

A Validation of the National Centers for Environmental Prediction's Short Range Ensemble Forecast

Andrew Hamm

Oklahoma Weather Center Research Experience for Undergraduates Program
NOAA/National Severe Storms Laboratory, Norman, Oklahoma
Northland College, Ashland, WI

Kimberly Elmore

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

Research Experience for Undergraduates Final Project

31 July 2003

Corresponding author address:

Andrew Hamm
Northland College
1411 Ellis Avenue
Ashland, WI 54806
hamma01@northland.edu

ABSTRACT

This paper investigates the performance of soundings generated from the National Centers for Environmental Prediction's Short Range Ensemble Forecast (NCEP SREF). The NCEP SREF is an operational ensemble forecast model with 15 members. Rank histograms are used as the primary tool to investigate consistent bias problems as well as ensemble dispersal. For the period spanning 1 May 2003 and 19 July 2003, nine different locations scattered about the continental U.S. are validated with rawinsonde data. Ensembles modified by a lagged bias correction and ensembles modified by both a lagged bias correction and the addition of observational errors are considered.

Rank histograms constructed from the unmodified ensemble imply either severe bias problems in the ensemble or a significantly underdispersed ensemble, depending on the variable examined, forecast time, pressure level, and location. Because forecasts between the different locations are poorly correlated, the assumption of independence is acceptable and rank histograms for each location are merged into combined rank histograms for all cities for a given variable, forecast time, and pressure level to produce adequate sample sizes. Combined rank histograms constructed from the bias corrected ensemble are U-shaped, which may be caused either by an under-dispersed ensemble, a non-homogeneous bias structure, or observational errors. However, including observational errors with the bias correction often results in uniform, or occasionally over-dispersed, rank histograms. Analysis of other factors, including the non-homogeneous biases of the ensemble, is shown to help understand the combined rank histograms. Without the bias correction, this ensemble is of limited utility, but the lagged bias correction greatly enhances the ensemble performance.

1. Introduction

An ensemble forecast is a collection of forecasts that verify at the same time. Each member may consist of identical models initialized with different, but equally plausible, initial conditions, different models, or identical models with differing parameterizations (Sivillo et al. 1997). An ensemble may contain as few as two members, but typically contain many more. Among the objectives of ensemble forecasting is to improve forecasting skill on a case by case basis, which is forecast accuracy (Murphy, 1993), through averaging.¹

The ensemble evaluated here is provided by the National Centers for Environmental Prediction (NCEP). This ensemble contains 15 individual models, which can be equally divided into three model families. The three model families in this ensemble are the Eta, the Kain-Fritsch Eta, and the regional spectral model, or RSM. All of the models in this ensemble are run with 48 km horizontal grid resolution¹. The perturbations for these models are generated by the breeding method¹. This study evaluates the ensemble to determine if the distribution of forecasts matches the distribution of verifying observations as a function of height.

One quality of an accurate forecast is a match between the distributions of the forecasts and verifying observations. By definition, an ensemble provides a range of possible forecast values for a certain variable at a given location and time; however, only an accurate ensemble will consistently forecast an acceptable range of values (e.g., Sivillo et al. 1997). An example of an ensemble forecast of the 850-hPa temperature at a given location 15 hours in the future may range from 290 K to 295 K. An ensemble forecast output can also be processed to a probabilistic forecast (e.g., Hamill 2001). Ranges of forecast values and probabilistic forecasts are two of the most powerful products provided by ensemble forecasts.

¹ Source: <http://wwwt.emc.ncep.noaa.gov/mmb/SREF/FCST/DOC/present4/main.htm>,

Rank histograms are among the tools used to evaluate the performance of this ensemble. The first step in constructing a rank histogram is to rank the individual forecasts of an n -member ensemble for a given variable at a given location, pressure level, and forecast time, from lowest value to highest value, forming a sorted list (e.g., Hamill 2001). Next, the verifying observation is included in the sorted list, which now contains $n+1$ values. There are $n+1$ possible ranks for the observation in this sorted list; the rank of the verifying observation is defined as one if its value is less than the ensemble member with the lowest rank and $n+1$ if its value is greater than the ensemble member with the highest rank. A rank histogram is constructed by summing the rank of the verification, or tally, over all days into one histogram for one given variable, location, pressure level, and forecast time (Hamill, 2001). The x-axis on a rank histogram consists of the observation rank, while the y-axis consists of the total number of days that the observation had that rank.

Interpretation of rank histograms may provide insight into the accuracy of the ensemble. A uniform rank histogram may imply that both the ensemble and the verification are drawn from indistinguishable distributions, whereas a non-uniform rank histogram implies that the ensemble and verification are drawn from different distributions (Hamill 2001). Rank histograms may allow for a quick diagnosis of problems within an ensemble, especially with the variability of the ensemble. For example, a rank histogram with high frequency counts at both extremes suggests one of several problems (Fig. 1). The ensemble may be underdispersive (e.g., Hamill and Colucci, 1997), there may be errors in the observational data (Hamill 2001), or systematic errors in the forecast may be present (Hamill and Colucci, 1997). Another possibility is a combination of these problems is creating the observed shape in the rank histogram.

Another problem is indicated when a rank histogram has high frequency counts near one extreme and low frequency counts near the other extreme; this is a symptom of a consistent bias in the ensemble. A rank histogram with high frequency counts near the center, but low frequency counts at the extremes suggests too much variability in the ensemble; a more precise ensemble may be necessary. Rank histogram uniformity may be tested statistically with a chi-squared goodness-of-fit test (e.g., Hamill and Colucci, 1997). Rank histograms that are consistently uniform suggest an accurate ensemble. This significance test has been effectively used in interpretations of rank histograms in past studies (Hamill and Colucci, 1997).

Ensembles have been evaluated using rank histograms in the past; these studies provide motivation for this investigation (Hamill and Colucci, 1997). Hamill and Colucci (1997) focus not only on the accuracy of the ensemble, but also on a comparison between a 15-member ensemble, with less resolution than the current ensemble, and a 29-km mesoeta model. They also considered precipitation, which is not considered here.

Section 2 of this paper presents the data used for this research. Section 3 presents the results of this research. Rank histograms, along with modifications to the ensemble, will be presented in this section. Section 4 will provide a conclusion, including some discussion into the applications of this research.

Section 2. Data and Methodology

This study utilizes verification soundings and model soundings. The verification data comes from the Forecast Systems Laboratory's web site for radiosonde data, which is currently <http://raob.fsl.noaa.gov>; the model data comes from NCEP. Model data are specific to the location of interest. Nine radiosonde sites, distributed about the continental US, were selected as locations to evaluate the validation of this ensemble (Fig. 2). Model data span 1 May 2003

through 18 July 2003, while verification data span 2 May 2003 through 19 July 2003. Days are missing due to either missing forecast data or verification data. NCEP's ensemble forecast is run at 0900 UTC and 2100 UTC every day with forecasts out to 63 hours; only the 0900 UTC forecast cycle is used here. Hence, forecast soundings for 15 hours, 39 hours, and 63 hours are validated because these forecasts all verify at 0000 UTC, for which verification soundings are available.

The soundings in both data sets contain pressure, temperature, dewpoint, wind direction, wind speed, and geopotential height. The model soundings have 50 hPa vertical resolution, while the verification soundings are linearly interpolated to every 50 hPa. The lowest pressure level used for most sites is 950 hPa (except 900 hPa for OTX and TUS, and 800 hPa for RIW), while the highest pressure level used is 150 hPa for all sites. The soundings contain dewpoint, which is converted to mixing ratio, wind speed and direction, which are converted to u and v components, temperature and geopotential height, all of which are used in the construction of the rank histograms.

