# An approach to fill in missing data from satellite imagery using data-intensive computing and DINEOF

**José Roberto Lomelí-Huerta**[1]**, Juan Pablo Rivera-Caicedo**[2]**, Miguel De-la-Torre**[1]**, Brenda Acevedo-Juárez**[3]**, Jushiro Cepeda-Morales**[4]**, and Himer Avila-George**[1]

[1] **Departamento de Ciencias Computacionales e Ingenierías, Universidad de Guadalajara, Ameca, Jalisco, México**
[2] **CONACYT-UAN, Secretaría de Investigación y Posgrado, Universidad Autónoma de Nayarit, Tepic, Nayarit, México**
[3] **Departamento de Ciencias Naturales y Exactas, Universidad de Guadalajara, Ameca, Jalisco, México**
[4] **Centro Nayarita de Innovación y Transferencia de Tecnología A. C., Universidad Autónoma de Nayarit, Tepic, Nayarit, México**

Corresponding author:
Himer Avila-George[1]

Email address: himer.avila@academicos.udg.mx

## ABSTRACT

This paper proposes an approach to fill in missing data from satellite images using data-intensive computing platforms. The proposed approach merges satellite imagery from diverse sources to reduce the impact of the holes in images that result from acquisition conditions: occlusion, the satellite trajectory, sunlight, among others. The amount of computation effort derived from the use of large high-resolution images is addressed by data-intensive computing techniques that assume an underlying cluster architecture. As a start, satellite data from the region of study are automatically downloaded; then, data from different sensors are corrected and merged to obtain an orthomosaic; finally, the orthomosaic is split into user-defined segments to fill in missing data, and then filled segments are assembled to produce an orthomosaic with a reduced amount of missing data. As a proof of concept, the proposed data-intensive approach was implemented to study the concentration of chlorophyll at the Mexican oceans by merging data from MODIS-TERRA, MODIS-AQUA, VIIRS-SNPP, and VIIRS-JPSS-1 sensors. Results reveal that the proposed approach produces results that are similar to state-of-the-art approaches to estimate chlorophyll concentration but avoid memory overflow with large images. Visual and statistical comparison of the resulting images reveals that the proposed approach provides a more accurate estimation of chlorophyll concentration when compared to the mean of pixels method alone.

## INTRODUCTION

Since the first satellite photographs in the 1940s, followed by missions that include Landsat and Suomi NPP from NASA, and Sentinel from ESA, just to mention three of the most popular, satellite imagery has been improved to the point of becoming daily-use information. Moreover, together with the increase of use, challenges have been emerged, presenting an increasing demand for computational resources and algorithms. Nowadays, images of Earth are commonly used to study the atmosphere, land, and oceans, presenting an increasing use in daily life activities. Applications of satellite imagery range from weather forecasting (Sato et al., 2021), monitoring natural disasters (Said et al., 2019), survey phytoplankton size structure impacts as an ecological indicator for the state of marine ecosystems (Gittings et al., 2019), among many others. Presently, various sensors provide different temporal, spatial, and spectral resolutions to study oceans' evolution. The Moderate-Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) sensors are well known in the community, mainly

because of data's continuous and free availability. Data from these sensors have been available since 1999 and are still in operation (Hu et al., 2010). The MODIS sensors are in orbit aboard the Terra and Aqua satellites (NASA, 2020). On the other hand, the VIIRS sensors are aboard the Suomi National Polar-Orbiting Partnership (SNPP) and the Joint Polar Satellite System (JPSS-1) (Kramer, Herbert J., 2020). Both MODIS and VIIRS sensors are provided with a set of bands commonly employed to study the oceans (Datla et al., 2016). Regardless of the application, processing data from satellites exhibit various challenges that are difficult for the analysis related to physical phenomena (Rodriguez-Ramirez et al., 2019). Moreover, one of the most representative issues emerge from acquisition conditions that produce incomplete data from the whole scene, either caused by occlusion (*e.g.* clouds) or the trajectory of the satellite at the acquisition moment (Zhang et al., 2018).

As reported by the scientific community, some approaches to fill in missing data employ machine learning techniques to merge information from multiple sources within the same set of sensors (Zhang et al., 2018). Other approaches use regression models to repair single spatial satellite images, presenting a tradeoff between accuracy and computational effort (Zhang et al., 2015). Furthermore, one of the most widely employed methods in oceanography to fill in missing data is `DINEOF` (Data INterpolating Empirical Orthogonal Function) (Liu and Wang, 2018; Liu and Wang, 2019). Examples of the use of DINEOF to study chlorophyll are disseminated in literature. For instance, Jayaram et al. (2018) implemented the interpolation functions of the orthogonal data to restore the levels of chlorophyll-*a* (Chl-*a*) at the Arabic sea between 2000 and 2015, using the MODIS sensor. Jayaram et al. (2021) compute the chlorophyll concentration using DINEOF to fill in gaps produced by clouds, using the Ocean Colour Monitor-2 (OCM-2) onboard Oceansat-2 satellite for the period 2016–2019 over the northern Indian Ocean. On the other hand, Alvera-Azcárate et al. (2011) implemented the data restoration with DINEOF using the time series with a single variable (monovariate), and several variables (multivariate approach). More recently, Alvera-Azcárate et al. (2021) reported a Suspended Particulate Matter reconstruction combining Sentinel-2 and Sentinel-3 imagery using DINEOF. The advantage of such a combination allows us to retain both the high spatial resolution of the Sentinel-2 data while increasing the temporal resolution from Sentinel-3 data. DINEOF was also employed by Bouchra et al. (2011) to restore the total suspended matter between Belgium and United Kingdom coasts, using MODIS data acquired between 2003 and 2006. Restored data were compared against the measurements *in-situ* of total suspended matter collected by the Cefas (Centre for Environment Fisheries and Aquatic Sciences); for factor calibration, a linear regression model was employed, considering the highest observed measurements as the reference values. Additionally, during the atmospheric correction, MODIS data pixels were labeled according to the quality of the restoration: those pixels within a $5 \times 5$ window that present inconsistencies over the Cefas time series were labeled as doubedly or low quality. Finally, DINEOF was used to compute missing data, and atypical values were assessed using spatial coherence.

Despite the approach employed to fill in missing data, and the study region, the challenges remain. Furthermore, improvements in computational efficiency and accuracy are still required to produce reliable studies. Indeed, the high computational cost required to analyze multi-temporal and multi-resolution data provided by satellite platforms is far from being solved (Babbar and Rathee, 2019). In particular, DINEOF is based on empirical orthogonal functions (EOF) to reconstruct missing data in a set of geophysical data through the calculus of the dominant modes of variability within satellite data (Beckers and Rixen, 2003). The DINEOF's amount of computation increases with the size of the input images and may be impractical with a large number of high-resolution images. Therefore, DINEOF is usually used to process images with a low spatial resolution of small geographical areas (GHER, 2020).

This research paper proposes a novel approach for automated hole filling in satellite imagery. The first novelty of the proposed approach compared to previous works is that it uses data from different sensors, while previous works used data from the same set of sensors. Another novelty in our proposal is the way the filling of missing data was performed, which is carried out by chaining three different strategies: (1) The first data with which the gaps in the images are filled comes from the fusion of four data sources (MODIS-TERRA, MODIS-AQUA, VIIRS-SNPP, and VIIRS-JPSS-1); (2) The next step consists of estimating the missing pixels close to those obtained in the previous step, for which the nearest neighbor approach using multivariate interpolation is employed; and (3) Empirical orthogonal functions are used to fill in the last missing data. Finally, the proposed approach uses an intensive computing strategy to avoid memory overflow when processing high-resolution images. For proof of concept, the detection of chlorophyll over the exclusive economic zone of Mexico (EEZM) is analyzed.

