EASTERN REGION TECHNICAL ATTACHMENT
NO 98-4
JUNE, 1998

Stratification and Mixed Model MOS Techniques to Predict Maximum Temperatures at Columbia, SC

Mark DeLisi
NOAA/National Weather Service Forecast Office
West Columbia, SC

Editor's Note: The author's current affiliation is NWSFO Mt. Holly, NJ

1. INTRODUCTION

This study was undertaken to answer three questions: 1) Is it possible, given the changing nature of the Eta model, to develop an objective operational temperature forecast based at least in part on that model? 2) What is the relative skill of Eta model-based predictors versus Nested Grid Model (NGM) model-based predictors? 3) Does stratification of the data into cold-air damming and non-cold-air damming days significantly improve forecast skill? These questions were addressed by developing maximum temperature equations to predict day one and day two maximum temperatures at the Columbia, South Carolina National Weather Service Forecast Office (CAE).

Multiple linear regression was used on a limited number of temperature related variables from the 0000 UTC run of the Eta model and the (NGM) to develop equations to predict day one and day two maximum temperatures. Records in the dependent data set were stratified by whether the corresponding day was a cold-air damming (CAD) day. Three types of local equations (LOC) were developed: 1) a maximum temperature equation for day one or day two when CAD was occurring; 2) maximum temperature equations for day one when CAD was not occurring; 3) maximum temperature equations for day two when CAD was not occurring. For 2) and 3), a separate LOC was developed for the hot season, the cool season, and the two transition seasons. One CAD LOC was developed for the entire year. This resulted in nine separate equations. All LOC were evolutionary in nature. Their forecasts were verified from August 21, 1996 through November 17, 1997 against NGM and Aviation Model Output Statistics (FWC and FAN, respectively) 0000 UTC day one and day two maximum temperature forecasts. Seasons were combined for purposes of verification.

2. DATA

Data were obtained from the Office of Systems Operations server in Gridded Binary Data format and examined with the Personal Computer Gridded Interactive Display and Diagnostic System (PCGRIDDS) and the General Meteorological Package Analysis and Rendering Program after converting the data into PCGRIDDS and General Meteorological Package (GEMPAK) formats. Four parameters were chosen as potential independent variables from the Eta model: 1) 1000 to 850 mb thickness (m), 2) 1000 mb temperature (°C), 3) B015 (surface to surface pressure minus 30 mb) boundary layer temperature (°C), and 4) 850 mb relative humidity. Four corresponding parameters were chosen as potential independent variables from the NGM model: 1) 1000 to

850 mb thickness (m), 2) 1000 mb temperature (°C), 3) S982 (sigma layer 982) boundary layer temperature (°C), and 4) 850 mb relative humidity. Values were from the 0000 UTC run of each model for the 24 hour (day one) and 48 hour (day two) forecast projections. In addition, the 0600 UTC day one observed temperature was selected as a potential independent variable. The dependent variable was the maximum temperature between 1200 and 0000 UTC (°F).

Data were stratified into two groups. The first group was comprised of days that could be classified as cold-air damming (CAD) days. CAD days were defined as days when there was an east or northeast flow at the surface, a southeast to southwest flow just above the surface layer, overcast skies throughout the period from 1200 to 0000 UTC, at least a trace of rain during that period, and a temperature difference of no greater than 12°F between the 0600 UTC temperature and the maximum temperature between 1200 and 0000 UTC.

The second group was comprised of all other days. CAD days occur relatively infrequently at CAE, so data from CAD days were collected from five sites including CAE. The other four sites were the National Weather Service Office at Greer, SC; and private contract offices at Charlotte, NC; Augusta , GA; and Florence, SC. This was done to increase the number of records in the dependent data set. The pooling of data in this fashion did render the records in the CAD dependent data set less than completely independent of each other. Only values from day one were retained for the CAD data set (even though the CAD equation was used for day one and day two), and data were collected without regard to season.

On non-CAD days, data were collected on nearly a daily basis for CAE only. Since consecutive days are not meteorologically independent of each other, this also had the effect of rendering the records in the non-CAD dependent data set less than completely independent of each other.

In the case of the data from the non-CAD days, data were further divided by season. The cool season data were from December 15 through March 14, the first transitional season data were from March 15 through May 14, the hot season data were from May 15 through September 14, and the second transitional season data were from September 15 through December 14.

