11A6 An Evaluation of Ensemble MOS Temperature Forecasts from the Medium Range Ensemble Forecast System

By

Richard H. Grumm[1] and Joe Villani
National Weather Service Office,
State College, PA 16803
and
Robert Hart
The Florida State University
Tallahassee, Florida

1. INTRODUCTION

Uncertainties in initial conditions and the growth of model errors in time introduce large uncertainties in weather forecasts at longer ranges. In general terms, the predictability increases as the scale of the feature of interest increases (Dalcher and Kalnay 1987; Droegemeier 1997). The use of an ensemble of initial conditions and an ensemble of forecast outcomes is one method to account for uncertainty in weather forecasting. There is a plethora of published research about ensemble forecasting and ensemble forecast methodologies (Du et al. 1997; Zhang and Krishnamurti 1997; Roebber et al. 2005). The value of consensus as a skillful forecast tool has been demonstrated for years (Woodcock and Engel 2005; Fritsch et al 2000;Vislocky and Fritsch 1995). The concept of consensus forecasts using Model Output Statistics (MOS: Glahn and Lowry 1972) was demonstrated by Vislocky and Fritch (1995).

In 1999 the Meteorological Development Laboratory (**MDL:**formerly Techniques Development Laboratory**)** began producing extended range MOS bulletins from the 0000 UTC cycle of the National Centers for Environmental Predictions (NCEP) Medium Range Ensemble Forecast System (MREF:Toth et al. 1997;Tracton and Kalnay 1993). These MOS bulletins were produced for the operational high resolution deterministic Global Forecast System (GFS) run, the MREF control run, and the 5 positively and 5 negatively perturbed forecast members. A total of 12 complete MOS bulletins were produced. A consensus mean forecast bulletin was also produced allowing users to easily compare the deterministic GFS based MOS (hereafter GFS-MOS) to the ensemble mean and range of critical forecast parameters, such as temperatures and probabilities of precipitation. A fire at the NCEP super computing center on 27 September 1999 led to a temporary loss of these bulletins. The 12 individual ensemble member bulletins were back in production by 2001 but the production of the ensemble mean forecast bulletin did not resume until 2004. In September 2001, the National Weather Service in State College and the Pennsylvania State University began producing ensemble MOS bulletins in real-time.

The value of MOS in weather forecasting was first demonstrated by Glahn and

[1]Corresponding author address: Richard H. Grumm, NOAA/NWS, 227 W. Beaver Avenue, State College, PA 16801

Lowry (1972). These techniques are still employed today, consisting 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 (National Weather Service 1971).The LFM-MOS, which was implemented in 1976 (Gerrity 1977) was used for nearly 20 years until the discontinuation of the LFM-MOS on 28 February 1996.

MOS equations were adapted to run from output from the LFM (1976) 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 was 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.

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 insure consistency and at least comparable accuracy to the operational MOS products. This was an important aspect of MOS implementations as new models and model changes were increasing dramatically in the late 1980s and 1990s.

Vislocky and Fritsch (1995) demonstrated that a blend or consensus of the less skillful LFM-MOS with the NGM-MOS produced a more skillful forecast than either of the two products. In a later study, Vislocky and Fritsch (1997) demonstrated the skill of consensus MOS in the National Collegiate Forecast contest. A simple blend of NGM-MOS and AVN-MOS 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 purpose of this paper is to evaluate the value of producing a consensus forecast for the extended range GFS based MOS data. The basic concept is similar to the production of consensus MOS forecasts first demonstrated by Visclocky and Fritsch (1995). This paper is divided into three sections. The first section describes the methods and data used in this study, including means to evaluate skill. The second section presents the results of the study, and the final section discusses and summarizes the results.

2. METHOD

i. data used in this study

In September 2001, all 12 MDL ensemble MOS bulletins were decoded to produce an ensemble MOS product. Table 1 lists the 12 MOS bulletins used to produce the ensemble MOS product. All MOS bulletins used were retrieved as basic text formatted products. The ensemble product included the variables listed in Table 2. The product was called Ensemble MOS (ENSMOS) and was made available on the world-wide web in both a graphical and text format in late September 2001.

