11A3 An Evaluation of short-term MOS and short-term Ensemble MOS Temperature Forecasts

By

Richard H. Grumm[1], Ron Holmes, and Joe Villani
National Weather Service Office,
State College, PA 16803

1. INTRODUCTION

Model Output Statistics (MOS) in weather forecasting was first demonstrated by Glahn and Lowry (1972). Their technique, still employed today, consisted of statistical relationships between predictands and variables. The variables were derived from numerical model output at discrete forecast times and the predictands were sensible weather elements such as maximum and minimum temperatures, dew points, cloud amounts, surface winds, and the probability of precipitation. Regression was employed to determine the value of the predictand from the model forecast variables. Initially MOS was based off the sub-synoptic advection model (SAM) and the primitive equation model (PEM). Glahn and Lowry (1972) verified their MOS forecasts and concluded that it was a useful technique in weather forecasting. Glahn and Bocchieri (1976) tested MOS equations on the Limited-Area Fine Mesh Model (LFM) forecasts of probabilities of precipitation (PoP). The LFM forecasts were comparable to PEM forecasts and facilitated the implementation of LFM PoP forecasts. The LFM was implemented in 1971.The LFM-MOS, was implemented in 1976 (Gerrity 1977), and was used for nearly 20 years until the discontinuation of the LFM-MOS on 28 February 1996.

MOS equations were adapted to run using output from the LFM (Gerrity 1977) and Nest Grid Model (NGM: Phillips 1979). Jacks and Rao (1985) examined LFM-based MOS temperature forecasts for Albany, New York from 1975-1981. They found a general warm and cold bias for low and high temperatures respectively. In a later study, Jacks et al. (1990) verified a wide range of NGM-MOS and LFM-MOS products. In May, 1987, the National Weather Service (NWS) implemented perfect prog equations to produce statistical forecasts from the NGM (Jensenius et al.1987). The NGM-MOS was instituted to replace the NGM-perfect prognosis in June of 1989 (Jacks et al. 1990). From a temperature forecasting perspective, the NGM-MOS was about equal in skill to the LFM-MOS guidance. However, for fields such as winds, clouds, and precipitation probabilities, the NGM-MOS showed some forecast skill advantage over the LFM-MOS product. This was likely the result of the finer detail and improved accuracy in prediction of the large scale flow by the higher resolution NGM compared to the older and coarser LFM.

[1]Corresponding author: Richard H. Grumm 227 W. Beaver Ave, State College, PA 16803

