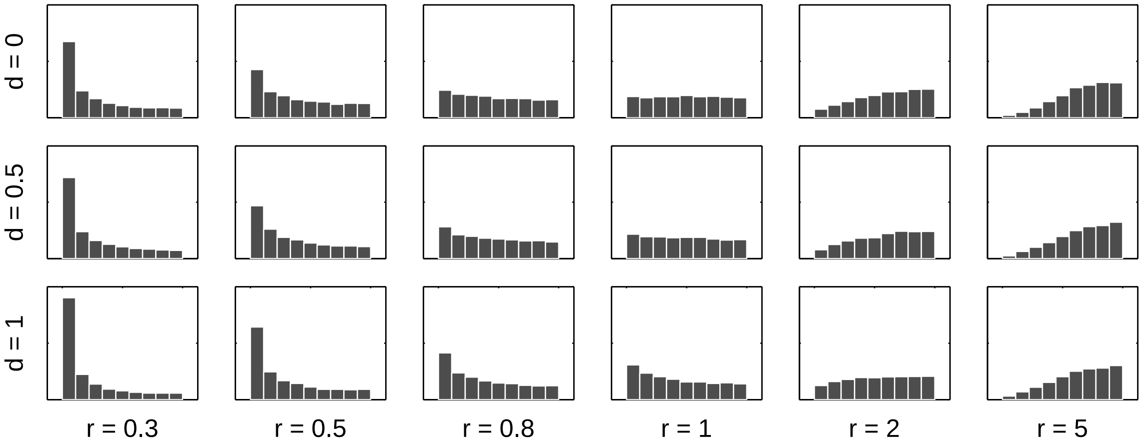

Figure 4: Simulation study for the minimum spanning tree (MST) rank histogram. The observation is bivariate standard normal. Ensemble forecasts of size $m = 8$ are sampled from a bivariate normal distribution with mean vector $\boldsymbol{\mu}_{\text{ens}} = (d, 0)'$ and covariance matrix $\boldsymbol{\Sigma}_{\text{ens}} = \text{diag}(r, r)$. The rows correspond to distance $d = \|\boldsymbol{\mu}_{\text{ens}}\| = 0$, 0.5 and 1 between the distribution centers, and the columns to variance ratio $r = \sigma^2_{\text{ens}}/\sigma^2_{\text{obs}} = 0.3$, 0.5, 0.8, 1, 2 and 5.

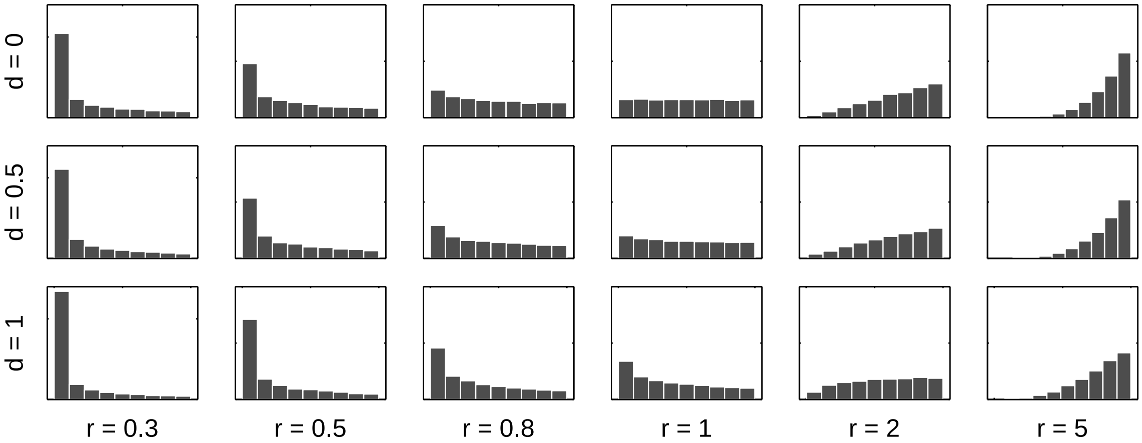

Figure 5: Simulation study for the Box density ordinate transform (BOT) histogram. The observation is bivariate standard normal. The density forecast is bivariate normal with mean vector $\boldsymbol{\mu}_{\text{fcst}} = (d, 0)'$ and covariance matrix $\boldsymbol{\Sigma}_{\text{fcst}} = \text{diag}(r, r)$. The rows correspond to distance $d = \|\boldsymbol{\mu}_{\text{fcst}}\| = 0$, 0.5 and 1 between the distribution centers, and the columns to variance ratio $r = \sigma^2_{\text{fcst}}/\sigma^2_{\text{obs}} = 0.3$, 0.5, 0.8, 1, 2 and 5.

(b') Compute minimum spanning trees: For $j = 0, 1, \ldots, m$, find the minimum spanning tree (MST) of the set $\{\boldsymbol{x}_k^* : k \in \{0, 1, \ldots, m\} \setminus \{j\}\}$ and its length, $l_j > 0$.

(c') Find MST rank: The MST rank, r, is the rank of l_0 within the pooled sample of MST lengths, with any ties resolved at random. Specifically, if

$$s^< = \sum_{j=0}^{m} \mathbb{I}(l_j < l_0) \qquad \text{and} \qquad s^= = \sum_{j=0}^{m} \mathbb{I}(l_j = l_0),$$

the MST rank, r, is chosen from a uniform distribution on the set $\{s^< + 1, \ldots, s^< + s^=\}$. It is an integer between 1 and $m + 1$.

In the applications below, we omit the standardization, and use the Euclidean distance in constructing MSTs.[6] Given a set of m points in $\mathbb{R}^d$, a spanning tree is a collection of $m - 1$ edges such that all points are used. The spanning tree with the smallest length is the minimum spanning tree (MST, Kruskal 1956). Briefly, given an ensemble forecast, we find the MST rank by tallying the length of the MST that connects the m ensemble members within the combined set of the $m+1$ lengths of the ensemble-only MST and the m MSTs obtained by substituting the observation for each of the ensemble members. Ties are rare, and randomization typically can be avoided. If the ensemble members and the observation are exchangeable, the lengths are exchangeable and the MST rank is uniform.

The respective rank histogram plots the empirical frequency of the MST ranks. Deviations from uniformity can be interpreted diagnostically, but the interpretation differs from that of the Talagrand diagram, and the construction fails in dimension $d = 1$. For an underdispersed or biased ensemble, the lowest MST ranks are overpopulated; for an overdispersed ensemble, the highest ranks occur too often. This is illustrated in the simulation study in Figure 4. For each panel, 10,000 ensemble forecasts were sampled.

Viewed slightly differently, the MST rank provides a center-outward ordering for the combined set of ensemble members and the observation, similarly to the way in which statistical depth functions operate (Zuo and Serfling 2000). Oja and Randles (2004, Section 2.3) describe other approaches towards multivariate ranks. Another possible connection to multivariate analysis is via the Friedman and Rafsky (1979) test, which employs MSTs of pooled sample points. Similarly, multivariate ranks relate to the multi-dimensional Smirnov two sample test (Bickel 1969).

