Verification of Model Output Statistics forecasts associated with the North American Mesoscale model upgrade

Results of work done during the Student Career Experience Program tour with the Statistical Modeling Branch (SMB) of the Meteorological Development Lab (MDL) of the National Weather Service (NWS), Silver Spring, MD, under the guidance of Mark Antolik & Kathy Gilbert, NWS/MDL/SMB

I. Introduction and Motivation

For the past four decades, the model output statistics (MOS) technique has been in operation and forecasters and meteorologists globally have used forecasts generated through MOS. MOS is a statistical approach to forecasting which involves determination of a relationship between a predictand and numerical weather prediction (NWP) model output for a variety of variables (predictors) at the same projection times (Glahn and Lowry 1972). Through the years, MOS has provided reliable forecasts, predicting quantities that the underlying NWP model is unable to predict. Specifically, the MOS technique can account for local effects on certain elements that the coarser resolution NWP model cannot detect (Antolik and Baker 2009).

Usually, statistical forecasting systems, such as MOS, rely on many years of stable development data for optimal performance. However, due to frequent changes in the underlying structure, physics, and dynamics of the NWP models, this kind of data is difficult to obtain. Fortunately, developers at the National Weather Service's (NWS) Meteorological Development Laboratory (MDL) have been able to show that reliable MOS forecasts can be produced using as little as just two years of data from models which are evolving (Antolik and Baker 2009).

Due to the possible degradation in the accuracy of MOS that may result from NWP model upgrades, it is beneficial to perform a "parallel verification," in which an old model's MOS equations are applied to the new model's output, and forecasts using these data are produced. These new forecasts are compared to the old forecasts and this process can show how well the existing MOS system would perform if the current MOS equations were kept in place. From these tests, a decision to upgrade the equations for all or even just some of the MOS elements can be made.

In September 2011, the National Centers for Environmental Prediction (NCEP) made an upgrade to the current operational North American Mesoscale (NAM) forecasting system. Specifically, upgrades were made to both the physics and dynamics packages of the Nonhydrostatic Multi-scale Model (NMM) portion of the NAM. These changes to the model, now called NMM-b, transitioned the formerly Weather Research Forecast (WRF)-based NMM, to the new NOAA Environmental Model System (NEMS)-based NMM-b. Just like its predecessor, the NMM-b runs four times per day (every 6 hours) and has an overall 12 km horizontal grid spacing domain with forecasts made out to 84 hours. However, the new NMM-b has four additional fixed nests, or sub-grids, within the domain that run out to 60 hours each: a 4km resolution conterminous United States (CONUS) nest, a 6km Alaska nest, and 3km Hawaii and Puerto Rico nests. Additionally, there is also a new single moveable nest run out to 36 hours for fire weather forecast support (DiMego and Rogers 2011).

The above changes in the NAM model provide the primary motivation for this project. This paper discusses how the current NAM MOS system would perform given the NMM-b implementation using a variety of statistical skill and accuracy scores. The tested elements, scores, results, and conclusions are described in the following sections.

II. Data and Methods

For this project, forecasts from the previously operational NAM MOS system and forecasts from those equations applied to the parallel NMM-b model output for the same time period were compared. Additionally, in the past, since the NAM has exhibited a marginal forecast advantage over the GFS since its inception, the NAM/NMM-b verifications were also compared to the GFS MOS to examine if this advantage would still hold after the NMM-b implementation. Operational and parallel forecasts for 2m temperature and dew point, local maximum and minimum, wind direction, wind speed, and 6 and 12 hour probability of precipitation (POP) were verified. These elements were chosen because past verifications have shown them to be most susceptible to underlying model changes. All parameters were evaluated at 335 sites (Figure A-1) spanning CONUS (300 sites), Alaska (30 sites), and Hawaii and Puerto Rico (5 sites). To further evaluate the regional effects of warm and cool season temperature and cool season dew point, the CONUS was broken up into 6 regions: Northeast, Southeast, North Central, South Central, Northwest, and Southwest, comprising of 50 stations each. The verifications were performed for two 6-month seasons: warm (April 1 - September 30, 2010), and cool (October 1, 2010 - March 31, 2011).

The 2m temperature, dew point, wind speed and direction, and 6 and 12 hour POP forecasts were verified out to 84 hours, at 3 hour time intervals. Local maximum temperature was verified at 30, 54, and 78 hours, while local minimum temperature was verified at 42, 66, and 90 hours. For each parameter, four plots were created: CONUS, Alaska, Hawaii/Puerto Rico, and overall results from all sites. For warm and cool season 2m temperatures and cool season dew point, six additional graphs – one for each CONUS region – were generated.

To assess the accuracy of the NAM MOS equations as applied to the NMM-b output, we used a compilation of several scores. For the 2m temperature, dew point, and maximum and minimum temperatures, accuracy was tested using two scores: mean absolute error (MAE) and the mean algebraic error (or bias). Wind speed accuracy was tested using MAE and the Heidke skill score (HSS) while wind direction accuracy was assessed using MAE and the cumulative relative frequency (CRF) of errors less than 30 degrees. Finally, for the POP categorical forecasts, we calculated the Brier skill score (see Table A-1 for score explanations). All calculations were done using previously developed FORTRAN programs and plots were made using Microsoft Excel.

III. Results and Discussion

In each of the text descriptions and figures of this section, "NAM" represents the previously operational NAM MOS, while "NMM-b" represents the NAM MOS equations as applied to the NMM-b model output. "GFS" refers to the GFS MOS system.

