ENSEMBLE FORECAST BIAS CORRECTION

Angeline G. Pendergrass[1]
National Weather Center Research Experiences for Undergraduates
University of Miami
Coral Gables, FL

Kimberly L. Elmore[2]
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma /
National Severe Storms Laboratory
Norman, OK

[1] Corresponding author address: 1476 Woodscreek Circle, Crystal Lake, IL 60014, a.pendergrass@umiami.edu
[2] Corresponding author address: 1313 Halley Circle, Norman, OK 73069, kim.elmore@noaa.gov

Abstract

This study investigates two bias correction methods, lagged average and lagged linear regression, for individual members of ensemble forecasts. Both methods use the forecast bias from previous forecasts to predict the bias of the current forecast at every station. Also considered is the training period length that results in the smallest forecast error.

Ensemble forecast and verification data span 23 July through 15 September 2003. The data are organized into a mini-ensemble composed of 5 models and 30 days. This mini-ensemble is corrected using each method of correction for training period lengths between 3 and 12 days. The resulting bias, mean absolute error, RMS error, and inter-quartile range of the corrected forecasts are then compared.

Forecasts corrected with the lagged linear regression method are less biased but have more variance than those corrected with the lagged average method.

1. Introduction

Ensemble forecasts are currently used to predict atmospheric variables such as temperature, dewpoint, and wind speed. They are useful because they can provide more information than deterministic forecasts. An ensemble forecast is made up of several members, each of which are individual solutions (Manousos 2004). Each ensemble member is either generated by a different numerical model of the atmosphere, has its own initial conditions, or has different governing physics than other members from the same numerical model (Hamm 2003). The resulting ensemble is generally more accurate if each ensemble member is bias corrected. Hamm (2003) determined that "bias correction is an important post-processing step for the NCEP SREF," but bias correction methods have not been themselves examined in previous studies. The purpose of this study is to investigate two methods of bias correction, the lagged average (Stensrud and Yussouf 2003) and a lagged linear regression correction.

The training period is defined as the number of days used for the bias correction. A training period that is too short may not provide enough information about the bias characteristics of each model, and a training period that is too long may not account for short-term variations or trends. The optimal training period may depend on the numerical model, geographic location, season, and parameter.

The rest of this paper is organized as follows: ensemble forecast data is described in section 2; the methodology through which the bias correction methods and training period lengths were calculated and compared is described in section 3; results of the study are described in section 4; section 5 contains a discussion of implications, drawbacks, limitations and interesting results; and section 6 contains conclusions.

2. Data

For this study, a combination of 5 members was selected from an overall data set of 31 members. The data cover 55 days, from 23 July to 15 September 2003.

The entire set of 31 ensemble members comes from three sources: National Center for Environmental Prediction (NCEP), National Severe Storms Lab (NSSL), and Forecast Systems Laboratory (FSL). The ensemble members generated by NSSL and FSL are experimental, and so frequently unavailable. Twenty-one of the ensemble members are based on variants of the Eta model (Mesinger; Janjic et al. 1988), seven on the Regional Spectral Model (RSM), two on the Rapid Update Cycle (RUC) model, and two on the Weather Research and Forecast (WRF) model.

The initial 31 ensemble members are not all used because many of the experimental members are frequently missing (Fig. 1). To obtain a contiguous set of forecasts, members that miss runs for a large proportion of days are excluded; then, days that have a very small proportion of runs are excluded (Fig. 2).

The winnowing process continues until there is a contiguous block of member runs. The most reliable members come from the NCEP Short-Range Ensemble Forecast (SREF) because this is an ensemble that is almost operational.

Forecast data are provided on the AWIPS 212 grid, which has an 80 km grid spacing. But, forecasts are verified at specific locations (observation sites). Hence, bi-linear interpolation is used to interpolate forecast value to specific, verifying locations.

Initial conditions for all the members used here use initial perturbations generated from the breeding method (Toth and Kalnay 1993). The breeding method is used to

generate perturbations that will grow most rapidly. In most cases, a single model can yield five members: a control run, two positively perturbed members and two negatively perturbed members.

Twenty-two members come from the NCEP SREF. All SREF members are initialized at 0600 UTC each day and use 32-km grid spacing. Seven members are from the RSM, and use two different physical parameterizations between them (only one of which is used). The RSM-SAS (Simple Arakawa-Shubert convection (Arakawa and Schubert 1974)) uses all five initial condition perturbations (RSM1, RSM2, RSM3, RSM4, and RSM5). The first 5 RSM members follow the naming convention used throughout this data. RSM1 uses the control initial conditions, RSM2 and RSM 3 use the negatively perturbed initial conditions, and RSM4 and RSM5 use positively perturbed initial conditions.

