3.2 THE NEW AVN/MRF MOS DEVELOPMENT AND MODEL CHANGES: A VOLATILE MIX?

Mary C. Erickson*, J. Paul Dallavalle, and Kevin L. Carroll
Meteorological Development Laboratory
Office of Science and Technology
National Weather Service, NOAA
Silver Spring, Maryland

1. INTRODUCTION

The U.S. National Weather Service (NWS) has over 30 years of experience in employing statistical regression to produce objective forecasts of a wide variety of sensible weather elements. For the most part, the Model Output Statistics (MOS) approach (Glahn and Lowry 1972) has been used to relate the observations of a given weather element to the output of numerical weather prediction (NWP) models and to produce the objective forecasts. While the skill, accuracy, and availability of these objective forecasts have improved throughout this period, certain attributes of the technique continue to pose a challenge. For instance, an accepted requirement of the MOS approach has been the existence of an "appropriate" developmental sample to produce the predictive relationships. The length of the sample, the underlying climate represented by the sample, and the NWP characteristics included in the sample all influence the usefulness of the resulting forecast equations. As advances in the complexity and resolution of NWP models are made with increasing frequency, some have questioned whether statistical forecasts can keep pace, given the limitations the sample requirement poses, or, indeed, whether direct model output could be used in lieu of the interpretive guidance. In 1996, as we began major modifications to our MOS development system and forecast equations, we discussed this challenge (Dallavalle 1996) and the continued benefits of the MOS technique. Now, with the implementation of the new MOS system in 2000 (Glahn and Dallavalle 2002), we revisit this discussion and focus on the impact to the MOS system of model changes introduced in 2001. Specifically, we discuss recent changes made in the Aviation (AVN) and Medium-Range Forecast (MRF) runs of the Global Spectral Model (GSM) and the impact of those changes on the MRF-based MOS guidance.

2. BACKGROUND

On May 30, 2000, the NWS implemented the new MOS-2000 system. New AVN- and MRF-based MOS guidance packages were produced and disseminated to the user community. The initial implementations of both packages contained only a subset of the complete weather element guidance. During the subsequent year, MDL developed extensive sets of equations and implemented the guidance for additional weather elements. In contrast to prior AVN and MRF MOS products, the new guidance packages were available for a large variety of forecast elements, for a full range of forecast projections, and for over 1000 sites in the contiguous U.S., Alaska, Hawaii, and Puerto Rico. Dallavalle and Erickson (2000) and Erickson and Dallavalle (2000) discuss the AVN and MRF alphanumeric products in more detail, and Erickson and Carroll (1999) provide a thorough discussion of the MRF MOS development process.

The new MOS equations were able to take advantage of more recent historical samples of both the AVN and MRF models, as well as important model improvements and higher resolution model archives, than used in prior developments. Both AVN and MRF output (out to 192 hours after 0000 UTC) were archived at a grid resolution of 95.25 km at 60EN. This represents a doubling of the resolution (190.5 km at 60EN) used by the existing NGM MOS system (Jacks et al. 1990, Dallavalle et al. 1992), and quadruple the resolution used in the development of the original MRF MOS system (381 km at 60EN) (Jensenius et al. 1995). For both the new AVN and MRF MOS systems, GSM output from April 1997 to March 2000 was used. Forecasts from the National Centers for Environmental Prediction (NCEP) reanalysis project (Kalnay et al. 1996) for every 5th day from 1992 through 1996 were also included in the new MRF development. Changes to the GSM, including increases in resolution, physics updates, and data assimilation modifications continued throughout the historical sample period. One 5-week sample of model forecasts from the summer of 1998, which was the period between a major GSM implementation and subsequent modification (Derber et al. 1998), was eliminated due to its anomalous nature. However, we judged that most of the model changes did not significantly affect the statistical characteristics of the sample. For more complete information on GSM changes made over the past 10 years see:

http://sgi62.wwb.noaa.gov:8080/research/model_changes.html.