## A COMPUTER-INTENSIVE APPROACH TO FILL IN MISSING DATA

The proposed data-intensive computing approach to fill in the missing geophysical data from satellite imagery comprises three main conceptual modules: (1) Automatic satellite data download, (2) Satellite data merging, and (3) Filling in missing satellite data using an intensive computer approach. Each module considers the output of the previous one, and their operation is detailed in the sections below. The whole process is depicted in Fig. 1.
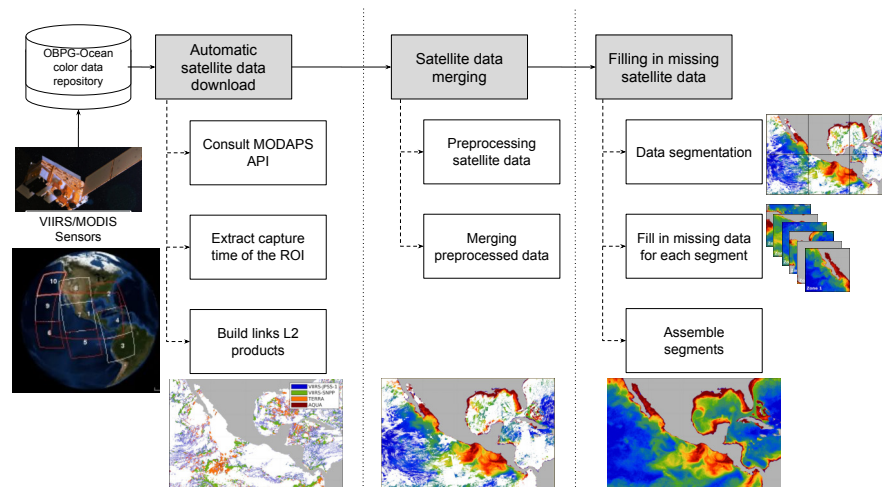
**Figure 1.** Proposed approach for filling in missing satellite data. First, L2 data from the region of interest (ROI) are downloaded. Satellite data are then merged using a 2-step strategy, estimating missing pixels as the average of at least three neighbors in same-day images from different sensors. Finally, missing data is filled in using a data-intensive approach that takes advantage of segmented ROIs and DINEOF

### Automatic satellite data download

The automatic satellite data download module is designed to continuously survey changes in the satellite repository and retrieve the most recent satellite imagery from the region of interest (ROI). Without loss of generality, it is assumed that data are retrieved from the OBPG-Ocean color data repository, but other platforms may be configured with the same behavior. The three steps established in this module are listed below.

1. The first step consists in querying the repository to retrieve the schedule of both sensors (MODIS and VIIRS): the time when they passed over the zone of study.

2. In the second step, the schedule information is processed to extract the precise hour when the satellite acquired the region of interest (ROI).

3. Finally, the links to the levels L2 products are built in the third step, and the download process starts. The resulting L2 products are stored in a user-specific path that is accessed by the other two modules.

As a result, the module for satellite data download retrieves the high-resolution images from the configured sensors, corresponding to the ROI at a determined date.

### Satellite data merging

Every time L2 satellite data corresponding to the ROI are downloaded, a new daily high-resolution image is created by merging data from the selected sensors according to the application. The procedure to create the combined image involves the two steps described in subsections below (preprocessing satellite data and merging preprocessed data).

*Preprocessing satellite data*

The first novelty of the proposed approach is that data from different sensors are used when performing satellite data fusion. Preprocessing data consists of creating the orthomosaics for each daily scene: one for each sensor $(I_1, I_2, \cdots I_L)$. Operations like spatial resampling or scaling are required in some cases to prepare the raw data to assemble a single orthomosaic for each of the $L$ sensors. Subsequently, each orthomosaic is processed to fill missing data during the merging phase; a $m \times m$ sliding window is applied to the $p$ empty pixels in the image that accomplish with the criteria of having at least three neighbors (*i.e.*, three pixels with data). Such a criterion was established to avoid simple information duplicity of close pixels. The new value $p_x$ of an empty pixel is computed using Eq. 1.

$$p_x = \frac{\sum_{i=1}^{n} p_i}{n},$$  (1)

where $p_x$ is the missing data pixel, $p_i$ is one of the $n$ neighbor pixels with data within the $m \times m$ sliding window, considering $n \geq 3$.

*Merging preprocessed data*

In this step, the preprocessed orthomosaics are merged. In order to obtain the combined image at a selected date (day), the orthomosaic with most data related to chlorophyll is first chosen and tagged as *base-image* ($I_b$). Then, it is necessary to define the order of the processing of each orthomosaic. The ordering criteria considers the root-mean-square error (RMSE) between the base-image and each of the remaining orthomosaics ($I_r$), assigning higher priority to the orthomosaics with lower RMSE, see Eq. 2.

$$RMSE(I_r) = \sqrt{\frac{1}{N} \sum_{r=1}^{3} (I_b - I_r)^2}$$  (2)

where $I_b$ is the base-image, $I_r$ corresponds to each of the other images, and $N$ is the number of valid pixels in both images (i.e., $I_b$ and $I_r$).

Once the priority is established, data from the four images are combined considering $I_b$ as the baseline and following the order of priority given by the RMSE: each missing pixel with coordinates $(x, y)$ in $I_b$ is substituted with the pixel from $I_r$ with the highest priority on the same position. If none of the $I_r$ images contains data, it is considered as a missing pixel. Finally, an adjustment is applied to reduce the impact of differences in acquisition conditions from each sensor, such as different acquisition times and the zone dynamics (currents and winds). Such an adjustment between images $I_b$ and $I_r$ was applied using the Inverse Distance Weighting (IDW) to the four nearest pixels in directions $(-x, x, -y, \text{and } y)$. In essence, the resulting high-resolution image $I_M$ produced by merging the sensor-wise orthomosaics $\{I_1, I_2, I_3, I_4\}$ incorporate the information from all sensors, and hence, includes fewer gaps than any of the individual orthomosaics.

## Filling in missing satellite data

As a second contribution, the proposed approach is able to process high-resolution images of large study areas. After preprocessing, the merged orthomosaic $I_M$ still remains with gaps, and DINEOF is employed to compute and fill in the gaps. In order to address this problem with high-resolution images from wide areas of study, the data-intensive approach is divided into the following three steps: (1) *data segmentation*, (2) *fill in missing data for each segment*, and (3) *assemble the segments*. The strategy to fill in missing data is shown in Figure 2, and each step is detailed in sections below.

