5. Empirical Orthogonal Functions

The purpose of this chapter is to discuss Empirical Orthogonal Functions (EOF), both in method and application. When dealing with teleconnections in the previous chapter we came very close to EOF, so it will be a natural extension of that theme. However, EOF opens the way to an alternative point of view about space-time relationships, especially correlation across distant times as in analogues[1]. EOFs have been treated in book size texts, most recently in Jolliffe (2002), a principal older reference being Preisendorfer(1988). The subject is extremely interdisciplinary, and each field has its own nomenclature, habits and notation. Jolliffe's book is probably the best attempt to unify various fields. The term EOF appeared first in meteorology in Lorenz(1956). Zwiers and von Storch(1999) and Wilks(1995) devote lengthy single chapters to the topic.

Here we will only briefly treat EOF or PCA (Principal Component Analysis) as it is called in most fields. Specifically we discuss how to set up the covariance matrix, how to calculate the EOF, what are their properties, advantages, disadvantages etc. We will do this in both space-time set-ups already alluded to in Eqs (2.14) and (2.14a). There are no concrete rules as to how one constructs the covariance matrix. Hence there are in the literature matrices based on correlation, based on covariance etc. Here we follow the conventions laid out in Chapter 2. The postprocessing and display conventions of EOFs can also be quite confusing. Examples will be shown, for both daily and seasonal mean data, for both the Northern and Southern Hemisphere. EOF may or may not look like teleconnections. Therefore, as a diagnostic tool, EOFs may not always allow the interpretation some would wish. This has led to many proposed 'simplifications' of the EOFs, which hopefully are more like teleconnections.

However, regardless of physical interpretation, since EOFs are maximally efficient in retaining as much of the data set's information as possible for as few degrees of freedom as possible they are ideally suited for empirical modeling. Indeed EOFs are an extremely popular tool

[1] A pair of analogues are two states in a geophysical system, widely separated in time, that are very close.

these days. A count in a recent issue of the Journal of Climate shows at least half the articles using EOFs in some fashion. Moreover, EOFs have some unique mind-twisting properties, including bi-orthogonality. The reader may not be prepared for bi-orthogonality, given the name empirical orthogonal functions and even Jolliffe's working definition given below.

5.1 Methods and definitions

5.1.1 Working definition:

Here we cite Jolliffe (2002, p 1). "The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principle components, which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in all of the original variables." The italics are Jolliffe's. PCA and EOF analysis is the same.

5.1.2 The Covariance Matrix

One might say we traditionally looked upon a data set f(s,t) as a collection of time series of length n_t at each of n_s gridpoints. In Chapter 2 we described that after taking out a suitable mean { } from a data set f(s,t), usually the space dependent time-mean (or 'climatology'), the covariance matrix Q can be formed with elements as given by (2.14):

$$q_{ij} = \sum_t f(s_i, t)\, f(s_j, t) / n_t$$

where s_i and s_j are the ith and jth point (gridpoint or station) in space. The matrix Q is square, has dimension n_s, is symmetric and consists of real numbers. The average of all q_{ii} (the main diagonal) equals the space time variance (STV) as given in (2.16). The elements of Q have great appeal to a meteorological audience. Fig.4.1 featured two columns of the correlation version of Q in map form, the NAO and PNA spatial patterns, while Namias(1981) published all columns of Q (for seasonal mean 700 mb height data) in map form in an atlas.

5.1.3. The Alternative Covariance Matrix

One might say with equal justification that we look upon f(s,t) alternatively as a collection of n_t maps of size n_s. The alternative covariance matrix Q^a contains the covariance in space between two times t_i and t_j given as in (2.14a):

$$q^a_{ij} = \sum_s f(s, t_i)\, f(s, t_j) / n_s$$

where the superscript a stands for alternative. Q^a is square, symmetric and consists of real numbers, but the dimension is n_t by n_t, which frequently is much less than n_s by n_s, the dimension of Q. As long as the same reference {f} is removed from f(s,t) the average of the q^a_{ii} over all i, i.e. the average of main diagonal elements of Q^a, equals the space-time variance given in (2.16). The average of the main diagonal elements of Q^a and Q are thus the same.
The elements of Q^a have apparently less appeal than those of Q (seen as PNA and NAO in Fig.4.1). It is only in such contexts as in 'analogues', see Chapter 7, that the elements of Q^a have a clear interpretation. The q^a_{ij} describe how (dis)similar two maps at times t_i and t_j are.

When we talk throughout this text about reversing the role of time and space we mean using Q^a instead of Q. The use of Q is more standard for explanatory purposes in most textbooks, while the use of Q^a is more implicit, or altogether invisible. For understanding it is important to see the EOF process both ways.

5.1.4 The covariance matrix: context

The covariance matrix typically occurs in a multiple linear regression problem where f(t, s) are the ***predictors***, and any dummy ***predictand*** y (t), $1 <= t <= n_t$, will do. Here we first follow Wilks(1995; p368-369). A 'forecast' of y (denoted as y*) is sought as follows:

$$y^*(t) = \sum_s f(s, t)\, b(s) + \text{constant}, \qquad (5.1)$$

where b(s) is the set of weights to be determined. As long as the time mean of f was removed, the constant is zero.

The residual $U = \sum_t \{ y(t) - y^*(t) \}^2$ needs to be minimized.

