3.4 DETERMINING THE MINIMUM NUMBER OF ENSEMBLE MEMERS NEEDED FOR STATISTICAL POSTPROCESSING

John L. Wagner Jr*
NOAA/NWS
Meteorological Development Laboratory, Silver Spring, MD

1. INTRODUCTION

Ensemble forecast systems (EFS) are a popular method that provide forecasters with an estimate of the uncertainty in the future state of the atmosphere. Statistical postprocessing techniques have been used to calibrate the reliability of an ensemble forecast from global models for a number of weather elements (Raftery et al 2005; Krzysztofowicz and Evans 2008; Unger et al 2009). As EFS increase the number of members and more EFS become available, more resources are required for postprocessing.

The Meteorological Development Laboratory (MDL) of NOAA's National Weather Service has developed a statistical postprocessing technique called Ensemble Kernel Density MOS (EKDMOS). EKDMOS produces reliable and accurate probabilistic guidance for 2-m temperature, 2-m dewpoint, daytime maximum temperature, and nighttime minimum temperature. The operational version of EKDMOS uses 42 North American Ensemble Forecast System (NAEFS; Candille 2009) members.

This extended abstract documents a series of sensitivity studies to help determine the minimum number of ensemble members needed to calibrate the ensemble mean and spread, while conserving limited computational resources. Statistical postprocessing techniques also require a representative reforecast sample for calibration when a model update occurs. Ensemble members from NCEP's Global Ensemble Forecast System (GEFS) were selected for this work since an expected model upgrade in spring 2015 will require a redevelopment of EKDMOS equations. Recent discussions within NOAA have suggested that a 5-member reforecast sample would provide adequate information to support the operational needs of existing statistical postprocessing techniques. Results from this study will validate this recommendation.

*Corresponding author address: John L Wagner
1325 East-West Hwy, Silver Spring MD 20910
E-mail: John.L.Wagner@noaa.gov

2. DATA

2.1 *GEFS Model Data*

Three years of GEFS data from April 1, 2011 through March 31, 2014 were used for this study and were divided into three warm seasons (April through September) and three cool seasons (October through March) for development and verification. K-fold cross validation was performed by rotating through our dataset, using two years of data for development while withholding the third year for independent verification. A set of 334 stations were chosen for development. These stations are located throughout the conterminous United States (CONUS), Alaska, Hawaii, and U.S. territories and are known by MDL to be reliable reporting stations. EKDMOS equations were developed for 2-m temperature, 2-m dewpoint, daytime maximum temperature, and nighttime minimum temperature. Predictors useful for predicting the temperature suite of elements were offered for regression. The predictors chosen most frequently were model 2-m temperature and model 2-m dewpoint. Three-hourly equations were generated for 2-m temperature and dewpoint out to 192-hours, and then 6-hourly equations were generated through 264-hours. Daily daytime maximum temperature and nighttime minimum temperature equations were generated through day 16.

2.2 *Observation Datasets*

An observation dataset was created from an archive maintained by MDL for the temperature suite of elements. A set of "extreme" temperature observations was also created for verification. This was done by taking 2-m temperature observations from 1980 through 2009 for our 334 stations, ranking the observations in a rolling 5-day sample for each day of the year (-2 days, +2 days), and computing the climatological 5% and 95% values. The observations from our April 2011-March 2014 sample were then compared to these climatological values and were only kept if the value fell within one of the 5% tails.

3. METHOD

3.1 EKDMOS Technique

The EKDMOS technique uses ensemble members to produce probabilistic forecasts for the temperature suite of weather elements. EKDMOS starts by using multiple linear regression with the ensemble mean values used for development (Glahn 2009). A second regression step is performed using the spread of the bias corrected ensemble members in our development sample to create a spread-skill relationship specific to each station (Veenhuis and Wagner 2012; Veenhuis 2013). The regression equation is applied to each ensemble member in our independent dataset and kernel density estimation is used to create a single probabilistic density function (PDF). The spread-skill equation is used to adjust the spread of the PDF to match the expected error. The PDF is then converted to a cumulative distribution function (CDF). An EKDMOS forecast consists of 11 points from the CDF along with the ensemble mean.

3.2 Test Cases

In order to test EKDMOS with fewer than 21 members, ensemble means were created using different numbers of ensemble members. While gathering data for these tests, two assumptions were made. First, it was assumed that any dataset, whether it be from a reforecast or from MDL's archive, would always contain the GEFS control member. It is also assumed that the lowest numbered ensemble members would be available. Ensemble mean values were generated using 11 members (control, members 1-10), 7 members (control, members 1-6), 5 members (control, members 1-4), and 3 members (control, members 1-2). These means were used to develop MOS equations that were then applied to all 21 GEFS members in our independent dataset. Kernel density estimation was used to combine each set of members into a PDF. Spread-skill equations were also developed from our development sample for each ensemble set. These equations were used to correct the spread of the corresponding PDFs prior to converting them to CDFs.

