SECOND MOMENT CALIBRATION AND COMPARATIVE VERIFICATION OF ENSEMBLE MOS FORECASTS

Bruce A. Veenhuis* and John L. Wagner

Meteorological Development Laboratory
Office of Science and Technology
National Weather Service
Silver Spring, Maryland

1. INTRODUCTION

Ensemble forecasting systems are routinely run at many operational meteorological forecasting centers. Ensembles provide useful forecast guidance; however, near-surface weather element forecasts derived directly from ensembles typically contain systematic biases. Moreover, ensembles are often under-dispersive. That is, too frequently the verifying observations fall outside the predicted envelope. Statistical post processing can improve the utility of ensembles for operational forecasters by correcting these deficiencies. Numerous post processing techniques have been proposed including Ensemble Dressing (Roulston and Smith 2003, Wang and Bishop 2005), Bayesian Model Averaging (BMA; Raftery et al. 2005, Wilson et al. 2007), Nonhomogenous Gaussian Regression (NGR; Gnieting et al. 2005, Wilks and Hamill 2007), and Ensemble Regression (Unger et al. 2009). One of the earliest techniques, Model Output Statistics (MOS; Glahn and Lowery 1972), has been used by the National Weather Service's (NWS) Meteorological Development Laboratory (MDL) for decades to statistically post process deterministic numerical models. Glahn, et al. (2009a), developed a MOS-based technique called Ensemble Kernel Density MOS (EKDMOS) which is applicable to ensemble forecasting systems. EKDMOS generates a statistically reliable forecast cumulative density function (CDF) from the ensemble.

An important component of probabilistic forecasting for continuous weather elements is determining the expected skill of the ensemble-mean. Intuitively, the ensemble-mean should be less accurate when the ensemble-members spread widely and more accurate when they tightly cluster around a solution. Specifying the correct spread is referred to as 2nd moment calibration. The original EDKMOS technique outlined by Glahn, et al. (2009a) used an empirically determined spread adjustment factor to calibrate the 2nd moment. The method produced statistically reliable results over a large number of cases. However, the technique lacked station specificity. More recently MDL has improved EKDMOS by adapting a spread adjustment technique proposed by Grimit and Mass (2007). Specifically, we develop spread-skill relationships which model the expected skill of the ensemble-mean as a linear function of the ensemble spread. Several other studies have also used the Grimit and Mass (2007) adjustment technique. Kolczynski, et al. (2009) applied a technique called Linear Variance Calibration to model wind forecast uncertainty, and Eckel, et al. (2011) used a comparable method to calibrate a global ensemble and assess forecast ambiguity.

This paper provides a brief overview of EKDMOS and documents our recent improvement to the 2nd moment calibration procedure. We use our improved EKDMOS technique to generate probabilistic forecast guidance from a global multi-model ensemble and present verification results for 2-m temperature, dewpoint, daytime maximum temperature, and nighttime minimum temperature.

2. DATA

We used numerical model output from the North American Ensemble Forecast System (NAEFS; Toth et al. 2005). The NAEFS is a suite of 42 ensemble-member forecasts combining the Canadian Meteorological Centre's (CMC) Global Environment Multiscale (GEM) Model and NCEP's Global Ensemble Forecast System (GEFS). The operational centers distribute a similar set of model outputs from their respective ensembles on the same 1x1 degree grid. The GEM and GEFS are each composed of a control run and 20 ensemble-members. MDL maintains an archive of operational NAEFS forecasts from July 2007 to present.

Station-based observations were taken from an archive maintained by MDL. A set of 2303 stations distributed throughout the conterminous United States, Canada, Alaska, Hawaii, and Puerto Rico was used for development and testing.

