

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

CAREN MARZBAN^{1,2}, RANRAN WANG¹, FANYOU KONG³, STEPHEN LEYTON⁴

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 is a standard tool for assessing the quality of ensemble forecasts. Like all measures, though, in certain situations it conveys misleading information. Here, first, simulated data are employed to illustrate how rank histograms are affected by 1) temporal correlations, and 2) correlations between ensemble members, in addition to the (better-studied) effect of 3) bias, and 4) over/under dispersion. Then, the rank histograms for realistic temperature and wind speed forecasts at 90 stations across the continental US are computed. It is found that even after the contribution of each of the four factors has been accounted for, the rank histograms for temperature forecasts are generally U-shaped across the US. This suggests that the distribution of observed temperatures differs from that of the ensemble forecasts, not in their bias and variance, but in higher moments of the distributions. An analysis of the geographic distribution of the rank histograms suggests that forecast reliability (as measured by rank histogram) is generally constant across the US, with a few stations (e.g. in Florida) displaying a different behavior. As for wind speed, the conclusions are similar to that of temperature forecasts, except that the rank histograms are generally more uniform.

1. Introduction

Given an observation, and an M -member ensemble of corresponding forecasts, the quality of the forecasts 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 measure 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 $1, 2, \dots, M+1$. Given a *set* of observations and forecasts, then, one can compare the histograms 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. 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.¹

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, for which the whole distribution can be described by only two parameters - the mean and the standard deviation. He 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

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

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 shape of a RH.

The comparison of the RH with the uniform distribution may be performed visually or statistically. Most commonly, one simply looks at a RH and assesses its flatness. Alternatively, a statistical test may be performed to make the comparison more rigorous. Either way, however, the conclusions are apt to be misleading if there is any correlation in the data. For example, a temporal correlation in the time series of the observed or forecast values, can affect both the shape of the RH as well as the results of a statistical test. Additionally, a correlation *between* the ensemble forecasts can also have similar adverse effects. The effect, on RH, of both types of correlation are examined in this paper.

It is important to point out that a positive correlation between ensemble members is qualitatively a different phenomenon than that of under-dispersive forecasts. Similarly, for negative correlations and over-dispersive forecasts. A RH tests whether forecasts and observations are drawn from the same distribution, and it is expected to look flat if the samples are drawn randomly and independently. There are, therefore, multiple explanations for a non-flat RH. A U-shaped RH, for example, may be due to the population of forecasts having a smaller variance than that of the observations, i.e., due to under-dispersive forecasts *within* each ensemble members' forecast; but it can also be due to a correlation *between* ensemble members. In short, deviations from uniformity in a RH may be attributed to differences in the populations of forecasts and observations, or to a correlation between the ensemble members, or to both.

All of the misleading behaviors described in Hamill (2001) are exacerbated in the presence of correlations. For example, 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 can be “systematic.” In short, a temporal correlation can lead to a RH which is unambiguously non-uniform, in spite of the reliability of the forecasts. Correlations between ensemble members further exaggerate this effect. One remedy which helps to bring some visual interpretability back to the RH is to supplement the RH with some measure of sampling variability. In the present work, for example, the RH is displayed as a series of boxplots. Correlations, then, typically affect the size of the boxplots, but the eye can readily assess the flatness of the resulting “RH.”²

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 (1995), 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 exceedingly 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

²Here, both the traditional RH and its boxplot variation are referred to as RH.

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. An example of this method is demonstrated by Wilks (2004; section 4) for the specific case of a first-order autoregressive processes. The current work does not deal with such adjustments, because no statistical testing is performed; however, see the Discussion section. The realistic data sets examined here are sufficiently long in duration to obviate the need for a statistical test. The analysis of the simulated data sets, however, does effectively assess statistical significance, by virtue of examining the sampling distribution of the frequencies of the ranks in a RH.

The outline of this paper is as follows: Simulated data are employed to illustrate how the shape of the RH changes with respect to 4 factors: 1) bias, 2) under- or over- variability within each ensembles' forecasts, 3) correlations between ensemble members, and 4) temporal correlations. The role of the first two factors is well-known, but the effect of the latter two is less studied. It is shown that temporal correlations can be rendered harmless (i.e., no longer visually misleading), if one plots the boxplot version of the RH. It is also shown that a (positive) correlation between ensemble members can manifest itself as a U-shaped RH, i.e., similar to over-dispersive forecasts. Turning to real data, the RHs for temperature forecasts at the 90 stations are computed, first without any preprocessing of the data. Then, the RHs are re-produced after bias-removal, followed by another set of RH based on data wherein forecasts and observations have equal variance (and mean). A final set of RHs is produced for "whitened" forecasts wherein the ensemble members are arranged to be uncorrelated. The last set of RHs allows one to examine the contribution to deviations from uniformity which may be attributed to differences in higher moments of the distribution of forecasts

and observations. A geographic analysis of these RHs also reveals some useful information. The same analysis is performed for forecasts of wind-speed, leading to similar conclusions.

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 4 factors. The real data sets consist of 20 years (1987-2006) 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 1 observed value and 10 forecast values from by a 10-member ensemble.

The 20-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 dataset used in this study. 48-h reforecasts, initialized at 0000 UTC, were produced every five days starting from 1 January 1987 through 31 December

2006. WRF outputs were written every 3 hours. Selected forecast variables such as temperature, wind speed, and accumulated precipitation 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 5.

As mentioned above, hourly surface observations over the 20 year period are obtained for the 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, which would otherwise render that hour completely unusable for statistical analysis.

3. Method

The main goal of this article is to produce RHs for realistic forecasts for a number of stations across the US. To render the RHs more interpretable, however, the contributions from four factors are isolated: contribution from 1) bias, 2) differences in the variance of observations and forecasts within ensemble members, 3) correlations between ensemble members, and 4) temporal correlations.

It is known that bias leads to RHs which display a trend (Hamill 2001). As for the contribution of variance to the shape of the RH, if the variance of forecasts from a given ensemble member is smaller (larger) than that of the observations, one generally obtains a U-shaped (dome-shaped) RH (Hamill 2001). The following subsections examine the role of

correlations.

a. Temporal Correlation

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 “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.

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 nothing but 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 a set of forecasts and observations with $\phi = 0.7$. Due to the unambiguous trend in the 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.

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 RH is consistent with the uniform distribution.³ In short, whereas the top RH would lead to the wrong conclusion that the forecasts are biased, the bottom RH suggests the correct conclusion that the forecasts and observations are from the same distribution.

In order to avoid being visually misled by sampling variability, the RHs produced here are of the boxplot variety, at least for the simulation study. For the real data, boxplots

³See the discussion section.

are not displayed because estimates of the ϕ parameter, based on real data, place it in the 0.3 ± 0.1 range for all 90 stations. We have confirmed through simulations that such small ϕ values lead to RHs which are not significantly affected by sampling variability - at least, not at a visually noticeable level. The discussion section addresses some of the issues that would arise if one does desire a boxplot version of RH for real data.

b. Correlated Ensembles

The correlation between the ensemble members is more difficult to simulate. For the sake of completeness a few options are described in the Appendix, which also shows that the task is made even more complex if one insists on including a temporal correlation in the simulation. In other words, simulating the behavior of the RH in the presence of *both* correlations is difficult. However, it can be argued that it is not necessary to examine both correlations, jointly. As discussed in the previous section, temporal correlations manifest themselves in RHs only through sampling variability. We have, in fact, confirmed this even in the presence of correlations between ensemble members. As such, it is sufficient to examine the correlation between ensemble members, when each member produces independently and identically distributed forecasts (i.e., with $\phi = 0$).

