Empirical Orthogonal Function Analysis

(Principal Component Analysis)

(1) A Geometric and Intuitive Introduction

(2) Some related mathematics: Maximizing Variance

(3) Motivation for EOFs from Statistical Forecasting

(4) Use of EOFs as a Diagnostic / Analysis Tool

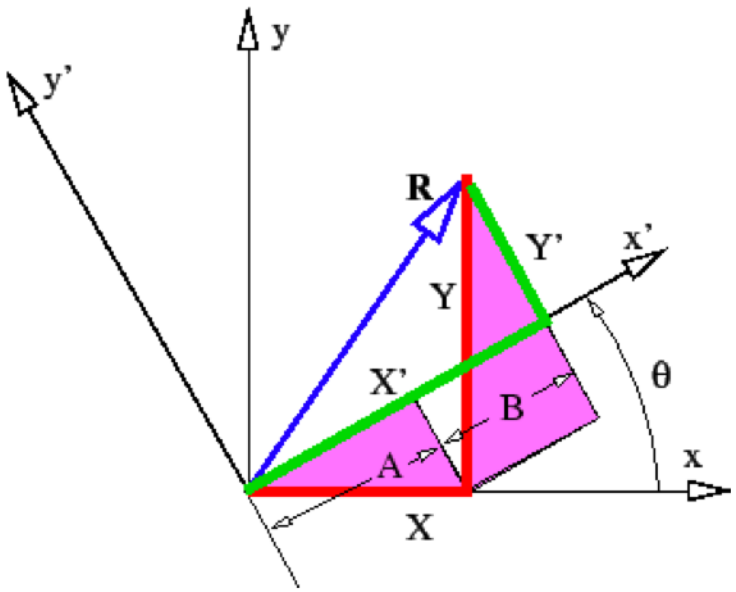


Figure 4: Definition figure for rotated coordinate system. The vector **R** has components *X* and *Y* in the unprimed coordinate system and components *X'* and *Y'* in the primed coordinate system.

A (2 dimensional) vector **R** may be represented in terms of the coordinates (X, Y), which are projections on the *orthogona*l unit vectors **x** and **y** as shown above. The condition of orthogonality for **x** and **y** is written as $\mathbf{x} \bullet \mathbf{y} = 0$, where the "•" indicates the ordinary dot product. The set of unit vectors **x** and **y** are called a **basis set**. (All **bold** quantities are vectors.)

Alternately, we can express the coordinates of the vector **R** in terms of the coordinates (X', Y'), the projections onto the *orthogonal* unit vectors **x'** and **y'** as shown above. This is an equally valid representation of the vector **R**, but in a rotated coordinate system.

Any orthogonal coordinate system may be used: The problem at hand may dictate that one system leads to certain advantages.

The coordinates depend on the basis (or unit) vectors:

$\mathbf{R} = X\,\mathbf{x} + Y\,\mathbf{y} = X'\,\mathbf{x}' + Y'\,\mathbf{y}'$

Using **x**,**y** as basis vectors, the coordinates are (X,Y)

Using **x**’,**y**’ as basis vectors, the coordinates are (X’,Y’)

Atmospheric States are routinely represented as Vectors

(Dimension is much higher than 2)

Grid Points: If there are N grid points, four variables (u, v, T, q) at L levels, and surface pressure P (in the case of the primitive equations), the (very big) vector would look something like:

$(u_{11}, u_{12}, \dots, u_{1N}, u_{21}, u_{22}, \dots, u_{2N}, \dots, u_{L1}, u_{L2}, \dots, u_{LN},$

$v_{11}, v_{12}, \dots, v_{1N}, v_{21}, {}_{22}, \dots, v_{2N}, \dots, v_{L1}, v_{L2}, \dots, v_{LN},$

$T_{11}, T_{12}, \dots, T_{1N}, T_{21}, T_{22}, \dots, T_{2N}, \dots, T_{L1}, T_{L2}, \dots, T_{LN},$

$q_{11}, q_{12}, \dots, q_{1N}, q_{21}, q_{22}, \dots, q_{2N}, \dots, q_{L1}, q_{L2}, \dots, q_{LN},$

$P_1, P_2, \dots, P_N)$

where u_{ki} is the variable **u** at level **k**, grid point **i.** Above, the different levels are color coded: **Level 1**, **Level 2**, **… , Level N.**

Often we only use one field at one level (e.g. 500 hPa height) to represent the atmospheric state.

