On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind-speed Forecasts from Fine-scale Ensemble Reforecasts

Caren Marzban[1,2], Ranran Wang[1], Fanyou Kong[3], Stephen Leyton[4]

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

[2] *Applied Physics Laboratory, University of Washington, Seattle, Washington*

[3] *Atmospheric Technology Services Company, Norman, Oklahoma*

[4] *Duke Energy Corp., Charlotte, North Carolina*

ABSTRACT

The rank histogram (RH) is a visual tool for assessing the reliability of ensemble forecasts, i.e., the degree to which the forecasts and the observations have the same distribution. But it is already known that in certain situations it conveys misleading information. Here, it is shown that a temporal correlation can lead to a misleading RH, but such a correlation contributes only to the sampling variability of the RH, and so it is accounted for by producing a RH which explicitly displays sampling variability. A simulation is employed to show that the variance within each ensemble member (i.e., climatological variance), the correlation between ensemble members, and the correlation between the observations and the forecasts, all have a confounding effect on the RH, making it difficult to use the RH for assessing the climatological component of forecast reliability. It is proposed that a "Residual" Q-Q plot (denoted R-Q-Q plot) is better suited than the RH for assessing the climatological component of forecast reliability. Then, the RH and R-Q-Q plots for temperature and wind-speed forecasts at 90 stations across the continental US are computed. A wide range of forecast reliability is noted. For some stations, the non-reliability of the forecasts can be attributed to bias and/or under- or over- climatological dispersion. For others, the difference between the distributions can be traced to lighter or heavier tails in the distributions, while for other stations the distributions of the forecasts and the observations appear to be completely different. A spatial signature is also noted and discussed briefly.

1. Introduction

Given a set of observations, and an M-member ensemble of corresponding forecasts, the quality of the forecast can be assessed in at least two ways: their accuracy, and their reliability. The former gauges the degree to which the forecasts agree with the observations, on a case-by-case basis. The latter generally measures whether or not the observations and the forecasts could have come from the same distribution or process. Under the null hypothesis that they do, the rank of the observation with respect to the M forecasts should follow a uniform distribution over the integers 1,2,..., M+1. Given a set of observations and forecasts, then, one can compare the histogram of the ranks - called a Rank Histogram (RH) - with a uniform distribution. If the RH is inconsistent with a uniform distribution, then one may conclude that the observations and the forecasts could not have come from the same distribution. Three common deviations from uniformity are referred to as trend, U-shaped, and dome-shaped.[1] A trend in the RH typically implies a bias in the forecasts. A U-shaped RH arises if the observation is generally outside of the range of forecasts, thereby giving it either high or low rank. By contrast, a dome-shaped RH occurs if the rank of the observations is generally in the mid-range values. The terms under-dispersive and over-dispersive are used to describe the corresponding ensembles. Forecasts leading to a flat RH are expected to be reliable.

Johnson and Bowler (2009) discuss how the notion of reliability consists of two independent components - one that assesses the ensemble variance and another which gauges the

[1] In this work, no distinction is made between U-shaped RH and V-shaped RH as described by Jolliffe and Primo (2008).

climatological variance of the forecasts.[2] They propose two conditions which must be met for forecasts to be reliable 1) Ensemble dispersion must be equal to the mean squared error of the forecasts, and 2) the climatological variance of the forecasts must be equal to that of the observations. Generally, RHs are used for assessing the former component or reliability, and in this paper a different plot (denoted R-Q-Q) is used for assessing the climatological component.

Hamill (2001) points out a number of situations wherein reliable forecasts can give rise to a non-flat RH, and conditions under which a flat RH may be produced from forecasts which are known to come from a different distribution than the observations. For example, Hamill considers normal distributions, and shows that a uniform RH may arise when observations are drawn from a standard normal, while the forecasts are taken from a mixture of normal distributions with different means and standard deviations. In other words, even though the observations and forecasts are drawn from different distributions, the RH may be mostly uniform-looking. He also shows that a U-shaped RH may arise not only when the forecasts may be under-dispersed, but also when they are conditionally biased. It is also shown that non-random sampling may falsely lead to a uniform-looking RH. In short, the shape of a RH is affected by numerous factors, thereby making it difficult to interpret. The consequences of this conclusion in the context of probabilistic forecasts are examined by Gneiting, Balabdaoui, and Raftery (2007).

Three other factors which affect the shape of the RH and therefore obfuscate its interpretation are 1) temporal correlation, 2) correlation between ensemble members, and 3) correlation between ensemble members and observation. Temporal correlations affect the

[2] We are grateful to Dan Wilks for pointing out this paper to us.

shape of the RH and therefore its ability to assess ensemble dispersion. As shown here, the latter correlations affect the interpretability of the RH in terms of climatological dispersion, but not ensemble dispersion (Wilks 2010).

Temporal correlations exist in all forecasts, by virtue of the dynamics of the underlying processes. Temporal correlations also affect the resulting probabilistic forecasts (Gneiting, Balabdaoui, and Raftery 2007), but a method for taming the visual effects of temporal correlations is to supplement the RH with a measure of the sampling variability of the frequency of each rank - for example, boxplots. Here, both the traditional RH and its boxplot variety are referred to as RH.

The strength of the correlation between ensemble members depends on whether the ensemble consists of different realizations (e.g., initial conditions) of the same model, or if it is a multi-model ensemble composed of truly different models. In practice, many ensembles, including the one examined here, consist of a mixture of the two types. As for the correlation between ensemble members and the observation, note that it ought not affect an assessment of reliability (as defined here), because the latter refers only to the extent that the forecasts and observations come from the same distribution. But, as shown below, it does affect the RH.

The correlation between ensemble members is qualitatively a different phenomenon than that of under- or over-dispersion. Indeed, as shown here, a correlation between ensemble members may give rise to a U-shaped or dome-shaped RH even when the variance *within* each ensemble member (i.e., climatological variance) is equal to the variance of the observation. In other words, a deviation from uniformity in a RH may be attributed to two different sources: variance within, and correlation between ensemble members; and it is difficult to determine the contribution of each. Moreover, it is shown that this ambiguity is further confounded because it also depends on the correlation between the forecasts and the observation. All of these phenomena are shown below. The conclusion is that the RH cannot be uniquely interpreted or diagnosed in terms of climatological dispersion. Wilks (2010) examines the RH's ability to diagnose ensemble dispersion.

A tool for assessing the climatological component of reliability of ensemble forecasts is the Q-Q plot (Wilks 2006), wherein the quantiles of one quantity (Y) are plotted against the quantiles of another (X). If the Q-Q plot displays a linear pattern, then the y-intercept and the slope of that pattern generally assess the relative location (e.g., mean) and relative climatological variance of the two quantities, respectively. A linear and diagonal pattern, therefore, suggests that the two quantities come from the same distribution. If the slope of the linear pattern is greater (less) than 1, then Y has a larger (smaller) variance than X. A vertical shift up (down) suggests that Y has a larger (smaller) mean than X. Deviations from linearity are also interpretable. A U-shaped (dome-shaped) pattern corresponds to a situation where Y is positively (negatively) skewed relative to X. An S-shaped pattern implies that Y has lighter tails than X; and heavy tails in Y (relative to X) produce a Q-Q plot that turns up for larger quantiles and down for lower quantiles.[3]

Evidently, the Q-Q plot provides similar diagnostic information as the RH, and this has been noted previously (Hamill 2001). But the two diagrams assess different components of

[3] Note that the curves in a Q-Q plot are necessarily monotonic, and as such cannot be truly U-shaped, dome-shaped, or even S-shaped. These terms are being used here only as a concise description of the general shape of the curves. For example, the description "U-shaped" technically refers to a curve with larger slope at the ends compared to the middle.

the reliability. One difference is that whereas a single RH is produced for an M-member ensemble, M Q-Q plots are required for such an ensemble. This is not a prohibitive feature of Q-Q plots, because one can display all M Q-Q plots on a single figure. More importantly, the Q-Q plot is not affected by the above-mentioned correlations. (The effect of temporal correlations on Q-Q plots is not addressed here.) Therefore, in assessing the reliability of realistic forecasts, although the RH may provide some useful information, it is advisable to examine Q-Q plots as well.

In order to better utilize the geometric landscape of the Q-Q plot, the plot is revised slightly; instead of plotting quantiles of Y versus quantiles of X, it is proposed to plot the difference or the "Residual" of the two quantiles, on the y-axis. Loosely speaking, this R-Q-Q plot is a Q-Q plot which has been rotated clockwise by 45°.

In an earlier version of this work (Marzban et al. 2010), it was reported that the RH should be abandoned in favor of the R-Q-Q plot, because of the confounding effects of the above-mentioned correlations on the RH. That conclusion is false, and originates from a misapplication of the RH to the assessment of climatological dispersion; Wilks (2010) shows that the RH does correctly assess ensemble dispersion even in the presence of correlations. Therefore, given that the RH and the R-Q-Q plot gauge different facets of forecast quality, both are examined here.

The outline of this paper is as follows: Simulated data are employed to illustrate how the shape of the RH changes with respect to temporal correlations, correlations between ensemble members, and the correlation between ensemble members and the observation. It is shown that temporal correlations can be rendered harmless (i.e., visually not misleading), if one displays the sampling variability of the histogram with boxplots. It is also shown that the correlation between ensemble members, and the correlation between them and the observation, have a confounding effect on the RH, while the R-Q-Q plot is not adversely affected by these correlations. Turning to real data, the boxplot version of the RHs for temperature and wind-speed forecasts at 90 stations are computed. It is found that the RHs are mostly U-shaped plus trend (i.e., not dome-shaped), as anticipated from the simulation study. Then to properly diagnose both components of reliability of the forecasts, RH and R-Q-Q plots are produced for all 90 stations. The geographical characteristic of the RHs and the R-Q-Q plots is also examined briefly.

