**EOF Calculations and Data Filling from Incomplete Oceanographic Datasets**

J. M. BECKERS and M. RIXEN

*University of Liège, Liège, Belgium*

(Manuscript received 5 October 2001, in final form 9 April 2003)

**ABSTRACT**

The paper presents a new self-consistent method to infer missing data from oceanographic data series and to extract the relevant empirical orthogonal functions. As a by-product, the new method allows for the detection of the number of statistically significant EOFs by a cross-validation procedure for a complete or incomplete dataset, as well as the noise level and interpolation error. Since the proposed filling and analysis method does not need a priori information about the error covariance structure, the method is self-consistent and parameter free.

1. **Introduction**

Oceanographic and meteorological applications of empirical orthogonal decompositions have been commonly used for several decades and are well documented in textbooks (e.g., Preisendorfer 1988). Empirical orthogonal functions (EOFs) are however not a tool specific to the ocean sciences, but are widely used in a number of scientific domains under different names: proper orthogonal decomposition (POD) or Karhunen–Loève decompositions are mostly used in a functional framework [working on mathematical continuous functions; e.g., Lumley (1970); Holmes et al. (1996)], whereas the matrix versions (working on discrete datasets) are called proper orthogonal modes (POM), principal component analysis (PCA), factor analysis (e.g., Hair et al. 1998), or empirical orthogonal function analysis.

In oceanography, EOF analysis has been used for several kinds of applications, among them objective analysis (or optimal interpolation) of in situ data; for time-dependent 3D datasets, classic objective analysis methods are generally intractable and the use of vertical EOFs and horizontal objective analysis of their amplitudes drastically reduces the amount of data during the costly optimal interpolation, but takes into account the strong vertical correlations contained in the EOFs (Holbrook and Bindoff 2000). Other applications of EOFs include statistical comparison between data and model results (e.g., Beckers and MEDMEX 2002), data compression (e.g., Pedder and Gomis 1998), analysis of variability (e.g., Hendricks et al. 1996), and filtering (e.g., Vautard et al. 1992). Reducing the size of a problem by projecting (complex) reference model equations onto a small number of principal components (D’Andrea and Vautard 2001) is also a powerful application of EOF analysis.

Each of these applications basically exploit the fact that EOFs allow a decomposition of a signal into a series of data-based orthogonal functions, which are designed so that only a few of these functions are needed to optimally reconstruct the initial signal.

Contrary to most other domains, in oceanographic data and in particular satellite data analysis, both noise in the data and missing data points are a concern and must be dealt with: in the datasets, there is a combination of missing data due to misfunctions of instruments, inaccessible data (e.g., due to cloud coverage), instrumental error bars (generally small, except for altimetric data for which the precision of a few centimeters can be in the range of the dynamical signal), and noise related to small-scale features not resolved by the observational network (at least all features below the Nyquist frequency).

The aim of the present paper is to analyze the problem of missing or unreliable data in oceanographic datasets and their EOF analysis. We will present the method on classic EOF (allowing for complex data), although generalizations of the present method are possible, as for example to extended EOFs, delayed EOFs, etc. (e.g., Emery and Thomson 1998); singular spectrum analysis (e.g., Vautard et al. 1992); or other transformations (e.g., Elmore and Richman 2001). For the sake of clarity and
without loss of generality, we also focus on data given in a uniform spatiotemporal grid; for nonuniform grids, the approach of North et al. (1982) or Shen et al. (1998) (basically replacing the discrete problem by a continuous integral problem, allowing weighting of the data according to their associated spatial or temporal coverage) can be used to work numerically in a discrete uniform space.

We assume thus that we have a matrix $X$ containing the observations, which is arranged such that the element $i, j$ of the matrix is called $(X)_{ij}$ and is given by the value of the field $f(r, t)$ at location $r$ and moment $t$: $$(X)_{ij} = f(r_i, t_j).$$

The field $f$ is an observational field and thus contains all errors (instrumental, unresolved structures, etc). We can then write the matrix as a succession of $n$ column vectors, $$X = (x_1, x_2, \ldots, x_n),$$
each of the column vectors $x_i$ being the discrete state vector of size $m$ at moment $t_j$. The state vector can be a 1D field, an ordered set of values from a 2D satellite image, etc. Values of the state vector do not need to be real but can accommodate complex values, allowing complex EOF on 2D vectors, Fourier transforms, Hilbert transforms, etc. (e.g., Emery and Thomson 1998).

EOFs can then be calculated easily by standard singular value decomposition (SVD) techniques (e.g., Lawson 1976, Hanson 1974), which are based on methods solving the following eigenvalue problem:

$$XX^*u = \rho_i u$$

and

$$X^*v = \rho_i v,$$

the first system being of dimension $m$ (as the eigenvector $u$ and the discrete state variable $x$) and the second one of dimension $n$ (as the eigenvector $v$). Here, $u$ is often referred to as the spatial EOF, and $v$ as the temporal EOF, while $\rho$ is the so-called singular value: $X^*$ is the adjoint (conjugate transposed) of the matrix $X$.

As an alternative to solving Eqs. (3) and (4), we can solve instead

$$XX^*u = \rho^2 u$$

and

$$X^*v = \rho^2 v,$$

showing that the spatial modes $u$ are the eigenvectors of the time-averaged (or temporal) covariance matrix $XX^*$, while $v$ are the eigenvectors of the spatially averaged covariance matrix $X^*X$.

Eigenvectors are normalized, and by virtue of Eqs. (5) and (6), orthogonal:

$$u^*u = \delta_{ij}, \quad v^*v = \delta_{ij},$$

where $\delta_{ij}$ is the Kronecker symbol (with a value of 1 if $i = j$ and 0 otherwise). For the spatial patterns $u_i$, we retrieve orthogonality, while for the temporal pattern $v_i$, the orthogonality is sometimes referred to as the time series of the $i$th EOF being uncorrelated to the other EOFs.

