3.6

ENSEMBLE DYNAMIC MOS

Peter P. Neilley
Weather Services International, Inc.
Billerica, MA

William Myers and Gregory Young
National Center for Atmospheric Research
Boulder, CO

1. Introduction

Model Output Statistics (MOS) have been used to derive forecasts of surface weather parameters from numerical weather prediction (NWP) models for over 30 years. Following the pioneering work of Glahn and Lowry (1972), the MOS technique determines relationships between explicit NWP model data and weather parameters needed to produce a weather forecast. Climatological data, recent observations and other data external to the NWP model are also often used in deriving the relationships. In most circumstances, the relationships are linear, with multivariate linear regression being the most common approach used to generate the relationships. Variants on the approach include the use of logistic regression (Glahn et al., 1991) when predicting binary parameters (e.g. occurrence of precipitation) and generalized additive models (Vislocky and Fritch, 1995a).

During the early era of NWP, the use of MOS was essential in producing quantitatively useful guidance of the near-surface weather from the models. Direct extraction of data from the models typically yielded large discrepancies between the forecast and observed surface weather. There were several reasons for this. First, the models had relatively poor horizontal and vertical resolution, and thus did not explicitly produce forecasts sufficiently near at the location of interest. Second, the model's physics generally were insufficient to accurately simulate the weather immediately near the earth's surface where physics plays a paramount role in determining the weather. Third, some parameters that are important components of a weather forecast were not available directly from the NWP model output. For example, early NWP models did not include explicit ice physics and thus could not predict the occurrence of snow or freezing rain, neither could they explicitly predict the occurrence of thunder.

Corresponding author address: Peter P. Neilley, WSI, 900 Technology Park, Billerica, MA 01821-4167. email: pneilley@wsi.com

Today, despite tremendous advances in the quality of NWP models, the use of MOS continues to play a paramount role in the production of forecast guidance. The National Weather Service currently publishes MOS forecasts based on the NGM model and two versions of MOS based on each of the AVN and MRF runs of the global spectral model. It is perceived that MOS forecasts continue to be superior to those generated by direct extraction from NWP data (see e.g. http://205.156.54.206/tdl/synop/results.htm for current NWS statistics), although no recent comprehensive study of such is known.

It is general practice that at least two years of model data and observations be used to derive stable MOS relationships (Jacks et al., 1990; Vislocky and Fritch 1995b). The length of the period depends on the ability to find meaningful relationships between the predictors and predictands. Noise in the data introduced by errors in the NWP model's forecasts and errors in observations can adversely affect the ability to find such relationships and thus the length of the required training period. The degree of noise also is dependent on location, season and forecast parameter with binary parameters (e.g. occurrence of precipitation) having especially noisy datasets.

For strict derivation and application of MOS, the NWP model must be static during both the training period as well as during the subsequent period that the MOS relationships are applied. Otherwise, a paradox can arise in which improvements in the model lead to worse MOS forecasts. For example, if a MOS relationship acts to remove a systematic bias found in the training set of NWP data, then subsequent improvements that reduce the NWP model's bias will result in a biased MOS forecast until the MOS relationships can be rederived.

Today, NWP models are highly dynamic, with the major modeling centers making improvements in physics, numerics and resolution on a near continuous basis. Thus, strict application of the traditional MOS technique can be questionable. One approach for overcoming the ever-changing model base is to use of the so-called perfect-prog method. In this technique, the MOS relationships are based not on explicit NWP model data but on *observations* of the variables that a NWP model can predict. The relationships are then applied to NWP forecasts under the assumption that the model "perfectly" predicts the atmosphere. As the model is improved, and its predictions become more perfect, then the perfect-prog MOS forecasts are also improved. However, the technique is limited to use only those predictors for which observations are available. Thus, many complex model variables (e.g. cloud water concentration, surface fluxes, vertical velocity) are not available for use in perfect-prog MOS even though they may contain significant insight to a MOS forecast.