This meant there were nine sets of data that would result in nine local equations (LOC). Eight sets of data were associated with non-CAD days, and one set of data was associated with CAD days. For the non-CAD days, there was a set of data for each day one and day two of each season. For the CAD days, there was a set of data for day one without regard to the season.

Data used to derive the LOC were collected from January 15, 1996 through October 15, 1997. Data used to verify the LOC were from August 21, 1996 through November 17, 1997. For each equation, data used for verification were not included in the dependent data set used for equation development. As the size of the data sets grew, equations were re-derived incorporating previously independent data into the dependent data set.

During the course of the study, modelers at the National Center for Environmental Prediction (NCEP) made changes to the Eta model. It was noticed (no evidence presented

here) that the changes effected the Eta-derived thermal variables used by the study. Specifically, values of Eta-derived thermal variables were lower for similar maximum temperatures at the end of the study than at the beginning of the study. The NGM, however, has been a static model for some time. The static nature of the NGM was used to make adjustments to Eta output so that pre-modification and post-modification Eta data would not be precluded from use in the same dependent data set.

Since the NGM is a static model, it was assumed that any change to the difference between corresponding Eta and NGM parameters from one year to the next was due to a change in the Eta parameter. Therefore, if the change to the former difference was computed, it could be added to the Eta parameter from the earlier year to account for the latter change.

The difference between the value of an Eta parameter and the corresponding NGM parameter was regressed on observed maximum temperature for a season. The same was done for the same season from the previous year. The difference between the resultant equations was itself an equation (the adjustment equation). That equation was comprised of 1) a predictor variable (observed maximum temperature) with an associated coefficient and 2) a constant which estimated the change in the difference between the Eta and NGM variables when the maximum temperature was 0°F.

If the change in the difference between the Eta and NGM parameters was related to maximum temperature, then the coefficient from the adjustment equation should have been significantly different from zero. If the coefficient was not significantly different from zero, but the constant was, then the indication was of a significant change in the difference between the Eta and NGM parameters over the course of a year that was not related to maximum temperature.

Maximum temperatures were used with the adjustment equation to compute adjustments to Eta parameter values from the earlier season, without regard to whether the coefficient was significantly different from zero. Adjustments were made to the Eta thermal parameters from the eight non-CAD data sets. Mainly for the purpose of conserving time and effort, no adjustments were made to the CAD data set, and no adjustments were made to 850 mb relative humidity values.

Before making the adjustment to each Eta parameter, the study assessed qualitatively the two mean Eta/NGM temperature differences and the resultant regression equations. If it was determined that they were not similar, then an adjustment was made to the earlier season Eta data. In fact, adjustments were made to all earlier season Eta thermal data (non-CAD cases) except for the second transitional season day one and day two 1000 mb and B015 temperatures.

3. EQUATION DERIVATION

The study used a commercially available statistical software package to perform variable selection and multiple linear regression to arrive at LOC. A complete explanation of these processes can be found in Draper and Smith (1981). A less rigorous but more meteorologically oriented explanation of the regression process can be found in Wilks (1995).

The variable selection process was designed

to achieve two goals. The first goal was to maximize the amount of variance in the dependent variable, observed maximum temperature, explained by the independent variables. The second goal was to minimize the amount of bias in the resultant maximum temperature forecasts.

Since there was some dependence between the records in all of the dependent data sets, no equation (that did not have to account for a quadratic feature in the residuals) was retained as an LOC if any of the independent variables had associated p-values in excess of .01. P-values can range from zero to one, and with regard to independent variables, those with better predictive value have lower p-values. In some cases, the residuals from the equations determined by the variable selection process did not behave in a way consistent with the assumptions of multiple linear regression. Specifically, they either did not exhibit homoscedasticity (equal variance along the range of predicted values) or a normal distribution. When either was the case, a transformation of the dependent variable was required to render the residuals homoscedastic and normally distributed. In a couple of instances (the second transitional season equations), a quadratic nature to the residuals necessitated that higher order terms of the independent variables be included in the equations. In the second transitional season LOC (day one and day two), some of the terms had associated p-values in excess of .01.

Derived LOC always explained greater than .80 of the variance in the dependent variable, usually explained greater than .90 of the variance in the dependent variable, and rarely explained greater than .95 of the variance in the dependent variable.

At a minimum, 45 records were required before a LOC was derived. Still, the original nine LOC were derived from small numbers of records. Therefore, they were rederived as the number of records increased. Each of the nine LOC was rederived at least twice during the course of the study. Only the latest LOC are given below. They compute the forecast maximum temperature in °F.

