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

J.-M. Beckers, A. Barth, A. Alvera-Azcárate

Ocean Sci. Discuss., 3, 735–776, 2006
www.ocean-sci-discuss.net/3/735/2006/

Ocean Science Discussions

Papers published in *Ocean Science Discussions* are under open-access review for the journal *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 – Accepted: 19 June 2006 – Published: 10 July 2006

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

Abstract

We present an extension to the Data INterpolating Empirical Orthogonal Functions (DINEOF) 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 later 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 be perfectly known (e.g., Rixen et al., 2000; Gomis et al., 2001). In practise ad hoc parametric correlation functions are 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, 2006[1]),

[1] Alvera-Azcárate, A., Barth, A., Beckers, J. M., and Weisberg, R. H.: Multivariate Recon-

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 krigging methods in the framework of computational fluid dynamics and was found to be more accurate than the latter for high temperal resolution and not too large 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.

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 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

struction of Missing Data in Sea Surface Temperature, Chlorophyll and Wind Satellite Fields, J. Geophys. Res., submitted, 2006.

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

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

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 $\mathbf{d}$ that minimises the expected error variance in location r is given by

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

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

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

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

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

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

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

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

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

with **c** being the covariance between data points and the point of interpolation and **B** the field-covariance matrix also called background error 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 homogenous 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 **I** is the identity matrix. In the following, the signal variance is

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

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.

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 **X**. Each column of **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

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

where **U** contains on each of its columns one of the spatial patterns of the EOFs, the pseudo-diagonal matrix **Σ** the singular values and **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 thoroughly 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.

DINEOF provides as result a Singular Value Decomposition of the data matrix $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ 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

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

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

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

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

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. Note that if the initial matrix was complete and contained homogenous noise, we would have $\sum_{i=1}^{n}\rho_i^2 = mn\,(\sigma^2+\mu^2)$ and

$$\sum_{i=1}^{N}\rho_i^2 = mn\,\sigma^2. \tag{13}$$

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, but Eq. (13) remains valid. On the other hand, 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

$$\mu^2 = \frac{1}{mn - M} \sum_{x_{ij} \text{not missing}} \left(x_{ij}^2 - {x_{ij}^r}^2 \right) \tag{14}$$

using only the original data values x_{ij} and the reconstruction x_{ij}^r 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[2]:

$$\mathbf{D} = \frac{1}{n} \mathbf{X} \mathbf{X}^{\mathrm{T}}. \tag{15}$$

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

$$\mathbf{L} = \frac{1}{\sqrt{n}} \mathbf{U}^N \mathbf{\Sigma}^N, \tag{16}$$

[2]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.

which 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 $\mathbf{V}^{N\mathrm{T}}\mathbf{V}^{N}=\mathbf{I}$, to the field covariance

$$\mathbf{B} = \frac{1}{n}\mathbf{X}^{N}\mathbf{X}^{N\mathrm{T}} = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{17}$$

since we assumed that the first N EOFs contain signals and the remaining EOFs some noise. Note that this rejection of higher EOFs is coherent with the fact that to accurately estimate higher EOFs, very large sample sizes are needed (e.g., North et al., 1982).

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 uncorrellated. 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

$$\mathbf{R} = \mu^2\,\mathbf{I}, \tag{18}$$

where μ^2 is given by Eq. (14).

Having now $\mathbf{R}$, the covariance matrix of the noise unexplained by the first N EOFs and the field coviance 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[3] 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

$$\mathbf{B} = \begin{pmatrix}\mathbf{L}_p\\ \mathbf{L}_m\end{pmatrix}\begin{pmatrix}\mathbf{L}_p^{\mathrm{T}} & \mathbf{L}_m^{\mathrm{T}}\end{pmatrix} = \begin{pmatrix}\mathbf{L}_p\mathbf{L}_p^{\mathrm{T}} & \mathbf{L}_p\mathbf{L}_m^{\mathrm{T}}\\ \mathbf{L}_m\mathbf{L}_p^{\mathrm{T}} & \mathbf{L}_m\mathbf{L}_m^{\mathrm{T}}\end{pmatrix}, \tag{19}$$

[3]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.

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

$$\mathbf{B}_p = \mathbf{L}_p \mathbf{L}_p^{\mathrm{T}}. \tag{20}$$

The row i of

$$\begin{pmatrix} \mathbf{L}_p \mathbf{L}_p^{\mathrm{T}} \\ \mathbf{L}_m \mathbf{L}_p^{\mathrm{T}} \end{pmatrix} \tag{21}$$

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

$$\varphi_i = \mathbf{i}^{\mathrm{T}} \mathbf{L}_p^{\mathrm{T}} \left(\mathbf{B}_p + \mathbf{R} \right)^{-1} \mathbf{d}. \tag{22}$$

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

$$\mathbf{x}_p = \mathbf{L}_p \mathbf{L}_p^{\mathrm{T}} \left(\mathbf{B}_p + \mathbf{R} \right)^{-1} \mathbf{d}. \tag{23}$$

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

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

We see that we can calculate the analyzed field in all points written in a compact form [4]:

$$\mathbf{x} = \begin{pmatrix} \mathbf{L}_p \\ \mathbf{L}_m \end{pmatrix} \mathbf{L}_p^{\mathrm{T}} \left(\mathbf{L}_p \mathbf{L}_p^{\mathrm{T}} + \mu^2 \mathbf{I}_p \right)^{-1} \mathbf{d} = \mathbf{L}\, \mathbf{L}_p^{\mathrm{T}} \left(\mathbf{L}_p \mathbf{L}_p^{\mathrm{T}} + \mu^2 \mathbf{I}_p \right)^{-1} \mathbf{d}. \tag{25}$$