2.3 Tests for exchangeability

The above tools can be subsumed and interpreted within the general framework of tests for exchangeability. Suppose that the function

$$F : \mathbb{R}^d \times \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{m \text{ times}} \longrightarrow \mathbb{R}$$

[6] Of course, other choices are possible and sometimes essential (Gombos et al. 2007). The MST rank histogram is based on distances and therefore depends strongly on the units used. For the multivariate rank histogram, this dependency is much less pronounced.

is invariant under permutations of its final m arguments. Given an ensemble forecast $\boldsymbol{x}_1, \ldots \boldsymbol{x}_m \in \mathbb{R}^d$ and the realizing observation $\boldsymbol{x}_0 \in \mathbb{R}^d$, let

$$z_j = F(\boldsymbol{x}_j; \boldsymbol{x}_{-j}), \qquad j = 0, 1, \ldots, m,$$

where $\boldsymbol{x}_{-j}$ denotes the set $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \setminus \{\boldsymbol{x}_j\}$. The simplified specification with a set-valued second argument is justified by the invariance property of the function F. If the observation and the ensemble members are exchangeable, the (possibly randomized) rank of z_0, computed with the observation in the first argument, among the pooled set $\{z_0, z_1, \ldots, z_m\}$ is uniform on $\{1, \ldots, m+1\}$. To assess exchangeability, we compute a rank for each individual forecast case, collect ranks, and check the resulting rank histogram for uniformity.

If the function F is a coordinate projection for the first argument, this approach yields the univariate Talagrand diagram. If we use standardized values and let

$$F(\boldsymbol{x}_j; \boldsymbol{x}_{-j}) = \sum_{k=0}^{m} \mathbb{I}(\boldsymbol{x}_k \preceq \boldsymbol{x}_j),$$

we recover the multivariate rank histogram. If the value of F is the length of the MST for the second set-valued argument, the MST rank histogram emerges. There are many other choices for F, such as linear combinations of coordinate projections of the first argument, or multivariate measures of spread applied to the second argument, and there is no obvious answer to the question which option one might prefer in applications.

2.4 Calibration checks for density forecasts

In the case of a density forecast, calibration requires that the realizing observation is indistinguishable from a random draw from the predictive density. This can be interpreted in terms of exchangeability, in that the members of a simple random sample drawn from the predictive distribution and the verifying observation are exchangeable. The above tools apply if we generate an ensemble forecast by sampling from the predictive density.

However, there are some other tools available, which are based on the predictive density itself or a function thereof. In the univariate case, calibration can be assessed using the probability integral transform (PIT). This is simply the value that the predictive cumulative distribution function (CDF) attains at the observation. If the observation is drawn from the predictive distribution, which is an ideal and desirable situation, and the predictive distribution is continuous, the PIT has a uniform distribution on the unit interval $[0, 1]$. Calibration then is checked empirically, by plotting the histogram of the PIT values and checking for uniformity (Dawid 1984; Diebold, Gunther and Tay 1998; Gneiting, Balabdaoui and Raftery 2007). The PIT histogram is typically used informally as a diagnostic tool and can be interpreted in the same way as its discrete analogue, the Talagrand diagram. In the multivariate case, one possible approach is to consider PIT histograms for projections and scan for 'interesting' (that is, non-uniform) directions (Ishida 2005), perhaps similarly to projection pursuit algorithms (Huber 1985).

In what follows, we focus on genuinely multivariate approaches. The most direct multivariate analogue of the PIT fails in that it is not uniform, even if the observation is drawn from the predictive distribution (Genest and Rivest 2001). However, a slightly less direct transform which was proposed by Diebold, Hahn and Tay (1999) remains uniform. This is a stepwise procedure, in which the univariate PIT is first applied to the CDF of the first component, then to the conditional CDF of the second component given the first, and so on. The resulting PIT is uniform on the unit hypercube in $\mathbb{R}^d$ (Rosenblatt 1952; Brockwell 2007). In practice, the method involves checks for uniformity in $\mathbb{R}^d$, and it depends on the ordering of the components. Clements and Smith (2000, 2002) and De Gooijer (2007) studied a variant for density forecasts in $\mathbb{R}^2$ which is based on the product (or the ratio) of the first (unconditional) and the second (conditional) PIT value.

The Box density ordinate transform (BOT) was proposed by Box (1980) and O'Hagan (2003) as a model check tool. If the predictive density for a future quantity is p and $\boldsymbol{x}_0$ materializes, our version of the BOT is defined as

$$u = 1 - \mathrm{pr}\,(p(\boldsymbol{X}) \leq p(\boldsymbol{x}_0)),$$

where $\boldsymbol{X}$ is a random vector with density p. For example, if p is a d-variate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then

$$u = 1 - \chi^2_d((\boldsymbol{x}_0 - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}))$$

equals one minus the CDF of a chi-square distribution with d degrees of freedom when evaluated at the standardized observation (Box 1980). It is easy to see that if $p(\boldsymbol{X})$ has a continuous distribution and $\boldsymbol{x}_0$ is distributed according to p, then u is standard uniform. If $p(\boldsymbol{X})$ has discrete components, a suitably randomized version of the BOT remains uniform (cf. Brockwell 2007; Czado, Gneiting and Held 2007), but power is reduced.

Figure 5 illustrates the behavior of the BOT in a simulation study which uses the same scenario as that in Figures 3 and 4, except that density forecasts are issued. Like the MST rank, the BOT provides a center-outward ordering, in that outlying observations tend to lead to low values, and inlying observation to high values. Consequently, the results are similar to those for the MST rank and the histograms can be interpreted analogously.

2.5 Marginal calibration

We now consider what Gneiting, Balabdaoui and Raftery (2007) referred to as marginal calibration. The underlying idea is very simple. If we assume that the ensemble members and the realizing observation are exchangeable, and composite over forecast cases in a steady state, we expect the empirical distribution of the ensemble values and the empirical distribution of the realizations to be statistically compatible. In the case of a density forecast, the ensemble corresponds to a random sample from the predictive density. A marginal calibration diagram plots and compares the two empirical distributions, and can be used diagnostically to unmask dispersion errors and forecast biases. In the case study below we show and discuss a simplified example, which displays ensemble mean forecasts and realizing observations of surface wind vectors.

3 Assessing sharpness

The term sharpness refers to the concentration of the predictive distributions, which is a property of the forecasts only. There is a strong dependence on the units used and components that are incommensurable, or incomparable in magnitude, call for standardization. As noted above, the sharper the probabilistic forecast, the less the uncertainty, and the sharper, the better, subject to calibration.

In the case of an ensemble forecast for a univariate quantity, this can be paraphrased simply: The smaller the ensemble spread, the sharper, and the sharper, the better, subject to proper coverage.[7] To quantify ensemble spread, one typically uses the ensemble range or the ensemble standard deviation. In the multivariate case, various measures have been proposed to quantify the sharpness of an ensemble forecast, such as the volume of its convex hull or bounding box (Weisheimer, Smith and Judd 2005; Judd, Smith and Weisheimer 2007), or the root mean squared Euclidean distance between the ensemble members and the ensemble mean vector (Stephenson and Doblas-Reyes 2000).

In the statistical literature, measures of scatter and spread have been discussed by Bickel and Lehmann (1979), Oja (1983) and Shaked and Shanthikumar 1994), among others. Our preferred measure is the determinant sharpness

$$\mathrm{DS} = (\det \mathbf{\Sigma})^{1/(2d)}, \tag{3}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ is the covariance matrix of an ensemble or density forecast for a quantity in $\mathbb{R}^d$. This generalizes the univariate standard deviation and applies to ensembles of size $m > d$ as well as to density forecasts, assuming that the predictive density has finite second moments.

4 Proper scoring rules as omnibus performance measures

Scoring rules provide summary measures in the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. We take scoring rules to be negatively oriented penalties that a forecaster wishes to minimize. Specifically, if the forecaster quotes the predictive distribution P and $\boldsymbol{x} \in \mathbb{R}^d$ materializes, the penalty is $s(P, \boldsymbol{x})$. We write $s(P, Q)$ for the expected value of $s(P, \boldsymbol{X})$ when $\boldsymbol{X} \sim Q$. In practice, scores are reported as averages over comparable sets of probabilistic forecasts, and we use upper case to denote a mean score, say

$$S_n = \frac{1}{n} \sum_{i=1}^{n} s(P_i, \boldsymbol{x}_i),$$

where P_i and $\boldsymbol{x}_i$ refer to forecast case $i = 1, \ldots, n$. Mean scores are used to rank and compare competing forecasting techniques informally or via tests (Diebold and Mariano 1995; Jolliffe 2007).