*Data segmentation*

The merged orthomosaic $I_M$ comprises the whole ROI to be monitored, which may be computationally unmanageable, depending on the area of study and the computer to process DINEOF. Thus, $I_M$ is evenly divided into $J \times K = NS$ smaller manageable size segments $\{I_{i,j}^S\}$:

$$I_M = \begin{bmatrix} I_{1,1}^S & I_{1,2}^S & \cdots & I_{1,K}^S \\ I_{2,1}^S & I_{2,2}^S & \cdots & I_{2,K}^S \\ \vdots & \vdots & \ddots & \vdots \\ I_{J,1}^S & I_{J,2}^S & \cdots & I_{J,K}^S \end{bmatrix}$$  (3)
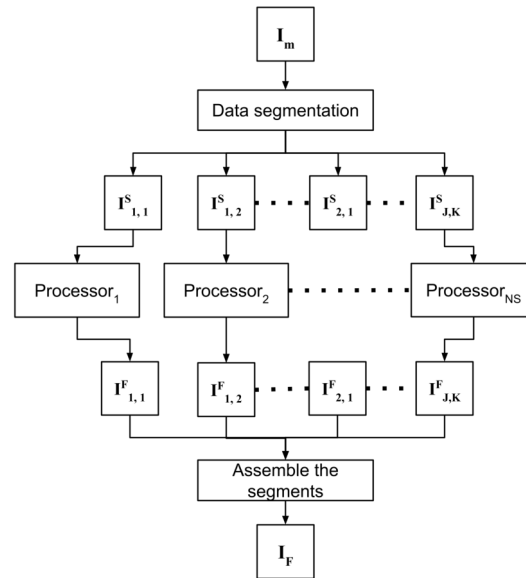
**Figure 2.** Filling in missing satellite data takes advantage of parallel processing to independently process previously divided segments and assemble results in a single orthomosaic

The user-defined values for $J$ and $K$ should be selected according to the computational resources available to execute DINEOF, and indirectly define the size of the segments $\{I_{j,k}^{S}\}$. Inspired by binary search, the orthomosaic may be evenly divided into $2 \times 2$, $4 \times 4$, $8 \times 8$, and so on. As soon as the computer system is able to process the images, the divisions are fixed and the monitoring process configured.

*Fill in missing data for each segment*

Filling in missing data is a parallel process that is independently applied to all segments in which the merged image was divided (see Fig. 2). Using a massive processing configuration (*e.g.* a computer cluster or a multiprocessor computer) is advantageous to accelerate the complete process. In this step, each segment $I_{j,k}^{S}$ is filled in, and the resulting filled segments $I_{j,k}^{F}$ are stored for posterior processing.

*Assemble segments*

At the final module, the resulting $I_{j,k}^{F}$ segments are assembled in the same order that was divided $I_M$, to obtain a new $I_F$ orthomosaic without holes.

$$
I_F = \begin{bmatrix} I_{1,1}^{F} & I_{1,2}^{F} & \cdots & I_{1,K}^{F} \\ I_{2,1}^{F} & I_{2,2}^{F} & \cdots & I_{2,K}^{F} \\ \vdots & \vdots & \ddots & \vdots \\ I_{J,1}^{F} & I_{J,2}^{F} & \cdots & I_{J,K}^{F} \end{bmatrix}
\tag{4}
$$

A distinct but equivalent way to define the number of segments is to establish the size of each segment $I_{j,k}^{F}$, assuming all segments are the same size. The size of $I_{j,k}^{F}$ corresponds to a 2-element tuple $(width, height)$ that define the number of pixels per side, considering the ratio between width and height of the segment to be the same of the ratio between the width and the height of the orthomosaic $I_M$: $\frac{width(I_{j,k}^{F})}{height(I_{j,k}^{F})} = \frac{width(I_M)}{height(I_M)}$.

## Computational complexity analysis

The application of *DINEOF* to a sequence of $T$ orthomosaics requires to assembly a $L \times T$ matrix, with $L = width \times height$ representing the number of pixels in $I_M$. After that, the resulting matrix is standardized, and the optimal number of empirical orthogonal functions (EOFs) are by the convergence of a validation process that depends on Singular Value Decomposition (SVD) computation. The computation of SVD

189  is in the order $O(LT^2)$, and the validation process depends on the maximum number of iterations ($Q$)
190  employed to find the optimal number of EOFs. Thus, the whole computation of *DINEOF* for a sequence
191  of $T$ orthomosaics is in the order $O(QLT^2)$, with typical values of $L \gg T$ and $L \gg Q$: the number of
192  pixels usually greatly exceeds the time frame $T$, as well as the iterations $Q$. Consequently, a significant
193  reduction in the number of pixels $L$ per segment $I_{j,k}^S$ causes a consequent reduction in the total number of
194  operations.

## STUDY CASE: CHLOROPHYLL ON THE EEZM

196  The study case used for proof of concept was designed to monitor the Chl-*a* over a wide sea area: EEZM.
197  Data from the MODIS and VIIRS sensors were combined to obtain L2 products with the least amount of
198  missing data. The importance of monitoring the Chl-*a* is related to the dynamics of phytoplankton, which
199  provides the information to predict the impact of climate change in ocean ecosystems. Phytoplankton is
200  composed of microscopic algae and other photosynthetic organisms that inhabit the surface of oceans,
201  rivers, and lakes. These microorganisms constitute the primary source of energy in aquatic systems due to
202  their photosynthetic capacity (Winder and Sommer, 2012), and their contribution to preserving the climate
203  balance and the biogeochemical cycle in such ecosystems (Hallegraeff, 2010). For some decades now, the
204  Chl-*a* has been widely used to estimate phytoplankton's biomass in surface water using satellite-based
205  methods (Gomes et al., 2020; Kramer and Siegel, 2019; O'Reilly et al., 1998). Such usage is given in
206  view of the fact that the Chl-*a* is the main photosynthetic pigment of phytoplankton. In fact, Chl-*a* is
207  used as a photoreceptor and gives the green color to the phytoplankton, and various studies have settled
208  the fundamentals of the impact of Chl-*a* with light reflectance of water bodies, especially in the visible
209  light and close infrared regions of the electromagnetic spectrum (Gitelson, 1992; Dall'Olmo and Gitelson,
210  2005; Yacobi et al., 2011).

### Study area

212  The area of study selected for the analysis and proof of concept corresponds to the EEZM, and covers the
213  sea region close to the seashore (CONABIO, 2022). The distance covered by the EEZM is up to 370.4
214  km from the continental and insular seacoast. The surface area of the EEZM is one of the greatest in the
215  world and is estimated to be 3 269 386 $km^2$.

216  The complete satellite images are required to study Chl-*a* concentrations, *e.g.* images without missing
217  data over the area of study. The size of such a huge area makes the task prohibitive for the computer
218  facilities available for experimentation. Bands 8 to 16 from the MODIS sensors were used, corresponding
219  to wavelengths from 405 to 877 nm and a spatial resolution of 1 km. These bands are mainly employed
220  for Ocean Color and Phytoplankton and Biogeochemistry. On the other hand, the VIIRS sensor provides
221  measurements from water, land, and atmosphere, with a temporal resolution of 12 hrs for day and night
222  ocean data acquisition.

### Computational details on experiments

224  A multiprocessor computer with distributed memory was used to run the experiments. The so called
225  Perseo computer, is part of the computing network of the *Centro Nayarita de Innovación y Transferencia*
226  *de Tecnología A.C., México*. The Perseo cluster is provided with 388 processing cores, 1,280 GB Ram,
227  356 TB permanent storage, and runs the `CentOS 7.0` operating system. The proposed approach
228  was implemented using a combination of scripts written in `Python` and Matlab 2018. In particular,
229  the automated *image download module* written in `Python`, and the whole processing code written in
230  `Matlab` are freely available to download through GitHUB: `https://github.com/jroberto37/`
231  `fill_missing_data.git`. The download script takes advantage of the geolocation products that
232  include `MOD03`, `MYD03`, `VNP03MODLL`, and `VJ103DNB`, for sensors MODIS-TERRA, MODIS-AQUA,
233  VIIRS-SNPP, and VIIRS-JPSS-1 respectively.

