On Evaluation of Ensemble Forecast Calibration Using the Concept of Data Depth

MAHSA MIRZARGAR

Department of Computer Science, University of Miami, Coral Gables, Florida

JEFFREY L. ANDERSON

National Center for Atmospheric Research,[a] Boulder, Colorado

(Manuscript received 9 September 2016, in final form 16 January 2017)

ABSTRACT

Various generalizations of the univariate rank histogram have been proposed to inspect the reliability of an ensemble forecast or analysis in multidimensional spaces. Multivariate rank histograms provide insightful information about the misspecification of genuinely multivariate features such as the correlation between various variables in a multivariate ensemble. However, the interpretation of patterns in a multivariate rank histogram should be handled with care. The purpose of this paper is to focus on multivariate rank histograms designed based on the concept of data depth and outline some important considerations that should be accounted for when using such multivariate rank histograms. To generate correct multivariate rank histograms using the concept of data depth, the datatype of the ensemble should be taken into account to define a proper preranking function. This paper demonstrates how and why some preranking functions might not be suitable for multivariate or vector-valued ensembles and proposes preranking functions based on the concept of simplicial depth that are applicable to both multivariate points and vector-valued ensembles. In addition, there exists an inherent identifiability issue associated with center-outward preranking functions used to generate multivariate rank histograms. This problem can be alleviated by complementing the multivariate rank histogram with other well-known multivariate statistical inference tools based on rank statistics such as the depth-versus-depth (DD) plot. Using a synthetic example, it is shown that the DD plot is less sensitive to sample size compared to multivariate rank histograms.

1. Introduction

A rank histogram serves as one of the most widely used techniques for evaluating the reliability or calibration of an ensemble forecast (Anderson 1996; Hamill 2001). The nonparametric nature of the univariate rank histogram makes it a flexible tool that can be used for assessing the calibration; it can also provide insight into the nature of miscalibration (if any). However, like any other evaluation technique, a rank histogram also has its associated drawbacks and restrictions.

One of the main drawbacks of the rank histogram is that it cannot be trivially generalized for multivariate ensemble forecasts. Various generalizations of the concept of rank function, as the main building block of univariate rank histograms, have been proposed, such as the minimum-spanning-tree (MST) ranking (Smith and Hansen 2004), average ranking (Thorarinsdottir et al. 2016), and band depth ranking (Thorarinsdottir et al. 2016). The MST rank histogram has been carefully studied before (Gombos et al. 2007). Recently, Wilks (2017) also performed a careful study of the utility of various multivariate calibration metrics in various covariance miscalibration scenarios. However, a thorough study of the properties of multivariate rank histograms based on the notion of data depth is still missing.

This paper aims to study various properties of data-depth-based prerank functions. Through an illustrative example, we demonstrate that a multivariate prerank function must satisfy an important property (i.e., affine invariance) to be correct. Moreover, the patterns in a multivariate rank histogram should be interpreted with

[a] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author e-mail: Mahsa Mirzargar, mirzargar@cs.miami.edu

DOI: 10.1175/MWR-D-16-0351.1

care as the implications can be different from what a univariate rank histogram suggests. The rest of the paper proceeds as follows. The next section gives an overview of the idea behind the rank histogram and its generalizations. Section 3 discusses various considerations when using multivariate rank histograms via examples. Section 4 provides an example using real ensemble forecasts for wind direction in the U.S. Pacific Northwest. We summarize our findings and conclude in section 5.

2. The rank histogram and its generalizations

The univariate rank histogram (Anderson 1996; Hamill 2001) and its variations (Gneiting et al. 2008; Wilks 2004; Thorarinsdottir et al. 2016) are some of the most widely used techniques for evaluating the reliability or calibration of an ensemble forecast. The popularity of the rank histogram for empirical evaluation of ensemble forecasts is mainly due to its nonparametric nature. If an ensemble forecast is calibrated, the observation would be *indistinguishable* from the ensemble members. In other words, the ensemble forecast and the observation are both random samples from the same probability distribution.

For constructing a univariate rank histogram, Anderson (1996) proposed to use the rank statistics of the observation when pooled with members of an ensemble forecast to evaluate calibration. If the observation is a random sample from the same distribution as the ensemble forecast, it has an equal probability of receiving any specific rank value (i.e., it is indistinguishable from any of the ensemble members). Therefore, the histogram of the rankings of a set of observations can be used to assess the calibration. The uniformity of a rank histogram is a necessary (but not sufficient) condition for calibration of an ensemble. The uniformity of the rank histogram can be tested quantitatively using a class of hypothesis testing techniques called goodness-of-fit (GOF) tests. However, the utility of the results of such tests are quite sensitive to sample size and deviation patterns (Elmore 2005). Therefore, in application, the assessment of the uniformity of a rank histogram is oftentimes performed qualitatively by visual exploration.

In addition to assessing the calibration, specific types of deviation can be associated with common types of miscalibration of ensemble forecasts, such as underdispersion, overdispersion, and the presence of some kind of bias in a univariate ensemble forecast. A univariate rank histogram with a ∪ shape indicates underdispersion, a peaked rank histogram indicates overdispersion, and a rank histogram with a slope indicates the presence of a bias.

The main building block of a rank histogram is the concept of ranking. The concept of ranking is well defined for univariate values but more complex for multivariate values. Among the first generalizations of the rank histogram to multivariate values is the MST rank histogram (Smith and Hansen 2004). The main idea of the MST rank histogram is to define a univariate metric for points in dimensions greater than one and use this univariate metric to find the ranking of the observation when pooled with the ensemble forecast. The MST rank histogram uses the sum of the length of line segments forming the minimum spanning tree of individual dimensions for ranking. However, this metric does not reduce to a univariate ranking. Moreover, Gombos et al. (2007) demonstrated that the MST rank histogram is sensitive to the choice of the norm used to construct the minimum spanning tree and can potentially result in misinterpretation of the quality of an ensemble forecast (Wilks 2004; Gombos et al. 2007).

Another extension of the univariate rank histogram can be built upon more stable concepts of ranking for multivariate data points widely studied in the statistics literature (Liu et al. 1999). In multivariate data analysis, the ranking is often defined using the concept of data depth. Given an ensemble of data drawn from a distribution F, data depth quantifies how central (or deep) a particular sample is within the cloud of the sampled data. The samples near the center of the ensemble are assigned higher depth values whereas samples farther away from the rest of the ensemble are assigned low depth values and can be considered as potential outliers. Therefore, sorting the depth values can provide a *center-outward* ordering of an ensemble of sampled data. In other words, data depth analysis provides a preranking function $\rho: \mathbb{R}^k \mapsto \mathbb{R}_+$ that can be used to define an ordering for members of a k-dimensional ensemble. More specifically, given an ensemble $\mathscr{E} := \{x_1, \ldots, x_m\}$, $x_i \in \mathbb{R}^k$, a preranking function $\rho: \mathbb{R}^k \mapsto \mathbb{R}_+$ can be defined as follows:

$$\rho_j = \sum_{i=0}^{m} \mathbb{I}(x_i \preceq x_j), \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function and the definition of $\preceq$ depends on the definition of data depth used.

