Bias Removal and Model Consensus Forecasts of Maximum and Minimum Temperatures Using the Graphical Forecast Editor

Jeffrey T. Davis
National Weather Service Office
Tucson, Arizona

1. Introduction

Operational Numerical Weather Prediction (NWP) models have inherent biases that need to be removed either objectively or subjectively before used in official National Weather Service (NWS) forecasts. The recent shift of the NWS to the Interactive Forecast Preparation System (IFPS) to create and distribute digital weather forecasts has led to more of a dependency on the direct use of NWP models. This dependency on raw model output is mainly due to the gridded format needed to initialize digital weather forecasts. The primary tool for creating these digital forecasts is the Graphical Forecast Editor (GFE) (Lefebvre, 1995). The GFE has two design features that account for the objective and subjective adjustments of the native NWP model grid. The front-end feature is referred to as "Smart Initialization" which is used to objectively derive sensible weather elements and downscale the coarser resolution models to a higher 2.5 km or 5 km grid spacing. The second feature is interactive and accounts for both subjective and objective adjustments through the use of simple graphical editing tools and "Smart Tools".

The baseline GFE design lacks the capability to objectively remove model biases in the initialization process. This is not to say that the GFE should have this capability, but high resolution, bias-corrected model grids are not currently included in the overall IFPS design. As a result, forecasters are tasked with subjectively correcting for these model biases using the GFE tools. Mass (2003) pointed out that this is a poor use of human resources when an objective bias removal technique would likely produce better results than any subjective attempt. Although some manual adjustments may always be required by a forecaster, most of the bias removal can be accomplished in the initialization or model post-processing stages of the IFPS.

The need for bias corrections arises from the many sources of systematic errors in NWP modeling systems. Over the past 30 years, the Model Output Statistics (MOS) approach (Glahn and Lowry, 1972) has been successfully used to improve upon model output through bias removal and statistical correction for selected sites. One drawback of MOS is that it requires a long training period of archived model fields from an unchanged or static model. Today, modeling centers make frequent changes to numerical procedures, physics, and resolution of NWP models. To overcome this ever-changing model base, other techniques more dynamic in nature are being investigated. Wilson and Vallèe (2002) describe an updateable MOS system used in Canada which was specifically designed to adapt to model changes. Mao et al. (1999) developed a similar technique that updated daily and relied on only the most recent 2 to 4 weeks of model and observational data.

Stensrud and Skindlov (1996) showed that a much simpler method of using a 7-day running mean bias correction could improve upon model grid-point forecasts of maximum temperature. Recently, Steed and Mass (2004) experimented with several different spatial techniques of applying bias removal to forecasts of temperature from a mesoscale model. For a couple of these methods they utilized different interpolation techniques to distribute bias calculations at verifying stations on the model grid. Another method applied a simple domain average from observed biases at all verifying sites. They also looked at a removal technique which used the Rapid Update Cycle (RUC) initial analysis as ground truth for calculating model biases at each grid-point. Their results indicated that each removal method performed nearly equally as well as the others during the winter months of 2003 to 2004 in the Pacific Northwest. They also found that a removal method using a 2-week running bias had the least amount of error compared to periods of 1, 3, 4 and 6 weeks.

In addition to correcting for model biases, the nature of weather prediction entails uncertainties that forecasters subjectively account for when making a deterministic forecast of sensible weather. Several studies have shown that averaging two or more different numerical forecasts to produce a consensus is more accurate in the long run than a single forecast (Verret and Yacowar, 1989, Vislocky and Fritsch, 1995). Furthermore, Etherton (2003) demonstrated the usefulness of combining MOS and bias-corrected model output from a short-range NWP ensemble using a weighted average. Stensrud and Yussouf (2003) also found that the simple mean of bias-corrected 2-meter temperatures from a 23 member multimodel ensemble was as accurate as the Nested Grid Model (NGM) MOS for sites in New England.

Stensrud and Yussouf (2003) also demonstrated the added value to users when using ensemble probabilities in a simple cost-loss model; thus, pointing out the advantage of using multiple guidance sources over a single forecast. Considerable information can be extracted from all the guidance sources to measure forecast uncertainties. Summary products similar to that of traditional NWP ensemble systems can be produced in the form of a mean, standard deviation, range, and extremes as well as raw or calibrated probabilities for various thresholds or categories. This extra information can be used directly in an automated probabilistic forecast system or as objective guidance in the IFPS forecasting process.

This paper describes an approach being explored to improve first-guess grids of maximum and minimum temperatures using the GFE which can be transferred to other elements. The method attempts to incorporate both aspects of bias removal and forecast uncertainty. The technique uses a simple 7-day running mean error correction and a lagged ensemble of bias-corrected and gridded station MOS to create a blended forecast. The Global Forecast System (GFS) and Eta models along with their associated MOS forecasts are used in this study. The premise of the approach is that a consensus or blended forecast from two or more different bias-corrected guidance sources is more skillful than a single guidance product. Preliminary results are examined for the feasibility and usefulness of the constructed error feedback and blending system in short-range prediction.