In addition to making the ENSMOS available in near-real time, the data were archived. In 2003 these data were placed in a relational database to facilitate verification of the individual MOS bulletins and the ENSMOS product. The current database contains a table for each of the 12 MOS bulletins forecasts and a table of select ENSMOS products. The current verification is limited to 12-hour temperatures and probability of precipitation forecasts.

The database allows for easy and automated production of temperature verification statistics including the bias, the mean-absolute error (MAE), and root-mean squared error (RMSE). The Grid Analysis and Display System software (GrADS; Doty and Kinter 1995) was used to produce graphical products of the skill measures. The displays were produced at each station and stratified by season. The 4 primary seasons were defined as winter (December-February), spring (March-May), summer (June-August), and autumn (September-November).

The common displays, showing all 12 members plus the consensus used a simple color scheme. All positively and negatively perturbed members were plotted in red and blue respectively. The operational GFS MOS was plotted in thick black, the ensemble control run was plotted in green, and the ensemble mean or consensus forecast, was plotted in gray. For brevity, comparisons are primarily limited to the GFS-MOS, the control MOS (hereafter CONMOS), and the ensemble MOS.

In addition to the traditional skill scores, defined below, tests were conducted to determine how often the observed temperature fell within the range of the ensemble members. Frequencies were computed to determine the percentage of time the observed temperature 1) was colder than the coldest ensemble member, 2) warmer than any ensemble forecast member, and 3) was within the range of the ensemble MOS forecasts.

ii. measures of skill

The bias was computed using the simple mean error as :

$$\text{BIAS} = \Sigma(F - O)/n \qquad (1)$$

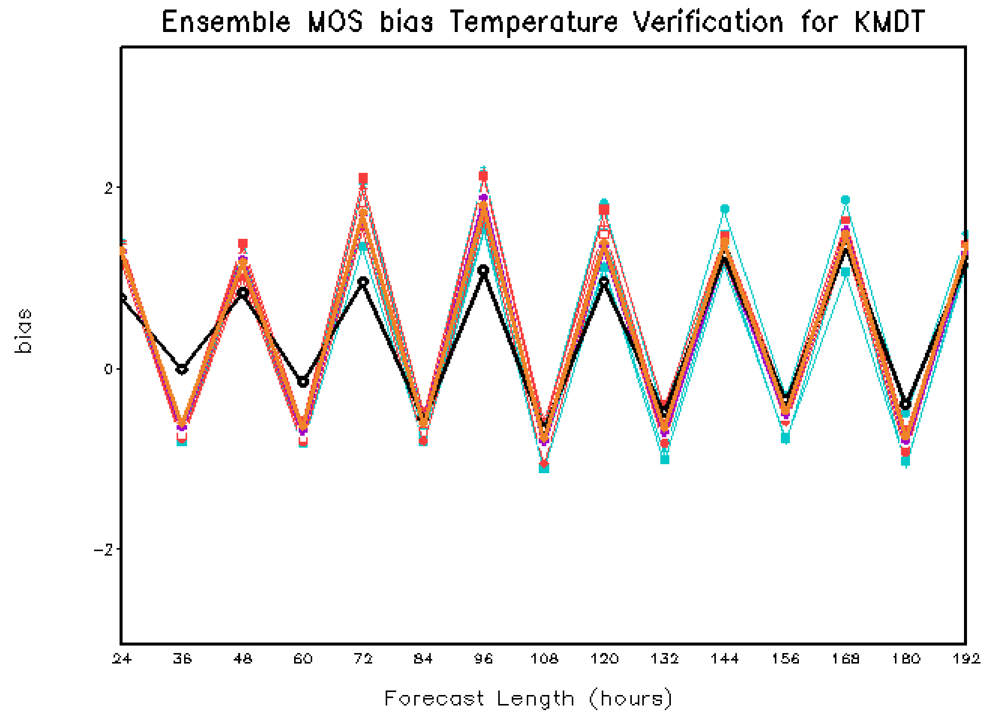

Figure 1. Bias scores for all GFS MOS members from 1 December 2004 through 28 February 2005 at Harrisburg, Pennsylvania (KMDT). Positively perturbed members are shown in red, negatively perturbed members are shown in blue. The thick black lines shows the high resolution, deterministic GFS-MOS, the thick purple line shows the low-resolution ensemble control run, and the thick gold line shows the ensemble blend or consensus forecast