Various definitions of data depth have been studied in the data analysis literature including multivariate points (Liu 1990), functional data (López-Pintado and Romo 2009), isocontours/isosurfaces (Whitaker et al. 2013), and multivariate trajectories (Mirzargar et al. 2014; López-Pintado et al. 2014). Recently, Thorarinsdottir et al. (2016) provided a definition of a preranking function based on the concept of functional band depth for d-dimensional ensembles. The functional band depth prerank function can be defined as follows:

$$\rho^{\mathrm{bd}}(x)=\frac{1}{d}\sum_{k=1}^{d}\sum_{1\le i_1<i_2\le m}\mathbb{I}[\min\{x_{i_1k},x_{i_2k}\}\le x_k\le\max\{x_{i_1k},x_{i_2k}\}], \tag{2}$$

where $\mathbb{I}$ represents an indicator function, $x_i=(x_{i_1},\ldots,x_{i_d})$ represents a d-dimensional ensemble member, $\{x_{i_1},x_{i_2}\}$ represents a random subset of size two from the ensemble, and m represents the size of the ensemble (Thorarinsdottir et al. 2016). Likewise, different definitions of data depth can be used to define preranking functions to generate multivariate rank histograms for other datatypes, for instance, to define rank histograms for ensembles of pathlines or isocontours.

The uniformity of the rank histogram is a necessary condition for reliability of an ensemble but it is not sufficient (Hamill 2001). Various studies have shown the challenges in using and interpreting rank histograms to evaluate calibration. Hamill et al. (2000) and Hamill (2001) provided interesting scenarios in which the uniformity of the univariate rank histogram is misleading. Most of those scenarios are directly applicable to generalizations of univariate rank histograms. Wilks (2004) reviewed various shortcomings of the MST rank histogram. Thorarinsdottir et al. (2016) showed that generalization of the rank histogram using functional band depth can be insightful to detect miscalibration due to differences in correlation structure of d-dimensional ensembles. However, the interpretation of different patterns in a multivariate rank histogram should be handled with care.

3. Interpreting multivariate rank histograms

In this section, we provide various examples to carefully study some important considerations that bear on the use of multivariate rank histograms with a specific focus on the multivariate rank histograms defined based on the notion of data depth. Our examples focus on demonstrating the similarities and differences between the interpretation of data-depth-based rank histograms compared to a univariate rank histogram and other definitions of multivariate rank histograms.

a. *Data type and invariance property*

An important consideration in using rank histograms is the *data type* of the ensemble. Data type refers to the type of the values in an ensemble (e.g., univariate values, multivariate points, time series, vector values). The data type of the ensemble has implications for how one can properly define the concept of ordering. This topic has received some attention before. For instance, Gneiting et al. (2007, 2008) defined a multivariate preranking function for multivariate ensembles and demonstrated its application using vector-valued ensembles for wind prediction application (Gneiting et al. 2007, 2008). For two multivariate values $\mathbf{x}$, $\mathbf{y}\in\mathbb{R}^d$, Gneiting et al. proposed to use the following rule for ordering:

$$\mathbf{x}\le\mathbf{y}\quad\text{if and only if}\quad x_i\le y_i\quad\text{for all}\quad i=1,\ldots,d. \tag{3}$$

When used for an ensemble of multivariate points, this definition uses the number of points on the lower-left side of an ensemble member to represent its prerank. Although not explicitly mentioned, this definition has some similarity to the definition of half-region depth defined for ensembles of functions (López-Pintado and Romo 2011). The half-region depth uses the region above (epigraph) and below (hypograph) the graphical representation of a function to define an ordering for an ensemble of functions.

Further studies on the preranking function defined in Eq. (3) have confirmed an identifiability problem with this ordering concept before (Pinson and Girard 2012; Thorarinsdottir et al. 2016). From the data-analysis perspective, this preranking function violates an important property called *affine invariance*. Affine invariance is an important property that needs to be satisfied in order to derive accurate and proper order statistics (Serfling and Zuo 2000). Affine invariance means that the ranking should not depend on the underlying coordinate system. In other words, the ranking should not change by affine transformation of the ensemble. The following example demonstrates a simple scenario in which the affine invariance property is violated using Eq. (3).

Consider a bivariate ensemble forecast of size 8 that follows a bivariate normal distribution where one of the variables is constrained to stay positive. Furthermore, assume that the observation is following another bivariate normal distribution (with the same constraint) with the same mean but a smaller standard deviation in one direction. That is, the ensemble forecast is overdispersed. Figure 1a shows the ensemble and the observations over 2500 realizations. Figure 1c shows the corresponding multivariate rank histogram using Eq. (3). As expected, the multivariate rank histogram is peaked, which indicates the overdispersion of the ensemble.

We now consider a simple affine transformation of the ensemble. Figure 1b shows the same set of realizations rotated −45°. Figure 1d demonstrates the resultant multivariate rank histograms. Note that the axis-dependent nature of the multivariate preranking function resulted in the sensitivity of the rankings to the orientation of the