Once the eigenvectors and eigenvalues are known [the number of eigenvalues different from 0 is the rank $q = \min(m, n)$], we can decompose the initial matrix as follows:

$$X = UDV^* = \sum_{k=1}^q \rho_i u_i v_i^*,$$

where we have for matrix $X$, defined matrices $U$ and $V$ so that they have as columns $i$ the eigenvectors $u_i$ and $v_i$, respectively, corresponding to the singular value $\rho_i$. Matrix $D$ is then a rectangular matrix whose only nonzero values are on its diagonal and $(D)_{ij} = \rho_i \delta_{ij}$. For the development, we use the classic convention of ordering the eigenvalues in descending order (from the larger to the smaller ones). The interest in SVD decomposition compared to any other orthogonal decomposition is because if instead of using all $q$ eigenvalues and eigenvectors, we use only the first $k$ of them, it can be shown (e.g., Preisendorfer 1988) that for a given number of modes $k$, no other choice of base vectors $u_i$ leads to a better approximation of the real matrix $X$ than the truncated series obtained with the $k$ first EOFs. This means that when retaining only 1 mode, the corresponding EOF (with its evolution) is the one that on average allows for the best approximation of the real matrix $X$ by a given spatial structure (the spatial EOF) whose amplitude evolves in time. A measure of how well an EOF mode represents the signal is contained in the singular values. Indeed, by construction

$$\text{trace}(XX^*) = \text{trace}(X^*X) = ||X||_2^2 = \sum_{k=1}^q \rho_i^2$$

is a measure of the total variance (also sometimes called energy) in the system. The ratio $f_i = \rho_i^2 / \sum_{k=1}^q \rho_k^2$ is thus a measure of the variance contained in mode $k$ compared to the overall energy and one often says that mode $k$ explains 100 $f_i\%$ of the variance and that the first $K$ modes explain 100 $\sum_{k=1}^K \rho_k^2 / \sum_{k=1}^q \rho_k^2\%$ of the total variance. This ratio is often the basis for deciding the number of EOFs to retain in a given decomposition. A typical choice is to retain those modes that, when summed up, explain 95% of the signal. Another choice for optimal truncation is the so-called Preisendorfer and Overland criteria (Preisendorfer and Overland 1982); in this case, a Monte Carlo method generating random matrices of the same size decides when pure noise is likely to be interpreted as a signal if the series is not truncated.

To finish the general description of EOF decompositions, we can mention that when working on EOFs or covariances, some a priori background field or trend is subtracted from the data series (e.g., Kantha and Clay-
son 2000), notably to allow for an easy interpretation of the eigenvalues in terms of the energy of the variability [see Eq. (9)]. For the sake of symmetry in terms of space or time coordinates, we assume here that the a priori average we subtract from the field (in order to work with anomalies) is the spatiotemporal average of all available data.

2. Missing data

The SVD method presented above assumes that the matrix $X$ is perfectly and completely known. Currently, when dealing with missing or unreliable data, one fills in the missing data by objective analysis methods or optimal interpolation (e.g., Houseago-Stokes 2000); of course the problem with such an approach is that one needs information about the correlation function and the signal/noise ratio of the data for the interpolation (e.g., Rixen et al. 2000; Gomis et al. 2001). Since the purpose of using EOFs is often for the data analysis itself, another preliminary data analysis is not very natural and introduces a priori information into the subsequent EOF analysis. This filling method (and its variants that use simpler interpolation methods) is the most commonly used option to estimate the missing data values, because it consists of applying a classic analysis tool (e.g., Wunsch 1996) as preprocessing, before using any SVD or EOF analysis package.

Another method for dealing with missing data consists of calculating the eigenvalues and eigenvectors of the correlation or covariance matrix $C$, where the covariance is calculated using only the available data (Boyd et al. 1994; Kaplan et al. 1997; von Storch and Zwiers 1999). This however does not necessarily lead to a semipositive defined covariance matrix (since it is no longer possible to write $C = X^* X$, which would lead to an always positive defined form $x^* C x$, with $x$ being a column vector). Therefore its eigenvalues can be negative. Since the energy of the system is related to the trace of the covariance matrix that is equal to the sum of its eigenvalues, having negative eigenvalues means that other eigenvectors will have higher eigenvalues than in reality, thus overestimating the corresponding amplitude of the EOF. In addition, once the eigenvalues and eigenvectors are calculated, one still faces the problem of deciding which EOFs are relevant. Due to the presence of negative singular values, the selection criteria can no longer rely on the relative importance of the eigenvalue, but should also take into account the effect of the negative eigenvalues on the distribution of the larger eigenvalues. Moreover, using a temporal instead of a spatial covariance matrix (based on averaging the available data only) will generally lead to different eigenvalues (and thus singular values) though the mathematical counterpart for the correlation matrix based on $XX^*$ or $X^* X$ gives identical results. Another way of interpreting this can be based on the observation that the EOF coefficients (the product of the singular value and the temporal EOF) of a gappy data vector cannot simply be obtained by a dot product, but must be estimated by a least squares approach (von Storch and Zwiers 1999).

Since both of the approaches (filling or correlation calculated on the sole available data) most commonly used contain some subjective components (choice of interpolation methods and interpolation parameters, level of significant EOFs, . . .), any method that would not need such additional choices could be of interest. The development of such a method is the purpose of the present work and is presented in the following way. First we present the interpolation method and analyze how it modifies the data at the missing data points in section 3a. Then we will tackle the problem of deciding on the optimal number of EOFs to be retained (section 3b). We will validate the method on two datasets: a synthetic one for objective quantifications of the errors of the method (section 4a), and then a real Advanced Very High Resolution Radiometer (AVHRR) dataset in section 4b. Finally, conclusions are drawn in section 5.

