(Statistical Forecasting: with NWP).
Notes from Kalnay (2003), appendix C

Postprocessing of Numerical Model Output to Obtain Station Weather Forecasts: MOS and Adaptive Regression based on simple Kalman Filter

If the numerical model forecasts are skillful, the forecast variables should be strongly related to the weather parameters of interest to the "person in the street" and for other important applications. These include precipitation (amount and type), surface wind, and surface temperature, visibility, cloud amount and type, etc. However, the model output variables are not optimal direct estimates of local weather forecasts. This is because models have biases, the bottom surface of the models is not a good representation of the actual orography, and models may not represent well the effect of local forcings important for local weather forecasts. In addition, models do not forecast some required parameters, such as visibility and probability of thunderstorms.

In order to optimize the use of numerical weather forecasts as guidance to human forecasters, it has been customary to use statistical methods to "post process" the model forecasts and adapt them to produce local forecasts. In this Appendix we discuss two of the methods that have been used for this purpose.

1) Model Output Statistics (MOS)

This method, when applied under ideal circumstances, is the gold standard of NWP model output post processing (Glahn and Lowry, 1972, Carter et al, 1989). MOS is essentially multiple linear regression where the predictors h_{nj} are model forecast variables (e.g., temperature, humidity or wind at any grid point, either near the surface or in the upper levels), and may also include other predictors such as astronomical or geographical parameters (e.g., latitude, longitude and time of the year) valid at time t_n . The predictors could also include past observations. The predictand y_n is a station weather observation (e.g., maximum temperature or wind speed) valid at the same time as the forecast. In other words, MOS produces an improved forecast, typically reducing the forecast bias, and becoming closer to climatology as the length of the forecast increases. In MOS, like in any statistical regression, the quality of the results improves with the quality and length of the training data set used to determine the regression coefficients b_j.

In order to underline the relationship between MOS and Adaptive Regression (based on Kalman Filter) we use for MOS the notation that Kalnay (2003) uses for Kalman Filter, with J predictors h_{nj}.

The *dependent* data set used for determining the regression coefficients is

$$\begin{aligned} y_n &= y(t_n), \quad n = 1,\dots,N \\ h_{nj} &= h_j(t_n), \quad n = 1,\dots,N;\ j = 1,\dots,J \end{aligned} \qquad (1.1)$$

where we consider one predictand y_n as a function of time t_n and J predictors h_{nj}.

The linear regression (forecast) equation is

$$\hat{y}_n = b_0 + \sum_{j=1}^{J} b_j h_{nj} = \sum_{j=0}^{J} b_j h_{nj}, \qquad (1.2)$$

where for convenience the predictors associated with the constant term b_0 are defined as $h_{n0} \equiv 1$. In linear regression the coefficients b_j are determined by minimizing the sum of squares of the forecast errors over the training period (e.g., Wilks, 1995). The sum of squared errors is given by:

$$SSE = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} e_n^{\,2} \qquad (1.3)$$

Taking the derivatives with respect to the coefficients b_j and setting them to zero we obtain:

$$\frac{\partial SSE}{\partial b_j} = 0 = \sum_{n=1}^{N} (y_n - \sum_{l=0}^{J} b_l h_{nl}) h_{nj},\ j = 0,1,...,J \qquad (1.4)$$

or

$$\sum_{n=1}^{N} \left[h_{jn}{}^T y_n - h_{jn}{}^T \sum_{l=0}^{J} h_{nl} b_l \right] = 0,\ j = 0,...,J \qquad (1.5)$$

where $h_{jn}^T = h_{nj}$. Eqs. (1.5) are the “normal” equations for multiple linear regression that determine the linear regression coefficients $b_j,\ j = 0,...,J$. In matrix form, they can be written as

$\mathbf{H}^T \mathbf{y} = \mathbf{H}^T \mathbf{H} \mathbf{b}$ or

$$\mathbf{b} = \left(\mathbf{H}^T \mathbf{H} \right)^{-1} \mathbf{H}^T \mathbf{y} \qquad (1.6)$$

where

$$\mathbf{H} = \begin{bmatrix} 1 & h_{11} & ... & h_{1J} \\ 1 & h_{21} & ... & h_{2J} \\ ... & ... & h_{nj} & ... \\ 1 & h_{N1} & ... & h_{NJ} \end{bmatrix},\ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ ... \\ b_J \end{bmatrix},\ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{bmatrix} \qquad (1.7)$$

are, respectively, the dependent sample predictor matrix (model output variables, geographical and astronomical parameters, etc.), the vector of regression coefficients, and the vector of predictands in the dependent sample. $\hat{\mathbf{y}} = \mathbf{Hb}, \mathbf{e} = \mathbf{y} - \mathbf{Hb}$ are the linear predictions and the prediction error respectively in the dependent sample. The *dependent* estimate of the error variance of the prediction is $s_e{}^2 = \frac{SSE}{N - J - 1}$ since the number of degrees of freedom is *N-J-1*. This indicates that one should avoid over fitting the dependent sample by ensuring that $N >> J$. For independent data, the expected error can be considerable larger than the dependent estimate s_e^2 because of the uncertainties in estimating the coefficients b_j .