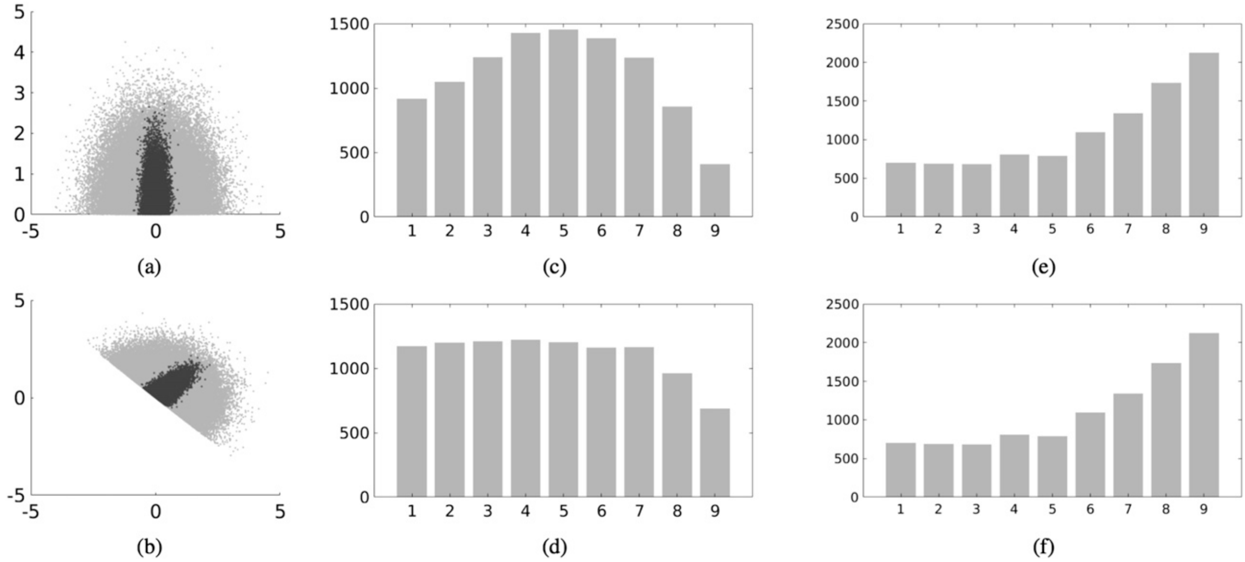


FIG. 1. Synthetic example to study the effect of affine transformation on the multivariate rank histogram proposed by Gneiting et al. (2008). In this example, the ensemble forecast is overdispersed. (a) The visualization of a bivariate ensemble forecast and the observations over 2500 realizations. The ensemble members are in light gray and the observations are in dark gray. (b) The set of data points in (a) is rotated −45°. (c) The multivariate rank histogram using the ensemble demonstrated in (a). (d) The multivariate rank histogram using the ensemble demonstrated in (b). Notice that rotation of the data points resulted in significant change in the rankings due to the violation of the affine invariance property. Using a proper ordering concept that satisfies affine invariance such as simplicial depth would result in consistency of the rank histogram irrespective of the orientation (or coordinate system). (e),(f) These panels are identical to each other and demonstrate the rank histogram for (a) and (b), respectively, where simplicial depth has been used for ranking. In comparison to the second column, the rankings are insensitive to affine transformation.

data. Hence, the rank histogram in this figure shows a different pattern of deviation even though the data have not changed. Figure 2 demonstrates the problem in a simpler setting using a small ensemble.

This problem can be avoided by using a preranking function that satisfies the affine invariance property and provides appropriate ordering for multivariate points. In what follows we provide an introduction to the definition of simplicial depth that is affine invariant. To the best of our knowledge, the concept of simplicial depth has not been utilized to generate multivariate rank histograms before.

b. *Simplicial depth*

Simplicial depth is one of the traditional and well-studied definitions of data depth for multivariate points (Liu and Singh 1992; Liu 1990; Rousseeuw and Ruts 1996). The formal definition of simplicial depth is as follows.

Let P denote a probability distribution in $\mathbb{R}^d$. The simplicial depth of a point $x \in \mathbb{R}^d$ is defined as the probability that x resides in a random simplex formed by $d+1$ distinct independent random points from P (Liu 1990):

$$D_P(x) = \mathrm{Prob}(x \in B), \quad x \in \mathbb{R}^d, \tag{4}$$

where $B = S_{d+1} := \Delta(x_1, \ldots, x_{d+1})$ is the d-dimensional simplex formed by considering the convex hull of $d+1$ random points from P. The closer the point is to the center of the distribution, the higher the probability of it being contained in a random simplex. Similarly, the simplicial depth values can be approximated for an ensemble of d-variate points as follows.

Given a set of n d-variate data points, $\mathscr{E} = \{x_1, \ldots, x_n\}$, consider all possible subsets of size $d+1$ points from the set. There exist $C(n, d+1)$ such subsets where $C(n, k)$ denotes the k combination of a set of size n. Construct the simplex formed by each subset and check whether x falls inside that simplex. Then, the empirical simplicial depth value for each $x \in \mathscr{E}$ can be evaluated by replacing the probability in Eq. (4) with an expected value and estimating its value using the sample mean (Liu 1990). That is,

$$\begin{aligned} D_{P_n}(x) &= \mathbb{E}[\mathbb{I}(x \in S_{d+1})] \\ &\approx \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < \cdots < i_{d+1} \le n} \mathbb{I}(x \in S_{d+1}), \end{aligned} \tag{5}$$

where $\mathbb{E}[\cdot]$ denotes expectation, $\{i_1, \ldots, i_{(d+1)}\}$ represent a subset of size $d+1$ from the ensemble, S_{d+1} represents the simplex formed by the set of $d+1$ random points from the observations, and $\mathbb{I}(\cdot)$ denotes an indicator function. In other words, the empirical depth

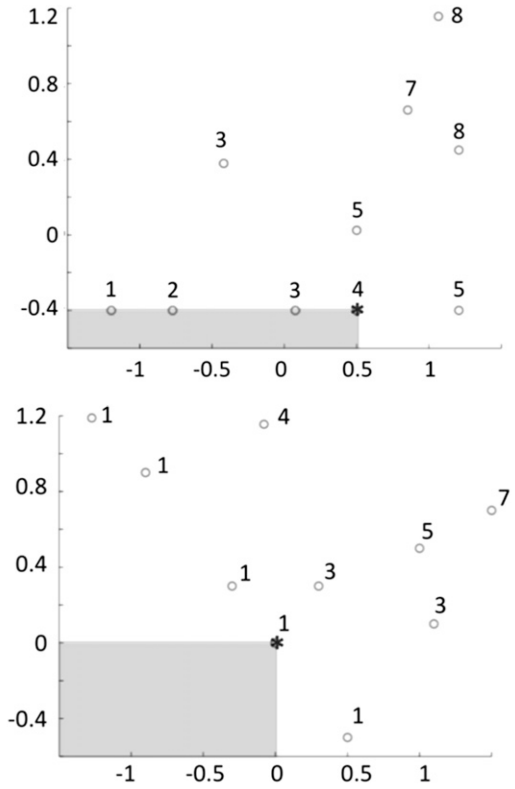


FIG. 2. The preranking function defined in Eq. (3) is not affine invariant. That is, the affine transformation of the data can change the prerank values as demonstrated here. The ensemble in the second row is a rotated version of the ensemble in the first row. The prerank value for each ensemble member is marked on top of the data point. Note that the prerank values represent the number of points that are on the lower left of each ensemble member (including itself). The gray box demonstrates the points that have *lower* (pre) rank than the point demonstrated by an asterisk. Notice that the rank values for the point demonstrated by an asterisk (and other ensemble members) has changed in the second row.

value of each observation in this case is the proportion of the simplices that contain x.

By sorting depth values one can provide a center-outward ordering of the ensemble. It is important to note that unlike the preranking function defined in Eq. (3), simplicial depth is invariant with respect to affine transformation, and hence, can provide a proper ordering for ensembles of multivariate points. Figures 1e and 1f demonstrate the multivariate rank histogram for the synthetic example presented earlier. Unlike multivariate rank histograms presented in Figs. 1c and 1d, the shape of the rank histogram stays the same after the rotation of the ensemble. Note that the simplicial depth analysis provides a center-outward ordering, and hence, the rank histogram in Figs. 1e and 1f includes many high ranks due to the overdispersed nature of the ensemble.