234  For preprocessing, the Graph Processing Tool (GPT) from the Sentinel Application Platform (SNAP)
235  was used to create the orthomosaics and project the sine wave system's data to the WGS-84. Segmentation
236  was performed over the merged high-resolution image to speed up the process of filling chlorophyll
237  concentration data, following the `NetCDF` format. The maximum area of the segments is defined in
238  the system configuration parameters and automatically establishes the number of segments in which the
239  image is divided. The `*.gher` binary files are then generated with their respective mask of the zone that

**Table 1.** Parameters employed in the evaluation of `DINEOF`

| Parameter | Description | Range |
|---|---|---|
| alpha | Parameter specifying the strength of the filter | [3 5 10 20 50] |
| numit | Number of iterations for the filter | [0.1 0.3 0.5] |
| nev | The maximum of number of modes you allow to compute | 20 |
| neini | The minimum number of modes you want to compute | 1 |
| ncv | The maximal size for the Krylov subspace | 35 |
| tol | The threshold for Lanczos convergence | 1.0e-8 |
| nitemax | The maximum number of iteration allowed for the stabilization of eofs obtained by the cycle | 300 |
| toliter | Precision criteria defining the threshold of automatic stopping of DINEOF iterations | 1.0e-3 |
| rec | For complete reconstruction of the matrix | 0 |
| eof | Writing the left and right modes of the input matrix | 0 |
| norm | Activate the normalization of the input matrix | 0 |
| seed | Seed to initialize the random number generator | 243435 |

is not processed (*e.g.*, land), as well as its `time` file that allows activating the filtering of the temporal covariance matrix.

The `*.gher` and `time` files generated at segmentation are then used to execute `DINEOF`, employing the configuration parameters shown in Table 1. The proposed algorithm rewrites such parameters in a file with the `*.init` extension. Afterward, the file is read by the `DINEOF` program, which computes the missing data for each of the segmented data series. In the fill-in missing data step, `DINEOF` generates a time series without holes for each segment, and it is stored in `*.gher` file format. Finally, the orthomosaic is reconstructed using the segmented high-resolution images without missing data. The resulting image is written in `NetCDF` format.

### *Evaluation in cloudy scenarios*

For validation, a free of the holes data set was generated for the time frame from January 2017 to December 2019. The data set was composed of 36 complete high-resolution images (without missing pixels), with a spatial resolution of 1 km from the four sensors. The images were composed with the 30 Chl-*a* daily images from each month. As an example, Fig. 3(a) shows that the Chl-*a* composed image corresponding to January 2018 does not present black or white regions, corresponding to zones with missing data.

Three scenarios were prepared to test the system under occlusion conditions by adding different levels of synthetic clouds to the composed images. The three levels of missing data were arbitrarily selected to represent different typical scenarios that are common in real data. Figs. 3(b), 3(c), and 3(d) show the composed image corrupted with the synthetic cloud masks, covering 20%, 30%, and 50% respectively. The generation of synthetic masks was based on real cloud images from the same scene at different dates (*e.g.* climate conditions), and the percentage of clouds was computed based on pixel counts. Regarding the cloud coverage in Fig. 3(b), a few clouds scarcely cover different regions of the sea, shaping natural clouds. Increasingly dense clouds are shown in Figs. 3(c), and 3(d), according to the corresponding percentage of the cloud masks.

## EXPERIMENTAL RESULTS

### Satellite data merging

The sensibility of the preprocessing satellite data module to the size of the sliding window was studied using nine different square windows: $m \times m$ windows with $m = \{3, 5, 7, 9, 11, 15, 21, 31 \text{ and } 51\}$. In this sensibility test, the base image employed for each sensor was composed by the sequence of images for February 2018; and missing samples were generated using the images for February the 1st, 2018, for each sensor.

Table 2 shows the RMSE, the percentage of filled data, the computation time that was employed in the nine different window sizes, and the mosaicking SNAP function. The mosaicking results are the reference for data coverage previous to the application of the sliding window. According to Table 2, the window sizes $5 \times 5$ and $7 \times 7$ presented the lowest RMSE value when compared to other window sizes. However, the latter showed a higher percentage of data coverage (28.71% against 25.52%), although the processing time increases according to the window size.
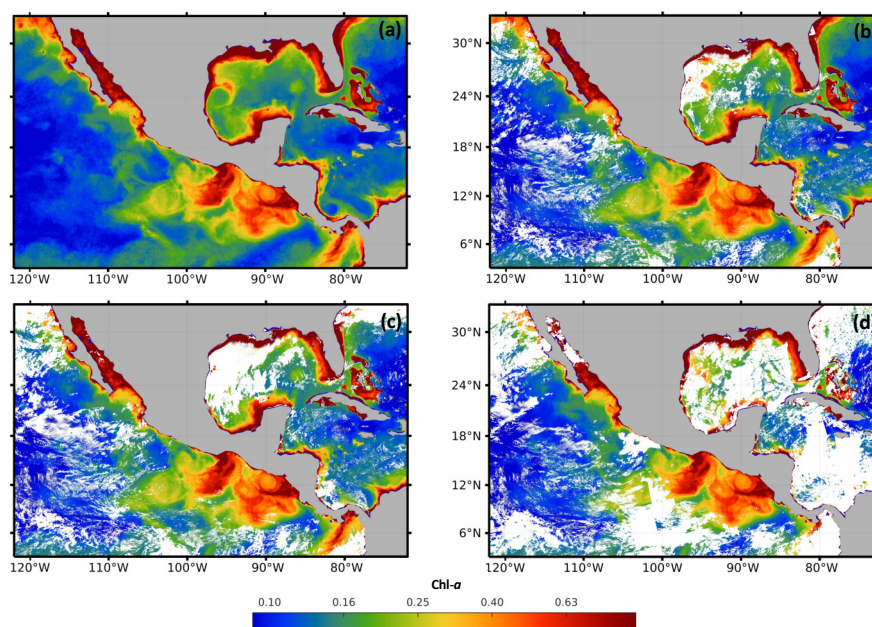
**Figure 3.** Maps of cloud masks in Mexico economic exclusive zone (EEZM). (a) Chl-*a* image composed by 30 scenes (January 2018), (b) Mask with 20% clouds, (c) Mask with 30% clouds and (d) Mask with 50% clouds

**Table 2.** Results of the preprocessing module in terms of RMSE, percentage of data coverage in the resulting orthomosaic, and the preprocessing time in seconds. Bold numbers symbolize the best results when distinct window sizes are compared

|  | Mosaicking | $3 \times 3$ | $5 \times 5$ | $7 \times 7$ | $9 \times 9$ | $11 \times 11$ | $15 \times 15$ | $21 \times 21$ | $31 \times 31$ | $51 \times 51$ |
|---|---|---|---|---|---|---|---|---|---|---|
| RMSE | 0.429 | 0.419 | **0.408** | **0.408** | 0.421 | 0.437 | 0.471 | 0.513 | 0.582 | 0.674 |
| Data coverage | 17.35% | 20.98% | 25.53% | **28.72%** | 30.57% | 32.31% | 35.01% | 37.96% | 41.37% | 45.56% |
| Time (s) | 7.753 | 8.748 | 13.402 | 17.583 | 19.314 | 21.640 | 28.685 | 40.228 | 60.736 | 138.663 |