Recently, Mao et al. (1999) used a new technique to overcome this problem. Instead of relying on two years of data to construct the MOS relationships, they attempted to derive linear regression relationships using only recent (2-4 weeks) model data and observations. Their regressions were updated daily thus allowing the system to adjust to changes in the base model automatically. In their specific application, they used a version of the regional spectral model (RSM) to drive MOS temperature forecasts and found skill comparable (but slightly less) than forecasts from the National Weather Services NGM-based MOS forecasts. It is unclear whether the difference is skill was due to differences in the underlying quality of RSM and NGM models, or to limitations in their technique. However, their specific method for determining the regression equations and their technique for extracting maximum and minimum temperature forecasts from time series of temperature forecasts were both sub-optimal and may have played a role in the comparison results.

Part of the reason that Mao et al. found encouraging results in their short training period MOS application stems from the fact that NWP models have improved considerably over the years. This has led to a considerable reduction in the variance (or "noise") between model forecasts and observations. With skillful predictions from the NWP model, it is much easier to find useful relationships between the model and observations. Thus, the use of at least several years of data to derive the MOS relationships may no longer be necessary. However, Mao et al.'s experiments were limited to short-range temperature forecasting and it remains unclear if similar results would be obtained on other variables or longer forecast lead times where considerably more variance between the forecasts and observations occurs.

Other than overcoming the issue of a changeable NWP model environment, a second potential advantage to a MOS forecast system that only uses recent data to derive the statistical relationships, is that the resulting relationships may be more applicable to the ongoing weather regime. In traditional MOS, the relationships are derived from a much broader set of data that presumably spans a wide range of weather regimes resulting in relationships that are more applicable to the climatological norm. For example, during drought conditions, it is conceivable that the MOS forecasts derived using just the recent (drought) data will be more skillful than MOS forecasts based on average conditions. Conversely, forecasts derived during one weather regime may yield less skillful results during changes in the regime and so which approach yields more skillful forecasts remains to be seen.

2. Dynamic-MOS

Experiments have been conducted on extending the work of Mao et al. by developing an operational forecast system referred to as dynamic-MOS. The basic concept is similar to Mao et al.'s in which MOS relationships are derived and continuously updated using just recent NWP model data. In dynamic-MOS, forecasts are made for a broad set of variables including temperature, dewpoint, wind, clouds, probability of precipitation, precipitation type and thunder. For the initial experiments, a longer training period (~12 weeks) was used compared than Mao et al. for two reasons. First, the system is forecasting some binary variables (e.g. probability of precipitation), which inherently will contain significant noise. Second, the system is attempting to make long-range forecasts and thus noise will be introduced as the model forecasts diverge from reality. No attempt has been made to study changes in dynamic-MOS's skill as a function of training period length.

New forecasts are created by the system upon receipt of new NWP data while the dynamic-MOS relationships are updated weekly using a multivariate linear-regression technique. Available computer time was the only limitation to the frequency of the dynamic-MOS relationship update cycle. Forecasts were produced for several thousand sites worldwide and forecasts were generated using each of NCEP's ETA, AVN and MRF models. The period covered by each dynamic-MOS forecast was determined by the length of base model's published data.

In general, we have found results similar to that reported by Mao et al. with the skill of dynamic-MOS being similar or slightly less than the skill of comparable NWS MOS products (i.e. the AVN and MRF MOS products; no NWS ETA MOS product is available). However, climatology and other semi-static predictors often used in NWS MOS products were purposely excluded[1] in dynamic-MOS. This resulted in a relative decrease in the skill of the dynamic-MOS system relative to NWS MOS, particularly for the longer forecast times. Since in many regards our dynamic-MOS system and results are similar to that of Mao et al., we will not focus on many details of its implementation and results here.

3. *A Pitfall of Dynamic-MOS*

[1] Climatology was considered in a post-processing of the Dynamic-MOS forecasts. The Dynamic-MOS system discussed here only refers the forecasts prior to the application of the post-processing.

