3.8

LOOKING FAR BACK VS. LOOKING AROUND ENOUGH: OPERATIONAL WEATHER FORECASTING BY SPATIAL COMPOSITION OF RECENT OBSERVATIONS

Claudia Tebaldi*
National Center for Atmospheric Research

1 INTRODUCTION

The need for operational weather prediction systems to produce accurate forecasts of numerous surface quantities in real time at thousands of locations all around the globe, has for a long time now required the efficient automation of this task. Starting in the '70s, objective statistical methods have been developed to "interpret" the output of numerical weather prediction (NWP) models, whose cornerstones may be identified as the *perfect prog* and *model output statistics* (MOS) techniques (Glahn and Lawry, 1972). It has become common for the weather service and other weather prediction centers to rely especially on MOS techniques. A long history of observations of the parameters that are to be forecasted, and of model output valid at the time when the former were observed, are matched and a statistical model is fitted, often consisting of a multiple linear regression. When the forecast is needed, the model output valid at the forecast time is combined according to the statistical model developed, whose coefficients are estimated by the history available (see Glahn et al., 1991, Vislocky and Fritsch, 1995 and Wilks, 1995 for an overview of current research and additional references on MOS). There are several shortcoming to this approach, that we list in detail in Section 2. They all relate to the "frozen quality" of the long history required to estimate MOS parameters.

In an operational system developed by the Research Applications Program at the National Center for Atmospheric Research (Mahoney, 2001a and 2001b) an attempt to deviate from the MOS paradigm has been proposed and the skill of the system's forecasts successfully verified with respect to the traditional MOS approach. In a nutshell, a continuously updated MOS relation is estimated on the basis of recent history at the station where the forecast is needed. This approach requires recent observations and matching model output in a number sufficient to estimate a stable statistical relation. For operational purposes "sufficient" is currently identified by one hundred days' worth of data.

Together with the advantages offered by this approach, however, several problems arise, and we propose to obviate them by what we call *spatial composition* of observations and model output. Very recently, sophisticated hierarchical state-space models have been developed in the area of weather prediction, whose skills have been amply demonstrated (Wikle et al., 1998, Nott et al., 2001). However, the estimation of their parameters requires complex and lengthy Montecarlo simulations, a luxury we cannot afford when developing large scale, operational, real time products. Section 3 details what we mean here by spatial composition, and explains why we think that, even in its simplicity, it may be effective in addressing the problems pointed out for MOS.

Section 4 describes the specific ingredients of our analysis: the spatial domain under study, the quantities to forecast, the models whose output is used in the prediction stage. Also the explanation of the verification methods adopted can be found there. Section 5 summarizes and discusses the results. Conclusions and proposed future directions follow in Section 6.

2 MOS AND DYNAMICALLY UPDATED MOS

The goal of MOS methods is to to estimate a robust relationship between the quantities computed by a numerical model and the weather parameters to be predicted, on the basis of a history of such data. The relationship is usually defined for a specific location and season. It is immediately clear how the necessity of "historical data" for both model and observed quantities introduces a rigidity that may be counterproductive: Numerical models undergo a constant refinement, in terms of resolution (spatial and temporal) and parameterization. Thus, the values of a specific quantity in the model output may be regarded as the realization of a process that is by no means stationary over the years of the model operation, and whose relation with

* *Corresponding author address*: Claudia Tebaldi, National Center for Atmospheric Research, Research Applications Program, Boulder, CO 80307-3000; email: tebaldi@ucar.edu

the quantity to be predicted is not as stable as its estimation "once and for all" may imply. Also, new models are put into operation, for which the history necessary to develop MOS equations is simply not available, until enough operational time has been logged. It may also be argued that averaging the relation between a forecast quantity and model quantities over a long history of seasons may wash out low frequency variations underlying many physical processes. For example, effects of current ENSO signals, for example, may be lost when applying laws whose coefficients have been statistically determined by averaging conditions over a large number of seasons.

For all these reasons, an appealing alternative consists of recursively estimating a new relation between predictand (observed) and predictors (model output) at a specific site by "looking back" at a recent history. On the basis of these recent data points a regression is dynamically estimated and updated.