3. THE EKDMOS TECHNIQUE

We developed MOS equations using the ensemble-means following Glahn, et al. (2009a). MOS uses forward screening multiple linear regression to relate numerical model predictors to verifying observations (i.e., predictands). Typically model fields closely related to the predictand are chosen. For example, model 2-m temperature and geopotential

*Corresponding author address:
Bruce A. Veenhuis, Meteorological Development Laboratory, 1325 East-West Highway, Silver Spring, MD, 20910; email
bruce.veenhuis@noaa.gov

thicknesses are the most common predictors in MOS 2-m temperature equations. Harmonics such as the cosine and sine of the day of the year are also offered as predictors and are important in the later projections. Forecasts were stratified into warm (April 1 – September 30) and cool (October 1 – March 31) seasons. We developed MOS equations for each projection, station, cycle, season, and forecast element. Separate equations were developed for the GEFS and GEM models. We applied the equations to the ensemble-members within each ensemble forecasting system. Kernel density fitting (Wilks 2006) was used to construct a probability density function (PDF) from the member forecasts. We used a normal kernel with a standard deviation equal to the MOS equation predicted standard error. The resulting PDF is our uncalibrated EKDMOS forecast.

To perform 2nd moment calibration we created spread-skill relationships that relate the ensemble-member standard deviation to the ensemble-mean standard error. This technique follows Grimit and Mass (2007). We first post processed each ensemble-member by applying either the GEFS or GEM-based MOS equation. Using the post processed members, for each forecast case *i* we calculated the ensemble-mean ($\overline{f_i}$) using

$$\overline{f_i} = \frac{1}{K}\sum_{k=1}^{K} f_{ik}, \qquad (1)$$

where K is the number of ensemble-members and f_{ik} is the ensemble-member *k.* In addition, we calculated the ensemble-member standard deviation (β_i) using

$$\beta_i = \sum_{k=1}^{K}\frac{\left(f_{ik}-\overline{f_i}\right)^2}{K-1}. \qquad (2)$$

The cases were sorted by the ensemble-member standard deviation from lowest to highest and grouped into equal case count bins. For each bin *b* we calculated the ensemble-mean standard error (α_b) and the bin-averaged ensemble-member standard deviation ($\overline{\beta_b}$) following

$$\alpha_b = \sqrt{\frac{\sum_{i\in b}(\overline{f_i}-o_i)^2}{N_b-1}}, \qquad (3)$$

and

$$\overline{\beta_b} = \frac{1}{N_b}\sum_{i\in b}\beta_i. \qquad (4)$$

Here $i \in b$ indicates summation over the cases within bin *b*, N_b is the bin case count, and o_i is the verifying observation. A linear regression line was fit to the derived data points providing a continuous function that relates the ensemble-mean standard error to the ensemble-member standard deviation. Testing revealed that approximately 100 cases per bin were required to develop a stable, monotonically increasing relationship. Because we had limited data available for development, four bins were used. To ensure reasonable relationships, we required the regression line slope parameter to be greater than 0. In addition, we used an *F*-test to determine if the slope parameter was statistically different from 0. We required there to be less than a 25% probability that the slope was positive solely due to chance. If either criterion was not met, we rejected the spread-skill relationship and substituted the standard error estimate from the original MOS equation.

Figure 1 illustrates how spread-skill relationships are developed. The example is for the 72-hour 2-m temperature forecast at the Baltimore-Washington International Airport, KBWI. The ensemble-member standard deviation is plotted on the abscissa while the error of the ensemble-mean is plotted on the ordinate. Each dot represents one dependent forecast case. In Figure 1a the vertical dashed lines are the breakpoints used to bin cases. We calculated the ensemble-mean standard error for the cases within each bin. The value for each bin is plotted as a black square in Figure 1b. The black line is the spread-skill relationship fit to the derived data points.

4. VERIFICATION RESULTS

In order to increase our independent sample size we performed cross validation (see Wilks). We developed MOS equations and spread-skill relationships using two years of dependent data and verified using the remaining independent year. All results presented below are for three years of cool season data covering the period 1 October 2007 – 31 March 2010. The verification statistics were computed using the full list of 2303 stations.