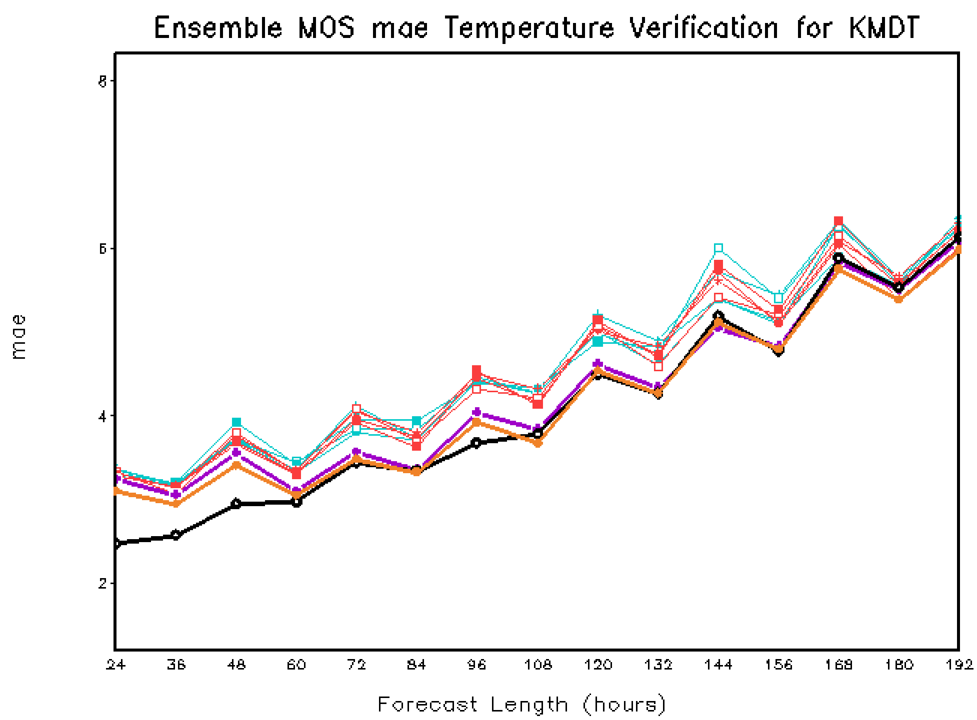

Figure 2. As in Figure 1 except showing mean-absolute errors for Harrisburg.

The MAE was computed as:

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

And the RMSE was computed as:

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

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

Figures 1-3 show the ENSMOS verification for Middletown, Pennsylvania (KMDT) showing the BIAS, MAE, and RMSE respectively for the winter of 2004-2005. The bias (Fig. 1) has seriated appearance due to diurnal fluctuations. Generally, there is a larger bias for all forecasts valid at 0000 UTC compared to forecasts valid at 1200 UTC. Initially, the GFS-MOS has a smaller bias through around 96 hours. At longer ranges, the GFS-MOS has a warm bias. The ENSMOS, the mean of all forecasts, has a bias that represents the average of all forecasts and is therefore along the center of the pack. Interestingly, the low-resolution control run has a central bias tendency similar to that shown by the consensus forecasts. The MAE and RMSE show that for the first 72 hours, the high resolution GFS-MOS has the smallest MAE and RMSE. The negatively perturbed members appear to have the overall larger MAE's and RMSE's. The ENSMOS has a smaller error than the perturbed and control members at all time periods and is of comparable skill to the GFS-MOS after 120 hours. The fact that at least one positively perturbed member shows more skill than the GFS-MOS and the ENSMOS at longer ranges suggests the validity of using an ensemble MOS technique at longer forecast ranges.

Though not shown, at all 6 MOS sites in central Pennsylvania, the largest RMSE and MAE's were associated with negatively perturbed members. At Altoona, Bradford (Fig. 4), and Johnstown, the ENSMOS often had slightly smaller MAE and RMSE values than the GFS-MOS. The MAE data at Bradford show that both the CONMOS and ENSMOS had smaller MAE's than the GFS-MOS. The main advantage was in the minimum temperature forecasts. These data also displayed the overall trend for larger errors with the negatively perturbed members. Errors at Williamsport (not shown) were similar to Harrisburg.