2. Methodology

The flexible configuration design of the GFE allows for the expansion of mutable databases and rapid prototyping through the use of the Python scripting language. These features make the GFE ideal for the construction of the error feedback and forecast blending system used in this study. The server software for the GFE stores and manages the different models that comprise the ensemble of grids used in the consensus forecasts. GFE procedures written in Python calculate the model forecast errors used in the bias corrections and derive the consensus forecasts of maximum and minimum temperatures based on each model's past performance. In addition to the consensus forecast, ensemble style products such as the arithmetic mean, range, spread, and extreme grids as well as raw probabilities for thresholds are calculated. The schematic for the modified GFE configuration is shown in Figure 1.

2.1 Initialization

The first step in the process uses the GFE "Smart Initialization" to downscale the coarser model resolutions to a grid spacing of 2.5 km or 5 km. For temperature, a high resolution terrain dataset is used to adjust model lapse rates based on elevation. This results in a more detailed temperature grid that follows closely the topography. There is no attempt to objectively add value to the raw model output in this downscaling step. The maximum and minimum temperatures are derived from the hourly temperatures at model time steps of usually 3 or 6 hours. LeFebvre et al. (2002) provides a more comprehensive description of the GFE initialization algorithms.

2.2 MOS Adjustments

After the models have been downscaled to the IFPS resolution, MOS point forecasts for both the GFS and Eta are interpolated on a grid using a popular collection of GFE tools called "MatchGuidance" (Barker, 2004). These tools decode the MOS bulletins and run an objective analysis which uses the downscaled model grids as a first-guess field. Corrections are applied to the first-guess field so that the MOS values match at specified grid-points. The final product used is the MOS-adjusted grids of maximum and minimum temperatures.

2.3 Storage of Model Grids and Forecast Errors

Following the downscaling and MOS adjustments, the gridded forecasts are archived in separate model databases with elements stored by cycle and projection times. The model cycles valid at 0000, 0600, 1200, and 1800 UTC are archived. The temperature elements are stored at projection times of same day (Day+0), next day (Day+1), second day (Day+2), and in some cases for the third day (Day+3). For the Eta MOS-adjusted grids, only the 0000 and 1200 UTC model cycles are available.

The performance of the forecasts can be evaluated based on the "Record of Analysis" once the grids have been archived. For this study, another popular GFE tool called "MatchObsAll" (Barker, 2004) is used for the "Record of Analysis". The tool extracts hourly temperature observations from various sources and runs the same objective analysis used in the MOS adjustments. The background field for the analysis is an average of the Eta model forecast and the analysis from the previous hour. The initial

"Record of Analysis" for maximum and minimum temperatures is derived from the analysis of hourly temperatures. An additional adjustment is made to these initial analysis grids by re-running the objective analysis using observed values at cooperative observer sites (COOP) taken from the Regional Temperature and Precipitation (RTP) summaries. This final adjustment results in the nearest grid-point to the COOP station matching the observed value. The "Record of Analysis" is considered ground truth and forms the basis for the model forecast error calculations. The forecast errors are stored by model cycle and projection time for at least 7 days and are used in the bias calculations.

2.4 Bias Removal

A period of 7 days is used to calculate the running bias at each grid-point to be applied to the downscaled NWP model output. No attempt is made to compare the performance of the 7-day running bias to shorter or longer periods. The 7-day period is assumed to be enough time to capture a useful model bias and short enough to respond quickly to changes in the model and/or GFE initialization algorithm. In addition to model and initialization changes, a 7-day period is expected to adapt more rapidly to transitioning weather regimes in comparison to a longer period of 2 weeks or 1 month.

After the model has been downscaled to the IFPS resolution, the 7-day running mean error is calculated based on the model cycle and forecast times. Figure 2 shows how the magnitude of the forecast error can change by cycle and projection times. The running bias is applied to the model output at individual forecast times out to two days (Day+2) and in some cases to the third day (Day+3). As a result, each model run and lead time will have a unique bias value which is subtracted from the model grid. Further, each grid-point will have a unique bias value as well.

2.5 Consensus Forecasts and Summary Grids

The approach used in this study incorporates two different but valid bias removal methods (mean error-correction and MOS) taken from the same model to construct an ensemble of forecasts. Both the GFS and Eta bias-corrected and MOS-adjusted grids are used for the consensus forecasts of maximum and minimum temperatures. The consensus forecast is fed back into the system and treated as an additional model. To increase the number of forecasts without adding more models, a lagged system is constructed by using previous model forecasts all valid at the same verification time. Figure 3 illustrates the layout of the time-lagged model grids.