3. EOF-based interpolation

We now present a method aimed at filling in the gaps in the datasets and determining the structure and number of significant EOFs.

Basically, since EOFs are supposed to summarize in a very efficient way the signal and are also supposed to detect coherent features, it is tempting to use the EOFs themselves for interpolations. If we knew the dominant EOFs of a given field and their amplitudes, their combination would also give us the value of the field at the missing data points.

a. Interpolation and adaptation of the EOFs

Large-scale EOFs should not be influenced by local changes in the values of a few points, and so one could try to calculate the EOFs based on a matrix in which a first rough estimate of the variable is used in the missing data points. Then, once the larger-scale EOFs and their amplitudes are estimated, they can serve to calculate the value of the field at the missing points. Then, of course, the EOFs themselves can be reevaluated and the process can be repeated until convergence. One of the problems, which will of course arise, is the decision on the number $N$ of relevant EOFs to be retained to recompose the signal at the missing data points. We will come to this question later. Formally the method can then be described as follows:

- let $I$ be the ensemble of discrete points where data are missing; that is, when $(i, j) \in I$, we do not have data or they are unreliable; $I$ is of course a subset of $[1, m] \times [1, n]$, and the number of missing data points is $n_I$;
- $X_i$ is the true field we would like to recover;
• \( R \) is a noise field (unresolved features, instrumental noise, etc. . .);
• \( X_0 = X_i + R \) is the true field with superimposed noise;
• \( X_i \) is the observed field with gappy data (typically cloud coverage for satellite images) and zero values at those locations; and
• \( X_o \) is the observed field where missing data were filled in (also called the analyzed field in analogy with objective analysis methods).

For any of these matrices, \( X_o \) is the reconstruction of the corresponding matrix by the first \( N \) EOFs, thus filtering out the contribution of the higher EOFs (it should be noted that for each of the different matrices, the EOFs are different).

We will use the same sub- and superscripts for other variables (e.g., singular values) when we need to distinguish between those that are relevant to the real signal data, the noisy signal, signals with missing information, etc.

Our proposed algorithm performs as follows. First for all missing data points \([i, j] \in I\) we put a value of 0 in the matrix (unbiased guess by assuming that the average value we subtracted was unbiased) to get the matrix \( X_o \).

Then we use the SVD decomposition of this matrix to get access to a first estimation of the spatial and temporal EOFs \( \mathbf{U} \) and \( \mathbf{V} \) as well as their amplitudes (the diagonal elements of the matrix \( \mathbf{D} \) containing the singular values):

\[
\mathbf{UDV}^* = X_o. \tag{10}
\]

The interpolated value at a missing data point is then easily obtained by the truncated EOF series:

\[
(X_o)_{ij} = (\mathbf{UD}_N^N)_{ij} = (X_o)_{ij}
= \sum_{k=1}^{N} \rho_k(u_k)(v_k^*) = (X_o)_{ij}, \quad (i, j) \in I. \tag{11}
\]

The matrix \( \mathbf{U}_N \) is the matrix consisting only of the first \( N \) spatial EOFs. We thus get the truncated reconstruction of the matrix at the missing data points. We now have an analyzed field \( X_a \), given by

\[
X_a = X_o + \delta X, \tag{12}
\]

where the matrix \( \delta X \) is zero everywhere, except at the missing data points.

Once the missing data points are replaced by their new values, we can apply the same procedure again, since the EOFs as well as their amplitudes will be modified by the changes at the missing data points. The following set of operations is thus performed until convergence:

\[
\mathbf{UDV}^* = X_o \quad \text{and} \quad (X_o)_{ij} = (\mathbf{UD}_N^N)_{ij} = (X_o)_{ij}
= \sum_{k=1}^{N} \rho_k(u_k)(v_k^*), \quad (i, j) \in I. \tag{14}
\]

It is instructive to analyze how the changes from one iteration to the next are structured. To do so, we assume that the changes are very small so that we can apply the perturbation theory of appendixes A and B.

If we start from the initial guess of zero anomalies at the missing data point, it can be shown (appendix A) that the variance of the dominant modes will generally be increased, whereas the less dominant modes will see their amplitude decreased. This is coherent with the interpretation that the initial guess overestimates "high frequency" signals and underestimates "large scale" signals by disregarding any local information around the missing data point.

We can also look at how the EOF structures are modified. In fact, it appears that the EOFs are slightly rotated so as to diminish the contribution to the dominant EOFs from the "noisy" EOFs partly introduced by the initial filling by zero values. We indeed observe in the examples (and show the mechanism in appendix A) that not only do we have a tendency to decrease the energy of higher modes, but that we subtract part of the corresponding higher EOF structure from the dominant EOF. This will, as will be seen in the examples, lead to a better representation of the real EOF structure.

In appendixes A and B it is also shown that if two EOFs have similar singular values, the iterations will strongly affect their representation, which could lead to large variations and convergence problems during the iterations. In this case, however, the EOFs do not have a proper meaning anyway, since any combination of two eigenvectors associated with the same eigenvalue is still an eigenvector. In this case (which was not encountered in the examples used later) iterations could be limited to a prescribed value.

b. Determination of the optimal number of EOF

We now have an algorithm that iteratively modifies EOFs and their amplitudes so as to achieve a self-consistent set of EOFs and interpolated data. However, the number \( N \) of relevant EOFs is yet undetermined. Subjective or objective criteria based on Monte Carlo methods could be applied (Preisendorfer and Overland 1982), but this would (a) introduce some subjective parameters or (b) introduce a criteria for truncation purely based on the matrix dimensions and not at all on the power spectra of the singular values of the given data. Furthermore, we would neglect a new possibility introduced by the method, the possibility to interpolate. Indeed, once we can interpolate missing data, we can turn toward cross-validation techniques (e.g., Brankart and Brasseur 1996). We can discard some additional points for validation by using the algorithm to interpolate these values too. Then we can compare the interpolated values to the data we put aside and have a measure of the interpolation error! Then of course finding the optimal number of EOFs is straightforward if we accept that the optimal number of EOFs as the one that minimizes the
interpolation error. The procedure of a cross validation in the determination of the number of relevant EOFs is then the following. We set aside a random set of valid data typically at least 30 data points in order to have a robust statistical estimate of the interpolation error; but even better that a small percentage of the data points with more than 30 points should be used. The interpolation error estimate is then simply the rms distance between the interpolated field at these points and the data there.

