3.7

COMBINING FORECASTS FOR SUPERIOR PREDICTION

Gregory S. Young*
National Center for Atmospheric Research, Boulder, Colorado

1 INTRODUCTION

Automated forecasting systems often can create several forecasts of the same weather parameter (e.g. by generating linear regression equations to predict maximum temperature based on the output of several different NWP models). Research has shown that a combination of the forecasts usually produces a prediction that is superior to a single forecast (Vislocky and Fritsch 1995, Brown and Murphy 1996).

Traditional statistical methods, such as multiple linear regression, can be used to "weight" the various inputs. Unfortunately, the collinearity inherent in such forecasts often leads to instability and poorly conditioned design matrices. Other methods, such as partial least squares (Garthwaite 1994) and principal component regression (Jackson 1991), avoid this problem by forming linear combinations of the original predictors that are orthogonal. Thus, the inversion of a nearly singular matrix is avoided. The gradient or steepest descent method (Forsythe 1977) treats the suite of forecasts as a system of equations to minimize. In this case, the goal is to minimize squared error. The weights are adjusted for each set of forecasts and observations to move the final estimate closer to the minimum. Gradient descent does not find the combination of forecasts to give the absolute minimum, as opposed to linear regression. Rather, it gradually approaches the minimum.

The purpose of this research is to investigate these methods of combining forecasts. For now the scope is narrow, only considering forecasts of maximum temperature for the following day, although other weather parameters will be considered in the future. This study focused on the following methods: multiple linear regression (MLR), principal components regression (PCR), partial least squares (PLS), gradient descent (GD), and a simple average. These methods are used to combine forecasts from a variety of sources including Model Output Statistics (MOS) and climatology.

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

2 METHOD

Assuming familiarity with the method of linear regression, the somewhat less well known techniques of PCR, PLS and GD will be outlined in brief.

2.1 Principal Components Regression

As the name implies, PCR begins with the multivariate statistical concept of principal component analysis (Jackson 1991). Linear combinations of the original predictors are generated that form an orthogonal basis for the predictor space. Further, these components are ordered such that a projection of the original predictors on the first component results in a vector whose variance is a maximum among all possible choices of components. A projection on the second component gives a vector whose variance is second to only the projection on the first component. Principal component analysis amounts to a rotation of coordinate axes to a new coordinate system. The first few components often explain the majority of the variability in a system.

In PCR, the original predictor variables are transformed by projecting them onto the principal components and these are then regressed against the responses:

$$\mathbf{T} = \mathbf{XD} \tag{1}$$

where $\mathbf{T}$ is the matrix of new predictor values, $\mathbf{X}$ is the $n \times m$ matrix of original values, and $\mathbf{D}$ is the matrix of principal components. The new set of predictors is uncorrelated and therefore contains no redundant information. Thus, the regression equation is of the form

$$Y = \beta_0 + \beta_1 T_1 + \cdots + \beta_k T_k \tag{2}$$

where T_i is a column of the matrix $\mathbf{T}$, a linear combination of the predictors.

The primary advantages of PCR are the ability to summarize information in the original predictors as linear combinations, perhaps simplifying the model by representing the variables in a lower dimensional space, and numerical stability. A regression that uses all of the principal components will be equivalent to a regression on all of the original variables. Usually, only the first

few principal components are used to avoid over-fitting. Due to the diagonal nature of $\mathbf{T}'\mathbf{T}$, it is more numerically stable to obtain the coefficient estimates from

$$b_t = (\mathbf{T}'\mathbf{T})^{-1}(\mathbf{T}'Y) \tag{3}$$

than

$$b_x = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'Y) \tag{4}$$

if $\mathbf{X}$ is poorly conditioned.

PCR is not without its disadvantages. When the components are determined, only $\mathbf{X}$ is considered. Hence, there is not necessarily any correlation between the first few components of $\mathbf{X}$ and Y. It is possible that the last principal component could be perfectly correlated with the response while the others are uncorrelated.