a. *Warm Season*

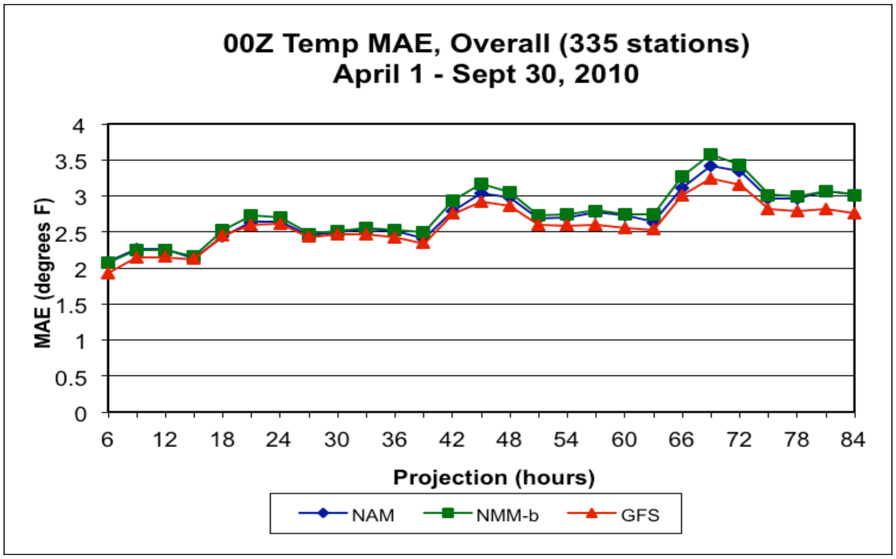

Figure 1: Warm season temperature MAE for all sites

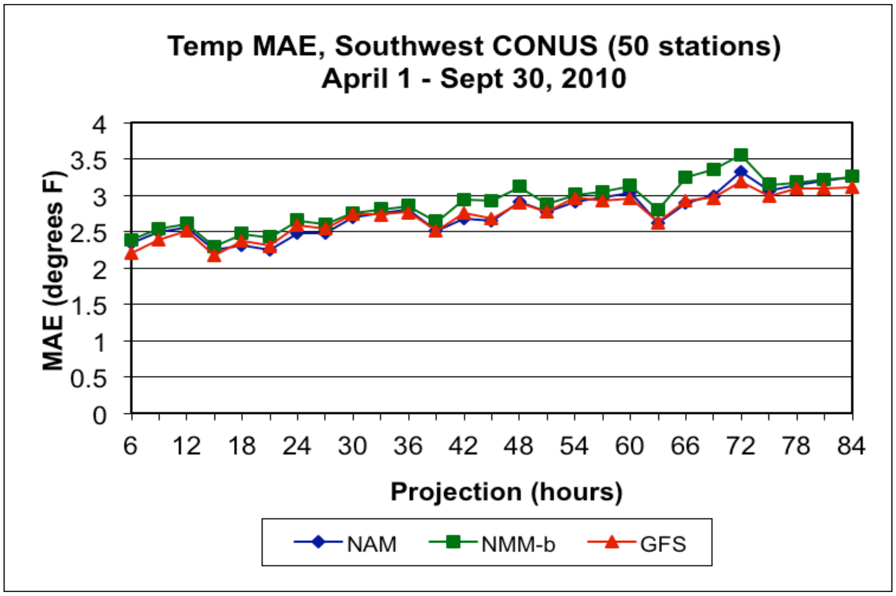

Figure 2: Warm season temperature MAE for southwest CONUS region

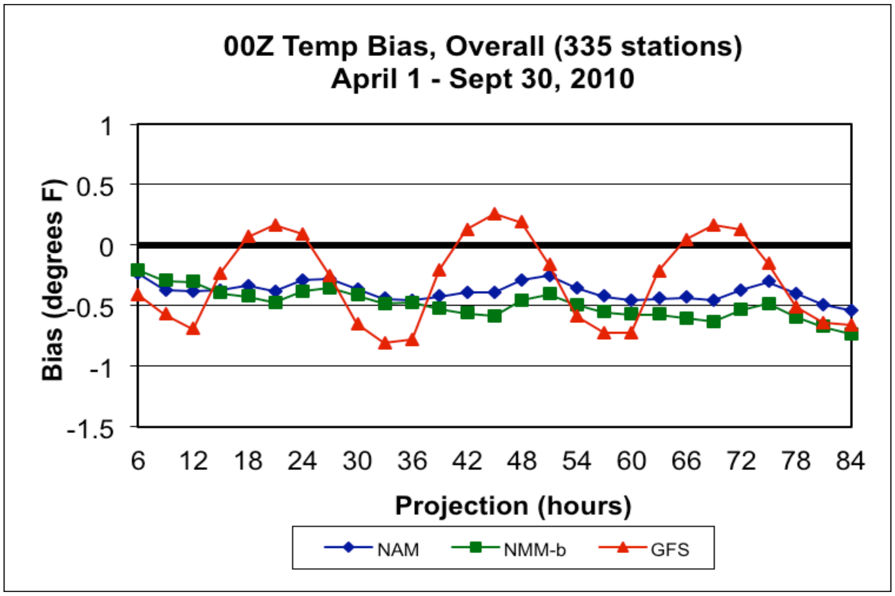

Figure 3: Warm season temperature bias for all sites

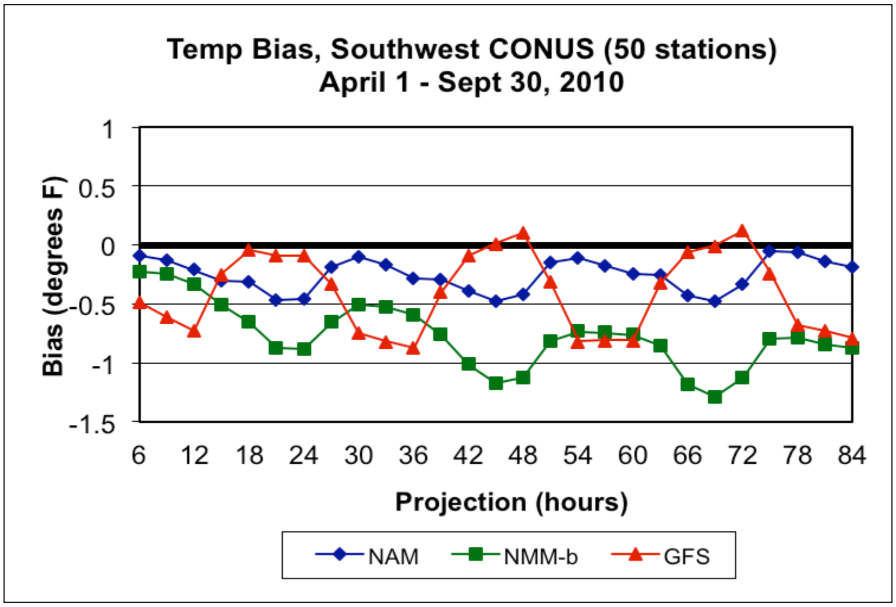

Figure 4: Warm season temperature bias for southwest CONUS region

For warm season temperature, the verifications showed that MAE and bias differences between the NAM and NMM-b systems were overall very similar for all regions. NMM-b MAE was the same or slightly worse than NAM with differences less than 0.25 degrees (Figure 1). The largest difference between the two is seen in the southwest US region, although those differences are still less than 0.5 degrees (Figure 2). The NAM and NMM-b both exhibit a cool bias, with NMM-b performing just a bit worse by about 0.2 degrees (Figure 3). Looking at the temperature bias broken up into the six CONUS regions, we see NMM-b bias is worse than the NAM in the southwest and southeast, while it is better in the northeast and north central regions. In the south central, and northwest regions, NMM-b bias is very similar to NAM. Differences between the two are less than about 1 degree, with the greatest contrast in the southwest (Figure 4). When compared to the GFS MOS system, NAM and NMM-b both have smaller biases and also exhibit a less pronounced diurnal cycle that is slightly out of phase.

