Data Interpolating Empirical Orthogonal Functions (DINEOF): a tool for geophysical data analyses.

Aida Alvera-Azcárate [a,b,*], Alexander Barth [a,b], Damien Sirjacobs [a], Fabian Lenartz [a], Jean-Marie Beckers [a,b]

[a] *GHER-AGO, University of Liège, Belgium*

[b] *National Fund for Scientific Research, FRS-FRNS, Belgium*

Abstract

An overview of the technique called DINEOF (Data Interpolating Empirical Orthogonal Functions) is presented. DINEOF reconstructs missing information in geophysical data sets, such as satellite imagery or time series. A summary of the technique is given, with its main characteristics, recent developments and future research directions. DINEOF has been applied to a large variety of oceanographic variables in various domains of different sizes. This technique can be applied to a single variable (monovariate approach), or to several variables together (multivariate approach), with no complexity increase in the application of the technique. Error fields can be computed to establish the accuracy of the reconstruction. Examples are given to illustrate the capabilities of the technique. DINEOF is freely offered to download, and help is provided to users in the form of a wiki and through a discussion email list.

Key words: Missing data reconstruction, DINEOF, software distribution

1 Introduction

Geophysical data sets, such as those obtained from satellites, often contain gaps (missing values) due to the presence of clouds, rain, or simply due to

* Corresponding author. Address: GHER-AGO, University of Liège. Allée du 6 Août 17, Sart Tilman. 4000 Liège. Belgium.

Email address: a.alvera@ulg.ac.be (Aida Alvera-Azcárate).

URL: http://modb.oce.ulg.ac.be (Aida Alvera-Azcárate).

incomplete track coverage. When dealing with time series of *in situ* data, it is also common to have gaps due to battery failings, weather-related shutdowns of equipment, maintenance and repairing of sensors, etc. There is a rising need, however, for complete data sets at the global, regional and local scale: several analysis methods, input for hydrodynamic models and data visualization are examples of applications where complete data sets are preferred or even necessary. In this paper, the technique called DINEOF (Data Interpolating Empirical Orthogonal Functions) is described. DINEOF reconstructs missing data in spatial fields and time series based on EOFs. An overview of its capabilities and recent developments is given.

An important part in scientific research is the possibility for any scientist to repeat an experiment. Computer-based techniques can be carefully described to allow an interested scientist to re-implement the necessary code to repeat the experiment. However, the availability of tools for specific analyses, and source code of a given technique, allows scientists to promptly and efficiently apply a given technique to their specific data set. A package containing the code of DINEOF is freely available for download, to use and modify to the user needs.

This paper is structured as follows: section 2 gives an overview of the technique, along with its most recent developments. In section 3 we discuss the availability of the DINEOF package and the help offered to users, in the form of a discussion group and help web pages. Section 4 contains the conclusions.

2 Description of DINEOF

2.1 Review of basic method

DINEOF has been developed at the GHER (Geo-Hydrodynamics and Environmental Research) in recent years. The bases of the method are established in [4], which consist on an iterative method to calculate the field values at missing positions: first, a spatial and temporal mean is removed from the analyzed data, and the missing data are initialised to zero. Then one EOF is calculated from this field, and the missing data is replaced with the values obtained by the EOF decomposition. This procedure is repeated until a convergence criterion for the missing data values, specified by the user, is reached. After that, two EOFs are calculated, and the whole procedure is repeated. Then follows three EOFs, etc. The total number of EOFs to be calculated is determined by cross-validation: a few valid data (usually 1% of the total data) are taken apart at the beginning of the procedure, and flagged as missing. At each EOF iteration, the value calculated from the EOF series is compared to the actual

value of these flagged data. The number of EOFs that minimises this difference is retained as the optimal number to reconstruct the data set.

DINEOF was improved in [2] by including an efficient EOF solver developed by [10]. This addition allowed for the reconstruction of large data sets, and was applied to the Sea Surface Temperature (SST) of the Adriatic Sea. Comparison with a global Optimal Interpolation (OI) technique showed that DINEOF was up to 30 times faster in solving the missing data problem, and with a similar accuracy when compared to *in situ* data.