2.2 Partial Least Squares

PLS is similar to PCR. The main difference is that while PCR ignores the response in the creation of components, PLS considers them in conjunction with the predictors. The goals of the two techniques are essentially the same, to explain the variability in Y as a linear function of components. The regression equation is identical to Equation 2 although the manner in which the T_i's are created differs.

The PLS algorithm is given as described by Garthwaite (1994). To simplify notation, let U_1 be the centered values of Y and V_{1j} be the centered values of X_j (the columns of $\mathbf{X}$). To obtain the first component T_1, regress U_1 against each V_{1j}. For $j = 1, ..., m$, the resulting set of regression equations is

$$\hat{U}_{1(j)} = b_{1j}V_{1j} \tag{5}$$

where $b_{1j} = \mathbf{v}'_{1j}\mathbf{u}_1/(\mathbf{v}'_{1j}\mathbf{v}_{1j})$. A weighted average is then taken of the various estimates of U_1:

$$T_1 = \sum_{j=1}^{m} w_{1j}b_{1j}V_{1j}. \tag{6}$$

While T_1 is undoubtedly useful in predicting U_1 (and Y) there is likely more information included in $\mathbf{X}$. That additional information can be estimated by the residuals of regressing V_{1j} on T_1. Let these residuals be denoted by V_{2j}. Also, the additional variability in Y that is not explained by T_1 can be estimated by regressing U_1 on T_1. Let these residuals be denoted by U_2. The next component, T_2, is a linear combination of the V_{2j} and should be useful in estimating U_2. It is constructed in exactly the same way as T_1.

The procedure extends iteratively to construct the components $T_2, ..., T_p$. Consider a component T_i that has been constructed from U_i and V_{ij} $(j = 1, ..., m)$. In order to create the T_{i+1}, V_{ij} $(j = 1, ..., m)$ is regressed against T_i to yield $\mathbf{t}'_i\mathbf{v}_{ij}/(\mathbf{t}'_i\mathbf{t}_i)$ as the regression coefficients. Thus, $V_{(i+1)j}$ is defined by

$$V_{(i+1)j} = V_{ij} - [\mathbf{t}'_i\mathbf{v}_{ij}/(\mathbf{t}'_i\mathbf{t}_i)]T_i. \tag{7}$$

Its sample values are $\mathbf{v}_{(i+1)j}$, the regression residuals. U_{i+1} is similarly defined as

$$U_{i+1} = U_i - [\mathbf{t}'_i\mathbf{u}_i/(\mathbf{t}'_i\mathbf{t}_i)]T_i, \tag{8}$$

and its sample values are $\mathbf{u}_{i+1}$. The remaining variability in Y is U_{i+1} and the remaining information in X_j is $V_{(i+1)j}$. Continuing the procedure as before, U_{i+1} is regressed against each $V_{(i+1)j}$. A set of j new predictors is produced of the form $b_{(i+1)j}V_{(i+1)j}$, where

$$b_{(i+1)j} = \mathbf{v}'_{(i+1)j}\mathbf{u}_{i+1}/(\mathbf{v}'_{(i+1)j}\mathbf{v}_{(i+1)j}). \tag{9}$$

As in Equation 6, a linear combination of these predictors is formed, giving the next component

$$T_{i+1} = \sum_{j=1}^{m} w_{(i+1)j}b_{(i+1)j}V_{(i+1)j}. \tag{10}$$

After the desired number of components has been generated, they are related to Y by ordinary least squares, yielding a regression equation (Equation 2).

Different weighting schemes have been used in PLS. The two most prominent are $w_{ij} = 1/m$ and $w_{ij} = \mathbf{v}'_{ij}\mathbf{v}_{ij}$. The latter sets $w_{ij} \propto \text{var}(V_{ij})$ and will be used in this study.

The actual implementation of PLS was performed in a slightly different manner than was presented here. The technique is described by Martens and Naes (1989) as the orthogonal scores algorithm and was carried out using code from Denham (1995).

