Some Observations of Temperature Forecasts Made using BOIVerify

William Martin
National Weather Service Office
Glasgow, Montana

1. Introduction

For various reasons, numerical model forecasts of surface temperatures are inaccurate, though they are typically “in the ball park”. To get more value out of such model forecasts, it is natural to try to correct such forecasts for bias. For example, if it is found that morning low forecasts are typically several degrees too low, then this bias can be corrected for, generating more accurate final output. With more experience, one might find that the bias is also a function of wind speed and cloud-cover, leading to a more complicated bias correction scheme. Model Output Statistics (MOS) methods generalize this process in a moderately sophisticated manner. In MOS (e.g., Carroll, 2005), a desired forecast quantity like temperature is linearly regressed against a number of model-forecast quantities such as 2-meter temperature, low-level thickness, relative humidity, and against some non-model parameters such as the sine of the day of year and observed temperature.

The current MOS package for the GFS model is segmented into two seasons (Oct- Mar and Apr-Sep) and used observations and model runs from 7 years (1997-2003) to develop the best-fit regression equations. For each station for which equations were developed, approximately 1200 observing days were used, producing statistically stable regression equations. A couple disadvantages of the MOS approach are the considerable number of model runs required to develop the equations, and the need to redevelop the equations if the model is changed in any significant way. MOS has proven to be very valuable to forecasters, and the desirability of implementing any change to an operational model so as to improve its forecasts, needs to be balanced against the harm that may be done to the MOS if the equations are not redeveloped for the revised model.

As useful as MOS has proven to be, forecasters still notice that the statistically unbiased MOS sometimes appears to be biased over short periods of time. For example, the previous week’s MOS temperatures might be noticed to be consistently warmer than the observations. This leads to the concept of short term bias-correction (BC) in which an attempt is made to remove this short term bias.

Obviously, if MOS is truly unbiased (which would be the case if the training data—the 1997-2003 years for the GFS MOS—were representative) then BC must sometimes fail. If MOS is too warm in some weather regimes, then it must be too cool in others, in order for the overall MOS to be unbiased. The reason MOS sometimes shows short-term bias is not clear, but is generally thought to be situation dependent. Weather is very complicated, and even though a large number of parameters are offered to the regression system for generating the MOS equations, these parameters probably can not account for every persistent situation that may actually develop.

BC must fail sometimes, and it is believed that it probably fails when there is a regime change in the weather. For example, in a period of calm and sunny weather MOS might be found to be persistently too low. Correcting this bias would produce superior forecasts as long as this regime persisted. If the weather changed to, say, overcast and windy, the MOS bias could be in the opposite direction. Using BC the day after a regime change would produce an inferior forecast. The solution to this problem is to be selective and to only use BC for forecasts of weather regimes similar to those for which the bias was found. The potential problem with this solution is that it is not necessarily clear if a regime change has occurred. The weather is different in some way almost every day. Since it is not known why MOS is biased the way it is in any given regime, it is not known what would need to change to cause its bias to change. Perhaps the wise approach is to definitely reject BC when there is a large change in the weather (however that can be decided) and to use it cautiously when the weather seems to be stable.

The BOIVerify verification program (Barker, 2007) has been developed by the Western Region of the National Weather Service. This program automatically archives forecast grids from various models and human generated forecasts as well as analysis grids for periods typically up to 50 days in the past. The bias of any of the forecast grids can be calculated and examined using the program. Also, the program routinely calculates BC grids based on the previous 30 days, and these grids are made available for use by forecasters. The 30 day learning period was chosen so that a large enough number of forecast-verification pairs could be used to produce a stable, low-noise BC grid (compare this with the 1200 points used by GFS MOS). An alternative approach currently used by NCEP for the SREF is to weight recent biases more heavily. This allows a sufficiently large number of data pairs for noise reduction, while producing bias more consistent with the most recent regime.

In what follows, we will look at several examples of using BOIVerify to compare the official forecast errors with that from raw model, MOS, and BC grids.