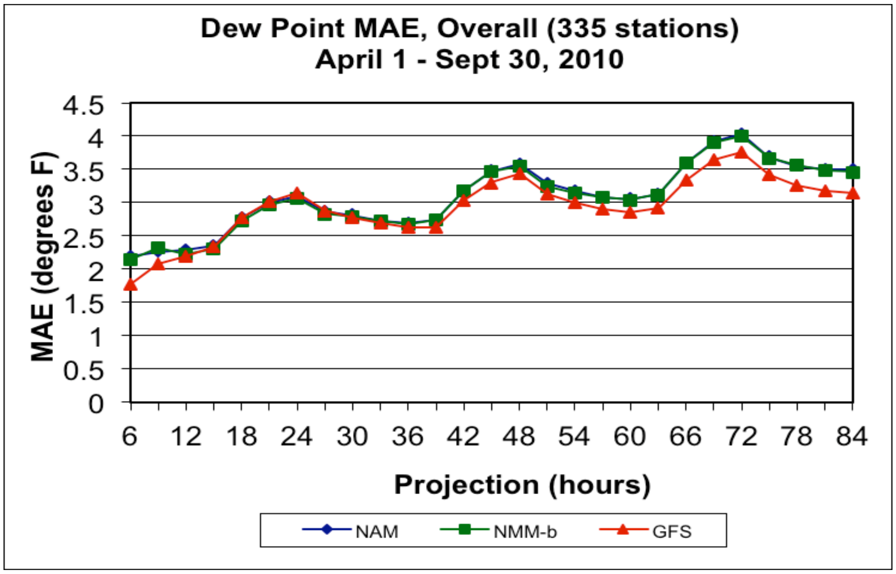

Figure 5: Warm season dew point MAE for all sites

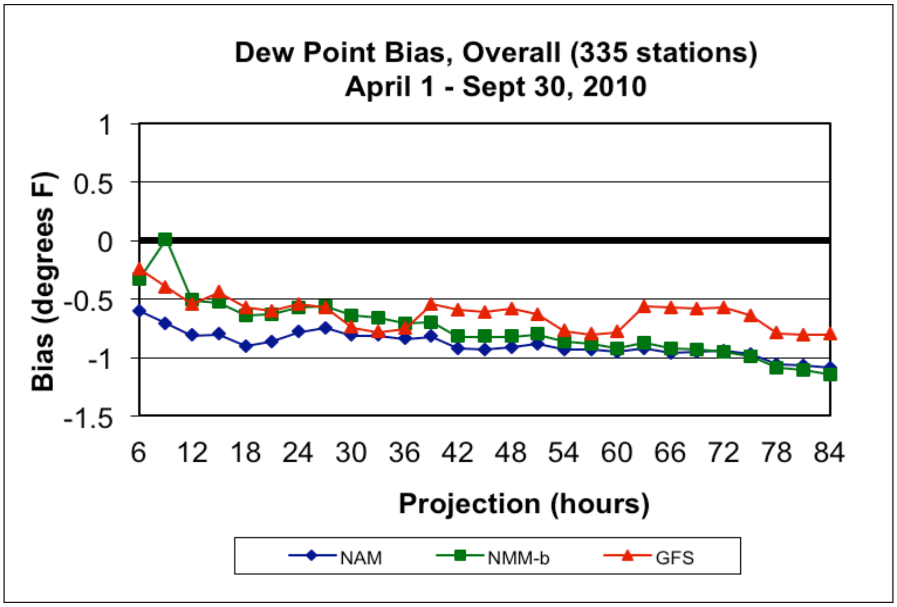

Figure 6: Warm season dew point bias for all sites

Warm season dew point verifications generally show the same trend as temperature. For overall and CONUS regions, NAM and NMM-b MAE are essentially similar with very small differences of less than about 0.2 degrees (Figure 5). All three models show a consistently dry bias with the NMM-b bias just a bit better than NAM (Figure 6). Once again, differences are very small – less than 0.5 degrees. In the Alaska and Hawaii/Puerto Rico regions, however, the NMM-b biases are a bit worse than NAM but only by less than 0.5 degrees. Across all regions, the GFS appears to perform similarly to or a little better than both the NAM and NMM-b.

