Outline:

1. Introduction to regression and correlation analyses
    1a. Overview of the mathematics
    1b. Application of regression and correlation analyses to climate data sets

2. Introduction to EOF analysis
    2a. Overview of the underlying mathematics of EOF analysis
    2b. A short example of EOF analysis in two dimensions
    2c. Application of EOF analysis to climate data sets

3. Introduction to spectral analysis (to be done).
    3a. Overview of the mathematics
    3b. Application to climate data.

A brief overview of statistical methods used in climate dynamics

In this chapter, we introduce the reader to the three statistical tools most commonly used in the climate dynamics literature: regression and correlation analyses; empirical orthogonal function (EOF) analysis, and spectral analysis. The goal of the chapter is to provide the reader with the requisite background needed to interpret the application of the tools in the climate literature.

The practical application of the tools to real-world data is complicated by the effects of missing data, uneven spatial and temporal sampling, variable grid sizes, persistence, etc. The reader is referred to Hartmann (2006) for a more detailed treatment on how to apply the tools in research.

1. Introduction to regression and correlation analyses

Regression and correlation analyses are arguably the most commonly used, most commonly misused, but most important statistical tools applied in the atmospheric sciences. In this section we will briefly outline the mathematics that underlie the regression and correlation coefficients and review the application of the coefficients to climate data.

1a. Overview of the mathematics

The basic idea of regression/correlation analysis is to predict the behavior of one variable (the predictand) based on fluctuations in one or more related variables (the predictors). It is possible to estimate the behavior of the predictand using multiple predictors (multiple regression), and it is also possible to predict the behavior of the predictand based on nonlinear relationships with the predictors. In this section, we consider only the case of *linear* regression based on a *single* predictor. In practice, the following discussion will help you to interpret most of the results you will run across in the climate literature.

Imagine you have two variables measured as a function of time: *y(t)* and *x(t)*. In the case of linear regression based on a single predictor, we are interested in predicting fluctuations in *y(t)* assuming a linear relationship with fluctuations in *x(t)*. That is:

$$1)\ \hat{y}(t) = a_1 x(t) + a_0$$

where $\hat{y}(t)$ denotes the *linear least squares* best fit of *y(t)* to *x(t)*. We'll learn how to find the linear least squares best fit shortly. But first, let's consider the physical relationships between the parameters *x(t)*, *y(t)* and $\hat{y}(t)$.

Figure 1 shows an example of a scatter plot for values of *y(t)* plotted as a function of *x(t)*. A single dot in the figure thus corresponds to values of *x(t)* and *y(t)* at a single timestep *t*. In the example, it is clear *x(t)* and *y(t)* exhibit a strong linear relationship, such that higher values of *x(t)* tend to be associated with higher values of *y(t),* and vice versa. The solid line in Fig. 1 denotes the linear least squares best fit to the data as given by Eq. 1.

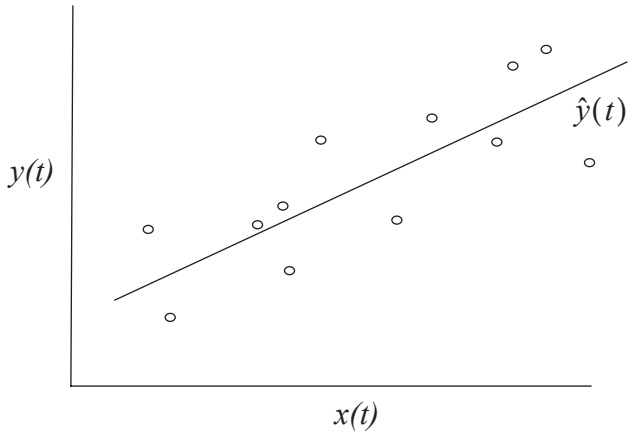

Fig. 1. Sample scatter plot for values of x(t), y(t) and the linear least squares best fit given by $\hat{y}(t)$.

Figure 1 includes two important pieces of information regarding the relationship between *x(t)* and *y(t)*. First, the slope of line $\hat{y}(t)$ provides an estimate of how much *y(t)* changes as a linear function of *x(t)*. Second, the grouping of the data about line $\hat{y}(t)$ tells us how much of the total variability in *y(t)* is described by the linear relationship with *x(t)*. For example, if the cloud of data is closely spaced about the line, then a large fraction of the variability in *y(t)* is predicted by $\hat{y}(t)$, but if the dots are scattered widely about the line, then only a small fraction of the variability in *y(t)* is predicted by $\hat{y}(t)$.

How do we solve for line $\hat{y}(t)$? First, we assume *x(t)* is known with precision. This is not always a precise assumption, but is reasonable for the types of analyses considered in this book. Second, we define an error function which quantifies the difference between the estimate of $\hat{y}(t)$ and the actual values of *y(t)*:

$$2)\ Q = \sum_{i=1}^{N} (\hat{y}(i) - y(i))^2$$

Substituting from 1):

$$3)\ Q = \sum_{i=1}^{N} (a_1 x(i) + a_0 - y(i))^2$$

In 2) and 3), Q is an estimate of the error between $\hat{y}(t)$ and *y(t),* and N is the number of time steps in both *x(t)* and *y(t)*. The differences between $\hat{y}(t)$ and *y(t)* are squared because:1) Squaring the differences ensures the error is always positive definite. Otherwise, the error will change sign from one data point to the next, and the average error will be close to zero. 2) Squaring the error ensures the minimization of 2) is a linear problem, as will be shown below.