The EKDMOS technique was also applied using all 21 GEFS members for development. This represents the standard EKDMOS approach and served as a baseline for comparison. EKDMOS equations were also created using only the control member. Since we are not able to develop a spread-skill equation from only one member, the EKDMOS spread adjustment technique (Glahn et al, 2009) was used to correct the spread of the control member PDFs.

4. VERIFICATION

Verification scores were produced for the first moment and for the second moment of each set of GEFS CDFs. Cool season 2-m temperature results will be shown here. Scores were also calculated for 2-m dewpoint, daytime maximum temperature and nighttime minimum temperature for warm and cool seasons, as well as for the warm season 2-m temperatures. Results were found to be similar to the ones discussed here.

4.1 First Moment Verification

The mean absolute error (MAE) and bias scores were calculated to compare the skill of forecasting the first moment. The ensemble mean was used to verify the first moment since it is the deterministic part of the EKDMOS forecast. The skill of the ensemble median will be shown using the continuous ranked probability score (CRPS), which has the MAE of the median as a component (Hersbach 2000).

MAE results for cool season 2-m temperatures are shown in figure 1. Since over-plotting often makes it difficult to determine any differences among the tests, the percent improvement from the 21-member baseline test are also shown in figure 1. Figure 2 shows the forecast bias for the same tests. All plots show very little dependence on development ensemble size.

4.2 Second Moment Verification

The second moment of the forecast PDF was verified using the CRPS and probability integral transform (PIT) histograms. While PIT histograms do visualize the reliability of a forecast system, it is often difficult to compare the PIT histogram of one system with another. Consequently, the squared bias in relative frequency (SQBIAS; Glahn et al 2009) was calculated.

Figure 3 shows the CRPS scores for cool season 2-m temperature. Again, the percent improvement score is shown to help detect any differences from the baseline test not seen due to over-plotting. Much like the MAE results, the CRPS results show that estimates of the second moment depend little on the development ensemble size.

Figure 4 shows the PIT histograms for all tests for the 48-hour cool season temperatures and figure 5 shows results for the 192-hour temperatures. All histograms are relatively flat, except for excess at the extremes, showing overall good reliability for each test. Results for all other time steps were produced and were found to show similar reliability.

Figure 6 shows the SQBIAS scores for each test. Previous tests have shown that SQBIAS values less than 0.04 can be considered insignificant. Values from all tests are shown to be within this range, confirming the good reliability shown in the PIT histograms.

4.3 Extreme Temperature Verification

Verification scores were also computed using our "extreme" temperature observation dataset. This dataset represents some of the toughest forecasts from our 3-year sample and arguably some of the most useful guidance for our users. MAE scores are shown in figure 7. Unlike the scores shown in figure 1, we can now determine the difference between the test cases, particularly after 84-hours. Results show that developing with larger ensembles leads to slightly better MAE scores in the extended ranges. The percent improvement plot shows that only the control member test deviates from the baseline test by more than 5 percent. Figure 8 shows the bias scores for the "extreme" observation dataset. With the exception of the control member test, we again see very little difference in scores between test cases.

The CRPS scores in figure 9 are very similar until about 120-hour, where once again we see the benefit of having more ensemble members in our development sample. The percent improvement plot shows that the 5-member test does not degrade from our baseline test by more than about 5 percent, while the control member test degrades by more than 10 percent in the extended ranges. Figure 10 shows the SQBIAS score. Note that the scores are no longer within the noise range. Again we see that developing with more members leads to improved reliability in the extended ranges.

5. SUMMARY AND CONCLUSION

Statistical postprocessing techniques like EKDMOS typically use all available ensemble members when developing probabilistic guidance. Bulk statistics show than when fewer ensemble members are used for development, there is no significant difference in the first or second moment verification scores. It is not until we look at "extreme" temperature verification that we see any differences in verification scores, which become noticeable mainly after day 4. Recent discussions within NOAA on the GEFS reforecast configuration have suggested rerunning with 5 ensemble members to coincide with the next model upgrade. This number appears to be adequate for EKDMOS and similar postprocessing techniques. This study focused on the temperature suite of weather elements. Further study is required to see if this recommendation is suitable for other elements.

6. ACKNOWLEDGEMENTS

The author would like to thank Greg Zylstra and Bruce Veenhuis of the EKDMOS team and Matthew Peroutka for their help in this work.

7. REFERENCES

Candille, Guillem, 2009: The multiensemble approach: the NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665.

Glahn, Bob, Matthew Peroutka, Jerry Wiedenfeld, John Wagner, Greg Zylstra, Bryan Schuknecht, and Bryan Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268.

Hersbach, Hans, 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Krzysztofowicz, Roman and W. Britt Evans, 2008: Probabilistic forecasts from the national digital forecast database. *Wea. Forecasting*, **23**, 270–289.

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174.

Unger, David A., Huug van den Dool, Edward O'Lenic, and Dan Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379.

