Chapter 5
Generalizations: Rotated, Complex, Extended and Combined EOF

5.1 Introduction

We have seen in the last section that the difficulty in identifying real physical patterns from EOF stems from their orthogonal nature. Orthogonality translates into the fact that typical patterns appear in secondary (higher order) EOF. Very often the first EOF has little structure, the second has a positive and a negative center, the third more centers and so on, in a way so as to maintain orthogonality.

In fact, the overall structure of the EOF is often determined by the geometrical shape of the domain chosen and different data on the same domain, with different covariance relations, may actually result in similar EOF. Moreover, in the preceding chapter we have seen the issue of sensitivity to partitioning the analysis domain into subdomains and we can interpret it as a case of the overall sensitivity to the domain shape. Another point is that EOF are obtained by trying to maximize the amount of total variance explained by a single mode. It is possible that the resulting optimized modes are difficult to interpret physically, either because the real relation is localized and the EOF are spreading it, creating artificial nonlocal relations, or because the EOF are so close in terms of eigenvalue separation, that the numerical techniques cannot really distinguish between them.

The risk of creating misleading or illusory representations of relations within the data is particularly troubling. The most common situation in which these malfunctioning can arise is when data represent localized variances. In this case the EOF will try to fit globally the domain under consideration, with as few modes as possible, generating first EOF (low order modes) with very large structures. For instance, it is possible that the first mode indicates anomalies all of the same sign, whereas the data do not indicate that there is ever a time when all stations were reporting such a one-sign pattern. This may be a fiction created by the EOF trying to maximize the variance explained. This is another way of realizing that the EOF have no way to guess the physical relations within the data.

5.2 Rotated EOF

The problems mentioned in the previous section do not mean that the EOFs must be ditched. They simply indicate that much care must be taken in producing and interpreting EOF and it is not simply a matter of using some canned routines out of a package.

Unfortunately we cannot produce a recipe for the intelligent use of EOF, but some techniques have been devised to make a misuse of EOF less probable. Rotation is one of them.

Simply put, we need to transform the EOF to another system of coordinates, exploiting the freedom to choose a different basis in the data space. We have seen from (4.3) that the EOF can be seen as the eigenfunctions of the covariance or correlation matrix, that is

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^*. \tag{5.1}$$

The basic indeterminacy in the EOF can be seen if we consider a similarity transformation that changes the matrix $\mathbf{S}$ into a different matrix. More precisely, let $\mathbf{T}$ be an $n \times n$ nonsingular matrix. Then we can write

$$\mathbf{G} = \mathbf{TST}^{-1}, \quad \text{that is} \quad \mathbf{S} = \mathbf{T}^{-1}\mathbf{GT}.$$

Substituting in (5.1) we obtain $\mathbf{T}^{-1}\mathbf{GT} = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^*$, and the eigenvalue decomposition of the non-Hermitian matrix $\mathbf{G}$ is

$$\mathbf{G} = (\mathbf{TU})\boldsymbol{\Sigma}^2(\mathbf{U}^*\mathbf{T}^{-1}). \tag{5.2}$$

The similarity of $\mathbf{G}$ and $\boldsymbol{\Sigma}^2$ (cf. Sect. 2.7) implies that the transformed matrix $\mathbf{G}$ has the same eigenvalues as $\mathbf{S}$. However, the eigenfunctions have been transformed according to

$$\mathbf{U_T} = \mathbf{TU}. \tag{5.3}$$

Owing to the non-orthogonality of $\mathbf{T}$, the columns of the new matrix $\mathbf{U_T}$ are correlated. In the derivation above, the aim is to exploit the freedom of choosing $\mathbf{T}$ so as to add some constraints to the EOF. The purpose is to alleviate some of the problems discussed earlier, especially the problem of the generation of patterns that cannot be reconciled with the expected physical relations in the data.

Whenever $\mathbf{T}$ is non-orthogonal, the transformed EOF represent an oblique coordinate system for the given data. Therefore, in this case, they are called *oblique* EOF, and each of them has a nonzero projection (correlation) on each other. On the other hand, if $\mathbf{T}$ is an orthogonal matrix, we can refer to it as a rotation matrix, $\mathbf{Q}$, and we write

$$\mathbf{U_Q} = \mathbf{QU}.$$

The choice of rotation is arbitrary, although we may want to require that some "nice" pattern feature be emphasized after rotation. In the most popular case, we
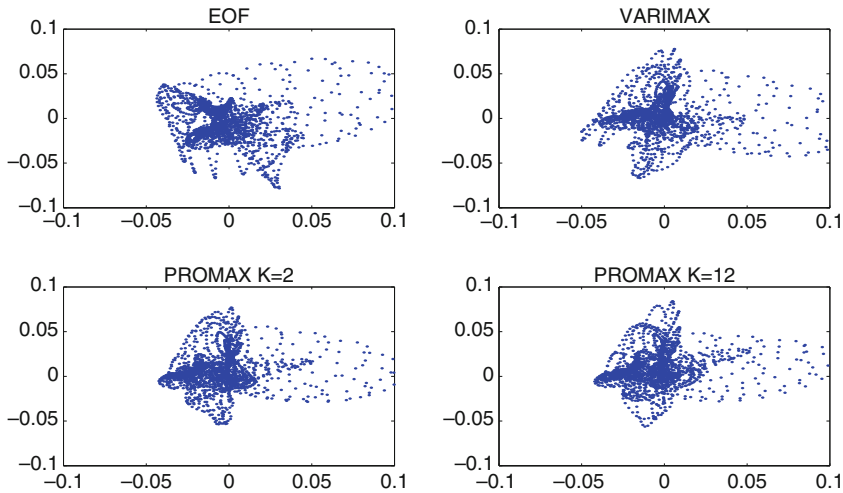
Fig. 5.1 Simple structure. The plots are scatterplot of the value of the EOF for each spatial points by pairs of modes, in this case EOF1 and EOF2. The standard EOF yield a situation in which there are several points where both modes have nonzero values (*top left panel, bottom right quadrant*). Rotated EOFs reduce such an effect, aligning the point values along the axis (*top right panel*), the maximum effect is obtained by relaxing the constraint of orthogonality with the oblique modes of PROMAX (*bottom panels*)

require that the spatial variance be concentrated in as few points as possible, to obtain *simpler* patterns. The definition of what simple is, is far from being straight-forward. The issue has been discussed at length in the specialized literature, but a rigorous definition of "simple" is elusive. A set of empirical principles to describe the properties of simple structures has been proposed, but the overall philosophy of the existing school of thought essentially reduces to trying to concentrate the co-efficients of the EOF in few modes, in such a way that for each variable, i.e. for each spatial point in our examples, only a very small number of EOF is needed to explain the variance. This is easily checked by producing scatter plots of the EOF modes two at a time. Figure 5.1 shows an example of such a plot, shown here for the case of the marine SST in our test dataset. The EOF are not well separated in the sense that there are several points for which both EOF1 and EOF2 have nonzero elements, as it can be seen by the alignment of the points along the diagonal of the third quadrant in the top left panel of Fig. 5.1. These points are positions where both EOF1 and EOF2 are needed to describe the variance in that point. Ideally, we would like to separate as much as possible the variance in such a way that distinct EOF describe most of the variance at separate points. This means that different EOF have to reach large values in different places, so that if EOF1 has large values in certain points then the higher modes must have small values in those same points. The top left panel of Fig. 5.1 shows that the EOF do not guarantee this property: there is a whole class of points where both EOF1 and EOF2 have important amplitudes.

Orthogonal Rotations

We can exploit the arbitrariness in the coordinate system definition to try to transform the original EOF to another coordinate system that yields a better separation of the EOF for each spatial points, i.e. simple structure. In principle the transformation is quite unrestricted, but orthogonality is a desirable quantity if we want to separate the variance of the field under examination.

The following panel (top right) shows a popular method to introduce some simple structure and maintain orthogonality. This result can be achieved by requiring that **Q** is such that the new pattern minimizes a functional, sometimes called the *simplicity functional*, that provides a distribution measure of spatial variance. The choice of the functional is very important and there are no rules to prescribe it, but a very popular choice is the VARIMAX method, where the functionals are chosen so that the rotated patterns maximize the functional

$$F_R(u_1, u_2, \ldots, u_r) = \sum_{k=1}^{r} f(u_k), \text{ with } \quad f(u_k) = \frac{1}{n} \sum_{i=1}^{n} p_{ik}^4 - \frac{1}{n^2} \left(\sum_{i=1}^{n} (p_{ik}^2) \right)^2,$$

or with

$$f(u_k) = n \sum_{i=1}^{n} \left(\frac{p_{ik}}{h_i} \right)^4 - \left(\sum_{i=1}^{n} \left(\frac{p_{ik}}{h_i} \right)^2 \right)^2,$$

