Ocean Sci., 2, 183–199, 2006
www.ocean-sci.net/2/183/2006/

Ocean Science

DINEOF reconstruction of clouded images including error maps – application to the Sea-Surface Temperature around Corsican Island

J.-M. Beckers[1,3], **A. Barth**[2], **and A. Alvera-Azcárate**[2]

[1]GeoHydrodynamics and Environment Research, MARE, University of Liège, Sart-Tilman B5, 4000 Liège, Belgium
[2]College of Marine Science, University of South Florida, 140 7th Avenue South, St. Petersburg, Florida 33701, USA
[3]Honorary Research Associate, National Fund for Scientific Research, Belgium

Received: 23 May 2006 – Published in Ocean Sci. Discuss.: 10 July 2006
Revised: 25 September 2006 – Accepted: 16 October 2006 – Published: 18 October 2006

Abstract. We present an extension to the Data INterpolating Empirical Orthogonal Functions (DINEOF) technique which allows not only to fill in clouded images but also to provide an estimation of the error covariance of the reconstruction. This additional information is obtained by an analogy with optimal interpolation. It is shown that the error fields can be obtained with a clever rearrangement of calculations at a cost comparable to that of the interpolation itself. The method is presented on the reconstruction of sea-surface temperature in the Ligurian Sea and around the Corsican Island (Mediterranean Sea), including the calculation of inter-annual variability of average surface values and their expected errors. The application shows that the error fields are not only able to reflect the data-coverage structure but also the covariances of the physical fields.

1 Introduction

When dealing with a data set containing missing or unreliable data, a general approach to fill in the missing data is the use of objective-analysis methods, in particular optimal interpolation (OI), (e.g., von Storch and Zwiers, 1999; Gomis and Pedder, 2005). The latter leads to an interpolated field with minimal expected error variance, certainly a desirable property. The optimality of the approach relies however on the assumption that correlation functions and the signal/noise ratio of the data are perfectly known (e.g., Rixen et al., 2000; Gomis et al., 2001). In practise ad hoc parametric correlation functions are often used and parameters in the best case are only calibrated for the specific data set, so that optimality in the statistical sense is rapidly lost.

When a series of clouded images is to be filled in, the repeated observation on a single grid can be exploited to improve the specification of the covariance functions. This was done in the development of the Data INterpolating Empirical Orthogonal Functions method (DINEOF) (Beckers and Rixen, 2003; Alvera-Azcárate et al., 2005; Alvera-Azcárate et al., 2006, where the time series of images provided a mean to calculate principal components of incomplete data as eigenvectors of a covariance matrix, and simultaneously filling in the missing data. The extension to an EOF decomposition version known as Singular Spectrum Analysis (e.g., Vautard et al., 1992) was also used to reconstruct time-series of river discharges (Kondrashov et al., 2005) and tidal gauge data (Bergant et al., 2005). The DINEOF interpolation was shown to provide similar results than optimal interpolation, being however incomparably faster. Also, DINEOF does not need any a priori information, contrary to OI in its most widely used form with prescribed covariance functions. The DINEOF method has also been compared to kriging methods in the framework of computational fluid dynamics and was found to be more accurate than the latter for high temporal resolution and not too large (i.e. typically 50%) data gaps (Gunes et al., 2006). DINEOF is however up to now hampered by the fact that contrary to OI, no local error estimates at each grid point can be provided. Only a global error can be calculated by DINEOF exploiting a cross-validation technique, while OI allows to draw spatial error maps (e.g., Shen et al., 1998). The present paper aims at closing the gap, providing local error maps for DINEOF. As a byproduct, it will be shown how OI can be combined with DINEOF calculations so that when using covariance matrix estimations from DINEOF it reduces drastically the calculations needed by standard OI.

Since the validation of the DINEOF analysis itself has already been performed thorougly in previous papers, we will focus here on the error fields instead. The paper is organized as follows. In Sects. 2 and 3 we formulate OI and DINEOF. We then show in Sect. 4 that a very efficient least-square fit of EOF amplitudes to an observed subset of data is equivalent to

Correspondence to: J.-M. Beckers
(jm.beckers@ulg.ac.be)

Published by Copernicus GmbH on behalf of the European Geosciences Union.

an OI if the filtered covariance matrix of DINEOF is used as the ad hoc covariance matrix of OI. This result is then used in Sect. 5 to use the statistically derived error estimates of OI as error fields for DINEOF. The method is then tested on a data set consisting of AVHRR Sea-Surface Temperature (SST) in the Mediterranean Sea around Corsica (Sect. 6). This section proves the efficiency of the method and the relevance of the error fields. The conclusions finish with some suggestions of additional improvements that could be included in the DINEOF tool.

2 Optimal interpolation

Optimal Interpolation (e.g., Daley, 1991) aims at minimising the expected error variance ϵ^2 at a given position r of the interpolated field φ compared to the true field φ_t

$$\epsilon^2(\mathsf{r}) = \overline{[\varphi(\mathsf{r}) - \varphi_t(\mathsf{r})]^2}, \tag{1}$$

with $\bar{\varphi}$ being the average of φ in a statistical sense, i.e., for repeated realisations. All fields are considered anomalies so that their averages are zero, and if considered adequate, trends or cycles can be removed prior to any treatment. The linear combination of the N_d available data d_i located in r_i, i=1, ..., N_d and grouped into a column vector $\boldsymbol{d}$ that minimises the expected error variance in location r is given by

$$\varphi(\mathsf{r}) = \sum_{i=1}^{N_d} w_i(\mathsf{r})\ d_i = \mathbf{w}^{\mathsf{T}}\boldsymbol{d} = \mathbf{c}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{d}, \tag{2}$$

where $^{\mathsf{T}}$ indicates a transposed matrix or vector and where we define a covariance matrix $\mathbf{D}$ between data points

$$\mathbf{D} = \overline{\boldsymbol{d}\,\boldsymbol{d}^{\mathsf{T}}} \tag{3}$$

and the covariance $\mathbf{c}$ of all data points with the target field at the point r in which the interpolation is calculated:

$$\mathbf{c} = \overline{\varphi_t(\mathsf{r})\,\boldsymbol{d}}. \tag{4}$$

The expected error variance itself is minimal and has the following value

$$\min \epsilon^2(\mathsf{r}) = \overline{\varphi_t(\mathsf{r})^2} - \mathbf{c}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{c}, \tag{5}$$

directly providing the error estimates in any desired location r after analysis by Eq. (2). In order for the method to be applicable, there remains to determine the covariances involved in the formulation.

In standard OI, decomposing the data d_i=ϵ_i+$\varphi_t(\mathsf{r}_i)$ as the sum of observational (or representativity) errors and the true field, the covariance matrix $\mathbf{D}$ is the sum of the observational error-covariance matrix $\mathbf{R}$ and the target field-covariance matrix $\mathbf{B}$ assuming observational errors and the target field to be uncorrelated. An element i, j of $\mathbf{B}$ is then given by $\overline{\varphi_t(\mathsf{r}_i)\varphi_t(\mathsf{r}_j)}$ and similarly for the observational error. Introducing decomposition $\mathbf{D}$=$\mathbf{R}$+$\mathbf{B}$ into Eq. (2) leads to the classical optimal interpolation formula

$$\varphi = \mathbf{c}^{\mathsf{T}}\,(\mathbf{B} + \mathbf{R})^{-1}\,\boldsymbol{d}, \tag{6}$$

with $\mathbf{c}$ being the covariance between data points and the point of interpolation and $\mathbf{B}$ the field-covariance matrix also called background error covariance matrix containing covariances between data locations. The latter is generally calculated from predefined correlation functions depending on the distance between data points (e.g., Emery and Thomson, 1997). For uncorrelated and homogeneous data errors of variance μ^2, the corresponding error-covariance matrix has the simplified diagonal form

$$\mathbf{R} = \mu^2\,\mathbf{I}, \tag{7}$$

which is used in most applications and where $\mathbf{I}$ is the identity matrix. In the following, the signal variance is

$$\sigma^2 = \left\langle \overline{\varphi_t(\mathsf{r})^2} \right\rangle, \tag{8}$$

where <> stands for a spatial average and σ^2/μ^2 is the signal/noise ratio.

Now suppose we look at a single image and would like to interpolate the missing data under clouds. The classical approach would be to define a covariance function, estimate a signal to noise ratio and then apply the OI algorithm. In its original and statistically optimal form, this would require the inversion of a matrix of size N_d=m_p, m_p being the number of unclouded or present pixels. This inversion can be quite time-consuming: a SeaWiFS scene of 1000×2000 pixels with 50% cloud coverage would require the inversion of a system of 10^6 equations with 10^6 unknowns. This is a major challenge since the matrix to be inversed is not banded. Therefore, optimal interpolation is in most cases downgraded by using only data points within a given distance from the point in which to interpolate, neglecting teleconnections.

3 DINEOF

DINEOF, instead of using the direct minimisation of expected error covariance as the objective of the interpolation, uses data-based principal components (called EOFs hereafter) to infer the missing data. To do so, we realise that EOFs can be obtained from a Singular Value Decomposition (SVD) representation of the data matrix $\mathbf{X}$. Each column of $\mathbf{X}$ contains a satellite image stored as a column vector of m pixels, and a pixel of such an image is the data $x_{i,j}$. We suppose we have n images (j=1, ..., n). Then the SVD decomposition reads

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}, \tag{9}$$

where $\mathbf{U}$ contains on each of its columns one of the spatial patterns of the EOFs, the pseudo-diagonal matrix $\mathbf{\Sigma}$ the singular values and $\mathbf{V}$ the temporal components. The SVD decomposition is then truncated to the first N EOFs and provides a filtered version of the data, also at the missing data points. This provides therefore the interpolated values. To calculate EOFs via an SVD, the data matrix needs however to be complete; but to infer the missing data we must know the EOFs, a circular dependence which of course results in an iterative method described in more details in Beckers and Rixen (2003) and Alvera-Azcárate et al. (2005). The number of EOFs to retain in the truncation is obtained by a cross-validation technique, adding artificial clouds in some locations and using as a global error estimate the RMS (root mean square) distance between the known values and the reconstructed ones under the artificial clouds. The optimal number of EOFs is then the one that minimises this error estimate. This method was thorougly tested in Alvera-Azcárate et al. (2005), where a set of 105 images on the Adriatic Sea was reconstructed and compared to in situ data. The method was numerically optimised using a Lanczos solver for the SVD decomposition (Toumazou and Cretaux, 2001), which allows to apply the technique to large sets of data. The accuracy of the method was checked against a classical OI reconstruction. The error obtained by DINEOF was smaller than with OI (0.95°C vs. 2.4°C using 452 independent in situ observations for validation) and DINEOF was able to make the reconstruction of the data set nearly 30 times faster than with OI.

Completely blank images cannot be reconstructed by the original DINEOF approach because no statistical measure can be derived from it. If time-correlation is know, similarly to OI, this can however be exploited to reconstruct images at any time by time-interpolating the EOF amplitudes, using time-correlation functions similar to OI correlation functions.