A key element of the dynamic-MOS system is the selection of a multivariate regression equation that will be used for the MOS forecasts. Depending upon the exact number of predictors considered, millions of different regression equations are possible. The adjusted R-squared metric (e.g. Devore, 1982) is a common means of identifying which regression is most likely to yield skillful forecasts. In dynamic-MOS, a combination of multiple regression identification techniques (e.g. forward substitution, back substitution, pairwise substitution, etc.), matrix conditioning, condition numbers and adjusted R-squared are used to identify an ideal regression. Despite the care used to ensure that high-quality regression is selected, occasionally a bad regression is used in the forecast system. Here, a bad regression is loosely defined as one that results in a forecast with unacceptably large error, possibly even physically unrealistic values. Even infrequent occurrences of forecasts based on bad regressions can have significant impact on mean forecast skill statistics of the dynamic-MOS system. The NWS avoids the use of bad regression in their operational MOS products by manually screening all selected regression solutions prior to their application. However, in an automated and continuously updated Dynamic-MOS system, such manual scrutiny is not practical.

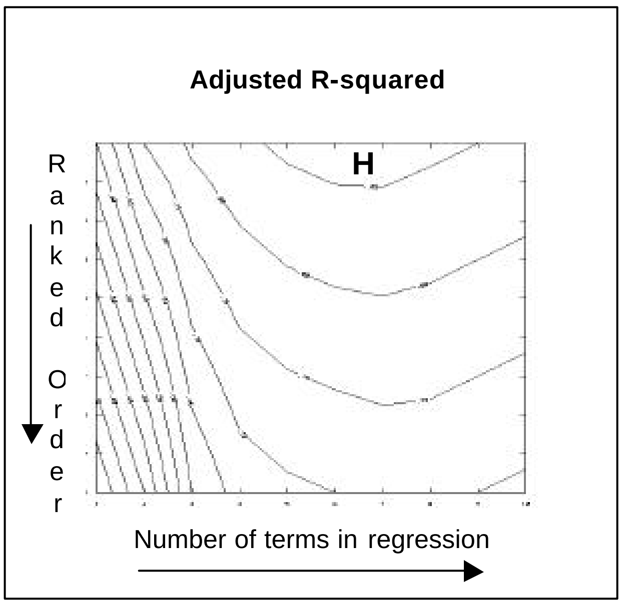

Fig. 1. Contour plot of the adjusted R-sqaured as a function of the number of terms in the regression and the ascending ranked order of the regression skill. The contour interval is 0.05.

By far the most common cause of bad regressions in our dynamic-MOS system is the use of extrapolated predictors. When a relatively short training period is used to derive the regression equations, it is possible that the range of values of some predictor in the training set is much less than normally seen. If such a regressor is included in the final regression selected, then a bad regression can result. Application of the bad regression with a regressor value far outside the range witnessed during training (but still normal) can yield a bad forecast. This is referred to as regressor extrapolation.

A simple solution to the regressor extrapolation problem is to store the range of regressor values used during the derivation of the regression equation, and then apply the regression only when the forecast values of each regressor fall within a tolerance of the values seen during derivation. This technique certainly avoids the use of extrapolated regressors but does result in a missing forecast, which may be equally unsatisfying. However, if a backup regression equation were available that does not include an extrapolated regressor, then a missing forecasts could be avoided

4. *Ensemble Dynamic-MOS*.

Since one may not know beforehand which term of a regression might be subject to regressor extrapolation, then at least as many backup regressions as terms in the original regression must be stored in order to avoid a missing forecast. Since the possibility exists that more than one regressor in the original regression may have extrapolation, or that a regressor in the backup regression may also be subject to extrapolation, then a large number of alternative regression equations may be needed in order to avoid missing forecasts.

In Dynamic-MOS, where 30 or more regressors are often considered during the derivation of each regression, and millions of combinations of those regressors are possible, it is likely that a sufficiently large set of suitable backup regressions can be identified. For example, Fig. 1 shows a typical set of contours of the adjusted R-squared values of the best ten regressions as a function of the number of terms in the regression. In this case, regressions with more than three terms have a relative uniform distribution of the adjusted R-squared values. This implies that most of the variance in the data is explained by the leading three terms and that the regression is over determined with many more predictors of similar skill available for selection. There is little significant difference between the best regression and the second best regression and so on. In fact, the identification of the best regression may not be meaningful as there is a large set of equally skillful regressions. Therefore, as this case is typical, there usually are ample sets of backup regressions to select from.

