Estimation of secondary phytoplankton pigments from satellite observations using self-organizing maps (SOMs)

Roy El Hourany, Marie Abboud-abi Saab, Ghaleb Faour, Olivier Aumont, Michel Crépon, Sylvie Thiria

JGR Oceans

RESEARCH ARTICLE
10.1029/2018JC014450

Key Points:
- The dynamic of phytoplankton communities can be observed while reconstructing accessory pigment variability from ocean color observations
- Self-organizing maps allow an accurate retrieval of different phytoplankton pigment concentrations from satellite observations
- A consistent global approach is established to estimate accessory pigment concentrations along with the uncertainties

Correspondence to:
R. El Hourany,
roy.elhourany@locean-ipsl.upmc.fr

Citation:
El Hourany, R., Abboud-Abi Saab, M., Faour, G., Aumont, O., Crépon, M., & Thiria, S. (2019). Estimation of secondary phytoplankton pigments from satellite observations using self-organizing maps (SOMs). *Journal of Geophysical Research: Oceans*, *124*, 1357–1378. https://doi.org/10.1029/2018JC014450

Received 7 AUG 2018
Accepted 6 FEB 2019
Accepted article online 7 FEB 2019
Published online 27 FEB 2019

Estimation of Secondary Phytoplankton Pigments From Satellite Observations Using Self-Organizing Maps (SOMs)

Roy El Hourany[1,2], Marie Abboud-Abi Saab[3], Ghaleb Faour[2], Olivier Aumont[1], Michel Crépon[1], and Sylvie Thiria[1,4]

[1]IPSL/LOCEAN, Sorbonne Université (Université Paris VI, CNRS, IRD, MNHN), Paris, France, [2]National Center for Remote Sensing, National Council for the Scientific Research, Beirut, Lebanon, [3]National Center for Marine Sciences, National Council for the Scientific Research, Batroun, Lebanon, [4]Observatoire de Versailles Saint-Quentin-en-Yvelins (OVSQ) Versailles Saint-Quentin-en-Yvelines University, Versailles, France

Abstract This study presents a method for estimating secondary phytoplankton pigments from satellite ocean color observations. We first compiled a large training data set composed of 12,000 samples; each sample is composed of 10 in situ phytoplankton high-performance liquid chromatography (HPLC)-measured pigment concentrations, GlobColour products of chlorophyll-a concentration, and remote sensing reflectance (Rrs(λ)) data at different wavelengths, in addition to advanced very high resolution radiometer sea surface temperature measurements. The resulting data set regroups a large variety of encountered situations between 1997 and 2014. The nonlinear relationship between the in situ and satellite components was identified using a self-organizing map, which is a neural network classifier. As a major result, the self-organizing map enabled reliable estimations of the concentration of chlorophyll-a and of nine different pigments from satellite observations. A cross-validation procedure showed that the estimations were robust for all pigments ($R^2 > 0.75$ and an average root-mean-square error = 0.016 mg/m^3). A consistent association of several phytoplankton pigments indicating phytoplankton group specific dynamic was shown at a global scale. We also showed the uncertainties for the estimation of each pigment.

Plain Language Summary The knowledge of phytoplankton variability is essential to the understanding of the marine ecosystem dynamics and its response to environmental changes. This paper presents a new approach to estimate phytoplankton pigment concentrations from satellite observations by using an artificial neural network, the so-called self-organizing map. This neural network was calibrated using a large data set of in situ pigment observations from oceanic cruises along with ocean color satellite data provided by the Globcolour project and advanced very high resolution radiometer sea surface temperature. This approach allows an accurate estimation of phytoplankton pigment concentrations and their related uncertainties. Moreover, the method allows to reproduce the spatio-temporal variability of pigment concentration and the dynamics of phytoplankton groups. A particular attention is given to the Southern Ocean whose phytoplankton communities are specific.

1. Introduction

Marine ecosystems are a major sink for atmospheric CO_2 (Häder et al., 2014). The net transfer of CO_2 from the atmosphere to the oceans and then sediments is mainly a direct consequence of the combined effect of the water solubility (physical pump) and the biological pump (Hülse et al., 2017). The biological carbon pump is a key natural process and a major component of the global carbon cycle that regulates atmospheric CO_2 levels, transferring both organic and inorganic carbon fixed by phytoplankton in the euphotic zone to the ocean interior (Chisholm, 1995; Hülse et al., 2017). Understanding the response of the biological carbon pump to global change is required to accurately predict the future impact of the increase in atmospheric CO_2 due to the human activities (Passow & Carlson, 2012). The dissolution of anthropogenic CO_2 in the ocean and the subsequent formation of carbonic acid have already resulted in a decrease of 0.1 pH unit and will continue to lower pH by an additional 0.2–0.3 pH units by the end of the century (Bhadury, 2015). This decline in ocean pH is referred to as ocean acidification (Orr et al., 2005). At the same time, warming should increase the mean surface temperatures by an average of 3 °C, leading to longer periods of stratification with fewer deep mixing events (Sarmiento et al., 2004) and a less efficient physical pump. Increased stratification

is expected to lead to nutrient limitation and an increase in average irradiance in the euphotic layer, where phytoplankton grow (Schulz et al., 2007). The knowledge of the space and time heterogeneity of phytoplankton abundance is essential to understand the marine ecosystem dynamics and responses to environmental changes (Mann & Lazier, 2006).

Phytoplankton are distributed among a large number of groups of microscopic photosynthesizing protists and cyanobacteria, which contribute nearly 50% to the total primary production of Earth by fixing about 50 Gt carbon per year (Baumert & Petzoldt, 2008). In order to understand and quantify the phytoplankton abundance and characteristics at the surface, in situ methods evolved over the past years passing from microscope count method to flow cytometer and high-performance liquid chromatography (HPLC) pigment diagnosis. In fact, HPLC enables the identification of 25 to 50 phytoplankton pigments within a single analysis for which each phytoplankton group is associated with specific diagnostic pigments (DPs). Therefore, HPLC measurements are now widely used to determine phytoplankton species in situ. However, these methods are time-demanding with a low cost-effectiveness and their applications are limited to accessible zones where seawater sampling can be made. Besides, it has been shown that many pigments overlap between unrelated groups leading to misinterpretation. Nevertheless, these methods are considered as references for phytoplankton identification.

In the last decade, remote sensing of surface optical properties has provided synoptic views of the abundance and distribution of sea surface constituents; the downward sunlight interacts with the seawater through backscattering and absorption in such a manner that the upward signal transmitted to the satellite contains information related to the composition of seawater. In the open ocean far from the coast (case 1 waters), the sunlight mainly interacts with the phytoplankton cell shape (backscattering), pigments (absorption), and debris (e.g., colored dissolved organic matter). Therefore, the light seen by the satellite sensor contains information on phytoplankton. Ocean color has been an effective platform to estimate the chlorophyll-a (Chla) concentration in surface waters, providing synoptic measurements over the world ocean (Antoine et al., 1996; Behrenfeld et al., 2005; Behrenfeld & Falkowski, 1997; Longhurst et al., 1995; Westberry et al., 2008).

Besides Chla, many other pigments in phytoplankton interact with the sunlight, such as chlorophyll-b (Chlb), chlorophyll-c (Chlc), and photosynthetic carotenoids (PSCs), or in protecting Chla and other sensitive pigments from photodamage, such as photoprotective carotenoids (PPC). Some pigments only occur in specific phytoplankton groups and are thus indicator pigments for their identification, for example, fucoxanthin in diatoms and peridinin in dinoflagellates (Letelier et al., 1993; Vidussi et al., 2001). The identification of these pigments by remote sensing would provide an unprecedented spatio-temporal distribution (Kostadinov et al., 2017) that would give powerful insights on the phytoplankton composition, light absorption, physiological state (Behrenfeld & Boss, 2006), and the dynamics of the marine food web and marine productivity (Bracher et al., 2017). This was early recognized that detection of major characteristics of the phytoplankton community from remote sensing images was a major challenge in ocean optics.

Phytoplankton absorption bears the imprints of different types of pigments and can be measured by optical measurements. Several recent studies have investigated the potential of using continuous optical data to derive surface concentrations of pigments other than Chla. Chase et al. (2013) decomposed a large global data set of hyperspectral particulate absorption measurements into Gaussian function components and assessed the magnitude of specific Gaussian functions in relation to the absorption by specific pigments or pigment groups. The method provided robust results for obtaining concentrations of Chla, Chlb, Chlc, phycoerythrin, PPC, and PSC. Organelli et al. (2013) used a multivariate approach applied to fourth-derivative spectra of phytoplankton or particulate absorption data to retrieve Chla, seven DPs, and the corresponding phytoplankton size classes at the Boussole site in the Mediterranean Sea. Pan et al. (2010) developed empirical algorithms based on reflectance ratios to approximate key phytoplankton pigment concentrations. The band-ratio algorithms were developed from radiometric measurements collocated with pigment data measured in northeastern U.S. coastal waters. This algorithm has successfully derived the concentrations of Chla, Chlb, Chlc, and nine different carotenoids. However, such band-ratio algorithms require a very large database (>400 collocations with satellite data) from a specific region to derive robust regionalized algorithm. Bracher et al. (2015a) developed models to estimate phytoplankton pigments from hyperspectral in situ and satellite measurements of remote-sensing reflectance (Rrs(λ)), which were applied to the Atlantic Ocean. These models were based on empirical orthogonal function analysis of normalized Rrs(λ) spectra.

In the present work, we propose to use self-organizing maps (SOMs; Kohonen, 2013), in order to evidence the relationship between satellite and in situ data measured at the ocean surface. The SOM are unsupervised neural classifiers, which were successfully applied in remote sensing for pattern extraction (Ehsani & Quiel, 2008; Gorricha & Lobo, 2012; Hu & Weng, 2009; Iskandar, 2010; Richardson et al., 2003) and more specifically to ocean color measurements for aerosol characterization (Diouf et al., 2013; Niang et al., 2006), radiance spectra classification (Ainsworth & Jones, 1999), and for phytoplankton absorption spectra analysis (Chazottes et al., 2006, 2007). This study aims at providing a global, accurate, and robust method that can be used to retrieve the HPLC pigments composition from the remote sensing signal and consequently permits to track the phytoplankton dynamics. The use of SOM gives the ability to automatically assign different patterns of satellite signal and the associated pigment composition based on the similarities in terms of shape and amplitude. In addition, the method allows us to take into account a larger in situ data set in a very efficient manner in terms of processing time, with a higher flexibility and reliability.

The paper is articulated as follows. In section 2, we briefly describe the global ocean database co-located with satellite data ($\mathbf{D_{pigment}}$) used to calibrate the SOM we developed to estimate the pigment concentration. The SOM and the calibration procedure are described in section 3. In section 4, we present the results and the validation of the method. In section 5, we estimate the pigment concentration in the global ocean. The results are discussed in section 6. A conclusion is presented in section 7.

2. Materials

The proposed method is based on a statistical clustering of complementary available remotely sensed parameters and in situ HPLC pigment data by using a SOM. For that, we used a rigorous data set of in situ HPLC pigment data collocated with satellite data, regrouping a large diversity of encountered situations and multiple parameter combinations between 1997 and 2014. Our approach is based on the spectral diversity of the water leaving signal, in conjunction with the variability of the phytoplankton pigments. The application of such approach will allow the spatio-temporal reconstruction of pigments. In this section, both the data and the methodology are described and explained.

2.1. GlobColour Data

To extend existing time series beyond that provided by a single satellite sensor, the ESA initiated the GlobColour project (http://www.globcolour.info/) to develop a satellite based ocean color data set to support global carbon-cycle research. It aims at satisfying the scientific requirement for a long (10+ years) time series of consistently calibrated global ocean color information with the best possible spatial coverage. This has been achieved by merging data from Sea-viewing Wide Field-of-view Sensor (SeaWiFS), Moderate Resolution Imaging Spectroradiometer (MODIS), Visible Infrared Imaging Radiometer Suite (VIIRS), Medium Resolution Imaging Spectrometer (MERIS), and Ocean and Land Colour Instrument (OLCI).

The GlobColour project provides a continuous data set of merged level 3 (Mapped, 4 km) daily remote sensing reflectance (Rrs(λ)). This product is generated for each instrument, using the corresponding level 2 data. The merged Rrs(λ) are then computed as the weighted average of all single-sensor products. The 547- to 560-nm bands are submitted to a specific processing just before averaging to prepare a more consistent merging between the sensors. These bands were fitted to the Rrs(555) of SeaWiFS. The main reason behind this choice is that SeaWiFS is widely considered as the highest quality sensor with the best match to in situ observations and is commonly used in peer literature (Belo Couto et al., 2016). The primary cause for using this data set was to increase the number of match-ups between HPLC in situ data and satellite parameters.

Basically, ocean-color sensors measure the diffuse sunlight backscattered by the ocean. The principle of detecting phytoplankton groups from space relies on their spectral contributions to the Rrs(λ), which in turn is determined by the spectral absorption (a, m^{-1}) and backscattering (b_b, m^{-1}) coefficients of the ocean (pure water and various particulate and dissolved matters) using this simplified formula (Morel & Gentili, 1996):

$$\mathrm{Rrs}(\lambda) = G \times b_b(\lambda)/(a(\lambda) + b_b(\lambda)) \quad (1)$$

where G is a parameter mainly related to the geometry of the situation (sensor and solar angles) but also to environmental parameters (wind, inherent optical properties, and aerosols).

Table 1
Compilation of HPLC Databases and Their Contributions in Terms of the Sampling Period, Zone, and Number of Observations

Campaign/data set	Zone	Period	N Obs
MAREDAT[a]	Global ocean	1997–2008	10,340
NOMAD[b]	Global ocean	1997–2003	1,080
GeP&Co[c]	Pacific-Atlantic-Indian Oceans	July 2000 to September 2002	1,205
Polarstern[d]	Atlantic ocean	October 2007 to May 2010	396
Labrador[e]	Labrador Sea	2005–2014	253
Tara Oceans Expedition[f]	Global ocean; Med	September 2009 to October 2013	410
SeaBASS[g]	Global ocean	1997–2014	1,836

Note. HPLC, high-performance liquid chromatography.
[a]Luo et al. (2012) [b]Werdell and Bailey (2005). [c]Dandonneau et al. (2004). [d]Bracher et al. (2015a). [e]Fragoso et al. (2016). [f]Pesant et al. (2015). [g]https://seabass.gsfc.nasa.gov/.

The contribution of the phytoplankton to the Rrs(λ) can be explained by its pigment content, which absorbs the light at specific wavelengths, and its physical structure, which scatters the light as a function of the wavelength.