Rank histograms are constructed for each variable, location, forecast time, and pressure level. Rank histograms for all locations for a given variable, forecast time, and pressure level are combined, under the assumption that the selected radiosonde sites are statistically independent, even though this is not strictly the case. All sites are separated by at least several hundred kilometers, which should severely limit any spatial correlation (Hamill and Colucci, 1997). Statistical independence between sites is necessary to interpret a combined rank histogram for a given variable, forecast time, and pressure level (Hamill 2001). These combined rank histograms contain more tallies, which reduces problems associated with the small sample sizes in the individual site rank histograms (Hamill and Colucci, 1997).

The range of the ensemble is examined to obtain an understanding of the changes in the spread of the ensemble as the variable, location, forecast time, and pressure level are changed. The 95% confidence interval for the range of the ensemble over all days is obtained through a t-test.

Section 3. Results

The range describes the spread of the ensemble and can be used to determine how the ensemble output changed as the variable, location, forecast time, and pressure level changed. A high range suggests a large spread in the ensemble and high uncertainty, while a low range suggests a small spread in the ensemble and lower uncertainty. The 95 percent confidence interval for the range, which is defined as the minimum value subtracted from the maximum value, is calculated and plotted, along with the mean range, as a function of height for a given variable, location, and forecast time. Figure 3 shows an example of the range of the ensemble for temperature for all forecast times at CHH, MFL, and IAD. There are many pressure levels where at least two of the confidence intervals for different locations intersect for the same variable and forecast time, which suggests that the ranges at these different locations are statistically equal at these pressure levels where the intersections occur. The range of the ensemble increases as forecast lead time increases, regardless of variable, forecast time, and location. The range of the ensemble changes as height increases, regardless of variable, forecast time, and location.

Three sets of rank histograms are examined: one set uses raw model data, another set uses a bias correction applied to ensemble members, and a final set adds noise to the model data. These modifications are discussed by Hamill and Colucci (1997) and Hamill (2001), respectively. The bias correction is used to improve the performance of this ensemble, while the

addition of noise to the model data is only used to counteract observational errors and provide a proper validation.

For mixing ratio from the raw ensemble, most of the tallies in the combined rank histograms fall into the first rank at pressure levels near the surface, which implies a moist bias; these rank histograms are U-shaped from 700 hPa to 500 hPa for 15 and 39-hour forecasts, which implies either an under-dispersed ensemble, a biased ensemble, or systematic errors in the forecast (Fig. 4). An under-dispersed ensemble is implied in most of the 63-hour forecast combined rank histograms for mixing ratio; exceptions include the pressure levels from 350 hPa to 250 hPa, where a moist bias in the ensemble is implied. For temperature, most of the tallies in the combined rank histogram are contained in the first rank, which implies a warm bias (Fig. 4), except for U-shaped rank histograms in the 63 hour forecasts from 500 hPa through 300 hPa. For geopotential height, most of the tallies in the combined rank histograms fall in the first rank, which implies a positive bias in the ensemble. In the combined rank histograms for u and v components of the wind, an underdispersive ensemble is implied over all pressure levels and forecast times (Fig. 4). Individual site rank histograms are examined to find a possible explanation for the shapes of these combined rank histograms.

When the individual site rank histograms for the u and v components of the wind are examined, an under-dispersed ensemble is not the only explanation for the U-shaped combined rank histograms. For the u component of the wind, except for 900 hPa, the individual site rank histograms for Miami for all forecast times and at pressure levels near the surface, a negative bias (most tallies in the 16 rank) in the ensemble is implied (Fig. 5), while an under-dispersed ensemble is implied by the individual site rank histograms above 650 hPa. In Spokane, a positive bias (most tallies in the 1 rank) in the ensemble is implied by the individual site rank

histograms for all pressure levels and forecast times, except at 900 hPa and at 800 hPa through 500 hPa, where an under-dispersed ensemble is implied by the individual site rank histograms (Fig. 5). In Minneapolis, a positive bias in the ensemble is implied by the individual site rank histograms near the surface (Fig. 5), except 900 hPa, and a negative bias in the ensemble is implied above 350 hPa. In Tucson, a positive bias in the ensemble is evident at pressure levels near the surface (Fig 5), except for 900 hPa, while a negative bias in the ensemble is implied above 500 hPa. Individual site rank histograms for the v component of the wind typically show various biases in the ensemble at different locations and different pressure levels. The biased combined rank histograms will not be examined further because a bias in the ensemble is clearly the cause of the shape of the rank histogram; possible explanations for the U-shaped rank histograms will be covered in the next section

Each model family could contain a different bias. Hence, the 95% confidence interval and the mean bias for a given family is calculated over all days for a given variable, location, forecast time, and pressure level. Figure 6 shows an example of how the biases for the three model families can change as a function of height. Differences in the biases among the three model families at different heights are noted for different variables, locations, and forecast times. Differing biases among the model families comprising the ensemble can complicate the interpretation of the rank histograms (Hamill, 2001).

Another feature of note is the rank histograms, both individual site and combined, for the 150-hPa level. These combined rank histograms imply both positive and negative biases, regardless of the combined rank histograms for other pressure levels for the same variable and forecast time. An example of a 150-hPa level combined rank histogram that is different from other rank histograms at different pressure levels for the same variable and forecast time is the

combined rank histogram for temperature for a 63-hour forecast (Fig. 7). One explanation for the shapes of these rank histograms involves the inability of models to provide an accurate forecast at this level in the atmosphere. Another possible explanation involves our inability to accurately observe the high levels of the atmosphere.

Many rank histograms, both individual and combined, imply non-homogeneous biases in the ensemble; a bias correction is the first modification to the model data. A mean seven-day lagged bias is calculated every day for each individual model, variable, location, forecast time, and pressure level. This lagged bias is added to the forecasted value for the given model, current day, correct pressure level, variable, location, and forecast time.

Various lag intervals were examined. Qualitatively, a seven-day lagged bias is more effective than a four, five, or six day lagged bias. Lag intervals greater than seven days are not considered. The shapes of many of the rank histograms change from sloped to U-shaped when this seven-day lagged bias correction is applied (Fig. 8). This simple bias correction proves a very effective way to remove bias from the ensemble. However, the rank histograms still show underdispersion. Based on Hamill's (2001) work, the apparent underdispersion may be illusory.

According to Hamill (2001), it might be necessary to add random noise to the model data to counteract random, nonbiased errors in the verification data. Hence, noise is added to the model soundings based on error estimates in Zapotocny et. al (2000). The rank histograms created from this ensemble with the noise addition are significantly different from rank histograms without noise. Rank histograms that include observational error do not suffer underdispersion as often as uncorrected rank histograms. In some cases, the resulting rank histograms suggest an overdispersive ensemble; nearly all of these rank histograms with overpopulated middle ranks were constructed from 15-hour forecasts (Fig. 9). However, nearly

all of the combined rank histograms for mixing ratio still show a significantly under-dispersed ensemble; these U-shaped rank histograms may be the result of the difficulty to accurately predict and measure moisture in the atmosphere. Far fewer combined rank histograms for any other variable were U-shaped when both modifications were added to the model data. In fact, many of the combined rank histograms are nearly uniform, as shown by a chi-squared goodness-of-fit test with a rejection criterion equal to 0.05 (Table 2). None of the combined rank histograms passed a chi-squared goodness-of-fit test before any modifications were added to the ensemble. The null hypothesis for a chi-squared goodness-of-fit test is that the given distribution is statistically uniform, while the alternative distribution is that the given distribution is not uniform. Over 47% of the combined rank histograms with the 15-hour forecast lead time did not reject the null hypothesis. Nearly 51.5% of the combined rank histograms with the 39-hour forecast lead time did not reject the null hypothesis. Over 41% of the combined rank histograms with the 63-hour lead time did not reject the null hypothesis. All of the rank histograms from the ensemble without any modifications rejected the null hypothesis and were not statistically uniform; the modifications to the ensemble produced an increase of 46.57% in the number of rank histograms that did not reject the null hypothesis and are considered to be statistically uniform, regardless of forecast time. Many of the rank histograms that rejected the null hypothesis had a p-value that was near 0.05, except for the rank histograms constructed for mixing ratio. Most of the mixing ratio rank histograms had p-values that were very close to zero.

