How to best combine statistical-empirical relationships to downscale seasonal forecasts

Problem presented by

Christopher Nankervis

Weather Logistics

ESGI107 was jointly hosted by
The University of Manchester
Smith Institute for Industrial Mathematics and System Engineering

with additional financial support from
KTN Ltd
Natural Environment Research Council
Manchester Institute for Mathematical Sciences

Report author

Matthew Cooks (University of Manchester)

Executive Summary

Fine-scale seasonal average weather forecasts are produced by Weather Logistics from global weather predictors and measurements using a novel empirical model. Forecasts of this kind allow better management of agricultural risks and food supply chain operations. Uncertainties arise at each step in the forecast process and propagate through to affect the final fine-scale forecasts; quanitifying this uncertainty is crucial for allowing better decision making and cost-benefit analyses.

The ares of investigation posed to the study group were to identify the best global predictors to improve the forecasts and to quantify the uncertainty that propagtes through the down-scaling process of the forecasting.

Kalmann filters were investigated to attempt to improve the predictive capabilities of the time series measurements of, for example, the El Nino index. This method reduces the influence of noise in the measurements and consequently the uncertainty in the data input into the empirical models.

The second area looked at was correlating various global measurements to the temperature in the North-west of England to attempt to identify the most important at different times of year. It was found that the large scale predictors that are most important in the summer may not be as efficient in the winter and a different set of predictors may need to be identified to predict the weather during the colder months. A weighted linear combination of a summer and winter predictor was derived which could offer a method of combining two different seasonal relations into one empirical relation valid throughout the year.

Version 1.0
June 10, 2015
iv+10 pages

Contributors

Colm Connaughton (University of Warwick)
Aikaterini Kaouri (Univeristy of Cyprus)
Pierluigi Cesana (University of Oxford)
William Parnell (University of Manchester)
Dave Abrahams (University of Manchester)

Contents

1 Introduction and Problem Overview

(1.1) Forecast downscaling is implemented by Weather Logistics by combining a set of regional information from global weather predictions with measurements; from satellites, in-situ observations and re-analyses. Fine-scale seasonal climate predictions, describing average weather conditions, are then produced in conjunction with output uncertainties on a 25km surface grid. Forecasts of this kind allow better management of agricultural risks and food supply chain operations. The company adopts an empirical (diagnostic) model that combines independent datasets from around the globe, which differs from a conventional regionally nested model.

(1.2) Operational seasonal predictions are produced by extracting and processing measurements and re-analyses; summarising natural variability in atmospheric, oceanic and land-surface variables in broad horizontal areas while also ensuring that any decadal climate trends are removed. Monthly to seasonal cycles generally offer the best predictive skill for seasonal forecasting and a selection process is first applied to identify these climate variables. Broad descriptions, in terms of character and spatial area, are used to describe global predictors. Those that best correlate with the monthly to seasonal variability monitored at historical weather stations are selected and empirical-statistical relationships formed to demonstrate how they are linked. The company then applies these relationships to the same predictors, which are obtained from seasonal forecast models.

(1.3) The problem is that uncertainties arise at each step in the forecast process: such as selecting broad features of the upper-level winds (jet stream). Jet stream attributes are summarised and condensed into a forecast index offering commercial potential at least on a diagnostic level. Operational skill is, however, diminished by overstatement of the predictive capabilities of jet stream dynamics, both its realistic characterisation and mechanistic proof in the process for generating a new climate index. Other prediction uncertainties arise from the use of mid-latitude sea-surface temperature data, particularly for the North Atlantic that offers potential for winter temperature forecasting. Summer rainfall forecast is also troublesome, despite accurate representations of tropical ocean dynamics.

(1.4) Several areas of focus have been identified; assessed using time series of predictor measurements and forecast data: selection of jet stream parameters to form a predictive index, mechanistic studies to assess the validity of the seasonal forecast process, and refining and lowering uncertainties in numerical weather predictions.

(1.5) The company wishes to improve estimates of predictand uncertainties; allowing better decision making and unbiased calculation of risk. This will allow better cost-benefit analyses for its supplied industries; such as for the (re)insurance and in the trading of soft commodities. While the company understands the uncertainties of the measurements and re-analyses, they are unsure about how these transcribe when downscaling is applied to operational forecasts.