2.3 Gradient Descent

The idea of GD or steepest descent has been around for over 150 years (Forsythe 1977). The goal is to locally minimize a function of n variables. Changing our notation from the linear regression framework, the combined forecast is of the form

$$\hat{Y} = \frac{\sum_{j=1}^{m} w_j X_j}{\sum_{j=1}^{m} w_j} + b \tag{11}$$

where w_j is a weight associated with forecast j and b is an overall bias term. In this context, the squared error of the combined forecasts is to be minimized. The GD method is used to update the weights on the individual forecasts to move the squared error of the final combined forecast to a minimum. The elements in response vector Y and design matrix X are considered

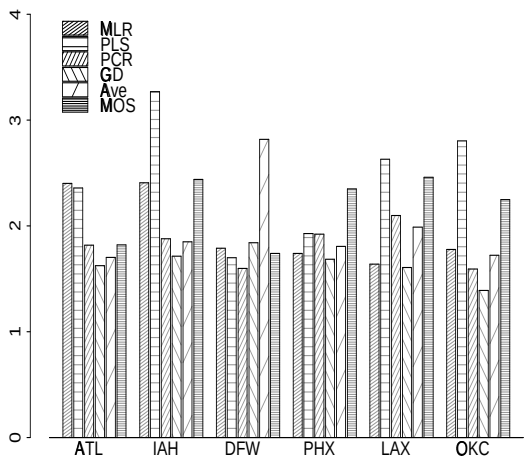

Figure 1: RMSE (in degrees C) for the five methods and NGM MOS for daily model refits. The sites shown are Atlanta, Houston, Dallas, Phoenix, Los Angeles, and Oklahoma City.

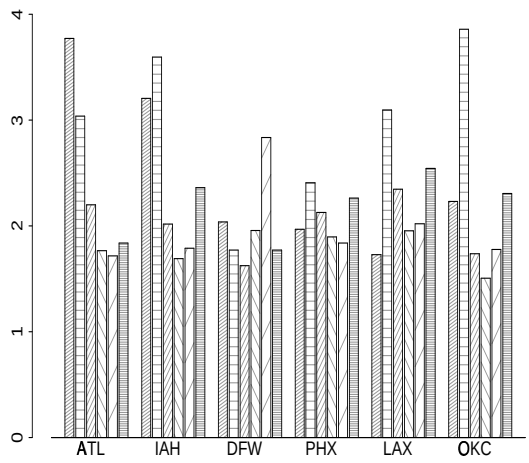

Figure 2: RMSE for weekly model refits for Atlanta, Houston, Dallas, Phoenix, Los Angeles, and Oklahoma City.

individually as in a time series, instead of all together. The weights are updated at each corresponding set of the response and forecasts.

Using calculus, the following equations are obtained for updating the parameters at time step t ($t = 1, ..., n$).

$$\Delta w_i^t = \zeta(Y^t - \hat{Y}^t)(X_i^t - \hat{Y}^t + b^t) \qquad (12)$$

and

$$\Delta b^t = \zeta(Y^t - \hat{Y}^t). \qquad (13)$$

The constant parameter ζ specifies the magnitude of the "steps" that are allowed. The changes in the w_i and b at step t are used to update the values at time $t+1$. Additional restrictions were placed on the weights requiring that they sum to one and be positive in value.

Assuming that the error surface is fairly stable, GD can produce very good estimates with a rather short training set. While it does not find the solution to give the absolute minimum squared error (as in regression), it will be "close" given an adequate amount of time. It also is very fast computationally and does not involve any matrix calculations.

3 INPUT FORECASTS

The input forecasts for this study came from a variety of sources. One was a climatological forecast based on 30 years of observational data. Several NWS MOS products were used, namely those derived from the NGM, MRF, and AVN models.

Also, several dynamic MOS (DMOS) forecasts were included (Mao et al. 1999). These DMOS forecasts are created from regression equations derived from a single model over a fixed period of time. The time period in this study was 100 days. Four DMOS forecasts were included, two for the ETA model (00Z and 12Z) and two for the AVN (00Z and 12Z).

One of the main interests in generating these combined forecasts is the ability to assimilate a large number of input forecasts effectively. Computation time may be increased, but the combined forecast should not significantly degrade as a result of including an additional input forecast, even if it is of poor quality.

