Linking satellites to genes with machine learning to estimate major phytoplankton groups from space

Roy El Hourany, Juan Pierella Karlusich, Lucie Zinger, Hubert Loisel, Marina Lévy, Chris Bowler

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Linking satellites to genes with machine learning to estimate major phytoplankton groups from space

Roy El Hourany[1], Juan Pierella Karlusich[2,3], Lucie Zinger[2,4], Hubert Loisel[1], Marina Levy[5], and Chris Bowler[2]

[1]Univ. Littoral Côte d'Opale, Univ. Lille, CNRS, IRD, UMR 8187, LOG, Laboratoire d'Océanologie et de Géosciences, F 62930 Wimereux, France
[2]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[3]FAS Division of Science, Harvard University, Cambridge, MA
[4]Naturalis Biodiversity Center, 2300 RA Leiden, The Netherlands
[5]Sorbonne Université, LOCEAN-IPSL, Laboratoire d'Océanographie et du Climat ; Expérimentations et Approches Numériques, CNRS, IRD, MNHN, 75005 Paris, France

Correspondence: Roy El Hourany (roy.elhourany@univ-littoral.fr), Marina Levy (marina.levy@locean.ipsl.fr), Chris Bowler (cbowler@biologie.ens.fr)

Abstract. Ocean color remote sensing offers two decades-long time series of information on phytoplankton abundance. However, determining the structure of the phytoplankton community from this signal is not straightforward, and many uncertainties remain to be evaluated, despite multiple intercomparison efforts of the different available algorithms. Here, we use remote sensing and machine learning to infer the abundance of seven phytoplankton groups at global scale based on a new molecular method from *Tara* Oceans. Our dataset is to our knowledge the most comprehensive and complete, available to describe phytoplankton community structure at global scale using a molecular marker that defines relative abundances of all phytoplankton groups simultaneously. The methodology shows satisfying performances to provide robust estimates of phytoplankton groups using satellite data, with few limitations regarding the global generalization of the method. Furthermore, this new satellite-based methodology allows a valuable global intercomparison with the pigment-based approach used in in-situ and satellite data to identify phytoplankton groups. Nevertheless, these datasets show different, yet coherent information on the phytoplankton, valuable for the understanding of community structure. This makes remote sensing observations excellent tools to collect Essential Biodiversity Variables and provide a foundation for developing marine biodiversity forecasts.

1 Introduction

In marine ecosystems, the production of organic matter (i.e., productivity) relies largely on phytoplankton. These unicellular photosynthetic microorganisms are evolutionarily diverse and exhibit a wide range of cell morphologies, sizes, photosynthetic accessory pigments, elemental requirements, and biogeochemical and trophic functions (Pierella Karlusich et al., 2020). They play a key role in regulating ocean biogeochemistry (Fuhrman, 2009), including the export of organic matter to the deep ocean (Guidi et al., 2009; Tilman et al., 2014), which contributes to the modulation of atmospheric CO2 levels and climate.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

In a continuously changing environment, it is important to investigate the potential impacts of increased climatic variation and the effect of environmental fluctuations on planktonic biodiversity and marine ecosystem functioning (Ibarbalz et al., 2019; Henson et al., 2021). Such investigation requires the acquisition of high-resolution, real-time, and global scale data on phytoplankton community structure that can additionally inform about the state of its associated ecosystem functions often referred to as Essential Biodiversity Variables (Pereira et al., 2013). Numerous studies aimed at understanding or predicting global marine phytoplankton patterns based on various in-situ techniques, from microscopy to DNA sequencing-based methods (Hillebrand and Azovsky, 2001; Irigoien et al., 2004; Smith, 2007; Rodríguez-Ramos et al., 2015; Powell and Glazier, 2017; Righetti et al., 2019; Dutkiewicz et al., 2020; Pierella Karlusich et al., 2020). However, this existing knowledge relies on highly fragmented, spatially disparate, and temporally punctual observations, limited by the challenges of in situ data collection.

Ocean color remote sensing is an effective alternative to observe the global spatio-temporal distribution of phytoplankton at the sea surface at a high resolution. Since 1978, ocean color satellites have been quantifying chlorophyll-a (Chla) concentrations as a proxy of phytoplankton biomass (O'Reilly et al., 1998; Sathyendranath et al., 2014). This focus continues to the present with Chla concentration being by far the most utilized product from ocean-color satellites (Sathyendranath et al. (2014), IOCCG report N 14). It is only recently that these images have begun to be used to retrieve additional information about phytoplankton communities, such as their size structure, and their taxonomic or functional composition. This interest has paralleled the incorporation of the concept of phytoplankton functional types (PFT) into studies of a range of ecological and biogeochemical problems (Le Quéré et al., 2005; Hood et al., 2006). Functional types are defined according to the scientific questions being considered and the observational capabilities available or required to address them. The approaches span from categories related to biochemical processes (e.g., silicifiers, calcifiers) and physiological adaptations towards environmental factors (e.g., light, nutrients, turbulence) to practical categories that can be quantified with a particular analytical technique (e.g., pigment types) (IOCCG report N 14). Many of these efforts have focused on specialized algorithms to detect a single taxon with distinctive optical characteristics (Brown, 1995; Iglesias-Rodríguez et al., 2002), while other algorithms have also targeted a variety of phytoplankton based on pigments (diagnostic pigment analysis, DPA) (Alvain et al., 2005, 2008; Uitz et al., 2006; Aiken et al., 2009; Bracher et al., 2009; Hirata et al., 2011; Chase et al., 2020; Ben Mustapha et al., 2013). These algorithms can detect concentrations of three size classes of pico-, nano-, and microplankton (Phytoplankton size class, PSC, (Uitz et al., 2006; Hirata et al., 2011; Chase et al., 2020) or flag the dominant functional group within a total of five (Alvain et al., 2005, 2008; Ben Mustapha et al., 2013).

Pigment-based phytoplankton groups have received increasing interest from the ocean color community over the past decade due to the existence of large datasets of HPLC measurements with long time series and broad spatial coverage. The DPA approach was first proposed by Vidussi et al. (2001), based on the use of phytoplankton pigment information measured by high-performance liquid chromatography (HPLC) analysis as an alternative to more costly in-situ methods. This approach is based on the association of secondary phytoplankton pigments with broad taxonomic phytoplankton groups. Pigments contained within phytoplankton taxonomic groups are in turn assumed to be associated with one of the three size classes. The method of using diagnostic accessory pigments was further developed by Uitz et al. (2006), who applied weighting coefficients to diagnostic pigments to describe the respective proportion of three Phytoplankton Size Classes (PSC) to total Chla. From this

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

study, several applications to satellite data emerged, linking DPA to remote sensing (Uitz et al., 2006; Hirata et al., 2008, 2011; Soppa et al., 2014; Di Cicco et al., 2017; Organelli et al., 2013; El Hourany et al., 2019b, a; Xi et al., 2020). However, the reliance of the DPA on links between pigments and phytoplankton taxa, as well as the size range of different phytoplankton taxonomic groups, is not trivial due to the presence of certain pigments across different phytoplankton size and taxa (Brewin et al., 2014; Chase et al., 2020). This aspect may compromise the relevance of satellite images to retrieve reliable/meaningful taxonomic or functional EBVs for this biological compartment. However, the limitation of this signal remains so far not properly quantified.

One of the main reasons for the current uncertainties related to the relevance of satellite data to monitor planktonic diversity lies in the fact that current observational taxonomic and functional data on phytoplankton that could be used to validate the method is highly fragmentary and often obtained with inconsistent methodologies. In the following work, we address this limitation by using *Tara* Oceans' phytoplankton observations. These data are, to date, the most comprehensive and harmonized data available on the phytoplankton taxonomic community structure on a global scale, as obtained from metagenomics reads of a single-copied gene present across all phytoplanktonic groups, an approach that provides an unbiased picture of phytoplankton cell abundances (Pierella Karlusich et al., 2022). We employ these data alongside satellite-derived parameters to train an unsupervised machine learning algorithm to discern the non-linear relationship between the phytoplankton taxonomic community structure and the optical signal perceived by the Ocean color satellite sensors, together with the physical environment. This new methodology allowed us to monitor the spatio-temporal variability of seven phytoplankton groups between 1997 and 2021. Furthermore, a comparison is performed between this new algorithm and available satellite products (El Hourany et al., 2019a; Xi et al., 2020) which are based on the DPA pigment approach, to highlight common patterns, confidence, and limitations to estimating phytoplankton community structure using different sets of information.

2 Materials

In this section, different datasets are presented, each playing an important role in either the training of the algorithm or the validation and evaluation of global patterns of the outputs. Therefore, three datasets are employed. The first is the input dataset which allies between the in-situ information on phytoplankton groups from the *Tara* Oceans expedition and their associated satellite matchups. The second corresponds to an independent dataset based on global in-situ HPLC measurements, used to compare the outputs of the presented algorithm and the DPA approach to estimate phytoplankton groups. And last, the third dataset compiles two operational satellite-derived products on phytoplankton groups' abundance and therefore are used to compare large scale patterns and evaluate any discrepancies.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

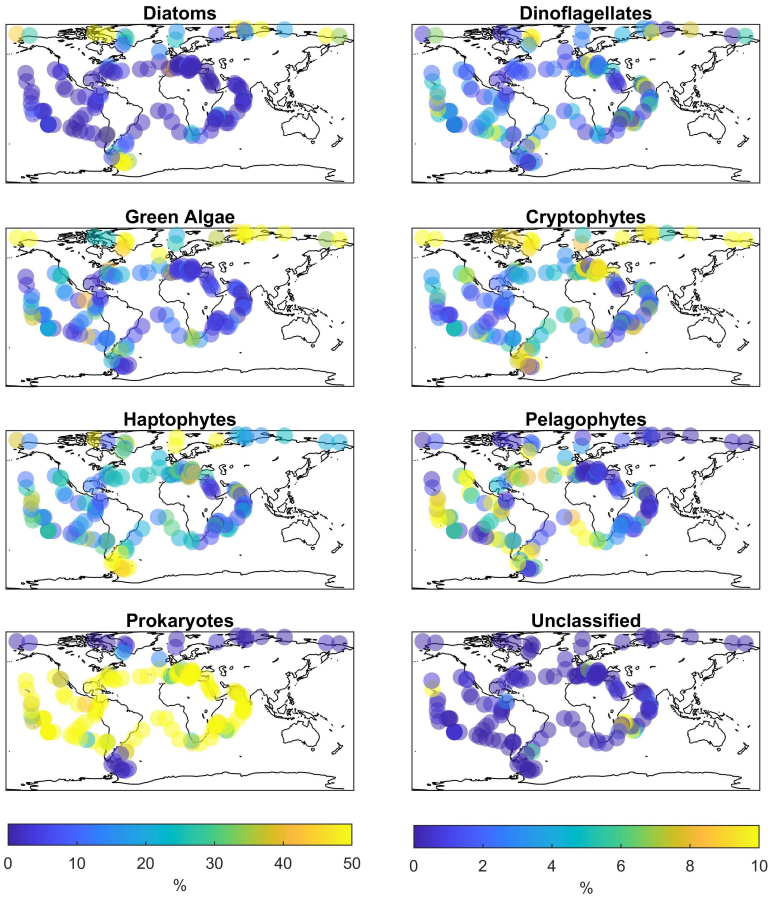

Figure 1. Global biogeographical patterns of marine phytoplankton based on *psbO* metagenomic reads from seawater samples collected by *Tara* Oceans expeditions. Note that different color scales are used for the most abundant groups (left panels) and for less abundant groups (right panels).

