International Journal of Mathematics and
Computer Science, **16**(2021), no. 1, 345–356

Reconstruction of Chlorophyll-a Data by Using DINEOF Approach in Sepanggar Bay, Malaysia

Fatin Nadiah Binti Mohamed Yussof[1], **Normah Binti Maan**[1], **Mohd Nadzri Bin Md Reba**[2]

[1]Department of Mathematics
Faculty of Science and Technology
Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia

[2]Faculty of Geoinformation and Real Estate
Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia

email: fatinnadiah5@gmail.com, normahmaan@utm.my

(Received July 7,2020, August 7, 2020)

Abstract

Loss of spatial data with a long gap is a significant limitation for remote sensing analyses using satellite-based monitoring of oceanography. This limitation could not be ignored as it may affect the subsequent analysis and modeling of the data. Hence, this gap needs to be improved by filling the spatial gap in the satellite datasets. In this research, Data Interpolating Empirical Orthogonal Functions (DINEOF) is applied to fill the spatial gap and has successfully worked in the reconstruction of missing data of chlorophyll-a for monitoring harmful algal blooms (HABs) in Sepanggar Bay located at coastal water of Kota Kinabalu, Malaysia. The original chlorophyll-a pixels are used to assess the accuracy of the predicted data. Then, the DINEOF model is compared with the Spatio-temporal Kriging model for validation purposes. The results obtained show that the DINEOF model

Key words and phrases: Chlorophyll-a, DINEOF, Spatial Long Gaps, Spatio-temporal Kriging.
AMS (MOS) Subject Classifications: 62D10.
ISSN 1814-0432, 2021, http://ijmcs.future-in-tech.net

has the highest Pearson correlation coefficient, 0.9940 and the smallest values of Root Mean Square Error (RMSE) and Mean Absolute Deviation (MAD) which are 0.2770 and 0.0155 respectively. Therefore, this proved that the DINEOF model is more effective for filling spatial long gaps.

1 Introduction

Monitoring oceanography variables such as chlorophyll-a by satellite sensors are widely being used nowadays because of their broad coverage in time and space [1] not to mention gaining the data quickly. In contrast, in situ, samplings impose a high cost and are time-consuming. However, satellite image data frequently caused spatial data loss due to clouds covering the measurements of ocean reflectance and effects from sunlight, atmospheric correction, adjacency from land, and bottom reflectance [2, 3]. It is important to fill the gaps because that would affect the modeling of the data and, as a result, lead to unreasonable inference and inaccurate results [4]. The missing data cannot be simply deleted since that would affect the originality and norm of the data [5]. Therefore, suitable methods to interpolate missing data need to be applied in order to solve this problem. The obstacles faced by the missing data are to find the correct way to predict the missing data [6, 7]. There are three types of missing data, These are Missing At Random (MAR), Missing Completely At Random (MCAR), and Missing Not At Random (MNAR) [8]. For MAR, the available data will be representative of the whole population. MCAR happens when the events leading to the missing data are independent of both observable and unobservable parameters. In contrast, MNAR occurs when the missing data are dependent on each other, which is one or more factors that are impossible to quantify and identify [9]. There are many types of interpolation methods used to overcome the missing data of Spatio-temporal data sets [10, 11, 12, 13, 14, 15, 16] which commonly include the construction of data based on pixel neighborhoods [17], optimal interpolation (OI) [18], Kriging [19] and empirical orthogonal functions (EOF) [20]. The Data Interpolating Empirical Orthogonal Functions (DINEOF) method turned out to be an excellent interpolation method, among the EOF methods, for high cloud coverage [21]. DINEOF has been applied to SST [3, 22, 23, 24], chlorophyll-a [25, 26, 28], turbidity [1] and total suspended matter (TSM) [6]. Generally, the methods consist of three categories, which are temporal interpolation, spatial interpolation, and Spatio-temporal interpolation. The objective of this paper is to reconstruct missing data of chlorophyll-a con-

centration in Sepanggar Bay located in the coastal area of Kota Kinabalu, Sabah, Malaysia, by applying the DINEOF method. Five years of MODIS derived chlorophyll-a time series is used in this paper. The accuracy of the DINEOF model is compared with another interpolation methods known as the Spatio-temporal Kriging which we will discuss later in this paper.

