Submitted for the Proceedings of the 2nd MERIS -(A)ATSR workshop, held in Frascati, 22-26 September 2008, ESA Special Publication SP-666

Cloud filling of total suspended matter, cholorphyll and sea surface temperature remote sensing products by the Data Interpolation with Empirical Orthogonal Functions methodology, application to the BELCOLOUR-1 database.

Damien Sirjacobs [(1)], Aida Alvera-Azcárate [(1)], Alexander Barth [(1)], YoungJe Park [(2)], Bouchra Nechad [(2)], Kevin Ruddick [(2)] and Jean-Marie Beckers [(1)]

[(1)] GeoHydrodynamic and Environmental Research, MARE center; University of Liège; Allée de la Physique 7, B5; 4000 Sart-Tilman; Belgium. Email: d.sirjacobs@ulg.ac.be

[(2)] Management Unit of the Mathematical Model of the North Sea; Royal Belgian Institute of Natural Sciences; Avenue Guledelle 100, Bruxelles, Belgium

1 ABSTRACT

Space-time filling of the gaps in satellite data archives is an important step for the improvement of various marine ecosystem studies. The Data Interpolation with Empirical Orthogonal Functions methodology (DINEOF) allows calculating missing data in geophysical datasets without requiring a priori knowledge about statistics of the full data set [1]. It was successfully applied to SST reconstructions as in [1] and [2]. Here, the DINEOF reconstruction method is applied to surface chlorophyll a (CHL), total suspended matter (TSM) and sea surface temperature (SST) data over the Southern North Sea and English Channel obtained from the BELCOLOUR archive.

2 INTRODUCTION

The objective of this study is to demonstrate a method for reconstruction of complete space-time information for surface CHL and TSM from an archive of satellite imagery. The resulting data is better suited for applications such as algae bloom detection or for providing light forcing for ecosystem modeling. In the longer term comparison of satellite data with reconstructed fields will contribute to the quality control of satellite data by highlighting suspect or extreme data.

Temporal coverage of OC data is not enough. Wide swath polar-orbiting ocean colour remote sensors acquire data with near-global coverage of the world's oceans and seas every few days. For example, Belgian waters at 51°N are imaged by MODIS-Aqua every day and by MERIS on average every 3-4 days. However, this maximal temporal coverage is greatly reduced by clouds and by sunglint. The usable data is further reduced for environmental conditions where derived products are considered to be of unacceptable quality because of various processing problems particularly relating to atmospheric correction (adjacency effects, high aerosol optical thickness, absorbing aerosols, cloud-edge and cloud shadow, low sun or viewing zenith angle, etc.). This temporal coverage, although far superior to shipborne sampling methods, is insufficient for many applications.

For example, for CHL based products, the development of harmful algae blooms can occur rapidly, over the period of a few days, and a satellite-only detection system may be rendered completely inoperative if this coincides with a cloudy period. As another example, TSM products are used by ecosystem modelers to control the light forcing in simulations designed to hindcast or forecast eutrophication as function of anthropogenic nutrient inputs [3]. These models require complete spatio-temporal data fields as input. Moreover it is important that such inputs contain as much of the high frequency variability as possible since TSM dynamics, such as the clearing of the water column by settling after a storm event, may be responsible for triggering algae blooms. More generally, users of satellite data products, such as marine scientists investigating conditions at specified sampling locations, prefer to receive a continuous time series of data rather than the gappy series typically provided directly from optical remote sensors. There is, therefore, a strong user demand for complete time series and cloud-free maps of CHL and TSM products. This is the primary motivation for the present study which has the objective of generating spatio-temporally complete 3D (horizontal space and time) fields of surface CHL and TSM from a collection of individual instantaneous images of these fields as retrieved from MERIS and MODIS.

