6.4

ARE MODEL OUTPUT STATISTICS STILL NEEDED?

Peter P. Neilley

And

Kurt A. Hanson

Weather Services International, Inc.
400 Minuteman Road
Andover, MA 01810

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 optimized relationships between parameters predicted by a NWP model and desired forecast quantities. Multivariate linear regression is the most commonly used statistical technique to determine the relationships, but other techniques such as generalized additive models (Vislocky and Fritch, 1995) and logistic regression for predicting discrete phenomena (Glahn et al, 1991) have been used amongst others.

MOS serves two primary purposes. First, the MOS technique provides forecasts of quantities that may not be explicitly predicted by the model. Examples include precipitation type forecasts and probabilistic forecasts of precipitation, thunder, fog, etc. from the deterministic NWP data.

The second purpose of MOS is to reduce the mean error of the raw model forecasts. MOS improves over the raw model data by two mechanisms: bias removal and statistical correction. The statistical correction is achieved when systematic errors in the model forecasts occur relative to the forecast parameters. For example, MOS can improve temperature forecasts in a circumstance where the model tends to predict temperatures that are too cold on cloudy days and too warm on sunny days.

The ability of MOS to provide statistical corrections depends on the ability of identify significant model-observation correlations. In order to do so, such correlations must be larger than those that may arise due to the random, stochastic noise inherent in the forecast system. As NWP models have improved over the decades, both the systematic and stochastic errors in the model have been reduced. However, it is likely that the stochastic errors have not decreased as quickly as the systematic errors since many of the stochastic errors are from non-meteorological/modeling sources (e.g. observation errors). These external errors are not reduced by improvements in the NWP model. Assuming the systematic errors of the model forecasts have reduced faster than the stochastic errors, then it is increasingly difficult to find statistically significant relationships between the forecast data and observations. This implies that either (a) the dataset used to derive the MOS relationships must be longer in order to obtain forecasts with equal statistical characteristics or (b) the fractional improvement of the MOS forecasts over the raw NWP model data must decrease. As NWP models are under near constant revision, using longer periods of data to develop regression solutions is not always warranted. Therefore, it seems reasonable to speculate that the fractional improvement of MOS forecasts over raw NWP data might indeed be decreasing.

Today, the National Weather Service routinely produces MOS forecasts based of the GFS (MAV and MEX MOS) ETA and NGM NWP models. Recent skill scores show that these NWS MOS products continues to provide improved forecasts over raw model extract (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. Further, it is not known if the measured improvement is derived mostly from either the bias correction or statistical correction or both.

It is the purpose of this study to ascertain whether or not statistical correction (as opposed to bias correction) of NWP model data should continue to play a material role in the MOS process of improving the NWP forecasts. If statistical correction is found not to be substantial, then one would question if a sophisticated regression-based MOS system is still needed.

2. Formulation of Problem

The multivariate linear regression approach to MOS forecasting of parameter F uses an equation of the form

$$F_{MOS} = A + c_1 * X_1 + c_2 * X_2 + \ldots\ldots c_n X_n \qquad (1)$$

Where A is the intercept, c_i are the regression coefficients, X_i are the predictors and n is the number of predictors used in the equation. The predictors are usually values extracted from the NWP model, but in the case of NWS MOS products can also include recent observations and climatological values.

Mao *et al.* (1999) used an alternative MOS formulation in which the regression is used to predict corrections to the raw model's forecast rather than forecast the particular value directly. This formulation is equivalent to forcing the regression to contain the model's raw forecast as a predictor with a coefficient of 1, i.e.

$$F_{MOS} - F_{NWP} = A + c_1*X_1 + c_2*X_2 + \ldots\ldots c_nX_n \qquad (2)$$

(after moving the model's forecast term F_{NWP} to the LHS) where the terms A, c_i and X_i are defined as in (1) but generally have different values in practice. The LHS of (2) represents a regression-based correction to the NWP's model forecast of F. Mao *et al.* referred to this term as the model calibration.