where the p_{ik} are the grid-point values of the kth EOF patterns (u_k) that we are trying to rotate, the h_i are the point-by-point standard deviations (communalities) of the r patterns we are rotating; see, e.g., Harman (1976). Both forms aim at maximizing the spatial variance of the EOF modes by concentrating the point values toward zero or one: the first example is known as the *raw VARIMAX* and the second one as the *normal VARIMAX*. Below is a possible simple Matlab implementation of the normal VARIMAX procedure, closely following Harman (1976).

```
function [coef,u,varrotated]=eofrot(z,ind,index)
%
% Algorithm for eof rotation with the varimax method
% Inputs:
%         z                      Data Matrix
%         ind                    Index for domain
%         index                  indeces of EOFs to be rotated
% Outputs:
%         coef                   Cofficient for the rotated EOFs
%         u                      Rotated Eof in ascending order
%         varrotated             Variance explained by rotated EOFs

[npoints,ntime]=size(z);          %  Time and space points
[uneof,ss,vneof]=svd(z,0);        %  Unrotated EOF for variance
totvar = sum(diag(ss.^2));        %  calculation
```

```
 lding = uneof(:,index);
 sl    = diag(ss(index,index).^2);
 varexpl = sum(sl)/totvar;          % Relative variance explained
                                    % by the unrotated modes
 [n,nf]=size(lding);
 b=lding;
 hl=lding*(diag(sl));

 hjsq=diag(hl*hl');
 hj=sqrt(hjsq);                     %  Normalize by the communalities
 bh=lding./(hj*ones(1,nf));

 Vtemp=n*sum(sum(bh.^4))-sum(sum(bh.^2).^2); % VARIMAX functional
                                             % to be minimized
 V0=Vtemp;
 for it=1:10;                       % Number of iterations
   for i=1:nf-1;         %  Program cycles through 2 factors
     for j=i+1:nf;
       xj=lding(:,i)./hj;           % notation here closely
       yj=lding(:,j)./hj;           % follows harman
       uj=xj.*xj-yj.*yj;
       vj=2*xj.*yj;
       A=sum(uj);   B=sum(vj); C=uj'*uj-vj'*vj;   D=2*uj'*vj;
       num=D-2*A*B/n;           den=C-(A^2-B^2)/n;
       tan4p=num/den;      phi=atan2(num,den)/4; angle=phi*180/pi;
       if abs(phi)> eps;
        Xj=cos(phi)*xj+sin(phi)*yj; Yj=-sin(phi)*xj+cos(phi)*yj;
        bj1=Xj.*hj;                      bj2=Yj.*hj;
        b(:,i)=bj1;                      b(:,j)=bj2;
        lding(:,i)=b(:,i);               lding(:,j)=b(:,j);
       end
     end
   end;
   lding=b;          bh=lding./(hj*ones(1,nf));
   Vtemp=n*sum(sum(bh.^4))-sum(sum(bh.^2).^2);% Update functional
   V=Vtemp;
   if abs(V-V0)<.0001;break;else V0=V;end;
 end;

 for i = 1:nf           % Reflect vectors with negative sums
     if sum(lding(:,i)) < 0
         lding(:,i) = -lding(:,i);
     end
 end
 Arot=lding ;                       % rotated eof
 coef=z'*Arot(:,1:nf);              % time series for rotated eof

 for i=1:nf
    varex(i) = sum(var(coef(:,i)*Arot(:,i)')*(ntime-1));
 end

 varexplrot = sum(varex)/totvar;   zvar=sum(var(z')*(ntime-1));
 [varex,I]=sort(varex);   % Sort in decreasing order of variance
 Arot=Arot(:,I);   Arot = fliplr(Arot); varex = flipud(varex');
```

```
varunrotated = sl/totvar;      varrotated   = varex/totvar;
u=zeros([96*48 nf]);
u(ind,1:nf) = Arot(:,1:nf); % Rotated EOF in mapping formats

end
```

The previous function performs a rotation using orthogonal rotations; the sister function `eofpromax` uses oblique rotations to maximize the spatial variance and can be found in the book Website. The Matlab Statistics Toolbox ((matlab7)) includes a few routines to compute the EOFs (Principal components) and the rotated factors; see Exercise 1 below.

The pictures in Figs. 5.2 and 5.3 show the difference between unrotated and rotated EOF, in this case after a normal VARIMAX rotation has been used. The rotation has been applied to the first ten modes. We can see how the rotation tends to separate the original EOF in a spatial sense. The first unrotated mode (top panel, Fig. 5.2), for instance, is composed of centers of activity, i.e. relative maxima and minima for the patterns of the EOF, that are distributed across the North American continent and the North Atlantic extending well into the European continent.

The rotated equivalent (top panel, Fig. 5.3) shows the emergence of a pattern that is more confined to the North American sector, with small or no amplitude elsewhere. The variation over Europe and Asia is picked up by the higher modes, represented here by modes 3 and 10, that instead tend to accumulate amplitude over the regions where there is little or no amplitude for mode 1. The separation is not perfect, as it can be noticed that mode 3 still has some amplitude in the central Pacific, in correspondence of the centers of mode 1. The effect is larger on the higher modes, and the rotated mode 10 is now more concentrated over Asia, showing a clear pattern from India to the Mediterranean. It is not possible to give a general rule on when rotation is necessary. It is found that when the EOF modes are very close together, i.e. the separation in the eigenvalues is not great, then rotation can disentangle the modes in the previous case between the Pacific and Atlantic modes.

The rotated modes can still be used to decompose the variance, in the sense that each of them explains a certain portion of the variance that can be attributed only to that mode, since the rotated EOF are still mutually orthogonal. The rotated EOF can then be ranked in order of percentage of explained variance.

The issue of rotation is still not widely accepted. Some investigators think that rotation should become the standard and therefore recommend to rotate all modes before attempting an interpretation, others are less convinced especially because of the ad hoc choices of the simplicity functional. In general, rotated EOF are more stable than the conventional vectors since they introduce another constraint that can be used to distinguish between eigenvectors. The well separated rotated EOF are therefore more resilient and then show less sensitivity to the errors that we have discussed in the previous chapters.
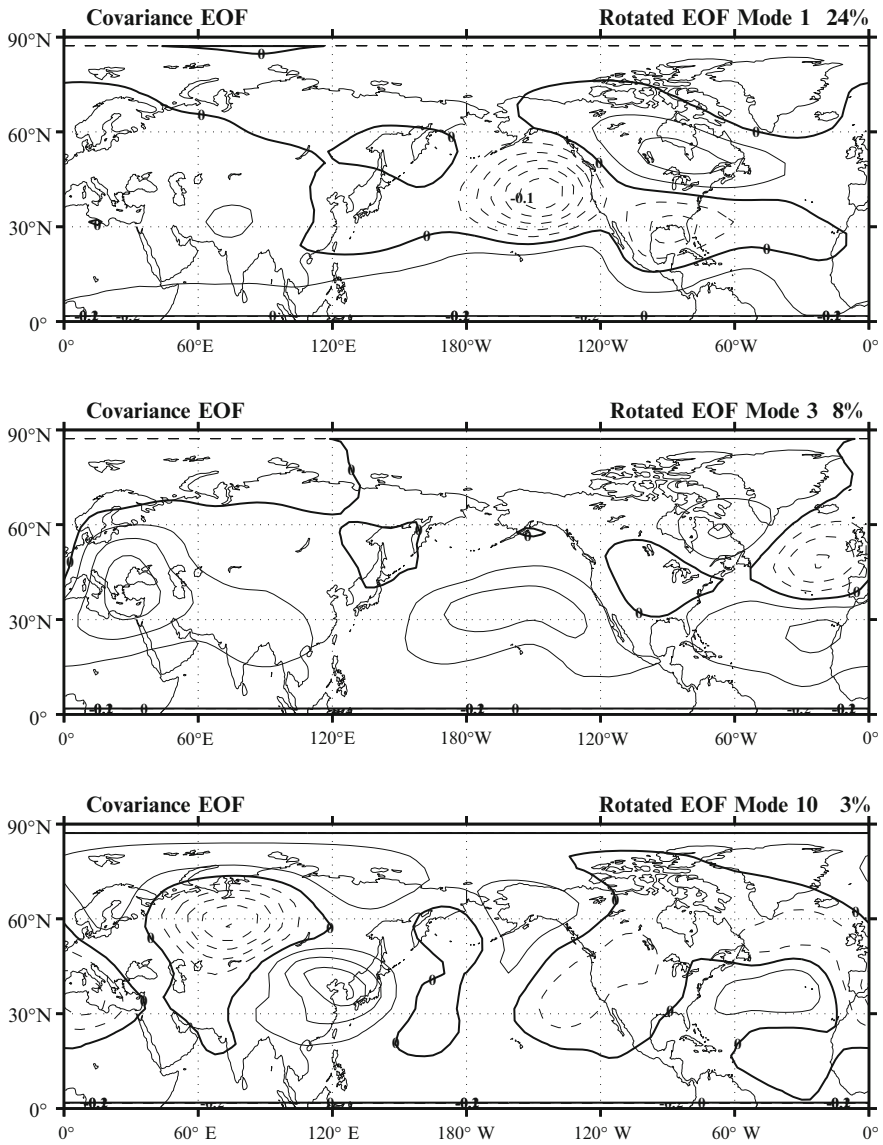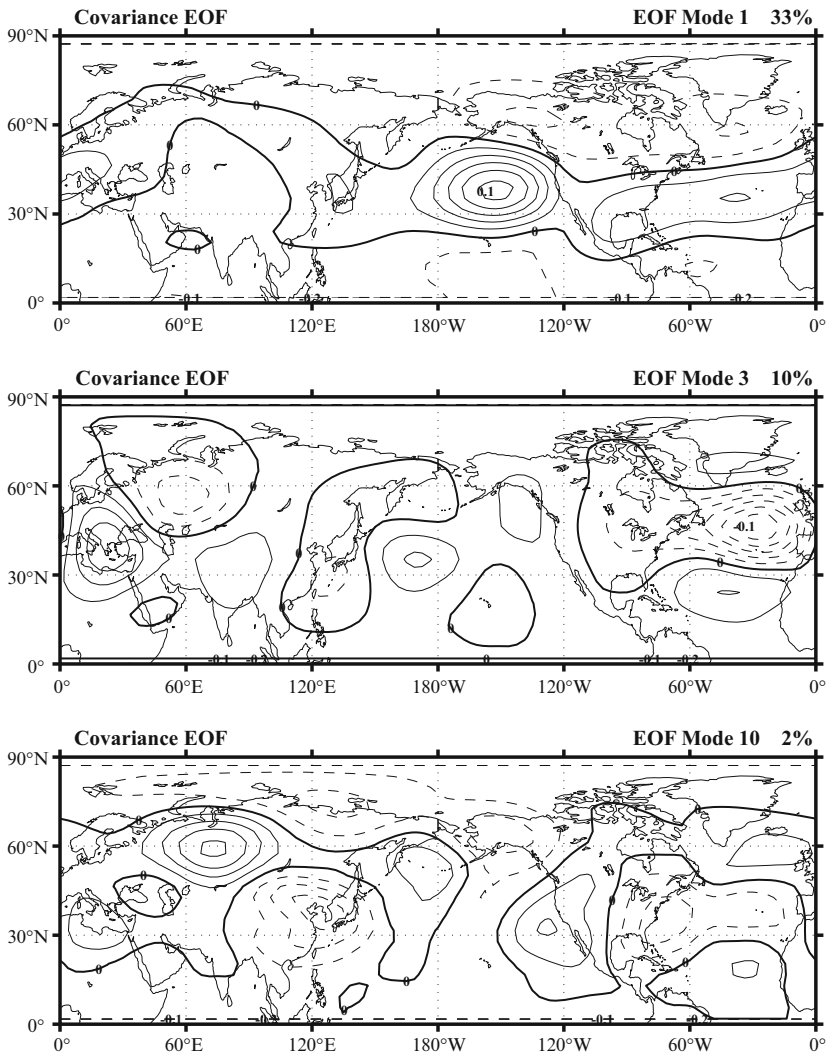
Covariance EOF   **Rotated EOF Mode 1   24%**