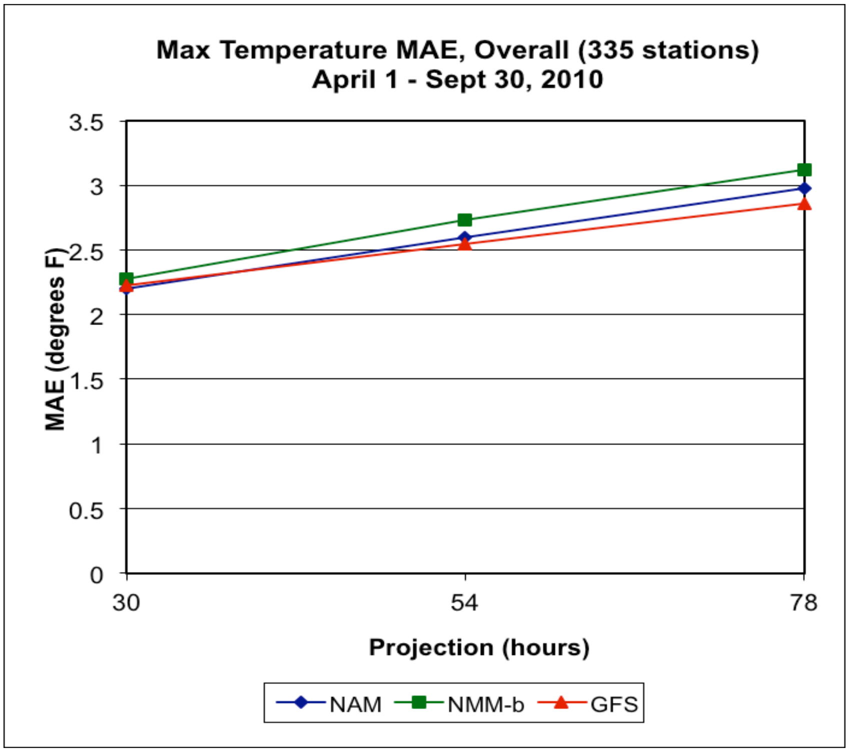

Figure 7: Warm season local maximum temperature MAE for all sites

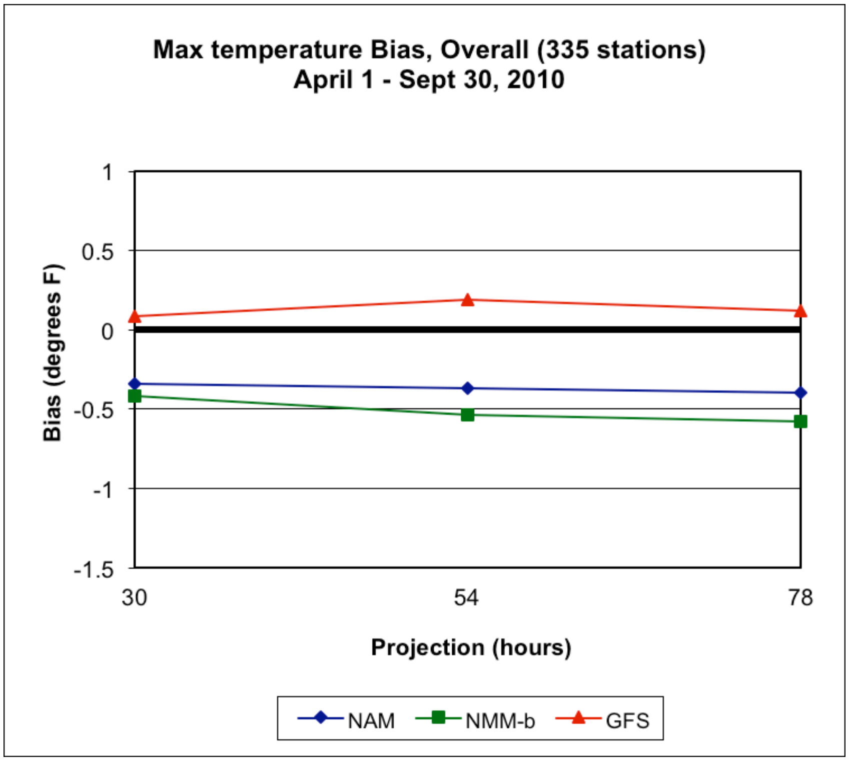

Figure 8: Warm season local maximum temperature bias for all sites

For local maximum and minimum temperatures, the NMM-b MAE is very similar to, or marginally worse than the NAM across all regions with very small differences of less than 0.25 degrees (Figure 7). The NMM-b biases across most regions are cool and either similar to, or slightly worse than the NAM, with differences of about 0.5 degrees or less (Figure 8). The max temperature GFS however, forecasts warm with biases a bit better than both NAM and NMM-b, while for the minimum temperature, the GFS is cool and worse than the NAM. GFS MAE values are effectively the same as the NAM and NMM-b.

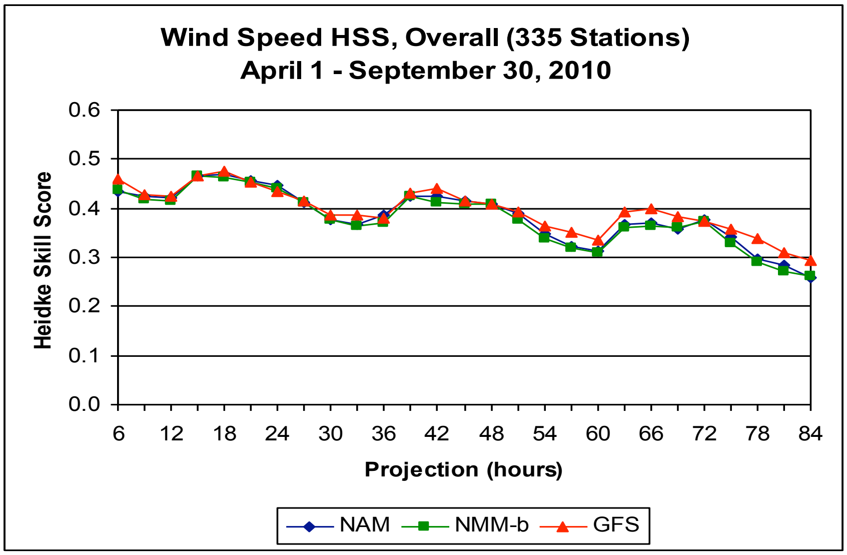

Figure 9: Warm season wind speed Heidke skill score for all sites

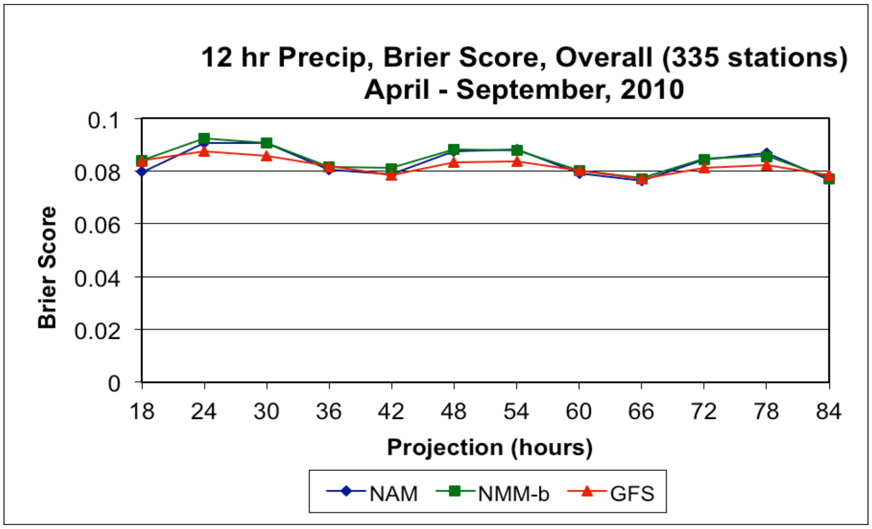

Figure 10: Warm season 12-hour POP Brier score for all sites

For wind speed and wind direction, the results suggest very little to no loss in skill with the NMM-b model implementation (Figure 9). The 6 and 12-hour POP analyses show the same results, with just a slight loss in skill following the implementation (Figure 10). Comparing these elements of the NAM and NMM-b to the GFS MOS, we see very small and relatively insignificant differences.

As a whole, the warm season verification results show insignificant degradation of MOS associated with the NMM-b changes and no reason for redevelopment of equations at this point. However, since synoptic-scale dynamic forcing is usually more pronounced during the winter months, differences between the NAM MOS and these equations applied to NMM-b model output are more likely to arise. For this reason, in order to do a complete analysis and make any kind of conclusions, we must also look at the 6-month cool season verifications.