The use of cloud filling techniques in ocean colour imagery is much less developed then in sea surface temperature imagery, perhaps because the satellite data has become easily available only recently or perhaps because CHL retrieval is notoriously more error-prone than SST retrieval. Examples of cloud filling of CHL images are provided in [4]. Use of a Kriging approach for cloud-filling of MERIS CHL imagery is described in [5]. Aspects of spatial and temporal interpolation of ocean colour data are addressed in [6] in the context of merging of global CHL data from missions such as SeaWiFS and MODIS. Simple

interpolation/replacement techniques using nearby pixels in space or time are used by [7] to fill cloudy MODIS imagery.

3 DATA

3.1 The BELCOLOUR database

Satellite images used in this study are extracted from the BELCOLOUR database [8]. This archive includes level 2 parameters such as chlorophyll *a* concentration CHL and total suspended matter TSM from SeaWiFS (1997-2004), MODIS-Aqua (2002-present) and MERIS (2002-present) for the North Sea [48.5°N-60°N, 4°W-9°E]. Sea surface temperture data are also archived from MODIS-Aqua (2002-present). These data, originally provided in the scan coordinates, are re-sampled in an equi-rectangular projection with a 1km spatial resolution to facilitate the use in applications including the DINEOF analysis. In this step, unreliable pixels are masked using the level 2 flags.

3.2 Selected data

The present study focuses on the subregion of the English Channel and the Southern Bight of the North Sea [48.5°N-52.5°N, 4°W-5°E] and was based on 595 MERIS TSM and CHL images (2003-2006) and on 3376 MODIS TSM and SST images (2002-2006). In the original datasets, the temporal proportion of missing data calculated for each sea pixel ranges from 75 to 100 % (without distinction of the reason : pixel out of satellite track, cloud cover, unreliable data).

4 METHODOLOGY

4.1 Pre-processing

In order to avoid the production of some artefacts in the EOFS, some limitations have to be set on the acceptable spatio-temporal proportion of missing data as in [2]. Prior to Dineof treatment, it was chosen to eliminate each image holding less then 5 % of the expected data. This reduced the number of exploitable images to 356 for MERIS and 1291 for MODIS. The same elimination criteria was applied in time, excluding thus from the study all pixels holding less then 5 % of valid data through the temporal dimension. After this first selection, the MODIS TSM data set presented a sligthly lower proportion of missing data than the MERIS dataset (69% against 73%).

In order to enhance the sensibility of the DINEOF analysis to the spatio-temporal variations of CHL and TSM data occuring in the lower part of the ranges, the base 10 logarithm of the data was taken instead of direct units. This scale change prior to the analysis also prevents any reconstructed pixels to reach negative values of the direct unit scale. The background field is calculated as mean value observed in each pixel over all selected images. This field is then deduced from the dataset in order to provide DINEOF with anomalies around the mean local value measured in base 10 logarithm.

4.2 DINEOF: described methodology

When having only cloud-free images, a very efficient way to synthesize the information contained in a collection of scenes is the use of empirical orthogonal functions (EOFs, also called principal components in other research domains). These functions have some interesting properties: when only one EOF is used, this EOF is on average the closest to all images, when multiplied for each image by appropriate amplitude. Hence it is the best possible approximation of all images using only one spatial pattern (or EOF) and an amplitude for each image. With two EOFs, it can be shown that no other combination of two patterns can provide a better approximation to all images than these two. In general the first N EOFs are therefore the best way to summarize the information content of all images if only N pattern can be stored. Each image is then replaced by a filtered version in which the basic patterns are linearly combined with amplitudes corresponding to each image. When images are sequential in time, the amplitudes can be interpreted as a time evolution of the spatial patterns amplitude and we will refer to them as temporal modes. The practical calculation of the EOFs can be performed by a singular value decomposition of the data matrix X. To construct the data matrix, each scene is stored as a one-dimensional array and corresponds to a column of the matrix X. The SVD decomposition thus provides 3 matrixes as in Eq. 1, giving direct access to the spatial patterns (colmumns of matrix U), the temporal evolution of these patterns (columns of V) and their overall amplitude (S).

$$X = U \,.\, S \,.\, V^T \qquad (1)$$

The amplitudes are generally stored by decreasing importance so that when using not all EOFs but only the first N, we neglect the smallest contributions. In this case, the truncated representation is given in Eq. 2, where the matrices on the right hand side only contain N columns corresponding to the N EOFs retained.

$$X_a = U_N \,.\, S \,.\, (V_N)^T \qquad (2)$$

Retaining only these dominant EOFs filters out some information from the scenes and it is customary to quantify the filtering effect by providing the explained variance when retaining N EOFs. This quantity is

generally expressed as a percentage of the total variance (information content) of the original data.

