CHAPTER 9

An Operational Interpolated Ocean Colour Product in the Mediterranean Sea

Gianluca Volpe[1], Bruno Buongiorno Nardelli[1,2], Simone Colella[1], Andrea Pisano[1], and Rosalia Santoleri[1]

[1]*Istituto di Scienze dell'Atmosfera e del Clima, CNR, Roma, Italy,* [2]*Istituto per l'Ambiente Marino Costiero, CNR, Napoli, Italy*

A novel technique to interpolate satellite ocean colour data has been developed and calibrated in the framework of the European MyOcean2 project and successively implemented within the Copernicus Marine Environment Monitoring Service (CMEMS) specifically for the Mediterranean Sea products. The methodology is based on the Data Interpolating Empirical Orthogonal Functions technique, which interpolates data voids from Empirical Orthogonal Function (EOF) modes iteratively estimated as characteristic spatial patterns. Here, this method is extended to take into account the temporal correlation between the observations. A higher-dimensional approach is followed by using a temporal sequence of daily images to build the state vector and thus the observation matrix used to compute the EOFs. An ad-hoc smoothing procedure is also applied to resulting 2-dimensional fields to filter out spurious signals and provide consistent spatial reconstructions. Several tests are performed on a dataset at 4 km resolution to calibrate the technique and to assess, among other issues, the most convenient number of images to be included in the state vector. The final CMEMS product at 1 km resolution is then validated with the independent chlorophyll data collected during dedicated oceanographic surveys between 1997 and 2015 across the entire Mediterranean basin.

Introduction

Numerous marine environmental and scientific applications require complete data time series of bio-geophysical parameters measured at fixed locations with a high spatial coverage, such as those provided by satellite optical remote sensing (e.g., Ferreira et al., 2011). They include, for example, the assimilation of observations in numerical biogeochemical models, the operational detection of harmful algae blooms, as well as the monitoring and assessment of marine ecosystem status by national and international institutions in order to comply with international rules/legal acts (e.g., for implementation of the European Marine Strategy Framework Directive) and/or to investigate the ecosystem functioning and its response to human and natural pressures (e.g., Volpe et al., 2012). However, cloud cover, low repetitiveness of satellite passes, and the relative narrow satellite swaths all prevent a full exploitation of optical satellite measurements.

Volpe, G., et al., 2018: An Operational interpolated ocean colour product in the Mediterranean Sea. In "*New Frontiers in Operational Oceanography*", E. Chassignet, A. Pascual, J. Tintoré, and J. Verron, Eds., GODAE OceanView, 227-244, doi:10.17125/gov2018.ch09.

As a consequence, optical sensors can only provide a partial view of sea surface, and interpolation techniques are needed to overcome the sparseness and uneven temporal coverage of satellite data.

The most common approach to operationally interpolate sea surface temperature (SST) or sea surface height (SSH) data is the optimal interpolation (OI) or analogous approaches (e.g., Martin et al., 2012). These are implemented in several processing centres for data production at both global and regional scales and within real-time operational chains and for offline processing (Le Traon et al., 1998; Reynolds et al., 2007; Buongiorno Nardelli et al., 2010, 2013; Donlon et al., 2012; Roberts-Jones et al., 2012). However, different approaches have been proposed to fill in the gaps present in ocean colour imagery (e.g., Chlorophyll-a, Total Suspended Matter, etc.), either based on kriging/optimal interpolation (e.g., Saulquin et al., 2011) or iterative methods (Alvera-Azcárate et al., 2009, 2015; Sirjacobs et al., 2011). In particular, the Data Interpolating Empirical Orthogonal Functions (DINEOF) method has been proposed and tested for the interpolation of optical satellite data on regular high-resolution spatial grids, even at very high temporal sampling (Beckers and Rixen, 2003; Miles and He, 2010; Sirjacobs et al., 2011; Alvera-Azcárate et al., 2015; Liu and Wang, 2016). This method allows the extraction of the dominant spatial patterns observed in a data time series through an iterative approach, while simultaneously filling in the missing data. DINEOF presents some interesting advantages as compared to more classical approaches (such as optimal interpolation), especially when working with ocean colour data. In fact, different scales of variability and background concentrations are generally associated with the coastal and open sea domains, along with their respective seasonal cycles. The underlying hypothesis of the standard OI techniques for the estimation of the field covariance is that the field is stationary and generally characterized by isotropy (Bretherton et al., 1976). This assumption may potentially lead to artefact propagations of coastal signals offshore, in the presence of extended cloud cover or when analysing different seasons. Adapting the OI covariance parameter estimation to bypass these assumptions would then require unpractical computational steps. These hypotheses are automatically relaxed when using DINEOF, as this technique directly identifies the dominant patterns and the main sources of variability through the calculation of the field principal components. In its original formulation, DINEOF is based on a purely spatial state vector (Beckers and Rixen, 2003). Consequently, the reconstruction of long time series by projecting sparse observations on dominant spatial modes may eventually lead to temporal discontinuities when large fractions of the original images are empty. To address this issue, a filtering of the covariance matrix has been suggested allowing for the reduction of spurious signals and to obtain more realistic reconstructions (Alvera-Azcárate et al., 2009). Here we propose a different approach, namely to augment the DINEOF technique by using a time-lagged extended Empirical Orthogonal Function (EOF) analysis (e.g., Weare and Nasstrom, 1982) instead of standard EOF in the iterative processing. This technique has been developed in the framework of the European project MyOcean2 and successively implemented within the Copernicus Marine Environment Monitoring Service (CMEMS) to provide both Near Real-Time/Delayed Time (NRT/DT) and reprocessed datasets (REP) over the Mediterranean Sea.

Operational applications often require NRT data, but the accuracy of NRT ocean colour data suffers the lack of up-to-date ancillary information, such as meteorological and ozone data, which are crucial for the atmospheric correction and are generally available only after a few days from the satellite overpass. Despite the lower quality of NRT data, they still represent the only available observations of the sea state and, therefore, they are routinely produced. Once the ancillary data are available, NRT data are reprocessed and the DT data are produced with a more reliable scientific quality. Several experiments have thus been carried out to define the optimal configuration for both products. These have been performed on a test dataset, varying the length of the data time series used to build the state vector, the lag considered to interpolate the single daily image and the number of EOF modes used to reconstruct the field.

A complete description of the DINEOF algorithm calibration, as well as of the validation of the final product with independent measurements, is provided in the next section (Data and Methods). As a matter of comparison, the next section includes an overview of the OI procedure currently used to produce daily gap-free (Level-4) SST data over the Mediterranean Sea and available through CMEMS. In the second subsection, the method operationally used to interpolate ocean colour data in the context of CMEMS is presented, while the different tests are shown and discussed in the final subsection.

Data and Methods

This section provides the details of the methodology for field interpolation using both the OI and the DINEOF approaches, with a special focus on the procedures followed to obtain both DINEOF input and output. The pre-processing of the Level-3 original data is a crucial part of the method and is essential to building the data matrix input to the interpolation procedure. Figure 9.1 summarizes the entire interpolation scheme.

Optimal Interpolation

The OI procedure is a mature and consolidated technique to reconstruct gap-free, 2-dimensional fields for oceanographic variables, such as SST, SSH, SSS (sea surface salinity), and ocean colour, as demonstrated by the growing body of literature on the topics. This section deals with the description of the regional and fine-tuned OI scheme, which is operationally used in the context of CMEMS to produce daily SST data at 1 km resolution over the Mediterranean Sea. The same OI parameterization and algorithm used for SST data interpolation are used here for the ocean colour chlorophyll images under the assumption that they (ocean colour and SST) show the same scales of spatial and temporal variability (Volpe et al., 2012). In particular, we use a space-time OI procedure to fill in data voids.