Veenhuis, Bruce A., 2013: Spread calibration of ensemble MOS forecasts. *Mon. Wea. Rev.*, **141**, 2467–2482.

--, and J. L. Wagner, 2012: Second moment calibration and comparative verification of ensemble MOS forecasts. *21st Conf. on Probability and Statistics in the Atmospheric Sciences.* New Orleans, LA, Amer. Meteor. Soc., 2.3. [Available online at https://ams.confex.com/ams/92Annual/webprogram/Paper197489.html]

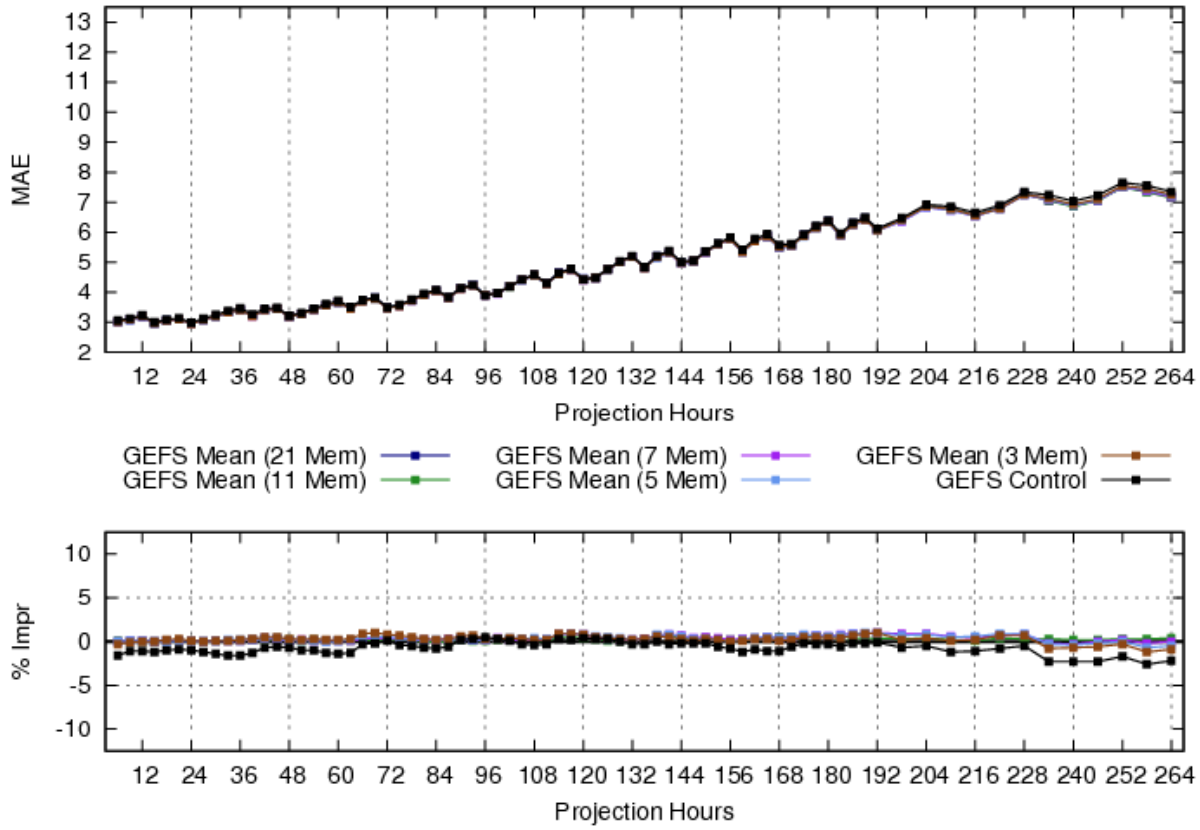

Figure 1. Mean absolute error (top) and percent improvement from the baseline 21-member mean (bottom) for cool season 2-m temperature.