The algorithm to determine the optimal number of EOFs and the corresponding interpolated data and EOFs is thus the following. We first apply the interpolation method with a single EOF retained until convergence. We can then calculate the error estimate. A second EOF is now taken into account in the interpolation method (starting of course with the interpolated data based on the single converged EOF) and interpolated until convergence with two EOFs. Then we can again calculate the error estimate. If the error starts to increase, the procedure stops; otherwise, we continue with more and more EOFs. To make sure that the process does not stop at a local minimum, we could also continue the procedure until reaching a prescribed number of EOFs, never more than \( q \), and retain the optimal set. Once the number of EOFs that minimizes the error estimate is found, we can fine-tune the calculation of the EOFs and data interpolation by finally injecting into the interpolation procedure the data that were set aside for the cross validation. In this case, one should of course start the final interactive procedure with the interpolated data obtained with the optimal number of EOFs during the cross validation. In the case where there are no missing data, the proposed cross-validation technique reduces to finding the optimal number of EOFs and then in a last step performing the normal EOF decomposition on the data, knowing the optimal truncation.

With the cross-validation technique, we can thus find the optimal number of EOFs and an error estimate of the interpolation procedure (the rms distance between interpolated data and the validation data).

We now will show how the algorithm works on two datasets: a synthetic one and an AVHRR SST dataset in the Adriatic Sea.

### 4. Test cases

#### a. Synthetic data

A first exercise will be done on a synthetic dataset of a spatial 1D problem with time evolution. The field is shown in Fig. 1 in \((x, t)\) space. The signal consists of a sum of a deterministic field and a completely random one, a normal distribution of known intensity (standard deviation), and zero mean. The deterministic part is given as follows:

\[
(X_0)_{ij} = \sin(x_i) \sin(t_j) + \sin(2.1x_i) \sin(2.1t_j)
\]

+ \sin(3.1x_i) \sin(3.1t_j) + \tanh(x_i) \cos(t_j)
+ \tanh(2x_i) \cos(2.1t_j) + \tanh(4x_i) \cos(0.1t_j)
+ \tanh(2.4x_i) \cos(1.1t_j) + \tanh(x_i + t_j)
+ \tanh(x_i + 2t_j),
\]

with \(x_i = i2\pi/m, t_j = j2\pi/n\).

The objective is to retrieve from \(X\), the best estimates of \(X\) and its relevant empirical orthogonal functions. In this synthetic dataset, we thus know the true solution and the noise, so that we can make all necessary statistical comparisons and error analyses. Here, we decide to normalize the error and we can choose either \(\|X\|\) (when known, as in our synthetic example), or \(\|X\|\) if only observations are available. Thus, a normalized error of 0.01 means an error with less than 1% energy of the signal. In the present example we normalize by \(\|X\|\).

We first apply our method to a case with a 20% noise level and 10% cloud coverage as illustrated in Fig. 1.

From the analysis of the singular values in this case (Table 1) the following points are clear:

- The noise added to the data increases all singular values \(\rho^2\) compared to the unperturbed field \((\rho^2)\), most significantly those from the fifth singular value on (cf. lines 1 and 2 in Table 1). This means that the first few singular values are slightly modified by the noise, or equivalently, that some of the noise structure was interpreted as EOFs signals. In view of this, there will be no chance to exactly recover the first four singular values \(\rho\).
- The singular values obtained from the first guess underestimate the energy of the first modes (because we impose zero values, not taking into account the correlation between neighbors) and overestimates the higher EOFs’ energy.
- The reconstructed data \((\rho^2)\) have energy levels for the EOFs much in line with the energy of the EOFs including the noise interpreted as signal \((\rho^2)\), but the cross validation was able to detect that the optimal number of EOFs was four. Then, cutting beyond EOF4 eliminates the noise associated with the higher EOFs but it is impossible to eliminate the noise interpreted as signal in the first EOFs. Therefore the reconstructed data have the energy level of the first EOFs comparable to the noisy field.

We can also ask ourselves if the EOF structures are better represented after the interpolation method was applied. For mode 1, we have the following measures of errors on the EOF structure: \(|u_0 - u_0^*| = 0.0027\), \(|u_0 - u_0^*| = 0.0035\), \(|u_0 - u_0^*| = 0.0005\), \(|u_{02} - u_{02}^*| = 0.0042\), and \(|u_{02} - u_{02}^*| = 0.0059\), which shows that the reconstructed EOF \(u_0\) is a better representation of the filtered EOF than the initial guess \(u_0\) and that it is closer to the EOF including the interpreted noise than to the filtered EOF. Therefore we can conclude that the interpolation clearly helped to improve the EOF representation.
Then we modify the importance of the cloud coverage and noise level, in order to assess if the method systematically improves the estimates of the EOFs. Different errors are shown in Table 2 for different combinations of (uniform but random) cloud coverage and noise/signal ratios.

From Table 2 we can see the following.

- The truncated EOF reconstructions from the analyzed field \( \mathbf{X}_a \) are always a better representation of the real truncated EOF representation \( \mathbf{X}_t \) than the initial guess EOF representation \( \mathbf{X}_0 \) (comparing the last two columns), except when there are no clouds, in which case both are identical.
- The untruncated analyzed field \( \mathbf{X}_a \) is always a better representation of the true untruncated field \( \mathbf{X}_t \) than the initial guess \( \mathbf{X}_0 \), (columns 4 and 5), except again in the case when there are no clouds, in which case they are identical.
- The truncated EOF series are always a better representation of the true field than the untruncated series, be it the initial guess solution (cf. columns 5 and 6) or the analyzed field (cf. columns 3 and 4).