$\partial U / \partial b(s) = 0$ leads to the "normal" equations, see Eq 9.23 in Wilks(1995), given by:

$$Q \;\; b = a,$$

where Q is the covariance matrix and a and b are vectors. The elements of vector a consist of $\sum_t f(t, s_i)\, y(t) / n_t$. Since Q and a are known, b can be solved for, in principle.

Note that Q is the same for any y. (Hence y is a 'dummy'.)

The above can be repeated alternatively for a dummy y(s)

$$y^*(s) = \sum f(s, t)\, b(t) \qquad (5.1a)$$

where the elements of b are a function of time. This leads straightforwardly to matrix Q^a. (5.1a) will be the formal approach to constructed analogue, see chapter 7.

Q and Q^a occur in a wide range of linear prediction problems and Q and Q^a depend only on f(s,t), here designated as the predictor data set.

In the context of linear regression it is an advantage to have orthogonal predictors, because one can add one predictor after another and add information (variance) without overlap, i.e. new information not accounted for by other predictors. In such cases there is no need for backward/forward regression and one can reduce the total number of predictors in some rational way. We are thus interested in diagonalized versions of Q and Q^a (and the linear transforms of f(s,t) underlying the diagonalized Qs).

5.1.5 EOF through eigen-analysis

In general a set of observed f(s,t) are not orthogonal, i.e. $\sum f(s_i, t) f(s_j, t) / n_t$ and $\sum f(s, t_i) f(s, t_j) / n_s$ are not zero for $i \neq j$. Put another way: neither Q nor Q^a are diagonal. Here some basic linear algebra can be called upon to diagonalize these matrices and transform the f (s,t) to become a set of uncorrelated or orthogonal predictors. For a square, symmetric and real matrix, like Q or Q^a, this can be done easily, an important property of such matrices being that all eigenvalues are positive and the eigenvectors are orthogonal. The classical eigenproblem for matrix B can be stated:

$$B e_m = \lambda_m e_m \quad (5.2)$$

where e is the eigenvector and λ is the eigenvalue, and for this discussion B is either Q or Q^a. The index m indicates there is a set of eigenvalues and vectors. Notice the non-uniqueness of (5.2) - any multiplication of e_m by a positive or negative constant still satisfies (5.2). Often it will be convenient to assume that the norm $|e|$ is 1 for each m.

Any symmetric real matrix B has these properties:

1) The e_m's are orthogonal

2) $E^{-1} B E$, where matrix E contains all e_m, results in a matrix Λ with the elements λ_m at the main diagonal, and all other elements zero. This is one obvious recipe to diagonalize B (but not the only recipe!).

3) all $\lambda_m > 0$, m=1, ...M.

Because of property 1 the $e_m(s)$ are a basis, orthogonal in space, which can be used to express:

$$f(s, t) = \Sigma \, \alpha_m(t) e_m(s) \quad (5.3)$$

where the $\alpha_m(t)$ are calculated, or thought of, as projection coefficients, see Eq (2.6). But the $\alpha_m(t)$ are orthogonal by virtue of property #2. It is actually only the 2nd step/property that is needed to construct orthogonal predictors. (In the case of Q, execution of step 2 implies that the time series $\alpha_m(t)$ (linear combinations of original f (s,t)) become orthogonal and $E^{-1} Q E$ diagonal.) Here we thus have the very remarkable property of bi-orthogonality of EOFs - both $\alpha_m(t)$ and $e_m(s)$ are an orthogonal set. With justification the $\alpha_m(t)$ can be looked upon as basis functions also, and (5.3) is satisfied when the $e_m(s)$ are calculated by projecting the data onto $\alpha_m(t)$.

We can diagonalize Q^a in the same way, by calculating its eigenvectors. Now the transformed maps (linear combinations of original maps) become orthogonal due to step 2, and the transformed time series are a basis because of property 1. (Notation may be a bit confusing here, since, except for constants, the e's will be α's and vice versa, when using Q^a instead of Q.) One may write:

$$Q\, e_m = \lambda_m\, e_m \quad (5.2)$$

$$Q^a\, e_m{}^a = \lambda_m{}^a\, e_m{}^a \quad (5.2a)$$

such that the e's are calculated as eigenvectors of Q^a or Q. We then have

$$f(s, t) = \Sigma\, \alpha_m(t) e_m(s) \qquad (5.3)$$

$$f(s, t) = \Sigma\, e_m{}^a(t) \beta_m(s) \qquad (5.3a)$$

where the α's and β's are obtained by projection, and the e's are obtained as eignevector. When ordered by EV, $\lambda_m = \lambda_m{}^a$, and except for multiplicative constants $\beta_m(s)=e_m(s)$ and $\alpha_m(t)=e_m{}^a(t)$, so (5.3) alone suffices for EOF.

Note that $\alpha_m(t)$ and $e_m(s)$ cannot both be normed at the same time while satisfying (5.3). This causes considerably confusion. In fact all one can reasonably expect is:

$$f(s, t) = \Sigma\, \alpha_m(t)/c \;\; e_m(s)*c \qquad (5.3b)$$

where c is a constant (positive or negative). (5.3b) is consistent with both (5.3) and (5.3a). Neither the polarity, nor the norm is settled in an EOF procedure. The only unique parameter is λ_m.