[4] The reader used to data assimilation can recognise the analysis $\mathbf{x}=\mathbf{B}\mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R} \right)^{-1} \mathbf{d}$ 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

$$\mathbf{x} = \mathbf{L}\left(\mathbf{L}_p^{\mathrm{T}}\mathbf{L}_p + \mu^2\,\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathrm{T}}\mathbf{d}. \tag{26}$$

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

$$\mathbf{L}_p\mathbf{a} = \mathbf{d} \tag{27}$$

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

$$\mathbf{a} = \left(\mathbf{L}_p^{\mathrm{T}}\mathbf{L}_p + \mu^2\,\mathbf{I}_N\right)^{-1}\mathbf{L}_p^{\mathrm{T}}\mathbf{d}. \tag{28}$$

This leads directly to Eq. (26) when the reconstruction uses the weights $\mathbf{a}$ to combine EOFs everywhere. Hence this is equivalent to OI. The major advantage of Eq. (26) 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_pN^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_pN^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 matrixes based only 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. (17), the analysis in those filters are performed in the low-dimensional error subspace instead of the space containing the observation space.

In practise, in order to construct the matrix to invert, there is no need to partition the matrixes into missing and non-missing data points: it is sufficient to use the EOF values only where data are present. The product $\mathbf{L}_p^{\mathrm{T}}\mathbf{L}_p$ 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 $\mathbf{L}^{\mathrm{T}}\mathbf{L}$, the use of a simple flag indicating missing data allows to disregard the corresponding contributions and a direct calculation of $\mathbf{L}_p^{\mathrm{T}}\mathbf{L}_p$.

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 are 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

$$\epsilon^2 = \overline{\varphi_t(\mathbf{r_i})^2} - \mathbf{i}^{\mathrm{T}}\mathbf{L}_p^{\mathrm{T}}\left(\mathbf{L}_p\mathbf{L}_p^{\mathrm{T}} + \mu^2\mathbf{I}_p\right)^{-1}\mathbf{L}_p\mathbf{i}, \tag{29}$$

but we prefer the equivalent form,

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

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

$$\overline{\varphi_t(\mathbf{r_i})^2} = \mathbf{i}^{\mathrm{T}} \mathbf{i}. \tag{31}$$

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

$$\mathbf{C} = \mathbf{I} - \left(\mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p + \mu^2 \mathbf{I}_N \right)^{-1} \mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p = \mu^2 \left(\mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p + \mu^2 \mathbf{I}_N \right)^{-1}. \tag{32}$$

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

$$\epsilon^2 = \mathbf{i}^{\mathrm{T}} \mathbf{C} \mathbf{i}, \tag{33}$$

demanding mN^2 operations to form the matrix products in $\mathbf{C}$ and N^3 operations to invert as before. If for some reason, the square root of the covariance matrix is needed, we can use the eigenvector (or SVD) decomposition,

$$\mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p = \mathbf{W}_p^{\mathrm{T}} \mathbf{\Lambda}_p \mathbf{W}_p, \tag{34}$$

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

$$\mathbf{C} = \mu^2 \mathbf{W}_p^{\mathrm{T}} \left(\mathbf{\Lambda}_p + \mu^2 \mathbf{I}_N \right)^{-1} \mathbf{W}_p. \tag{35}$$

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

$$\mathbf{C} = \mathbf{C}^{1/2} \left(\mathbf{C}^{1/2} \right)^{\mathrm{T}} \tag{36}$$

is therefore:

$$\mathbf{C}^{1/2} = \mu \mathbf{W}_p^T \left(\mathbf{\Lambda}_p + \mu^2 \mathbf{I}_N \right)^{-1/2}. \quad (37)$$

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 $\mathbf{L}_p^T \mathbf{L}_p$ 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.

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

$$\mathbf{E} = \mathbf{L}\mathbf{C}\mathbf{L}^T = \mu^2 \mathbf{L} \left(\mathbf{L}_p^T \mathbf{L}_p + \mu^2 \mathbf{I}_N \right)^{-1} \mathbf{L}^T = \mathbf{S}\mathbf{S}^T, \quad (38)$$

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 $\mathbf{E}$.

To have an idea of amplitude of the analysis error, we can scale the involved matrices on the following ground: The inner matrix to invert 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. (16) we have

$$\mathbf{L}_p^T \mathbf{L}_p \sim \frac{m_p}{m} \frac{1}{n} \mathbf{\Sigma}^N \mathbf{\Sigma}^N. \quad (39)$$

Two extreme situations are worth analysing

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

$$\mathbf{C} \sim \mu^2 \left(\mathbf{L}_p^T \mathbf{L}_p \right)^{-1}, \quad (40)$$

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

$$\mathbf{E} \sim \mu^2 \mathbf{L} \left(\mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p\right)^{-1} \mathbf{L}^{\mathrm{T}}. \tag{41}$$

Formally we use a pseudo-inverse should the inversion become singular. Using the definition (16), this leads to

$$\mathbf{E} = \mu^2 \frac{1}{n} \mathbf{U}^N \mathbf{\Sigma}^N \left(\mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p\right)^{-1} \mathbf{\Sigma}^N \mathbf{U}^{N\mathrm{T}} \sim \mu^2 \frac{m}{m_p} \mathbf{U}^N \mathbf{U}^{N\mathrm{T}}. \tag{42}$$

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

$$\bar{e^2} \sim \mu^2 \frac{N}{m_p}. \tag{43}$$

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. We will come back to this issue later. 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 less degrees of freedom a system has, the easier is it to reconstruct missing points from data in unclouded points.