In this study, we used Rrs(λ) at four different wavelengths (412, 443, 490, and 555 nm). These Rrs(λ) were downloaded between 1997 and 2014. Each of these Rrs(λ) depends on several biogeochemical and physical factors such as the influence of phytoplankton pigments on the variability of Rrs(443), Rrs(490), and Rrs(555), especially the 490-nm wavelength due to the maxima of absorption of several secondary pigments near this wavelength.

We also added the Chla concentration, which gives an important information on the total phytoplankton abundance. The Chla was estimated with the OC5 algorithm (Gohin, 2011) and is provided on a daily basis via the GlobColour portal. The use of the 4 Rrs(λ) along with the Chla OC5 in the classification procedure will help to identify the nonlinear relationship between Rrs(λ)-Chla-pigments and then to accurately estimate the pigment composition. The Rrs(λ) data were the primary variables to be used in this approach, evaluating the optical relationship with the corresponding pigment composition. This is mainly due to the diversity of the Rrs(λ) spectra that will be analyzed by the SOM in terms of shape and amplitude. Adding the Chla OC5 as a second parameter allows to take into account the inter-relationship between auxiliary pigments and Chla and the relationship between Chla OC5 and Rrs(λ; on which the OC5 algorithm is based).

2.2. Advanced Very High Resolution Radiometer Sea Surface Temperature Data

Since the pigment concentrations are characterized by a well-defined seasonal cycle, the use of sea surface temperature (SST) in the algorithm of pigment retrieval permits to better fit the relationship between the in situ HPLC and satellite data (Pan et al., 2013). Therefore, adding SST should promote the identification of pigments covarying with the seasonal variability of this physical factor. The SST data were downloaded at 4-km resolution and at a daily frequency between 1997 and 2014, using a product of the advanced very high resolution radiometer (AVHRR) instruments aboard National Oceanic and Atmospheric Administration (NOAA) polar-orbiting satellites: version 5.3 level 3 global 4-km sea surface temperature (Casey et al., 2010; Saha et al., 2018; https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:AVHRR_Pathfinder-NCEI-L3C-v5.3). Current retrieval algorithms for SST from AVHRR are based largely upon the multichannel sea surface temperature algorithm (McClain et al., 1985), which may be written as

$$\mathrm{SST} = A_1 + A_2{}^{*}T_4 + Y\,(T_4 - T_5) \tag{2}$$

where A_1 and A_2 are constants determined through a least squares fit to in situ data; T_4 and T_5 are brightness temperatures as derived from channels 4, 10.3–11.3 μm and 5, 11.5–12.5 μm; and Y is a weighting factor based on the knowledge of known absorption coefficients (Emery et al., 1994).

Table 2
Phytoplankton Groups and Their Associated Major Pigments

Phytoplankton size class	Phytoplankton functional group	Pigment
Microphytoplankton (20 to 200 μm)	Diatoms[a]	Fucoxanthin (Fuco)
	Dinoflagellates[b]	Peridinin (Perid)
Nanophytoplankton (2 to 20 μm)	Prymnesiophytes[c]	19'HexFucoxanthin (19HF)
	Chromophytes[c]	19HF and 19'ButFucoxanthin (19BF)
	Chlorophytes[c]	Chlorophyll-b (Chlb)
	Cryptomonads[d]	Alloxanthin (Allo)
Picophytoplankton (<2 μm)	Cyanobacteria-Synechoccocus[e]	Zeaxanthin (Zea)
	Prochlorococcus[e]	Divynil-chlorophyll-a and -b (DVChla and DVChlb)

[a]Jeffrey (1980). [b]Jeffrey and Hallegraeff (1987). [c]Wright and Jeffrey (1987). [d]Gieskes and Kraay (1983). [e]Guillard et al. (1985).

2.3. HPLC Pigment Data Set

In parallel to the remotely sensed data, the in situ HPLC database used in this work is a compilation of different published data sets such as MAREDAT, NOMAD, and SeaBASS and several oceanic campaigns such as GEP&CO, TARA Ocean Expedition, Polarstern, and Labrador Sea expeditions (Table 1). This database gathers 10 different pigments: total Chla, Divynil-Chla (DVChla), Chlb, Divynil-Chlb (DVChlb), 19'Hexfucoxanthin (19HF), 19'Butfucoxanthin (19BF), Fucoxanthin (Fuco), Peridinin (Perid), Alloxanthin (Allo), and Zeaxanthin (Zea). Allo and Zea are both included in PPC pigments, while Fuco, Perid, 19HF, and 19BF are PSC pigments.

All pigments except Chla are considered as DPs. A diagnostic pigment analysis can be applied to classify phytoplankton types from HPLC pigment data (Table 2; Vidussi et al., 2001). Diagnostic pigment analysis defines an ensemble of DP for specific phytoplankton groups that can be quantified with respect to the sum of all the relative DP concentrations (i.e., DP/ΣDP) to estimate the relative abundance of a specific phytoplankton group. This approach has been used in different studies at both global and regional scales (Di Cicco et al., 2017; Hirata et al., 2011; Marty et al., 2002; Mayot et al., 2017; Sammartino et al., 2015; Uitz et al., 2006; Vidussi et al., 2001).

2.4. The Experimental Database

A total of 15,520 HPLC duplicate-free samples were compiled in this database, limited to the first optical depth which is about 15–35 m (D'Ortenzio & d'Alcalà, 2009). The duplicates were identified by comparing the coordinates (latitude, longitude), the date (year, month, and the day), and the depth. Vector data with similar or close values within these parameters were averaged. Due to their different optical properties (Arrigo et al., 1998; Fenton et al., 1994; Mitchell et al., 1991), 3,761 HPLC samples originating from the Antarctic Ocean were excluded (latitude>50°S). Afterward, satellite matchups were retrieved by extracting the nearest available pixel to the 11,759 remaining HPLC sample coordinates in a 3 × 3 box, within a time lapse of ±1 day. To specialize the applicability of the method to oceanic waters, which are characterized by low to moderate Chla levels, in particular in comparison with coastal waters, the 95% percentile of the data was used, which corresponds to a threshold at 3 mg/m^3 for the Chla values. High values above this percentile were flagged and replaced by a missing value in the remaining data. For consistency, the 95% percentile was then used to flag all high values of each variable. At this stage, 853 measurements with more than 14 missing variables were eliminated from the database. The resulting database contains 10,906 sample, from which 7,594 samples are collocated with satellite images between 1997 and 2014. The geographic location of this database samples is presented in Figure 1.

Despite the scattered missing values, the remaining data contain information, which may efficiently contribute to the calibration of the SOM, and this is due to the learning algorithm dedicated to the SOM method that will take into account the intervariable relationship (see section 3.1).

Finally, the experimental database ($\mathbf{D_{pigment}}$), that will serve to train the SOM, has a dimension of 10,906 × 16, where 10,906 are the number of samples (individuals) and 16 are the number of components (10 HPLC pigments: in situ Chla, DVChla, Chlb, DVChlb, 19HF, 19BF, Fuco, Perid, Allo, and Zea, and six satellite-derived variables: OC5 Chla, 4 Rrs(λ), and SST).

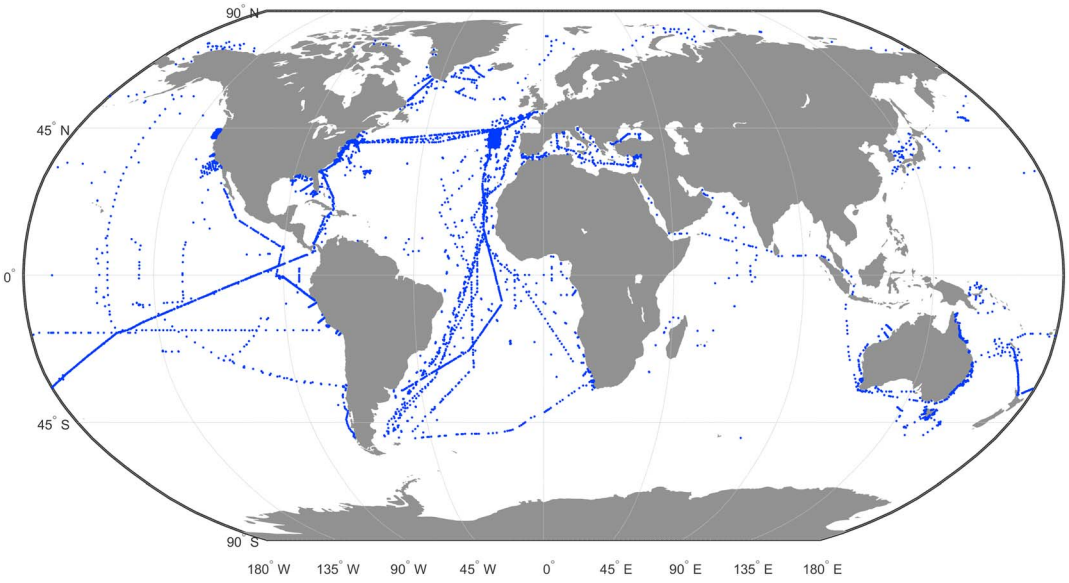

Figure 1. Geographic location of the high-performance liquid chromatography in-situ measurements.

3. The Proposed Method

3.1. SOM: The General Concept

The SOM algorithms (Kohonen, 2013) constitute powerful nonlinear unsupervised classification methods. They are unsupervised neural classifiers, which have been commonly used to solve environmental problems (Cavazos & Cavazos, 1999; Liu, 2005; Liu et al., 2006; Niang et al., 2006; Reusch et al., 2007; Richardson et al., 2003). SOM aims at clustering vectors of a multidimensional database (**D**) into classes represented by a fixed network of neurons (the SOM map). The SOMs are defined as an undirected graph, usually a rectangular grid of dimension $p \times q$. This graph structure is used to define a discrete distance (denoted by δ) between the neurons of the map, which present the shortest path between two neurons. Moreover, SOM enables the partition of **D** in which each cluster is associated with a neuron of the map and is represented by a prototype that is a synthetic multidimensional vector (the referent vector **w**). Each vector $\mathbf{z_i}$ of **D** will be assigned to the neuron whose referent **w** is the closest, in the sense of the Euclidean Norm, and will be called the projection of the vector $\mathbf{z_i}$ on the map. A fundamental property of a SOM is the topological ordering provided at the end of the clustering phase: two close neurons on the map represent data that are close in the data space. Indeed, the neurons are gathered in such a way that two close vectors of **D** are projected on two *relatively* close neurons (with respect to δ) on the map. The estimation of the referent vectors **w** of a SOM and the topological order is achieved through a minimization process in which the referent vectors **w** are estimated from a learning data set (the DPIG database in the present case). The cost function is of the form

$$J^T_{\mathrm{SOM}}(\chi, \mathrm{W}) = \sum_{\mathrm{zi}\in \mathrm{D}} \sum_{\mathrm{c}\in \mathrm{SOM}} K^T(\delta(c, \chi(z_i)))\|z_i - w_c\|^2 \quad (3)$$

where c ϵ SOM indices the neurons of the SOM map, χ is the allocation function that assigns each element $\mathbf{z_i}$ of **D** to its referent vector $\mathrm{w}_{\chi(\mathrm{z_i})}$. $\delta(\mathrm{c}, \chi(\mathrm{z_i}))$ is the discrete distance on the SOM between a neuron c, and the neuron allocated to observation $\mathbf{z_i}$ and $K^T a$ kernel function parameterized by **T** (where **T** stands for *temperature* in the scientific literature dedicated to SOM) that weights the discrete distance on the map and decreases during the minimization process.

This cost function takes into account the proper inertia of the partition of the data set **D** and ensures that its topology is preserved.

SOMs have frequently been used in the context of completing missing data (Jouini et al., 2013), so the projected vectors $\mathbf{z_i}$ may have missing components. Under these conditions, the distance between a vector $\mathbf{z_i}$ ϵ **D** and the referent vectors **w** of the map is the Euclidean distance that considers only the existing components

(the truncated distance or TD hereinafter). The use of the TD allows to take into account the information embedded in the incomplete data described in the section 2.

3.2. Construction of the SOM

3.2.1. Phase 1: Training Phase

In the present study, the SOM map is constituted by a two-dimensional rectangular grid ($200 \times 100 = 20{,}000$ referents) trained using $\mathbf{D_{pigment}}$, minimizing the $J^{T}_{\mathrm{SOM}}(\chi, W)$ cost function. In order to equitably distribute the weights through the training procedure, the 16 parameters were normalized with their variances. So each parameter contributes to build the SOM. Using a number of neurons larger than that of the training data set allowed to refine the discretization of $\mathbf{w}$ and therefore to obtain a more accurate pigment estimation. Several experiments were made to find the ideal SOM size and have shown a significant increase of the general performance of the method at estimating pigment concentrations when the number of neurons increases to a certain extent (5,000; 10,000; and 20,000 neurons).

In the case of 20,000 neurons map, at least half of the neurons of the SOM have captured a sample of the database, which permitted to define a referent vector $\mathbf{w}$ for these neurons. The second half of the neurons defined their $\mathbf{w}$ through the topological ordering using the equation (3). In other terms, the discrete distance $\delta(c, \chi(z_i))$ between the neighboring neurons and the kernel K^{T} was used to determine the referent vector $\mathbf{w}$ of each neuron that has not captured any data (Sarzeaud & Stephan, 2000). This proves the interest of the topological order provided by the SOMs maps, which is used to accurately interpolate referent vectors for neurons, which have not captured any data.

At the end of the training phase, each neuron of the SOM (denoted SOM-Pigments) was therefore associated with a referent vector $\mathbf{w_k}$ constituted of 16 components, with $k \in \{1 \ldots 20{,}000\}$.

3.2.2. Phase 2: Pigment Retrieval

In the second phase, which is an operating phase, we estimate the pigment concentrations using the different satellite images. The six ocean satellite observations (four Rrs(λ), satellite Chla, and SST) of a pixel P_j are projected onto the SOM. Doing so, the projected parameters are normalized with the corresponding variances of $\mathbf{D_{pigment}}$ to maintain an equal weight among the parameters and are assigned with the closest best matching neuron using the TD defined in section 3.1. At the end of the assignment phase, each pixel is associated with a referent vector $\mathbf{w_k}$ corresponding to the best matching neuron, which includes the 10 pigment concentrations.

For this phase, level 3 mapped 4 km daily images of SST, Chla, and Rrs(λ) at four wavelengths (412, 443, 490. and 555 nm) were used to estimate the pigment concentrations.