In order to find the best fit of $\hat{y}(t)$ to the data, we want to minimize the error given by 2) and 3) (hence the phrase, *linear least squares*). This is done by taking the derivative of 3) with respect to the coefficients a_1 and a_0 and setting the resulting equations equal to zero:

4) $\frac{dQ}{da_0} = 0 = 2\sum_{i=1}^{N}(a_1x(i) + a_0 - y(i)) = a_1\sum_{i=1}^{N}x(i) + a_0N - \sum_{i=1}^{N}y(i)$

5) $\frac{dQ}{da_1} = 0 = 2\sum_{i=1}^{N}(a_1x(i) + a_0 - y(i)) \bullet x(i) = a_1\sum_{i=1}^{N}x(i)^2 + a_0\sum_{i=1}^{N}x(i) - \sum_{i=1}^{N}x(i)y(i)$

dividing through by *N*, and recalling that the mean is given by $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x(i)$ allows us to rewrite 4) and 5) as:

6) $\bar{y} = a_1\bar{x} + a_0$

7) $\overline{xy} = a_1\overline{x^2} + a_0\bar{x}$

Eq. 6) and 7) comprise two equations and two unknowns. The solutions for a_1 and a_0 are:

8) $a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$

9) $a_0 = \bar{y} - a_1\bar{x}$

At this point, we've solved for the line that gives us the linear least squares best fit to the data. But a_1 is typically expressed in terms of departures about the time mean. To do this, first divide variables *x(t)* and *y(t)* into their time means and deviations about their time means:

10) $x(t) = \bar{x} + x(t)'$ and $y(t) = \bar{y} + y(t)'$

From which it follows that:

11) $\overline{xy} = \overline{(\bar{x} + x)(\bar{y} + y)} = \overline{\bar{x}\bar{y} + \bar{x}y' + x'\bar{y} + x'y'} = \bar{x}\bar{y} + \overline{x'y'}$

and

12) $\overline{x^2} = \bar{x}^2 + \overline{x'^2}$

Substituting from 11) and 12) into 8), the expression for a_1 can be rewritten as:

13) $a_1 = \dfrac{\overline{x'y'}}{\overline{x'^2}}$

The numerator in Eq. 13 should be readily recognizable as the covariance between *x(t)* and *y(t)*, and the denominator as the variance of *x(t)*. Thus the slope of the line that gives us the linear least squares best fit of *y(t)* to *x(t)* is equal to the covariance between *x(t)* and *y(t)* divided by the variance of *x(t)*. The corresponding *y*-intercept is given by Eq. 9.

Now that we know how to find the slope of the line that gives us the best linear fit between *x(t)* and *y(t)*, how do we find the corresponding "goodness" of the fit? The goodness of the fit can be defined as the fraction of variance in *y(t)* explained by $\hat{y}(t)$: i.e., if the data are closely spaced about the line in Fig. 1 (that is, the observations given by *y(t)* closely resemble the linear fit to *x(t)* given by $\hat{y}(t)$), then a substantial fraction of the variance in *y(t)* is described by $\hat{y}(t)$, and vice versa.

The fraction of variance in *y(t)* explained by $\hat{y}(t)$ is mathematically equal to the variance of $\hat{y}(t)$ divided by the variance of *y(t)*. The variance of $\hat{y}(t)$ is equal to:

14) $\frac{1}{N}\sum\left(\hat{y}(t)-\bar{\hat{y}}\right)^2$

where the summation implicitly denotes the sum over the *N* timesteps in the data.

Since $\bar{\hat{y}} = a_1\bar{x} + a_0 = \bar{y}$, 14) can be rewritten:

15) $\frac{1}{N}\sum(\hat{y}(t)-\bar{y})^2$

Similarly, the variance of *y(t)* is equal to:

16) $\frac{1}{N}\sum(y(t)-\bar{y})^2$

So the fraction of variance in *y(t)* explained by $\hat{y}(t)$ is equal to 14) divided by 16). For reasons explained below, this ratio is given the shorthand r^2 in the statistics literature. Dividing 14) by 16), and substituting from 1), 9), and 10):

17) $r^2=\frac{\sum(\hat{y}(t)-\bar{y})^2}{\sum(y(t)-\bar{y})^2}=\frac{\sum(a_1x(t)+a_0-\bar{y})^2}{\sum y'(t)^2}=\frac{\sum(a_1x(t)+\bar{y}-a_1\bar{x}-\bar{y})^2}{\sum y'(t)^2}=\frac{\sum(a_1x'(t))^2}{\sum y'(t)^2}$

Substituting from 13) and exploiting the definition of the mean:

18) $r^2=\frac{\sum(a_1x'(t))^2}{\sum y'(t)^2}=\frac{\sum\left(\frac{x'(t)y(t)'}{x'(t)^2}\right)^2\sum x'(t)^2}{\sum y'(t)^2}=\frac{\sum(x'(t)y'(t))^2}{\sum x'(t)^2\sum y'(t)^2}=\frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}}$

The quantity r^2 is commonly used in the climate literature. As noted above, it corresponds to fraction of variance explained by a linear least squares fit between two variables. r^2 always lies between 0 and 1, and a value of 1 means that 100% of the variance in *y(t)* is explained by the linear fit to *x(t)*.