Section 4. Conclusion

The distributions of the forecasts and verifying observations are clearly distinct for the raw ensemble output. However, the models within this ensemble have biases, which are not

always equivalent. Many of the combined rank histograms for geopotential height, mixing ratio, and temperature imply a significant bias in the ensemble, while a small number of other combined rank histograms for the same variable and forecast time, but at different pressure levels, imply an under-dispersed ensemble. Individual site rank histograms clearly reveal non-homogeneous biases within the ensemble. Hence, the combined rank histograms that resulted from the individual site rank histograms become ambiguous. For example, the combined rank histograms for the u and v components of the wind imply an under-dispersed ensemble, while the individual site rank histograms for these two variables imply different biases in the ensemble at different locations and different pressure levels. In addition, the model families exhibit different biases at different heights for different variables, locations, and forecast times, which complicates rank histogram interpretation. According to Hamill (2001), when rank histograms, constructed from an ensemble whose members have different biases, are combined, the resulting rank histogram may erroneously suggest an under-dispersed ensemble, which appears to be the case here for u and v . Yet, even though the ensemble has different biases at different locations and pressure levels, the ensemble may still be under-dispersed and the shape of the combined rank histogram may be a result of the opposite biases and the lack of variability in the ensemble.

However, when the lagged bias correction is applied to the model data, many of the combined rank histograms change from implying a bias to implying an under-dispersed ensemble. Some combined rank histograms that imply a severely under-dispersed ensemble before the bias correction change to imply a less-severe underdispersion problem when the lagged bias correction is added to the model data. Even though many of the combined rank histograms still imply an under-dispersed ensemble, the performance of the ensemble appears to

improve with the addition of the lagged bias correction. Hence, applying the bias correction is an important post-processing step for the NCEP SREF.

There is a significant change in the appearance of both the combined rank histograms and the individual site rank histograms after observational noise is added to the model data. However, too much noise will result in rank histograms that are incorrectly uniform or incorrectly imply an over-dispersive ensemble, and too little noise will not effectively counteract the errors inherent in the verifying observations. However, these modifications clearly improve the validation statistics for the ensemble. Before modification, none of the ensembles pass a chi-squared goodness-of-fit test, but after modification, 47.06% of the 15-hour forecasts, 51.47% of the 39-hour forecasts, and 41.18% of the 63-hour forecasts passed a chi-squared goodness-of-fit test. If the noise addition is appropriate, these rank histograms imply that if a bias correction is applied to this ensemble, the performance of the ensemble is, in fact, quite good.

Without a bias correction, the ensemble provides poor forecast guidance. The addition of a bias correction to the ensemble appears to significantly enhance the ensemble's utility. This is not the case when accounting for observational errors. Noise is only used to assess the accuracy of the ensemble by counteracting the errors that cannot be avoided in the collection of the observational data; including observational errors is inappropriate operationally because, ideally, a forecast should be made for weather, not observations. The goal of an individual forecast is to accurately predict the actual weather, not the errors in the observations. Observational errors are used only in assessing the statistical performance of the ensemble.

Acknowledgements. This material is based on work supported by the National Science Foundation under Grant No. 0097651. I would like to thank Kim Elmore for his patience and

support all summer. I would like to thank all of the people at the National Severe Storms Laboratory and Oklahoma University who make this Research Experience for Undergraduates possible, especially Daphne Zaras, the director.

REFERENCES

- Hamill, Thomas M. 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*: **Vol. 129, No. 3**, pp. 550–560.
- , and Colucci, Stephen J. 1997: Verification of Eta–RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*: **Vol. 125, No. 6**, pp. 1312–1327.
- Murphy, Allan H. 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*: **Vol. 8, No. 2**, pp. 281–293.
- Sivillo, Joel K., Ahlquist, Jon E., Toth, Zoltan. 1997: An Ensemble Forecasting Primer. *Weather and Forecasting*: **Vol. 12, No. 4**, pp. 809–818.
- Zapotocny, Tom H., Nieman, Steven J., Menzel, W. Paul, Nelson, James P., Jung, James A., Rogers, Eric, Parrish, David F., DiMego, Geoffrey J., Baldwin, Michael, Schmit, Timothy J. 2000: A Case Study of the Sensitivity of the Eta Data Assimilation System. *Weather and Forecasting*: **Vol. 15, No. 5**, pp. 603–622.