For our experiments, we used multivariate, least-squares regression to find the predictors and coefficients in (2). The regression is done against the errors in the NWP model's forecast rather than against the observations of the forecast variable. For this reason, we refer to this as Regression Against Model Error or RAME and hence write (2) simply as

$$F_{RAME} = A + c_1*X_1 + c_2*X_2 + \ldots\ldots c_nX_n \qquad (3)$$

The term A in (3) has two separate components: the mean model error in forecasting F and the intercept of the regression. These two terms can be treated and computed separately:

$$F_{RAME} = A_e + A_i + c_1*X_1 + c_2*X_2 + \ldots\ldots c_nX_n \qquad (4)$$

Where A_e is the mean model bias and A_i is the regression intercept. To facilitate our analysis, we split the terms in F_{RAME}, rearrange (4) and simplify as:

$$F_{MOS} = F_{NWP} + A_e + F_{REG} \qquad (5)$$

where F_{REG} represents the contribution due to regression

$$F_{REG} = A_i + c_1*X_1 + c_2*X_2 + \ldots\ldots c_nX_n \qquad (6)$$

The purpose of this study is to assess the relative contributions of the terms in (5) to the value of the MOS forecast. We do so incrementally, by first assessing the error in the model forecast F_{NWP}, and then looking at the reduction in error by the terms A_e and F_{REG} subsequently.

When extracting forecast values from the NWP model, some simple algorithms have been applied to the model data depending on the forecast variable. For the variables studied here, the following algorithms were applied:

(a) 2 m temperature: *Corrected from NWP elevation to actual station elevation using a lapse rate dependent on the model's lapse rate in the boundary layer*

(b) Probability of Precipitation: *A simple logistic function based on 1000-700mb mean RH and total precipitation.*

(c) 10 m wind speed: *No algorithm applied*.

By comparing the skill of bias-corrected forecasts extracted from the NWP model these using simple extraction algorithms to that improved upon using regression we are assessing whether or not "intelligent" forecast extraction from the NWP data is a sufficient means to derive an optimal forecast from the model, or whether a more rigorous regression process is needed.

3. Experimental Methodology

To solve for F_{REG} in (5), we follow the dynamic-MOS approach discussed by Neilley *et al*. (2001). We collect a recent history of observation and model data for each forecast site. About 150 predictors are extracted from the model for use overall, of which about 35 are made available to the regression engine for each individual forecast variable.

For these experiments, model data were extracted from the 0000 UTC runs of the NCEP GFS model. We used forecasts from the 1 deg resolution output of this model, which are currently available from NCEP through day 8 of the forecast period. Data were collected from January through June 2003. Data from the first 100 days of this period were used to derive regressions (i.e. Eq. 6) and model biases (A_e), while the remaining days were used as an independent sample to evaluate the results. Eighty of the largest cities in the continental U.S. were used as forecast sites in this study. Individual experiments (regressions, biases, etc.) were computed for each forecast time of the day (00, 03, 06, …. 21 UTC) but results were aggregated over all times.

A series of experiments was conducted for each forecast variable, in which a minimum allowable value of the adjusted R-squared parameter (DeVore, 1982) was imposed on the regression. The threshold was systematically increased from 0.1 to 0.7 and individual results tabulated. When the best regression from the training dataset did not have an adjusted R-squared value that exceeded the imposed threshold, no regression was selected and the term F_{REG} was set to 0 in (5). Results were stratified as a function of the minimum threshold. For comparison purposes, a full regression forecast of F was also computed in a manner similar to standard MOS applications. This allows us to ascertain whether or not the formulation of (5), and in particular, the requirement that the

coefficient on the direct model extract term be uniformly 1 significantly affected the results.

Results were derived from applications of the regressions and biases computed from the training dataset to the subsequent ~50 days in the dataset. Separate results were computed for each forecast variable (2 m temperature, probability of precipitation and wind speed) and for each day into the forecast period (1-8). Here we show results for Day 2 (tomorrow) and aggregated over all days.

For each forecast made, an error was computed. Mean, mean absolute, and root mean square errors were computed over the entire validation period. Here we focus on the RMS errors as these sufficiently characterize the broader set of measures. A number of supporting statistics were computed and are presented below.