The corresponding quantity *r:*

19) $r=\frac{\overline{x'y'}}{\sqrt{\overline{x'^2}}\sqrt{\overline{y'^2}}}$

is referred to as the coefficient of correlation (or correlation coefficient), and varies from -1 to 1. Negative values of *r* denote an out of phase relationship between *x(t)* and *y(t)*, and vice versa. As evidenced in 19), the correlation coefficient is equal to the covariance between *x(t)* and *y(t)* divided by the product of the standard deviations of *x(t)* and *y(t)*.

The quantity r is more commonly used than r^2 in the climate literature because it is directly related to the statistical significance of the linear relationship between two time series. For a detailed discussion of assessing significance using the correlation coefficient, see the discussion in Hartmann (2006). For the purpose of this textbook, it is sufficient to know that the t-statistic for a particular correlation coefficient and sample size is given as:

$$20)\quad t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

where N is the effective sample size (i.e., the sample size taking into account persistence inherent in the data), and the significance a particular value of t can be found from standard lookup tables included in any elementary statistics textbook.

Note that from 19) and 13), the regression coefficient can also be expressed in terms of the correlation coefficient:

$$21)\quad a_1 = r \cdot \frac{\sqrt{\overline{y'^2}}}{\sqrt{\overline{x'^2}}}$$

Hence the regression coefficient of *y(t)* on *x(t)* is equal to the corresponding correlation coefficient multiplied by the ratio of the standard deviations of *y(t)* and *x(t)*. The regression coefficient thus contains information about the correlation coefficient (i.e., the goodness of the fit) and the relative amplitudes of *y(t)* and *x(t)*. In the special case where *x(t)* and *y(t)* have equal standard deviations (e.g., if both time series are standardized), the correlation coefficient and the regression coefficient are equal.

1b. Application of regression analysis to climate data sets

Calculating the correlation and regression coefficients is straightforward. As is the case with the use of any statistical tool in climate dynamics, the skill lies in the interpretation of the results.

Below we use regression analysis to examine the structure of global sea surface temperature anomalies associated with one of the most important patterns of variability in the climate system: the El-Nino/Southern Oscillation (ENSO) phenomenom. We will examine "anomalous" data, that is data in which the seasonal cycle (and the diurnal cycle, if the data is sampled on timescales shorter than a day) has been removed from the data at all grid points. This is commonly done if we are interested in examining variability in the climate system that is not related to the seasonal cycle. Note that if the seasonal cycle is not removed from climate data prior to using correlation or regression analysis, most variables will exhibit high correlations with remote locations simply because all points on the globe are impacted to some extent by the annual cycle in solar heating.

The details of ENSO are discussed at length in Chapter xx. For the purpose of this discussion, you need only know that the dominant oceanic feature of ENSO is a periodic warming and cooling of sea-surface temperatures in the eastern tropical Pacific. Month-to-month and year-to-year variations in ENSO can be simply identified by averaging sea-surface temperature anomalies in the equatorial band between roughly the dateline and the coast of South America.

Figure 2 shows an example of a time series of ENSO based on an index of sea-surface temperature anomalies averaged between the dateline and 90W, and between 5S and 5N. Major warm ENSO events are evidenced by warm SST anomalies in the eastern tropical Pacific, such as around 1973, 1982, 1988, and 1998.

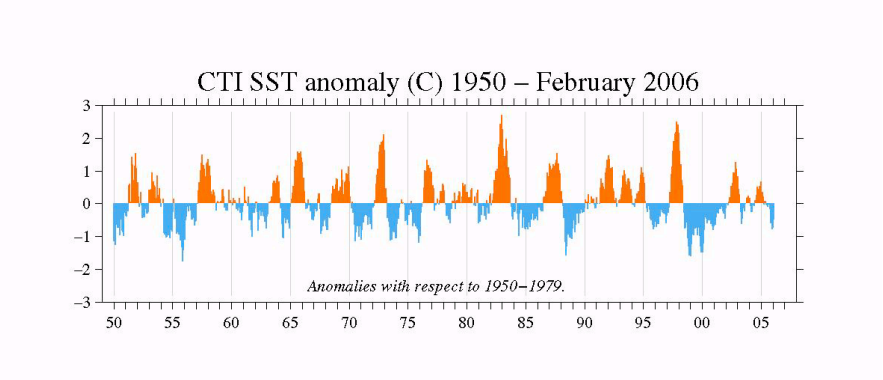

Fig. 2. Time series of SST anomalies averaged between 5S and 5N and 180-90W. Units K. Courtesy Todd Mitchell, Univ. of Washington.

Let's suppose you wish to answer the question: "how do SSTs vary throughout the globe in association with the ENSO cycle?". There are many ways to approach this problem using regression analysis. One is to regress and correlate SSTs at points throughout the globe onto the time series in Fig. 2. In this case, the resulting correlations will vary between -1 and 1 (as they must), and the resulting regression coef-

ficients will be in units of K (local change in SST) per K change in SSTs in the eastern tropical Pacific. Another, and more commonly used, approach is to regress SSTs at points throughout the globe onto *standardized* values of the time series in Fig. 2. To standardize a time series, you simply subtract the mean from the time series (in the case of anomalous data, the mean will already be zero) and then divide the time series by its standard deviation. The correlations based on the standardized index are unchanged from the case considered above, but the corresponding regression coefficients are now in units of K (local change in SST) per *standard deviation* change in SSTs in the eastern tropical Pacific. This is a particularly useful approach, since one standard deviation corresponds to a typical fluctuation in the base index time series.