²⁷⁸ In order to compare the results of the mosaicking and evidence the advantages of applying the sliding
²⁷⁹ window, Fig. 4 shows the results of the application of the method to high-resolution images for February
²⁸⁰ $1^{st}$, 2018. The four images in Fig 4 correspond to the central region of the Gulf of Mexico, with coordinates
²⁸¹ $North = 30.40^o$, $South = 18.60^o$, $East = -83.30^o$ and $West = -94.90^o$. Fig. 4(a) shows the original
²⁸² image from sensor VIIRS-JPSS-1. Fig. 4(b) shows the spatial distribution of the pixels from both methods
²⁸³ and the pixels that were filled with the proposed approach. Fig. 4(c) shows the results of the mosaicking
²⁸⁴ function from the SNAP software; and Fig. 4(d) shows the results obtained with the window size $7 \times 7$. A
²⁸⁵ visual comparison of Fig. 4(d) and Fig. 4(a) evidences the advantage of using the sliding window to reduce
²⁸⁶ the amount of missing data, even when compared to commercial software (Fig. 4(c)). Images from the
²⁸⁷ four sensors were complemented with different percentages of missing data: 67.72% for MODIS-AQUA,
²⁸⁸ 65.93% for MODIS-TERRA, 58.55% for VIIRS-SNPP, and 58.29% for VIIRS-JPSS-1.
²⁸⁹ Fig. 5 shows the orthomosaics obtained for each sensor after preprocessing downloaded samples and
²⁹⁰ applying the $7 \times 7$ sliding window. Due to the differences in trajectories and climate conditions at the
²⁹¹ overflight time, all orthomosaics present quite different areas of missing samples. For example, Fig. 5(a)
²⁹² and Fig 5(b) corresponding to MODIS Aqua and MODIS Terra respectively, have a band of missing data
²⁹³ at the center of the image, but with distinct orientations. On the other hand, Fig. 5(c) and Fig 5(d) that
²⁹⁴ correspond to VIIRS JPSS-1 and SNPP, do not present clear missing data patterns. Such differences favor
²⁹⁵ the exploitation of the different sources to obtain a more complete resulting orthomosaic $I_M$. Once the
²⁹⁶ four orthomosaics were generated, data from the VIIRS-JPSS-1 sensor was selected as the base image in
²⁹⁷ the merging preprocessed data module. The orthomosaic from the VIIRS-JPSS-1 sensor was chosen as
²⁹⁸ the base image ($I_b$) because it presents a lower percentage of missing data than the other sensors. Then,
²⁹⁹ the final merged $I_M$ is processed with DINEOF with the orthomosaics from previous days, as described in
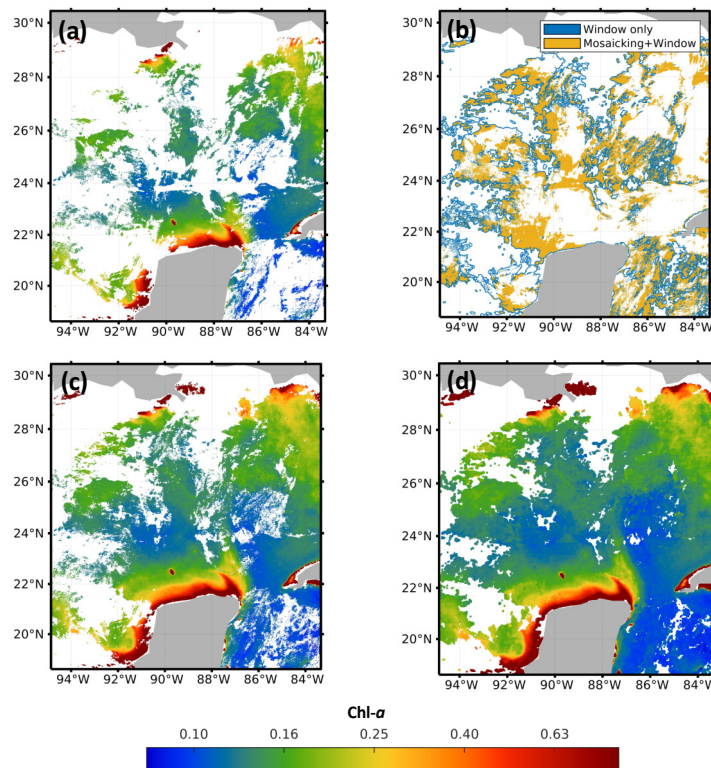
**Figure 4.** Impact of the pre-filling with sliding windows in the region of the Gulf of Mexico. (a) JPSS-1 original, (b) Filling zones, (c) Mosaicking and (d) Windows $7 \times 7$.

the following section.

**Filling in missing satellite data**

After preprocessing was applied to the whole temporal series 2018-2019, the impact of the three hyper-parameters was evaluated on the proposed system: (1) `alpha`, (2) `numit`, and (3) `time` (see Table 1). The values of the hyperparameters were explored through the application of DINEOF, after splitting the preprocessed orthomosaic into six independent zones (see Fig. 6). Such a division favors the analysis of Chl-*a*'s behavior either in coastal zones or deep sea, separated from coasts. For example, zone 4 presents deep seawater with a concentration of Chl-*a* that differs from the concentration in zone 5, which is closer to coasts. On the other hand, a high concentration of Chl-*a* can be observed close to the coasts in zones 1 to 4 and 6. In that sense, Fig. 6 presents the six zones in which the area of study was divided for the evaluation of the DINEOF tuning parameters.

The proposed approach was evaluated at two different levels. First, at the adjustment of internal hyperparameters of DINEOF, where image segmentation was adapted to $2 \times 3$ sub-images, and hyper-parameters *alpha* and *numit* were evaluated according to ranges in Table 1. The search for the more suitable hyperparameters for DINEOF was conducted by running two experimental designs: one for `time`=30, and another for `time`=60. During the adjustment process, the RMSE was estimated for the distinct possible values of `alpha` and `numit`, considering the ranges established in Table 1.

The resulting expected error obtained through the search process is shown in Fig. 7. The first and third rows of Fig. 7 (images a, b, c, g, h, and i) represent the expected error for the temporal series with `time`=30. Similarly, the second and fourth rows of Fig. 7 (images d, e, f, j, k, and l) represent the expected error for the temporal series with `time`=60. In all images from Fig. 7, the horizontal axis represents the `alpha` parameter, the vertical axis represents the `numit` parameter, and the color of the cells represents the expected error computed with DINEOF. The color scale is shown at the bottom of the same figure. According to Fig. 7(b) and 7(e), the highest expected error was attained at zone 2, either with `time`=30 or `time`=60. And in those cases, the values of `alpha`=0.1, and `numit`=3.0 present a
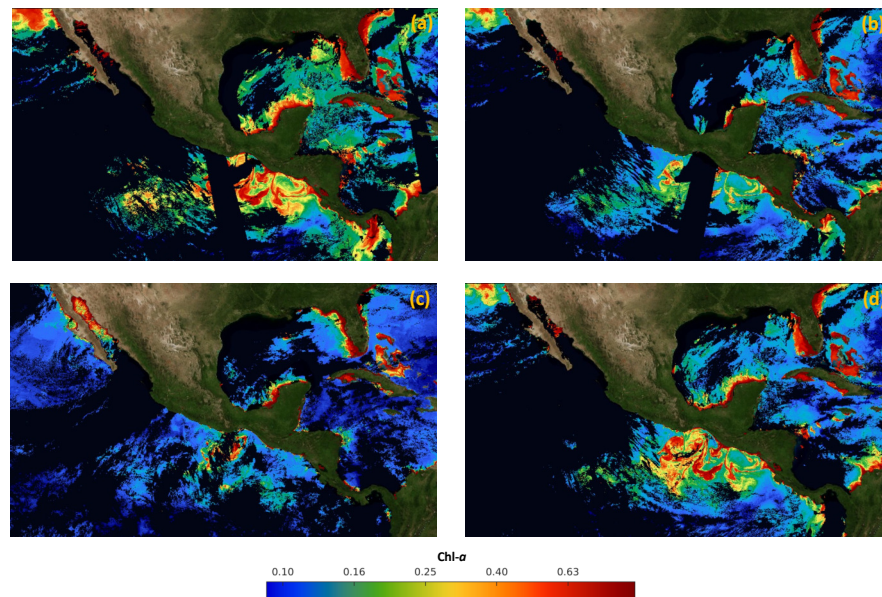
**Figure 5.** Maps of orthomosaics in Mexico economic exclusive zone (January 1st 2019). (a) MODIS Aqua orthomosaic, (b) MODIS Terra orthomosaic , (c) VIRSS JPSS-1 orthomosaic and (d) VIRSS SNPP orthomosaic.

lower expected error, *e.g.*, seems to be favorable in both scenarios. On the other hand, Fig. 7(g) and 7(j) exhibit the lowest expected error, regardless of the value assigned to both `alpha` and `numit`, as well as the timeframe. In the rest of the cases and regardless of the time frame, the lowest expected error is attained with `alpha`=0.1 and `numit`=3.0, and they were fixed for the application of the segmented fill in the algorithm.