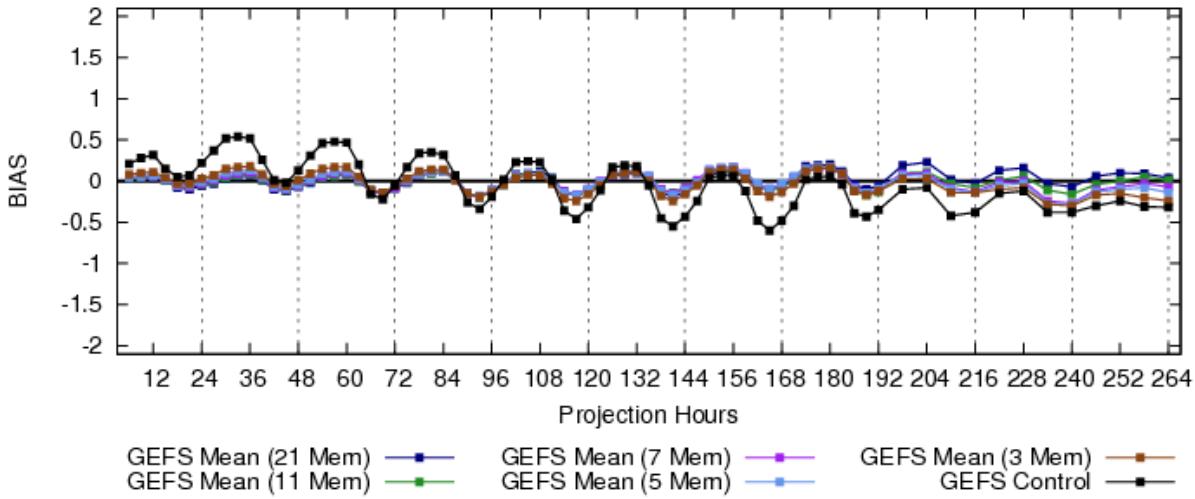

Figure 2. Bias for cool season 2-m temperature.

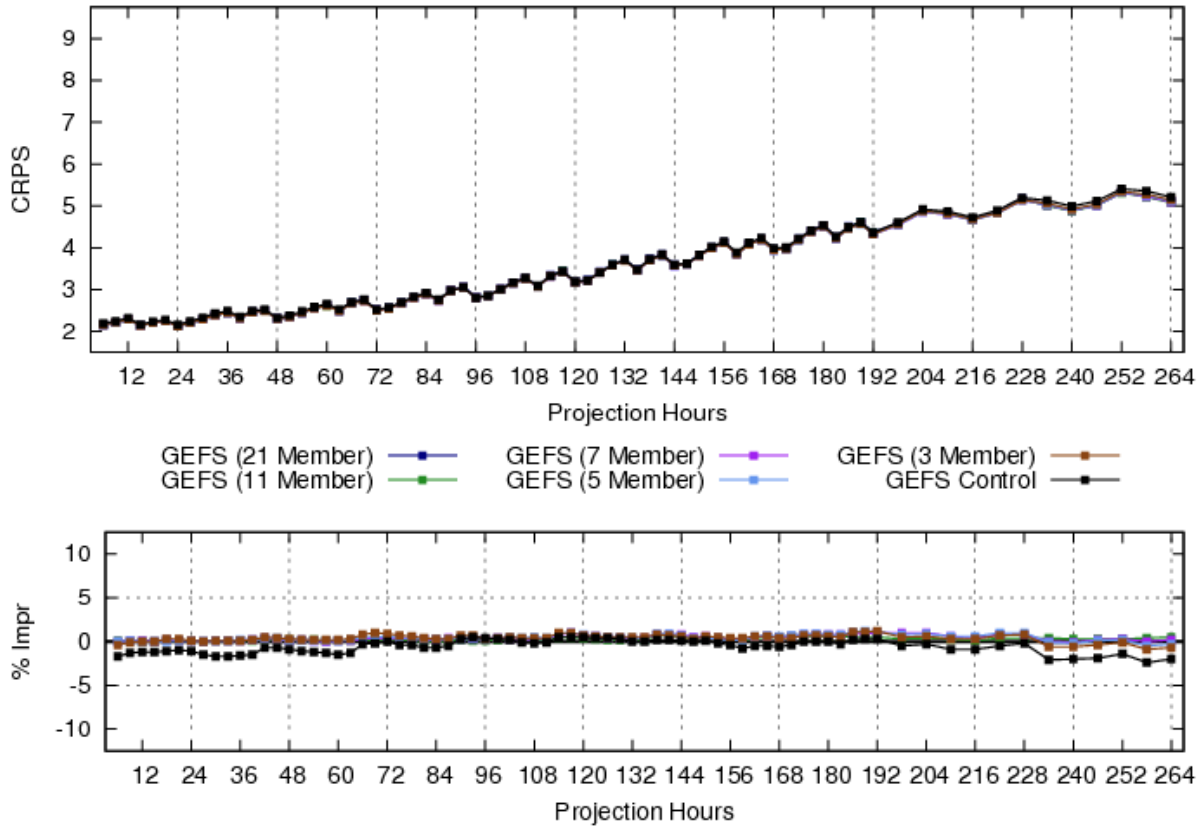

Figure 3. Continuous ranked probability score (top) and percent improvement from the baseline 21-member mean (bottom) for cool season 2-m temperature.

