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

Roy El Hourany[1], Juan J. Pierella Karlusich[2,3], Lucie Zinger[2,4], Hubert Loisel[1], Marina Levy[5] & 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

Correspondance : R. ~~El Hourany (roy.elhourany@univ-littoral.fr), M. Levy (marina.levy@locean.ipsl.fr), C. Bowler (cbowler@biologie.ens.fr).~~El Hourany (roy.elhourany@univ-littoral.fr), M. Levy (marina.levy@locean.ipsl.fr) and C. Bowler (cbowler@biologie.ens.fr)

Abstract

Ocean color remote sensing ~~offers~~has been used for more than two decades ~~long time series of information on phytoplankton abundance. However, determining the structure of the~~ to estimate primary productivity. Approaches have also been developed to disentangle phytoplankton community structure based on spectral data 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~~ space, in particular when combined with in situ measurements of photosynthetic pigments. Here, we propose a new ocean color algorithm to derive the relative cell 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~~, as well as their contribution 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.~~total chlorophyll-a (Chla) at the global scale. Our algorithm is based on machine learning and has been trained using remotely-sensed parameters (reflectance, backscattering and attenuation coefficients at different wavelengths, plus temperature and Chla) combined with an omics-based biomarker developed using *Tara* Oceans data representing a single-copy gene encoding a component of the photosynthetic machinery that is present across all

Formatted: Numbering: Restart each page

Formatted: English (United States)

Formatted: English (United States)

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

phytoplankton, including both prokaryotes and eukaryotes. It differs from previous methods which rely on diagnostic pigments to derive phytoplankton groups. Our methodology provides robust estimates of the phytoplankton community structure in terms of relative cell abundance and contribution to total Chla concentration. The new generated datasets yield complementary information about different aspects of phytoplankton that are valuable for assessing the contributions of different phytoplankton groups to primary productivity and inferring community assembly processes. 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 CO_2 levels and climate.

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

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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; 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; Alvain et al., 2008). 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; Ben Mustapha et al., 2013; Alvain et al., 2008).

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 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., 2019a, b; 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.

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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.

The production of organic matter (i.e., productivity) in marine ecosystems 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) and global climate, partly through the absorption of atmospheric CO_2 and export of carbon to the deep ocean (Guidi et al., 2009; Tilman et al., 2014, Tara Ocean Foundation et al., 2022).

In order to investigate the potential impacts of environmental changes on marine ecosystem functioning (Ibarbalz et al., 2019; Henson et al., 2021), high-resolution, real-time, and global scale data on phytoplankton community structure are required (Pereira et al., 2013). However, existing knowledge about the global distribution of phytoplankton communities from in-situ observations is highly fragmented, spatially disparate, and temporally punctual. It is furthermore limited by both the challenges of in situ data collection and by the associated costs of measurement techniques, which range from microorganism imaging, flow cytometry, to DNA sequencing (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).

Ocean color remote sensing offers an interesting alternative to map the global distribution of phytoplankton communities at the sea surface at a high spatio-temporal resolution. Since 1978, ocean color satellites have been used to observe the concentration of the main phytoplankton pigment, chlorophyll-a (Chla), considered as a proxy of phytoplankton biomass (O'Reilly et al., 1998; Sathyendranath et al., 2014). Recently, ocean color data have also been used to gain information about phytoplankton communities, such as their size structure, and their taxonomic or functional composition. This interest has facilitated the integration 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 correspond to categories linked to biogeochemical processes (e.g., silicifiers, calcifiers) and physiological adaptations to environmental factors (e.g., light, nutrients, turbulence), or to more practical categories quantified using a particular analytical technique (e.g., pigment types) (IOCCG report N 14). Specialized algorithms applied to ocean color data have consequently been developed to detect specific taxa with distinctive optical characteristics (Brown, 1995; Iglesias-Rodríguez et al., 2002), or the abundance of phytoplankton functional types and size classes (Alvain et al., 2005; 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; Alvain et al., 2008).