Within OI, the single interpolated pixel is obtained as a linear combination of the observations (e.g., chlorophyll anomalies with respect to the first guess) directly weighted with their correlation (or equivalently their covariance) to the interpolation pixel and inversely with their cross-correlation and error. The estimation of the covariance of all observations is computationally very demanding and operationally unfeasible. To overcome these limitations, the covariance is generally assumed to be well represented by a characteristic functional form. Furthermore, only a subset of the available observations concurs to the estimation of the interpolated value, depending on the spatial and temporal *influential radius* to the interpolation point, i.e., the maximum distance (in both space and time) over which the observations are deemed useful to the interpolation. To this aim, we use an *influential spatial radius* of 20 km and a *temporal window* of 10 days centered on the day that has to be interpolated (as defined in Buongiorno Nardelli et al., 2013).

Following Marullo et al. (2007) and Buongiorno Nardelli et al. (2013), the covariance, C, is directly estimated through the functional form:

$$C(\Delta r, \Delta t) = e^{-\frac{\Delta t}{\tau}} e^{-\frac{\Delta r}{\lambda}},$$

with Δr being the relative spatial distance between observation and interpolation pixels and Δt their time difference; τ and λ are the temporal and spatial de-correlation lengths and are set as three days and 5 km, respectively. These space (5 km) and time (three days) decorrelation lengths are, in turn, defined by the field specific temporal and spatial scales of variability (see Figure 9.4a for the temporal scale and Buongiorno Nardelli et al. (2013) for the spatial scale). When no direct observations are available to the interpolation, for example due to prolonged cloud cover, the first guess is used as interpolation value to fill in the data gaps. Here, as first guess we use the daily climatology at the same spatial resolution as the ocean colour data, as derived from the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS, McClain et al., 2004).

DINEOF Interpolation

Processing modes

Here, the two configurations (NRT and DT) for interpolating the daily chlorophyll data are described. For filling in missing data, both schemes require a data time series to get geophysical information from. Intuitively and from the biogeochemical point of view, the optimum would be to center the data to be interpolated within the time series, e.g., half of the data time series in the future and half in the past with respect to the day that has to be interpolated. This approach should ensure all the data time series to be relevant for the reconstruction of the field; this configuration is hereafter referred to as OPT and is used within CMEMS for the REP processing mode. In the context of operational oceanography, the main goal is to minimize the time required to produce the gap-free daily field while keeping its scientific robustness and reliability. Time minimization means shifting the data that has to be interpolated towards the future end of the time series, in order to provide users (e.g., data assimilation into ecosystem models) with a reasonably recent field. The main difference between NRT and DT is that NRT requires a data time series, with the most recent day being the current day that needs to be interpolated. While in the DT configuration, the same input data is reprocessed after a defined time lag (TL) so that it not only benefits from the information contained in these last days but also from the better quality of the data processed with updated ancillary information. Another important difference between the two approaches is the way the input data matrix is built and given to the interpolation procedure. Moreover, the structure of the input data matrix becomes crucial to accounting for both the spatial and the temporal variability of the data time series and to include them into the interpolated field. In the input data matrix, the rows are used to identify the spatial variability, and they are used by the interpolation procedure to build the EOF spatial patterns.

DINEOF input data

Two satellite datasets are involved in this work, the testing and operational datasets at 4 km and 1 km spatial resolution, respectively. Since computational time depends on the dimension of the data, the 4 km-resolution data are used for testing purposes. The full resolution dataset is then used to implement the final configuration in order to provide users with appropriate statistics associated with the operational product. The 4 km dataset is made from the SeaWiFS, while the 1 km-resolution data are derived from the CMEMS operational processing chain. Each of the two datasets (testing and operational) has its own reference climatology at 4 and 1 km spatial resolution, respectively.

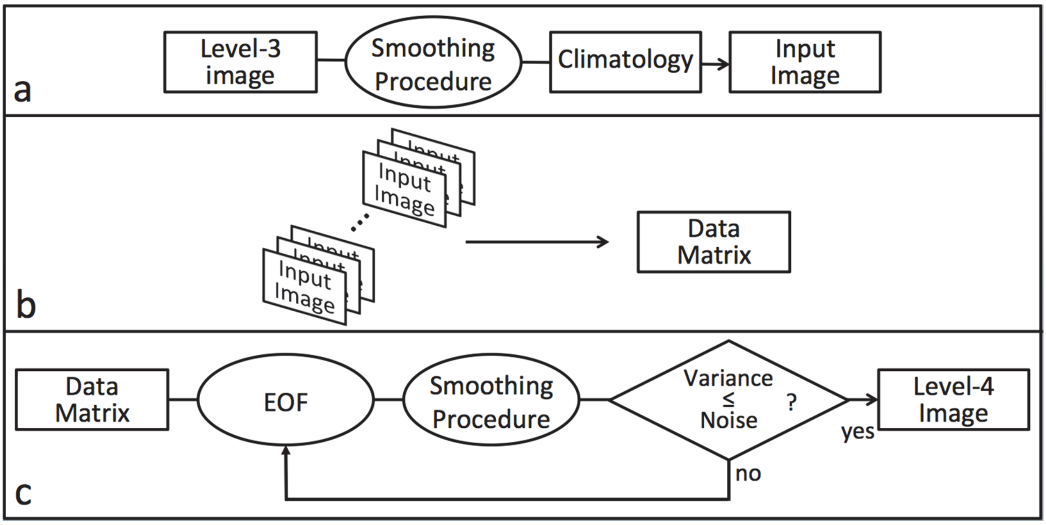


Figure 9.1. Interpolation scheme. Panel a shows that the original Level-3 image has to be merged with the climatology through a smoothing procedure to obtain the input image, which is then used to build the data matrix (panel b), input to the interpolation procedure. The data matrix enters the EOF calculations, and spatial modes are merged with the original data matrix via the smoothing procedure until the explained variance does not exceed that of the noise.

The testing dataset derives from the SeaWiFS Level-0 data that have been processed up to Level-3 with the SeaWiFS Data Analysis System (SeaDAS) software package version 7.0 available from NASA website (http://seadas.gsfc.nasa.gov). The Mediterranean- and sensor-specific Ocean Colour 4-band algorithm (MedOC4, Volpe et al., 2007) for chlorophyll retrieval in Case-1 waters is then applied to the resulting remote sensing reflectance. To ensure the most possible reliability to data, apart from the shallow water and turbid water flags, all masking criteria provided by the SeaDAS software package are applied (McClain et al., 1995). Single chlorophyll swath maps are remapped at a nominal spatial resolution of 4 km on an equirectangular grid covering the Mediterranean domain (30°N–46°N; 6°W–36.5°E).