2 Materials and Methods

2.1 Location of Study

The study area is in the coastal water of Kota Kinabalu, Sabah, which is Sepanggar Bay (Fig. 1). The freshwater inflows in this region come from the Inanam and Menggatal rivers together with the factory waste and domestic sewage. The nearby inland is the terrestrial area and there are reclaimed areas and a seaport at the southern side of Sepanggar Bay. On the north-western side of the bay, there is Gaya Island which acts as a shield while in the eastern side there is an aquaculture project for broodstock. As a result, the productivity of the phytoplankton varies temporally and spatially.

2.2 Data Sets

The data used in this study is a week composite of chlorophyll-a concentration obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite sensors covering the Sepanggar Bay in the coastal waters of Kota Kinabalu, Sabah ($115.2^0N, 116.2^0N, 5.9^0N, 6.9^0N$) during the four-year period 2015-2018. The chlorophyll-a images were obtained in MODIS level 3 from the National Aeronautics and Space Administration (NASA) Ocean Color WEB server with a 4.00 km resolution. The dimensions of the data are 494 x 178 pixels. Among 190 images, there exist images which undergo extreme cloud coverage of up to 95 %. These types of images that consist of less than 5% of data is unreliable because it might affect the quality of the reconstruction since they do not provide useful information [1]. Hence, only 178 images are used in this study.

2.3 Data Interpolating Empirical Orthogonal Functions (DINEOF)

This method is developed for the reconstruction of missing data in oceanographic data sets [29]. Let X be the initial matrix of dimension m x n,

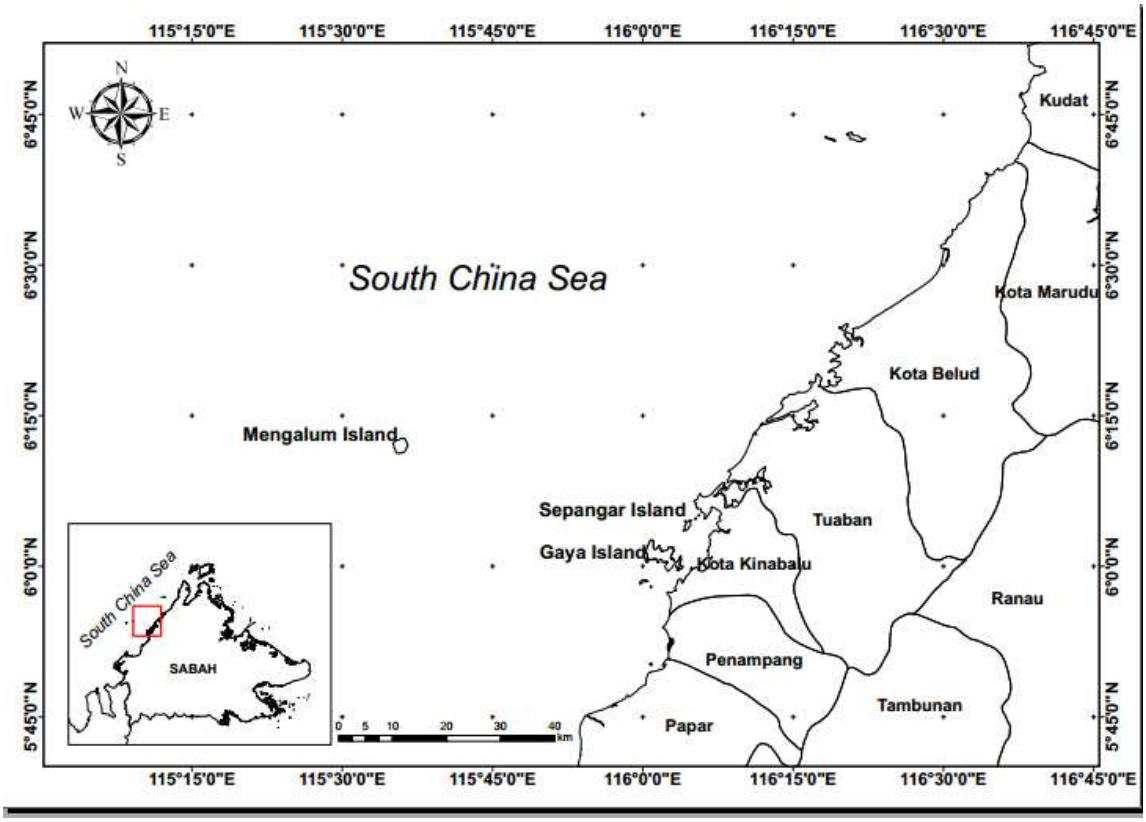

Figure 1: Location of the study area in coastal water of Kota Kinabalu, Sabah.