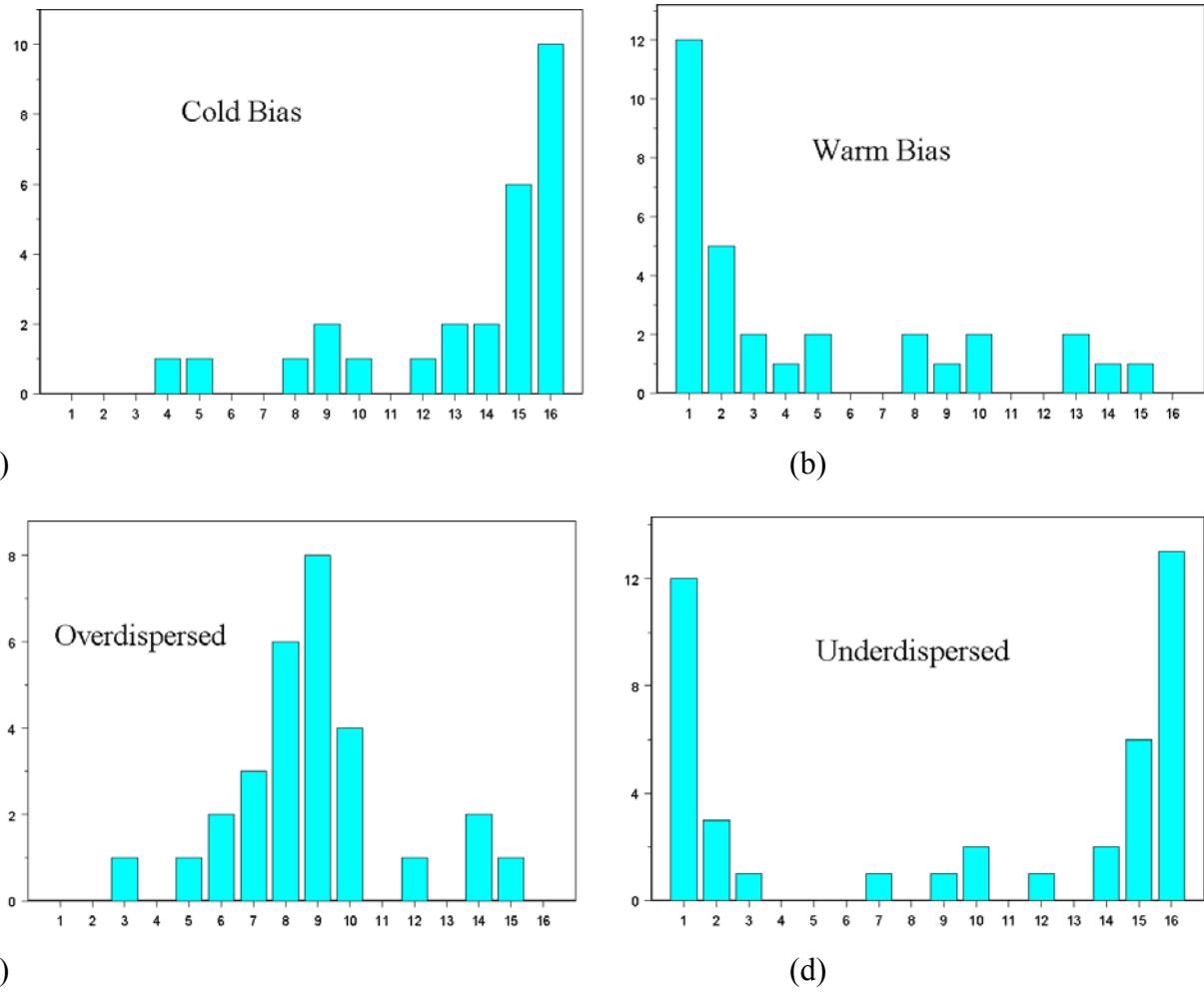


Figure 1. Examples of the four main shapes of rank histograms and the problem with the ensemble from which each rank histogram was constructed. (a). Example of a rank histogram constructed from an ensemble with a cold bias. (b). Example of a rank histogram constructed from an ensemble with a warm bias. (c). Example of a rank histogram constructed from an over-dispersed ensemble. (d). Example of a rank histogram constructed from an under-dispersed ensemble.

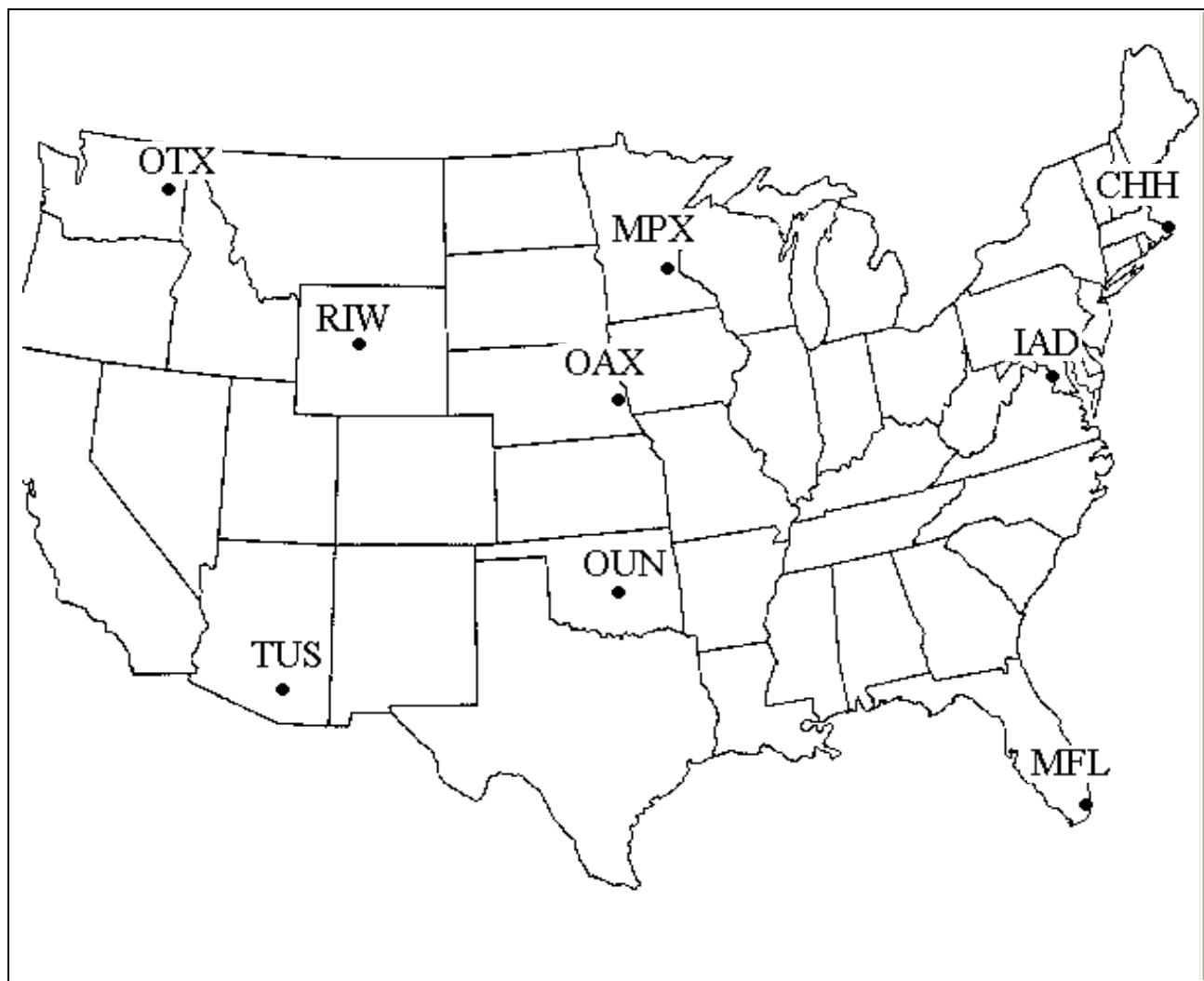


Figure 2. Map of the United States showing the locations of the nine radiosonde sites.

