7.6 A Validation of the NCEP SREF

Andrew J. Hamm[1] and Kimberly L. Elmore[2]

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.

The ensemble evaluated here is provided by the National Centers for Environmental Prediction (NCEP). This ensemble contains 15 individual members, which are equally divided into three model families: 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 resolution. The perturbations for these models are generated by the breeding method. 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). Ideally, an ensemble forecast can be used to generate 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.

Rank histograms are among the tools used to evaluate the performance of this ensemble (Hamill 2001). A rank histogram is constructed by summing the rank of the verification relative to the individual ensemble members 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.

Unfortunately, rank histograms are not necessarily subject to unique interpretations. 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 with high frequency counts at both extremes suggests one of several problems. The ensemble may be under dispersive (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), or a combination of these.

A consistent bias is indicated when a rank histogram has high frequency counts near one extreme and low frequency counts near the other. 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 (though do not guarantee) that the distribution of the ensemble forecasts matches the distribution of the verifying observations.

Past ensemble evaluations using rank histograms provide motivation for this investigation (Hamill and Colucci, 1997). Hamill and Colucci 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.

The rest of this paper is organized as follows: section 2 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 provides a conclusion, including some discussion into applications of this research.

2. Data and Methodology

This study utilizes verification soundings and model soundings. The verification data come from the Forecast Systems Laboratory's web site for rawinsonde data, which is currently http://raob.fsl.noaa.gov; the model data come from NCEP. Model data are specific to the location of interest. Nine rawinsonde sites, distributed about

[1] Oklahoma Weather Center Research Experience for Undergraduates, Norman, OK, and Northland College, Ashland WI.

[2] Cooperative Institute for Mesoscale Meteorological Studies, Univ. of Oklahoma/National Severe Storms Laboratory, Norman, OK.
Corresponding Author Address: 1313 Halley Circle, Norman, OK 73069, kim.elmore@noaa.gov

the continental US, are selected as locations for ensemble validation (Fig. 1). 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.

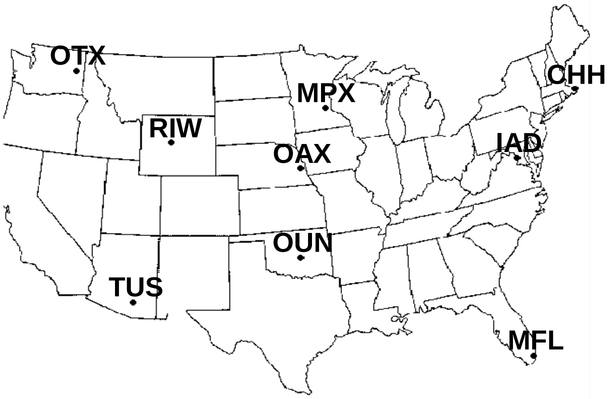

Figure 1. Map showing verification sounding site locations.

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).

3. Results

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 range describes the spread of the ensemble and may be used to describe how the ensemble behaves as a function of parameter, location, forecast time, and pressure level. A large range indicates a large spread in the ensemble and relatively more uncertainty, than a small range. Figure 2 shows an example of the mean range of the ensemble for temperature for all forecast times at CHH, MFL, and IAD. The range of the ensemble increases with forecast lead time regardless of variable and location. Range changes as height increases, regardless of variable, forecast time, and location.

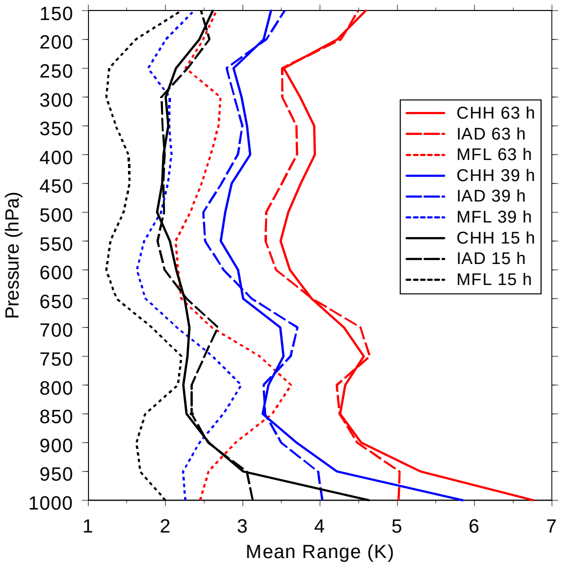

Figure 2. Mean range of ensemble temperature forecasts for CHH, IAD, and MFL for 15 h (black), 39 h (blue) and 63 h (red) forecasts.