Since there is only one set of bi-orthogonal functions, it follows that during the above procedure Q and Q^a are simultaneously diagonalized, one explicitly, the other implicitly for free. It is thus advantageous in terms of computing time to choose the covariance matrix with the smallest dimension. Often, in meteorology $n_t << n_s$. Savings in computer time can be enormous.

5.1.6 Explained variance EV

The eigenvalues can be ordered: $\lambda_1 > \lambda_2 > \lambda_3 ..... > \lambda_M > 0$. Moreover:

$$\sum_{m=1}^{M} \lambda_m = \sum_{i=1}^{n_s} q_{ii} / n_s = STV$$

The λ_m are thus a spectrum, descending by construction, and the sum of the eigenvalues equals the space time variance.

Likewise

$$\sum_{m=1}^{M} \lambda_m = \sum_{k=1}^{M} q^a_{kk} / n_t = STV$$

The eigenvalues for Q and Q^a are the same. The total number of eigenvalues, M, is thus at most the smaller of n_s and n_t

In the context of Q one can also write: explained variance of mode m (λ_m) = $\Sigma\ \alpha^2_m(t) / n_t$ as long as |e|=1. Jargon: mode m 'explains' λ_m of STV or $\lambda_m / \Sigma\ \lambda_m$ *100. % EV

The notion EV requires reflection. In normal regression one explains variance of the predictand y. But here, in the EOF context, we appear to explain variance in f (s,t) the predictor. Indeed EOF is like self-prediction. (5.1) is still true if the dummy y(t) is actually taken to be the time series of f(s,t) at the m-th point: f(s_m ,t). (5.1) then reads:

$$f^*(s_m, t) = \sum_s f(s, t)\, b(s, s_m) \qquad (5.1)$$

where b (s, s_m), for a fixed point s_m, is the regression coefficient between s_m and any s ($1 <= s <= n_s$) and the EOFs are found by minimizing U

$$U = \sum_{m=1}^{n_s} \sum_t \left(f(s_m, t) - \sum_s f(s, t)\, b(s, s_m) \right)^2 \qquad (5.4)$$

Now f (s, t) is both predictor and predictand. At this point the order m=1, 2 etc is arbitrary. As per (5.2) and executing E^{-1} B E the expression

Σ f (s, t) b (s, s_m) is transformed to $\Sigma\ \alpha_m(t) e_m(s)$

s

For any truncation N ($1<=N<M<=n_s$) the N functions retained are maximally efficient in EV because they minimize U.

5.2 Examples

Fig.5.1 gives an example of an EOF calculation. Shown are the first four EOF following explicit diagonalization of Q by step 2, i.e. solving (5.2) for B=Q and ordering by EV. The maps show $e_m(s)$ for m=1, 4; the time series underneath each map are $\alpha_m(t)$, m=1,4 as per (5.3), where t=1948 thru 2005, 58 values at annual increment. The example corresponds exactly to the seasonal mean JFM 500 mb data on the 20N-pole domain used already in chapter 4. The time mean removed is for 1971-2000, the current WMO climate normal. The time mean of each time series has thus zero mean over 1971-2000. The maps $e_m(s)$ are normed, see Appendix I/inset I for details, and the physical units are in the time series (gpm). One can thus see the amplitude of the time series going down with increasing m and decreasing EV. The EV of the first four EOFs is 23.0, 18.5, 9.3 and 8.1% respectively, for a total of nearly 60%. Indeed the first few modes explain a lot of the variance, as is the general idea of 'principal components' analysis. EOF mode 1 and 2 still look like NAO and PNA, but explain somewhat more variance than the EOT counterparts in Fig.4.4. Moreover we did not have to identify some gridpoint as basepoint a-priori as we had to for Teleconnections and EOT[2]. While modes 3 and 4 explain much less variance than modes 1 and 2, they do have the attractive looks of dispersion on a sphere. The patterns shown are all orthogonal to each other (and to all 54 patterns not shown as well), but often in a complicated way (not like sin/cos). Likewise, the time series have zero inner product, but in a complicated way. For instance, the first two time series appear correlated by eye in the low frequencies, but this is compensated (and much harder to see for the human eye) by negative

---

[2]However, note that we mention a base point (seed=65N,50W) on top of each map. This is related to the initial guess used to start an iteration towards the EOFs, see appendix/inset 2. The initial guess is irrelevant, except for the polarity and the speed of convergence to EOF.

correlation on the inter-annual time-scale.

Fig 5.2 and 5.3 are EOFs for *daily* 500mb data, for NH and SH respectively during 1998-2002 (DJF or 450 days). We use only 5 years here, but still have many more realizations than for seasonal means (58). The contrast between the two hemispheres could hardly be larger. In the NH, standing wave patterns looking like sweeping combinations of NAO and PNA dominate even daily data albeit at much reduced EV compared to seasonal means. In the SH EOFs on daily data suggest domination of surprisingly simple harmonic-like wave motion in the west to east direction. This is certainly consistent with the relative success of EWP forecasts in the SH described in Chapter 3. It is amazing that so many variations of what looks basically like zonal wavenumber four can be spatially orthogonal. (For exactly four waves along 50S there would be only two orthogonal arrangements). The time series in Fig 5.2 and 5.3 are for 450 daily points with zero mean and 4 discontinuities (at the end of Feb in '98, '99, 00 and '01). Even the daily time series show some low-frequency behavior with periods of 10-20 (or more) days of one polarity. Often the leading EOFs are thought of as displaying the large scale low-frequency behavior, and most often they have been calculated from time-filtered data (e.g. monthly means) to force this to be the case. But this may not always be necessary - after all EOFs target the variance and given that the atmosphere has a red spectrum both in time and space it should be no surprise that low-frequency large-scale components show up first.