However, if a large set of backup regressions is computed and stored, then an alternative approach to quality control of the dynamic-MOS forecasts is possible. If a forecast based on each of the stored regression equation is computed, then an ensemble of Dynamic-MOS forecasts results. By applying suitable methods to the ensemble of forecasts, bad forecasts could be identified and removed from the ensemble and a final forecast constructed from the remaining members. The

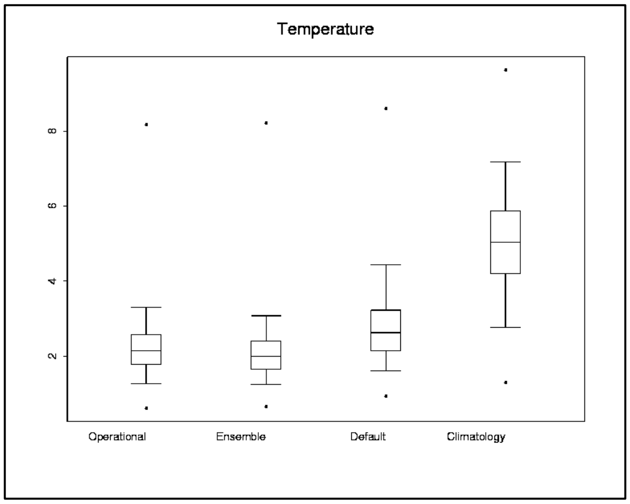

Fig. 2. Box plots of the normalized error distribution of 24-hr temperature forecasts using (a) the leading regression equation, (b) the ensemble dynamic-MOS approach, (c) direct model extraction of the 2-m temperatures and (d) a climatology forecast. Distributions are shown in standard deviations. The box plots show the median (center line of box), 25th and 75th percentiles (outside edges of box), 5th and 95th percentiles (wiskers) and extreme outliers (dots).

technique has the potential of solving not only the extrapolated forecast problem (indirectly), but eliminating bad regressions arising for other reasons (e.g. near colinearity in the training data). It also has the potential for improving the skill of forecasts made in the absence of any bad regressions by application of ensemble technique.

5. Experiments and Results.

Experiments have been conducted on an ensemble Dynamic-MOS scheme based the NCEP ETA model. For these experiments, an ensemble of regression equations that contained at least two different regressors in each equation were determined and stored for each of about 1,500 forecast sites. For each site, forecast lead-time and forecast variable, an ensemble of Dynamic-MOS forecasts was computed using each of the stored regression equations. The simple median of the ensemble forecasts was then selected as the final forecast. Distributions of forecast errors from the ensemble-median forecasts were compared to the error distributions associated with using thea single leading regression. Statistics were compiled using one month of forecasts during the spring of 2000.

Fig. 2 shows box plots of the normalized, absolute error distributions in the 24-hr temperature forecasts. The distributions cover all sites and days in the experimental period. Forecast errors were normalized using the local standard deviation to facilitate the combination of errors from sites with different climatologies. Box plots are shown for the scheme using just the leading regression equation and for the ensemble scheme. Box plots from forecasts generated by directly extracting the 2 m temperature from the ETA model and for climatology are also shown for comparison.

These results show that the ensemble dynamic-MOS scheme produced the best temperature forecasts on average with a reduced median forecast error compared to using the single regression technique. The ensemble scheme also reduced the occurrence of large-error forecasts somewhat. The very large (nearly 8 standard deviation error) remains in all of the forecast systems and is associated with a highly anomalous event that was poorly forecast by the ETA and is not the result of bad regressions. The plot also shows that the use of dynamic-MOS continues to have a significant advantage in improving forecasts over direct model extraction.

Fig. 3 and 4 show similar box plots but for forecasts of fractional cloudiness and wind speed respectively. These cases much more dramatically illustrate the potential gains in forecast skill and reduction in large forecast error than can be achieved through the use of a ensemble dynamic MOS approach. In these cases, median forecast error was reduced by 20%-40% and the 95th percentile largest forecast errors were reduced by almost half.