DINEOF provides as result a Singular Value Decomposition of the data matrix $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$ where $\mathbf{\Sigma}$ contains the singular values ρ_i (ordered as usual with decreasing amplitude) on the diagonal and where $\mathbf{U}$ and $\mathbf{V}$ are normalized according to

$$\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}, \tag{10}$$

$$\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}. \tag{11}$$

We do however only consider the N first EOFs to be significant so that the truncated SVD is our best estimate of the field:

$$\mathbf{X}^N = \mathbf{U}^N\mathbf{\Sigma}^N\mathbf{V}^{N\mathsf{T}}, \tag{12}$$

where $\mathbf{U}^N$ is a $m\times N$ matrix with N columns containing the first N spatial EOFs, $\mathbf{V}^N$ is a $n\times N$ matrix with N columns containing the first N temporal EOFs and $\mathbf{\Sigma}^N$ a diagonal matrix of size $N\times N$ containing the first N singular values ρ. The truncated SVD expansion defines the reconstruction $x^r_{i,j}$ of the field.

Hence we consider that the N retained modes contain the signal and that the remaining modes contain noise. This noise is not necessarily the noise in the sense of OI, because it might still contain some components of the signal. The method is simply not able to distinguish those from noise during the cross-validation. Because the spectrum of the rejected modes is flat, we can however be relatively confident in the separation process. Note that this rejection of higher EOFs is also coherent with the fact that to accurately estimate higher EOFs, very large sample sizes are needed (e.g., North et al., 1982).

If the initial matrix was complete and contained homogeneous noise, we would have

$$\sum_{i=1}^{n}\rho_i^2 = mn\,(\sigma^2+\mu^2) \tag{13}$$

$$\sum_{i=1}^{N}\rho_i^2 = mn\,\sigma^2. \tag{14}$$

This first equality simply expresses the SVD property which states that the sum of all singular values squared equals the total variability of the data (e.g., von Storch and Zwiers, 1999). The second one states that the first N modes account for the variability of the signal if we call the first N modes the signal.

For the matrix with M missing data, we cannot base the calculation of the noise value on the singular values, because the reconstruction is only valid for the first N EOFs. However Eq. (14) remains valid because the first N EOFs are supposed to be correctly specified, otherwise the whole DINEOF reconstruction is at doubt.

To estimate the noise (or variability that is not reconstructed by the EOF expansion), we have a series of points for which data are available before reconstruction (where there are no clouds). The noise can thus be evaluated as the difference between the original values x and the filtered ones x^r

$$\mu^2 = \frac{1}{mn-M}\sum_{x_{ij}\,\text{not missing}}\left(x_{ij}^2 - {x_{ij}^r}^2\right) \tag{15}$$

using only the original data values x_{ij} and the reconstruction x^r_{ij} in the $nm-M$ not missing data points.

4 Least-square fits and Optimal Interpolation

We will now use the covariance matrix from the DINEOF decomposition in an Optimal Interpolation approach. Instead of using a prescribed covariance matrix for OI, we can invoke the ergodic theorem and replace statistical averages by time averages if a sufficiently large amount of images are available. Hence the covariance matrix can be based on our SVD decomposition and the covariance between each couple of

grid points is now calculated as an average over the n images instead of an infinite statistical ensemble[1]:

$$\mathbf{D} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}. \tag{16}$$

This is, however, not a very good estimate of the covariance matrix because we only trust the first N EOFs. If we define scaled spatial EOFs

$$\mathbf{L} = \frac{1}{\sqrt{n}}\mathbf{U}^N\mathbf{\Sigma}^N, \tag{17}$$

where $\mathbf{L}$ is a matrix with N columns, each of which is the spatial EOF scaled by the singular values and (for convenience) by $1/\sqrt{n}$. The N retained significant EOFs lead therefore, exploiting the truncated SVD decomposition and $\mathbf{V}^{N\mathsf{T}}\mathbf{V}^N = \mathbf{I}$, to the field covariance

$$\mathbf{B} = \frac{1}{n}\mathbf{X}^N\mathbf{X}^{N\mathsf{T}} = \mathbf{L}\mathbf{L}^{\mathsf{T}}, \tag{18}$$

since we assumed that the first N EOFs contain signals and the remaining EOFs some noise.

As already mentioned, the observation error covariance cannot be determined by our DINEOF expansion because the higher EOFs are not significant. But if the explained variance is well captured by $\mathbf{B}$, we can try to model the observational errors as being uncorrelated. Knowing the total variance of the data and the reconstructed field variance, we can estimate the noise. In other words, the observational error variance μ^2 is taken to be the variance not retained within the EOF expansion. Assuming the observational error uncorrelated we therefore would model

$$\mathbf{R} = \mu^2\,\mathbf{I}, \tag{19}$$

where μ^2 is given by Eq. (15). The assumption that the errors are uncorrelated might be questioned for satellite data, where atmospheric corrections and associated errors are likely to contain spatial correlations. In this case a non-diagonal matrix should be used, exactly as in the original OI method. We then face however a) the problem of specifying the correlation functions and b) the problem of inverting the covariance matrices. In some cases (e.g., Barth et al., 2006), when errors are correlated at a prescribed scale L, an intermediate complexity approach can be implemented: first, OI is performed with the full error-covariance matrix to serve as a reference solution; then the analysis is repeated with a diagonal error-covariance matrix where the diagonal is, compared to the non-diagonal version, inflated by a factor r . It turns out numerically that when the inflation factor $L^2/\Delta x/\Delta y$ is used, the analysis is closest to the reference solution. This is readily understood in terms of number of "independent data" used during the analysis step. For the present paper, we therefore assume that we can take into account correlated noise by inflating the diagonal form of the error-covariance matrix, the calibration of which will be performed in Sect. 6.3. Should there be better information on the error-covariance structures, a full matrix $\mathbf{R}$ should be used, exactly as in OI.

Having now $\mathbf{R}$, the covariance matrix of the noise unexplained by the first N EOFs and the field covariance matrix $\mathbf{B}$, we can use standard OI on a single image to interpolate everywhere, including missing points and data covered points. Here we assume the points are ordered[2] and the first m_p grid points are present and the remaining $m-m_p=m_m$ are missing. We partition the covariance matrix correspondingly

$$\mathbf{B} = \begin{pmatrix}\mathbf{L}_p\\ \mathbf{L}_m\end{pmatrix}\begin{pmatrix}\mathbf{L}_p^{\mathsf{T}} & \mathbf{L}_m^{\mathsf{T}}\end{pmatrix} = \begin{pmatrix}\mathbf{L}_p\mathbf{L}_p^{\mathsf{T}} & \mathbf{L}_p\mathbf{L}_m^{\mathsf{T}}\\ \mathbf{L}_m\mathbf{L}_p^{\mathsf{T}} & \mathbf{L}_m\mathbf{L}_m^{\mathsf{T}}\end{pmatrix}, \tag{20}$$

where $\mathbf{L}_p$ contains for example the first m_p rows of $\mathbf{L}$, i.e., the EOF values at points for which data are available.

The covariance matrix between data points is then simply

$$\mathbf{B}_p = \mathbf{L}_p\mathbf{L}_p^{\mathsf{T}}. \tag{21}$$

The row i of

$$\begin{pmatrix}\mathbf{L}_p\mathbf{L}_p^{\mathsf{T}}\\ \mathbf{L}_m\mathbf{L}_p^{\mathsf{T}}\end{pmatrix} \tag{22}$$

can be written as $\mathbf{i}^{\mathsf{T}}\mathbf{L}_p^{\mathsf{T}}$, where $\mathbf{i}$ is column array of dimension N×1 containing the values of the N scaled EOFs at grid point i (irrespectively if whether or not the data are missing). We can easily interpret $\mathbf{i}^{\mathsf{T}}\mathbf{L}_p^{\mathsf{T}}$ as the covariance $\mathbf{c}^{\mathsf{T}}(\mathrm{r}_i)$ used in OI. The analysis in point i then provides

$$\varphi_i = \mathbf{i}^{\mathsf{T}}\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{B}_p + \mathbf{R}\right)^{-1}\boldsymbol{d}. \tag{23}$$

In particular for all points with data, we can construct the vector of the analyzed field $\mathbf{x}_p$:

$$\mathbf{x}_p = \mathbf{L}_p\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{B}_p + \mathbf{R}\right)^{-1}\boldsymbol{d}. \tag{24}$$

Similarly, for all missing data points, according to Eq. (2), we must use the covariance between data and missing points applied to the $\left(\mathbf{B}_p+\mathbf{R}\right)^{-1}\boldsymbol{d}$ to calculate

$$\mathbf{x}_m = \mathbf{L}_m\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{B}_p + \mathbf{R}\right)^{-1}\boldsymbol{d}. \tag{25}$$

We see that we can calculate the analyzed field in all points written in a compact form [3]:

$$\begin{aligned}\mathbf{x} &= \begin{pmatrix}\mathbf{L}_p\\ \mathbf{L}_m\end{pmatrix}\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{L}_p\mathbf{L}_p^{\mathsf{T}} + \mu^2\,\mathbf{I}_p\right)^{-1}\boldsymbol{d}\\ &= \mathbf{L}\,\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{L}_p\mathbf{L}_p^{\mathsf{T}} + \mu^2\,\mathbf{I}_p\right)^{-1}\boldsymbol{d}.\end{aligned} \tag{26}$$

[1] Having removed the data mean, the denominator should be $n-1$ for the estimation of the covariance matrix, but the final interpolation result is independent of this scaling.

[2] This is not a restrictive hypothesis, in practise it amounts to use indirect indexing in matrices rather than to perform a sorting before application of the method.

[3] The reader used to data assimilation can recognise the analysis $\mathbf{x}=\mathbf{B}\mathbf{H}^{\mathsf{T}}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathsf{T}}+\mathbf{R}\right)^{-1}\boldsymbol{d}$ where $\mathbf{H}$ is the observation matrix, here containing only a mask of zeros and ones.

Now, assuming the inverse matrix involved in the calculation exists and because of Eq. (A2) from the appendix, this is equivalent to

$$\mathbf{x} = \mathbf{L}\left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\,\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathsf{T}}\boldsymbol{d}. \tag{27}$$

We will now show that this is nothing else than a regularised least-square fit to the first N EOFs trying to find the N components of amplitude column vector $\boldsymbol{a}$ so that $\mathbf{x}$=$\mathbf{L}\boldsymbol{a}$. Indeed, minimizing the distance of the data points to the linear combination of scaled EOFs by solving the (in general overdetermined) problem

$$\mathbf{L}_p\boldsymbol{a} = \boldsymbol{d} \tag{28}$$

is a classical problem (e.g., Lawson and Hanson, 1974) and its regularised solution is

$$\boldsymbol{a} = \left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\,\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathsf{T}}\boldsymbol{d}. \tag{29}$$

This leads directly to Eq. (27) when the reconstruction uses the weights $\boldsymbol{a}$ to combine EOFs everywhere. Hence this is equivalent to OI. The major advantage of Eq. (27) compared to OI is its reduced calculation cost. The matrix inversion asks for N^3 operations in the least-square fit and m_p^3 in standard OI (typically N=20 while m_p=10^6 for satellite images). The construction of the matrix to invert is proportional to $m_p N^2$ for the least-square fit and the remaining matrix multiplications ask for mN operations. Since $m, m_p \gg N$ the dominant cost is $m_p N^2$, several orders of magnitude smaller than m_p^3 for a standard OI.