4.1 Propriety

In keeping with a landmark Test discussion paper (Winkler 1996), we stress the importance of propriety. A proper scoring rule is designed such that it does not provide any incentive to the forecaster to digress from her true beliefs, in the following sense (Savage 1971; Bröcker and Smith 2007; Gneiting and Raftery 2007). Suppose that the forecaster's best judgement is the predictive distribution Q. The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote her true belief, $P = Q$, if

$$s(Q, Q) \leq s(P, Q)$$

with equality if and only if $P = Q$. A scoring rule with this property is said to be strictly proper. If $s(Q, Q) \leq s(P, Q)$ for all P and Q, the scoring rule is said to be proper. Propriety is an essential

[7] The range of an m-member ensemble provides a prediction interval with nominal coverage $(m - 1)/(m + 1)$.

property of a scoring rule, ensuring that it addresses calibration and sharpness simultaneously (Winkler 1977, 1996).

4.2 The energy score as a multivariate generalization of the continuous ranked probability score

The continuous ranked probability score is arguably the most versatile scoring rule for probabilistic forecasts of a univariate scalar variable. It is defined as

$$\begin{aligned} \mathrm{crps}(P, x) &= \int_{-\infty}^{\infty} (F(y) - \mathbb{I}\{y \geq x\})^2 \, \mathrm{d}y && (4) \\ &= E_P|X - x| - \frac{1}{2} E_P|X - X'|, && (5) \end{aligned}$$

where F is the CDF associated with the predictive distribution P, and X and X' are independent random variables with distribution P. This scoring rule can be traced back to Matheson and Winkler (1976) and is increasingly being used in the meteorological community (Hersbach 2000; Candille and Talagrand 2005; Wilks 2006, p. 252). Gneiting and Raftery (2007) showed the equality of the standard form (4) and the kernel score representation (5). Grimit, Gneiting, Berrocal and Johnson (2006) reviewed properties and uses of the continuous ranked probability score and introduced an analogue for circular variables.

To assess probabilistic forecasts of a multivariate quantity, we propose the use of the energy score

$$\mathrm{es}(P, \boldsymbol{x}) = E_P\|\boldsymbol{X} - \boldsymbol{x}\| - \frac{1}{2} E_P\big\|\boldsymbol{X} - \boldsymbol{X}'\big\|, \tag{6}$$

where $\| \cdot \|$ denotes the Euclidean norm and $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors with distribution P. This is a direct generalization of the continuous ranked probability score (5), to which the energy score reduces in dimension $d = 1$. Gneiting and Raftery (2007) showed its propriety and noted a number of generalizations that include non-Euclidean versions.

If $P = P_{\mathrm{ens}}$ is an ensemble forecast of size m, the evaluation of the energy score is straightforward. The predictive distribution P_{ens} places point mass $1/m$ on the ensemble members $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d$, and (6) reduces to

$$\mathrm{es}(P_{\mathrm{ens}}, \boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^{m} \|\boldsymbol{x}_j - \boldsymbol{x}\| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{7}$$

If $P = \delta_{\boldsymbol{\mu}}$ is the point measure in $\boldsymbol{\mu} \in \mathbb{R}^d$, that is, a deterministic forecast, the energy score reduces to the Euclidean norm of the error vector, in that

$$\mathrm{es}(\delta_{\boldsymbol{\mu}}, \boldsymbol{x}) = \|\boldsymbol{\mu} - \boldsymbol{x}\|. \tag{8}$$

Thus, the energy score provides a direct way of comparing deterministic forecasts, discrete ensemble forecasts and density forecasts using a single metric that is proper. If closed form expressions for the expectations in (6) are unavailable, as is often the case for density forecasts, we employ Monte

Carlo methods.[8]

The Euclidean version of the energy score does not make any distinction between the components of the forecast vector. Hence, if the components are incommensurable, or incomparable in magnitude, a standardization may be advisable or necessary, depending on the application at hand. In our case study on wind vectors the components are alike, and we work with the original non-transformed, east-west and north-south wind components. The energy score and the Euclidean error then are in the unit of meters per second.[9]

4.3 Scores for density forecasts

If the predictive distribution P for a multivariate quantity has density function p, some other types of proper scoring rules become available.

Perhaps the most widely known is the logarithmic score, which is defined as

$$\mathrm{logs}(P, \boldsymbol{x}) = -\log p(\boldsymbol{x}).$$

Allowing for affine transformations, this is the only 'local' proper scoring rule, in the sense that it depends on the predictive distribution only through the value $p(\boldsymbol{x})$ that the predictive density attains at the observation (Bernardo 1979). The logarithmic score has many appealing properties (Bröcker and Smith 2007; Gneiting and Raftery 2007) but can incur practical difficulties, such as infinite scores.

Good (1971) and Matheson and Winkler (1976) studied the quadratic score and the spherical score, defined as

$$\mathrm{qs}(p, \boldsymbol{x}) = -\,2\,p(\boldsymbol{x}) + \|p\|^2 \qquad \text{and} \qquad \mathrm{sphs}(P, \boldsymbol{x}) = -\,\frac{p(\boldsymbol{x})}{\|p\|},$$

respectively, where $\|p\|^2 = \int (\,p(\boldsymbol{y}))^2\,\mathrm{d}\boldsymbol{y}$. For example, if p is a bivariate normal density with component standard deviations σ_1 and σ_2 and correlation coefficient ρ, then

$$\|p\|^2 = \frac{1}{4\pi\,\sigma_1\sigma_2\,(1-\rho^2)^{1/2}}.$$

Dawid and Sebastiani (1999) studied proper scoring rules that depend on the predictive distribution only through its first and second (joint) moments. Like the energy score, they can be used both for ensemble forecasts and density forecasts. In contrast, the logarithmic, quadratic and spherical scores apply to density forecasts only.

[8] When computing the energy score for bivariate normal or related predictive densities, we replace (6) by the computationally efficient Monte Carlo approximation

$$\widehat{\mathrm{es}}(P, \boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{x}_i - \boldsymbol{x}\| - \frac{1}{2\,(k-1)}\sum_{i=1}^{k-1}\|\boldsymbol{x}_i - \boldsymbol{x}_{i+1}\|,$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is a simple random sample of size $k = 10,000$ from the predictive density.

[9] Suppose that $\boldsymbol{\mu}$ is a point forecast and the vector $\boldsymbol{x}$ realizes. The Euclidean error $\|\boldsymbol{\mu} - \boldsymbol{x}\|$ provides a measure of deterministic forecast performance that is appealing, because it can be interpreted as the special case (8) of the energy score. Alternatively, the squared Euclidean error $\|\boldsymbol{\mu} - \boldsymbol{x}\|^2$ can be used to assess point predictions of a vector-valued quantity; see, for example, Malmberg, Holst and Holst (2008). The squared Euclidean error has a desirable theoretical property, in that the predictive mean is the point forecast that minimizes the expected squared Euclidean error. In future work, we intend to address these issues in more detail.

There is no automatic choice of a proper scoring rule to be used in any given situation, unless there is a unique and well defined underlying decision problem. In many types of applications probabilistic forecasts have multiple simultaneous uses, and it may be appropriate to use a variety of diagnostic tools and scores, to take advantage of their distinct emphases and strengths.

5 Case study: Probabilistic forecasts of surface wind vectors

For several reasons, surface wind is an important meteorological variable to consider when undertaking forecast verification studies on regional scales.[10] The assessment of surface wind predictions addresses a central purpose of limited-area numerical weather prediction models, namely their ability to simulate local circulations (Rife and Davis 2005). Prior work has addressed the evaluation of probabilistic forecasts of wind speed (Gneiting et al. 2006; Gneiting, Balabdaoui and Raftery 2007) or wind direction (Grimit et al. 2006), but not of wind vectors. This omission is addressed in the subsequent case study.