b. *Cool Season*

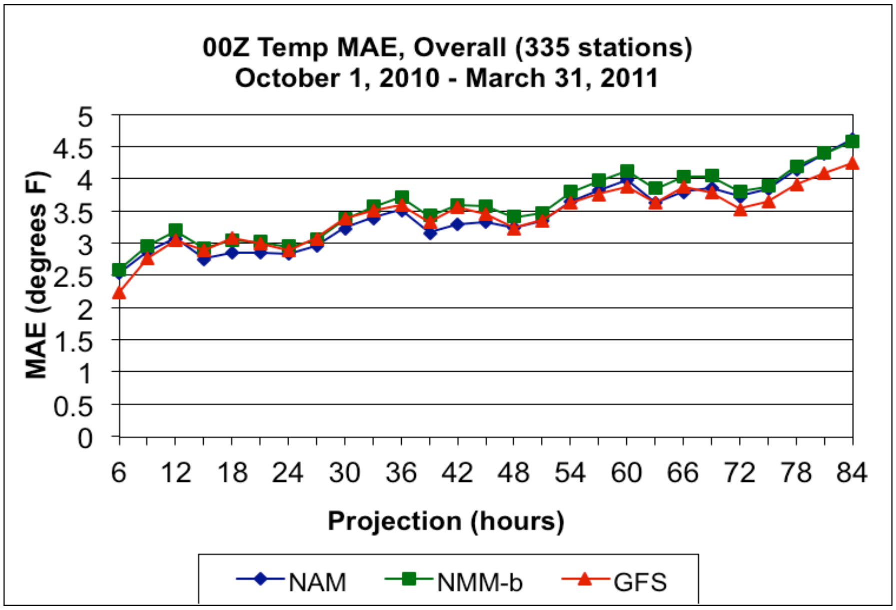

Figure 11: Cool season temperature MAE for all sites

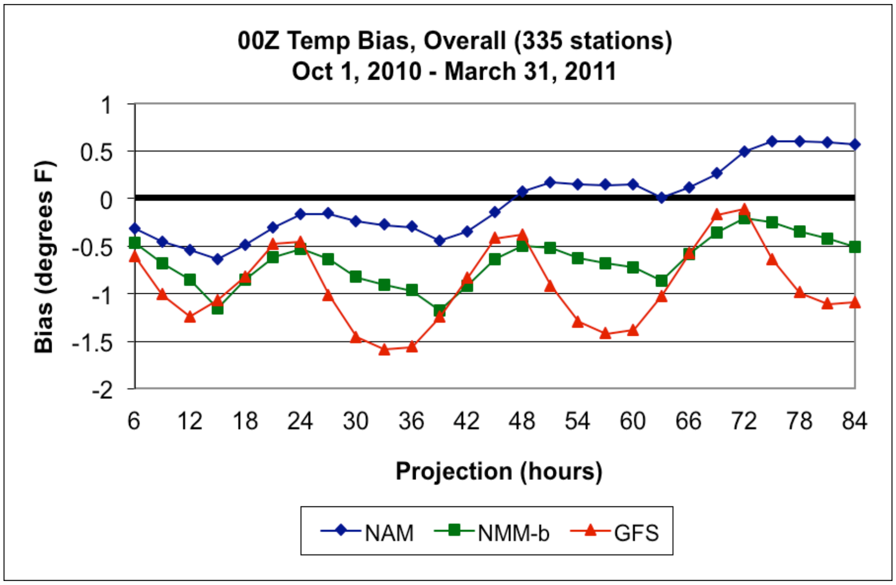

Figure 12: Cool season temperature bias for all sites

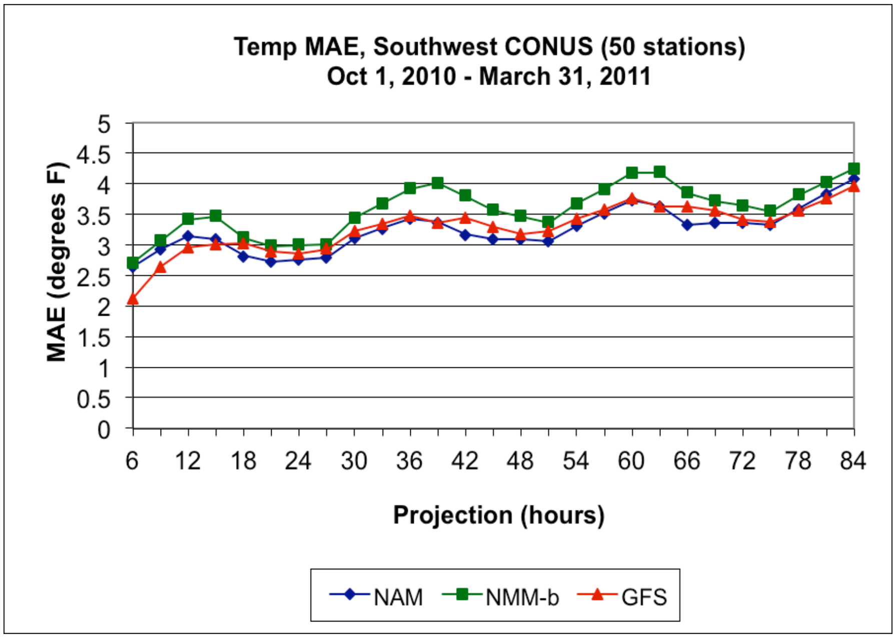

Figure 13: Cool season temperature MAE for southwest CONUS region

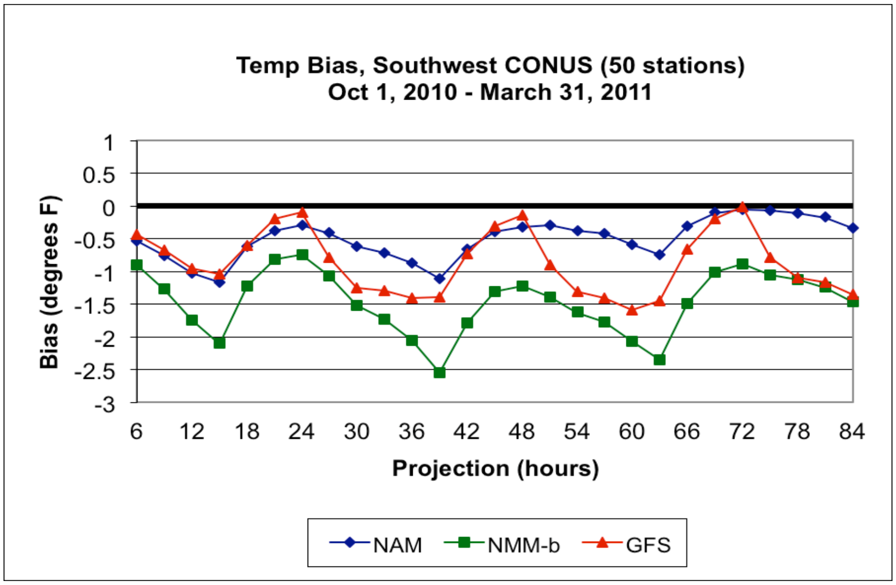

Figure 14: Cool season temperature bias for southwest CONUS region

The cool season 2m temperature results show NMM-b MAE values slightly worse than the NAM across all regions analyzed with differences less than half a degree (Figure 11). Additionally, the NMM-b output exhibits a cool bias, and is worse than the NAM with differences between the two approaching about a degree (Figure 12). Regionally, we also see NMM-b MAE performing worse than the NAM, with the greatest differences in the southwest of about 0.5 degrees (Figure 13). The most alarming of this contrast in the southwest is the almost 2 degree bias difference between the NAM and NMM-b outputs (Figure 14). Looking at MAE, the GFS appears to perform worse than the NAM, as expected, but better than NMM-b. For biases, GFS MOS performs similar to, or worse than, the NMM-b except in the southwest region of the US where we see better biases from the GFS compared to the NMM-b.