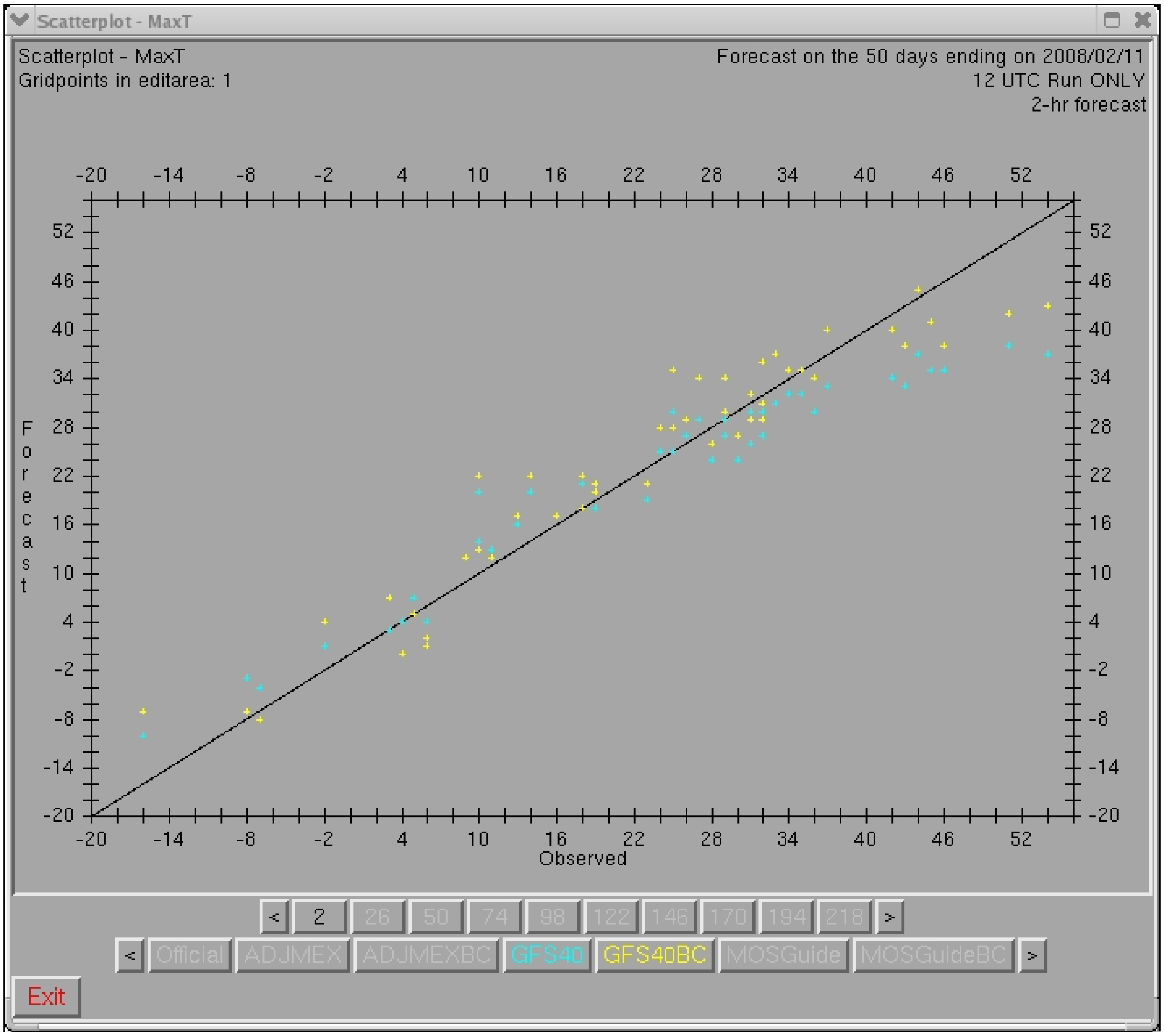
2. Non-linear bias for extreme forecasts

Figure 1 below shows a subtle, though persistent feature of non-linear bias that neither MOS nor BC can account for. Figure 1 is a scatter plot of 50 nominal 2-hr forecasts of maximum temperature for the Glasgow, Montana airport from the raw GFS model and the BC GFS model. It is possible to run such scatter plots for all the 15334 grid points in the forecast area; however, the verification grid is based on a relative few observations at specific points with the rest of the values in the grid being analyzed from these observations, and we chose to verify against an actual observation point.