With fixed hyperparameters, at the second level of evaluation, the RMSE was computed for distinct areas of the segments on the three cloudy scenarios (*e.g.* 20%, 30%, and 50% clouds). Four segmentation levels were considered for $I_M$ in order to parallelize the process, with image segments represented by the triplets $j \times k \times t$, with $j$ and $k$ as described in Section Filling in missing satellite data; and $t$ representing the time in trimesters. The segmentation levels correspond to $312 \times 187 \times 12$, $625 \times 374 \times 12$, $1250 \times 749 \times 12$, and $2500 \times 1498 \times 12$. However, the computer configuration employed to run the software was not able to completely run the system with the latter configuration due to memory overflow.

Table 3 presents the average time and RMSE that were obtained after the execution of the experimentation with the aforementioned segment sizes. According to Table 3, the computation of the missing data showed a better performance when the segments size and the amount of data to estimate were rather small compared to the size of $I_M$. In fact, the lowest RMSE was attained when $I_M$ was divided into $8 \times 8$ segments in the three cloud scenarios. In the hardest scenario, with 50% of missing data due to clouds, the proposed approach achieved an RMSE of 0.43, which was lower than all other feasible cases.

**Table 3.** Average time and RMSE obtained after the application of the proposed approach with distinct segment sizes. Bold numbers symbolize the lowest RMSE

| Segment size | $312 \times 187 \times 12$ | | $625 \times 374 \times 12$ | | $1250 \times 749 \times 12$ | |
|---|---|---|---|---|---|---|
| Cloud test | Time ($\sigma$) | RMSE ($\sigma$) | Time ($\sigma$) | RMSE ($\sigma$) | Time ($\sigma$) | RMSE ($\sigma$) |
| 20% | 7.52 (4.59) | 0.45 (0.11) | 43.41 (28.82) | **0.36 (0.08)** | 187.17 (111.09) | 0.37 (0.08) |
| 30% | 10.42 (27.84) | 0.47 (0.12) | 43.21 (27.76) | **0.35 (0.07)** | 175.84 (90.69) | 0.36 (0.06) |
| 50% | 13.86 (31.77) | 0.52 (0.16) | 58.36 (42.02) | **0.43 (0.17)** | 186.72 (130.17) | 0.44 (0.16) |

Finally, Fig. 8(a) presents the merged image $I_M$, created with the data acquired by the MODIS and VIIRs sensors on January 1st, 2019. On the other hand, Fig. 8(b) presents the final result of the proposed
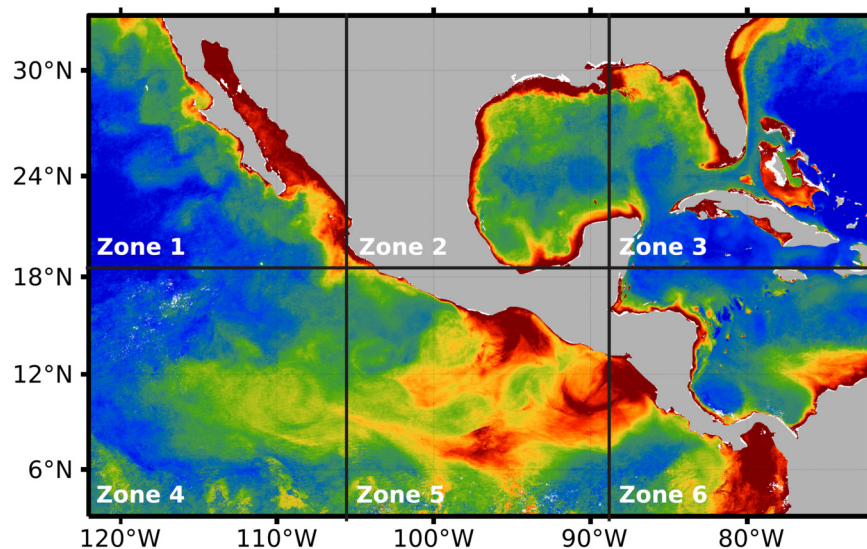
**Figure 6.** Segmentation of the area of study, performed automatically by the proposed approach

approach, which was created with satellite data from June 1st, 2018 to May 5th, 2019. It can be observed that there are no holes or missing data, and it is ready to create time-series related to the concentration of Chl-*a*.

### Computational complexity analysis

As an example that follows the case study, processing the sequences with the original size ($2500 \times 1498 \times 12$) requires as many operations as $O(300 \times 3,745,000 \times 12^2)$. On the other hand, by applying the proposed segmentation strategy, the task is divided into sixteen sequences of $312 \times 187 \times 12$, which require as many operations as $O(300 \times 58,344 \times 12^2)$. Following this example, Figure 9 presents the number of operations expected for each segment size, showing the effect of the split of the whole image into segments. Figure 9, it is evident the direct relationship between the size of the segments and the number of operations required to complete the segmented *DINEOF*: the smaller the segments, the fewer operations are required. However, it has to be pointed out that sensitivity analysis is important to verify the performance. As previously shown in Table 3, for the experimental settings analyzed in this paper, the best segment size found during experimentation was $625 \times 374 \times 12$. This analysis is suggested to be conducted on distinct scenarios, according to the particular requirements of the context.

### CONCLUSIONS

This research paper introduced a new efficient approach to filling in missing high-resolution data from different satellite sensors. The proposed approach consists of three general steps: (1) Automatic satellite data download, (2) Satellite data fusion, and (3) Filling of missing satellite data using a computationally intensive approach. The first novelty of the proposed approach is that data from different sensors are used when performing satellite data fusion. As a second contribution, an approach is introduced to be able to process high-resolution images of large study areas. The proposed approach divides the orthomosaic into segments that DINEOF can process; the segments can be processed in parallel using a computer cluster; next, the orthomosaic is reconstructed without missing data. Finally, an analysis of the computational effort of the proposed approach was performed and it was found to be of quadratic order.