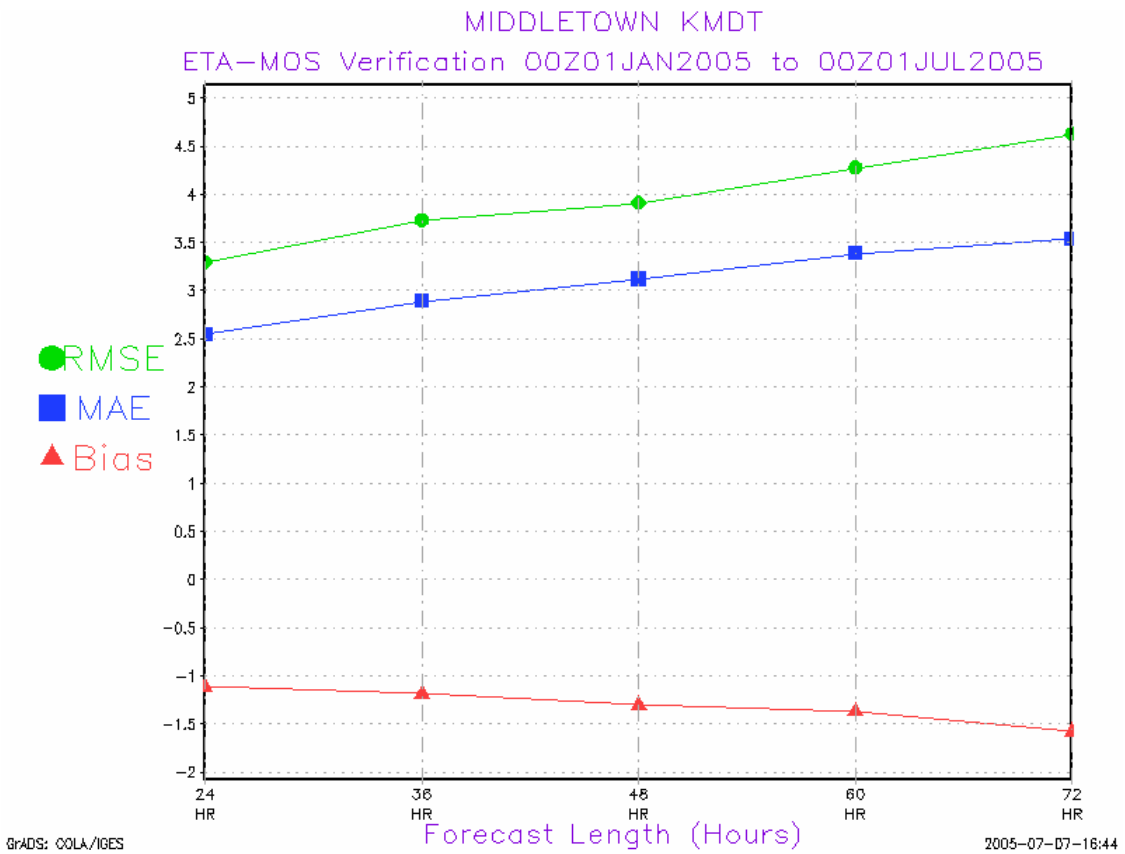

Figure 1. ETA-MOS temperature verification for Middletown (KMDT) for the period spanning 0000 UTC 1 January 2005 through 0000 UTC 1 July 2005. Bias, mean-absolute error (MAE), and root-mean square error (RMSE) are shown. Data are displayed by forecast length.

Erickson et al. (1991) demonstrated how new MOS equations were implemented in the upgraded Regional Analysis and Forecast System (RAFS). The NGM was the core forecast model of the RAFS. This paper showed how MOS had to be run and tested in parallel against the model changes to ensure consistency and at least comparable accuracy to the operational MOS products. This was an important aspect of MOS implementations as it provided a means to develop MOS with evolving models.

Vislocky and Fritsch (1995) demonstrated that a blend of the less skillful LFM-MOS with the NGM-MOS produced a more skillful forecast than either of the individual products. In a later study, Vislocky and Fritsch (1997) demonstrated the skill of consensus MOS in the National Collegiate Forecast contest. A simple NGM-MOS and AVN-MOS blended product was better than 97% of the forecasters in the contest. This ensemble-like product also used output from the Eta and NGM along with recent surface observations. This experiment paved the way for more ensemble MOS products. Woodcock and Engel (2005) demonstrated the improvements over MOS-based forecasts using operational consensus forecasts.

The concept of blending or producing an ensemble of MOS products, first demonstrated by Visclocky and Fritsch (1995) served as the basis for the development of a short-term ensemble MOS product using the operational MOS

products produced by the National Centers of Environment Prediction (NCEP). Short-term ensemble MOS is defined here as MOS products of 60-hours or shorter in length. This paper provides an evaluation of MOS forecasts and compares an ensemble blend to the three operational MOS products. This paper is divided in three sections. The first section describes the data and methods used. The second section provides some results, and the third section summarizes these results.

2. METHOD

MOS bulletins from the NGM, Eta, and the Global Forecast System (GFS) were collected in real-time. These bulletin's are also known as FWC, MET, and MAV bulletins. These data were decoded and stored in a relational database. The NGM-MOS[2] and GFS-MOS are stored by their bulletin names. Thus NGM-MOS and GFS-MOS images will show the bulletin names FWC-MOS and MAV-MOS respectively. The newer ETA-MOS is not identified by its bulletin name. The data included maximum and minimum temperatures; 3-hourly temperatures; 3-hourly dew points, wind speed, wind direction; probabilities of precipitation (POP) for 6-, 12-, and 24-hour forecasts, and weather type. These data were then extracted from the database to produce a consensus or blended MOS product based on these short-term MOS forecasts. The product was called the short-term ensemble MOS (STE-MOS). Initially, the 3 MOS products were averaged using equal weights at each location. However, verification showed that the GFS-MOS had highest skill and the NGM-MOS had the lowest skill. As a result, the weightings used in this study were 4, 3, and 1 for the GFS, Eta, and NGM MOS products respectively.

The Eta-MOS and NGM-MOS are produced twice daily at 0000 and 1200 UTC. The GFS-MOS is produced four times daily at 0000, 0600, 1200, and 1800 UTC. This facilitated the production of a lagged STE-MOS product using the 0600 and 1800 UTC blended with the older 0000 and 1200 UTC data. Stratifying the data by forecast cycle allowed for testing to see if any particular forecast cycle was more skillful than another. Stratifying the data by verification time (0000 and 1200 UTC) helped determine a bias in forecasting high and low temperatures.

In this study, images show the performance of the MOS products over the past 6 months. Tables will show data for the winter months only. This concept was employed to show the flexibility of the data base and highlight some seasonal MOS error trends.

The database allows for easy and automated production of verification statistics including mean-absolute error (MAE), bias, and root-mean squared errors. GrADS was used to produce graphical products of the skill measures including mean-absolute error (MAE), bias (BIAS), and root-mean square errors (RMS). The displays were produced at each station in Pennsylvania. The data were stratified by model to include NGM-MOS, Eta-MOS, GFS-MOS, and the STE-MOS. Data were also plotted by

[2] The legacy database contains the decoded "FWC" and "MAV" products from MDL thus images will show FWC-MOS and MAV-MOS, terms used interchangeably with NGM-MOS and GFS-MOS respectively. The image names are derived automatically from the database.