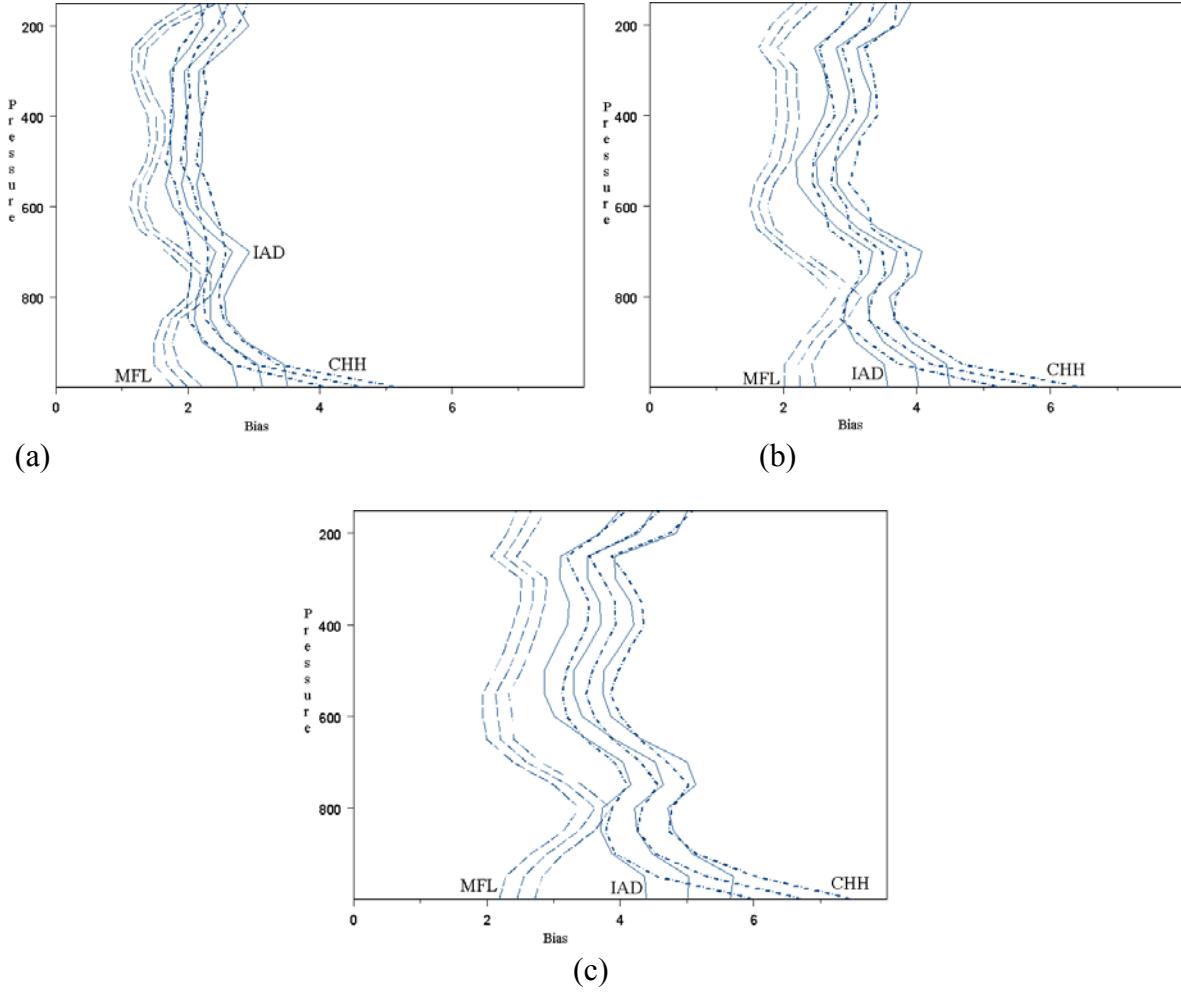


Figure 3. Changes in the range of the ensemble at CHH, IAD, and MFL and several different pressure levels for temperature for 15, 39, and 63-hour forecasts. The three lines for each model family correspond to the 95% confidence interval and mean bias. (a). Range of the ensemble for a 15-hour forecast. (b). Range of the ensemble for a 39-hour forecast. (c). Range of the ensemble for a 63-hour forecast.

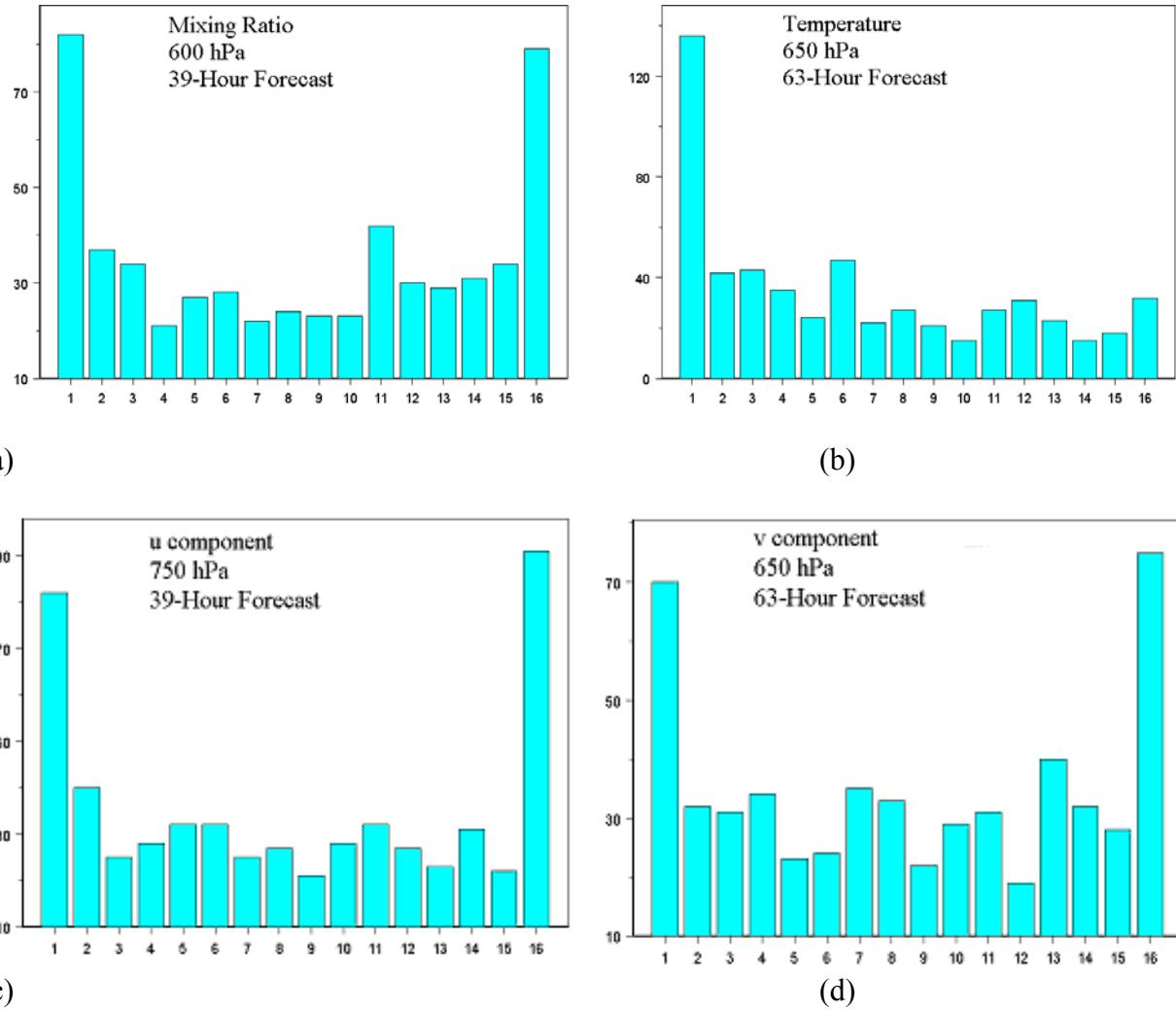


Figure 4. Combined rank histograms constructed from the ensemble without any modifications.

(a). Combined rank histogram possibly implying an under-dispersed ensemble for mixing ratio at 600 hPa for a 39-hour forecast.. (b). Combined rank histogram implying a warm bias in the ensemble for temperature at 650 hPa for a 63-hour forecast. (c). Combined rank histogram possibly implying an under-dispersed ensemble for u component at the 750 hPa level for a 39-hour forecast. (d) Combined rank histogram implying an under-dispersed ensemble for v component at 650 hPa for a 63-hour forecast.