The gain is due to the fact that we can factorize the data-based covariance matrix because of the SVD decomposition found by DINEOF. Using covariance matrices based only on available data (Boyd et al., 1994; Kaplan et al., 1997; von Storch and Zwiers, 1999; Eslinger et al., 1989) or prescribed covariance functions leads to a full matrix $\mathbf{B}$ and the need to invert the $m_p \times m_p$ matrix.

Error subspace based Kalman filters such as the Reduced Rank Square Root Filter (Verlaan and Heemink, 1997), the Singular Evolutive Extended Kalman filter (Pham et al., 1998) and the Ensemble Square Root Kalman Filter (Evensen, 2004) use an equivalent approach. Since the model error covariance can be decomposed in a similar way as Eq. (18), the analyses in those filters are performed in the low-dimensional error subspace instead of the space containing the observation space. The special structure of the error covariance matrix in DINEOF seems to indicate that it is less general than OI. This is true insofar as the specification of covariance matrices is left open for choice. In reality these covariance matrices should be the real ones, and DINEOF estimates $\mathbf{B}$ from the data themselves.

In practise, in order to construct the matrix to invert, there is no need to partition the matrices into missing and non-missing data points: it is sufficient to use the EOF values only where data are present. The product $\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p$ is for example simply obtained by creating an $N \times N$ matrix using $\mathbf{L}$ with a mask of zeros in missing data points. Even simpler, in the loops which perform the product $\mathbf{L}^{\mathsf{T}}\mathbf{L}$, the use of a simple flag indicating missing data allows to disregard the corresponding contributions and a direct calculation of $\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p$.

5 Error fields

In Alvera-Azcárate et al. (2005) we observed that the least-square fit approach and DINEOF are very close in terms of results. Hence we can use the error-estimates of OI as a proxy for the error-fields of DINEOF, with a subsequent a posteriori verification that the difference between OI and DINEOF reconstruction is smaller than those error fields. To calculate the error field, we would rather like to apply a method similar to the least-square fit instead of an equivalent standard OI error calculation because of the dramatically different problem size. In OI, the error in a given point can be assessed by the analysis of the covariance between this point and data points, see Eq. (5). For a grid point i (located in r_i), this is normally performed as

$$\epsilon^2 = \overline{\varphi_t(\mathbf{r_i})^2} - \mathbf{i}^{\mathsf{T}}\mathbf{L}_p^{\mathsf{T}}\left(\mathbf{L}_p\mathbf{L}_p^{\mathsf{T}} + \mu^2\mathbf{I}_p\right)^{-1}\mathbf{L}_p\mathbf{i}, \tag{30}$$

but we prefer the mathematically equivalent form (see Appendix A),

$$\epsilon^2 = \overline{\varphi_t(\mathbf{r_i})^2} - \mathbf{i}^{\mathsf{T}}\left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\mathbf{i}, \tag{31}$$

leading to a much smaller matrix to be inverted. The local field variance can be estimated as the diagonal component i of $\mathbf{B}$ which is nothing else than

$$\overline{\varphi_t(\mathbf{r_i})^2} = \mathbf{i}^{\mathsf{T}}\mathbf{i}. \tag{32}$$

Then, all we have to do is to calculate once and for all for a given image

$$\begin{aligned}\mathbf{C} &= \mathbf{I} - \left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p \\ &= \mu^2\left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\mathbf{I}_N\right)^{-1}.\end{aligned} \tag{33}$$

To calculate $\mathbf{C}$, we need to invert a matrix of the rather small size $N \times N$ and from there we calculate the error variance in each grid point as the quadratic form

$$\epsilon^2 = \mathbf{i}^{\mathsf{T}}\mathbf{C}\mathbf{i}, \tag{34}$$

demanding mN^2 operations to form the matrix products in $\mathbf{C}$ and N^3 operations to invert as before.

One could also calculate the error-covariance matrix $\mathbf{E}$ of the analysis, from which the local error field is retrieved along the diagonal:

$$\mathbf{E} = \mathbf{L}\mathbf{C}\mathbf{L}^{\mathsf{T}} = \mu^2\mathbf{L}\left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p + \mu^2\mathbf{I}_N\right)^{-1}\mathbf{L}^{\mathsf{T}} = \mathbf{S}\mathbf{S}^{\mathsf{T}}, \tag{35}$$

where $\mathbf{S}=\mathbf{L}\mathbf{C}^{1/2}$ has only N columns and allows therefore an efficient storage and manipulation of the information contained in $\mathbf{E}$. The square root of $\mathbf{C}$ can be calculated efficiently and be re-evaluated almost at no cost for different values of μ if required (see Appendix B).

For very weak or very strong noise we can show (see Appendix C) that the relative error behaves as follows:

– For small μ

$$\frac{\bar{\epsilon^2}}{\sigma^2} \sim \frac{\mu^2}{\sigma^2}\frac{N}{m_p}. \tag{36}$$

– For large μ

$$\frac{\bar{\epsilon^2}}{\sigma^2} \sim 1 - \frac{\sigma^2}{\mu^2}\frac{m_p}{N}. \tag{37}$$

In both situations the factor $\mu^2 N/(m_p\sigma^2)$ appears, which can be interpreted as the ratio of observational errors ($\mu^2\mathbf{I}_N$) versus the background error captured by the EOFs ($\mathbf{L}_p{}^{\mathsf{T}}\mathbf{L}_p$) and hence the relative weights in the analysis step.

In any case, we are now in a position to calculate error estimates in each grid point according to Eq. (34), with a total cost that is proportional to mN^2, both for the construction of $\mathbf{C}$ and the error calculation. As before, in practise, the calculation of $\mathbf{C}$ can be done by an adequate flagging of operations during matrix multiplications instead of preliminary partitioning. In summary, we calculate first the DINEOF decomposition, then an extremely fast objective analysis of each image based on a reformulation into a small least-square fit problem using the DINEOF based covariance matrices, and finally we can generate the OI error map of each image at almost no additional cost compared to the analysis itself.

In addition to the error fields, the error-covariance matrix can also be calculated, particularly efficiently when the square root of $\mathbf{C}$ is calculated. The SST error covariance is for example a necessary information for the calculation of the uncertainty of spatial averages, such as the estimation of the ocean surface heat content. This application can benefit of the DINEOF cloud free SST to integrate over the entire domain. But the estimation of the error variance of the total heat content not only necessitates the error variance but also the error covariance since the error tends to be correlated in space. If $\bar{\phi}=\frac{1}{m}\sum_i x_i$ is the spatial average value of the analysed field, the associated error-variance e^2 is indeed

$$e^2 = \frac{1}{m^2}\sum_{ij} E_{i,j} \tag{38}$$

where $E_{i,j}$ are the covariances found in the error-covariance matrix $\mathbf{E}$ of the analysis. Note that when the errors are homogeneous and uncorrelated, the error-variance of the mean is the local error divided by the number of data points.

We have however still to prove that the use of covariance matrices based on DINEOF leads to physically acceptable results. To do so, we will now test the method on a large data set of SST.

6 Application to Sea-Surface Temperature around the Corsican Island

The method will now be tested in the Mediterranean Sea around Corsica. The circulation in the Ligurian Sea describes a permanent cyclonic gyre, which is more intense in winter (Larnicol et al., 1995). This intensification and the increased transport through the straits can be explained in terms of steric sea level slope and is mainly due to net sea-surface heat flux (Vignudelli et al., 2000).

Two northward currents surrounding the coast of Corsica, the West Corsican Current (WCC) and the East Corsican Current (ECC), join in the Ligurian Sea and form the Northern Current (NC). The NC seasonal cycle is modulated by variations in volume and heat content of the ECC and WCC, and presents its highest transport values in winter (Vignudelli et al., 2003). It has been shown (Orfila et al., 2005) that the SST seasonal cycle in the Ligurian Sea is linked to the North Atlantic Oscillation, which can affect the strength of the winter season. The NC is mainly formed by warm modified Atlantic water, which is separated from the colder central basin by the Liguro-Provençal front. On average, the waters in the central Ligurian Sea are colder than those nearer the coast, the two being separated by the front (see Medar climatology for example, Rixen et al., 2005).

The NC flows south-westward following the French and the Spanish coasts along the continental slope, completing the cyclonic loop. The signal of the NC extends from the north of Corsica to as far as the Catalan Sea (e.g., Astraldi et al., 1999; Millot, 1999). The main currents contributing to the circulation in the Ligurian Sea can be seen in Fig. 1. In the Tyrrhenian Sea, east of Corsica and Sardinia, the orographic effect of the two islands induces a windstress that is responsible for a general cooling east of the Bonifacio strait between the Islands (e.g., Millot and Taupier-Letage, 2005) and a dipole (anticyclonic/cyclonic) structure of the circulation (e.g., Astraldi et al., 1994).

6.1 Description of the data set

AVHRR Pathfinder version 5 SST data from 1 January 1995 to 31 December 2004 have been taken from the Jet Propulsion Laboratory web site (ftp://podaac.jpl.nasa.gov). The data are daily averaged SST maps, and only nighttime passes are used in this study, to avoid daytime surface heating. A region covering the waters around the Corsican Island, in the northwestern Mediterranean Sea has been chosen because of available data and research projects going on in the region (see Fig. 1). Only images containing at least 5% of valid

data are retained, with a maximum of m=5995 data points for a cloud-free image, each data point representing a grid box of 4 km×4 km. From the initial 3653 images, n=2640 are retained using this criteria (about 72% of the initial data). The mean cloud coverage of this data set is 55.2%.

The time and space average of the SST data has been substracted from the observations. With this definition of anomalies, the first EOF will represent the seasonal cycle. The benefit of including the seasonal cycle is that the reconstruction will be able to represent the modulation of the seasonal cycle (like the 2003 heatwave or 2004, the year "without summer" in Europe). But we recognize that there are also drawbacks. If only a few data points are present, the method could produce a winter SST distribution in summer. This is one of the reasons why we do not attempt to reconstruct SST fields with less than 5% data. After taking this precaution, we did not observe such problems.

6.2 SST estimation

This 10-year record of SST data has been reconstructed using DINEOF. We chose such a long SST time series in order to be able to represent some (recurrent) mesoscale features. For the cross-validation, a set of initially present points is set aside and considered as missing. The reconstruction of these points is then compared to their initial value, to establish the error of the reconstruction. Usually, the cross-validation points are chosen randomly from the whole data set, but in this work we used clusters of points with the shape of real clouds extracted from the initial cloudy data set. These points represent more realistically the missing data, so the error of their reconstruction reflects more accurately the actual error of the reconstruction. We randomly chose clouds from the data set and add them to the 50 cleanest images, to be sure that the data masked were initially present. About 4.4% of the initially present data were masked in this way, and this 4.4% of data were used in the cross-validation to find the number of optimal EOFs minimising the error of the reconstruction.