If we had cloud-free images, EOFs could be calculated easily and an approximate representation of each image obtained as a truncated combination of a few EOFs. Hence we can imagine to use this combinationof EOFs for points in which we do not have data to interpolate the missing data there. Of course we have a circular dependence because the calculation of EOFs requests a set of cloudless images and the interpolation of the missing data requests the knowledge of the EOFs. To solve this problem, an iterative method was implemented in the DINEOF package:

Assuming we know the first EOF, we can estimate the missing data value at any location with this EOF. Once we have this value, the EOF can be recalculated and so on until convergence. Then a second EOF is taken into account with the same approach, before going to a third and so on.

There remains to initialise the iterative process and to decide when to stop adding EOFs to the reconstruction. The first point is easily dealt with by putting a first guess of zero anomalies in the missing data points, while the number of retained EOFs is fixed by a cross-validation technique: a few data points are set aside by adding virtual clouds on some scenes and an rms misfit between the reconstruction and the data set aside is calculated for each reconstruction. The number of EOFs retained is then naturally the one that leads to the minimal misfit. For more details we refer to [1] and [2].

4.3 Production of complete fields at regular time steps and extraction of multitemporal averages

Once the EOFs are defined by DINEOF, they can be exploited to regenerate full fields at any intermediate moments when no satellite images were acquired, by assuming that a linear interpolation of the temporal eofs is a valid estimate of their evolution. In the present work, full fields were produced at daily intervals for the whole period. For MODIS, this temporal resolution is generally comparable to the frequency of exploitable images and is thus meaningfull, except in some winter periods. Thus DINEOF treatment of MERIS products in the North Sea area allows producing weekly averaged fields as seasonal climatologies studies.

5 RESULTS

Optimal reconstructions (minimising the global error estimator) were obtained by DINEOF when synthetising the original signal into 8 modes for MERIS CHL and into 18 modes for MERIS TSM. The variability of these original signals explained by the EOF synthesis reached 93.5 % for CHL and 97.2 % for TSM. For the MODIS TSM dataset, the 97.5 % of the original variability of the signal could be synthetised into 14 modes, with less weight on the first mode comparativelly to MERIS TSM, revealing thus that the secondary TSM dynamic modes are better captured as consequence of the higher frequency of image acquisition and of the lower global proportion of missing data of the MODIS dataset. The MODIS SST dataset could be synthetised into 13 modes explaining 98 % of the input signal variability, of which 67% only by the first mode, underlying the strong seasonal pattern of the SST dynamic for this area.

5.1 Background fields and 3 dominants EOFs retained for MERIS TSM

For MERIS TSM, the background field is illustrated in Fig.1, the 3 dominants spatial EOFs are illustrated in Fig.2, Fig.3 and Fig.4 together with the corresponding parameter ‘varex’ representing the variability of the original signal explained by each EOF, whereas the associated temporal modes are illustrated in Fig.5.

The TSM background field shows a general raising gradient towards the coasts, an inverse correlation with water depth, and a clear influence of large estuaries.