Figure 3 shows the result of regressing SST anomalies at points throughout the globe onto standardized values of the index in Fig. 2. At all grid points, the numerical values can be thought of as corresponding to the slope of line given by a_1 in Fig. 1, but for the case where standardized values of the index in Fig. 2 lie along the abscissa, and local SST anomalies lie along the ordinate.

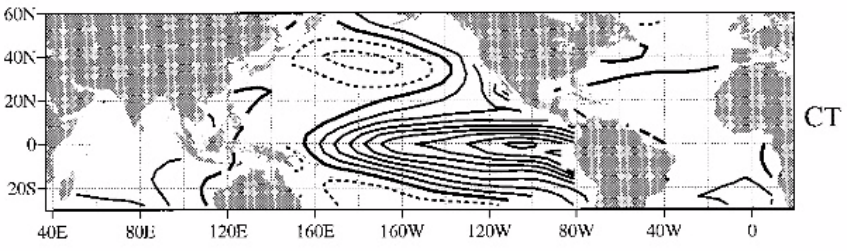

Fig. 3. Regression of SST anomalies onto standardized values of the index in Fig. 2. Contour interval is 0.1 K. From Zhang et al. 1997.

The results in Fig. 3 tell us that a one standard deviation change in SSTs in the eastern tropical Pacific (the warm phase of the ENSO cycle) is associated with warmer than normal SSTs that extend through a broad region extending from 20 N to 20 S, and from 160 E to the South American coast. The amplitudes of the SST anomalies exceed 1 K in the eastern tropical Pacific but are indistinguishable from 0 K in the western tropical Pacific. The figure also tells us that the warm phase of the ENSO cycle is associated with SSTs that are roughly 0.2 K colder than normal in the North Pacific.

The results in Fig. 3 show the amplitude of changes in SSTs associated with ENSO, but do not reveal the significance of the results. To do this, you could either plot the corresponding correlation coefficients in an adjacent panel (recall the correlations

are equal to Fig. 3 multiplied by the ratio of the standard deviations of the local SSTs and the index used as a basis for the regression). Alternatively, you could shade regions in Fig. 3 where the associated correlation coefficient exceeds the 95% level based on 20).

A few final points about regression (and correlation) analysis before moving on to EOF analysis:

- Regression analysis is a "linear" method. Thus in the example above, the cold phase of the ENSO cycle is constrained to have anomalies equal and opposite to those shown in Fig. 3. To identify "nonlinear" aspects of the relationship between ENSO and the global SST anomalies, it is common to composite (i.e, average) SST anomalies during warm and cold events separately.

- Regression analysis reveals statistical relationships only, and does not prove cause and effect. In the above example, we know from numerical experiments that the SST anomalies in, for example, the North Pacific, are a remote response to anomalies in tropical climate, and not vice versa (this will be discussed more in Chapter xx). But the nature of the causal relationship can not be inferred from the results in Fig. 3 alone.

- The significance of a correlation depends on the effective sample size (Eq. 20), but the effective sample size is almost always smaller than the number of data points used in the analysis. For example, a fifty year record of monthly-mean SST anomalies averaged over the Atlantic basin may include 50x12=600 time steps, but in reality the time series includes only a handful of independent samples because of the substantial memory inherent in area-averaged SSTs. Always be careful and critical when interpreting correlations based on limited sample sizes.

2. Introduction to EOF analysis

Empirical Orthogonal Function (EOF) or equivalently Principal Component (PC) analysis isolates structures that explain maximum variance in a data set. Like regression analysis, it is one of the most commonly used statistical tools in the climate literature. We will be showing the results of EOF analysis occasionally throughout the textbook.

Below we briefly review the key mathematics that underlie EOF analysis, provide a short example of EOFs in two dimensional data, and discuss how to interpret the results of EOF analysis as they appear in the literature. The goal here is to give you enough information to be able to interpret the results of EOF analysis. The details of how to compute your own EOFs are deferred to textbooks on statistical analysis, including, for example, von Storch and Zwiers (1998) and Hartmann (2006).

2a. Overview of the underlying mathematics of EOF analysis

The following demonstrates how to find the state vector that explains the largest possible fraction of total variance in data matrix A_{MxN}, where in this case M represents the sampling dimension (time) and N the structure dimension (space). In the vernacular of the climate dynamics literature, the resulting $1xN$ state vector is referred to as the "leading EOF" of the data. For example, data set A_{MxN} might reflect sea level pressure (SLP) anomalies recorded at M timesteps and N grid points throughout the globe. In this case, the state vector corresponding to the leading EOF represents the pattern of SLP anomalies that explains more variance in the global SLP field than any other state vector.