Mean absolute error (MAE) plots are shown in Figures 2 through 5 for 2-m temperature, dewpoint, daytime maximum temperature, and nighttime minimum temperature. In each figure, the MAE of the NAEFS EKDMOS mean forecast is plotted as the red line. For comparison we include the MAE for the operational GFS MOS (blue lines). The GFS MOS forecasts are taken from an archive maintained by MDL. The GFS MOS underwent a major update in March 2010, which is not reflected in our verification sample. We include the GFS MOS to approximately gauge the expected increase in skill afforded by NAEFS EKDMOS. For 2-m temperature and dewpoint, the NAEFS EKDMOS MAE is lower than the GFS MOS at all projections except at the 6-h and 9-h forecasts. The GFS MOS utilizes a persistence predictor at the earliest projections. Due to the delayed availability of the NAEFS data, EKDMOS does not use a persistence predictor, likely explaining the difference in accuracy. Examining the MAE plots for daytime maximum temperature and nighttime minimum temperature we see that NAEFS EKDMOS is more accurate. At the 8-day lead time, NAEFS EKDMOS provides approximately one day increase in accuracy. For example, the NAEFS EKDMOS 198-hour daytime maximum temperature forecast is almost as accurate as the 174-hour GFS MOS forecast.

To assess statistical reliability, probability integral transform (PIT) histograms are shown in Figure 6. An in-depth explanation of PIT histograms and their interpretation is provided by Hamill (2001) and Glahn, et al. (2009a). Mound-shaped PITs indicate over-dispersion while U-shaped PITs indicate under-dispersion. Ideally, the PIT should be uniformly flat with the height of each bin equal to 1. We see that for all forecast elements considered, the PITs are generally flat indicating the NAEFS EKDMOS forecasts are statistically reliable.

Compared to our original methodology, the spread-skill calibration technique produces greater day-to-day variability in the predicted spread. In Figure 7, we show histograms of the predicted 80% credible interval (CI) width for the 102-h 2-m temperature forecast at the Baltimore-Washington International Airport. Here we have normalized the 80% CI width by dividing by the mean value. The values along the abscissa can be interpreted as the percent difference from the average predicted spread. For example, a value of 1.2 implies the 80% CI is 20% wider than average, while a value 0.8 indicates the 80% CI is 20% narrower. Results for our original calibration technique, shown in Figure 7b, are labeled Spread Adjustment. Comparing spread-skill with spread adjustment, we see than the spread-skill histogram is much broader. Examining the full set of stations, we found that spread-skill produced greater spread variability at 92% of the stations and that on average the degree of spread variability doubled. Results for other weather elements and projections were similar.

This increased day-to-day spread variability does not degrade statistical reliability significantly. To show this, we constructed PIT histograms using forecasts stratified by spread. At each station, we sorted forecasts by the 80% CI width and grouped them into three equal case-count spread categories. Using station-pooled results, we constructed PITs for each spread category. An example is presented in Figure 8 for the 102-hour 2-m temperature forecast. The PITS are generally flat, indicating statistical reliability.

5. DISCUSIONS AND CONCLUSIONS

MDL has used the EKDMOS technique to generate accurate and statistically reliable forecast guidance from the NAEFS. MDL has adapted EKDMOS to use the 2nd moment calibration technique proposed by Grimit and Mass (2007). The method improves the day-to-day spread variability while maintaining statistical reliability. This will increase the value of NAEFS EKDMOS forecast for end users.

NAEFS EKDMOS should be operationally implementation on the NCEP Central Computing System (CCS) by the end of the 2012 fiscal year. The implementation will use spread-skill calibration and include forecasts for 2-m temperature, dewpoint, daytime maximum temperature, and nighttime minimum temperature. Forecasts will be generated at stations and analyzed to 2.5 km grids covering the CONUS and Alaska using the BCDG technique (Glahn et al., 2009b). Experimental NAEFS EKDMOS products are currently hosted online at http://www.mdl.nws.noaa.gov/~naefs_ekdmos. The website provides forecast images, meteograms for select METAR stations, and gridded forecasts in a GRIB2 format. Once implemented, operational gridded NAEFS EKDMOS guidance will be generated twice daily from the 0000 and 1200 UTC runs of the NAEFS. Gridded forecast will be made publically available via the NWS's National Digital Guidance Database (NDGD). Further details regarding the NDGD can be found online (http://www.nws.noaa.gov/ndgd/index.shtml).

7. REFERENCES

Eckel F. A., M. S. Allen, and M. C. Sittel: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting*, accepted.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11,** 1203–1211.