One drawback of the technique lies in the truncated EOF series used for the reconstruction. As only a limited number of EOFs are retained, information at the small scales (such as small vortices or filaments at front edges) is sometimes lost in the reconstruction. These features have a small variability, given their transient nature, and are therefore not included in the most dominant EOFs, unless a sufficiently large number of small scale events are present recurrently at the same location in the data set. This problem can be addressed in some cases by increasing the length of the data set, since a larger number of small scale events might then be present, so the structure of the associated EOFs can be better estimated. As a consequence, a higher number of EOFs are generally retained when the length of the data set increases.

DINEOF can be applied to any variable and in any location of the world ocean. [4], [2], [3] and [9] worked with SST in the Adriatic and Ligurian seas. DINEOF was applied to chlorophyll data in the northern Adriatic Sea by [8]. SST, chlorophyll and winds in the Gulf of Mexico were used in [1]. Other variables are suitable to be used: figure 1 shows an example for total suspended matter (TSM) in the English Channel. Two images extracted from a total of 360 (spanning 4 years) are shown. The reconstructed fields retain the information present in the original data. In zones obscured by clouds, DINEOF is able to reconstruct mesoscale information thanks to the correlated information extracted by the EOF series.

Figure 2 shows an example of how the correlated information is used by DINEOF. Three snapshots of cloudy SST in the central Mediterranean Sea, from a 6-month reconstruction, are shown along with the DINEOF reconstruction. A cold filament is partially seen south of Sicily on the 24 and 26 July 2000. DINEOF retains this information and the reconstructed series of SST shows a coherent sequence, with the cold filament evolving with time. This filament is then also visible on the 25 July 2000, when it was completely obscured by clouds in the original image.

Although all applications mentioned above use spatial maps of satellite data, DINEOF can also successfully reconstruct time series. As an example, figure 3 shows monthly time series of *Posidonia oceanica* leaf area in the Calvi Bay (Corsica). Twelve years were used in the reconstruction.

2.2 Recent developments

DINEOF can be used with multivariate data sets. Different variables can be used together in order to increase the information of a given phenomenon. For example, [1] showed that the concurrent use of SST and chlorophyll *a* concentration, or SST, chlorophyll *a* concentration and wind data in the West Florida Shelf improved the reconstruction of events like the upwelling of cold subsurface water due to the action of the wind. An upwelling has a clear signal in SST (cold water), chlorophyll (high concentration of nutrients) and wind (high winds flowing southwards along the west Florida coast). Information of a given variable at different time steps (for example, SST at time t with winds at time $t-2$) showed also good results, because of the response time in SST to wind changes. [5] and [6] used a technique similar to DINEOF, using singular spectrum analysis instead of EOFs, and showed that the time-lag information helps to improve the reconstruction of missing data.

An advantage of multivariate DINEOF compared to other missing data reconstruction techniques lies in the non-parametric nature of the technique. No information about the covariance between different variables is necessary, facilitating the analyses and decreasing the subjectivity that is often required to compute such fields. An example of a multivariate DINEOF reconstruction is given in figure 4.

Recently developed is the possibility of computing error fields from the analysed fields [3], allowing the user to quantify the confidence in the reconstructed data. For example, zones with high cloud coverage will have higher errors in the reconstruction. An example is given in figure 5. Three years of SST in the Black Sea (2003-2005) have been reconstructed using DINEOF. Figure 5 shows the initial data and results obtained on 19 June 2004. Almost the totality of the Black Sea is covered by clouds on that date, except for a band in the eastern part of the basin. The error field reflects this distribution, with lower errors in that zone. The error fields are calculated using the EOFs obtained by DINEOF as the background error covariance. Then an OI approach is used

to compute the estimated error at each point of the reconstruction.

3 DINEOF availability

3.1 Source code and Linux and Windows packages

An important aspect in any scientific development is the possibility for other scientists to reproduce the results obtained with a specific technique. DINEOF consists of a series of Fortran 90 routines that have been assembled and are freely available under the GNU General Public License. The complete package can be downloaded at `http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF`. In this web page, the latest version of the software is found, tested by the developers and including the most recent developments. These routines can be compiled and modified to the user's needs. The software ARPACK [7] for solving eigenvalue problems is needed, and if netCDF support is required, then the netCDF library has to be installed as well.