As discussed in the Appendix, in designing a simulation of the correlation between ensemble members, it is not entirely clear how the observations should be correlated with the forecasts from each of the ensemble members. On the one hand, given that a RHS is designed to be a measure of reliability, and not accuracy, one may argue that a correlation between the observations and the forecasts should not be introduced. Here, this ambiguity

is (partially) resolved by an examination of real forecasts. Specifically, the pooled covariance matrix (across all 90 stations) is computed from real data, and is used in the simulation.

To be more specific, a sample is drawn from a 10-dimensional (multivariate) normal distribution with all the mean and variance parameters set to 0 and 1, respectively. All off-diagonal elements of the covariance matrix are set to the pooled estimates obtained from real data. This assures that 1) the simulated forecasts are bias-free, because they, and the observations, have zero mean; 2) the variance of each ensemble member is equal to that of other ensemble members, and to that of the observations (i.e., equal to 1); and 3) the ensemble members are correlated in a realistic manner.

This scheme for deciding how observations should be correlated with the forecasts also addresses the question of how the ensembles themselves should be correlated; the correlation between simulated ensemble members is inherited from that of the real ensemble forecasts. In this way, the simulated observations differ from the simulated forecasts only in terms of the correlation between them. ⁴

The top panel in Figure 4 shows the resulting RH. Evidently, a correlation between ensemble members renders the RH U-shaped. It is important to point out that this U-shaped behavior is not due to the under- or over-dispersion, because all the variances are identical (i.e., 1). This U-shape behavior is due to correlation. It is not surprising that

⁴Here the full covariance matrix is used in the simulation. However, it is also possible to replace the elements corresponding to the correlation between observations and forecasts with 0. The resulting simulation would assure that the observations are not correlated with the forecasts. We have confirmed that the results are similar to those reported here; the main reason for the similarity is that a RH is insensitive to the accuracy of forecasts.

a correlation between ensemble members gives rise to a RH resembling one that is due to under-variability within each ensemble members' forecasts; but, as seen here, the reasons are very different.

c. Parsing the RH

The above simulations illustrate the effect of correlations on the RH. However, given that the main purpose of this article is to interpret RHs for real data, it behooves one to isolate, and account for, the contribution to RH from these correlations, as well as from bias and variance.⁵ To that end, for each of the 90 stations a sequence of RHs is constructed, where the contributions of bias, variance, and correlations between ensembles, are systematically filtered out.

The contribution of bias is readily accounted for by simply subtracting from each forecast the long-term (climatological) mean of the corresponding ensemble member's forecasts. The observations can be "centered" in a similar fashion. This insures that all ensemble forecasts and the observations have a mean of zero.

To control for the effect of under- or over-variability, it is sufficient to divide the above difference by the respective standard deviations. This "standardization" assures that the forecasts and observations have a mean of zero, and a standard deviation of 1.

Finally, in order to isolate the contribution from the correlation between ensemble mem-

⁵These are the first three of the four aforementioned factors affecting the shape of a RH. The fourth factor - temporal correlation - affects the RH only through sampling variability. And, as argued above, the relatively small size of the estimated AR(1) parameters suggests that temporal correlation is not a major source of concern.

bers, one can “whiten” the data. The details of whitening can be found in (Bishop 1996) and (Fukunaga 1990). Suffice it to say that it is a transformation which maximally decorrelates the ensemble members. Said differently, whitening assures that the forecasts from the ensembles are orthogonal. The effectiveness of whitening can be demonstrated on the correlated forecasts whose RH is shown in Figure 4 (top); whitening the forecasts gives rise to the RH shown in the bottom panel of Figure 4. Evidently, whitening effectively filters out the contribution, to the RH, from correlation between ensemble members.

4. Results of Application

Figures 5, 6, 7, and 8 show the RHs for “raw,” centered, standardized, and whitened temperature data, respectively. From Figure 5 it can be seen that the RHs for nearly all stations are non-uniform, displaying U-shaped and trend features. The three exceptions are KLNK, KICT, and KOKC, for which the RHs do not display either. In fact, within sampling variability these three RHs appear to be mostly flat. An visual examination of the trends suggests that almost all stations suffer from negatively-biased forecasts, with the exception of KLAX and KSAN which have a positive bias. All of these conclusions are qualitative; for example, one may argue that the RH for KSAC is also consistent with a uniform RH. However, the superposition of trends and U-shaped RHs makes it difficult to draw an unambiguous conclusion.

Figure 6 shows the RHs for centered data, i.e. when bias has been removed. Evidently, and as expected, the trends have been removed from the RHs. The shape of the RHs can be classified into three types: 1) flat, 2) U-shaped, and 3) “Ends” (Jolliffe and Primo 2008),

i.e., flat apart from two peaks at the lowest and highest ranks. The most flat RHs belong to KLAS and KPHX. It is noteworthy that the RH for these two stations, prior to the removal of bias, is dominated by a trend, making it difficult to see that the forecasts for these stations actually produce reasonably flat RHs after bias correction. Another pattern is that the stations along the Eastern and Western coasts generally produce Ends RHs, while those in the interior tends to be more U-shaped, or at least less Ends-looking. However, again, in order to better diagnose these RHs, it is important to remove the effects of over- and under-dispersion within each ensemble member's forecasts.

Interestingly, controlling for this within-variability does not significantly affect the RH (Figure 7). In other words, standardizing the forecasts and observations leads to RHs (Figure 7) which are almost identical to those produced from centered data (Figure 6). This suggests that the U-shaped or Ends behavior of these RHs in Figure 6 is not due to under- or over-dispersion of the forecasts from each ensemble.

Finally, Figure 8 shows the RH after the forecasts are further whitened. The most striking feature of these RHs is that most are still U-shaped, although to a far-lesser degree than prior to whitening. Some even show signs of a trend (e.g., KMIA). Additionally, a few stations (e.g., KLAX, KSAN, KSFO) display only a single-peak at the lowest ranks, on an otherwise flat RH. Given that bias, and variance, and correlation have been explicitly removed from the data, there are two possible explanations for these behaviors.

It is known that sampling from a negatively-biased distribution half the time, and from a positively biased distribution the remaining half, will lead to a U-shaped RH; Hamill (2001) calls this "conditional bias." It is easy to show that changing the proportion from half, leads to the appearance of a trend. In other words, it is possible that the deviations from uniformity

are due to conditional bias (or even conditional variance). This is one explanation.

The second explanation follows when one notes that deviations from uniformity imply that the forecasts and observations come from different distributions. Given that the difference between the distributions cannot be in their mean, nor in their variance-covariance parameters, it follows that the difference in the two distributions is in higher moments of the distributions. This conclusion would be unfounded if the distribution of temperature (or wind-speed) were normal, for then the distribution would be fully described by the mean and variance-covariance parameters. However, an examination of the distributions, via qqplots (not shown here), does suggest that the distributions are, in fact, not normal. In short, the second explanation for the non-uniform RHs is that the distribution from which the forecasts are drawn is in fact different from that of the observations, with the difference being in some higher (than 2) moment of the distribution.

Before examining the RH for wind-speed, note that whitening can be employed, as a post-processing step, to render the forecasts more reliable. This is similar to the ideas of Hamill and Colucci (1997; 1998), and Wilks (2006) wherein post-hoc corrections are examined for calibrating forecasts. In spite of the unreliability remaining in the forecasts even after whitening, the RHs are more uniform than those based on “raw” or standardized data.

Figures 9-12 shows the analogous RHs for wind speed. The conclusions are mostly the same, with a few exceptions. For example, Southwestern stations such as KLAX, KSAN, KSFO display a positive bias - in the opposite direction to the rest of the stations, which convey a negatively-biased forecasts. By contrast to temperature, the RHs for wind-speed in Northwestern stations display no obvious trend.