The consensus forecast of maximum and minimum temperatures is a weighted average of all forecasts based on each model's past performance. There is a wide range of methods, such as linear regression, gradient descent, partial least squares, and fuzzy logic, that can be used to determine the blending weights for each model. However, for this initial step, a simple weighting scheme is used. A 7-day running Mean Absolute Error (MAE) for each model cycle and forecast time is calculated, and a relative weight is given to each forecast based of the linear weighting function shown in figure 4. A fixed relative weight of .70 is applied to the most recent model cycle based on the premise that the latest forecast is more skillful than the older ones.

To show forecast uncertainty and identify errors in the feedback and blending system, grids similar to that of a traditional NWP ensemble prediction system are prepared. The simple arithmetic mean is calculated which can also exhibit predictive skill in the long run. The spread about the mean is also made available, providing some insight into forecast uncertainties and analysis errors propagating through the system. For example, Figure 5 shows a large spread in the region of the Grand Canyon which can be tied to bad data in the objective analysis of observed high temperatures. Along with the spread, high and low extreme grids as well as the range are generated to provide additional information on the variability of the solutions. As a final advantage of using multiple forecasts, raw probabilities for exceeding temperature thresholds are computed. However, there is currently no attempt to calibrate or determine the reliability of these probabilistic forecasts.

3. Preliminary Results

To look at the feasibility and usefulness of the bias removal and model blending methods, the performance of both need to be evaluated and compared to a benchmark. For the purpose of this study, three main questions are investigated: (1) Will a simple bias removal technique improve upon direct model output? (2) Will a weighted consensus forecast from multiple guidance sources be more accurate than a single forecast? (3) How do both methods compare to MOS forecasts?

To answer these questions and simplify the process, a point-based approach is used instead of a grid-based approach to validate the methodologies. The forecast domain covers most of Arizona into parts of southeast California and extreme western New Mexico. Thirteen sites ranging in elevation from 59 feet below Mean Sea Level (MSL) to 7, 078 feet MSL are used in the validation dataset. Each site corresponds to the nearest grid-point of a GFS and Eta MOS station. Figure 6 shows the forecast domain and locations of the 13 sites overlaid on the IFPS 2.5 km resolution terrain. The same objective analysis used for the “Record of Analysis” and the gridded station MOS insures that observed temperatures are in sync at grid-points with the MOS bulletins. Since the Eta MOS is only available for two of the four model cycles, the 1200 and 0000 UTC runs are used. Forecasts of maximum and minimum temperatures at lead times valid for the next day (Day+1) and second day (Day+2) during the months of May through July 2004 are evaluated.

3.1 Bias-Corrected Forecasts

Preliminary results indicate that the bias-corrected forecasts of maximum and minimum temperatures have smaller Mean Absolute Error (MAE) values for all sites than those produced by the GFS and Eta direct model output. The percentage of improvement by the bias-corrected forecasts over the direct model output as measured in terms of the MAE is given by:

$$\text{Improvement}(\%) = ((MAE_{raw} - MAE_{bias}) / MAE_{raw}) \times 100,$$

Where MAE_{raw} is the MAE for the direct or raw model output and MAE_{bias} represents the bias-corrected values. For the 13 sites, the bias-corrected forecasts show improvement

over the direct output for both the GFS and Eta models. In general, results show over a 50 percent improvement at most sites. The only exception to this is for maximum temperature forecasts from the Eta model in which a less than 40 percent improvement is achieved. This lower percentage for the Eta model might suggest that the simple bias removal technique is less useful as the direct model output becomes more accurate. The highest percentage of improvement is at stations above 5,000 feet. For example, the 0000 UTC bias corrections for maximum temperatures at Flagstaff, Arizona (~7,000 feet) show a 73 percent improvement over the raw output for the GFS and 80 percent for the Eta model. Table 1 shows the average MAE values in degrees Fahrenheit (F) and average percentage of improvement taken from all 13 sites.

3.2 Bias-corrected forecasts vs. MOS

The bias-corrected forecasts for maximum temperature compared to MOS show mixed results. The GFS bias-corrected MAE is slightly lower than the GFS MOS for the 1200 UTC cycle at both lead times. Otherwise, the MOS MAE values for maximum temperature are lower than the bias removed forecasts. In the case of minimum temperatures, MOS has lower MAE values than the bias-corrected forecasts for most of the sites in the study. Table 2 shows the average MAE values for the MOS and bias-corrected forecasts for both model cycles at lead times of Day+1 and Day+2.

One thing to take into account when interpreting these results is the issue of comparing a model grid-box average forecast with a point forecast such as MOS. In the case of this point-based evaluation, the results may favor MOS. Nonetheless, the preliminary results are encouraging when considering the dynamic nature of NWP models today. The significant time lag between a new or changed NWP model and the availability of the associated MOS product is typically a few years. Given this perspective, the bias-corrected forecasts for maximum temperature compare favorably with both the GFS and Eta MOS. Although the bias removal method for minimum temperatures improves upon the direct model output, the technique does not provide desirable results when compared to MOS.