Eta-KF (Kain-Fritsch convective scheme (Kain and Fritsch 1998)) and Eta-BMJ (Betts-Miller-Janic convective scheme) each use all 5 perturbations. These are, respectively, EKF 1-5 and EBM 1-5.

The last two members used are from the Eta model with varying initial conditions and physical parameterizations. Eta-RAS-Mic (ETA4) uses an experimental Ferrier Microphysics with Relaxed-Arakawa Shubert convective parameterization with positively perturbed pair two initial conditions. Eta-KF-CON (ETA5) uses an experimental Ferrier microphysics with more frequent calls to cloud water condensation and ice deposition and Kain-Fritsch convective parameterization, and it uses positively perturbed initial conditions.

3. Methodology

a. Preparing the data

The first task in the project is ingesting the data into S-Plus (Insightful Corp. 2004). The data are organized according to the source model, station identifier, date, and variable. The evaluated variables are 2-m temperature, 2-m dewpoint, and u and v wind components. The verification data are organized similarly.

While individual model runs start at varying times, the goal is for all of them to constitute a single ensemble. To accomplish this, data are categorized by verification time. Because some ensemble members start at 1200 UTC, this is used as the time chosen for the start of the entire ensemble forecast each day. The earliest forecast from the latest member is for 6 h, thus the first forecast from the ensemble each day verifies at 1800 UTC that same day. Output is available every 3 h after the initial 6 h forecast, through 48 h, or 1200 UTC on day 2, though only the first 24 h are evaluated here. However, each ensemble forecast is made up of member forecasts for various lengths of time. For example, a 12 h ensemble forecast is made up of 5 members with a 12 h forecast starting at 1200 UTC, 13 members with an 18 h forecast starting from 0600 UTC, and 4 members with a 24 h forecast starting from 0000 UTC the previous day. This introduces additional diversification into the ensemble by having members with initial conditions when the atmosphere was in different states.

Bias correction requires that each forecast be accompanied by a verification value or observation. Thus, missing data are problematic. The most common cause of missing data in this study is missing model runs. To facilitate processing, members and days are eliminated until there is a contiguous data set with no missing members. A result of this

method of data reduction is that the available forecasts are not continuous in time (Fig. 2). Yet, even these steps do not ensure contiguous data, because observations are frequently missed, which results in missing verification values.

b. Analysis

Two bias correction methods are evaluated in this study: lagged average and lagged linear regression. Both bias correction methods use the bias from forecasts on previous days to calibrate each member individually, at each location. To compare these two approaches, mean bias, mean absolute error (MAE), and root-mean-square error (RMSE) are used. Bias is the difference between the corrected forecast and its verification, MAE uses the mean absolute value of the bias, and RMSE is the mean of square root of the sum of the squared bias. Forecasts from New England are used to evaluate the two methods so they can 1) be easily compared with Stensrud and Yussouf (2003), and 2) to reduce the computational burden.

The lagged average bias correction method has been used in the past (e.g. Stensrud and Yussouf 2003), and is a simple bias correction where prior biases over some training period are averaged, and this average is applied as a correction to the current forecast (Fig. 4a).

In contrast, the lagged linear regression method uses is a least-squares line to model the trend in the bias of the forecasts over the training period at each location. Both methods are tested for training periods between 3 and 12 days (Fig. 4b).

For lagged linear regression, at least two data points are necessary or a line cannot be calculated. Yet, with only two data points, least squares has no meaning. As a result,

the training period must be at least 3 days (preferably longer). Missing data are ignored such that a missing day reduces the effective length of the training period. In no case does the training period contain less than three days. For a given training period, locations are excluded for which the number of days is less than there are for a given training period. The number of locations with sufficient data for the lagged linear regression method increases with longer training periods (Fig. 3).

4. Results

To compare the bias correction methods, mean biases, MAEs, and RMSEs from all locations are plotted as a function of training period length and also of forecast time for each variable. Each plot shows the mean and inter-quartile range (IQR) of the error statistic for each method. The bias distributions from each correction method are also compared. Analysis of the bias distributions for all forecasts at a specific training period length for one variable at time 12 (verifying at 0000 UTC) shows differences between bias distributions of forecasts corrected with lagged average and lagged linear regression (Fig 5). Biases from the lagged average method are more narrowly distributed than those from the lagged linear regression method for smaller training period lengths. But, for longer training period lengths, these bias distributions look similar. This is the case for all variables.

