8 Methods in Short-Term Climate Prediction

1.climo, 2.persist, 3.OCN, 4. Local regress, 5 Non-local regress, 6 Composites, 7 regress at pattern level, 8 NWP, 9 Consolidation, 10 Other method, 11 Methods not used, App I, App II

8.7 Regression on the Pattern Level

Most empirical methods in short-term climate prediction are nowadays based on multiple linear regression 'on the pattern level'. A primitive example is as follows. Suppose we have two data sets, f(s,t) called the Predictor, and g(s,t) called the Predictand. One can perform two stand alone EOF analyses of f and g, and then do the regression between the time series of the leading modes in the predictand and predictor data sets. Klein and Walsh(1984) made an in depth comparison of regression between EOF mode time series on the one hand and regression between the original data at gridpoints on the other - this was in the context of 'specification' (as discussed for instance in Ch 7.3). Using modes is efficient, and cuts down on endless choices, but it may not always help the skill.

For a more general approach we first discuss the time lagged covariance matrix.

8.7.1 The time lagged covariance matrix

When we have two data sets, f(s,t) called Predictor, and g(s,t) called Predictand, one can define the elements of the time lagged covariance matrix C_{fg} as

$$c_{ij} = \sum_{1}^{n_t} f(s_i,t)\, g(s_j,t+\tau)/n_t \qquad (8.5)$$

where n_t is the number of time level, a time mean of f and g was removed, and τ is the time lag. C, non-square in general, thus contains the covariance between the predictor at any place in its domain, and the predictand anywhere in its domain - local as well as non-local. (From the time lag in g our intention is clear: to predict g from f. However, some analyses below (CCA, SVD) do not go beyond establishing associations between f and g, leaving in the middle who predicts

whom. Most texts on SVD and CCA thus do not show a time lag)

Associated with c_{ij} there is also a correlation

$$\rho_{ij} = n_t * c_{ij} / \mathrm{sqrt}(\Sigma f^2(s_i,t) \Sigma g^2(s_j,t+\tau)) \qquad (8.5a)$$

If g and f were the same data set, and the time lag is zero, C would be the square Q, as per Eq (2.14). Along with C_{fg} we also need Q_f and Q_g below. ($Q_f = C_{ff}(\tau=0)$). Given how Q was manipulated to calculate EOF (presented in chapter 5 as 'self prediction') one may surmise that C can be used to relate patterns&time series in the predictor field to patterns&time series in the predictand field. Indeed, C, in its various renditions depending on prefiltering, truncation, orthogonality constraints, organization of input data sets, etc, is among the most studied in short term climate prediction. Instead of the role played by the notion explained variance (EV) in EOFs, the target of calculating coupled patterns/time series is often in 'explaining' the co-variance of f and g. Because covariance can be negative, the target is often taken to be 'squared covariance' (SC). i.e. the fraction of Σc_{ij}^2, where summation is over all i and j, that can be explained by 1, 2 or m coupled 'modes'.

Without any truncation or constraint C is set up to create any imaginable regression between f and g, so as to minimize the rmse of the prediction of g, on the dependent data that is used to compute C. Here lies a very significant problem. With so many predictors f(s,t), it is hard to avoid overfitting[1]. C contains the correlation of everything with everything. The overfit is combated by severe truncation at the pattern level. This reduces the subjective nature of choosing predictors.

Somewhere in C also lie the methods we already discussed before, like local persistence and local and non-local regression. The reason to present these simpler methods separately and upfront is twofold. First we may easily lose local effects when applying truncation at the pattern level, i.e the very high persistence in temperature in San Diego California would not make it into a

[1]Standard texts on regression should be consulted to find methods of exploratory regression that can avoid overfit in most cases.

pattern method until hundreds of modes are admitted. Secondly, C is calculated without any physical intuition. The local effects approach can more easily be defended on physical grounds.

8.7.2 CCA, SVD and EOT2

In chapter 5 we presented EOFs of the data set f(s,t) as:

$$f(s,t) = \sum_{m=1}^{M_f} \alpha_m(t)\, e_m(s) \qquad (8.6)$$

where both the time series and spatial patterns are orthogonal. Eq (8.6) still gives a complete representation of f as long as either the time series or the spatial patterns are orthogonal, and M_f is large enough. Likewise we have for the predictand:

$$g(s,t+\tau) = \sum_{m=1}^{M_g} \beta_m(t+\tau) d_m(s) \qquad (8.6a)$$

Coupling the modes among the two data sets f and g, which have the same number of time levels but possibly different spatial domains (also $M_f \neq M_g$), will be discussed below in terms of the properties of $\alpha_m(t)$ and $\beta_m(t+\tau)$ and $d_m(s)$ and $e_m(s)$ respectively. In any of the methods below orthogonality is maintained in either time or space (not both), so the coupled modes allow projection of future data and/or partial rebuilding of f and g themselves with a set of modes, and the notion explained variance (not optimal obviously) within each data set still applies.