3.3 Model Consensus Forecasts

The preliminary results for the consensus forecasts are promising for the maximum temperatures. Both the 0000 and 1200 UTC consensus forecasts valid at Day+1 and Day+2 have lower MAE values than the bias-corrected and MOS forecasts. However, the minimum temperature consensus forecasts show no improvement over the individual forecasts. In some cases, the consensus minimum temperatures exhibit higher MAE values than both the bias-corrected and MOS forecasts. Refer to Table 2 for the average MAE value comparisons.

4. Summary

An approach for improving maximum and minimum temperature forecasts is being explored using the capabilities of the GFE. The method is based on a simple bias removal technique and a weighted consensus forecast from multiple guidance sources. For a 3 month period from May through July 2004, preliminary results suggest that significant improvements to the GFS and Eta direct model output can be achieved using a simple

7-day running bias correction. Further, results show that a weighted consensus forecast is on average as good as or better than a single guidance forecast for maximum temperatures. Little or no improvement in accuracy is achieved for consensus forecasts of minimum temperatures.

Although an error feedback and blending system is not a new concept or practice, more work needs to be done before conclusions on the usefulness of this approach can be made. However, the study points out the feasibility of the bias removal method to improve upon first-guess grids of maximum and minimum temperatures in the GFE. The study also demonstrates the potential flexibility of the design to adapt to model and algorithm changes as well as handling the addition of new models and/or an improved "Record of Analysis".

Future work will focus on adding the RUC model to the ensemble of forecasts and expanding the number of core elements to temperature, dewpoint temperature, sky cover, and probability of precipitation. A more sophisticated method of determining the weights of each model for the consensus forecasts will be explored. Work on calibrating and validating the reliability of the probabilistic forecasts derived from the time-lagged multiple guidance sources is also planned.

5. Acknowledgments

Special thanks to Pamela Elslager for her work in comparing forecast temperatures from *The Weather Channel's* automated DiCast system to forecasts made by the Tucson NWS office. Her in-house verification work of the DiCast (a Dynamic MOS and Blending System) provided the justification for this initial study. Also, thanks to Brian Francis, Erik Pytlak, and Glen Sampson for their comments and review of this paper.

6. References

Barker, T., 2004: Match guidance version 1.3. [Available online from http:// www.mdl.nws.noaa.gov/~applications/STR/data/ 798/MatchGuidance_1.3.doc]

_____, 2004: MatchObsAll version .97. [Available online from http:// www.mdl.nws.noaa.gov/~applications/STR/data/754/MatchObsAll097.doc]

Etherton, B.J., 2004: Model consensus and ensemble weighting for spot forecasts. [Available online from http://ams.confex.com/ams/pdfpapers/68170.pdf]

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.

LeFebvre, T.J., 1995: Operational forecasting with AFPS. *Preprints 11th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Dallas, Amer. Meteor. Soc., 249-254.

_____, T.J., M.B. Romberg, T. Hansen, 2002: Initializing gridded fields from numerical models. *Preprints Interactive Symposium on AWIPS*, Orlando, Amer. Meteor. Soc., 41-45.

Mao, Q., R.T. McNider, S.F. Mueller, and H-M. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting*, **14,** 190-202.

Mass, C.F, 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting*, **18**, 75-79.

Steed, R.C., and C.F. Mass, 2004: Bias removal on a mesoscale forecast grid. [Available online from http://www.mmm.ucar.edu/mm5/workshop/ws04/Session2/ Steed.Rick.pdf]

Stensrud, D.J., and J.A. Skindlov, 1996: Gridpoint predictions of high temperature from a Mesoscale model. *Wea. Forecasting*, **11**, 103-110.

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.

Verret, R., and N. Yacowar, 1989: Improvement of numerical weather element forecasts by combining forecasts from different procedures. *Preprints, 11th Conf. on Probability and Statistics*, Monterey, CA, Amer Meteor. Soc., 58-63.

Vislocky, R.L., and J.M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. Bull. *Amer. Meteor. Soc*.**, 76**, 1157-1164.

Wilson, L.J., and M. Vallèe, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206-222.