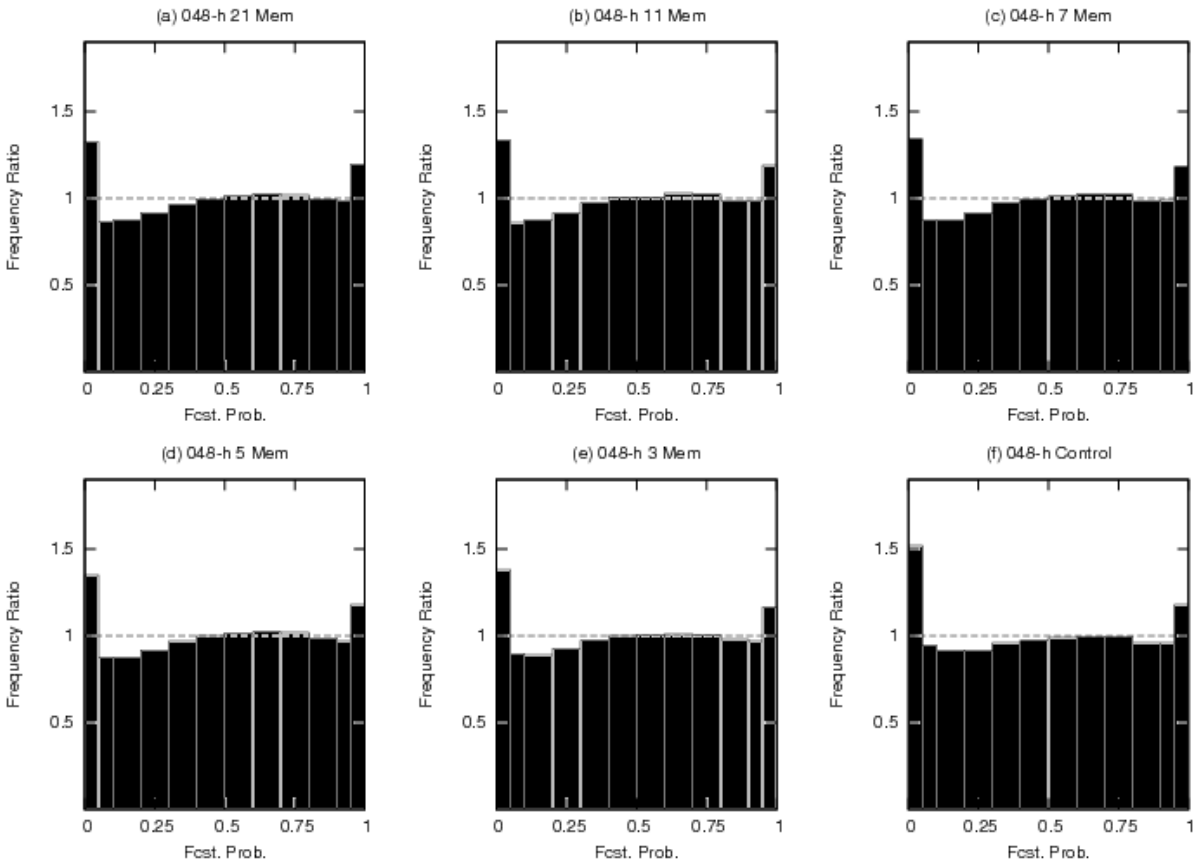

Figure 4. Probability integral transform (PIT) histograms for 48-hour cool season 2-m temperature.