4 RESULTS

A total of 18 different domestic locations were chosen for the study, corresponding to large cities. The variable of interest was maximum temperature at day 1 and a total of nine input forecasts were used in this stage (future work will be to include different weather parameters and more input forecasts). The study period extended from July 3, 2001 through September 30, 2001. On average, about 80 days of usable data were obtained for each of the sites. Three of the methods (PLS, PCR, and GD) required constants to be supplied. For PCR, the first principal component was used to determine the regression and the first three components

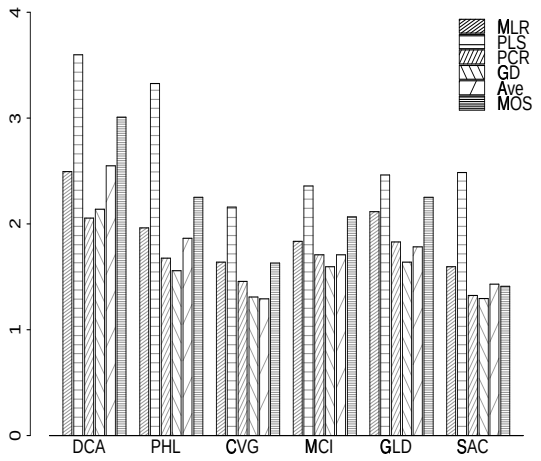

Figure 3: RMSE for daily refits for Washington DC, Philadephia, Cincinnati, Kansas City, Goodland KS, and Sacramento.

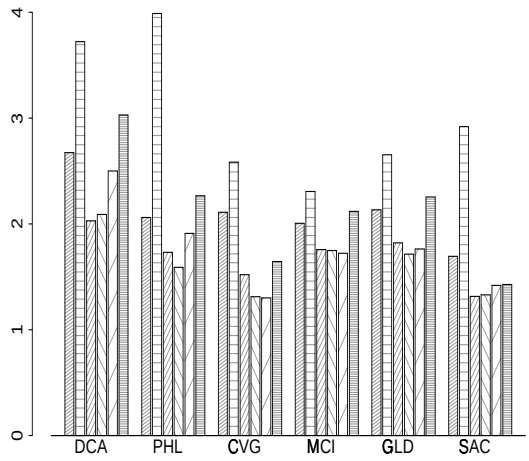

Figure 4: RMSE for weekly refits for Washington DC, Philadephia, Cincinnati, Kansas City, Goodland KS, and Sacramento.

were used for PLS. In GD, the step size (ζ) was chosen to be 0.01.

Two other variables of interest were the amount of past data to be used and the time between the refitting of the models (or updating of the weights). A long history of every input forecast and corresponding observations could be cumbersome to store for several weather parameters over hundreds or thousands of sites. In a real-time system, it also may be computationally prohibitive to refit the model every day. GD has an advantage on both of these counts in that it is very efficient in updating the weights and does not require a history of the actual forecasts, only the weights associated with them from the previous step.

Two analyses were run, one for refitting the models every day and one refitting once a week. The same data history of 30 days was maintained for both. MOS generated from the 12Z NGM run was included for comparison.

The results were fairly consistent for the various analyses. Unsurprisingly, using the more frequent update cycle improved prediction for all of the methods. Table 1 shows the average RMSE (across all sites) in degrees Celsius for the five methods and NGM MOS for both daily and weekly refittings. Note that GD performed the best for each refit cycle. PCR was the best of the regression techniques and was comparable to the simple average for both cycles. PLS consistently performed the worst. NGM MOS was exceeded by GD, PCR, and the average for both time frames. The other MOS forecasts fared worse (RMSE of 2.84 for MRF MOS and 3.02 for AVN MOS).

Method	MLR	PCR	PLS	GD	Ave	MOS
daily	2.00	1.78	2.62	1.62	1.85	2.07
weekly	2.34	1.90	3.02	1.71	1.86	2.08

Table 1: Average RMSE for daily and weekly model updates for forecasts of day 1 maximum temperature (in degrees C).