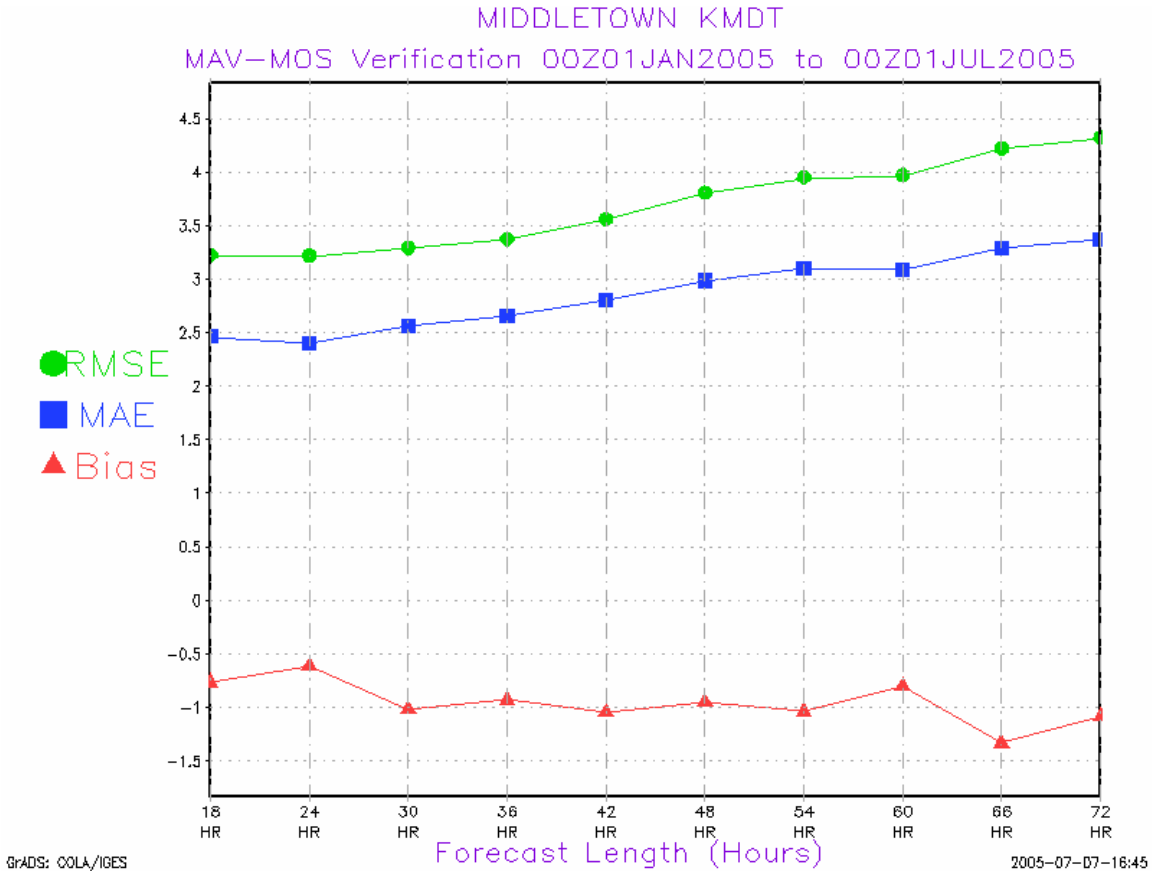

Figure 2. As in Figure 1 except GFS-MOS for Middletown with forecasts to 72-hours. The database table is called MAV and the image name shows "MAV-MOS" which is based on the GFS model.

forecast length showing errors, including MAE, bias, and RMSE at each forecast time. To identify high and low temperature errors, the data were further stratified by verifying cycle. This facilitated evaluating the bias and error differences between high and low temperature forecasts.

The bias was computed using the simple mean error as :

$$BIAS = \Sigma(F - O)/n, \qquad (1)$$

the MAE was computed as:

$$MAE = \Sigma(abs(F - O))/n, \qquad (2)$$

and the RMSE was computed as

$$STD = (1/n\Sigma((F - O))^2)^{1/2} \qquad (3)$$

where F is the forecast value and O is the observed value. The summations were taken from n=0 to n=n over the time periods indicated in the figures and tables.

3. RESULTS

Verification of ETA-MOS and GFS-MOS data for Middletown (KMDT) is shown in Figures 1 and 2 respectively. These data show the ETA-MOS and GFS-MOS bias, MAE, and STD of the temperature forecasts at KMDT for the forecast lengths shown. An overall cold bias is seen in Eta-MOS data. The Eta-MOS MAE ranges from 2.5 at 24 hours to around 3.2 at 72 hours. The bias is smaller in the GFS-MOS and in the short-length forecasts; the GFS-MOS had a smaller MAE than the Eta-MOS. These data suggest that the GFS-MOS temperature forecasts are more skillful