Since each atmospheric state (call it **S**) is a vector, we can use alternate sets of orthgonal basis vectors.

Empirical Orthogonal Functions (EOFs) are one choice of orthogonal basis vectors. (They are have special properties - which we will get to !)

Each EOF is represented by a set of grid points of the different variables in the state S.

The coordinates of any atmospheric state are then the Principal Components, which are just like the coordinates (X', Y') in the picture, only in N dimensions instead of 2.

- We can write any 2 dimensional vector **R** in the basis set of **x** and **y** as:

$$\mathbf{R} = X\,\mathbf{x} + Y\,\mathbf{y}$$ (where **x** and **y** are 2 dimensional)

-We can write any atmospheric state **S** using EOFs **e** as:

$$\mathbf{S} = P_1\,\mathbf{e_1} + P_2\,\mathbf{e_2} + P_3\,\mathbf{e_3} + P_4\,\mathbf{e_4} + \ldots$$

The new "unit vectors" are the EOF patterns **e**, the new coordinates are the "Principal Components" P, which depend on time.

Time Dependence

- If the 2 dimensional vector **R** depends on time, and we express it in terms of a fixed basis set of **x** and **y**, then we have:

$$\mathbf{R}(t) = X(t)\,\mathbf{x} + Y(t)\,\mathbf{y}$$ (where **x** and **y** are 2 dimensional),

so that the coordinates X and Y depend on time.

-We can write any atmospheric state **S** (t) using EOFs **e** as:

$$\mathbf{S} = P_1(t)\,\mathbf{e_1} + P_2(t)\,\mathbf{e_2} + P_3(t)\,\mathbf{e_3} + P_4(t)\,\mathbf{e_4} + \ldots$$

The new "unit vectors" are the fixed EOF patterns **e**, the new coordinates are the "Principal Components" P, *which depend on time*.

Orthogonality

- The two unit vectors x and y are orthogonal, and have unit length (*orthonormal)*:

$\mathbf{x} \bullet \mathbf{y} = 0$ $\qquad$ $\mathbf{x} \bullet \mathbf{x} = 1$ $\qquad$ $\mathbf{y} \bullet \mathbf{y} = 1$

$x_1 y_1 + x_2 y_2 = 0$ $\qquad$ $x_1 x_1 + x_2 x_2 = 1$ $\qquad$ $y_1 y_1 + y_2 y_2 = 1$

-EOFs **e** (in N-dimensiona) are also orthonormal. Take the nth EOF:

$$\mathbf{e}_n \bullet \mathbf{e}_n = e_{1n}\, e_{1n} + e_{2n}\, e_{2n} + \ldots + e_{Nn}\, e_{Nn} = 1$$

$$\mathbf{e}_\alpha \bullet \mathbf{e}_\alpha = 1 \qquad (\alpha = 1, \ldots, N)$$

$$\mathbf{e}_\alpha \bullet \mathbf{e}_\beta = 0 \qquad (\alpha, \beta \text{ not equal})$$

or:

$$\mathbf{e}_\alpha \bullet \mathbf{e}_\beta = \delta_{\alpha\beta}$$

Empirical Orthogonal Function Analysis

(Principal Component Analysis)