2.1 Input dataset

2.1.1 Metagenomic read abundance of the *psbO* gene

The *psbO* gene encodes the manganese-stabilizing protein, of around 270 amino acids, a core subunit of photosystem II (PSII) and unique to organisms carrying out oxygenic photosynthesis. The *psbO* gene is single copy in the vast majority of Eukaryotes and Prokaryotes. The reads mapping *psbO* were retrieved from the metagenomes and used as a proxy of phytoplankton relative cell abundance Pierella Karlusich et al. (2022). Given that five major organismal size fractions were collected by *Tara* Oceans (0.22-3um, 0.8-5um, 5-20um, 20-180um, 180-2000um), we formatted the data into major phytoplankton groups based on their

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

taxonomy. For every station, we pooled the results obtained for the five-size fractions into a single aggregated sample. We discarded the sampling stations where collected size fractions did not cover the full range of sizes so as not to bias the results. We then split this composite dataset composed by samples collected at 211 different stations into seven main phytoplankton groups based on DNA reads taxonomic assignation: diatoms (Bacillariophyta, hereafter referred to as Diat), dinoflagellates (Dinoflagellata, Dino), green algae (Chlorophyta, Green), haptophytes (Haptophytina, Hapto), pelagophytes (Pelagophyceae, Pelago), cryptophytes (Cryptophyta, Crypto), and prokaryotes mainly corresponding to Cyanobacteria (Fig. 1). The *psbO* read abundances of these main groups, which cover oligotrophic to eutrophic waters (Chla from 0.01 to 10 mg m^{-3}) were expressed as relative abundance (%) in relation to the total number of reads. Phytoplankton that were not assigned to any of the seven cited groups (Unclassified) represented less than 5% of the total phytoplankton community.

2.1.2 Satellite-derived Dataset

We used ocean color satellite data from the Globcolour project (R2019, full archive reprocessed, 2020), which consists of creating and maintaining a long time series of ocean color data from satellite measurements (from 1997 to the present). This database is the result of the fusion of data from various satellite sensors: 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). Nine Globcolour products were used, at daily and at 4km spatio-temporal resolution: Remote sensing reflectances at 4 wavelengths (Rrs412, Rrs443, Rrs490, and Rrs555), satellite Chla (product CHL1), light attenuation coefficient at 490 nm (Kd490), photosynthesis available radiation (PAR), Normalized fluorescence light height (NFLH) and particulate backscattering at 443 nm (bbp). In addition, we used Climate Change Initiative Sea Surface Temperature (SST) data at 4 km resolution and at daily frequency available from the Copernicus Marine Services (CMEMS) portal.

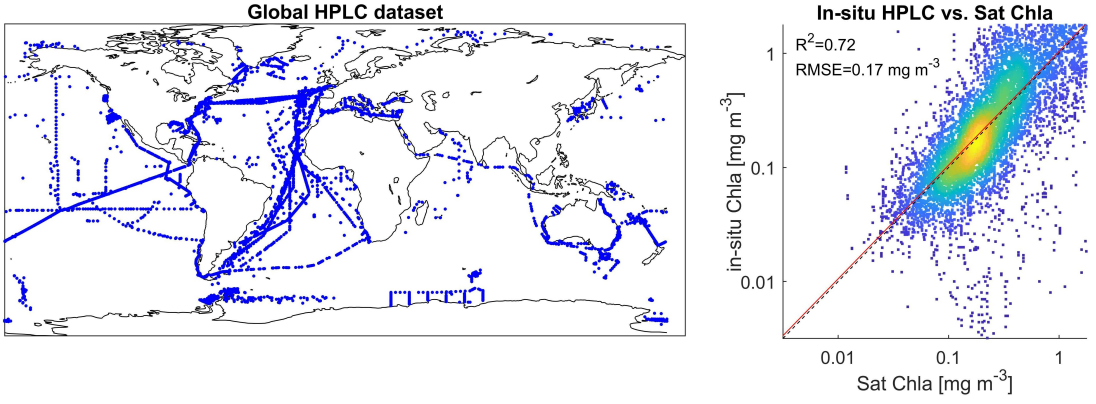

Figure 2. Geographical location of the global HPLC dataset stations regrouping observations from 1997 and 2014. The right panel represents a comparison between in-situ HPLC Chla measurement and its matchup using Globcolour Chla product.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Table 1. Percentage of missing values within the initial database (D)

Tara Oceans	***psbO***	**Sat**									
D (211 stations)		Chla	Rrs 412	Rrs 443	Rrs 490	Rrs 555	SST	bbp	Kd490	NFLH	PAR
Percentage of missing values	31	18	45	43	43	43	30	55	53	37	14

2.1.3 Structure of the Training and test databases

The initial dataset (D) is constituted of 211 *Tara* Oceans *psbO* observations of the relative abundance of the seven defined phytoplankton groups with the associated matchups of 10 satellite-derived parameters (Chla, SST, 4 Rrs (412, 443, 490, and 555 nm), NFLH, Kd490, PAR, and bbp). Even though the values are negligeable, the unclassified phytoplankton fraction was also added to ensure coherence and preservation of the total phytoplankton pool. The matchups between satellite observations and in-situ observations were selected by considering 3x3 pixel boxes around the in-situ coordinates and +/- 1 day around the day of the in-situ measurement. Relative *psbO*-based abundances were multiplied by the in-situ value of Chla measured as well at each *Tara* Oceans station, expressing every phytoplankton group as a function of a Chla fraction. All variables were normalized by their variance to homogenize weights. The hypothesis behind this is that the phytoplankton community should be treated as a whole, and therefore, the variability of each phytoplankton group is dependent on each other in a relative way. In parallel, D presents missing values (table 1); satellite missing values are usually linked to cloud coverage or masked data due to coastal/ice influence. However, missing values within the in-situ observation are due to a lack of measurements of a given station. And since D contains a low number of observations (211 stations), every observation is valuable. Therefore, to tackle these limitations, it is essential to use a non-linear multivariate regression algorithm that can deal with missing values and allow a robust generalization in the case of limited observations.

2.2 Global Phytoplankton HPLC pigment dataset

In order to compare the output of our methodology with an independent dataset, a global HPLC dataset has been compiled, regrouping 12 000 HPLC observations originating from several HPLC datasets between 1997 and 2014 (Fig. 2): MAREDAT, NOMAD, SeaBASS, and other oceanographic campaigns: Labrador, Gep&co, Polarstern Luo et al. (2012); Werdell and Bailey (2005); Dandonneau et al. (2004); Bracher et al. (2015); Fragoso et al. (2016); Peloquin et al. (2013). This HPLC dataset was collocated with satellite Globcolor and CCI matchups. This HPLC dataset contains the most used phytoplankton pigments to identify major phytoplankton groups; Fucoxanthin (Fuco), Peridinin (Perid), Alloxanthin (Allo), Zeaxanthin (Zea), Chlorophyll-b (Chlb), 19'-Hexanoyloxyfucoxanthin (19HF), 19'-Butanoyloxyfucoxanthin (19BF). The different phytoplankton groups and their associated pigments are shown in table 2. The objective of this dataset is to allow an independent global comparison of diagnostic pigments (DPA) vs. Satellite estimated phytoplankton groups. An average Chla fraction value for each phytoplankton group was calculated using the three sets of coefficients presented in table 2 (Uitz et al., 2006; Soppa et al., 2014; Brewin et al., 2015).

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Table 2. Phytoplankton groups and size classes associated with their diagnostic pigments and coefficients

Phytoplankton size class	Phytoplankton group	Diagnostic Pigment (DP)	Coefficients (a)*		
			Uitz et al., 2006	Soppa et al., 2014	Brewin et al., 2015
Micro	Diatoms	Fucoxanthin (Fuco) (Jeffrey, 1980)	1.41	1.55	1.51
	Dinoflagellates	Peridinin (Perid) (Jeffrey, 1980; Jeffrey and Hallegraeff, 1987)	1.41	0.41	1.35
Nano	Haptophytes	19'-Hexanoyloxyfucoxanthin (19HF) (Wright and Jeffrey, 1987)	1.27	0.86	0.95
	Green algae	Chlorophyll-b (Chlb) (Vidussi et al., 2001)	0.35	1.17	0.85
	Cryptophytes	Alloxanthin (Allo) (Gieskes and Kraay, 1983)	0.6	2.39	2.71
	Pelagophytes	19'-Butanoyloxyfucoxanthin (19BF) (Wright and Jeffrey, 1987)	1.01	1.06	1.27
Pico	Prokaryotes	Zeaxanthin (Zea) (Dandonneau et al., 2004; Guillard et al., 1985)	0.86	2.04	0.93
Coefficients based on global HPLC dataset corresponding the sum of the weighted diagnostic pigments to the total Chla; Chla=$\sum a$DP					

Tara Ocean's HPLC measurements (Pesant et al., 2015) were excluded from this global dataset. Since these HPLC measurements were performed on the same stations as for *psbO*, they are used in this study to evaluate the correspondence between pigments and phytoplankton groups.

2.3 Intercomparison with operational PFT satellite products

It is essential in this study to assess the consistency of the results of the presented method and to identify potential differences in emerging patterns by comparing them to existing products. For this matter, two satellite products have been identified:

2.3.1 Satellite-derived phytoplankton Chla fraction (CMEMS/Globcolour dataset)

This multi-sensor product contains the concentration of each phytoplankton functional type (expressed in terms of Chla concentration fraction) based on the Xi et al. (2020) algorithm, processed from 1997 till present. This algorithm allows an estimate of the Chla concentration of diatoms, dinoflagellates, haptophytes, green algae, and prokaryotes. The algorithm was implemented using HPLC-based phytoplankton groups (based on DPA approach, Soppa et al. (2014)) and satellite reflectance in the visible spectrum (bands comprise between 400 and 681 nm) with empirical orthogonal function (EOF). This dataset is found on the CMEMS portal (product number: OCEANCOLOUR_GLO_BGC_L3_MY_009_103).