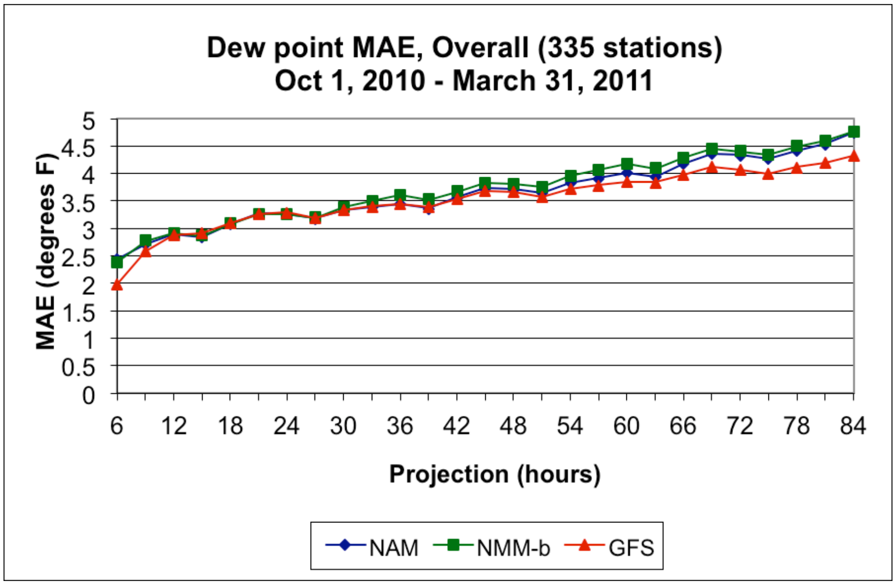

Figure 15: Cool season dew point MAE for all stations

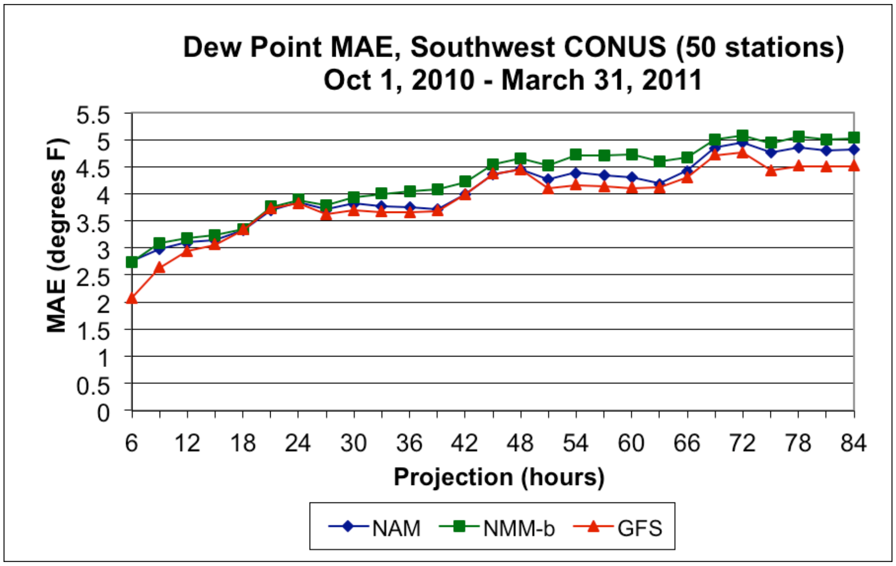

Figure 16: Cool season dew point MAE for southwest CONUS region

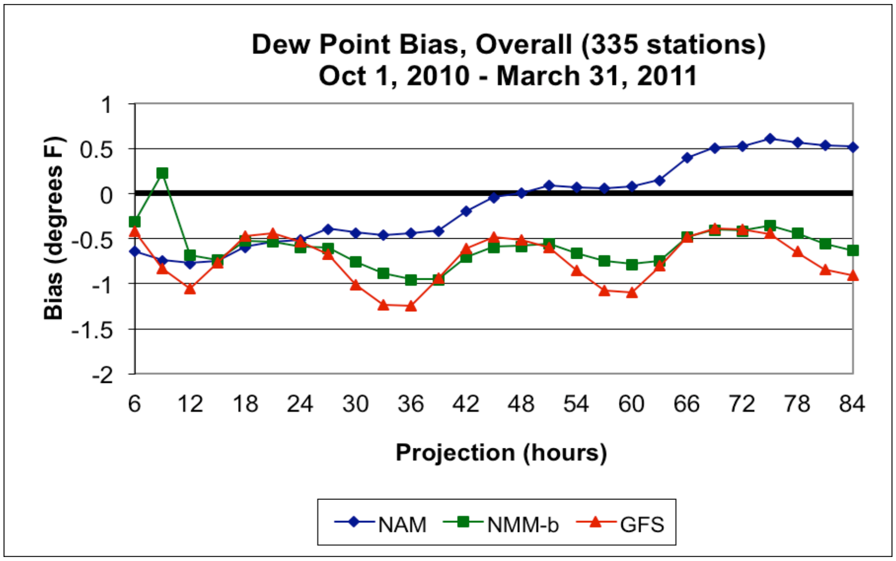

Figure 17: Cool season dew point bias for all stations

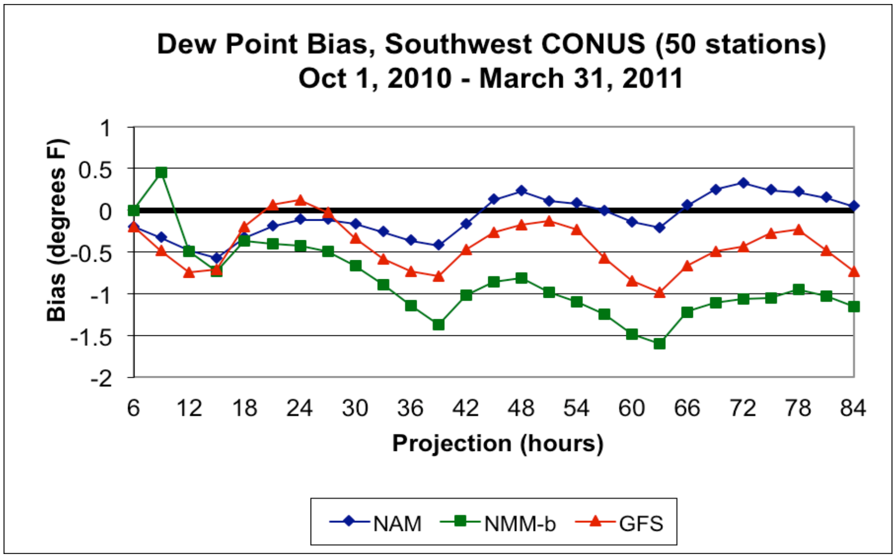

Figure 18: Cool season dew point bias for southwest CONUS region

Overall, the NMM-b dew point MAE values were very similar to, or a little higher than the NAM, with small differences of less than about 0.25 degrees (Figure 15). The greatest of these differences occurs in the southwest CONUS region where differences approach 0.5 degrees (Figure 16). We also see an overall dry NMM-b bias across all regions worse than that of the NAM (Figure 17) – in some cases by almost 1 to 1.5 degrees, especially in the southwest CONUS region (Figure 18). Comparing to the GFS, we see GFS MAE values similar to NMM-b in early projections, and slightly better than NMM-b in later projections. GFS bias values were also close to, or even a bit worse, than the NMM-b and also have a much stronger diurnal cycle. This holds true in all regions except the southwest, where the GFS actually out performs the NMM-b, but is still worse than the NAM.