7. Figures

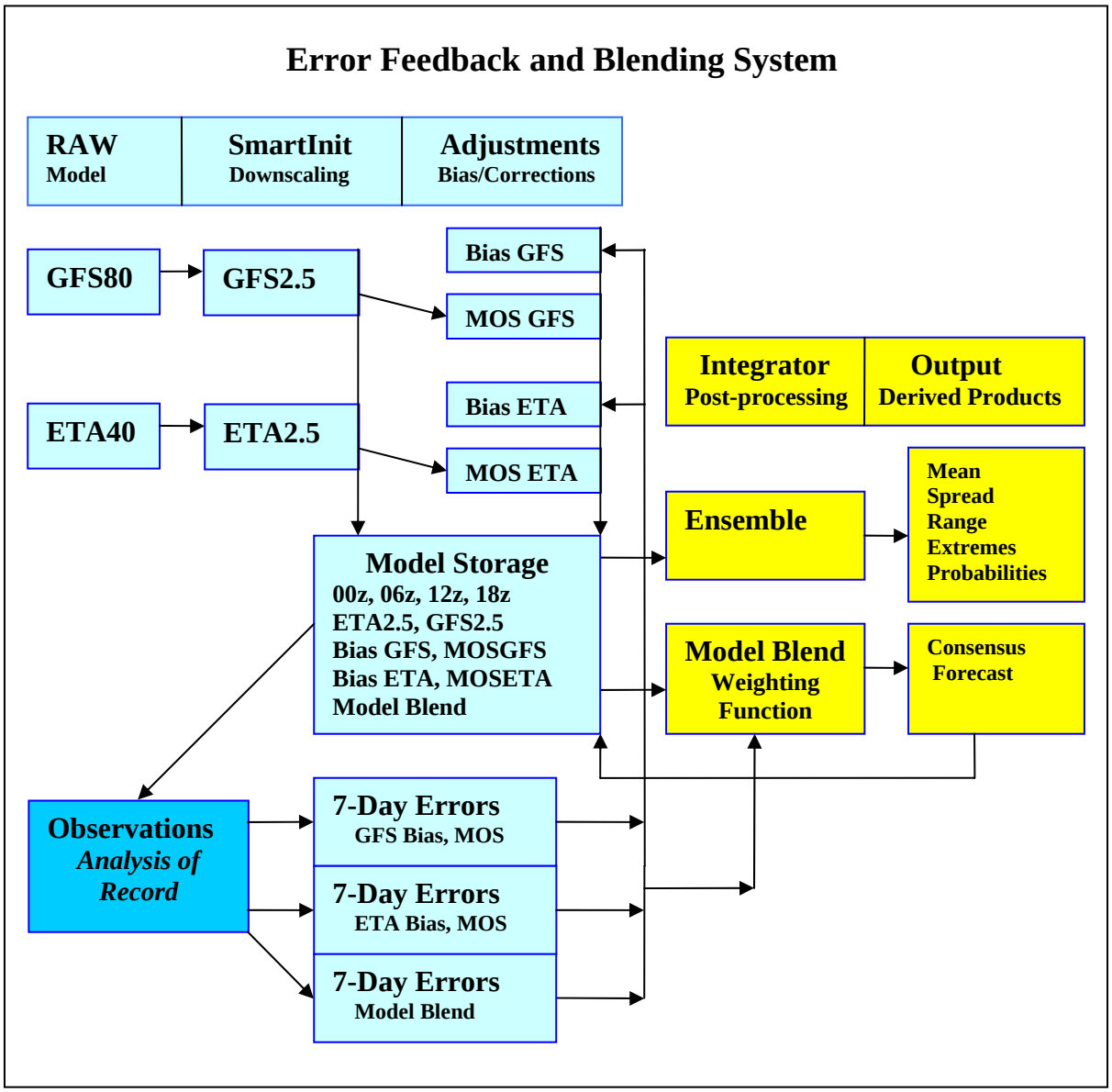

Figure 1. Schematic showing the modified GFE configuration for the error feedback and blending methods used in this study.