The plan view display of the BIAS, MAE, and RMSE for the period of 1 January 2004 through 31 December 2004 for 120 hour forecasts is shown in Figure 5. These data show that the high resolution GFS-MOS and CONMOS are of comparable skill. At several sites in western Pennsylvania, the CONMOS had slightly better skill scores than the higher resolution GFS-MOS. Though not shown, similar results were found at all forecast lengths.

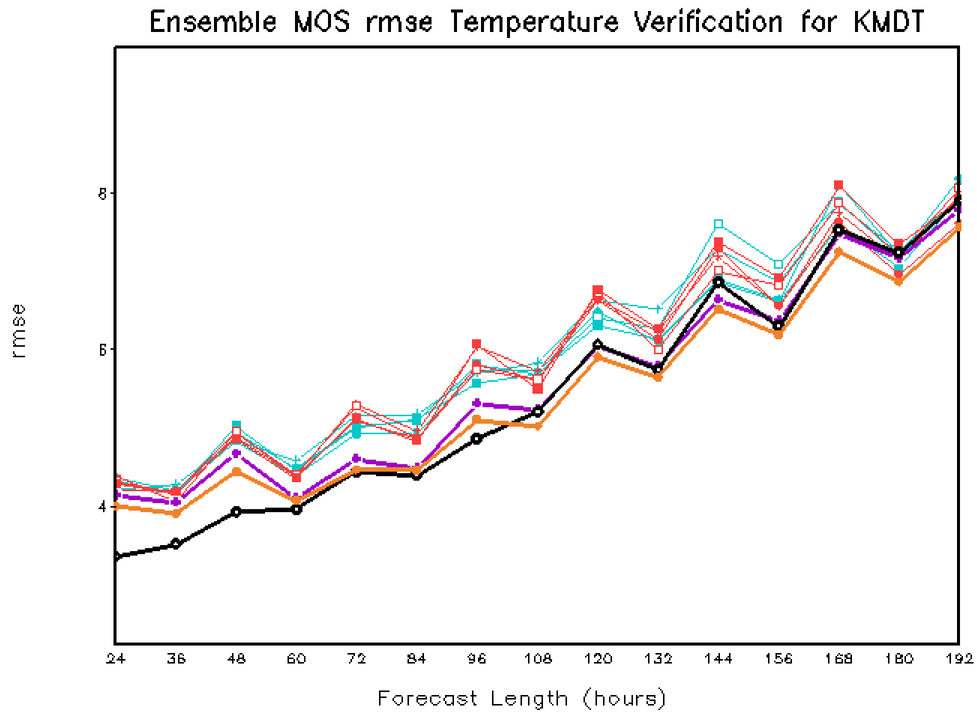

Figure 3. As in Figure 1 except showing root-mean square errors for Harrisburg.

Table 3 shows the frequency when observed temperatures fell within, above, and below the range of the ensemble MOS temperatures forecasts. These data are valid for Harrisburg, Pennsylvania (KMDT). This type of data was examined for other sites across Pennsylvania. Similar to Harrisburg, at most sites, the observed temperature fell within the forecast range 40 to 60% of the time at all forecast projections. There was a slight tendency for the observed temperature to be lower than all ensemble members more often than for the observed temperature to be warmer than the all ensemble members. The overall warm bias is reflected in these data.

4. CONCLUSIONS/DISCUSSION

Verification of temperature forecasts showed that at locations such as Williamsport and Harrisburg, the high resolution GFS-MOS was more skillful for the first 24-96 hours at forecasting temperatures. In western Pennsylvania, the ENSMOS and CONMOS often had lower MAE and RMSE values than the GFS-MOS at all time periods. An examination of short-term MOS products (not shown) revealed a large warm bias in low temperature forecasts in western Pennsylvania.

At most sites, the higher resolution GFS-MOS has an advantage in the forecasts from 12-96 hours. This suggests that the coarser models are often not as skillful at these time ranges. This demonstrates the value of having a high resolution model and the need to consider weighting ensemble forecasts stronger with the more skillful deterministic model.

An encouraging result is that at longer ranges, a perturbed member can have lower MAE’s and RMSE’s than either the GFS-MOS and the ENSMOS. This suggests that at longer ranges, the operational model is not routinely the most skillful model.