(1) A Geometric and Intuitive Introduction

(2) Some related mathematics: Maximizing Variance

(3) Motivation for EOFs from Statistical Forecasting

(4) Use of EOFs as a Diagnostic / Analysis Tool

Rewriting the expansion slightly, we have

$S_i(t) = P^{(1)}(t)\, e_i^{(1)} + P^{(2)}(t)\, e_i^{(2)} + P^{(3)}(t)\, e_i^{(3)} + P^{(4)}(t)\, e_i^{(4)} + \ldots$

where we have used the subscript i to denote the ith component of the vectors **S** and $\mathbf{e}^{(1)}$, $\mathbf{e}^{(2)}$ … and have introduced the time dependence explicitly. More formally, this can be written as:

$$S_i(t) = \sum_{\alpha} P^{(\alpha)}(t)\, e_i^{(\alpha)}$$

The next step in defining the set of orthogonal functions e is to define the *covariance matrix* C_{ij}

If we assume that the time mean of S vanishes at each grid point, that is:

$$\overline{S}_i = 0$$

where the overbar denotes the time mean, then the covariance matrix is just given as:

$$C_{ij} = \overline{S_i S_j}$$

This is a symmetric matrix, which has positive definite eigenvalues λ, each associated with an eigenvector **e**. The set of eigenvectors is orthogonal, and can be taken to be ortho-normal:

$$\sum_j C_{ij}\, e_j^{(\alpha)} = \lambda^{(\alpha)}\, e_i^{(\alpha)}$$

$$\sum_i e_i^{(\alpha)} e_i^{(\beta)} = 0 \quad if\ \alpha \neq \beta$$

$$\sum_i e_i^{(\alpha)} e_i^{(\alpha)} = 1 \quad for\ all\ \alpha$$

The vectors **e** are just like the unit vectors (**x**,**y**) or (**x'**,**y'**) we introduced earlier, but they may have a higher dimension. In fact their dimension is N, the number of grid points.

We use the ortho-normal properties of the eigenvectors **e** to derive an expression for the coefficients P:

$$S_i = \sum_\beta P^{(\beta)} e_i^{(\beta)}$$

$$\sum_i e_i^{(\alpha)} S_i = \sum_\beta P^{(\beta)} \sum_i e_i^{(\alpha)} e_i^{(\beta)} = P^{(\alpha)}$$

The coefficient P for any "mode" β is just the dot product of the original vector S (representing an atmospheric field) with the corresponding "unit vector" **e**

The coefficients P are the *Principal Components*. *They are the new coordinates*.

The vectors (patterns) **e** are the *Empirical Orthogonal Functions (EOFs). They are the new unit vectors*.

The N diagonal elements of the matrix C_{ij} are just given by:

$$C_{11}, C_{22}, C_{33} \cdots or$$

$$\overline{S_1 S_1}, \overline{S_2 S_2}, \overline{S_3 S_3},$$

These diagonal elements clearly give the temporal variance at each grid point.

Thus the sum of the temporal variance over all grid points, or V ("the variance"), is thus the Trace of the matrix, which by a linear algebra theorem is equal to the sum of the eigenvalues:

$$V = \sum_{\alpha} \lambda^{(\alpha)}$$

$$V = \sum_i \overline{S_i S_i} = \sum_i \overline{\sum_\alpha P^{(\alpha)} e_i^{(\alpha)} \sum_\beta P^{(\beta)} e_i^{(\beta)}} =$$

$$\sum_\alpha \sum_\beta \overline{P^{(\alpha)} P^{(\beta)}} \sum_i e_i^{(\alpha)} e_i^{(\beta)} = \sum_\alpha \overline{P^{(\alpha)} P^{(\alpha)}}$$

From these two results for the variance V, we see that it can be written as a sum of terms, each term corresponding to one of the "modes" α. We thus say that the EOFs "explain" variance, one mode at a time.

If we order the EOFs and PCs so that the mode (1) has the largest eigenvalue, the mode (2) has the next largest eigenvalue, and so on, it becomes clear that a *truncated* representation of the field (keeping just a small number M of "modes") may still capture much of the overall variability of the field S.