– At the other extreme, for very large noise

$$\mathbf{E} = \mathbf{L} \left(\mathbf{I} + \frac{1}{\mu^2} \mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p\right)^{-1} \mathbf{L}^{\mathrm{T}} \sim \mathbf{L} \left(\mathbf{I} - \frac{1}{\mu^2} \mathbf{L}_p^{\mathrm{T}} \mathbf{L}_p\right) \mathbf{L}^{\mathrm{T}} \tag{44}$$

which using the same reasoning yields

$$\mathbf{E} = \frac{1}{n}\mathbf{U}^N\mathbf{\Sigma}^N\mathbf{\Sigma}^N\mathbf{U}^{N\mathrm{T}} - \frac{m_p}{mn^2}\frac{1}{\mu^2}\mathbf{U}^N\left(\mathbf{\Sigma}^N\right)^4\mathbf{U}^{N\mathrm{T}}. \tag{45}$$

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

$$\bar{\epsilon^2} \sim \sigma^2\left(1 - \frac{\sigma^2}{\mu^2}\frac{m_p}{N}\right). \tag{46}$$

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 ($\mathbf{L}_p{}^{\mathrm{T}}\mathbf{L}_p$) and hence the relative weights in the analysis step:

$$\frac{\mathrm{tr}\left(\mu^2\mathbf{I}_N\right)}{\mathrm{tr}\left(\mathbf{L}_p^{\mathrm{T}}\mathbf{L}_p\right)} \sim \frac{\mu^2 N}{\sigma^2 m_p} \tag{47}$$

In this last equation we used Eq. (39) and that the sum of the leading N eigenvalues is $nm\,\sigma^2$.

In summary

– For small μ

$$\frac{\bar{\epsilon^2}}{\sigma^2} \sim \frac{\mu^2}{\sigma^2}\frac{N}{m_p}. \tag{48}$$

– For large μ

$$\frac{\bar{\epsilon^2}}{\sigma^2} \sim 1 - \frac{\sigma^2}{\mu^2}\frac{m_p}{N}. \tag{49}$$

In any case, we are now in a position to calculate error estimates in each grid point according to Eq. (33), with a total cost that is proportional to mN^2, both for the construction of **C** and the error calculation. As before, in practice, the calculation of **C** can be done by adequate flagging of operations during matrix multiplications instead of preliminary partitionning. 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 matrixes, 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 **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

$$e^2 = \frac{1}{m^2}\sum_{ij} E_{i,j} \tag{50}$$

where $E_{i,j}$ are the covariances found in the error-covariance matrix **E** of the analysis. Note that when the errors are homogenous 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 matrixes 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 cyclonic gyre, which is more intense in winter and is mainly due to curl of the wind stress (Larnicol et al., 1995). 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 seasonal cycle on 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. The NC flows south-westward following the French and the Spanish coasts along the continental slope. 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 characteristics of 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).

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 (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.

6.2 SST estimation

This 10-year record of SST data has been reconstructed using DINEOF. 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 a 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 regions obscured by the clouds is, the less 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

OSD
3, 735–776, 2006

EGU

in the Ligurian Sea.

6.3 Error estimation

Equation (14) 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 **R**. However such a non-diagonal error-covariance matrix **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 **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 redundant (or strongly correlated) observations:

$$\mathbf{R} = r\mu^2 \mathbf{I}. \tag{51}$$

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

$$r \sim \frac{L^2}{\Delta x \Delta y} \tag{52}$$

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

$$\bar{\epsilon^2} \sim \mu^2 \frac{NL^2}{m_p \Delta x \Delta y} \sim \mu^2 \frac{NL^2}{S} \tag{53}$$

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 instance. The ratio $N/(S/L^2)$ is thus a measure on how well the N EOFs could be captured by the S/L^2 independed 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 approaches were tested. 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 clouded regions.

Here we see how the square root matrix of **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.

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 for the correlation length of the observational error of 29 km is found.

The question arises which from the two approaches is the more realistic one. A possible criteria is a comparison with the correlation length of the SST anomalies, which should be larger than the correlation length of the observational error just found. Independently from the cross-validation error, we estimated 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 $e^{-1} \sim 0.37$ is 80 km (the chosen threshold is based on correlation function of the type $e^{-d/L}$ where d is the distance). This is in agreement with both error length scales, since we can expect 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. Hence we present the results from the second 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 the order of 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).

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 be smaller than this error estimation. The RMS difference between both fields is 0.17°C and indeed smaller than the average error:

$$\sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\epsilon_{ij}^2} = 0.24°\mathrm{C}. \tag{54}$$

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

$$y_{ij} = \frac{\left(x_{ij}^r - x_{ij}^{r,(\mathrm{OI})}\right)^2}{\epsilon_{ij}^2}. \tag{55}$$

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 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 **E** given in Eq. (38):

$$\mathbf{x}_j^r \sim \mathcal{N}(\mathbf{m}_j^r, \mathbf{E}_j), \tag{56}$$

$$\mathbf{x}_j^{OI} \sim \mathcal{N}(\mathbf{m}_j^{OI}, \mathbf{E}_j), \tag{57}$$