3.2.3. Phase 3: Cross Validation

In order to evaluate the performance of the SOM on a global scale, the statistical parameter R^2 and root-mean-square error (RMSE) have been calculated via the following cross-validation procedure implemented for each estimated pigment:

- Step 1: The initial data set was randomly segmented into 20 different partitions: 95% of the data served as a training set and 5% as a test set.
- Step 2: At each iteration, the SOM was trained with the training set.
- Step 3: We applied the pigment retrieval procedure described in phase 2.
- Step 4: The pigment concentrations estimated from the retrieval procedure were compared to the observed test set.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{Obs}_i - \mathrm{Est}_i)^2}{n}} \tag{4}$$

3.2.4. Phase 4: Uncertainty Calculation and Quality Control

In order to compute the uncertainty on the estimated pigments, an ascending hierarchical clustering was applied to the SOM in order to cluster similar neighboring neurons into 500 Big-clusters (approximately 40 neurons per cluster). Then, for each Big-cluster j ($j = 1 \ldots 500$), a standard deviation vector ($\mathbf{Std_j}$) was calculated, whose components x_i ($i = 1 \ldots 16$) are the 16 parameters of $\mathbf{D_{pigment}}$. The standard deviation for component x_i belonging to cluster j is denoted $\mathbf{Std_j}$ (x_i) and is estimated from the values of the neurons

clustered within the Big-cluster *j*. As a result, each neuron *k* belonging to the Big-cluster *j* will be associated with its $\mathbf{Std_j}$ and will be denoted $\mathbf{Std_k}$

$$\mathrm{Std_k(xi)} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{xi}-\overline{\mathrm{xi}})^2}{n}} \tag{5}$$

where n is the number of neurons k in the Big-cluster j, xi is a component ($i = 1 \ldots 16$) of a neuron k, and $\overline{\mathrm{xi}}$ is the mean of the component x_i computed from the neurons in Big-Cluster j:

$$\overline{\mathrm{xi}}_j = \frac{1}{n}\sum_{i=1}^{n}\mathrm{xi} \tag{6}$$

Moreover, in the operational phase, a quality control was applied based on the difference between the Rrs(λ) spectrum projected on the SOM map and its corresponding selected neuron. A mask (M) is generated for each pixel p using the four Rrs (λ) as follows:

$$\Delta\mathrm{Rrs}(\lambda)p = \frac{\mathrm{Rrs}(\lambda)_p - \mathrm{Rrs}(\lambda)_k}{\mathrm{Std}_k(\mathrm{Rrs}(\lambda))} \tag{7}$$

where λ represents the wavelengths 412, 443, 490, and 555 nm; Rrs(λ) is the value of Rrs(λ) at wavelength λ for pixel p, $\mathrm{Rrs}(\lambda)_k$ is the value of Rrs(λ) at wavelength λ of the selected neuron k, and $\mathrm{Std}_k(\mathrm{Rrs}(\lambda))$ is the standard deviation for Rrs(λ) as described above. $\Delta\mathrm{Rrs}(\lambda)p$ is a function of $\mathrm{Std}_k(\mathrm{Rrs}(\lambda))$; it is the ratio of the difference between the Rrs(λ) of p and its associated neuron k and the $\mathrm{Std}_k(\mathrm{Rrs}(\lambda))$.

The mean of the four $\Delta\mathrm{Rrs}(\lambda)p$ is attributed to each pixel of the resulting image;

$$\overline{\Delta\mathrm{Rrs}(\lambda)}_p = \frac{\sum_{\lambda=412,443,490,555}^{4}\Delta\mathrm{Rrs}(\lambda)_p}{4} \tag{8}$$

The quality control mask (denoted M_p) was defined using a threshold at $\overline{\Delta\mathrm{Rrs}(\lambda)}_p = \pm 2$ STD. Therefore, values of $\overline{\Delta\mathrm{Rrs}(\lambda)}_p$ above this threshold are rejected. This allows to filter the pixels with abnormal Rrs(λ) spectrum that were not observed within the $\mathbf{D_{pigment}}$ used to train the SOM.

4. Results

4.1. Analysis of the SOM Organization and Topology

Once the training phase of the SOM is done, the relationship between the in situ pigments and those estimated from the satellite observations was assessed. A correlation analysis was performed between the parameters within the initial database and within the SOM. As a result, the correlations are efficiently preserved (Table 3). Therefore, the SOM is representative of the initial database $\mathbf{D_{pigment}}$, reproducing accurately the relationship between the parameters.

Besides, the 16 components affected to the neurons of the SOM seemed to be well organized; diagonal gradients are found for most of the parameters (Figure 2). Such results denote a coherent relationship between the variables.

The organization of the pigments on the SOM with respect to the satellite data helps understanding the link between the phytoplankton pigments and the satellite signal. The in situ and satellite Chla are closely related, which may explain the satisfying correlation between satellite and in situ Chla data ($r = 0.70$, pVal < 0.001). The Fuco follows a diagonal gradient on the SOM (top left, bottom right) and thus exhibits an organization close to that of both in situ and satellite Chla. As a consequence, there is a strong relationship between these two pigments (Chla in situ vs. Fuco: $r = 0.80$, Chla OC5 vs. Fuco: $r = 0.67$, pVal < 0.001). The Chlb and Perid mainly exhibit a vertical gradient (top to bottom), while the Allo is organized with high values simultaneously on the top and left sides of the SOM map. 19HF and 19BF are both covariant ($r = 0.77$, pVal < 0.001) and present a diagonal gradient (top right, bottom left), while both Chla and Fuco present a

Journal of Geophysical Research: Oceans

10.1029/2018JC014450

Table 3
Comparison Between the Correlation Matrix for the 16 Variables of the Initial Database (White), and the Correlation Matrix of the 16 Variables Within the SOM After the Training Phase (gray)

		SOM															
In situ data	Var	Chla	DVChla	Chlb	DVChlb	19HF	19BF	Fuco	Perid	Allo	Zea	Chla OC5	Rrs(412)	Rrs(443)	Rrs(490)	Rrs(555)	SST
	Chla		−0.14	0.52	−0.09	0.51	0.36	0.79	0.53	0.52	0.00	0.70	−0.47	−0.48	−0.30	0.33	−0.32
	DVChla	−0.18		−0.08	0.18	−0.09	0.01	−0.15	−0.09	−0.16	0.38	−0.17	0.12	0.09	0.07	−0.07	0.20
	Chlb	0.39	−0.06		0.02	0.39	0.38	0.37	0.34	0.39	0.02	0.41	−0.44	−0.44	−0.25	0.23	−0.31
	DVChlb	−0.08	0.23	0.09		−0.07	0.01	−0.09	−0.03	−0.07	0.13	−0.10	−0.01	−0.02	−0.03	−0.08	0.12
	19HF	0.55	−0.15	0.35	−0.04		0.78	0.37	0.40	0.33	0.01	0.31	−0.42	−0.45	−0.37	0.07	−0.36
	19BF	0.38	−0.05	0.37	0.02	0.77		0.22	0.33	0.27	−0.04	0.14	−0.37	−0.40	−0.35	−0.02	−0.39
	Fuco	0.81	−0.17	0.36	−0.07	0.41	0.29		0.52	0.47	−0.08	0.69	−0.44	−0.46	−0.31	0.28	−0.29
	Perid	0.54	−0.10	0.31	−0.02	0.45	0.38	0.48		0.35	−0.01	0.44	−0.38	−0.40	−0.31	0.14	−0.18
	Allo	0.52	−0.19	0.34	−0.04	0.40	0.24	0.44	0.31		−0.14	0.40	−0.36	−0.36	−0.24	0.16	−0.25
	Zea	−0.02	0.39	−0.03	0.15	0.06	0.02	−0.04	0.01	−0.13		−0.01	0.07	0.06	0.14	0.14	0.32
	Chla OC5	0.74	−0.19	0.34	−0.09	0.37	0.20	0.67	0.42	0.39	−0.01		−0.56	−0.55	−0.27	0.50	−0.25
	Rrs(412)	−0.48	0.10	−0.41	−0.01	−0.48	−0.43	−0.45	−0.38	−0.36	−0.01	−0.58		0.97	0.68	−0.14	0.44
	Rrs(443)	−0.49	0.09	−0.40	−0.02	−0.50	−0.44	−0.46	−0.40	−0.35	−0.02	−0.57	0.97		0.81	0.01	0.45
	Rrs(490)	−0.30	0.08	−0.19	0.00	−0.38	−0.35	−0.29	−0.28	−0.23	0.08	−0.30	0.68	0.81		0.51	0.39
	Rrs(555)	0.30	−0.07	0.20	−0.04	0.09	0.02	0.25	0.14	0.15	0.13	0.43	−0.12	0.03	0.51		0.02
	SST	−0.34	0.21	−0.26	0.10	−0.37	−0.42	−0.31	−0.21	−0.23	0.20	−0.28	0.42	0.43	0.34	−0.05	

Note. This comparison clearly indicates how self-organizing map (SOM) preserved the correlations between variables. 19BF, 19'ButFucoxanthin; 19HF, 19'HexFucoxanthin; Allo, alloxanthin; Chla, chlorophyll-a; Chlb, chlorophyll-b; DVChla, Divynil-Chla; DVChlb, Divynil-Chlb; Fuco, fucoxanthin; Perid, peridinin; SST, sea surface temperature; Zea, zeaxanthin.

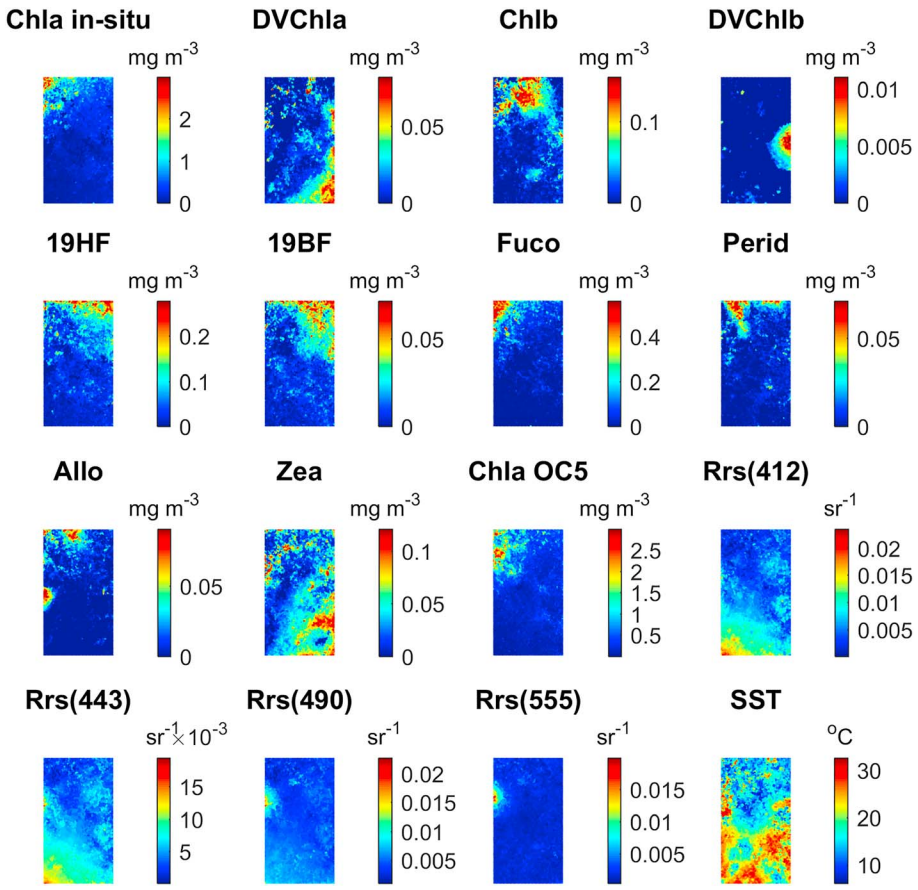

Figure 2. Organization of the 16 variables on the self-organizing map (SOM) map after the training phase. Each map represents the values recorded by the neurons of the SOM for the 16 variables. The topological organization of these neurons reflects the inter-variable relationship.

gradient following the other diagonal. DVChla and Zea present a diagonal gradient (bottom right, top left), and DVChlb is maximum in the middle of the right side of the map.

Most of the DPs follow a negative gradient of SST ($r < -0.21$, pVal < 0.001), while Zea, DVChlb, and DVChla covary with SST (high values of Zea, DVChlb, and DVChla coincide with high SST values; $r > 0.1$, pVal < 0.001). Noting that Zea is an indicator of Cyanobacteria communities and that both DVChlb and DVChla are DP of Prochlorococcus (Cyanobacteria), this link suggests a correlation between SST and Cyanobacteria spatio-temporal patterns. In a general view, the SOM topology seems to follow a certain regionalization; Chla, Chlb, 19HF, 19BF, Fuco, Perid, and Allo pigment components show high concentrations at the top of the SOM, whereas low concentrations are found at the bottom. In contrast, DVChla, DVChlb, and Zea display an opposite structuring. This organization promotes a subdivision into oligotrophic waters (bottom) where cyanobacteria dominate and more productive waters (top) where the other species are abundant.

4.2. Cross-Validation Results

Table 4 shows the cross-validation results of the 10 pigments for each experiment with an increasing number of neurons, and Figure 3 presents the scatterplots that compare the observed pigment concentrations versus the estimated values for the test sets using 20,000 neurons. The RMSE and the R^2 coefficients between the estimated and the observed pigments both improve when a higher number of neurons are specified; For example, the accuracy of the Fuco prediction shifts from an R2 of 0.5 to 0.78 when the SOM is increased from 5,000 to 20,000 neurons. Overall, when 20,000 neurons are prescribed, the R^2 coefficients for the different pigments range between 0.75 and 0.89 with an average RMSE of 0.012 mg/m^3. Furthermore, the inspection of the scatterplots (Figure 3) shows that there is no bias in the estimation procedure.

Table 4
Statistical Results of the Cross-Validation Result for the 10 Pigments

	N Neur = 20,000		N Neur = 10,000		N Neur = 5,000			
	R^2	RMSE (mg/m^3)	R^2	RMSE (mg/m^3)	R^2	RMSE (mg/m^3)	pVal	N Obs
Chla $_{SOM}$	0.84	0.22	0.79	0.23	0.55	0.24	0.001	4179
DVChla	0.77	0.01	0.56	0.02	0.48	0.02	0.001	2502
Chlb	0.85	0.01	0.69	0.02	0.64	0.02	0.001	2900
DVChlb	0.89	0.001	0.49	0.002	0.43	0.01	0.001	383
19HF	0.75	0.02	0.73	0.03	0.58	0.03	0.001	4284
19BF	0.79	0.01	0.68	0.01	0.61	0.01	0.001	4192
Fuco	0.87	0.02	0.76	0.02	0.50	0.02	0.001	4382
Perid	0.80	0.01	0.35	0.02	0.54	0.01	0.001	3113
Allo	0.76	0.01	0.50	0.02	0.41	0.01	0.001	2215
Zea	0.79	0.01	0.74	0.01	0.54	0.02	0.001	4262

Note. Non–log-transformed data were evaluated to calculate the root-mean-square error (RMSE).