In Figs 5.2 and 5.3 the anomalies were formed first by subtracting an harmonically smoothed daily Z500 climatology based on 1979-1995 (Schemm et al 1997). This leaves a considerable non-zero mean anomaly for a time series as short as 15 months (D, J or F) during 1998-2002. As a 2nd step we subtracted the time mean across the 15 months to arrive at Figs 5.2 and 5.3. Without removing the time mean the 1st EOF in the SH is a (nearly) zonally invariant pattern with a nodal line at 55S and a time series that is positive for most of the 5 years - such 'annular' variations are well established for the SH (Hartman 1995) and are known as the Southern Annular Mode or Antarctic Oscillation (AAO). Apparently the SH had stronger than average westerly mid-tropospheric flow in a band centered at 55S during 1998-2002.[3]

One should note that the EOFs (Fig. 5.1) and the EOTs presented in Fig. 4.4 are very similar. In this case EOF analysis does tell us something about teleconnections. The NAO and PNA renditions in Fig.4.4 are naturally orthogonal in space and very dominant in EV, so adding the constraint of spatial orthogonality will not change the results too much for these two modes. Usually the similarity between EOF and EOT is much less, and this could make interpretation of EOF (if one wants to see teleconnections) difficult. For instance in Fig.5.2 the leading EOFs are not readily seen as pure PNA and NAO, even though the EOT counterparts (not shown) would.

5.3 Simplification of EOF - EOT

Jolliffe (2002, p) gives a list of methods to simplify EOFs, as practiced in various fields. The methods include 'rotation', 'regionalization', and EOT. Why the need for simplification? Often the blame is placed on the bi-orthogonality, which is allegedly too constraining or too mathematical to allow physical interpretation in many cases. In fact, this means that the way EOFs are calculated, as per (5.2), is overkill. There is no need to calculate the eigenvalues of Q or Q^a if the only purpose is to create orthogonal predictors and to diagonalize either Q or Q^a (not both at the same time). All simplification of EOF methods appears to come down to relaxing orthogonality in one dimension (time or space) while maintaining orthogonality in the other. The EOT method, as described in chapter 4, has orthogonal time series, thus diagonalizes Q, but the elements appearing on the diagonal are not the eigenvalues of Q. Notice that if the outer summation m=1, n_s was not applied, a trivial solution presents itself for Eq (5.4): b $(s, s_m) = 1$ for $s = s_m$ and b=0 for all other s. This procedure yields the EOT's described in Chapter 4. The time series of the EOT attached to gridpoint s_m is $f(s_m, t)$ and thus trivially explains all variance at s_m. However, $f(s_m, t)$ explains far more variance of the whole field (STV) than it would if it just

[3] The name AAO was invented later as the counterpart for the AO (Thompson and Wallace 1998), which has, however, since been renamed Northern Annular Mode.

explained the variance at one point. This, of course, is because of the non-zero correlation between point m and most other points. The EOTs are not unique, one can start with any gridpoint and proceed with any choice for a next gridpoint[4]. Ordering m by EV makes sense and drives the EOT closer to EOF. Upon ordering, the matrix Q has been transformed to a diagonal matrix M (and thus orthogonal time series) with elements μ_m, where μ_m decreases with m. Although μ_m are not the eigenvalues of Q we still have $\Sigma\ \mu_m$ = STV. No spatial basis exists. So even though an equation like (5.3) is still valid, it is satisfied without the e's being orthogonal and the α's are not projection coefficient (i.e. should not and cannot be calculated by projection data onto the e's).

In the alternative case we have

$$U = \sum_{m=1}^{n_t} \sum_{s} \left(f(s, t_m) - \sum_{t} f(s, t)\, b(t, t_m) \right)^2 \qquad (5.4a)$$

In the same way as described before (5.4a) diagonalizes Q^a to form M^a (and thus orthogonal maps, which are linear combinations of the original maps), and $\Sigma\ \mu_m^a$ = STV. The time series are not orthogonal. On a mode by mode basis μ_m^a does not equal μ_m (in general). Although an equation like (5.3) is valid, the α's are not orthogonal.

The alternative EOTs for JFM Z500 mean over 20-90N are shown in Fig 5.4. The alternative (or 'reverse') EOTs start with the observed field in 1989. At 18.3 % EV this observed field explains more of the variance of the 58 fields combined than any other. After regressing 1989 out of the data set the 'once reduced' observed field in 1955 emerges as the next leading EOT, etc. The normalization described in Appendix 1 was applied for display purposes, so we can compare Fig. 5.4 to Figs 4.4 and 5.1 in that the spatial patterns have been forcibly normed. While the higher order modes all have a year assigned to them (like seed=1974), they look less and less

[4]EOT is quite free, while EOF is completely constrained by bi-orthogonality. One can even chose time series from outside the domain of the data set of from another data set (as in Ch 8.7) and start EOT analysis that way.

like the observations in that year (because modes 1 to m-1 were regressed out). The alternative calculation has 50% EV for 4 modes and the results do not look much like Fig 4.4, but are sweeping large scale fields nevertheless. The time series show a high positive value for the seed year. The leading alternative EOTs are 'real' in the sense that they were actually observed at some point in time, a statement that cannot be made about the PNA, NAO (Fig. 4.4) or EOFs. The utility of reverse EOT should become clearer in the chapter on (constructed) analogues.