5.1 University of Washington mesoscale ensemble

Over the past decade, a short-range ensemble weather forecast system for the North American Pacific Northwest has been under development at the University of Washington, with the goal of producing calibrated probabilistic forecasts for weather parameters at regional scales. The original configuration was that of a five member ensemble using the MM5 numerical weather prediction model and a nested, limited-area grid configuration that centers on the states of Washington and Oregon (Grimit and Mass 2002). The horizontal grid spacing of the inner nest was 12 km and 33 levels were used in the vertical. Large-scale analyses from several operational forecast centers provided the necessary initial conditions and lateral boundary conditions. Beginning in the autumn of 2002, the size of the ensemble was increased to eight members using additional global analyses and forecasts and named the University of Washington (UW) mesoscale ensemble (Eckel and Mass 2005). The ensemble is run to up to three days lead time beginning at 4:00 pm local time each day. Real-time forecast products over the Pacific Northwest are available online at `http://www.atmos.washington.edu/~ens/uwme.cgi`.

The evaluation period of this study begins on 31 October 2002 and extends through 31 March 2004. This period encompasses two cool seasons (October–March) during 2002–03 and 2003–04 and one warm season (April–September) during 2003. Station-based observations of surface wind were acquired in real-time from over two dozen networks operated by local, state, federal and foreign agencies as well as a few independent organizations (Mass et al. 2003). All together, the mesoscale observation network includes approximately 800 wind observing locations scattered through the Pacific Northwest. We restrict our verification study to the sites that had more than 300 observations during the evaluation period. There were 565 such stations, with a minimum of 301, a median of 382 and a maximum of 425 observations. Model surface wind component forecasts at the four grid-box centers surrounding each station were bi-linearly interpolated to the observation location and then rotated from grid-relative to north-relative. Forecasts and observations represent two-minute averages of wind components between 3:58 pm and 4:00 pm local time.

[10] Meteorological stations typically measure surface winds at a height of 10 m. Throughout the paper, we refer to 10 m winds as surface or near-surface winds.

Table 1: Probabilistic forecasting methods in our case study. All three methods are employed in ensemble forecast and density forecast types.

Probabilistic Forecasting Method	Acronym
University of Washington mesoscale ensemble	UW
Error dressing	ED
Climatology	CLI

Several probabilistic forecasting methods for surface wind vectors are now assessed, including the raw ensemble forecast, postprocessed ensemble forecasts and climatological forecasts, and using both ensemble forecast and density forecast types.

5.2 Ensemble forecasts

We first describe our ensemble forecasting methods, which are based on the approaches listed in Table 1. In doing so, we let $\boldsymbol{x}_{ij}$ denote the verifying wind vector at site i on day j.

The **University of Washington (UW)** ensemble is the aforementioned eight-member, standard University of Washington mesoscale ensemble described by Eckel and Mass (2005). The respective ensemble mean (point) forecast for the wind vector at station i on day j is

$$\boldsymbol{\mu}_{ij}^{\mathrm{UW}} = \frac{1}{8} \sum_{k=1}^{8} \boldsymbol{\mu}_{ijk}^{\mathrm{UW}}, \tag{9}$$

where the average extends over the ensemble members $\boldsymbol{\mu}_{ij1}^{\mathrm{UW}}, \ldots, \boldsymbol{\mu}_{ij8}^{\mathrm{UW}}$.

The **error dressing (ED)** ensemble combines the UW ensemble with statistical information, by adding historical error vectors to the station and site specific UW ensemble mean forecast. Roulston and Smith (2003) coined the term ‘dressing’ for this general approach. To create an ED ensemble forecast with m members that is valid at site i and day j, we sample errors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ from the empirical distribution of local UW ensemble mean errors $\boldsymbol{e}_{ij'} = \boldsymbol{x}_{ij'} - \boldsymbol{\mu}_{ij'}^{\mathrm{UW}}$, where the index j' ranges over all available dates within the evaluation period.[11] We restrict attention to sites with more than 300 forecast cases, so at each site i, at least 300 such error vectors are available. The wind vectors $\boldsymbol{\mu}_{ij}^{\mathrm{UW}} + \boldsymbol{e}_1, \ldots, \boldsymbol{\mu}_{ij}^{\mathrm{UW}} + \boldsymbol{e}_m$ then form the ED ensemble. This is a straightforward statistical postprocessing technique that addresses both forecast biases and dispersion errors. We apply it in a fashion that is akin to cross-validation and tailored to a methodological case study. Of course, real-time forecasting requires the sampling of past errors only.

The **climatological (CLI)** ensemble at site i samples m vectors from the wind observations $\boldsymbol{x}_{ik}$, where the index k ranges over the site-specific observation dates. We stress, once again, that our goal is an initial assessment of raw, statistically postprocessed and climatological ensembles in a methodological pilot study. Real-time forecasting calls for the use of more sophisticated techniques, and our verification results provide guidance in this direction.

A summary of the wind vector forecast performance and an overall comparison of the three forecast techniques is conveyed in Table 2 and Figure 6. All results are temporally and spatially

[11] For simplicity, we do not exclude the error at day $j' = j$.

Table 2: Mean energy score (ES) and mean Euclidean error (EE) in $\mathrm{m} \cdot \mathrm{s}^{-1}$, multivariate rank histogram discrepancy (Δ) and mean determinant sharpness (DS) for ensemble forecasts of surface winds over the Pacific Northwest, aggregated over stations with more than 300 available cases.

Ensemble Forecast	ES	EE	Δ	DS
UW	2.95	3.41	0.64	0.54
ED	2.32	3.29	0.04	6.95
CLI	2.09	3.01	0.12	6.29

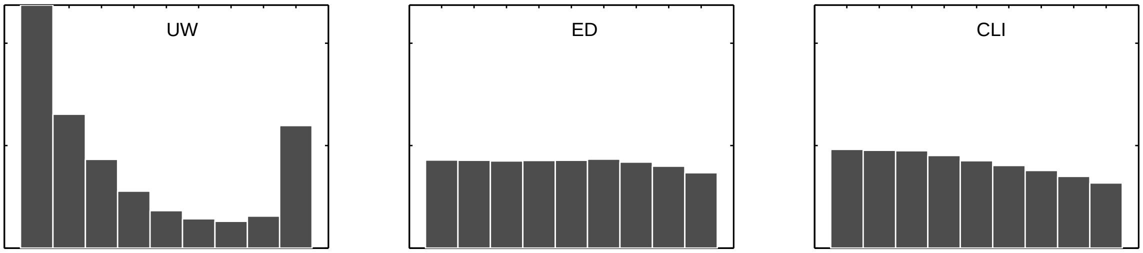

Figure 6: Multivariate rank histograms for ensemble forecasts of surface winds over the Pacific Northwest, aggregated over stations with more than 300 available cases. The respective discrepancy measure Δ is shown in Table 2.

aggregated, using the 565 stations with more than 300 forecast cases during the evaluation period. To facilitate the comparison, the multivariate rank histograms are based on ensembles with $m = 8$ members. For the ED and CLI ensembles, these 8 members are picked at random.

The mean energy score (7) serves as omnibus performance measure for the ensemble forecasts. We also show the mean Euclidean error (8) for the respective ensemble mean forecast, which is a summary measure of deterministic forecast performance. Not surprisingly, the postprocessed ED technique outperforms the raw UW forecast. The CLI ensemble performs best, despite ignoring information from numerical weather prediction models. How can this be explained? The multivariate rank histograms and the respective discrepancy measure (1) show that the UW ensemble is severely underdispersed, while the ED and CLI ensembles are calibrated. The determinant sharpness measure (3) indicates that on average the CLI ensemble is sharper than the ED ensemble. In keeping with the principle of maximizing sharpness subject to calibration, the CLI ensemble outperforms the ED ensemble. The UW ensemble is by far the sharpest; however, it is uncalibrated, resulting in a deterioration of the energy score.