2.3.2 Satellite-derived phytoplankton pigments

SOM-Pigments (El Hourany et al., 2019a) is a machine learning-based algorithm that allows the estimation of phytoplankton pigment concentrations in oceanic waters from satellite ocean color data (Chla, Rrs at four wavelengths: 412, 443, 490 and 555nm) and SST. this algorithm is based on the use of Self-Organizing maps (SOM) and was calibrated using a global HPLC dataset which includes 10 phytoplankton pigment concentrations: Chlorophyll-a (Chla), Divynil-Chlorophyll-a (DVChla), Chlorophyll-b (Chlb), Divynyl-Chlorophyll-b (DVChlb), 19'Hexfucoxanthin (19HF), 19'Butfucoxanthin (19BF), Fucoxanthin (Fuco), Peridinin (Perid), Alloxanthin (Allo), Zeaxanthin (Zea). The results of the cross-validation procedure scored a regres-

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

sion coefficient of 0.75 and an average RMSE of 0.016 mg.m-3. Using pigment concentrations, it is possible to determine the relative abundance of several phytoplankton groups (table 2).

The SOM-Pigments algorithm allowed us to estimate the 10 phytoplankton pigment concentrations cited above from satellite data on a global scale between 1997 and 2021. These pigments were used alongside the coefficients of Soppa et al. (2014) in table 2 to estimate five phytoplankton groups: Diatoms, dinoflagellates, haptophytes, green algae, pelagophytes, cryptophytes, and prokaryotes. We selected the above-mentioned set of coefficients for comparability reasons.

3 Methods

To extract the most information of the above-mentioned datasets, several machine learning algorithms were used in this study. Developing an operational algorithm that estimates from satellite information the phytoplankton groups' abundance was done using an unsupervised neural network called Self-Organizing map (SOM). The use of SOM and topology-constrained organization allowed us to uphold the non-linear relationships between phytoplankton group composition and satellite data through topology conservation. We tested different sets of learning hyperparameters and several combinations of satellite predictors to identify the optimal configuration of our algorithm. Once the training and the validation procedure have been done, the algorithm (called SOM-psbO) is operational and could be applied to satellite images to predict the phytoplankton groups' abundance on a global scale and daily from 1997 till the present. Following that, to identify global large scale patterns upheld by SOM-psbO, a second algorithm was used, the Ascending Hierarchical clustering algorithm. This latter permitted us to emphasize the predominant data structure learned by the SOM-psbO and characterize phytoplankton biomes. To explain the potential divergence between DPA approach and *psbO* measurements, the last analysis serves to highlight the cumulative importance of a pigment composition to estimate a phytoplankton group abundance through a Random Forest approach. In the following section, each methodology and algorithm are explained in detail.

3.1 Self-Organizing map applied to *Tara* Oceans *psbO* data (SOM-psbO)

3.1.1 Training of the SOM-psbO

The SOM algorithm (Kohonen, 2013) aims to cluster a multidimensional database (D) into classes represented by a fixed network of neurons (the SOM map). This network is used to define a discrete distance 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. Each observation of D will be assigned to the closest neuron, in the sense of the Euclidean Norm (EN). 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 so that two close observations of D (in the sense of EN) are assigned to two relatively close neurons on the map. The estimation of a neuron's vector and the topological order is achieved through a minimization process depending on the distance between the neuron and its assigned observation.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

SOMs have frequently been used in the context of completing missing data (Jouini et al., 2013). Under these conditions, the distance between an observation within D and the neuron's vectors 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 for considering the information embedded in the incomplete data.

Several experiments were made to find the ideal SOM size and have shown a significant increase in the general performance of the method at estimating pigment concentrations when the number of neurons increases to a certain extent. Using a number of neurons larger than the training data set allows discretization to be refined. In this case, some of them will capture a sample of the database, which permits to define a referent vector w for these neurons. When the neuron did not capture any data observation, the discrete distance between the neighboring neurons is used to determine the referent vector w of each neuron that has not captured any data (Sarzeaud and Stephan, 2000; El Hourany et al., 2019a). This leads to preserving the topological order provided by new interpolated neurons.

Following that, the best map size was evaluated while calculating two sets of metrics for each increasing map size:

- A quantification error and a topographic error: the quantification error represents the difference between an observation D and its closest neuron. This error is monitored during the training procedure till it reaches stability at a minimum value with increasing training epochs. This is where the training should stop to prevent overfitting. However, the topographic error is a representation of having, for each observation of the database, distant best-matching neurons. This quantity is important to monitor in order to ensure the preservation of the topological order within the SOM map with an increasing number of interpolated neurons.

- Mean regression coefficient and RMSE: in this case, for each given map size, D is used to cross-validate the SOM map. This is done using a one-leave-out procedure, where each observation of D is used iteratively either as a test or for training. Therefore, at each iteration of the cross-validation procedure, we calculate the closest neuron to the test observation on basis of its satellite variables only and associate these latter with the neuron's seven phytoplankton groups vector. When all the observations were used as a test, we calculate a mean R^2 and an RMSE, associated with the given size map, while comparing the estimated and predicted phytoplankton group values. Therefore, we define the best size map as the size at which all errors are at their lowest and the R^2 at its highest.

Therefore, we define the best size map as the size at which all errors are at their lowest and the R^2 at its highest.

3.1.2 Combination of satellite variables

In order to choose the best set of satellite data to estimate the phytoplankton groups, several combinations of satellite parameters were tested following the training and cross-validation procedure described above. Ten combinations were undergone, and the results of the cross-validation tests were presented in Fig. 3.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

3.1.3 Operational phase

After the training phase is concluded and the best SOM configuration is set, the algorithm becomes operational. In the following, the operational algorithm should be designated as SOM-psbO.

In the second phase, which is the operational phase, we estimate the phytoplankton group variability using different satellite images. The set of ocean satellite observations of a pixel is projected onto the SOM-psbO. In doing so, the projected parameters are normalized with the corresponding variances of D to maintain an equal weight among the parameters and are assigned with the closest best-matching neuron using the truncated distance. At the end of the assignment phase, each pixel is associated with a referent vector corresponding to the best matching neuron, which includes the seven phytoplankton groups as a function of Chla fraction. In order to regain information on each group's relative abundance, each Chla fraction of a neuron is divided by the total Chla value of this same neuron. Since the training was undergone using the whole phytoplankton structure at once, alongside the total Chla information, the SOM-psbO allows the inherent structure of the data to be preserved. For this phase, level 3 mapped 4 km daily images were used to estimate the phytoplankton group concentration at the same spatio-temporal resolution.

However, since our dataset D is limited to a short number of data, naturally occurring cases might be missed, and not considered while using such a dataset. A robust quality evaluation of the output of this method should be quantified to prevent abnormal predictions to cases that are not seen within dataset D. A reliability index was calculated between 1997 and 2021 by testing the set of values obtained for a given pixel against the values in the original dataset; if the value of a variable falls out of the bond defined by the +-2 Std around the mean value of the distribution, the variable is marked as distant. This is performed on all ten satellite variables per pixel. The reliability index is calculated by dividing the number of distant variables by the total number of existing variables to give an insight into how distant a pixel can be. With this definition, low values of the reliability index indicate that the method is reliable, and more care should be given to regions where the reliability index is larger.

3.2 Ascending Hierarchical Clustering applied on SOM-psbO

A hierarchical ascending clustering (HAC) was used on SOM-psbO's neurons; The reason behind this further clustering on the neurons is to emphasize major non-linear relationships observed in the database; in this case, the HAC is used to describe potential phytoplankton community biomes across the global ocean.

The HAC is a bottom-up algorithm for dataset clustering. The HAC starts from individuals and combines them according to their similarity (with respect to the chosen distance) to obtain new clusters. The exact number of biomes is not known a priori but at the end of the SOM+HAC procedure which suggests several possibilities of a number of clusters to be taken into account. A compromise is made between the number of clusters we can explain from a physical point of view and the number of clusters we need to include the maximum of information embedded in the dataset. This procedure has been used with success in several studies (Reygondeau et al., 2014; Richardson et al., 2003; Rossi et al., 2014; Sawadogo et al., 2009; El Hourany et al., 2021). At the end of the HAC clustering phase, each neuron of the SOM-psbO will be associated with a cluster. The association of several neurons in a cluster will allow us to identify common phytoplankton community structures, and therefore characterize

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

phytoplankton biomes. Upon applying SOM-psbO as described in the operational phase section, each pixel of a satellite image will be associated with a cluster.

3.3 Evaluation of the importance of pigments to estimate phytoplankton groups using random forest

Each *psbO*-derived phytoplankton group's abundance was associated with its corresponding HPLC pigments measurement performed on the same *Tara* Ocean's station. The importance of pigments to predict phytoplankton groups was computed using a Bagged Random Forest algorithm (number of learners set to 200), following the permutation-based importance method.

The bagged random forest algorithm is a set of Decision Trees, each constituted of internal nodes and leaves. In the internal node, the selected feature (i.e. pigment in this case) is used to make a decision on how to divide the dataset into separate sets with similar responses in terms of a given phytoplankton group. Since this algorithm is used in a case of regression, the decision is evaluated while monitoring the error decrease between the real phytoplankton group abundance and the predicted one, which corresponds to the value of a divided set. The permutation-based importance method will randomly shuffle each pigment and compute the change in the model's performance to predict the abundance of a phytoplankton group.

Using this method, a pigment composition of the seven major phytoplankton pigments cited in table 2 was tested to predict the abundance of each *psbO*-derived phytoplankton group, and therefore estimate their importance. The concentration of each pigment was evaluated in terms of pigment ratios, a ratio relative to the sum of all pigments' concentration, and in parallel, the *psbO*-derived relative abundance was used.

4 Results and discussion

4.1 Cross-validation, performances, and spatial limitation of the SOM-psbO algorithm

Ten combinations of satellite parameters were explored to estimate phytoplankton groups from *Tara* Oceans data (Fig. 3). The best combination of satellite parameters was: Chla, SST, PAR, Rrs at 4 wavelengths, bbp and Kd490. While using this set of satellite predictors, several SOM sizes were tested and the best set of maps with n ranging between 210 and 330 neurons was chosen based on the concordance of the maximum regression coefficient between estimated and observed phytoplankton values as well as the minimum error values of Quantization and topographic error, and the global RMSE related to all phytoplankton groups (Fig. 4).

The cross-validation of the best combination shows a performance of an average R^2 of 0.71 distributed between a maximal R^2= 0.86 for the green algae, and a minimal R^2 = 0.56 for the dinoflagellates and cryptophytes (Fig. 5, table 3). Upon summing all Chla fractions, the cross-validation analysis shows a satisfying agreement between estimated total Chla and in-situ values (R^2= 0.87) and therefore a preservation of the initial phytoplankton quantity expressed in total Chla.