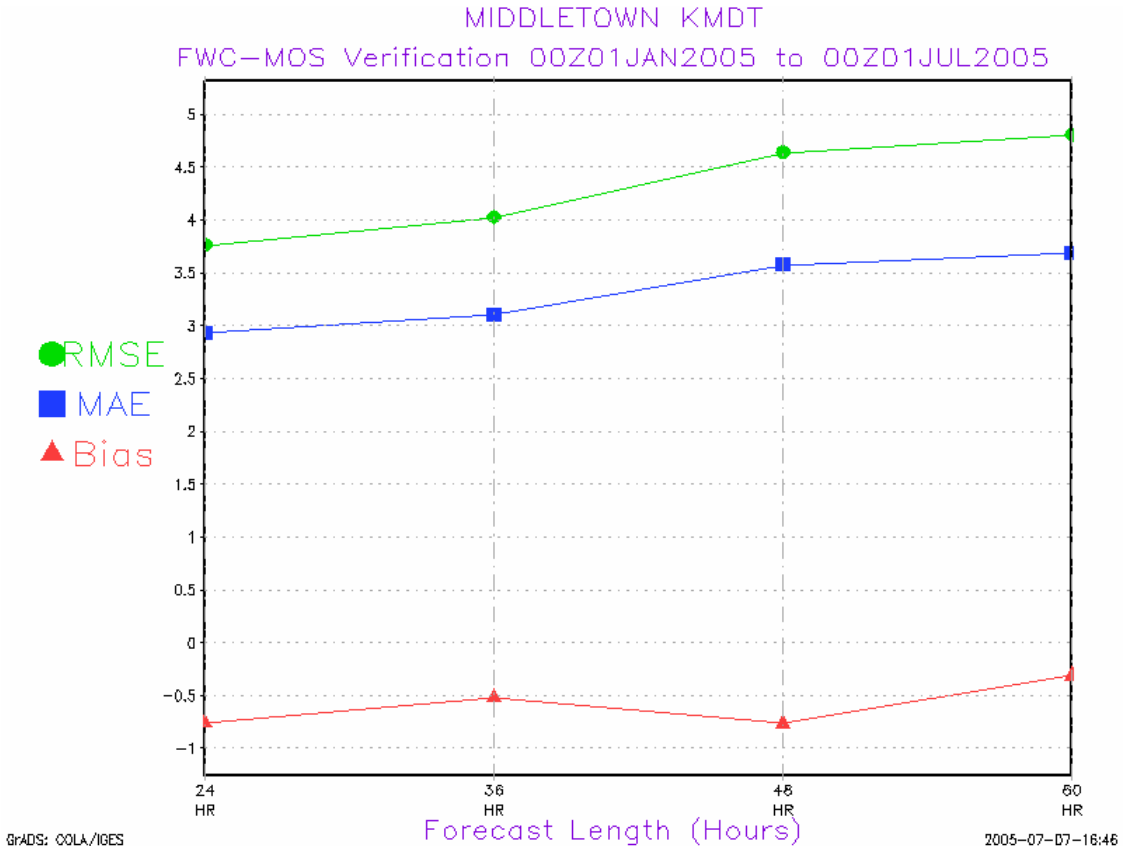

Figure 3. As in Figure 1 except NGM-MOS for Middletown. Similar to the GFS-MOS, this product is also known as FWC-MOS and the image name is derived from the database.

than those produced by the Eta-MOS. For comparison purposes, the NGM-MOS (also known as FWC guidance) is shown in Figure 3. These data show that overall, the NGM-MOS has larger errors than the Eta-MOS and GFS-MOS products at all forecast ranges. Only the 24 hour forecasts are of comparable skill to those produced by the other two MOS guidance products.

The STE-MOS for Middletown is shown in Figure 4. Overall, the MAE is smaller than all ETA-MOS and NGM-MOS based temperature guidance. Initially, 0600 and 1800 UTC MOS products were not used to compute an STE-MOS product. The impact of producing a 0600 and 1800 UTC STE-MOS products reduces STE-MOS skill. This is due to the effect of lagging in 12-hour old ETA- and NGM-MOS data. Data shown here reflect the more skillful 0000 and 1200 UTC based STE-MOS data. Overall, the STE product shows the benefits, at KMDT in using a consensus or ensemble approach.

The 30-day GFS-MOS for Bradford (KBFD) is shown in Figure 5. Overall, these data show less skill at KBFD than at KMDT. Though not shown, Eta-MOS and NGM-MOS showed a comparable degradation in forecast skill at KBFD. A few things of note include the large warm bias at KBFD relative to KMDT (see Fig. 2). This warm bias diminishes with forecast length. The overall MAE at KBFD was larger at all forecast times when compared to KMDT. Similar errors were found at other MOS sites.