As proof of concept, the proposed approach was applied to fill in the holes in satellite imagery, using data from MODIS sensors aboard the Terra and Aqua satellite platforms and VIIRS aboard the SNPP and JPSS-1 satellites. The multi-sensor fusion approach implements geospatial techniques such as sliding averages and inverse weights interpolation with the squared distance and approaches to adjust the differences in the time each platform overflights the zone of study. Results showed that the proposed approach overcomes the traditional averaging strategy. When compared to the proposed approach, the traditional average strategy produces zones in which the values of Chl-*a* are overestimated. In contrast,
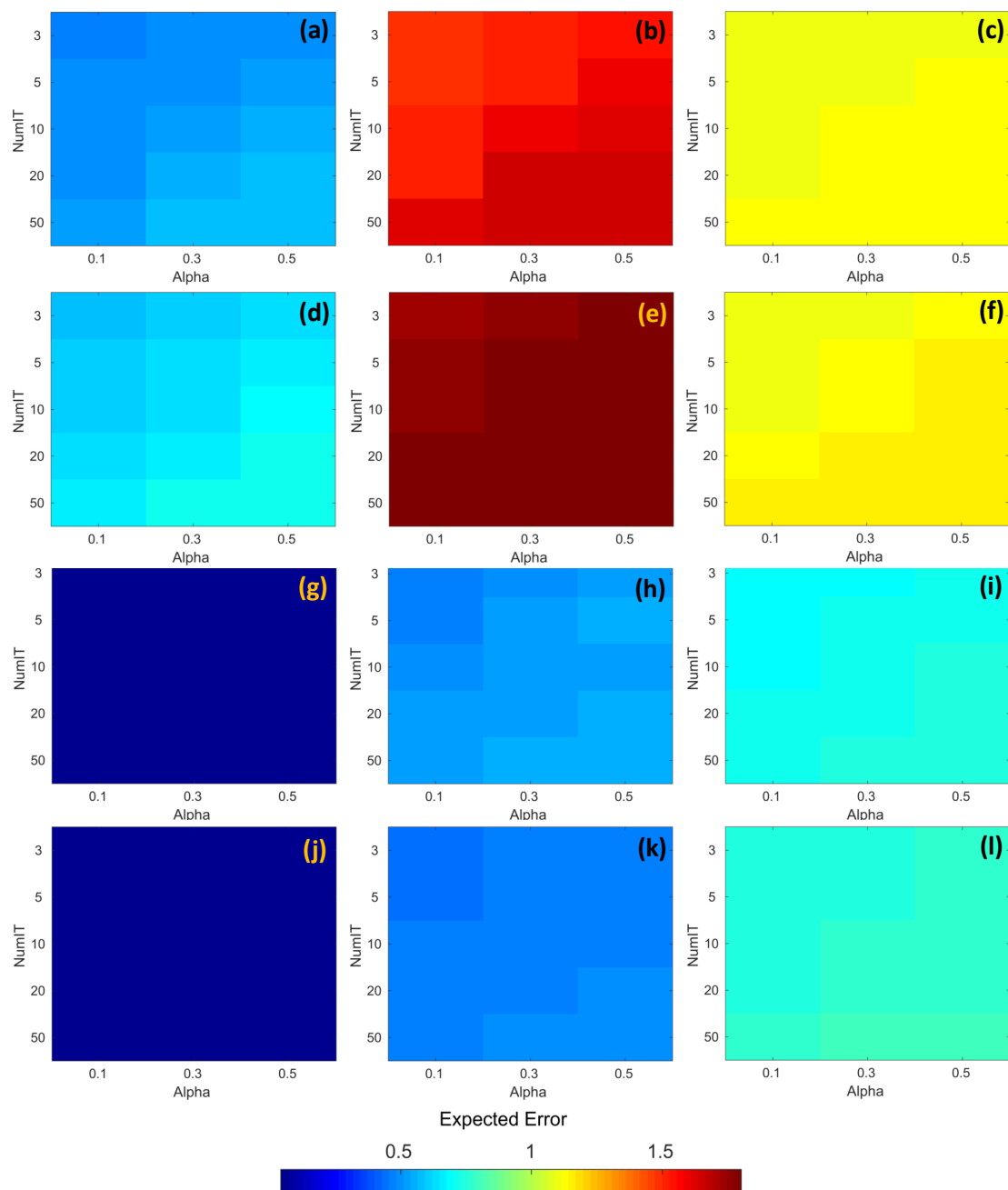
**Figure 7.** RMSE corresponding to the application of DINEOF for different values of `alpha` and `numit`, at distinct zones and time frames; the colorbar represents the RMSE. (a) Zone 1 / time=30, (b) Zone 2 / time=30, (c) Zone 3 / time=30, (d) Zone 1 / time=60, (e) Zone 2 / time=60, (f) Zone 3 / time=60, (g) Zone 4 / time=30, (h) Zone 5 / time=30, (i) Zone 6 / time=30, (j) Zone 4 / time=60, (k) Zone 5 / time=60 and (l) Zone 6 / time=60.

377   the proposed approach preserves the oceanic structures required to study the ocean dynamics related to
378   the currents and winds.

379      Although the DINEOF method is widely employed in several remote sensing studies to fill in missing
380   data, the direct process of high-resolution images with DINEOF results in a high computational cost in
381   terms of memory and computing power. For that reason, the proposed approach divides the merged image,
382   and each segment is processed separately using DINEOF. Finally, the processed segments are assembled
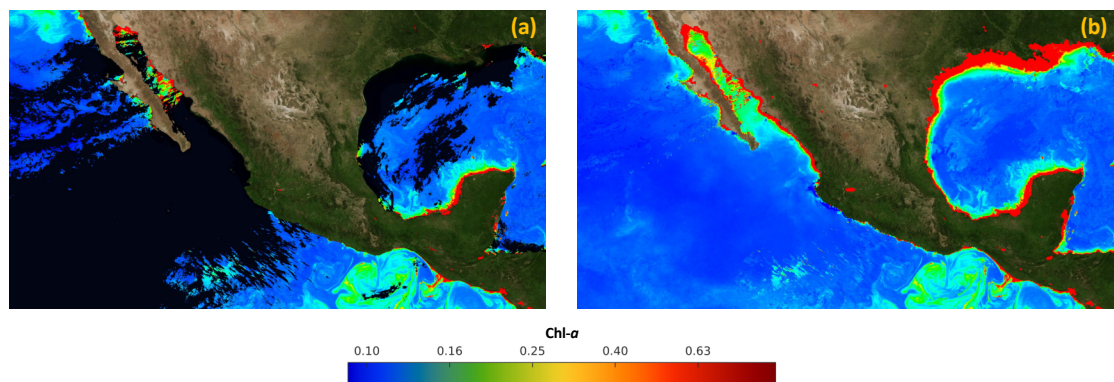
**Figure 8.** Results of the process of filling in missing data in satellite images. (a) Merged image ($I_M$). (b) Final Image without missing data ($I_F$).
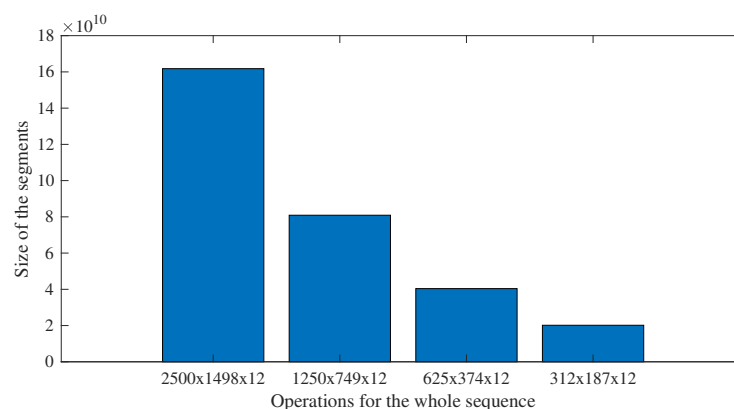
**Figure 9.** Number of operations expected as the segments are reduced, or equivalently, the number of segments are augmented.

without missing data. The process of adjusting the input parameters for DINEOF endorses its performance to fill in missing data through the time series analysis. The analysis of the results shows, in general terms, that the best outcomes are obtained with low values of `alpha` = 3 and `numint` = 0.1. The proposed approach was implemented using `Python 3` and `Matlab R2018b`, which enables the automation of repetitive tasks, including automated download of satellite data from the OBPG-Ocean Color Data for levels L2 and L3 of sensors MODIS and VIIRS. The characteristics of `Python` support the link between diverse platforms previously developed for satellite imagery: multiplatform and multiparadigm.