$m > n$, where m is the spatial dimension and n is the temporal dimension. The data might contain some unknown values due to the missing data. In order to compute the EOF decomposition, the Singular Value of Decomposition (SVD) technique was used in which the missing data is introduced as an initial guess. The equation can be defined as follows:

$$X = USV^T, \tag{2.1}$$

where U is the spatial EOFs with dimension p x h, the pseudo-diagonal matrix S with dimension h x h signifies singular values and V as the temporal EOFs with dimension q x h. The matrix X rank depends on the value of h where $h \leq min(p, q)$. To obtain the best estimate of the field by reconstruction method, choose the spatial and temporal EOFs that are the most significant. Eigenvector decomposition calculates the corresponding vector

of the k biggest singular values.

$$XX^T u_i = \rho_i^2 u_i, \tag{2.2}$$

where the ith column of U is denoted as u_i and ρ is the perspective singular value, $i = 1, ..., k$. The following equation can be applied:

$$Av_i = \rho_i^2 v_i u_i = \frac{Xv_i}{\rho_i} \tag{2.3}$$

in which $A = X^T X$ is real symmetric of q x q matrix DINEOF described as:

1. Data sets were stored in a p x q matrix form where p represents the number of pixels and q is the number of images. In this study, the matrix size applied was 494 x 178. For the reconstruction process, a random of 5% of the valid data (no. of pixels) was initialized to 0, where we assumed the data as missing to be used in the cross-validation for the unbiased guess. However, the missing data were 'flagged' to distinguish them from those existing points on the mean. Then, the corresponding Spatio-temporal mean of the initial matrix was subtracted from the initial matrix. The matrix is denoted by X and is applied throughout the whole method.

2. By using the first k EOFs, the first estimate of the singular values and singular vectors are obtained. Then, the elements of matrix X are replaced by the values obtained with the EOF series.

$$X_i, j = \sum_{p=1}^{k} \rho_p (u_p)_i (v_p^T)_j \tag{2.4}$$

   The steps are repeated until the given k achieved its convergence.

3. The new value of the missing data is obtained by recomputing the EOFs after an improved guess has been introduced for the missing data. The steps are as follows:

   - There is a number of estimates at the end after the convergence is achieved with EOFs. The best number of estimates is obtained by cross-validation.
   - A random number of data set is applied for the cross-validation technique to get the best number of EOFs, which gives the minimum value of error between data set aside and the values obtained at these points with the reconstruction method.

- The whole procedure is repeated once the optimal number N of EOFs is obtained. The data set aside for cross-validation is also included by only considering the first N EOFs. Then, compute the final value of the missing data.
- The DINEOF method is implemented by using the 3.0 Linux binary available through GHER (Geo-Hydrodynamics and Environmental Research) by the University of Liege [30] while MATLAB software is applied for data handling and analysis.

2.4 Spatio-temporal Kriging

Kriging background: Generally, the Kriging method only applies to spatial data where the spatial data concerns the location of the data and the distance between location points is taken into account. Let be the spatial data set of an attribute z at a position where U is a vector of spatial coordinates. The Kriging method is used to combat the missing values of z at a set of m locations. Basically, Kriging is a method based on regression using observed surrounding data points, weighted according to covariance values.
Spatio-temporal Kriging: Chlorophyll-a concentration characteristics change over time and space. Chlorophyll-a concentration is also correlated with concentrations of the past and future. Therefore, it is necessary to consider the time dimension in the Kriging model for better estimation. Spatio-temporal Kriging takes into account the spatial location $\rho_\alpha = (x_\alpha, y_\alpha)$ with the temporal timestamp t_i . In contrast, ordinary Kriging only considers one spatial dimension where $\rho_\alpha = x_\alpha$ and only utilizes temporal information for imputation in which is the mile marker.
Chlorophyll-a is formulated as $Q(\mu_\alpha, t_i);\ \alpha = 1, 2, ..., n; i = 1, 2, ..., m$ in the space-time framework. The covariance of this model is similar to spatial models where it is the difference of the variance of the mean squared between data separated by a given spatial and temporal lag (h_s, t_s) :

$$C(h_s, h_t) = E[(z(u_\alpha, t_\alpha - z(u_\alpha + h_s, t_\alpha + h_t))^2]. \tag{2.5}$$

The experiment semivariogram is computed to half of the covariance in order not to be different from the common practice in spatial statistics:

$$\hat{\gamma}_{s,t}(h_s, h_t) = 1/2E[(z(u_\alpha, t_\alpha) - z(u_\alpha + h_s, t_\alpha + h_t))^2]. \tag{2.6}$$