Three sets of combined (all locations combined) 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 remove bias errors in the ensemble, while the addition of noise to the model data is used to address observational errors.

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. 3). 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 counts 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. The combined rank histograms for u and v components of the wind are U-shaped, which implies an under dispersive ensemble at all levels and forecast times. Individual site rank histograms are examined to find a possible explanation for the shapes of these combined rank histograms.

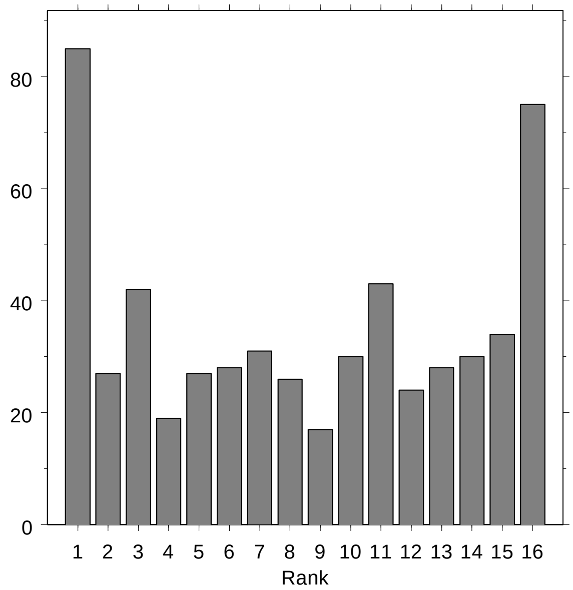

Figure 3. Combined rank histogram for 63 h mixing ratio forecasts at 600 hPa without bias or observational noise correction.

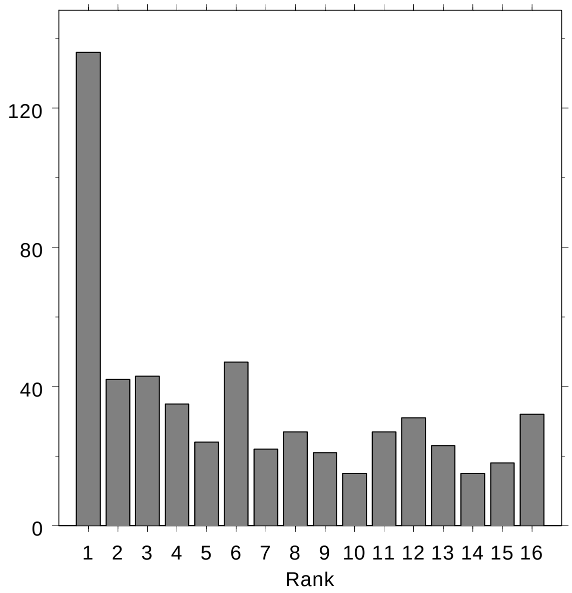

Figure 4. Same as Fig. 3, but for 39 h temperature forecasts at 750 hPa.

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, FL (MFL) for all forecast times and at pressure levels near the surface, a negative bias (most counts in rank 16) in the ensemble is implied (Fig. 5), while an under-dispersed ensemble is implied above 650 hPa. In Spokane, WA (OTX, Fig. 6), a positive bias (most counts in the 1^{st} rank) in the ensemble is implied 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. In Minneapolis, MN (MPX, Fig. 7), a positive bias in the ensemble is implied near the surface, except 900 hPa, and a negative bias in the ensemble is implied above 350 hPa. In Tucson, AZ (TUS, Fig. 8), a positive bias in the ensemble is evident at pressure levels near the surface, except for 900 hPa, while a negative bias in the ensemble is implied above 500 hPa. Individual site rank histograms for the v component typically show characteristics similar to the u component at all locations and pressures various. Biased combined rank histograms are not discussed further because bias in the ensemble is