Many Atmospheric Fields are Dominated by Large Scales (examples: Geopotential Height, Temperature, Winds)

Most of the day-to-day variability of maps of height, temperature, winds, can be captured by a relatively few EOF basis functions

Specifically, M ~ 10 components can capture most of the variability of a field with N >> M components!

EOFs have been designed to be the most efficient way to capture the temporal variance using as few basis functions as possible - M dimensions are important in some sense, not N.

Homework Problem

Prove that any two distinct PC's are linearly independent:

$$\overline{P^{(\alpha)} P^{(\beta)}} = 0 \; if \; \alpha \neq \beta$$

Hint: First use the expansion: $P^{(\alpha)} = \sum_i e_i^{(\alpha)} S_i$

for both PCs, (although you need to use a different dummy index for the two summations), and then use the definition of e as the eigenvector of the covariance matrix.

Empirical Orthogonal Function Analysis

(Principal Component Analysis)

(1) A Geometric and Intuitive Introduction

(2) Some related mathematics: Maximizing Variance

(3) Motivation for EOFs from Statistical Forecasting

(4) Use of EOFs as a Diagnostic / Analysis Tool

One of the motivations for the use of Empirical Orthogonal Functions in Meterology was in the context of statistical forecasting.

When choosing a statistical model, one of the big challenges is to pick as few *predictors* as possible which are good in explaining much of the past data (lowest mean squared error) as possible.

EOF analysis provides a rational method for doing this.

Simple example: You want to predict the surface temperature at one station for each day, based on the regional map of 850 hPa temperature and sea-level pressures on the previous day. This could lead to a huge number of predictors, which might fit the training data very well, but lead to lousy forecasts! On the other hand, using say the first 2 PCs of each of the two fields as predictors really reduces the number of predictors considerably.

Another Example: If one uses leading two PCs of a field as the *predictands*, then one automatically gets an entire field predicted as $P^{(1)}E_i^{(1)} + P^{(2)}E_i^{(2)} + \ldots$

"Downscaling" Example from ECMWF: Indian Monsoon JJAS Rainfall

(Franco Molteni's Lecture)

-You want to forecast pattern of rainfall over continental India the upcoming Monsoon seasons.

- ECMWF Seasonal Forecasts do a poor job of rainfall prediction and variability over the land, but a much better job over the adjacent oceanic regions. How to make use of this?

- Compute EOFs and PCs of *observed* rainfall from past data over a wide region encompassing India land points and adjacent oceanic regions. Then use the fact that the EOFs provide a **perfectly good ortho-normal basis set** to project the Forecast rainfall on this basis set. This means computing the coefficients C:

$$C^{(\alpha)} = \sum_i E_i^{(\alpha)} F_i$$

for a few α, where the EOFs E are observed, and F is the forecast field. The coefficients C are not the PCs, but they are express the forecast rainfall field over the wide region in a perfectly good basis set.

-These coefficients may be considered a forecast of the PCs: the fact that the sum over grid points i includes the oceanic region (where the forecast model is good) may help to make this a decent forecast of the PCs.

- Now you reconstruct the rainfall pattern over the region using the observed EOFs using your estimated PCs:

$$G_i = \sum_i E_i^{(\alpha)} C^{(\alpha)}$$

where G is your new (or "downscaled") prediction of the grid point rainfall over the wide region.

Empirical Orthogonal Function Analysis

(Principal Component Analysis)

(1) A Geometric and Intuitive Introduction

(2) Some related mathematics: Maximizing Variance

(3) Motivation for EOFs from Statistical Forecasting

(4) Use of EOFs as a Diagnostic / Analysis Tool

Several ways in which EOFs can help us understand weather and climate variability:

(1) As a tool for understanding - interpreting individual patterns as having a physical / dynamical meaning

(2) EOF patterns provide a set of unit vectors (or basis functions) to evaluate the same field provided by another source - this can be very useful for statistical downscaling (eg Dr. Molteni's example for ECMWF seasonal forecasts)

(3) As a way to reduce the dimension of our fields (drastically) for use in further analysis, such as Singular Value Decomposition or Canonical Correlation Analysis