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

$$\mathbf{d}_j = \mathbf{x}_j^r - \mathbf{x}_j^{OI} \sim \mathcal{N}(\mathbf{m}_j^r - \mathbf{m}_j^{OI}, 2\mathbf{E}_j). \tag{58}$$

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

$$\tilde{\mathbf{S}} = \sqrt{\frac{n}{2}}\mathbf{C}^{-1/2}\mathbf{\Sigma}^{N-1}\mathbf{U}^{N\mathrm{T}}, \tag{59}$$

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

$$\tilde{\mathbf{S}}\mathbf{E}_j\tilde{\mathbf{S}}^{\mathrm{T}} = \frac{1}{2}\mathbf{I}_N. \tag{60}$$

The transformed variable follows therefore:

$$\mathbf{z}_j = \tilde{\mathbf{S}}\mathbf{d}_j \sim \mathcal{N}(\tilde{\mathbf{S}}(\mathbf{m}_j^r - \mathbf{m}_j^{OI}), \mathbf{I}_N). \tag{61}$$

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

$$\mathbf{m}_j^r = \mathbf{m}_j^{OI}. \tag{62}$$

In this case we would accept the alternative hypothesis H1:

$$\mathbf{m}_j^r \neq \mathbf{m}_j^{OI}. \tag{63}$$

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

$$\mathbf{z}_j \sim \mathcal{N}(0, \mathbf{I}_N). \tag{64}$$

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, $\overline{z}$. This mean is smaller than the critical $z_{\alpha/2}$ value used in a two-sided z-test for α=0.05.

$$|\overline{z}|\sqrt{Nn} = 0.93 < z_{\alpha/2} = 1.96\,. \tag{65}$$

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} follow 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:

$$\sum_{i,j} z_{ij}^2 = 3703 < 28\,644\,. \tag{66}$$

where the z-values are summed over time and over 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 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 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. 5) 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. 6. 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. 7) 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 analised image is 5 days, and the SST on 30 December is notably colder than on 26 December. The DINEOF reconstruction on 30 December 2003 (Fig. 8) presents smoother values and a more realistic SST distribution on the western and northern Ligurian Sea. Both analysis are similar east of the Corsican island, in the Thyrrenian 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 is similar to the reconstruction of Fig. 8 (image not shown).

6.5 Inter-annual variability

As an example of the DINEOF's 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-days cut-off low pass filter (Fig. 9). The seasonal cycle shows an asymmetric behavior: 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. 10. 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. 11) has been computed from Eq. (50). 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. 10. 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 fraction of missing data is 0.92.

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. 10 has an rms value of 0.22°C, much higher than the error estimate found in Fig. 11. Note that if we had also taken only the available data to calculate the confidence interval for the mean for 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.

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 interannual variability of

the spatial means. It was shown that this approach allowed the isolation of interannual 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 analyze 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.

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. 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 MAREXXX.

Appendix A

Useful matrix identities

–

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

–

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

provided the inverse matrix exists and **I** is an identity matrix (with 1 on the diagonal) of appropriate dimension.

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. 736, 740, 741, 746

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. 752

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. 756

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., accepted, 2006. 754

Beckers, J.-M. and Rixen, M.: EOF calculations from incomplete oceanographic data sets, J. Atmos. Ocean Technol., 20, 1839–1856, 2003. 736, 740
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. 737
Boyd, J., Kennelly, E., and Pistek, P.: Estimation of EOF expansion coefficients from incomplete data, Deep Sea Res., 41, 1479–1488, 1994. 745
Daley, R.: Atmospheric Data Analysis, Cambridge University Press, 1991. 738
Emery, W. and Thomson, R.: Data analysis methods in physical oceanography, Pergamon, 1997. 739
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. 745
Evensen, G.: Sampling strategies and square root analysis schemes for the EnKF, Ocean Dynamics, 54, 539–560, 2004. 746
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. 736
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. 736
Gunes, H., Sirisup, S., and Karniadakis, G.: Gappy data: To Krig or not to Krig?, J. Comput. Phys., 212, 358–382, 2006. 737
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. 745
Kondrashov, D., Feliks, Y., and Ghil, M.: Oscillatory modes of extended Nile River records (A.D. 622–1922), Geophys. Res. Lett., 3, L10 702, doi:10.1029/2004GL022156, 2005. 737
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. 752
Lawson, C. and Hanson, R.: Solving Least square problems, Prentice Hall, 1974. 745
Millot, C.: Circulation in the Western Mediterranean Sea, J. Mar. Syst., 20, 423–442, 1999. 752

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. 752
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. 743
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. 752
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. 746
Rixen, M., Beckers, J.-M., Brankart, J.-M., and Brasseur, P.: A numerically efficient data analysis method with error map generation, Ocean Modelling, 2, 45–60, 2000. 736
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. 756
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. 737
Toumazou, V. and Cretaux, J.-F.: Using a Lanczos Eigensolver in the Computation of Empirical Orthogonal Functions, Mon. Wea. Rev., 129, 1243–1250, 2001. 741
Vautard, R., Yiou, P., and Ghil, M.: Singular spectrum analysis: a toolkit for short, noisy chaotic signals, Physica D, 58, 95–126, 1992. 737
Verlaan, J. and Heemink, A.: Tidal Flow Forecasting using Reduced Rank Square Root Filters, Stochastic Hydrology and Hydraulics, 11, 349–368, 1997. 746
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. 752
von Storch, H. and Zwiers, F.: Statistical analysis in climate research, Cambdridge University Press, 1999. 736, 745