Upon removing the effect of bias, Figure 10 shows that the RHs for all 90 stations

are very similar. As in the RHs for temperature forecasts, the RHs for wind-speed are mostly unchanged when the effect of variance is also removed (Figure 11). Whitening the forecasts leads to the RHs shown in Figure 12. These RHs are more uniform than those for temperature forecasts, suggesting that whitening of wind-speed forecasts can readily lead to near-reliable forecasts. Clearly, some deviation from uniformity persists, and the same explanations offered for the temperature RHs can be applied here.

5. Summary and Discussion

Temperature and wind-speed fine-scale (15 *km* grid-spacing) forecasts generated from a 20-year reforecast analysis, involving 10 ensemble members, are examined for 90 stations across the continental US. In order to render the rank histograms (RH) interpretable, the contributions from four factors are isolated. The factors are 1) bias, 2) variance within each ensemble member, 3) correlation between ensemble members, and 4) temporal correlation. It is argued that temporal correlations can be conveyed by producing a boxplot version of the RH. The other three factors are controlled by centering, standardizing, and whitening the data. After accounting for the contribution from each factor, any remaining deviation from uniformity can be attributed to a difference between the population of forecasts and observations in higher moments of the distributions. It is also argued that these transformations can be performed as a post-processing step which can render forecasts more reliable. The methodology is then applied to real temperature and wind-speed forecasts. The main conclusions are:

- 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 can give rise to a U-shaped RH, reminiscent of under- or over-variability, but for different reasons.
- The RH for realistic temperature forecasts at nearly all 90 stations display a U-shaped pattern, even after bias, variance, and correlation have been accounted for. This suggests that the distribution of forecasts differs from that of the observations in higher moments of the distribution.
- Temperature and wind-speed forecasts in the Northwest are mostly bias free (or, at least, less biased than in the rest of US).
- Forecasts in Southwestern stations have the opposite bias than those in the rest of the US.
- Postprocessing the forecasts via whitening (and standardizing) can improve the reliability of forecasts.

Examining the contribution of different facets of the forecasts to the RH is consistent with the view of a RH as an 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 a chi-square 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, although by virtue of being based on a diagram (i.e., a histogram), it is more diagnostic than most other omnibus tests which usually lead to a single p-value.

One of the contributions of this work is the proposal that RHs should be displayed in terms of boxplots. The added benefit of visually displaying sampling variability, however, is accompanied by two potential sources of misinterpretation. The first is that one may interpret the consistency of the boxplots with a uniform distribution (e.g., Figure 3, bottom) as evidence in *support* of uniformity. As in traditional hypothesis testing, this would be an incorrect interpretation of the sampling distribution. The correct conclusion is that the data *do not contradict* the hypothesis of uniformity. For example, The bottom panel in Figure 3 simply suggests that uniformity cannot be ruled out; it does not imply that the true distribution is uniform.

Displaying a sequence of boxplots in a single figure, as in the RHs examined here, has another potentially misleading consequence. Consider only two boxplots, for example. Such comparative boxplots are useful only when the cases within each boxplot are independent of the cases in the other boxplots. In the absence of independence - a situation known as matched or paired data - the comparison of the boxplots can be misleading. For instance, two boxplots with significant overlap might suggest that the two groups are statistically equivalent. However, if the data are in fact paired, then this conclusion is unjustified, because it may be that the cases in one boxplot are consistently higher than the cases in the other, on a case-by-case basis, in which case the correct conclusion would be that the groups are statistically distinct. With only two groups, i.e., two boxplots, it is sufficient to examine only the boxplot of the difference between the paired data. But with multiple groups (11 in the RHs), this solution does not apply. A more sophisticated test of independence between

groups is called for. For the current work, no systematic test of independence is performed on the 11 boxplots appearing in each RH. However, the correlation coefficient for the frequencies in adjacent boxplots was examined for a few instances, and no notable correlation was found. In short, the relative position of the boxplots in the RHs is likely to be informative, or at least not misleading.

As mentioned previously, this work does not emphasize statistical testing, but focuses on diagnosing the shape of the RH. The main reason is that such tests are almost guaranteed to result in statistical significance because of the extensive length of the time series. For example, the uniformity of the RH was assessed by a chi-squared test. The test was performed on the standardized data, as well as on the whitened data. Critical values of the chi-squared statistic are given in Wilks (2004) 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 the 3.0 to 21.0 range, for an *alpha*-level of 0.05 (Wilks, 2000). The observed values of the chi-squared statistic for temperature data are in the 200-500 range for standardized data, and 100-200 for whitened data⁶. The analogous values for wind speed are 300-500, and 35-68, respectively. All of these observed chi-squared values are much larger than the critical values, and so, uniformity of the RHs can be rejected. Again, this is mostly a consequence of the long time series, and is not surprising based on a visual examination of the RHs. The important question, however, addressed in this paper, is *how* one should interpret the deviation from uniformity.

⁶All of these ranges are the interquartile range.

Although in this work it has been unnecessary to quantify the sampling variability of the RH, it is possible to do so even for real data, where the population is not known. One can employ a resampling scheme (Efron and Tibshirani 1998). However, given that the data are temporally correlated, a resampling scheme must which accounts for that correlation. There exist such schemes - known as block bootstrap (Bühlmann 2002; Lahiri 2003; Politis, Romano, and Wolf 1999), but the extra computational effort is not called for, because of the relatively long time series generated from a 20-year data set. For this reason, the RHs for the realistic temperature forecasts are of the traditional type (i.e., without boxplots). A finer analysis, which does assess sampling variability will be done at a future time,

Additional future work involves a partitioning of the analysis by season. As mentioned previously, conditional bias is one explanation for the U-shaped RHs found at the end of the analysis here. It will be interesting to see if conditioning the analysis on the seasons will reduce this non-uniformity. It will also be interesting to distinguish between different variations on the U-shape, namely, V-shape, and Ends (Jolliffe and Primo 2008). Finally, although the geographic distribution of the RHs is examined here very briefly , a more substantive examination, as well as possible explanations for the observed geographic distribution will be proposed elsewhere.

Acknowledgments.

Don Percival is acknowledged for providing the R code for generating an AR(1) process. Nate Hiers helped in producing the ensemble reforecast dataset. Lynn Rose and Mark Bradford provided helpful discussions throughout the project. The computation 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.

6. Appendix: Simulating data with both correlations

In this appendix, it is shown that there are several ways of simulating ensemble forecasts in a way that includes a temporal correlation and a correlation between ensemble members; each way has certain advantages and disadvantages. For example, for an ensemble consisting of 10 members, it may be tempting to draw samples from a 10-dimensional multivariate normal distribution, with a correlation matrix with 1's along the diagonal, and ρ everywhere else; ρ would be interpreted as the correlation coefficient between any two ensemble members. However, such a matrix is not positive definite, and so, does not qualify as a correlation matrix. Another difficulty arises in deciding how the observed value (i.e., the verification), $y(t)$ should be related to the 10 ensemble forecasts, $x_i(t)$, $i = 1, \dots, 10$. Ideally, one would expect the ensemble members to be uncorrelated with one another, but correlated with the observations (if the forecasts have any accuracy in predicting the observed quantity). In practice, one would expect some sort of correlation between all 11 quantities, if the forecasts have any accuracy in predicting the observed quantity.