For those users with little or no experience in compiling programs from the source code, a pre-compiled, ready-to-use package is also available. This package exists for Linux and Windows platforms, and requires no knowledge of Fortran nor any compilation.

3.2 Help pages in form of Wiki

At the same web page, `http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF`, a complete guide to assist the new user through the installation and use of DINEOF is also available. Written in the form of a wiki to allow for its rapid edition by developers and users, it guides the user through the process of downloading, installing and using the software in various platforms. Installation instructions for ARPACK and netCDF are also given. A list of references is updated with the most recent publications. DINEOF comes with an example data set (original data and reconstruction), so users can check the correct installation of the package.

3.3 Discussion group

For questions related to the installation, use, or any scientific discussion about DINEOF, a discussion group has been set up at `http://groups.google.com/group/dineof`. Users can, through e-mails, ask the discussion group about advise on how to install DINEOF might any problem arise. The developers of DINEOF, or any other user willing to assist, can answer the questions posted to the group. The benefit of a discussion group is that new users can see the history of emails sent previously and find an answer to their problems, and that, as the discussion group grows, more people have the experience to assist others in the installation and application of DINEOF, making feedback quicker and richer. If an error is found by a user, it can be signaled to the discussion group: the developers can correct the mentioned error, and other users can be directly aware of its existence.

4 Conclusions

DINEOF is an EOF-based technique to reconstruct geophysical data sets such as satellite images and time series. Missing data, such as those caused by the presence of clouds in some satellite data, can be reconstructed with the help of the correlated information present in the data set. It is therefore important to have sufficiently long time series of data to allow for an accurate reconstruction. DINEOF has been applied to a wide variety of variables (SST, chlorophyll *a*, winds, TSM, seagrass leaf area, etc) and domains (Adriatic Sea, Ligurian Sea, Mediterranean Sea, Black Sea, southeast US coast, etc). This parameter-free technique allows for a rapid reconstruction of missing data without the need of *a priori* information about the data set, such as the covariance between variables or the correlation length. This makes DINEOF easy to use, and allows for an accurate reconstruction of missing data. DINEOF reconstructions have been compared with independent data in several applications, giving similar or even better results than those obtained with other reconstruction methods such as optimal interpolation. Recent developments (such as multivariate reconstruction and the generation of error fields) have been explained, and examples given.

DINEOF is freely available to download at `http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF` under the GNU General Public License, and users can modify the code to their needs. The development of new features in DINEOF is in part the result of the feedback obtained through the discussion group, so user interaction is an important part in the process of improving DI-

NEOF. The developers then benefit also from the availability of the software, in the form of feedback from the users.

Future directions in the development of DINEOF include the application of the technique to very small and very large domains, two problems that need different approaches: large domains need high computing resources, therefore we need to optimize the resources needed by DINEOF. Small domains contain information at very small scales, however DINEOF retains mainly information at the meso and large scales. Modifications to improve the reconstruction of small scale information will be needed for such applications.

Acknowledgments

The National Fund for Scientific Research (FRS-FNRS), Belgium is acknowledged for funding the postdoctoral positions of A. Alvera-Azcárate and A. Barth. J.-M. Beckers is Honorary Research Associate at the FNRS, Belgium. The SST and wind data were obtained from the Physical Oceanography Distributed Active Archive Center (PO.DAAC) at the NASA Jet Propulsion Laboratory, Pasadena, CA, (`http://podaac-www.jpl.nasa.gov/`). The MERIS TSM data were supplied by the European Space Agency under Envisat AOID698, and made available by the MUMM through the BELCOLOUR-1 database, online at `http://www.mumm.ac.be/BELCOLOUR/EN/Products/index.php`. Kevin Ruddick, Bouchra Nechad and Youngje Park are acknowledged for sharing these data within the context of the project RECOLOUR (REconstruction of COLOUR scenes - SR/00/111), funded by the Belgian Science Policy (BELSPO) in the frame of the Research Program For Earth Observation "STEREO II". Sylvie Gobert, Gilles Lepoint, and the personnel at STARESO (Calvi) are acknowledged for providing the *Posidonia oceanica* data.