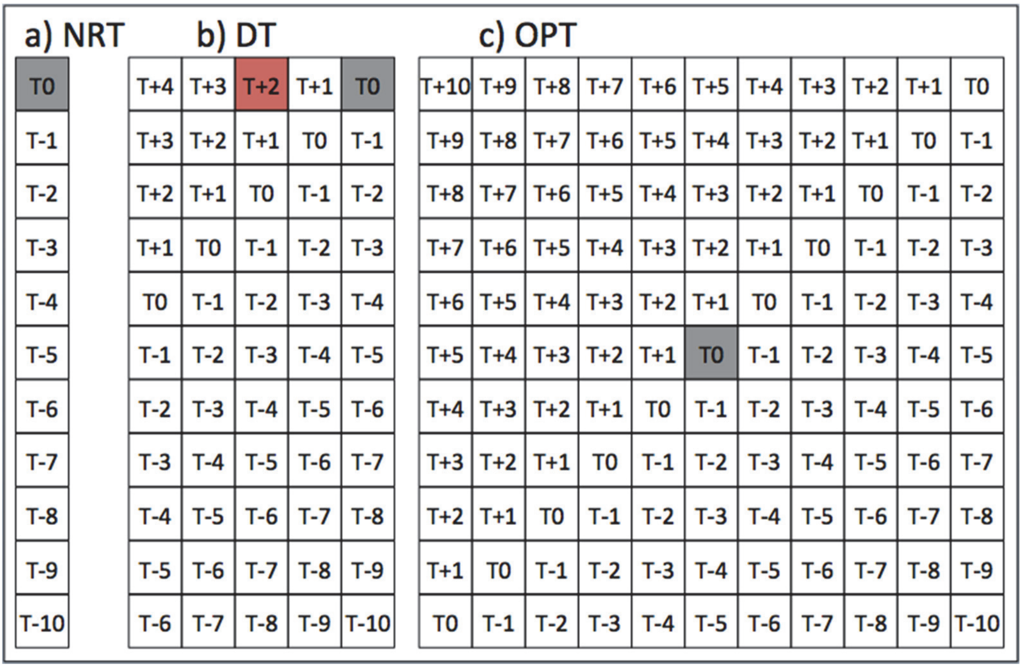
The operational dataset is the multi-sensor data derived from CMEMS over the Mediterranean Sea that merges all available sensors at any given time (Volpe et al., 2017;

http://marine.copernicus.eu/documents/QUID/CMEMS-OC-QUID-009-038to045-071-073-78-079-095-096.pdf). With the merged product, the number of daily valid observations increases by roughly 20% with respect to their single-sensor contributors, without the introduction of any significant source of uncertainty (Volpe et al., 2017).

Figure 9.1a shows the schematics of the pre-processing that any satellite images undergo before being ingested into the interpolation processor. Daily images are transformed into their base-10 logarithm to account for chlorophyll lognormal distribution. To avoid spurious and noisy signals in the DINEOF output due to the low quality of the input images, a filtering procedure is routinely applied to daily input images. The filtering procedure is made of two parts. It first checks for the existence of isolated pixels and, if they are missing, set them to a predefined missing value. Afterwards and with the assumption that ocean colour data do not vary much on the pixel distance level, the procedure checks for the existence of isolated missing pixels to fill them using the median value of its surrounding good pixels. As EOF requires complete input data time series, missing values within this filtered daily image are set to respective climatological values via a procedure that smoothens out spurious spatial gradients. These images constitute the single building blocks of the input data matrix.

The input data matrix

As mentioned, regardless of the processing mode and hence of the input data structure, the interpolation requires a data time series to get geophysical information from. This paragraph explains the differences in the structure of the input data matrix among the three processing modes that are operationally used in the CMEMS processing chain: NRT, DT, and OPT. Under the NRT configuration, the input data matrix is built such that each row, the state vector, corresponds to the single daily sea domain. In this way, the columns represent the single sea pixel time series; the NRT represents the standard DINEOF configuration, which only accounts separately for space and time covariance (meaning that dominant modes are identified exclusively either as spatial patterns or as temporal patterns). Conversely, extended EOFs used in DT and OPT configurations consist of a temporal sequence of spatial patterns and therefore effectively account for the full space-time covariance. Thus, the DT configuration differs from the NRT in such a way that each row, rather than being one single daily sea domain, is made up of multiple subsequent daily sea domains (Figure 9.2). This configuration enables the temporal variability to be taken into account when computing the EOF modes. Starting from the top row, the first image on the right is the image that has to be interpolated and refers to time T0; all other images in this row are more recent than that, referring to times T+1, T+2, up to T+TL. When seeking data, the operational procedure first looks for the correct L3 product. If, for any reason, this product is missing the climatology is taken.



a) NRT

T0
T-1
T-2
T-3
T-4
T-5
T-6
T-7
T-8
T-9
T-10

b) DT

T+4	T+3	T+2	T+1	T0
T+3	T+2	T+1	T0	T-1
T+2	T+1	T0	T-1	T-2
T+1	T0	T-1	T-2	T-3
T0	T-1	T-2	T-3	T-4
T-1	T-2	T-3	T-4	T-5
T-2	T-3	T-4	T-5	T-6
T-3	T-4	T-5	T-6	T-7
T-4	T-5	T-6	T-7	T-8
T-5	T-6	T-7	T-8	T-9
T-6	T-7	T-8	T-9	T-10

c) OPT

T+10	T+9	T+8	T+7	T+6	T+5	T+4	T+3	T+2	T+1	T0
T+9	T+8	T+7	T+6	T+5	T+4	T+3	T+2	T+1	T0	T-1
T+8	T+7	T+6	T+5	T+4	T+3	T+2	T+1	T0	T-1	T-2
T+7	T+6	T+5	T+4	T+3	T+2	T+1	T0	T-1	T-2	T-3
T+6	T+5	T+4	T+3	T+2	T+1	T0	T-1	T-2	T-3	T-4
T+5	T+4	T+3	T+2	T+1	T0	T-1	T-2	T-3	T-4	T-5
T+4	T+3	T+2	T+1	T0	T-1	T-2	T-3	T-4	T-5	T-6
T+3	T+2	T+1	T0	T-1	T-2	T-3	T-4	T-5	T-6	T-7
T+2	T+1	T0	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8
T+1	T0	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9
T0	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9	T-10

Figure 9.2. Graphical representation of the input data matrix for a) NRT, b) DT, and c) OPT processing modes. Single squares refer to the single daily sea pixel domain. The image that has to be interpolated is referred to as T0 and is highlighted in grey for each configuration. All squares marked with T-1 refer to data images of one day earlier than the one that has to be interpolated. It is possible to see that a single data file is used more than once with the purpose of accounting for the data temporal variability. In all three processing modes, each row represents the state vector of the input data matrix.

One aspect that must be taken under consideration is the position of the image that has to be interpolated within the input data matrix. Here, two possibilities are examined: one in which the image that has to be interpolated is centered in the state vector (red square in Figure 9.2b, and referred to as DT$_{\mathrm{C}}$), and the other as shown by the grey square of Figure 9.2b. The latter constitutes the operational way of interpolating the Level-3 fields in DT mode and is referred to as DT$_{\mathrm{L}}$. However, it must be noted that since the method is based on a statistical approach, the most appropriate configuration is linked to the availability of observations within the input data matrix, and, therefore, cannot be determined a priori. The operational configuration assumes the data voids to be more or less equally distributed within the input data matrix. A future version of this approach should dynamically individuate the best configuration aimed at minimizing the distance between T0 and the most clear sky days to allow the contribution of the observations to be more relevant than that of the climatology.

The procedure applied to build the input data matrix is exactly the same, whether the input data are from SeaWiFS (4 km) or from CMEMS (1 km). This means that all results obtained from the testing phase, and hence about the fine-tuning of the methodology, can be realistically applied to the operational context.