The R^2 coefficient between the Chla estimated by the SOM ($Chla_{SOM}$) and the in situ Chla is equal to 0.84 with an RMSE of 0.24 mg/m^3. When predicting $Chla_{SOM}$, the SOM provides a tool to validate the spatial reconstruction by comparing the output with the satellite $Chla_{OC5}$ images.

This comparison performed between daily images shows a good agreement between the two Chla, scoring an R^2 of 0.85, and an RMSE of 0.13 mg/m^3 (Figure 4). Besides, a slight overestimation of 0.01 mg/m^3 by the $Chla_{SOM}$ compared to $Chla_{OC5}$ was highlighted for Chla values less than 1 mg/m^3. A saturation is observed at values near 3 mg/m^3, and it is mainly due to the limitation of the Chla values of the learning database, which were used to train the SOM. On the spatial scale, the Chla SOM reproduces the patterns initially observed in the original product, which proves that the reconstruction is efficient.

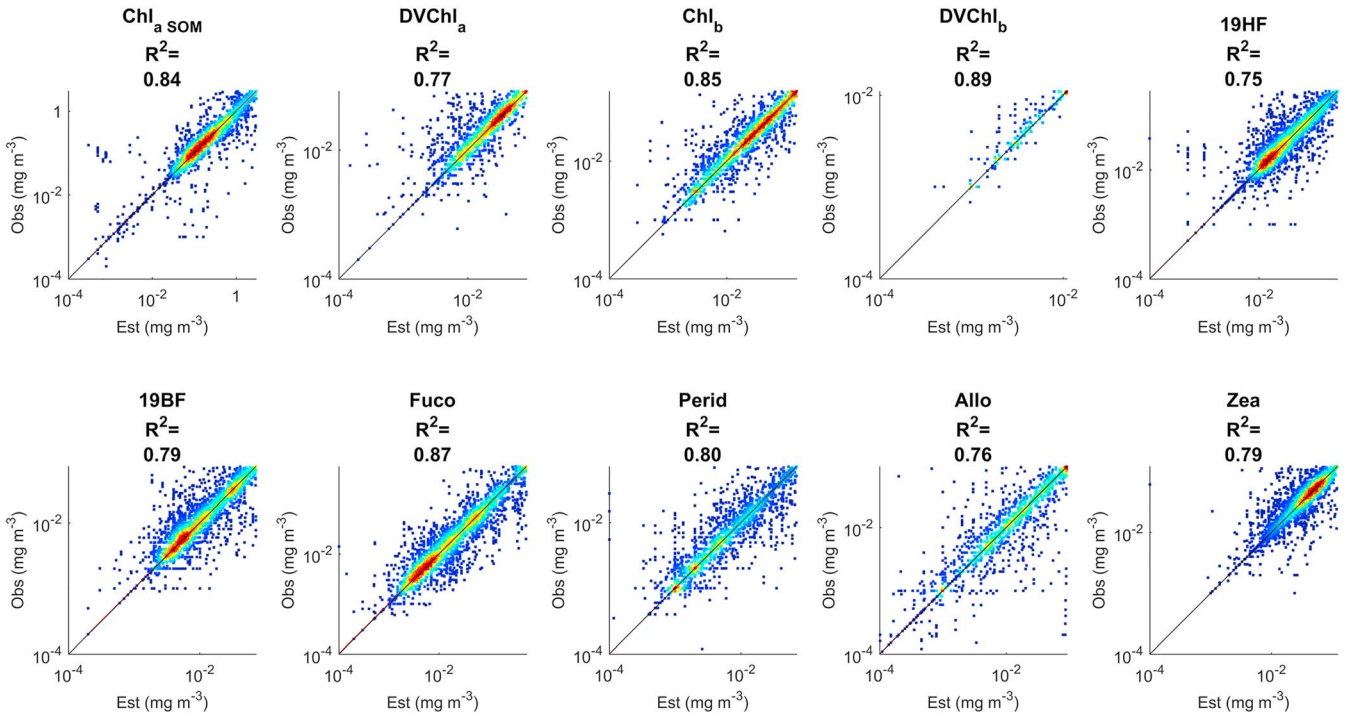

Figure 3. Scatter plots illustrating the cross-validation cumulative results for the 10 pigments, estimated versus observed. The R^2 above each scatter plot represents the regression coefficient.

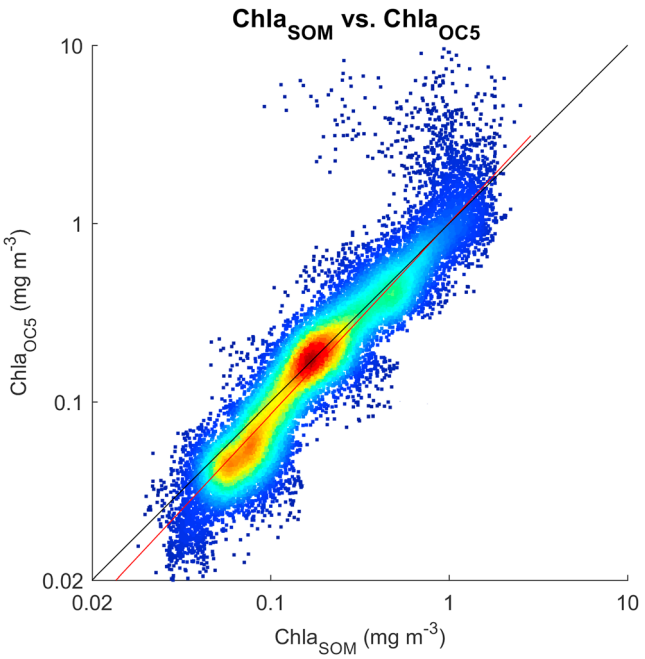

Figure 4. Comparison between $Chla_{SOM}$ and $Chla_{OCS}$ daily values showing a good agreement ($R^2 = 0.85$, root-mean-square error = 0.13 mg/m^3).

5. Estimation of Pigment Concentrations in the Global Ocean

The method described in section 3 has been used to generate daily images of pigments concentrations at global scale using satellite reflectance, Chla, and SST products presented in section 2 (Figures 5–7).

Furthermore, a quality control mask was determined from the mean difference $\overline{\Delta Rrs(\lambda)}_p$ between the satellite Rrs(λ) spectrum projected on the SOM map and the Rrs(λ) spectrum of the best matching neuron (see section 3, phase 4 for a detailed description; Figure 5b). To quantify the quality of the pigments concentrations estimated by our method, a 10-year climatology (2006–2016) of $\overline{\Delta Rrs(\lambda)}_p$ has been generated and is displayed in Figure 8.

Values of $\overline{\Delta Rrs(\lambda)}_p$ remain relatively low over most of the ocean both in winter and in summer. However, the Southern Ocean exhibits high values (above ±2 STD) in winter, which may even exceed ±4 STD in some areas. In addition to that, the southern subtropical gyres of the Pacific Ocean are also characterized by relatively high uncertainties that do not yet exceed drastically 2 STDs. A 2-year (2012–2013) daily analysis was performed in order to evaluate the frequency of pixels that are characterized by high values of $\overline{\Delta Rrs(\lambda)}_p$ during that period. The near shore pixels have a 60% frequency to be flagged mainly due to the fact that most case 2 water data were excluded from the initial database. In the open ocean, the frequency drops to less than 20% except in the Southern Ocean where this frequency is higher, reaching 45% in the circumpolar current. This quite large uncertainty in that region is explained by its specific optical characteristics as well as by the exclusion of the data collected in the Southern Ocean from the training database and the very limited availability of satellite observations (less than 50 observations per pixel through a 2-year period). Consequently, one may call into question the reliability of the SOM-pigments to predict accurately the pigment concentrations in the Southern Ocean during the winter season. Yet the 2-year frequency analysis shows that some

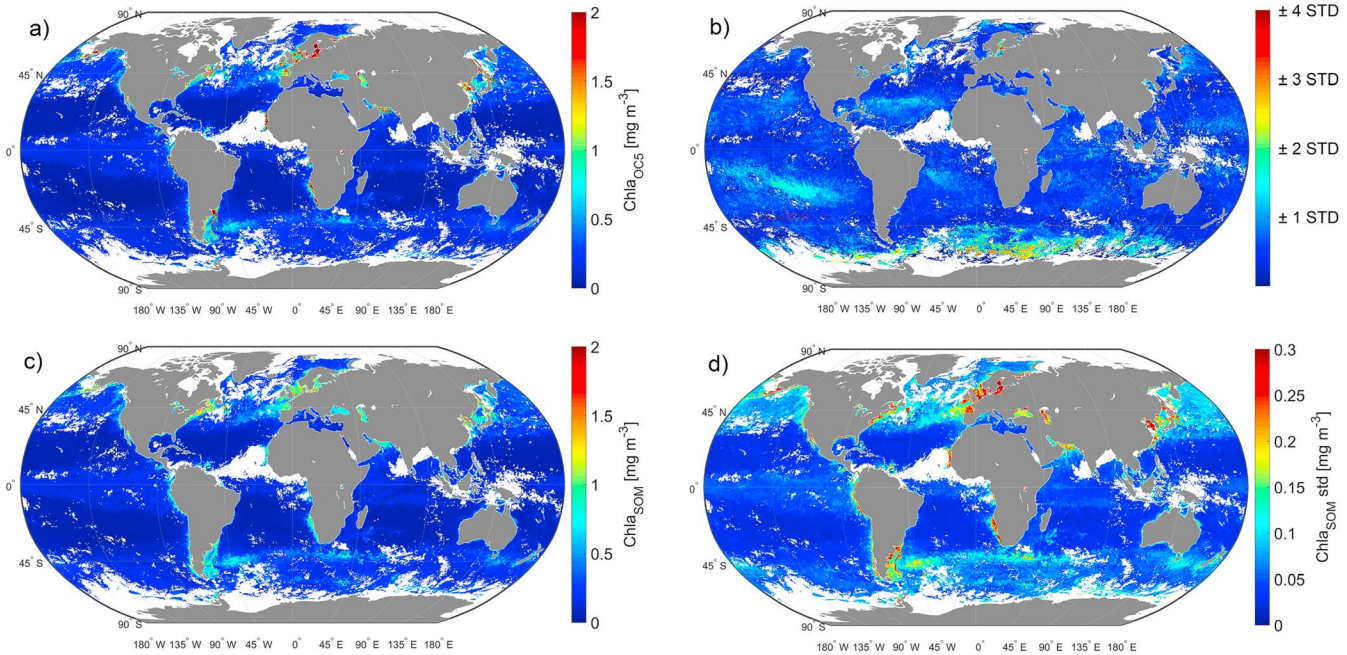

Figure 5. Eight-day composite (2–9 April 2017) of (a) Chla OC5, (b) $\overline{\Delta Rrs(\lambda)}_p$, (c) chlorophyll-a (Chla) self-organizing map (SOM), and (d) Chla SOM uncertainties.

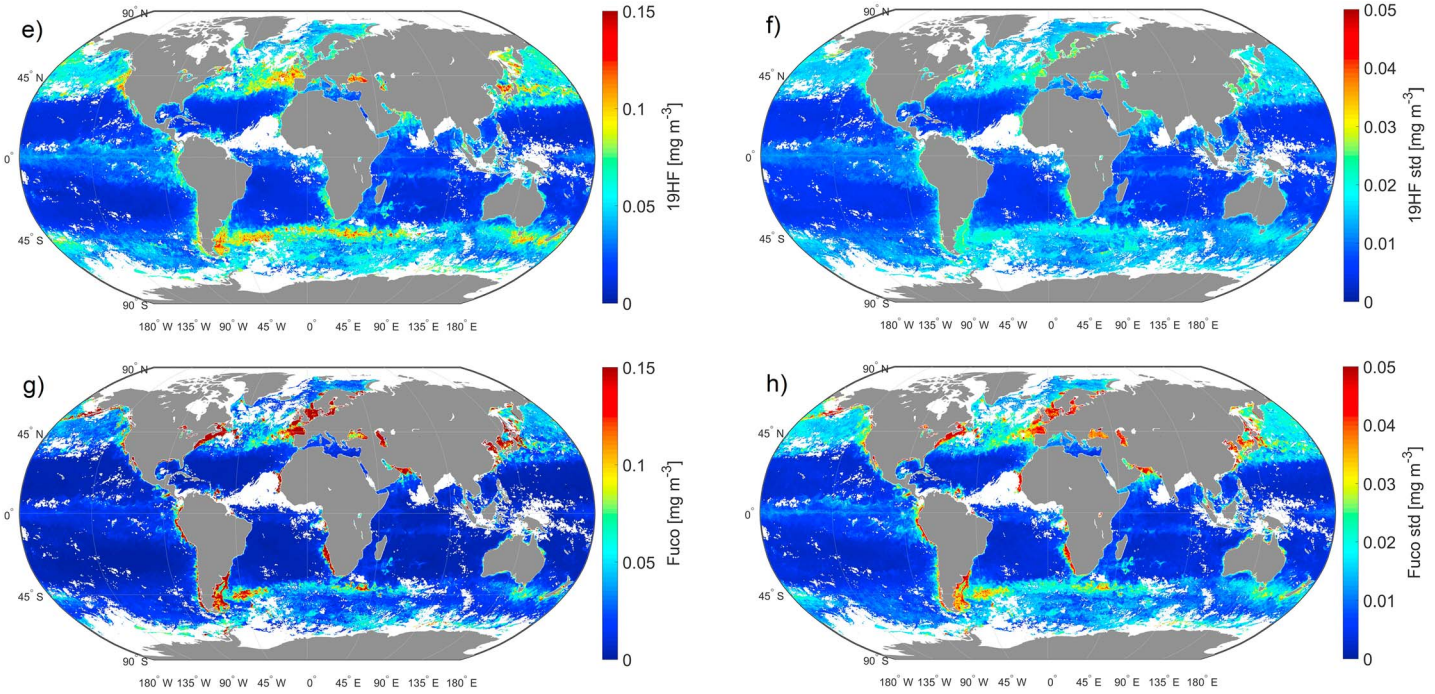

Figure 6. Eight-day composite (2–9 April 2017) of (e) 19HF, (f) 19HF uncertainties, (g) Fuco, and (h) Fuco uncertainties.

observations in the Southern Ocean fall under the threshold. They mainly correspond to situations that do exist in other regions of the ocean and thus that have been accounted for in the training phase. Therefore, in this case, the estimated pigments should be reliable. In the following, pixels with a ±2-STD deviation are flagged to insure a relevant representation while tracking stable patterns of six major pigments concentration (Chla, DVChla, Chlb, 19HF, Fuco, and Zea), in winter (December–February) and summer (July–