With the current reality of frequent numerical model improvements, several techniques were employed in the development of the MOS system to attempt to develop robust forecast equations which would accommodate the mixed development samples. MDL's archive of GSM data, for example, was established on a standardized grid chosen to approximate the resolution of NCEP's 1E and 2.5E latitude/longitude grids. This approach allowed us to keep our sample somewhat immune to future resolution changes. In addition, all model predictors were smoothed to reduce volatility, and station-specific relative frequencies of several weather elements were included in equa-

* *Corresponding author address*: Mary C. Erickson, National Weather Service/OST/MDL, 5200 Auth Road, Room 706, Camp Springs, MD 20746-4304; e-mail: mary.erickson@noaa.gov

tions to provide valuable station climatic information independent of the model forecasts. Each of these steps was taken to improve our ability to make skillful synoptic-scale forecasts, yet reduce the sensitivity of our system to model changes. However, note that many highly correlated predictors, such as 2-m temperature or dewpoint, were not eliminated from the equations despite their sensitivity to NWP enhancements. Evaluations of the MOS guidance on independent data demonstrated overall improvement in the new MOS system over the older guidance.

3. THE MODEL CHANGED!

In the spring of 2001, one year after the new MOS system was implemented, a new suite of GSM changes was scheduled for implementation. Changes to the model physics (modifications to the cloud condensate and cumulus momentum mixing) and analysis (refinement of the hurricane relocation algorithm) were made to improve circulation patterns in both the extratropics and the tropics, and reduce the false alarm rate for tropical storms (Moorthi et al. 2001). For evaluation purposes, NCEP ran the revised MRF in a parallel mode to the operational version of the MRF, and made the resulting forecasts available to MDL starting in March 2001. Years before, MDL had participated in a similar evaluation of the impact of changes in the Regional Analysis and Forecast System (RAFS) on the NGM MOS (Erickson et al. 1991) and found the process quite helpful. Parallel tests provide information as to which sensible weather elements are likely to be affected by the upcoming model changes, and give the opportunity to evaluate shifts in key model predictors. In the case of the spring 2001 GSM changes, three one-month evaluations were conducted, and the results were presented to the NCEP modelers and scientific advisors throughout the NWS to assist in critically reviewing the changes.

To generate parallel MOS forecasts, the new operational MOS equations were applied to this parallel run of the MRF. The resulting forecasts were compared to the MOS forecasts issued operationally. Forecasts of maximum and minimum temperature (max/min), probability of precipitation (PoP), and total sky cover were included in the evaluation. The first test, made for March 2001, revealed a drop in accuracy of the max/min forecasts which appeared to be the result of a distinct shift in the bias characteristics of the MRF low-level thermal fields over North America. The evaluations of the PoP and sky cover forecasts did not reveal any deterioration in skill. Figure 1 shows the mean absolute errors (MAE) of both the operational and parallel MRF MOS max/min forecasts from 0000 UTC for March 2001. For most forecast projections, the bias (forecast - observation) cooled by approximately 2EF from the operational to the parallel MOS guidance (Fig. 2). This feedback was provided to NCEP modelers who discovered an error in the snow albedo algorithm that was contributing to the degradation. On April 14, 2001, this problem was fixed, and a new version of the GSM began running as the parallel model. In addition, reruns of the MRF for February 2001 were

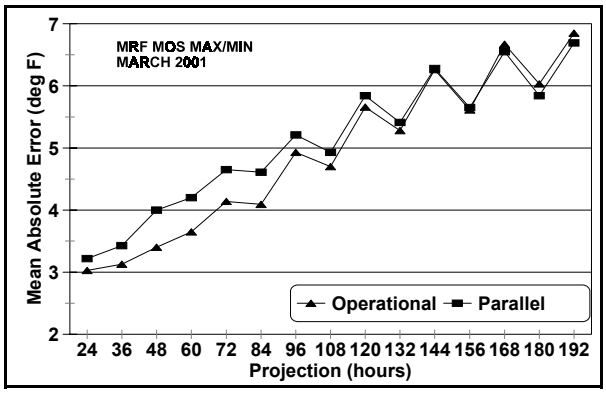

Figure 1. Mean absolute error of operational and parallel MRF MOS max/min temperature forecasts for March 2001. Forecasts were verified for approximately 330 sites in the contiguous U.S., Alaska, Hawaii, and Puerto Rico.

made so that a month of winter season forecasts could be evaluated. As a second test, the February 2001 MOS forecasts based on the parallel data were compared to operational forecasts. This evaluation indicated a slight improvement in the parallel daytime max and nighttime min forecast skill in terms of overall MAE (Fig. 3); however, the bias still shifted from a general 1EF warm bias to a 0.5EF cool bias as shown in Fig. 4. In Alaska, the bias was affected more dramatically, introducing a 1.5EF cool bias and degradation in the MAE after 120 hours (not shown). Finally, a third test evaluated the MOS temperature forecasts based on parallel runs made during the April 14 to May 15 period. These results provided additional evidence that a cool bias (Fig. 5) and some degradation in accuracy might result from the proposed changes to the MRF.