Let e_1 denote the state vector we are seeking. By definition, if e_1 explains the largest possible fraction of variance in data set A, then e_1 has the maximum possible "resemblance" to the ensemble of state vectors that make up A. The resemblance between e_1 and the data matrix A at a given timestep t can be quantified as the projection (or, equivalently, the inner product) between the two vectors, that is:

$$1) \sum_{j=1}^{N} A(t,j)e_1(j)$$

Where N is the number of spatial grid points in the data matrix. Before extending 1) to include the resemblance between e_1 and A at all time steps, three aspects of the analysis are worth special attention. First, the result of 1) can be either positive or negative. Thus e_1 may project strongly onto A at individual timesteps, but if the polarity of the projection varies from timestep to timestep, then the average of 1) over the entire data matrix will be very small. For this reason, the quantity in 1) is squared before being averaged over the data matrix. (Similarly reasoning is used when minimizing the root-mean-square error in linear regression, as discussed in

the previous section). Second, it is typical to force the length of state vector e_1 to be one (that is, the projection of e_1 with itself equals one). In this way, only the direction of e_1 and not its amplitude affects the projection. Third, the means are typically subtracted from the time series in A_{MxN} before performing the analysis. Otherwise the results in 1) are dominated by the mean of the data, whereas we are principally interested in variability about the mean.

The net resemblance between e_1 and the data matrix A can thus be quantified as the average over all time steps of the squared projection of e_1 onto A:

$$2)\ \lambda = \frac{1}{M}\sum_{i=1}^{M}(A(i,j)e_1(j))^2 = \frac{1}{M}(Ae_1)^2 = \frac{1}{M}e_1^T A^T A e_1$$

Since 2) provides a measure of the resemblance between e_1 and A, the leading EOF (i.e., e_1) is found by simply maximizing the expression for λ. We'll get back to the maximization of λ in a moment, but first a quick comment on the matrix $\frac{1}{M}A^T A$.

$\frac{1}{M}A^T A$ is a square matrix with dimensions NxN. First, consider the elements along the diagonal of $\frac{1}{M}A^T A$. Element (1,1) corresponds to the summation:

$$3)\ \frac{1}{M}\sum_{i=1}^{M}A(i,1)A(i,1)$$

where, as you'll recall, M is the number of time steps in the data. Since the means of the columns in A are zero, the quantity in 3) is equal to the variance of the time series at grid point $N=1$. Similarly, element (2,2) is the variance at grid point $N=2$, etc. So the diagonal of $\frac{1}{M}A^T A$ contains the variance of the time series at all grid points in A.

Now, consider the off-diagonal elements in $\frac{1}{M}A^T A$. Element (1,2) corresponds to the summation:

4) $\frac{1}{M}\sum_{i=1}^{M} A(i,1)A(i,2)$

which is the covariance between the time series at grid points *N=1* and *N=2* in *A*. Thus the off-diagonal elements in $\frac{1}{M}A^TA$ correspond to the covariances between time series at different grid points in the data matrix. Note that element (1,2) is identical to element (2,1), etc., and thus $\frac{1}{M}A^TA$ is a square *symmetric* matrix.

The matrix $\frac{1}{M}A^TA$ is referred to as the "temporal covariance" matrix of *A*. If we denote the temporal covariance matrix as C_{NxN}, then we can rewrite 2) as:

5) $\lambda = \frac{1}{M}e_1^TA^TAe_1 = e_1^TCe_1$

Multiplying both sides by e_1:

6) $Ce_1 = \lambda e_1$

Eq. 6) corresponds to a classical eigenvalue problem from mathematics. What does the expression tell us about e_1? It tells us that e_1 must be an eigenvector of the temporal dispersion matrix $\frac{1}{M}A^TA$. Since we are seeking to maximize λ, 6) also tells us that e_1 is the eigenvector of $\frac{1}{M}A^TA$ which has the largest possible eigenvalue λ.

In summary, the state vector e_1 that explains the largest fraction of variance in data matrix *A* can be found by:

1) solving for the temporal covariance matrix $\frac{1}{M}A^TA$;

2) eigenanalyzing $\frac{1}{M}A^TA$;

3) choosing the eigenvector of $\frac{1}{M}A^TA$ that has the largest eigenvalue.

As noted earlier, in the climate dynamics literature the eigenvectors of the temporal covariance matrix are called the Empirical Orthogonal Functions (EOFs) of the data. The eigenvector corresponding to the largest eigenvalue of the covariance matrix (e_1 in the above discussion) is referred to as the leading EOF, and is the pattern that explains the largest fraction of the variance in the data matrix. The eigenvector associated with the second largest eigenvalue of the covariance matrix is referred to as the second EOF, and is the pattern that explains the second largest fraction of variability in the data, etc.

The time series that describe the temporal evolution of the EOFs are referred to as the principal component (PC) time series. The PC time series can be found by either 1) projecting the original data matrix onto the vectors corresponding to the EOFs, or 2) eigenanalyzing the spatial covariance matrix $\frac{1}{N}AA^T$. (The mathematics of the two operations are identical).

Additional insight into EOFs can be derived by rewriting 6) as:

7) $CE = EL$

where C_{NxN} is again the temporal covariance matrix of A_{MxN}, but E is the matrix of all N eigenvectors (e), and L is a diagonal matrix with all N eigenvalues (λ) organized along its diagonal. From 7) it is evident that EOF analysis transforms the data into a new coordinate space where the new covariance matrix (L) is diagonal. Hence, the EOF analysis "diagonalizes" the temporal covariance matrix. The fraction of variance explained by eigenvector e_i is thus equal to eigenvalue λ_i divided by the trace of L (i.e., the sum of all eigenvalues).