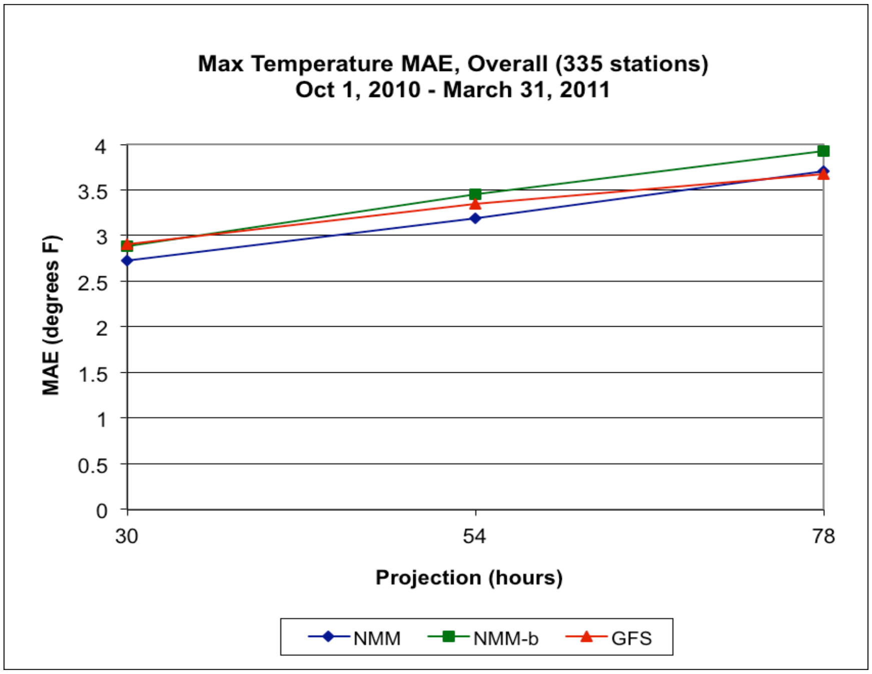

Figure 19: Cool season local maximum temperature MAE for all stations

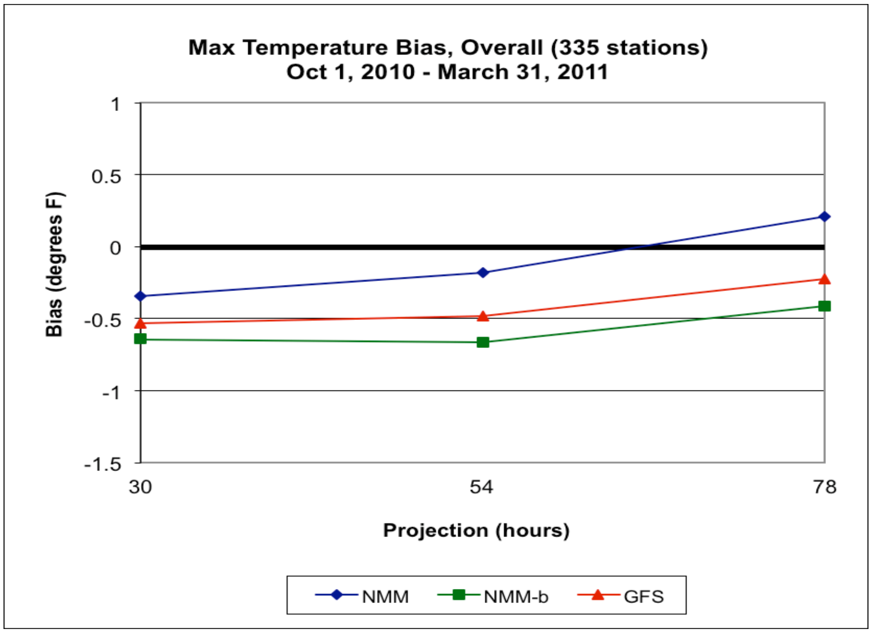

Figure 20: Cool season local maximum temperature bias for all stations

Looking at the local maximum and minimum temperature cool season results, the MAE values for the NMM-b are once again similar to, or very marginally worse than, those of the NAM with very small differences of less than 0.5 degrees (Figure 19). Compared to NAM, the NMM-b does have a cool bias greater than NAM by less than a degree (Figure 20). As was seen with the temperature and dew point results, the GFS seems to perform just a little bit better than the NMM-b but still not quite as good as the NAM.