On May 16, 2001, the new suite of GSM changes became operational. While tests indicated that a substantial cool bias in low-level thermal fields had replaced a warm bias in the operational forecasts in parts of North America, other evaluations indicated improvements in summer hemisphere circulation forecasts, removal of

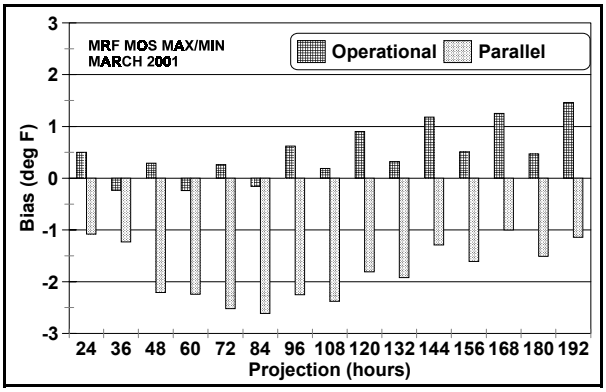

Figure 2. Same as Fig. 1, except for the bias of the MOS forecasts.

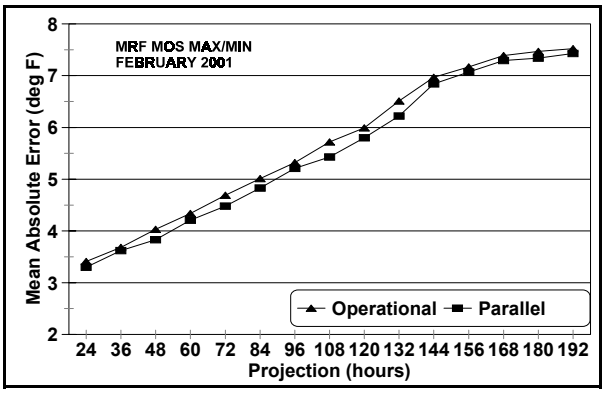

Figure 3. Same as Fig. 1, except for MOS forecasts for February 2001.

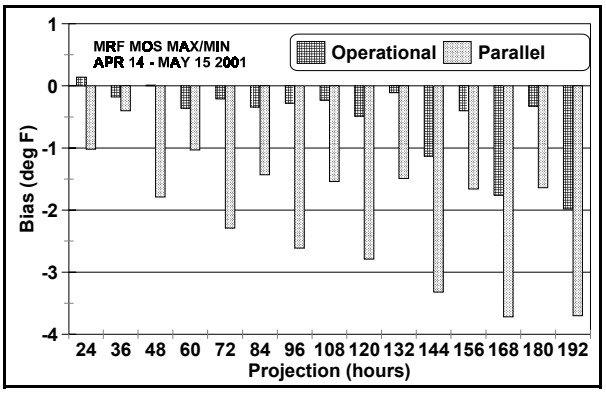

Figure 5. Same as Fig. 2, except for MOS forecasts for April 14 to May 15, 2001.

tropical storm false alarms, and improvement in tropical circulation forecasts (H. Pan 2001, personal communication). The removal of the false alarms in hurricane situations outweighed the bias concern; thus, the decision was made to go ahead with the changes.