To evaluate the quality of the representation of the inter-variable relationships within the input data by the SOM, the correlation coefficients and phytoplankton groups values distributions were compared between the referent vectors (constituting the neurons) of the SOM and D. This analysis showed that the correlation coefficients were not altered within the SOM, as well as

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

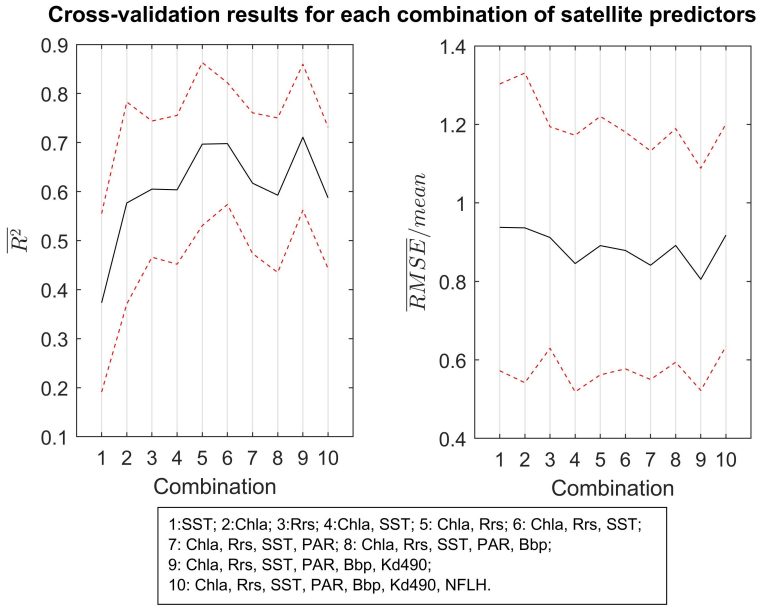

Figure 3. Cross-validation results expressed in terms of mean regression coefficient and mean relative RMSE for each combination of satellite parameters and for all the phytoplankton groups. The red dashed line delimits the standard deviation of the performances regarding different phytoplankton groups. The lowest error is shown at sensitivity test 9, regrouping the parameters Chla, bbp, Kd490, SST, and Rrs at 4 wavelengths.

the shape of the distribution of the values while compared to the initial dataset. These results highlight the capacity of SOM to preserve the characteristics of D after the training procedure (Fig. 6).

However, regarding the limited size of the initial dataset, one should be cautious when applying SOM-psbO to the global satellite data. Through the quality control described in section 3 and performed on each pixel of the daily satellite image between 1997 to 2021, a global map was generated to illustrate the extent of the applicability of this method (Fig. 7). Regions of low confidence can be identified where more than 40% of the pixels were masked throughout the time series between 1997 and 2021. These regions are mainly shown in coastal and turbid waters, and in the south pacific gyre, and are characterized either by very high or very low Chla values. This result is expected since the SOM algorithm is not able to extrapolate beyond the values' distribution of the initial dataset. Furthermore, moderate confidence regions can be defined in which around 25% of the pixels fall out of the accepted bonds. And these regions are mainly concentrated in high latitudes, especially in the Southern Ocean, mainly due to the limited number of data points that sampled the area and the particular optical characteristics of that region (Mitchell et al., 1991).

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Figure 4. Panel 1) Quality assessment based on the quantization and topographic error related to the training of the SOM as a function of increasing SOM size. In parallel, panel 2) represents the average regression coefficient and the root mean squared error as a function of increasing SOM size, calculated through a "one-leave out" cross-validation procedure between satellite-derived and in-situ *psbO* values. The dashed black and red line corresponds respectively to the R^2 = 0.54 and the RMSE=0.029 mg m^{-3} using the "K-nearest neighbor" algorithm.

Table 3. Results of the cross-validation of SOM-psbO, and the HPLC-based validation based on different metrics; regression coefficient (R^2), root-mean-squared-error (RMSE), mean absolute error (MAE) and the Spearman correlation coefficient (Rsp).

Phytoplankton group	**Cross-validation: Tara Oceans**				**HPLC-based Validation: Global**			
	R^2	RMSE [mg m^{-3}]	MAE [mg m^{-3}]	Rsp	R^2	RMSE [mg m^{-3}]	MAE [mg m^{-3}]	Rsp
Diatoms	0.71	0.074	0.038	0.76	0.78	0.29	0.20	0.85
Dinoflagellates	0.56	0.010	0.006	0.57	0.55	0.033	0.015	0.55
Green Algae	0.86	0.026	0.031	0.67	0.54	0.064	0.059	0.57
Haptophytes	0.66	0.043	0.023	0.82	0.57	0.11	0.085	0.68
Prokaryotes	0.81	0.089	0.050	0.66	0.20	0.064	0.15	0.23
Cryptophytes	0.84	0.095	0.021	0.73	0.50	0.072	0.031	0.61
Pelagophytes	0.56	0.019	0.010	0.56	0.49	0.02	0.018	0.58
Chlorophyll-a	0.87	0.231	0.133	0.89	0.71	0.54	0.45	0.81

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

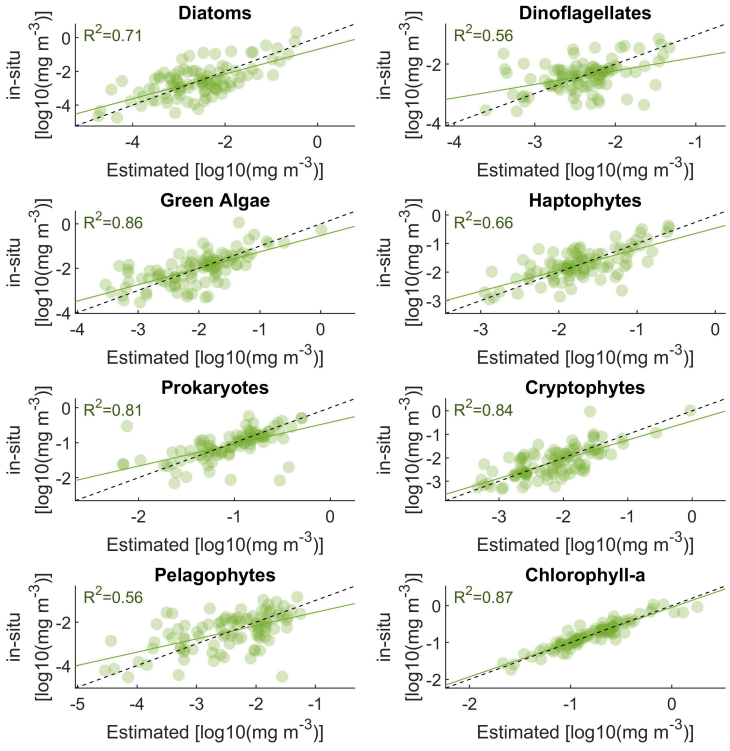

Figure 5. Results of the "one-leave out" cross-validation procedure. Each observation is used iteratively as a train or as a test set until all observations serve as tests.

4.1.1 Independent validation using global HPLC pigment dataset

The global in-situ HPLC dataset was used to estimate Chla fractions for each phytoplankton group. This dataset was compared to its matching phytoplankton group's Chla fraction estimated using SOM-psbO and satellite data (Fig. 8). Evaluating the sum of Chla fractions and comparing it with in-situ Chla can be considered as a baseline evaluation of this method. This comparison shows a satisfying correspondence scoring an R^2=0.7 with an RMSE of 0.17 mg m^{-3}. A relatively good correspondence is noted for the diatoms, showing an R^2=0.78 between in-situ and SOM- *psbO*'s diatoms Chla fraction. Moderate correspondence is noted for dinoflagellates, green algae, haptophytes, cryptophytes, and pelagophytes, with an R^2 ranging between 0.57 and 0.49. The prokaryotes had the lowest correspondence between both outputs. The comparison between DPA-based phytoplankton groups and SOM-psbO's estimates is highly uncertain. It compares two types of information indicating the same phytoplankton group, with different underlying assumptions about how to define and describe a certain group. For some of the groups, these results are coherent with the cross-validation performances of SOM-psbO. Diatom Chla fraction, is well captured

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

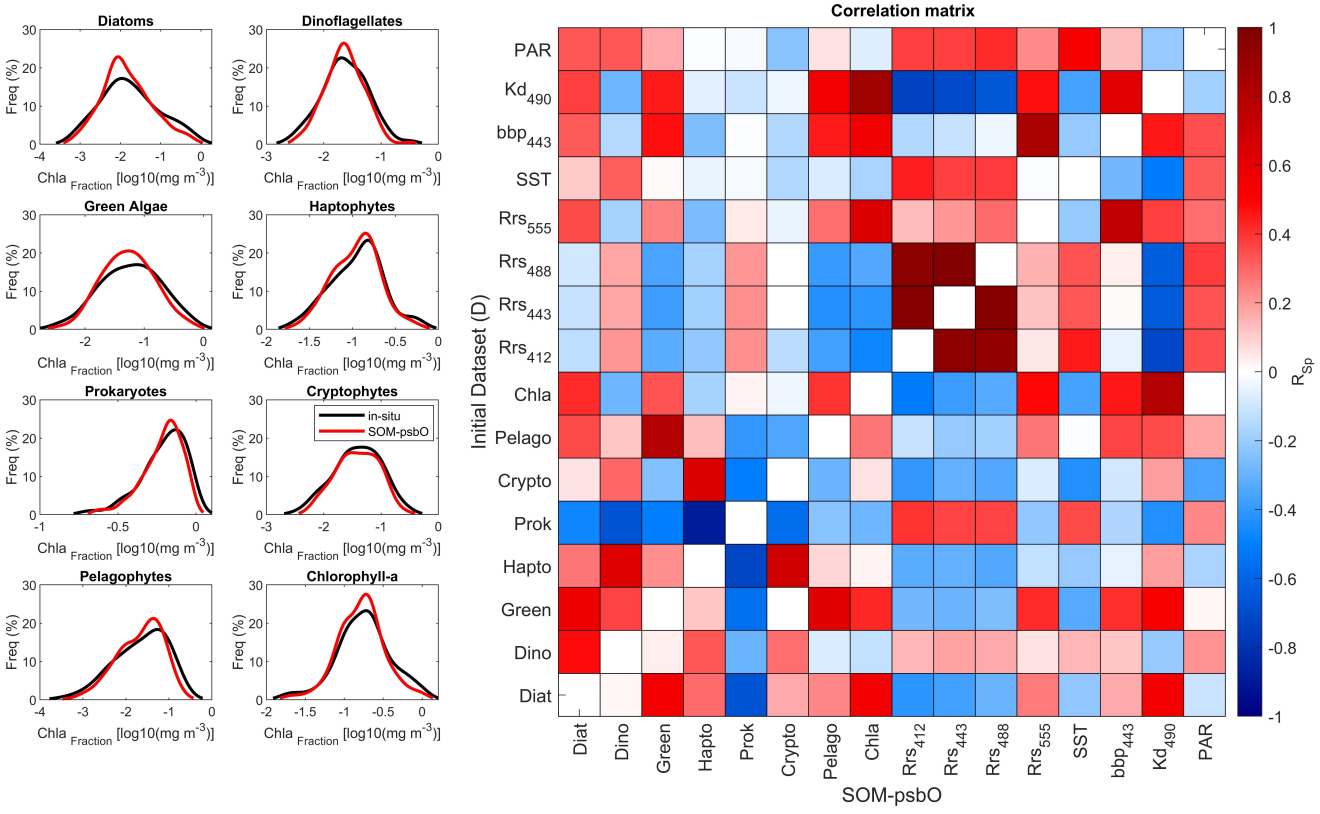

Figure 6. Evaluation of the preservation of the initial dataset's characteristics; Left panel) distribution of the values of *psbO*-derived Chla fraction for each phytoplankton group and the total Chla in the initial dataset D and SOM neurons (n=270). Right panel) Spearman correlation coefficient matrix comparing intra-correlations in the initial dataset D and SOM neurons.