Covariance EOF   **Rotated EOF Mode 3   8%**

Covariance EOF   **Rotated EOF Mode 10   3%**

Fig. 5.2 Conventional EOF for the test data sets Z500

Fig. 5.3 Rotated EOF according to the normal VARIMAX method for the test data sets Z500. Also shown is the variance explained by the rotated mode

Exercises and Problems

1. Given the set of data

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & -1 & 1 \\ -1 & 1 & 0 & 1 & 1 \\ 0 & 0 & -1 & 0 & -1 \\ -1 & 1 & 1 & 1 & 1 \\ -2 & 1 & 1 & 1 & 1 \end{pmatrix},$$

   compute the first two EOF and the rotated EOF with VARIMAX (use the MAT-LAB functions `princomp` and `rotatefactors`).

   The command L = princomp(X); *yields*

$$\mathbf{L} = \begin{pmatrix} -0.6756 & -0.4171 & 0.6047 & -0.0192 & 0.0600 \\ 0.2740 & 0.1259 & 0.4752 & 0.0060 & -0.8266 \\ 0.3931 & -0.5676 & 0.0620 & 0.7157 & 0.0847 \\ 0.3233 & 0.4582 & 0.6261 & 0.0680 & 0.5374 \\ 0.4577 & -0.5273 & 0.1125 & -0.6948 & 0.1310 \end{pmatrix},$$

   and the subsequent command [L1,T]=rotatefactors(L(:,1:2)); *gives*

$$\mathbf{L}_1 = \begin{pmatrix} -0.7908 & 0.0714 \\ 0.2948 & -0.0636 \\ -0.0259 & -0.6900 \\ 0.5335 & 0.1728 \\ 0.0500 & -0.6964 \end{pmatrix}.$$

   After rotation, it is possible to better decompose the data variance among the two principal components: the first data column and also somehow the forth column, are well represented by the first component. On the other hand, the third and fifth data columns are well represented by the second (rotated) principal component.

Non-orthogonal Rotations

The main conceptual difficulty with rotations is the fact that again we are forcing a condition on the data that we do not know whether it is reasonable to enforce. On the other hand the freedom of changing coordinate system includes transformations of type (5.3) that are not orthogonal, therefore we can ask whether it is possible to use a modal decomposition that does not require orthogonality from the start. By

removing the orthogonality constraint, we are left with a large selection of possible transformations.

The method aims at identifying a transformation of a preliminary standard EOF pattern to achieve simpler structure. The transformation matrix is obtained by solving an oblique Procrustes problem. This mathematical problem can be stated as follows: Given matrices $\mathbf{A}$ and $\mathbf{B}$ of size $n \times m$ with $\mathbf{A}$ having full column rank, find a matrix $\mathbf{T}$ satisfying

$$\mathbf{B} = \mathbf{AT} + \mathbf{E}$$

such that the Frobenius norm of the error matrix $\mathbf{E}$,

$$\|\mathbf{E}\|_F^2 = \text{trace}((\mathbf{B} - \mathbf{AT})^*(\mathbf{B} - \mathbf{AT})), \tag{5.4}$$

is minimized. $\mathbf{B}$ is often called the *target* matrix. The matrix $\mathbf{T}$ can be found as the only critical point of the function to be minimized, that is, as the solution of

$$\frac{\partial}{\partial \mathbf{T}}(\|\mathbf{E}\|_F^2) = -2\mathbf{A}^*\mathbf{B} + 2\mathbf{A}^*\mathbf{AT} = 0.$$

Solving for $\mathbf{T}$ yields

$$\mathbf{T} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{B}.$$

The interpretation of the problem is relatively simple. The successful solution of the Procrustes problem is the identification of a linear relation between two sets of data. In case $\mathbf{A}$ is not full column rank, $\mathbf{A}^*\mathbf{A}$ is singular and $\mathbf{T}$ cannot be determined as outlined above. However, a (non-unique) minimizing solution can always be obtained by recurring to the pseudoinverse of $\mathbf{A}^*\mathbf{A}$ (cf. end of Sect. 2.8).

The PROMAX method uses the Procrustes problem to obtain a simple structure solution. The basic idea is to create a "simple" target matrix and then use a Procrustes transformation to obtain an oblique set of modes that have a more insightful structure than the original modes. The observation that orthogonally rotated modes, such as those obtained by VARIMAX, are usually a good deal simple themselves suggests that the VARIMAX modes can be used as starting point. Therefore, each element of the target matrix $\mathbf{B}$ can be defined as

$$b_{ij} = \frac{|v_{ij}|^k}{v_{ij}}, \tag{5.5}$$

where v_{ij} are the VARIMAX pattern values in each spatial point, previously normalized. The Procrustes problem is then formulated with $\mathbf{B}$ as target matrix and with $\mathbf{V}$ the VARIMAX pattern matrix as the data matrix

$$\mathbf{B} = \mathbf{VT} + \mathbf{E},$$

with solution

$$\mathbf{T} = (\mathbf{V}^*\mathbf{V})^{-1}\mathbf{V}^*\mathbf{B}.$$

The oblique patterns are then given by

$$\mathbf{V}_{\text{promax}} = \mathbf{VTD},$$

where the matrix $\mathbf{D}$ scales the oblique modes to unit length, namely

$$\mathbf{D}^2 = \text{diag}(\mathbf{T}^*\mathbf{T})^{-1}.$$

The definition of the target matrix as a power of the original pattern (cf. the exponent k in (5.5)) is an attempt to emphasize the differences between maxima and minima, to obtain a simpler structure in which intermediate values are unfavored. The value of the parameter k is arbitrary, but there is a difference in the sensitivity of the modes according to the shape of the sought after real pattern. If we expect a strong pattern with large variations between its extreme values, then k should be set to a low number. In practice, $k = 2$ or $k = 4$ are often used.

The comparison between the standard and rotated modes is shown in Fig. 5.4 for the first mode in the test data set for Z500. The orthogonal VARIMAX rotation results in an intense pattern, better localized, as we have seen in the preceding pictures. The PROMAX solution (lower panels) obtains patterns even more localized on North America, but we can notice one of the problems with PROMAX, especially if a large value of k is selected (bottom panel is for $k = 12$). The construction of these modes tends to polarize the spatial variability, concentrating the variance in smaller regions. The modes have fewer peaks, but of larger amplitude. We can see that for $k = 12$ the centers are more intense, even in regions where the EOF or the VARIMAX showed little amplitude. This example emphasizes that simple structure in principle that does not necessarily imply more meaningful modes. Figure 5.1 shows that from the point of view of simple structure, i.e. the polarization and separation of the pattern values in space, we are getting better every time. The concept of simple structure is therefore a very useful concept, but it cannot be considered as the only guiding principle.

Oblique modes have not found a widespread usage in data analysis, perhaps because of the parametric freedom, but also because they cannot be used to separate the variance.