One alternative, which also allows for temporal correlations, would be to allow the observed values and the forecasts from **each** ensemble member, to be drawn from a bivariate

normal distribution with mean = 0, variance = 1, and correlation = ρ . In other words, $y(t)$ and $x_i(t)$ can be drawn from a standard bivariate normal with correlation ρ . It can be shown that the conditional distribution of the $x_i(t)$, given $y(t)$, also follows the normal distribution, but with mean = $\rho y(t)$, and variance = $(1 - \rho^2)$. Given that the mean parameter depends on $y(t)$, one can allow this parameter to follow an AR(1) process. In other words, $y(t)$ can be drawn from an AR(1) process, and for each value of $y(t)$ a sample of 10 $x_i(t)$ is drawn. In this way one can simulate data on 10 ensembles and corresponding observations, controlling the correlation among the 11 quantities through ρ , and the temporal correlation through ϕ .

Another alternative solution would be to 1) Draw the observations from an AR(1) process; 2) Generate the forecasts for the first member by adding a “small” zero-mean normally distributed sample to the observed values. 3) Repeat step 2 for the other members; 4) standardize each member’s forecasts to have mean = 0 and variance = 1. In this way, each ensemble member’s forecast is from an AR(1) process and is also correlated with all other members. Meanwhile, the AR(1) observations are also correlated with the forecasts.⁷ In this scheme the correlation between the ensemble members is implicitly controlled by the variance of the normal sample added to the first member. A variance of zero, leads to completely correlated members, while a large variance results in less correlated members.

One may object that the RH for such a system with correlated members is guaranteed to be non-uniform because the distribution from which the forecasts are drawn is not the same as that of the observations; that the former is a mixture of normals with different variances, while the latter is single standard normal. That objection would be valid were it

⁷This situation assumes that the forecasts have some accuracy. This assumption does not affect the generality of the analysis here because a RH tests the reliability of the forecasts, not their accuracy.

not for two facts: a) that the sum of normal variates is still a normal variate; and b) step 5 in the procedure assures that forecast and observation distributions have the same mean and variance (0 and 1, respectively). As such, any deviation of a RH from uniformity must be attributed to correlations between samples, and not differences in the two populations.

References

- 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, Blabdaoui and Raftery (TR 483)), do things similar to us and to Hamill. They even consider serial dependence with $\phi=0.5$ in Fig 3. More is on page 13 of their paper. They also say that Fruhwirth-Schnatter 1996 and Diebold et al (1998) also consider independence.
- 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**, 711-724.
- Jolliffe, I. T., C. Primo, 2008: Evaluating rank histograms using decompositions of chi-square test statistics. *Mon. Wea. Rev.*, **136**, 2133-2139.
- Lahiri, S.N., 2003: *Resampling methods for dependent data*. Springer, New York.

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

normal and nonnormal state space models, *Environmental and Ecological Statistics*, **3**, 291309.

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: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.*, **13**, 243-256.

List of Figures

1	Model domains, with the horizontal grid spacing of 135, 45, and 15 <i>km</i> for the outer, intermediate, and inner domain, respectively.	27
2	AR(1) time series with $\phi = 0.7, 0.3, 0, -0.3, -0.7$, from top to bottom, respectively.	28
3	A traditional RH for a single realization (top), and the boxplot variety based on multiple realizations (bottom) of forecasts and observations drawn from a multivariate AR(1) process, with no correlation between ensemble members.	29
4	RH for simulated forecasts and observations, with correlation between ensemble members, before (top) and after (bottom) whitening.	30
5	RHs for temperature forecasts across 90 stations nationwide.	31
6	Same as Figure 5, after bias has been removed.	32
7	Same as Figure 5, after bias and variance have been removed.	33
8	Same as Figure 5, after bias, variance, and correlation between ensemble members have been removed.	34
9	RHs for wind-speed forecasts across 90 stations nationwide.	35
10	Same as Figure 9, after bias has been removed.	36
11	Same as Figure 9, after bias and variance have been removed.	37
12	Same as Figure 9, after bias, variance, and correlation between ensemble members have been removed.	38

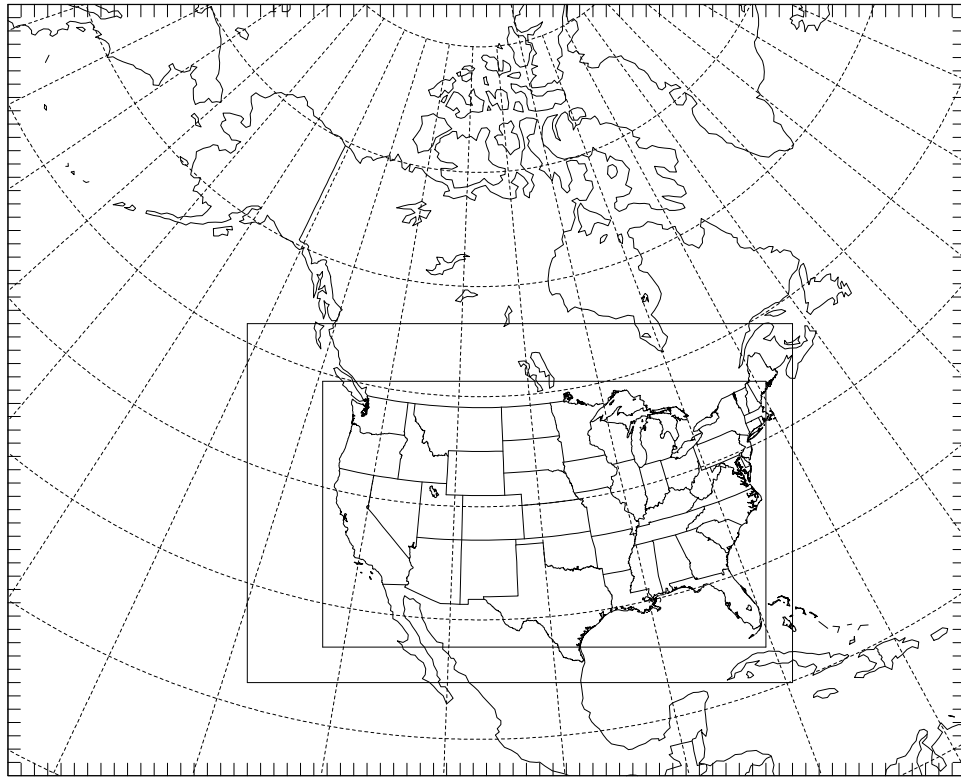


FIG. 1. Model domains, with the horizontal grid spacing of 135, 45, and 15 km for the outer, intermediate, and inner domain, respectively.

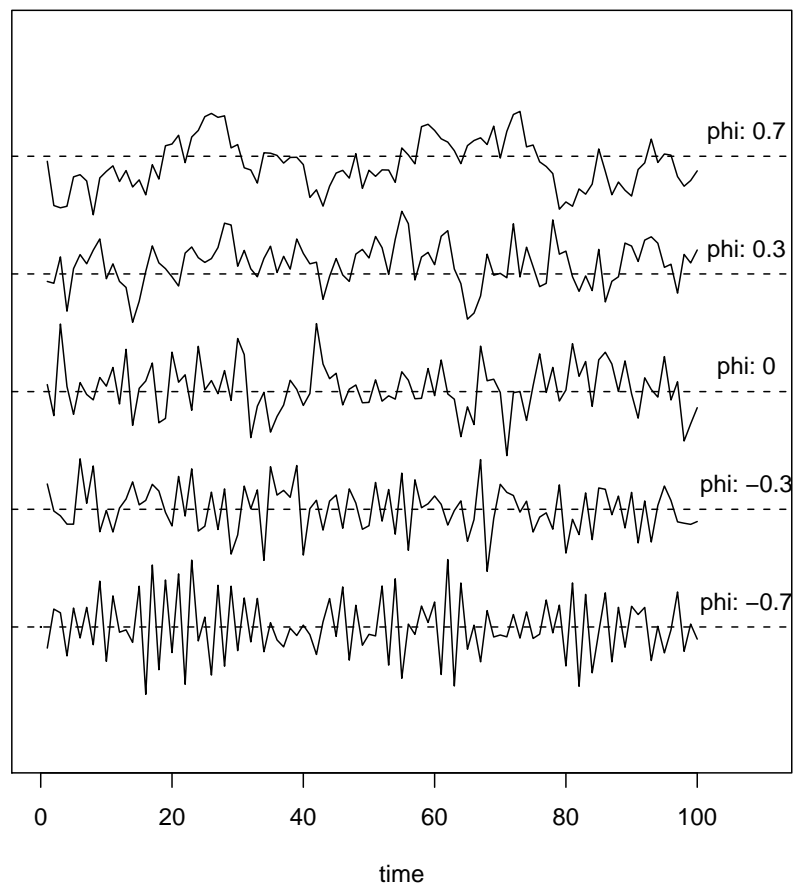


FIG. 2. AR(1) time series with $\phi = 0.7, 0.3, 0, -0.3, -0.7$, from top to bottom, respectively.