Plots of mean and IQR of these distributions allow comparison of distribution characteristics (Fig. 6). For smaller training periods (about 3 to 7 days), forecasts corrected with lagged linear regression have mean biases smaller than those corrected

with lagged average, but larger IQR. In many cases, this IQR is larger than that for uncorrected forecasts. Forecasts corrected with the lagged average method usually have a smaller mean bias and smaller IQR than the uncorrected forecast bias distributions. With longer training periods, the IQR of for the lagged linear regression decreases, while increasing for the lagged average method. Both methods yield similar IQR for the longest (10 to 12 days).

For most variables, MAE is lower for forecasts corrected with the lagged average method than those with the lagged linear regression method (not shown). Lagged average MAE increases with training period length, while lagged linear regression MAE decreases with training period length. RMS errors show similar trends (not shown).

While these bias error distributions for long training periods look similar for both methods, a Komolgorov-Smirnov goodness-of-fit test is used to investigate further. Bias distributions from the 12 h forecast, valid at 0000 UTC on day 1, for all variables are investigated. Between the two methods, al distributions are distinct at $p = 0.05$. Within each method, only the v-wind biases corrected with the lagged average method for training period lengths of 5 and 6, and also for training periods 10 through 12, are indistinguishable at $p = 0.05$.

Plots of error statistics as a function of forecast time (Fig. 7) show forecast bias and IQR increase with forecast time for both methods, as do the uncorrected forecast biases. This indicates that overall skill decreases for longer forecasts. The only variable that does not display this general increase over the first 24 h is v wind component.

Overall biases for the v component behave differently form the others over the experimental period. Unlike the other variables, the resulting bias of v component

forecasts using the lagged average method shows a persistent negative bias, and is often worse than no correction at all. In contrast, the lagged linear regression method does a good job at removing systematic bias (Fig. 8a). A time series of uncorrected biases reveals that over the experimental period (23 July to 9 September, in this case), uncorrected forecast biases drift from positive to negative (Fig. 8b). This trend is corrected by the lagged linear regression method, but the lagged average method frequently predicts that the bias will be more positive than observed and so introduces a systematic error in the ensemble. Thus, in this case the lagged linear regression is clearly more desirable than the lagged average.

5. Discussion and Conclusions

This study is small in scale in that it only studies an ensemble of 5 members over 73 stations in New England, and for less than two months of the warm season. An investigation on a larger scale could produce more general results. In general, forecasts corrected with the lagged linear regression method are less biased but have more variability than those corrected with the lagged average method. Differences in bias hold for all training period lengths, but differences in variability decrease as training period increases.

Effects of variables other than training period on the bias correction methods should be studied as well. Some variables that may affect the bias corrections are geographic location and season. Also, training periods longer than 12 days may be useful.

To know which of these methods yields the best ensemble performance requires a careful application of rank histograms (Hamill 2001). Time constraints prevented this next, obvious step.

Though this investigation does not yield conclusive results about which of the two methods is better, there is enough evidence to allow reasonable speculation. Hamm (2003) found that bias corrections can reduce the under-dispersiveness of ensembles. The lagged linear regression has potential to do this, because it often increases the variability (IQR) of members. The main purpose of bias correction is to remove bias from the ensemble as a whole. Since lagged linear regression does a good job of removing bias from members on average, it will probably help remove it from ensembles as well.

The effect of lagged average and lagged linear regression bias corrections on ensemble performance is not evaluated here. As stated earlier, evaluating ensemble performance requires, as a start, analysis of rank histograms (Hamill 2001).

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. 0097651. I would like to thank everyone at the National Severe Storms Laboratory and the University of Oklahoma for hosting the Research Experience for Undergraduates (REU) program that allowed me to participate in this research, especially Daphne Zaras, who directs the program. I am grateful to Lance Maxwell for driving all of the REU students around and planning everything for us.

References

Arakawa, A. and W. H. Schubert, 1974: Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *J. Atmos. Sci.*, **31,** 674-701.