5.3 Complex EOF

We have seen how conventional and rotated EOF can be employed to identify patterns that optimize the explanation of the variance. EOF identify the dominant pattern, but the information on the time evolution is only implicitly included into the evolution of the coefficients. Data that contain oscillations in time or in space and time as a propagating signal, are very common in applications. In Sect. 4.5.1 we have seen an example in which the standard EOF have been applied to an ideal example of a propagating wave. The signature of the propagation is visible in the
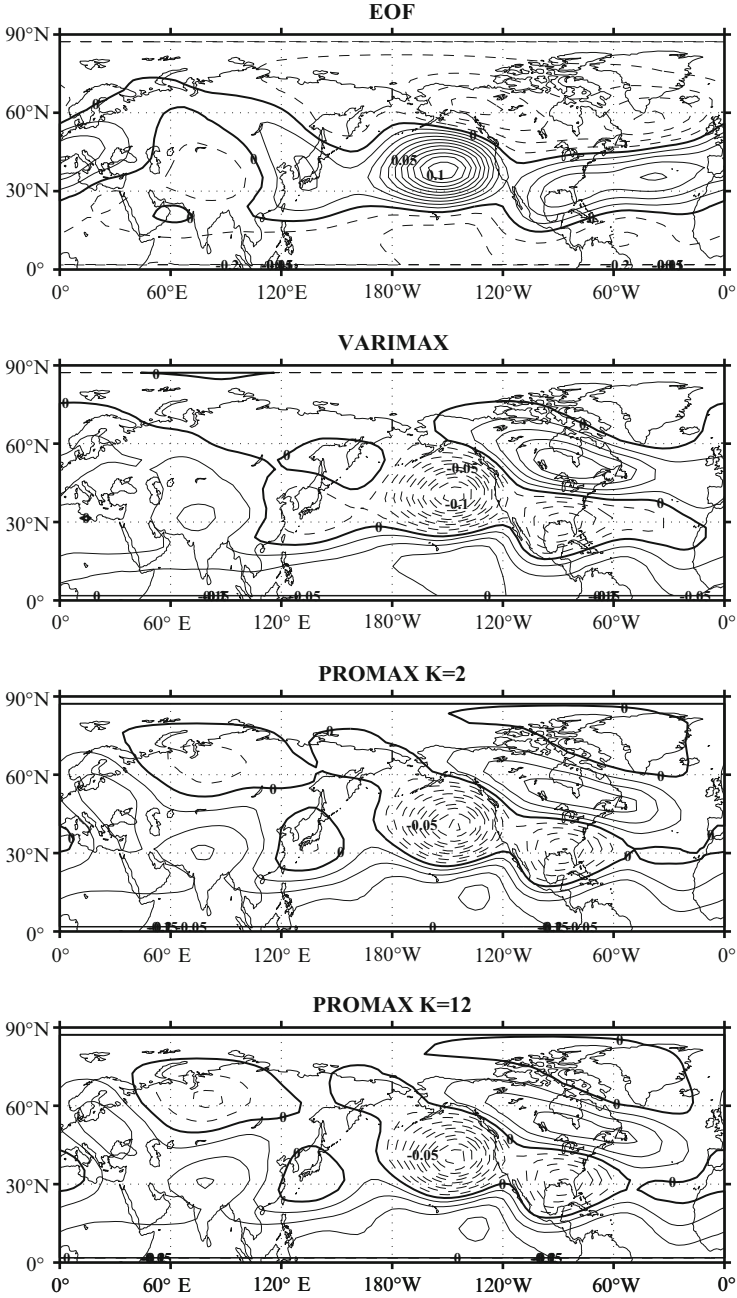
Fig. 5.4 Conventional, rotated and PROMAX (*oblique*) EOF for the test data sets Z500

EOF, but it requires some indirect interpretation. The presence of propagation is indicated by two modes whose patterns are in quadrature, namely the relative maxima and minima of one pattern correspond to the zero lines of the other and the two EOF explain a similar amount of variance (see Fig. 4.13).

The variations of the coefficients in time (top panel of Fig. 5.5) show a periodic behavior in time. There is a shift in time corresponding to a quarter of a wavelength between EOF1 and EOF2. A quarter wavelength shift in time is the phase lag typical of a harmonic wave of the form

$$V(\mathbf{x}, t) = \Re[U(\mathbf{x})e^{-i\omega t}] = \Re[U(\mathbf{x})(\cos(\omega t) + \iota \sin(\omega t))]. \qquad (5.6)$$

Therefore, the variation in time of the EOF coefficient seems to identify a kind of variability that can be expressed as a harmonic wave with real part EOF1 and imaginary part EOF2. The EOF analysis has been able to find couples of modes that are strongly linked, in fact they may be part of the same physical system.

Waves are a pervasive physical phenomenon so it is not surprising that the EOF's feature of detecting propagating modes has raised considerable interest. On the other hand, it is also true that this capability is a sort of byproduct of the general property of EOF to maximize variance. Would it be possible to sharpen the EOF definition so as to go after propagating modes?
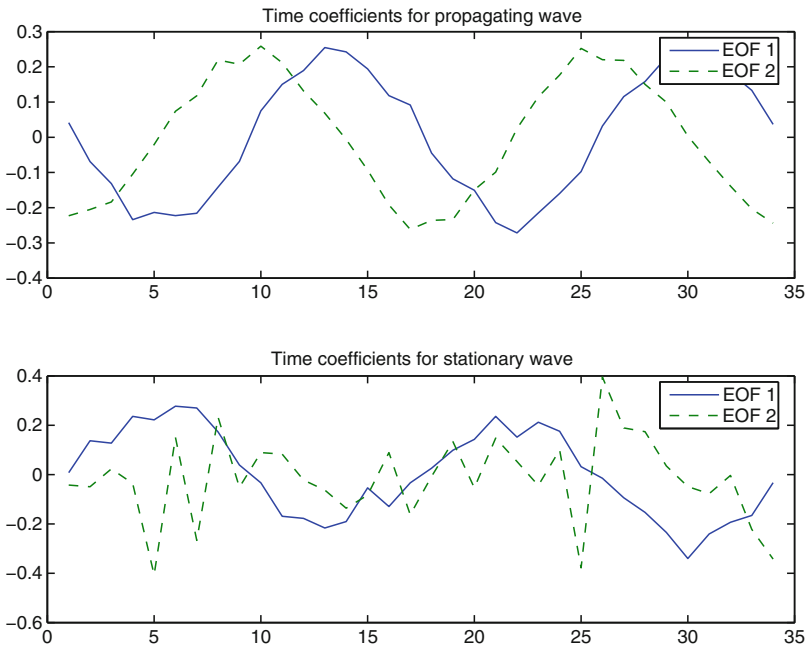
Fig. 5.5 EOF coefficients of the example in Sect. 4.5. *Top panel*: a propagating wave. *Bottom panel*: a stationary wave

We have seen that the quarter wavelength shift is a peculiar phase relation that indicates propagation. Can we find a way to enhance the modes that are in that particular phase relation? One possibility is to change the available data to stress the phase relation we are looking for; in our case we can expand the data by adding a new data set obtained by shifting all data by one quarter wavelength. This is a mathematical procedure that can be performed by *Hilbert transform*. The analytical definition of the transform is

$$\hat{f} = H[f(x,t)] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{s(\tau)}{t - \tau} d\tau$$

where the integral is to be understood to be a Cauchy principal value to avoid the singularities at infinity and at $t = \tau$. In practice, the transform of discrete signal is performed using a discrete Fourier transform (Hahn 1996)

$$\hat{f} = H[f(x,t)] = \sum_{\omega} f_H(x,\omega)e^{-2\pi i \omega t}, \quad f_H(x,\omega) = \begin{cases} i g(x,\omega) & \text{for } \omega > 0 \\ 0 & \text{for } \omega = 0 \\ -i g(x,\omega) & \text{for } \omega < 0. \end{cases}$$

where $g(\omega)$ is the discrete Fourier transform of f. The Hilbert transform shifts the data series a quarter period to obtain a new, augmented, data series of complex data,

$$\mathbf{X}_C = \mathbf{X} + i H(\mathbf{X}),$$

where the real part contains the original data and the imaginary data the quarter period shifted data. Let us assume that $\mathbf{X}_C$ has been detrended, so that its mean is zero. The variance is thus given by the sum of the diagonal elements of the following matrix

$$\mathbf{X}_C \mathbf{X}_C^* = \mathbf{X}\mathbf{X}^* + H(\mathbf{X})^* H(\mathbf{X}) + i \left(\mathbf{X} (H(\mathbf{X}))^* - H(\mathbf{X})\mathbf{X}^* \right). \quad (5.7)$$

Therefore, the variance of the new data set $\mathbf{X}_C$ is twice the variance of the original data series, as the imaginary term does not contribute to the variance. However, the balance in the imaginary term is rather delicate and it often happens that in real cases affected by noise, the variance is only *approximately* twice the original variance of the data. Complex EOF defined through Hilbert transforms will therefore try to optimize variance using patterns that are complex and whose real and imaginary parts are shifted by a quarter period. Below is a Matlab implementation of this procedure, that was used to generate later plots.

```
function [u,lam,v,proj]=ceof(z,indf,nmode,nproj)
%
% Compute complex EOF of z and expand it for nmode modes
%
% Inputs:
%       z Data Matrix
%       indf        Index for the data
```

```
%       nmode      Number of EOF to return
%       nproj      Number of EOF to generate projections
% Outputs:
%       u          EOF arrays (nspace x nmode)
%       lam        variance explained   (ntime)
%       v          Unnormalized EOF coefficients
%       proj       Projection on the nmode EOF
%
resol = [96 48];
zh=hilbert(z);
[uu,ss,v]=svd(zh,0);

lam = diag(ss).^2/sum(diag(ss).^2);   % Explained variances

u=zeros([resol(1)*resol(2) nmode]);   % Keep Only first modes
u(indf,1:nmode)=uu(indf,1:nmode);
proj=zh'*uu(:,1:nproj);    % Compute projections

return
```

Figure 5.6 shows the first complex EOF (CEOF 1) for the case of the analytical wave of Sect. 4.5.1. The top panels show the real and imaginary parts of the first mode and they display the familiar shape in quadrature one with the other. The real and imaginary parts of the coefficient are also shifted one quarter wavelength. We can see that the CEOF has recovered the propagating wave hidden in the noise.