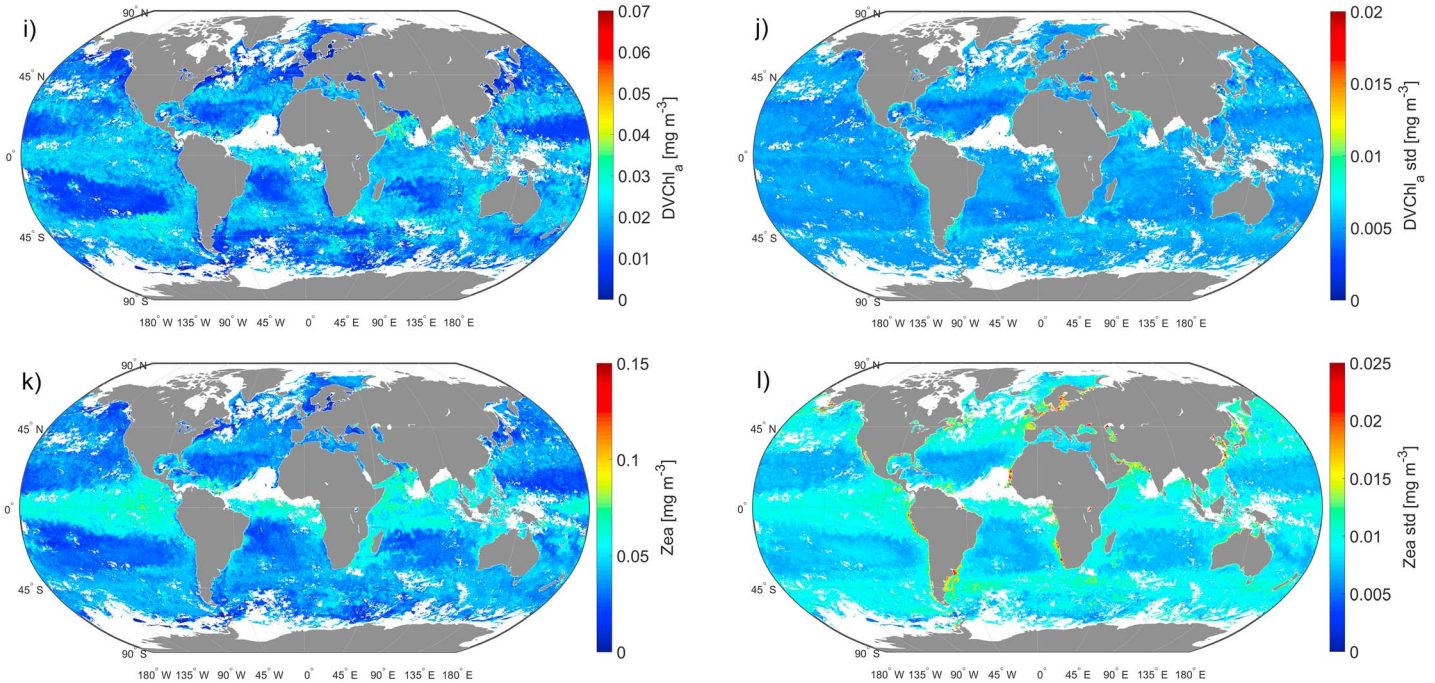

Figure 7. Eight-day composite (2–9 April 2017) of (i) DVChla, (j) DVChla uncertainties, (k) Zea, and (l) Zea uncertainties.

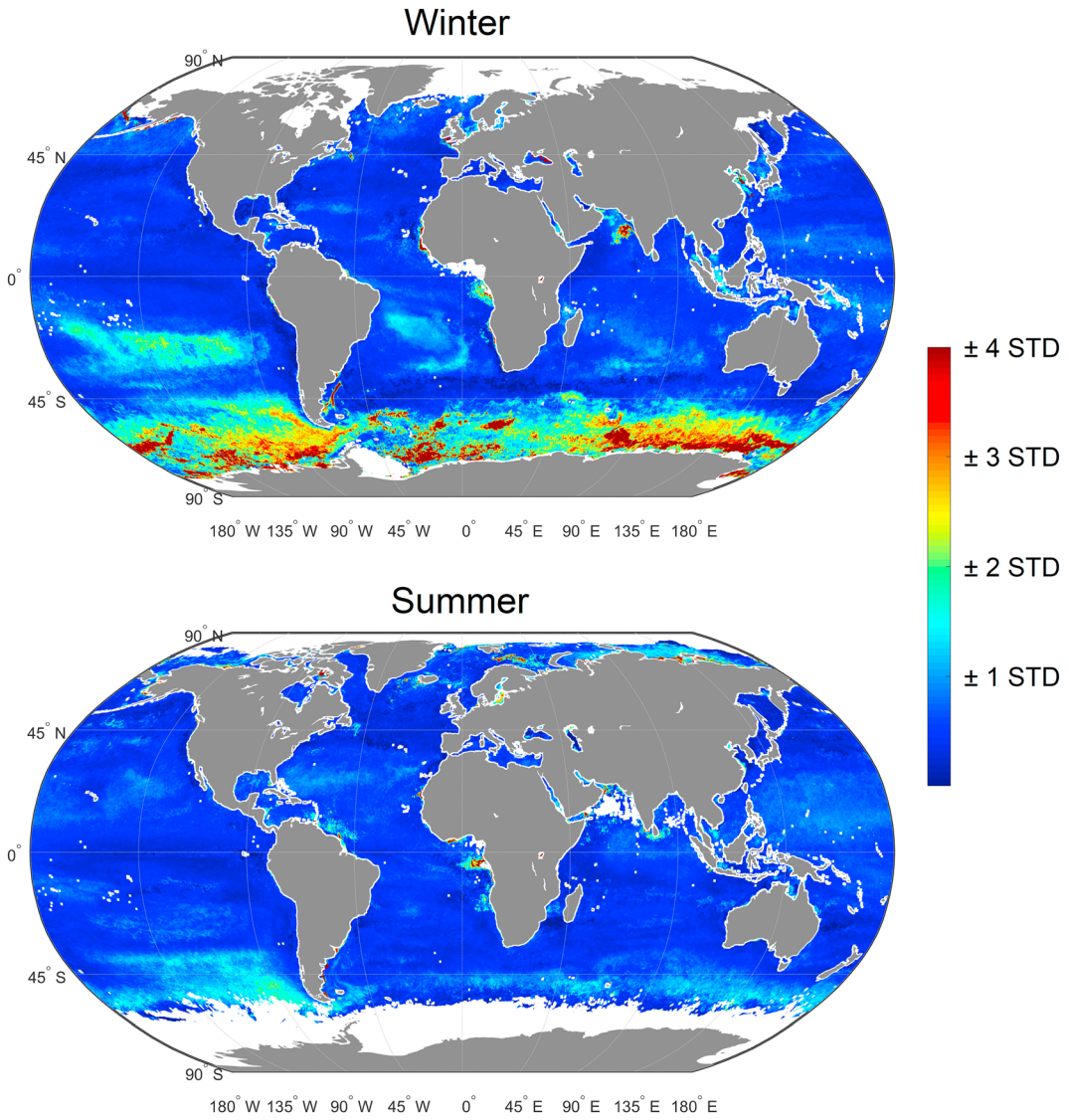

Figure 8. Seasonal climatology of the quality control mask as a function of STD.

September). Seasonal climatology composites were calculated and are represented in Figures 9 and 10 for the period ranging between 2006 and 2016.

The results show specific patterns for every pigment. We note that the dynamics of Fuco is similar to that of Chla: High concentrations are observed in the Antarctic Sea in the winter season, and a pronounced maximum of Fuco >0.2 ± 0.05 mg/m^3 is observed in the northern part of the Atlantic and of the Pacific in summer. High concentrations of Fuco >0.2 ± 0.05 mg/m^3 are also found in Eastern Boundary Upwelling Systems such as off Peru, California, and in the Benguela upwelling. This close relationship between Fuco and Chla was expected from the analysis of the SOM topology (see section 4.1).

19HF (Figure 10) presents a spatio-temporal variability with stable patterns above 45°N (winter) and 45°S (summer) at the level of the subpolar/temperate interface. High concentrations of >0.15 ± 0.05 mg/m^3 of 19HF and >0.05 ± 0.001 mg/m^3 of 19BF are recorded in winter above 45°S associated with the circumpolar current and in summer around 45°N.

Chlb shows a pronounced variability in the subpolar regions, north of 45°N and south of 45°S latitudes. In winter, the Antarctic Sea shows higher concentrations of Chlb; while in summer, higher concentrations are observed in the northern part of the Atlantic and the Pacific, reaching values above 0.12 ± 0.03 mg/m^3.

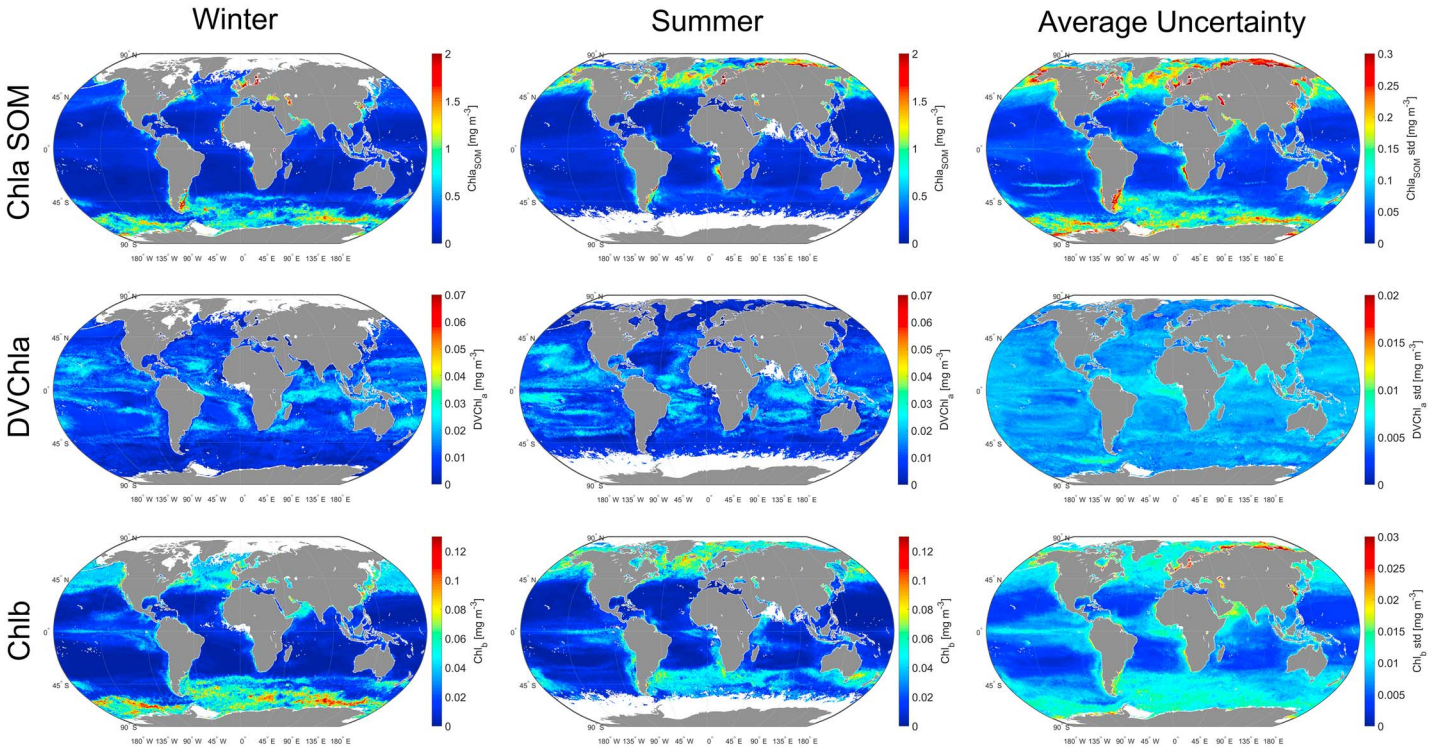

Figure 9. Seasonal climatology (winter-summer) generated between 2006 and 2016 of the estimated pigment concentrations via self-organizing map (SOM) and their corresponding uncertainties for Chla SOM (top line), DVChla (middle line), Chlb (bottom line).

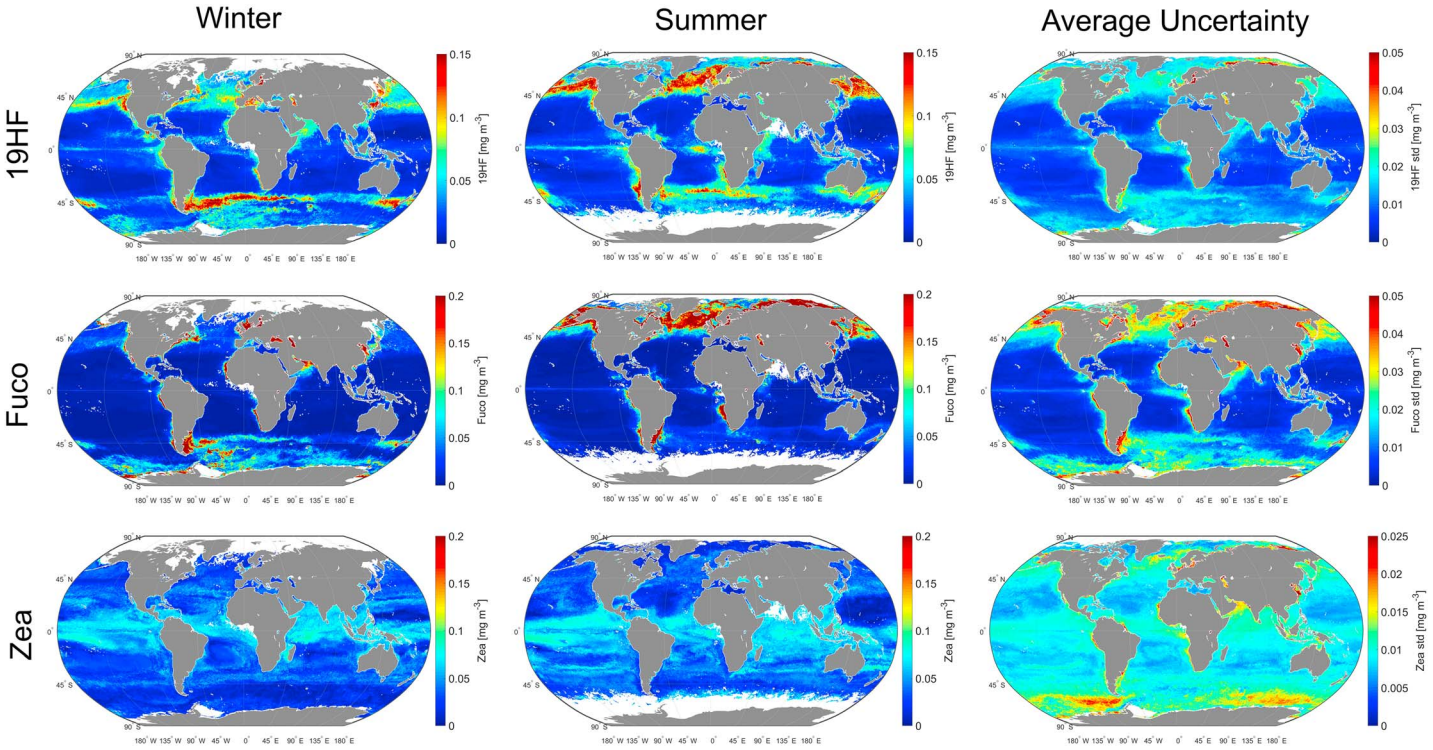

Figure 10. Seasonal climatology (winter-summer) generated between 2006 and 2016 of the estimated pigment concentrations via self-organizing map (SOM) and their corresponding uncertainties for 19HF (top line), Fuco (middle line), and Zea (bottom line).