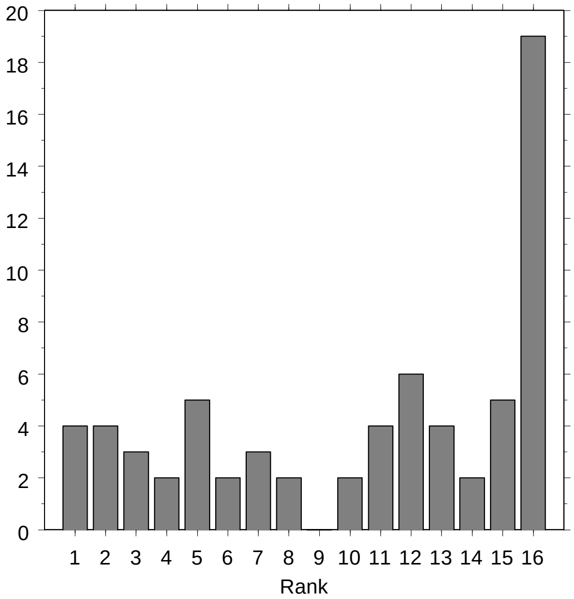

Figure 5. Individual rank histogram for MFL 15 h forecast of the u component

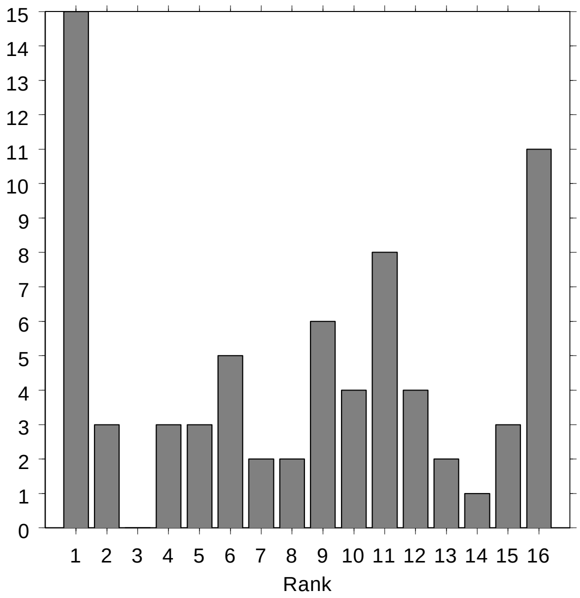

Figure 6. As in Fig. 5, but for OTX.

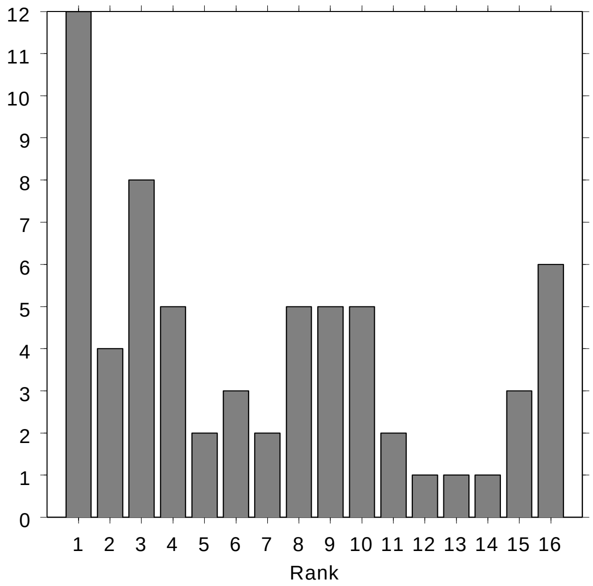

Figure 7. As in Fig. 5, but for MPX.

clearly responsible for the shape of the rank histogram. Possible explanations for the U-shaped rank histograms are given 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. Figure9 shows an example of the different bias characteristics for each model family. Differences in the biases among the three model families at different heights exist for different variables, locations, and forecast times. Different biases among the model families comprising the