The best way to estimate the skill of MOS (or any statistical prediction) that can be expected when applied to *independent* data is to perform *cross-validation*. This can be done by reserving a portion (such as 10%) of the dependent data, deriving the regression coefficients from the other 90%, and then applying it to the unused 10%. The process can be repeated 10 times with different subsets of the dependent data to increase the confidence of the cross-validation, but this also increases the computational cost.

It is clear that for a MOS system to perform optimally, several conditions must be fulfilled:

a) The training period should be as long as possible (several years).
b) The model-based forecasting system should be kept unchanged to the maximum extent possible during the training period.
c) After training, the MOS system should be applied to future model forecasts that also use the same unchanged model system.

These conditions, while favorable for the MOS performance, are not favorable for the continued improvement of the NWP model, since they require “frozen” models. The main advantage of MOS is that if the conditions stated above are satisfied, it achieves the best possible linear prediction. Another advantage is that it naturally takes into account the fact that forecast skill decreases with the forecast length, since the training sample will include, for instance, the information that a 1-day model prediction is on the average considerably more skillful than a 3-day prediction. The main disadvantage is that MOS is not easily adapted to an operational situation in which the model and data assimilation systems are frequently upgraded.

Typically, MOS equations have 10-20 predictors chosen by forward screening (Wilks, 1995). In the US NWS, the same MOS equations are computed for a few (4-10) relatively homogeneous regions in order to increase the size of the developmental database.

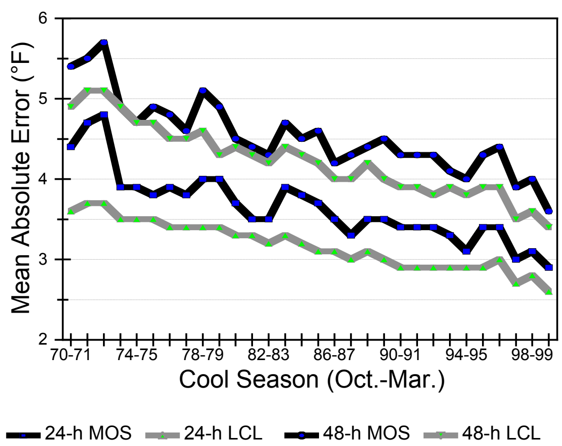
In order to stratify the data into few but relatively homogeneous time periods, separate MOS equations are developed for the cool season (October to March) and the warm season (April to September). As shown in Table 1 in the Adaptive Regression section, MOS can reduce very substantially the errors in the NWP model forecasts, especially at short lead times. At long lead times, the forecast skill is lost, so that the MOS forecast becomes a climatological forecast and the MOS forecast error variance asymptotes to the climatology error.

The error variance of an individual NWP forecast, on the other hand, asymptotes to twice the climatological error variance, plus the square of the model bias so that the advantage of MOS over the original NWP forecast increases substantially with the length of the forecast.

Figure 1 shows the evolution of the error in predicting the maximum temperature by the statistical guidance (MOS) and by the local human forecasters (LCL). The human forecasters skill in the 2-day forecast is now as good as the one-day forecast was in the 1970's. The human forecasters bring added value (make better forecasts) than the MOS statistical guidance, which in turn is considerably better than the direct NWP model output. Nevertheless, the long-term improvements are driven mostly by the improvements in the NWP model and data assimilation systems (Chapter 1 of Kalnay, 2003).

In summary, the forecast statistical guidance (and in particular MOS) adds value to the direct NWP model output by objectively interpreting model output to remove systematic biases and quantifying uncertainty, predicting parameters that the model does not predict, and producing site-specific forecasts. It assists forecasters providing a first guess for the expected local conditions, and allows convenient access to information on local model and climatology conditions.

Fig. 1: Evolution of the mean absolute error of the MOS guidance and of the local official NWS forecasts (LCL) averaged over the US (Courtesy of J. Paul Dallavalle and Valery Dagostaro from the US NWS).



2) Adaptive Regression based on a simple Kalman Filter (AR)

Adaptive Regression based on Kalman Filtering has also been widely used as a postprocessor. In MOS or in other statistical prediction methods such as nonlinear regression or neural networks, the regression coefficients are computed from the dependent sample, and are not changed as new observations are collected until a new set of MOS equations are derived every 5 or 10 years. Because the regression coefficients are constant, the order of the observations is irrelevant in MOS, so that older data have as much influence as the newest observations used to derive the coefficients.

In Adaptive Regression, the Kalman Filter equations (Kalnay, 2003, Chapter 5, Section 6) are applied in a simple, sequential formulation to the multiple regression coefficients $\mathbf{b}_k = \mathbf{b}(t_k)$, whose values are *updated* every time step t_k, rather than keeping them constant as we did in MOS (eq 1.2 $\hat{y}_n = b_0 + \sum_{j=1}^{J} b_j h_{nj} = \sum_{j=0}^{J} b_j h_{nj}$)

$$\hat{y}_k = \sum_{j=0}^{J} b_j(t_k)h_{kj} = \begin{bmatrix} 1 & h_{k1} & \ldots & h_{kJ} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_J \end{bmatrix}_k = \mathbf{h}_k^T \mathbf{b}_k \tag{1.8}$$

Recall that Kalman Filtering consists of two steps. In the first step, starting from the analysis at time t_{k-1}, we forecast the values of the model variables (in this case the coefficients $\mathbf{b}_k$) and their error covariance at time t_k. In the second step, the Kalman weight matrix is derived, and, after obtaining the observations at time t_k, the model variables and error covariance are updated, obtaining the analysis at time t_k, which is more accurate than the forecast.