Climatology

The general scope of earth observation is to know the value of a parameter (e.g., the chlorophyll concentration, the diffuse attenuation coefficient of light) at any time and location. However, when observing ocean colour, clouds may prevent the single pixel to be seen for long time periods, especially during winter when the interpolation becomes even more important. In this context, the only source of information relevant for the field interpolation is the climatology, which turns out to be more reliable than the data referring to periods too far away in time from the one that has to be interpolated. It is known that the decorrelation timescale in ocean colour data in the Mediterranean Sea is around a few days. Thus, it is generally safe to assume that the most plausible expectation of such a value will be the average (or the median value) of that parameter from past observations, i.e., the climatology field, with a certain degree of variability (e.g., within a few standard deviations from the average). The climatology is obtained from the 13 years of SeaWiFS data using the MedOC4 regional algorithm for chlorophyll (Volpe et al., 2007). As mentioned, this daily field has the same cylindrical projection and spatial resolution, 1 km and 4 km, as the testing and the operational fields, respectively. To reduce the impact of the short-scale variability, these climatology maps are created using all data falling into a moving temporal window of ± 5 days. One of the main purposes of a climatology field is to serve as a reference, and therefore it is expected to be as reliable as possible, thus avoiding biases caused by single incorrect pixel values. To overcome these possible biases, a filtering procedure is applied to the entire SeaWiFS time series by removing all isolated pixels and by filling in all isolated missing pixels using the near-neighbourhood approach. The resulting daily climatology time series includes the pixel-scale standard deviation, the average, the median, the modal, the minimum, and the maximum values.

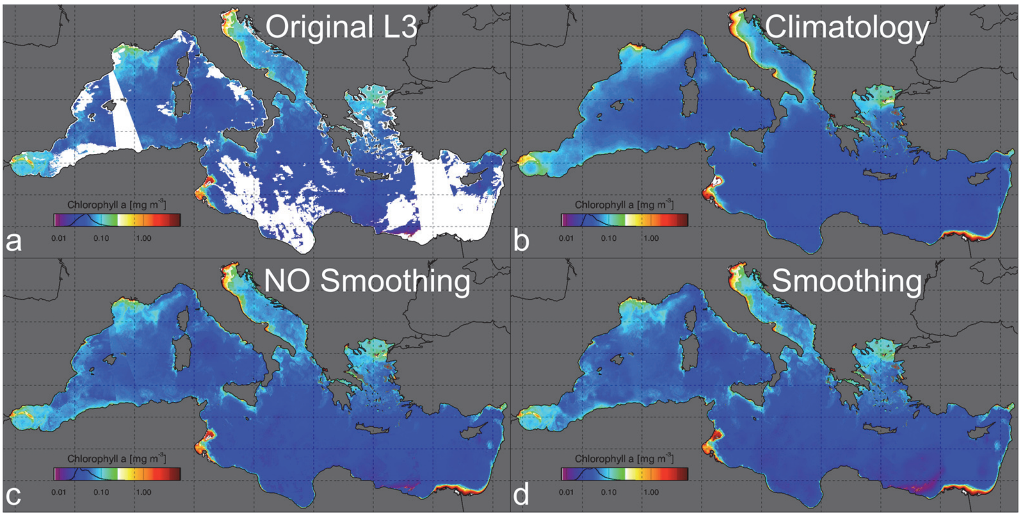


Figure 9.3. Example of the impact of the smoothing procedure used to merge the original Level-3 data (panel a) downloaded from the CMEMS website and referring to the merged product between MODIS-Aqua and NPP-VIIRS collected on the 28 July, 2017, and relevant daily climatology (panel b). The output image when the two fields are merged without (panel c) any smoothing procedure and by applying SP (panel d). The colour palettes contain the distribution histogram of each image. In panel a, white areas refer to missing data either because of the clouds or because out of the orbits of the satellites.

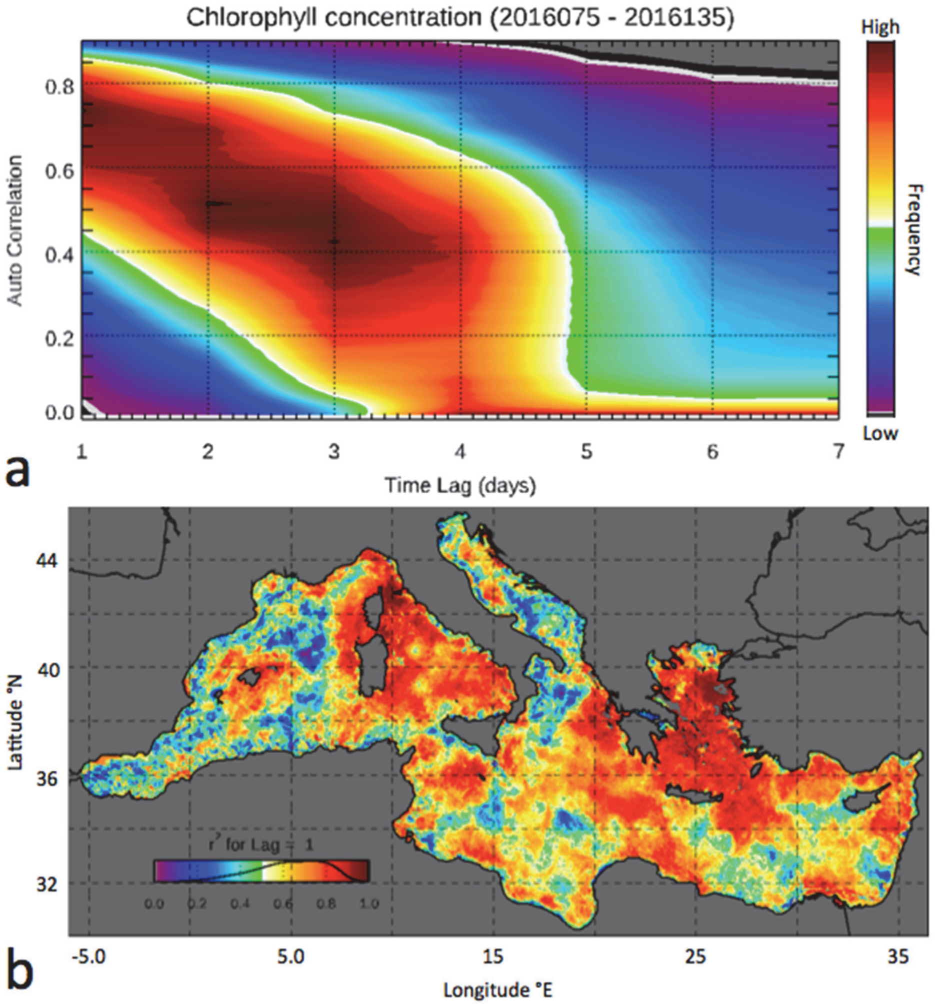


Figure 9.4. Panel a: Frequency distribution of the chlorophyll autocorrelation (r2) as a function of time lag, computed with two months of L4 data. Panel b: autocorrelation map computed with the same two-month data with a time lag of one day.

Smoothing procedure

Smoothing procedure (SP) allows for minimizing the occurrence of spurious gradients due to the merging of two fields of different origin (e.g., observations and climatology) that might contain information at different temporal and spatial scales. SP is applied anytime two fields need to be merged together: it is applied when L3 and climatology are merged together, in correspondence with the first EOF run or for the subsequent runs when the holes in the original L3 daily maps are filled in with the EOF modes. The procedure smooths out the differences computed in correspondence with common observations, that is, where both observations and climatology (or the reconstructed field) co-occur. This procedure allows for filling gaps in while keeping the original observations and without introducing any spurious gradients. Figure 9.3 shows an example of a typical two-dimensional field with and without the application of the smoothing procedure. The impact of using SP is particularly evident south of the Balearic Islands, where the original L3 and the reference climatology differ significantly giving origin to unrealistic spatial gradients that are efficiently erased by the application of the smoothing function. Moreover, the distribution histogram of the SP output (Figure 9.3d, inside the colour palette) is much closer in shape than the original L3, whereas the one in Figure 9.3c (without the application of SP) is very similar to the one of the climatology (Figure 9.3b). This highlights the importance of using a procedure that does not significantly alter the original data distribution.