Not only can regular EOT be calculated in any order for gridpoint m, one can also manipulate the results. If someone does not like a portion of EOT#1, for instance the NAO related covariance over east Asian in Fig.4.4, one can modify EOT#1 by blanking out this area (forcing zeros). This would be an example of surgical regionalization. After this start one can continue to find the 2nd EOT, possibly blank out more areas etc. This is the basis for Smith and Reynolds' (2003; 2004) Reanalysis for Sea-Level Pressure and SST for century long periods. Smith and Reynolds also found by example that EOTs, for their SST/SLP data sets, are 'nearly the same' as rotated EOF as arrived at by a popular recipe called 'Kaiser varimax'. The principle objective of rotation of EOFs is to obtain more regionalized 'simple' structures that are also more robust to sampling uncertainty as mimicked by leaving one or a few years in or out of the data set. The maps in Fig.4.4 change little upon adding one year, while the maps in Fig.5.1, (certainly the modes 3 and 4) can change beyond recognition. Much work on rotation methods was done by Richman(1986), see also Richman and Lamb(1985). There are two types of rotation, leaving either the temporal or the spatial orthogonality intact (and sacrificing the other). There is an associated loss in EV for the leading modes, making the components less principal. The rotation performed commonly on atmospheric/oceanographic data is essentially to 'rotate' from Fig.5.1 to roughly Fig.4.4, although not precisely so, relaxing orthogonality in space. The precise results of this rotation may be dependent upon the specified number of modes retained in the rotation (O'Lenic and Livezey 1988) and the rotation recipe. Since Fig.4.4 was obtained without setting the number of modes (in rotation), and even without a rotation recipe (like varimax) EOT seem an easy, natural and quick way to achieve many of the stated goals of rotation. Regular EOTs are more robust, simpler and more regionalized than EOFs. The alternative EOTs ‘regionalize’ in the time domain, i.e. try to maximize projection on a flow that happened at a certain time.

Appendix II shows how EOFs can be calculated one by one by an iteration procedure. A good starting point for iteration are EOTs because they are already close to EOFs. Fig. 5.5. shows the EOFs one obtains by starting from the alternative EOTs (and normalization as per Appendix I). We obviously obtain the same EOFs as already shown in Fig.5.1, but we show them for a number of reasons: 1) To show by example that there is only one set of EOFs, and one can iterate to those EOFs from drastically different directions, 2) The polarity, while arbitrary, is set by the first guess. For example EOF1 in Fig.5.5 results from whatever the anomalies were in 1989, while in Fig.5.1 the polarity of EOF1 relates to the starting point being a basepoint near Greenland, 3) The first EOT after EOF1 is removed differs from the first EOT after EOT1 is removed. While in Fig.5.4 1955 is the 2nd year chosen, in Fig.5.5 it is 1948. Comparing Fig.5.1 to 5.5 some modes comes out in the same (opposite) polarity, but within the framework of (5.3a/b) these are obviously the same functions in both in time and space.

5.4 Discussion of EOF

5.4.1 Summary of procedures and properties

Fig.5.6 is an attempt to summarize in one schematic the various choices a researcher has and the operations one can perform. Obviously, one needs a data set, follow the two options (Q and Q^a), see the three possibilities EOT, alternative EOT and EOF. In all three cases one can have a complete empirical orthogonal function representation of the data. Only in the latter case is an eigenanalysis called for. The EOFs can be ‘rotated’ in the direction of EOT. From any initial conditions one can iterate towards the gravest EOF.

5.4.2 The spectrum

Fig.5.7 shows the EV as a function of modenumber for seasonal JFM Z500. For the EOF line (red) this is also a spectrum of $\lambda(m)$. The EV by mode, left scale, decreases with m because of the ordering. The cumulative EV (right scale) obviously increases with m and is less noisy. The EOFs are more efficient than either version of EOT but never by more than a few % (10% cumulative). The cumulative EV lines for EOT and EOF are best separated at about m=8 for this data set. But as we add more modes the advantage of EOF decreases. EOFs have a clear advantage only for a few modes. For either EOF or EOT only two modes stand out, or can be called principal components, already at mode 3 we see the beginning of a continuum of slowly decreasing poorly separated EV values. In this case the regular EOT are slightly more efficient than alternative EOT, but this is not generally true. This relative advantage depends on the number of points in time and space. If one imagines random processes taking place at a large number of gridpoints, the regular EOT would be very inefficient, on the order of EV=100%/n_s per mode, while the alternative EOT cannot be less efficient than 100%/n_t

5.4.3 Interpretation of EOF

The interpretation of EOFs in physical terms is rarely straightforward. Adding to the difficulties are the vagaries of the procedure and display. The details of the covariance matrix, the exact domains used, the weighting of non-equal area grids etc varies. Even the display convention is confusing. Eq. (5.3) consists of a time series with physical units multiplying a spatial field of non-dimensional regression coefficients. The latter are the spatial patterns of the EOF but many authors have displayed instead the correlation between the time series and the original data. While this may look better, these are not the EOFs. Nevertheless, in spite of these problems, IF (if!) there is an outstanding mode (like ENSO in global SST or the NAO in monthly or longer time mean sea-level pressure) any of the techniques mentioned will find them. Problems only arise with the less than principal modes, poorly separated from each other in EV etc, On the other hand why bother interpret such modes?