As implied by the name of the analysis technique, EOFs comprise an orthogonal basis set. That is, the EOFs of a data set are linearly uncorrelated with each other. Mathematically, this is because the eigenvectors of any square symmetric matrix are orthogonal. In practice, the orthogonality constraint is key as it allows us to decompose the climate system into a series of structures and time series that are uncorrelated with each other. But the orthogonality constraint also means higher order EOFs are strongly impacted by the mathematics of the operation, as each successive EOF must be orthogonal to all EOFs with higher eigenvalues. For example: the

leading EOF has no orthogonality constraints placed on it; the second EOF must be orthogonal to the first EOF; the third EOF must be orthogonal to the first two EOFs, etc. For this reason, you should recognize that EOFs higher than EOF 1 may reflect mathematical constraints as much as true physical structure in the data. Be wary of over-interpreting higher order EOFs.

2b. A short example of EOF analysis in two dimensions

Below we provide a simple example of the application of EOF analysis in two - dimensions.

Consider data matrix *A* with dimensions 1000x2. Matrix *A* may be thought of as consisting of two time series of length 1000. In the example below, both columns have a mean of zero and are correlated at a level of $r=0.58$. The regression coefficient between the columns is roughly one.

Figure 1 shows a scatter plot of the columns in *A* plotted as a function of the abscissa $(1,0)$ and the ordinate $(0,1)$ axes. Physically, the abscissa corresponds to fluctuations in the first column in A with no corresponding fluctuations in the second column of A, while the ordinate corresponds to fluctuations in the second column in A with no corresponding fluctuations in the first column of A.

What direction would you choose if you wanted to describe as much as possible of the variability in *A* using a single vector in the two-dimensional plane? It is visually apparent that you would not choose the abscissa or the ordinate: the abscissa captures variations in *A* along the horizontal direction in the plot, but does not account for the considerable amount of variability in the vertical direction; the ordinate captures variations in *A* along the vertical direction in the plot, but does not account for the considerable amount of dispersion in the horizontal direction.

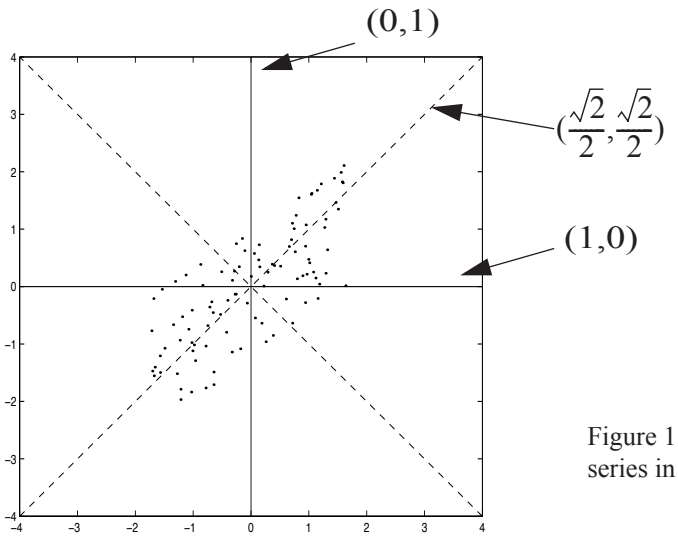

Figure 1. Scatter plot of the time series in sample data matrix *A*.

In fact, as evidenced in Figure 1, most of the variability in *A* lies not along the abscissa or ordinate, but along a coordinate axis that lies roughly 45 degrees above the abscissa. In vector notation, the direction corresponds to $(1,1)$, or in the case where the vector has length one, $(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$.

The direction spanned by vector $(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$ corresponds to the leading EOF of the data set. Physically, the vector $(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$ describes equal amplitude fluctuations in both columns. The second EOF of the data matrix is orthogonal to the first EOF (i.e, $[\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}]$), and describes out of phase fluctuations in the two columns. The two EOFs account for 88% and 12% of the variability in the data set, respectively.

Now see if you can predict the results of EOF analysis for different configurations of data set A_{1000x2}. First, imagine the time series in *A* correspond to monthly mean surface temperature data recorded at 1000 time steps at two midlatitude continental locations. If the annual cycle is not removed from the data, then much of the variability in the two time series will be dominated by the shared seasonal march from warm summer temperatures to cold winter temperatures. The two time series will thus be highly correlated simply by virtue of the fact that both stations exhibit simi-

lar seasonal cycles. In this case, the first EOF will correspond to equal amplitude fluctuations at both stations, the corresponding PC will be dominated by the seasonal cycle, and the first EOF/PC pair will account for a very large fraction of the total variance in the data set.

Now imagine that the time series in *A* correspond to monthly mean surface temperature data recorded at a continental and maritime location. As in the previous example, the two time series will be highly correlated by virtue of the fact both stations exhibit similar seasonal cycles. But in this case, the amplitude of the seasonal cycle will be much larger at the continental location. The first EOF will thus describe same signed variations at both stations, but the amplitude of the EOF will be somewhat larger at the continental station. The first PC will again follow seasonal cycle.