4. WHAT MAKES A GOOD SAMPLE?

After the implementation of the new GSM, we decided to take a look at the influence of the sample length and model continuity on the operational MRF MOS max/min forecast equations to better understand their response to the model changes. Because the warm season (April - September) temperature equations were developed from the 1992-96 reanalysis data and from 1997-99 MRF data, our sample spanned numerous versions of the GSM. Were the model characteristics so blended that we were losing some of the benefits of the MOS approach? How sensitive were the MOS forecasts to the historical samples included in the development? To answer these questions, we derived two sets of test equations to predict the max and min temperature in the warm season. The first set of test equations was derived by using MRF output from 1997 to 2000, thus eliminating the reanalysis data from 1992-1996 and adding one more year of recent model data to the sample. This test allowed us to look at the possibility that the differences in resolution, physics, and climate regime during the reanalysis period had adversely affected the statistical relationships. We developed a second set of test equations by using MRF output from 1999 and 2000. Major changes were made to the model during the warm season of 1998 (Derber et al. 1998, Caplan and Saha 1998), and isolating the 1999/2000 samples provided the most homogeneous data set in regards to the model physics. Figure 6 shows the MAE for forecasts generated by the two sets of test equations (1997-2000, 1999-2000) and the operational equations when applied to the parallel MRF (PARA) for April 14 - May 15, 2001. The MAE of the operational MRF MOS guidance (OPER) is also shown. Forecasts were verified for approximately 330 sites in the contiguous U.S., Alaska, Hawaii, and Puerto Rico. The max/min forecasts generated from the operational equations clearly outperformed the other three systems. The deterioration in the MOS guidance caused by the model changes is clear in the differences between the MAE of the OPER and PARA systems. The decrease in accuracy of the max temperature forecast is particularly evident. The MAE's of the 1999-2000 system are the greatest, demonstrating the risk, not only of a short sample, but also of a sample whose statistical characteristics are significantly different from those of the newest model. The forecasts of the

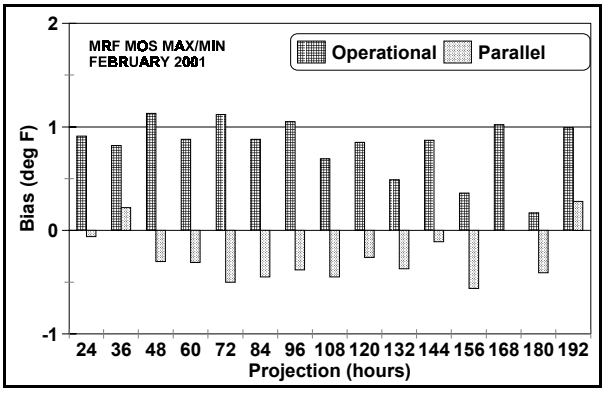

Figure 4. Same as Fig. 2, except for MOS forecasts for February 2001.

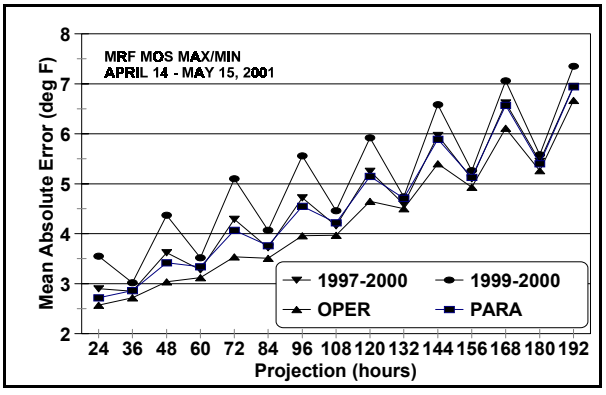

Figure 6. Mean absolute error of operational, parallel, and experimental MOS max/min temperature forecasts for the April 14 - May 15, 2001 period.

1997-2000 system were significantly more accurate than those of the 1999-2000 system, and almost as accurate as those of the PARA. It appears that a long training sample, including a variety of climate regimes and NWP changes, provided the most robust forecast system.

5. BIAS CORRECTION

Since the MOS max temperature forecasts were affected more severely than any other weather element, MDL also examined the differences in the thermal fields output directly from the operational and parallel versions of the MRF. Temperatures at constant pressure levels (1000, 950, 850, 500 mb) and various thicknesses were compared starting at the initial state to the 192-h projection after initial model run time. Although the operational and parallel versions of the MRF started out with nearly identical initial values, the parallel thermal fields diverged after only 24 hours, and were consistently cooler than the operational run. Figure 7 compares the 1000-mb model and observed 2-m temperature for the April 14 - May 15, 2001 test period. These results were consistent at various pressure levels and for various thicknesses, although the magnitude of the cold bias decreased at higher pressure levels. Admittedly, we had only a small sample of data (three 1-month samples) on which to base a judgment, but since the bias pattern seemed very consistent and isolated to thermal fields, we decided to test an algorithm to subtract the systematic error out of specific predictors. Our approach was based on practical considerations, our software system, and on the hypotheses that the systematic thermal error in the MRF established itself quickly, remained relatively constant with projection, and was larger than the random error. Given these hypotheses, we could estimate the systematic error inherent in the model physics by examining a single day's forecasts. This method would allow us to rederive equations using bias-corrected thermal predictors and nearly our full developmental sample without making extensive changes to our software system. The equations could then be applied to the output of the new GSM to lesson the effect of the model bias on the resulting MOS forecasts. The traditional MOS approach would involve collecting one or two seasons of new model data before developing linear regression equations to account for systematic error -- which might change with the next model modification.