2. Data Description

Simulated and real data sets are examined here. The former is designed to illustrate the effect on the RH of the aforementioned correlations. The real data sets consist of 19 years (1987-2005) of 48-hr forecasts of temperature, wind speed, and corresponding observations at 90 stations across the continental US (CONUS). The data are not daily but are taken at 5-day intervals, leading to 1387 cases. Each case consists of one observed value and ten forecast values from a 10-member ensemble.

The 19-year forecast dataset was generated retrospectively (i.e., reforecast) by using a regional fine-scale ensemble forecasting system based on the Weather Research and Forecast (WRF) model (WRF-ARW Version 3.0.1). The ensemble system consists of ten members, each with unique initial perturbations and varying physics options and land-use tables. The initial perturbations are bred vectors produced by 6-hourly breeding cycle, using the North America Regional Reanalysis (NARR) as background fields. In order for the fine-scale ensemble forecast system to have consistent lateral boundary perturbations, a three-domain, two-way nesting framework was employed (Figure 1). The outer domain reforecasts, with a $135\,km$ horizontal grid spacing, have fixed lateral boundary conditions (LBCs) from NARR. The ensemble reforecasts in the innermost domain, with a $15\,km$ horizontal grid spacing covering the CONUS, are the data used in this study. 48-h reforecasts, initialized at 0000 UTC, were produced every five days starting from 1 January 1987 through 31 December 2005. WRF outputs were written every 3 hours. Selected forecast variables such as temperature (in $^{\circ}C$, and wind speed (in knots) were interpolated to 90 weather stations across the CONUS. These are the stations examined in this paper, and their names and approximate locations are shown in Figure 7.

Hourly surface observations over the 19 year period are obtained for each of the 90 locations. Some data preprocessing has been performed to identify instances of bad or missing data, and then to replace them with surrogate values. Methods involving climatology, conditional climatology, persistence and nearest-neighbor are used to derive surrogate values for each weather parameter. The preprocessing of the hourly surface observations leads to a uniform dataset with no gaps.

It is worth mentioning that the simulations and the real data differ in one important way: In the former the ensembles are effectively unlabeled, and as such can be thought of as different perturbations of the initial conditions. The real data, however, are from a multi-model ensemble. The simulation is based on this simplistic assumption in order to allow for a transparent exposition. More realistic simulations may be considered that take into account differences between the ensemble members.

3. Method

The main goal of this article is to assess the climatological component of the reliability of realistic forecasts for 90 stations across the US. Given that reliability is normally assessed through RHs, an attempt is made to render the RHs as interpretable as possible. To that end, first, a simulation is set up to quantify how the RH varies with temporal correlation, correlation between ensemble members, and the correlation between the ensemble members and the observation. Then, RHs are computed for all 90 stations, for temperature and wind-speed forecasts. All RHs are computed in the boxplot variety in order to tame the visual effects of temporal correlations. Upon illustrating that the RHs cannot be used to interpret the climatological component of reliability, the analysis of reliability is performed in terms of R-Q-Q plots.

The effect of bias and variance across forecasts is examined by Hamill (2001). There, it is assumed that forecasts are drawn from a normal distribution with parameters μ and σ; while the observations are drawn from a standard normal; i.e.

$$\begin{aligned} y_i &\sim N(\mu, \sigma^2), \quad i = 1, 2, 3, ..., M \ , \\ y_0 &\sim N(0, 1) \ , \end{aligned} \tag{1}$$

where $y_i \ \ (i = 1, ..., M)$ refers to the forecast from the i^{th} ensemble, M denotes the number of members in the ensemble (here 10), and y_0 is the corresponding observation. For brevity, the index representing the case is not shown. In this framework, μ measures the bias of the ensemble, and σ its variability. Hamill examines the behavior of the RHs jointly, by varying μ and σ. He shows that bias leads to RHs which display a trend. He also shows that if the variance of forecasts is smaller (larger) than that of the observations, one generally obtains a U-shaped (dome-shaped) RH. These are common and standard ways of interpreting/diagnosing the RH.

Note that in Hamill's work the simulated forecasts are assumed to have no correlations of any kind. Then, ensemble variance and climatological variance are equal, both controlled by σ. As shown below, if the correlation between the ensemble members is not comparable to that between the forecasts and the observation, the above interpretations are no longer valid regarding the climatological variance.

a. Temporal Correlation

All of the misleading behaviors described in the Introduction are exacerbated in the presence of a temporal correlation. In the presence of a temporal correlation, a RH may appear to have a trend, suggestive of biased forecasts, even when there is no bias in the forecasts; this is illustrated below. The reason is that a temporal correlation can exaggerate the effects of sampling variability. Whereas sampling variability leads to "random" fluctuations about a uniform distribution if the forecasts are independent, in the presence of a temporal correlation the fluctuations become more "systematic." In short, a temporal correlation can lead to a RH which looks unambiguously non-uniform, in spite of the reliability of the forecasts. One remedy which brings visual interpretability back to the RH is to supplement it with some measure of sampling variability. Therefore, in the present work, the RH is displayed as a series of boxplots. Temporal correlations, then, typically affect the size of the boxplots, but the eye can readily assess the shape of the resulting pattern.

To examine the role of temporal correlations, a first-order autoregressive process is employed. In particular, a time series of length 1000 is generated via an AR(1) process, defined as a time series satisfying $x(t) = \phi\, x(t-1) + \epsilon(t)$, where $x(t)$ denotes the value of the time series at time t, and ϵ is a normally distributed random variable with mean = 0, and variance = 1. The parameter ϕ controls the temporal correlation. It can be shown that the autocorrelation function for such a process is given by ϕ^l, where l is the lag (i.e., the x-axis of the autocorrelation function). Figure 2 shows several examples of such time series, all with mean = 0 and variance = 1, but for different values of ϕ. Evidently, $\phi = 0$ displays no correlation, while a larger (and positive) ϕ results in a time series with persistent local trend over some time interval; the time series at any given time is likely to be near its state at a previous time. By contrast, an anti-correlated time series (with negative ϕ) yields a time series which tends to spend little time in any given region; it is likely to be far from where it was a unit of time ago if it is far from the mean.

AR(1) processes have been examined by Wilks (2004) in the context of statistical tests of uniformity performed on RHs. The existence of temporal (or auto-) correlation in the data renders most statistical tests misleading unless the correlation is accounted for in some manner. Here, the emphasis is on the visual features of the RH, as opposed to the quality of some statistical test performed on it. However, the shape can also be misleading in the presence of a temporal correlation. The main reason is that a temporal correlation can induce false patterns which are in fact only due to random sampling variability. In the absence of a temporal correlation, sampling variability generally leads to a RH which deviates from uniformity in some random fashion. In the presence of a temporal correlation, however, the fluctuations about a flat histogram are no longer random in appearance. Figure 3 (top) shows an example of the RH for a set of forecasts and observations with $\phi = 0.7$. This particular realization is somewhat extreme in terms of displaying a pronounced trend. Although other realizations may not be as extreme, due to the unambiguous trend in this RH, one may be tempted to conclude that the forecasts are biased. But this conclusion is known to be false, because the forecasts and observations are, in fact, taken from the same distribution. The non-uniform appearance of the RH is due to temporal correlation.

One method for taming the visual effects of sampling variability in a RH is to explicitly display that variability. For simulated data, it is relatively straightforward to estimate the variability. All that is necessary is multiple realizations of the AR(1) process. As such, one can estimate the distribution of the frequency of each rank in the RH. Said differently, one can generate the empirical sampling distribution of the frequency for each rank. Then a boxplot can be used to summarize that distribution. Consequently, a RH can be displayed as a set of boxplots, one for each rank. The bottom panel in Figure 3 shows the resulting RH for 20 different realizations from the same distribution which gives rise to the RH in the top panel. The boxplot version of the RH makes it abundantly apparent that the top RH is an extreme realization, and that the bottom RH is consistent with the uniform distribution. In short, whereas the top RH would lead to the wrong conclusion that the forecasts are biased, the bottom RH implies the correct conclusion that the forecasts and observations are from the same distribution.

For real data, boxplots are produced via bootstrap (Efron and Tibshirani 1998). In the traditional bootstrap the unit for re-sampling is an individual case, and the cases are assumed to be independent. Given the existence of temporal correlation in the times series, the traditional bootstrap cannot be used. Instead, the method of block bootstrap is used (Efron and Tibshirani 1998; Bühlmann 2002; Lahiri 2003; Politis, Romano, and Wolf 1999);

there the unit for resampling is a block of the time series, and it is the blocks which are assumed to be independent. In this work, the size of the block is taken to be a season (i.e., about 3 months), in order to assure that the blocks are not correlated through seasonal patterns. It has been confirmed that the results are relatively insensitive to the size of the block. An example of both the traditional and boxplot version of the RH for a single station (KLAX) are shown in Figure 4 (top row).

In order to avoid being visually misled by sampling variability, in the presence of temporal correlation, the RHs produced here are all of the boxplot variety. This tames the effect of temporal correlations.

b. Residual Q-Q Plots

The RH in Figure 4 is difficult to interpret. There is a hint of a trend, implying bias (i.e., forecasts higher than observations); and one may even argue that there is a slight U-shaped pattern, suggesting ensemble under-dispersion of the forecasts. Of course, none of these features are unambiguous.

The climatological component of reliability is assessed in the corresponding Q-Q plots also shown in Figure 4 (bottom, left). All are mostly linear, suggesting that the forecasts in each ensemble member appear to have approximately the same distribution as that of the observations. However, given that the slope of most of the Q-Q plots is larger than 1, it follows that the climatological variance (within) most ensemble members is larger than the variance of the observations. As such, one may say that the forecasts are over-dispersed in the climatological sense.

It is convenient to revise the Q-Q plot in order to expose more structure. Instead of plotting the quantiles of the forecasts versus that of the observed, one may plot the difference or the residuals of the quantiles (quantiles of forecast - quantiles of observed) versus the quantiles of the observed. The resulting plot is effectively a Q-Q plot about the diagonal line. In this version of the Q-Q plot, denoted R-Q-Q plot, bias is reflected as a vertical shift (as in the Q-Q plot), but over- or under-dispersion correspond to a positive or negative slope, respectively. The R-Q-Q plot for KLAX is shown in the bottom-right panel of Figure 4.