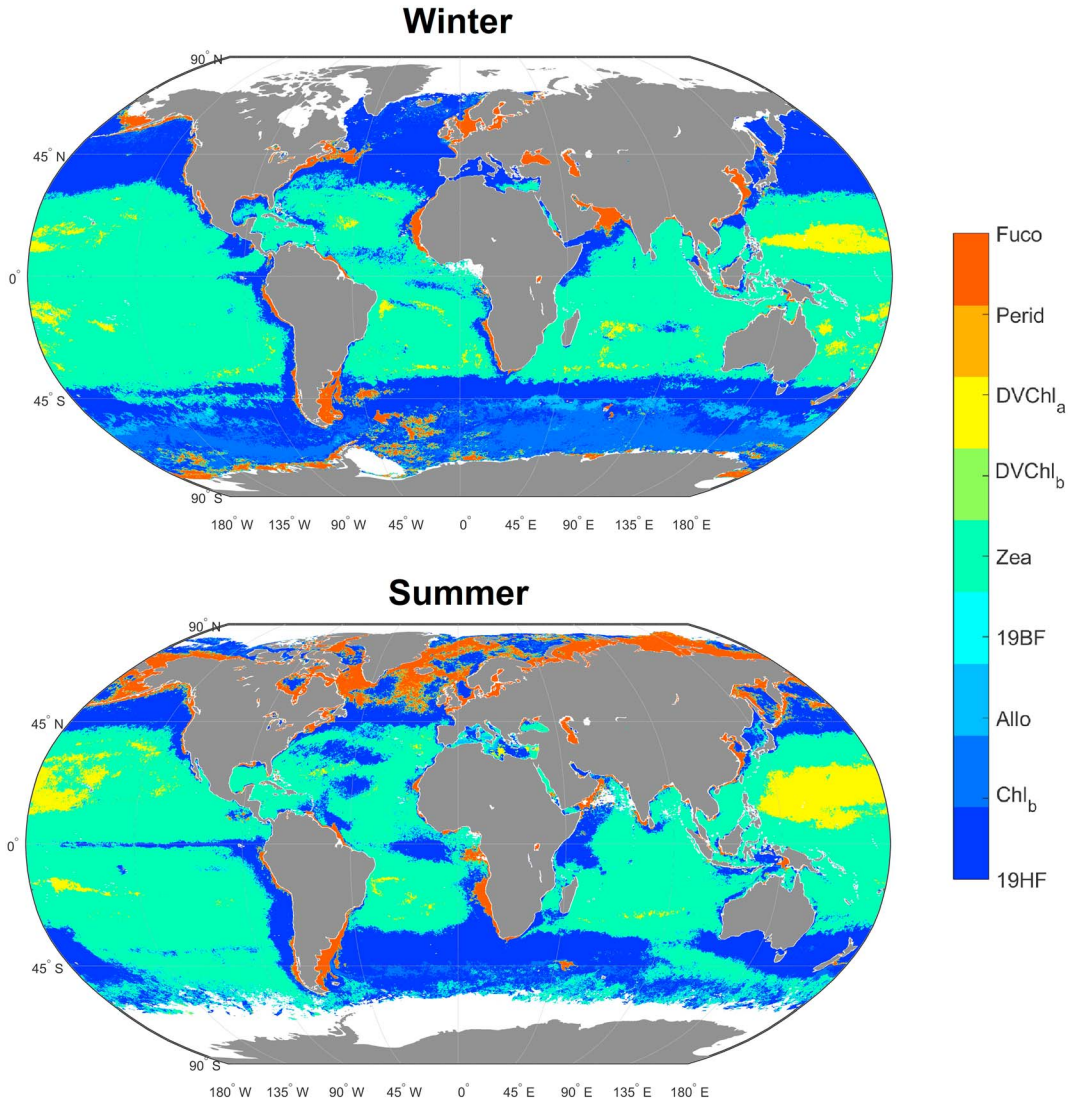

Figure 11. Global projection of the pigment with the maximum concentration at each pixel. The seasonal climotology was used to generate these MaxDP images.

Zea and DVChla pigments are restricted to the subtropical and tropical regions and both present minimal concentrations in the high latitudes throughout the year. Zea shows relatively constant concentrations around 0.1 ± 0.01 mg/m^3 between 30°N and 30°S. The largest concentrations are reached near the equatorial divergences. In contrast, DVChla is characterized by a more patchy distribution and tends to be maximum at about 0.04 ± 0.01 mg/m^3 along the boundaries of the subtropical gyres.

To further assess the seasonal patterns of these pigments, the pigment with the highest concentration per pixel is represented in Figure 11. This simple analysis is designed to estimate the spatio-temporal distribution of the dominant pigments at the global scale. However, this simple diagnostic should be cautiously interpreted since the pigment to chlorophyll or pigment to carbon ratios is highly variable. As a consequence, a dominant pigment does not necessarily mean that the associated species dominate the phytoplankton community.

First, the 19HF pigment seems to dominate in the northern area of the Pacific and the Atlantic and around the Circumpolar Current during both the winter and summer seasons. Meanwhile, Fuco is the most abundant pigment in coastal areas and in semienclosed and enclosed basins. During the winter season, this pigment dominates along the Antarctic coast and downstream of the islands in the Southern Ocean. In summer,

Fuco dominates in the northern Pacific and Atlantic coasts and around the arctic. Chlb maxima are observed mainly in the Southern Ocean in winter. Both DVChla and Zea are restricted to the tropical and subtropical regions and are almost absent at high latitudes. DVChla dominates in the subtropical gyres, while Zea is the most abundant pigment over large areas of the tropical domain, with the exception of the upwelling and coastal regions. Both pigments are characteristic of cyanobacteria, and their distributions are consistent with observations (Flombaum et al., 2013).

6. Discussion

The originality of our method is to model the relationship between satellite observations and phytoplankton pigments by partitioning a large database in a very large number of small clusters using the SOM. This efficient neural network clustering method amounts to model the multidimensional relationship between pigments and satellite observations by a piecewise continuous function. The clustering allows us to easily take into account the multifactorial aspect of the relationship and the different orders of magnitude of the parameters.

6.1. Comparison to Other Approaches Deriving Pigment Concentrations

Several attempts have been done to derive pigment concentrations from ocean color observations; Pan et al. (2010) developed Rrs(λ) band-ratio algorithms for retrieving pigment concentrations. These algorithms are represented by third-order polynomial functions using Rrs(λ) band ratio of either 490/550 nm or 490/670 nm (for SeaWiFS; for MODIS changed accordingly to MODIS bands 488 and 547 nm). They calibrated these functions by using satellite ocean color observations collocated with in situ pigment measurements. Validation of their results with collocated satellite (SeaWiFS and MODIS) reflectance data and pigment concentrations showed very accurate predictions for several pigments (Chla, Chlb, Perid, Fuco, Allo, and Zea). The RMSE ranged from 0.23 to 0.29, and the R^2 ranged from 0.65 to 0.90. This method was modified for the northern South China Sea using globally derived relationships and locally identified links between pigment concentrations and SST (Pan et al., 2013) . They achieved an accuracy similar to Pan et al. (2010). Compared to our SOM results, the quality of these estimations is similar. However, based on a larger data set, our SOM offers a robust tool to estimate pigment concentrations at the global scale.

Chase et al. (2013) derived concentrations of different chlorophyll types and several accessory pigments classified in two categories: PSC and PPC from a global data set of in situ hyperspectral particulate absorption measurements. This work was followed by another approach (Chase et al., 2017) in which they combined water leaving reflectance measurements and absorption signal to derive pigment concentrations. In both studies, they modeled the pigment absorptions by projecting the spectral signal on 12 Gaussian functions. In Chase et al. (2017), TChlb was estimated with an R^2 of 0.51 and PPC with an R^2 of 0.70 compared to HPLC measurements. Our SOM showed an improved retrieval of Chlb, DVChlb, Zea, and Allo with an R^2 of 0.85, 0.89, 0.79, and 0.76, respectively, noting that each pigment is estimated separately, not as a sum of PPC or PSC. This further indicates the robustness of our approach, characterizing the relationship between pigments and AOP (reflectance), as opposed to the IOPs used in their study. The estimation of IOP from AOP is based on an inversion model, which introduces additional uncertainties.

Bracher et al. (2015b) developed a method to assess pigment concentrations from continuous optical measurements. The method applied an empirical orthogonal function analysis to remote-sensing reflectance data derived from ship-based hyperspectral underwater radiometry combined with multispectral satellite data (using the MERIS Polymer product) measured in the Atlantic Ocean. Their results show a satisfactory prediction for several pigment concentrations from satellite data, with a R^2 of 0.25 for DVChla, 0.74 for 19BF, 0.68 for 19HF, 0.71 for Fuco, and 0.40 for Zea. Our prediction for these pigments is more accurate recording a R^2 of 0.77, 0.79, 0.75, 0.87, and 0.79, respectively.

Hirata et al. (2011) present a set of equations describing the phytoplankton size structures from Chla abundance in order to highlight the interpigment relationships; these authors used a global HPLC database of in situ secondary phytoplankton pigment concentrations to derive nonlinear relationships between phytoplankton size classes and Chla. These equations were also applied within the neurons of the SOM, and the results are shown in function of Chla in Figure 12. The underlying relationship between microphytoplankton, nanophytoplankton, picophytoplankton, and Chla derived from the SOM fits perfectly this

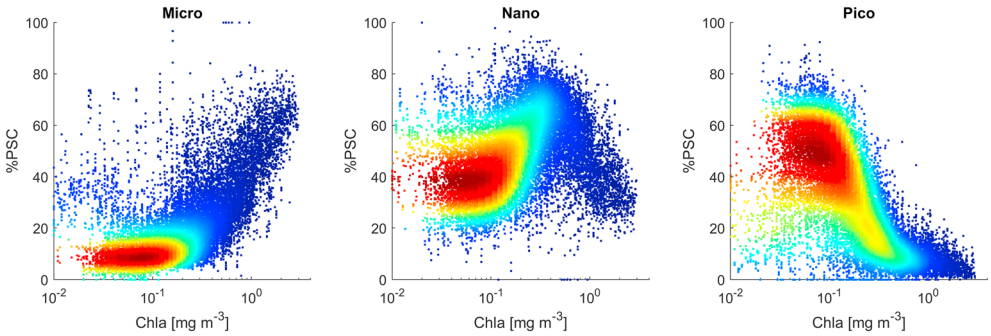

Figure 12. Interpigment relationships represented as percentage of phytoplankton size classes (microphytoplankton, nanophytoplankton, and picophytoplankton) in function of Chla, calculated within the neurons of the SOM-pigments. (%Micro = 1.41(NFuco + Perid)/ΣDP, %Nano = (Xn * 1.27Hex + 1.01Chlb + 0.35But + 0.6Allo)/ΣDP, %Pico = 1-(micro + nano); with ΣDP = 1.41Fuco + 1.41Perid + 1.27Hex + 1.01Chlb + 0.35But + 0.6Allo + 0.86Zea and Xn indicates a proportion of nanoplankton contribution in Hex (Hirata et al., 2011; Uitz et al., 2006).

described and analyzed by Hirata et al. (2011). For microphytoplankton, the fractional contribution to Chla monotonically increases with increasing Chla, whereas for picophytoplankton, this contribution monotonically decreases with increasing Chla while showing large variations. The fractional contribution of nanophytoplankton does not vary monotonically with Chla as found in microphytoplankton and picophytoplankton. Rather, the percentage of nanophytoplankton increases as Chla increases up to approximately 0.3 mg/m^3 but decreases as Chla further increases, resulting in a broad maximum between approximately 0.2–0.6 mg/m^3.

The comparison with other methods that have been proposed to retrieve the pigment concentrations from ocean color observations shows that the SOM method was able to model the nonlinear relationship between pigment concentrations and satellite-derived reflectance, SST, and Chla data and gave robust results. Meanwhile, the SOM was also able to reproduce the underlying interpigment relationships efficiently, while being consistent with the description of Hirata et al. (2011) in their analysis performed on in situ data.

6.2. Uncertainties and Quality Control

Besides the synoptic estimation of phytoplankton pigments, the SOM-Pigments allowed the quantification of uncertainties and to ensure a good quality control of the output product. The climatology revealed that most pigments have higher uncertainties with higher concentrations. The Fuco, 19HF, and Chlb uncertainties are important in eutrophic regions and near coastal zones. In contrast, DVChla and Zea were both characterized by higher uncertainties in oligotrophic areas and this is mainly due to the nonlinearity of the relationship between the pigments, the Rrs spectrum, and the phytoplankton community indicated behind.

Furthermore, upon the calculation of the quality control $\overline{\Delta \mathrm{Rrs}(\lambda)}_p$, patterns of high STD are recorded in the Southern Ocean during the blooms in winter, which indicates that predictions are highly uncertain in that region These high STDs can be linked to different biooptical properties, which have been suggested to occur in the Southern Ocean compared to other oceanic regions (Arrigo et al., 1998; Fenton et al., 1994; Mitchell et al., 1991). This led us to eliminate the Austral Ocean's data before the training phase. Furthermore, satellite observations are very limited in this region due to the sea ice coverage especially during the winter season. Yet once the abnormal data are flagged, The SOM procedure produces a coherent reconstruction of the pigment variability patterns, which are mainly modulated by a global relationship between pigment composition and the satellite signal, assessed by the SOM-Pigments.