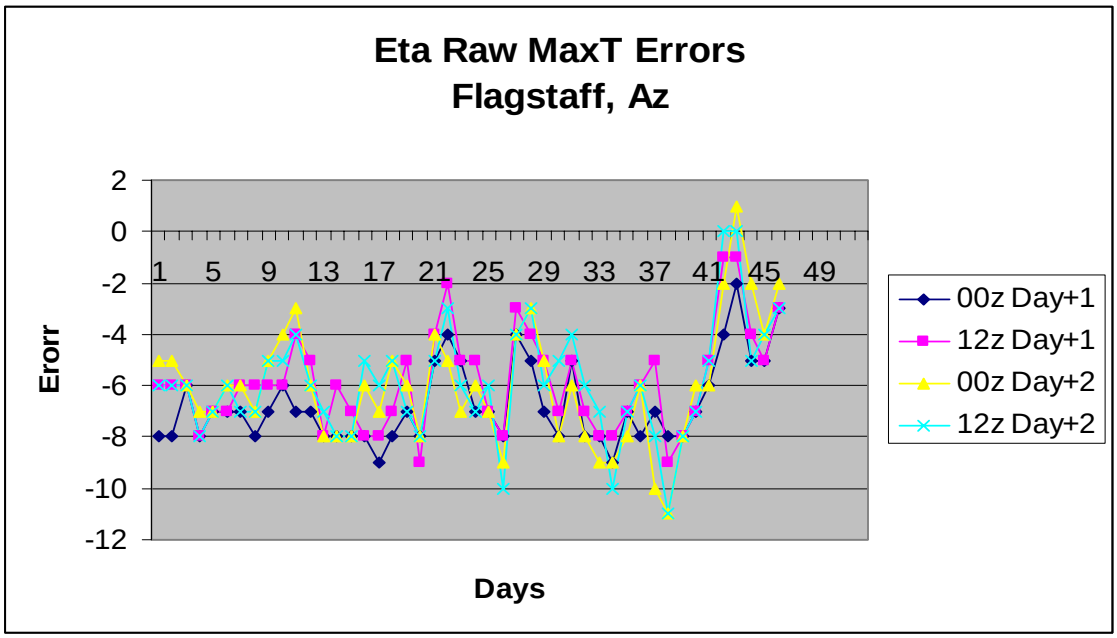

Figure 2. Shows the magnitude of change in forecast error by model cycle and projection time at Flagstaff, Arizona for MaxT.

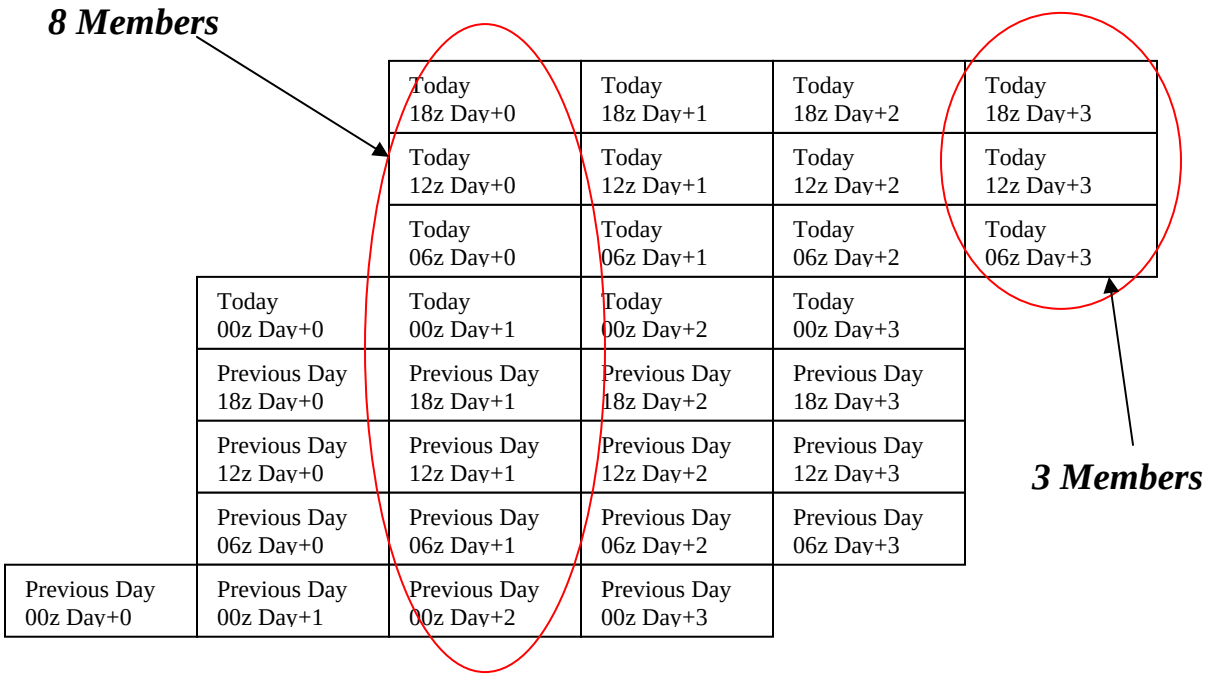

Figure 3. Illustrates the lagged model grids used in the consensus forecasts. In the current configuration for same day forecasts (Day+0) at certain model cycles, over 70 members can be available for blending.