______, M. R. Peroutka, J. Wiedenfeld, J. L. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009a: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268.

______, K. Gilbert, R. Cosgrove, D. P. Ruth, K. Sheets, 2009b: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529.

Gneiting T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,***133**, 1098-1118.

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, **135**, 203–221.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129,** 550–560.

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, A. Deng, 2009: Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *J. Appl. Meteor. Climatol.*, **48**, 2001–2021.

Raftery A. E., T. Gneiting, F. Balabdaoui, and M. Polakawski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133,** 1155-1174.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16-30.

Toth, Z., and Coauthors, 2005: The North American Ensemble Forecast System. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., CD-ROM, 11A.1.

Unger D. A., H. van den Dool, E. O'Lenic, D. Collins, 2009: Ensemble regression. *Mon Wea Rev.*, 137**, 7,** 2365-2379.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteo. Soc.*, **131** 956-986.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Elsevier/Academic Press, 627 pp.

_____, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135,** 2379-2390.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using bayesian model averaging. *Mon. Wea. Rev.*, **135**, 1364-1385.

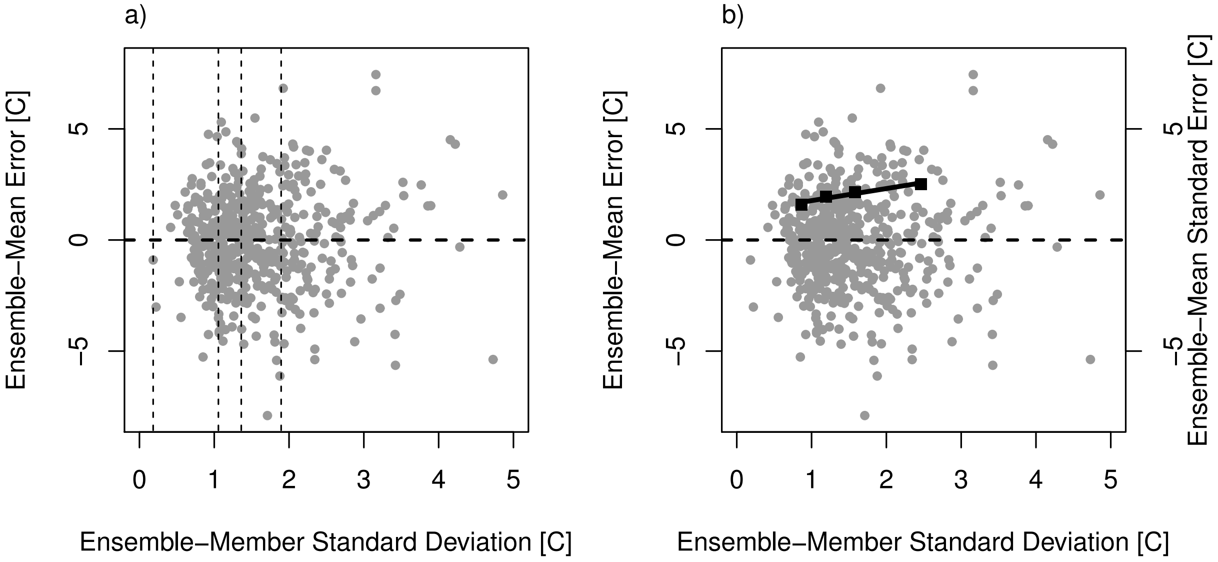


Fig. 1: Scatter plots illustrating spread-skill relationship development. The ensemble-member standard deviation is plotted along the abscissa while the error of the ensemble-mean is plotted along the ordinate. Each dot represents one forecast case. The vertical dashed lines (Fig. 1a) are the breakpoints used to bin cases. In Figure 1b the black squares are the bin-calculated values for the ensemble-mean standard error while the black line is the spread-skill relationship.