In addition to multivariate points, simplicial depth is also suitable to study unit-length vector-valued ensembles. The main intuition for using simplicial depth for vector-valued ensembles is that a unit-length 2D vector is a point on a circle. Therefore, the depth value of a unit-length 2D vector among an ensemble of unit-length 2D vectors can be defined using Eq. (4) where the band B in this case is the arc connecting any two random subsets of the ensemble on a circle (Liu and Singh 1992). Figure 3 demonstrates the concept of containment in a band in this case. Note that simplicial depth is applicable only to unit-length vectors. A proper generalization of the simplicial depth for ensembles of vectors with arbitrary length is an interesting and open problem.

To demonstrate the utility of the simplicial depth for evaluation of vector-valued ensembles, we consider an example similar to the one used by Gneiting et al. for wind prediction (Gneiting et al. 2007, 2008).

Proper prediction of the wind direction requires the simulation to have enough resolution and proper modeling of the topography of a region. Therefore, for a region with complicated topography, a coarse model cannot resolve the wind vector properly. For example, let us assume that we have a narrow valley running east–west so that winds are confined to go mostly east–west at a measuring station in the valley. In a coarse model that does not resolve the topography of the valley properly, winds will be much more homogeneous. Therefore, the ensemble would be overdispersed compared to the observations.

For this example, consider an ensemble of 2D wind vectors of size 19 following a normal distribution for the angles with a mean value of zero degree and standard deviation of $\pi/5$ and the observation is a random draw from another normal distribution with mean value of zero degree and standard deviation of $\pi/10$. Figure 4 demonstrates the multivariate rank histogram using the notion of simplicial depth for this synthetic example. The rank histogram has been generated over 2500 realizations. The multivariate rank histogram demonstrates the overdispersive nature of the ensemble forecast correctly. Although this example demonstrates the utility of simplicial depth for 2D wind vectors, a similar concept can be used for higher-dimensional vectors, for instance 3D vectors in fluid flow.

It is important to note that simplicial depth and the functional band depth prerank function (Thorarinsdottir et al. 2016) both satisfy affine invariance, and hence provide proper ordering. However, each of these prerank functions is suitable for different data types. Functional band depth prerank function (Thorarinsdottir et al. 2016) is suitable for time series and functional values, whereas simplicial depth is suitable for multivariate points and vector-valued ensembles.

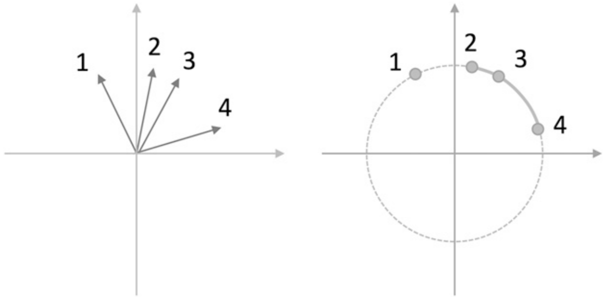


FIG. 3. The concept of simplicial depth can be used for ordering ensembles of unit-length 2D vectors directly by considering them on a unit circle (i.e., polar representation). (left) An ensemble of four unit-length vectors, and (right) the same ensemble has been represented as points on a circle. A vector is contained among two other vectors if it is on the smallest arc connecting the other two when represented as points on a circle. The solid line on the right-hand side highlights the smallest arc between 2 and 4 that contains 3 but not 1.

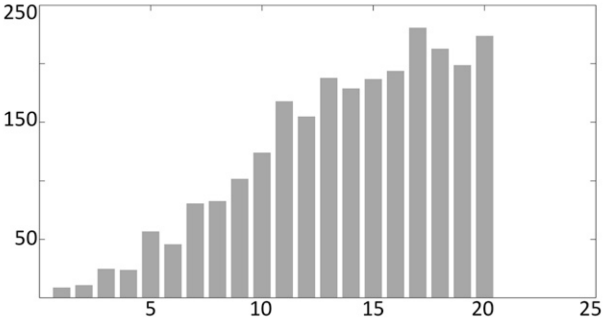


FIG. 4. Multivariate rank histogram for a synthetic example for wind direction prediction using simplicial depth for preranking.

c. Identifiability issue and the DD plot

Another important consideration in using multivariate rank histograms is the interpretation of the patterns of deviation, especially for the multivariate rank histograms developed based on the concept of data depth. Forecasters are accustomed to specific interpretation of such patterns. For instance, oftentimes forecasters associate a ∪-shaped rank histogram with the presence of underdispersion or underprediction. However, such interpretations can result in misleading conclusions when using multivariate rank histograms based on the concept of data depth. Recent work on multivariate rank histograms provided evidence for such misinterpretations (Thorarinsdottir et al. 2016; Wilks 2017).

Before we introduce an alternative graphical inference tool that tackles this problem, we present an example to demonstrate that in comparison to univariate rank histograms, multivariate rank histograms based on the concept of data depth have an inherent identifiability problem for underdispersed or biased ensembles. We use a synthetic example to demonstrate the typical emerging patterns in a multivariate rank histogram in the presence of bias, overdispersion, and underdispersion.

To generate a nontrivial yet interesting ensemble of functions, we have chosen to use a Gaussian process to define a distribution of functions as

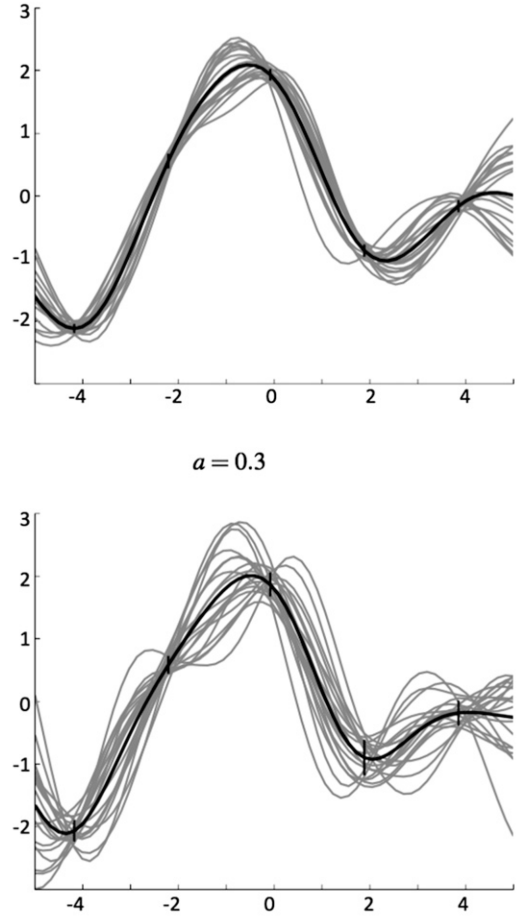
$$f(x) \sim \mathcal{N}[m(x), K(x,x')], \tag{6}$$

where $m(x)$ is the mean function and $K(x, x')$ is the covariance function. A Gaussian process is fully specified by fixing $m(x)$ and $K(x, x')$. For all examples in this section, we consider the mean function to be zero for simplicity and the covariance function to have the following generic form:

$$K(x,x') = \exp(-a\,|\,x - x'|^2), \tag{7}$$

where the parameter a controls the effective correlation length. The independent variable x can be considered as position or time. We used 32 uniform samples in the interval [−5, 5] for the independent variable to represent each function. To control the variability of the samples drawn from the Gaussian process, we have chosen to constrain a small set of data points along the sample paths (five points in Fig. 5). The constraining points are assumed to be imperfect where the noise associated with each data point follows a normal distribution (see Fig. 5 for an illustration). We follow the procedure introduced in Rasmussen and Williams (2006, their section 2.2) to sample from the Gaussian process described above. We generated an ensemble of size 19 by drawing sample paths from the Gaussian process introduced in Eq. (6). We chose $a = 0.4$ when drawing samples for the ensemble and $a = 0.3$ when drawing an observation. To generate an overdispersed ensemble, we chose to use a standard deviation of 0.2 at the constraining points and standard deviation of 0.1 for the observation (see Fig. 5). We switched the values of a and the standard deviation at constraining points in order to generate an underdispersed example. To generate an ensemble with bias, we simply shifted the functional values by a constant value of 0.23. We then generated the rank histograms using functional band depth for 2500 realizations.

Figure 6 demonstrates the resulting rank histograms. Notice that both underdispersed and biased ensembles result in a similar pattern in the rank histogram visualization. That is, a sloped rank histogram with numerous low ranks. The similar pattern of the rank histogram for underdispersed and biased ensembles stems from the fact that unlike univariate sorting, data-depth-based ranking is a *center-outward*



$a = 0.3$

$a = 0.4$

FIG. 5. A multivariate ensemble generated using a Gaussian process. The darker function corresponds to the mean of the ensemble, and the vertical lines correspond to one standard deviation of the noise introduced at the constraining points [(top) std = ± 0.1 and (bottom) std = ± 0.2]. By modifying the value of the standard deviation at the constraining points and the parameters of the covariance function, the dispersion can be controlled.

ordering (Thorarinsdottir et al. 2016; Wilks 2017). Therefore, by definition, the underdispersion and bias both result in numerous low rankings that are indistinguishable. It is worth mentioning that the same behavior has been reported for the MST rank histogram even though the MST rank histogram does not use a center-outward ordering (Gombos et al. 2007).

An important implication of this behavior is that a sloped multivariate rank histogram with many low ranks needs to be investigated further to better understand the nature of the miscalibration. In addition, a ∪-shaped multivariate rank histogram can no longer be interpreted as an indicator of underdispersion or bias. One can easily construct an example in which the combination of overdispersion and bias can generate a ∪-shaped multivariate histogram as follows.

Consider a bivariate ensemble forecast generated from a unimodal bivariate normal distribution:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{where} \quad \boldsymbol{\mu} = [2, 3], \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \qquad (8)$$

Assume that the observation is generated from a mixture model where the observation is a random draw from either of the following two distributions: $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ or $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_2)$ with equal probability where

$$\begin{cases} \boldsymbol{\mu}_1 = [2, 3] \\ \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix} \end{cases} \begin{cases} \boldsymbol{\mu}_2 = [2, 5] \\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \end{cases}. \qquad (9)$$

Note that the ensemble forecast is overdispersed compared to $\mathcal{N}_1$. Therefore, the observations drawn from $\mathcal{N}_1$ will mainly have high ranks. On the other hand, the ensemble is overdispersed and biased compared to $\mathcal{N}_2$, and, hence, the observations from $\mathcal{N}_2$ will be assigned low ranks. Figure 7 demonstrates the multivariate rank histogram using simplicial depth over 2500 realizations. Similarly a ∩-shaped multivariate rank histogram should not be interpreted as overprediction (or overdispersion) without further study of the nature of miscalibration.

The identifiability problem discussed above is a major drawback to multivariate rank histograms, and in particular, the center outward ordering introduces a unique identifiability problem for the data-depth-based multivariate rank histograms. However, the identifiability issue can be alleviated by augmenting a multivariate rank histogram with another data-depth-based diagnostic visualization called the DD plot.

DEPTH-VERSUS-DEPTH (DD) PLOT

The depth-versus-depth (DD) plot is a graphical inference tool that has been proposed in the statistical literature in order to graphically compare two ensembles and correspondingly the underlying distributions (Liu et al. 1999; Li and Liu 2004). Similar to rank histograms, specific patterns in the DD plot indicate specific kinds of differences between two ensembles, such as the presence of bias, overdispersion or underdispersion, and differences in the skewness and kurtosis. The definition of the depth-versus-depth (DD) plot is as follows.

Consider two ensembles in $\mathbb{R}^d$: $\mathscr{E}_1 = \{x_1, \ldots, x_n\}$ and $\mathscr{E}_2 = \{y_1, \ldots, y_m\}$. Each of these ensembles has its corresponding underlying probability distribution, F and G, respectively. A combined sample, $\{\mathscr{E}_1 \cup \mathscr{E}_2\}$, can be