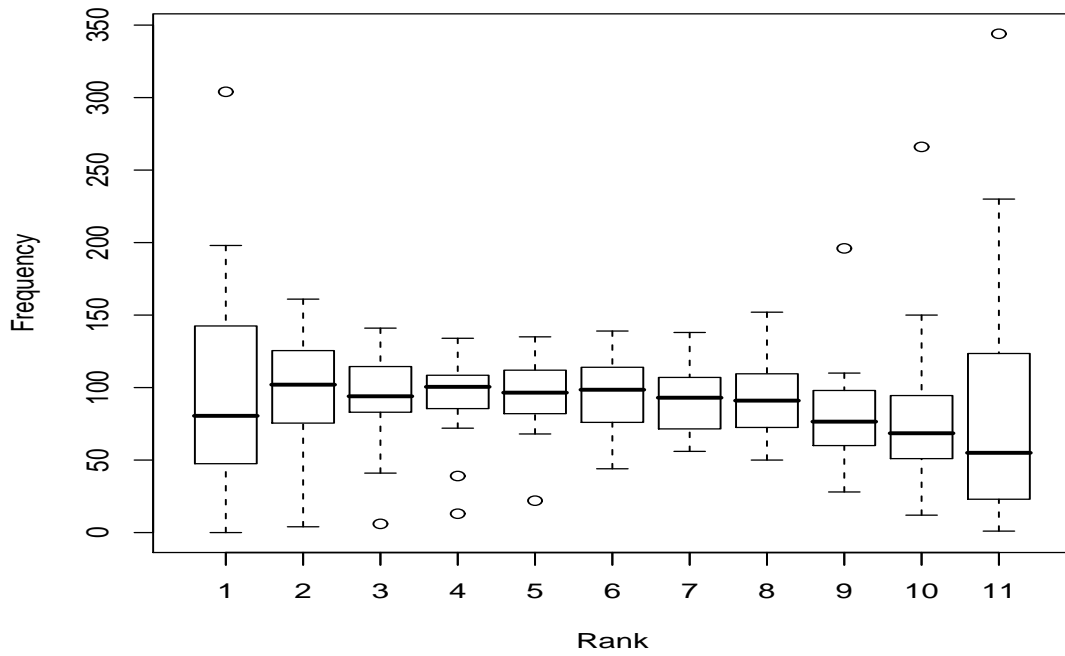
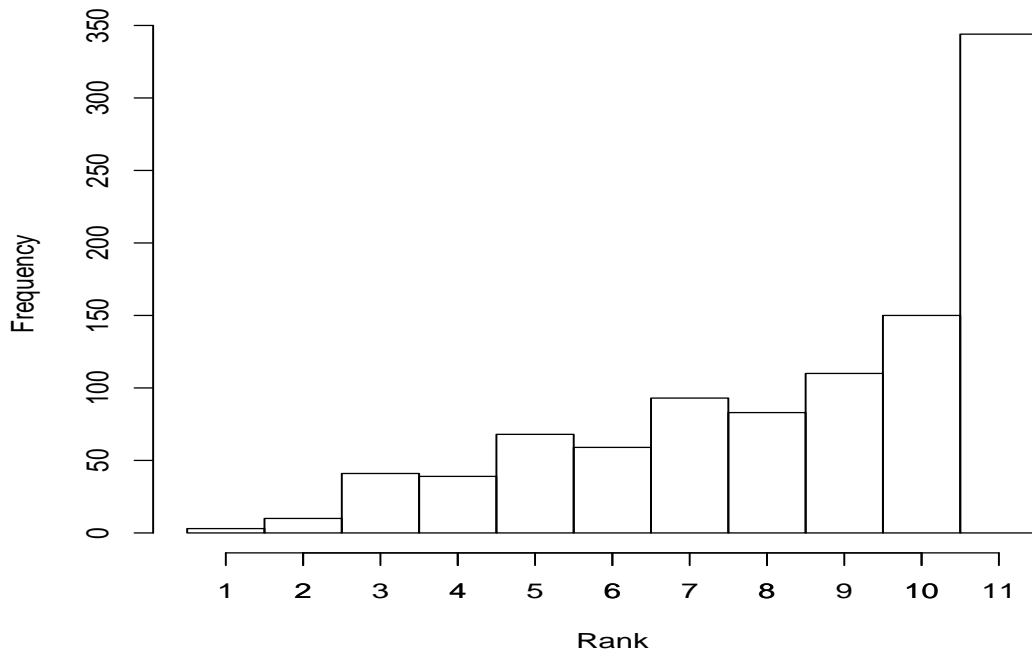


FIG. 3. A traditional RH for a single realization (top), and the boxplot variety based on multiple realizations (bottom) of forecasts and observations drawn from a multivariate AR(1) process, with no correlation between ensemble members.