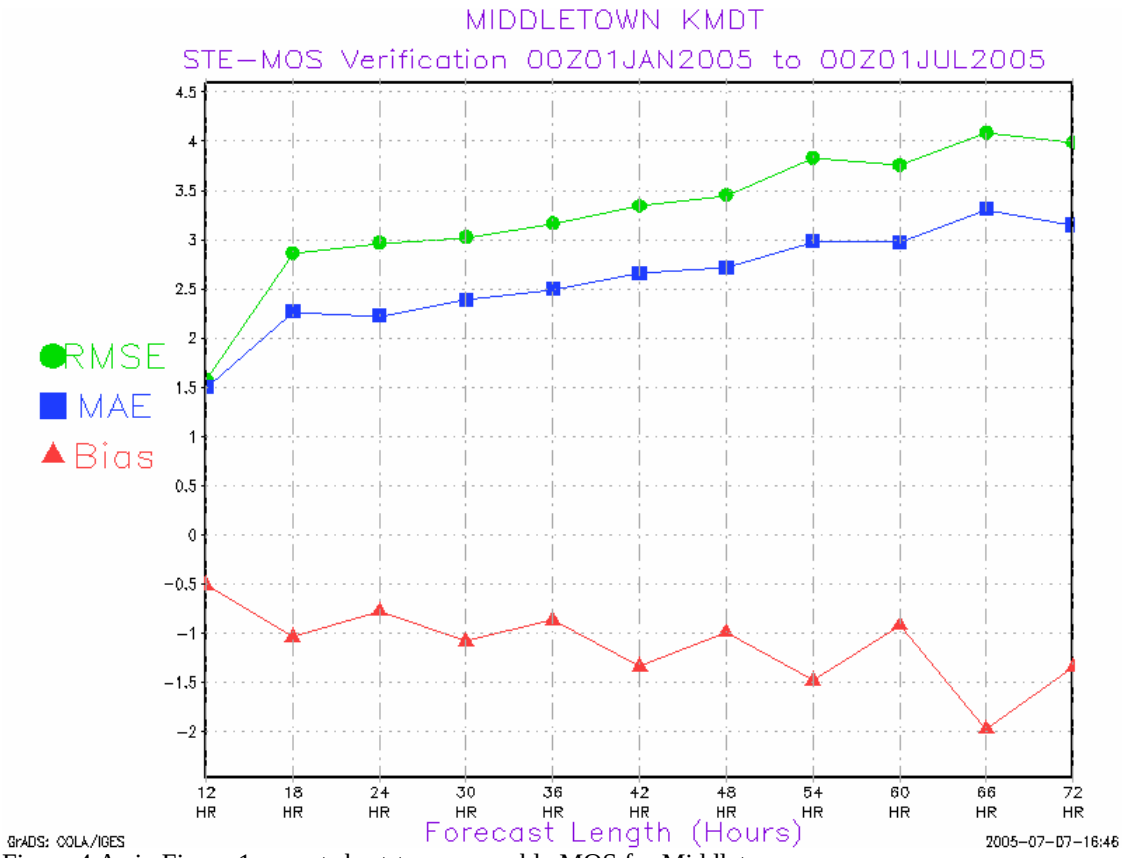

Figure 4 As in Figure 1 except short-term ensemble MOS for Middletown.

The skill scores for January 2004 and 2005 is shown in Table 1. These data show the GFS-MOS BIAS and MAE by forecast length for select stations. For the time period shown, the BIAS and MAE were generally lower at KMDT and KJST.

The STE-MOS data are shown in Table 2. These data, compared to Table 1 show that the GFS-MOS is the most skillful MOS product. Thus, blending the GFS-MOS with the less skillful MOS products produces and ensemble MOS of slightly less overall skill than the GFS-MOS.

Figure 6 shows the GFS-MOS temperature verification showing MAE by forecast length and forecast cycle. This allows comparison of each subsequent forecast cycle in 6-hour increments. For example, the 0000 UTC 36 hour forecasts show an MAE of 4.0, the subsequent 0600 UTC cycle shows an MAE of 3.8, the 1200 UTC cycle shows and MAE of 3.7 and the 1800 UTC cycle shows an MAE of 3.5. These data imply that each cycle offers an equal or better forecast than the previous forecast cycle for forecasts valid at the same time.

Figure 7 shows the GFS-MOS MAE and BIAS for forecasts verifying at 0000 and 1200 UTC by forecast cycle (0000,0600, 1200 and 1800 UTC). These data show that the MAE is around 3 for forecasts verifying at 0000 UTC and around 4 for forecasts verifying at 1200 UTC. There is a warm bias at both times with a larger warm bias in forecasts verifying at 1200 UTC. These data clearly show that over the period from 1 January to 1 July 2005, the GFS-MOS is more skillful at afternoon high temperature forecasts than morning low temperature