The missing value $Q^*(\mu, t)$ of the normal space-time Kriging system can be estimated as weighted average of values of surrounding locations:

$$Q^*(u, t) = \sum \lambda_{\alpha,i}(u, t)Q(u_\alpha, t_i) \text{ with } \sum \lambda_\alpha(u, t) = 1 \tag{2.7}$$

The weights $\lambda_{\alpha,i}(\mu_\alpha, t_i)$ assigned to each neighboring data point are calculated by minimizing the prediction variance:

$$\sigma^2(u,t) = Var[Q^*(u,t) - Q(u,t)], \tag{2.8}$$

while maintaining unbiasedness of the estimated value $Q^*(\mu, t)$.

3 Evaluation criteria

The accuracy of DINEOF method and Spatio-temporal Kriging is evaluated by using the Pearson correlation coefficient (r), mean absolute deviation (MAD) and root mean square error (RMSE). Suppose there is a number of missing data points n in the test data sets with Y^i_{act} as the ground truth for i^{th} missing data point and Y^i_{est} as the estimated value for the missing data point. The formula is as follows:

$$MAD = \frac{\sum_i^n |Y^i_{act} - Y^i_{est}|}{n} \tag{3.9}$$

$$RMSE = \sqrt{\frac{\sum_i^n (Y^i_{act} - Y^i_{est})^2}{n}} \tag{3.10}$$

4 Results and Discussion

Figure 2 illustrates the image of the original gappy image and also the reconstruction image by using the DINEOF method. Clearly, it can be seen that there is lots of missing details in the original image due to cloud cover. After applying the DINEOF method for reconstruction purposes, there is no gappy image. Then, the accuracy of the DINEOF method and the Spatio-temporal Kriging method are evaluated, as shown in Table 1. This is to compare the value of the Pearson Correlation Coefficient (r), RMSE, and MAD.
The accuracy of DINEOF method and Spatio-temporal Kriging method are evaluated as shown in Table 1. From Table 1, we can see that the DINEOF method gives the most outstanding result compared to Spatio-temporal Kriging method. The value of r is the highest which shows that the reconstructed missing data have high correlation with the actual value. RMSE and MAD value of DINEOF is the smallest than Spatio-temporal Kriging method. This defined that DINEOF method has high accuracy in recontructed the missing data compared to Spatio-temporal Kriging. This result also shows that

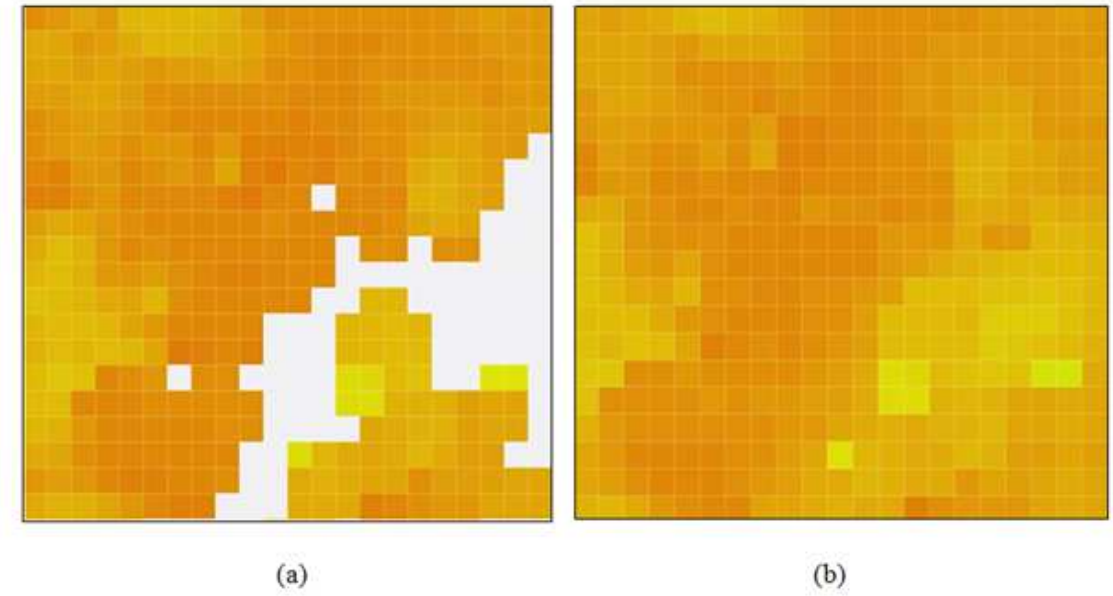

Figure 2: Week composite image for week 1 of January 2018. (a) Original gappy image (b) Reconstruction image