In Adaptive Regression, the “forecast” or first guess of the regression coefficients at t_k is simply that they are the same as the updated (analysis) coefficients at t_{k-1}. Their error covariance is the same as that estimated in the previous time step, plus an additional error introduced by this “regression forecast model”:

$$\begin{aligned} \mathbf{b}_k^f &= \mathbf{b}_{k-1}^a \\ \mathbf{P}_k^f &= \mathbf{P}_{k-1}^a + \mathbf{Q}_{k-1} \end{aligned} \tag{1.9}$$

Here $\mathbf{Q}_k = \mathbf{q}_k \mathbf{q}_k^T$ is the “regression model” error covariance (a matrix of tunable coefficients that is diagonal if we assume that the errors of the different coefficients are not correlated).

The Kalman gain or weight vector for adaptive regression is given by

$$\mathbf{k}_k = \mathbf{P}_k^f \mathbf{h}_k (\mathbf{h}_k^T \mathbf{P}_k^f \mathbf{h}_k + r_k)^{-1} \tag{1.10}$$

Note that for a single predictand, the forecast error covariance $\mathbf{h}_k^T \mathbf{P}_k^f \mathbf{h}_k$ and the observational error covariance $\mathbf{R}_k = r_k$ are both scalars, and computing the Kalman gain matrix does not require a matrix inversion.

At time t_k the observed forecast error or innovation $e_k = y_k^o - \mathbf{h}_k^T \mathbf{b}_k^f$ is used to *update* the regression coefficients:

$$
\begin{aligned}
\mathbf{b}_k^a &= \mathbf{b}_k^f + \mathbf{k}_k (y_k^o - \mathbf{h}_k^T \mathbf{b}_k^f) \\
\mathbf{P}_k^a &= (\mathbf{I} - \mathbf{k}_k \mathbf{h}_k^T) \mathbf{P}_k^f
\end{aligned}
\qquad (1.11)
$$

In summary, the adaptive regression algorithm based on Kalman Filtering can be written as:

$$
\begin{aligned}
y_k^f &= \mathbf{h}_k^T \mathbf{b}_{k-1}^a \\
\mathbf{P}_k^f &= \mathbf{P}_{k-1}^a + \mathbf{Q}_{k-1} \\
e_k &= y_k^o - y_k^f \\
w_k &= \mathbf{h}_k^T \mathbf{P}_k^f \mathbf{h}_k + r_k \\
\mathbf{k}_k &= \mathbf{P}_k^f \mathbf{h}_k w_k^{-1} \\
\mathbf{b}_k^a &= \mathbf{b}_{k-1}^a + \mathbf{k}_k e_k \\
\mathbf{P}_k^a &= \mathbf{P}_k^f - \mathbf{k}_k w_k \mathbf{k}_k^T
\end{aligned}
\qquad (1.12)
$$

where w_k is a temporary scalar defined for convenience. The two tuning parameters in the algorithm are r_k, the observational error covariance (a scalar), and $\mathbf{Q}_k$, the “regression model” error covariance (a diagonal matrix with one coefficient for the variance of each predictor if the errors are uncorrelated).

Unlike regression, MOS, or neural networks, Adaptive Regression is *sequential*, and gives more weight to recent data than to older observations. The larger $\mathbf{Q}_k$, the faster older data will be forgotten. It also allows for observational errors. This method can be generalized to several predictands, in which case the observation error covariance matrix may also include observational error correlations.

The following table compares a simple Kalman Filtering applied to the 24 hr surface temperature forecasts for July and August 1997 at 00Z, averaged for 8 different US stations, using as a single predictor the global model output for surface temperature interpolated to each individual station. It was found that after only a few days of spin-up, starting with a climatological first guess, and with minimal tuning, the AR algorithm was able to reach a fairly steady error level substantially better than the numerical model error, and not much higher than regression on the dependent sample which is the best possible regression result. Not surprisingly, MOS, using many more predictors, and several years of training, provides an even better forecast than this simple AR.

NWP (Global model)	Dependent Regression	Adaptive Regression	MOS
5.36K	2.67K	3.07K	2.29K

Table 1: RMS error in the forecast of the surface temperature at 00Z averaged for 8 US stations. In the Dependent Regression and Kalman Filtering, the only predictor used was the direct model prediction of the temperature interpolated to the station. The MOS prediction has more than 10 predictors and several years of training.

In summary, Kalman Filtering provides a simple algorithm for adaptive regression. It requires little training so that it is able to adapt rather quickly to changes in the model, and to long-lasting weather regimes. It is particularly good in correcting model biases. However, in general it is not as good as regression based on long dependent samples.