A dynamic MOS system like this has been developed at NCAR and tested over thousands of sites, for forecast times out to three days and predictions of the following list of weather parameters like max and min temperature, probability of precipitation in 3, 6 or 24 hrs intervals, amount of precipitation in 3 and 6 hrs intervals, temperature and dew point temperature, u and v wind components and speed. For each forecast time, parameter and site a linear regression is estimated on the basis of the most recent 100 days' worth of data.

The system is not without pitfalls, however. Because of failures in recording observations or in model runs, the history may be incomplete and the resulting number of degrees of freedom for estimating the regression coefficients may be insufficient. Also, when trying to estimate "rare events" like precipitation (especially at certain locations) or extreme conditions, the sampling at a single location may not provide enough instances of those events. The length of the time window, fixed at 100 days somewhat arbitrarily, may also introduce spurious seasonal signals in the relation estimated by the regression. It is especially true for some parts of the year and some geographical areas that climatic conditions vary dramatically in a 100 day span. A shorter time window may be less prone to register deeply different conditions, but it may not be adequate for the parameters of a multivariate regression to be accurately and robustly estimated.

Another limiting aspect of this approach has to do with fitting the regression to the single site. It is to be expected that higher and higher resolution NWPs, or regionally nested models, will be able to produce predictors at grid points of interest where no observation is available. The single site approach may only produce a forecast at these points by some kind of interpolation of either the observed quantities at surrounding sites or directly from the forecasted quantities at surrounding sites. We will see in the next section that the idea of spatial composition may offer an alternative solution to this problem as well.

3 WHO IS NEIGHBOR OF WHOM

Spatial composition consists of estimating the regression between predictand and predictors over a neighborhood of sites. By compounding contemporary observations at different sites we may be able to shorten the time window and still sum up enough data points to robustly estimate the parameters of a multivariate regression. The neighborhood can be defined by several alternative criteria. We may call neighbors, alternatively:

1. Sites within a specified distance of each other. Distance here is simply Euclidean distance computed with respect to latitude/longitude coordinates.
2. Sites whose latitude and elevation are not "too different" with respect to the quantity to be predicted.
3. Sites whose climatology hints at common dynamics operating with respect to the quantity to be predicted.

More specifically, since we are going to test this idea in the prediction of maximum temperature (maxT) and daily probability of precipitation (PoP), the definitions become:

- With respect to definition 1 and for both types of prediction (maxT and PoP) the radius varies over 1.5, 2, 5, 8, 10 degrees.
- With respect to definition 2 and for prediction of maxT, neighbors are those sites within 2 degrees of latitude and 250 meters elevation.
- With respect to definition 3 and for prediction of PoP, neighbors are sites whose climatology is within a certain range of variations of PoP. Precisely, sites are taken to be part of the same neighborhood if the pair PoP, ΔPoP between them corresponds to one of the rows in Table 1. We use a twenty-year climatology to evaluate these quantities for each site and form the neighborhoods accordingly. The rationale behind the rules in Table 1 is to account for the rarity of precipitation at certain sites, and the more common occurrence at others. The more usual precipitation is as an event, the more flexible can be the definition of

Table 1: *Given a certain value of PoP p at station A, station B belongs to A's neighborhood if its value of PoP is within a certain range Δp, established as a function p. The first column of the table lists the intervals for p and the corresponding values of Δp are listed in the second column*

p	Δp
$< 7\%$	1%
$7\% - 18\%$	2%
$18\% - 30\%$	3%
$30\% - 46\%$	4%
$46\% - 56\%$	5%
$56\% - 68\%$	6%
$> 68\%$	8%

similarity among sites, as reflected by the larger size of the interval for sites that register precipitation more commonly. This can also be seen as a reflection of hypothesizing a bernoulli distribution for the occurrence of precipitation, whose variance is a function of the value of the probability of a positive event.

As said before, the expectation is that by combining observations and predictors from several sites we can get away with a shorter time span of data, thus avoiding seasonal signals; we can collect more observations of rare events (e.g precipitation occurrence); and we can provide a set of estimates of regression parameters that could be used for sites within the neighborhood at which predictors values are available but observations are not. As mentioned in Section 2 this is likely the case when "sites" are in fact points on a regular grid at which model output is available but where there is no recording station. No interpolation is needed here, since the values of the predictors at the "missing station" site are fed to the regression coefficients that have been estimated for its neighborhood.