Being focused on extracting the signals that are shifted one quarter wavelength, the CEOF are very efficient at doing that, but at the same time the Complex EOF do not comparably perform if the oscillatory signal has a structure with a different phase relation. For instance, if the signal is stationary, namely it changes in time without a change of phase in space, like an oscillating beam, CEOF run into trouble. Propagation and stationarity are identified clearly in our ideal experiment by simple EOF (Fig. 5.5) because the stationary signal (bottom panel) shows no clear phase relation between the time series of the coefficient. Application of the CEOF to a stationary signal (Fig. 5.7) produces a spatial pattern that bears indication of the signal stationary nature. Only the real or imaginary component is now needed to give the spatial structure of a stationary signal, in this case the real part, whereas the other component is usually noise, without a clear pattern. It would appear that CEOF have successfully identified the signal, however if one looks at the time coefficient (bottom panel) it is possible to see that both time coefficients oscillate, pretty much in the same way as in the preceding propagating case. CEOF can only distinguish between spatial propagation and lack of it, implying the absence of spatial phase relations; in general, however, the inspection of the time coefficient alone is not sufficient to distinguish between them. As an example, in Fig. 5.6 it is possible to see that the variation of the spatial phase (the arrows in the panel) is organized and smooth, corresponding to the organized propagation. In contrast, in Fig. 5.7 the phase variation is disorganized and dominated by noise. This investigation can be somewhat difficult to perform with real data, where spatial phase relations are
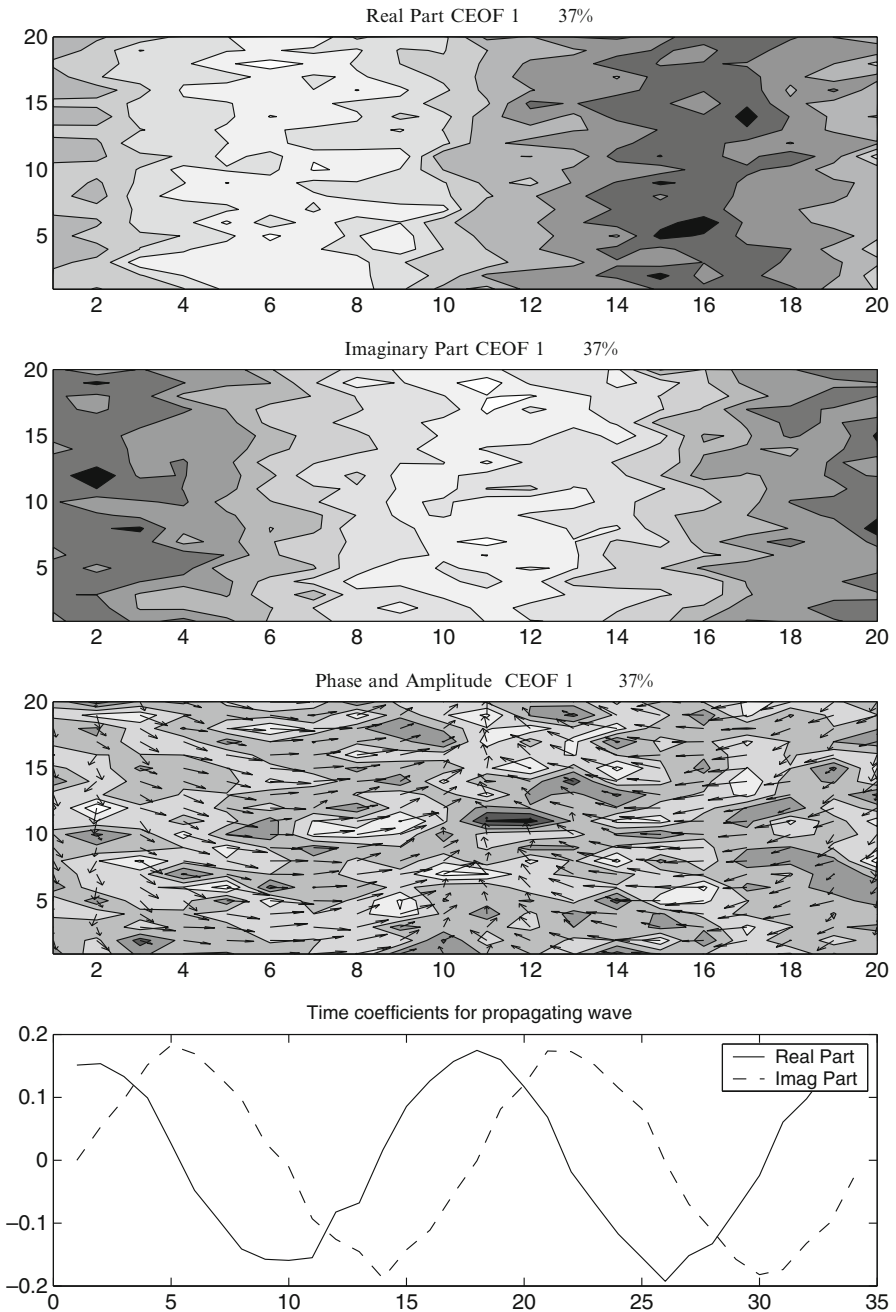
Fig. 5.6 First complex EOF of the analytical example. *Top panels*: spatial patterns of the real and imaginary parts, then the amplitude and phases of the mode. In the title, the explained variance is recorded. *Bottom panel*: time evolution of the coefficient
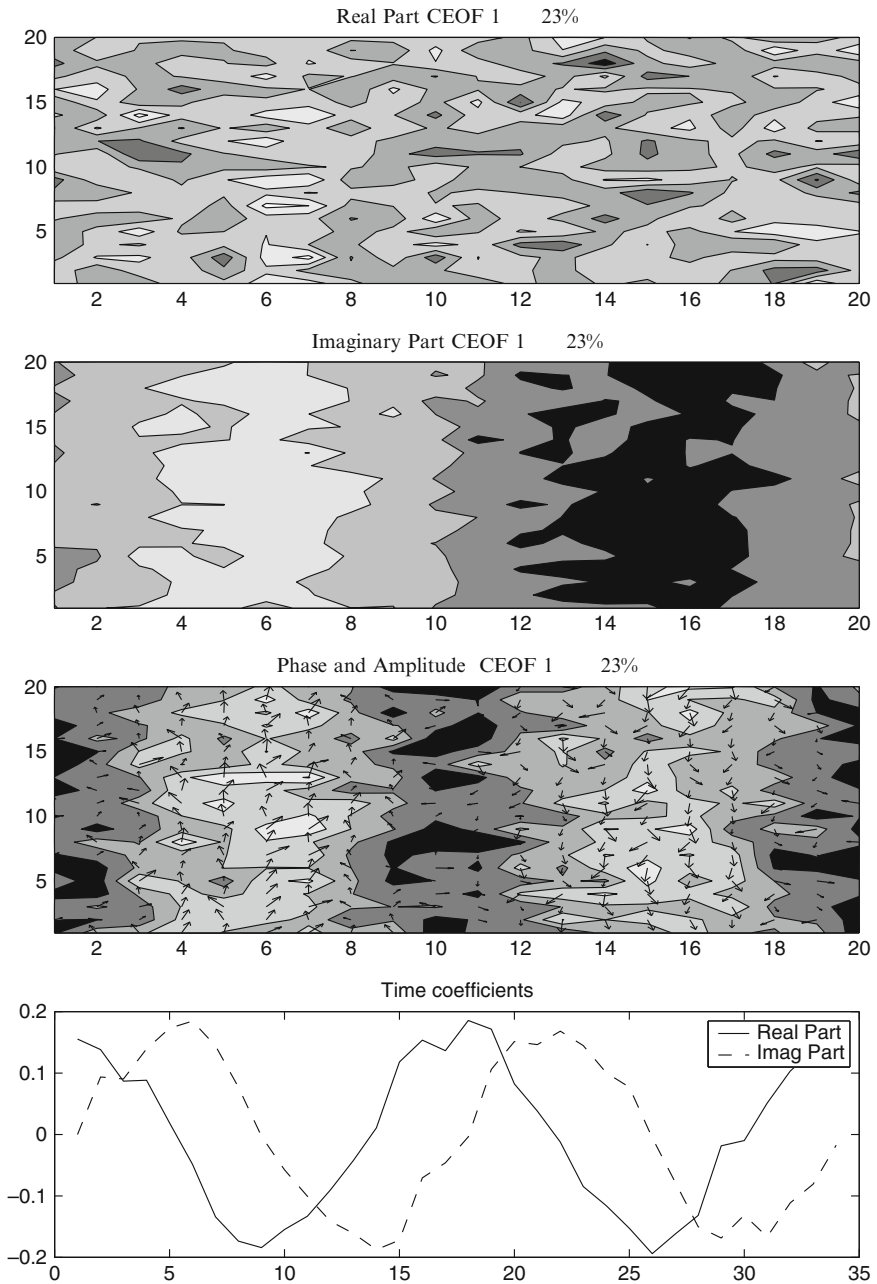
Fig. 5.7 As in Fig. 5.6 but for the case of a stationary wave

difficult to identify. In practice *Complex EOF* cannot be used to distinguish between propagating and non-propagating (i.e. stationary) oscillations.