5.4.4 Reproducibility (sampling variability)

We have described methods to calculate orthogonal functions from a given data set. To what extent are these the real EOFs of the population?. There are two different issues here. One is that the looks of a certain mode may change when the data set (a sample!) is changed slightly. Since EOF has been done often with a bias towards spatial maps and teleconnections this is a problem for the interpretation. Zwiers and von Storch(1999) are probably the best text in which sampling variability is addressed. The second important issue is the degree to which orthogonal functions explain variance on independent data. Even if modes are hard to reproduce they may continue to explain variance on independent data without too much loss, see examples in Van den Dool et al (2000). Hence the sampling errors are mainly a problem for the physical interpretation, not for representing the data concisely. Here the North et al(1982) rule regarding eigenvalues applies: Never truncate a λ_m spectrum in the middle of a shoulder (a shoulder or knee is an interruption in an otherwise regular decrease of eigenvalue with modenumber). Eigenvalues that are not well separated indicate possible problems with reproducibility (in terms of looks). Including all elements of the shoulder allows EV to be relatively unharmed on independent data. Recent work on this topic includes Quadrelli et al(2005).

5.4.5 Variations on the EOF theme

In the literature one can find related procedures, such as ‘extended’ EOFs (or ‘joint’ or ‘combined’ EOFs), abbreviated as EEOF. This means that 2 data sets h(s,t) and g(s,t) are merged into one f(s,t), then the EOF is done on f(s,t) as before. For example Chang and Wallace(1987) combined a precip and temperature US data set and did EOF analysis on the combined data. It is convenient (but not necessary) to have the two variables on the same grid (stations). In that case EEOF is like extending the space domain to double the size. While EEOF is methodologically the same as EOF, one needs to make decisions about the relative weighting of the participating fields - in the case of precip and temp one even has different physical units, so standardized anomalies may be an approach to place T&P on equal footing. If the number of gridpoints is different one could also adjust relative weighting to compensate. One has this problem already with EOF on a univariate height field from 20N-90N. Because the standard deviation varies in space some researchers prefer standardized anomalies, for instance to give the tropics a better chance to participate (Barnston and Livezey 1987). This goes back to using covariance versus correlation matrices. There is no limit to the number of data sets in EEOF. Some authors merge time lagged data sets to explore time lagged (predictive) connections through EEOF. When two data sets are unrelated the EEOFs are the same as the original EOFs in the participating data, and chosen in turn. The first EEOFs in meteorology are probably in Kutzbach(1967).

Another procedure used is called complex EOF. This variation on EOF is used to specifically find propagating modes. In chapter 3 we used sin/cos to diagnose propagation (and we could have called EWP 'complex'). But EOFs are not analytical and not known ahead of time, so it is unclear whether any structures will emerge that are '90 degrees out of phase' so as to suggest propagation. Nevertheless, the EOFs based on daily data in the SH (Fig. 5.3) certainly have the looks of propagating waves. Another way is to inspect the $\lambda(m)$ spectrum of regular EOF to search for pairs - in the case of propagation one expects two modes with nearly equal EV side by side. Such pairs (which may not be a case of poor separation) may indicate propagation, but this is not a sure sign. A more automated way of finding propagation is to transform f(s,t) into $f(s,t)+\sqrt{(-1)} * h(s,t)$, where h(s,t) is the Hilbert transform of f(s,t) (which is found by shifting all harmonics in f(s,t) over 90 degrees). Complex EOF is done on transformed data. This method has been tried at least since the early 1980's, see for instance (Horel 1984), Branstator(1987), Kushnir(1987) and Lanzante(1990).

5.4.6 EOF in models:

If EOFs are so efficient, why haven't they replaced arbitrary functions like spherical harmonics in spectral models (very widely used in NWP from 1980 onward)?. Essentially the use of EOF would allow the same EV for (far) fewer functions and be more economic in terms of CPU. EOFs are efficient in EV, but they are not efficient in CPU in transforming from a grid to coefficients and back. Spectral models did not become a success until the Fourier/Legendre transforms had been engineered to be very fast on a computer. Moreover, EOFs are not really functions, they are just a matrix of numbers, without easy (analytical) recipes for derivatives, interpolation etc. A further drawback is that the EV efficiency of EOFs is about anomalies relative to a mean state. I.e. one would need to make an anomaly version of an NWP model to exploit the strength of EOF, not a very popular research direction either. Still EOFs have been investigated (Rinne and Karhilla 1975) and if CPU is not the issue they are obviously as good as any other function or better (Achatz and Opsteegh 2003). A final point about EV efficiency is that while EOFs are the most efficient, this feature in reality often applies only to a few functions that truly stand out (the principal components). After the first two modes are accounted for, the next functions add all very small amounts of variance, see Fig 5.7. If numerical prediction requires truncation at say 99% of the observed variance, EOFs are barely efficient relative to more arbitrary functions. Only when severe truncation to a few modes is required EOF is useful. Yet another issue is that accuracy in the time derivative (the main issue in prediction) may require insight in unresolved scales, or at least the interaction between the resolved and unresolved scales. This is somewhat possible with analytical functions, but is not easy with EOFs (Selten 1993) who studied, as an alternative, EOF of the time derivatives.