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

The diagnostic pigment analysis method (DPA, Vidussi et al 2001) relies on the association of secondary phytoplankton pigments with different broad taxonomic phytoplankton groups. DPA classification was later refined by Uitz et al., (2006) who gave different weightings to the diagnostic pigments to retrieve three phytoplankton size classes (PSC) from total Chla. The advantage of this method is that phytoplankton pigments can be measured in a cost-effective manner through high performance liquid chromatography (HPLC). Today, large in-situ HPLC datasets are available with broad spatial and temporal coverage. These HPLC datasets have enabled the development of several DPA-based ocean color algorithms, which has made it possible to evaluate the abundance of different phytoplankton groups and size classes from ocean color satellite data (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., 2019a, b; Xi et al., 2020). However, the limitation of the DPA approach is that it is associated with large uncertainties in the classification of phytoplankton due to the presence of certain pigments in different phytoplankton taxa and cell size classes, which also vary with acclimation to light, temperature, and nutrient availability (Brewin et al., 2014; Chase et al., 2020).

In this work, we propose an alternate approach to develop an ocean color algorithm for phytoplankton group detection from in-situ metagenomic observations. The approach is ground-truthed on data collected by *Tara* Oceans, which constitutes the most comprehensive and harmonized molecular dataset available on phytoplankton taxonomic community structure on a global scale. More specifically, we used metagenomics reads to extract the global-scale distribution and abundance of the single-copy gene *psbO*, which is present across all phytoplankton groups and that provides an unbiased picture of phytoplankton cell abundances (Pierella Karlusich et al., 2022). We used these data, together with satellite-derived optical, physical and biogeochemical parameters to train an unsupervised machine learning algorithm able to discern the non-linear relationship between phytoplankton taxonomic community structure and data derived from satellites. This new algorithm allowed us to derive the spatio-temporal variability of seven phytoplankton groups between 1997 and 2021. We then compared the performance of this new algorithm with that of two previous DPA-based algorithms (El Hourany et al., 2019a; Xi et al., 2020).

2. Material

In this section, we present the datasets that were used for training the algorithm and for evaluating the outputs. The input dataset includes the in-situ distribution and abundance of phytoplankton groups inferred from metagenomics data from *Tara* Oceans and their associated satellite matchups. The outputs of the new algorithm are compared to a global dataset of in-situ HPLC diagnostic pigments, as well as with estimates from two DPA-based remote sensing algorithms.

2.1. Input dataset

2.1.1. ~~Metagenomic~~In-situ 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~~

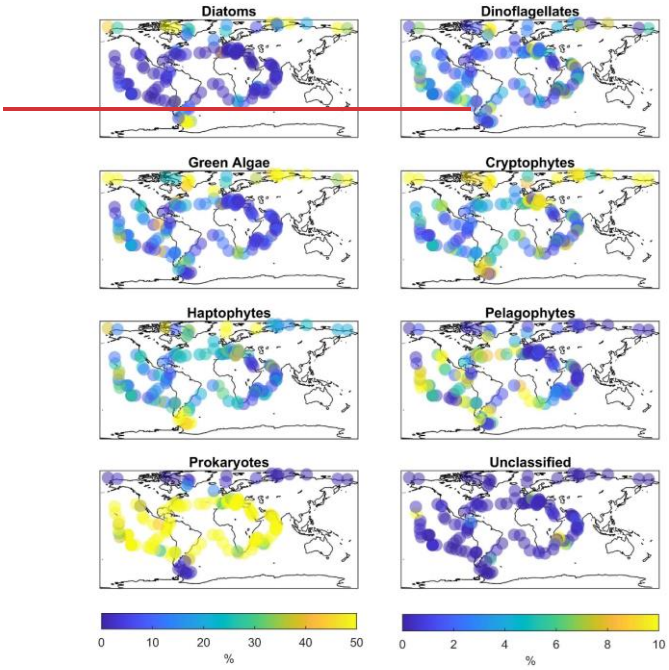
Formatted: Font color: Black

Formatted: Font: Not Bold, Font color: Black

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~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 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.~~



The *psbO* gene encodes the manganese-stabilizing protein, of around 270 amino acids, which constitutes a core subunit of photosystem II (PSII) and is unique to organisms carrying out oxygenic photosynthesis. The *psbO* gene is a single-copy gene in the vast majority of eukaryotes and prokaryotes. We used *psbO* reads from the metagenomes generated from the *Tara* Oceans expedition as a proxy of phytoplankton relative cell abundance (Pierella Karlusich et al., 2022). Among the 211 *Tara* Oceans stations, 145 stations sampled *psbO* reads in different ocean regimes from oligotrophic to eutrophic waters (Chla from 0.01 to 10 mg.m^{-3}, median at 0.3 mg.m^{-3}), from 2009 to 2013. Seawater samples were filtered in order to differentiate five planktonic size fractions (0.22-3um, 0.8-5um, 5-20 um, 20-180 um, 180-2000 um). For the purpose of this study, we pooled the five size fractions