As in RHs, the shape of Q-Q plots (and hence R-Q-Q plots) can be interpreted in terms of the shape of the distributions. An S-shaped R-Q-Q plot implies that the distribution of the forecasts has lighter tails than that of the observed. Some R-Q-Q plots resemble an S-shape that has been reflected about the x-axis; in that case, the distribution of the forecasts has heavier tails than that of the observed. The skewness of the underlying distributions is also reflected in R-Q-Q plots: If the distribution of the forecasts is positively (negatively) skewed relative to that of the observed, then the corresponding R-Q-Q plot is U- (dome-) shaped.

c. *Correlation Between Ensemble Members, and Between Ensemble Members and Observation*

In this section a simulation is set up to examine how the shape of the RH varies with the correlation between ensemble members, and that between ensemble members and observation. The two correlations are varied jointly.

To generate data from an M-member ensemble plus an observation, samples are drawn from an $(M + 1)$-dimensional multivariate normal distribution with a certain structure.

Specifically,

$$(y_0, y_1, y_2, ..., y_M) \sim MVN((0, \mu, \mu, ..., \mu), \Sigma) \ , \tag{2}$$

where the mean of the observations is set to 0, and all the M members of the ensemble are assumed to have the same mean, μ. Σ is the covariance matrix for the forecasts and observations, and with little loss of generality, it is parametrized to have the following structure:

$$\Sigma = \begin{pmatrix} 1 & R\sigma & R\sigma & R\sigma & ... & R\sigma \\ & \sigma^2 & r\sigma^2 & r\sigma^2 & ... & r\sigma^2 \\ & & \sigma^2 & r\sigma^2 & ... & r\sigma^2 \\ & & & \sigma^2 & ... & r\sigma^2 \\ & & & & & \sigma^2 \end{pmatrix} ,$$

where σ is the climatological variance of the forecasts (i.e. *within* each ensemble member), and it is assumed to be the same for all the members. The variance of the observation is set to 1. R measures the correlation between each member and the observation, and it too is assumed to be the same for all the members. Finally, r gauges the common correlation between ensemble members.

This model has four parameters μ, σ^2, R, and r. If $R = r = 0$, then the parameters μ and σ^2 control bias and dispersion, respectively, just as they do in Equation (1); and they affect the RH in the well-known manner described previously. It is worth emphasizing that σ controls the standard deviation *within* each ensemble member. The ensemble standard deviation as computed *across* the 10 ensemble forecasts, for each case, is not controlled in this simulation; its value is determined by the choice of σ, R and r.[4]

[4] The ensemble variance is given by $(1-r)\,\sigma^2$, and R enters the formulation through the comparison of this ensemble variance with the mean squared error of the forecasts.

To examine the dependence of the RH on these parameters two experiments are performed. In the first simulation, μ and σ are varied, just as they are in Hamill (2001), but for $R = r = 0.9$. This experiment allows one to examine how the standard interpretation of the RH (in terms of bias and climatological variance) is affected by the realistic situation where the forecasts and the observations are mutually and highly correlated.

In another simulation, μ and σ are set to 0 and 1, respectively, while R and r are varied. In this case, each ensemble member has the same mean and climatological variance as the observation, and so, one can examine the effect on the RH from only correlations between ensemble members and/or observation.

Figure 5 shows the results for the first simulation. This figure is the analog of Figure 1 in Hamill (2001), except here $R = r = 0.9$, whereas $R = r = 0$ in Hamill. Interestingly, when $\mu = 0, \sigma = 1$, the RH is uniform. Also, as in the $R = r = 0$ case, $\mu > 0$ manifests itself as a trend in the RH. However, whereas $\sigma > 1$ in Hamill's simulation gives rise to a dome-shaped RH, in the current simulation when $R = r = 0.9$ the RHs are mostly U-shaped. In other words, when there exists a correlation between ensemble members and an equal correlation between ensemble members and the observation, climatological over-dispersion of the forecast does not manifest itself as a dome-shaped RH. This conclusion persists for other values of $R = r$, although to a lesser degree (not shown here). It is worth mentioning that the values of R and r in realistic temperature data (below) are in the 0.9 range.

The results of the second simulation are shown in Figure 6. Recall that in this simulation, $\mu = 0$ and $\sigma = 1$, and R and r are varied. Evidently, even if R and r are nonzero, the RH is uniform as long as $R = r$. As such, the RH can still be used to interpret climatological dispersion. However, in realistic situations ensemble members often tend to be more similar to each other than to the observation. In this situation, $R < r$, and all of the RHs in Figure 6 are U-shaped. Therefore, one should not be surprised to find an abundance of U-shaped RHs in realistic situations, even when the climatological variance of the forecasts is equal to that of the observation. As for ensemble dispersion, its value is consistent with the U or dome shape of the RH; this issue is examined by Wilks (2010) using the multivariate model considered here. It is also interesting to point out that for small values of R (say 0.1), the uniformity of the RH is somewhat robust with respect to different values of r. However, when R is large (e.g., 0.7), then a small change in r from 0.7 to 0.9 quickly takes the RH from dome-shaped to U-shaped. (The blank panels in Figure 6 correspond to values of R and r which are not allowed if Σ is to be a covariance matrix.)

The main conclusion of these simulations is that if the correlation between ensemble members and the observation is generally different from that between ensemble members, then the RH cannot correctly assess the climatological component of reliability. For realistic ensembles, where both correlations are large (~ 0.9), the RH is then expected to be U-shaped *by default*, and so cannot assess the climatological component of forecast reliability. And if the ensemble forecasts are more similar to each other than to observation, then the RHs are still U-shaped, even though there exists no climatological over-dispersion. This expectation is confirmed in the next section, and for these reasons, the climatological component of reliability of the forecasts is assessed with R-Q-Q plots.

4. Results From Real Data

Figure 7 shows the RHs for the temperature forecasts. It can be seen that the RHs for nearly all stations display a combination of a trend and a U-shape. Almost all RHs have a positive trend, suggesting that the forecasts are negatively biased (i.e., lower than observed, on the average). One clear exception is KLAX for whom there is a negative trend; but as already discussed above, this trend does not imply that the forecasts are positively biased.

Figure 8 shows the R-Q-Q plots for all 90 stations. The most prominent feature across the 90 stations is that very few display a linear pattern; but when they do, the linear pattern is centered about the x-axis, and has no slope. In other words, only a few stations produce reliable forecasts in the climatological sense. They include KHSV, KGSO, and KRDU, with a few more scattered across the country; but even these have problems for extreme values, because the tails of the R-Q-Q plots tend to deviate from the horizontal line.

It is possible to continue with the analysis of the R-Q-Q plots in a qualitative fashion, acknowledging that the ensuing conclusions are only approximately true. For example, it is evident that there exists a negative slope associated with the Northwest region. This implies that the forecasts are generally climatologically under-dispersed. Many of these stations also display a downward vertical shift (e.g., KMFR), which implies that the forecasts are negatively biased (i.e., are lower than observed, on the average). Interestingly, the most Northwest station (KSEA) does not display much bias or dispersion error. There are also some stations on the Eastern coast which show this type of dispersion error (e.g., KORF, KACY).

Some stations show no sign of over- or under- climatological dispersion, but they do display a strong negative bias. Although stations of this type do appear in the Western region (e.g., KLAS), a majority of them occur in the Northeastern region. Stations in the middle of the US, as well as those in the Southeast, generally appear to have reliable forecasts, in the sense that they do not suffer from unambiguous bias or climatological dispersion error.

There is more information in these R-Q-Q plots which can help in better diagnosing the source of the errors. Figure 9 presents a sample of four R-Q-Q plots, demonstrating some of the additional information contained therein. The R-Q-Q plots for KBUF are relatively linear, and with a slope of zero, suggesting that the forecasts and observations come from the same distribution, different only in their mean parameter. The R-Q-Q plot for one ensemble member, however, does appear to fall on the x-axis. So, the distribution of the forecasts from that ensemble member is particularly similar to that of the observed. In this way, one can examine "the best" member(s) among the ensemble.

The shape of the R-Q-Q plots for KBIL suggests that the forecasts have a light tail relative to the observation. The U-shaped R-Q-Q plots for KDFW suggest that the corresponding forecasts are positively skewed relative to the observed; it is worth noting that two of the ensemble members have R-Q-Q plots with a different pattern. It would be interesting to examine these two members more closely. KSFO displays another informative feature; the vertical spread of the R-Q-Q plots between ensemble members is uncharacteristically small. In other words the ensemble members have comparable biases.

If one were to perform a qualitative ranking of the stations in terms of their temperature R-Q-Q plots, KMOB, KLNK, KDSM, KMSP, and KGSO would be among the best five. The worse five stations would be KGGW, KMSO, KBIL, KMFR, and KSFO, all in the Northwest region. It should be emphasized that given the multi-faceted nature of R-Q-Q plots, these rankings are necessarily qualitative.

Figure 10 shows the RHs for wind speed. Consistent with the simulation study performed above, the RHs are mostly U-shaped plus a trend. The correlation matrix for the ensemble forecasts (not shown) confirms the existence of a large correlation between ensemble members at all stations.

The R-Q-Q plots for wind-speed (Figure 11) display much less spatial structure than those of temperature (Figure 8). Two of the more interesting stations are KSFO and KLAX. The R-Q-Q plots for KSFO clearly imply that the forecasts and the observations come from different distributions. Moreover, there is very little spread across the R-Q-Q plots for the different ensemble members, implying that members are very similar to one another. In short, the ensemble members at KSFO are all consistently unreliable. Although the R-Q-Q plots for wind-speed at KSFO are similar to those of temperature, that is not the case for KLAX. The wind-speed forecasts at KLAX have the opposite problem than temperature forecasts. Whereas temperature forecasts are climatologically over-dispersed, wind-speed forecasts are climatologically uner-dispersed; and they clearly suffer from a negative bias, as well.