Cold Season:

Day One:

$$47.51 + 1.538(C) - .03110(D) \quad (1)$$

Day Two:

$$47.84 + 1.426(B) \quad (2)$$

First Transitional Season:

Day One:

$$[-15.71 + .1801(A) - .005323(H)]^2 \quad (3)$$

Day Two:

$$\exp[4.010 + .01897(C) - .0009042(H)] \quad (4)$$

Warm Season:

Day One:

$$38.81 + .7312(B) + 1.040(F) \quad (5)$$

Day Two:

$$[113.9 + 378.2(B) + 137.0(F) - 241.9(C)]^{1/2} \quad (6)$$

Second Transitional Season:

Day One:

$$[155720 - 72.03(C) - 6688(G) + 1206(C^2) + 1192(G^2) - 1387(C)(G)]^{1/3} \quad (7)$$

Day Two:

$$48.55 + .9842(C) - .008511(G) - .02558(C^2) - .01335(G^2) + .05813(C)(G) \quad (8)$$

Cold Air Damming:

Day One or Two:
11.62 + .7812(G) + .6234(I) (9)

where:

A = Eta 1000 to 850 mb thickness (m*10)
B = Eta 1000 mb temperature (°C)
C = Eta B015 temperature (°C)
D = Eta 850 mb relative humidity (%)
E = NGM 1000 to 850 mb thickness (m*10)
F = NGM 1000 mb temperature (°C)
G = NGM S982 temperature (°C)
H = NGM 850 mb relative humidity (%)
I = 0600 UTC observed temperature (°F)

4. VERIFICATION

As stated previously, independent data from the period August 21, 1996 through November 17, 1997 were used for verification. Occasionally, a LOC, an FWC, and/or FAN were not run or their forecasts were not recorded by the study. There were 316 LOC forecasts for day one (22 CAD days and 294 non-CAD days) and 301 LOC forecasts for day two (22 CAD days and 279 non-CAD days). There were 316 FWC forecasts for day one (22 CAD days and 294 non-CAD days) and 303 FWC forecasts for day two (21 CAD days and 282 non-CAD days). There were 319 FAN forecasts for day one (23 CAD days and 296 non-CAD days) and 306 FAN forecasts for day two (22 CAD days and 284 non-CAD days).

It is reemphasized that the LOC were rederived during the course of the verification process, but forecast maximum temperatures were not recomputed using updated equations. With regard to the CAD LOC equation, the verification process presents the LOC optimally. Consider that the forecaster will always be faced with the decision of whether to use the standard LOC or the CAD LOC. For purposes of verification, the study assumed that the forecaster would always make the correct decision.

Table 1 shows the variance in the observed maximum temperatures explained by the forecasts of the LOC, the FWC, and the FAN. This is done for day one forecasts, day two forecasts, and both days combined across non-CAD days, CAD days, and all days. For non-CAD days, variance explained by the forecasts of each of the equations is comparable. The same is true for all days. However, the LOC explain considerably more variance on CAD days than either the FWC or the FAN.

Table 2 shows the mean absolute error (MAE) of the LOC, the FWC, and the FAN for non-cold air damming days, CAD days, and all days across day one forecasts, day two forecasts and both day forecasts combined. Table 3 shows the biases in a similar manner.

For the MAEs and biases, a p-value of .05 corresponds to the 95 percent confidence level, and a p-value of .01 corresponds to the 99 percent confidence level. For non-CAD days, the LOC MAEs are not significantly less than the FWC MAEs at a p-value of .05 or less. The LOC MAEs are significantly less than the FAN MAEs at a p-value of .0175 (day one), zero (day two), and zero (both days combined).

For CAD days, the MAEs of the LOC are significantly less than those of either the FWC or the FAN at p-values at or close to zero (zero for day one, day two, and both days combined versus the FWC; .006 for day one, .0006 for day two, and zero for both days combined versus the FAN).

For all days, the MAEs of the LOC are again significantly less than those of either the FWC or the FAN at p-values at or close to zero (.0009 for day one, .0021 for day two, and zero for both days combined versus the FWC; .0016 for day one, and zero for day two and both days combined versus the FAN).

From Table 3, the only LOC bias that was significantly different from zero was the bias for CAD days across both days forecasts (p-value of .0157). All of the biases of the FWC and the FAN were significantly different from zero (p-value of zero in every case).