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

into a single aggregated sample, correcting by the different sampling volumes for each size fraction, and discarded stations where not all size fractions were available, to avoid biasing the results.

psbO data enabled us to taxonomically differentiate seven phytoplankton groups: diatoms, dinoflagellates, green algae, haptophytes, pelagophytes, cryptophytes, and prokaryotes (Cyanobacteria) (Fig. 1). The *psbO* read abundances of these seven groups are expressed as relative phytoplankton cell abundance (%). Phytoplankton that were not assigned to any of these seven groups (Unclassified) represented less than 5% of the total phytoplankton community.

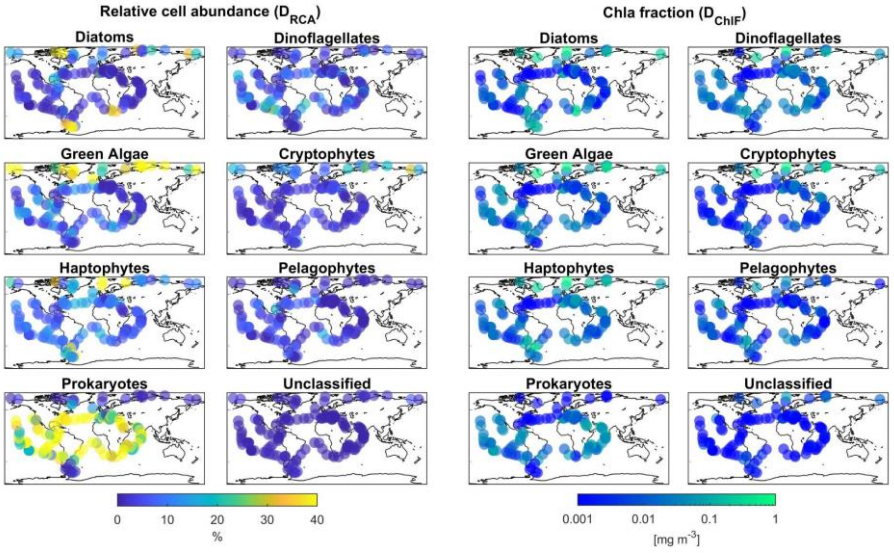


Figure 1. ~~***Figure 1.***~~ *Global biogeographical patterns of marine phytoplankton relative cell abundance and Chla fraction per group based on psbO* ~~*metagenomic*~~ *reads obtained from metagenomes from seawater samples collected* ~~*by*~~*during the* Tara *Oceans expeditions.* ~~*Note that different color scales are used for*~~

In addition to the use of *psbO* as a proxy of relative cell abundance, we also estimated the Chla proportion of the most abundant phytoplankton groups ~~*(left panels) and*~~ . For this, the relative *psbO* read abundances were weighted by their size fraction and then multiplied by the in-situ value of Chla measured at each *Tara* Oceans station. This conversion from *psbO* reads to Chla gives the contribution of each phytoplankton group to the total Chla, by accounting for cell size. We should note however that filters may retain cells smaller than the nominal pore size because of net clogging, being trapped in fecal pellets, as well as being present as symbioses and colonies. This has been observed with prokaryotic pico-sized cells such as *Synechococcus* and *Prochloroccocus* being over-represented in the 180-2000 um size fraction (Fig. 2). To minimize this impact, we based our size-weighting on 4 size-fractions, while excluding the 180-2000 um size range following the protocol in Sommeria-Klein et al., 2021. Chla fraction per group is expressed as follows:

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

$$Chla\ fraction_{PFT} = \mathrm{Chla}_{in-situ} * \frac{\sum_{s=1}^{4}\left(\frac{psbO_{PFT} * size_s}{\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}\right)}{\sum_{s=1}^{4}\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}$$

where PFT is a designated phytoplankton group, and size is the four used size fractions.

There are hence two levels of information derived from the molecular dataset; relative abundance of *psbO* reads as a proxy of relative cell abundance, and the fraction of Chla that each group represents. Both types of information have different implications. Chla is often used as a proxy of biomass, which is a relevant parameter for ~~*less abundant groups (right panels).*~~energy and matter fluxes (e.g., food webs, biogeochemical cycles), while cell abundance corresponds to species abundance for unicellular organisms, which is an important measure for inferring community assembly processes.