Finally, imagine that the data in *A* are in anomaly form, i.e., the seasonal cycle (and the diurnal cycle, if the data is sampled on timescales shorter than a day) has been removed from each location. In this case, the EOF analysis will not be dominated by the seasonal cycle (obviously), and thus the results tell you whether the two time series tend to vary in-phase, out-of-phase, or exhibit no relationship on timescales other than the seasonal cycle.

Note that since the seasonal cycle typically dominates variability in most climate records, it is common to compute EOFs for anomalous data. In the following section, we will briefly overview the EOFs of anomalous data recorded at hundreds of grid points and time steps. But before moving onto these results, a few final points about calculating EOFs:

- It is possible to calculate EOFs based on eigenanalysis of either the covariance matrix (as reviewed in this chapter) or the correlation matrix. In the case of the latter, the time series in the original data are standardized before computing the EOFs.

Typically, you will want to retain the amplitude of the original data and thus use the covariance matrix in the eigenanalysis. This is because the EOFs should exhibit spatially varying amplitudes consistent with the spatially varying amplitudes inherent in the original data (i.e., it doesn't make physical sense to allow geopotential heights to have same amplitude fluctuations at tropical and high latitudes). But there are cases where the use of the correlation matrix is preferable, particularly if the state space reflects different parameters (e.g., relative humidity, temperature, etc.) recorded at the same location.

- The statistical significance of an EOF is commonly approximated by the degree of separation between its eigenvalue and adjacent eigenvalues in the diagonalized covariance matrix (North et al. 1982). In this case, the error bars on eigenvalue λ_i are:

8) $$\Delta\lambda_i = \lambda_i \sqrt{\frac{2}{N}}$$

where $\Delta\lambda_i$ is the error bar on eigenvalue λ_i and N is the effective number of degrees of freedom in the data. The first EOF/PC pair is considered significant if the error bars on λ_1 do not overlap the error bars on λ_2, the second EOF/PC pair is considered significant if the error bars on λ_2 do not overlap the error bars on λ_1 and λ_3, etc.

In practice, N is difficult to estimate, so the significance given by 8) should be viewed only as an estimate of the reproducibility of a given EOF/PC pair. As always, the best way to assess reproducibility is to repeat the analysis for subsets of the data, e.g., if the EOF changes substantially between the first and second halves of the data, you should be duly concerned about its robustness.

2c. Application of EOF analysis to climate data sets

As is the case with all statistical tools, EOFs are generally easy to calculate using canned routines available in modern software packages. But the results of the analysis are not always easy to interpret physically. In this section, we will review the EOFs of three climate fields: global sea-surface temperature anomalies, Northern Hemisphere sea-level pressure anomalies, and Southern Hemisphere sea-level pressure anomalies.

Before we proceed to the results, a quick comment on the presentation of EOFs. One possible way to present the results of EOF analysis is to show the EOFs and PCs in the dimensionless units given by the eigenanalysis, i.e., both the EOFs and PCs are vectors with length one. This is a reasonable choice, but does not give the reader a clear sense of the spatially varying importance (i.e., amplitude) of a particular EOF.

An alternative is to take advantage of the fact the first EOF is equivalent to the projection of the data matrix onto the first PC. Since the projection is linearly related to the regression coefficient (see the discussion on the regression coefficient), it follows that the map formed by regressing the data onto the standardized leading PC has identical structure to the first EOF. The advantage of this method is that the amplitude of the resulting map is in units of change in the data per standard deviation change in the leading PC time series. Higher order EOFs will have weaker regression coefficients simply because they are increasingly less important in terms of variance explained in the data.

Note that if the data are weighted by the cosine of latitude (as they must be in the case of data gridded along meridians), then the regression of the data matrix onto the PC time series is not precisely identical to the first EOF. Nevertheless, the resulting regression maps are still very similar to the actual EOFs, and for this reason, are still referred to as EOFs in the literature.

On to the results.

The top panel in Fig. 2 is a reproduction from the section on regression analysis, and shows the regression of SST anomalies onto the cold-tongue index. The bottom panel in Fig. 2 shows the first EOF of global SST anomalies, presented as SST anomalies regressed onto standardized values of the leading PC time series. In the analysis, the global mean has been removed from the SST data at each grid point, otherwise the first EOF is dominated by the marked global warming of the past few decades.

What do the results in Fig. 2 tell you about large scale structure in the global SST field? They tell you that leading EOF of the global SST field corresponds to ENSO. Or put another way: the ENSO phenomenon is the leading pattern of variability in the global SST field.

The middle panel in Fig. 2 shows global SST anomalies regressed onto the leading PC time series of the tropical ocean area 30N-30S. Technically speaking, only the pattern between 30N and 30S corresponds to the leading EOF of the tropical SST field. But the key point of the result in the middle panel is that the leading EOF of the global SST field draws its variance from the tropical domain.