The plain distinguishing feature of Canonical Correlation Analysis (CCA) is that the correlation of $\alpha_m(t)$ and $\beta_m(t+\tau)$, denoted cor(m), is maximized - the modes are ordered such that cor(m)>cor(m+1) for all m. Within each data set we have for CCA

$$\sum_t \alpha_k(t)\ \alpha_m(t) = 0 \qquad \text{for } k \neq m \qquad \text{(CCA-1)}$$

$$\sum_t \beta_k(t+\tau)\ \beta_m(t+\tau) = 0 \qquad \text{for } k \neq m \qquad \text{(CCA-2)}$$

i.e orthogonal time series, and across the data sets:

$$\sum_t \alpha_k(t)\ \beta_m(t+\tau) = 0 \qquad \text{for } k \neq m \qquad \text{(CCA-3)}$$

$$\sum_t \alpha_k(t)\ \beta_m(t+\tau) = \text{cor}(m) \qquad \text{for } k = m \qquad \text{(CCA-3a)}$$

where summation is over time. The cor(m) can be found as the square root of the eigenvalues of the matrix $M = Q_f^{-1} C_{fg} Q_g^{-1} C_{fg}^T$ (or from $Q_g^{-1} C_{fg}^T Q_f^{-1} C_{fg}$). Note that CCA's maps are not orthogonal.

On the other hand, in a method often called singular value decomposition (SVD) the explained SC is maximized. For SVD[2] we have within each data set:

$$\Sigma_s e_k(s)\ e_m(s) = 0 \qquad \text{for } k \neq m \qquad \text{(SVD-1)}$$

$$\Sigma_s d_k(s)\ d_m(s) = 0 \qquad \text{for } k \neq m \qquad \text{(SVD-2)}$$

i.e. orthogonal maps, and across the data sets:

$$\Sigma \alpha_k(t)\ \beta_m(t+\tau) = 0 \qquad \text{for } k \neq m \qquad \text{(SVD-3)}$$

$$\Sigma \alpha_k(t)\ \beta_m(t+\tau) = \sigma(m) \qquad \text{for } k = m \qquad \text{(SVD-3a)}$$

where $\sigma(m)$ is the m'th singular value of C_{fg}. The SC explained by mode m is $\sigma^2(m)$.

Notice the (dis)similarities of SVD and CCA. CCA has orthogonal time series, SVD orthogonal maps. Properties (CCA-1) and (CCA-2) vs (SVD-1) and (SVD-2) appear to be a matter of space-time reversal, but this can not be stated for the 3rd property. The roles of cor(m) and $\sigma(m)$ appear similar. The notion 'SC explained' is sometimes also used for CCA, but does not relate trivially to cor(m). Theoretically it is possible that the first CCA mode describes a perfectly coupled f-g process of infinitesimal amplitude (high cor, low SC).

CCA and SVD are methods to find coupled modes, but they are not quite forecast methods. A regression between the $\alpha_m(t)$ and $\beta_m(t+\tau)$ is needed to forecast $\beta_m(t+\tau)$ given $\alpha_m(t)$.

An easy way of explaining both the idea and the actual application of methods like CCA and SVD to a forecast situation may be to use 'EOT2' - we used EOT in Chapters 4 and 5, but extend it here to 2 data sets. Specifically, we seek the position s_1 in space so that the time series $f(s_1,t)$ explains the most of the variance in the predictand data set g(s,t) at lag τ. I.e. we find i for which U(i) defined as

---

[2] We use the name SVD, even though we agree with Zwiers and Von Storch (1999) that it is unfortunate that the name of the method is confused with a basic matrix operation; they suggest Maximum Covariance Analysis.