Figures 1 through 6 display the RMSE for each of the methods by site, with daily refittings and weekly refittings. GD performed well for all of the sites. Perhaps most importantly, its combined forecasts were consistent. It was never the worst method for any of the sites and was most frequently the best. Site 3 (Dallas, TX) was the one site where each of the regression methods yielded a lower RMSE over the study period than GD or the average.

5 CONCLUSIONS AND FUTURE WORK

From this small study, the gradient descent method or a simple average seem to be reasonable choices for combining input forecasts into a single prediction. GD is particularly appealing in light of its computational efficiency and its abrogation of the need for a lengthy

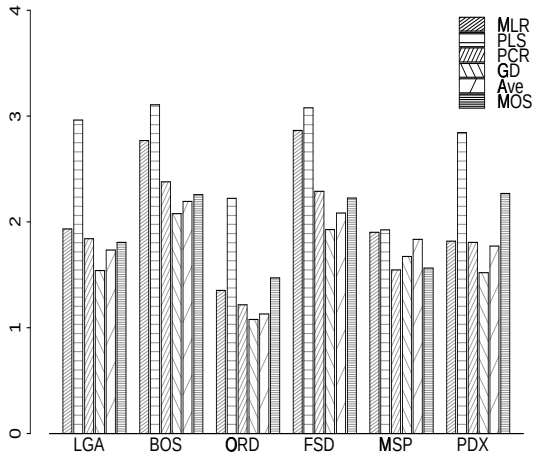

Figure 5: RMSE for daily refits for New York City, Boston, Chicago, Sioux Falls, Minneapolis, and Portland OR.

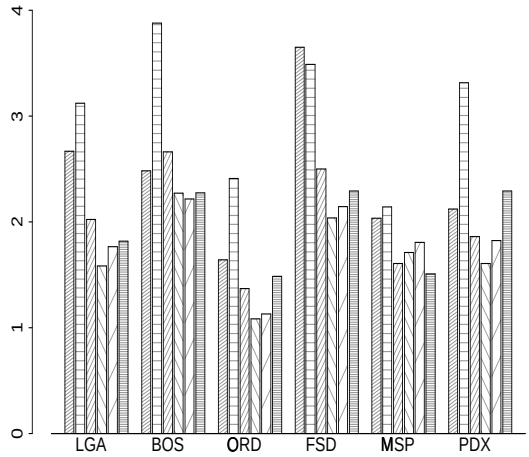

Figure 6: RMSE for weekly refits for New York City, Boston, Chicago, Sioux Falls, Minneapolis, and Portland OR.

data history. The simple average also performed quite well, only slightly worse than GD. Of the regression approaches, PCR performed the best. NGM MOS was outperformed by several of the methods, which further demostrates the value of a combined or consensus forecast. Much can be gained from combining forecasts in a coherent way.

Much work remains to be done on this project. Not only does the length of the study period need to be increased, other seasons should be examined. As more automated forecasts become available, certain methods may adapt better to the increased dimensionality of the problem. Several new NWS MOS products such as the new AVN MOS were not included. Day 1 maximum temperature is also considerably easier to predict than for instance, wind speed and direction. These challenging weather parameters should be considered. The simple average performed surprisingly well and some form of it may perform even better. For example, an average of only the MOS products could do very well, as illustrated in Vislocky and Fritsch (1995). Another criticism of combining forecasts is that sensitivity to extreme values is lost. Instances of extremes should be examined to check for such a dampening.

REFERENCES

Brown, B. G. and Murphy, A. H., 1996: Improving forecasting performance by combining forecasts: the example of road-surface temperature forecasts. *Meteorol. Appl.*, **3**, 257-265.

Denham, M.C., 1995 PLS Software obtained from Netlib repository.

Forsythe, G. E., Malcolm, M. A., and Moler, C. B. 1977 **Computer Methods For Mathematical Computations**, Prentice-Hall, London.

Garthwaite, P. H., 1994: An interpretation of partial least squares. *Journal of the American Statistical Association*, **89**, 122-127.

Jackson, J., 1991 **A User's Guide to Principal Components**, John Wiley & Sons, Inc.

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

Martens, H. and Naes, T., 1989 *Multivariate Calibration*, Wiley, Chichester.

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