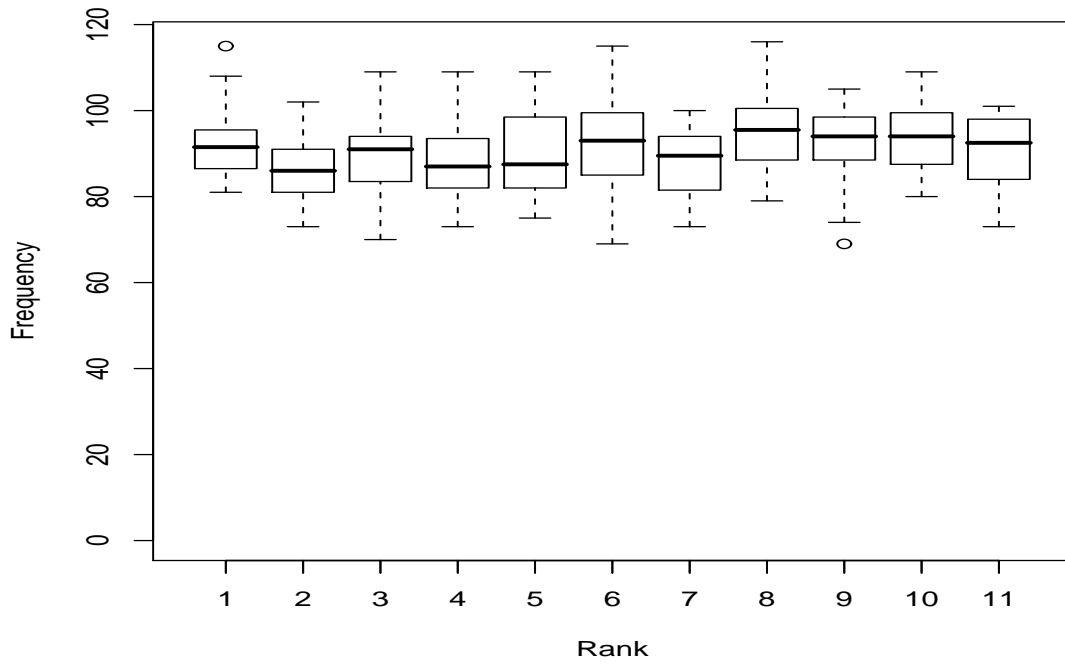
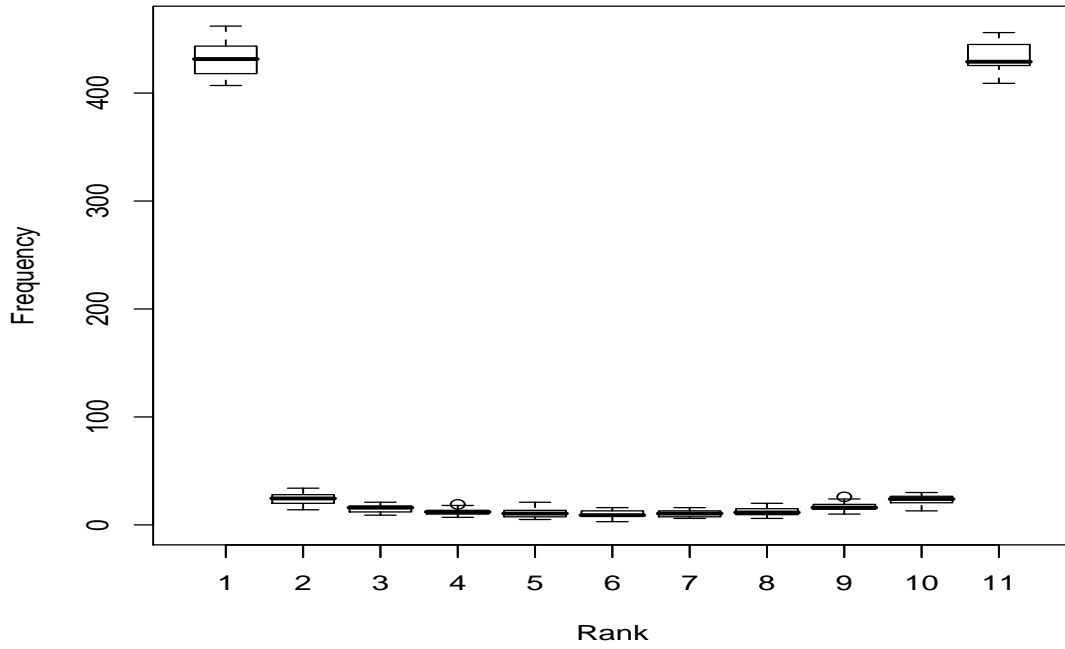


FIG. 4. RH for simulated forecasts and observations, with correlation between ensemble members, before (top) and after (bottom) whitening.

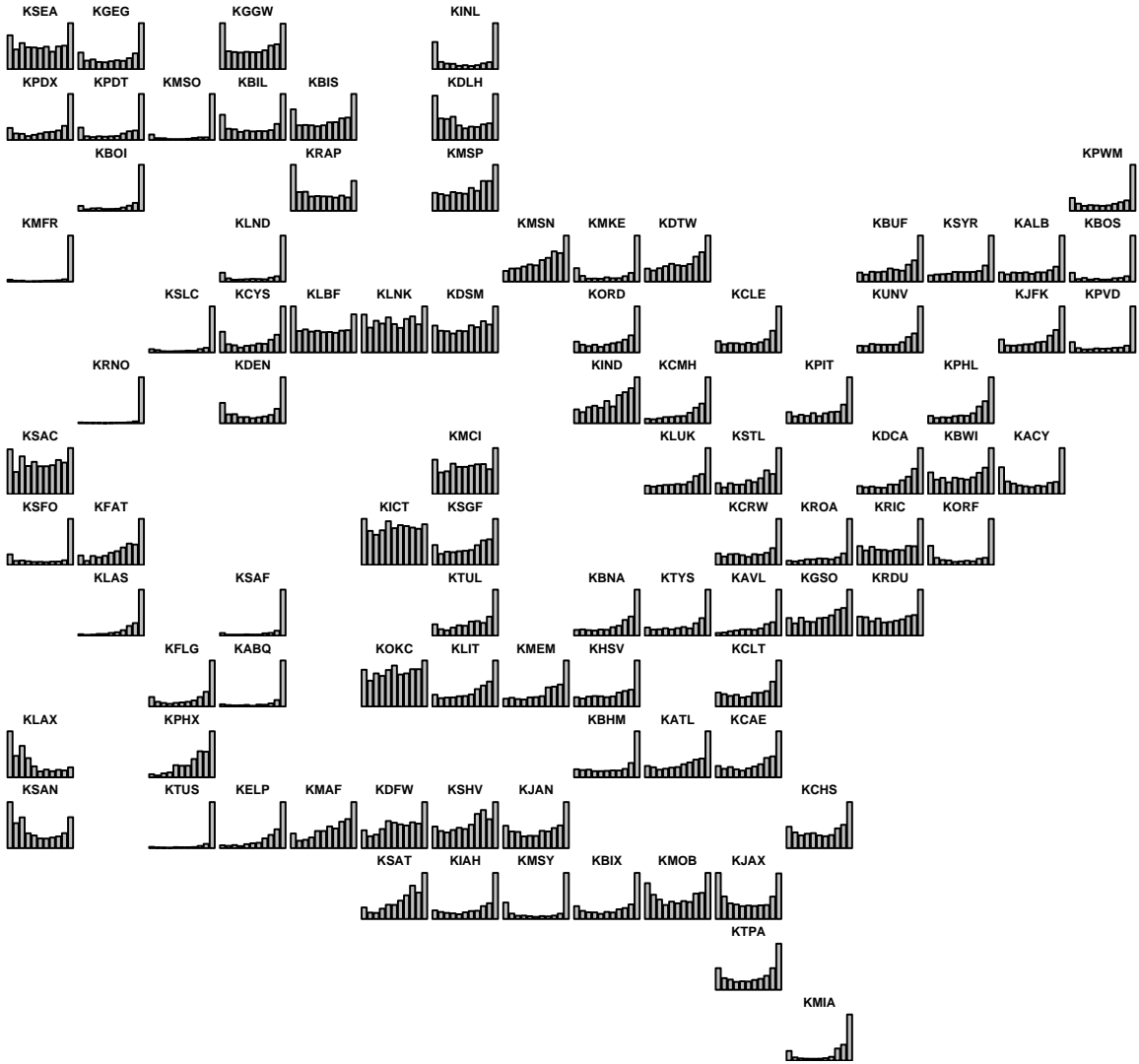


FIG. 5. RHs for temperature forecasts across 90 stations nationwide.