The non-linear feature of interest in Fig. 1 is the flattening of the scatter plot at both high and low observed temperatures. The observed data included a period in early January 2008 in which record high temperatures were observed (in the mid-50s F). The forecasts from both the model and the BC model were, in fact, well above climatology (low 40s F versus climatology of 20 F). Similarly, at the extreme cold end of the plot, the forecast of -10 F was short of the observed -16 F). Most of the forecasts are generally unbiased, except at the extremes.

As both BC and MOS are linear processes, they can not eliminate this kind of bias. Linear bias correction can take a scatter plot like Fig. 1 and rotate the points and translate the points so that they are, on average, closer to the line of perfect agreement. The forecasts at the extreme ends of Fig. 1 can not be improved by linear transformation without harming the other forecasts.

However, we have seen on other occasions this flattening of the verification curve for extreme events. If human forecasters realize that this happens, they can improve their forecasts non-linearly by only adjusting the forecasts of extreme events. Typically, if the model is forecasting temperatures 20 degrees off of climatology, there is a good chance it is under forecasting the anomaly.



Scatterplot - MaxT
Gridpoints in editarea: 1

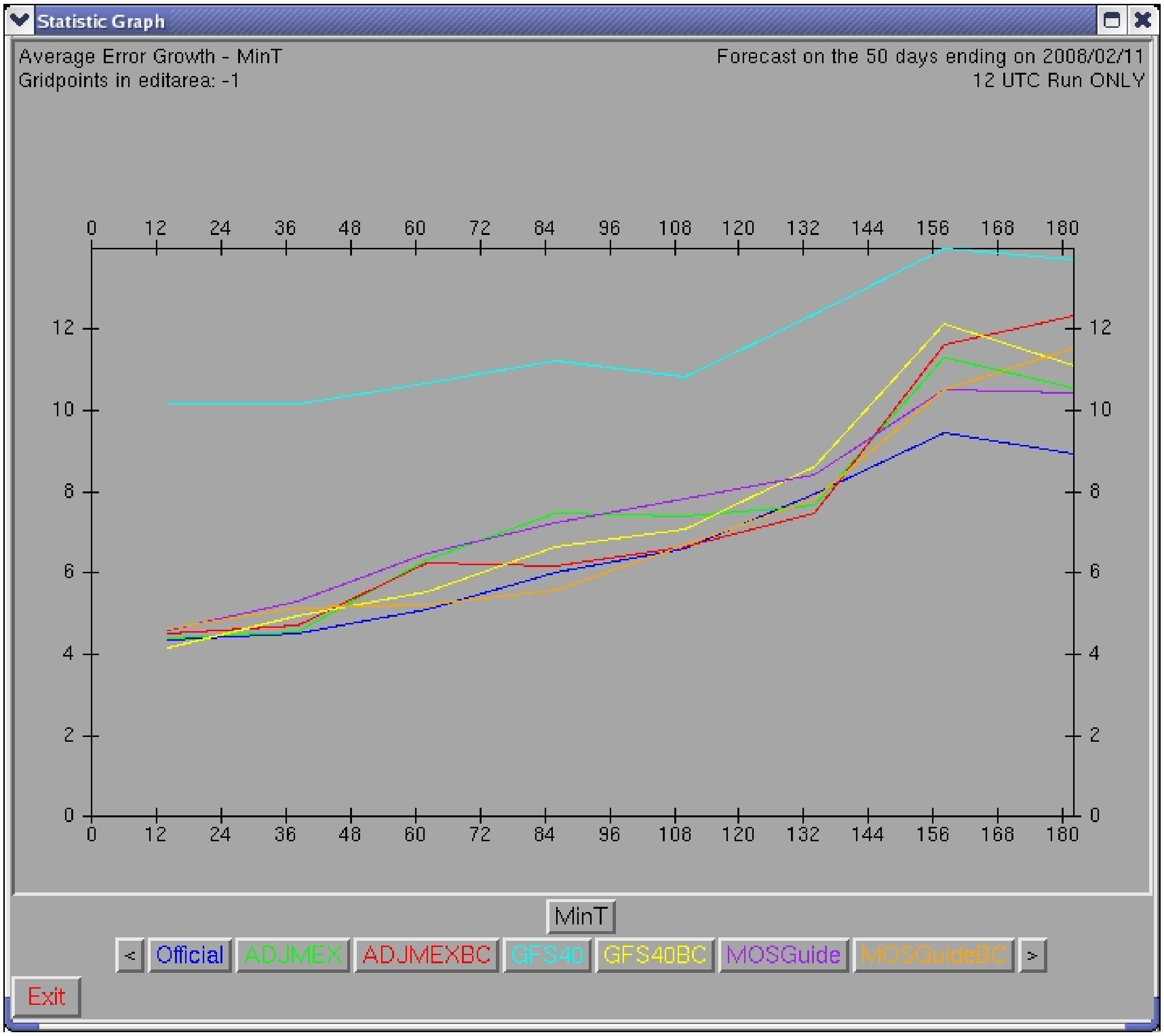
Forecast on the 50 days ending on 2008/02/11
12 UTC Run ONLY
2-hr forecast