DINEOF interpolation

The reconstruction of missing data is performed using the DINEOF method first developed by Beckers and Rixen (2003) and later used by Volpe et al. (2012). Volpe et al. (2012) provided some insight, in the ocean colour context, into the advantages of using the DINEOF approach to fill in missing data with respect to optimal interpolation. All three versions of this product (NRT, DT, and OPT) are here tested. In all configurations, the technique works as follows: starting from the day that has to be interpolated (grey square in Figure 9.2), the time series of the previous 10-day data is used to build a data matrix (Figure 9.2). Ten days are deemed to be a good compromise between the need for observations that tend to include as many images as possible and the chlorophyll temporal decorrelation scale in the Mediterranean Sea, to reduce the contribution of uncorrelated observations to the reconstruction of the missing data.

In this respect, Figure 9.4a shows an example of the autocorrelation temporal scales of variability at basin-scale computed with a two-month data time series of interpolated fields: correlation sharply decreases after a few days, reaching the null value in five days. Similarly, Figure 9.4b shows that there is a considerable spatial patchiness due to the short temporal data time series used to compute the statistics, thus it is not entirely representative of all of the scales that can be encountered in the basin.

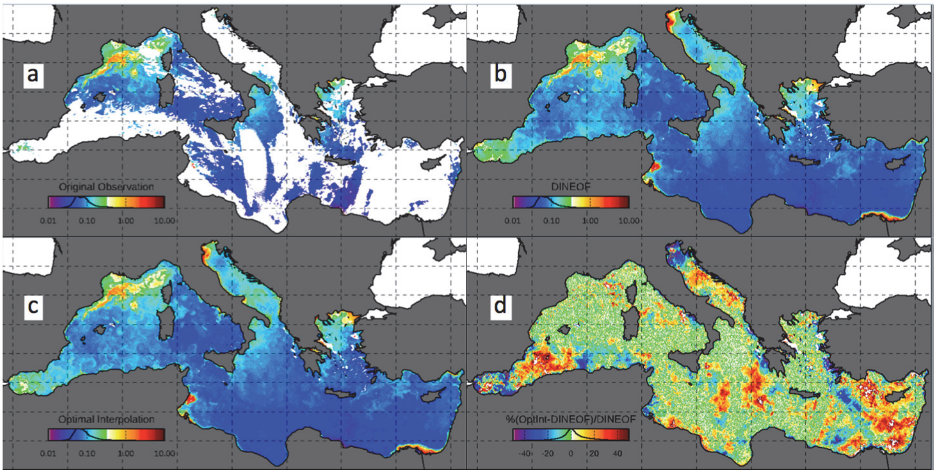


Figure 9.5. Panel a shows the daily merged product from MODIS-AQUA and NPP-VIIRS on 1 April, 2012. Panel b shows the Level-4, in DT mode. The entire data time series of both Level-3 and Level-4 is available on the CMEMS website (http://www.marine.copernicus.eu) and at http://gos.artov.isa.cnr.it. Panel c shows the same interpolated chlorophyll image using the OI procedure, with the settings shown in the second section. Panel d shows the difference (%) between the two interpolated fields.

All missing data are replaced with respective daily climatological sea pixels, and a mask matrix, namely “holes”, is built with zeros in correspondence with effective observations and ones where the climatology is used. This data matrix constitutes the input to the iterative EOF procedure, which uses the singular-value decomposition approach (SVD from the LAPACK library under the IDL environment and is equivalent to the EOF). After each iteration, the input data matrix to the next SVD iteration is built with original observations for holes=0, and with the field reconstructed from the SVD output of the previous iteration for holes=1. The reconstruction is carried out using the number of modes corresponding to the iteration number. Thus, after the first iteration the climatology is replaced with the first EOF mode, which is then replaced with the field reconstructed with the first two modes after the second iteration, and so on. To determine the number of modes that can be used effectively to reconstruct the final field, the cumulative variance is computed at each iteration and compared with the one determined by an independent EOF run performed using a data matrix of the same dimensions as the input data matrix and filled with random numbers. The variance explained by the first mode of this EOF identifies the noise. The interpolation procedure stops when the variance explained by the current iterative mode equals that of noise. It is important to stress that the number of modes used to reconstruct the final field can vary from one day to another or depending on the configuration (NRT, DT, OPT), because of the intrinsic variability of the observations used to build the input data matrix.

Moreover, as it will detailed further in a section below, the procedure smooths out differences between original observations and a reconstructed field the same way, as shown in Figure 9.3. This scheme prevents artificial gradients from being created and is particularly effective during periods of scant data availability, (e.g., winter, when the cloudy pixels are much more numerous than the clear sky pixels). Figure 9.5 shows an example of the CMEMS daily merged chlorophyll product and of its interpolation in DT mode. As a means of comparison and to show how robust the interpolation procedures used here are, the same image is provided applying the OI technique. Figure 9.5d shows that less than 15% of the pixels exceed 25% difference.

Results and Discussion

This section discusses how the final configuration described above is selected from among the many available options. As mentioned, the various tests performed are all aimed at addressing a series of different issues concerning both the way the input data matrix is built and how it is processed to achieve the Level-4. All tests are repeated any time a new issue is identified. In general, all tests are meant to improve the quality of the operational product so that when one specific test clearly tackles the issue that it was meant to address, the operational configuration is changed accordingly. The issues that need to be sorted out before the operational product can be regularly delivered to users are: 1) the number of EOF modes used to reconstruct the final field, 2) the number of days used to build the data matrix, 3) the structure of the data within the input matrix to account for the temporal variability, and 4) the position of the image that has to be interpolated within the input matrix.

The first of these issues has to do with determining the most appropriate number of EOF modes to be used in a field reconstruction. In theory, one would correctly expect more EOF modes to correspond to more information in the final reconstructed field. However, the use of higher EOF modes explaining lower variance could correspond to lower signal-to-noise information, which translates into a noisier final output (i.e., salt-and-pepper noise). With this in mind, we test the reconstruction of the final field with a number of EOF modes such that the variance can be explained by the higher mode to be lower than the one explained by the noise, or with a number of modes such that their cumulative variance can be explained to be at least 90%, 95%, or 99% of total variance, and lastly using all available EOF modes. The result (not shown) is that the root mean square error and bias between the original fields and the interpolated fields increases with the number of modes used to reconstruct the final field. Stopping the EOF iterations in correspondence with the variance explained by the noise becomes an operational requisite and it is thus implemented in the CMEMS operational processing chain.

The standard DINEOF approach only considers the spatial variability, and one of the novelties of this work is the fact that it aims at including the temporal variability in the interpolation scheme. As already discussed in a previous section, this is achieved by building the state vector as a sequence of daily images. The issue then becomes the number of daily data to be used for building the data matrix, input to the interpolation procedure. The rationale for such a question is intuitive and has to do with the fact that the EOF procedure needs the observed fields to get information from. The question becomes particularly relevant under persistent cloud cover that prevents the retrieval of the geophysical information. One would increase the length of the time series until enough information is available for the interpolation procedure. However, one constraint is that the temporal de-correlation scale of the field has to be interpolated. In the Mediterranean Sea, chlorophyll autocorrelation sharply decreases within a few days, so that the input data matrix should not contain data newer or older than one or two weeks of the data that has to be interpolated. In winter, when the number of available pixels dramatically decreases due to persistent cloud cover, the two-week limit may result in no available local data. To address this issue, daily observations and climatology are operationally blended together via the smoothing procedure (see next section and Figure 9.3). The only drawback to this approach is the different space and time scales of variability of the two kinds of data that, in correspondence with persistent and spatially-diffused cloud cover, results in images with lower patterns of variability.