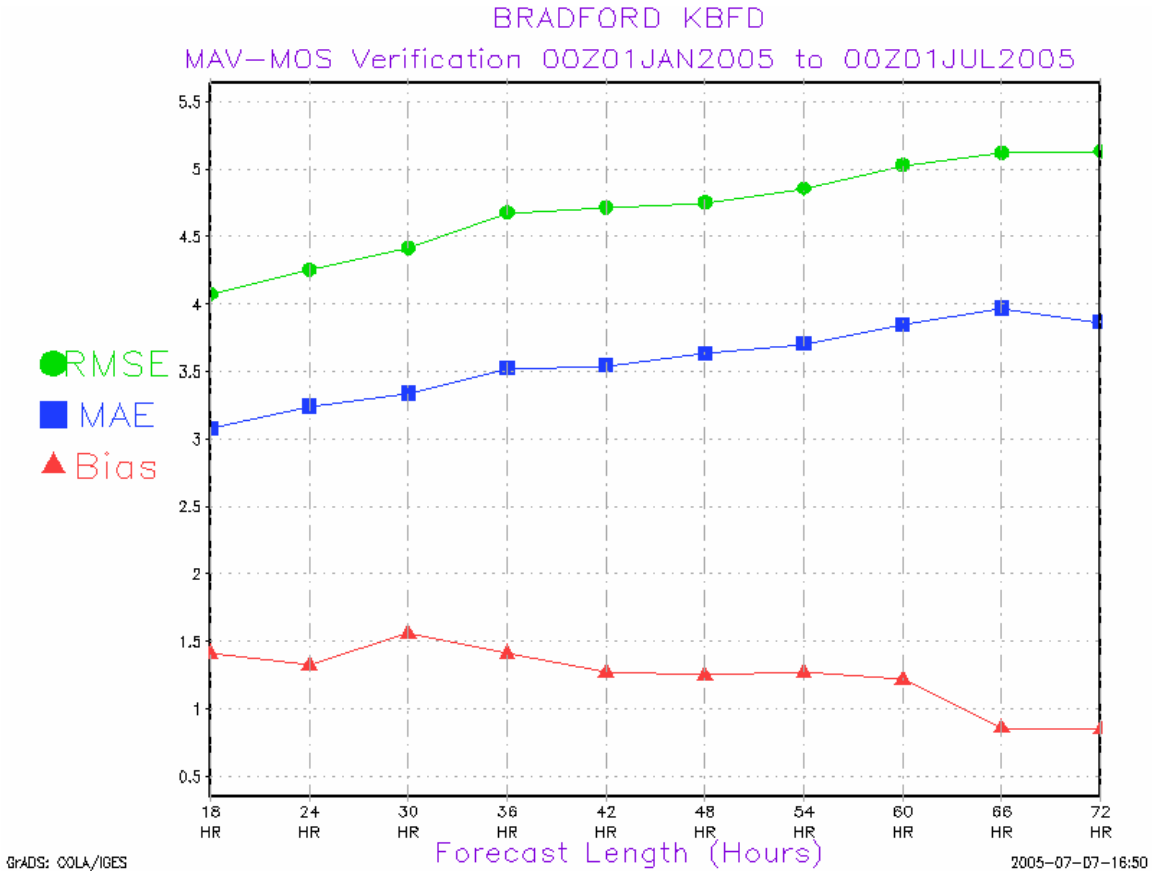

Figure 5. As in Figure 2 except GFS-MOS for Bradford (KBFD).

forecasts at KBFD. These data also suggest comparable skill from forecasts initialized at each cycle. Though not shown, with the exception of KMDT and Altoona (KAOO), this tendency for poor minimum temperature forecasts was evident at Bradford (Fig. 7), Williamsport (KIPT), Johnstown (KJST), and to a lesser degree, at University Park Airport (KUNV).

4. CONCLUSIONS

A study of MOS temperature forecasts was conducted to determine the MAE, bias, and RMSE of these forecast across central Pennsylvania. These data showed that overall the GFS-MOS was the most skillful MOS product. The older NGM-MOS was the least skillful MOS product. The Eta-MOS showed some seasonal variation in skill and at times was of comparable skill to the GFS-MOS. However in the long term the GFS-MOS was routinely the most skillful at temperature forecasting.

The STE product was more skillful than either the NGM-MOS or the Eta-MOS and offered little improvement over the more skillful GFS-MOS product. These results are not as promising as those shown by Vislocky and Fritsch (1995) where the ensemble mean MOS product was more skillful than the most skillful MOS product. In this study, the more skillful GFS-MOS relative to the Eta-MOS and older NGM-MOS was hard to improve upon using an ensemble technique. The results suggest that the NGM-MOS weighting should be decreased with increased forecast length. The STE-MOS offers the most benefit at forecasting low temperatures where all MOS biased products show diminished

skill and a pronounced bias at several stations.

Examining the data stratified by forecast verification time showed that the 0000 UTC forecasts were more skillful than the 1200 UTC forecasts. This suggests that the MOS equations are more accurate at high temperature than low temperature forecasting. At National verification sites, such as Harrisburg (KMDT) the overall MOS errors were lower than at other MOS sites. This suggests that the MOS equations have been tuned to these verification points. The MOS forecasts of low temperatures at Bradford (KBFD) showed large errors in low temperature forecasts with a distinct warm bias. KMDT had no significant bias in temperature forecasts while stations like KBFD and KAOO had significant bias in the forecasts. This was more significant in the low temperature forecasts.
From a forecast perspective, these data suggest that the GFS-MOS is the MOS product to beat. The overall skill of the GFS-MOS in forecasting maximum temperatures is better than the Eta-MOS and NGM-MOS forecasts. However, the Eta-MOS also shows considerable skill in this area. Forecasts of minimum temperatures show there is considerable variability and a lack of skill at certain sites which are accompanied by a strong warm bias at several sites. Isolating the conditions associated with these warm bias situations offers an opportunity of both further study and a means to improve upon MOS forecasts at locations where the strong warm bias and large MAE are present.

The data in Tables 1 & 2 and the graphs suggest seasonal trends in MOS verification. All MOS forecast products show smaller (larger) minimum (maximum) temperature error forecast errors in the warm season. For example, at Bradford (Figures 6 & 7) similar graphs for the winter months and by MOS product (not shown) indicated larger MAE's for forecasts valid at 1200 UTC and smaller MAE's for forecasts valid at 0000 UTC. Thus, during the cold season low temperature errors are larger than in the warm season. Similarly, there is a decrease in skill forecasting high temperatures in the warm season compared to the cold season. At the AMS meeting in August, seasonal and summarized results for 2005 will be presented.