Other reasons to use observed EOFs in models may include a desire to study the energetics, stability, and maintenance of EOFs in a complete dynamical framework (Schubert

1985). This works only if the model's EOFs are close to observed EOFs.

Models formulated in terms of observed EOFs only use spatial orthogonality. No one has considered it possible to use the temporal orthogonality. It follows that one might as well use alternative EOTs. This would allow empirical access to the time derivative.

All of the above was about functions in the horizontal. EOFs are also efficient for representation of data in the vertical. Because there are fewer other (efficient or attractive) methods for the vertical, EOFs have been used in operational models for the vertical direction (Rukhovets 1963) for many years.

5.4.7. More examples

In this chapter we have shown just a few examples. It is beyond the scope of this book to investigate EOF/EOT of all variables at all levels throughout the seasonal cycle, both hemispheres, daily as well as time filtered data. The Barnston and Livezey(1987) study remains unique in that sense. Two final examples of EOF calculation will suffice here, see Fig.5.8 and 5.9. Fig. 5.8 is an EOF analysis for global SST. The season chosen is OND, when the variance is near its seasonal peak, and the EV for the first mode is well over 30%. The first mode absolutely stands out and it is identified as the ENSO mode as it manifests itself in global SST data. Starting the iteration from 1975, the polarity is the opposite of a warm event. The time series on the right indicates all famous warm event years (1972, 1982, 1997) as large -ve excursions. Cold event years, except 1998, are harder to trace in the time series as they project on more than one mode, i.e. cold events are not just the opposite of warm events. The 2nd mode is very striking also and has a near uniform increase in the time series over 58 years, and indicates a pattern of warming SST in many oceans, except east of the dateline along the equator in the Pacific and in parts of the Southern Oceans. This mode had its strongest +ve projection in the 98/99 cold event. The 3rd mode is one of interdecadal variation, and a period of 30-35 years. At this point, however, modes explain only 5-6% of the variance.

Fig. 5.9 gives more information about both sensitivity of EOF to details and the possible impact of ENSO on mid-latitude. Fig.5.9 is identically the same as Fig. 5.5 except that we use streamfunction (ψ) instead of height (Z). Since the horizontal derivatives of ψ and Z are approximations to the same observed wind these two variables are closely related in mid-latitude. Under geostrophic theory $\psi=Z/f$, where f is the Coriolis parameter. At lower latitude the variations in ψ are thus more pronounced as those in Z, although not nearly as much as when using the correlation instead of the covariance matrix for Z. Fig. 5.9 shows the familiar NAO and PNA as modes # 2 and 3 at somewhat reduced variance (3 and 2% less than in Fig 5.5, respectively), but they are preceded by a mode we have not seen before, moreover explaining 32% of the variance. One can tell from the high projection in 1983 (the seed) and 1998 that this pattern is active during warm ENSO years. The main action is a deep low in the North Pacific near 40N and 160W and a like signed anomaly over the US Gulf coast. Together these ψ centers modulate the subtropical jets in both oceans. While the 1st mode looks loosely speaking like a Pacific North American Pattern it is in fact orthogonal to the 3rd pattern. Clearly, any statements on the impact of ENSO on the mid-latitudes, as per EOF analysis, requires careful study. Changing the domain size in Fig. 5.5. down to 10N, or the equator also has a large impact, because an ENSO mode would be first, followed by NAO and PNA.

5.4.8 Common misunderstandings

a) *The 1st EOF 'escapes' the drawbacks of being forced to be orthogonal because it is the first.* This opinion is wrong. There is no first EOF. All EOFs are simultaneously known by solving (5.2), and this is true even if only one of them is calculated. Ordering of EOFs is entirely arbitrary relative to the calculation method. The opinion in italics is, however, correct for EOT. For instance in Fig.5.4 the field observed in 1989 is the first alternative EOT. Issues of orthogonality come in only when chosing the second EOT.

b) *EOF patterns always show teleconnections*. This is wrong. They may or they may not. For instance, Fig.5.1 would suggest that the main center for the PNA in the Pacific is accompanied by a same signed anomaly in the Greenland/Iceland/Norway area. Fig. 4.1 shows that the simultaneous correlations between the Pacific and the Atlantic are extremely weak. While the EOTs in Fig 4.4 are faithful to teleconnections (defined by linear simultaneous correlation), the EOFs in Fig.5.1 are not (everywhere).

c) *EOF analysis forces warm and cold ENSO events to come out as each other's opposite*. While linearity has its drawbacks, this is not one of them. There is nothing against time-series that have very strong values in a few years to be compensated by weak opposite anomalies in many other years; i.e. nothing in the procedure forces strong opposing values in just a few years. Fig 5.8 demonstrates this for OND SST. The largest Pacific warm events are clear in the first mode (1972, 1982, 1997), but 1998 (a strong cold event) does NOT have a strong opposite projection in the first mode. In fact cold events can only be reconstructed using several modes, so the asymmetry noted by several authors (Hoerling et al 1997) is not butchered by the EOF procedure.

d) *EOFs describe stationary patterns only*. Geographically fixed patterns are no drawback in describing a moving phenomena. A pair of EOFs, each stationary in its own right, but handing over amplitude from one to the other as time goes on, describe a moving system. Daily unfiltered data, full of moving weather systems, can be described 100% by EOFs.