Table 1: Reconstructed result using DINEOF and Spatio-temporal Kriging

Method	**r**	**RMSE**	**MAD**
DINEOF	0.9940	0.2770	0.0155
Spatio-temporal Kriging	0.8795	0.3587	0.1512

the DINEOF method is more effective when filling long gaps missing data. Monitoring harmful algal bloom events by using satellite sensors encounter limitations with loss of spatial data. Therefore, in this study. DINEOF method is applied to reconstruct the missing data of chlorophyll-a in Sepanggar Bay located in the coastal water of Kota Kinabalu, Sabah. The data sets of week composite are obtained from the MODIS satellite sensors of level 3 is having a long loss of spatial gaps.

5 Conclusion

The findings show that the DINEOF method outperforms the Spatio-temporal Kriging method, where it can be seen that in the findings, the method has the

highest correlation coefficient compared to Spatio-temporal Kriging method. Besides that, it also has the smallest values of RMSE and MAD, which shows that the DINEOF method has high accuracy in filling long loss of spatial gaps. Therefore, the DINEOF method can be used for filling the gaps of missing data for satellite sensors.

Acknowledgments

This work was supported by Transdisciplinary Research Grant Scheme (TRGS) under the research program of Characterisation of Spatio-temporal Marine Microalgae Ecological Impact Using Multi-sensor Over Malaysia Waters for the specific sub-project of Mathematical Modelling of Harmful Algal Blooms (HABs) in Malaysian Waters (R.J130000.7809.4L854) funded by the Ministry of Education, Malaysia. The authors are also thankful to Universiti Teknologi Malaysia for providing the facilities for this research.

References

[1] Aida A. Azcárate , Alexander Barth, Michel Rixen, Jean M. Beckers, Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature, Ocean Modeling, **9,** no. 4, (2005), 325–346.

[2] Andrea Hilborn, Maycira Costa, Applications of DINEOF to satellite-derived chlorophyll-a from a productive coastal region, Remote Sensing, **10,** no. 9, (2018), 1449–1470.

[3] Damien Sirjacobs, Aida A. Azcárate, Alexander Barth, Geneviève Lacroix, YoungJe Park, Bouchra Nechad, Kevin Ruddick, Jean M. Beckers, Cloud filling of ocean color and sea surface temperature remote sensing products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions methodology, Journal of Sea Research, **65,** no. 1, (2011), 114–130.

[4] Min Deng, Zide Fan, Qiliang Liu, Jianya Gong, A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets, ISPRS International Journal of Geo-Information, **5,** no. 2, (2016), 13.

[5] Zhipeng Gao, Weijing Cheng, Xuesong Qiu, Luoming Meng, A missing sensor data estimation algorithm based on temporal and spatial correla-

tion, International Journal of Distributed Sensor Networks, **11,** no. 10, (2015), 1–10.

[6] Bouchra Nechad, Aida A. Azcaràte, Kevin Ruddick, Naomi Greenwood, Reconstruction of MODIS total suspended matter time series maps by DINEOF and validation with autonomous platform data, Ocean Dynamics, **61,** no. 8, (2011), 1205–1214.

[7] Andrew Gelman, Jennifer Hill, Data analysis using regression and multilevel/hierarchical models,Cambridge UniversityPpress, 2006.

[8] Donald B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons Inc., New York, 2004.

[9] Joseph L. Schafer, John W. Graham, Missing data: our view of the state of the art, Psychological methods, **7,** no. 2, (2002), 147.

[10] Celestino O. Galán, Fernando S. Lasheras, Javier C. Juez , Antonio B. Sánchez , Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions, Journal of Computational and Applied Mathematics, **311,** (2017), 704–717.

[11] Antonio M. Durán-Rosal, Cesar H. Martínez, Antonio J. Tallón-Ballesteros, Alfonso C. Martínez-Estudillo, Sancho S.Sanz , Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks, Ocean Engineering, **117,** (2016), 292–301.

[12] Sehyun Tak, Soomin Woo, Hwasoo Yeo, Data-driven imputation method for traffic data in sectional units of road links, IEEE Transactions on Intelligent Transportation Systems, **17,** no.6, (2016), 1762–1771.