The lowest error, 0.42°C, was obtained by using the N=11 leading EOFs. We found that the optimal number of EOFs for the reconstruction is sensitive to the distribution of the chosen cross-validation points. The larger the region obscured by the clouds is, the fewer EOFs are used for the reconstruction. This indicated that only certain EOF modes with sufficiently large scale features can be reliably reconstructed, while high-order EOF (representing small scales) cannot be estimated given the typical cloud size in the Ligurian Sea. In other words, mesoscale features covered by clouds are under-sampled from the start and cannot be reliably reconstructed and the optimal number of EOFs included only the 11 most dominant EOFs, discarding most of the mesoscale signal.

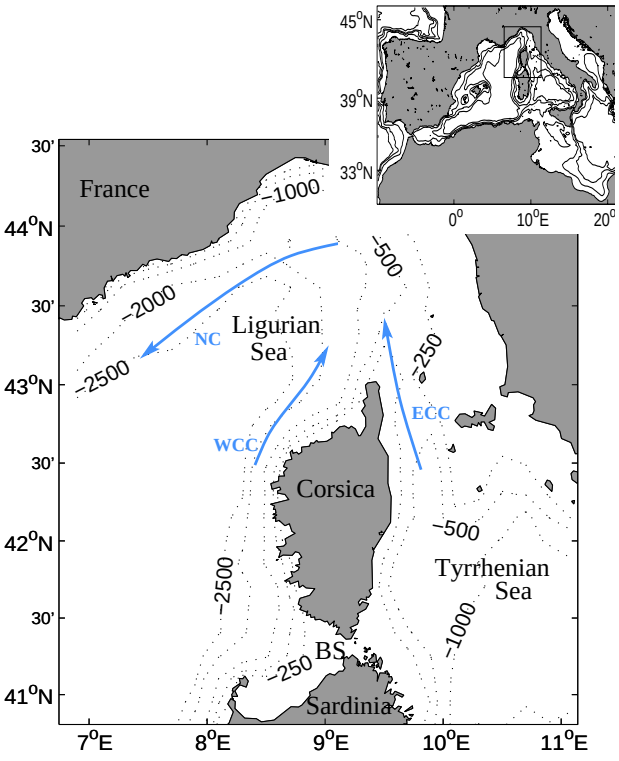

Fig. 1. Zone of interest around the Corsican Island in the Northwestern Mediterranean Sea. The Northern Current (NC), flowing southwestward is formed by the Western Corsican Current (WCC) and the Eastern Corsican Current (ECC). Strong frontal regions are associated with these currents, as the Liguro-Provençal Front, which follows the path of the NC. BS stands for Bonifacio Strait.

6.3 Error estimation

Equation (15) allows us to estimate the error variance from the variance filtered by the EOF reconstruction. First experiments revealed that the spatial error correlation of the SST observations could not be neglected and should be translated into a non-diagonal matrix $\mathbf{R}$. However, such a non-diagonal error-covariance matrix $\mathbf{R}$ would require the inversion of a $m_p \times m_p$ matrix. This matrix tends also to be more and more ill-conditioned if the correlation length is large. Computations with non-diagonal error covariance $\mathbf{R}$ are thus numerically prohibitive. In addition, it is not always clear how to specify off-diagonal terms. One straightforward way to circumvent this problem is to sub-sample the data such that the observations can be considered as independent. Another method is to retain the full observations data set, but to decrease the "weight" (i.e. increase the error variance) of the observations. It can be shown (e.g., Barth et al., 2006), that the error variance must be multiplied by the number r of

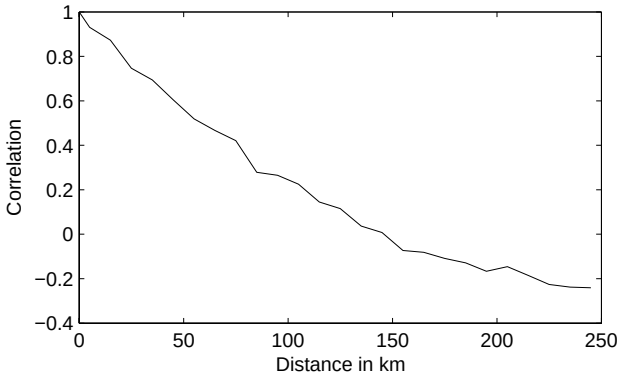

Fig. 2. Spatial SST correlation as a function of distance.

redundant (or strongly correlated) observations:

$$\mathbf{R} = r\mu^2 \mathbf{I}. \tag{39}$$

For a two-dimensional dataset, the factor r can be estimated by:

$$r \sim \frac{L^2}{\Delta x \Delta y} \tag{40}$$

where L is the correlation length of the observational error and Δx and Δy are the zonal and meridional resolution, respectively.

If we replace μ^2 by $r\mu^2$ in the asymptotic case for low noise we have

$$\bar{\epsilon^2} \sim \mu^2 \frac{NL^2}{m_p \Delta x \Delta y} \sim \mu^2 \frac{NL^2}{S} \tag{41}$$

where the surface S represents the observed area of the domain. S divided by L^2 is the number of really independent data used and it can be interpreted as the observed degrees of freedom or the number of EOF modes constrained by the observations at a particular time instant. The ratio $N/(S/L^2)$ is thus a measure on how well the N EOFs could be captured by the S/L^2 independent scalars present in the data set. Consequently, the more EOF modes are constrained, the smaller the average error will be.

It remains to determine the adequate value of r. Two methods were tested.

1 From the DINEOF cross-validation we already know that the error of the reconstruction of initially-missing points is 0.42°C. We used this information to calibrate the correlation length L (or equivalently the parameter r). Different length scales L were used until the error fields from the analysis gave on average a value of 0.42°C under the clouded regions. Here we see how the square root matrix of $\mathbf{C}$ could be of interest. Indeed, a change of μ^2 during the calibration process solely modifies the diagonal matrix, all other parts remaining unchanged. Hence the calculation of the error fields for another μ is extremely fast. From this procedure we obtained a correlation length for the observational error of 66 km and a parameter r=276.

2 In the second approach, a method similar to the cross-validation in DINEOF is used: using the same artificial clouds as for the cross-validation in DINEOF, the parameter r is calibrated until the difference between the optimally interpolated values under these artificial clouds is as close as possible to the observed ones. Note that for this approach we use a covariance matrix of DINEOF calculated also disregarding the same data points in order to be consistent with the DINEOF cross-validation. With this second approach, a value of 29 km is found for the correlation length of the observational error.

The question arises which of the two approaches is the more realistic one. Method 1 should provide the most coherent error estimates (because of the criteria for method 1) while method 2 should provide the best analysis (because of the minimisation of the error itself). A possible criteria to choose a method is a comparison with the correlation length of the SST anomalies. For in situ data, we would expect the correlation length of the observational error to be smaller than the correlation length of the SST anomalies. For satellite images, notably because of the atmospheric corrections, this might be questionable and scales might become comparable. In this case, DINEOF, without any information on these large scale errors, will interpret them as signal. Hence we expect that what we consider noise must have a correlation length smaller than the signal we reconstructed.

Independently from the cross-validation error, we estimated therefore the correlation function of the SST anomalies directly from the available data, where their spatial mean has been subtracted. This correlation function is shown in Fig. 2. The correlation length scale of the SST defined by the correlation threshold of e^{-1}∼0.37 is 80 km (the chosen threshold is based on a correlation function of the type e^{-d/L_a} where d is the distance. This simple approach can be justified since the correlation curve appears not to be affected by much noise and resembles an exponential. Even if the exact value of L_a depends on the different correlation functions that could be fitted, since the inflation approach itself is an approximation, we contented ourselves with an approximate value of L_a to which we can compare L).

The value we obtain is in agreement with both error length scales, and it seems that the SST length scale is larger (but of the same magnitude) than the SST error length scale. Both calibration methods for the correlation length of the observational error are thus not incoherent with the correlation length of the signal. Since in addition the analysed fields are very similar for both values and the error fields are not fundamentaly different (compare panels c and d of Fig. 3 with panels a and b of Fig. 4), no further optimisation seems necessary for the moment. Hence we present the results from the second

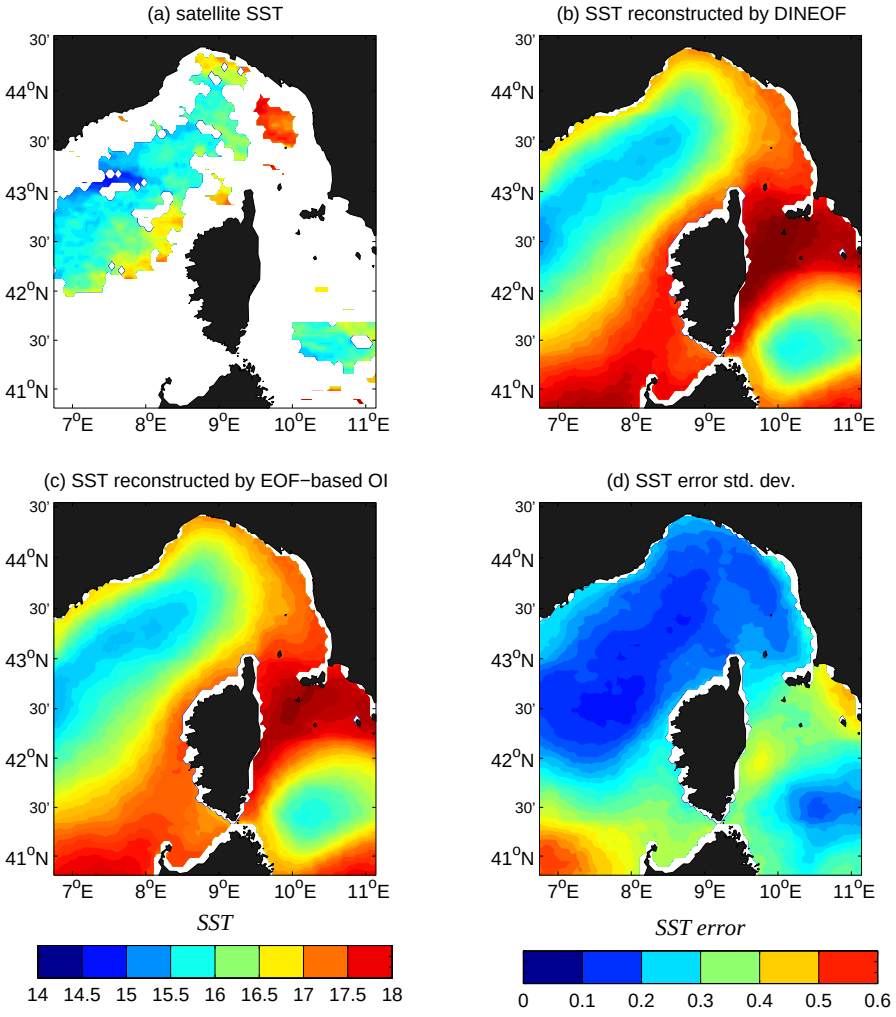

Fig. 3. Panel **(a)** is the observed SST on 15 November 1998. Panels **(b)** and **(c)** show the reconstruction by DINEOF and by optimal interpolation based on the same EOFs and a correlation length of 29 km. The estimated error standard deviation for the reconstruction is shown in panel **(d)**.

approach, based on the analysis error minimisation and leading to a clearer separation of the scales of noise (29 km) and signals (80 km). Note that the internal radius of deformation has a value of 4–7 km during winter in this region (e.g., Barth et al., 2005) leading to an associated wavelength of unstable motions of the order of 6 times the deformation radius (e.g., Cushman-Roisin , 1996), i.e. 25–44 km. The meanders of the Northern Current exhibit a typical length scale of 30 km to 60 km (e.g., Sammari et al., 1995).