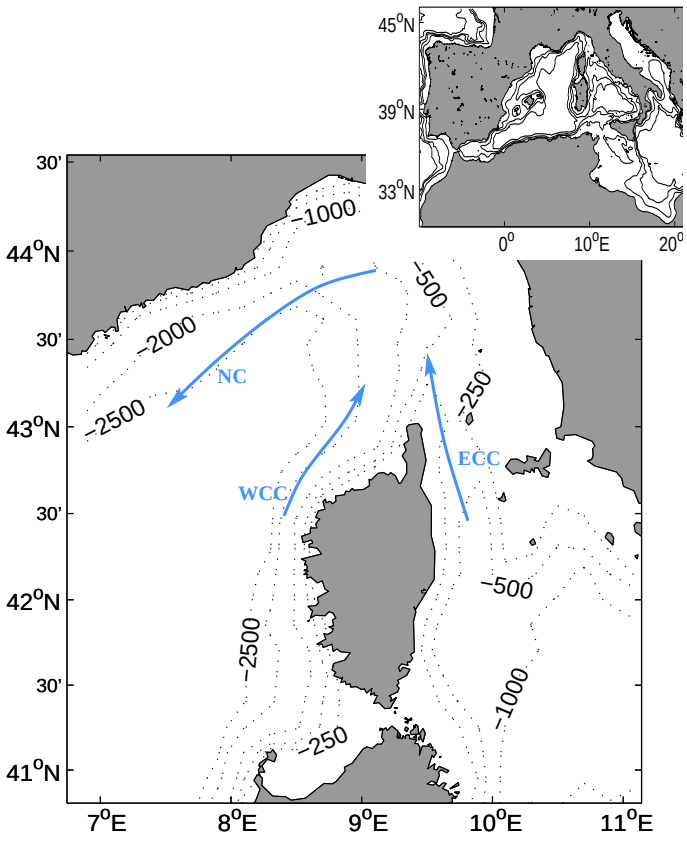

Fig. 1. Zone of interest around the Corsican island in the Northwestern Mediterranean Sea. The Northern Current (NC), flowing southwestward is formed by the Western Corsican Current (WCC) and the Eastern Corsican Current (ECC). Strong frontal regions are associated with these currents.

OSD
3, 735–776, 2006

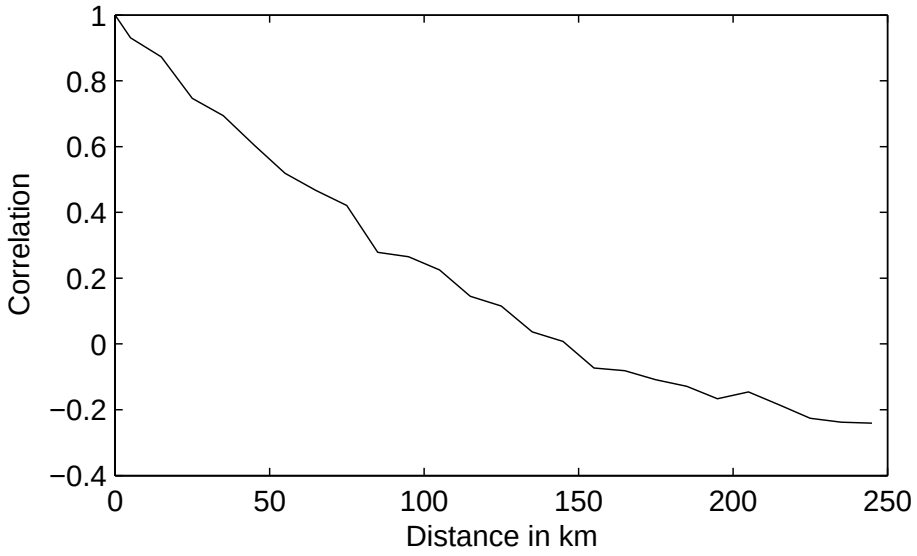

Fig. 2. Spatial SST correlation as a function of distance.