The first mode of MERIS TSM account for about 40 % of the signal and is clearly seasonal signal, being positive in winter and negative in summer: it shows general winter increase of surface TSM in most of the domain but particularly in shallow areas, and the opposite in summer. The contribution of this EOF in the most western and central part of the English Channel is opposite, with TSM positive contribution in summer and negative contribution in winter, relatively to the background field. The second mode is generally representing the dynamic of some summer local reduction (relatively to the previoussly explained signal) in the south-east coasts of England and in the coasts of Normandy in France, as in large part of the southern Bight of the North Sea. By opposition this mode describes a mecanism of increase of TSM in the western and central part of the English Channel, relatively to the dynamic explained by the previous mode. Third mode shows already complex spatio temporal modulations, and the complexity of interpretation of the modes raises as we look at the further retained modes, having progressively less weigth in the reconstruction of the complete signal.

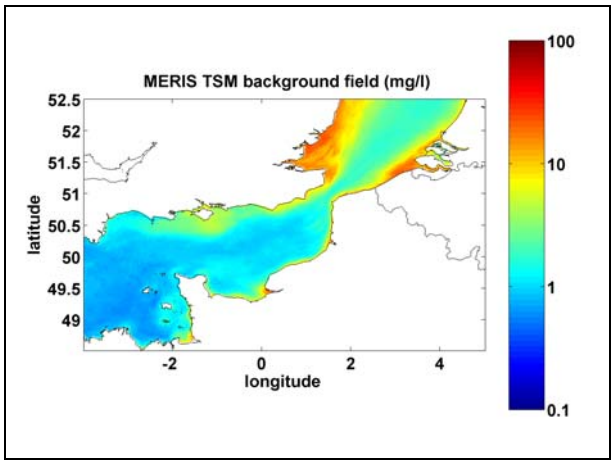

Figure 1. Background field obtained forMERIS TSM.

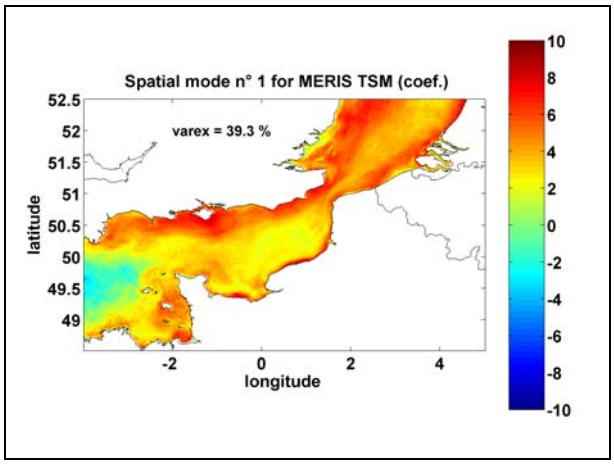

Figure 2. Spatial EOFs obtained for the first mode of MERIS TSM signal (map of coefficient of variation around the background field).

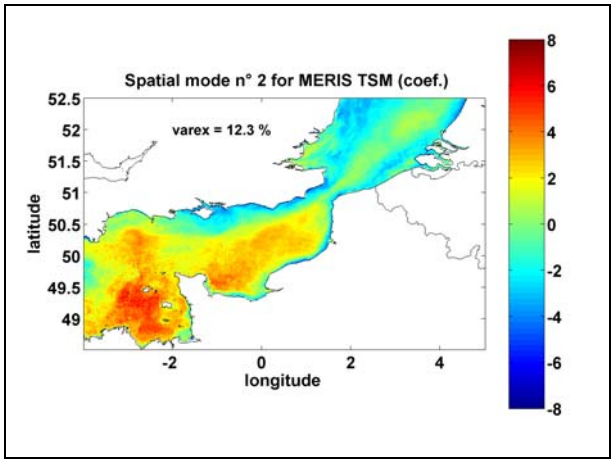

Figure 3. Spatial EOFs obtained for the second mode of MERIS TSM signal (map of coefficient of variation around the background field).