[13] Francesco Tonini, Whalen W. Dillon, Eric S. Money, Ross K. Meentemeyer, Spatio-temporal reconstruction of missing forest microclimate measurements, IEEE Transactions on Intelligent Transportation Systems, **218,** (2016), 1–10.

[14] Shreenivas Londhe, Pradnya Dixit, Shalaka Shah, Shweta Narkhede, Infilling of missing daily rainfall records using artificial neural network, IEEE Transactions on Intelligent Transportation Systems, **21,** no.3, (2015), 255–264.

[15] John Tipton, Mevin Hooten, Simon Goring, Reconstruction of spatio-temporal temperature from sparse historical records using robust probabilistic principal component regression, Advances in Statistical Climatology, Meteorology and Oceanography, **3,** no. 1, (2017), 1–16.

[16] Wenjie Ruan, Peipei Xu, Quan Z. Sheng, Nickolas JG Falkner, Xue Li, Wei E. Zhang, Recovering Missing Values from Corrupted Spatio-Temporal Sensory Data via Robust Low-Rank Tensor Completion, International Conference on Database Systems for Advanced Applications, (2017), 607–622.

[17] Brandon Casey, Robert Arnone, Peter Flynn, Simple and efficient technique for spatial/temporal composite imagery, International Society for Optics and Photonics, **6680,** (2007), 1–08.

[18] Andrew F. Bennett, Inverse modeling of the ocean and atmosphere, Cambridge University Press, 2005.

[19] Dagmar Müller, Estimation of algae concentration in cloud covered scenes using geostatistical methods, Proceedings of ENVISAT symposium held in Montreux, Switzerland, April 23-27, 2007.

[20] Marc H. Taylor, Martin Losch, Manfred Wenzel, Jens Schröter, On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data, Journal of Climate, **26,** no. 22, (2013), 9194–9205.

[21] Jean M. Beckers, Alexander Barth, Aida A. Azcárate, DINEOF reconstruction of clouded images including error maps? application to the Sea-Surface Temperature around Corsican Island, Ocean Science, **2,** no. 2, (2006), 183–199.

[22] Unai Ganzedo, Aida A.Azcarate, Ganix Esnaola, Agustin Ezcurra, Jon Saenz, Reconstruction of sea surface temperature by means of DINEOF: a case study during the fishing season in the Bay of Biscay, International journal of remote sensing, **32,** no. 4, (2011), 933–950.

[23] Yizhen Li, Ruoying He, Spatial and temporal variability of SST and ocean color in the Gulf of Maine based on cloud-free SST and chlorophyll reconstructions in 2003–2012, Remote sensing of environment, **144,** (2014), 98–108.

[24] Andrea C. Acosta, Carmen E. Morales, Samuel Hormazabal, Isabel Andrade, Marco A. Correa-Ramirez, Phytoplankton phenology in the coastal upwelling region off central-southern Chile (35 S–38 S): Time-space variability, coupling to environmental factors, and sources of uncertainty in the estimates, Journal of Geophysical Research: Oceans, **120,** no. 2, (2015), 813—831.

[25] Jason N. Waite, Franz J. Mueter, Spatial and temporal variability of chlorophyll-a concentrations in the coastal Gulf of Alaska, 1998–2011, using cloud-free reconstructions of SeaWiFS and MODIS-Aqua data, Progress in Oceanography, **116,** (2013), 179–192.

[26] Xiaoming Liu, Menghua Wang, Gap filling of missing data for VIIRS global ocean color products using the DINEOF method, IEEE Transactions on Geoscience and Remote Sensing, **56,** no. 8, (2018), 4464–4476.

[27] Aida A. Azcárate, Quinten Vanhellemont, Kevin Ruddick, Alexander Barth, Jean M. Beckers, Analysis of high frequency geostationary ocean color data using DINEOF, Estuarine, Coastal and Shelf Science, **159,** (2015), 28–36.

[28] Aida A. Azcárate, Alexander Barth, Damien Sirjacobs, Fabian Lenartz, Jean M. Beckers, Data Interpolating Empirical Orthogonal Functions (DINEOF): A Tool for Geophysical Data Analyses, Estuarine, Medit. Mar. Sci. Spec. Issue, (2011), 5–11.

[29] Jean M. Beckers, Michel Rixen, EOF calculations and data filling from incomplete oceanographic data sets, Journal of Almospheric and Oceanic Technology, **20,** no. 2, (2003), 1839–1856.

[30] Tapio Schneider, Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, Journal of climate, **14,** no. 5, (2001), 853–871.