2 The solution

2.1 Kalmann Filter

2.1.1 Introduction

(2.1.1) A Kalman filter is both a tool and a theoretical approach to make predictions on the state of a system based on an actual measurement. It tells us how to combine two (or more) different noisy estimates of the state of the system to produce a better estimate of the state than either estimates taken in isolation.

(2.1.2) In the current problem we have a time series of observations of the El Nino index, y_t, which we assume are subject to noise. Our second way of estimating the El Nino is to use a model which will also be subject to error. The generic state space model is as follows

$$\begin{aligned} y_t &= Z\alpha_t + \zeta_t, \quad \zeta_t = N(0, \sigma^2_{\zeta_t}), \\ \alpha_{t+1} &= M\alpha_t + R\xi_t, \quad \xi_t = N(0, \sigma^2_{\xi_t}). \end{aligned}$$

Here

- Z is the observation operator,
- α_t is the true state at time t (which cannot be directly observed),
- ζ_t is the observation error,
- M is a linear operator representing our model of how the *true state* evolves in time,
- ξ_t is the model error,
- R is an operator to perform moving averaging on the model error (if this is part of the model).

(2.1.3) The Kalman filter is a recursive algorithm to estimate the state α_t given the state at the previous time, α_{t-1}, and the observation y_t at time t:

$$\begin{array}{rclrcl} v_t &=& y_t - Z\alpha_t, & F_t &=& ZP_tZ^T + \sigma^2_\xi, \\ \alpha_{t+1} &=& M\alpha_t + K_t v_t, & P_{t+1} &=& MP_t(M - K_tZ)^T + R\sigma^2_\xi R^T, \end{array} \qquad (1)$$

where

- $K_t = MP_tZ^TF_t^{-1}$ is the Kalman gain matrix.
- P_t is the estimated covariance matrix of the hidden state α_t (calculating P_t as one goes along to estimate the uncertainty in the filtered time series.

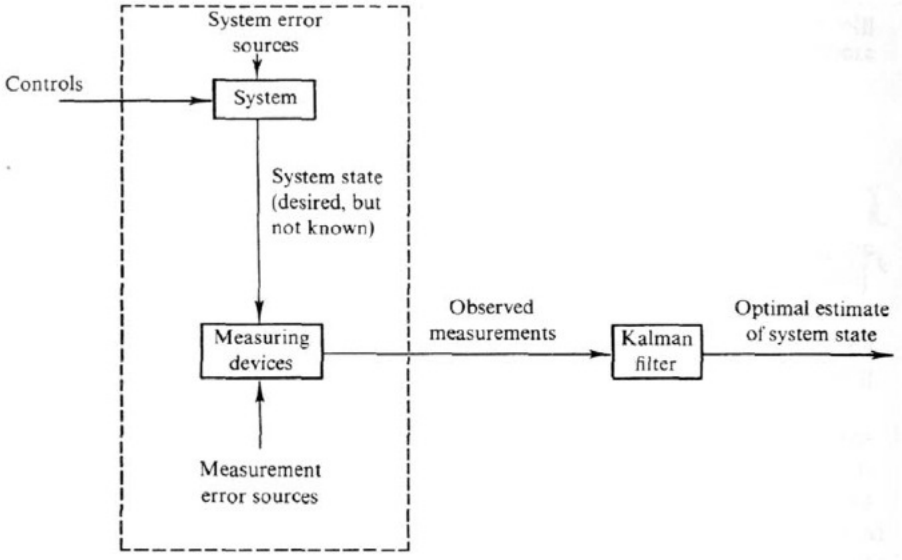
System error sources
Controls
System
System state (desired, but not known)
Measuring devices
Measurement error sources
Observed measurements
Kalman filter
Optimal estimate of system state

Figure 1: A flow chart showing the process involved in developing a better estimate of the state using a Kalman filter.