References

[1] Alvera-Azcárate, A., Barth, A., Beckers, J. M., Weisberg, R. H., 2007. Multivariate reconstruction of missing data in sea surface temperature, chlorophyll and wind satellite fields. Journal of Geophysical Research 112, C03008, doi:10.1029/2006JC003660.

[2] Alvera-Azcárate, A., Barth, A., Rixen, M., Beckers, J. M., 2005. Reconstruction of incomplete oceanographic data sets using Empirical Or-

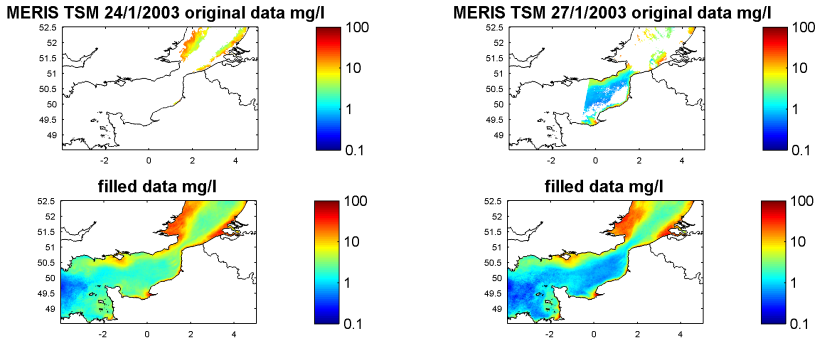

Fig. 1. TSM in the English Channel. Top row shows original data, and bottom row shows the DINEOF reconstructed data.

thogonal Functions. Application to the Adriatic Sea surface temperature. Ocean Modelling. 9, 325–346.

[3] Beckers, J.-M., Barth, A., Alvera-Azcárate, A., 2006. DINEOF reconstruction of clouded images including error maps. Application to the Sea Surface Temperature around Corsican Island. Ocean Science 2 (2), 183–199.

[4] Beckers, J.-M., Rixen, M., 2003. EOF calculations and data filling from incomplete oceanographic data sets. Journal of Atmospheric and Oceanic Technology 20 (12), 1839–1856.

[5] Kondrashov, D., Feliks, Y., Ghil, M., 2005. Oscillatory modes of extended Nile River records (A.D 622-1922). Geophysical Research Letters 32, L10702.

[6] Kondrashov, D., Ghil, M., 2006. Spatio-temporal filling of missing data in geophysical data sets. Nonlienar Processes in Geophysics 13, 151–159.

[7] Lehoucq, R. B., Sorensen, D. C., Yang, C., 1997. ARPACK user's guide: solution of large scale eigenvalue problems with implicity restarted Arnoldi methods, 1–152.
URL `http://www.caam.rice.edu/software/ARPACK/`

[8] Mauri, E., Poulain, P. M., Juznic-Zontac, Z., 2007. MODIS chlorophyll variability in the northern Adriatic Sea and relationship with forcing parameters. Journal of Geophysical Research 112, C03S11, doi:10.1029/2006JC003545.

[9] Mauri, E., Poulain, P. M., Notarstefano, G., 2008. Spatial and temporal variability of the sea surface temperature in the Gulf of Trieste between January 2000 and December 2006. Journal of Geophysical ResearchIn Press, doi:10.1029/2007JC004537.

[10] Toumazou, V., Cretaux, J. F., 2001. Using a Lanczos eigensolver in the computation of Empirical Orthogonal Functions. Monthly Weather Review 129 (5), 1243–1250.

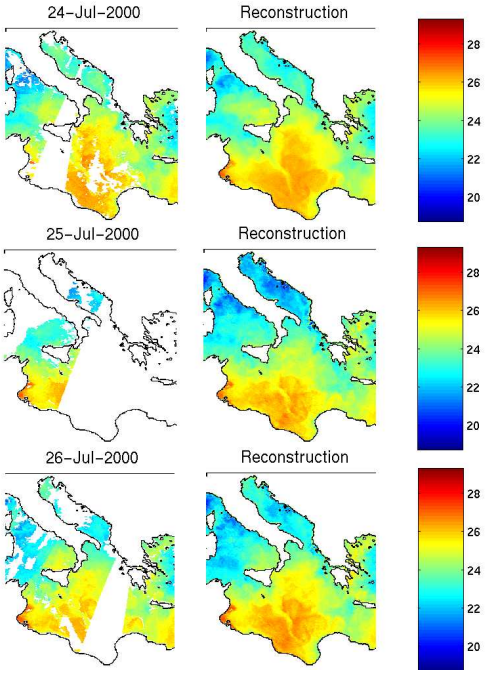

Fig. 2. SST in the central Mediterranean Sea. Left column shows original SST, with missing data, at three consecutive days. Right column shows the reconstructed SST for the same dates.