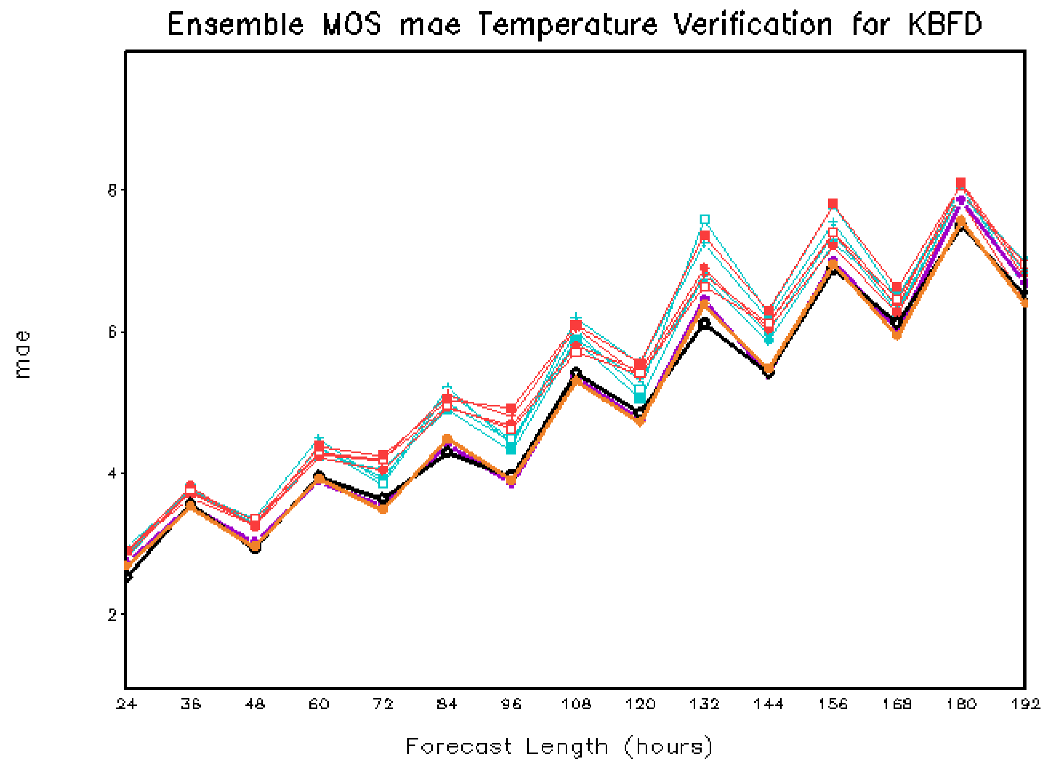

Figure 4. As in Figure 1 except mean-absolute errors for Bradford, Pennsylvania for the period 1 December 2004 through 28 February 2005.

The fact that observed temperatures often fall outside the range of the 12 members suggests that there is a lack of diversity in the current MREF system.

5. ACKNOWLEDGEMENTS

Meteorological Development Laboratory for providing data lost by our local real-time feed. This provided a complete dataset for 2004. The WFO in Taunton, Massachusetts for providing data to examine results in New England. We would like to thank Josh Watson of ER/SSD for reviewing and providing comments on this preprint.

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. *Mon. Wea. Rev*,**104**,1565–1572.

Dalcher, A. and E. Kalnay, 1987: Error Growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474-491.

Doty, B.E. and J.L. Kinter III,1995: Geophysical Data Analysis and Visualization using GrADS. Visualization Techniques in Space and Atmospheric Sciences, eds. E.P. Szuszczewicz and J.H. Bredekamp. (NASA, Washington, D.C.), 209-219.

Droegemeier, K.K., 1997: The numerical predictions of thunderstorms: Challenges, potential benefits, and results from real-time operational tests. WMO Bulletin, **46**, 324-336.

Du, J., S.L. Mullen and F. Sanders. 1997: Short-Range Ensemble Forecasting of Quantitative Precipitation. *Mon. Wea. Rev*,**125**,2427–2459.

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.

Fritsch, J.M., J. Hilliker, J. Ross and R. L. Vislocky. 2000: Model Consensus. *Wea. Forecasting*, **15**,571–582.

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. of App. Meteor.*,**11**,1203–1211.