The Preferred Structure of the Interannual Indian Monsoon Variability

DAVID M. STRAUS[1] and V. KRISHNAMURTHY[2]

Abstract—The leading empirical orthogonal function (EOF) of the June-Sept. mean, rotational horizontal wind at 850 hPa and 200 hPa (over the region 12.5°S–42.5°N, 50°E–100°E) from 56 years (1948–2003) of reanalysis (from the National Centers for Environmental Prediction) shows strong anti-cyclonic circulation at upper levels, strong Indian Ocean cross-equatorial flow and on-shore flow over western India at lower levels . The associated principal component (PC) is correlated at the 0.75 level with the seasonal mean observed Indian Monsoon rainfall (IMR). Composite differences of vertically integrated divergence

Table 1

EOF Variance. The space-time variance explained by the leading EOFs, in percent terms. The range of explained variance includes the uncertainty as estimated from NORTH et al. (1981). The correlation of the corresponding PC with the IMR is given in the last column. Values significant at the 5% level are bold faced

	Explained Variance (%)	Range of Exp. Var (%)	Correlation with IMR
EOF-1	25.2	30–20	**0.75**
EOF-2	15.3	18–12	−0.07
EOF-3	10.8	13–9	0.15
EOF-4	6.6	8–5	0.14

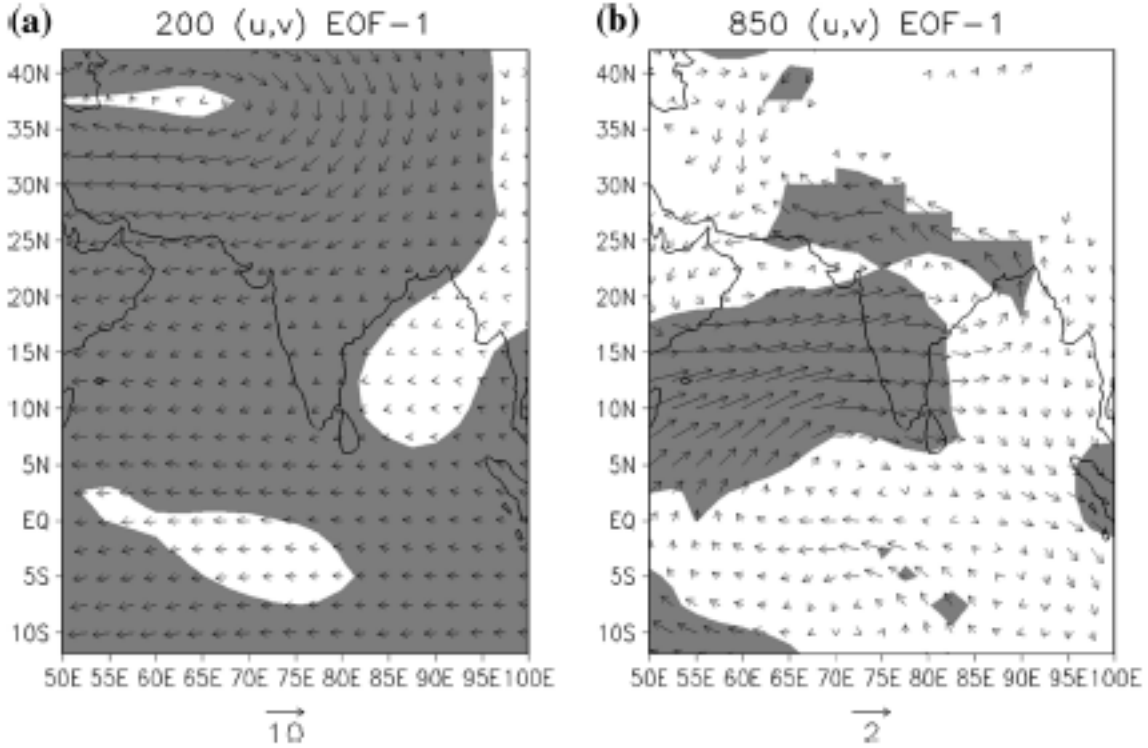
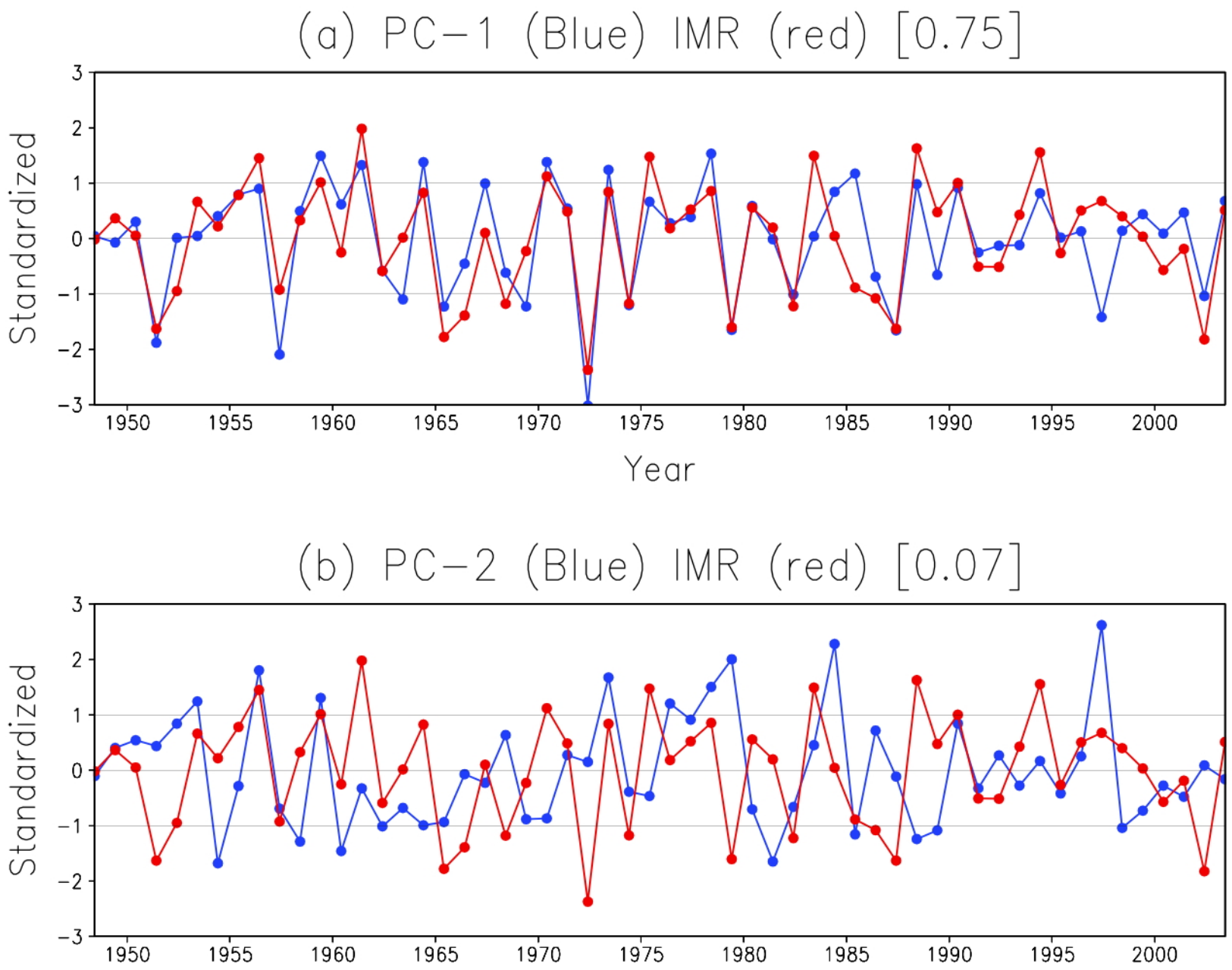


Figure 1
Composite differences of full horizontal winds for 200 hPa (a) and 850 hPa (b). Composite differences are defined as the difference between the mean of all (7) years for which PC-1 > one standard deviation (σ) and the mean of all (12) years for which PC-1 $< -\sigma$. Shading denotes 10% significance using a two-sided t-test. Reference vectors show the magnitude (in m/sec) in each case.