2 m Temperature (C)			
Days 1-8			
Forecast Type	**Adj R^2**	**RMSE**	**%**
F_{NWP}		2.96	
$F_{NWP} + A_e$		2.76	
$F_{NWP} + A_e + F_{REG}$	0.1	3.99	99
$F_{NWP} + A_e + F_{REG}$	0.2	3.96	96
$F_{NWP} + A_e + F_{REG}$	0.3	3.81	89
$F_{NWP} + A_e + F_{REG}$	0.4	3.72	75
$F_{NWP} + A_e + F_{REG}$	0.5	3.22	41
$F_{NWP} + A_e + F_{REG}$	0.6	2.77	12
$F_{NWP} + A_e + F_{REG}$	0.7	2.76	0
Regular Regression		3.13	

Table 4.1. The Root Mean Square Error (C) of several MOS 2 m temperature forecast experiements For each forecast type, the root mean square error (RMSE) is shown. For the forecast experiments that include a regression the percentage of time that a regression that exceeding the minimum allowed adjusted R^2 value is listed.

4. Results

4.1 *2 m temperature*. Table 1 shows the results of the various 2 m temperature forecast experiments averaged over all forecast days. The best performing forecast method in this case (based on the RMSE) was the bias-corrected model extracted temperature forecasts. The bias-correction reduced the error by about 8%. All attempts to improve the forecast skill relative to the bias-corrected model forecasts yielded worse results. Only for the cases with a relatively strict adjusted R^2 values greater than 0.6 did the results approach that of the bias-corrected model forecasts. However, in those cases, a small fraction (12% or less) of the cases actually had a non-zero regression computed. We note that the RMSE of the regular regression method yielded a result somewhat worse (3.13 C) than even the raw model forecasts (2.96 C).

2 m Temperature (C)			
Day 2			
Forecast Type	**Adj R^2**	**RMSE**	**%**
F_{NWP}		2.44	
$F_{NWP} + A_e$		2.34	
$F_{NWP} + A_e + F_{REG}$	0.1	2.54	100
$F_{NWP} + A_e + F_{REG}$	0.2	2.54	100
$F_{NWP} + A_e + F_{REG}$	0.3	2.53	88
$F_{NWP} + A_e + F_{REG}$	0.4	2.61	58
$F_{NWP} + A_e + F_{REG}$	0.5	2.66	50
$F_{NWP} + A_e + F_{REG}$	0.6	2.31	21
$F_{NWP} + A_e + F_{REG}$	0.7	2.34	0
Regular Regression		2.28	

Table 4.2. As in Table 4.1 but just for day 2 (tomorrow) of the forecast period.

Table 4.2 shows similar results but for just Day 2 of the forecast period. In this case, the best forecast was the regular regression. However, the general trends witnessed for days 1-8 are repeated here, with regression corrections to the bias-corrected model error generally yielding forecasts with larger errors than the bias-corrected model forecasts. Only for the most stringent adjusted R^2 threshold over 0.6 did the regression corrections improve the forecast quality. However, only 21% of those cases actually had non-zero regressions used.

4.2 *Probability of Precipitation*. Table 4.3 shows the results of probability of precipitation experiments. The general trends shown for the temperature forecasts are seen again in these results. (Scores for the regular regression method were not available at the time of publication). The best forecasts were found for the bias-corrected model forecasts. In this case, the bias correction added almost no value (RMSE decreased about 0.005) over the direct model extract. All attempts to improve the model forecasts using a regression correction yielded forecasts with lower skill. Results for individual days (not shown) have results with trends identical to the day 1-8 mean results shown.