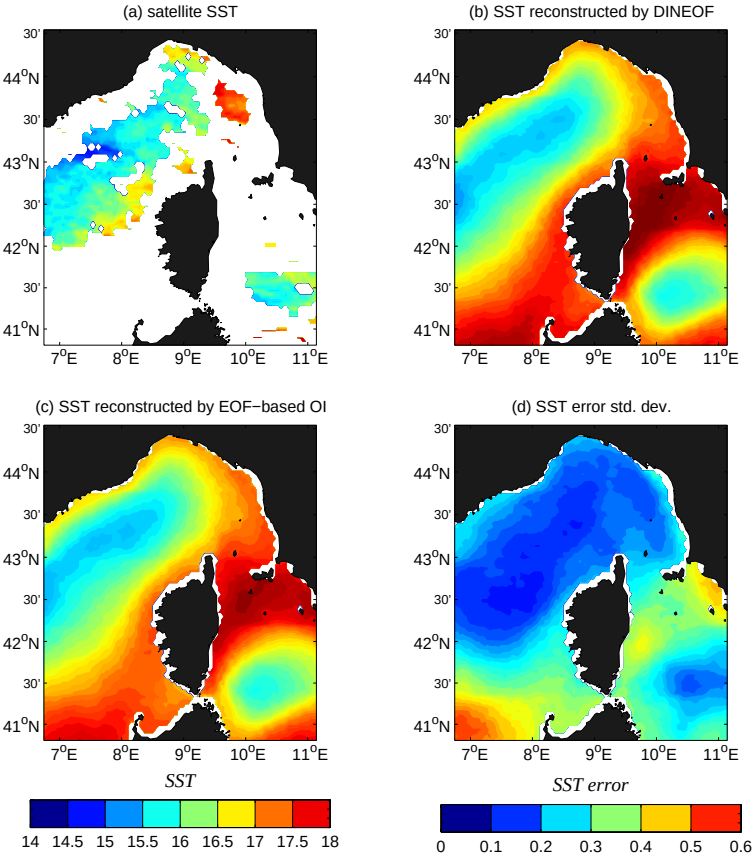

Fig. 3. Panel **(a)** is the observed SST on 15 November 1998. Panels **(b)** and **(c)** show the reconstruction by DINEOF and by optimal interpolation based on the same EOFs and a correlation length of 29 km. The estimated error standard deviation for the reconstruction is shown in panel **(d)**.

OSD
3, 735–776, 2006

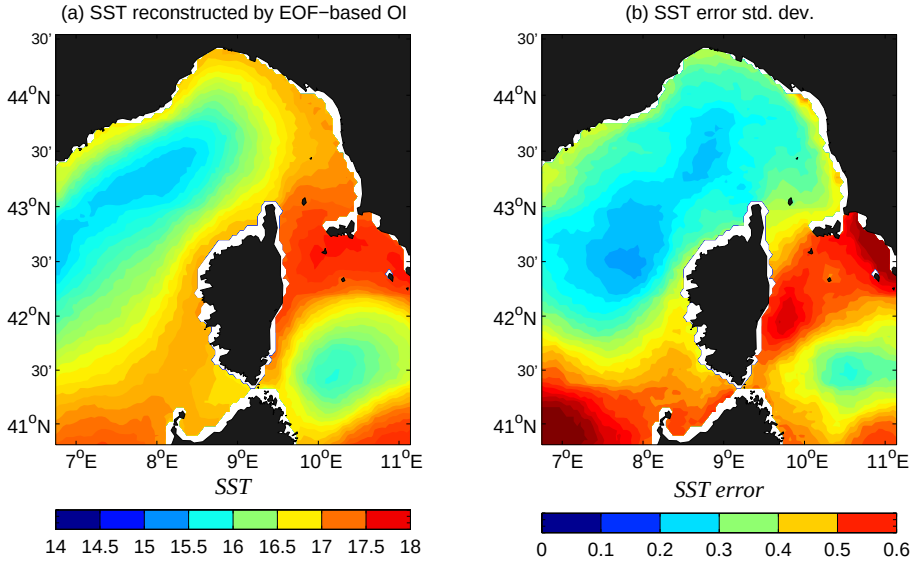

Fig. 4. Panel **(a)** SST on 15 November 1998 reconstructed by optimal interpolation using the same EOFs than DINEOF and a correlation length of 66 km. The estimated error standard deviation for this reconstruction is shown in panel **(b)**.

OSD
3, 735–776, 2006

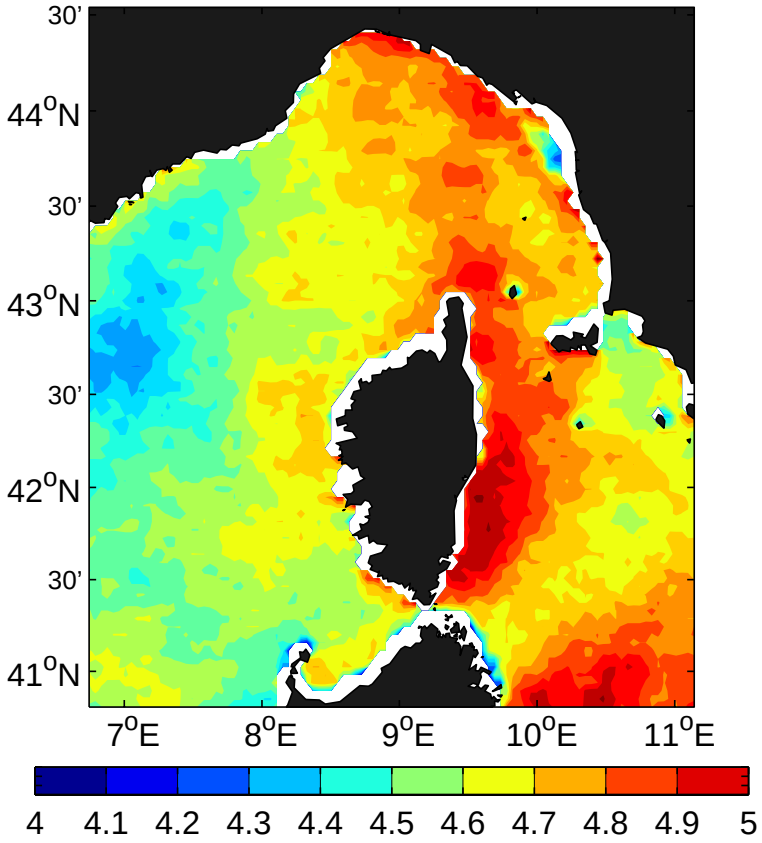

Fig. 5. Standard deviation of SST over the studied time period including the seasonal cycle.