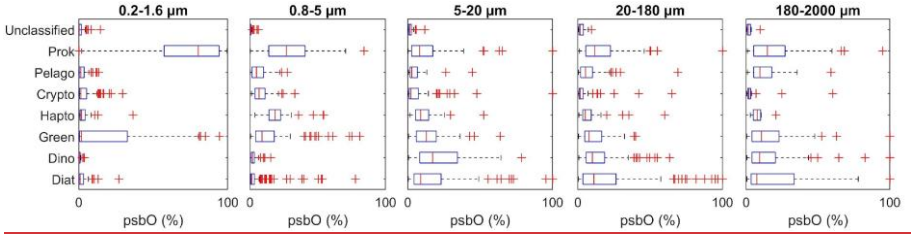


Figure 2. *Relative abundance of psbO reads as a proxy of phytoplankton group cell abundance observed in each size fraction. The boxplots represent the distribution of each group and each panel shows the different size fractions.*

2.1.2. Satellite ~~derived Dataset~~ datasets

We used ocean color ~~satellite data~~products 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~~ day, downloaded from the Globcolour portal. These products were constructed by merging 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~~We used sixteen Globcolour products ~~were used, at daily and at 4km spatio-temporal resolution:~~: Chlorophyll-a concentration (Chla), Remote sensing reflectances at ~~4~~11 wavelengths (~~Rrs412, Rrs443, Rrs490, and Rrs555), satellite Chla,~~412 till 670 nm), light attenuation coefficient at 490 nm (Kd490), photosynthesis available radiation (PAR), Normalized fluorescence light height (NFLH) and particulate backscattering at 443 nm (bbp). These products have daily and 4km spatio-temporal resolution. In addition, we used the Climate Change Initiative Sea Surface Temperature (SST) ~~data~~product at 4 km resolution and ~~at~~ daily frequency ~~available from~~distributed by the Copernicus Marine Services (CMEMS) portal.

2.2. HPLC datasets

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

To compare *psbO*-derived phytoplankton group distributions with more conventional, DPA-based products, we compiled a global HPLC dataset regrouping 12 000 HPLC observations from several HPLC datasets between 1997 and 2014 (Fig. 3): MAREDAT, NOMAD, SeaBASS, and other oceanographic campaigns: Labrador, Gep&co, Polarstern, BROKE-West, SAZ-Sense Voyage (Luo et al., 2012; Werdell and Bailey, 2005; Dandonneau et al., 2004; Bracher et al., 2015; Fragoso et al., 2016; Peloquin et al., 2013; Wright et al., 2010; de Salas et al., 2011). This HPLC dataset was collocated with satellite Globcolor and CCI matchups. It depicts the abundance of the pigments most widely used to identify major phytoplankton groups: Fucoxanthin (Fuco), Peridinin (Perid), Alloxanthin (Allo), Zeaxanthin (Zea), Chlorophyll-b (Chlb), 19'-Hexanoyloxyfucoxanthin (19HF), and 19'-Butanoyloxyfucoxanthin (19BF) (Table 1). To estimate Chla fraction for each phytoplankton group, namely diatoms, dinoflagellates, haptophytes, green algae, cryptophytes, pelgophytes and prokaryotes, diagnostic pigments were used. The Chla fraction per group is expressed by $Chla_{PFT} = (DP * \alpha)/\sum DP * \alpha$. Three sets of coefficients are proposed for a global ocean application and are presented in Table 1 (Uitz et al., 2006; Soppa et al., 2014; Brewin et al., 2015). Therefore we calculated an average Chla fraction value for each phytoplankton group using the three sets of coefficients.

Simultaneously, T*ara* Oceans HPLC measurements (Pesant et al., 2015), which are available for the same stations and sampling time as for *psbO*, were considered to evaluate the correspondence between pigments and *psbO*-derived phytoplankton groups.