Table 1. First seven singular values for exact field (\( \rho_t \)), noisy field (\( \rho_{pr} \)), clouded field filled with zeros (\( \rho_0 \)), and analyzed field (\( \rho_a \)).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_t )</td>
<td>107.692</td>
<td>35.877</td>
<td>34.715</td>
<td>31.924</td>
<td>8.263</td>
<td>1.395</td>
<td>0.289</td>
</tr>
<tr>
<td>( \rho_{pr} )</td>
<td>108.954</td>
<td>37.390</td>
<td>36.663</td>
<td>34.208</td>
<td>14.941</td>
<td>14.550</td>
<td>14.125</td>
</tr>
<tr>
<td>( \rho_0 )</td>
<td>98.331</td>
<td>34.451</td>
<td>33.419</td>
<td>32.376</td>
<td>16.678</td>
<td>15.918</td>
<td>15.683</td>
</tr>
<tr>
<td>( \rho_a )</td>
<td>108.786</td>
<td>37.567</td>
<td>36.581</td>
<td>35.174</td>
<td>14.317</td>
<td>13.600</td>
<td>13.062</td>
</tr>
</tbody>
</table>
The effect of repeating the iterations can be analyzed by comparing the calculations with a repeated iteration compared to a computation where iterations were limited to one application of the iterations (for the same number of EOFs). Errors $\|X_n - X\|^2$ and $\|X'_n - X'_n\|^2$ (which measure how well we really reproduce the significant part of the real data) were reduced typically by 30% when allowing the iterations to proceed.

Looking now qualitatively at the way the method recreates the fields from the incomplete information, we observe (Fig. 1) that for a moderate noise/signal level and a cloud coverage of 10%, the reconstruction performs quite well, except for some higher-frequency signals imposed by the noise (in this case, truncating the EOF series earlier gives worse results because too much information is then eliminated, which shows that the cross-validation technique was able to pick up the best truncation). Increasing the noise (Fig. 2) shows that indeed the noise is responsible for the higher-frequency signals in the reconstruction. For a noise/signal ratio of 0.1 (Fig. 3) even without cloud cover, the optimal EOF truncation filters only part of the noise. For lower noise, but a 50% cloud cover (Fig. 4), the reconstruction performs surprisingly well in view of the very fuzzy and incomplete original image. When the noise/signal ratio is decreased (Fig. 5), the reconstruction is still enhanced. For very low noise levels, cloud coverages of more than 75% (Fig. 6) still include the most relevant information that the EOF-based analysis method is able to extract. For even higher cloud coverage (Fig. 7) we see the problem that at some locations or moments there are no data at all. This clearly shows up in the EOF reconstruction and is due to the fact that no correlation information at all is available at these moments or points. Since this is the only information the EOF method uses (and no other a priori information in the correlation functions), it is impossible to infer the missing data when the whole time series is missing at some point (or one spatial pattern at one moment).

Concerning the detection of the optimal number of EOFs to retain, Fig. 8 shows the real error associated with the EOF truncation and error estimation from cross validation. In order to verify the stability of the truncation with respect to the random choice of the data put aside for cross validation, we repeated the experiment five times, with different random points and even a different number of data points set aside. In all cases the optimal number of EOFs detected was four, which corresponds to the minimum error. The difference in the absolute error estimation simply stems from the fact that, as already mentioned several times, the significant EOFs contain some interpreted noise as confirmed by the analysis of the singular values in the presence and absence of noise. Fig. 9 shows that the signal with noise has added some energy to the significant EOFs. We can also observe that on a subjective basis, the truncation

<table>
<thead>
<tr>
<th>$n$</th>
<th>N/S</th>
<th>$|X_n - X|^2$</th>
<th>$|X'_n - X'_n|^2$</th>
<th>$|X_0 - X|^2$</th>
<th>$|X_0 - X'_n|^2$</th>
<th>$|X'_n - X|^2$</th>
<th>$|X'_n - X'_n|^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.258</td>
<td>0.045</td>
<td>0.192</td>
<td>0.263</td>
<td>0.057</td>
<td>0.045</td>
<td>0.056</td>
</tr>
<tr>
<td>0.5</td>
<td>0.120</td>
<td>0.036</td>
<td>0.076</td>
<td>0.526</td>
<td>0.297</td>
<td>0.032</td>
<td>0.292</td>
</tr>
<tr>
<td>0.0</td>
<td>0.744</td>
<td>0.061</td>
<td>0.431</td>
<td>0.431</td>
<td>0.061</td>
<td>0.059</td>
<td>0.059</td>
</tr>
<tr>
<td>0.5</td>
<td>0.023</td>
<td>0.011</td>
<td>0.017</td>
<td>0.496</td>
<td>0.292</td>
<td>0.006</td>
<td>0.287</td>
</tr>
<tr>
<td>0.77</td>
<td>0.0015</td>
<td>0.033</td>
<td>0.033</td>
<td>0.758</td>
<td>0.671</td>
<td>0.033</td>
<td>0.671</td>
</tr>
<tr>
<td>0.83</td>
<td>0.00036</td>
<td>0.210</td>
<td>0.210</td>
<td>0.760</td>
<td>0.714</td>
<td>0.208</td>
<td>0.710</td>
</tr>
<tr>
<td>0</td>
<td>0.00036</td>
<td>0.00007</td>
<td>0.00037</td>
<td>0.00037</td>
<td>0.00007</td>
<td>0.00007</td>
<td>0.00007</td>
</tr>
</tbody>
</table>
with four EOFs would be the most natural, since only from EOF5 on are differences in the singular values between the noisy data and real data important. We also note that changing the random dataset put aside for the cross validation slightly modified the error estimate as well as the real error. For practical purposes, changing the random set of data put aside is thus an easy way to verify the robustness of the truncation. If for some reason there is then a doubt on the robustness, the cross validation presented here (simultaneously putting aside a number of data points) could be replaced by classical cross validation. The latter sets aside only a single data point (which makes the analysis less sensitive to the data elimination) and then repeats the experiment a sufficient number of times until a statistically significant measure of the error is obtained. This approach is then even more robust at the price of a higher number of analyses.