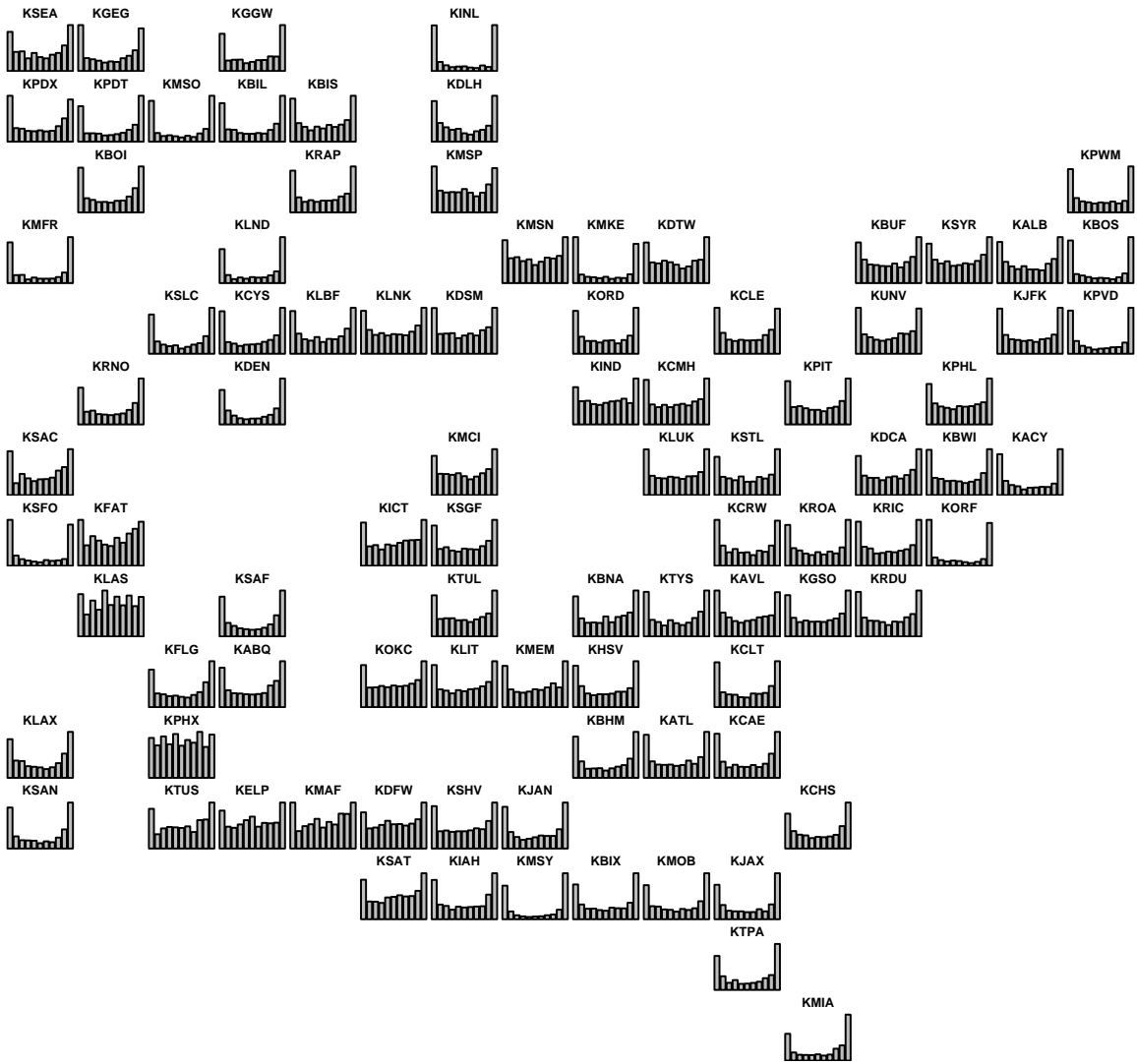


FIG. 6. Same as Figure 5, after bias has been removed.



FIG. 7. Same as Figure 5, after bias and variance have been removed.

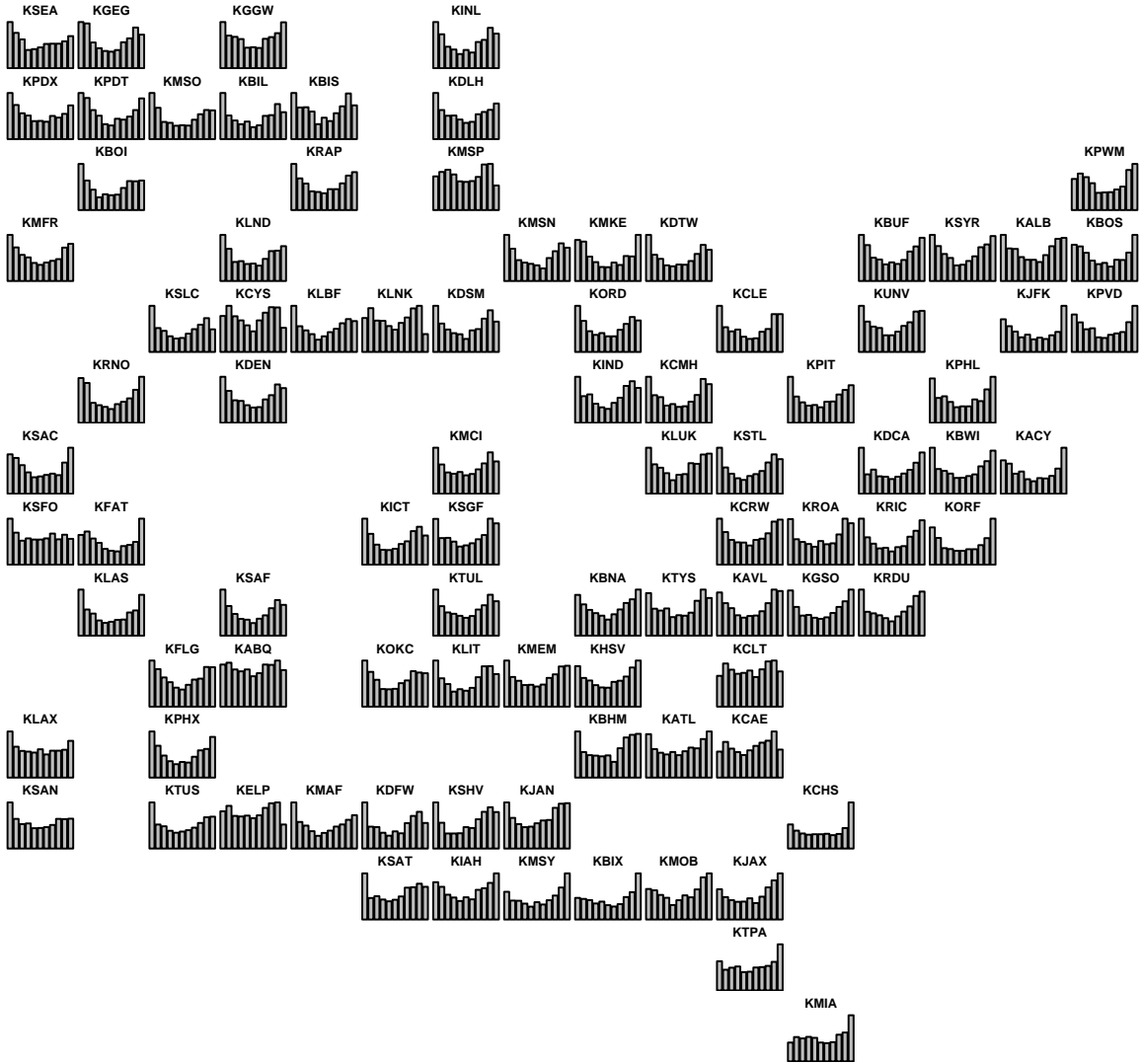


FIG. 8. Same as Figure 5, after bias, variance, and correlation between ensemble members have been removed.