For a more detailed discussion we refer to Section 5.4 below, where we discuss results at individual stations. Briefly, the ED ensemble tends to perform best if the terrain at a station is uniform. The CLI ensemble outperforms the others when winds at a station are heavily influenced by the surrounding terrain, resulting in local effects that the 12 km grid of the numerical weather prediction model cannot resolve. The latter situation is more prevalent in the diverse terrain of the Pacific Northwest, so overall the simplistic CLI technique appears superior.

Table 3: Mean energy score (ES) in $\mathrm{m \cdot s^{-1}}$, mean logarithmic score (LogS), mean quadratic score (QS) and mean spherical score (SphS) for density forecasts of surface winds over the Pacific Northwest, aggregated over stations with more than 300 available cases.

Density Forecast	ES	LogS	QS	SphS
UW	2.92		0.43	−0.05
ED	2.34	4.71	−0.01	−0.12
CLI	2.12	4.23	−0.04	−0.17

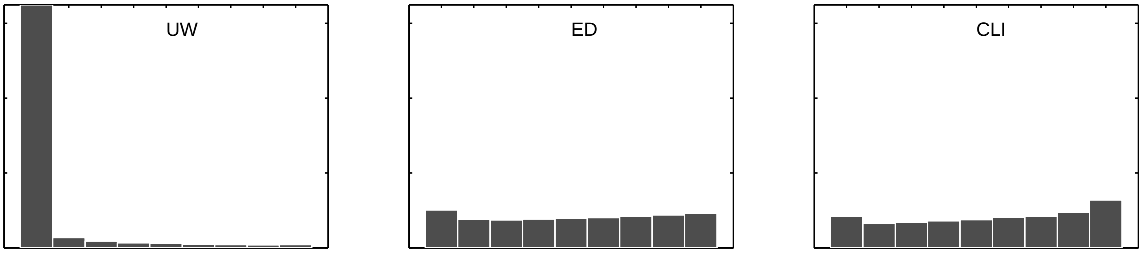

Figure 7: Box density ordinate transform (BOT) histogram for density forecasts of surface winds over the Pacific Northwest, aggregated over stations with more than 300 available cases.

5.3 Density forecasts

We now consider variants of the ensemble forecasts that take the simple form of smooth, bivariate Gaussian predictive densities.

To obtain a density version of the **University of Washington (UW)** forecast, we fit a bivariate normal density to the UW ensemble values, in a procedure that has been aptly described as ensemble smoothing (Wilson, Burrows and Lanzinger 1999; Wilks 2002). Specifically, the UW density forecast for the wind vector at station i on day j is

$$p_{ij}^{\mathrm{UW}} = \mathcal{N}_2(\boldsymbol{\mu}_{ij}^{\mathrm{UW}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{UW}}),$$

where $\boldsymbol{\mu}_{ij}^{\mathrm{UW}}$ and $\boldsymbol{\Sigma}_{ij}^{\mathrm{UW}}$, respectively, denote the mean vector and the covariance matrix of the ensemble members $\boldsymbol{\mu}_{ij1}^{\mathrm{UW}}, \ldots, \boldsymbol{\mu}_{ij8}^{\mathrm{UW}}$.

For the density version of the **ensemble dressing (ED)** method, we fit a bivariate normal density to the site-specific empirical distribution of forecast errors. The density forecast at station i on day j then is

$$p_{ij}^{\mathrm{ED}} = \mathcal{N}_2(\boldsymbol{\mu}_{ij}^{\mathrm{UW}} + \boldsymbol{\beta}_i^{\mathrm{ED}}, \boldsymbol{\Sigma}_i^{\mathrm{ED}}),$$

where $\boldsymbol{\beta}_i^{\mathrm{ED}}$ and $\boldsymbol{\Sigma}_i^{\mathrm{ED}}$ denote the mean and the covariance of the local UW ensemble mean errors $\boldsymbol{e}_{ik} = \boldsymbol{x}_{ik} - \boldsymbol{\mu}_{ik}^{\mathrm{UW}}$. The ED density forecast has a static covariance matrix at each unique location; it corrects for forecast biases and dispersion errors, and the mean remains dependent on the ensemble forecast.

For the density version of the **climatological (CLI)** forecast, we fit a bivariate normal density to the local empirical distribution of wind vectors, in that

$$p_{ij}^{\text{CLI}} = \mathcal{N}_2(\boldsymbol{\mu}_i^{\text{CLI}}, \boldsymbol{\Sigma}_i^{\text{CLI}}),$$

where $\boldsymbol{\mu}_i^{\text{CLI}}$ and $\boldsymbol{\Sigma}_i^{\text{CLI}}$ refer to the site-specific mean and covariance of the wind vector observations. At each unique location, the climatological forecast density is static over time.

Table 3 shows the mean energy score, mean logarithmic score, mean quadratic score and mean spherical score for the three types of density forecasts. The slight deterioration of the energy score, when compared to the results for the respective ensemble forecast types in Table 2, can likely be attributed to deviations from normality. This may suggest the use of non-Gaussian distributions, such as mixtures of bivariate normal densities, or skew-densities (Azzalini and Genton 2008). All four scores result in the same ordering, which places the CLI density forecast first, followed by the ED and UW forecast, except that the mean logarithmic score for the UW density forecast is missing. This is so because various forecast cases at various stations led to a numerically infinite logarithmic score, which happens in the case of a very sharp forecast and an outlying verifying wind vector and illustrates some of the practical difficulties with the logarithmic score. An alternative point of view is that these forecasts were poor beyond belief; hence, the infinite penalty is justified.

Figure 7 shows Box density ordinate transform (BOT) histograms for the three types of density forecasts. The ED and CLI density forecasts showed nearly uniform histograms, while the UW density forecast was severely underdispersed, as evidenced by the overrepresentation of low BOT values. These findings agree with and corroborate our interpretation of the rank histograms for the respective types of ensemble forecasts.

5.4 Ensemble forecasts at individual stations

We return to the ensemble forecasts described in Section 5.2 and discuss verification results at individual sites. To begin, we consider forecasts at Olympia Airport, Washington and West Vancouver, British Columbia. Olympia Airport lies in uniform terrain in the Puget Sound Lowlands of Western Washington. The station at West Vancouver is located near the Pacific Ocean and is heavily influenced by its surrounding terrain, as well as the land-water contrast. There were 397 wind observations at Olympia and 364 at West Vancouver.

Figure 8 shows marginal calibration diagrams for the UW ensemble mean forecast (9) at the two sites. Observations of wind speed and wind direction are quantized individually, as evidenced by the patterns in the plots. At Olympia, the numerical weather prediction model that underlies the UW ensemble reproduces the site-specific bivariate distribution of wind vectors, despite its 12 km grid resolution. At West Vancouver, actual wind vector observations are constrained to a small subset of the model state space, likely caused by channeling effects of nearby terrain.

Summary measures of predictive performance and multivariate rank histograms at the two stations are shown in Tables 4 and 5 and Figures 9 and 10. At Olympia, the ED ensemble performs best, showing both the lowest mean energy score and the lowest mean Euclidean error. The CLI and the ED ensemble have nearly uniform multivariate rank histograms; however, the ED ensemble is sharper, and it provides a better, state-dependent point forecast. The UW ensemble is sharp, but uncalibrated, and therefore penalized by the energy score. At West Vancouver, the CLI ensemble by far outperforms the UW and ED ensembles. Of course, this is unsurprising, given that the 12 km grid scale of the numerical weather prediction model is unable to resolve the local topography at this station.