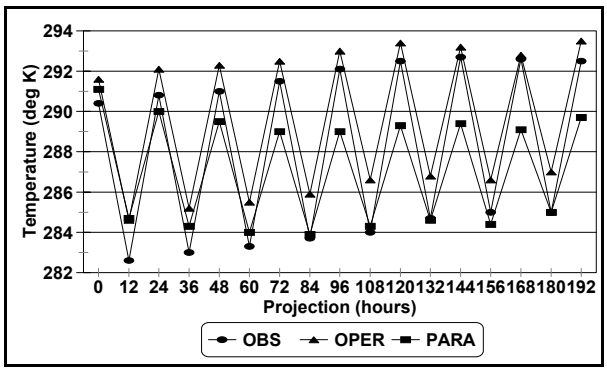

Figure 7. Average operational (OPER) and parallel (PARA) model 1000-mb temperature forecasts for the April 14 - May 15, 2001 period. Forecasts were interpolated to approximately 330 sites in the U.S. and Puerto Rico. Average observed 2-m temperature (OBS) valid at the same time and location is also shown.

The bias correction routine was based on the concept of removing the bias observed in the last model run (in the case of the MRF, each model run is 24 hours apart) from the current model forecast. The following formula was applied to each grid point of the thermal predictor:

$$T_{corr} = T_{proj} - [\, T_{prev} - T_0 \,]$$

In this equation, T_{corr} is the final corrected predictor value and T_{proj} is the original GSM predicted value at the projection a forecast is to be made. T_{prev} is the predictor value from a previous model run, and valid at time 0, and T_0 is the predictor value at the initial time 0. The term in brackets represents the bias, calculated at each grid point, that is removed from the GSM forecast field (forecast - observation). As an example, a "corrected" 72-h 1000-mb temperature forecast is computed by subtracting a bias from the original 72-h forecast. This bias is estimated by the difference between the 24-h 1000-mb temperature forecast made in the previous cycle, and today's initialized 1000-mb temperature.

The specific configuration of this bias correction experiment was quite simple. Only the 24-h bias was used as a correction, and temperatures at 1000, 925, 850, and 700 mb were corrected. Although the 2-m temperature was a key predictor in development, we were not able to include this bias corrected temperature in the equations since the 2-m fields at the initial time were not available in the archived parallel model output, and, therefore, the forecasts could not have been generated for evaluation. Because a bias correction based on one day introduces random error into the predictor variables, the corrected predictors were not as closely correlated to the max and min temperature as the uncorrected fields. Since the equations were developed on one sample for application to an independent sample with entirely different thermal characteristics, we forced the bias corrected fields into the equations, and removed all the uncorrected thermal fields as potential predictors.

Test equations were thus developed based on data from 1997 to 2000. (Reanalysis data could not be used because forecast and initial fields were only available every 5th day, eliminating our ability to properly calculate the 24-h bias correction.) Since the operational warm season equations were developed based on data from 1992 to 1999, we also developed a set of control forecast equations to demonstrate differences arising solely from the change in developmental sample. These equations, based on data from 1997 to 2000, were offered the same complete set of predictors used in the original development. Forecasts for the April 16 to May 15 time period were made from both sets of equations (BIAS and CNTL) and their accuracy and bias compared to those of the

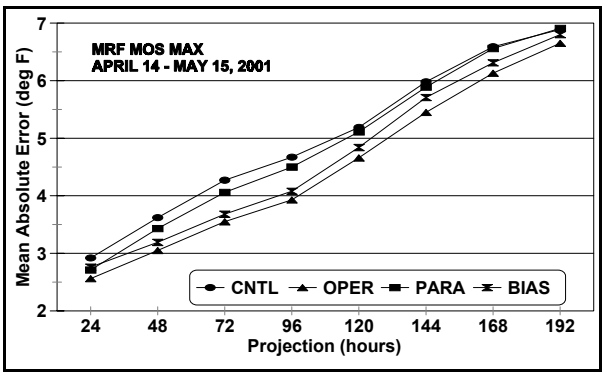

Figure 8. Mean absolute error of operational, parallel, control, and bias-corrected MRF MOS max temperature forecasts at approximately 330 stations for April 14 - May 15, 2001.