Operational data and verification can be found at :
http://nws.met.psu.edu/verification/index.htm

This site contains 7- and 30-day MOS verifications of MEX, STE, GFS (MAV), ETA (MET), and NGM (FWC) MOS data.

5. ACKNOWLEDGEMENTS

We would like to thank MDL for access to the MOS product and Kenneth Johnson of ER/SSD for reviews and comments during the preparation of this paper.

6. REFERENCES

Bermowitz, R.J., 1975: An Application of Model Output Statistics to Forecasting Quantitative Precipitation. *Mon. Wea. Rev.*, **103**,149–153.

Carter, G.M., and H.R. Glahn, 1976: Objective Prediction of Cloud Amount Based on Model Output Statistics**.** *Monthly Weather Review*,**104**,1565–1572.

Erickson,M,C, J. B. Bower, V J. Dagostaro, J. Dallavalle, E Jacks, J. S. Jensenius Jr. and J. C. Su. 1991:

Evaluating the Impact of RAFS Changes on the NGM-Based MOS Guidance**.** *Weather and Forecasting*,**6**,142–147.

Gerrity, J. P. 1977: The LFM model–1976: A documentation. *NOAA Tech. Memo.* **NWS NMC-60**, National Oceanic and Atmospheric Administration, U.S. Department of Commerce,68 pp.

Glahn, H.R and J.R. Bocchieri. 1976: Testing the Limited Area Fine Mesh Model for Probability of Precipitation Forecasting**.** *Mon. Wea. Rev.*, **104**,127–132.

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.

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.

Jacks, E., and S. T Rao, 1985: An Examination of the MOS Objective Temperature Prediction Model. *Mon. Wea. Rev.*,**113**,134–148

Jensenius, J.S. JR,G.M.Carter, J.P.Dallavalle, and M.C Erickson, 1987: Perfect Prog maximum/miniumum temperatures, probability of precipitation, cloud amount, and surface wind guidance based on the output from the Nested Grid Model (NGM). Tech. Procedure Bulletin 369, 12pp.

Karl, T.R., 1979: Potential Application of Model Output Statistics (MOS) to Forecasts of Surface Ozone Concentrations. *J. Appl. Meteor.*, **18**,254–265.

Klein, W.E., and G.A. Hammons, 1975: Maximum/Minimum Temperature Forecasts Based on Model Output Statistics. *Mon. Wea. Rev.*, **103**, 796–806.

Phillips, N. A., 1979: The Nested Grid Model. *NOAA Tech. Rep.*, **NWS 22**, National Oceanic and Atmospheric Administration, U.S. Department of Commerce ,80 pp.

Vislocky, R.L., and J. M. Fritsch, 1995: Improved Model Output Statistics Forecasts through Model Consensus. *Bull. Amer. Meteor. Soc.*,**76**,1157–1164.

Vislocky, R.L., and J. M. Fritsch, 1997: Performance of an Advanced MOS System in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*,**78** ,2851–2857.

Woodcock, F. and C. Engel, 2005: Operational Consensus Forecasts. *Wea. Forecasting*, **20**, 101-111.

Zurndorfer, E.A., J.R. Bocchieri, G.M. Carter, J. P. Dallavalle, D.B. Gilhousen, K.F. Hebenstreit, and D.J. Vercelli, 1979: Trends in Comparative Verification Scores for Guidance and Local

Aviation/Public Weather Forecasts.
Mon. Wea. Rev., **107**, 799–811.