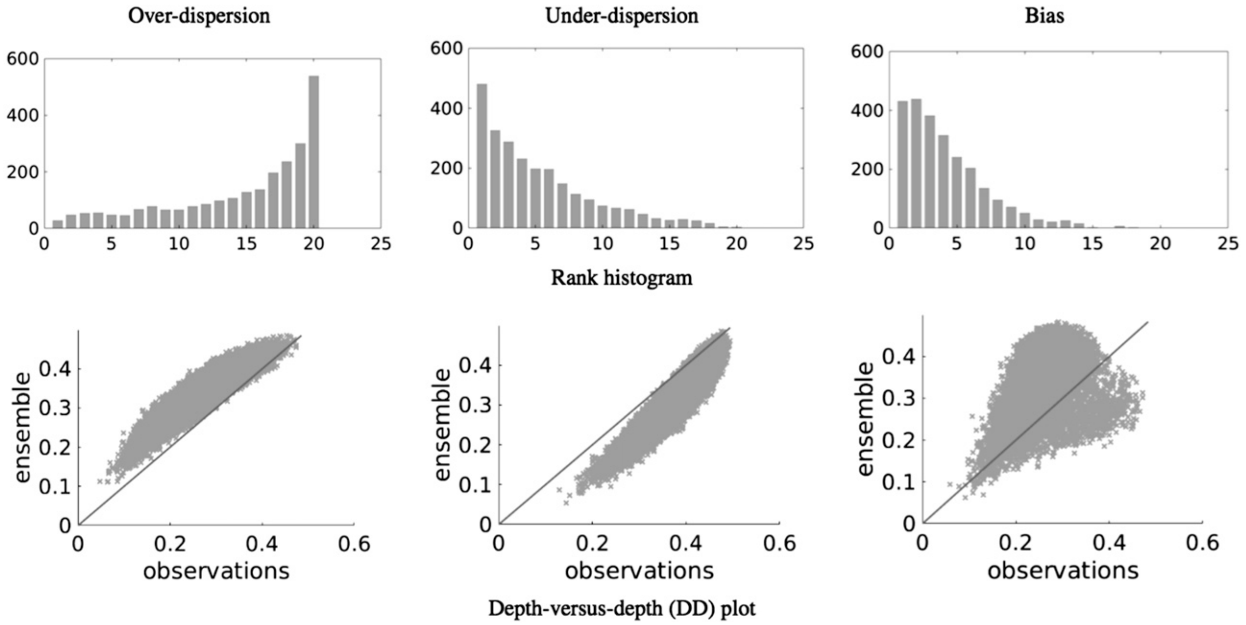


FIG. 6. In comparison to univariate rank histograms, multivariate rank histograms based on the concept of data depth have an inherent identifiability problem for underdispersed or biased ensembles. (top) Multivariate rank histograms generated for various types of miscalibration using functional band depth for forecast trajectories. (bottom) The DD plot for the same experiment. Unlike multivariate rank histograms, the DD plot can distinguish underdispersion from bias.

considered as pooling all the members from $\mathscr{E}_1$ and $\mathscr{E}_2$. If $\mathscr{E}_1$ and $\mathscr{E}_2$ have distinct members, $\{\mathscr{E}_1 \cup \mathscr{E}_2\}$ has $m + n$ elements. For each element in the combined sample, we can compute two depth values under the two corresponding empirical distribution F_n and G_m. By plotting the depth values of the combined sample, under the two corresponding empirical distribution, we can graphically compare two ensembles (or distributions). More specifically, the DD plot can be defined as in Liu et al. (1999):

$$\mathrm{DD}(F_n, G_m) = \{[D_{F_n}(x), D_{G_m}(x)], x \in \{\mathscr{E}_1 \cup \mathscr{E}_2\}\}, \tag{10}$$

where $D_{F_n}(x)$ and $D_{G_m}(x)$ correspond to empirical depth values with respect to F and G, respectively. Note that the DD plot is always a subset of $\mathbb{R}^2$ irrespective of the dimensionality of the original ensemble. It is easy to conclude that if $F = G$, the data depth values would be similar under each distribution and, hence, the plot of the depth values would closely follow the diagonal line from (0, 0) to (1, 1) in $\mathbb{R}^2$. Figure 8 demonstrates the typical patterns in the DD plot that are of interest for the evaluation of calibration.

The second row in Fig. 6 also demonstrates the DD plot visualization for the Gaussian process example described earlier using the functional band depth to compute the data depth values. In this case, $\mathscr{E}_1$ corresponds to the forecast ensemble and $\mathscr{E}_2$ corresponds to the observations. One can notice that unlike the multivariate rank histogram, the DD plot results in patterns consistent with Fig. 8 and can easily distinguish bias from underdispersion.

In addition to bias, underdispersion, and overdispersion, the DD plot is capable of revealing differences in skewness and kurtosis (Liu et al. 1999). Unlike previous patterns, revealing the difference in skewness and kurtosis requires that the two ensembles

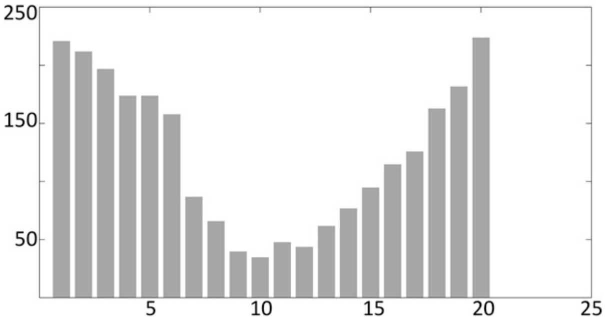
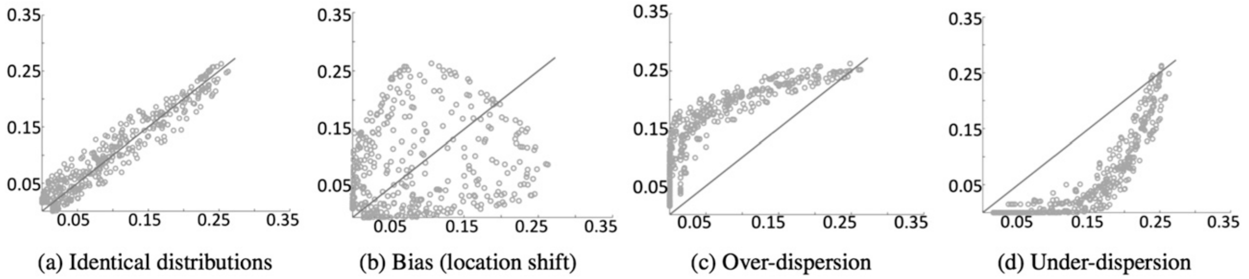


FIG. 7. A ∪-shaped multivariate rank histogram based on data depth analysis cannot be simply considered to indicate underdispersion. In this example, a controlled combination of bias and overdispersion has resulted in a ∪-shaped rank histogram.



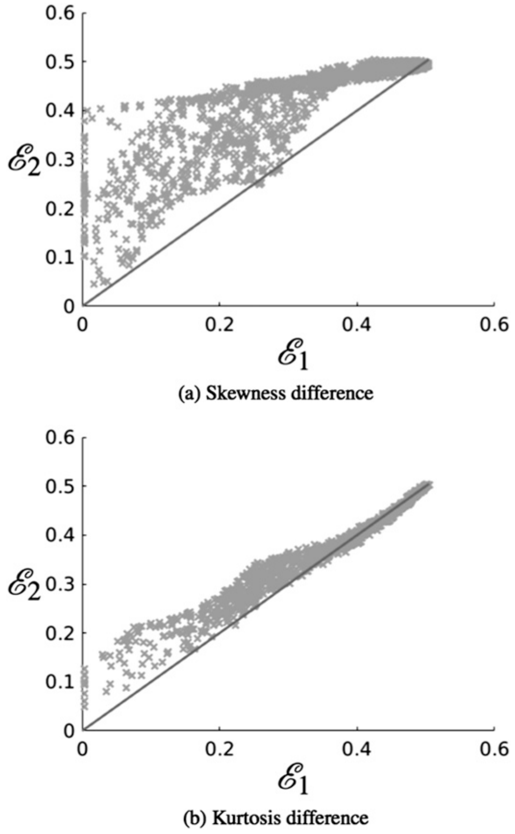
(a) Identical distributions (b) Bias (location shift) (c) Over-dispersion (d) Under-dispersion

FIG. 8. Each ensemble for this example is of size 200 drawn from a bivariate Gaussian distribution, and the simplicial depth was used to generate the DD plot. Specific patterns in the DD plot are associated with specific differences between the two ensembles.