A complex analysis of the test data set for SST yields the result shown in Fig. 5.8. This picture displays the second mode represented in its real and imaginary components. The top panel is the real part, showing a mode of variability concentrated in the equatorial area, the middle panel is the imaginary component of the mode. The
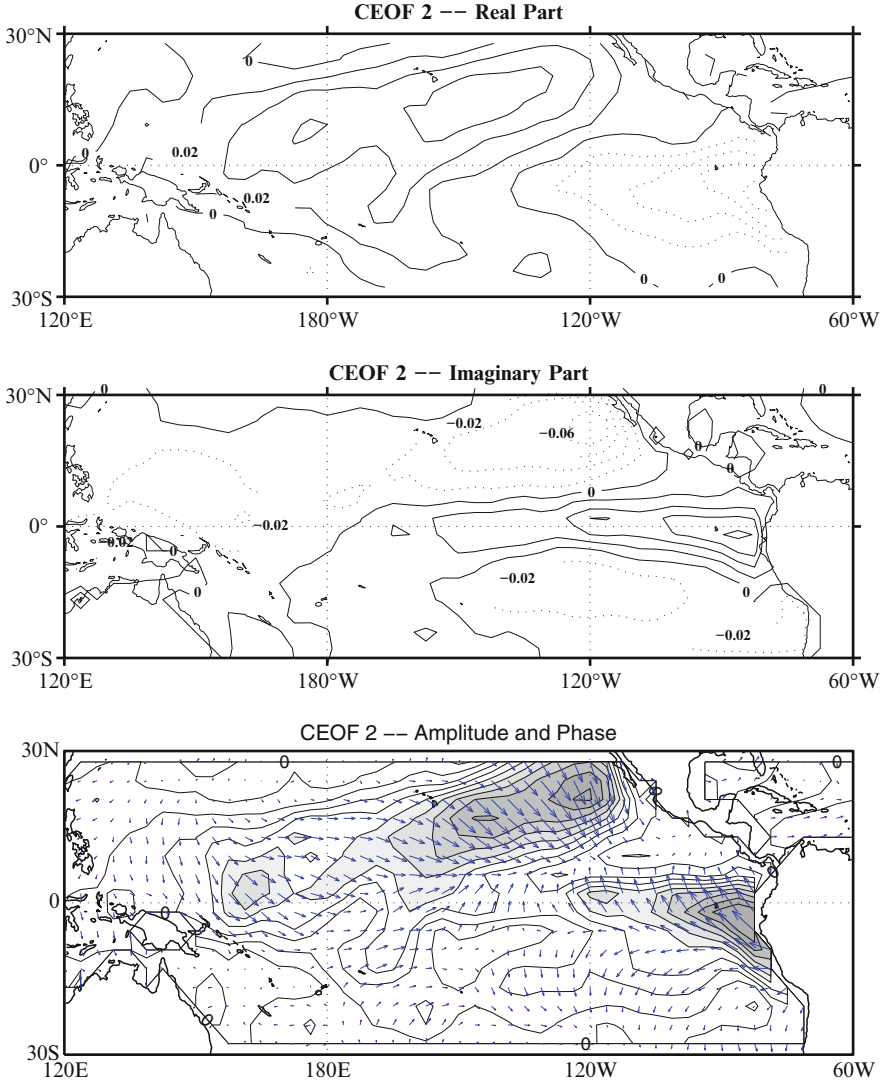
Fig. 5.8 Second complex EOF of the marine temperatures in the Pacific. *Top panel*: real component, *middle panel*: imaginary component, *lower panel*: amplitude and phase
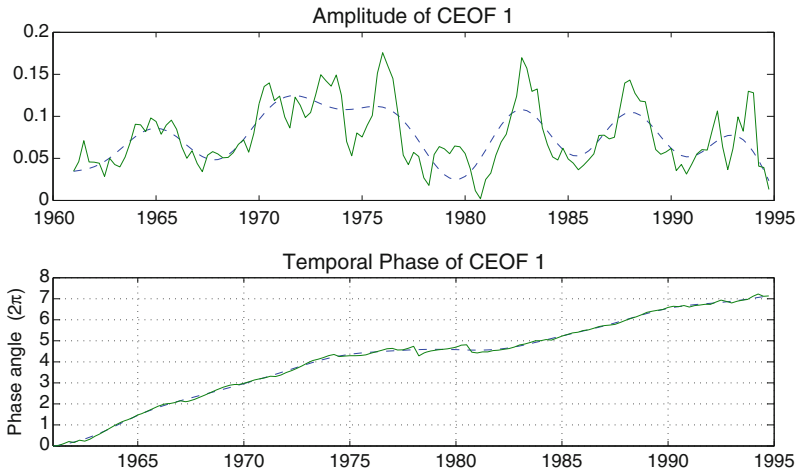
Fig. 5.9 Time series of the coefficient of the second Complex EOF shown in the previous picture. The *top panel* shows the amplitude of the complex coefficient, whereas the *bottom panel* shows the evolution of the unwrapped phase angle. The phase velocity is obtained as the derivative of the phase, showing an acceleration after 1980

bottom panel is the representation in amplitude and phase. The amplitude is concentrated in the east equatorial Pacific, the rotation of the phase indicates a phase velocity towards the west. Here the convention used is that the phase arrows point to the east if the real part is positive and the imaginary part is zero.

The time evolution of the mode coefficient is displayed in Fig. 5.9, indicating the periods of time in which such a mode is more or less energetic. The "unwrapped" phase, that is the phase of the time coefficient reduced to a single-value function by adding a factor 2π every time it crosses the zero line, also shows different phase speed from a period to the next.

Figure 5.10 shows the modal actual evolution, cycling through the real and imaginary parts with alternate signs. The reported field is only based on the reconstruction of the second mode, starting from 1980 onward. The picture shows that the CEOF indicates an oscillatory behavior that can also be aperiodic in time. There are periods in which oscillations are clearly visible, and periods where oscillations are quiescent and there is very little appearance of the mode. This is a good example of the capability of the CEOF to capture irregular oscillations.

5.4 Extended EOF

Complex EOF are based on the analysis of variance by taking into account the data time behavior. This is done by creating a new data set that includes the original data series and a new series that is shifted by a quarter wavelength. The Hilbert transform

COMPLEX EOF   *51%*
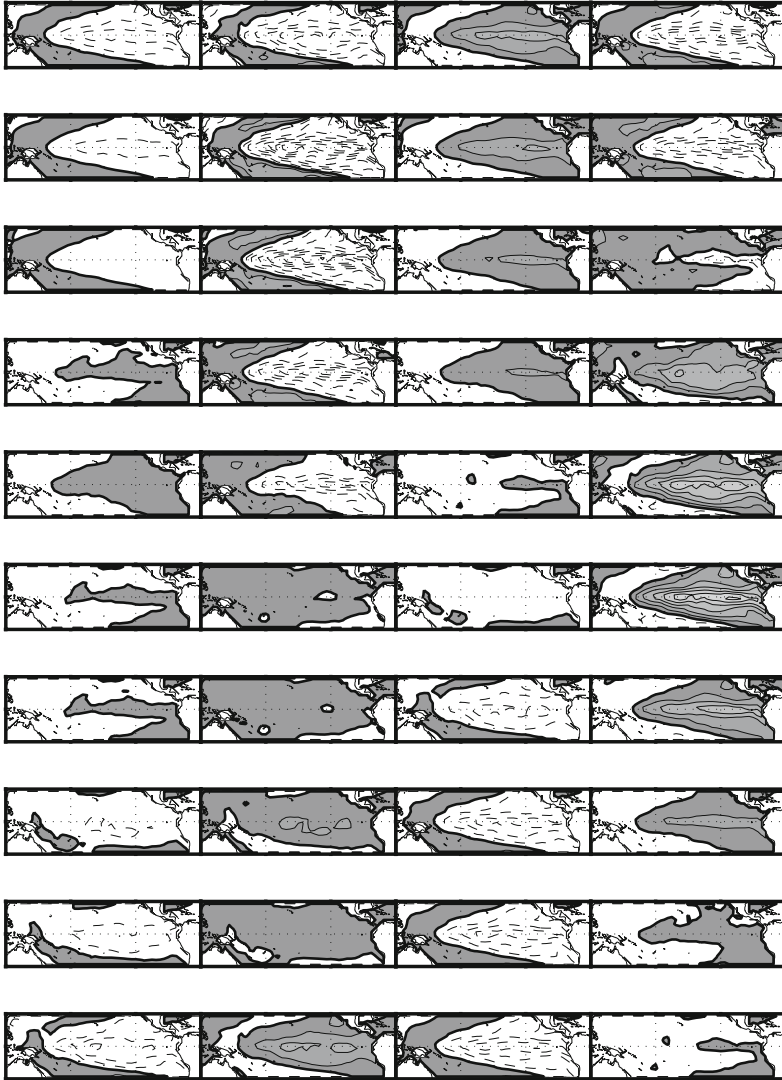
Fig. 5.10 Time evolution of the second for SST from Winter 1980 (*top left panel*) for each consecutive season. Time is increasing downwards and from left to right. The 1982–1983 El Niño event is visible in the first and second column on the left

makes the procedure very rigorous. However, it is sometimes desirable to use a less rigorous approach and to gain some flexibility in the process. *Complex EOF* change the state vectors in a way that the basic data are not the data at a given time, but the combination data at a single time plus data shifted one quarter wavelength in time. A possible alternative is to introduce a derivative EOF analysis that crudely realizes

this fact. This new method, often called the Extended EOF (EEOF), simply consists in extending the data set with repetitions of the time series suitably lagged. For the test cases we are using here it will mean to extend the data by adding several copies of the time series with proper time shifts, i.e.

$$\mathbf{X}_E = \begin{bmatrix} \mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_{m-2} \\ \mathbf{x}_2 \ \mathbf{x}_3 \ \cdots \ \mathbf{x}_{m-1} \\ \mathbf{x}_3 \ \mathbf{x}_4 \ \cdots \ \ \mathbf{x}_m \end{bmatrix}.$$

The basic observation vector at time n is given by

$$y_E(n) = \begin{bmatrix} \mathbf{x}_n \\ \mathbf{x}_{n+1} \\ \mathbf{x}_{n+2} \end{bmatrix}.$$

It is formed by $k+1$ fields, each showing the dominant mode of variations over the k lags. A single mode is then formed by several components each representing the spatial pattern for that phase of the lags. The trick is to include the lags that are important for reproducing possible oscillatory patterns. It is advisable to investigate the autocorrelation function to gather some indications of the number of lags that need to be included. The method is very flexible, the lags do not need to be consecutive. Instead of using three consecutive months like in the previous example, we could have chosen some three months in three months. In principle they do not even need to be equally distributed; arbitrary lags could be defined, but results would be extremely difficult to interpret. In practice it is advisable to use regularly spaced lags. The variance of the augmented series is a multiple of the variance of the original series and it is approximately $k+1$ times the original variance, so the amount of variance explained must be assessed against this augmented variance.