5.4.9 Closing comment

We entered this chapter in classical fashion, the Q matrix, 1-point teleconnections, regular EOT, and associated diagnostics but we want to leave it in the alternative fashion, emphasizing the Q^a matrix. The search for analogues (ch 7), the featured recipe to calculate degrees of freedom (ch 6), alternative EOT and the construction of analogues (ch7) all use Q^a. Linking orthogonal functions to specific moments in times past has one major advantage, i.e. direct access to the temporal evolution. If one expresses a specific state in the atmosphere as a linear combination of alternative EOTs one can easily make a forecast using a linear combination of the states that followed.

Inset or appendix I (applies to Figs 4.4, 5.1, 5.2, 5.3, 5.4, 5.5 and 5.9)

About the graphical presentation of EOT and EOF, units, normalization etc. Let's assume that a data set f(s,t) can be represented by (5.3) or (5.3a). We now discuss some postprocessing, which in a nutshell is a matter of finding a factor c by which to divide e and multiply α. The l.h.s. of (5.3) does not change in this operation:

$$f(s,t) = \sum_m \alpha_m(t) * c(m) \ e_m(s) / c(m) \qquad (5.3b)$$

$$f(s,t) = \sum_m \beta_m(s) * c^a(m) \ e_m^a(t) / c^a(m) \qquad (5.3c)$$

Factor c is a function of m, and, c is different depending on whether we start with normal or reverse set-up, hence the superscript a. The reasons for doing these extra manipulations are varied. One could wish to have unit vectors in either space or time, regardless of how the calculations is done. Another reason is to make it more graphically obvious that EOFs obtained by normal and reverse calculation are indeed identically the same. Here we present a postprocessing which makes, in all four (EOF/EOT, normal/ reverse) possible cases, the spatial maps of unit norm, and places the variance and physical units in the time series.

Thus note the following:

1) We plot maps and time-series consistent with (5.3). We do not plot correlations on the map (as is very customary), because e is really a regression coefficient (when the α_m carry the physical units and the variance).

2) To say that a mode m calculated under (5.3) is the same as mode m calculated under (5.3a) only means that $\alpha_m(t)\, e_m(s) = \beta_m(s)\, e_m^a(t)$; but $e_m(s)$ and $\alpha_m(s)$ are not the same, in general. But with the appropriate c and c^a applied $\alpha_m(t)*c(m) = e_m^a(t) / c^a(m)$ and $\alpha_m(s)*c^a(m) = e_m(s) / c(m)$

3) In order to force two identical modes to actually look identically the same, we do the following. We divide the map at each point by its spatial norm (and multiply the time series by this same norm so as to maintain the l.h.s. as in (5.3, (5.3c)). The spatial norm of $e_m(s)$ is defined as

$c(m) = \{ \sum e_m(s)\ e_m(s) \}^{½}$, similarly

$c^a(m) = \{ \sum \beta_m(s)\ \beta_m(s) \}^{½}$

where the sum is over space. This action would make all maps of unit norm and places the variance in the time series. The plots thus show, for example, $\alpha_m(t)*c(m)$ as the time series.

4) One can still tell how the calculation was performed, depending whether base-points or seed years are mentioned in the label.

It turned out to be unsatisfactory vis-a-vis the contouring package to plot the normalized maps, so for cosmetics, we divide the map at each point by its absolute maximum value - this procedure creates maps with a maximum value of +/-1 (nearly always +1).

Inset or Appendix II. Iteration

Often one needs to calculate only a few principal components. An insightful method is the so-called power method. Given Q, and an arbitrary initial state x_0 , one simply repeatedly executes $x_{k+1} = Q\ x_k$, k=0,1 etc. In view of $Q\ e_m = \lambda_m\ e_m$ (5.2), x_{k+1} will converge to the eigenvector associated with the largest eigenvalue. This is so because if x_0 contains at least minimal projection onto e_1 that projection will be multiplied by λ_1 and since λ_1 is larger than all other eigenvalues the projection onto e_1 will ultimately dominate x_k for large k. Once x_k has converged the eigenvalue is found as the multiplication factor that results from executing Q x_k . At each step of the iteration one may need to set the norm of x_k equal to that of x_0 for stability. This methods only fails to find the first eigenvector if the separation of λ_1 and λ_2 is very very small, i.e completely degenerate or pure propagation, a rare circumstance in practice. Once the first eigenvector is found one can proceed by removing the projection onto the estimate of e_1 from f(s,t), then recalculate a new Q from once reduced f(s,t). At this point the 2^{nd} eigenvalue should dominate. Etc. Convergence is often very quick. An arbitrary guess x_0 will do but the EOT's are obviously a better initial guess. The iteration thus described is a rotation of the EOTs in the direction of EOFs. Van den Dool et al(2000) does the iteration even without a covariance matrix.

End=chapter 5