Another issue is represented by the position of the data that has to be interpolated within the input data matrix. Figure 9.2 shows the input data matrices for the three operational processing modes; these are the result of the various tests in which the position of data to be interpolated is allowed to change within the state vector, from the most recent (NRT, the upper left square in any of the configurations shown in Figure 9.2) to the one centred in the state vector (DT_C, centred in the upper row in any of the configurations of Figure 9.2) or the one in the last position of the state vector (DT_L, the upper right square in any of the configurations of Figure 9.). These issues and their impact on the quality of the interpolated field are investigated through a matchup analysis against in situ observations (Figure 9.5).

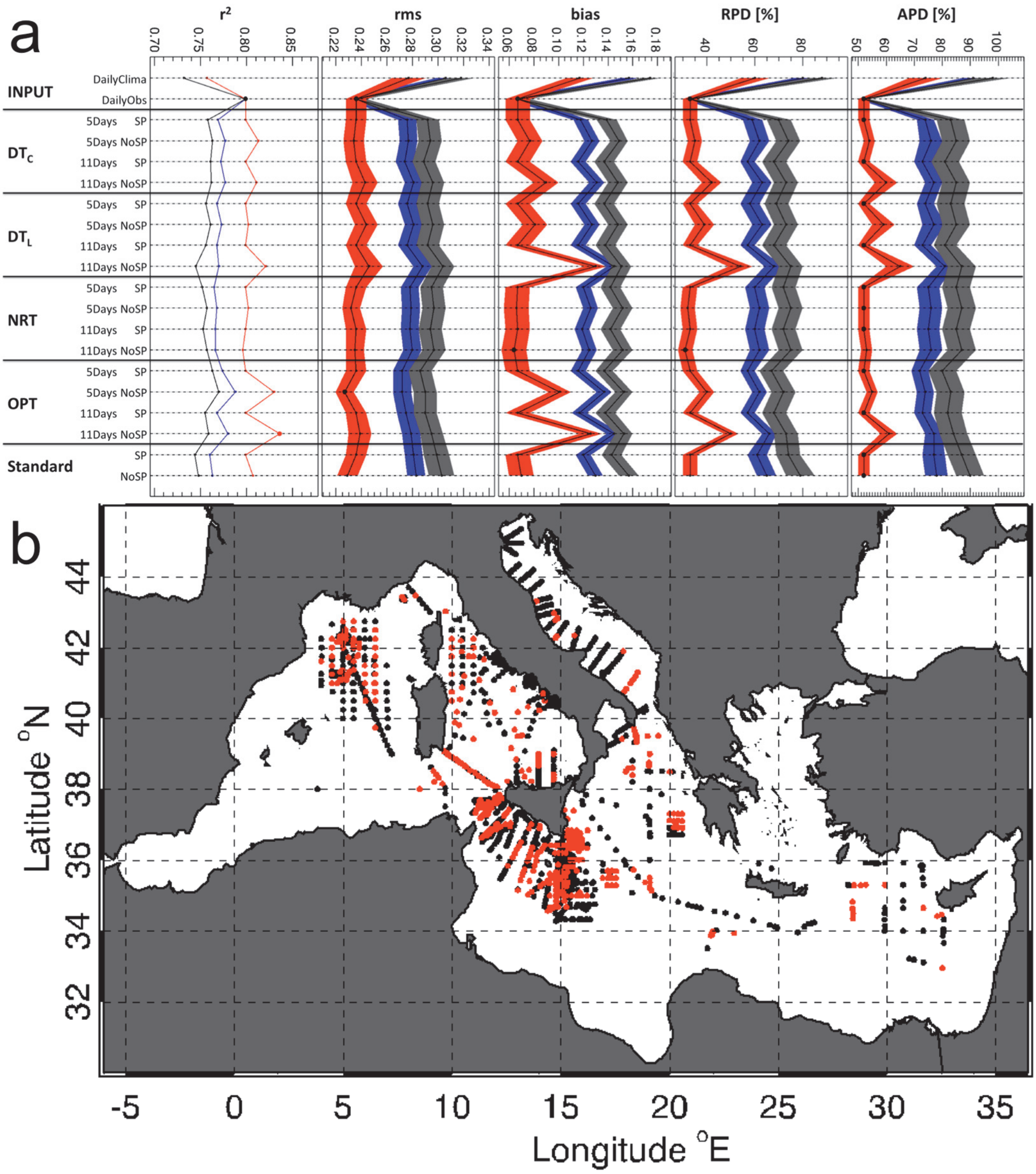


Figure 9.6. Panel a: Shaded areas correspond to ±1STD, computed via bootstrap method, provide the confidence level of the associated statistics. The three coloured shaded areas refer to all matchups (blue), matchups in correspondence with missing original pixels only (grey), and in correspondence with valid original pixels only (red). The five plots are the correlation coefficient, r2, the root mean square error, rms, the bias, and the relative and absolute differences (%) between satellite-derived and in situ chlorophyll concentration. For details on the various tests shown on the y-axis see the main text. The labels 5Days and 11Days refer to the length of the state vector, whereas SP and NoSP refer to the use or non-use of the smoothing procedure. The number of the three matchup datasets is 1139, 504 and 1643, respectively, and their spatial distribution is shown in panel b. Red points in panel b refer to matchups in correspondence with valid original pixels only and black points indicate those with interpolated pixels only.

Figure 9.5b shows the spatial distribution of the matchup points, which consists of two sub-datasets with one (1139 stations) made with the in situ data in correspondence with missing initial observations, and the other (504 stations) with in situ data in correspondence with valid initial observations. This separation allows discerning, above all, the added value of the interpolation procedure. Figure 9.5 shows the results of the matchup analysis in terms of correlation coefficient, root mean square error, bias, and the relative and absolute differences (%) between in situ observations and the interpolation outputs from the various tests. Both the input and output of the interpolation procedure are compared with the in situ observations, and this provides a means for precisely quantifying the impact of the method. Moreover, to address the significance of the results, all calculations are performed using the bootstrapping method in which the statistics are computed 1000 times over half of the validation dataset only, at any time. This method allows the estimation of the statistics variability providing their confidence level.

The first important and non-trivial result is the fact that in all configurations the interpolated fields behave better than the climatology. This is not straightforward, as the climatology enters as first guess the interpolation procedure in roughly 70% of the entire matchup dataset (1139 initial missing pixels against the total number of matchups of 1643). Apart from a few tests, e.g., DT_L 11Days NoSP, OPT 5Days NoSP and OPT 11Days NoSP, all other tests exhibit pretty much the same level of performance when compared with their in-situ counterparts. This means that the interpolation scheme is able to capture the input data matrix variability and to project it into the final output.

Results associated with the application of the smoothing procedure are generally better than when the procedure is not applied (all the NoSP runs). Moreover, the statistics associated with the matchup dataset built only with the valid original pixels (red lines in Figure 9.5) show a similar behavior to those of the missing original pixels (grey lines in Figure 9.5), despite their absolute difference. This highlights the importance of the smoothing procedure, which became a requisite of the operational processing and is currently implemented in the CMEMS chain. The uncertainty contribution of the interpolation procedure is on average less than 30% (Figure 9.5). Despite the above-mentioned theoretical considerations, Fig. 9.6 does not provide definitive evidence that the DT and OPT configurations are unequivocally better than the NRT or standard DINEOF, showing that the use of a multi-day state vector (basically all runs but standard) only slightly improves the statistics. Thus, what surely drives the quality of the output is the amount and quality of the original observations in the input data matrix.