Concerning the present synthetic dataset, we have to mention that in reality cloud coverage is of course not random and that when larger patches of points are missing some degradations could appear. On the other hand, we did not implement any algorithm here to discard spatial points with too few data (or moments with full cloud coverage), so that some lines or columns in the reconstruction clearly show up as being underdetermined.

For comparison with other methods that deal with missing data, we use a simple method that is classically used and that linearly interpolates the missing data in space. No interpolation in time is done, since most of the time, satellite images are treated individually during the filling procedure. We can then compare our EOF-based interpolation and EOF truncation (Fig. 10) with the linearly interpolated field and its associated truncated EOF representation (Fig. 11, where the same number of EOFs is used). In the latter case, the systematic effects of linear interpolation are still visible in the EOF truncated (and hopefully filtered) reconstruction. One could argue that the improvement made by using the
new method is not obvious, but besides the fact that the methods allow to detection of the optimal number of EOFs, a quantitative analysis of the error reveals the advantage of the new method. In the case presented here the error $\|X - \hat{X}\|^2$ decreases from 0.067 to 0.024, while the error $\|X - \hat{X}\|^2$ decreases from 0.064 to 0.020 when using the EOF interpolation rather than a spatial linear interpolation.

b. Real satellite data

As an illustrative example to demonstrate how the method behaves on real data and nonuniformly distributed cloud coverage, we will use an AVHRR dataset available from Ifremer (and available online at http://satftp.soest.hawaii.edu/adriatic/Adriatic/html/tecdoc.htm) from over the Adriatic Sea.

Seventy-one images of 400 pixels each centered in the middle of the Adriatic Sea (Fig. 12) from Julian day 130 to 294 of were retained for the present testing purposes with images at night (around midnight or 0200 LT) in order to avoid problems related to skin temperature.

As seen from the cloudy sequence (Fig. 13), the cloud coverage is not at all uniform, some pictures being more than half covered by clouds, and noisy patterns in the data are obvious. We therefore expect the present sequence to be a difficult test for the method. After application, the method selects an optimal truncation with nine EOFs. When reconstructing the data with these nine EOFs (Fig. 14), a clearly filtered sequence is visible, while the regions where clouds were filled in are almost impossible to detect (except maybe for one or two pictures in the middle of the sequence, where a persistent large cloud coverage can be seen). The benefit of updating the EOFs is clearly seen by looking at Fig. 15.

\footnote{Optimized versions with larger datasets are surely possible in view of the computational cost being roughly $[\min(n, m)^2](m + n)$.
Indeed, the original EOFs are neither able to eliminate the cloud signature nor the noise in the data. As in the synthetic data example, the choice of the optimal truncation by the cross-validation error estimator is quite robust (Fig. 16); however there is an indication that the error sequence contains some local minima. Our interpretation of the EOF modifications is also confirmed: Fig. 17 shows that the most significant EOF has increased energy compared to the initial guess, while the higher modes have a decreased energy after readjustments. Looking at the evolution of the temperature at a given location for the initial data and the reconstructed version in Fig. 18, the effects of both filtering and interpolation by the EOFs are obvious.

5. Discussion

Two improvements to classic EOF analysis were presented herein:

1) An automatic, parameter-free, and self-consistent filling mechanism was proposed. Compared to a classic interpolation method (as optimal interpolation), the use of the EOFs generated by the data themselves has the advantage of not needing any a priori knowledge of a correlation function or correlation length. Moreover, since it is based solely on the data, inhomogeneities or nonisotropic behaviors are automatically taken into account (as long as they are in the available data) during the interpolation.
2) A cross validation that allows for establishing the optimal cutoff in EOF development (number of relevant EOFs to be retained) and the prediction error in the filled data gaps (from the cross validation estimator) were introduced.

Here computational efficiency considerations were not dealt with, since if only a few eigenvalues are needed, efficient eigenvalue solvers exist (compared to objective analysis used to fill the missing data [cost proportional to \(nm^3\) if done on space]), the present interpolation/EOF decompositions are less time consuming [roughly, \((mn)\left\{\min(m,n)\right\}^{2.5}\) for the analysis of the full range of possible \(N\) without optimization]. Moreover, the number of iterations could probably be reduced by using a better first guess. One could, for example, use any other classical method (be it filling methods or estimations of EOFs from the data correlation matrix) before actually starting the iterations.

We can also recall that with an enhanced and self-consistent filling mechanism, we are also improving the representation of the EOFs, meaning that any method relying on EOFs (e.g., some Kalman filters or predictions tools based on satellite images) will likely benefit from this better representation. For the method to have a chance to work, one needs, for each moment, at least a sufficiently large number of data points (otherwise one should drop the whole picture) and for each spatial point a sufficient amount of data in time (otherwise one should discard the point from the analysis). In both cases, disregarding a complete picture or a pixel can easily be dealt with. In the case where an image at a given moment is not available amounts to using nonuniform time steps between the data, so the methods to deal with nonuniform data can then be used.