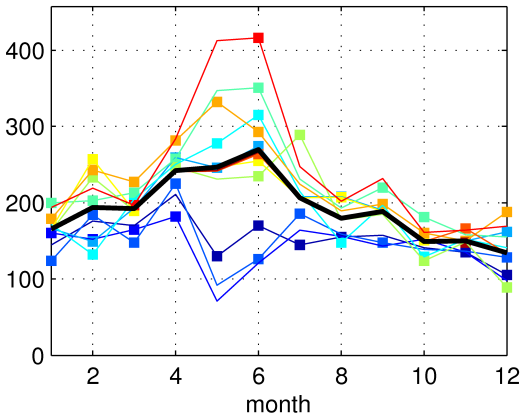

Fig. 3. Example of time series reconstruction. Leaf area (in m^2) of *Posidonia oceanica* in the Calvi area (Corsica) from 1992 to 2005. Each color represents a different year. Squares show months when leaf area was measured, and months without squares have been reconstructed with DINEOF. In total, 55% of the data is initially missing.

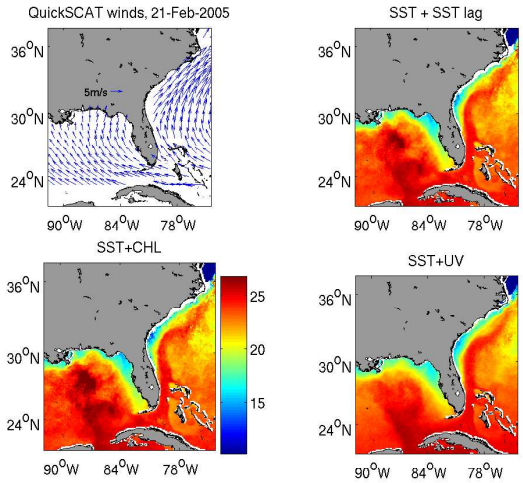

Fig. 4. Example of a six-month multivariate reconstruction with DINEOF on 21 February 2005. Top left panel shows the wind field,reconstructed along with SST; top right panel shows SST reconstructed with a 1-day lagged SST; bottom left panel shows SST reconstructed along with chlorophyll *a* concentration; bottom right panel shows SST reconstructed along with the wind field.

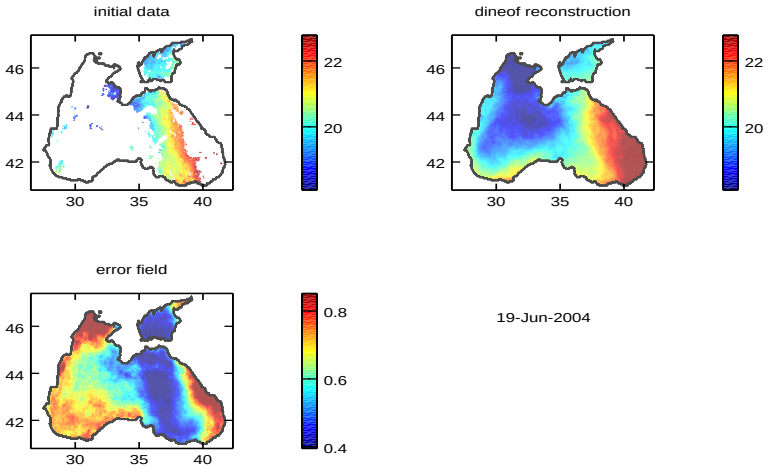

Fig. 5. Example of error field as obtained with DINEOF. Top left figure shows the initial cloudy data on 19 June 2004. Top right figure shows the reconstructed data for the same date. Bottom figure shows the error field. The error variance (taken to be the variance not retained by the EOF basis used by DINEOF) is 0.33°C.