2.1.2 The model

(2.1.4) The first step is to come up with a model M. In order to do this we look at the ACF of the data: the ACF of the raw data decays slowly thus indicating that the series has long term memory (or non-stationarity). We then calculated the differenced time series:

$$y_t^* = y_t - y_{t-1}.$$

Plotting the ACF of y_t^* indicates that the ACF of the difference decays within two or three lags. The series y_t^* is therefore stationary and should be plausibly modelled by a simple autoregressive moving average process

$$y_{t+1}^* = \phi_1 y_t^* + \phi_2 y_{t-2}^* + \theta_1 \xi_{t-1} + \xi_t,$$

(here according to an ARIMA(2,1) process). This is equivalent to the following state space from

$$\begin{aligned} y_t &= (1,1,0)\alpha_t + \zeta_t, \\ \alpha_{t+1} &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & \phi_1 & 1 \\ 0 & \phi_2 & 0 \end{pmatrix} \alpha_t + \begin{pmatrix} 0 \\ 1 \\ \theta_1 \end{pmatrix} \xi_{t+1}, \end{aligned} \tag{2}$$

where the hidden state is

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \xi_t \end{pmatrix}.$$

The next step is to estimate the appropriate values for the parameters $\phi_1, \phi_2, \theta_1, \sigma_\zeta^2, \sigma_\xi^2$ in order to make the model (2) a good simulation of the observed data y_t.

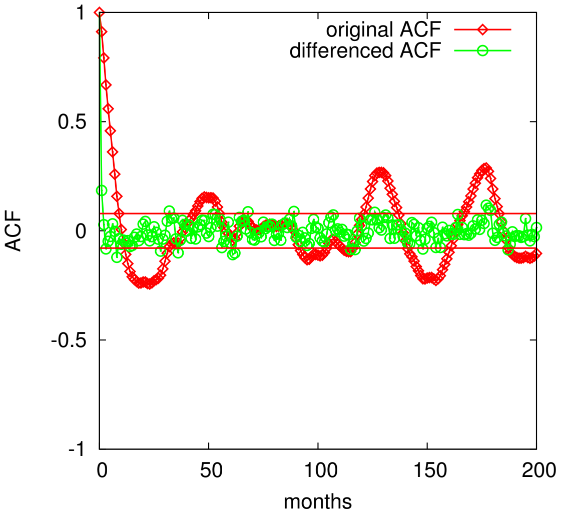

Figure 2: ACF of the El Nino time series

(2.1.5) Since we have a model for the series we can now calculate, in principle, the probability $P(y_1, \ldots, y_n)$ of observing the sequence of observations $(y_1, \ldots, y_n)$ for any given values of the parameters $\phi_1, \phi_2, \theta_1, \sigma_\zeta^2, \sigma_\xi^2$. Since the model (2) goes one step at a time:

$$P(y_1, \ldots, y_n) = P(y_n | Y_{n-1}) p(Y_{n-1}),$$

where Y_i is shorthand for the sequence $\{y_1, y_2, \ldots, y_i\}$ by recursion. Therefore:

$$P(Y_n) = \Pi_{i=1}^n p(y_i | Y_{i-1}).$$

Part of the Kalman filter is that it gives an optimal, in the sense of minimal variance, normally distributed estimate of y_t given the sequence $\{y_1 y_2, \ldots y_{t-1}\}$ of previously observed values. Therefore,

$$P(y_i | Y_{i-1}) = N(\alpha_i, F_i).$$

Thus, the probability or *likelihood* of the observed data $\{y_1, \ldots, y_n\}$ given the model (2) is written as a function of the model parameters is

$$L(\phi_1, \phi_2, \theta_1, \sigma_\zeta^2, \sigma_\xi^2) = \Pi_{t=1}^n \frac{1}{\sqrt{2\pi F_t}} e^{-\frac{1}{2} \frac{(y_y - \alpha_t)^2}{F_t}}.$$

It can be noted that this is explicitly computed from (1). The log-likelihood is

$$\log L = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^n \log F_t + \frac{(y_t - \alpha_t)^2}{F_t}.$$

The values of the parameters which best fit the data are obtained by maximising $\log L$ with respect to the parameters

$$\{\phi_1^*, \phi_2^*, \theta_1^*, \sigma_\zeta^*, \sigma_\xi^*\} = \arg \max_{\{\phi_1,\phi_2,\theta_1,\sigma_\zeta^2,\sigma_\xi^2\}} \log L.$$

We did this maximisation numerically in Mathematica using the *downhill simplex* algorithm. The results are:

- $\phi_1^* \approx 0.57$,
- $\phi_2^* \approx -0.29$,
- $\theta_1^* \approx 0.54$,
- $\sigma_\zeta \approx 0.19$,
- $\sigma_\xi \approx 0.22$.