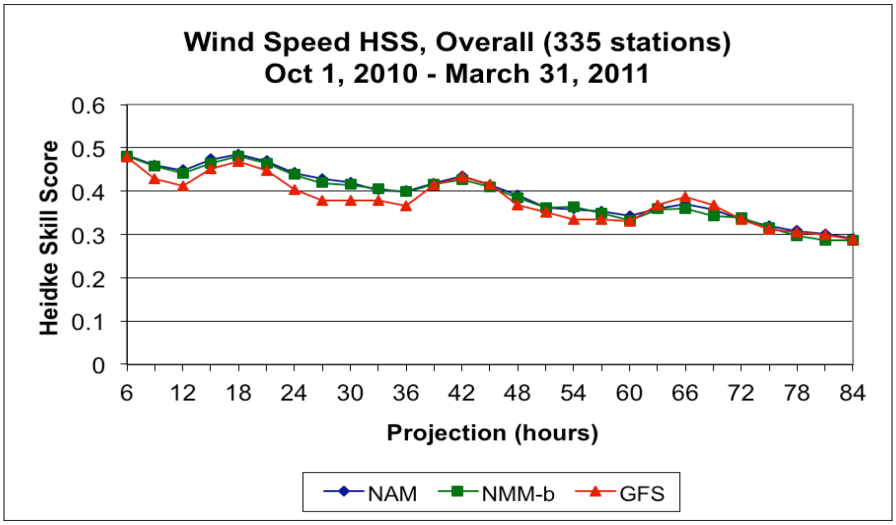

Figure 21: Cool season wind speed Heidke skill score for all stations

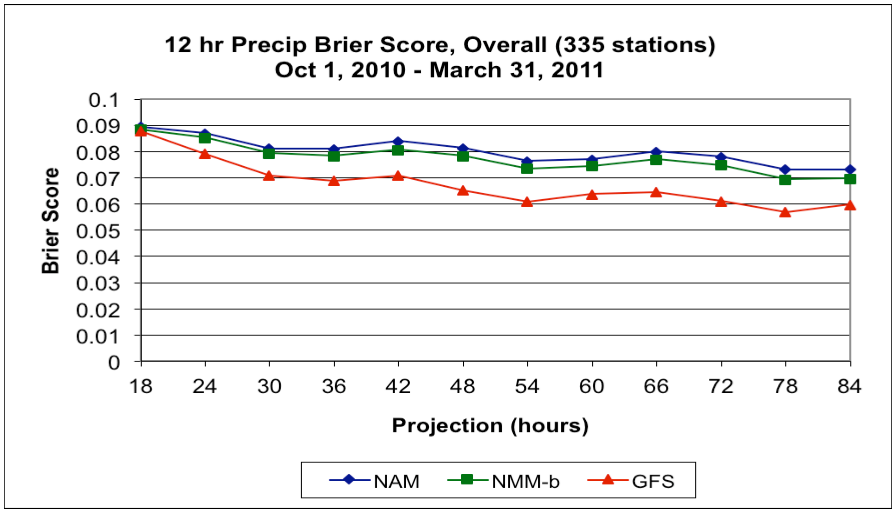

Figure 22: Cool season 12-hour POP Brier score for all stations

For wind speed/direction and POP we see just about the same results as we saw with the warm season tests. For all scores, differences between the NAM and NMM-b MOS systems are minimal (Figures 21 & 22). For wind speed and direction, the performance differences between the GFS MOS and NAM MOS are small; in most cases the NAM MOS outperforms the GFS just a little bit. For both 6 & 12 hour POP, the GFS MOS is slightly worse than the NAM MOS. The wind and POP results suggest that impacts to the MOS for these elements after the implementation will be insignificant and we see no reason for a redevelopment of these equations.

The cool season results indicate to us that the NMM-b implementation may have more damaging effects to the MOS than was originally anticipated after the warm season tests, especially for the temperature elements. For this reason, a temperature equation redevelopment was suggested. For wind speed, direction, and 06/12-hour POP elements, we believe impacts to the MOS will be insignificant and we see no reason for a redevelopment of these equations.

IV. Conclusions

After completion of the verifications, we see that the resulting effects on the existing NAM MOS system for the cool season are similar to those of the warm season, although they are slightly more pronounced in some cases, especially for temperature-related weather elements (2m temp, dew point, max and min). Additionally, the previous NAM MOS performance advantage over the GFS MOS seems to be lost with the implementation of the NMM-b. For these reasons, the decision to upgrade the MOS equations for the four temperature elements was made. Since there was only one year of parallel NMM-b data available, we chose to use a dependent sample consisting of three years of operational NAM data in addition to the 1 year of available NMM-b data in order to increase the sample size for the regression analysis. Even with the inclusion of just one year of actual NMM-b output, we believe these new equations should perform better than the original NAM MOS equations, given the results of previous work with short-sample NAM MOS systems (i.e. Antolik and Baker 2009). Equation redevelopment was performed for both seasons, and these equations were implemented experimentally at NCEP in August of 2011. Once enough warm and cool season data is available, ongoing verifications at MDL will show the skill of the new NAM MOS system.

References:

Antolik, M. S., and M. N. Baker, 2009: On the ability to develop MOS guidance with short dependent samples from an evolving numerical model. Preprints, *23rd Conference on Weather Analysis and Forecasting,* Omaha, NE, American Meteorological Society, 6A.1.

DiMego, G., and E. Rogers. 2011, "Q4FY11 NAM Upgrade Package Decision Brief", NCEP Director Briefing, Camp Springs, MD. Accessed on November 27, 2011 from http://www.emc.ncep.noaa.gov/mmb/briefings/ under the heading 'NAM Upgrade to NEMS/NMMB – Implementation Decision Brief (18 October 2011)'.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11, 1203 – 1211.

WRF MOS Evaluation. Last updated July 14, 2006. Retrieved November 27, 2011 from http://www.nws.noaa.gov/tdl/synop/wrfmoseval.htm.

Appendix

Score	Element	Equation	Meaning

MAE	Temp, dew point, max, min, wind speed, wind direction	$\frac{1}{n}\sum_{i-1}^{n}\lvert F-O\rvert_i$
		n = # of forecasts
		F = forecast value
		O = observation value
			Average of the absolute differences between the forecast and actual observation. Higher MAE = lower accuracy.

Bias	Temp, dew point, max, min	$\frac{1}{n}\sum_{i-1}^{n}(F-O)_i$
		n = # of forecasts
		F = forecast value
		O = observation value
			Positive value indicates a warm (T) or wet (T_d) bias, a negative value indicates a cool (T) or dry (T_d) bias.

HSS	Wind speed	$\frac{H-E}{T-E}$
		H = # of hits
		E = expected hits by chance
		T = total # of cases
			Proportion of correct forecasts. Value of 1 indicates perfect forecasting, 0 indicates no skill.

CRF	Wind direction	Percentage of forecast errors less than 30 degrees. Values between 0 and 1 (100%). A higher value indicates better accuracy	

Brier	POP	$\frac{1}{n}\sum_{i-1}^{n}(F-O)_i^2$
		n = # of forecasts
		F = probability between 0 and 1
		O = observation (set to 1 if event occurred, 0 if event did not occur)
			Used for categorical elements such as precipitation. The squared accumulant is subtracted from 1 and compared to a reference forecast. Higher score represents higher skill.

Table A-1: Summary of scores used, their equations, meanings, and the elements relevant to each score (WRF MOS evaluation).

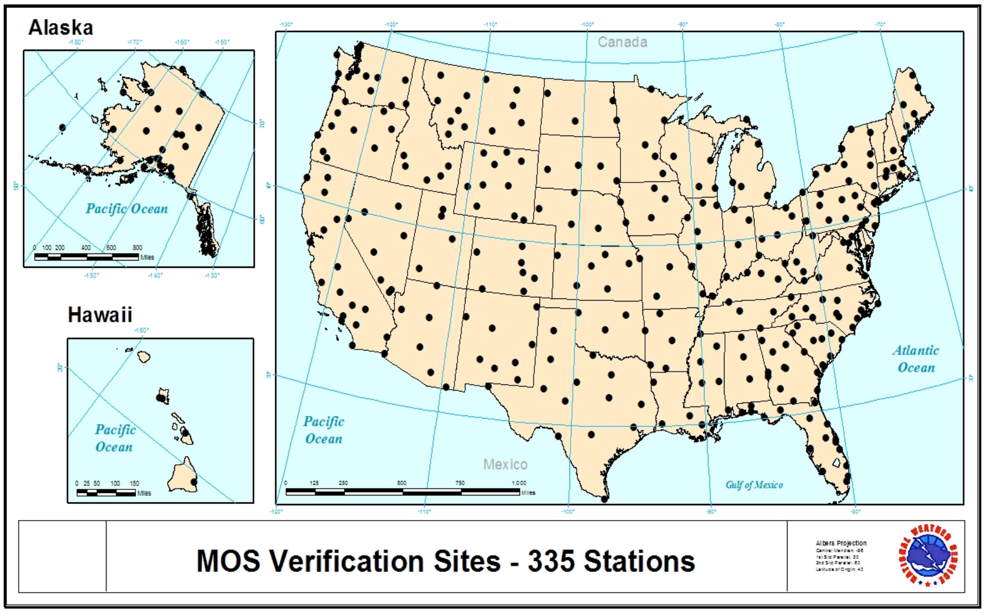

Figure A-1: Locations of the 335 standard SMB MOS verification sites for this project (not including two sites in Puerto Rico)