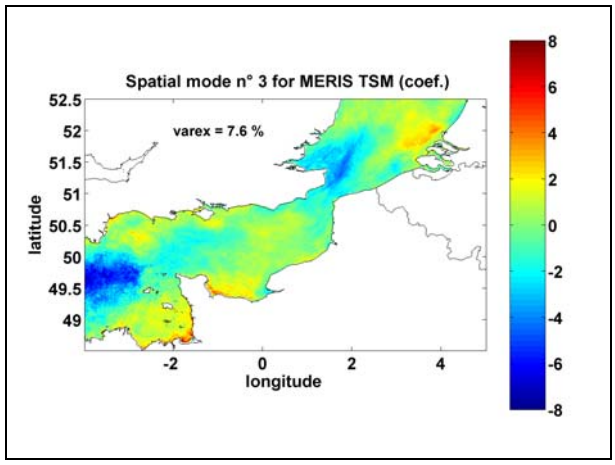

Figure 4. Spatial EOFs obtained for the third mode of MERIS TSM signal (map of coefficient of variation around the background field).

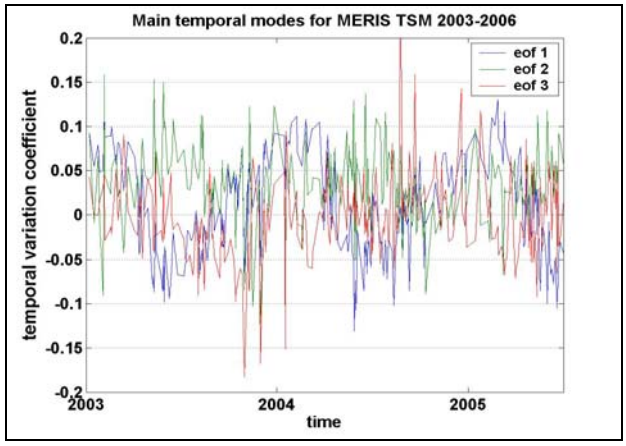

Figure 5. Temporal EOFs obtained for the 3 first modes of MERIS TSM signal (coefficients of variation of weight of each mode in the signal reconstruction, zoom on the period 01/2003-06/2005).

5.2 Multitemporal averages and time series extraction at reference stations.

Weekly and monthly averaged fields were produced from daily fields reconstruction for MERIS TSM and CHL products, as for MODIS TSM and SST products. As illustration for MERIS TSM and CHL, weekly averaged seasonal signals were extracted at two reference stations: the Scheelde turbidity maximum station in the Belgian shelf, and the CEFAS buoys "West G" in the U.K. waters (Fig.6 and Fig.7). These time series shows well the strong seasonal TSM dynamics and the onset of the spring CHL bloom corresponding to the sharp reduction of TSM, higher TSM and CHL concentration of the Scheelde plume station regarding to the 'WestG' station, as the unusually intense spring bloom event observed in the

Scheelde plume in 2003 due to a combination of unusual levels of PAR light and Phosphorus concentrations as described in [9].

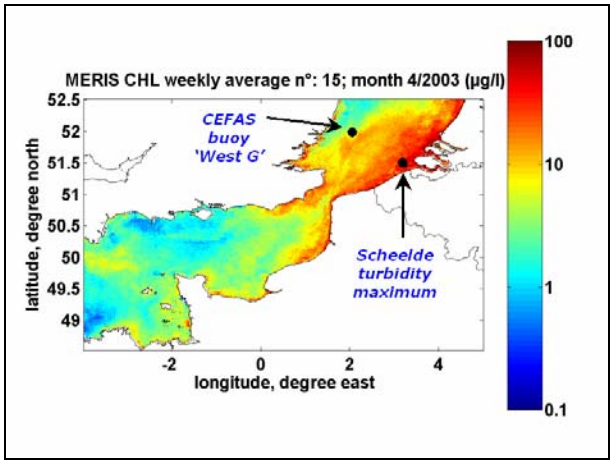

Figure 6. Localisation of the'CEFAS West G' and 'Scheelde Turbidity Maximum' stations on a weekly averaged reconstruction showing the intense bloom event of spring 2003.