Because the error estimates for method 2 are generally lower than for method 1, we will also be more severe using method 2 when deciding whether the difference between the OI and DINEOF reconstructions fall within the error bars or not. In order to confirm the validity of our approach consisting in taking the error fields from OI as error fields for DINEOF, the RMS difference between SST estimations from OI and DINEOF should indeed be smaller than this error estimation. The RMS difference between both fields is 0.17°C and indeed smaller than the average error:

$$\sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}^2} = 0.24°\text{C}. \tag{42}$$

We also computed the difference between DINEOF SST (x^r) and the OI SST ($x^{r,(\mathrm{OI})}$) scaled by the error estimation:

$$y_{ij} = \frac{\left(x_{ij}^r - x_{ij}^{r,(\mathrm{OI})}\right)^2}{\epsilon_{ij}^2}. \tag{43}$$

In 93% of the data points the scaled difference is lower than 1. This means that for the vast majority of the data points, the difference between both reconstructions is smaller than the

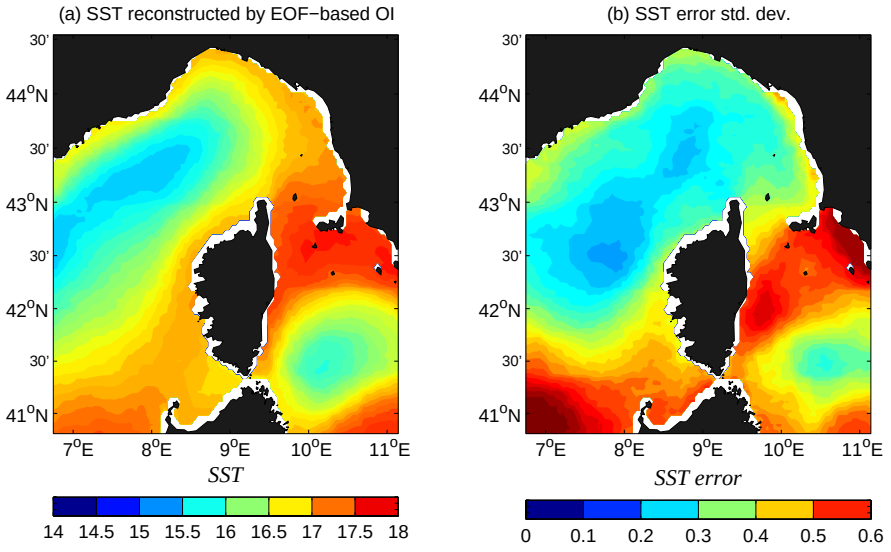

Fig. 4. Panel **(a)** SST on 15 November 1998 reconstructed by optimal interpolation using the same EOFs than DINEOF and a correlation length of 66 km. The estimated error standard deviation for this reconstruction is shown in panel **(b)**.

error estimation. However, this analysis takes only the error variance into account. The error estimation method also provides the error covariance. This enables us to establish the significance of the difference between reconstructions knowing the spatial correlation of the error. We assume that both reconstructions are a realisation of the Gaussian distributed random variable with possibly different means but the same covariances, i.e. the error covariance $\mathbf{E}$ given in Eq. (35):

$$\mathbf{x}_j^r \sim \mathcal{N}(\mathbf{m}_j^r, \mathbf{E}_j), \tag{44}$$

$$\mathbf{x}_j^{OI} \sim \mathcal{N}(\mathbf{m}_j^{OI}, \mathbf{E}_j), \tag{45}$$

where j is the temporal index for the image under consideration. The difference also follows a Gaussian distribution:

$$\mathbf{d}_j = \mathbf{x}_j^r - \mathbf{x}_j^{OI} \sim \mathcal{N}(\mathbf{m}_j^r - \mathbf{m}_j^{OI}, 2\mathbf{E}_j). \tag{46}$$

In order to transform this distribution into a normal one, we introduce the matrix $\tilde{\mathbf{S}}$:

$$\tilde{\mathbf{S}} = \sqrt{\frac{n}{2}}\mathbf{C}^{-1/2}\mathbf{\Sigma}^{N-1}\mathbf{U}^{N\mathsf{T}}, \tag{47}$$

which transforms the covariance matrix of the difference $\mathbf{d}_j$ into the identity matrix:

$$\tilde{\mathbf{S}}\mathbf{E}_j\tilde{\mathbf{S}}^{\mathsf{T}} = \frac{1}{2}\mathbf{I}_N. \tag{48}$$

The transformed variable follows therefore:

$$\mathbf{z}_j = \tilde{\mathbf{S}}\mathbf{d}_j \sim \mathcal{N}(\tilde{\mathbf{S}}(\mathbf{m}_j^r - \mathbf{m}_j^{OI}), \mathbf{I}_N). \tag{49}$$

We will examine if the difference between both reconstructions is significant to reject the null-hypothesis (H0):

$$\mathbf{m}_j^r = \mathbf{m}_j^{OI}. \tag{50}$$

In this case we would accept the alternative hypothesis H1:

$$\mathbf{m}_j^r \neq \mathbf{m}_j^{OI}. \tag{51}$$

Under the null-hypothesis, the transformed variable z follows a normal distribution.

$$\mathbf{z}_j \sim \mathcal{N}(0, \mathbf{I}_N). \tag{52}$$

Now we can test if our sample $\mathbf{z}_j$ has a mean significantly different from zero. We compute the average of z_j over all EOF modes and over time, $\bar{z}$. This mean is smaller than the critical $z_{\alpha/2}$ value used in a two-sided z-test for α=0.05.

$$|\bar{z}|\sqrt{Nn} = 0.93 < z_{\alpha/2} = 1.96\,. \tag{53}$$

This statistical test shows that the averaged difference between the OI reconstruction and the DINEOF reconstruction are not sufficiently large to be statistically significant.

The previous test measured the magnitude of the bias. We can also perform a test based on the L2-norm. Under the null-hypothesis the sum of the squared z_{ij} follows a χ^2-distribution with Nn=29 040 degrees of freedom. If this sum exceeds the critical value of 28 644, then the null-hypothesis must be rejected. But in our case, this value is again below this threshold:

$$\sum_{i,j} z_{ij}^2 = 3703 < 28\,644\,. \tag{54}$$

where the z-values are summed over time and over the EOF modes. Both tests show that the null-hypothesis cannot be rejected. This does not prove, however, that the hypothesis H0 is true. If there is any difference in the reconstructions $\mathbf{m}_j^r$

and $\mathbf{m}_j^{OI}$, then the difference is so small that it could not be detected by the current sample. But the fact that we are using a large sample of 2640 images (corresponding to 10 years of data) gives us confidence that if there is any difference between both reconstructions, it must be small. Therefore we conclude that the OI reconstruction and the DINEOF reconstruction are sufficiently close for the OI-derived estimation to be also a valid error estimation for the DINEOF reconstruction.

6.4 Validation

The validation of the DINEOF analysis itself against in situ data has been already presented in previous papers. The verification of error fields is much more delicate and in reality it amounts at validating the error-covariance matrix which needs a large amount of data. Nevertheless, we can attempt to validate the error estimates by using in situ observations and the data set aside during the cross-validation. We calculated the difference of the reconstruction $\mathbf{x}^r$ with the original data $\mathbf{x}$ at the cross-validation points. In principle, this distance should be scaled by the error-covariance matrix (Eq. (49)) as we did when we compared OI and DINEOF. In order to simplify the analysis, we can just calculate, neglecting the correlations of the errors,

$$\mathbf{z} = \mathrm{diag}(\epsilon)^{-1}(\mathbf{x}^r - \mathbf{x}) \tag{55}$$

where $\mathrm{diag}(\epsilon)$ is a diagonal matrix whose elements are the local standard deviations of the errors predicted by Eq. (34) at the data locations. For method 1, we get, using all available cross-validation points, an RMS value of 1.07, close to the expected value of 1. The difference might be due to the correlations we neglected. Hence we repeated the calculation not using all cross-validation points, but 200 randomly chosen points. We can expect these data to be independent and when calculating the RMS value of Eq. (55), we obtain 0.996, which is a nearly perfect match. For method 1, the error fields are therefore perfectly coherent with the actual distribution of differences between the reconstruction and the data set aside under the artificial clouds. Since method 1 was designed to provide the most coherent error fields, this is the best we can expect.

For method 2, the RMS value is now 1.88, suggesting that with the smaller value of the error-correlation length, we generally underestimate the errors. Looking into the details of the distribution, it appears that mostly the small errors are underestimated. This can be explained by observing that the error correlation length basically controls the base error in regions with large data coverage and low errors as in Eq. (41). The variance scales as L^2, so that standard deviation scales as L at low noise. This is confirmed by comparing the error fields of Figs. 3 and 4 in the cloud-free regions. Hence for method 2, which has an error-correlation length roughly half of that of method 1, we underestimate the small errors by a factor 2, hence the RMS value of 1.88 for the normalized error.

In view of the previous analysis, method 1 should be preferred for coherent error maps. The question that now comes into mind is how the error estimates compare to in situ data. For this exercise, 845 near-surface (1 m depth) temperature observations have been extracted from the MFSTEP database (http://www.ifremer.fr/mfstep/), 122 data points corresponding to unclouded points and the rest to clouded regions. For the validation of the error fields, we have to be conscious that we are now comparing different temperature measurements. Up to now we have only dealt with satellite data, and when speaking about errors, we only took into account the noise on the actual information at the satellite sensor and subsequent interpolation errors under the clouds. In this sense, the error maps are error maps for pure satellite information. When comparing with in situ data, we have to keep in mind that the latter represent a different thing and compared to the satellite we will add representativity errors and possibly bias. Also, correlated errors, due to atmospheric correction that DINEOF has interpreted as signal, will now show up as actual errors. Hence the error maps provided up to now are not a direct measure of the errors compared to the in situ data.

To make them comparable, we can first note that the difference between analysed and original satellite data at cloud-free points where we have in situ data gives us an estimate of the noise in the satellite data (0.22°C) that we cannot interpret as signal with the interpolation method. A comparison of original satellite SST and in situ data then provides a measure of error now including representativity and bias compared to the previous one. This yields an RMS value of 0.687°C.