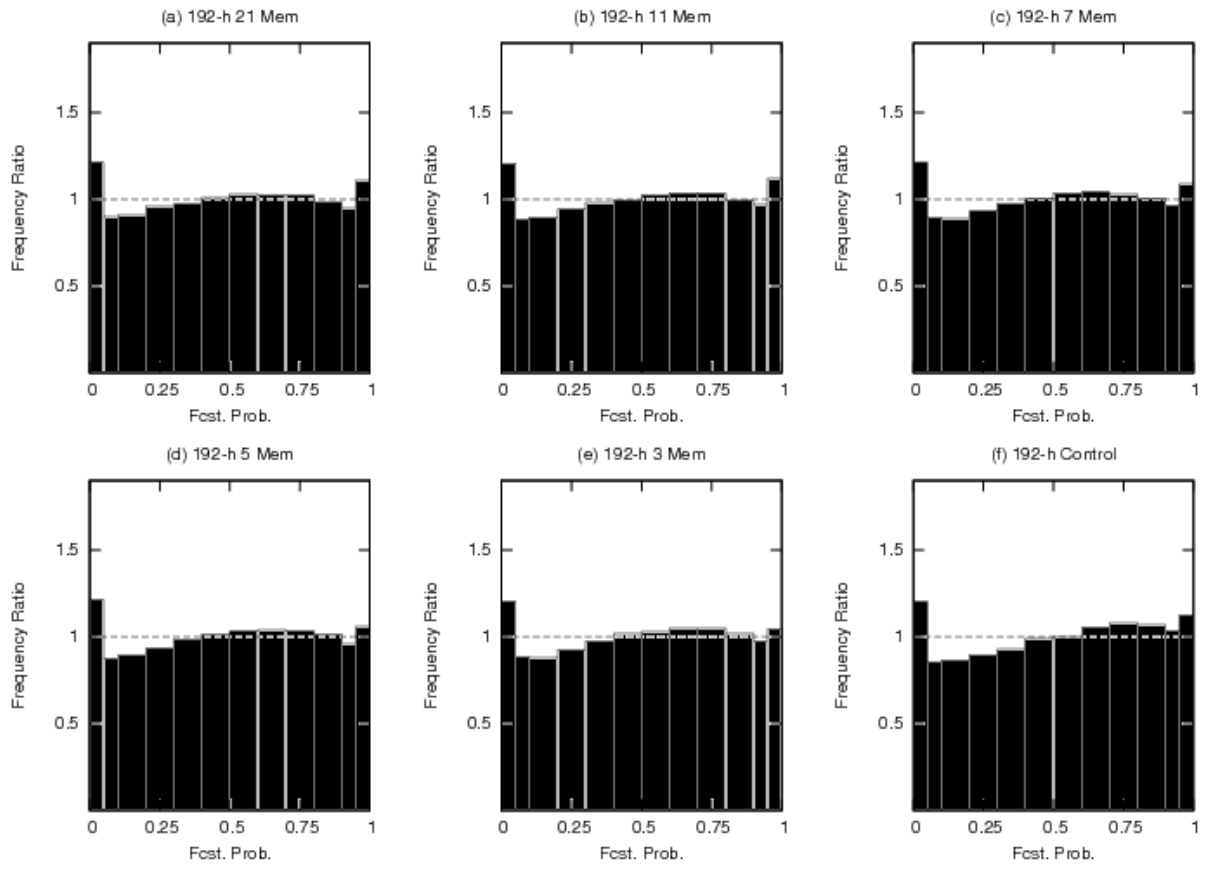

Figure 5. Probability integral transform (PIT) histograms for 48-hour cool season 2-m temperature.

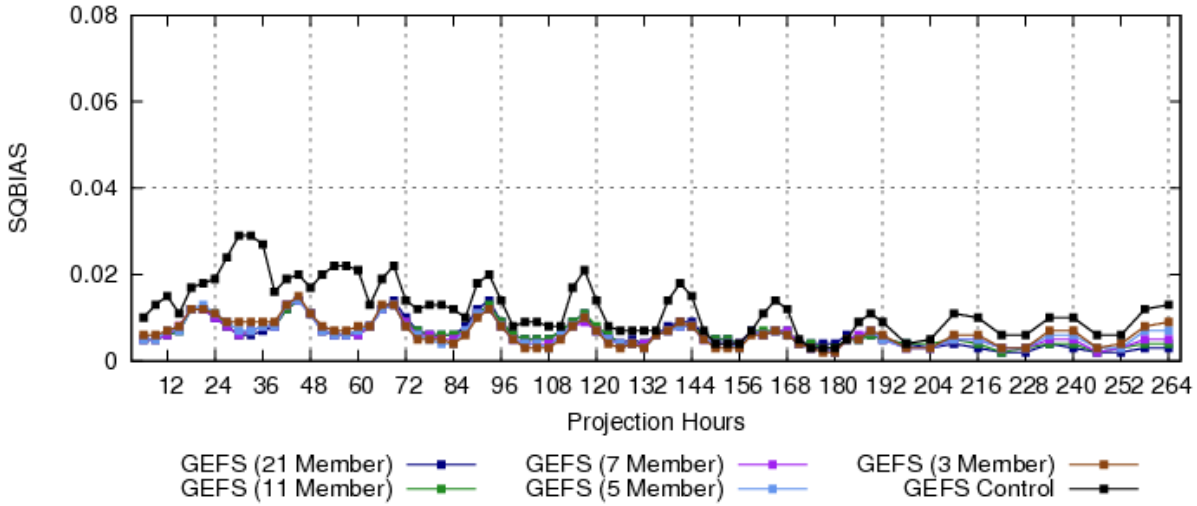

Figure 6. Squared bias in relative frequency (SQBIAS) for cool season 2-m temperature.