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

Hamm, A. J. a. K. L. E., 2003: A Validation of the NCEP SREF. *Preprints, 20th AMS Conference on Weather Analysis and Forecasting and 16th AMS Conference on Numerical Weather Prediction*, Seattle, WA, American Meteorological Society, CD-ROM, WAF/NWP 7.6.

Insightful, C., 2004.

Kain, J. S. and J. M. Fritsch, 1998: Multiscale Convective Overturning in Mesoscale Convective Systems: Reconciling Observations, Simulations, and Theory. *Mon. Wea. Rev.*, **126,** 2254-2273.

Manousos, P., cited 2004: Ensemble Prediction Systems: A training manual targeted for meteorologists wanting to know more about the ensemble technique. [Available online from http://www.hpc.ncep.noaa.gov/ensembletraining/.]

Mesinger, F., Z. I. Janjic, S. Nikovi, D. Gavrilov, and D. G. Deaven, 1988: The Step-Mountain Coordinate: Model Description and Performance for Cases of Alpine Lee Cyclogenesis and for a Case of an Appalachian Redevelopment. *Mon. Wea. Rev.*, **116,** 1493-1520.

Stensrud, D. J. and N. Yussouf, 2003: Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over New England. *Mon. Wea. Rev.*, **131,** 2510-2524.

Toth, Z. and E. Kalnay, 1993: Ensemble Forecasting at NMC: The Generation of Perturbations. *Bull. Amer. Meteor. Soc.*, **74,** 2317-2330.

Captions

Fig. 1. Model runs available at all times. Gaps in data necessitated winnowing of data set.

Fig. 2. Days of data used in study (2003).

Fig. 3. Locations with enough data to use the lagged linear regression correction method. Bottom axis is training period length.

Fig. 4a. Schematic of lagged average bias correction method for training period length of 10. Triangle is predicted bias for the current forecast, and the actual bias is just above it.

Fig. 4b. Schematic of lagged linear regression bias correction method for training period of 10 days. Plus-sign is predicted bias for the current forecast from the lagged linear regression correction method. Actual bias is just below it.

Fig. 5. Left column is lagged average method; right column is lagged linear regression. Top histograms are training period length 3, bottom are 12. Vertical line is 0. Plots for forecast time 12 (verifies 0000 UTC). 5a. is corrected temperature forecast bias, b. is dewpoint, c. is u-wind, and d. is v-wind.

Fig. 6. Forecast bias as a function of training period length.

Fig. 7. Forecast bias as a function of time.

Fig. 8a. v-wind forecast bias as a function of time. Notice that the lagged average method produces forecasts with a consistently negative bias.

Fig. 8b. v-wind EKF2 forecast biases at SFM (Sanford, Maine). Notice that they decrease on average throughout the time the forecasts were studied.