A simple Matlab implementation of the Extended EOFs approach follows.

```
function [u,lam,v,proj]=eeof(z,indf,nmode,nproj)
%
%Compute Extended EOFs of matrix z and expand it for nmode modes
% Use  3 lags
% Inputs:
%           z Data Matrix
%           indf    Index for the data (from the reading routine)
%           nmode   Number of EOF to return
%           nproj   Number of EOF to generate projections
% Outputs:
%           u       EOF arrays (nspace x nmode)
%           lam     variance explained   (ntime)
%           v       Unnormalized EOF coefficients
%           proj    Projection on the nmode EOF
%
resol = [96 48];
[np,nt]=size(z);
```

```
lags=3;       nmode=2;
zh=ones((lags+1)*np,nt-lags);
zh(1:np,:) = z(:,1:nt-lags);
zh(np+1:2*np,:) = z(:,2:nt-lags+1);
zh(2*np+1:3*np,:) = z(:,3:nt-lags+2);
zh(3*np+1:4*np,:) = z(:,4:nt-lags+3);
[uu,ss,v]=svd(zh,0);

lam = diag(ss).^2/sum(diag(ss).^2);   % Explained variances

u=zeros([resol(1)*resol(2) 4]);       % Only first mode
uc=zeros(np,4);
for i=1:4
    uc(:,i) = uu((i-1)*np+1:i*np,nmode);
end
u(indf,1:4)=uc(:,1:4);

proj=zh*uu(:,1:nproj);      % Compute projections

return
```

The example reported in Fig. 5.11 shows the result of applying an EEOF analysis to the tropical SST. The lags have been defined to the seasonal means of the SST and three seasonal lags have been used. It is possible to see how the main pattern of variations are captured.

Exercises and Problems

1. Show that the diagonal terms of the imaginary term in (5.7) do not contribute to the variance of the field.
2. Show that the total variance of the EEOF time series is approximately $k + 1$ times the original one, and that the approximation gets better as the number of time observations increases.
3. Construct the time evolution for the first mode for the EEOF technique.

5.5 Many Field Problems: Combined EOF

The extension of the EOF analysis to the time domain shows that the concept is more general than we may have thought. The logical path that we have followed to go into the time domain has exploited the freedom to change the rules of compositions of the data fields. We have generated other ways to analyze variance by arranging/transforming the data differently. The extension we have made in the previous section was mainly in the time variable, but we can use the freedom to change the definition of the data vectors to explore the variation of combined fields. We can, for instance decide to define a new data set by putting together the height and SST

EEOF Mode 2   Season 1   *13%*

Season 2
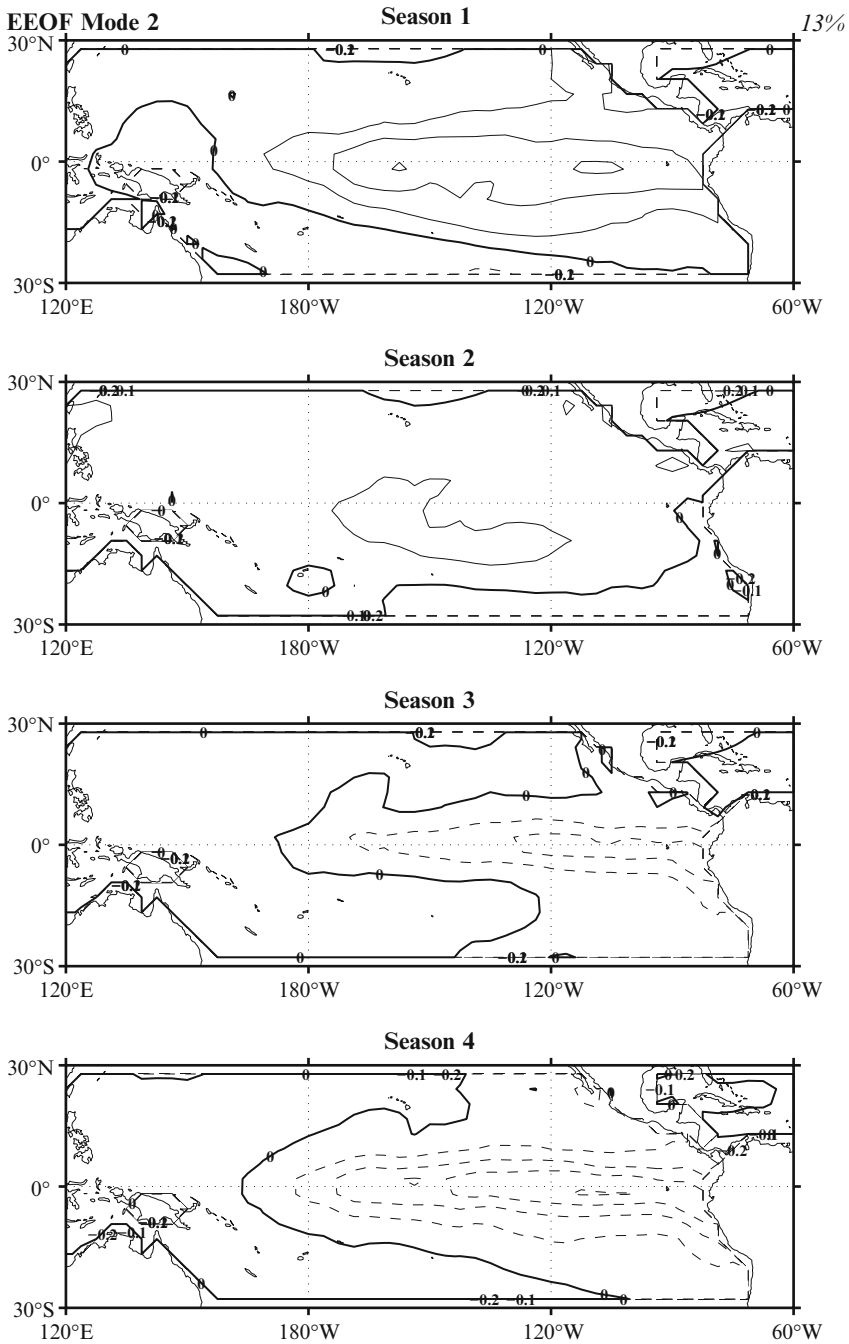
Season 3

Season 4

Fig. 5.11 First EEOF mode for the SST data set. The analysis has been performed by season, using three lags of one season each. The picture depicts the evolution of the mode through four consecutive seasons. The amount of variance explained by the mode is referred to the total variance of the augmented series

data. The data matrix can then be written as

$$
\mathbf{Y}_n = \begin{bmatrix} \mathbf{z}_1 \cdots \mathbf{z}_n \\ \mathbf{s}_1 \cdots \mathbf{s}_n \end{bmatrix},
$$

where the data are arranged in such a way to keep the time correspondence between the different fields, so that fields at the same time are put in the same column. We can also use the data matrix for the fields $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m]$ and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m]$ so that the new combined data matrix Y becomes

$$
\mathbf{Y} = \begin{bmatrix} \mathbf{Z} \\ \mathbf{S} \end{bmatrix}.
$$

Assuming zero mean, we can compute the covariance matrix for the combined field as

$$
\mathbf{YY}^* = \begin{bmatrix} \mathbf{Z} \\ \mathbf{S} \end{bmatrix} [\mathbf{Z}^*, \mathbf{S}^*] = \begin{bmatrix} \mathbf{ZZ}^* & \mathbf{ZS}^* \\ \mathbf{SZ}^* & \mathbf{SS}^* \end{bmatrix}, \tag{5.8}
$$

showing that the total variance of the combined field is the sum of the variance of the composing fields.

The two data sets can have different geographic extensions, though they must have the same number of time levels. There is also no limitation in the number of fields that are patched together in this way. We can put in the same data space three or four different fields, in principle there is no limit. This a very useful and rather unique feature of the combined EOF. There are several situations when this may be convenient. For instance, when treating tropical air-sea phenomena it is often useful to look for combined modes of variations of wind stress, SST, Outgoing Lonwave Radiation (OLR), precipitation, clouds, etc. The combined EOF is the only method that allows a simultaneous considerations of the possible modes of variation of different variables.

The combination of fields in this way requires some care to handle different units and quantities. Different data have widely different numerical values corresponding to the different units that are used to measure them. These differences could generate systematic deviations in the resulting patterns that do not correspond to real variability patterns. The problem can be overcome by transforming the data to values of the same order of magnitude by using suitable scales, making the data adimensional. The simplest way is to divide the data by constants that represent typical value for that variable. For instance, in our case we could use a temperature scale of 300 K, and a geopotential height scale of 5000 m, that would change all the data values to order one. Another possibility is to normalize them by the point-by-point standard deviation, in a similar way to what was done in Sect. 4.4.1. In the first case the scaling is simply equivalent to a multiplication by a constant and the covariance structure is not modified, so we get the Combined Covariance EOS, in the latter case the covariance structure is modified and we get Combined Correlation EOF.