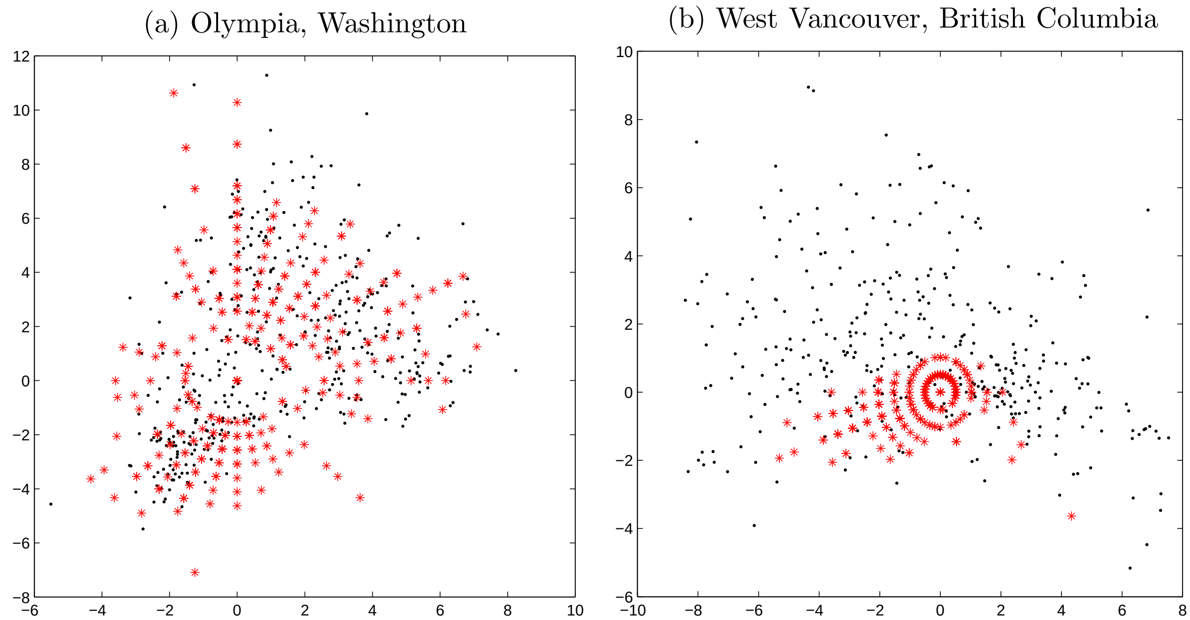

Figure 8: Marginal calibration diagram for (a) Olympia, Washington and (b) West Vancouver, British Columbia. Each red star represents an observation; each black square a UW ensemble mean forecast.

Figure 11 provides an overall assessment of the ensemble forecast types at the 565 individual sites considered here. Panels (a), (b) and (c) show the energy score for the UW, ED and CLI ensembles. The UW ensemble shows considerable spatial variation in the energy score, which the postprocessed ED ensemble tends to diminish. The CLI ensemble shows lower (better) energy score at stations in terrain that constrains wind direction, and higher scores in uniform terrain. Panel (d) shows which of the three methods performs best, as judged by the energy score. Only the postprocessed ED and the climatological CLI ensemble are competitive. The ED ensemble generally performs best in uniform terrain, including parts of the Pacific Coast, Southeastern Oregon and Southeastern Washington; otherwise, the CLI ensemble tends to be superior.

It is apparent that advanced statistical postprocessing techniques need to be developed that apply to ensemble predictions of surface winds. The ED and CLI techniques are simplistic; yet, our verification results can inform future efforts. The superiority of the ED ensemble at sites in uniform terrain, such as Olympia Airport, suggests the application of more versatile statistical postprocessing techniques at this type of stations. One possible approach would be based on a bivariate implementation of Bayesian model averaging (Raftery, Gneiting, Balabdaoui and Polakowski 2005). In less uniform terrain, climatological prior information and dynamic information from ensembles of numerical weather prediction models need to be merged in novel ways. Krzysztofowicz (2004) proposed a Bayesian framework for doing this, and a bivariate implementation might be feasible. Probabilistic forecasts on grids, rather than at individual stations, pose additional challenges, which we hope to address in future work.

Table 4: Same as Table 2, but for ensemble forecasts at Olympia, Washington only.

Ensemble Forecast	ES	EE	Δ	DS
UW	2.37	2.80	0.59	0.51
ED	1.95	2.75	0.21	2.75
CLI	2.27	3.27	0.19	3.27

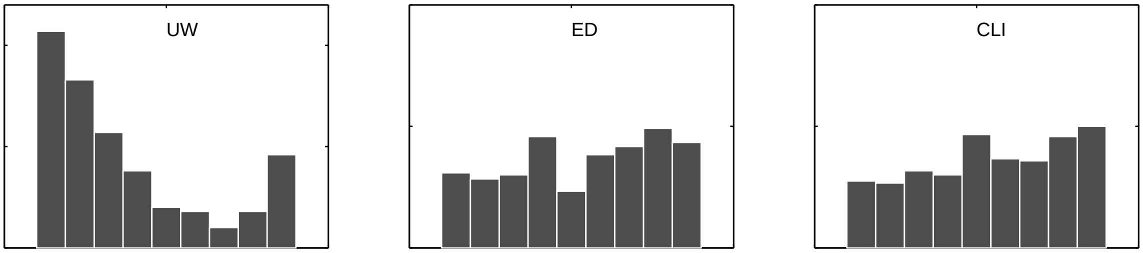

Figure 9: Multivariate rank histograms for ensemble forecasts at Olympia, Washington. The respective discrepancy measure Δ is shown in Table 4.

Table 5: Same as Table 2, but for ensemble forecasts at West Vancouver, British Columbia only.

Ensemble Forecast	ES	EE	Δ	DS
UW	3.33	3.83	0.63	0.63
ED	2.58	3.66	0.30	8.43
CLI	0.83	1.20	0.41	0.77

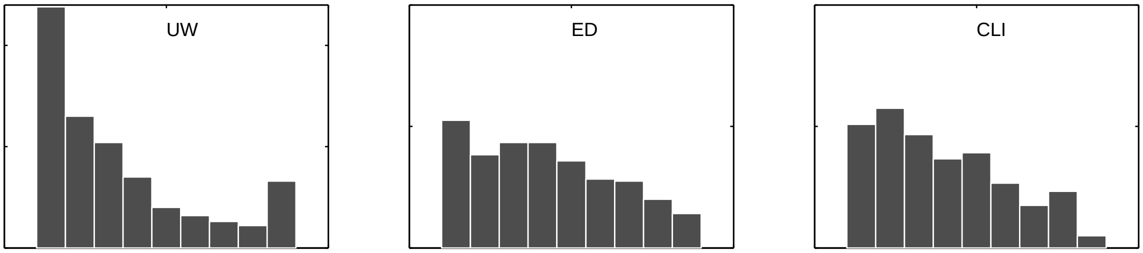

Figure 10: Multivariate rank histograms for ensemble forecasts at West Vancouver, British Columbia. The respective discrepancy measure Δ is shown in Table 5.