Now that the model parameters are known, performing the filtering is simply a case of plugging Eqs. (1.2) into the Kalman filter (1.1) and iterating forward. To initialize the filter we used

$$\alpha_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

As a result there is an initial transient which dies away quite quickly as the filter assimilates the observations (this ad hoc initialisation could be improved in principle but we didn't look into this).

2.1.3 Forecasting

(2.1.6) To run the filter in forecast mode we simply iterate the Kalman filter equations (1.1) into the future with $K_t = 0$. We tried to do a three month forecast. The resulting three month forecast is a incredibly accurate compared to the measured data with the peak around month 130 captured almost exactly. Similarly the smaller peak around months 150 and the approximately constant negative region in the initial 120 months are both captured by the prediction.

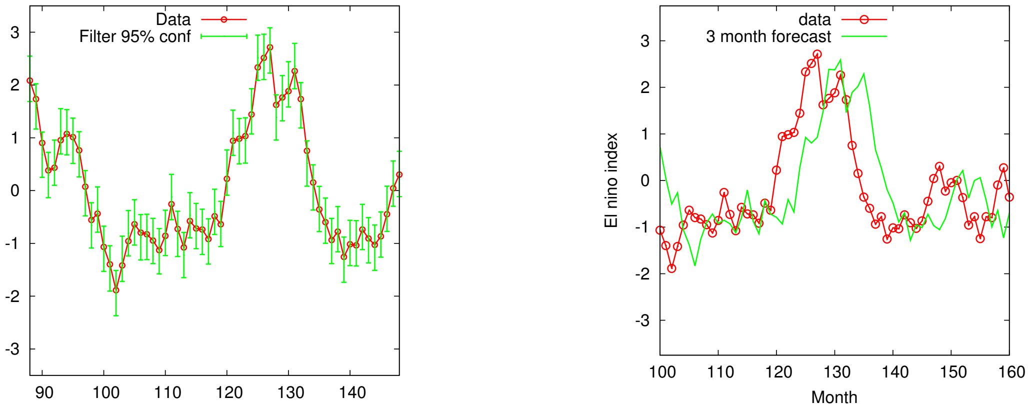

Figure 3: LEFT: estimation of states; RIGHT: forecast. Note here a mismatch due to an indexing error in our preliminary version of the numerical model.

2.2 Empirical Relationships

(2.2.1) We now attempt to use the existing data provided by weather logistics in order to make long term weather forecasts by deriving empirical relationships between large scale predictors and the resulting localised weather. A vast amount of data was available to us spanning many decades and every region of the UK. We chose to restrict out attention to the northwest of England as this was where the study group was being held and we focused on trying to predict the temperature in this region.

2.2.1 Identifying the Key Predictors

(2.2.2) Figure 2.2.1 shows the correlation coefficient between 6 of the large scale predictors and the monthly averaged temperature in the northwest of England over a 40 year period. The first five properties are proposed relevant characteristics of the jet stream. The first, J, is a measure of the maximum wind speed achieved by part of the jet stream along its full length and is given in terms of an index between 1 and 10. The position of the jet stream is averaged over two large regions over the Atlantic ocean, one over the west Atlantic, and the second close to the UK. The position and standard deviation, a measure of the width of the jet, are calculated in both domains and are labelled $P1$, $P2$, σ_1 and σ_2, respectively. The fifth predictor is the sea surface temperature around the UK.