4 DATA AND METHODS

Figure 1 shows the data sites used in the experiment. We apply the idea of neighborhoods (in its different incarnations as explained in Section 3) to each one of the sites, for both maxT and PoP prediction, for lead times of one and two days, and for four different NWP model outputs (AVN and ETA models, output produced at 0 and 12 UTC). Details on the models and their parameterizations can be found at `http://weather.unisys.com/model/details.html`.

For the prediction of maxT a multivariate linear regression is fitted between observed maxT at the site(s) and model output values. For PoP we use a multivariate logistic model (McCullagh and Nelder, 1983). See the Appendix for a list of the predictors in the regressions. We need to address the issue of overfitting, which arises when a multivariate model with a potentially high number of predictors is to be fitted to a limited number of observations. Our choice is to first rank the potential predictors by decreasing absolute value of their correlation with the predictand, and then choose the first $n/20$ predictors as terms of the regression, where n is the total number of data points available for regression estimation. We also cap the number, allowing no more than ten predictors in the regression. This is important in that we are going to estimate the performance of the different models by the mean square error (MSE) computed on a set of data points kept aside in the stage of parameter fitting. Overfitting of the training set by estimating a regression model with a large number of terms is a clear and present danger, and translates into very poor out-of-sample performance.

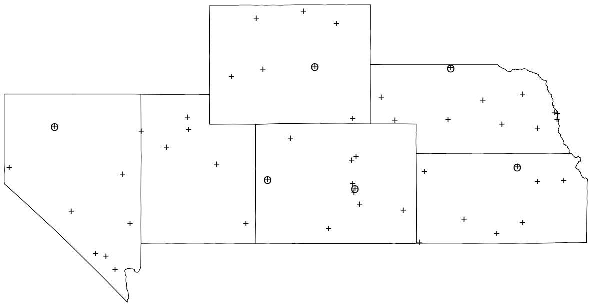

Figure 1: *The sites whose observations consitute the predictand values in our dataset. Model output – constituting the predictors set – has been interpolated to the sites' location. The sites indicated by a circle are used for the experiment on the spatial vs. temporal window's relevance for MSE control.*

We perform the out-of-sample exercise in two respects. We divide the data available at the site for which predictions are to be evaluated into a training set (in sample) and a test set (out of sample). We include the former in the estimation (together with contemporary observations from the neighborhood) but we evaluate performances only on the latter. This is what we mean by "train-test verification". Alternatively, we leave the observations at the site out of the estimation altogether, and still evaluate the prediction at the site on the test set. This way we mimic the case where a site provides predictors but no observed values of the predictand. This is defined as "leave-site-out verification".

5 RESULTS

We present results in the form of boxplots of MSE (also defined Brier score when a probability forecast is concerned) for the two forecasted quantities (maxT or PoP), four types of model output (AVN and ETA at 0Z, AVN and ETA at 12Z), and two lead times (one or two days ahead). These boxplots represent the distribution of MSE over a number of sites that varies between 35 and 40, since for some combination of the above factors (quantity, model, lead time) a few sites do not have enough data points to be considered for the single site regression, which is our benchmark in assessing the performance of the different definitions of neighborhood.

For each site we keep out of the training set the last 30 days of available history. The remaining 100 data points are either used for estimating the single site regressions, or combined with the contemporary data points at neighbor sites. Once the regression parameters are estimated, the 30 points that have been left aside are predicted upon, and the resulting MSE computed. In the case of the leave-site-out verification the 100 points for the site in exam are not included in the training set. The MSE is always computed on the same most recent 30 days of history. This particular set of results was obtained at the end of September 2001, with observations and model output spanning the most recent 4 months.

Notice how quite consistently across the different panels in Figure 2 through Figure 9, the boxplot trend is parabolically shaped, with a minimum that corresponds to the regression estimates resulting from pooling together sites within five to eight degrees distance. This result provides an indication that, at least on average, something is to be gained by simply adding observations from other sites, all other factors being equal, up to a certain distance. Beyond this point we are probably introducing noise rather than useful information in the regression. Here we present only plots for one day ahead prediction. Two days ahead predictions show the same kind of trend.