Figure 1. Scatterplot of 2-hr forecast versus observed maximum temperatures from the GFS model (blue) and BC GFS model (yellow) from the 50 model runs ending 2/11/2008. The forecast point for verification was the Glasgow airport (KGGW).

3. Superiority of human forecasts at long range

Figure 2 plots the average error in minimum temperature forecast versus forecast lead time over 50 days for the KGGW location. The dark blue line is the official forecast and the light blue line is the raw GFS forecast. All the other lines are various corrected forecasts derived from the GFS model. An outstanding feature of this plot is the poor performance of the raw GFS model, which has errors typically 6 degrees F worse than any other forecast. Another feature, and one we have seen fairly consistently, is the superiority of the human official forecast to all others at the longest lead times (144-180 hours). The reason for this is not at all obvious. Humans are almost entirely dependent on models at such long forecast lead times, at it is not clear how they could be superior to models. After conversations with forecasters in the office, we suspect that their skill is derived from use of the GFS ensemble.

It is also note-worthy that the BC gridded MOS forecasts (MOSGuideBC, orange line) arguably had the best forecast in the 72-120 hour range, though the improvement is slight relative to the official or other corrected models. However, we note that this plot is for all forecasts in a 50 day period. Over a long period of time, BC is expected to be no better than MOS for reasons stated in the introduction. Selective use of BC can't be reflected in a plot such as this, except to the extent that human forecasters may be selectively using BC to improve their forecasts. Presumably, if regime selectivity can be achieved, BC grids would have better forecasts than that indicated by Fig. 2.


Average Error Growth - MinT
Gridpoints in editarea: -1

Forecast on the 50 days ending on 2008/02/11
12 UTC Run ONLY

Figure 2. Average error of minimum temperature forecast at KGGW versus forecast lead time for the 50 12 UTC forecasts ending on 2/11/2008 for the human forecast (blue), the raw GFS model (light blue), and various MOS, and BC grids.