by this latter, and the values agree with the ones estimated using HPLC observations, however, we note a major over-estimation within the HPLC DPA method. For prokaryotes, the satisfying cross-validation performances of SOM-psbO lead us to say that the use of zeaxanthin as an indicator of cyanobacteria's abundance may not be entirely representative of this group.

4.2 Global patterns of satellite-derived phytoplankton groups using SOM-psbO

Using satellite data, we generated a daily database of the relative abundance of the seven focus phytoplankton groups from 1997 to 2021. From this satellite-derived dataset, we computed the seasonal relative abundance patterns for each phytoplankton group (Fig. 9).

In terms of relative abundance, we could distinguish two largely dominant groups with antagonist spatial distributions: haptophytes and prokaryotes. This latter dominates largely in tropical regions all year round, reaching a relative abundance of 70% in subtropical gyres. Such environments suffer from ultra-oligotrophy. In conditions where low nutrients and high surface

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

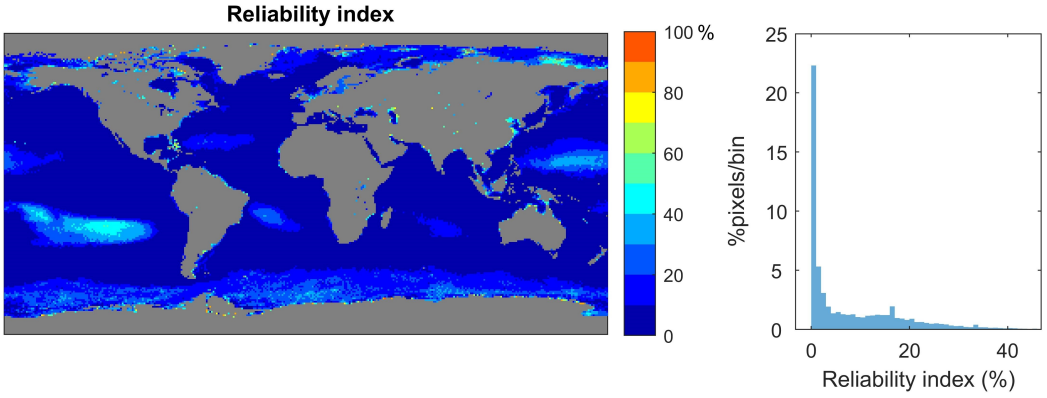

Figure 7. Applicability of the satellite *psbO*-based method; Geographical (left) and values distribution (right) of the reliability index calculated between 1997 and 2021 by testing the set of values obtained for a given pixel against the values in the original dataset (D). Low values of the reliability index indicate that the method is reliable, and more care should be given to regions where the reliability index is larger.

stratification prevail, picophytoplankton groups such as cyanobacteria strive with their high biovolume-to-size ratio making them adequate to dominate such regions (Chisholm, 1992; Raven, 1998). Haptophytes are largely abundant at mid and high latitudes and in the equatorial region, showing a maximum relative abundance of 40% in the Southern Ocean from September to March and in the northern hemisphere from March to September.

Other groups have a mid-ranged relative abundance, such as Diatoms and Green algae, representing each up to 30% of the total phytoplankton, blooming in the same way as Haptophytes, and dominating at high latitudes. High-latitude regions are characterized by high nutrient resources and exceptional seasonal variability of light intensity. This leads to an increase in phytoplankton groups with larger cell sizes, among them, the diatoms which are considered the most efficient and productive group among all the phytoplankton community (Loreau, 1998; Loreau and Hector, 2001). The three last phytoplankton groups have relative abundances barely exceeding 10% of the total phytoplankton community. Dinoflagellates and pelagophytes are observed mostly at mid and subtropical latitudes and cryptophytes in coastal areas and high latitudes.

From a qualitative perspective, the information captured by the SOM-psbO was clustered into six groups, each characterized by a particular remote sensing reflectance spectrum, in response to a phytoplankton community structure (Fig. 10). This application identifies clusters that are dominated by Eukaryotes (Diatoms, Haptophytes, and Green Algae, C1 and C2), a transitional cluster where Prokaryotes dominate with significant eukaryotic existence (C3), and three other clusters highly dominated by Prokaryotes (C4, C5, and C6). Based on the distribution of these clusters on a global scale, we can define several biomes, citing: C1 is centered in subtropical gyres, C2 in transitional zones such as mid-latitude regions as well as the equatorial region, C4 is shown on the edge of the subtropical Antarctic front, C5 for Polar and high latitudes and C6 for coastal and eutrophic waters. This structuration into parallel and transitional biomes supports the important effect of the latitudinal physical gradients on the structuration of the phytoplankton community (Ibarbalz et al., 2019), such as light availability and temperature.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

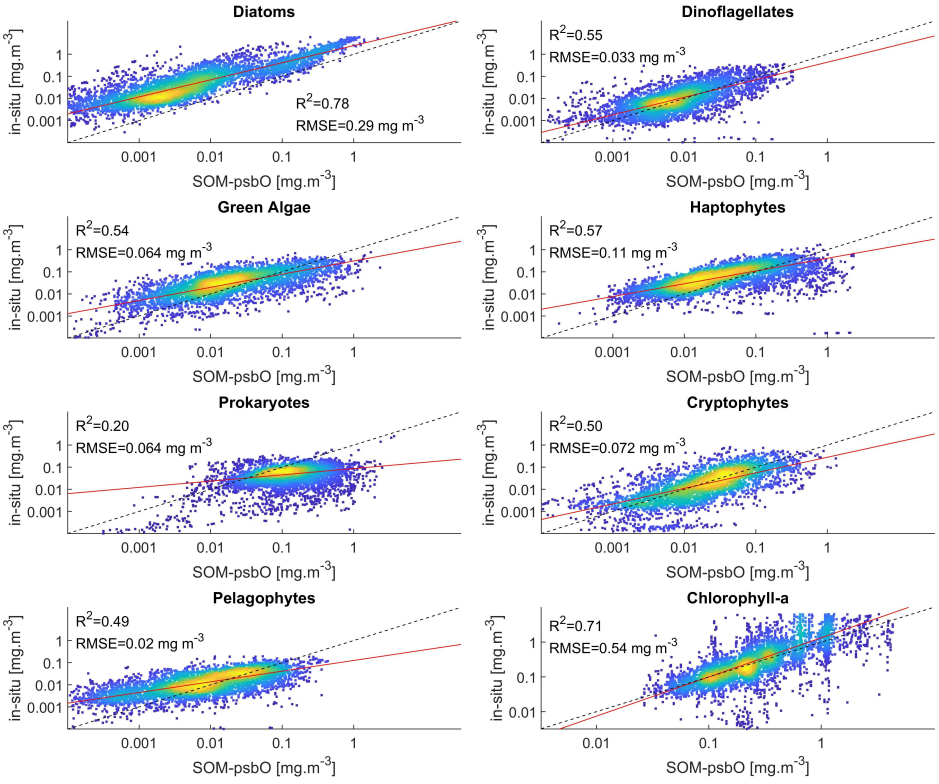

Figure 8. Evaluation of SOM-psbO method while comparing the outputs to in-situ global HPLC-based phytoplankton group measurements (DPA approach).

These patterns are in agreement with global phytoplankton studies in-situ (Ibarbalz et al., 2019; Sommeria-Klein et al., 2021) and satellite estimates (Alvain et al., 2005, 2008; Hirata et al., 2011; Ben Mustapha et al., 2013; El Hourany et al., 2019a; Xi et al., 2020).

4.3 Intercomparison of satellite-derived phytoplankton groups products

A comparison was performed between SOM-psbO's output, and two operational products based on Xi et al. (2020) and El Hourany et al. (2019a) algorithms, based on five phytoplankton groups common to all three outputs: Diatoms, Dinoflagellates, green algae, Haptophytes, and Prokaryotes. The annual patterns show a good agreement between all three satellite-derived phytoplankton estimates (Fig. 11).