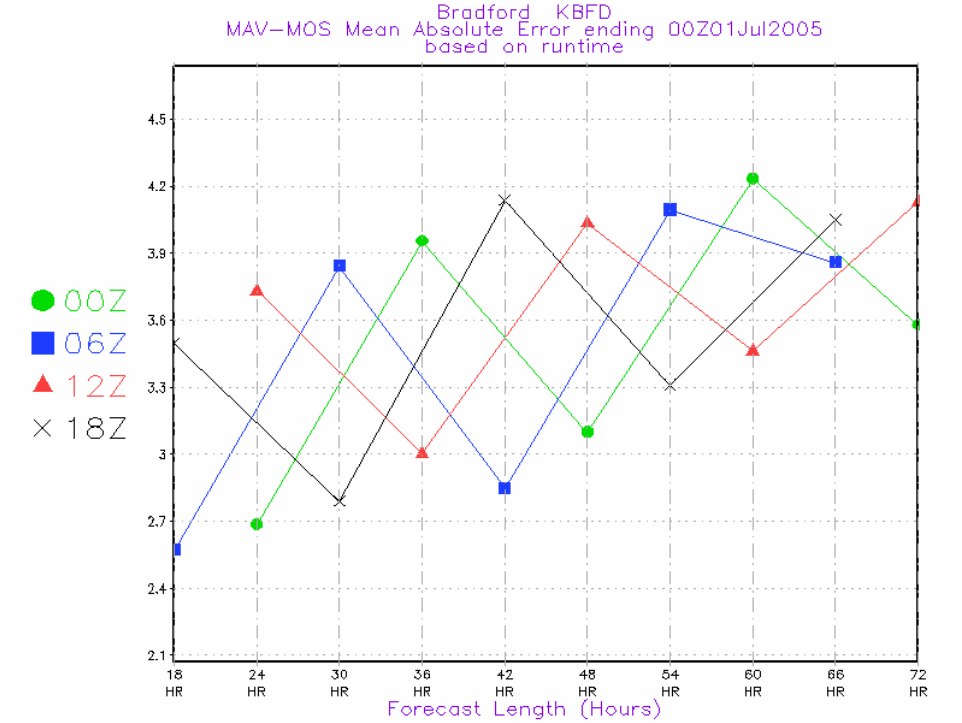

Figure 6. GFS-MOS temperature verification showing MAE by forecast length for the period of 0000 UTC 1 January 2005 through 0000 UTC 1 July 2005. Each line shows the MAE by forecast cycle. This allows comparison of each subsequent forecast cycle in 6-hour increments.

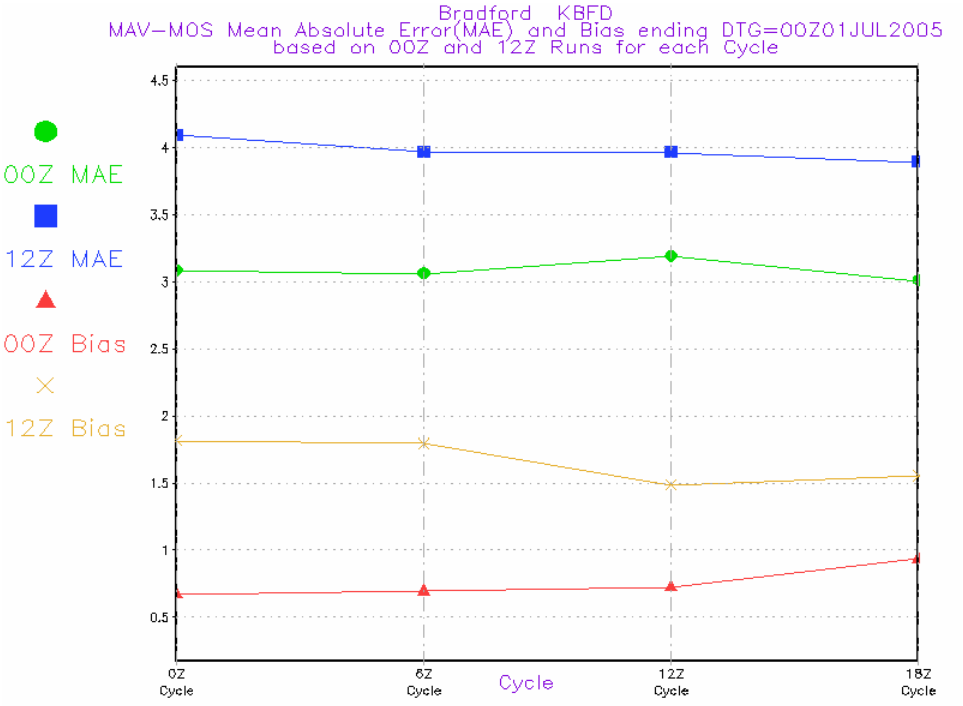

Figure 7. MAV temperature forecasts for Bradford showing MAE and BIAS for forecasts verifying at 1200 and 0000 UTC daily for forecasts initialized at 0000, 0600, 1200, and 1800 UTC.

January 2004-2005 GFS MOS Errors

TIME	KBFD		KMDT		KAOO		KIPT		KJST	
Forecast Length	BIAS	MAE	BIAS	MAE	BIAS	MAE	BIAS	MAE	BIAS	MAE
24	1.62	3.52	0.53	2.97	1.41	3.28	1.09	3.11	0.22	2.51
36	1.79	3.60	0.69	2.95	1.62	3.40	1.23	3.55	0.33	3.17
48	1.82	3.80	0.49	3.22	1.46	3.55	1.25	3.49	0.30	3.10
60	2.47	4.75	0.75	3.54	1.92	4.17	1.29	3.77	0.66	3.69

Table. 1 Global Forecast System (GFS) based MOS errors for January 2004 and 2005. Data shown include the Station, bias, and mean absolute error by station and forecast length (hours).

January 2004-2005 STE MOS Errors

TIME	KBFD		KMDT		KAOO		KIPT	
Forecast Length	BIAS	MAE	BIAS	MAE	BIAS	MAE	BIAS	MAE
24	1.89	3.49	1.34	3.14	1.96	3.29	1.66	3.07
36	1.94	3.56	1.14	3.50	2.31	3.38	1.95	3.55
48	2.05	3.91	1.10	3.51	2.16	3.63	1.69	3.68
60	2.96	4.56	1.16	3.49	2.89	4.39	1.79	3.87

Table. 2 Short-term ensemble MOS errors for January 2004 and 2005. Data shown include the Station, bias, and mean absolute error by station and forecast length (hours).