6 hr Probability of Precipitation			
Days 1-8			
Forecast Type	**Adj R^2**	**RMSE**	**%**
F_{NWP}		0.44	
$F_{NWP} + A_e$		0.44	
$F_{NWP} + A_e + F_{REG}$	0.1	0.49	89
$F_{NWP} + A_e + F_{REG}$	0.2	0.47	65
$F_{NWP} + A_e + F_{REG}$	0.3	0. 46	30
$F_{NWP} + A_e + F_{REG}$	0.4	0.44	9
$F_{NWP} + A_e + F_{REG}$	0.5	0.44	0
$F_{NWP} + A_e + F_{REG}$	0.6	0.44	0
$F_{NWP} + A_e + F_{REG}$	0.7	0.44	0
Regular Regression		NA	

Table 4.3. As in Table 4.1 but just for days 1-8 of the 6-hr probability of precipitation forecasts.

4.3 *Wind Speed*. Table 4.4 shows the results using wind speed as the forecast variable. These results are considerably different than the temperature and precipitation results shown earlier. Here, the regressions universally reduce the error of the forecasts and perform best for relatively low values of the adjusted R^2 threshold. The skill of the forecasts actually decreases as the adjusted R^2 value increases reflecting the reduced percentage of forecasts in which no regression is computed causing the (worse) bias-corrected model forecasts to be used soley. The individual day's (not shown) results mirrored the trends seen in the days 1-8 means shown.

Wind Speed (m s^{-1})			
Days 1-8			
Forecast Type	**Adj R^2**	**RMSE**	**%**
F_{NWP}		2.14	
$F_{NWP} + A_e$		2.24	
$F_{NWP} + A_e + F_{REG}$	0.1	1.93	100
$F_{NWP} + A_e + F_{REG}$	0.2	1.93	100
$F_{NWP} + A_e + F_{REG}$	0.3	1.93	93
$F_{NWP} + A_e + F_{REG}$	0.4	1.99	71
$F_{NWP} + A_e + F_{REG}$	0.5	2.11	36
$F_{NWP} + A_e + F_{REG}$	0.6	2.19	9
$F_{NWP} + A_e + F_{REG}$	0.7	2.23	3
Regular Regression		NA	

Table 4.3. As in Table 4.1 but just for days 1-8 of the wind speed (m s^{-1}) forecasts.

5. Summary and Discussion.

In this study, we have attempted to determine the relative skill of the various components of a MOS forecast and in particular to determine when multivariate linear regression adds value to the forecasts compared to more simple model extract and correction techniques. We have compared various measures of skill in forecasts derived from days 1-8 of the NCEP GFS model, and focused on the RMSE results here.

Based on the results of our experiments, we have concluded that multivariate least-squares regression does not improve the quality of forecasts in many circumstances. The results appear strongly dependent on forecast variable with 2 m temperature and probability of precipitation forecasts showing the least value in the regressions while the regressions had the most value in wind speed forecasts.

The value of least-squares linear regression is not obvious as a general technique to improving NWP model forecasts. Our results imply that in developing a general NWP model output post-processing system, using simple algorithms and bias-correction techniques often can yield forecasts that are comparable if not superior to forecasts derived from the more complex regression system. We speculate that this conclusion is the result of the fact that as NWP model forecasts are becoming increasing skillful, it is becoming increasing difficult for statistical methods to find consistent correlations between the model data and observations.

We note that the regression method used here relied on a relatively small training dataset (100 day) compared to the size often used in NWS MOS regression systems. It is not clear if a longer training dataset would yield results more favorable for the regression technique. One caveat of using a longer training dataset is the lack of stability in the NWP model over the years making the applicability of regressions derived from long datasets unclear.

Finally we note that the NWS often uses climatological predictors in formulating their MOS equations. We did not consider such factors here. As NWP solutions tend to diverge from reality during the later portion of the forecast period, forecasts that are near the climatological norm will tend to yield RMS errors that are lower. Hence, it is a common artifact of NWS MOS forecasts that they trend towards the climatological norm as the forecast period progresses. If we had used climatological predictors in our regression system here, we would have expected considerably better results for the regressions. However, including the climatological predictors dampens the magnitude of events predicted in the NWP model which can lower the "value" of the MOS forecasts even if their RMS errors are reduced.

6. 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.

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.

Neilley, P., W. Myers and G. Young, 2001: Dynamic Ensemble MOS. Proceedings of the 20th AMS Conf. On Prob and Statistics.

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.