Some differences between the estimated quantities of Chla phytoplankton groups fraction are noted: For diatoms, the outputs based on El Hourany et al. (2019a) exhibit higher Chla Diat values, and the ones based on Xi et al. (2020) show low values

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

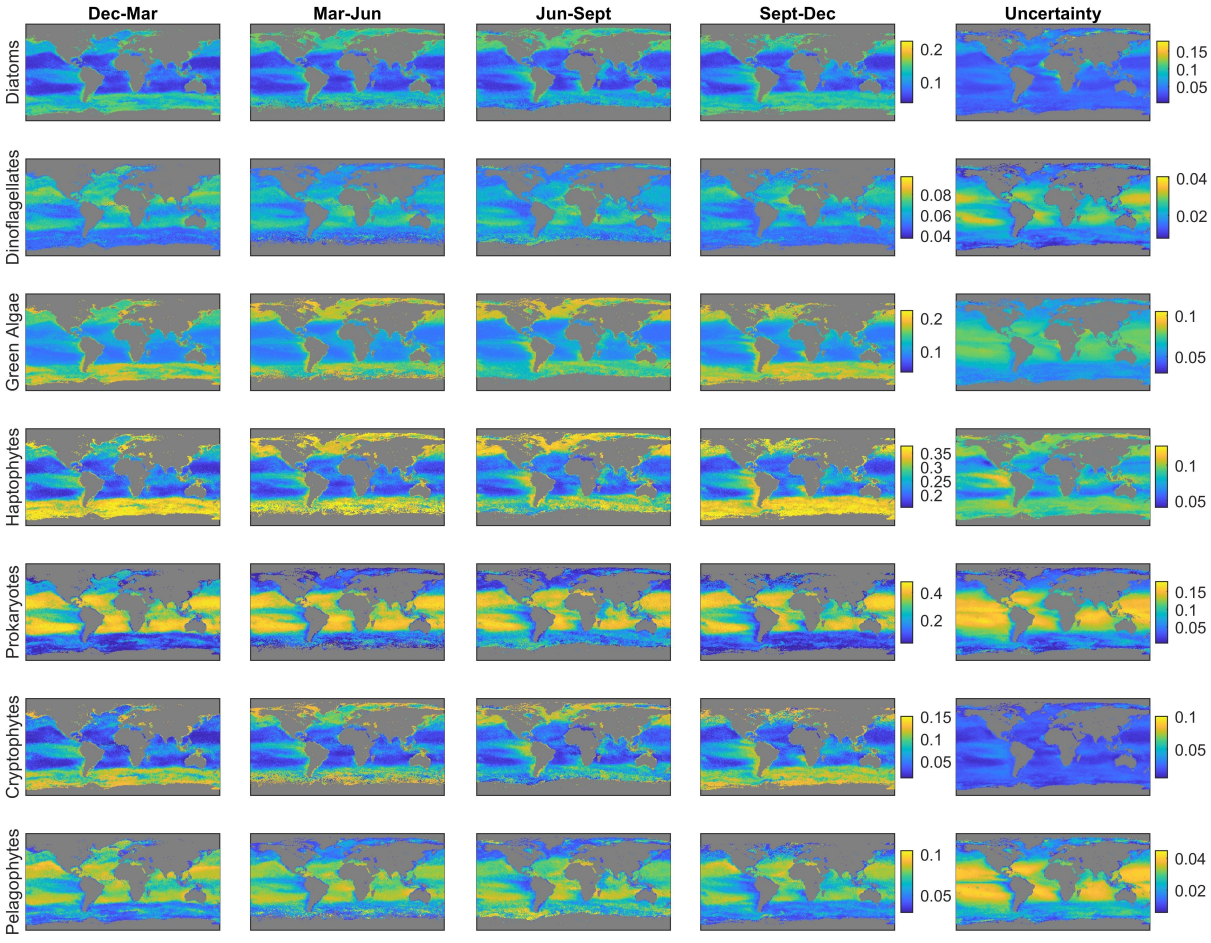

Figure 9. Mean seasonal relative abundances of the seven satellite psbO-derived phytoplankton groups (compiled using data from 1997-2021). The uncertainty related to each group represents the relative standard deviation between all estimated values at different initialization of the SOM and at each "one-leave out" cross-validation iteration.

near the equatorial latitudes. However, for SOM-psbO, the diatoms Chla fraction shows an increase at equatorial latitudes, mainly highlighting the upwelling activity in this latitudinal band. For green algae and haptophytes, SOM-psbO records higher concentrations than the other two products which are relatively matching. For prokaryotes, the outputs of Xi et al. (2020) show higher estimation, accentuated near the arctic and the equatorial region. And last, as for the dinoflagellates, the SOM-Pigments method shows lower values of Chla Dino, especially in subtropical gyres.

Addressing the differences between the outputs referring to the same phytoplankton group is not a straightforward task. Two methods are based on DPA approach, the latter is not trivial due to uncertainties related to the choice of pigments to

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

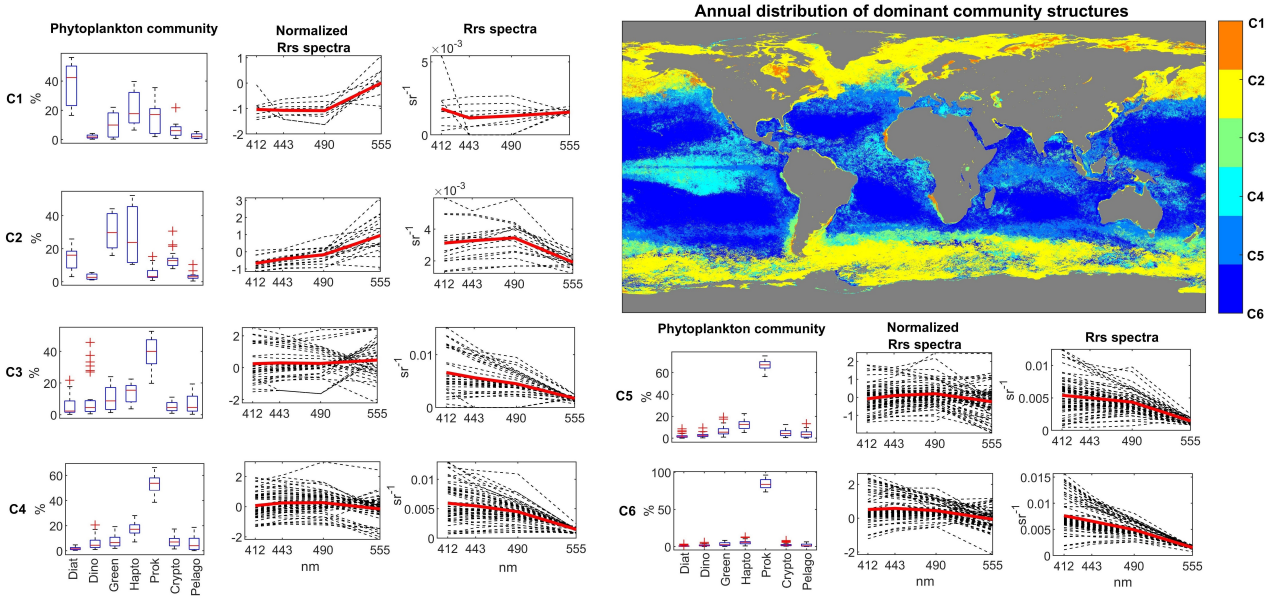

Figure 10. Satellite-derived biomes of phytoplanktonic communities, obtained by unsupervised clustering (Hierarchical clustering) on the SOM's referent vectors. The normalized and original Rrs spectrum was also derived to characterize each cluster's optical signature. The global map shows the most frequent community structure recorded during the 1997-2021 period.

delimit certain groups. Indeed, several studies showed that the DPA approach tends to overestimate some groups such as diatoms (Brewin et al., 2014; Chase et al., 2020). This approach may compromise the relevance of satellite images when used. However, the added value of such an approach resides in the availability of the large HPLC dataset allowing the development of robust algorithms. On the other hand, the method described in this paper and the generated outputs are based on a complete and harmonized database on the phytoplankton taxonomic community structure on a global scale; an approach that provides an unbiased picture of phytoplankton cell abundances. But the major limitation of this approach is the low number of observations which makes the global generalization of such a method a major challenge.

The random forest analysis, performed using the in-situ *Tara* Oceans' *psbO* and HPLC measurements, highlights the need for a multivariate approach to predict the phytoplankton community structure from pigments (Fig. 12). Therefore, the variability of each group is best explained not only by one diagnostic pigment but the pigment composition as a whole. It is important to consider how natural variability may impact the interpretation of the pigment composition in terms of phytoplankton community structure. Pigment ratios not only vary with phytoplankton composition, but also with their acclimation to light, temperature, and nutrients.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

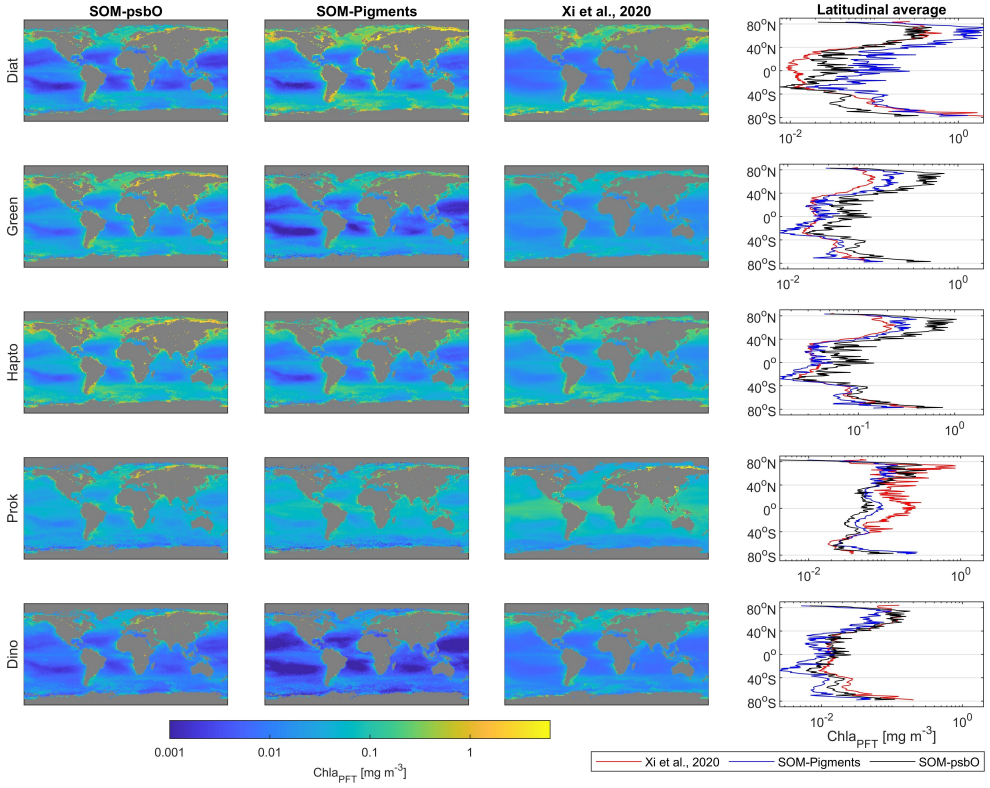

Figure 11. intercomparison of five satellite-derived phytoplankton group Chla fractions based on SOM-psbO, SOM-Pigments (El Hourany et al., 2019a), and Xi et al. (2020) algorithms. The average per latitude of each Chla fraction is calculated to reveal latitudinal patterns.