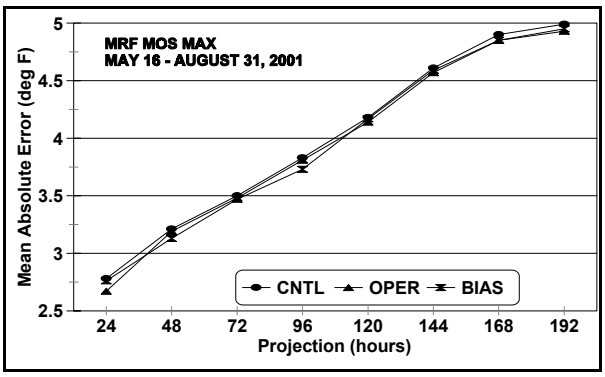

Figure 10. Same as Fig. 8, except for forecasts from May 16 to August 31, 2001.

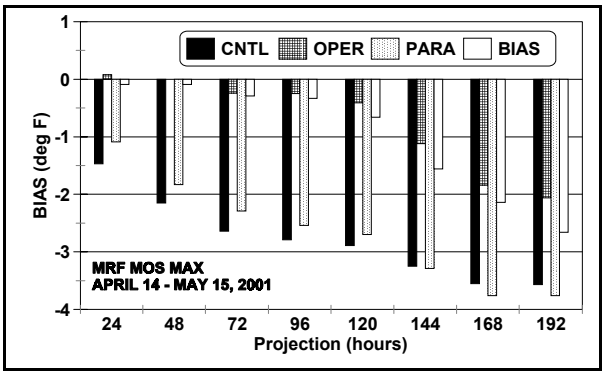

Figure 9. Same as Fig. 8, except for bias.

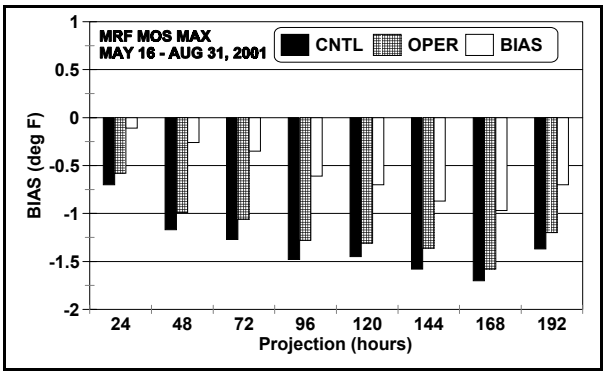

Figure 11. Same as Fig. 9, except for forecasts from May 16 to August 31, 2001.

operational forecasts (OPER) as well as the forecasts produced by applying the operational equations to the parallel GSM output (PARA). Figure 8 shows that the bias corrected forecasts were more accurate than the control and parallel forecasts, and nearly as accurate as the operational forecasts at most projections. The cold bias of the MOS forecasts was also reduced by approximately 1.5EF at all but the earliest projection (Fig. 9). The drop-off in the usefulness of the technique with increasing projection is likely due both to the fact that the bias was based on the error found in the 24-h forecasts and may not have fit the 144- to 192-h forecasts well, and that the increase in random error in the model forecasts with increasing projection overwhelmed any systematic bias. These results were encouraging and gave hope that a relatively simple short-term solution could be applied to adapt the temperature guidance to the change in model characteristics.

Since our test equations included the drawbacks of being tested on only one month of data and requiring the removal of one of the best predictors (the 2-m temperature), we continued to evaluate the approach before implementing corrected forecast equations operationally. In addition, we were not sure if this model trait would continue as we moved into the summer season. The verification results for the May 16 - August 31 period are shown in Figs. 10 and 11. These results indicate that the bias correction forecast equations continued to generate accurate and relatively unbiased guidance.