The use of the multi-image state vector, along with the position of the image that has to be interpolated within the input data matrix, are the two elements that mostly contribute to the variability among the three configurations (NRT, DT, OPT). In fact, Figure 9.6 shows, in correspondence with the areas where the original Level-3 presents data voids (panel a), that the impact of the climatology (panel b) on the interpolation outputs decreases from the NRT (panel d) to the DT (panel e) and to the OPT (panel f) configurations, even if its features are clearly visible in the input image (panel c). The three configuration outputs mainly differ in their correspondence with the Gulf of Lions, where the area of the bloom is represented by two distinct patches in the

NRT and DT and by a more uniform and coherent patch in the OPT run. This result is due to the fact that, in the OPT configuration, the space-time variability benefits of as many days in the future as they are in the past. There are several little features where the three outputs differ, such as the one in correspondence with the Rhodes Gyre, east of the Island of Crete, which varies in shape and to a lesser extent in its background intensity. Analogously, the Bonifacio Gyre visible in the input image (Figure 9.6c) and clearly derived from the climatology field is visible in the NRT and DT runs, but not in the OPT run (Figure 9.6f).

The operational product

The previous section showed that the best performing configurations for the three processing modes are those schematically drawn in Figure 9.2, by smoothing differences anytime two fields of different origin need to be merged together, and by stopping the iterations before the variance explained by the current mode reaches the level of the noise. As mentioned, this configuration is operationally implemented into the CMEMS processing chain through which the entire data archive has recently been reprocessed.

Processing Level	r^2	RMS	Bias	RPD	APD	N
Level-3	0.720	0.272±0.009	-0.029±0.009	17±6	56±5	784
Level-4	0.720	0.288±0.006	-0.062±0.006	8±2	53±2	2005
Level-4*	0.718	0.286±0.005	-0.061±0.006	8±2	53±2	1967
Optimal Interpolation	0.724	0.281±0.006	-0.040±0.006	16±4	58±3	1967

Table 9.1. Statistics associated with the matchup exercise performed over the entire Mediterranean Sea in situ dataset collected in several cruises carried out by CNR from 1997 to 2015. First and second rows refer to the entire operational Level-3 and Level-4 datasets. Statistics associated with matchups in correspondence with valid DINEOF and OI data are shown for comparison in the third and fourth rows.

Figure 9.7 and Table 9.1 show that the comparison of both Level-3 and Level-4 data with in situ observations does yield similar agreement. They also confirm the importance of the data interpolation in terms of field coverage (the number of matchups for the Level-4 is nearly triple as compared with the Level-3), without introducing significant sources of uncertainty. This result is quite different than the one achieved within the testing analysis, in which there is a net increase of about 30% uncertainty from Level-3 (red statistics in Figure 9.5a) and Level-4 (blue statistics in Figure 9.5a), while in the operational context the uncertainty difference between the two is of only a few percentages. As a means of comparison, days having corresponding in situ measurements are also processed with the OI method; the result of this matchup exercise, shown in Table 9.1, highlights that the two methods are absolutely equivalent at the scale of the single pixel statistics. In cases of prolonged cloud cover, however, the OI method is more susceptible to propagating features offshore giving rise to unphysical oceanographic structures (results not shown here). On the other hand, we observe that the DINEOF method, at least the way it is implemented here, is largely dependent on the climatology. This results in the two-dimensional output sometimes being unable to reproduce the small-scale processes, which can be better inferred from the OI technique if observations falling within the spatial and temporal influential radius are available.

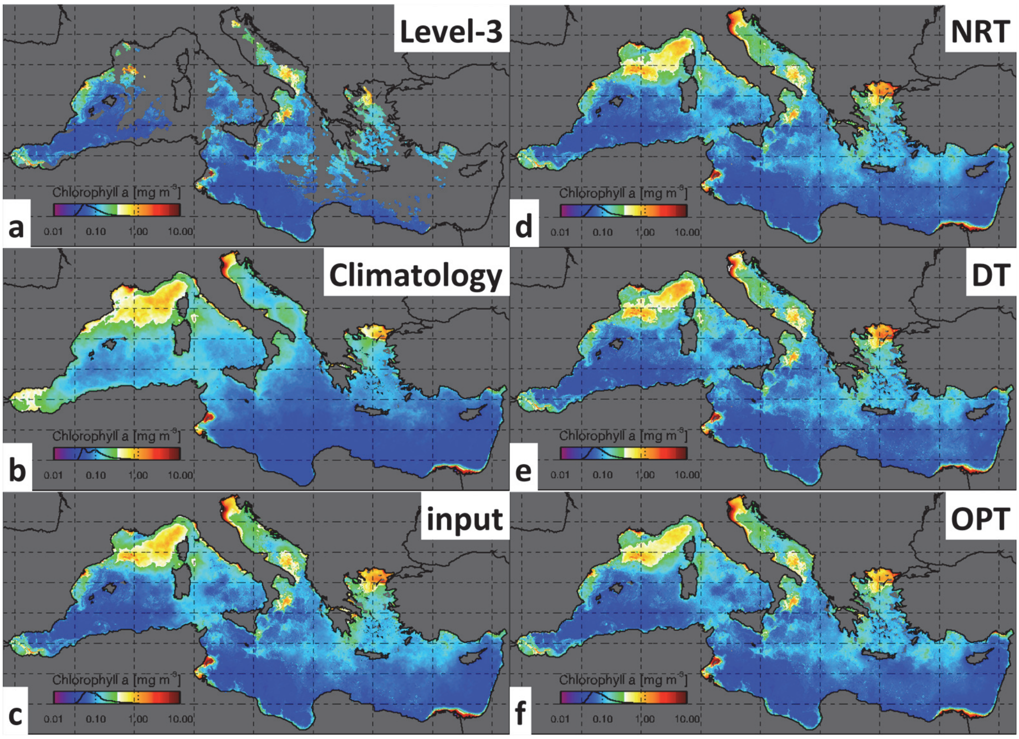


Figure 9.7. Example of the full chain that transforms the original observations (panel a), using the daily climatology (panel b), into the data that constitutes the single image (panel c) for the input data matrix. The outputs were determined using the input data matrix schemes of Figure 9.2a for NRT (panel d), Figure 9.2b for DT (panel e), and Figure 9.2c for OPT (panel e). These images refer to SeaWiFS daily products collected on April 3, 1998.

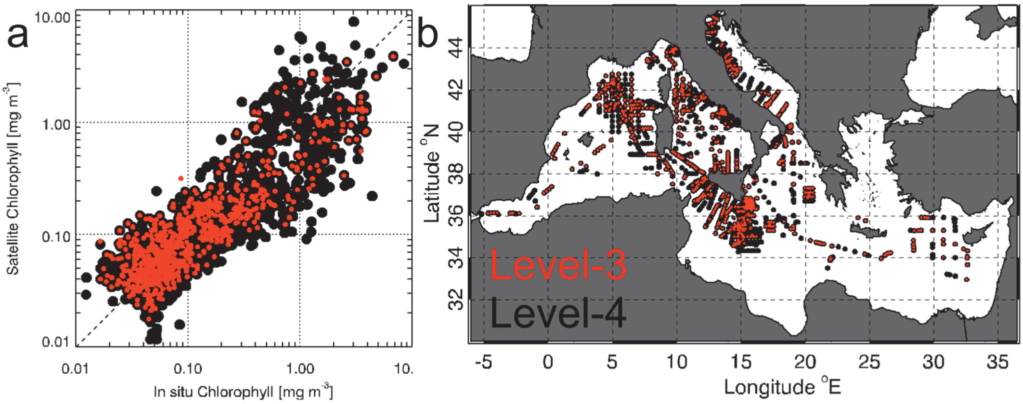


Figure 9.8. The matchup between in situ and satellite derived chlorophyll spans from 1997 to 2015, for both Level-3 (red) and Level-4 (black).