The difference between the two previous values provides therefore the minimal error $\epsilon_{\min}$ we are going to make for a cloudless image, now compared to the in situ values. The value of 0.45°C we obtain is clearly typical of the expected precision or incompressible base error, similar to those obtained in Marullo et al (2006) for the same region.

Now, when comparing the error map ϵ calculated with DINEOF with the actual differences between reconstructed satellite data and in situ data $\mathbf{x}^{\mathrm{is}}$, we have to scale by the sum of the interpolation error and incompressible base error, instead of the sole interpolation error:

$$\mathbf{z} = \mathrm{diag}(\epsilon + \epsilon_{\min})^{-1}(\mathbf{x}^r - \mathbf{x}^{\mathrm{is}}) \tag{56}$$

Here, in situ data can be considered independent so that the components of $\mathbf{z}$ should have a gaussian distribution with unit standard deviation. The RMS value calculated is 0.96 and the distribution is shown in Fig. 5, confirming our conclusions that the calculated error fields are meaningful.

6.5 Results

As an example of the reconstruction, Fig. 3 shows a SST snapshot on 15 November 1998 (panel a). The central part of the Ligurian Sea and a fraction of the Tyrrhenian Sea are

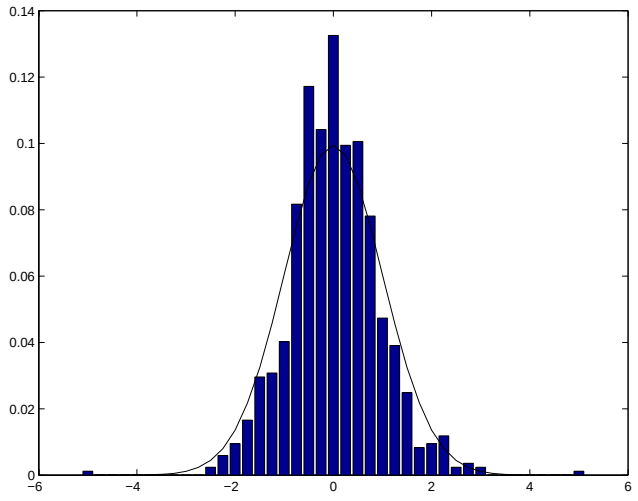

Fig. 5. The histogram shows the distribution of the difference between reconstructed SST and near surface in situ temperature, scaled by the error estimation including representativity errors. The solid line represents the theoretical normal distribution.

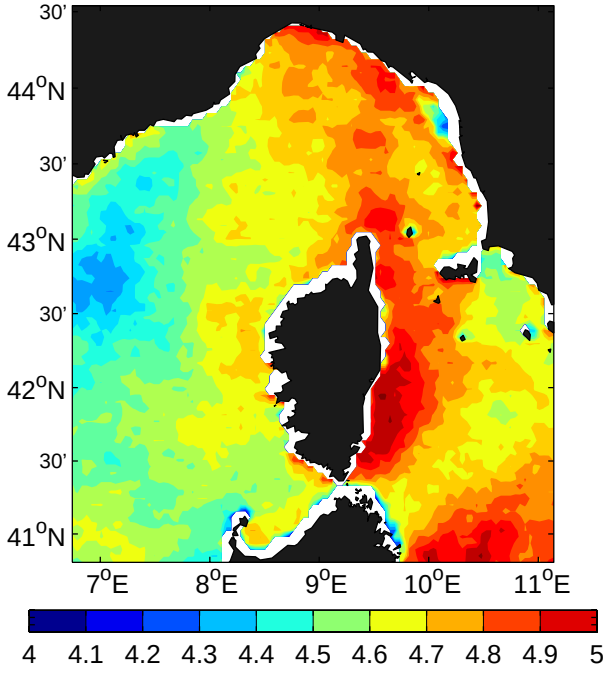

Fig. 6. Standard deviation of SST over the studied time period including the seasonal cycle.

present in the observed SST. As one would expect, the estimated error standard deviation is the lowest in those regions. The error increases gradually and is highest far away from the existing observations. Although the background error covariance is defined by global EOF modes and does therefore not include an explicit correlation length scale, the presented error estimation method was able to quantify the local effect of clouds on the error variance.

East of Corsica (approximatively at 42° N and 9°30′ E) the error estimation is relatively high despite the presence of observations nearby. The SST standard deviation over the studied time period (Fig. 6) is particularly high in this region. During summer this zone is warmer than e.g. the west coast of Corsica. The shallower depth of the east coast of Corsica shields this zone from the large-scale ocean current. This example shows that the error estimation takes also the variability of the field into account.

Although the Northern Current is covered by clouds in this snapshot, its SST signature has been reconstructed by DINEOF (panel b) and the OI method (panel c) using the error covariance the EOFs computed by DINEOF. It is unlikely that an OI method using an isotropic and homogeneous error covariance would be capable of reconstructing the Northern Current in a situation where very few data are available. To test this possibility, we have made a comparison between the DINEOF reconstruction and an isotropic OI reconstruction on 30 December 2003. The cloudiness at and around this date is especially high, with some days with no data at all, which makes it appropriate for our purposes. Using a time window of 5 days, observations from the 26 December 2003, 27 December 2003, 30 December 2003 and 4 January 2004 are available for the OI reconstruction, shown in Fig. 7. These four days present a mean cloud coverage of 76.7%. For the OI reconstruction, a spatial correlation length of 80 km (consistent with the correlation length for the Ligurian Sea found in Sect. 6.3) and a temporal correlation length of 3 days are used.

The OI reconstruction (Fig. 8) is degraded by the data on 26 December 2003, mostly in the western part of the Ligurian Sea. This image is the only one that presents a good data coverage, but the time difference between this image and the analysed image is 5 days, and the SST on 30 December is notably colder than on 26 December. This trend is clearly visible in the unclouded pixels, in particular east of the Strait of Bonifacio.

The DINEOF reconstruction on 30 December 2003 (Fig. 9) presents smoother values and a more realistic SST distribution on the western and northern Ligurian Sea. Both analyses are similar east of the Corsican Island, in the Tyrrhenian Sea, where most data are available. This example shows the ability of the global DINEOF analysis to produce better results than a standard isotropic OI reconstruction when only a few SST observations are present. The EOF-based OI reconstruction on this date (image not shown) is similar to the reconstruction of Fig. 9. This shows that OI for satellite data should use point-to-point (i.e., non-homogeneous, non-isotropic) correlation functions. DINEOF provides such point-to-point correlation, but with the added possibility of efficient inversion. We note that the correlation function of DINEOF do however eliminate some structures such as the

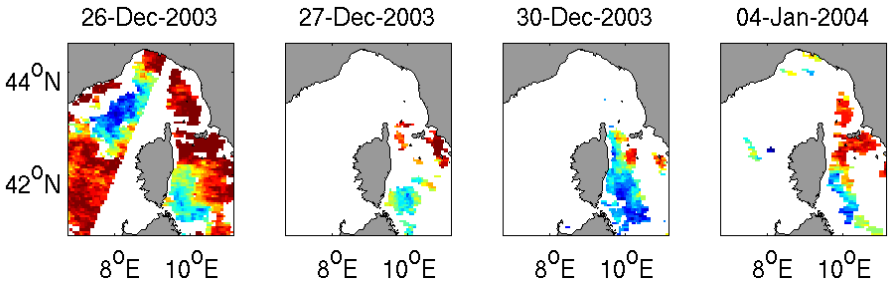

Fig. 7. Data used in the OI reconstruction of 30 December 2003. The colorbar is the same as in Fig. 8.

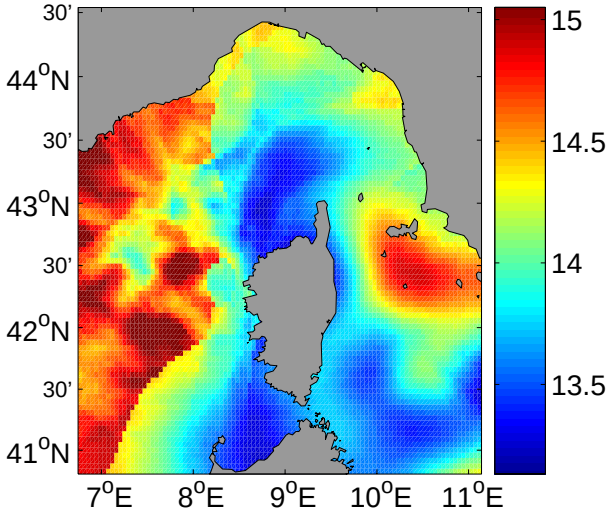

Fig. 8. Reconstruction of the SST on 30 December 2003 by Optimal Interpolation.

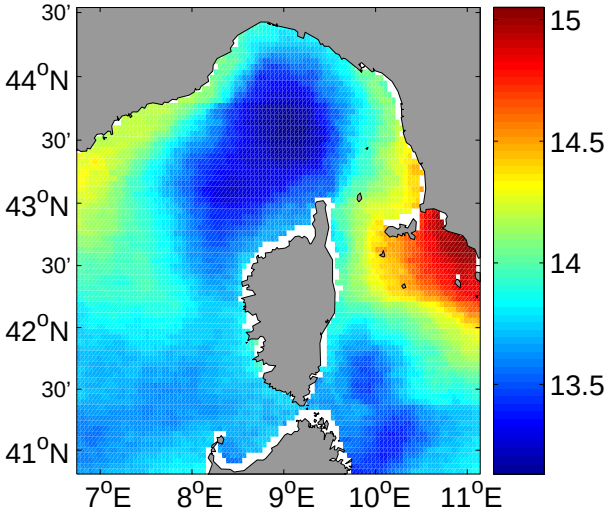

Fig. 9. Reconstruction of the SST on 30 December 2003 by DINEOF.

one near 43° N and 7.5° E, but OI also would filter out some of the signal.

6.6 Inter-annual variability

As an example of DINEOF application we assess if the accuracy of the reconstructed SST is sufficient to study inter-annual variability of the spatial averaged sea-surface temperature.

The average seasonal cycle has been computed from the reconstructed SST using all data from 1995 to 2005 filtered with 15-day cut-off low pass filter (Fig. 10). The seasonal cycle shows an asymmetric behaviour: while the mean temperature remains almost constant at the minimum temperature during January to March, the maximum temperature is only reached during a short period of time during August. The deviations from this seasonal cycle are shown in Fig. 11. The heatwave of 2003 affecting south Europe, in particular France, can be clearly seen from this time series. The error of the spatial mean SST (Fig. 12) has been computed from Eq. (38) using the error estimates of method 1. Since we can assume that the error of the seasonal cycle is negligible, the error estimate represents also the expected error of the temperature anomaly of Fig. 11. The expected error of the mean SST based on DINEOF is thus more than two orders of magnitude smaller than the inter-annual signal in SST of our studied domain. The reconstructed SST is therefore suitable to study inter-annual SST variability.