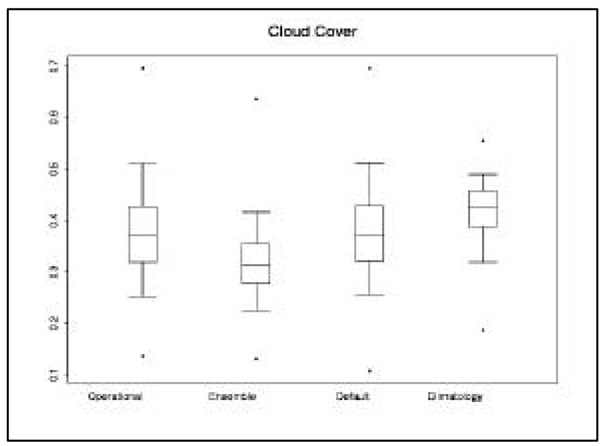

Fig. 3. As in Fig. 2 except for fractional cloudiness.

6. *Summary*.

A dynamic-MOS forecast system has been constructed similar in concept to that of Mao et al. In dynamic-MOS, relationships between the model variables and the forecast parameters are determined using a considerable shorter period of time than traditionally used in deriving MOS relationships. The relationships are continuously updating allowing the dynamic-MOS scheme to automatically adjust to changes in the underlying NWP model. The dynamic-MOS system has been used to produced forecasts for a broad set of forecast parameters and for thousands of sites around

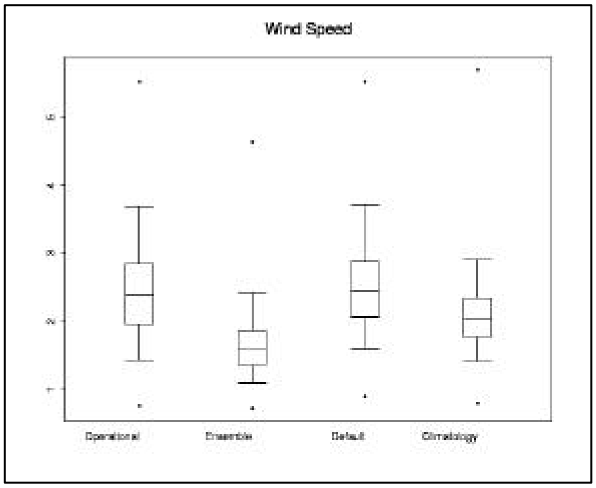

Fig. 4. As in Fig. 2 except for wind speed.

the world. It has been applied to the ETA, AVN and MRF models from NCEP. Results of our dynamic-MOS experiments were similar to those reported by Mao et al., with comparable or slightly less skill in the dynamic-MOS forecasts when compared to similar products from the NWS.

An ensemble extension of the Dynamic-MOS scheme was proposed and tested. In ensemble dynamic-MOS, a large set of regressions are computed and stored for each forecast parameter. The regression set is then used to produce and ensemble of dynamic-MOS forecasts. Some simple experiments of the technique using a month of forecasts from the ETA model showed that the ensemble technique can provide a significant improvement in the forecast skill when compared to the skill of the single best regression. Although the technique was proposed as a quality-control mechanism for identifying and eliminating bad regressions, the results shown here indicate that the scheme has the ability to not only reduce the occurrence of large forecast errors, but also in improving the overall skill of the dynamic-MOS forecast system.

7. References.

Devore, J. L. , 1982: Probability and statistics for engineering and the sciences. Brooks/Cole Publishing, Monterey, CA. 640 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.

Glahn, H. R., A. H. Murphy, L. J. Wilson, and J. S. Jensenius, 1991: Lectures of the WMO training workshop on the interpretation of NWP products in terms of local weather phenomena and their verification. WMO TD No. 421, Geneva.

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

Mao Q., R. T. McNider, S.F. Mueller, H. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea . Forecasting*., **14**, 190-202.

Vislocky, R. L. and J. M. Fritch, 1995a: Generalized additive models versus linear regression in generating probabilistic MOS forecasts of aviation weather parameters. *Wea. Forecasting*, **10**, 669-680.

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