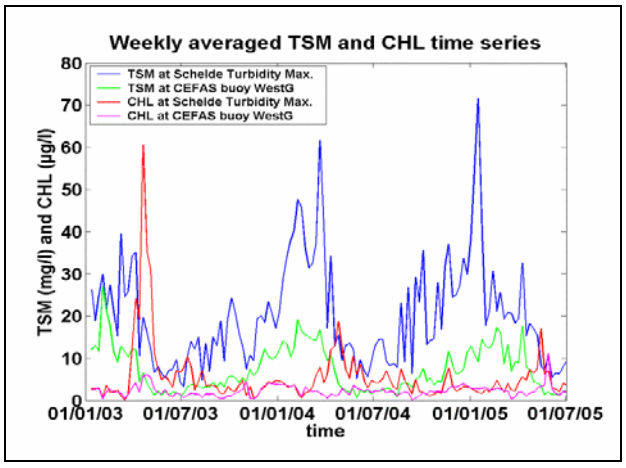

Figure 7. Weekly averaged time series of MERIS TSM and CHL DINEOF reconstructions at reference stations 'CEFAS West G' and 'Scheelde Turbidity Maximum'.

6 CONCLUSIONS

This study shows promising applications of the DINEOF methodology to ocean optical remote sensing data. It allowed to fill in a 4 years set of MERIS CHL and TSM products, as of MODIS TSM and SST with sets of EOFs representing from 93 to 98 % of the variability of the input signal. Weekly and Monthly averaged reconstructed fields were produced from regular daily reconstructions, underlying the interest of the global method for the establisment of precise surface water seasonal climatologies. These products were exploited by other studies for comparison with in situ data, as to attempt further improvements of marine ecosystem models using remote sensing products as forcings [3].

Any subregional or local multitemporal averages can be reproduced with the described DINEOF approach and provided to interested users, according to the objective of their study.

Estimation of the quality of the reconstructions were already promising, error maps were produced according to [10]. Finally the ability of the method to identify oultying data was illustrated, which will allow to orient further improvements of the reconstruction methodology on these products.

7 ACKNOWLEGMENTS

This work was realised in 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".
The MERIS products were supplied by the European Space Agency under Envisat AOID698.
NASA/Goddard Space Flight Centre is thanked for MODIS data.

8 REFERENCES

1. Beckers, J.-M. and Rixen, M. (2003). EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. Journal of Atmospheric and Oceanic Technology, 20:18391856.

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

3. Lacroix, G., Park, Y., Sirjacobs, D., Ruddick, K., Beckers, J.-M., Lancelot, C. (2008). Interannual variability of the spring bloom timing in the Southern North Sea investigated by MIRO&CO-3D and remote sensing. Advances in Marine Ecosystem Modelling Research Symposium (AMEMRII 2008).

4. Alvera-Azcárate, A., Barth, A., Beckers, J.-M. and Weisberg, R. H. (2007). Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields, *J. Geophys. Res.*, *112*, C03008, doi:10.1029/2006JC003660.

5. Müller, D. (2007). Estimation of algae concentration in cloud covered scenes using geostatistical methods. Proceedings of ENVISAT symposium held in Montreux, ESA SP-636, 2007.

6. IOCCG. (2007). Ocean-colour data merging. Gregg, W. (ed.). International Ocean-Colour Coordinating Group, report number 6, IOCCG, Dartmouth, Canada.

7. Casey, B., Arnone, R. and Flynn, P. (2007). Simple and Efficient Technique for Spatial/Temporal Composite Imagery. Published in the Proceedings of SPIE, Conference on Coastal Ocean Remote Sensing, v6680, held in San Diego, CA on 26-30 Aug 2007.

8. MUMM-RBINS. BELCOLOUR-1 database. Online at http://www.mumm.ac.be/BELCOLOUR/EN/Products/index.php (as of 25 september 2008).

9. Borges, A., Ruddick, K., Schiettecatte, L.-S. and Delille, B. (2008). Net ecosystem production and carbon dioxide fluxes in the Scheldt estuarine plume, BMC Ecology, 8:15, doi: 10:1186/1472-6785-8-15.

10. Beckers, J.-M., Barth, A. and 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.