Figures

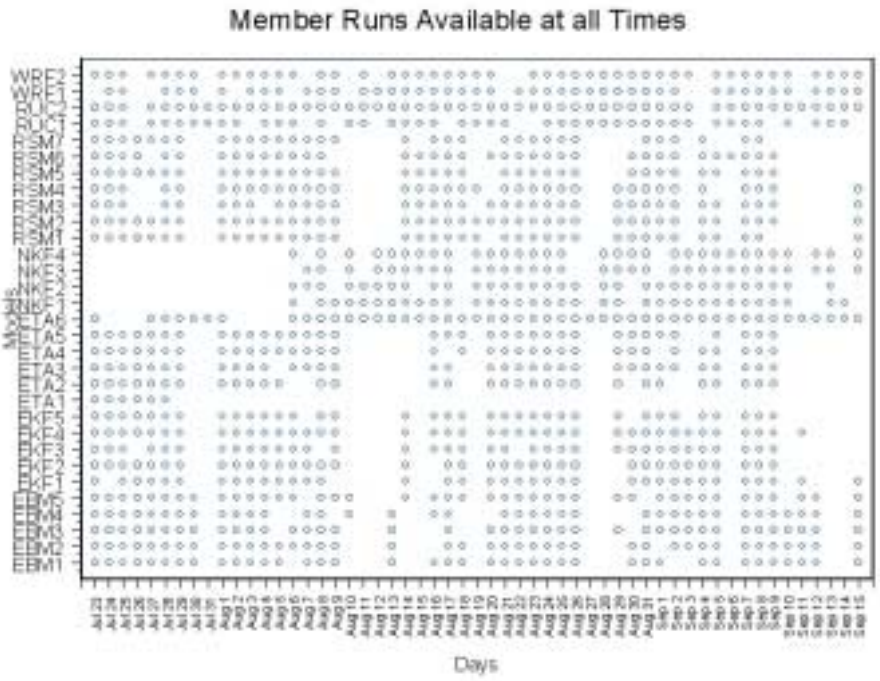

Figure 1

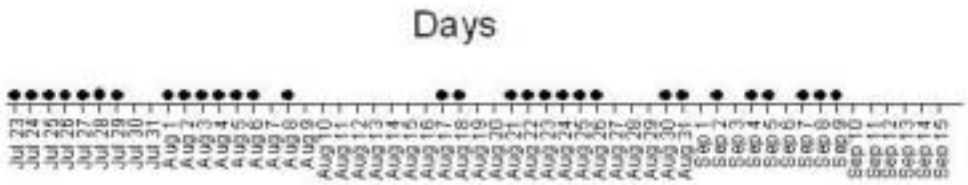

Figure 2

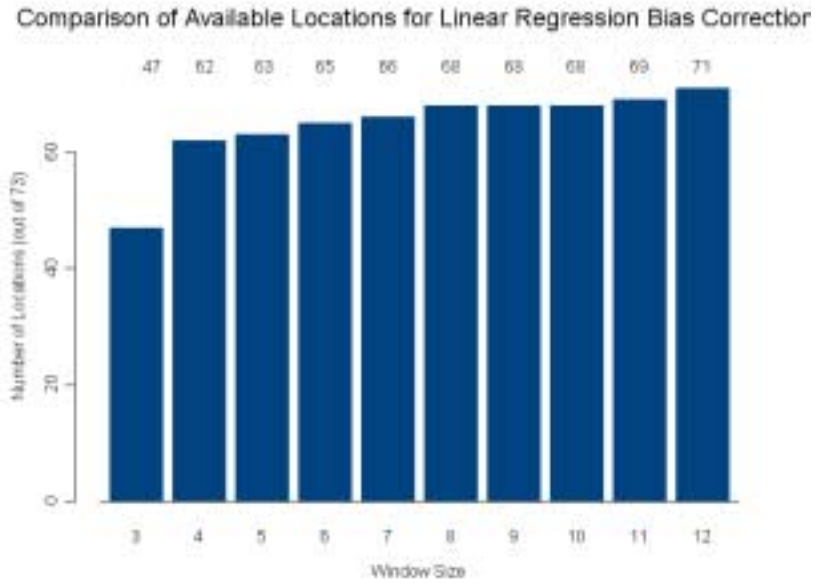

Figure 3

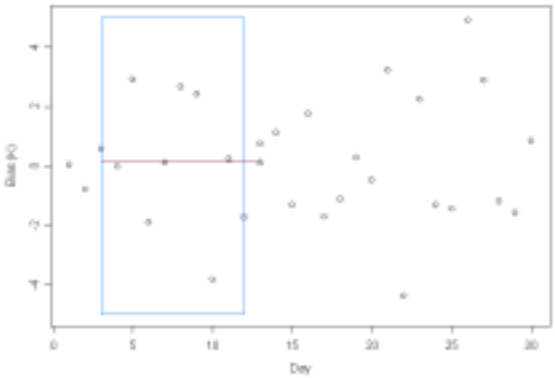

Figure 4a

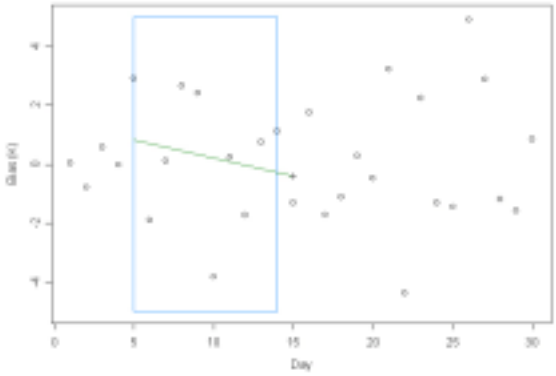