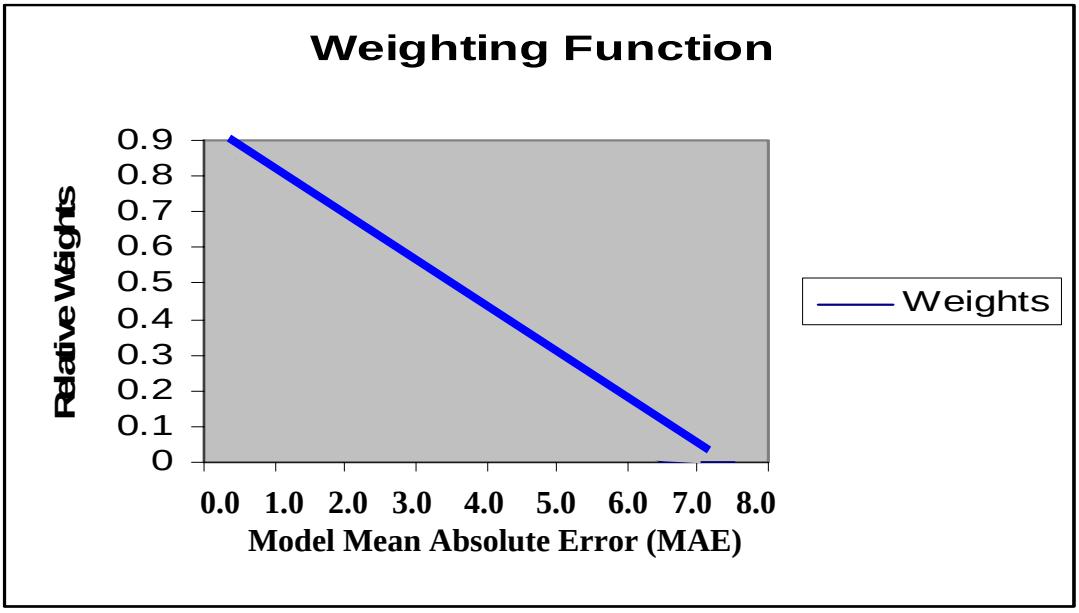

Figure 4. Shows the relative weights given to each model as a function of a 7-Day running MAE.

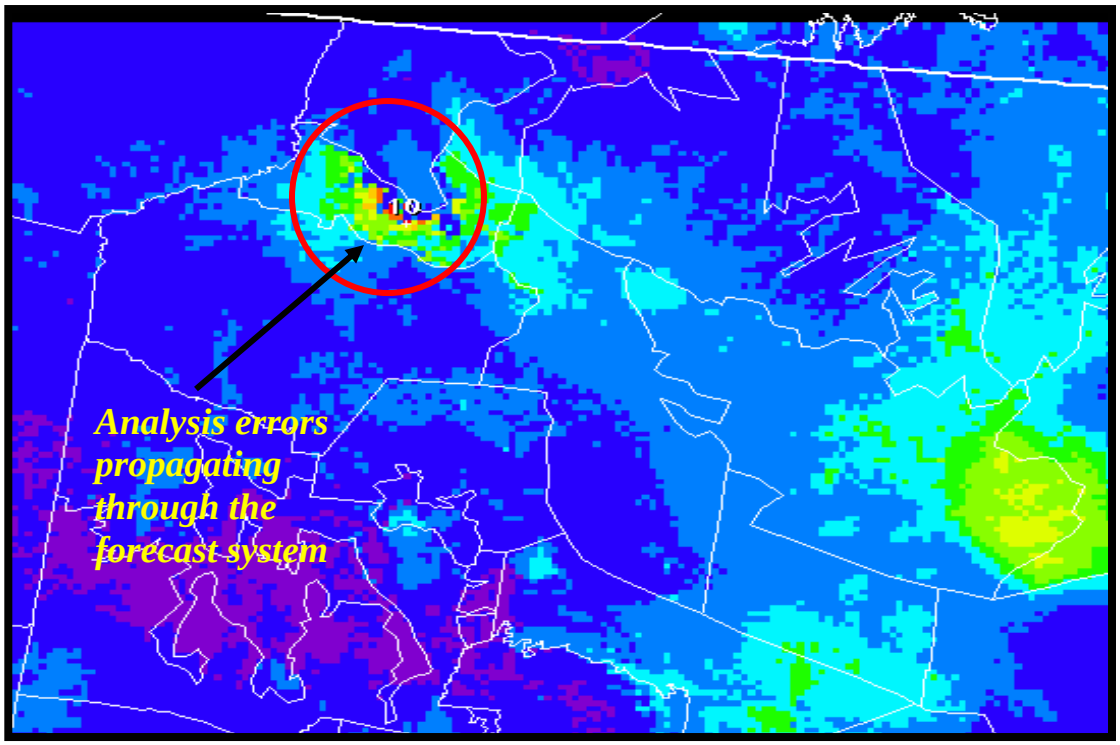

Figure 5. Shows an area of large spread about the ensemble mean in the Grand Canyon region of northern Arizona. The large spread is mainly tied to bad observed data in the objective analysis in which model biases are derived. These analysis errors can propagate through the forecast system.