A qualitative ranking of the stations in terms of their wind-speed R-Q-Q plots, would place KBOI, KTUS, KDFW, KIAH, and KTPA among the best five, and KMSY, KLAX, KSFO, KMKE, and KPWM, among the worse five.

A more complete analysis and interpretation of these results is beyond the scope of this article and is performed elsewhere.

5. Summary and Discussion

Traditionally, reliability is assessed with the rank histogram (RH). But in certain situations it can be misleading (Hamill 2001). Furthermore, reliability has two components even when the forecasts are unbiased: one referring to ensemble dispersion, and another gauging climatological dispersion. Here, in order to tame the effects of temporal correlations, the RH is revised to display sampling variability. Using simulated data, it is then shown that the RH is still uninterpretable in the climatological component of reliability because of the confounding effects of the correlation between ensemble members, and the correlation between ensemble members and the observation. RH and Residual Q-Q plot (denoted R-Q-Q plot) are employed to assess the reliability of temperature and wind-speed forecasts generated from a 19-year reforecast data set, involving 10 ensemble members, at 90 stations across the continental US. The conclusions of the study are as follows:

- In the presence of temporal correlation, a traditional RH can be highly misleading. However, the boxplot variety of the RH can render it interpretable.
- A correlation between ensemble members, or between ensemble members and the observation, renders the RH uninterpretable regarding the climatological component of reliability. For example, one can get a U-shaped or dome-shaped RH, reminiscent of under- or over- ensemble variability, even if the climatological variance of the forecasts is equal to that of the observation. Conversely, a uniform RH may arise even if the climatological variance is not equal to the variance of the observation. The effect of ensemble dispersion on the RH, within the context of the multivariate model of Equation (2), is examined by Wilks (2010).

- The RHs for realistic temperature and wind-speed forecasts at most stations are U-shaped (even after the data have been transformed to assure that the bias and climatological variance of the forecasts are equal to those of the observation; see the discussion, below). This, however, does not imply that the forecasts are under-dispersed.

- A modified ("Residual") version of the Q-Q plot is better suited to the assessment of climatological variability, because it assesses the reliability of each ensemble member separately, and so is not affected by the correlation between ensemble members.

- The R-Q-Q plots for temperature and wind-speed forecasts suggest that some stations suffer from climatological under-dispersion, while others display climatological over-dispersion. Most produce negatively biased forecasts (i.e., forecast < observed). Yet others have forecasts which appear to come from a distribution which is completely different from that of the observation, some of which can be explained in terms of heavier or lighter tails in the distribution of the forecasts.

- The geographic distribution of the RH and R-Q-Q plots for the various stations, suggests that there is a spatial structure underlying the reliability of the forecasts. For example, under-dispersion (both ensemble and climatological) of temperature forecasts is generally associated with the Northwestern stations, while Northeastern stations display mostly bias only.

Although not shown here, RHs have also been produced for centered, as well as for climatologically standardized data. In the former, the observations and the forecasts from each ensemble member are assured to have the same mean, namely zero. In the standardized data, the observations and the forecasts from each ensemble member are assured to have a mean of zero and a standard deviation (within ensemble member) of 1. For the standardized data, given that the forecasts have no bias, and have the same climatological variance as observation, one might expect uniform RHs. However, all of the 90 RHs are found to be U-shaped. It is worth mentioning that removing bias alone has a drastic effect on most of the 90 RHs, mostly in the form of removing trends from the RH, and leaving mostly U-shaped RHs. But, interestingly, assuring that the ensemble members all have the same climatological variance (within ensemble member) does not significantly change the RHs, leaving most U-shaped. This suggests that the U-shaped behavior of the RHs is not due to under- or over- climatological dispersion of the forecasts from each ensemble, but due to ensemble dispersion.

There is yet another explanation for U-shaped RHs. It has been pointed out that such RHs may be due to observational error (Anderson 1996, Hamill 2001). This can be seen readily by noting that observational error simply increases the variance of the observations with respect to the "true" (error-free) observations. So, even if the RH corresponding to error-free observations were uniform, that of the real observations would still be U-shaped because the variance of the observations would be magnified by the observational errors. Similarly, observational bias would affect the trend conveyed in the RH. In short, the presence of observational error can adversely affect the shape of the RH. In most situations where observational errors cannot be quantified, there is nothing one can do to remedy this "defect" of the RH. Given that R-Q-Q plots are also sensitive to differences in mean and variance between forecasts and observations, they too are affected by observational errors. As such, any U-shaped R-Q-Q plot may in fact be due to error in the observations.

Apart from the visual effects on the RH, correlations also affect statistical tests of uniformity, because lack of independence generally reduces the degrees of freedom in a data set. As such, correlations can affect the significance of a statistical test. The chi-squared test, for example, proposed by Wilks (2004), Anderson (1996), and Hamill and Colucci (1997), assumes that the data are independently and identically distributed (iid). Violations of this assumption in real data can lead to either uncharacteristically large p-values (suggesting that the RH is consistent with a uniform distribution), or misleadingly small p-values (implying that the RH is inconsistent with a uniform RH). More sophisticated statistical tests, as proposed by Elmore (2005), and Jolliffe and Primo (2008), make the same iid assumption, and are therefore, subject to generating misleading conclusions. One way to control the effect of temporal correlations on statistical tests is to adjust the level of significance so as to reflect the reduction in the degrees of freedom. Although the current work focuses on the visual assessment of reliability, the uniformity of the RHs is assessed by a chi-squared test anyway. The test is performed on the standardized data. Critical values of the chi-squared statistic are given in Wilks (2004, section 4) for different values of the ϕ parameter describing an AR(1) process. Fitting an AR(1) model to the temperature data for the various stations yields estimated ϕ values ranging from 0.7 to 0.9 (depending on the station). The range of estimated ϕ values for wind speed is 0 to 0.2. The critical values of the chi-squared statistic for such ϕ values are in the 3.0 to 21.0 range, for an α-level of 0.05. The interquartile range of the observed values of the chi-squared statistic for the standardized temperature data is 200-500, and the analogous values for wind speed are 300-500. All of these observed chi-squared values are much larger than the critical values, and so, uniformity of the RHs can be rejected.

Examining the contribution of different facets of the forecasts to the RH is consistent with the view of a RH as a qualitative but omnibus test of the equality of two distributions. Common tests usually compare two distributions in terms of specific moments. For example, Student's t-test is a test of means, while an F test can be used to test equality of two variances. By contrast, the Kolmogorov-Smirnov test compares two distributions without any reference to specific moments or parameters of the distributions. As such, it is said to be nonparametric (Wasserman 2007); such tests generally have power against all alternatives. The RH is effectively such a test, albeit a qualitative one. One may argue that by virtue of being a diagram (i.e., a histogram), it is more diagnostic than omnibus tests which usually lead to a single p-value. However, all tests - qualitative or not - are misleading in some situations, and, as shown here, the RH is no exception.

Acknowledgments.

Dan Wilks, Thomas Hamill, Don Percival, Jon Wellner, and Kim Elmore are acknowledged for invaluable contribution. Nate Hiers helped in producing the ensemble reforecast dataset. Lynn Rose and Mark Bradford provided helpful discussions throughout the project. The computational resources for producing ensemble reforecast data are provided by the National Center for Computational Sciences (NCCS) funded by DOE, and by the National Institute for Computational Sciences (NICS), a joint center of University of Tennessee and Oak Ridge National Laboratory, funded by National Science Foundation. This research is supported by DTRA SBIR grant HDTRA1-07-C-0122.

References

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1529.

Bishop, C. M., 1996: *Neural networks for pattern recognition.* Clarendon Press, Oxford, pp. 482.

Bühlmann, P., 2002: Bootstraps for time series, *Statistical Science*, **17**, 52-72.

Efron, B., and R. J. Tibshirani, 1998: *An introduction to the bootstrap.* Chapman & Hall, London.

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

Fukunaga, K., 1990: *Introduction to statistical pattern recognition.* San Diego, Academic Press. 602 pp.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Jour. of Royal Statistical Society: Series B (Statistical Methodology)*, **69(2)**, 243-268.

Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

Hamill, T.M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711724.

Jolliffe, I. T., C. Primo, 2008: Evaluating rank histograms using decompositions of chi-square test statistics. *Mon. Wea. Rev.*, **136**, 2133-2139.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.*, **137**, 1717-1720.

Lahiri, S.N., 2003: *Resampling methods for dependent data.* Springer, New York.

Marzban, C., R. Wang, F. Kong, S. Leyton, 2010: Rank Histograms and Correlations. 20th Conference on Probability and Statistics in the Atmospheric Sciences; The 2010 Annual Meeting of the American Meteorological Society, Atlanta.

Politis, D.N., J.P. Romano, and M. Wolf 1999: *Subsampling.* Springer, New York.

Wasserman, L., 2007: *All of nonparametric statistics*, Springer.

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, Second Edition. Academic Press, San Diego, CA, 467 pp.

Wilks, D. S., 2010: On the reliability of the rank histogram. *Mon. Wea. Rev.*, this same issue.