(a) PC−1 (Blue) IMR (red) [0.75]
Standardized
3
2
1
0
−1
−2
−3
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
Year
(b) PC−2 (Blue) IMR (red) [0.07]
Standardized
3
2
1
0
−1
−2
−3
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000

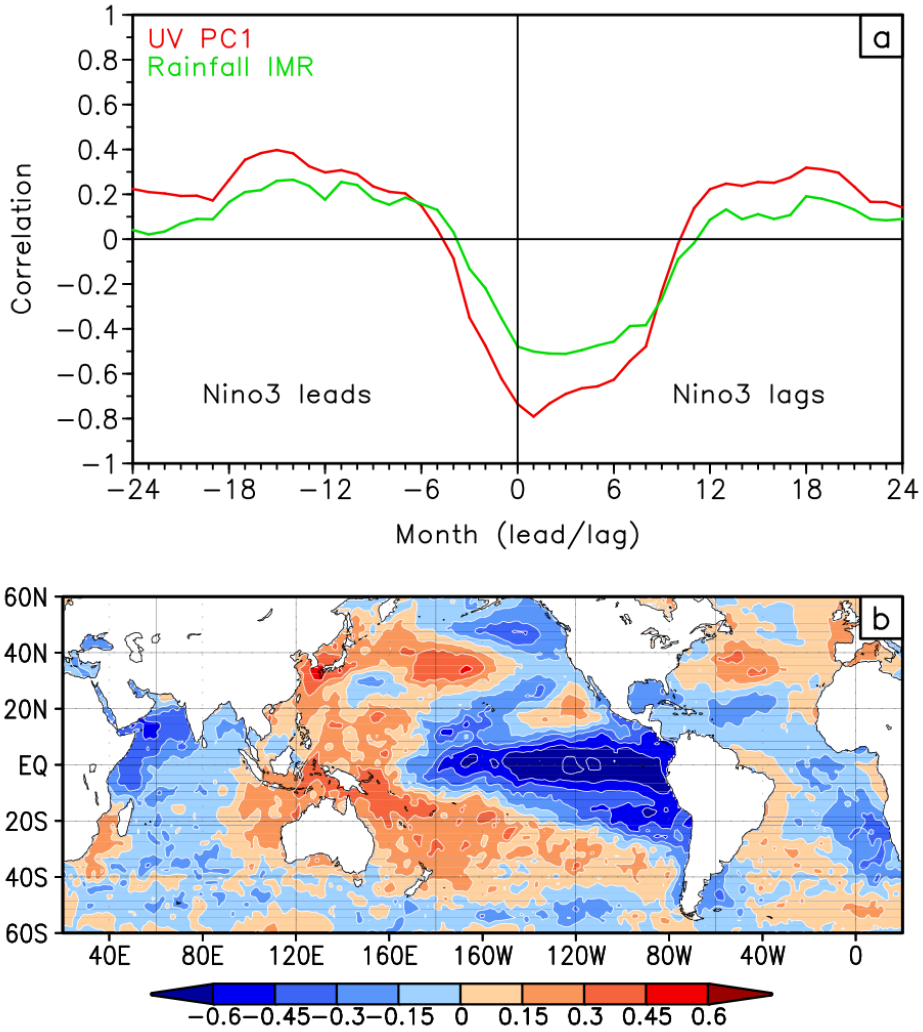


Fig. 6. (a) Lag correlation of monthly anomalies of NINO3 index with PC-1 (red) and JJAS seasonal anomaly of IMR index (green) for the period 1948-2003. (b) Point correlation of PC-1 with JJAS seasonal anomaly of SST.