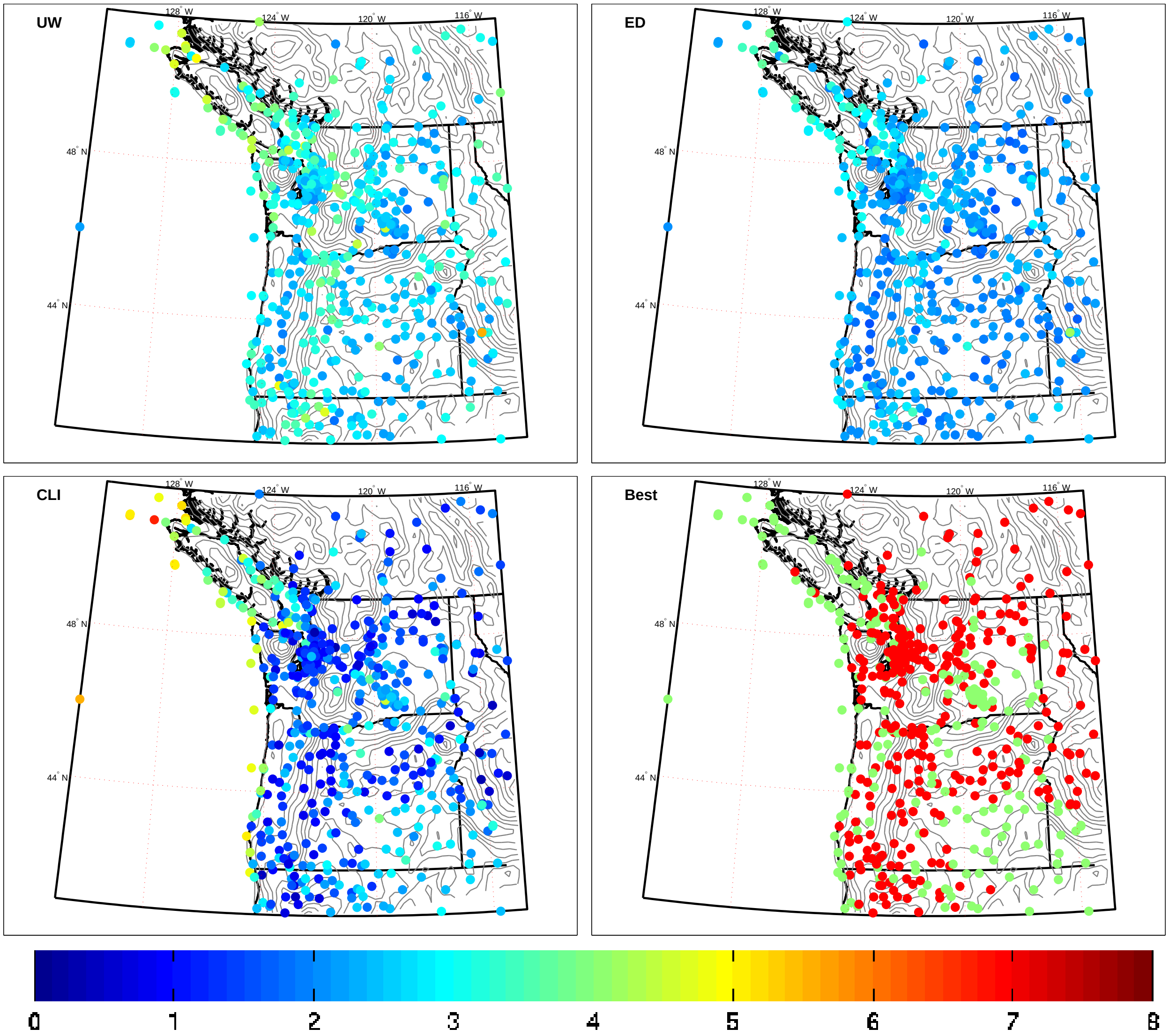

Figure 11: Mean energy score in $\mathrm{m \cdot s^{-1}}$ at individual stations for ensemble forecasts with (a) the UW, (b) the ED and (c) the CLI technique. Panel (d) shows which forecast performed best in terms of the energy score, with green and red representing the ED and CLI ensembles, respectively. The color bar applies to the energy scores in panels (a), (b) and (c).

6 Closing remarks

We have discussed tools for the assessment of probabilistic forecasts of vector-valued quantities. Our work is in the tradition of diagnostic forecast evaluation (Murphy, Brown and Chen 1989; Murphy and Winkler 1992; Gneiting, Balabdaoui and Raftery 2007), which emphasizes the need to understand the ways in which predictive models fail or succeed. In the case of ensemble forecasts, we have explored the use of the multivariate rank histogram, which is a direct generalization of the univariate Talagrand diagram, to check calibration. The determinant sharpness generalizes the univariate standard deviation as a measure of spread. The energy score is a proper multivariate generalization of the univariate continuous ranked probability score; its key uses lie in the ranking and comparison of competing forecasting techniques. In the context of density forecasts, we have shown that the Box density ordinate transform (BOT) can be used successfully to assess calibration. R code can be provided upon request.

In a meteorological case study, we have applied tools that include rank and BOT histograms, marginal calibration diagrams and proper scoring rules to diagnose strengths and deficiencies of probabilistic short-range forecasts of surface wind vectors over the Pacific Northwest. While the flavor of our work is diagnostic, inferential approaches are feasible and have been explored both in economic and in meteorological applications (Diebold and Mariano 1995; Hamill 1999; Clements 2005; Jolliffe 2007).

The same set of tools applies very generally, whenever statistical models for multivariate quantities need to be critiqued, evaluated and compared (Box 1980; O'Hagan 2003). Frequently, models can be fitted in cross-validation mode, and can be assessed based on the quality of the ensuing predictive distributions. Simplicity, generality and interpretability are attractive features of the toolbox presented here, which applies in parametric as well as non-parametric settings, and does not require models to be nested, nor be related in any way. We believe that these tools can provide guidance in a wealth of applied statistical problems for multivariate continuous data, ranging from the evaluation of probabilistic forecasts to model criticism, model comparison and model choice.

Acknowledgements This research has been supported by the National Science Foundation under Awards ATM-0724721 and DMS-0706745, by the Joint Ensemble Forecasting System (JEFS) under Subcontract S06-47225 from the University Corporation for Atmospheric Research (UCAR), and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. Leonhard Held acknowledges support by the Swiss National Science Foundation. Larissa Stanberry acknowledges technical support by the Department of Educational Psychology at the University of Washington. We thank Jeff Baars, Peter Bickel, Tom Hamill, Subhash Kochar, Cliff Mass, Tony O'Hagan, Adrian Raftery, McLean Sloughter and an anonymous reviewer of a related paper for providing data, comments and/or references.

References

Anderson, JL (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J Climate 9:1518–1525

Azzalini A, Genton MG (2008) Robust likelihood methods based on the skew-t and related distributions. Int Stat Rev 76:106–129

Bernardo JM (1979) Expected information as expected utility. Ann Stat 7:686–690

Berrocal VJ, Raftery AE, Gneiting T (2007) Combining spatial statistical and ensemble information in probabilistic weather forecasts. Mon Wea Rev 135:1386–1402

Besag J, Green P, Higdon D, Mengersen K (1995) Bayesian computing and stochastic systems. Stat Sci 10:3–66

Bickel PJ (1969) A distribution free version of the Smirnov two sample test in the p-variate case. Ann Math Stat 40:1–23

Bickel PJ, Lehmann EL (1979) Descriptive statistics for nonparametric models IV. Spread. In: Contributions to statistics, Jureckova J (ed.), Academia, Prague, pp 33–40

Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. J Roy Stat Soc Ser A, 143:383–425

Brockwell AE (2007) Universal residuals: A multivariate transformation. Stat Prob Lett 77:1473–1478

Bröcker J, Smith LA (2007) Scoring probabilistic forecasts: The importance of being proper. Wea Forecasting 22:382–388

Candille G, Talagrand O (2005) Evaluation of probabilistic prediction systems for a scalar variable. Quart J Roy Meteorol Soc 131:2131–2150

Clements MP (2005) Evaluating econometric forecasts of economic and financial variables. Palgrave Macmillan, Basingstroke, Hampshire

Clements MP, Smith J (2000) Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. J Forecasting 19:255–276

Clements MP, Smith J (2002) Evaluating multivariate forecast densities: A comparison of two approaches. Int J Forecasting 18:397–407