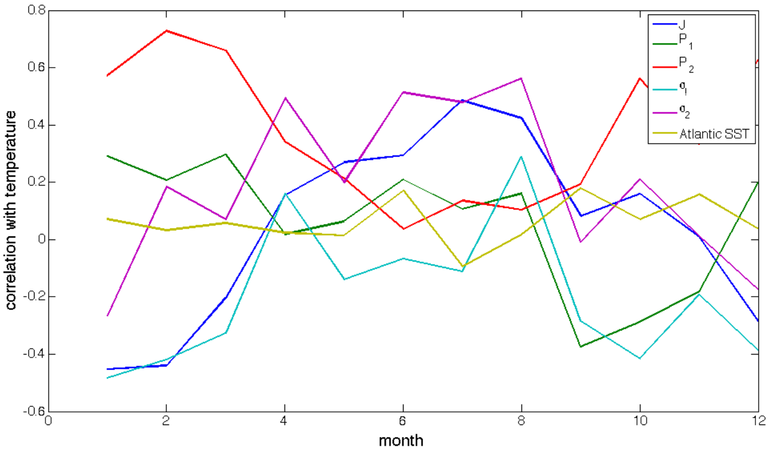

Figure 4: The monthly averaged correlation coefficients between 6 predictors and the temperature in the northwest of England over 40 years.

(2.2.3) The first observation of Figure 2.2.1 is the high monthly variability in some of the predictors, especially σ_2. This is not believed to be a observable property and so we first calculated correlation coefficients over two month periods in an attempt to smooth out the osciallations. We further chose 3 of the predictors to focus on based on the size of the correlation coefficient. The jet stream strength shows good correlation in both the winter and the summer months. The position and standard deviation of the jet close to the UK have good correlation in winter and summer respectively with little to none in the other seasons. The two monthly averaged correlation coefficients for these 3 predictors are shown in Figure 2.2.1 by the solid lines with the one monthly averages shown by the dashed lines for comparison. Taking the two monthly averages has the desired effect of smoothing out the osciallations in the previous figure.

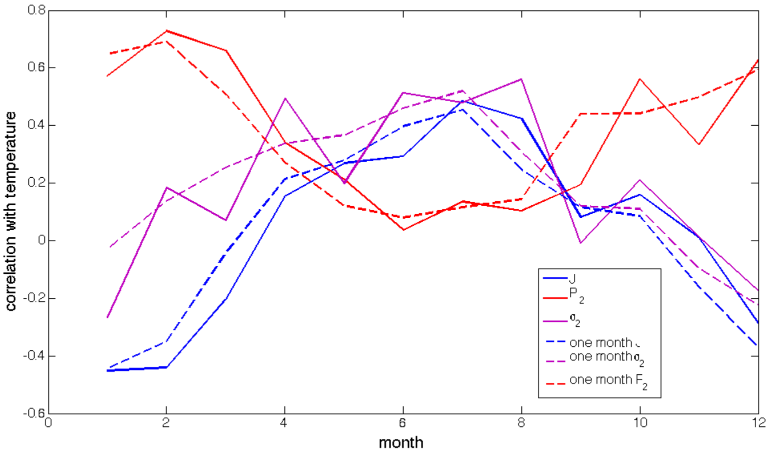

Figure 5: The two monthly averaged correlation coefficients between the chosen 3 predictors and the temperature in the northwest of England over 40 years.

(2.2.4) It is interesting that the position of the jet stream is important in the winter but that the standard deviation becomes much more significant in the summer. To explain this we calculated the maximum and minimum positions of the jet stream in the 40 year period that we had the data for. This is shown in Figure 2.2.1. The figure shows that there is a much larger variation in the latitudinal position of the jet stream in the winter months than in the summer. Regardless of the standard deviation, if the jet stream is positioned at a latitude of 58^o in the winter it is unlike to ever be wide enough to stretch as far south as 40^o. In the summer, however, there is only a 10ø variation in the position and a wide jetstream will lie over a very similar region. If the jet stream is very narrow then this 10^o difference will become much more significant. For this reason, the standard deviation becomes a more important predictor in the summer months. The intercorrelation between the standard deviation and the position in the summer months is not explored but we expect the position will become more important if th standard deviation is small.