List of Figures

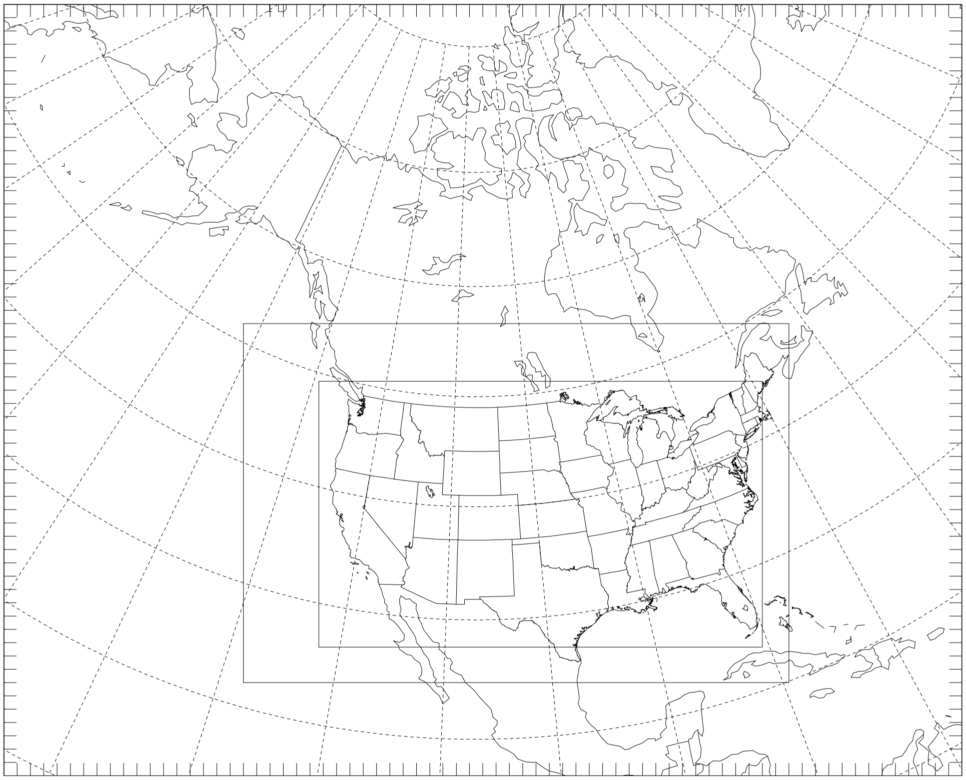

Fig. 1. Model domains, with the horizontal grid spacing of 135, 45, and 15 km for the outer, intermediate, and inner domain, respectively.

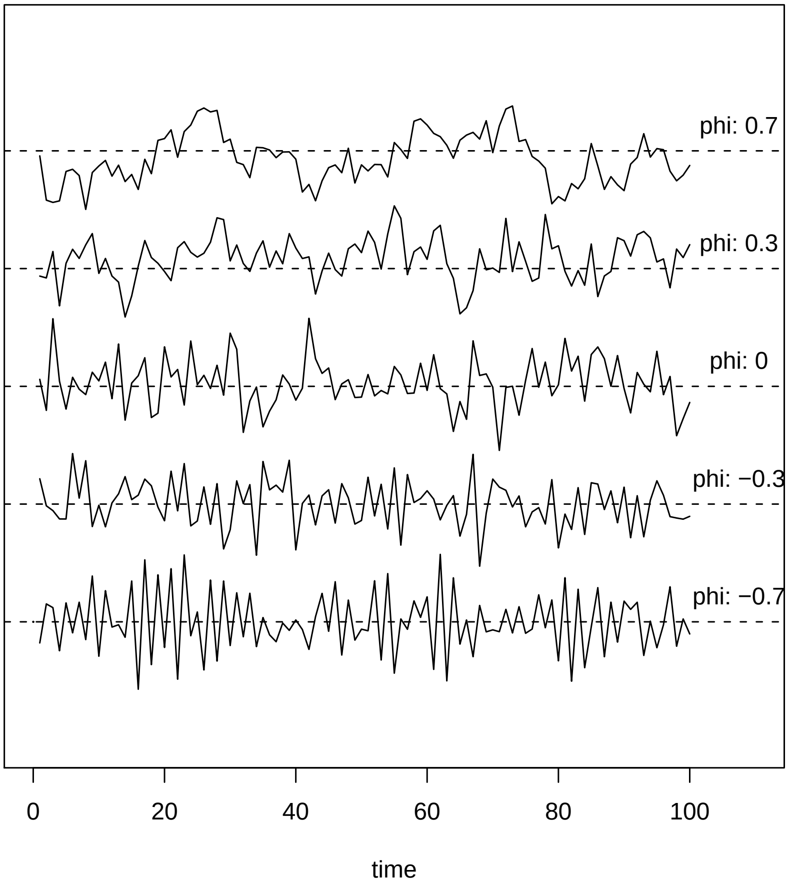

FIG. 2. AR(1) time series with $\phi = 0.7$, 0.3, 0, -0.3, -0.7, from top to bottom, respectively.

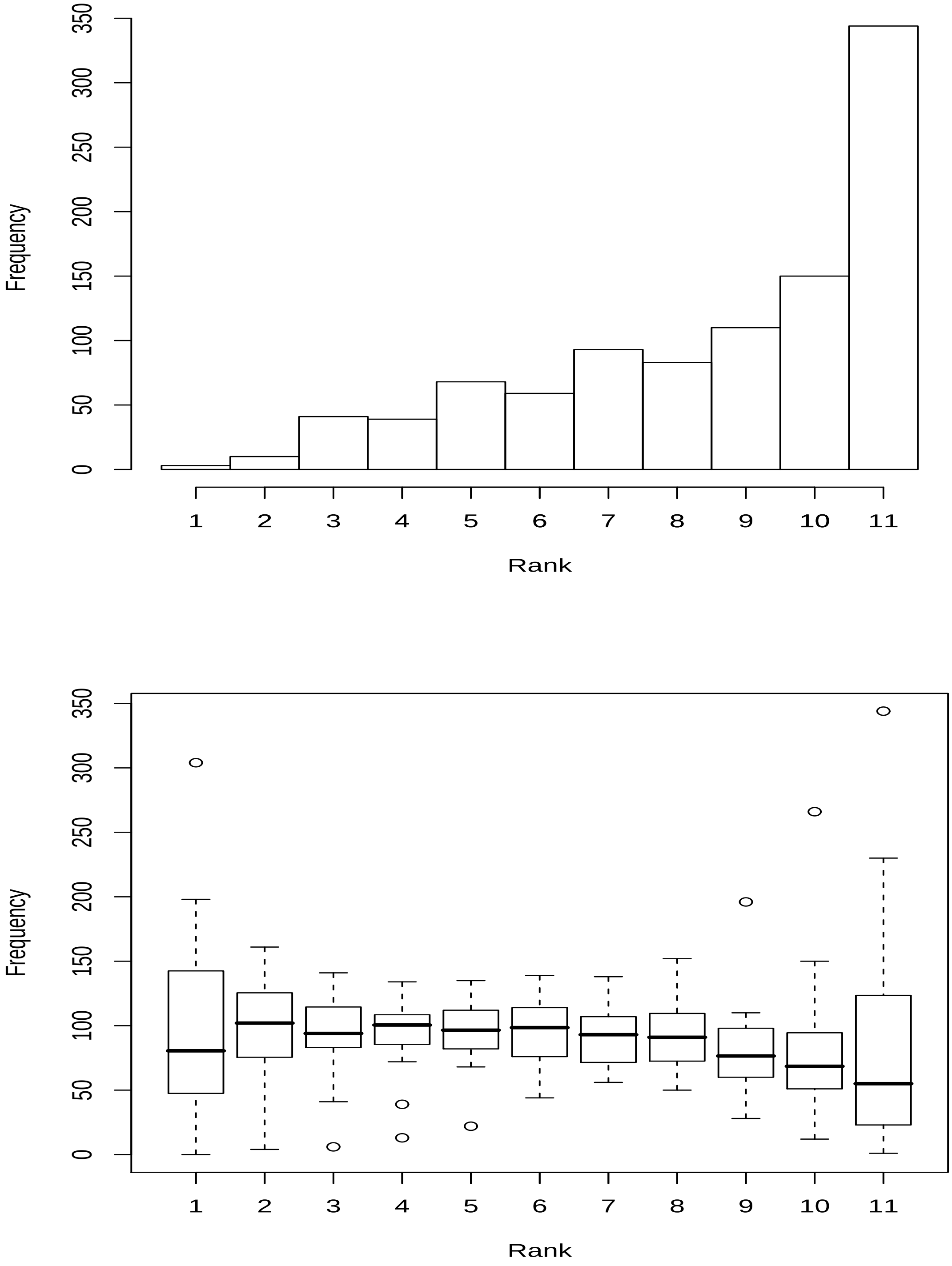

FIG. 3. A traditional RH for a single realization (top), and the boxplot variety based on multiple realizations (bottom) of reliable forecasts and observations drawn from a multivariate AR(1) process, with no correlation between ensemble members.