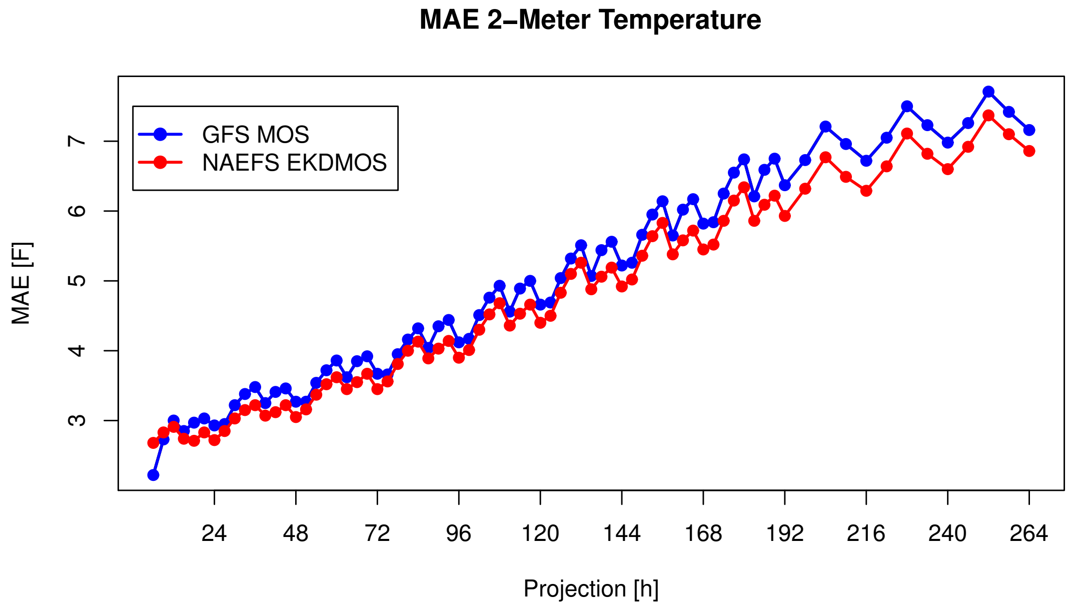


Fig. 2: Cool season 2-m temperature MAE plotted by projection.

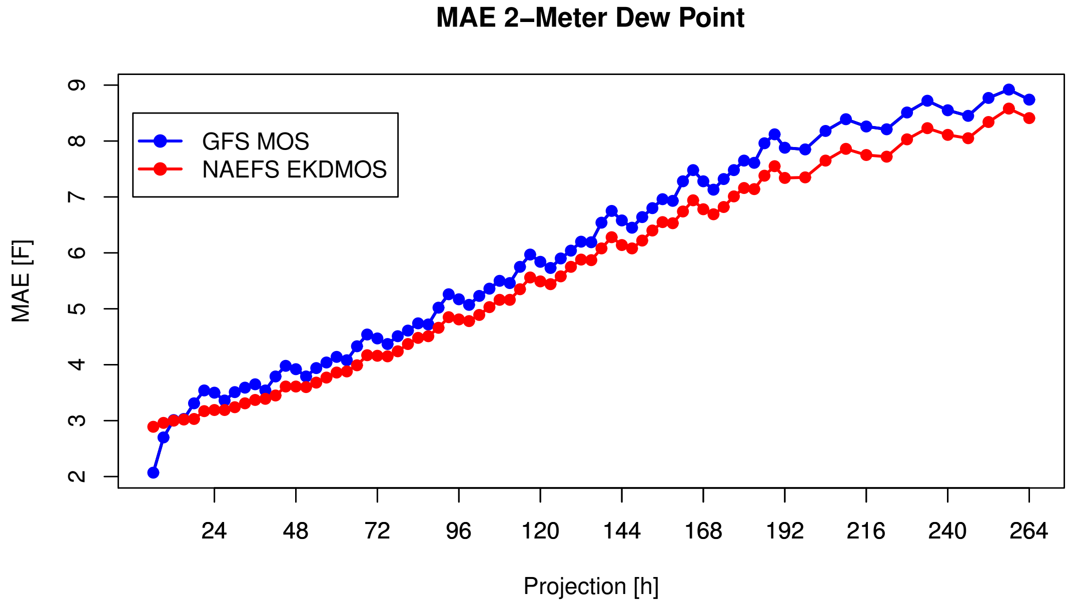


Fig. 3: Same as Figure 2 but for dewpoint.