2.2.2 Deriving Empirical Relations

(2.2.5) Weather Logistics currently only use a single relationship between the predictors and the weather which is used for all seasons. We propose two temperature predictors, one for the summer and one for the winter,

$$\begin{aligned} T_W &= T_W(J, P_2), \\ T_S &= T_S(J, \sigma_2), \end{aligned}$$

which, for simplicity, depend linearly on the two independent variables

$$\begin{aligned} T_W &= a_1(J - a_2)(a_3 P_2 + a_4), \\ T_S &= b_1(J - b_2)(b_3 \sigma_2 + b_4). \end{aligned}$$

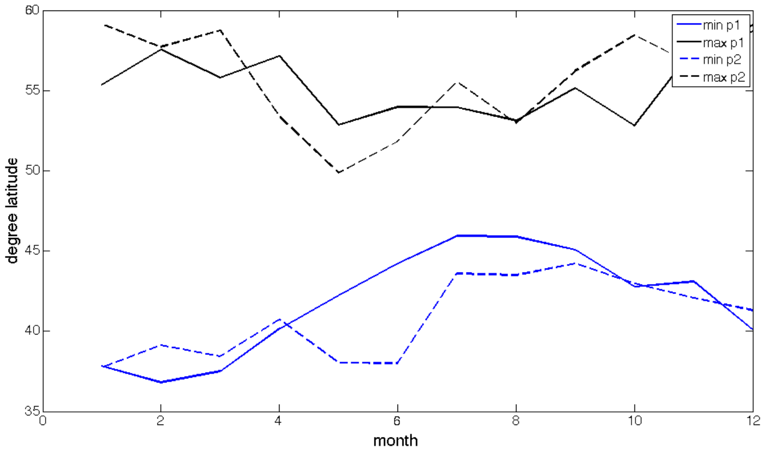

Figure 6: The maximum and minimum position of the jet stream in the two domains over the Atlantic.

The a_i and the b_i are found by minimising the errors between these predictors and the data using the L^2 norm.

$$\min_{a_i} ||T_W - T||_2,$$
$$\min_{b_i} ||T_S - T||_2.$$

Here T is the actual temperature at the same time as the predictors are measured.

(2.2.6) We label the numerical number of the month by N_m starting with 1 for January. An all season predictor is formed by taking a weighted sum

$$T(J, P_2, SD_2) = \alpha(N_m)T_W(J, P_2) + \beta(N_m)T_S(J, SD_2).$$

where

$$\alpha(N_m) = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{12}(N_m - 1.5)\right),$$
$$\beta(N_m) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi}{12}N_m\right).$$

These relations are chosen so that the peak in the cosine corresponds to the peak in correlation, as seen in Figure 2.2.1, of P_2 and σ_2, respectively.

(2.2.7) The two monthly correlation coefficients of the all season predictor are shown in Figure 2.2.2. The correlation of the individual predictors are shown by the dashed lines and the summer, winter and all season predictor are shown by the solid lines. In the summer months the summer predictor which combines the predictive power of the jet stream strength and standard deviation offers a notable improvement on the correlation with temperature than the two individual predictors independently. This figure could be improved on further by adding additional predictors. The winter predictor correlates slightly better than just the position of the jet stream but not by much. This is rather surprising since both the position and strength of the jet stream each correspond very well and we expected a marked improvement when combining them. A possible explanation for this is that there is some correlation between the strength and position

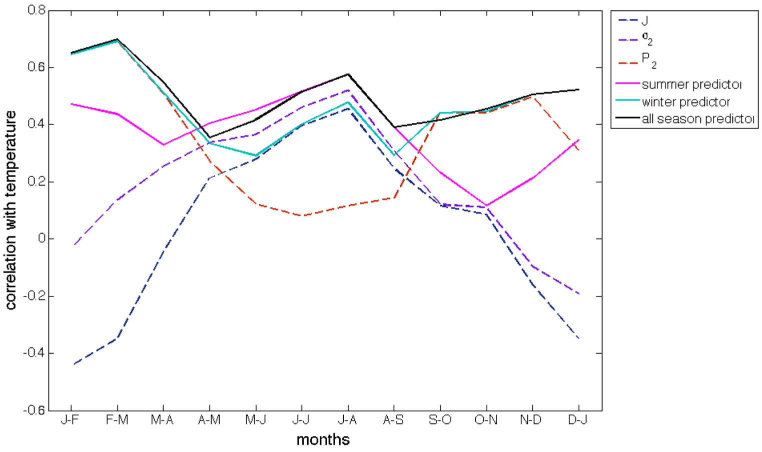

Figure 7: Correlation of the combine predictors.

of the jet stream and although the two both correlate with temperature there is no additional information that is obtained by combining them.