Czado C, Gneiting T, Held L (2007) Predictive model assessment for count data. Tech. Rep. no. 518, Dept. of Statistics, University of Washington

Dawid AP (1984) Statistical theory: The prequential approach. J Roy Stat Soc Ser A 147:278–292

Dawid AP, Sebastiani P (1999) Coherent dispersion criteria for optimal experimental design. Ann Stat 27:65–81

De Gooijer JG (2007) Power of the Neyman smooth test for evaluating multivariate forecast densities. J Appl Stat 34:371–381

Delle Monache L, Hacker JP, Zhou Y, Deng X, Stull RB (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. J Geophys Res 111:D24307, doi:10.1029/2005JD 006917

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–263

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts: With applications to financial risk management. Int Econ Rev 39:863–883

Diebold FX, Hahn J, Tay AS (1999) Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. Rev Econ Stat 81:661–673

Eckel FA, Mass CF (2005) Aspects of effective short-range ensemble forecasting. Wea Forecasting 20:328–350

Friedman JH, Rafsky LC (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Stat 7:697–717

Genest C, Rivest LP (2001) On the multivariate probability integral transform. Stat Prob Lett 53:391–399

Gneiting T (2008) Editorial: Probabilistic forecasting. J Roy Stat Soc Ser A, 171:319–321

Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. Science 310: 248–249

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. J Roy Stat Soc Ser B 69:243–268

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Wea Rev 33:1098–1118

Gneiting T, Larson K, Westrick K, Genton MG, Aldrich E (2006) Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) method. J Am Stat Assoc 101:968–979

Good IJ (1971) Comment on 'Measuring information and uncertainty' by Robert J. Buehler. In: Foundations of statistical inference, Godambe VP, Sprott DA (eds.), Holt, Rinehart and Winston, Toronto, pp 337–339

Gombos D, Hansen JA, Du J, McQueen J (2007) Theory and applications of the minimum spanning tree rank histogram. Mon Wea Rev 135:1490–1505

Granger, CWJ (2006) Preface: Some thoughts on the future of forecasting. Oxford Bull Econ Stat 67S:707–711

Grimit EP, Mass CF (2002) Initial results of a mesoscale short-range ensemble system over the Pacific Northwest. Wea Forecasting 17:192–205

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Quart J Roy Meteorol Soc 132:2925–2942

Hamill TM (1999) Hypothesis tests for evaluating numerical precipitation forecasts. Wea Forecasting 14:155–167

Hamill TM (2001) Interpretation of rank histograms for verifying ensemble forecasts. Mon Wea Rev 129:550–560

Hamill TM, Colucci SJ (1997) Verification of Eta-RSM short-range ensemble forecasts. Mon Wea Rev 125:1312–1327

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea Forecasting 15:559–570

Huber PJ (1985) Projection pursuit. Ann Stat 13:435–475

Ishida I (2005) Scanning multivariate conditional densities with probability integral transforms. Center for Advanced Research in Finance, University of Tokyo, Working Paper F-045

Jolliffe IT (2007) Uncertainty and inference for verification measures. Wea Forecasting 22:637–650

Jolliffe IT, Stephenson DB (2003) Forecast verification: A practitioner's guide in atmospheric science. Wiley, Chichester

Judd K, Smith LA, Weisheimer A (2007) How good is an ensemble at capturing truth? Using bounding boxes for forecast evaluation. Quart J Roy Meteorol Soc 133:1309–1325

Kruskal JB (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Amer Math Soc 7:48–50

Krzysztofowicz R (2004) Bayesian processor of output: A new technique for probabilistic weather forecasting. Extended abstract no. 4.2, 17th Conference on Probability and Statistics in the Atmospheric Sciences

Malmberg A, Holst J, Holst U (2008) A real-time assimilation algorithm applied to near-surface ocean wind fields. Environmetrics 19:319–330

Mass CF, Albright M, Ovens D, Steed R, MacIver M, Grimit E, Eckel T, Lamb B, Vaughan J, Westrick K, Storck P, Colman B, Hill C, Maykut N, Gilroy M, Ferguson SA, Yetter J, Sierchio JM, Bowman C, Stender R, Wilson R, Brown W (2003) Regional environmental prediction over the Pacific Northwest. Bull Am Meteorol Soc 84:1353–1366

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Management Sci 22:1087–1096

Murphy AH, Winkler RL (1992) Diagnostic verification of probability forecasts. Int J Forecasting 7:435–455

Murphy AH, Brown BG, Chen YS (1989) Diagnostic verification of temperature forecasts. Wea Forecasting 4:485–501

National Research Council (2006) Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. The National Academies Press, Washington

O'Hagan A (2003) HSSS model criticism. In: Highly structured stochastic systems, Green PJ, Hjort NL and Richardson S (eds.), Oxford University Press, pp 423–444

Oja H (1983) Descriptive statistics for multivariate distributions. Stat Prob Lett 1:327–332

Oja H, Randles RH (2004) Multivariate nonparametric tests. Stat Sci 19:598–605

Palmer TN (2002) The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Quart J Roy Meteorol Soc 128:747–774

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Wea Rev 133:1155–1174

Rife DL, Davis CA (2005) Verification of temporal variations in mesoscale numerical wind forecasts. Mon Wea Rev 133:3368–3381

Rosenblatt M (1952) Remarks on a multivariate transformation. Ann Math Stat 23:470–472

Roulston MS, Smith LA (2003) Combining dynamical and statistical ensembles. Tellus Ser A 55:16–25

Savage LJ (1971) Elicitation of personal probabilities and expectation. J Am Stat Assoc 66:783–801

Shaked M, Shanthikumar JG (1994) Stochastic orders and their applications. Academic, Boston

Shephard N (1994) Partial non-Gaussian state space. Biometrika 81:115–131

Smith LA (2001) Disentangling uncertainty and error: On the predictability of nonlinear systems. In: Nonlinear dynamics and statistics, Mees AI (ed.), Birkhäuser, Boston, pp 31–64

Smith LA, Hansen JA (2004) Extending the limits of ensemble forecast verification with the minimum spanning tree histogram. Mon Wea Rev 132:1522–1528

Stephenson DB, Doblas-Reyes FJ (2000) Statistical methods for interpreting Monte Carlo forecasts. Tellus Ser A 52:300–322

Stigler SM (1975) The transition from point to distribution estimation. Bull Int Stat Inst 46:332–340

Talagrand O, Vautard R, Strauss B (1997) Evaluation of probabilistic prediction systems. In: Proceedings of a workshop held at ECMWF on predictability, 20–22 October 1997, European Centre for Medium-Range Weather Forecasts, Reading, pp 1–25

Timmermann A (2000) Density forecasting in economics and finance. J Forecasting 19:231–234

Weisheimer A, Smith LA, Judd K (2005) A new view of seasonal forecast skill: Bounding boxes from the DEMETER ensemble forecasts. Tellus Ser A 57:265–279

Wilks DS (2002) Smoothing forecast ensembles with fitted probability distributions. Quart J Roy Meteorol Soc 128:2821–2836

Wilks DS (2004) The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Mon Wea Rev 132:1329–1340

Wilks DS (2006) Statistical methods in the atmospheric sciences (2nd ed). Elsevier Academic, Amsterdam

Wilson LJ, Burrows WR, Lanzinger A (1999) A strategy for verification of weather element forecasts from an ensemble prediction system. Mon Wea Rev 127:956–970

Winkler RL (1977) Rewarding expertise in probability assessment. In: Decision making and change in human affairs, Jungermann H, de Zeeuw G (eds.), D. Reidel, Dordrecht, pp 127–140

Winkler RL (1996) Scoring rules and the evaluation of probabilities. Test 5:1–60

Zuo Y, Serfling R (2000) General notions of statistical depth functions. Ann Stat 28:461–482