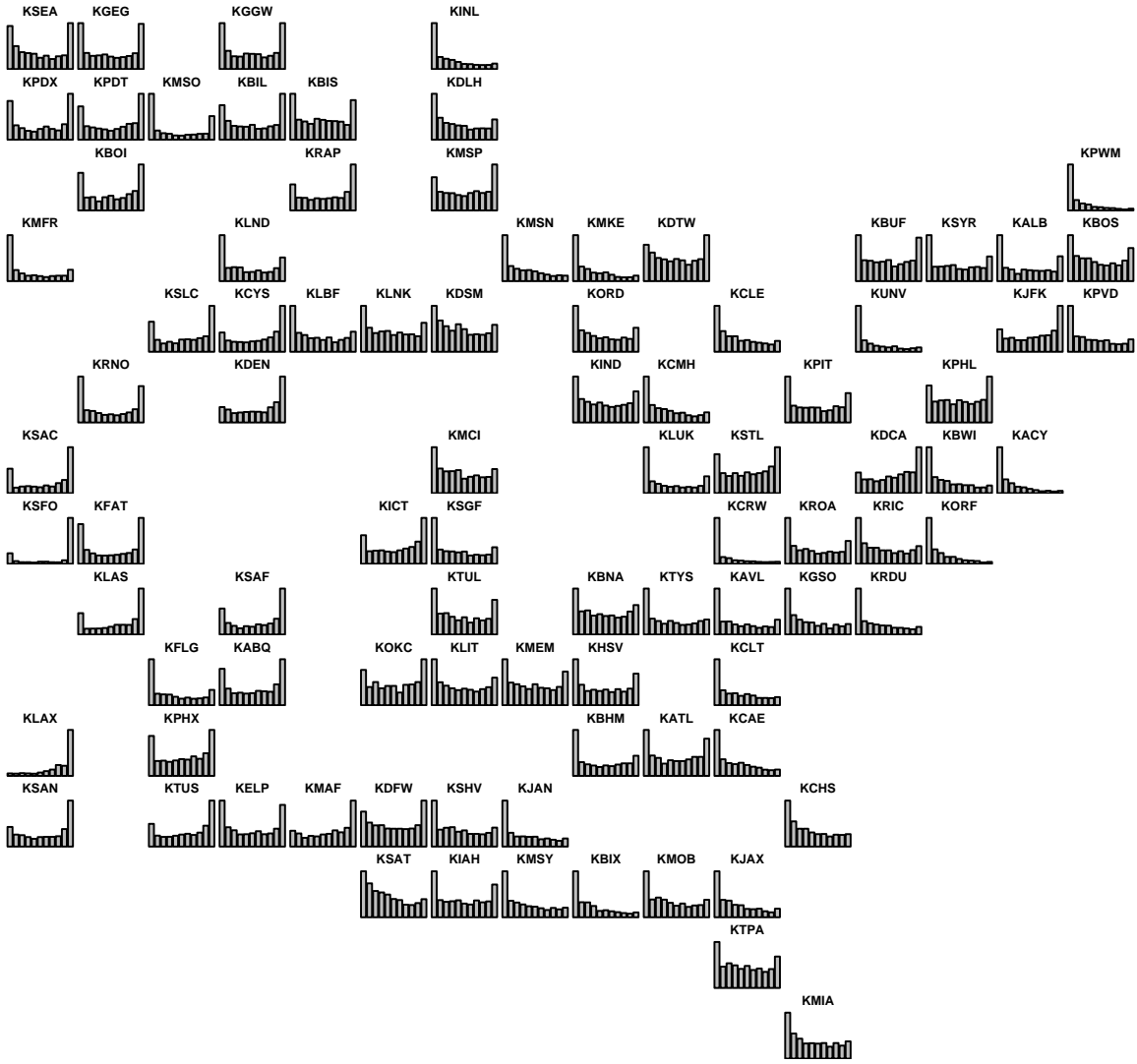


FIG. 9. RHs for wind-speed forecasts across 90 stations nationwide.



FIG. 10. Same as Figure 9, after bias has been removed.



FIG. 11. Same as Figure 9, after bias and variance have been removed.

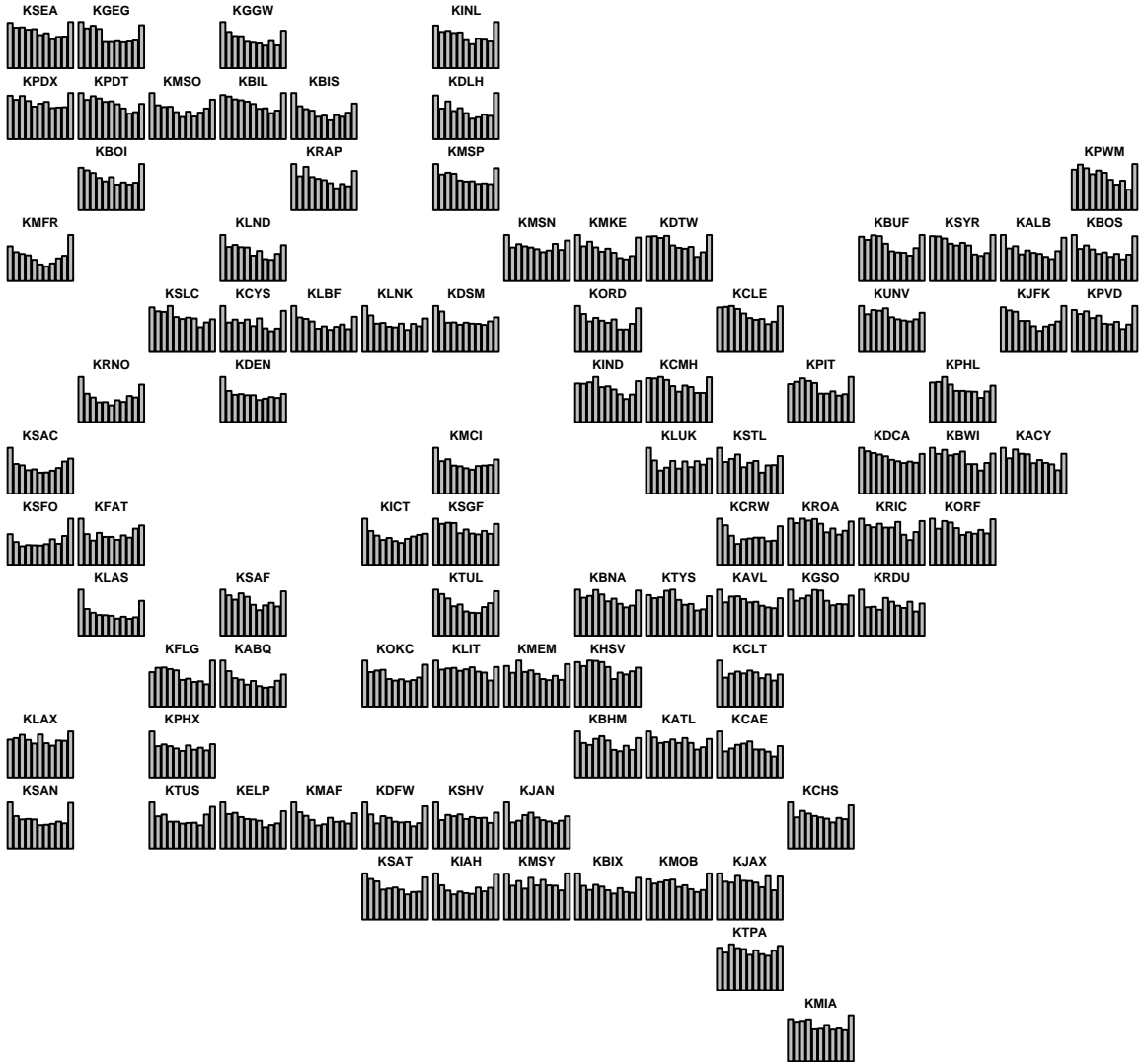


FIG. 12. Same as Figure 9, after bias, variance, and correlation between ensemble members have been removed.