As future work, the proposed approach may be applied to distinct scenarios that may provide evidence of the efficiency of the proposed data-intensive approach. Between the application areas, one of the most relevant may be satellite-guided fishing through phytoplankton monitoring. In fact, phytoplankton constitutes the basic nourishment for small fishes, crustaceans, and other sea life forms that are the base food for larger fishes and sea mammals. Although the data-intensive approach was evaluated on sea monitoring, land applications may also be benefited from its efficiency. Additionally, with these results, it may be interesting to combine the proposed algorithm to address applications that incorporate artificial intelligence (*i.e.* forecasting, Etc.). Finally, incorporating the algorithm in open libraries may favor the comparison with future proposals in this research area.

## ACKNOWLEDGMENTS

This research was financed by the project: CONACYT-INEGI 290628. We thank the GHER group at the University of Liège for providing DINEOF; NASA OBPG for providing the satellite data used in experiments.

## REFERENCES

Alvera-Azcárate, A., Barth, A., SirJacobs, D., Lenartz, F., and Beckers, J. (2011). Data Interpolating Empirical Orthogonal Functions (DINEOF): a tool for geophysical data analyses. *Mediterranean Marine Science*.

Alvera-Azcárate, A., Barth, A., Troupin, C., Beckers, J., and Van Der Zande, D. (2021). Creation of high resolution suspended particulate matter data in the north sea from sentinel-2 and sentinel-3 data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 7446–7448.

Babbar, J. and Rathee, N. (2019). Satellite image analysis: A review. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6.

Beckers, J.-M. and Rixen, M. (2003). Eof calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and oceanic technology*, 20(12):1839–1856.

Bouchra, N., Aida, A.-A., Kevin, R., and Naomi, G. (2011). Reconstruction of MODIS total suspended matter time series maps by DINEOF and validation with autonomous platform data. *Ocean Dynamics*.

CONABIO (2022). Zona Económica Exclusiva de México. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, official website: `http://www.conabio.gob.mx/informacion/metadata/gis/contdv250_zeemgw.xml?_xsl=/db/metadata/xsl/fgdc_html.xsl&_indent=no` (accessed on March 2, 2022).

Dall'Olmo, G. and Gitelson, A. (2005). Effect of bio-optical parameter variability on the remote estimation of chlorophyll-a concentration in turbid productive waters: experimental results. *Applied optics*, 44(3):412–422.

Datla, R., Shao, X., Cao, C., and Wu, X. (2016). Comparison of the Calibration Algorithms and SI Traceability of MODIS, VIIRS, GOES, and GOES-R ABI Sensors. *Remote Sensing*, 8:1–26.

GHER (2020). Data INterpolating Empirical Orthogonal Function (DINEOF). Official website: `http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF/` (accessed on April 3, 2020).

Gitelson, A. (1992). The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing*, 13(17):3367–3373.

Gittings, J. A., Brewin, R. J., Raitsos, D. E., Kheireddine, M., Mustapha, O., Jones, B. H., and Hoteit, I. (2019). Remotely sensing phytoplankton size structure in the Red Sea. *Remote Sensing of Environment*, 234:111387.

Gomes, P., Valente, T., Geraldo, D., and Ribeiro, C. (2020). Photosynthetic pigments in acid mine drainage: Seasonal patterns and associations with stressful abiotic characteristics. *Chemosphere*, 239:124774.

Hallegraeff, G. M. (2010). Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge 1. *Journal of phycology*, 46(2):220–235.

Hu, C., Lee, Z., Ma, R., Yu, K., Li, D., and Shang, S. (2010). Moderate resolution imaging spectrora-diometer (modis) observations of cyanobacteria blooms in taihu lake, china. *Journal of Geophysical Research: Oceans*, 115(C4).

Jayaram, C., Pavan Kumar, J., Udaya Bhaskar, T. V. S., Bhavani, I. V. G., Prasad Rao, T. D. V., and Nagamani, P. V. (2021). Reconstruction of gap-free OCM-2 chlorophyll-a concentration using DINEOF. *Journal of the Indian Society of Remote Sensing*, 49(6):1419–1425.

Jayaram, C., Priyadarshi, N., Kumar, J. P., Bhaskar, T. V. S. U., Raju, D., and Kochuparampil, A. J. (2018). Analysis of gap-free chlorophyll-a data from MODIS in Arabian Sea, reconstructed using DINEOF. *Taylor & Francis*.

Kramer, S. J. and Siegel, D. A. (2019). How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans*, 124(11):7557–7574.

Kramer, Herbert J. (2020). National Polar-orbiting Partnership (SUOMI). Official website: `https://earth.esa.int/web/eoportal/satellite-missions/s/suomi-npp` (accessed on April 3, 2020).

Liu, X. and Wang, M. (2018). Gap filling of missing data for viirs global ocean color products using the dineof method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4464–4476.

Liu, X. and Wang, M. (2019). Filling the gaps of missing data in the merged viirs snpp/noaa-20 ocean color product using the dineof method. *Remote Sensing*, 11(2):178.

NASA (2020). Moderate Resolution Imaging Spectroradiometer (MODIS). Official website: `https:`

458 //modis.gsfc.nasa.gov/ (accessed on April 3, 2020).

459 O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., and
460 McClain, C. (1998). Ocean color chlorophyll algorithms for seawifs. *Journal of Geophysical Research:*
461 *Oceans*, 103(C11):24937–24953.

462 Rodriguez-Ramirez, R., Sánchez, M. G., Rivera-Caicedo, J. P., Fajardo-Delgado, D., and Avila-George,
463 H. (2019). Automating an image processing chain of the sentinel-2 satellite. In Mejia, J., Muñoz, M.,
464 Rocha, Á., Peña, A., and Pérez-Cisneros, M., editors, *Trends and Applications in Software Engineering*,
465 pages 216–224, Cham. Springer International Publishing.

466 Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., and Conci, N. (2019). Natural
467 disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications*,
468 78(22):31267–31302.

469 Sato, M., Takahashi, Y., Kubota, H., Noda, A., Hamada, J., and Lopez, G. V. C. (2021). Quasi-Real
470 Time Monitoring of Lightning and Weather in the Philippines and Western North Pacific for the Severe
471 Weather Intensity Prediction. In *EGU General Assembly Conference Abstracts*, EGU General Assembly
472 Conference Abstracts, pages EGU21–13950.

473 Winder, M. and Sommer, U. (2012). Phytoplankton response to a changing climate. *Hydrobiologia*,
474 698(1):5–16.

475 Yacobi, Y., Moses, W., Kaganovsky, S., Sulimani, B., Leavitt, B., and Gitelson, A. (2011). Nir-red
476 reflectance-based algorithms for chlorophyll-a estimation in mesotrophic inland and coastal waters:
477 Lake kinneret case study. *Water research*, 45(7):2428–2436.

478 Zhang, J., Clayton, M. K., and Townsend, P. A. (2015). Missing data and regression models for spatial
479 images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1574–1582.

480 Zhang, Q., Yuan, Q., Zeng, C., Li, X., and Wei, Y. (2018). Missing data reconstruction in remote sensing
481 image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions*
482 *on Geoscience and Remote Sensing*, 56(8):4274–4288.