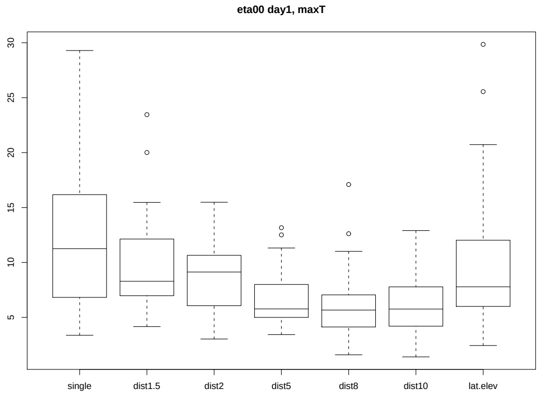

Figure 2: *Distribution of MSE for maxT (computed at each site over 30 point predictions) over the sites of Figure 1. "Single" corresponds to single-site regression, "dist1.5" through "dist10" correspond to regressions trained over neighborhoods of radius 1.5 through 10 degrees. "lat.elev" corresponds to regressions trained over neighborhoods defined in terms of latitude and elevation differences (see text). One day lead time. ETA model output at 0 UTC.*

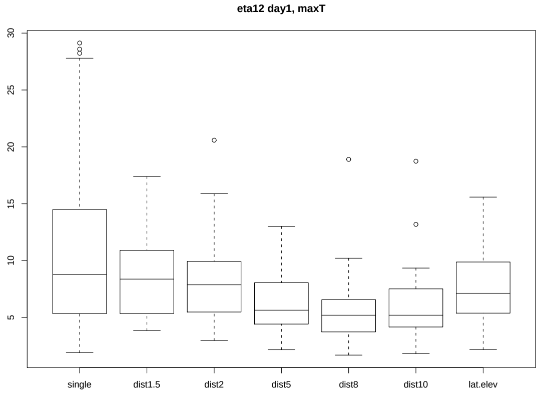

Figure 3: *As in Figure 2. ETA model output at 12 UTC.*

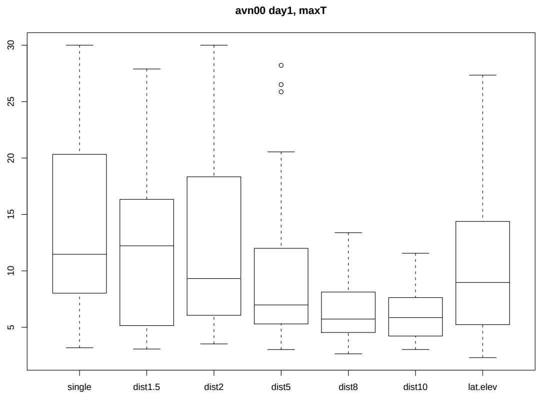

Figure 4: *As in Figure 2. AVN model output at 0 UTC.*

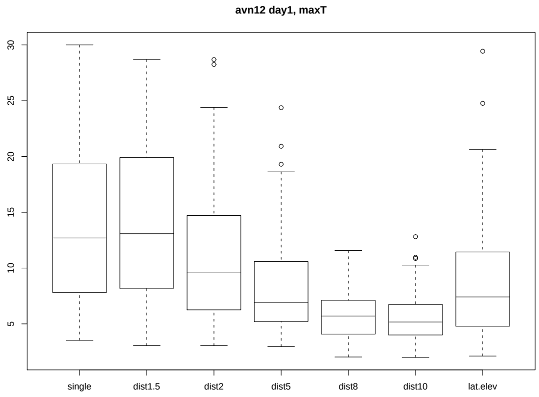

Figure 5: *As in Figure 2. AVN model output at 12 UTC.*

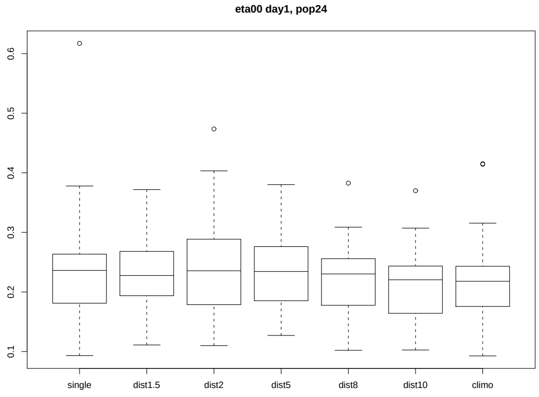

Figure 6: *As in Figure 2, for PoP, one day lead time. "climo" corresponds to regressions trained over neighborhoods defined by Table 1 in text. ETA model output at 0 UTC.*

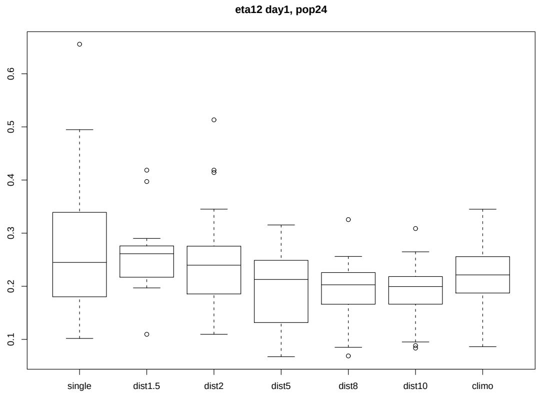

Figure 7: *As in Figure 6, ETA model output at 12 UTC.*

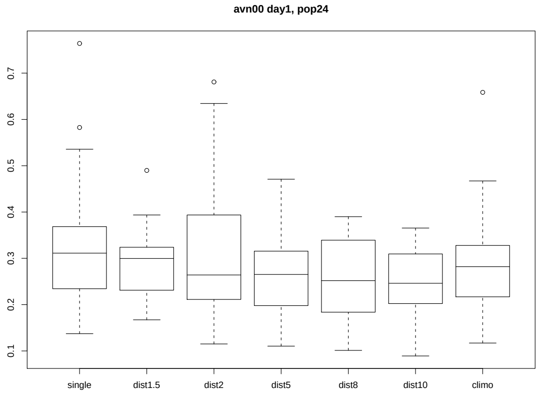

Figure 8: *As in Figure 6, AVN model output at 0 UTC.*

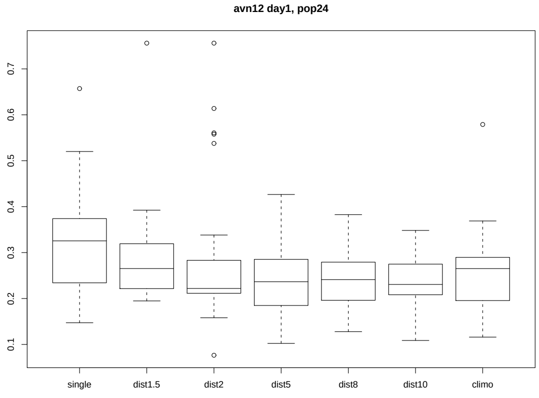

Figure 9: *As in Figure 6, AVN model output at 12 UTC.*

We now turn to the exercise in leave-site-out verification. Figure 10 shows some results specific a particular site and the prediction of maxT. The predicted values of maxT are here plotted against the true values. Those pertaining to the site left out of the training set are indicated by dark crosses in the plots. The cloud of light shaded dots represents the (true,predicted) pairs of the training set. The prediction quality is summarized in the value of the MSE at the top of each panel ('mseis' for the training set, 'mseos' for the test set) and seems to improve as we add points to the training set by enlarging the neighborhood around the station, as indicated by the title and the 'thicker and larger' cloud of dots. The line in the plots is the 45 degree line, the reference line where the dots and crosses would lie in case of perfect prediction. Such results are common across the majority of sites and all combinations of factors, again, but we cannot show more examples because of limited space. THere are a few cases where adding more sites does not improve the forecast, and we show one of them in Figure 11, but the majority of cases is exemplified by Figure 10.