2.2.3 Predictive Capability

(2.2.8) There seems to be a reasonable correlation between our all season predictor and the temperature in the northwest of England. The question remains whether the predicitons of the jet stream strength, position and standard deviation are sufficiently accurate to offer comparable correlation. We had access to predictions of these quantities from a 13 year period and the simulations were started at 3 monthly intervals and run for 3 months at a time.

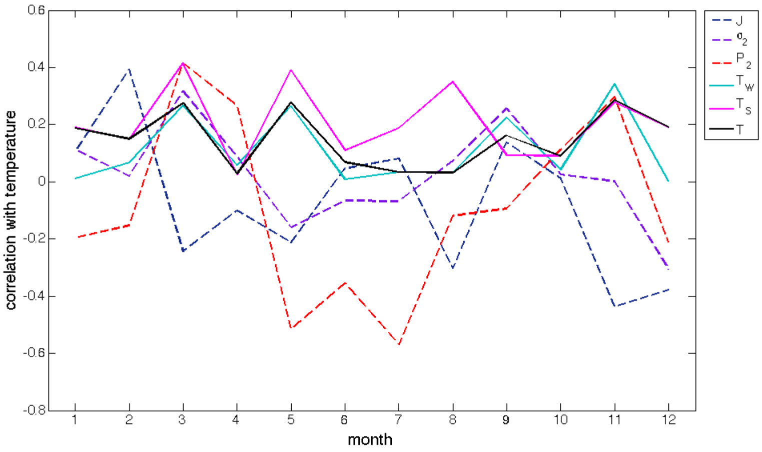

Figure 8: The correlation between the predicted values of the predictors against measured temperature in the northwest of England.

(2.2.9) The correlation between the predicted values of the predictors is shown in Figure 8.

Initally, the graph appears to show little correlation but there does appear to be roughly 3 monthly peaks in the correlation. These are due to the method used to obtain the predictions of the predictors; the simulations are initiated every 3 months and so the intermediate months are one, two and three month forecasts with the one month forecasts corresponding to the peaks in the correlation because of their better predicitve capabilities. The two month and three month forecasts of the predictors are less accurate and consequenty correlate less well with the observed temperatures. Another limitation is that we only had access to 13 years of predictions which is not necessarily sufficient to obtain statistically significant results.

(2.2.10) Despite the afore mentioned limitations, the summer predictor appears to work fairly well in May and August (months 5 and 8 respectively) which roughly correspond to the one month forecasts. The winter predictor works best in March and November but definitely has more success at the end of the year than at the start. The jet stream seems to be most accurately predicted in the winter months but not in the summer. Interestingly, the predicted position seems to correlate better with the temperature in the summer than the actual measurements did.

3 Conclusions

(3.0.11) Applying the Kalman filter to the El Nino index turned out to be incredibly effective and an accurate 3 month forecast was produced which matched to a high degree of accuracy with the measurements. This approach could be applied to further indexes and data sets to improve the overall predctive capabilities of Weather logistic's empirical forecasting model.

(3.0.12) The study into the correlation between different global predictors and the downscale temperature measurements in the North-west of England identified the possibility that different predictors may be necessary at different times of year. The example found was that there is more variability in the latitudinal position of jet stream in the winter months than in the summer and as a result the temperature was less influenced by position and more by the width or standard deviation of the jet stream.

(3.0.13) A further area of work that was identified but not investigated due to time constraints was that there is likely to be a time delay in the influence of some global predictors to the downscale weather. The El Nino is measured in the east Pacific and is unlikely to have immediate consequences on the UK's climate; often it is linked to the winter weather in the UK despite occuring the summer months. The predictors investigated in Section 2.2 correlated measurements with same-time UK temperature and as a result the predictors that were found to have to the most effect were the measurements taken closest to the UK. The position of the jet stream over the west Atlantic was found to have little effect using this method but it could be that there is a 2-4 week delay in its influence. This is suggested as future work along with studying the effect of the Kalmann filter in the earlier sections which was only applied to the El Nino index which was found not to have an immediate effect on the UK temperature.