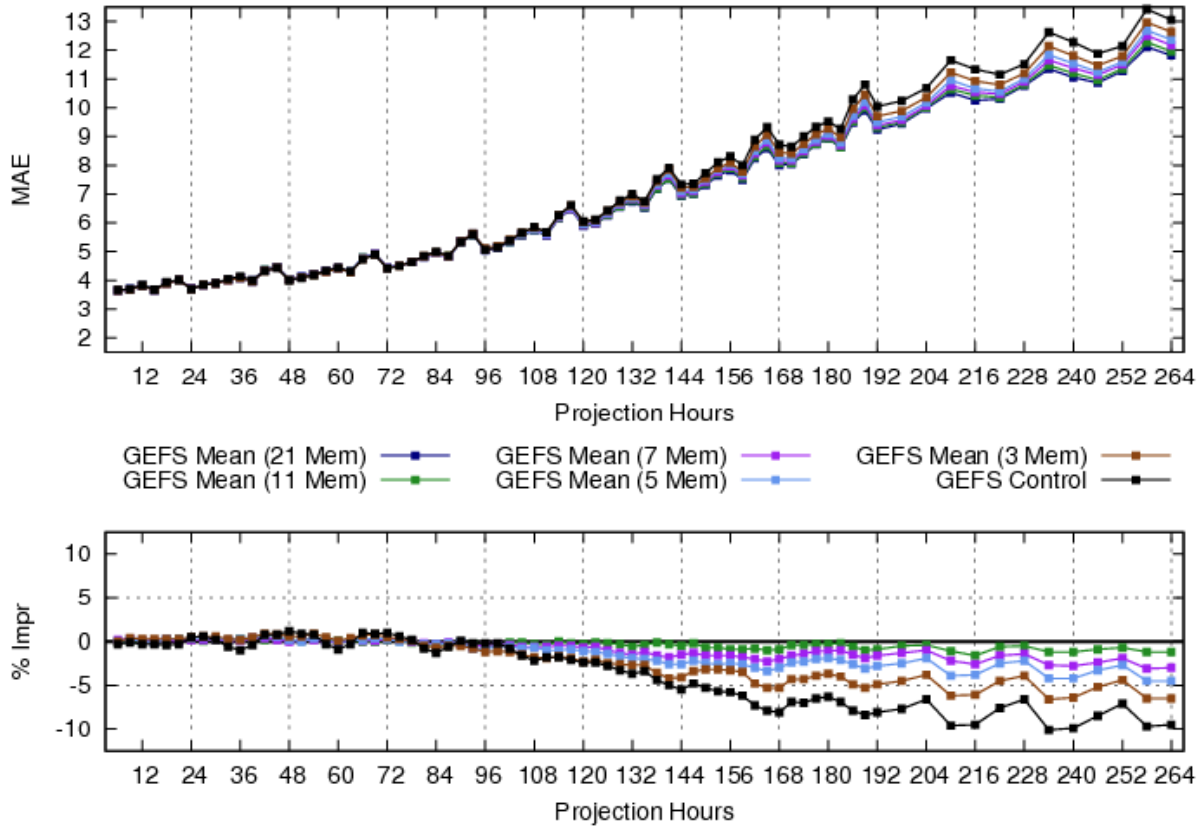

Figure 7. Mean absolute error (top) and percent improvement from the baseline 21-member mean (bottom) for cool season 2-m temperature. Observations from the 5% climatological tails were used for verification.

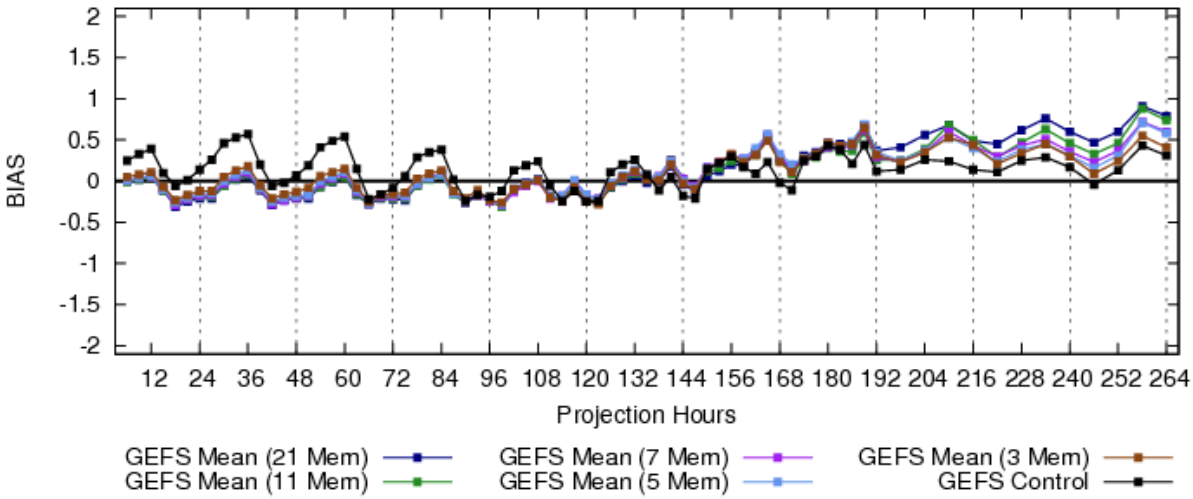

Figure 8. Bias for cool season 2-m temperature. Observations from the 5% climatological tails were used for verification.