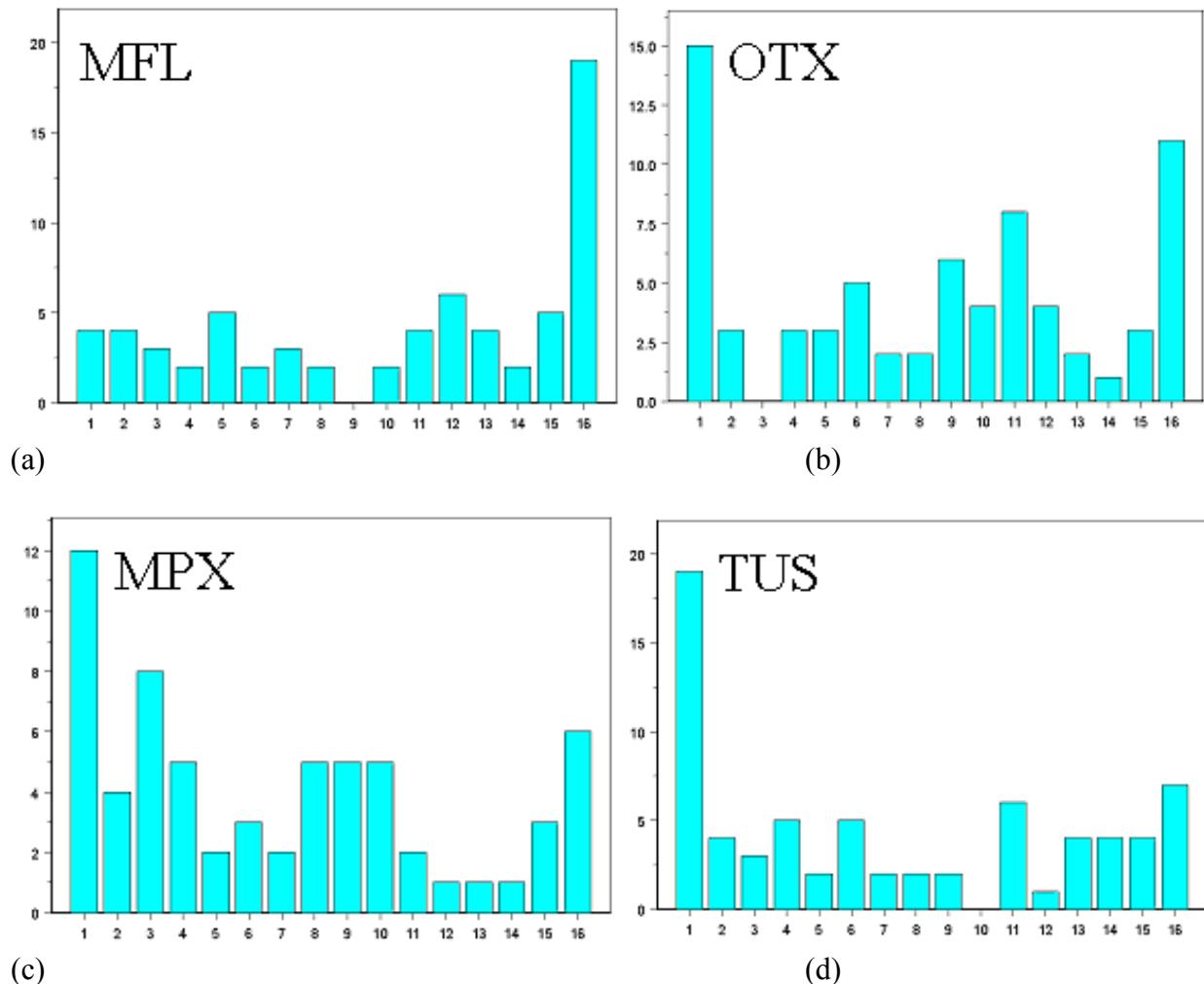
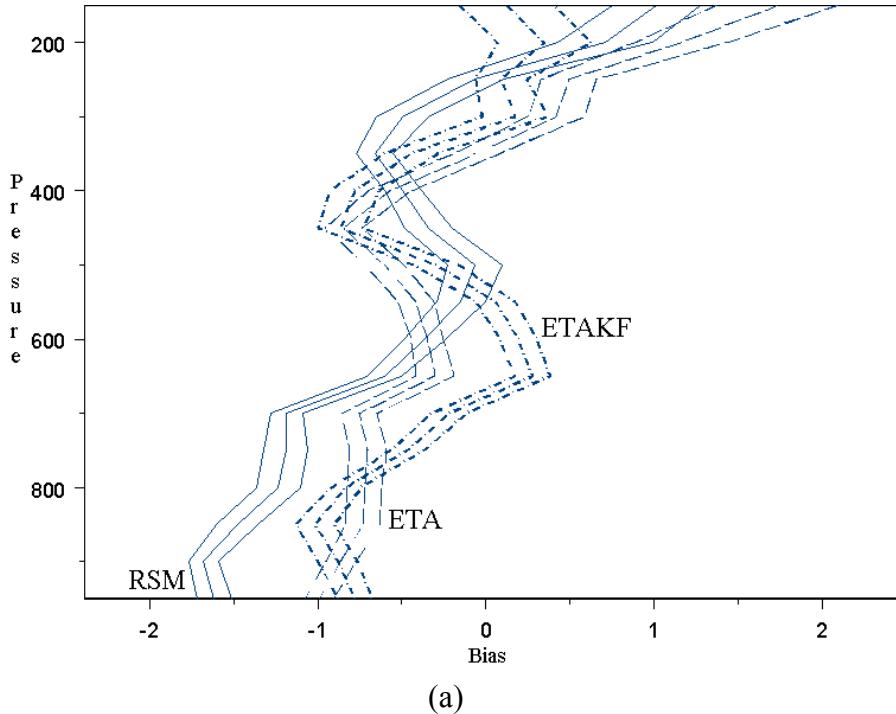
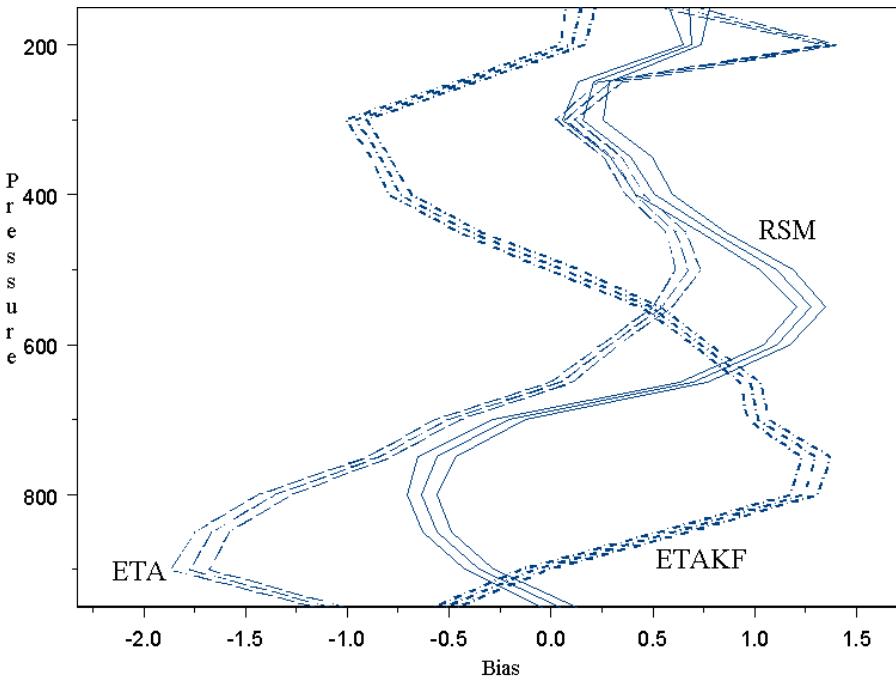


Figure 5. Individual site rank histograms constructed from the ensemble without any modifications for four sites for the u component at 800 hPa for a 15-hour forecast. (a). Individual site rank histogram for Miami, FL. (b). Individual site rank histogram for Spokane, WA. (c). Individual site rank histogram for Minneapolis, MN. (d). Individual site rank histogram for Tucson, AZ.