$$U(i) = \Sigma_j \left(\rho^2_{ij} * \Sigma_t \, g^2(s_j, t+\tau) / n_t \right) \qquad (8.7)$$

is maximum. Having found s_1 we take $f(s_1,t)$ to be the first mode's time series of *both* f and g expansions, i.e. $f^{explained}(s, t) = a(s_1, s)\, f(s_1, t)$, and $g^{explained}(s, t+\tau) = b(s_1, s)\, f(s_1, t)$, where $a(s_1, s)$ is the regression coefficient to predict $f(s,t)$ from $f(s_1, t)$ and $b(s_1, s)$ is the regression coefficient to predict $g(s,t+\tau)$ from $f(s_1, t)$. The spatial patterns in (8.6) are thus: $e_1(s)= a(s_1, s)$ and $d_1(s)= b(s_1, s)$. Note that $b(s_1, s)$ is proportional to the correlation defined in Eq(8.5) and used in Eq (8.7). We then seek the 2nd point in the once reduced data sets $f^{reduced}(s, t) = f(s, t) - a(s_1, s)\, f(s_1, t)$, and $g^{reduced}(s, t+\tau) = g(s, t+\tau) - b(s_1, s)\, f(s_1, t)$, to find s_2 etc. This procedure has many of the properties of CCA, specifically the identities (CCA-1), (CCA-2) and (CCA-3/3a), the latter with cor(m)=1 for all modes. (Oddly, EOT2 actually 'beats' CCA on producing the highest correlation between the time series.) EOT2 has at least two notions of relevance, the EV in data set f, and the EV in data set g. The latter is what is maximized, albeit under the constraint that we use a single time series of f at one point in space (rather than linear combinations of f at various points). There does not appear to be a particular need for the explained SC, after all the target of the prediction is EV in g.

Making a forecast of g is easy. For the first mode we need the observation of f at s_1, then multiply by $b(s_1, s)$. Subsequent modes are similar, but f has to be m-1 times reduced for the mth mode.

The reader will not be surprised that there is an 'alternative' lagged covariance matrix given by

$$c^a_{ij} = \Sigma\, f(s, t_i)\, g(s, t_j+\tau) / n_s \qquad (8.5b)$$

where summation is in space. Here we consider inner products of maps of fields f and g at times t_i and $t_j+\tau$. At first sight this definition is possible only if the domain and gridpoints for f and g are the same. However, this discrepancy is resolved by first executing EOFs on f and g individually and thinking of s in (8.5b) as the mode number. We now pick the one f map at time t_i which

maximizes the variance explained in g, an expression similar to (8.7) but reversing the roles of time and space. This single map then acts as $e_1(s)$ for f and $d_1(s)$ for g. There are two time series, which are regression coefficients $a(t_1, t_i)$ to predict $f(s, t_i)$ from $f(s, t_1)$ and $b(t_1, t_i)$ to predict $g(s, t_i+\tau)$ from $f(s, t_1)$. This alternative EOT2 route leads to the expansion (8.6) and (8.6a) with the properties (SVD-1) and (SVD-2) but not (SVD-3). The alternative EOT2 has again two notions of relevance, the EV in data set f, and the EV in data set g. The latter is not only what is maximized[3], but is the purpose of the regression.

The two EOT versions that closely bracket CCA (regular EOT2) and SVD (alternative EOT2) come with either 2 maps and one time series (nearest CCA) or one map and two time series (nearest SVD). From this it appears that SVD is subject to more orthogonality constraints than CCA - after all (CCA-3) follows trivially when there is only one time series to begin with, but (SVD-3) does not follow automatically from having a single map (d=e).

Note that when admitting too many modes CCA/SVD goes in the direction of multiple linear regression. Obviously, truncation is necessary for reaping the benefits of regression at the pattern level.

Much information about SVD and CCA can be found in Bretherton et al(1992), Newman and Sareshmukh(1997) and Zwiers and von Storch(1999). Wilks(1995) provides a good discussion of CCA.

CCA was not used much in meteorology until Barnett and Preisendorfer(1987). The main methodological twist in their paper is a prefiltering step where both f and g are truncated to just a few EOFs before calculating C. (Moreover, the EOF associated time series are standardized, as in a version of the Mahalanobis norm (Stephenson 1997)) The prefiltering greatly reduces CCA's susceptibility to noise. The prefiltering also makes the practical difference between SVD and CCA in many instances very small. Additionally Barnett and Preisendorfer(1987) applied their adjusted

[3] under the constraint that we use maps of f at one point in time (rather than linear combinations at various times).

CCA to the seasonal forecast and had the predictor data set cover four antecedent seasons. This method and this particular predictor lay-out has been popularized by Barnston(1994) and his work found short-term climate prediction application on nearly all continents (Johansson et al (1998) for Europe; Thiaw et al(1999) for Africa; Hwong et al(2001) for Korea, Shabbar and Barnston(1996) for Canada, He and Barnston(1996) for tropical Pacific Islands and Barnston and Smith(1996) for the whole globe). While SVD is often mentioned in one breath with CCA, and widely used in research (Waliser et al 1999; Wu and Dickinson 2005) there appear to be far fewer real-time forecast applications based on SVD. CCA is also applied as a method to correct errors in GCM predictions (Smith and Livezey 1999; Tippett et al, 2005)

As a diagnostic tool SVD or CCA may be as difficult to use as EOF, i.e. the patterns in the predictor and predictand data set may or may not be revealing the underlying physics. Plenty of examples of patterns are found in Barnston(1994). Newman and Sardeshmukh (1997) show the failure (to a certain extent) of SVD to discover that vorticity and streamfunction are linear transforms of each other. Zwiers and Von Storch(1999) also provide several examples.