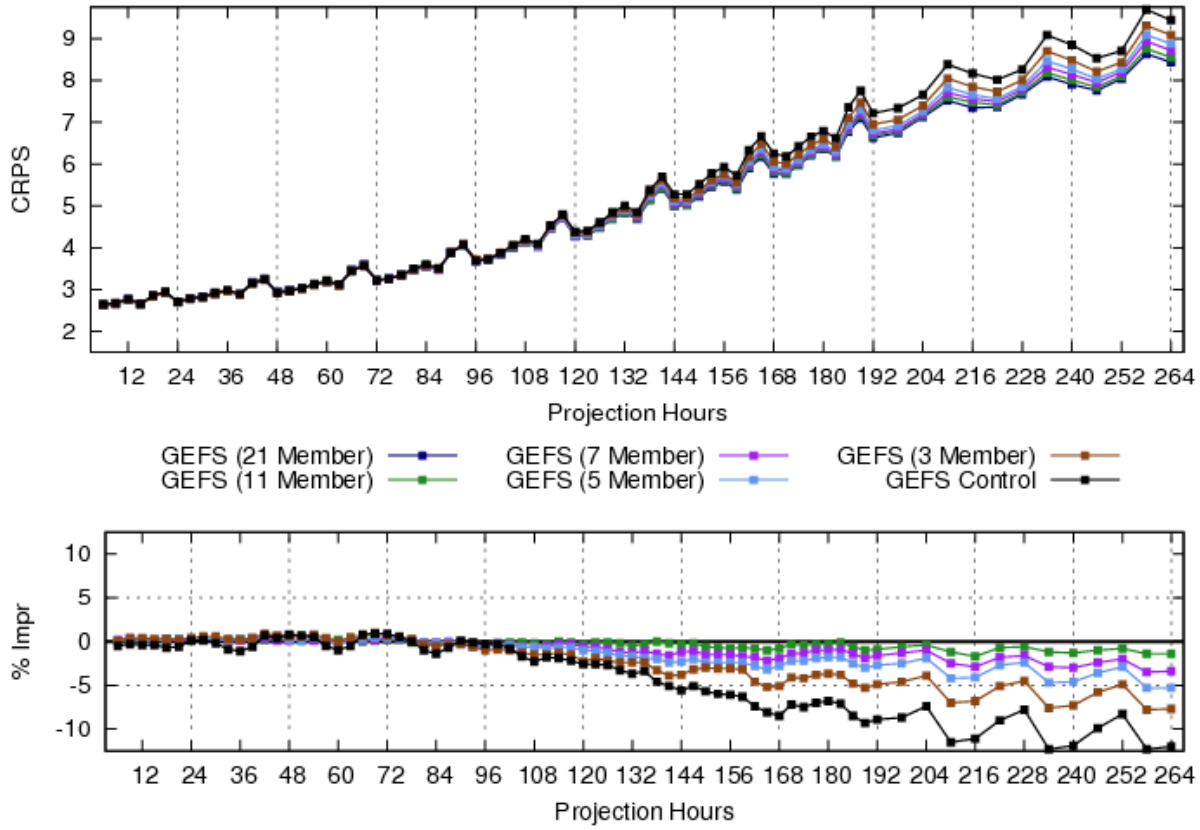

Figure 9. Continuous ranked probability score (top) and percent improvement from the baseline 21-member mean (bottom) for cool season 2-m temperature. Observations from the 5% climatological tails were used for verification.

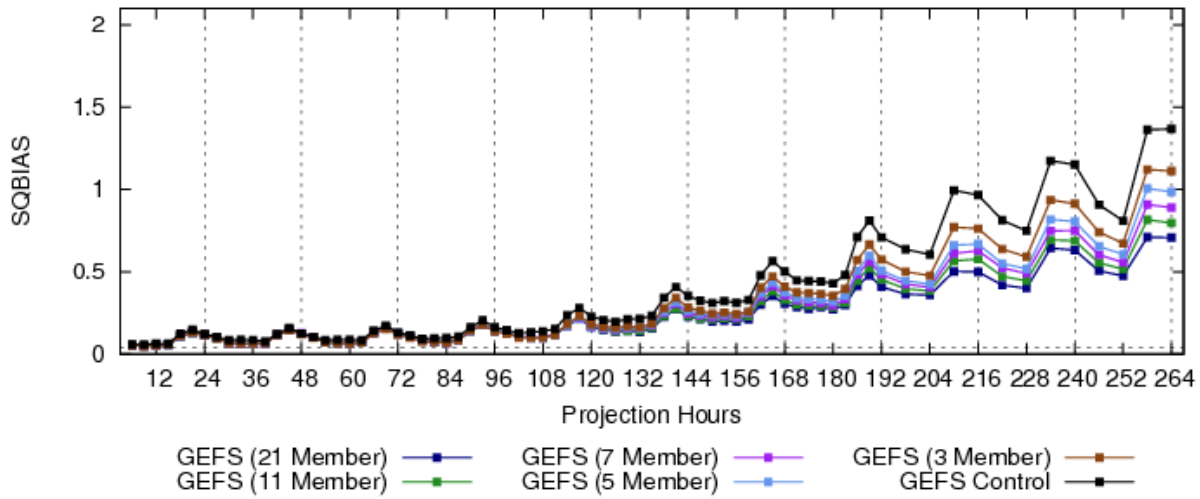

Figure 10. Squared bias in relative frequency for cool season 2-m temperature. Observations from the 5% climatological tails were used for verification.