As for the relationship between length of time window and size of neighborhoods, Table 2 exemplifies the findings of the following exercise, applied to all combinations of factors. For six sites, located across the area under study and labeled by circles in Figure 1, we vary the size of the neighborhood by adding one station at a time, closest first. Along the other dimension, for a given set of neighbors we vary the number of days included in the training set, from two weeks to 90 days. For each combination the MSE is computed on the - by now famous - 30 days kept out of the training set (because of space limits here we show only a subset of rows from the original table, i.e. several stations are added when comparing one row to the following).

The results in the table — again indicative of more general results — indicate the necessity of including at least 30 days to reach a one-digit-value of MSE. Once this number of time-data points is included, improvement can be achieved by keeping the neighborhood small and augmenting the data history, but even smaller values of MSE can be achieved by extending the training to sites farther away. After a certain point, enlarging the neighborhood doesn't seem to provide better forecasts, while increasing the time series length does continue to lead to improvements. It is possible to regard this result as a promising one. In cases when we want to use a recently introduced model or a recently recording station, the possible unavailability of a long history seems to be counteracted by the availability of data in a neighborhood.

Table 2: *Values of MSE for out-of-sample prediction of maxT at a specific site. The spatial dimension varies vertically, the time span of the data used for the regression varies orizontally. The values along the row labelled* 0 *are derived from the single site regression.*

	time span of data (days)					
distance (deg)	14	21	30	45	60	90
0	43.6	34.6	31.3	21.0	11.0	9.3
1.1	41.4	46.1	19.2	11.1	10.2	9.7
2.2	32.6	26.9	16.2	10.5	9.8	9.4
3.3	24.2	9.7	6.1	7.4	6.8	6.6
3.8	27.6	10.3	6.8	7.2	6.6	6.3
4.9	15.2	9.5	10.6	6.5	5.9	5.6
7.1	15.9	11.0	9.5	6.1	5.4	5.0
7.9	14.7	15.1	9.1	6.2	5.5	5.1
9.5	15.5	14.8	9.2	6.5	5.9	5.3

6 CONCLUSIONS

It may have been more interesting to see dramatic and robust differences in the optimality of the definitions of neighborhood for different sites, but it is certainly operationally easier to handle if the analysis suggests that a single definition of neighborhood may improve predictions fairly consistently over a large spatial domain. Specifically, our analysis seems to indicate that a neighborhood of sites within 5 degrees (approximately 500 kilometers) may provide useful information for fitting regressions of maxT and PoP in order to forecast these quantities out to 24 or 48 hours. We also found that to a certain degree one may trade-in historical information for spatial information, and maintain a small mean square error of the prediction. There are a large number of caveats, however, and a correspondingly large number of questions deserving further investigation. Our study has been limited with respect to the geographical area and season considered, quantities predicted, and lead time of the prediction. We are going to address all these issues in future work, where different seasons and forecast quantities, and long-term prediction as well as short-term will be studied. In addition, other areas of the globe will constitute new test cases. For now, we anticipate that the results we found by 'randomly' choosing this particular season and by predicting maximum temperature and probability of precipitation at numerous sites over a geographical territory that is quite diverse, are representative of a larger set of circumstances.

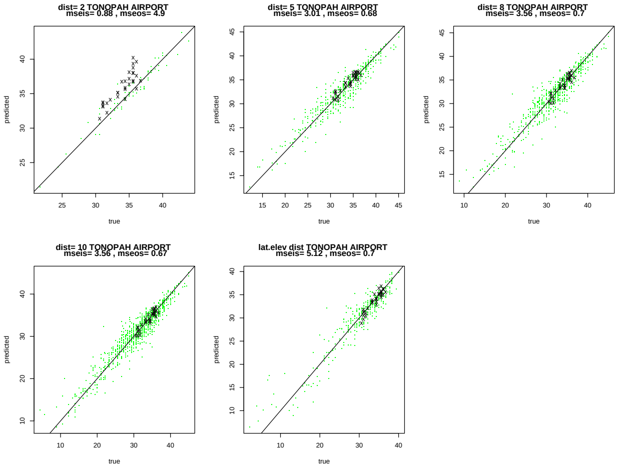

Figure 10: *Prediction at one site by a regression estimated only on its neighbors. The larger the neighborhood the better. Fitted vs true values of maxT. See text for details.*