6.3. Spatio-temporal Variability of Phytoplankton Pigments

The phytoplankton communities are unequally distributed throughout the ocean in response to several physical and biochemical factors. The latter stimulate some phytoplankton groups to prosper over others. Thus, the prediction of phytoplankton pigments via SOM-Pigments clearly represents how phytoplankton communities are clustered at the global scale. The objective of this section is not to discuss in details the retrieved phytoplankton groups' distribution but to illustrate some major features.

In upwelling and deep mixing regions, Fuco concentrations are elevated, followed by 19HF. In these regions characterized by a high input in nutrients, the variability of Fuco is largely explained by the opportunistic nature of diatoms. The latter species are the most responsive to sporadic changes in abiotic factors such as an increase in the nutrient load (Fogg, 1991) and tend to thrive in nutrient-rich and turbulent regions (Tréguer et al., 1995). Furthermore, in the Southern Ocean, high concentrations of Fuco are retrieved by our method. The latter result suggests that the island mass effect mainly benefits to diatoms as evidenced by in situ observations (Blain et al., 2008; Korb et al., 2008).

In temperate and subpolar environments, diatoms and nanophytoplankton are abundant, especially during the blooming season. As stratification increases with time and heating, nutrients become depleted in the surface layer, the grazer community develops, and the production of diatoms declines in favor of nanophytoplankton (Holligan et al., 1993; Iglesias-Rodríguez et al., 2002; Lochte et al., 1993). Therefore, in contrast with diatoms, nanophytoplankton remain abundant all year long, as evidenced by the dominance of nanophytoplankton's associated pigments (19HF, 19BF, Allo, and Chlb) in both winter and summer.

Conversely, as indicated by the quasi-stationary patterns of Zea and DVChla, smaller cells such as Cyanobacteria and other picophytoplankton species are favored by more constant environmental conditions mainly observed in the tropical and subtropical regions throughout the year. The equatorial Pacific is characterized by the presence of an upwelling driven by northeast and southeast trade winds. Upwelling of cool water results in a large supply of nutrients to the surface layer (Chavez, 1996; Chavez & Barber, 1987). Despite a constantly favorable light/nutrient regime, Chl biomass and primary production are relatively low considering the nutrient levels in that area (Barber & Chavez, 1991). This HNLC characteristics appears to be primarily due to iron limitation (Barber & Chavez, 1991; Coale et al., 1996) and, to a lesser extent, to regulation by silicate (Dugdale & Wilkerson, 1998) and grazing (Smetacek, 1999). These environmental conditions favor a phytoplankton community dominated by small cells (Chavez, 1996; Moon-van der Staay et al., 2000).

The results of the SOM can be compared to the output of the PHYSAT method (Alvain et al., 2005, 2008; Ben Mustapha et al., 2013). This method offers a tool to identify dominant phytoplankton functional type by exploiting satellite reflectance anomalies at several wavelengths. The global phytoplankton-type patterns provided by PHYSAT shows a qualitative agreement with our Max DP analysis. The global distribution of phytoplankton groups is characterized by the dominance of Synechococcus and Prochlorococcus groups in oligotrophic tropical waters, where Zea and DVChla are the most abundant pigments. Nanoeucaryotes and diatoms prevail in the eutrophic waters of high latitudes coinciding with higher 19HF/Fuco concentrations. In the PHYSAT climatologies, diatom blooms are clearly visible in areas characterized by strong upwelling conditions. Similarly, in our study, Fuco is the most abundant pigment in these regions. Therefore, we claim that our method offers the potential to investigate the spatial and temporal patterns of the phytoplankton community, in particular the variability of its dominant groups that compose this community.

7. Conclusion

We present robust estimations of the concentrations of various phytoplankton pigments by using SOMs learned on a set of satellite derived Chla, remote sensing reflectance data, surface temperature, and collocated pigment concentrations. In our study, it was shown that the SOM has efficiently modeled the relationship between the phytoplankton pigment concentrations and satellite data, which enables reliable estimation of the concentration of 10 different pigments (Chla, DVChla, Chlb, DVChlb, 19HF, 19BF, Fuco, Perid, Allo, and Zea). The method proves to be applicable for estimating concentrations of not only Chla but also of other pigments. Cross-validation results indicate that estimations were robust for all pigments ($R^2 > 0.75$ and an average RMSE = 0.016 mg/m^3). The large database used to develop and calibrate the SOM led to a satisfying estimation over a wide spatio-temporal scale. Therefore, a consistent picture of several phytoplankton pigments indicating group-specific behavior on a global scale was shown, revealing also the uncertainties associated with each pigment.

Besides long records of satellite data provided by Globcolour, the SOM-Pigments can be applied to study variability and change of overall phytoplankton and physiological responses to environmental variables by generating global images of pigment concentration with daily, weekly, monthly, and climatological

temporal resolution. This method can also be applied on broader areas to study the dynamic of phytoplankton at mesoscale, which involves phytoplankton pigments different than these used in this study. For that, further regional studies should be conducted to evaluate the robustness of the SOM and its capacity of reconstructing the spatio-temporal variability of the phytoplankton dynamic.

Acknowledgments
This work was supported by the National Council for Scientific Research-Lebanon, in the frame of a doctoral fellowship, codirected between the National Center for Remote sensing-CNRSL, Lebanon; the Sorbonne Université, Faculty of Sciences-Paris VI; and the LOCEAN, France. This work was also supported by the National Center for Space Studies (CNES), France, under the TOSKA project (2017-2018). The different merged satellite ocean color data were obtained from the GlobColour project portal (www.globcolour.info). AVHRR Pathfinder Level 3 Daily Daytime SST Version 5.3 data set were obtained from (http://doi:10.7289/V52J68XX/). We acknowledge the different sources of the HPLC pigment data set: MAREDAT, POLERSTERN data, Labrador Sea expeditions data, and Tara Oceans Expedition data, all available on https://pangaea.de/, GeP&Co database accessed at http://www.obs-vlfr.fr/proof/vt/op/ec/gep_co/gep.htm), and finally the NOMAD: NASA bio-Optical Marine Algorithm Dataset, and the numerous campaigns found on the NASA SeaBASS portal were accessed at (https://seabass.gsfc.nasa.gov/). Following best practices, the SOM-pigments was deposited into a public domain repository accessible at https://github.com/RoyElHourany/SOM-Pigments. Prerequisite software library SOM Toolbox 2.0 for Matlab is required, implementing the self-organizing map algorithm, Copyright (C) 1999 by Esa Alhoniemi, Johan Himberg, Jukka Parviainen, and Juha Vesanto and accessible at https://github.com/ilarinieminen/SOM-Toolbox.

References

Ainsworth, E. J., & Jones, I. S. F. (1999). Radiance spectra classification from the ocean color and temperature scanner on ADEOS. *IEEE Transactions on Geoscience and Remote Sensing*, *37*(3), 1645–1656. https://doi.org/10.1109/36.763281

Alvain, S., Moulin, C., Dandonneau, Y., & Bréon, F. M. (2005). Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery. *Deep-Sea Research Part I: Oceanographic Research Papers*, *52*(11), 1989–2004. https://doi.org/10.1016/j.dsr.2005.06.015

Alvain, S., Moulin, C., Dandonneau, Y., & Loisel, H. (2008). Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochemical Cycles*, *22*, GB3001. https://doi.org/10.1029/2007GB003154

Antoine, D., André, J.-M., & Morel, A. (1996). Oceanic primary production: 2. Estimation at global scale from satellite (coastal zone color scanner) chlorophyll. *Global Biogeochemical Cycles*, *10*, PA3213. https://doi.org/10.1029/95GB02832

Arrigo, K. R., Worthen, D., Schnell, A., & Lizotte, M. P. (1998). Primary production in Southern Ocean waters. *Journal of Geophysical Research*, *103*(C8), 15,587–15,600. https://doi.org/10.1029/98JC00930

Barber, R. T., & Chavez, F. P. (1991). Regulation of primary productivity rate in the equatorial Pacific. *Limnology and Oceanography*, *36*(8), 1803–1815. https://doi.org/10.4319/lo.1991.36.8.1803

Baumert, H. Z., & Petzoldt, T. (2008). The role of temperature, cellular quota and nutrient concentrations for photosynthesis, growth and light–dark acclimation in phytoplankton. *Limnologica - ecology and Management of Inland Waters*, *38*(3–4), 313–326. https://doi.org/10.1016/J.LIMNO.2008.06.002

Behrenfeld, M. J., & Boss, E. (2006). Beam attenuation and chlorophyll concentration as alternative optical indices of phytoplankton biomass. *Journal of Marine Research*, *64*(3), 431–451. https://doi.org/10.1357/002224006778189563

Behrenfeld, M. J., Boss, E., Siegel, D. A., & Shea, D. M. (2005). Carbon-Based Ocean productivity and phytoplankton physiology from space. *Global Biogeochemical Cycles*, *19*, GB100610. https://doi.org/10.1029/2004GB002299

Behrenfeld, M. J., & Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography*, *42*(1), 1–20. https://doi.org/10.4319/lo.1997.42.1.0001

Belo Couto, A., Brotas, V., Mélin, F., Groom, S., & Sathyendranath, S. (2016). Inter-comparison of OC-CCI chlorophyll-aestimates with precursor data sets. *International Journal of Remote Sensing*, *37*(18), 4337–4355. https://doi.org/10.1080/01431161.2016.1209313

Ben Mustapha, Z., Alvain, S., Jamet, C., Loisel, H., & Dessailly, D. (2013). Automatic classification of water-leaving radiance anomalies from global SeaWiFS imagery: Application to the detection of phytoplankton groups in open ocean waters. *Remote Sensing of Environment*, *146*, 97–112. https://doi.org/10.1016/j.rse.2013.08.046

Bhadury, P. (2015). Effects of ocean acidification on marine invertebrates—A review. *Indian Journal of Geo-Marine Sciences*, *44*(4), 454–464. Retrieved from http://nopr.niscair.res.in/bitstream/123456789/34717/1/IJMS. pdf (Accessed: 6 June 2018).

Blain, S., Quéguiner, B., & Trull, T. (2008). The natural iron fertilization experiment KEOPS (KErguelen Ocean and Plateau compared Study): An overview. *Deep Sea Research Part II: Topical Studies in Oceanography*, *55*(5–7), 559–565. https://doi.org/10.1016/j.dsr2.2008.01.002

Bracher, A., Bouman, H. A., Brewin, R. J. W., Bricaud, A., Brotas, V., Ciotti, A. M., et al. (2017). Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development. *Frontiers in Marine Science*, *4*, 1–15. https://doi.org/10.3389/fmars.2017.00055

Bracher, A., Taylor, M. H., Taylor, B., Dinter, T., Röttgers, R., & Steinmetz, F. (2015a) Phytoplankton pigments, hyperspectral downwelling irradiance and remote sensing reflectance during POLARSTERN cruises ANT-XXIII/1, ANT-XXIV/1, ANT-XXIV/4, ANT-XXVI/4, and Maria S. Merian cruise MSM18/3., PANGAEA. https://doi.org/10.1594/PANGAEA.847820

Bracher, A., Taylor, M. H., Taylor, B., Dinter, T., Röttgers, R., & Steinmetz, F. (2015b). Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations. *Ocean Science*, *11*(1), 139–158. https://doi.org/10.5194/os-11-139-2015

Casey, K. S., Brandon, T. B., Cornillon, P., & Evans, R. (2010). *The past, present, and future of the AVHRR Pathfinder SST program, in Oceanography from Space* (pp. 273–287). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-8681-5_16

Cavazos, T., & Cavazos, T. (1999). Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. *Journal of Climate*, *12*(5), 1506–1523. https://doi.org/10.1175/1520-0442(1999)012<1506:LSCACT>2.0.CO;2

Chase, A., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., Dall'Olmo, G., et al. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography*, *7*(2013), 110–124. https://doi.org/10.1016/j.mio.2014.02.002

Chase, A. P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of phytoplankton accessory pigments from hyperspectral reflectance spectra: Toward a global algorithm. *Journal of Geophysical Research: Oceans*, *122*, 9725–9743. https://doi.org/10.1002/2017JC012859

Chavez, F. P. (1996). Forcing and biological impact of onset of the 1992 El Niño in Central California. *Geophysical Research Letters*, *23*(3), 265–268. https://doi.org/10.1029/96GL00017

Chavez, F. P., & Barber, R. T. (1987). An estimate of new production in the equatorial Pacific. *Deep Sea Research Part A. Oceanographic Research Papers*, *34*(7), 1229–1243. https://doi.org/10.1016/0198-0149(87)90073-2

Chazottes, A., Bricaud, A., Crépon, M., & Thiria, S. (2006). Statistical analysis of a database of absorption spectra of phytoplankton and pigment concentrations using self-organizing maps. *Applied Optics*, *45*(31), 8102. https://doi.org/10.1364/AO.45.008102

Chazottes, A., Crépon, M., Bricaud, A., Ras, J., & Thiria, S. (2007). Statistical analysis of absorption spectra of phytoplankton and of pigment concentrations observed during three POMME cruises using a neural network clustering method. *Applied Optics*, *46*(18), 3790–3799. https://doi.org/10.1364/AO.46.003790

Chisholm, S. W. (1995). The iron hypothesis: Basic research meets environmental policy. *Reviews of Geophysics*, *33*(S2), 1277–1286. https://doi.org/10.1029/95RG00743

Coale, K. H., Johnson, K. S., Fitzwater, S. E., Gordon, R. M., Tanner, S., Chavez, F. P., et al. (1996). A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean. *Nature Publishing Group*, *383*(6600), 495–501. https://doi.org/10.1038/383495a0

Dandonneau, Y., Deschamps, P.-Y., Nicolas, J.-M., Loisel, H., Blanchot, J., Montel, Y., et al. (2004). Seasonal and interannual variability of ocean color and composition of phytoplankton communities in the North Atlantic, equatorial Pacific and South Pacific. *Deep Sea Research Part II: Topical Studies in Oceanography*, *51*(1–3), 303–318. https://doi.org/10.1016/j.dsr2.2003.07.018

di Cicco, A., Sammartino, M., Marullo, S., & Santoleri, R. (2017). Regional empirical algorithms for an improved identification of phytoplankton functional types and size classes in the Mediterranean Sea using satellite data. *Frontiers in Marine Science*, *4*, 126. https://doi.org/10.3389/fmars.2017.00126

Diouf, D., Niang, A., Brajard, J., Crepon, M., & Thiria, S. (2013). Retrieving aerosol characteristics and sea-surface chlorophyll from satellite ocean color multi-spectral sensors using a neural-variational method. *Remote Sensing of Environment*, *130*, 74–86. https://doi.org/10.1016/j.rse.2012.11.002