Acknowledgments. European projects MEDAR (MAS3-CT98-0174) and SOFT (EVK3-CT-2000-00028) supported this work. M. Rixen presently benefits from a Marie-Curie fellowship at the Southampton Oceanographic Center (SOC). The National Fund for Scientific Research, Belgium, is acknowledged for the financing of supercomputer and the position of the first author. Two anonymous reviewers made constructive comments and are acknowledged. This is MARE publication MARE003. (Color images can be downloaded from http://modb.oce.be/personal/beckers/publications/cv.html.)

**APPENDIX A**

**Transfer of Variance between Different EOFs and Changes of Structure during Iterations**

At the first iteration with 0 values at the missing data points, we get eigenvectors and singular values \(\rho\) that satisfy

\[
X_i v_i = \rho_i u_i \quad \text{and} \quad (A1)
\]

\[
X_i^* u_i = \rho_i v_i \quad (A2)
\]

When the matrix on which the SVD is performed is perturbed by using Eq. (12), the eigenvalues and eigenvectors according to appendix B, and the singular values \(\rho\), are modified by the amount \(r\):

\[
X_i v_i = \rho_i u_i \quad \text{and} \quad (A1)
\]

\[
X_i^* u_i = \rho_i v_i \quad (A2)
\]
Fig. 11. Linear interpolation and EOF truncation: N/S = 0.15, n = 0.50 cloud cover. (left) clouded field interpolated linearly in space (but not in time), and (right) EOF truncated field.

\[ 2r_i = v_i^* \delta X^* u_i + u_i^* \delta X v_i, \]  

(A3)

but since

\[ (\delta X)_{ij} = \sum_{(i,j) \in I} \sum_{k=1}^N \rho_k(u_{ik})(v_{jk}) \delta_{ij}, \]  

(A4)

we can rewrite

\[ 2r_i = \sum_{(i,j) \in I} \sum_{k=1}^N \rho_k(u_{ik})(v_{jk}) \delta_{ij}, \]  

\[ + (u_i^*),_j (v_i^*),_j \delta_{ij}(v_i),_j, \]  

\[ = \sum_{(i,j) \in I} \sum_{k=1}^N \rho_k(u_{ik})(v_{jk}) \delta_{ij}(v_i),_j, \]  

\[ + (u_i^*),_j (u_i^*),_j (v_i^*),_j (v_i^*),_j, \]  

(A5)

The change in the EOF fields can also be assessed in this way and we find the following after using the development of appendix B and some matrix manipulation:

\[ (\rho_i^2 - \rho_j^2) \alpha_i^e = \sum_{(i,j) \in I} \sum_{k=1}^N \rho_k((v_i^*),_j (v_i^*),_j (u_i^*),_j (u_i^*),_j), \]  

\[ + \rho_k(u_i^*),_j (u_i^*),_j (v_i^*),_j (v_i^*),_j, \]  

(A6)

where \( \alpha_i^e \) is basically the transfer of information from mode \( l \) into mode \( i \). Mode \( i \) is therefore mostly influenced by the modes with similar singular values (\( \rho_i \approx \rho_j \)) or strong interactions between modes at the missing data points.

To further understand how these modifications act, suppose that we have a single missing data point, that we use only one EOF (\( N = 1 \)), and that we are at the first iteration of our interpolation process (we made the EOF decomposition of \( X_0 \)), so that at the missing data point before updating we have \( (X_0)_{ij} = 0 \). But then the first singular value changes according to Eq. (A5) is

\[ r_i = \rho_1((v_1^*),_j (v_1^*),_j (u_1^*),_j (u_1^*),_j) \geq 0. \]  

(A7)

This shows that the first EOF number will increase the energy of mode 1 after the interpolation is done, except in the case when the corresponding EOF is zero.

Fig. 12. Location of the AVHRR images rectangle in the Adriatic Sea used in this study.
at the missing data point. Otherwise the term is always positive and simply the product of the temporal and spatial EOF amplitude at the missing data point. This interpretation shows that by replacing the missing data value by the first EOF guess, we add energy to the system, which is simply the energy of the first EOF at that location.

For the other singular values, we can exploit the fact that

\[ (X_{ij}) = \sum_{k=1}^{c} \rho_k(u_i) (v_j^*) \quad \text{or equivalently} \]

\[ (X_{ij}) = 0 = \sum_{k=1}^{c} \rho_k(u_i) (v_j^*) \quad \text{(A8)} \]
Fig. 14. Reconstructed sequence of images in Fig. 13 with optimal number of EOFs and the proposed algorithm including the changed EOFs.

\[ \rho_i(u_i, v_i) = \frac{1}{\sqrt{\lambda_i}} \rho_i^{(u)}(v_i) + \rho_i^{(v)}(u_i), \quad (A9) \]

so that for \( i > 1 \)

\[ 2r_i = \rho_i[(v_i^*)_i, (u_i^*)_i, (u_i)_i] + (u_i^*)_i, (v_i^*)_i, (v_i)_i]. \quad (A10) \]

The contributions to this sum can have any sign, except the contribution \( k = i \), which is included here; the latter
Fig. 15. Reconstructed sequence of images in Fig. 13 with the optimal number of EOFs, but using unmodified EOFs from the first guess $x = 0$.

contribution is always negative (except again if the corresponding mode has a zero amplitude at the missing data point) and corresponds to the fact that initially, to force the value to zero, higher EOFs had seen their amplitude increased in order to be able to satisfy the initial guess (which is inconsistent with the large-scale EOFs present). Now the correction tries to reduce the energy of the corresponding mode. For the other contributions $k \neq i$, their amplitudes decrease with $k$ since the singular values are ordered with a decreasing amplitude. Furthermore, since on average, EOFs are decorrelated, the different contributions will compensate somehow, except the one for $k = i$, which is always negative. For a single mode, we thus expect mode 1 to
see its energy increased and the energy of the others decreased after the first replacement.