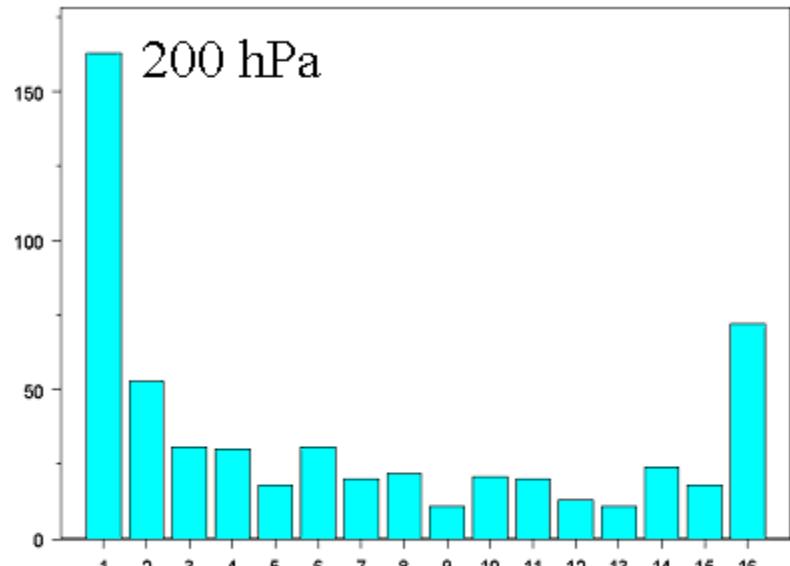


(a)

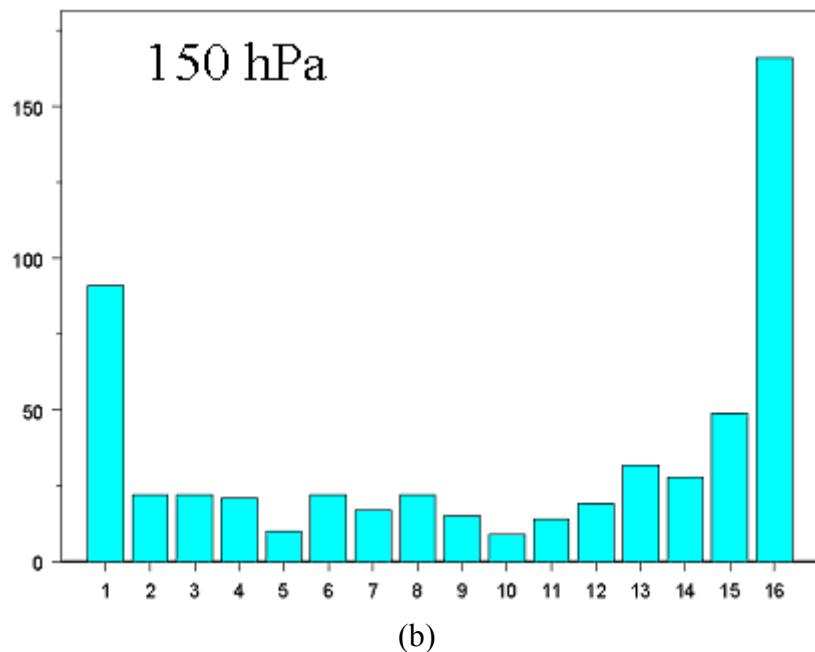


(b)

Figure 6. Changes in the bias of each model family as a function of height. The three lines for each model family correspond to the 95% confidence interval and mean bias. (a). An example of model families exhibiting similar biases is MFL for the u component for a 15-hour forecast. (b). An example of model families exhibiting different biases is MFL for temperature for a 63-hour forecast.

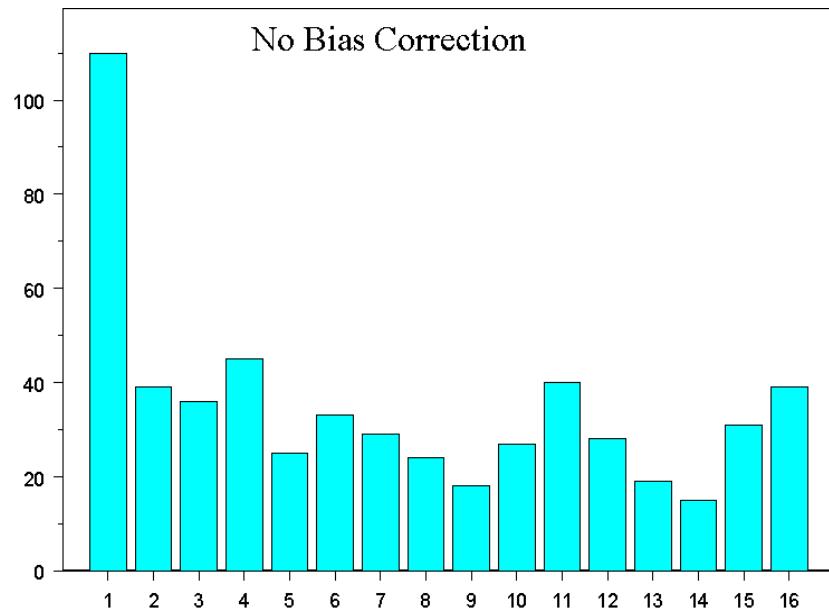


(a)

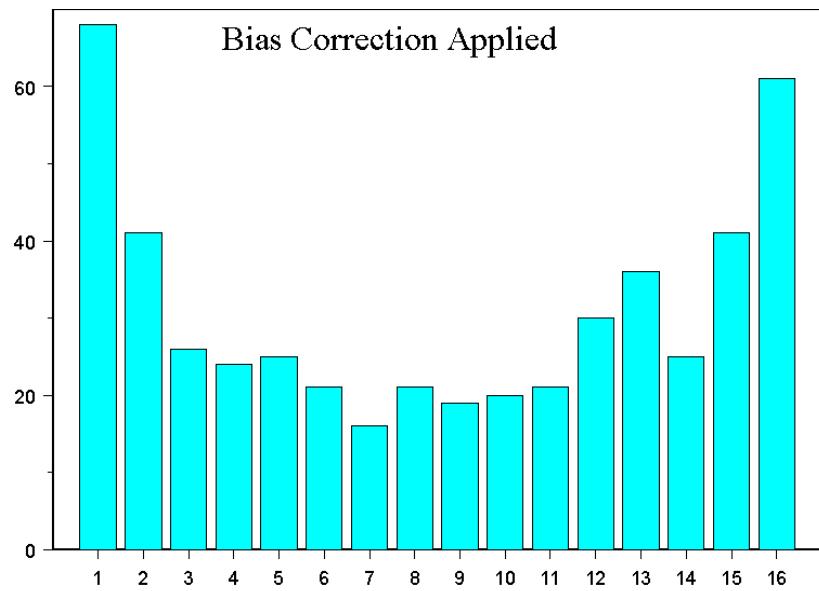


(b)

Figure 7. (a). Combined rank histogram for temperature at the 150-hPa pressure level for a 63-hour forecast. No modifications were made to the ensemble. (b). Combined rank histogram for temperature at the 150-hPa pressure level for a 63-hour forecast. No modifications were made to the ensemble.



(a)



(b)

Figure 8. Combined rank histograms for temperature at 700 hPa for a 63-hour forecast.

(a). Combined rank histogram constructed from the ensemble *without* the seven-day lagged bias correction.

(b). Combined rank histogram constructed from the ensemble with the seven-day lagged bias correction applied.