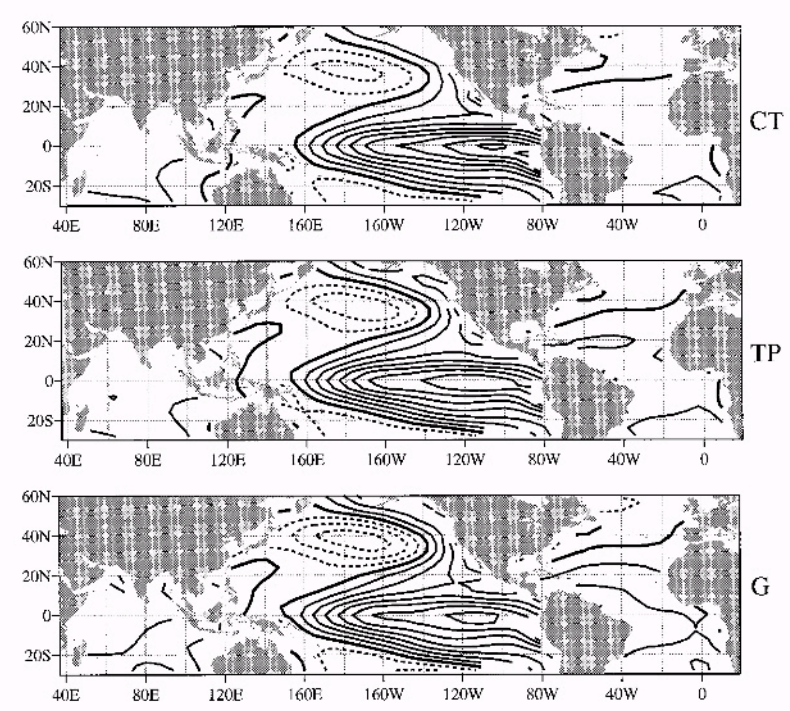

Fig. 2. SST anomalies regressed onto standardized values of (top) the cold tongue index; (middle) the leading PC time series of tropical SST anomalies; (bottom) the leading PC time series of global SST anomalies.

Figure 3 shows the leading EOFs of the NH (left) and SH (right) sea level pressure fields. Both patterns are characterized by coherent SLP anomalies throughout the polar regions with opposite signed SLP anomalies in middle latitudes. Thus in this case the EOF analysis tells you that the dominant patterns of variability in the SLP field of both hemispheres are both described by north-south vacillations in atmospheric mass between middle and high latitudes. The patterns in Fig. 3 have a strong annular, or ring-like, component, and are thus commonly referred to as the "annular modes" of climate variability. We will discuss the annular modes in more detail in Chapter xx.

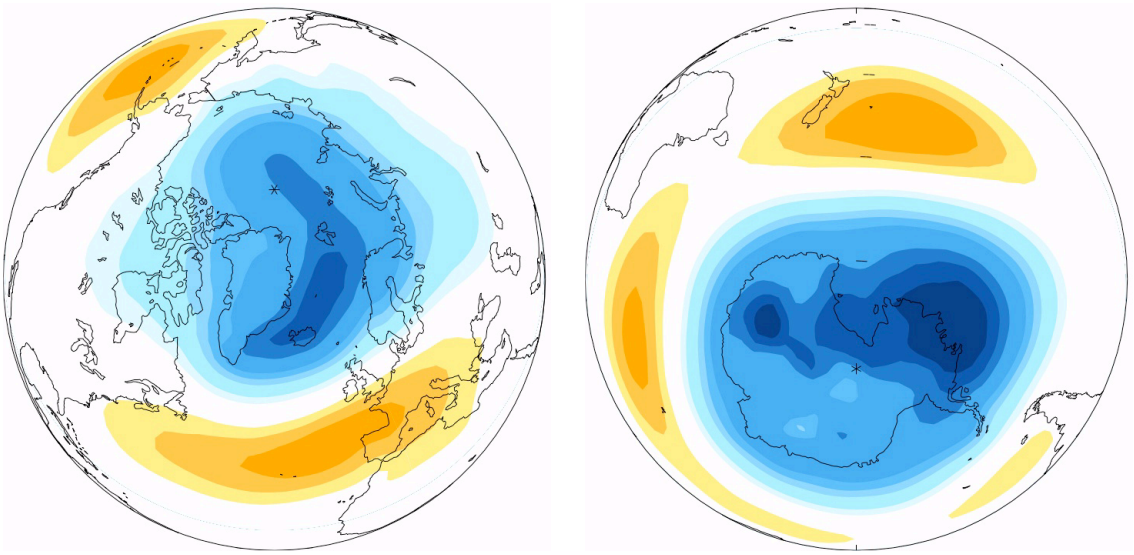

Fig. 3. SLP anomalies regressed onto the leading PC time series of the NH (left) and SH (right) SLP fields.

Possible problems.

1. Calculate the EOFs and PCs of the following 4x2 data matrix "by hand". Assume the sampling domain lies in the columns.

$$A = \begin{bmatrix} 2 & 4 \\ -3 & -6 \\ 1 & -2 \\ 0 & 4 \end{bmatrix}$$

a) First calculate the covariance matrix $\frac{1}{M}A^TA$. Explain why it is more efficient to eigenanalysis $\frac{1}{M}A^TA$ than $\frac{1}{N}AA^T$ in this case.

b) Find the eigenvalues and eigenvectors of $\frac{1}{M}A^TA$ (i.e., the EOF's of A). Scale the eigenvectors to length one.

c) Find the PC time series that correspond to the two EOFs by projecting the data matrix onto the leading EOFs. Hint: the leading PC can be found by solving Au_i, where u_i is the 2x1 leading EOF. Scale the PCs to length one.

d) Calculate the variance explained by each EOF/PC pair.