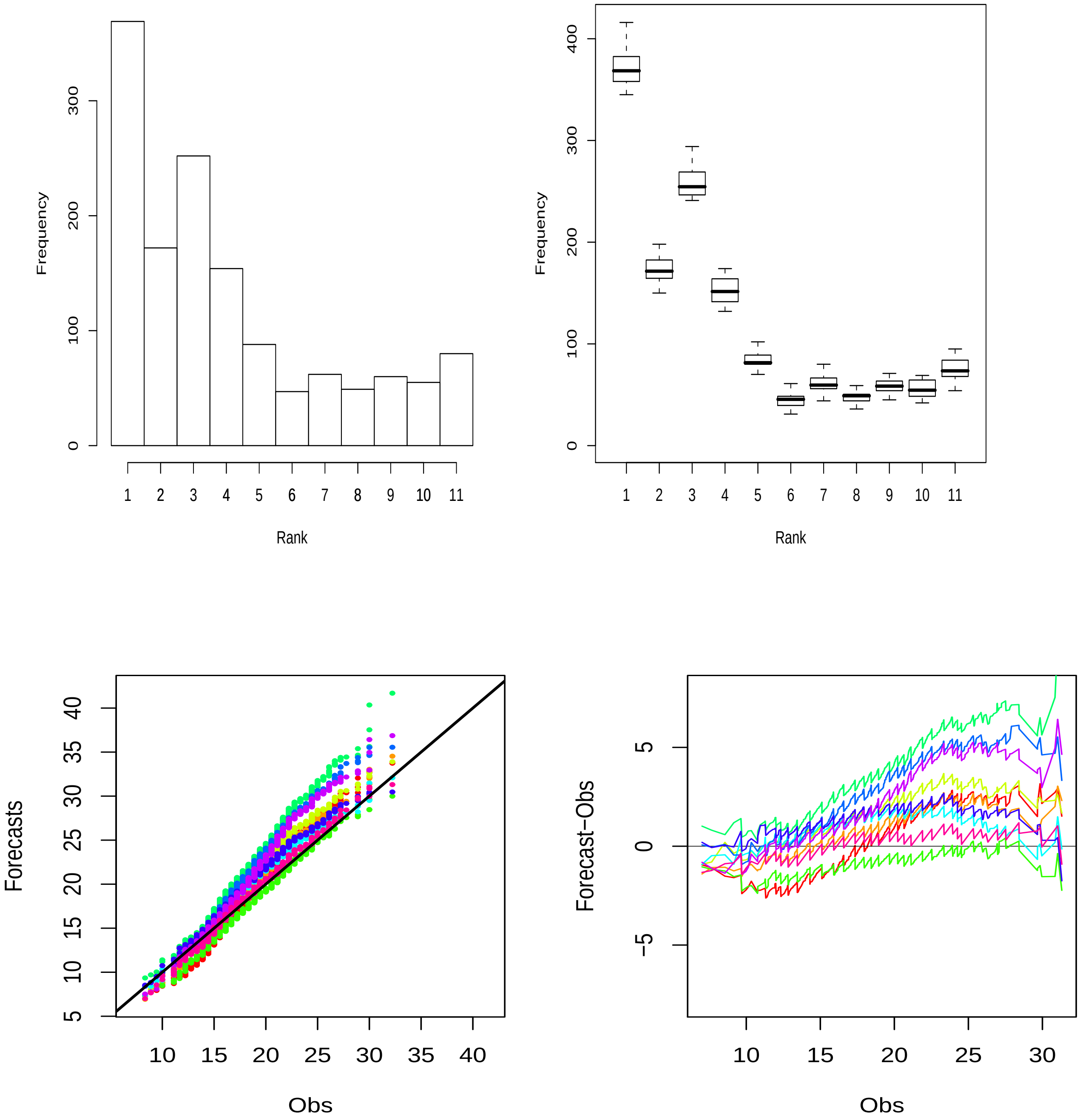

Fig. 4. RH of temperature at KLAX (top, left), and its boxplot version (top, right). Also shown are the Q-Q plots for all 10 ensemble forecasts (bottom, left), and the R-Q-Q plot described in text.

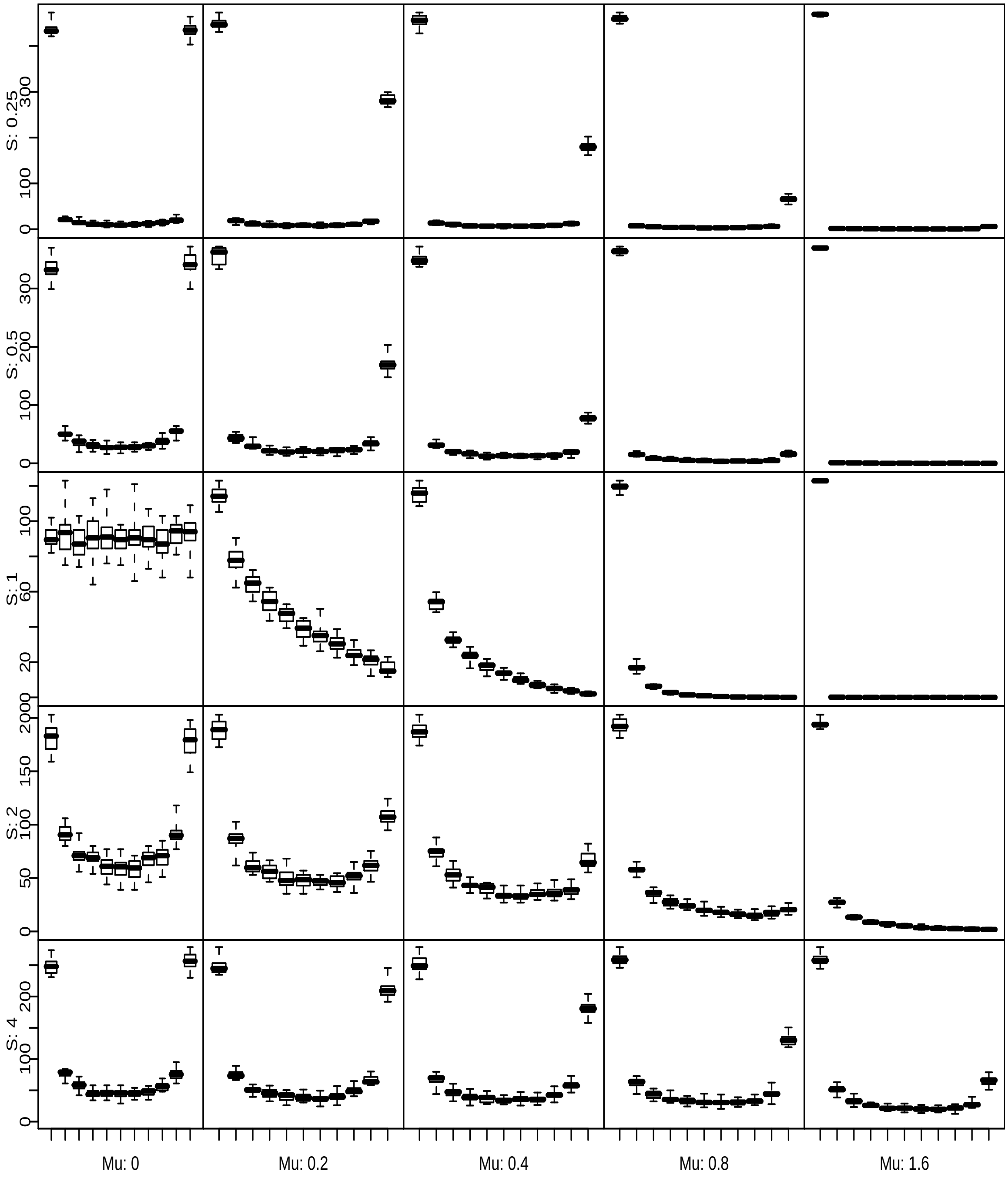

FIG. 5. Rank Histograms for different values of the common forecast mean μ (along x-axis) and variance within ensemble member σ (along y-axis), when there is a strong correlation between ensemble members and an equally strong correlation between the ensemble members and the observation (i.e., $R = r = 0.9$).

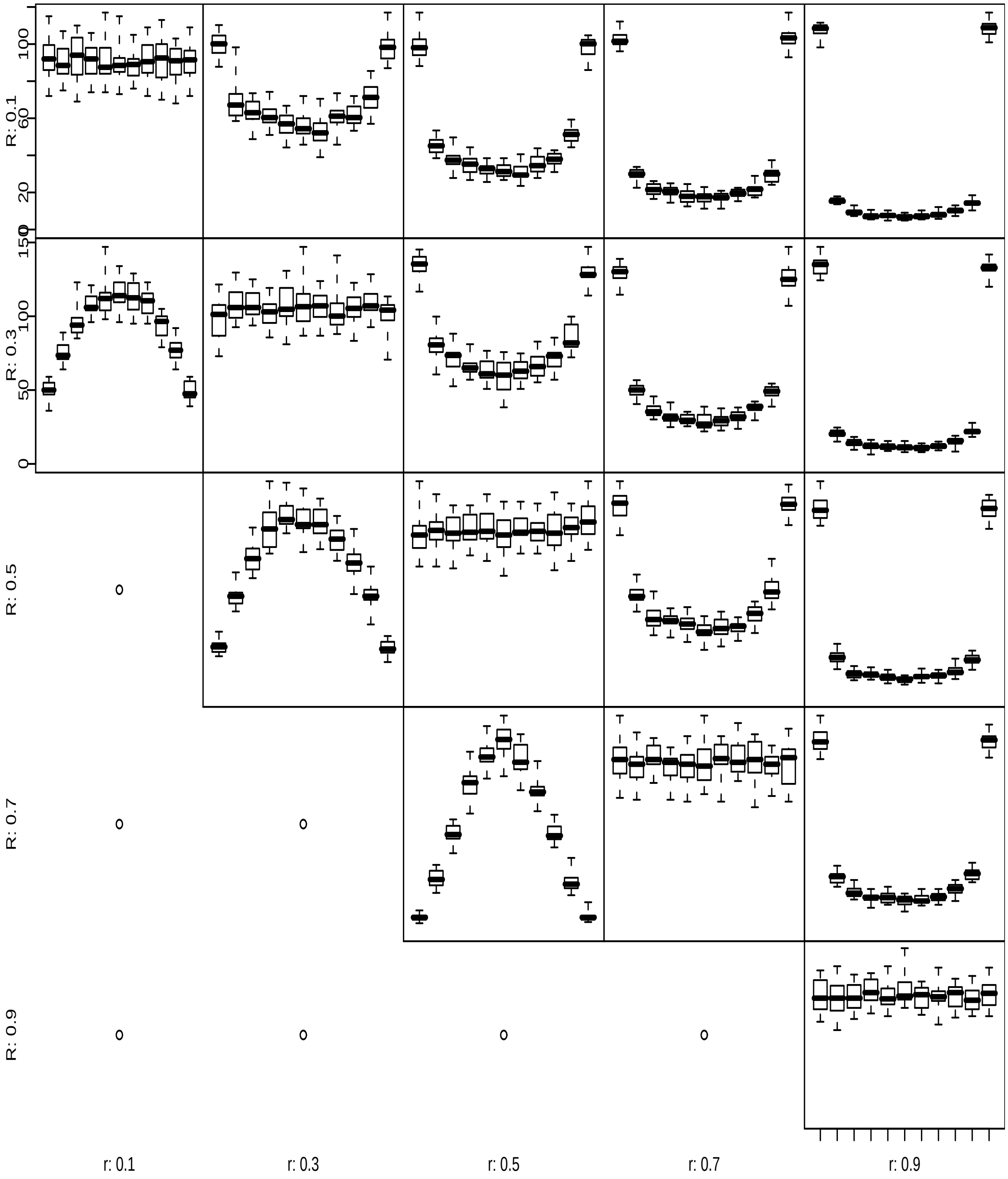

FIG. 6. Rank Histograms for different values of the correlation between ensemble members, r (along x-axis), and the correlation between ensemble members and the observation, R (along y-axis).