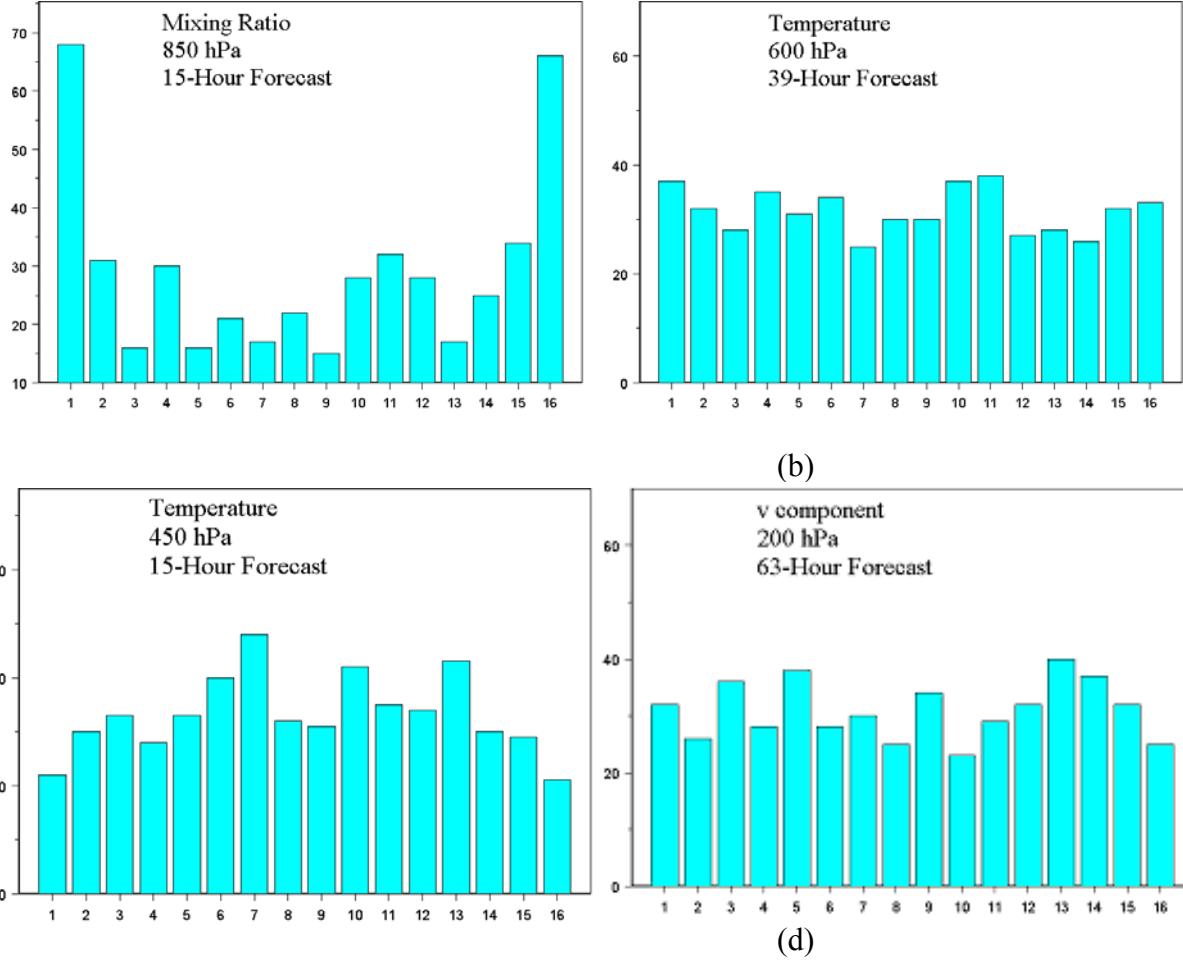


Figure 9. Combined rank histograms, all of which were constructed from the ensemble with both modifications. (a). Combined rank histogram for mixing ratio at 850 hPa for a 15-hour forecast. (b). Combined rank histogram for temperature at 600 hPa for a 39-hour forecast. (c). Combined rank histogram for temperature at 450 hPa for a 15-hour forecast. (d). Combined rank histogram for v component at 200 hPa for a 63-hour forecast.

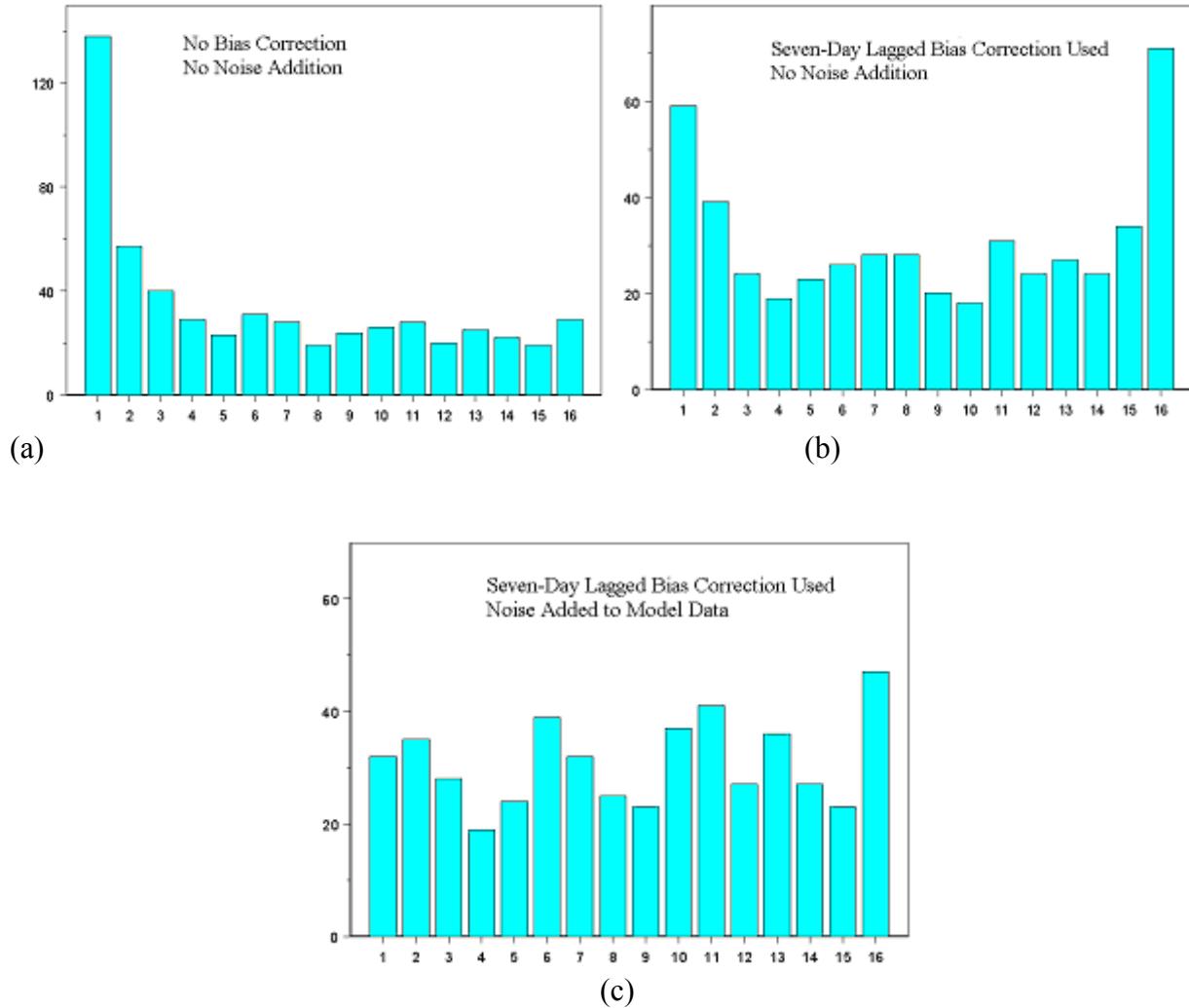


Figure 10. Combined rank histograms for temperature at 550 hPa for a 63-hour forecast.
 (a). Combined rank histogram constructed from the ensemble without the seven-day lagged bias correction or the addition of noise to the model data. (b). Combined rank histogram constructed from the ensemble with the seven-day lagged bias correction, but not the addition of noise to the model data. (c). Combined rank histogram constructed from the ensemble with both the seven-day lagged bias correction and the addition of noise to the model data.

Table 1. Results of the Chi-Squared Goodness of Fit Test that was completed on the combined rank histograms constructed from the ensemble with both modifications. The pressure levels are at the top of the table, while the names of the combined rank histograms are along the left side of the table. The null hypothesis for this significance test was that a uniform distribution and the rank histogram being tested are statistically equal; the alternative hypothesis is that there these two distributions are not equal. An X denotes where the null hypothesis was not rejected. A U denotes where the rank histogram was U-shaped. An O denotes where the rank histogram implies an over-dispersed ensemble.