Figure 4b

Temperature

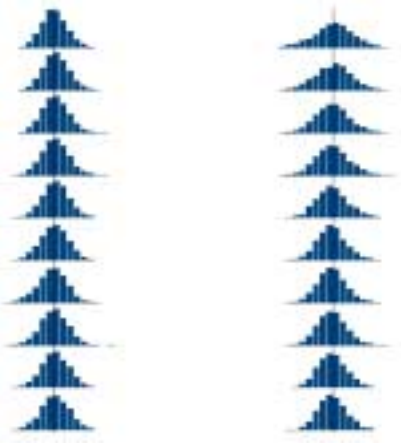

Figure 5a

Dewpoint

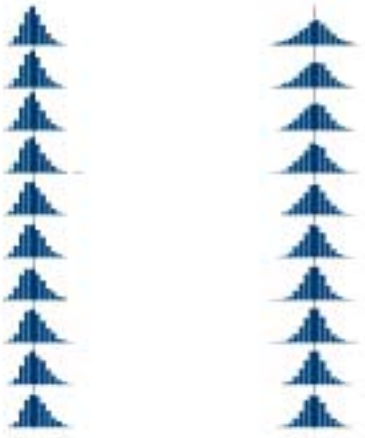

Figure 5b

u-wind

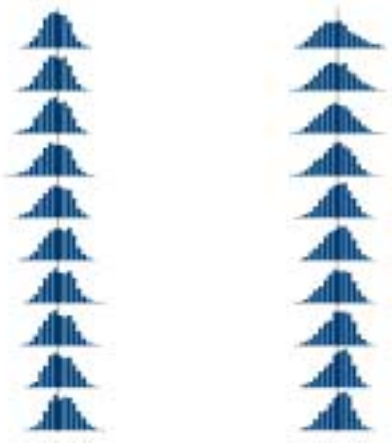

Figure 5c

v-wind