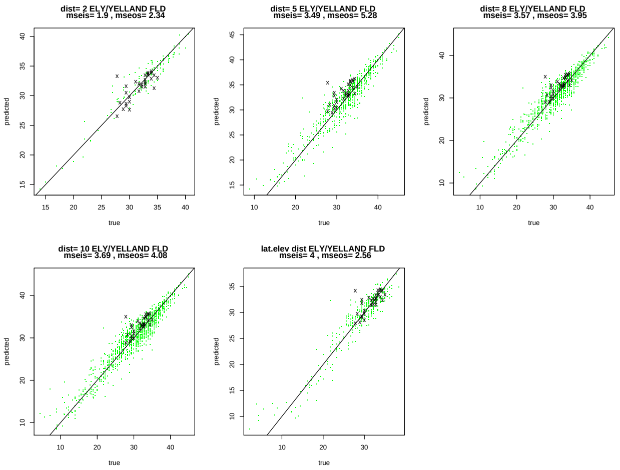

Figure 11: *Like Figure 10. But in this case enlarging the neighborhood does no good, it actually deteriorates the prediction.*

APPENDIX

List of predictors used in the regression of maxT. Many of them are to be considered as predicted by the NWP model at forecast time plus 6, 12 and 24 hrs earlier. The same is true for the predictors in the second list.

- temperature at surface, 925 mb, 700 mb;
- thickness 1000 - 700 mb, 850 - 700 mb;
- temperature difference 900 - 700 mb;
- u,v component of wind at 10m above surface;
- windspeed difference 1000 - 950 mb, 900 - 850 mb;
- cloud cover;
- relative humidity;
- latent/sensible heat flux;
- temperature advection 925mb;
- surface pressure;
- sin/cos of Julian day of the year.

List of predictors used in the regression of PoP.

- mean u,v component of wind 1000 - 700 mb;
- mean vertical velocity 850 - 500 mb;
- relative humidity;
- mean relative humidity 850 - 700 mb;
- precipitable water;
- 6hr probability of thunder;
- 24hr probability of precipitation;
- moisture advection at 925 mb;
- 24hr total precipitation accumulation;
- 24hr convective precipitation;
- 24hr stratiform precipitation;
- 24hr maximum CAPE;
- 24hr maximum CIN.

ACKNOWLEDGEMENTS

NCAR is sponsored by the National Science Foundation.

REFERENCES

Glahn, H.R. and D.A. Lowry, D.A., 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting, *Journal of Applied Meteorology*, 11, 1203-1211.

Glahn, H.R., Murphy, A.H., Wilson, L.J. and Jensenius, J.R.Jr., 1991: *Lectures Presented at the WMO Training Workshop on the Interpretation of NWP Products in Terms of Local Weather Phenomena and their Verification*, PSMP Report 34, World Meteorological Organization.

Mahoney, W.P., 2001a: An Advanced Weather Information Decision Support System for Winter Road Maintenance. *Preprints, 8th World Congress on Intelligent Transport Systems*, 30 September - 4 October 2001, Sydney, Australia.

Mahoney, W.P., 2001b: An Advanced Winter Road Maintenance Decision Support System. *Preprints, Intelligent Transportation Society of America (ITS) 2001*, 4 - 7 June 2001, Miami Beach, Florida.

McCullagh, P. and Nelder, J.A., 1983: *Generalized Linear Models.* Chapman and Hall, New York, NY.

Nott, D.J., Dunsmuir, W.T.M., Kohn, R. and Woodcock, F., 2001: Statistical Correction of a Deterministic Numerical Weather Prediction Model, *Journal of the American Statistical Association*, 96, 794-804.

Vislocky, R.L. amd Fritsch, J.M., 1995: Generalized Additive Models Versus Linear Regression in Generating Probabilistic MOS Forecasts of Aviation Weather Parameters, *Weather and Forecasting*, 10, 669-680.

Wikle, C.K., Berliner, L.M. and Cressie, N., 1998: Hierarchical Bayesian Space-Time Analysis, *Journal of Environmental and Ecological Statistics*,5, 117-154.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, San Diego, CA.