5. CONCLUSIONS

The LOC were developed from a comparatively small (both in terms of the number of potential independent variables and the number of records) data set. They were evolutionary in nature, and assuming the equations generally improved over the course of the study, results of verification are not results of verification of the best LOC exclusively. The records used to develop the LOC were not completely independent of each other. Nonetheless, the LOC outperformed the FAN over CAD days and non-CAD days. The LOC were competitive with the FWC on non-CAD days, and they outperformed the FWC on CAD days to the extent that when one compared the LOC to the FWC over all days, the LOC outperformed the FWC. Again, it should be said that the study assumed the correct LOC (CAD or non-CAD) would be used for purposes of verification. However, as NCEP's Eta model has been improving, it has been making CAD episodes easier to forecast. The results of this study's verification are therefore not nearly as idealistic as they would have been several years ago.

Given the results, it appears that it is possible to use Eta based parameters in a MOS technique even though the Eta model is evolutionary in nature. There is no direct proof presented here that the technique employed in this study helped to do this.

An inspection of the LOC shows that most of the variables retained were Eta parameters. If one assumes that the technique used to modify the Eta variables was of some use, then one is forced to discount the presence of the NGM 850 mb relative humidity parameters in the two first transitional season equations and the presence of the NGM S982 temperature parameter in the CAD equation when evaluating whether Eta-based parameters do a better job than NGM-based parameters. This is because no adjustments were made to the Eta 850 mb relative humidity parameter or any Eta parameters in the CAD cases, and so they were not afforded the possible benefit of the modification technique. With this in mind, ten out of 15 parameters retained were from the Eta model, four were from the NGM, and one was an observed value. It is not demonstrated that Eta-based parameters do a better job than NGM-based parameters in a MOS technique, but the indication is so.

Finally, it appears that something is gained by stratifying cases so that separate equations can be developed for relatively rare events that standard MOS equations do not forecast well, at least with regards to maximum temperature. Even though there is still some drawback in having to know in advance when to use such equations, the benefit of having them quite likely outweighs that constraint.

Acknowledgments. My thanks to Anthony Petrolito and Harry Gerapetritis at CAE, who helped me compile the data for this study. Thanks to Mary Erickson at the Techniques

Development Laboratory for her suggestion regarding the 0600 UTC day one temperature. Thanks also to Michael Cammarata, Science Operations Officer at WSFO CAE, for his valuable comments on this manuscript.

6. REFERENCES

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis.* 2d ed. John Wiley and Sons, 709 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Table 1: Variance in observed maximum temperatures at CAE explained by the day one, day two, and combined maximum temperature forecasts from the LOC, the FWC, and the FAN equations.

	Non-CAD Days			CAD Days			All Days		
	LOC	FWC	FAN	LOC	FWC	FAN	LOC	FWC	FAN
Day One	.9243	.9378	.9229	.8741	.6921	.6536	.9289	.9206	.9085
Day Two	.9239	.9234	.8941	.8030	.6308	.6329	.9272	.9061	.8765
Combined	.9240	.9307	.9086	.8362	.6609	.6378	.9279	.9134	.8926

Table 2: Mean absolute error (MAE) of the CAE day one, day two, and combined maximum temperature forecasts from the LOC, the FWC, and the FAN equations (°F).

	Non-CAD Days			CAD Days			All Days		
	LOC	FWC	FAN	LOC	FWC	FAN	LOC	FWC	FAN
Day One	2.56	2.71	2.93	2.09	7.64	4.61	2.53	3.06	3.05
Day Two	2.53	2.75	3.56	2.63	7.62	5.82	2.53	3.09	3.72
Combined	2.55	2.73	3.24	2.36	7.63	5.20	2.53	3.07	3.38

Table 3: Bias of the CAE day one, day two, and combined maximum temperature forecasts from the LOC, the FWC, and the FAN equations (°F).

	Non-CAD Days			CAD Days			All Days		
	LOC	FWC	FAN	LOC	FWC	FAN	LOC	FWC	FAN
Day One	-0.13	1.29	-1.51	0.82	7.45	4.00	-0.06	1.72	-1.12
Day Two	0.04	1.22	-1.70	1.36	7.33	4.82	0.14	1.64	-1.23
Combined	-0.05	1.25	-1.61	1.09	7.40	4.40	0.03	1.68	-1.17