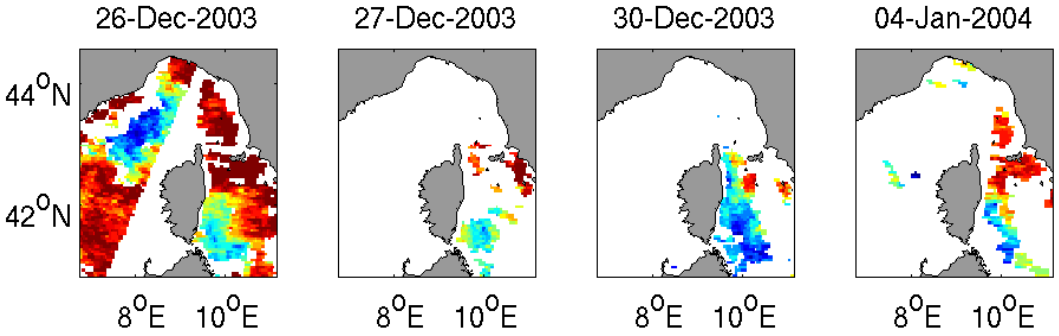

Fig. 6. Data used in the OI reconstruction of 30 December 2003. The colorbar is the same as in Fig. 7.

OSD
3, 735–776, 2006

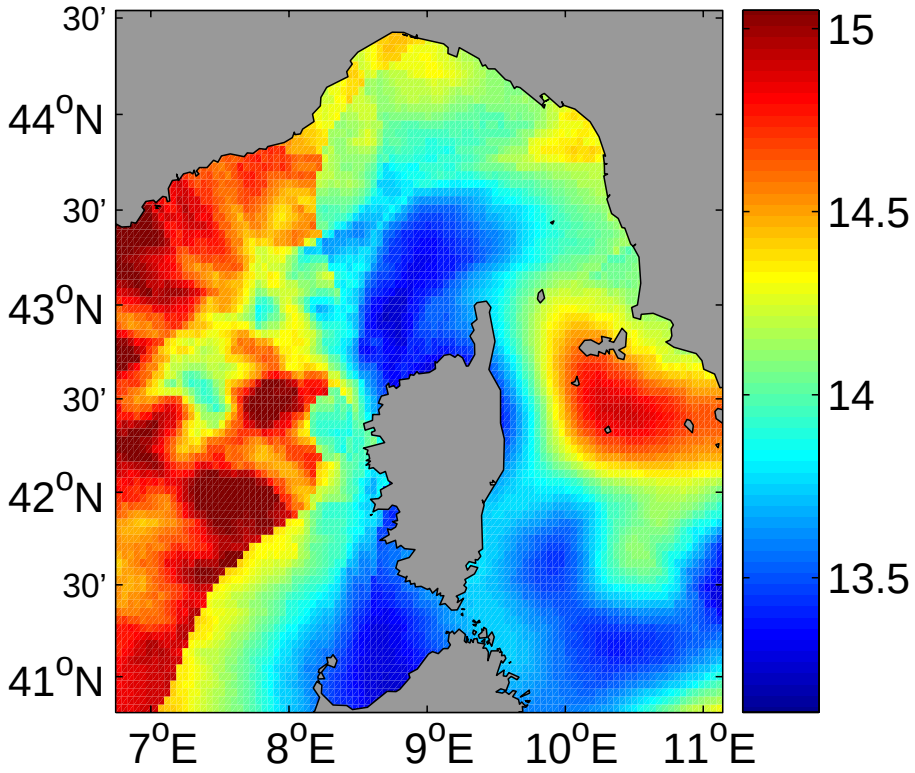

Fig. 7. Reconstruction of the SST on 30 December 2003 by Optimal Interpolation.

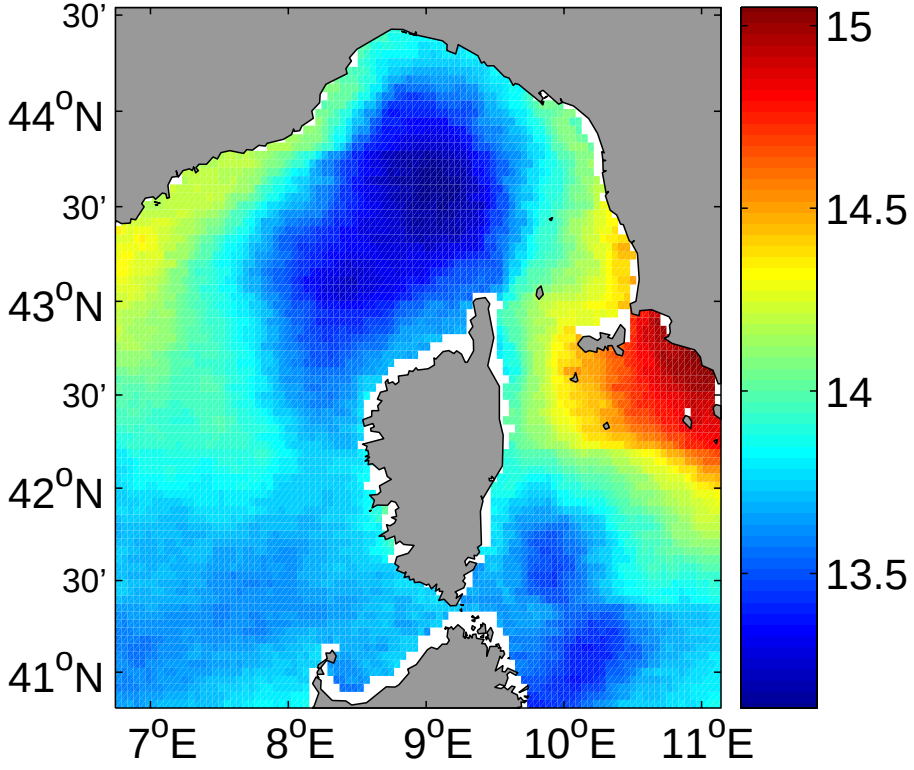

Fig. 8. Reconstruction of the SST on 30 December 2003 by DINEOF.