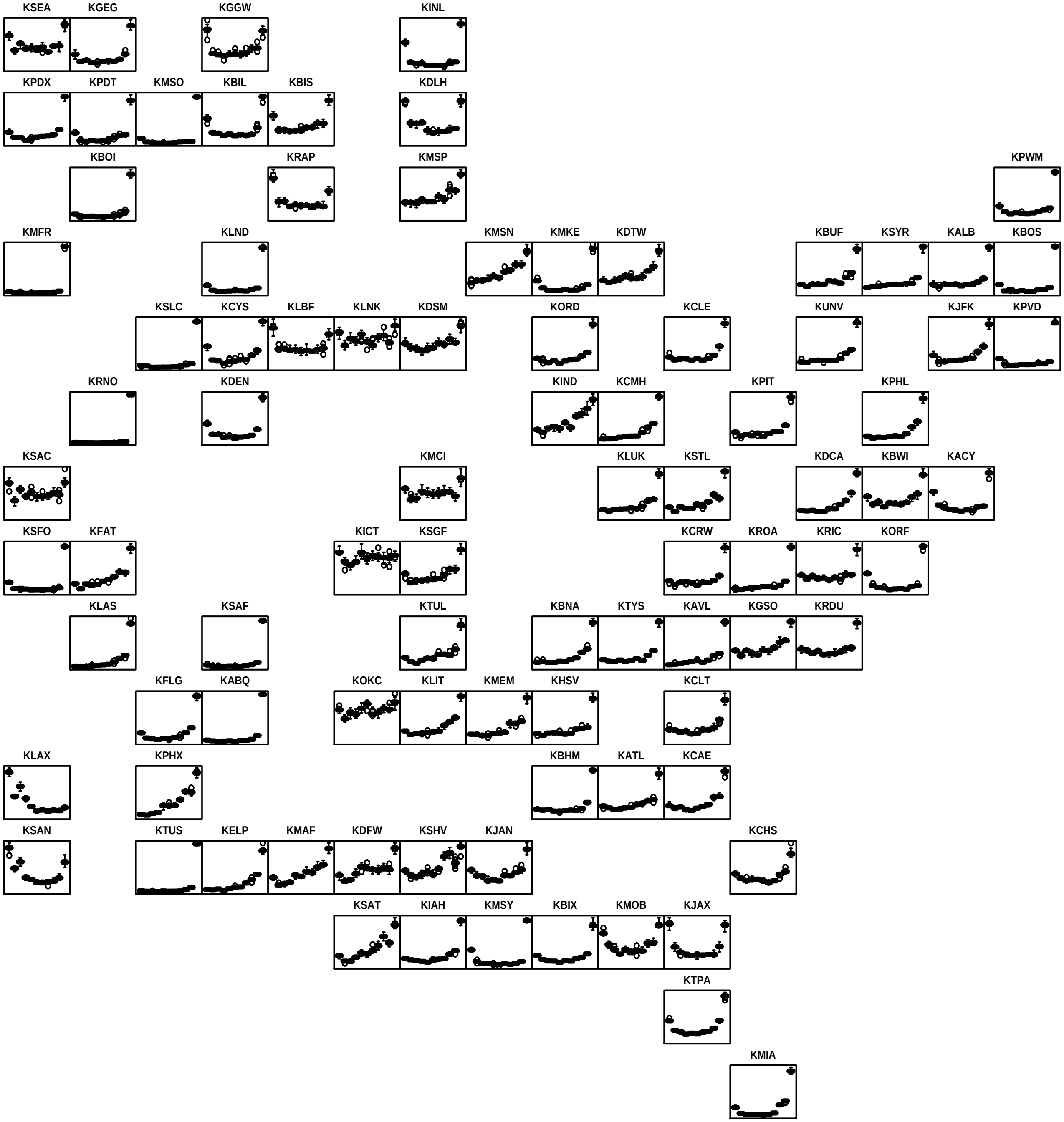

FIG. 7. RHs for temperature forecasts across 90 stations nationwide.

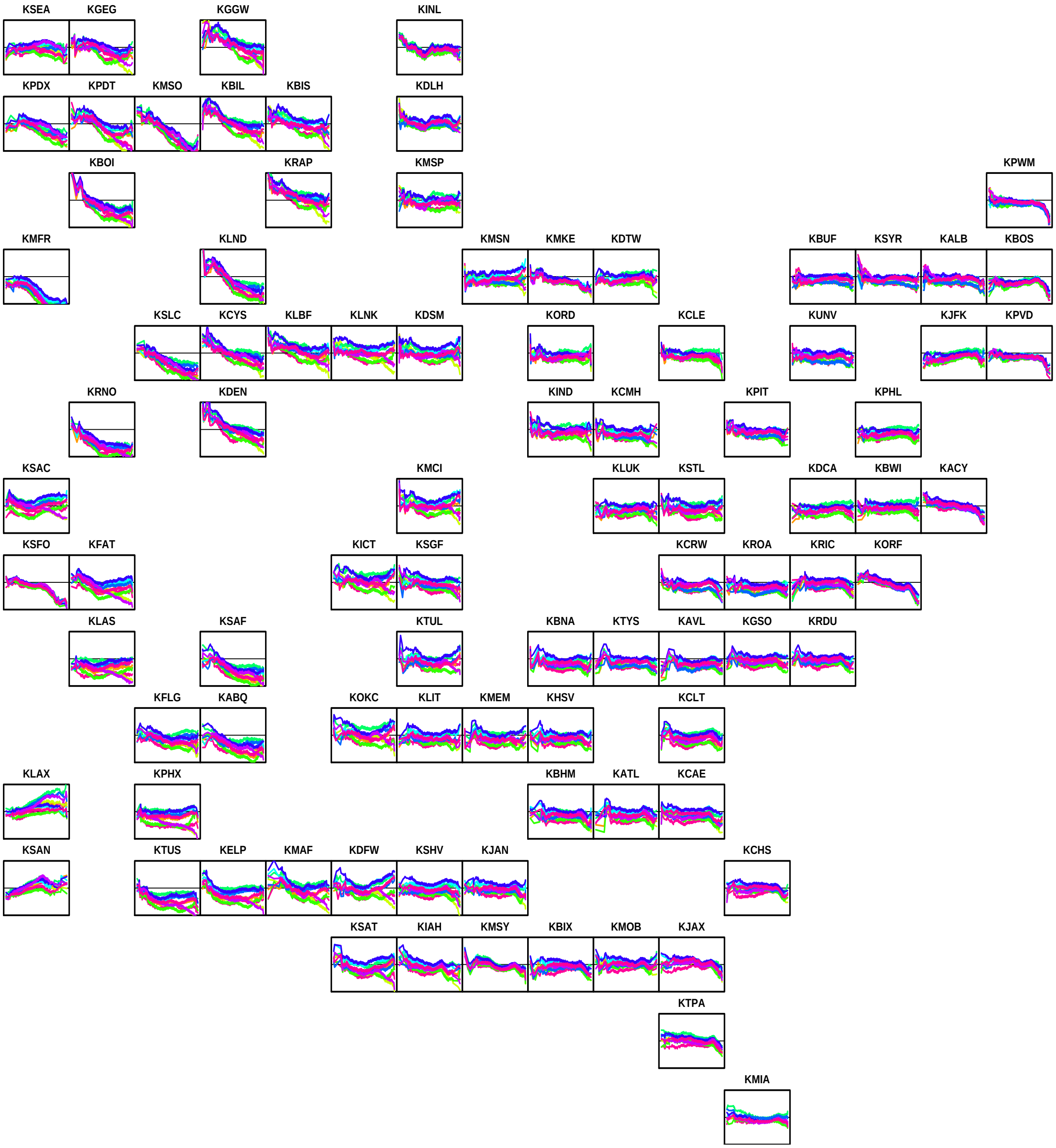

FIG. 8. R-Q-Q plots for temperature forecasts across 90 stations nationwide. The horizontal line is the x-axis, and the y-axis pertains to (forecast - observed), and so R-Q-Q plots below the x-axis imply a negative bias (i.e., forecast < observed). The range along the y-axis for all panels is $\pm 8°C$.