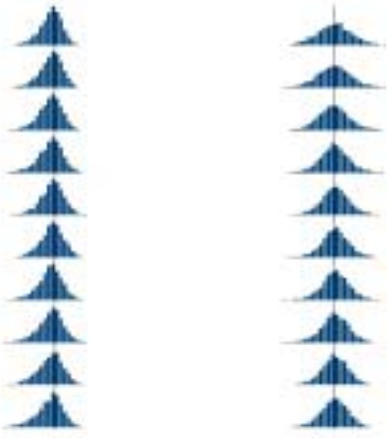

Figure 5d

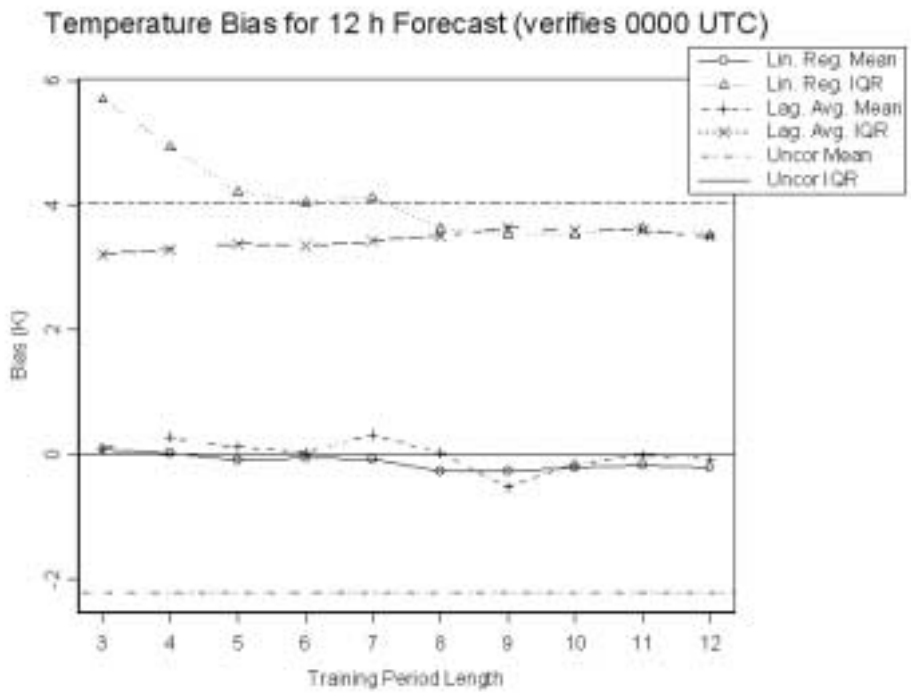

Figure 6

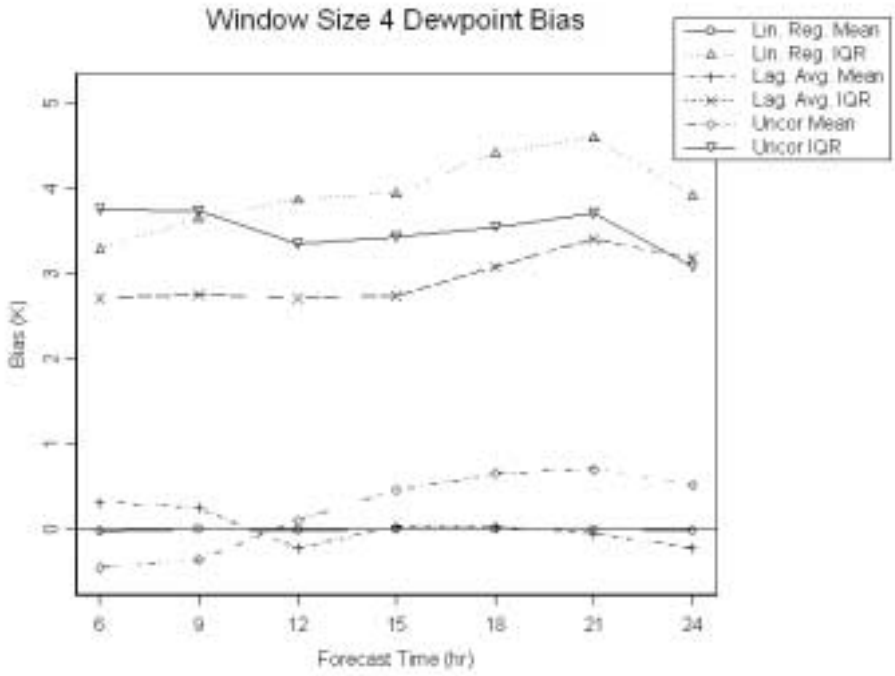

Figure 7

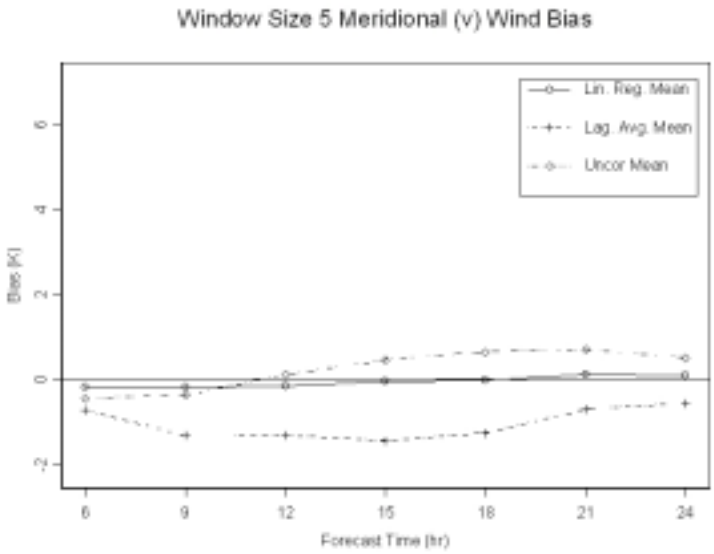

Figure 8a

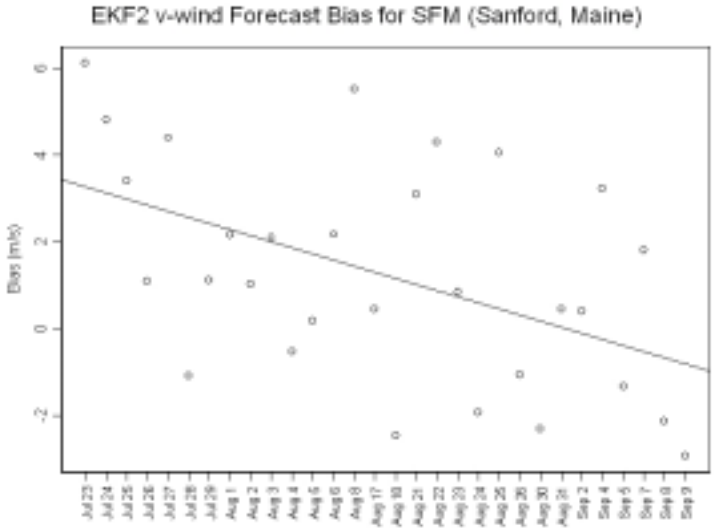

Figure 8b