Hart, K.A,W.J. Steenburgh , D. J. Onton, and A.J. Siffert. 2004: An Evaluation of Mesoscale-Model-Based Model Output Statistics (MOS) during the 2002 Olympic and Paralympic Winter Games. *Wea. Forecasting*,**19**,200–218.

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 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). Technical Procedure Bulletin 369, National Weather Service (NOAA), Silver Spring, Maryland,12pp.

Karl, T.R. 1979: Potential Application of Model Output Statistics (MOS) to Forecasts of Surface Ozone Concentrations. *J. of App. 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.

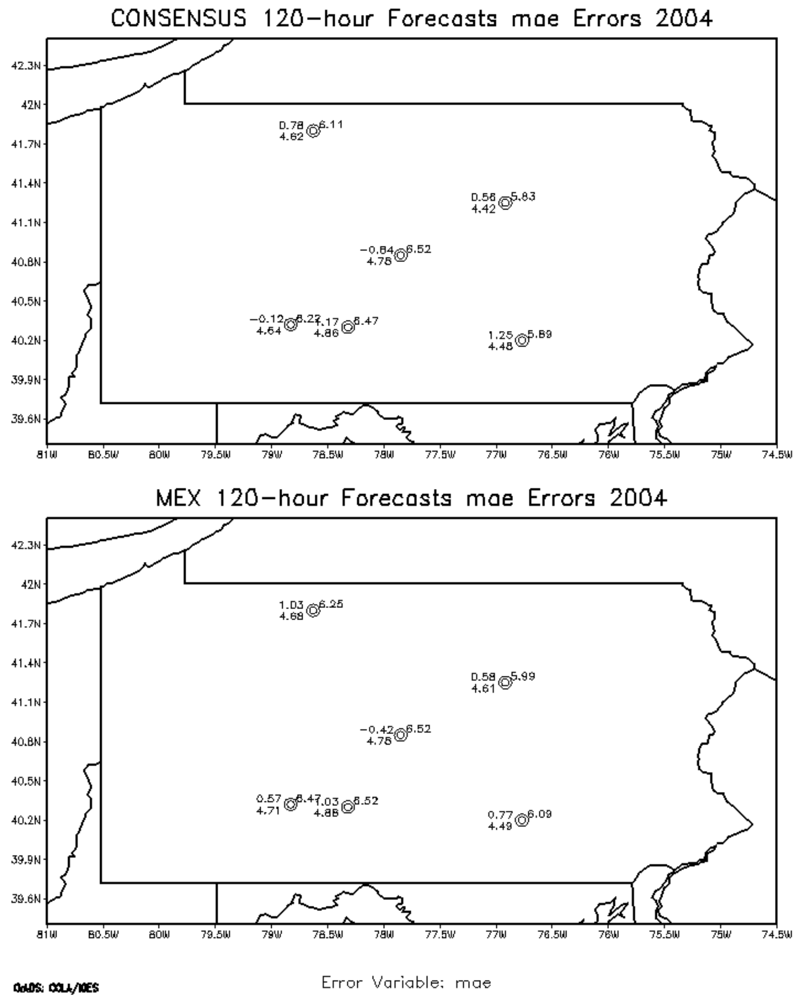

Figure 5. Plan view display of temperature error scores for all 120-hour forecasts issued from 1 January 2004 through 31 December 2004. Data include BIAS, MAE, and RMSE. BIAS is in the upper left, RMSE is in the upper right, and MAE is in the lower left site of the station. Upper panel shows the consensus scores and the lower panel the GFS-MOS (MEX) scores.

National Weather Service, 1971: The Limited-area Fine Mesh (LFM) model. NWS Tech. Proc. Bull., No. 67, NOAA, 11pp.

Phillips, N. A., 1979: The Nested Grid Model.*NOAA Tech. Report,* **NWS 22**, U.S. Department of Commerce, 80 pp.

Tracton, M. S. and E. Kalnay, 1993: Ensemble forecasting at NMC: Operational implementation. *Wea. Forecasting*, 8, 379-398.

Toth, Z., E. Kalnay, S. M. Tracton, R. Wobus and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting*, 12, 140-153.