The expected error is highly variable in time, but the low-passed error estimate (cut-off frequency of 15 days) reveals a seasonal cycle in the error estimation. The reconstruction has the highest error in winter and is about 25% more accurate during summer. The seasonality of the error estimation is due to the cloud coverage. The unfiltered error estimation correlates to 0.85 with the fraction of missing data. The correlation between the filtered error estimation and the filtered

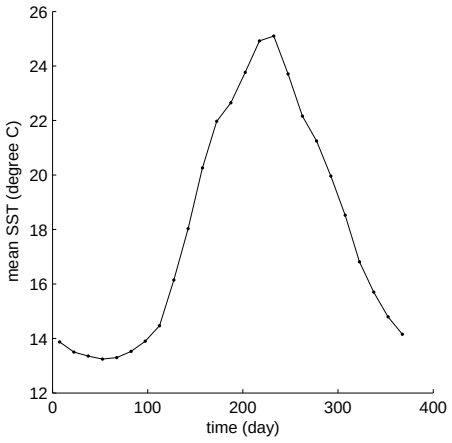

Fig. 10. Seasonal cycle of the spatially averaged SST using reconstructed SST from 1995 and 2005 and filtered with 15-days cut-off low pass filter.

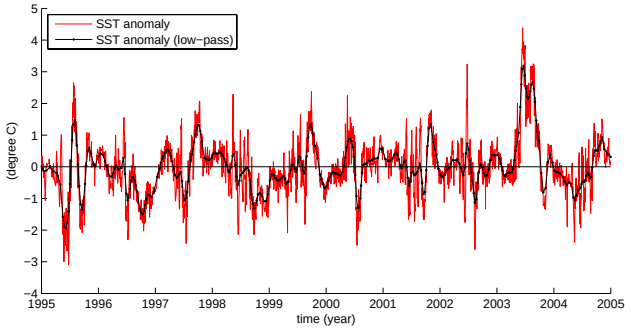

Fig. 11. Mean SST anomalies and filtered mean SST anomalies (15-days cut-off frequencies).

fraction of missing data is 0.92. It is rather the fact that for each image, there are more clouds in winter than in summer, that explains therefore the higher errors in winter.

If we had taken the simple approach to calculate the mean temperature using only available data, we would have obtained another time-series. The difference between the anomaly of the latter (compared to the same seasonal cycle) and our estimate of Fig. 11 has an RMS value of 0.22°C, much higher than the error estimate found in Fig. 12. Note that if we had also taken only the available data to calculate the confidence interval for the mean obtained by the simple approach, we would have found an expected error on the mean of 0.012°C. This is much lower than the actual error of the simple estimate and yet higher than the error we get on the DINEOF analysis of the mean. Clearly, the DINEOF approach provides better estimates of the mean and narrower associated error bars. Graphically, the time series is not very different, specially if the filtered time series are compared.

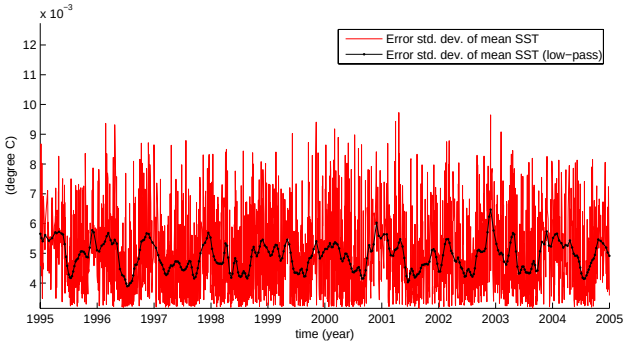

Fig. 12. Error estimate of mean SST and filtered error estimate (15-days cut-off frequencies).

filter is used. This is why we chose not to show them. The error estimates are however very different and in reality the brute force approach leads to a field that is outside the error bar of the DINEOF analysis. Stated differently: the brute force approach with its error bar for the mean around the signal has no intersection with the DINEOF version and its error bars (independently of the way we estimate this error: from the standard deviation of the data or from the typical error of sensors, 0.3°C , divided by the square root of the number of data). Even applying an inflation factor to this, which amounts to use only the number of independent data in the estimation of the error on the mean, the same conclusion holds. Hence we must admit that both time series are significantly different.

7 Conclusions

We presented a method that allows to complement the cloud filling method DINEOF with local error estimates. The approach uses the error estimates from optimal interpolation (OI), itself exploiting the covariance fields provided by DINEOF. Because of the factorisation of the covariance matrix also provided by DINEOF, OI can be performed as a least-square fit of EOF amplitudes, which drastically reduces computational requirements. The same approach can be exploited during the error calculations.

In the present paper we applied the method to the reconstruction of SST fields in the region around Corsican Island, including the calculation of the inter-annual variability of the spatial means. It was shown that this approach allowed the isolation of inter-annual variability with very small error bars.

In the present case, the difference between the analysis provided by OI and DINEOF was shown to be smaller than the error fields, justifying the use of the error field for both analyses.

Should the difference be too large in some applications, the present method still allows to provide error estimates, but only for the OI. The latter, however, still benefits from the covariance factorization of DINEOF.

Another possibility would be to adapt DINEOF so as to include OI in the iterations, using the covariance from the EOFs under calculation, as the method of estimating missing values. Such a hybrid approach would lead to a coherent set of EOFs, covariance matrix and error fields. This approach was not yet implemented because in the cases we tested, the difference between OI and DINEOF were too small to justify the additional complexification. Probably a more important point to analyse for further improvement is the inherent hypothesis of the method that cloud coverage is uncorrelated with the interpolated field. This can probably be justified for SST when clouds are not persistent but it is already more questionable for Chlorophyll which reacts rapidly to changes in insolation or storms associated with clouds. In this case, additional information from scatterometers and in situ could probably help improve the detection of patterns of variability in a multivariate approach.

We finish by mentioning that the definition of the region covered by the data imposes an a priori constraint on the scales that are analysed. Global EOFs on the chosen region do not honour the multi-scale nature of the ocean. While a global truncated EOF series is unable to represent small-scale variation, local EOFs ignore large-scale correlation (induced by processes such as ENSO, NAO,...). If the data coverage is uniformly dense, as it is the case for satellite data, long-range correlation can be neglected since the large-scale processes are well present in the data and "oversampled". For in situ data where coverage is highly non-uniform, one needs to include long and small-range correlations but we think that this is desirable even for satellite data.

One possibility to tackle a larger-scale problem would be to reconstruct global SST, e.g., at 1 degree resolution, then reconstruct the SST anomaly at 1/4 degree (observed SST minus reconstructed global SST) for each ocean basin independently, and then reconstruct the SST anomaly at 1/20 (observed SST minus basin-wide SST) for each sub-basin independently, and so on.... At each level one would introduce more and more small scale features.

Appendix A

Useful matrix identities

–

$$\left(\mathbf{A}+\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I}+\mathbf{V}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{A}^{-1} \quad \text{(A1)}$$

–

$$\mathbf{L}^{\mathsf{T}}\left(\mathbf{L}\mathbf{L}^{\mathsf{T}}+\mu^2\mathbf{I}\right)^{-1} = \left(\mathbf{L}^{\mathsf{T}}\mathbf{L}+\mu^2\mathbf{I}\right)^{-1}\mathbf{L}^{\mathsf{T}} \quad \text{(A2)}$$

provided the inverse matrix exists and $\mathbf{I}$ is an identity matrix (with 1 on the diagonal) of appropriate dimension.

Appendix B

Square root calculations

If for some reason, the square root of the covariance matrix is needed, we can use the eigenvector (or SVD) decomposition,

$$\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p = \mathbf{W}_p^{\mathsf{T}}\mathbf{\Lambda}_p\mathbf{W}_p, \quad \text{(B1)}$$

with $\mathbf{W}_p^{\mathsf{T}}\mathbf{W}_p$=$\mathbf{I}_N$ and $\mathbf{\Lambda}_p$ a $N\times N$ diagonal matrix, which leads to the following expression of $\mathbf{C}$:

$$\mathbf{C} = \mu^2\mathbf{W}_p^{\mathsf{T}}\left(\mathbf{\Lambda}_p+\mu^2\mathbf{I}_N\right)^{-1}\mathbf{W}_p. \quad \text{(B2)}$$

The square root matrix $\mathbf{C}^{1/2}$ defined as

$$\mathbf{C} = \mathbf{C}^{1/2}\left(\mathbf{C}^{1/2}\right)^{\mathsf{T}} \quad \text{(B3)}$$

is therefore:

$$\mathbf{C}^{1/2} = \mu\mathbf{W}_p^{T}\left(\mathbf{\Lambda}_p+\mu^2\mathbf{I}_N\right)^{-1/2}. \quad \text{(B4)}$$

Note that the matrix expression in brackets is a diagonal matrix and its square root involves only the square root of its diagonal elements. Because $\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p$ is of size $N\times N$, the SVD decomposition and subsequent calculation of the square root of $\mathbf{C}$ is essentially an inexpensive operation compared to the analysis.

Appendix C

Error behavior

To have an idea of the amplitude of the analysis error, we can scale the involved matrices on the following ground: the inner matrix to invert in Eq. (35) involves the grid points with data and is in fact a covariance matrix between EOF modes (on average over the points with data). Since on statistical average the EOFs are independent if all m points are available, the matrix behaves as a diagonal matrix of size N depending on the singular values ρ_i. If only m_p points are present, instead of having a vector product of full EOFs (that would have a unit norm by construction), the product Eq. (10) over m_p points scales as m_p/m . Therefore using Eq. (17) we have

$$\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p \sim \frac{m_p}{m}\frac{1}{n}\mathbf{\Sigma}^N\mathbf{\Sigma}^N. \quad \text{(C1)}$$

Two extreme situations are worth analysing

– If the noise is relatively small (compared to the variance of the data) we have

$$\mathbf{C} \sim \mu^2 \left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right)^{-1}, \tag{C2}$$

so that the error covariance matrix after the objective analysis with low noise behaves as

$$\mathbf{E} \sim \mu^2 \mathbf{L} \left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right)^{-1} \mathbf{L}^{\mathsf{T}}. \tag{C3}$$

Formally we use a pseudo-inverse should the inversion become singular. Using the definition (17), this leads to

$$\begin{aligned}\mathbf{E} &= \mu^2 \frac{1}{n} \mathbf{U}^N \mathbf{\Sigma}^N \left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right)^{-1} \mathbf{\Sigma}^N \mathbf{U}^{N\mathsf{T}} \\ &\sim \mu^2 \frac{m}{m_p} \mathbf{U}^N \mathbf{U}^{N\mathsf{T}}.\end{aligned} \tag{C4}$$

The average error over the grid is the trace tr($\mathbf{E}$) of the covariance matrix divided by the number of grid points m. Using the orthormality of the EOFs, this leads to

$$\bar{\epsilon^2} \sim \mu^2 \frac{N}{m_p}. \tag{C5}$$

In other words, the average expected error is the noise reduced by the factor depending on the EOF expansion and data points used. This is probably an overoptimistic finding, because in reality errors on the data are not independent and instead of m_p, there should appear the number of data with uncorrelated errors (see Sect. 6.3). From this analysis, we found that in the case of low observational errors, the expected error of the reconstruction is inversely proportional to the number of EOF chosen. This number characterizes the degrees of freedom in the system. Therefore, the fewer degrees of freedom a system has, the easier it is to reconstruct missing points from data in unclouded points.