Sampling Errors in the Estimation of Empirical Orthogonal Functions

GERALD R. NORTH, THOMAS L. BELL AND ROBERT F. CAHALAN

(Monthly Weather Review, July 1982)

ABSTRACT

Empirical Orthogonal Functions (EOF's), eigenvectors of the spatial cross-covariance matrix of a meteorological field, are reviewed with special attention given to the necessary weighting factors for gridded data and the sampling errors incurred when too small a sample is available. The geographical shape of an EOF shows large intersample variability when its associated eigenvalue is "close" to a neighboring one. A rule of thumb indicating when an EOF is likely to be subject to large sampling fluctuations is presented.

$$\delta\lambda_\alpha \equiv \epsilon l_\alpha^{(1)} \approx \lambda_\alpha (2/N)^{1/2}, \qquad (24)$$

Interpretation: Uncertainty in the eigenvalue is proportional to the eigenvalue itself. N = number of realizations, that is the number of maps going into the analysis.

Modes whose eigenvalues are separated by ***less*** than this uncertainty are ***not*** statistically distinct. The associated eigenvectors (unit vectors) are not really distinct - *any linear combination of the two is equally good!*

Statistical Significance vs. Physical Meaning

In this case, λ_2 is well enough separated from λ_1 and λ_3 by the North et al standard for mode 2 to be considered well separated from mode 1 and mode 3. Yet there is no clear physical interpretation for mode 2. *Statistical significance does not automatically confer physical meaning*.

"Degenerate Modes"

If two eigenvalues λ_α and λ_β are not well separated, the two associated modes are said to be *degenerate*. This means that the orthogonal directions represented by the two unit vectors $\mathbf{e}^{(1)}$ and $\mathbf{e}^{(2)}$ are not unique - any set of orthogonal directions in the plane defined by the two unit vectors is equally meaningful - this means that any linear combination of the two associated patterns is equally meaningful

Important Practical Tips

Area Weighting

Remembering that the "variance" V that is maximized is just the sum over all grid points of the temporal variance:

$$V = \sum_i \overline{S_i S_i}$$

This is not strictly speaking correct, since we would like the grid points to be area weighted. (An EOF on a regular lat-lon grid which includes high latitude regions will be very distorted because of this problem.) The fix is to either use an equal area grid (cumbersome in practice), or, for a regular lat-lon grid, to pre-weight all fields by the square root of the cosine of latitude before performing the analysis. This means defining a new set of fields: $\widehat{S}_i = \sqrt{\cos(\phi_i)}\, S_i$

so that the variance used becomes: $V = \sum_i \cos(\phi_i) \overline{S_i S_i}$

Compute the EOFs and PCs consistently using the new (weighted) fields, and then convert the EOFs back to the original grid point form by dividing out the weight at the end of the calculation: $S_i = \widehat{S}_i / \sqrt{\cos(\phi_i)}$

Units

Note that since the eigenvectors (EOFs) $\mathbf{e}_i^{(\alpha)}$ are normalized, they are dimensionless, and the units of the original field are assigned to the PCs. Sometimes it is convenient to have dimensionless PCs (in fact PCs which are standardized - that is have unit variance), and have the dimensions of the original fields associated with the EOFs or patterns. This is obtained by dividing each $P^{(\alpha)}$ by its standard deviation $\sigma^{(\alpha)}$, and multiplying each normal vector $\mathbf{e}_i^{(\alpha}$ by the same standard deviation:

$$S_i = \sum_{(\alpha)} P^{(\alpha)} e_i^{(\alpha)} = \sum_{(\alpha)} p^{(\alpha)} E_i^{(\alpha)}$$

$$p^{(\alpha)} = P^{(\alpha)} / \sigma^{(\alpha)}$$

$$E_i^{(\alpha)} = \sigma^{(\alpha)} e_i^{(\alpha)}$$