6. CONCLUSIONS AND FUTURE WORK

The sampling and bias correction experiments have shown that developmental samples containing a mixture of model evolutions can be used productively by the MOS system, but that short-term corrections to the MOS forecasts may still be required to adjust to changing model bias characteristics. With the increasing complexity involved in the parameterization algorithms used in NWP models, correction of the bias in the model itself seems to be a very delicate process. If modelers are adjusting these parameterizations regularly, it is difficult to respond to these changes without building adaptive tools directly into our MOS system. A direct approach to this problem is to further post-process the MOS forecasts themselves with a type of bias correction. Jensenius et al. (1992) used a method involving a fixed training sample to calibrate MRF-based perfect prog forecasts. MDL is currently testing a regression correction for MRF MOS temperatures based on the past 30 days of MOS forecasts and verifying observations for a particular station and projection. The resulting correction will be applied to the current

MOS temperature forecasts. The update of the bias correction in the test will be done monthly, but shorter time periods could be used. While this technique may be straightforward for temperature or wind forecasts, applying the approach to forecasts of less well behaved weather elements such as PoP, ceiling height, or precipitation type probabilities could be difficult due to the low frequency of events.

A second approach is to improve the flexibility of the MOS system through the predictors themselves. The bias-correction of thermal predictors was one attempt to improve flexibility, and we will continue to explore possible refinements to this technique to make it more robust. In addition, we have considered modifying the format in which certain model predictions are offered to the regression. Rather than directly offering the 1000-850 mb thickness, for example, the deviation of this thickness from a climatic normal, or even the variance of the deviation over a period of time may provide a useful way for the regression to accommodate consistent bias changes in the developmental sample, and to respond to differences in these fields once the equations are implemented. Finally, a consensus MOS forecast provides an entirely different way to take advantage of the variety of MOS packages to provide robust forecasts. Vislocky and Fritsch (1995) have shown that consensus consistently outperforms any particular set of objective guidance. With MOS packages based on MRF, AVN, NGM, Eta, and ensemble output available or scheduled, we certainly have enough variety to provide robustness.

7. REFERENCES

Caplan, P., and S. Saha, 1998: Further changes to the 1998 NCEP operational MRF analysis/forecast system. NWS Technical Procedures Bulletin No. 450, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 25 pp.

Dallavalle, J. P., 1996: A perspective on the use of Model Output Statistics in objective weather forecasting. Preprints 15th Conference on Weather Analysis and Forecasting, Norfolk, Amer. Meteor. Soc., 479-482.

_____, J. B. Bower, V. J. Dagostaro, D. T. Miller, and J. C. Su, 1992: Development of a new statistical weather forecast system. Preprints 12th Conference on Probability and Statistics in Atmospheric Sciences, Toronto, Amer. Meteor. Soc., 201-206.

_____, and M. C. Erickson, 2000: AVN-based MOS guidance - The alphanumeric message. NWS Technical Procedures Bulletin No. 463, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 12 pp.

Derber, J., H. Pan, J. Alpert, P. Caplan, G. White, M. Iredell, Y. Hou, K. Campana, and S. Moorthi, 1998: Changes to the 1998 NCEP operational MRF model analysis/forecast system. NWS Technical Procedures Bulletin No. 449, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 13 pp.

Erickson, M. C., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, E. Jacks, J. S. Jensenius, Jr., and J. C. Su, 1991: Evaluating the impact of RAFS changes on the NGM-based MOS guidance. Wea. Forecasting, 6, 142-147.

_____, and K. L. Carroll, 1999: Updated MRF-based MOS guidance: Another step in the evolution of objective medium-range forecasts. Preprints 15th Conference on Weather Analysis and Forecasting, Denver, Amer. Meteor. Soc., 190-195.

_____, and J. P. Dallavalle, 2000: MRF-based MOS guidance - The alphanumeric message. NWS Technical Procedures Bulletin No. 460, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 12 pp.

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.

_____, and J. P. Dallavalle, 2002: The new NWS MOS development and implementation systems. Preprints 16th Conference on Probability and Statistics in the Atmospheric Sciences, Orlando, Amer. Meteor. Soc., (in press).

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. Wea. Forecasting, 5, 128-138.

Jensenius, J. S., K. K. Hughes, and J. B. Settelmaier, 1992: Calibrated perfect prog temperature and probability of precipitation forecasts for medium-range projections. Preprints 12th Conference on Probability and Statistics in Atmospheric Sciences, Toronto, Amer. Meteor. Soc., 213-218.

_____, _____, and _____, 1995: The National Weather Service's AVN- and MRF-based statistical weather forecast systems. Preprints 14th Conference on Weather Analysis and Forecasting, Dallas, Amer. Meteor. Soc., 163-170.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteor. Soc., 77, 437-471.

Moorthi, S., H. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. NWS Technical Procedures Bulletin No. 484, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 13 pp.

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