Each mode is now a combination of the fields that have been used to create the combined data set. The mode describes the principal mode of variations of the combined data and it is not different from the EOF that we have described in the previous chapter. However, the various fields can be identified in the mode by reconstructing the different components with the corresponding order in the data field. In this sense, the combined EOF is a straight generalization of the EOF that can be considered as a one-parameter Combined EOF. A typical implementation is as follows

```
function [u,lam,v,proj]=combeof(zz,inds,indz,nmode,nproj)
%
% Compute combined EOF of matrix zz. The matrix zz contains
% the ordered fields to be combined, in this case Z and S.
% Inputs:
%          zz    Combined Data Matrix
%          inds       Index for the S data (ocean)
%          indz       Index for the Z data (atmosphere)
%          nmode      Number of EOF to return
%          nproj      Number of EOF to generate projections
% Outputs:
%          u          EOF arrays (nspace x nmode)
%          lam        variance explained    (ntime)
%          v          Unnormalized EOF coefficients
%          proj       Projection on the nmode EOF

resol = [96 48];  ss=resol(1)*resol(2);
[uu,ss,vv]=svd(zz,0);
lam = diag(ss).^2/sum(diag(ss).^2);  % Explained variances

ls=length(inds);     lz=length(indz);
u=zeros([ss nmode]);    v=zeros([ss nmode]);
for i=1:nmode
    u(indz,i)=uu(1:lz,i);
    v(inds,i)=uu(lz+1:lz+ls,i);
end

proj=zz*uu(:,1:nproj); % Compute projections

return
```

We have used the standard deviation normalization to produce Combined Correlation EOF of the height and SST fields showed in Figs. 5.12 and 5.13. The mode resembles very much the EOF obtained by performing the analysis of the SST or the height field alone. The pattern can be superposed almost exactly. It is possible to understand this effect by inspecting the structure of the combined data covariance matrix in (5.8). The structure is essentially given by a block matrix structure where the blocks are the covariance matrix of the component fields along the diagonal and the cross-covariance matrices of the fields in the off diagonal positions. Therefore, the diagonal terms express the internal variability of the fields, whereas the off-diagonal terms express the variance of one field that is related to the other field.
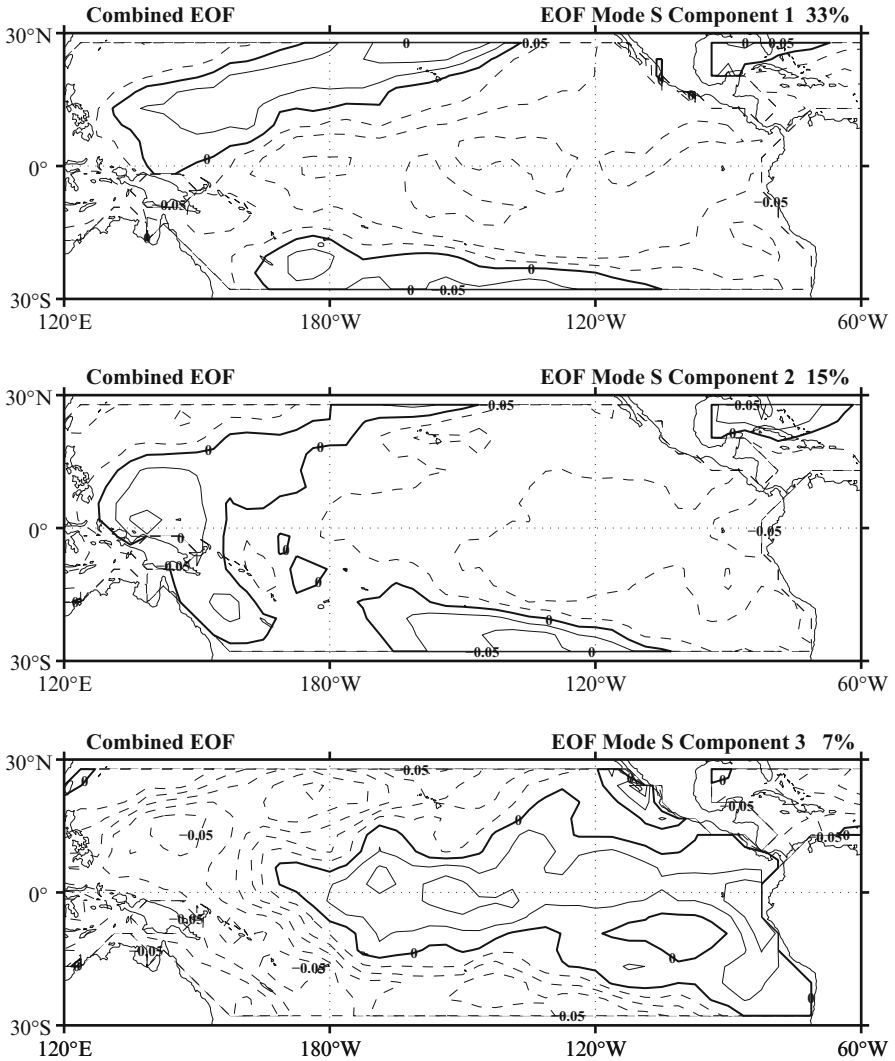
Fig. 5.12 The first three combined EOF modes for the Height-SST data set. Here is shown the SST component in descending order of explained total combined variance

The combined EOF will obtain the same EOF as the individual fields if the off-diagonal terms are small compared to the diagonal ones. This happens if the data fields are independent of each other and therefore the cross-covariance components are small; in this case, the structure of the combined covariance matrix is essentially dominated by the individual covariance of the fields. The combined EOF will be dominated by the autocovariance of each field if the internal variability of the fields less larger than the cross-covariance.
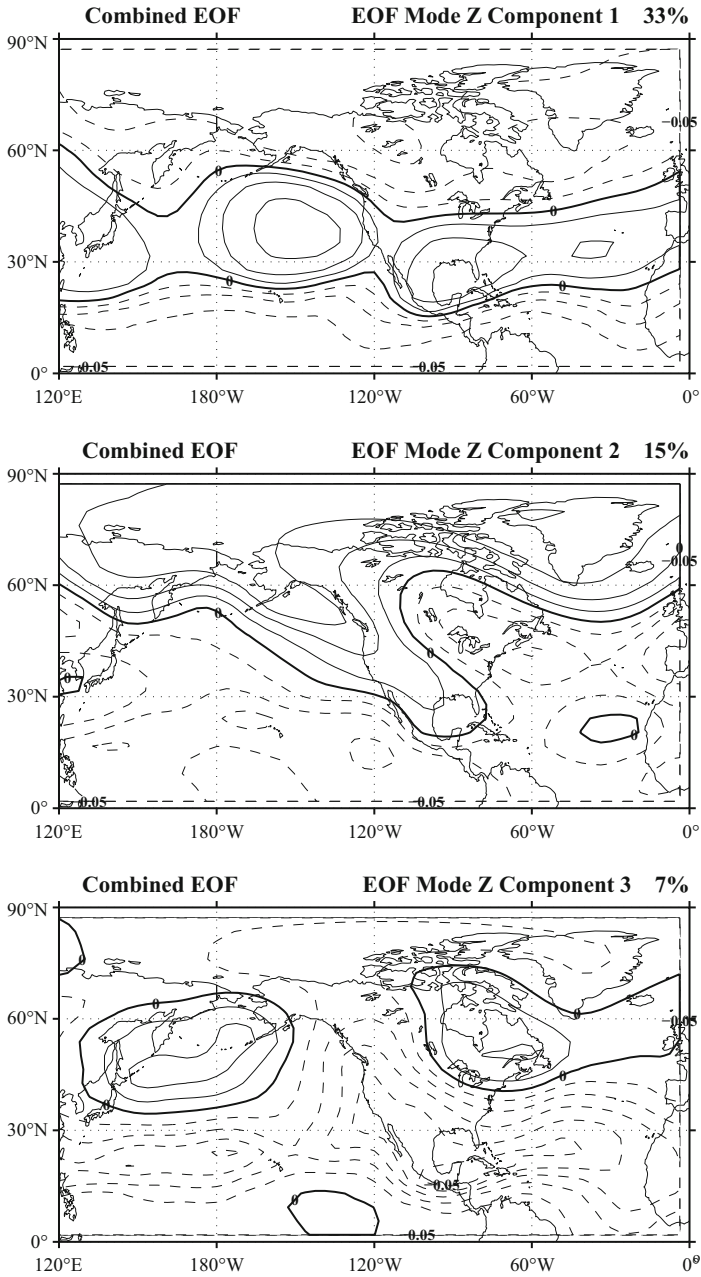
Fig. 5.13 The first three combined EOF modes for the Height-SST data set. Here is shown the Z component in descending order of explained total combined variance

This observation leads to the main weakness of the combined EOF: by mixing the autocovariance of each field and the cross-covariance of one field with the other, combined EOF cannot separate the patterns for the different kind of variability and one cannot tell the respective amount due to the autocovariance or to the cross-covariance. The Combined EOF mode will bear the imprint of both sectors of variability of a particular variable. It is a pity, because the cross-covariance could be extremely useful when one has to study coupled problems, like the air-sea interaction in the tropics. The Combined EOF cannot give a suitable help on this issue, but we will see in the following chapter that we can work out specific methods to address this exciting issue.