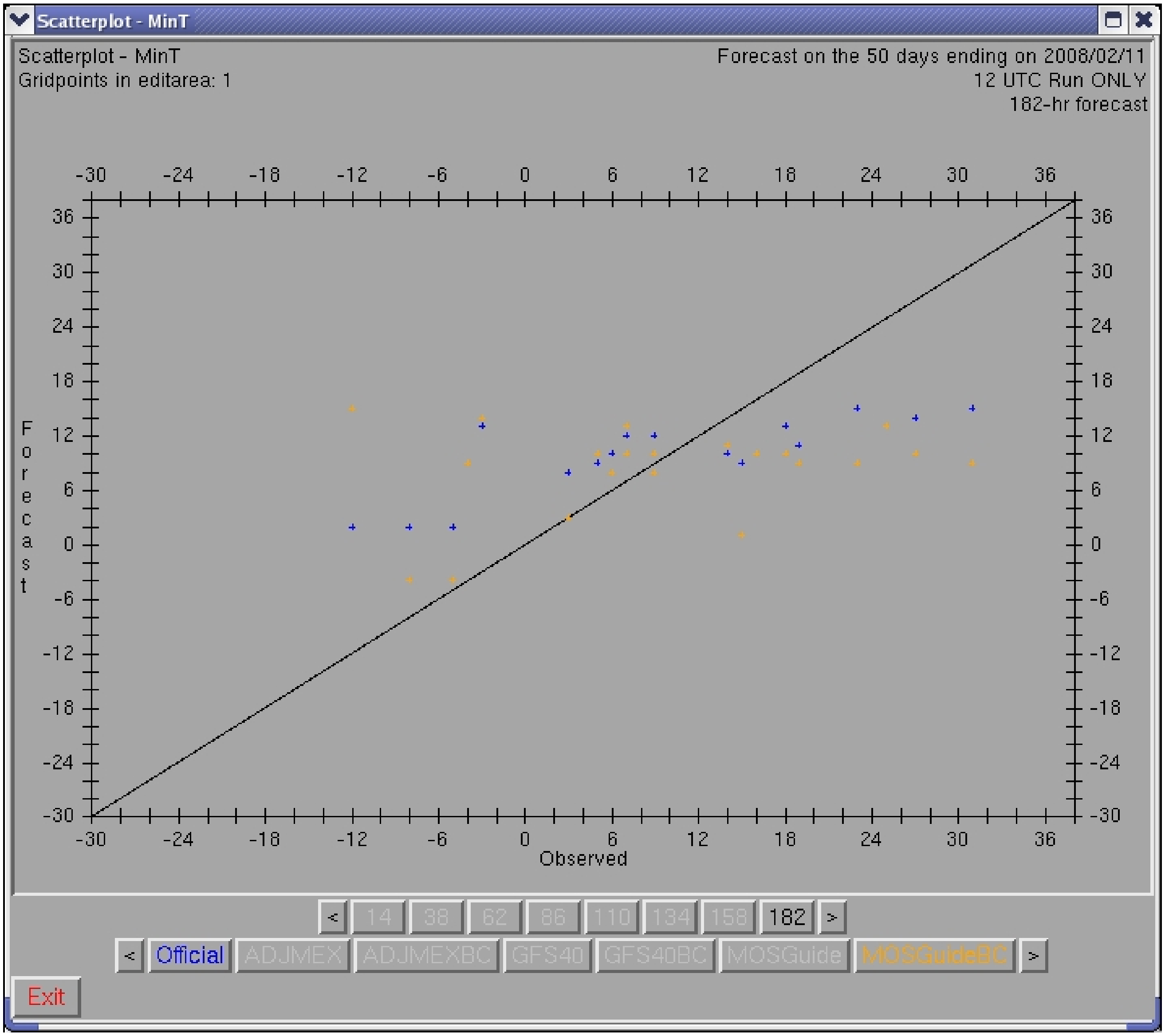
4. Trend towards climatology at long leads and the degradation of signal to noise ratio

Figure 3 shows a scatter plot of 16 forecasts of minimum temperature at station KGGW versus the observed minimum. The forecasts were made over a period of 50 days ending 2/11/2008 (for the 16 days for which both the official and gridded MOS were available). The forecasts have a 182 hour lead time and are from the Official grids and from the BC gridded MOS. This plot shows the typical flattening of the entire curve at long lead times (compare with Fig. 1). The climatological low this time of year is about 2 F, and at long lead times, models and humans tend towards this value. Model inaccuracy at long lead times is the reason the sine of the day of the year (which climatology is linearly related to) is used as one of the predictors for MOS equations. Models are so inaccurate at long lead times that climatology is one of the better predictors.