have the same location or center (i.e., removal of any bias) and similar scale or dispersion. To demonstrate this capability of the DD plot, we consider ensembles of bivariate points for which the notion of scale or dispersion is well defined (Liu et al. 1999). Figure 9a demonstrates the DD plot where the two ensembles have different skewness. For the first ensemble $\mathscr{E}_1$, we used 500 realizations from a lognormal distribution with the mean value of (0, 0) and the standard deviation of $\delta = 1$, and for the second ensemble $\mathscr{E}_2$, we used 500 realizations from a lognormal distribution with the mean value of (0, 0) and the standard deviation of $\delta = 0.2$. These two ensembles have different skewness. After equalizing the scale of the two ensembles using the procedure proposed in Liu et al. (1999, their section 2), we used simplicial depth to produce the DD plot. There exists some similarity between the DD plot in Fig. 6 (overdispersion) and the DD plot in Fig. 9a. In both cases, the points are above the diagonal line. However, unlike the overdispersed case, in the presence of skewness difference, the points no longer form a half-moon shape and are scattered asymmetrically. That is, the values are more spread toward the lower-left corner of the DD plot.

To demonstrate the case in which the ensembles have different kurtosis, we have chosen 500 realizations from a uniform distribution for $\mathscr{E}_1$ and 500 realizations from a bivariate Gaussian distribution with the mean value of (0, 0) and the standard deviation of one for $\mathscr{E}_2$. The DD plot corresponding to these two ensembles using simplicial depth has been depicted in Fig. 9b. In this case, the points on the left (i.e., lower depth values) are shifted to one side of the diagonal line. As we move toward the right side of the DD plot (i.e., higher depth values), the points are following the diagonal line more closely. The patterns revealed in Fig. 9 are consistent with the patterns provided for skewness and kurtosis differences in Liu et al. (1999).

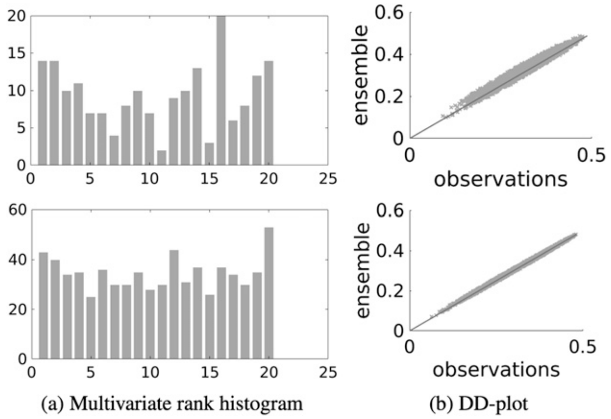
Finally, the DD plot is also less sensitive to ensemble size compared to rank histograms. The finite size of an ensemble can cause serious issues for the interpretation of a rank histogram (Hamill 2001). Figure 10 demonstrates an example where both the ensemble and observations are random draws from



(a) Skewness difference

(b) Kurtosis difference

FIG. 9. The DD plot reveals different patterns in the presence of (a) skewness difference and (b) kurtosis difference. The skewness difference looks similar to the overdispersion pattern. However, in the presence of skewness, the points no longer form a half-moon shape and are more spread toward the lower-left corner of the DD plot. In the presence of kurtosis difference, the points on the left are shifted to one side of the diagonal line. As we move toward the right portion of the DD plot, the points are following the diagonal line more closely.



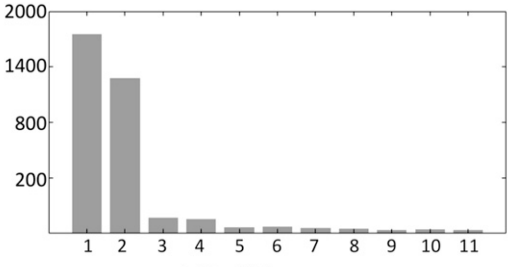
(a) Multivariate rank histogram (b) DD-plot

FIG. 10. The multivariate rank histogram and the DD-plot visualization of ensemble forecast trajectories and observations drawn from the same Gaussian process. (top) Generated using 200 realizations and (bottom) generated using 800 realizations. Note that the DD plot in both cases suggests the calibration of the ensemble forecasts with the observations, whereas the rank histogram fails to reveal the calibration due to sensitivity to the sample size.

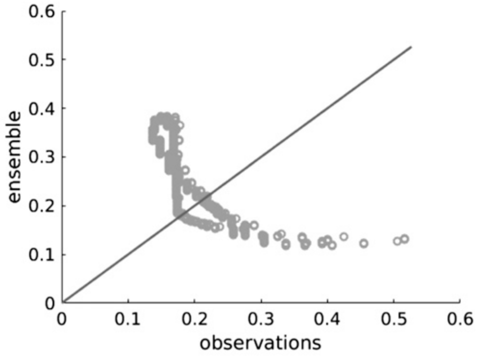
a single Gaussian process. Notice that the DD plot demonstrates the calibrated nature of the ensemble with only 200 realizations, whereas the rank histogram is far from uniform even with 800 realizations. It is important to note that both the DD plot and rank histogram use data depth. However, the goodness of fit to a uniform distribution (i.e., the underlying building block of a rank histogram) requires a large sample size to insure the reliability of a rank histogram (Elmore 2005).

4. Application to real ensemble data

In this section, we demonstrate the identifiability issue of the multivariate rank histogram discussed in section 3 for wind direction forecasts and observations for the Olympia, Washington, airport. The observations are from the hourly National Weather Service ASOS facility at the airport. The ensemble forecasts are from the NCAR experimental real-time convection-permitting ensemble prediction system (Schwartz et al. 2015; Sobash et al. 2016). An ensemble of 10 forecasts using a 3-km resolution WRF Model is initialized every day at 0000 UTC and forecasts are available hourly out to a lead of 48 h. For the verification example here, all lead times are lumped together into a single verification. When there are no missing observations, each ensemble forecast leads to the computation of 49 rank values including the initialization (analysis) time. The evaluation period begins at 0000 UTC 20 January 2016 and extends through 11 July 2016. We restrict our analysis during this period to hours for which there existed a valid wind observation value (i.e., we have eliminated the hours corresponding to missing data). There were approximately 2700 hourly observations during this period.