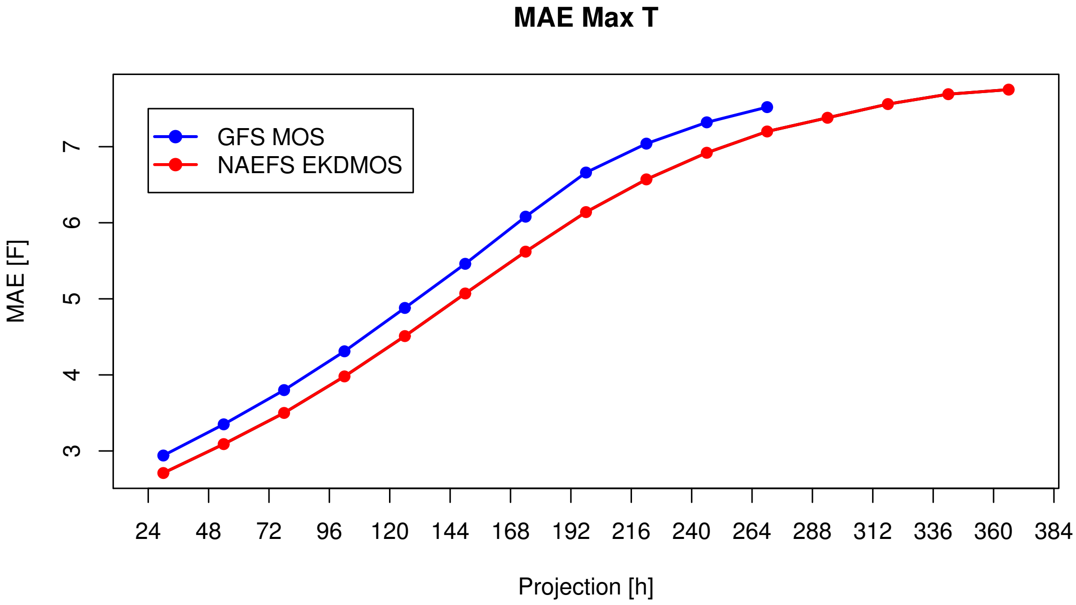


Fig. 4:. Same as Figure 2 but for daytime maximum temperature.