5 Conclusions

We found a remarkable congruence between the data derived from satellites and omics, despite the relatively small number of omics data. The link has been made possible due to machine learning techniques and the preservation of the data structure using Self-Organizing maps. The methodology showed satisfying performances to provide robust estimates of seven major phytoplankton groups, with few limitations regarding the global generalization of the method. The size of the training database is essential to provide a straightforward easy generalizable method. However, in this case, estimates in regions like the subtropical gyres should be interpreted with caution. As DNA sequencing costs continue to decrease and new expeditions generate molecular data from undersampled ocean regions, we expect the training datasets to increase in future years, and thus the accuracy of our method. Furthermore, this study presented a valuable global dataset of relative quantities of phytoplankton groups based on a new molecular method, leading to unique inter-comparison with the DPA-based approach used on in-situ

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

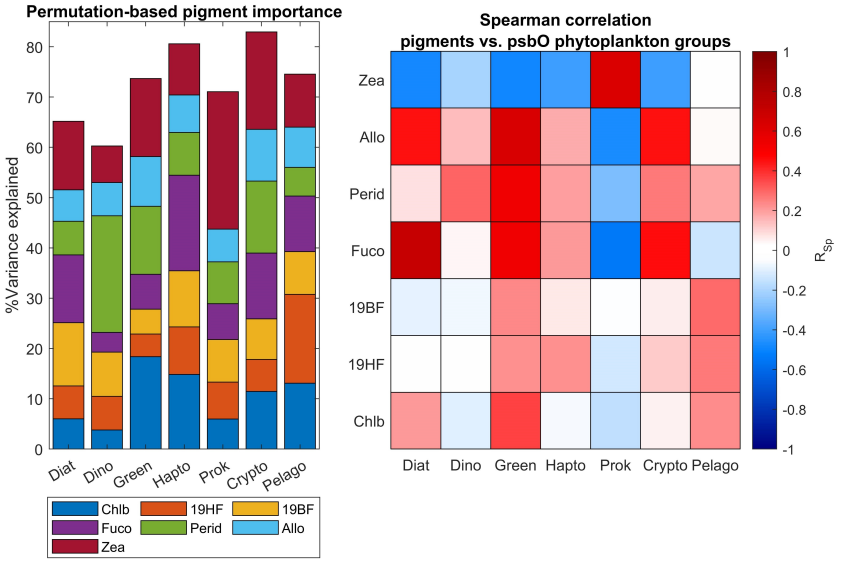

Figure 12. Evaluation of secondary pigments' importance regarding the estimation of phytoplankton groups. The left panel represents the percentage of the variance of each phytoplankton group explained by a set of frequently used phytoplankton secondary pigments. This analysis has been done using a random forest algorithm applied to the in-situ *Tara* Ocean's *psbO* and HPLC dataset. A Spearman correlation coefficient has been calculated between each pigment and the phytoplankton groups (right panel).

and satellite data to identify phytoplankton groups. Nevertheless, these datasets show different yet coherent information on the phytoplankton, valuable for the understanding of the community structure.

The methodology presented in this work provides a unique opportunity to observe in real-time and high-resolution the state of the major phytoplankton groups at the global scale. This makes remote sensing observations excellent tools to collect EBVs, play the role of broker between monitoring initiatives and decision-makers, and provide a foundation for developing marine biodiversity forecasts under different policy and management scenarios. To reach this objective, remote sensing data need inherently to be validated with in-situ observations as well. Few steps away from the PACE mission launch, a strategic climate continuity mission that will make global hyperspectral ocean color measurements possible. This will allow extended data records on ocean ecology and global biogeochemistry, revolutionizing the detection of phytoplankton communities from space. From the perspective of PACE, this study is a step towards further understanding the effect of environmental changes on phytoplankton community structure and diversity.

Code and data availability. psbO dataset: https://www.ebi.ac.uk/biostudies/studies/S-BSST761;

Globcolour dataset: https://www.globcolour.info/, https://hermes.acri.fr/.SSTCCIdataset: https://data.marine.copernicus.eu/product/SST_GLO_SST_L4_REP_OBSERVATIONS_010_024/description. Global HPLC pigment dataset: MAREDAT, POLERSTERN data, Labrador

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

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/php/x_datalist.php?xxop=gepco&xxcamp=gepco), 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-psbO will be deposited into a public domain repository accessible upon publication. Prerequisite software library SOM Toolbox 2.0 for Matlab is required, implementing the self-organizing map and Hierarchical Ascending Classification algorithm, Copyright (C) 1999 by Esa Alhoniemi, Johan Himberg, Jukka Parviainen, and Juha Vesanto and accessible at https://github.com/ilarinieminen/SOMToolbox. Matlab function for Random Forest algorithm was used to run the algorithm. MATLAB version R2020b, Statistics and Machine Learning Toolbox-Functions.

Author contributions. Conceptualization, RE, ML, CB. Methodology, RE. Validation, RE, JPK. Formal analysis, RE, JPK, ML, CB. Investigation, RE, JPK, LZ, HL, ML, CB. Resources, ML, CB. Data curation, JPK, RE. Writing-original draft preparation, RE, Writing-review and editing, RE, JPK, LZ, HL, ML, CB. Visualization, RE Supervision, ML, CB Project administration, ML, CB. Funding acquisition, RE, ML, CB.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Acknowledgements. The authors acknowledge the recommendations and guidance of Emmanuel Boss (Pr. At University of Maine) and Sylvie Thiria (Emeritus Pr. Sorbonne University). R.E. acknowledges CNES postdoc fellowship 2019-2021, CNES TOSCA 2020-2021, and Sorbonne University Emergence program 2021-2023 ML4BioChange. J.J.P.K. acknowledges postdoctoral funding from the Fonds Français pour l'Environnement Mondial. C.B. acknowledges ERC Advanced Award Diatomic (grant agreement No. 835067), and French Government 'Investissements d'Avenir' programs OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL Research University (ANR-11-IDEX-0001-02). This article is contribution number xxx of *Tara* Oceans.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

References

Aiken, J., Pradhan, Y., Barlow, R., Lavender, S., Poulton, A., Holligan, P., and Hardman-Mountford, N.: Phytoplankton pigments and functional types in the Atlantic Ocean: A decadal assessment, 1995-2005, Deep-Sea Research Part II: Topical Studies in Oceanography, 56, 899–917, https://doi.org/10.1016/j.dsr2.2008.09.017, 2009.

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

Alvain, S., Moulin, C., Dandonneau, Y., and Loisel, H.: Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view, Global Biogeochemical Cycles, 22, 1–15, https://doi.org/10.1029/2007GB003154, 2008.

Ben Mustapha, Z., Alvain, S., Jamet, C., Loisel, H., and Dessailly, D.: 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, https://doi.org/10.1016/j.rse.2013.08.046, 2013.

Bracher, A., Vountas, M., Dinter, T., Burrows, J. P., Röttgers, R., and Peeken, I.: Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data, Tech. rep., www.biogeosciences.net/6/751/2009/, 2009.

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

Brewin, R. J., Sathyendranath, S., Jackson, T., Barlow, R., Brotas, V., Airs, R., and Lamont, T.: Influence of light in the mixed-layer on the parameters of a three-component model of phytoplankton size class, Remote Sensing of Environment, 168, 437–450, https://doi.org/10.1016/J.RSE.2015.07.004, 2015.

Brewin, R. J. W., Sathyendranath, S., Tilstone, G., Lange, P. K., and Platt, T.: A multicomponent model of phytoplankton size structure, Journal of Geophysical Research: Oceans, 119, 3478–3496, https://doi.org/10.1002/2014JC009859, 2014.

Brown, C.: Global Distribution of Coccolithophore Blooms, Oceanography, 8, 59–60, https://doi.org/10.5670/oceanog.1995.21, 1995.

Chase, A. P., Kramer, S. J., Haëntjens, N., Boss, E. S., Karp-Boss, L., Edmondson, M., and Graff, J. R.: Evaluation of diagnostic pigments to estimate phytoplankton size classes, Limnology and Oceanography: Methods, 18, 570–584, https://doi.org/10.1002/LOM3.10385, 2020.

Chisholm, S. W.: Phytoplankton Size, Primary Productivity and Biogeochemical Cycles in the Sea, pp. 213–237, https://doi.org/10.1007/978-1-4899-0762-2_12, 1992.

Dandonneau, Y., Deschamps, P.-Y., Nicolas, J.-M., Loisel, H., Blanchot, J., Montel, Y., Thieuleux, F., and Bécu, G.: 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, 303–318, https://doi.org/10.1016/j.dsr2.2003.07.018, 2004.

Di Cicco, A., Sammartino, M., Marullo, S., and Santoleri, R.: 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, 2017.

Dutkiewicz, S., Cermeno, P., Jahn, O., Follows, M. J., Hickman, A. A., Taniguchi, D. A., and Ward, B. A.: Dimensions of marine phytoplankton diversity, Biogeosciences, 17, 609–634, https://doi.org/10.5194/BG-17-609-2020, 2020.

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

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

El Hourany, R., Abboud-Abi Saab, M., Faour, G., Mejia, C., Crépon, M., and Thiria, S.: Phytoplankton Diversity in the Mediterranean Sea From Satellite Data Using Self-Organizing Maps, Journal of Geophysical Research: Oceans, 124, 5827–5843, https://doi.org/10.1029/2019JC015131, 2019b.

El Hourany, R., Mejia, C., Faour, G., Crépon, M., and Thiria, S.: Evidencing the Impact of Climate Change on the Phytoplankton Community of the Mediterranean Sea Through a Bioregionalization Approach, Journal of Geophysical Research: Oceans, 126, e2020JC016 808, https://doi.org/10.1029/2020JC016808, 2021.

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

Fuhrman, J. A.: Microbial community structure and its functional implications, https://doi.org/10.1038/nature08058, 2009.

Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., Picheral, M., and Gorsky, G.: Effects of phytoplankton community on production, size, and export of large aggregates: A world-ocean analysis, Limnology and Oceanography, 54, 1951–1963, https://doi.org/10.4319/LO.2009.54.6.1951, 2009.

Henson, S. A., Cael, B. B., Allen, S. R., and Dutkiewicz, S.: Future phytoplankton diversity in a changing climate, Nature Communications 2021 12:1, 12, 1–8, https://doi.org/10.1038/s41467-021-25699-w, 2021.

Hillebrand, H. and Azovsky, A. I.: Body size determines the strength of the latitudinal diversity gradient, Ecography, 24, 251–256, https://doi.org/10.1034/J.1600-0587.2001.240302.X, 2001.

Hirata, T., Aiken, J., Hardman-Mountford, N., Smyth, T., and Barlow, R.: An absorption model to determine phytoplankton size classes from satellite ocean colour, Remote Sensing of Environment, 112, 3153–3159, https://doi.org/10.1016/J.RSE.2008.03.011, 2008.

Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., and Yamanaka, Y.: Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types, Biogeosciences, 8, 311–327, https://doi.org/10.5194/bg-8-311-2011, 2011.