(a) Rank histogram



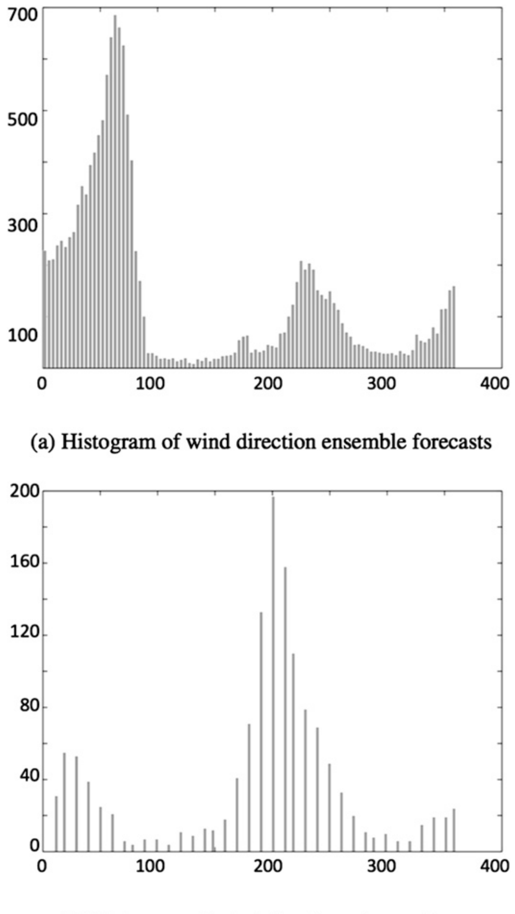
(b) DD-plot

FIG. 11. (a) Multivariate rank histogram of the wind direction angles at Olympia airport. (b) The DD-plot visualization of the same data. Both (a) and (b) suggest miscalibration of the ensemble forecast at this station, the DD plot in this figure demonstrates the presence of significant bias/shift, whereas the multivariate rank histogram has an identifiability issue. The one-sided pattern observed in (a) can indicate either underdispersion or bias.

Figure 11a shows the corresponding multivariate rank histogram using the simplicial depth preranking function. Note that the rank histogram is highly one sided, and suggests that the ensemble forecast is not calibrated. However, the nature of miscalibration is not clear in this case (see section 3). The misclassification can be due to underdispersion or bias. Figure 11b depicts the DD-plot visualization of this dataset. By comparing the pattern in Figs. 11b to 8, it is easy to conclude that the miscalibration is a result of the presence of significant bias in the ensemble. Visualizing the wind direction angles histogram also demonstrates the presence of significant bias in the ensemble forecast (see Fig. 12).

5. Conclusions

This paper studied and evaluated various properties of data-depth-based multivariate rank histograms. We



(a) Histogram of wind direction ensemble forecasts

(b) Histogram of wind direction observations

FIG. 12. The histogram of the wind direction data from Olympia Airport. (a) Histogram of the ensemble forecast angles retrieved from NCAR during the study period reported in section 4. The ensemble includes 10 members at each forecast hour. (b) Histogram of the wind direction angles observed at Olympia Airport during the same period. The shift in the peak of the two histograms suggests the presence of bias or shift in the ensemble forecast compared to the observed angles.

first reviewed an important property that a prerank function must satisfy in order to provide accurate rank statistics. We also reviewed various patterns that emerge in the presence of miscalibration and the identifiability issue of multivariate rank histograms. We proposed to complement the multivariate rank histogram using a well-known multivariate statistics inference tool based on rank statistics called the depth-versus-depth (DD) plot to alleviate the identifiability issue of multivariate rank histograms. Unlike the multivariate rank histogram, the DD plot does not have an identifiability problem in the presence of bias or underdispersion. Finally, we validated the relevance of the studies provided by demonstrating the identifiability issue using observed and forecasted wind directions at a specific U.S. Pacific Northwest station.

This paper studied and evaluated multivariate rank histograms from a qualitative and visual representation point of view based on the patterns that emerge. However, an interesting line of future work would be to conduct more quantitative studies in order to ensure the reliability of a multivariate rank histogram based on the notion of data depth.

Acknowledgments. We thank Glen Romine and Ryan Sobash for help compiling the ensemble forecasts. The first author is sponsored in part by National Science Foundation (NSF) Grant IIS-1212806. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Elmore, K. L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting*, **20**, 789–795, doi:10.1175/WAF884.1.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

——, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, **17**, 211–235, doi:10.1007/s11749-008-0114-x.

Gombos, D., J. A. Hansen, J. Du, and J. McQueen, 2007: Theory and applications of the minimum spanning tree rank histogram. *Mon. Wea. Rev.*, **135**, 1490–1505, doi:10.1175/MWR3362.1.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835–1851, doi:10.1175/1520-0493(2000)128<1835:ACOPFF>2.0.CO;2.

Li, J., and R. Y. Liu, 2004: New nonparametric tests of multivariate locations and scales using data depth. *Stat. Sci.*, **19**, 686–696, doi:10.1214/088342304000000594.

Liu, R. Y., 1990: On a notion of data depth based on random simplices. *Ann. Stat.*, **18**, 405–414, doi:10.1214/aos/1176347507.

——, and K. Singh, 1992: Ordering directional data: Concepts of data depth on circles and spheres. *Ann. Stat.*, **20**, 1468–1484, doi:10.1214/aos/1176348779.

——, J. M. Parelius, and K. Singh, 1999: Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Stat.*, **27**, 783–840, doi:10.1214/aos/1018031260.

López-Pintado, S., and J. Romo, 2009: On the concept of depth for functional data. *J. Amer. Stat. Assoc.*, **104**, 718–734, doi:10.1198/jasa.2009.0108.

——, and ——, 2011: A half-region depth for functional data. *Comput. Stat. Data Anal.*, **55**, 1679–1695, doi:10.1016/j.csda.2010.10.024.

——, Y. Sun, J. K. Lin, and M. G. Genton, 2014: Simplicial band depth for multivariate functional data. *Adv. Data Anal. Classif.*, **8**, 321–338, doi:10.1007/s11634-014-0166-6.

Mirzargar, M., R. T. Whitaker, and R. M. Kirby, 2014: Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Trans. Vis. Comput. Graph.*, **20**, 2654–2663, doi:10.1109/TVCG.2014.2346455.

Pinson, P., and R. Girard, 2012: Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy*, **96**, 12–20, doi:10.1016/j.apenergy.2011.11.004.

Rasmussen, C. E., and C. K. Williams, 2006: *Gaussian Processes for Machine Learning*. The MIT Press, 266 pp.

Rousseeuw, P. J., and I. Ruts, 1996: Algorithm as 307: Bivariate location depth. *J. Roy. Stat. Soc. Ser. C Appl. Stat.*, **45** (4), 516–526.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

Serfling, R., and Y. Zuo, 2000: General notions of statistical depth function. *Ann. Stat.*, **28**, 461–482, doi:10.1214/aos/1016218226.

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, doi:10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

Thorarinsdottir, T. L., M. Scheuerer, and C. Heinz, 2016: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.*, **25**, 105–122, doi:10.1080/10618600.2014.977447.

Whitaker, R. T., M. Mirzargar, and R. M. Kirby, 2013: Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Trans. Vis. Comput. Graph.*, **19**, 2713–2722, doi:10.1109/TVCG.2013.143.

Wilks, D. S., 2004: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.*, **132**, 1329–1340, doi:10.1175/1520-0493(2004)132<1329:TMSTHA>2.0.CO;2.

——, 2017: On assessing calibration of multivariate ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 164–172, doi:10.1002/qj.2906.