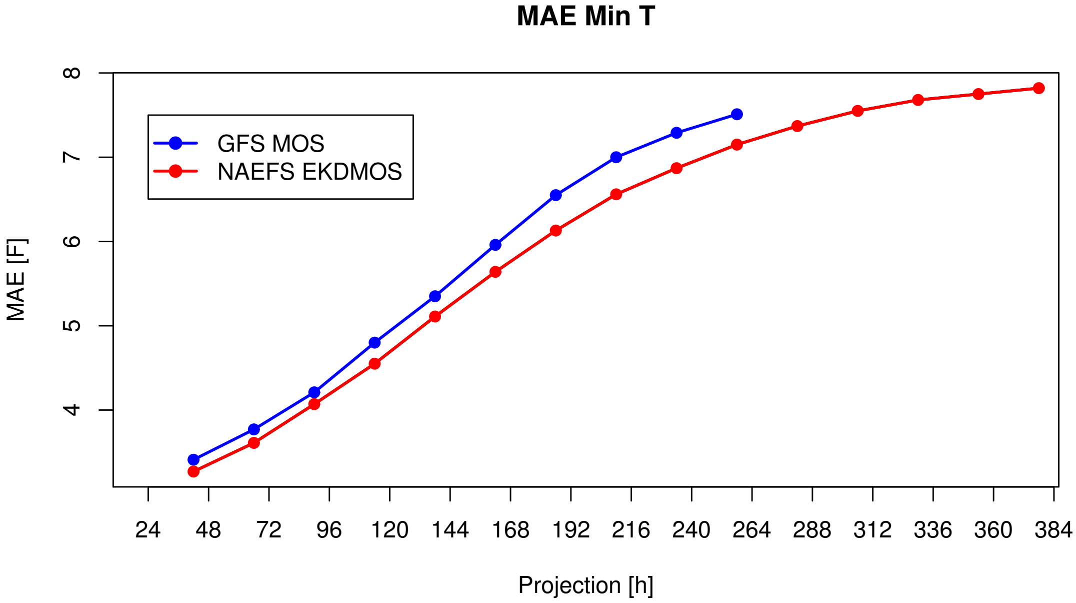
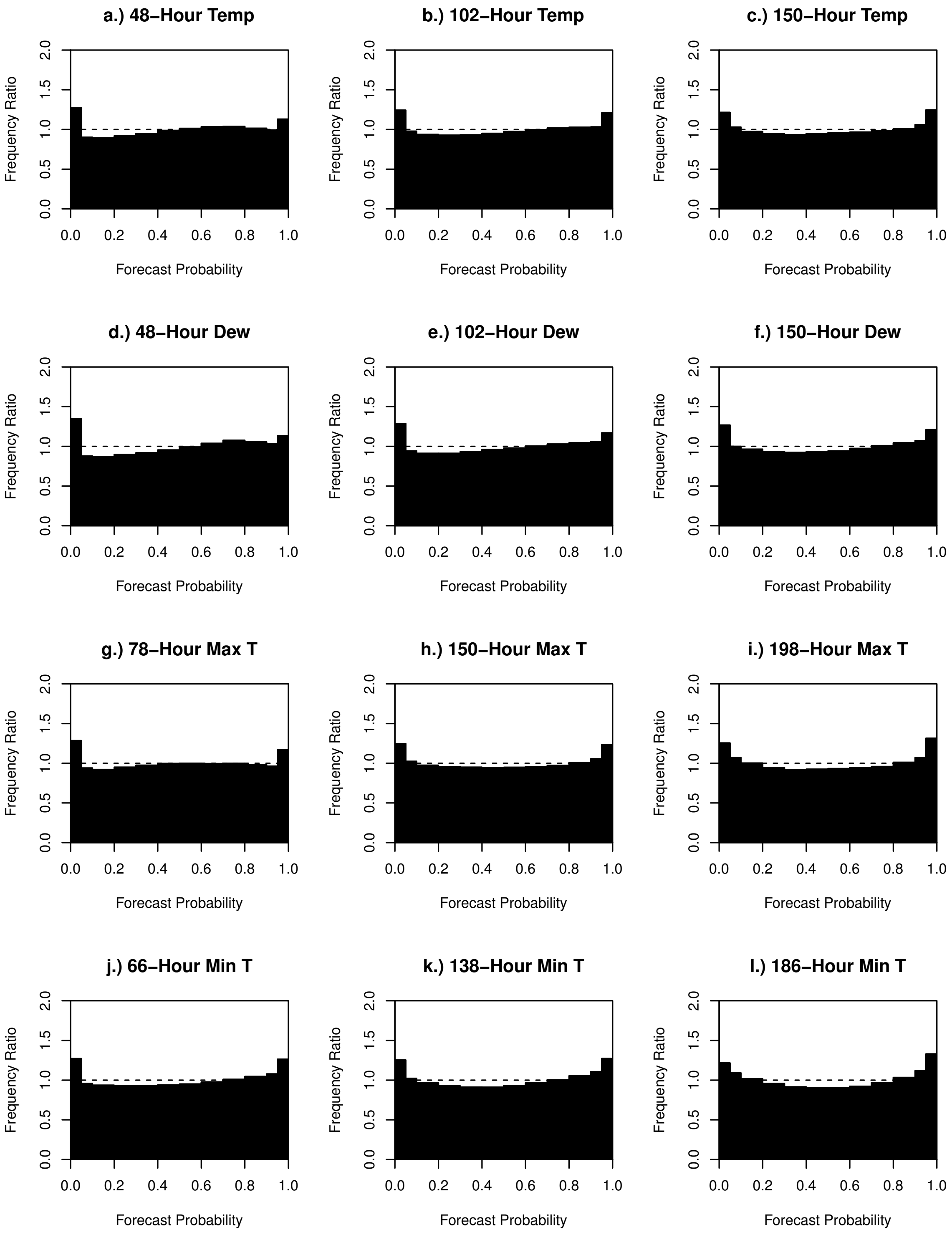