Dugdale, R. C., & Wilkerson, F. P. (1998). Silicate regulation of new production in the equatorial Pacific upwelling. *Nature*, *391*(6664), 270–273. https://doi.org/10.1038/34630

D'Ortenzio, F., & d'Alcalà, M. R. (2009). On the trophic regimes of the Mediterranean Sea: A satellite analysis. *Biogeosciences*, *6*(2), 139–148. https://doi.org/10.5194/bg-6-139-2009

Ehsani, A. H., & Quiel, F. (2008). Geomorphometric feature analysis using morphometric parameterization and artificial neural networks. *Geomorphology*, *99*(1–4), 1–12. https://doi.org/10.1016/j.geomorph.2007.10.002

Emery, W. J., Yu, Y., Wick, G. A., Schluessel, P., & Reynolds, R. W. (1994). Correcting infrared satellite estimates of sea surface temperature for atmospheric water vapor attenuation. *Journal of Geophysical Research*, *99*(C3), 5219. https://doi.org/10.1029/93JC03215

Fenton, N., Priddle, J., & Tett, P. (1994). Regional variations in bio-optical properties of the surface waters in the Southern Ocean. *Antarctic Science*, *6*(04), 443–448. https://doi.org/10.1017/S0954102094000684

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincon, J., Zabala, L. L., Jiao, N., et al. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences*, *110*(24), 9824–9829. https://doi.org/10.1073/pnas.1307701110

Fogg, G. E. (1991). The phytoplanktonic ways of life. *New Phytologist*, *118*(2), 191–232. https://doi.org/10.1111/j.1469-8137.1991.tb00974.x

Fragoso, G. M., Poulton, A. J., Yashayaev, I. M., Head, E. J. H., & Purdie, D. A. (2016). Spring phytoplankton communities of the Labrador Sea (2005–2014): pigment signatures, photophysiology and elemental ratios. *Biogeosciences Discussions*, *14*, 1–43. https://doi.org/10.5194/bg-2016-295

Gieskes, W. W. C., & Kraay, G. W. (1983). Dominance of Cryptophyceae during the phytoplankton spring bloom in the central North Sea detected by HPLC analysis of pigments. *Marine Biology*, *75*(2–3), 179–185. https://doi.org/10.1007/BF00406000

Gohin, F. (2011). Annual cycles of chlorophyll-a, non-algal suspended particulate matter, and turbidity observed from space and in-situ in coastal waters. *Ocean Science*, *7*(5), 705–732. https://doi.org/10.5194/os-7-705-2011

Gorricha, J., & Lobo, V. (2012). Improvements on the visualization of clusters in geo-referenced data using self-organizing maps. *Computers and Geosciences*, *43*, 177–186. https://doi.org/10.1016/j.cageo.2011.10.008

Guillard, R. R. L., Murphy, L. S., Foss, P., & Liaaen-Jensen, S. (1985). Synechococcus spp. as likely zeaxanthin-dominant ultraphytoplankton in the North Atlantic1. *Limnology and Oceanography*, *30*(2), 412–414. https://doi.org/10.4319/lo.1985.30.2.0412

Häder, D.-P., Villafañe, V. E., & Helbling, E. W. (2014). Productivity of aquatic primary producers under global climate change. *Photochemical & Photobiological Sciences*, *13*(10), 1370–1392. https://doi.org/10.1039/c3pp50418b

Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., et al. (2011). Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences*, *8*(2), 311–327. https://doi.org/10.5194/bg-8-311-2011

Holligan, P. M., Groom, S. B., & Harbour, D. S. (1993). What controls the distribution of the coccolithophore, Emiliania huxleyi, in the North Sea? *Fisheries Oceanography*, *2*(3–4), 175–183. https://doi.org/10.1111/j.1365-2419.1993.tb00133.x

Hu, X., & Weng, Q. (2009). Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks. *Remote Sensing of Environment*, *113*(10), 2089–2102. https://doi.org/10.1016/J.RSE.2009.05.014

Hülse, D., Arndt, S., Wilson, J. D., Munhoven, G., & Ridgwell, A. (2017). Understanding the causes and consequences of past marine carbon cycling variability through models. *Earth-Science Reviews*, *171*, 349–382. https://doi.org/10.1016/J.EARSCIREV.2017.06.004

Iglesias-Rodríguez, M. D., Brown, C. W., Doney, S. C., Kleypas, J., Kolber, D., Kolber, Z., et al. (2002). Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids. *Global Biogeochemical Cycles*, *16*(4), 1100. https://doi.org/10.1029/2001GB001454

Iskandar, I. (2010). Variability of satellite-observed sea surface height in the tropical Indian Ocean: Comparison of EOF and SOM analysis. *MAKARA of Science Series*, *13*(2). https://doi.org/10.7454/mss.v13i2.421

Jeffrey, S. W. (1980). Algal pigment systems. In *Primary productivity in the sea* (pp. 33–58). Boston, MA: Springer. https://doi.org/10.1007/978-1-4684-3890-1_3

Jeffrey, S. W., & Hallegraeff, G. M. (1987). Chlorophyllase distribution in ten classes of phytoplankton: a problem for chlorophyll analysis. *Marine Ecology Progress Series*, *35*, 293–304. https://doi.org/10.3354/meps035293

Jouini, M., Lévy, M., Crépon, M., & Thiria, S. (2013). Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method. *Remote Sensing of Environment*, *131*, 232–246. https://doi.org/10.1016/J.RSE.2012.11.025

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. https://doi.org/10.1016/J.NEUNET.2012.09.018

Korb, R., Whitehouse, M., Atkinson, A., & Thorpe, S. (2008). Magnitude and maintenance of the phytoplankton bloom at South Georgia: A naturally iron-replete environment. *Marine Ecology Progress Series*, *368*, 75–91. https://doi.org/10.3354/meps07525

Kostadinov, T. S., Cabré, A., Vedantham, H., Marinov, I., Bracher, A., Brewin, R. J. W., et al. (2017). Inter-comparison of phytoplankton functional type phenology metrics derived from ocean color algorithms and Earth System Models. *Remote Sensing of Environment*, *190*, 162–177. https://doi.org/10.1016/J.RSE.2016.11.014

Letelier, R. M., Bidigare, R. R., Hebel, D. V., Ondrusek, M., Winn, C. D., & Karl, D. M. (1993). Temporal variability of phytoplankton community structure based on pigment analysis. *Limnology and Oceanography*, *38*(7), 1420–1437. https://doi.org/10.4319/lo.1993.38.7.1420

Liu, Y. (2005). Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research*, *110*(C6), C06003. https://doi.org/10.1029/2004JC002786

Liu, Y., Weisberg, R. H., & Mooers, C. N. K. (2006). Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research*, *111*, C05018. https://doi.org/10.1029/2005JC003117

Lochte, K., Ducklow, H. W., Fasham, M. J. R., & Stienen, C. (1993). Plankton succession and carbon cycling at 47°N 20°W during the JGOFS North Atlantic Bloom Experiment. *Deep Sea Research Part II: Topical Studies in Oceanography*, *40*(1–2), 91–114. https://doi.org/10.1016/0967-0645(93)90008-B

Longhurst, A., Sathyendranath, S., Platt, T., & Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research*, *17*(6), 1245–1271. https://doi.org/10.1093/plankt/17.6.1245

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., et al. (2012). Database of diazotrophs in global ocean: Abundance, biomass and nitrogen fixation rates. *Earth System Science Data*, *4*(1), 47–73. https://doi.org/10.5194/essd-4-47-2012

Mann, K. H., Kenneth, H., & Lazier, J. R. N. (2006) Dynamics of marine ecosystems: biological-physical interactions in the oceans. Blackwell Pub. Retrieved from https://www.wiley.com/en-us/Dynamics+of+Marine+Ecosystems%3A+Biological+Physical+Interactions+in+the+Oceans%2C+3rd+Edition-p-9781405111188 (Accessed: 3 May 2018).

Marty, J.-C., Chiavérini, J., Pizay, M.-D., & Avril, B. (2002). Seasonal and interannual dynamics of nutrients and phytoplankton pigments in the western Mediterranean Sea at the DYFAMED time-series station (1991–1999). *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(11), 1965–1985. https://doi.org/10.1016/S0967-0645(02)00022-X

Mayot, N., D'Ortenzio, F., Uitz, J., Gentili, B., Ras, J., Vellucci, V., et al. (2017). Influence of the phytoplankton community structure on the spring and annual primary production in the northwestern Mediterranean Sea. *Journal of Geophysical Research: Oceans*, *122*, 9918–9936. https://doi.org/10.1002/2016JC012668

McClain, E. P., Pichel, W. G., & Walton, C. C. (1985). Comparative performance of AVHRR-based multichannel sea surface temperatures. *Journal of Geophysical Research*, *90*(C6), 11587. https://doi.org/10.1029/JC090iC06p11587

Mitchell, B. G., Brody, E. A., Holm-Hansen, O., McClain, C., & Bishop, J. (1991). Light limitation of phytoplankton biomass and macronutrient utilization in the Southern Ocean. *Limnology and Oceanography*, *36*(8), 1662–1677. https://doi.org/10.4319/lo.1991.36.8.1662

Moon-van der Staay, S. Y., van der Staay, G. W. M., Guillou, L., Vaulot, D., Claustre, H., & Medlin, L. K. (2000). Abundance and diversity of prymnesiophytes in the picoplankton coμmunity from the equatorial Pacific Ocean inferred from 18S rDNA sequences. *Limnology and Oceanography*, *45*(1), 98–109. https://doi.org/10.4319/lo.2000.45.1.0098

Morel, A., & Gentili, B. (1996). Diffuse reflectance of oceanic waters III: Implication of bidirectionality for the remote-sensing problem. *Applied Optics*, *35*(24), 4850–4862. https://doi.org/10.1364/AO.35.004850

Niang, A., Badran, F., Moulin, C., Crépon, M., & Thiria, S. (2006). Retrieval of aerosol type and optical thickness over the Mediterranean from SeaWiFS images using an automatic neural classification method. *Remote Sensing of Environment*, *100*(1), 82–94. https://doi.org/10.1016/J.RSE.2005.10.005

Organelli, E., Bricaud, A., Antoine, D., & Uitz, J. (2013). Multivariate approach for the retrieval of phytoplankton size structure from measured light absorption spectra in the Mediterranean Sea (BOUSSOLE site). *Applied Optics*, *52*(11), 2257–2273. https://doi.org/10.1364/AO.52.002257

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., et al. (2005). Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature Publishing Group*, *437*(7059), 681–686. https://doi.org/10.1038/nature04095

Pan, X., Mannino, A., Russ, M. E., Hooker, S. B., & Harding, L. W. Jr. (2010). Remote sensing of phytoplankton pigment distribution in the United States northeast coast. *Remote Sensing of Environment*, *114*(11), 2403–2416. https://doi.org/10.1016/J.RSE.2010.05.015

Pan, X., Wong, G. T. F., Ho, T. Y., Shiah, F. K., & Liu, H. (2013). Remote sensing of picophytoplankton distribution in the northern South China Sea. *Remote Sensing of Environment*, *128*, 162–175. https://doi.org/10.1016/j.rse.2012.10.014

Passow, U., & Carlson, C. A. (2012). The biological pump in a high CO_2 world. *Marine Ecology Progress Series*, *270*, 249–272. https://doi.org/10.2307/24876215

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans Data. *Scientific Data*, *2*, 150023. https://doi.org/10.1038/sdata.2015.23

Reusch, D. B., Alley, R. B., & Hewitson, B. C. (2007). North Atlantic climate variability from a self-organizing map perspective. *Journal of Geophysical Research*, *112*, D02104. https://doi.org/10.1029/2006JD007460

Richardson, A., Risien, C., & Shillington, F. (2003). Using self-organizing maps to identify patterns in satellite imagery. *Progress in Oceanography*, *59*(2–3), 223–239. https://doi.org/10.1016/J.POCEAN.2003.07.006

Saha, K., Zhao, X., Zhang, H., Casey, K. S., Zhang, D., Baker-Yeboah, S., et al. (2018). AVHRR Pathfinder version 5.3 level 3 collated (L3C) global 4km sea surface temperature for 1981–Present. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V52J68XX

Sammartino, M., di Cicco, A., Marullo, S., & Santoleri, R. (2015). Spatio-temporal variability of micro-, nano- and pico-phytoplankton in the Mediterranean Sea from satellite ocean colour data of SeaWiFS. *Ocean Science Discussions*, *11*(5), 759–778. https://doi.org/10.5194/os-11-759-2015

Sarmiento, J. L., Slater, R., Barber, R., Bopp, L., Doney, S. C., Hirst, A. C., et al. (2004). Response of ocean ecosystems to climate warming. *Global Biogeochemical Cycles*, *18*, GB3003. https://doi.org/10.1029/2003GB002134

Sarzeaud, O., & Stephan, Y. (2000). Data interpolation using Kohonen networks. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, *6*, 197–202.

Schulz, K., Neill, C., Riebesell, U., Schulz, K. G., Bellerby, R. G. J., Botros, M., et al. (2007). Enhanced biological carbon consumption in a high CO_2 ocean. *Nature*, *450*(7169), 545–548. https://doi.org/10.1038/nature06267

Smetacek, V. (1999). Diatoms and the Ocean Carbon Cycle. *Archiv für Protistenkunde*, *150*(1), 25–32. https://doi.org/10.1016/S1434-4610(99)70006-4

Tréguer, P., Nelson, D. M., Van Bennekom, A. J., Demaster, D. J., Leynaert, A., & Quéguiner, B. (1995). The Silica Balance in the World Ocean: A Reestimate. *American Association for the Advancement of Science*, *268*(5209), 375–379. https://doi.org/10.1126/science.268.5209.375

Uitz, J., Claustre, H., Morel, A., & Hooker, S. B. (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal Geophysical Research*, *111*, C08005. https://doi.org/10.1029/2005jc003207

Vidussi, F., Claustre, H., Manca, B. B., Luchetta, A., & Marty, J.-C. (2001). Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter. *Journal of Geophysical Research*, *106*(C9), 19,939–19,956. https://doi.org/10.1029/1999JC000308

Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sensing of Environment*, *98*(1), 122–140. https://doi.org/10.1016/j.rse.2005.07.001

Westberry, T., Behrenfeld, M. J., Siegel, D. A., & Boss, E. (2008). Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles*, *22*, G82024. https://doi.org/10.1029/2007GB003078

Wright, S. W., & Jeffrey, S. W. (1987). Fucoxanthin pigment markers of marine phytoplankton analysed by HPLC and HPTLC. *Marine Ecology Progress Series*, *38*, 259–266. https://doi.org/10.3354/meps038259