We spent some paragraphs explaining SVD, CCA etc because so much of the modern empirical work is along these lines. Regression on the pattern level is thought to take away the arbitrariness of correlating everything with everything. Although methodological details are hotly debated sometimes, the other choices may be more important than the exact method. For instance, which predictors, how far back in time, how many time levels, the domain for predictors and predictands, pre-filtering, truncation etc, may be more important than the exact CCA vs SVD method. The CCA at CPC and CDC, identically the same method, often give conflicting tropical Pacific SST forecasts. While we presented the above material as a strictly separated predictor f and predictand g, keep in mind that the data sets may be combined, i.e. fields of the predictand at an earlier time may be appended to f in order to forecast g. CCA has been used at both CPC and CDC for real time seasonal prediction; skill levels are at best (short lead JFM seasonal T&P) 0.3 - 0.35 correlation nationwide with regional variations that reflect the large impact of ENSO

(Barnston 1994; Quan et al 2005). The CCA modes suggest lesser influences from other tropical areas and mid-latitude oceans as well.

8.7.3 LIM, POP and Markov

Somewhat similar to CCA and SVD are the linear inverse model (LIM) and principal oscillation patterns(POP). The similarity is in the central role of the lagged covariance matrix as in (8.5), evaluated from data. However, both POP and LIM try to generalize the results for lag τ to all other lags by assuming an underlying theory. Following the Winkler et al(2001) notation one may assume a linear model given by

$$d\,x/dt = L\,x + R \qquad (8.8)$$

where x is the retained scales state vector, L is a linear operator and R is random forcing due to unresolved scales (possibly with structure in space). Vector x would for instance be a combination of data sets f and g. The solution to (8.8) is

$$x(t+\tau) = \exp(L\,\tau)\,x(t) + R', \quad (8.9)$$

where R' depends on the history of R. The operator L can be determined from data at a chosen lag τ_0, i.e. we evaluate C for lag τ_0. L is given by $C(\tau_0)\,C^{-1}(\tau=0)$, see Winkler et al (2001) for detail. The forecast for any lag τ is given by the first term in (8.9). The forecasts for τ_0 would, everything else being the same, be close to CCA's. But an analytical flavor is added because time evolution is implied. Moreover, it is possible to calculate the eigenvectors of the asymmetric L once and for all - they are structures evolving in time, and ultimately damped. By knowing the projection of the current initial state onto the known eigenvectors of L, the forecast can be made analytically and can be interrogated for diagnostic purposes, such as in deriving the optimal structure to produce an El Nino pattern 10 months later (Penland and Sardeshmukh 1995). This is similar to what we presented for CA (section 7.6), although CA has additional growth due to unstable normal modes.

Several examples of POP, including for MJO forecasts, are given in Zwiers and von

Storch(1999). Winkler et al's (2001) application is in the week2 forecast, while Penland pioneered LIM for seasonal SST forecasts, both in the Pacific (Penland and Magorian 1993) and Atlantic (Penland and Matrosova 1998). In all cases C is calculated from EOF truncated data, but the degree of truncation varies wildly.

A straightforward method has been presented in Xue et al(2000). In this paper the discretized version of (8.8) is used: $x(t+\tau) = C(\tau)\, C^{-1}(\tau=0)\, x(t)$, i.e. given an initial state x(t) and $C(\tau)\, C^{-1}(\tau=0)$ as determined from data, the forecast for lead τ can be made. No linear model is assumed, so the calculation has to be done for each τ seperately, and nothing connects the forecasts at two different τ, except to the extent the data suggest. No modes are calculated, neither eigenmodes of L (as in LIM/POP, see Eq 8.9), or M (CCA) nor singular vectors of C (as in SVD). This cuts down on interpretation. The problem is handled as multiple linear regression, however after extremely heavy truncation using extended EOF in the input data. Xue et al(2000) use sea-level height, wind stress and SST to forecast the same (sea-level height, wind stress and SST) in the tropical Pacific which appears to be a wise choice, since the methods has worked well in real time. They call their method a 'Markov' (MRK) method. CCA, SVD, and LIM, POP and MRK have options in truncation both in preparing the input data, and in truncating the modes calculated from C, L or M.