Hood, R. R., Laws, E. A., Armstrong, R. A., Bates, N. R., Brown, C. W., Carlson, C. A., Chai, F., Doney, S. C., Falkowski, P. G., Feely, R. A., Friedrichs, M. A., Landry, M. R., Keith Moore, J., Nelson, D. M., Richardson, T. L., Salihoglu, B., Schartau, M., Toole, D. A., and Wiggert, J. D.: Pelagic functional group modeling: Progress, challenges and prospects, Deep Sea Research Part II: Topical Studies in Oceanography, 53, 459–512, https://doi.org/10.1016/J.DSR2.2006.01.025, 2006.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., Labadie, K., Ferland, J., Marec, C., Kandels, S., Picheral, M., Dimier, C., Poulain, J., Pisarev, S., Carmichael, M., Pesant, S., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Poulton, N., Raes, J., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Pelletier, E., Bopp, L., Lombard, F., and Zinger, L.: Global Trends in Marine Plankton Diversity across Kingdoms of Life, Cell, 179, 1084–1097.e21, https://doi.org/10.1016/J.CELL.2019.10.008, 2019.

Iglesias-Rodríguez, M. D., Brown, C. W., Doney, S. C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P. K., and Falkowski, P. G.: Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids, Global Biogeochemical Cycles, 16, 47–1–47–20, https://doi.org/10.1029/2001GB001454, 2002.

Irigoien, X., Hulsman, J., and Harris, R. P.: Global biodiversity patterns of marine phytoplankton and zooplankton, Nature 2004 429:6994, 429, 863–867, https://doi.org/10.1038/nature02593, 2004.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Jouini, M., Lévy, M., Crépon, M., and Thiria, S.: 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, 2013.

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

Le Quéré, C., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Global Change Biology, 0, 051013014052 005–???, https://doi.org/10.1111/j.1365-2486.2005.1004.x, 2005.

Loreau, M.: Biodiversity and ecosystem functioning: A mechanistic model, Proceedings of the National Academy of Sciences, 95, 5632–5636, https://doi.org/10.1073/PNAS.95.10.5632, 1998.

Loreau, M. and Hector, A.: Partitioning selection and complementarity in biodiversity experiments, Nature 2001 412:6842, 412, 72–76, https://doi.org/10.1038/35083573, 2001.

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., Bonnet, S., Boström, K. H., Böttjer, D., Capone, D. G., Carpenter, E. J., Chen, Y. L., Church, M. J., Dore, J. E., Falcón, L. I., Fernández, A., Foster, R. A., Furuya, K., Gómez, F., Gundersen, K., Hynes, A. M., Karl, D. M., Kitajima, S., Langlois, R. J., LaRoche, J., Letelier, R. M., Marañón, E., McGillicuddy, D. J., Moisander, P. H., Moore, C. M., Mouriño-Carballido, B., Mulholland, M. R., Needoba, J. A., Orcutt, K. M., Poulton, A. J., Rahav, E., Raimbault, P., Rees, A. P., Riemann, L., Shiozaki, T., Subramaniam, A., Tyrrell, T., Turk-Kubo, K. A., Varela, M., Villareal, T. A., Webb, E. A., White, A. E., Wu, J., and Zehr, J. P.: Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates, Earth System Science Data, 4, 47–73, https://doi.org/10.5194/essd-4-47-2012, 2012.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. a., Carder, K. L., Garver, S. a., Kahru, M., and McClain, C.: Ocean color chlorophyll algorithms for SeaWiFS, Journal of Geophysical Research, 103, 24 937, https://doi.org/10.1029/98JC02160, 1998.

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

Peloquin, J., Swan, C., Gruber, N., Vogt, M., Claustre, H., Ras, J., Uitz, J., Barlow, R., Behrenfeld, M., Bidigare, R., Dierssen, H., Ditullio, G., Fernandez, E., Gallienne, C., Gibb, S., Goericke, R., Harding, L., Head, E., Holligan, P., Hooker, S., Karl, D., Landry, M., Letelier, R., Llewellyn, C. A., Lomas, M., Lucas, M., Mannino, A., Marty, J.-c., Mitchell, B. G., Muller-Karger, F., Nelson, N., Prezelin, B., Repeta, D., Smith Jr, W. O., Smythe-Wright, D., Stumpf, R., Subramaniam, A., Suzuki, K., Trees, C., Vernet, M., Wasmund, N., and Wright, S.: The MAREDAT global database of high performance liquid chromatography marine pigment measurements, Earth Syst. Sci. Data, 5, 109–123, https://doi.org/10.5194/essd-5-109-2013, 2013.

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D. S., McGeoch, M. A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J. P., Stuart, S. N., Turak, E., Walpole, M., and Wegmann, M.: Essential biodiversity variables, Science, 339, 277–278, https://doi.org/10.1126/SCIENCE.1229931/SUPPL_FILE/1229931.PEREIRA.SM.PDF, 2013.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Trouble, R., Dimier, C., and Searson, S.: Open science resources for the discovery and analysis of Tara Oceans data, Scientific Data, 2, https://doi.org/10.1038/SDATA.2015.23, 2015.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Pierella Karlusich, J. J., Ibarbalz, F. M., and Bowler, C.: Phytoplankton in the Tara Ocean, https://doi.org/10.1146/annurev-marine-010419-010706, 12, 233–265, https://doi.org/10.1146/ANNUREV-MARINE-010419-010706, 2020.

Pierella Karlusich, J. J., Pelletier, E., Zinger, L., Lombard, F., Zingone, A., Colin, S., Gasol, J. M., Dorrell, R. G., Henry, N., Scalco, E., Acinas, S. G., Wincker, P., de Vargas, C., and Bowler, C.: A robust approach to estimate relative phytoplankton cell abundances from metagenomes, Molecular Ecology Resources, 00, 1–25, https://doi.org/10.1111/1755-0998.13592, 2022.

Powell, M. G. and Glazier, D. S.: Asymmetric geographic range expansion explains the latitudinal diversity gradients of four major taxa of marine plankton, Paleobiology, 43, 196–208, https://doi.org/10.1017/PAB.2016.38, 2017.

Raven, J.: The twelfth Tansley Lecture. Small is beautiful: the picophytoplankton, Functional ecology, 12, 503–513, 1998.

Reygondeau, G., Irisson, J.-O., Ayata, S. D., Gasparini, S., Benedetti, F., Albouy, C., Hattab, T., Guieu, C., and Koubbi., P.: Definition of the Mediterranean Eco-regions and Maps of Potential Pressures in These Eco-regions, Tech. rep., Perseus Deliverable 1, http://www.perseus-net.eu/assets/media/PDF/deliverables/3336.6_Final.pdf, 2014.

Richardson, A. J., Risien, C., and Shillington, F. A.: Using self-organizing maps to identify patterns in satellite imagery, Progress in Oceanography, 59, 223–239, https://doi.org/10.1016/j.pocean.2003.07.006, 2003.

Righetti, D., Vogt, M., Gruber, N., Psomas, A., and Zimmermann, N. E.: Global pattern of phytoplankton diversity driven by temperature and environmental variability, Science Advances, 5, 6253–6268, https://doi.org/10.1126/SCIADV.AAU6253/SUPPL_FILE/AAU6253_SM.PDF, 2019.

Rodríguez-Ramos, T., Marañón, E., and Cermeño, P.: Marine nano- and microphytoplankton diversity: redrawing global patterns from sampling-standardized data, Global Ecology and Biogeography, 24, 527–538, https://doi.org/10.1111/GEB.12274, 2015.

Rossi, V., Ser-Giacomi, E., Lõpez, C., and Hernández-García, E.: Hydrodynamic provinces and oceanic connectivity from a transport network help designing marine reserves, Geophysical Research Letters, 41, 2883–2891, https://doi.org/10.1002/2014GL059540, 2014.

Sarzeaud, O. and Stephan, Y.: Data interpolation using Kohonen networks, Proceedings of the International Joint Conference on Neural Networks, 6, 197–202, 2000.

Sathyendranath, S., Aiken, J., Alvain, S., Barlow, R., Bouman, H., Bracher, A., Brewin, R., Bricaud, A., Brown, C. W., Ciotti, A. M., Clementson, L. A., Craig, S. E., Devred, E., Hardman-Mountford, N., Hirata, T., Hu, C., Kostadinov, T. S., Lavender, S., Loisel, H., Moore, T. S., Morales, J., Mouw, C. B., Nair, A., Raitsos, D., Roesler, C., Shutler, J. D., Sosik, H. M., Soto, I., Stuart, V., Subramaniam, A., and Uitz, J.: Phytoplankton functional types from Space, International Ocean-Colour Coordinating Group, Dartmouth, Nova Scotia, B2Y 4A2, Canada., ioccg; 15 edn., https://epic.awi.de/id/eprint/36000/, 2014.

Sawadogo, S., Brajard, J., Niang, A., Lathuiliere, C., Crepon, M., and Thiria, S.: Analysis of the Senegalo-Mauritanian upwelling by processing satellite remote sensing observations with topological maps., in: 2009 International Joint Conference on Neural Networks, pp. 2826–2832, IEEE, https://doi.org/10.1109/IJCNN.2009.5178623, 2009.

Smith, V. H.: Microbial diversity–productivity relationships in aquatic ecosystems, FEMS Microbiology Ecology, 62, 181–186, https://doi.org/10.1111/J.1574-6941.2007.00381.X, 2007.

Sommeria-Klein, G., Watteaux, R., Ibarbalz, F. M., Karlusich, J. J. P., Iudicone, D., Bowler, C., and Morlon, H.: Global drivers of eukaryotic plankton biogeography in the sunlit ocean, Science, 374, 594–599, https://doi.org/10.1126/SCIENCE.ABB3717/SUPPL_FILE/SCIENCE.ABB3717_MDAR_REPRODUCIBILITY_CHECKLIST.PDF, 2021.

Soppa, M. A., Hirata, T., Silva, B., Dinter, T., Peeken, I., Wiegmann, S., and Bracher, A.: Global retrieval of diatom abundance based on phytoplankton pigments and satellite data, Remote Sensing, 6, 10 089–10 106, https://doi.org/10.3390/rs61010089, 2014.

https://doi.org/10.5194/egusphere-2022-1421
Preprint. Discussion started: 14 December 2022

Tilman, D., Isbell, F., and Cowles, J. M.: Biodiversity and Ecosystem Functioning, Annu. Rev. Ecol. Evol. Syst, 45, 471–493, https://doi.org/10.1146/annurev-ecolsys-120213-091917, 2014.

Uitz, J., Claustre, H., Morel, A., and Hooker, S. B.: Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, J. Geophys. Res., 111, C08 005, https://doi.org/10.1029/2005jc003207, 2006.

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

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

Xi, H., Losa, S. N., Mangin, A., Soppa, M. A., Garnesson, P., Demaria, J., Liu, Y., D'Andon, O. H. F., and Bracher, A.: Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data, Remote Sensing of Environment, 240, 111 704, https://doi.org/10.1016/J.RSE.2020.111704, 2020.