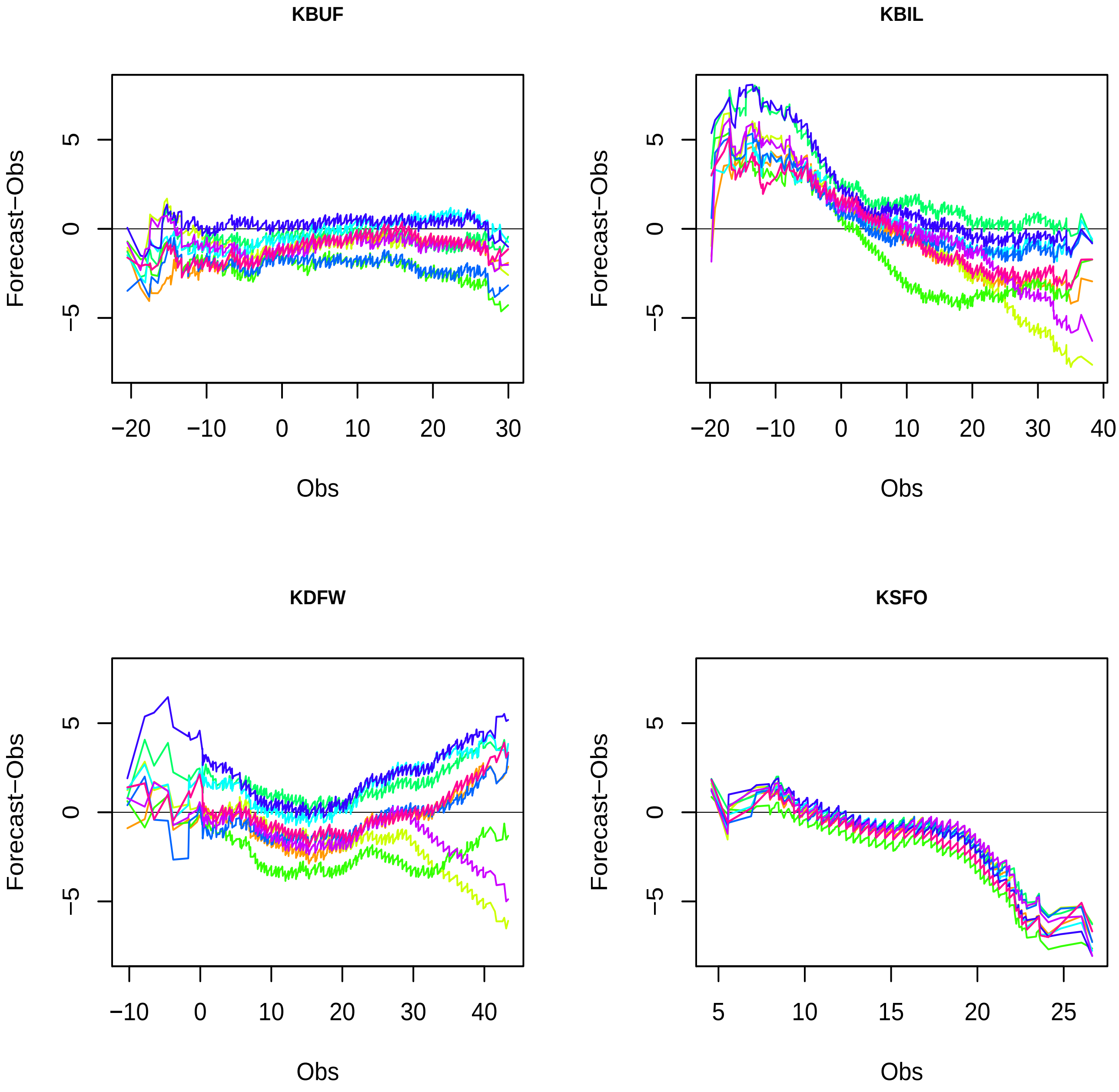

FIG. 9. R-Q-Q plot for temperature forecasts at four stations.

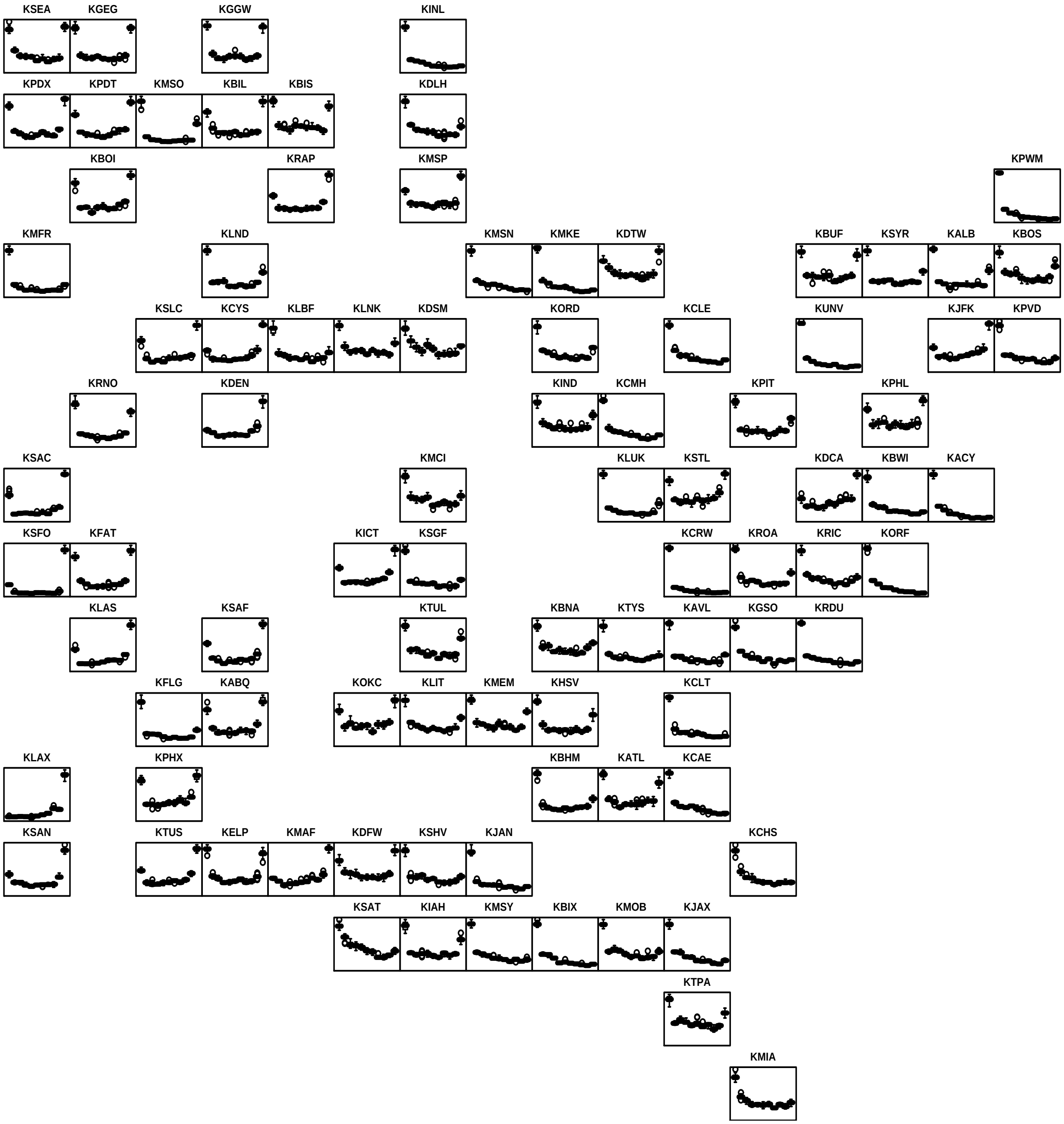

FIG. 10. RHs for wind-speed forecasts across 90 stations nationwide.

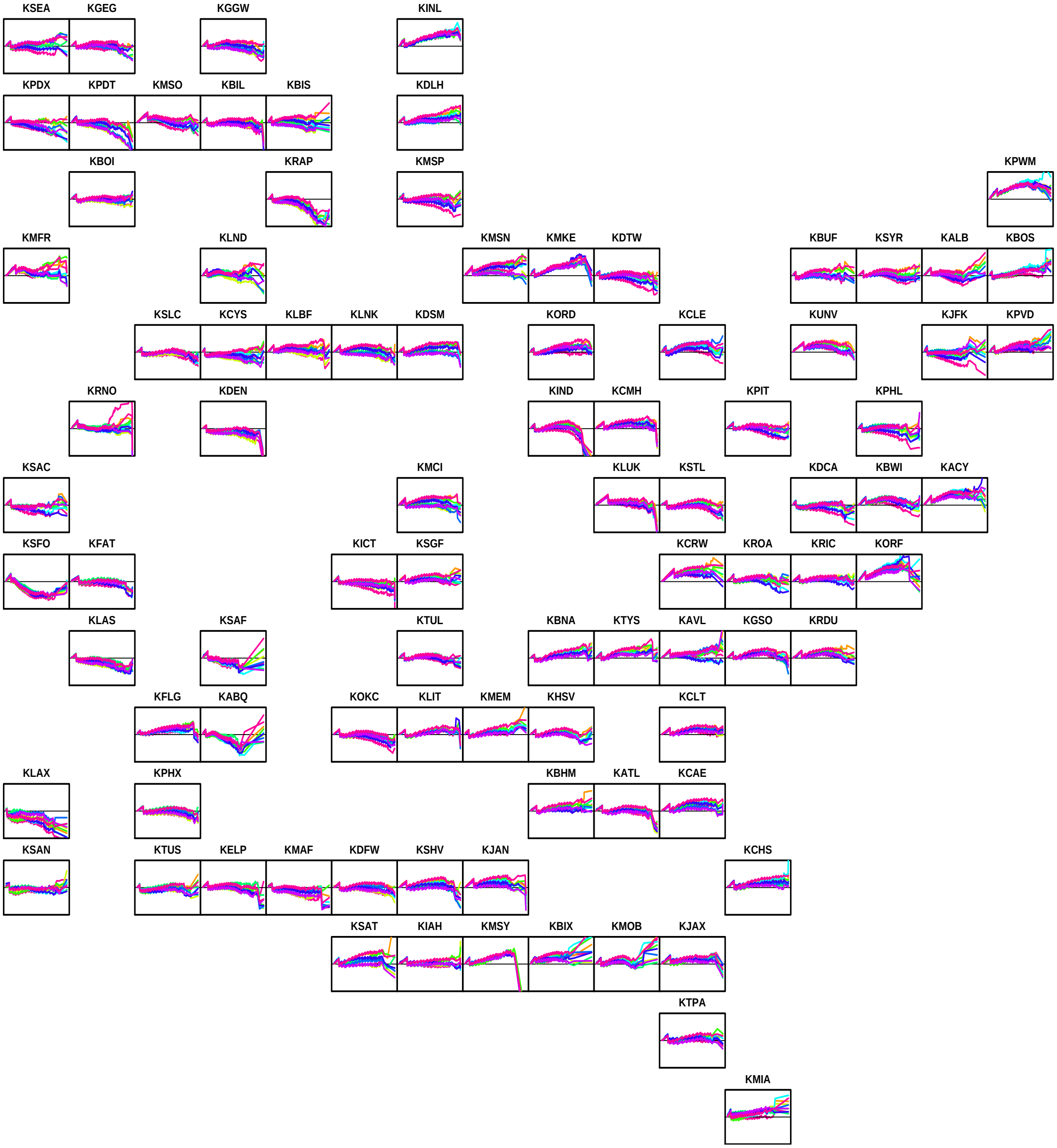

FIG. 11. Same as Figure 8, but for wind-speed forecasts. The range along the y-axis for all panels is $\pm 6\, m/s$.