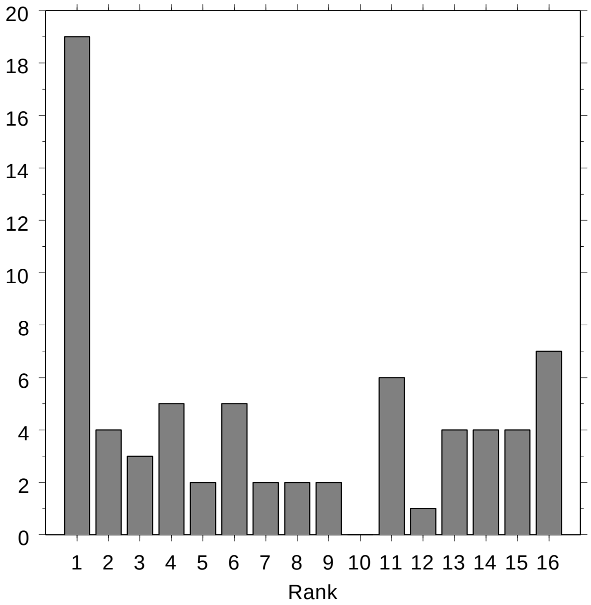

Figure 8. As in Fig. 5, but for TUS.

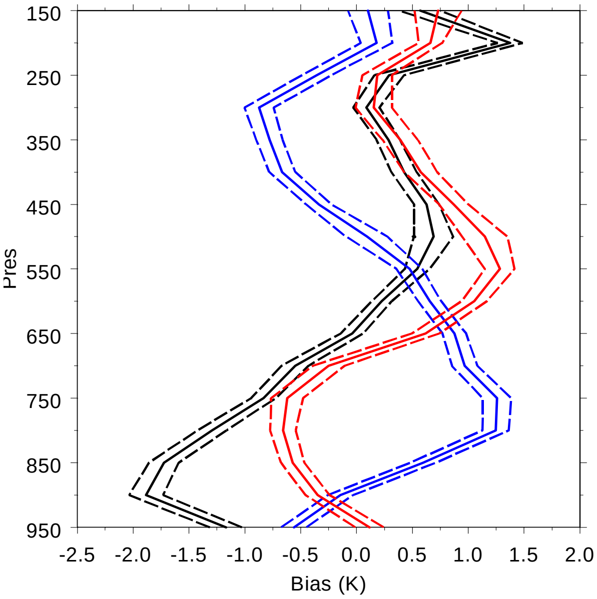

Figure 9. Mean temperature bias over the period for each model family: black is for the Eta, blue is for the EtaKF and red is for the RSM. Dashed lines show the 95% confidence interval based on a t-test.

ensemble complicate the interpretation of the rank histograms (Hamill, 2001).

Combined rank histograms for 150 hPa 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. 10). One

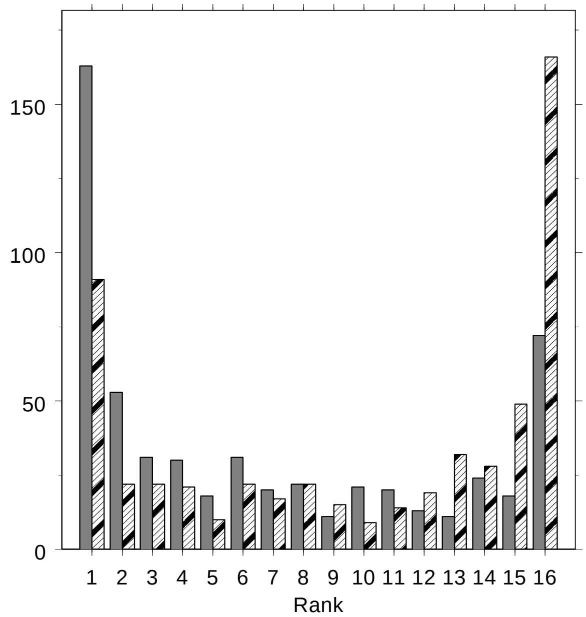

Figure 10. Combined rank histograms of temperature for 200 hPa (gray) and 150 hPa (diagonal hatching).

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 forecast value for the given model, current day, pressure level, variable, location, and forecast time.

Various lag intervals were tried and, 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. This simple bias correction proves a very effective way to remove bias from the ensemble. However, the rank histograms still show under dispersion. Hamill (2001) suggests that the apparent under dispersion in some cases may be illusory.