Fig. 5: Same as Figure 2 but for nighttime minimum temperature.



a.) 48–Hour Temp

b.) 102–Hour Temp

c.) 150–Hour Temp

d.) 48–Hour Dew

e.) 102–Hour Dew

f.) 150–Hour Dew

g.) 78–Hour Max T

h.) 150–Hour Max T

i.) 198–Hour Max T

j.) 66–Hour Min T

k.) 138–Hour Min T

l.) 186–Hour Min T

Fig. 6: PIT histograms for 2-m temperature (a-c), dewpoint (d-f), daytime maximum temperature (g-i) and nighttime minimum temperature (j-l).

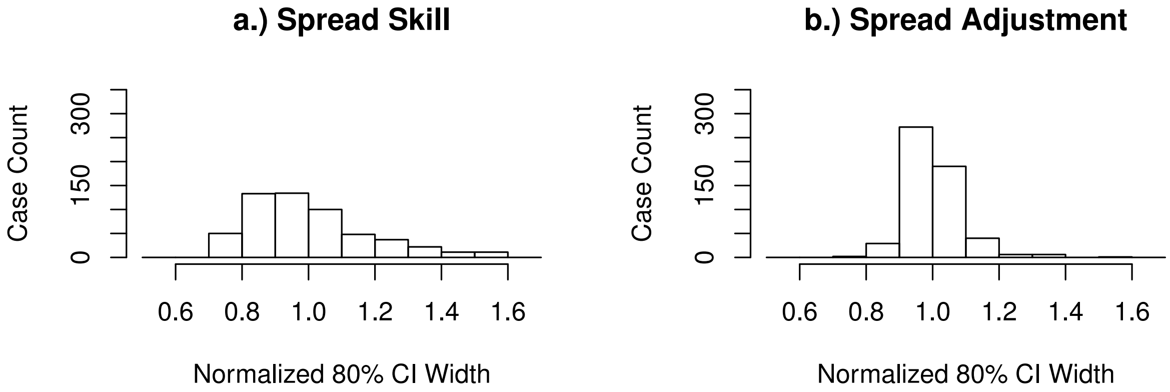


Fig. 7: Histograms of the predicted 80% credible interval (CI) width comparing spread-skill (a) with our original 2nd moment calibration called Spread Adjustment (b). The figures are for the cool season 102-h 2-m temperature forecast at the Baltimore-Washington International Airport.

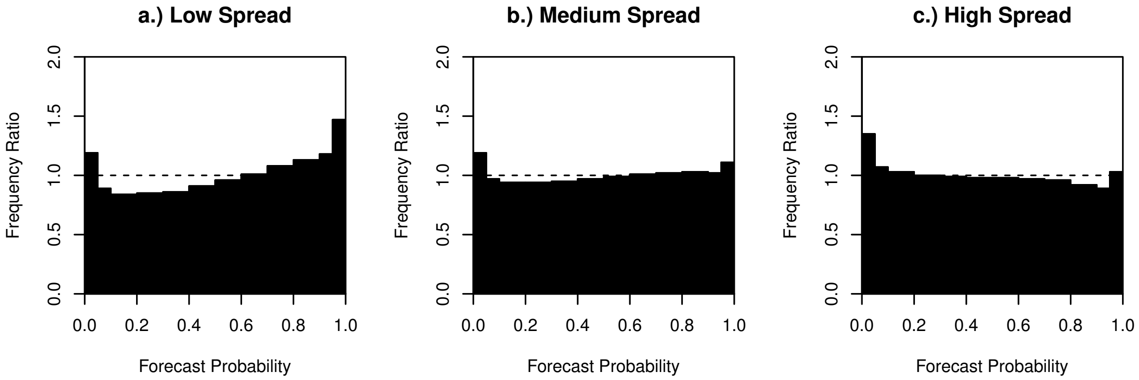


Fig. 8: PIT histograms for 102-h 2-m temperature. Forecasts have been stratified into low (a), medium (b), and high (c) spread cases using station-specific thresholds. The PITs were created using the full set of 2303 stations.