When we use $N$ modes, instead of using a single EOF for replacing the missing data values, the same analysis can be done. For a mode $i \leq N$ we have

$$2r_i = \sum_{k=1}^{N} \rho_i[(v_i \mu^*_j)(u_i \mu_j) + (u_i \mu^*_j)(u_i \mu_j)]$$

$$= \sum_{k=1}^{N} \rho_i[(v_i \mu^*_j)(v_i \mu_j) + (u_i \mu^*_j)(u_i \mu_j)]$$

$$+ \rho_i[(v_i \mu^*_j)(u_i \mu_j)].$$  \hspace{1cm} (A11)

The last term being always positive and the others of changing signs, we therefore expect the modes $i < N$ to see their energy increased.

For $i > N$, we can again use

$$\sum_{k=1}^{N} \rho_i[(u_i \mu^*_j)(v_i \mu_j)] = \sum_{k=N+1}^{r} \rho_i[(u_i \mu^*_j)(v_i \mu_j)] + (X_i \mu^*_j) \hspace{1cm} (A12)$$

to show that the contribution to the $r_i (i > N)$ is always decreased by the component $k = i$, while the other ones are of unspecified sign. Also statistically, when more missing data points are present, we expect that

$$\sum_{(i, j) \in I} (u_i \mu^*_j) \rightarrow 0, \quad i \in I \rightarrow [1, m] \hspace{1cm} (A13)$$

because of the orthogonality of the eigenvectors and the probability that when summing up all data points, the sum will cover the full range of $i$ values. The same applies for the temporal sum

$$\sum_{(i, j) \in J} (v_i \mu^*_j) \rightarrow 0, \quad j \in J \rightarrow [1, n]. \hspace{1cm} (A14)$$
We can also look at how the EOF structures are modified. In fact, it appears that the EOFs are slightly rotated so as to diminish the contribution to the dominant EOFs from the “noisy” EOFs partly introduced by the initial filling by zero values. Indeed for a single missing data point we have, according to Eq. (6),

\[(\rho_i^2 - \rho^2)\alpha_i = \sum_{k=1}^N \rho_k (\rho_k (v_k^*)^T (u_k^*) (u_k^*)^T (u_k)) \alpha_i + \rho_k (u_k^*)^T (u_k) (v_k^*)^T (v_k)^T \alpha_i \]  

+ \rho_k (u_k^*)^T (u_k) (v_k^*)^T (v_k)^T \alpha_i. \quad (A15)

**APPENDIX B**

**Perturbation Theory on Singular Value Decomposition**

A question that arises when dealing with matrices constructed from real data concerns the effects of noise in the data. Therefore we can analyze the problem of singular value decomposition (SVD) of a matrix subjected to perturbations. Suppose that we have found the solution of the SVD eigenvalue problem for an unperturbed matrix \(A\) so that we have

\[Av_i = \rho_i u_i \quad \text{and} \quad A^* u_i = \rho_i v_i \]  

or equivalently

\[AA^* u_i = \rho_i u_i \quad \text{and} \quad A^* A v_i = \rho_i v_i. \quad (B1)

Suppose we now perturb the matrix so that a new matrix \(A + \delta A\) is obtained with \(|\delta| \ll 1\). We can then search for eigenvalues and singular values of the form

\[u_i + \delta u_i, \quad v_i + \delta v_i, \quad \text{and} \quad \rho_i + \delta \rho_i. \quad (B7)

Neglecting higher-order terms\(^{a}\) in \(\delta\), using the fact that \(u_i, v_i, \text{and} \rho_i\) are solutions of the unperturbed problem, we obtain the following set of equations for the perturbations:

\[(A B^* + B A^*) u_i + A A^* p_i = 2 \rho_i r_i u_i + \rho_i^2 p_i, \quad \text{and} \quad (B8)\]

\[(A B^* + B A^*) v_i + A A^* q_i = 2 \rho_i r_i v_i + \rho_i^2 q_i. \quad (B9)\]

In addition, since the eigenvectors are orthogonal and normalized \(u_i^* u_i = \delta_{ii}\), and we have (by perturbing this orthonormality condition)

\[u_i^* p_i = -p_i^* u_i \quad \text{and} \quad v_i^* q_i = -q_i^* v_i. \quad (B10)\]

Multiplying Eq. (B8) by \(u_i^*\) leads, by using Eqs. (B2), (B1), (B10), and (B11), to

\[2r_i = v_i^* B^* u_i + u_i^* B v_i, \quad (B12)\]

which takes real values as we can easily verify. Since \(p_i\) is orthogonal to \(u_i\) by virtue of Eq. (B10), we can develop it as follows:

\[p_i = \sum_{k=1}^N \alpha_k u_k, \quad (B13)\]

Injecting this equation into Eq. (B2) and applying \(u_i^*\) to it leads, by again using Eqs. (B2), (B1), (B10), and (B11), to

\[\alpha_i = -\alpha_i^{(i)} = \frac{\rho_i v_i^* B^* u_i + \rho_i u_i^* B v_i}{\rho_i^2 - \rho_i^2}. \quad (B14)\]

which gives the contribution of mode \(i\) on the change to mode \(i\). If \(\rho_i \sim \rho_i\), the change in EOFs is likely to be large, reflecting the fact that for two identical eigenvalues, any combination of the corresponding eigenvectors is also an eigenvector. When the eigenvalues are close, it means that a perturbation can transfer the contribution between these two EOFs very easily. Similarly we can calculate the perturbations on \(v_i\):

\[q_i = \sum_{k=1}^N \beta_k v_k \quad \text{and} \quad (B15)\]

\[\beta_i = -\beta_i^{(i)} = \frac{\rho_i u_i^* B v_i + \rho_i v_i^* B^* u_i}{\rho_i^2 - \rho_i^2}. \quad (B16)\]

**REFERENCES**


---

\(^{a}\) This requires that \(AB^*\) is essentially different from 0 meaning that the perturbation matrix must not be orthogonal to the initial matrix.