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

Roebber, P.J, D. M. Schultz, B. A. Colle and D.J. Stensrud. 2004: Toward Improved Prediction: High-Resolution and Ensemble Modeling

Systems in Operations. *Wea. Forecasting*,**19**,936–949.

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, Engel C (2005) Operational Consensus Forecasts. *Wea. Forecasting.* **20**,101-111.

Zhang, Z.and T. N. Krishnamurti. 1997: Ensemble Forecasting of Hurricane Tracks**.** *Bull. Amer. Meteor. Soc.*,**78**,2785–2795.

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.

MDL Ensemble MOS Bulletins		
Member	**Bulletin For http retrieval**	**Bulletin Description**
MEX	mdl_mrfmex.txt	High resolution GFS MOS bulletin
C0MEX	mdl_ensc0mex.txt	MOS based off the control run of the MREF
p1mex	mdl_ensp1mex.txt	MOS based off the first positively perturbed member of the MREF
p2mex	mdl_ensp2mex.txt	MOS based off the second positively perturbed member of the MREF
p3mex	mdl_ensp3mex.txt	MOS based off the third positively perturbed member of the MREF
p4mex	mdl_ensp4mex.txt	MOS based off the fourth positively perturbed member of the MREF
p5mex	mdl_ensp5mex.txt	MOS based off the fifth positively perturbed member of the MREF
n1mex	mdl_ensn1mex.txt	MOS based off the first negatively perturbed member of the MREF
n2mex	mdl_ensn2mex.txt	MOS based off the second negatively perturbed member of the MREF
n3mex	mdl_ensn3mex.txt	MOS based off the third negatively perturbed member of the MREF
n4mex	mdl_ensn4mex.txt	MOS based off the fourth negatively perturbed member of the MREF
n5mex	mdl_ensn5mex.txt	MOS based off the fifth negatively perturbed member of the MREF

Table 1 List of medium range ensemble members available, the file names for data retrieval and a description of each MOS bulletin. All bulletins are available once a day based on the 0000 UTC forecast cycle.

ENSEMBLE MOS VARIABLES

VARIABLE	**DESCRIPTION**	ENSEMBLED	REMARKS
TMAX/TMIN	12-hour maximum and minimum temperatures	YES	Arithmetic averaged
TEMP	Temperature at specified hour	YES	Arithmetic averaged
DWPT	Dew point at specified hour	YES	Arithmetic averaged
POP12	12-hour probability of precipitation	YES	Arithmetic averaged
POP24	24-hour probability of precipitation	YES	Arithmetic averaged
CLOUDS	Cloud Amount Category (CLEAR, PARTLY CLOUDY, CLOUDY)	YES	Translated to integers then Arithmetic averaged
QPF12	12-hour quantitative precipitation category	YES	Arithmetic averaged
QPF24	24-hour quantitative precipitation category	YES	Arithmetic averaged
WIND	Wind speed	YES	Arithmetic averaged
TS12	12-hour thunderstorm probability	YES	Arithmetic averaged
TS24	24-hour thunderstorm probability	YES	Arithmetic averaged
TYPE	Weather Type	YES	Translated to integers then Arithmetic averaged

Table 2. List of available ensemble MOS variables. Table includes the variable name, a description of the variable, whether or not the variable is used to produce ensemble output, and a brief description of the ensemble method.

STATION	Forecast Length	Number of observations	Observed temperature within forecast range	Observed temperature higher than the forecast maximum	Observed temperature lower than the forecast minimum
KMDT	24	530	50	18	32
KMDT	36	521	45	29	26
KMDT	48	527	54	19	27
KMDT	60	518	49	27	24
KMDT	72	524	60	14	26
KMDT	84	515	56	25	19
KMDT	96	521	57	16	26
KMDT	108	512	58	25	17
KMDT	120	518	56	19	24
KMDT	132	509	53	26	20
KMDT	144	515	52	20	27
KMDT	156	506	49	28	23
KMDT	168	512	50	21	29
KMDT	180	503	44	32	23
KMDT	192	509	41	25	23

Table 3. Frequency (percent) of the time, by forecast length, that the observed temperature was within, above, and below the range of ensemble forecast value of temperature the 12-hour maximum or minimum temperature. Data are valid only for the maximum or minimum 12-hour temperature forecast for the period 1200 UTC 1 January 2004 through 1200 UTC 20 June 2005.