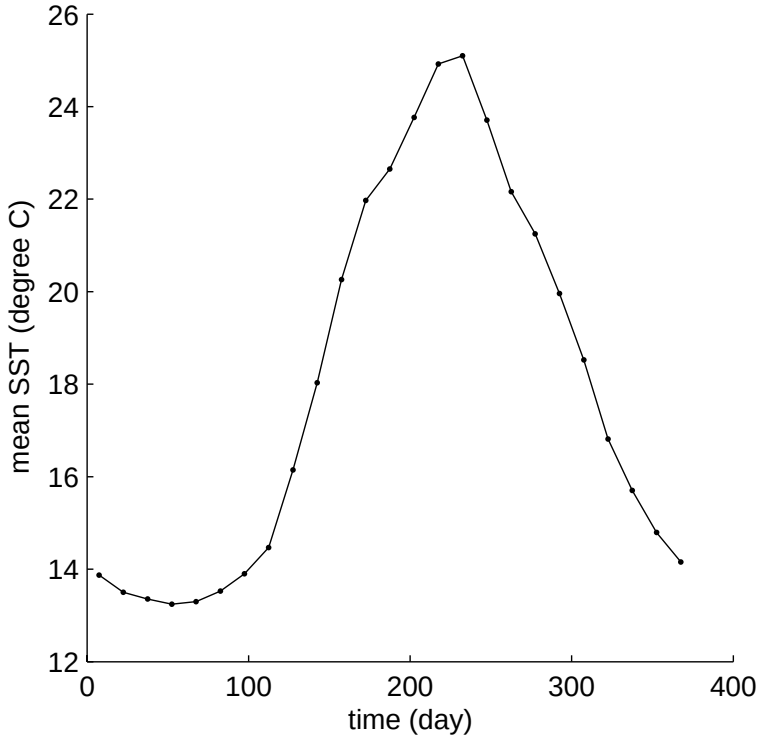

Fig. 9. Seasonal cycle of the spatially averaged SST using reconstructed SST from 1995 and 2005 and filtered with 15-days cut-off low pass filter.

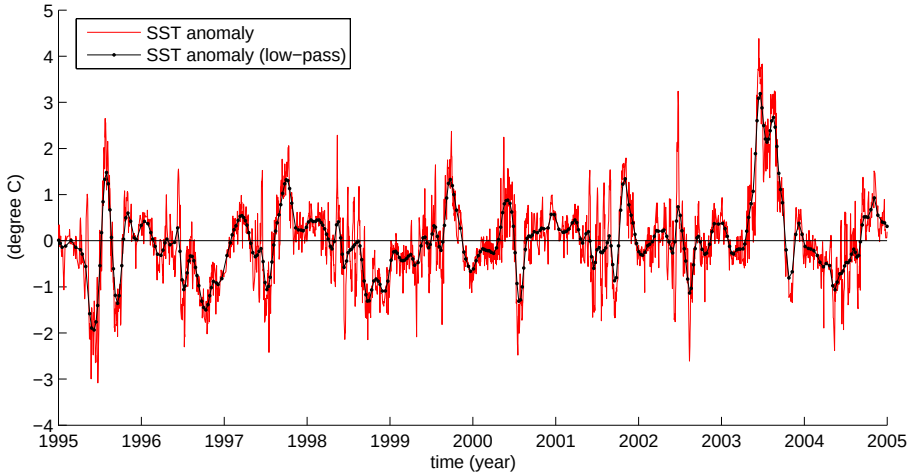

Fig. 10. Mean SST anomalies and filtered mean SST anomalies (15-days cut-off frequencies).

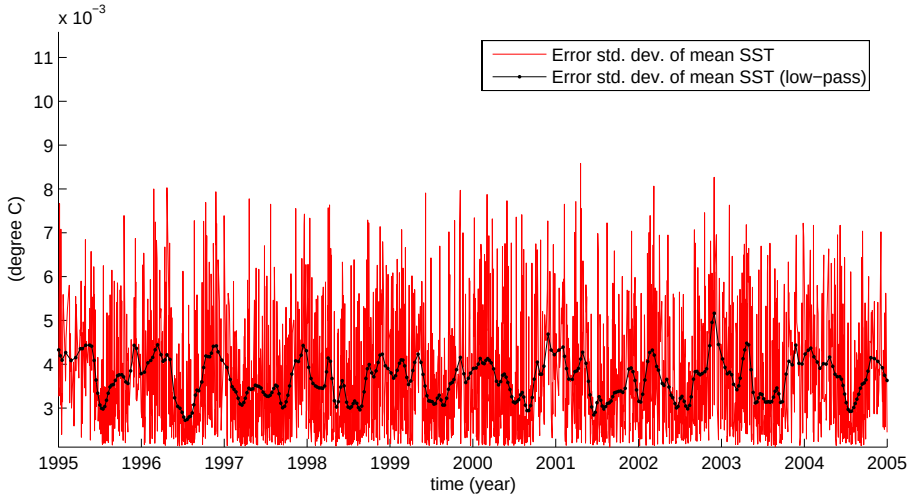

Fig. 11. Error estimate of mean SST and filtered error estimate (15-days cut-off frequencies).