Conclusions

The aim of this work is to describe and fully characterize the steps needed to achieve an operational ocean colour interpolated product over the Mediterranean Sea. This product is operationally provided to the user community through the CMEMS web portal. An important aspect that needs to be pointed out is the fact that the method described here is totally independent from the region of interest and from the product itself; it is general and it could reasonably be applied for the interpolation of other fields over other regions including the global ocean. An example is the L4 chlorophyll field over the Black Sea[2], which is operationally produced with this approach, in the context of CMEMS. One aspect that surely needs to be considered is the de-correlation temporal and spatial scales that should drive the length of the data time series used as input to the interpolation procedure.

The DINEOF method has been demonstrated to perform equally well as, and under given circumstances even better, than the *standard* OI procedure, which is widely used to fill in different oceanographic variables such as SST, SSS, and SSH.

To avoid the use of data time series that are too long, because they might be poorly correlated with the data that has to be interpolated, the input data matrix is built starting from relatively recent original observations with holes filled in with daily climatology, as first guess. This allows for reducing the length of the data time series, thus significantly saving computational resources without any appreciable drawbacks. The impact of the climatology fields is demonstrated to be negligible, as all configurations return statistics better than those obtained with climatology.

The use of climatology is made possible by the *ad hoc* smoothing procedure, which efficiently enables the merging of fields with different temporal and spatial scales without introducing unrealistic features. The overall impact of the smoothing procedure is that it allows the original observations to be kept in the final interpolated field. As a result, the interpolated field behaves exactly the same way as the original in correspondence with valid initial observations. An important point is that, since the original observations are kept as they are, their quality needs to be very carefully checked before entering the interpolation procedure.

The multi-image state vector is efficient when including the data space-time variability in the final interpolated fields and decreasing the weight that climatology has in determining the space-time distribution of the interpolated fields.

Acknowledgements

This work was funded by the European Integrated project MyOcean2 and by the Italian Flagship project, RITMARE. The results of this work are operationally implemented in the CMEMS processing chain. A special thanks to Clifford Watkins, Virginie van Dongen, Dakui Wang, and Natalia Tilinina for their early review of this work. An anonymous reviewer is also thanked for the useful comments that helped improve the work.

[2] http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details &product_id=OCEANCOLOUR_BS_CHL_L4_NRT_OBSERVATIONS_009_045

References

Alvera-Azcárate, A., A. Barth, D. Sirjacobs, and J.-M. Beckers (2009), Enhancing temporal correlations in EOF expansions for the reconstruction of missing data using DINEOF, Ocean Sci. Discuss., 6(2), 1547–1568, doi:10.5194/osd-6-1547-2009.

Alvera-Azcárate, A., Q. Vanhellemont, K. Ruddick, A. Barth, and J. M. Beckers (2015), Analysis of high frequency geostationary ocean colour data using DINEOF, Estuar. Coast. Shelf Sci., 159, 28–36, doi:10.1016/j.ecss.2015.03.026.

Beckers, J., and M. Rixen (2003), EOF Calculations and Data Filling from Incomplete Oceanographic Datasets*, J. Atmos. Ocean., 1–32.

Bretherton, F., R. Davis, and C. Fandry (1976), A technique for objective analysis and design of oceanographic experiments applied to MODE-73, Sea Res. Oceanogr., 23.

Buongiorno Nardelli, B., S. Colella, R. Santoleri, M. Guarracino, and A. Kholod (2010), A re-analysis of Black Sea surface temperature, J. Mar. Syst., 79(1–2), 50–64, doi:10.1016/j.jmarsys.2009.07.001.

Buongiorno Nardelli, B., C. Tronconi, a. Pisano, and R. Santoleri (2013), High and Ultra-High resolution processing of satellite Sea Surface Temperature data over Southern European Seas in the framework of MyOcean project, Remote Sens. Environ., 129, 1–16, doi:10.1016/j.rse.2012.10.012.

Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer (2012), The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, Remote Sens. Environ., 116, 140–158, doi:10.1016/j.rse.2010.10.017.

Ferreira, J. G. et al. (2011), Overview of eutrophication indicators to assess environmental status within the European Marine Strategy Framework Directive, Estuar. Coast. Shelf Sci., 93(2), 117–131, doi:10.1016/j.ecss.2011.03.014.

Liu, X. and M. Wang, (2016), Analysis of ocean diurnal variations from the Korean Geostationary Ocean Color Imager measurements using the DINEOF method, Estuarine Coastal Shelf Sci., 180, 230-241. doi:10.1016/j.ecss.2016.07.006

Martin, M. et al. (2012), Group for High Resolution Sea Surface temperature (GHRSST) analysis fields intercomparisons. Part 1: A GHRSST multi-product ensemble (GMPE), Deep Sea Res. Part II Top. Stud. Oceanogr., 77–80, 21–30, doi:10.1016/j.dsr2.2012.04.013.

McClain, C. R., Feldman, G. C., & Hooker, S. B. (2004). An overview of the SeaWiFS project and strategies for producing a climate research quality global ocean bio-optical time series. Deep-Sea Research Part II: Topical Studies in Oceanography, 51(1–3), 5–42. http://doi.org/10.1016/j.dsr2.2003.11.001

Miles T.N., He R., (2010), Temporal and spatial variability of Chl-a and SST on the South Atlantic Bight: Revisiting with cloud-free reconstructions of MODIS satellite imagery, Continental Shelf Research, 30 (18), pp. 1951-1962.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax (2007), Daily High-Resolution-Blended Analyses for Sea Surface Temperature, J. Clim., 20(22), 5473–5496, doi:10.1175/2007JCLI1824.1.

Roberts-Jones, J., E. K. Fiedler, and M. J. Martin (2012), Daily, Global, High-Resolution SST and Sea Ice Reanalysis for 1985–2007 Using the OSTIA System, J. Clim., 25(18), 6215–6232.

Saulquin, B., F. Gohin, and R. Garrello (2011), Regional objective analysis for merging high-resolution MERIS, MODIS/Aqua, and SeaWiFS chlorophyll-a data from 1998 to 2008 on the european atlantic shelf, IEEE Trans. Geosci. Remote Sens., 49(1 PART 1), 143–154, doi:10.1109/TGRS.2010.2052813.

Sirjacobs, D., A. Alvera-Azcárate, A. Barth, G. Lacroix, Y. Park, B. Nechad, K. Ruddick, and J.-M. Beckers (2011), Cloud filling of ocean colour and sea surface temperature remote sensing products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions methodology, J. Sea Res., 65(1), 114–130, doi:10.1016/j.seares.2010.08.002.

Le Traon, P. Y., F. Nadal, and N. Ducet (1998), An Improved Mapping Method of Multisatellite Altimeter Data, J. Atmos. Ocean. Technol., 15(2), 522–534.

Volpe, G., B. Buongiorno Nardelli, P. Cipollini, R. Santoleri, and I. S. Robinson (2012), Seasonal to interannual phytoplankton response to physical processes in the Mediterranean Sea from satellite observations, Remote Sens. Environ., 117, 223–235, doi:10.1016/j.rse.2011.09.020.

Volpe, G., S. Colella, V. Brando, V. Forneris, F. La Padula, J. Pitarch Portero, A. Di Cicco, M. Bragaglia, F. Artuso, R. Santoleri (2017), The Mediterranean component of the Copernicus Ocean Colour Thematic Assembly Centre: Level 2 to Level 3 processing, Remote Sens. Environ., submitted.

Weare, B. C., and J. S. Nasstrom (1982), Examples of Extended Empirical Orthogonal Function Analyses, Mon. Weather Rev., 110, 481–485, doi:10.1175/1520-0493(1982)110<0481:EOEEOF>2.0.CO;2.