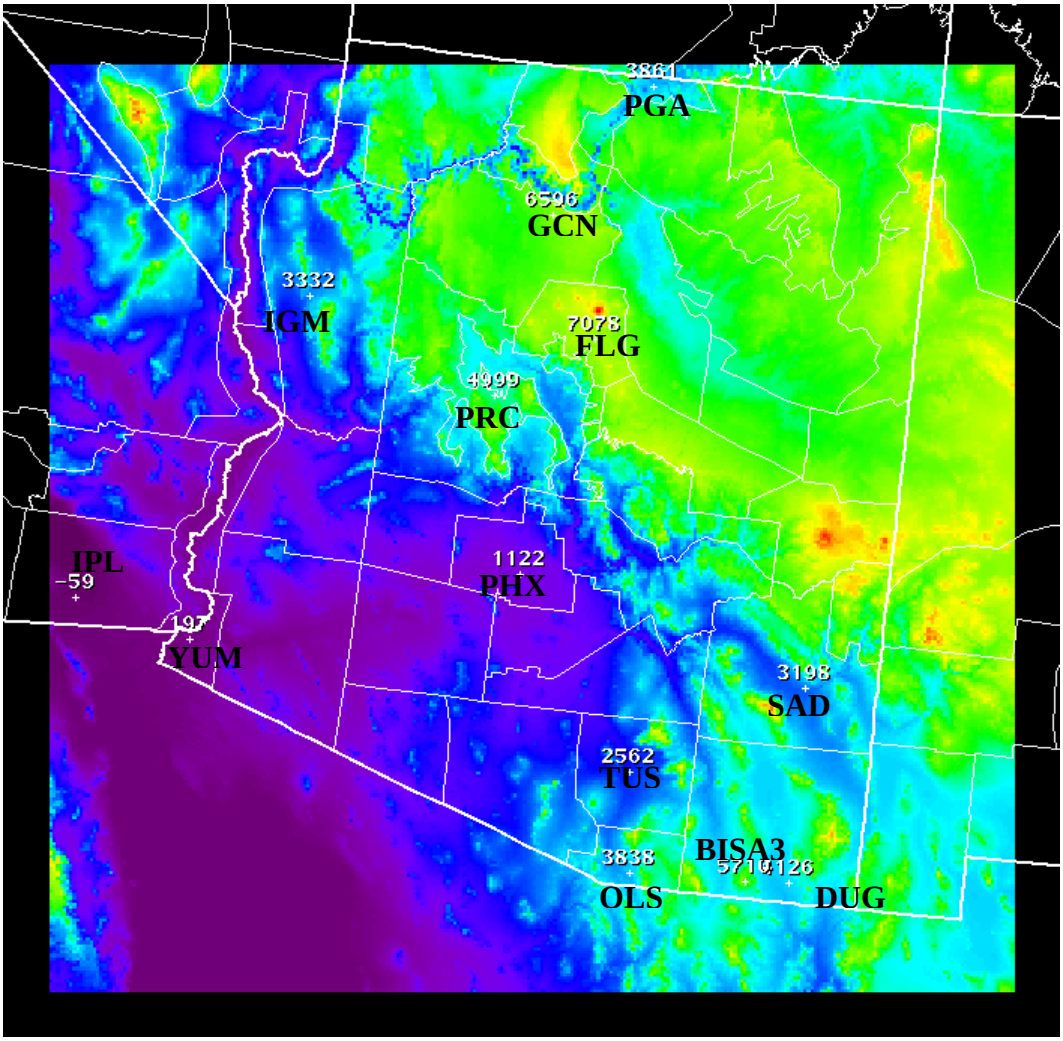

Figure 6. Locations of the 13 sites used in the validation.

8. Tables

Element	Cycle (UTC)	Lead (Day+)	GFS RAW	GFS Bias	% Imprv	Eta RAW	Eta Bias	% Imprv
Max T	00	1	7.40	2.45	67	4.64	1.98	57
Max T	00	2	6.41	2.02	68	3.51	2.22	37
Max T	12	1	6.41	1.98	69	3.48	2.15	38
Max T	12	2	6.08	2.35	61	3.59	2.33	35
Min T	00	1	12.82	4.84	62	9.82	3.27	67
Min T	00	2	11.67	4.49	62	10.05	3.57	64
Min T	12	1	12.92	5.25	59	9.81	3.36	66
Min T	12	2	13.22	5.06	62	10.28	3.54	66

Table 1. MAE values in degrees F. Percentage of improvement (%Imprv) for the bias-corrected over the raw model output. The Bias label is for the bias-corrected forecasts and RAW label represents the downscaled direct model output.

Element	Cycle (UTC)	Lead (Day+)	GFS Bias-COR	GFS MOS	Eta Bias-COR	Eta MOS	Model Consensus
Max T	00	1	2.45	2.11	1.98	1.94	1.87
Max T	00	2	2.02	2.15	2.22	2.04	1.82
Max T	12	1	1.98	2.17	2.15	1.97	1.78
Max T	12	2	2.35	2.46	2.33	2.12	2.10
Min T	00	1	4.84	2.77	3.27	2.74	3.32
Min T	00	2	4.49	3.16	3.57	2.97	3.78
Min T	12	1	5.25	3.03	3.36	2.51	3.16
Min T	12	2	5.06	2.98	3.54	2.84	3.52

Table 2. MAE values in degrees F for forecasts averaged from all 13 stations studied. The label Bias-COR represents the bias-corrected forecast.