To properly interpret rank histogram shape, the error distribution of the verifying observations must be considered. 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 under dispersion as often as uncorrected rank histograms (Fig. 11). In some cases, the resulting rank histograms suggest an over dispersive ensemble. Curiously, nearly all of apparently over-dispersive rank histograms come from 15-hour forecasts.

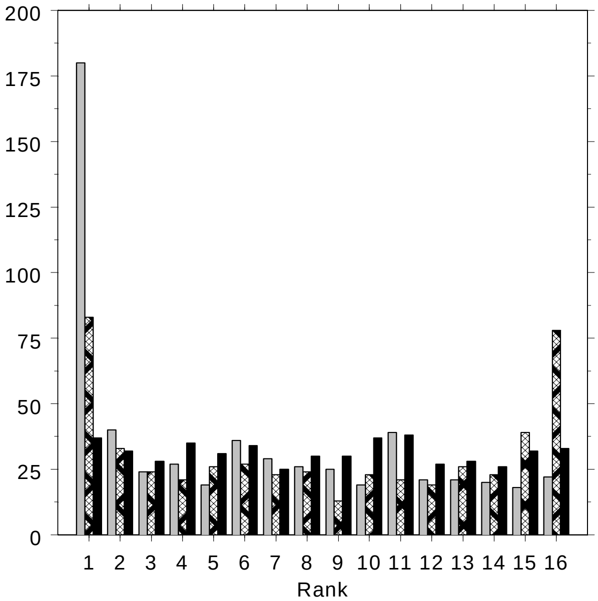

Figure 11. Rank histograms for 600 hPa temperature forecasts without bias or noise correction (gray), with bias correction only (hatched) and with both bias and noise corrections added (black).

Unfortunately, nearly all of the combined rank histograms for mixing ratio still show a significantly under-dispersed ensemble, which may result from difficulty in both accurately predicting and measuring moisture in the atmosphere.

Based on a chi-square goodness-of-fit test at p=0.05, many of the bias- and noise-combined rank histograms become nearly uniform (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. Excluding mixing ratio, 51% of the combined rank histograms with 15 h lead times did not reject the null hypothesis. For the 39 h lead time, 65% of the combined rank histograms failed reject the null hypothesis, and 53% of the combined rank histograms for the 63 h lead time did not reject the null hypothesis. Without the bias and noise corrections, the none of the rank histograms were consistent with a uniform distribution. 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.

4. Conclusions

The distributions of the forecasts and verifying observations are clearly distinct for the raw ensemble output. However, the models within this ensemble have unique biases. Many of the combined rank histograms for mixing ratio, u, v 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 result 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. 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 (Hamill, 2001) 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 opposing biases and a 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 under dispersion 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 better 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 included.These modifications clearly improve the validation statistics for the ensemble. Before modification, neglecting mixing ratio, none of the ensembles pass a chi-squared goodness-of-fit test for uniform distributions, but after modification, 51% of the 15-hour forecasts, 65% of the 39-hour forecasts, and 53% of the 63-hour forecasts passed a chi-squared goodness-of-fit test. Care must be exercised when adding observational noise. 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. The error characteristics of the verifying observation must be known. 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 performs poorly, but a simple bias correction added to each member 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 including unavoidable observation errors. Including observational errors is inappropriate operationally because a forecast is 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 should be used only in assessing the statistical performance of the ensemble.

5. Acknowledgements.

This work is supported by the National Science Foundation under Grant No. 0097651. 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, who directs the Program at the University of Oklahoma and NSSL.

6. References

Hamill, Thomas M. 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev,* **129,** 550-560.

———, and Colucci, Stephen J. 1997: Verification of Eta-RSM Short-Range Ensemble Forecasts. *Mon. Wea. Rev.,* **125,**. 1312-1327.

Murphy, Allan H. 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. and Forecasting,* **8,** 281-293.

Sivillo, Joel K., Ahlquist, Jon E., Toth, Zoltan. 1997: An Ensemble Forecasting Primer. *Wea. and Forecasting,* **12,** 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. *Wea. and Forecasting,* **15,** 603-622.