In Fig. 3, the MOSGuideBC grids show virtually no skill with the points scattered about a horizontal line, while the Official forecast has a small slope to it and a little skill. This is consistent with the superior performance of Official forecasts seen in Fig. 2.

One wonders, since BC can rotate points so as to put the line they are scattered about coincident with the line of perfect agreement, why not rotate the MOSGuideBC points of Fig. 3 about 45 degrees counter-clockwise, thus producing a dramatically better forecast. In fact, since the trend towards climatology is a very consistent feature of long-lead forecasts, one can wonder why BC hasn't already done this. A moment's thought will demonstrate why this doesn't work. The scatter plot for MOSGuideBC of Fig. 3 is basically too noisy for rotation to work. An additional BC rotation would be applied to each point as a change in slope parameter. Warm values would be made warmer and cool values would be made cooler. As all the values are basically the same plus noise, all that is accomplished is an amplification of noise. For example, in Fig. 3, the MOSGuideBC point with a forecast of 15 F versus a verification of -12 F is a warm forecast. Rotation would make this warm forecast even warmer and more inaccurate. Rotation only works if the slope of the points (the signal) is larger than the noise. This is the case for the Official forecasts of Fig. 3.

For the Official forecast, because there is some identifiable slope to the scatter plot, a little rotation might bring some improvement. This suggests that humans could do even a little bit better at long lead times by increasing their deviation from climatology.



Scatterplot - MinT
Gridpoints in editarea: 1

Forecast on the 50 days ending on 2008/02/11
12 UTC Run ONLY
182-hr forecast

Figure 3. Minimum temperature forecast at 182 hours lead time from the Official (blue) and BC gridded MOS (orange) for 16 forecasts made over 50 days ending 2/11/2008 versus observations at station KGGW.

5. Summary

This report has discussed several observations made using BOIVerify to analyze the accuracy of Official and BC grids:

--A non-linear effect that can’t be corrected by MOS or BC has been seen whereby model forecasts of large anomalies are under forecast.

--Humans consistently do better at long forecast leads than MOS or BC grids. The reasons for this are not clear, but could be do to the use of the GFS ensemble.

--Plots of error's growth (as Fig. 2) of BC forecasts understate the potential of BC because no regime selectivity is employed. BC should not improve on MOS unless there is selective use of it.

--At long lead times, both model and human forecasts trend towards climatology and become noisier. The decrease in signal-to-noise ratio renders BC less effective at these long forecasts.

Statistical correction can't fix a bad model forecast; it can only fine-tune a good forecast. Ultimately, models may improve to the point that statistical correction is not necessary. Any model error that is consistent is traceable to a systematic problem with the model that could be corrected. However, the great complexity of weather and the slow but steady progress that is being made in modeling, suggest that statistical correction will be around for awhile. Bias correction, in particular, shows promise, and formalizes what forecasters have probably been doing intuitively since models began. A critical issue for BC that probably needs more research is the identification of when to use it and when not to use it.

6. REFERENCES

Barker, Tim, 2007: *BOIVerify version 2.0: SmartTools, Procedures, and Scripts for maintaining a gridded verification database and displaying various statistics*. WFO Boise, ID

Carroll, Kevin L., 2005: GFS-based MOS temperature and dewpoint guidance for the United States, Puerto Rico, and the U.S. Virgin Islands. *MDL Technical Procedures Bulletin* No. 05-05, NOAA, U.S. Dept. of Commerce, 8 pp.