– At the other extreme, for very large noise

$$\begin{aligned}\mathbf{E} &= \mathbf{L}\left(\mathbf{I} + \frac{1}{\mu^2}\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right)^{-1} \mathbf{L}^{\mathsf{T}} \\ &\sim \mathbf{L}\left(\mathbf{I} - \frac{1}{\mu^2}\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right) \mathbf{L}^{\mathsf{T}}\end{aligned} \tag{C6}$$

which using the same reasoning yields

$$\begin{aligned}\mathbf{E} = \frac{1}{n}\mathbf{U}^N \mathbf{\Sigma}^N \; \mathbf{\Sigma}^N \mathbf{U}^{N\mathsf{T}} \\ - \frac{m_p}{mn^2}\frac{1}{\mu^2}\mathbf{U}^N \left(\mathbf{\Sigma}^N\right)^4 \mathbf{U}^{N\mathsf{T}}.\end{aligned} \tag{C7}$$

Taking the trace divided by m we recover an average error. The first term contains the first N squared singular values, that we can immediately relate to σ^2. The second term contains singular values to the fourth power. If we assume that the first N values are similar (and thus related to σ) we get

$$\bar{\epsilon^2} \sim \sigma^2 \left(1 - \frac{\sigma^2}{\mu^2}\frac{m_p}{N}\right). \tag{C8}$$

Here, because of the large noise, the relative error is of the order of 1, as should be expected.

In both asymptotic cases the factor $\mu^2 N/(m_p\sigma^2)$ appears, which can be interpreted as the ratio of observational errors ($\mu^2\mathbf{I}_N$) versus the background error captured by the EOFs ($\mathbf{L}_p{}^{\mathsf{T}}\mathbf{L}_p$) and hence the relative weights in the analysis step:

$$\frac{\mathrm{tr}\left(\mu^2\mathbf{I}_N\right)}{\mathrm{tr}\left(\mathbf{L}_p^{\mathsf{T}}\mathbf{L}_p\right)} \sim \frac{\mu^2 N}{\sigma^2 m_p} \tag{C9}$$

In this last equation we used Eq. (C1) and the fact that the sum of the leading N eigenvalues is $nm\,\sigma^2$.

Acknowledgements. European projects MFSTEP (EVK3-CT-2002-00075), EUR-OCEANS (European Network of excellence FP6 – Global change and ecosystems Contract number 511106) and Concerted action RACE (Communauté Française de Belgique) allowed to perform this work. The National Fund for Scientific Research, Belgium is acknowledged for the financing of a supercomputer. D. Gomis made helpful suggestions on error estimates in an early version of the paper and then provided, together with M. Rixen, S. Vignudelli, P. Cipollini and two anonymous reviewers, valuable comments on the discussion paper. The AVHRR Oceans Pathfinder SST data were obtained from the Physical Oceanography Distributed Active Archive Center (PO.DAAC) at the NASA Jet Propulsion Laboratory, Pasadena, CA (http://podaac.jpl.nasa.gov). This is MARE publication MARE097.

Edited by: P. Cipollini

References

Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers, J.-M.: Reconstruction of incomplete oceanographic data sets using Empirical Orthogonal Functions. Application to the Adriatic Sea, Ocean Modelling, 9, 325–346, 2005.

Alvera-Azcárate, A., Barth, A., Beckers, J. M., and Weisberg, R. H.: Multivariate Reconstruction of Missing Data in Sea Surface Temperature, Chlorophyll and Wind Satellite Fields, J. Geophys. Res., accepted, 2006.

Astraldi, M., Gasparini, G. P., and Sparnocchia, S.: The seasonal and interannual variability in the Ligurian-Provencal basin, in : The Seasonal and Interannual Variability of the Western Mediterranean Sea, edited by: La Violette, P. E., AGU Coastal and Estuarine Studies, 46, 93–113, 1994.

Astraldi, M., Balopoulos, S., Candela, J., Font, J., Gacić, M., Gasparini, G. P., Manca, B., Theocharis, A., and Tintoré, J.: The role of straits and channels in understanding the characteristics of Mediterranean circulation, Prog. Oceanogr., 44, 65–108, 1999.

Barth, A., Alvera-Azcárate, A., Rixen, M., and Beckers, J.-M.: Two-way nested model of mesoscale circulation features in the Ligurian Sea, Prog. Oceanogr., 66, 171–189, 2005.

Barth, A., Alvera-Azcárate, A., Beckers, J.-M., Rixen, M., and Vandenbulcke, L.: Multigrid state vector for data assimilation in a two-way nested model of the Ligurian Sea, J. Mar. Syst., in press, 2006.

Beckers, J.-M. and Rixen, M.: EOF calculations from incomplete oceanographic data sets, J. Atmos. Ocean Technol., 20, 1839–1856, 2003.

Bergant, K., Susnik, I., Strojan, M., and Shaw, A.: Sea Level Variability at Adriatic Coast and its Relationship to Atmospheric Forcing, Ann. Geophys., 23, 1997–2010, 2005, http://www.ann-geophys.net/23/1997/2005/.

Boyd, J., Kennelly, E., and Pistek, P.: Estimation of EOF expansion coefficients from incomplete data, Deep Sea Res., 41, 1479–1488, 1994.

Cushman-Roisin, B.: Introduction to Geophysical fluid dynamics, Prentice Hall, 1996.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, 1991.

Emery, W. and Thomson, R.: Data analysis methods in physical oceanography, Pergamon, 1997.

Eslinger, D., O'Brian, J., and Iverson, R.: Empirical Orthogonal Function Analysis of Cloud-Containing Coastal Zone Color Scanner Images of Northeastern North American Coastal Waters, J. Geophys. Res., 94, 10 884–10 890, 1989.

Evensen, G.: Sampling strategies and square root analysis schemes for the EnKF, Ocean Dynamics, 54, 539–560, 2004.

Gomis, D. and Pedder, M.: Errors in dynamical fields inferred from oceanographic cruise data: Part I. The impact of observation errors and the sampling distribution, J. Mar. Syst., 56, 317–333, 2005.

Gomis, D., Ruiz, S., and Pedder, M.: Diagnostic analysis of the 3D ageostrophic circulation from a multivariate spatial interpolation of CTD and ADCP data, Deep Sea Res., 48, 269–295, 2001.

Gunes, H., Sirisup, S., and Karniadakis, G.: Gappy data: To Krig or not to Krig?, J. Comput. Phys., 212, 358–382, 2006.

Kaplan, A., Kushnir, Y., Cane, M., and Blumenthal, B.: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperature, J. Geophys. Res., 102, 27 853–27 860, 1997.

Kondrashov, D., Feliks, Y., and Ghil, M.: Oscillatory modes of extended Nile River records (A.D. 622–1922), Geophys. Res. Lett., 3, L10702, doi:10.1029/2004GL022156, 2005.

Larnicol, G., Le Traon, P. Y., Ayoub, N., and De Mey, P.: Mean sea level and surface circulation variability of the Mediterranean Sea from 2 years of TOPEX/POSEIDON altimetry, J. Geophys. Res., 100, C12, 25 163–25 177, 1995.

Lawson, C. and Hanson, R.: Solving Least square problems, Prentice Hall, 1974.

Marullo, S., Nardelli, B., Guarracino, M.n and Santoleri, R.: Observing The Mediterranean Sea from space: 21 years of Pathfinder-AVHRR Sea Surface Temperatures (1985 to 2005). Re-analysis and validation, Ocean Sci. Discuss., 3, 1191–1223, 2006, http://www.ocean-sci-discuss.net/3/1191/2006/.

Millot, C.: Circulation in the Western Mediterranean Sea, J. Mar. Syst., 20, 423–442, 1999.

Millot, C. and Taupier-Letage, I.: Circulation in the Mediterranean Sea, in: The Mediterranean Sea, edited by: Saliot, A., The Handbook of Environmental Chemistry, 5, 29–66, Springer, 2005.

North, G., Bell, T., Cahalan, F., and Moeng, F.: Sampling Errors in the estimation of empirical orthogonal functions, Mon. Wea. Rev., 110, 699–706, 1982.

Orfila, A., Álvarez, A., Tintoré, J., Jordi, A., and Basterretxea, G.: Climate teleconnections at monthly time scales in the Ligurian Sea inferred from satellite data, Prog. Oceanogr., 66, 157–170, 2005.

Pham, D., Verron, J., and Roubaud, M.: A singular evolutive extended Kalman filter for data assimilation in oceanography, J. Mar. Syst., 16, 323–340, 1998.

Rixen, M., Beckers, J.-M., Brankart, J.-M., and Brasseur, P.: A numerically efficient data analysis method with error map generation, Ocean Modell., 2, 45–60, 2000.

Rixen, M., Beckers, J.-M., Levitus, S., Antonov, J., Boyer, T., Maillard, C., Fichaut, M., Balopoulos, E., Iona, S., Dooley, H., Garcia, M.-J., Manca, B., Giorgetti, A., Manzella, G., Mikhailov, N., Pinardi, N., Zavatarelli, M., and the Medar Consortium: The Western Mediterranean Deep Water: a proxy for global climate change, Geophys. Res. Lett., 32, L12608, doi:10.1029/2005GL022702, 2005.

Sammari, C., Millot, C., and Prieur, L.: Aspects of the seasonal and mesoscale variabilities of the Northern Current in the Western Mediterranean Sea inferred from the PROLIG-2 and PROS-6 experiments, Deep Sea Res., 42, 893–917, 1995.

Shen, S., Smith, T., Ropelewski, C., and Livezey, R.: An optimal regional averaging method with error estimates and a test using tropical pacific SST data, J. Climate, 11, 2340–2350, 1998.

Toumazou, V. and Cretaux, J.-F.: Using a Lanczos Eigensolver in the Computation of Empirical Orthogonal Functions, Mon. Wea. Rev., 129, 1243–1250, 2001.

Vautard, R., Yiou, P., and Ghil, M.: Singular spectrum analysis: a toolkit for short, noisy chaotic signals, Physica D, 58, 95–126, 1992.

Verlaan, J. and Heemink, A.: Tidal Flow Forecasting using Reduced Rank Square Root Filters, Stochastic Hydrology and Hydraulics, 11, 349–368, 1997.

Vignudelli, S., Cipollini, P., Astraldi, M., Gasparini, G. P., and Manzella, G.: Integrated use of altimeter and in situ data for understanding the water exchanges between the Tyrrhenian and Ligurian Seas, J. Geophys. Res., 105, 19 649–19 663, 2000.

Vignudelli, S., Cipollini, P., Reseghetti, F., Fusco, G., Gasparini, G., and Manzella, G.: Comparison between XBT data and TOPEX/Poseidon satellite altimetry in the Ligurian-Tyrrhenian area, Ann. Geophys., 21, 123–135, 2003, http://www.ann-geophys.net/21/123/2003/.

von Storch, H. and Zwiers, F.: Statistical analysis in climate research, Cambdridge University Press, 1999.