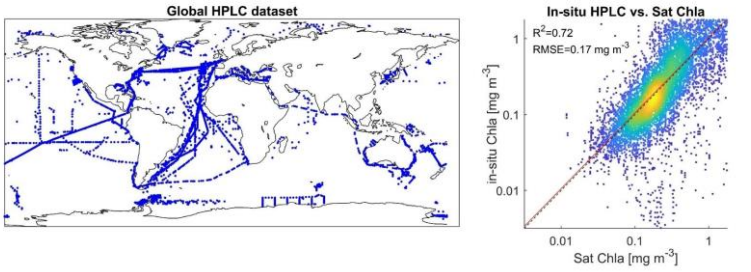


Figure 3: *Geographical location of the global HPLC dataset stations regrouping observations from 1997 to 2014. The right panel shows the correlation between in-situ HPLC Chla measurements and its matchup in the Globcolour Chla product.*

2.3. PFT satellite products

In order to compare the outputs of our method to those of existing DPA-based remote sensing algorithms, we used two of them :

2.3.1. CMEMS phytoplankton Chla fraction

This Globcolour 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 to present. This

Formatted: English (United States)

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

algorithm estimates the Chla concentration of diatoms, dinoflagellates, haptophytes, green algae, and prokaryotes. The algorithm was implemented using HPLC-based phytoplankton groups using the 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 distributed by CMEMS (product number: OCEANCOLOUR_GLO_BGC_L3_MY_009_103).

2.3.2. SOM 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 (SOMs), an unsupervised neural network, and was calibrated using the HPLC dataset described above.

The SOM-Pigments algorithm applied to Globcolour products allowed us to estimate the concentration of ten phytoplankton pigments (Chlorophyll-a (Chla), Divinyl-Chlorophyll-a (DVChla), Chlorophyll-b (Chlb), Divinyl-Chlorophyll-b (DVChlb), 19'Hexfucoxanthin (19HF), 19'Butfucoxanthin (19BF), Fucoxanthin (Fuco), Peridinin (Perid), Alloxanthin (Allo), Zeaxanthin (Zea)) at the global scale from 1997 to 2021. We then used the coefficients in Table 1 to convert pigments into the Chla concentration of five phytoplankton groups, namely diatoms, dinoflagellates, haptophytes, green algae and prokaryotes.

Table 1. *Phytoplankton groups and size classes associated with their diagnostic pigments and coefficients.*

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

3. Methods

Formatted Table

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Several machine learning algorithms were used in this study. The algorithm to estimate phytoplankton groups from satellite data was built using SOM (Kohonen, 2013) and topology-constrained organization. This allowed us to confirm the non-linear relationships between phytoplankton group composition and satellite data through topology conservation. Next, we used the Ascending Hierarchical clustering algorithm to identify the large scale patterns generated by SOM. This allowed us to emphasize the predominant data structure learned by SOM and to characterize phytoplankton biomes. Finally, to characterize the differences between the DPA- and *psbO*-based approaches, we used Random Forest models to highlight the cumulative importance of a pigment composition to estimate a phytoplankton group abundance. In the following section, each methodology and algorithm are explained in detail.

~~2.1.3.~~3.1. Structure of the ~~Training~~training and test databases~~:~~

The initial dataset (D) ~~is constituted~~consists of ~~211~~the 145 *Tara* Oceans ~~*psbO*~~ observations of ~~the~~*psbO* relative abundance of the seven defined phytoplankton groups ~~with~~, the Chla fraction per group, and the associated matchups of ~~10~~17 satellite-derived parameters (Chla, SST, 411 Rrs (412~~, 443, 490, and 555 nm~~ 645nm), NFLH, Kd490, PAR, and bbp). ~~Even though the values are negligeable, the~~The unclassified phytoplankton fraction was also ~~added~~considered, despite negligible values, 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~~ (El Hourany et al., 2019a, b).

We built two sub-datasets, the first (D_{RCA}) relating *psbO*-~~based abundances were multiplied by the in-situ value~~derived relative cell abundance of ~~Chla measured as well at each *Tara* Oceans station, expressing every~~the seven defined phytoplankton groups to the 17 satellite-derived parameters, and the second (D_{ChlF}) joining *psbO*-derived Chla fraction per phytoplankton group ~~as a function~~and the same 17 satellite-derived parameters. We then constructed two algorithms, using either D_{RCA} or D_{ChlF}, both based on the same SOM methodology described below. Following the positioning of ~~a Chla fraction.~~ Tara Ocean's stations, and the distribution of Chla values within both datasets (Fig. A2), both algorithms are suitable for case 1 waters applications (i.e. open ocean).

All variables were normalized by their variance to homogenize weights. The ~~hypothesis~~rationale behind this is that the phytoplankton community should be treated as a whole~~, and therefore~~; consequently, the variability of each phytoplankton group is dependent on each other in a relative way. ~~In parallel, D presents~~D_{RCA} and D_{ChlF} both present missing values (~~table 1); satellite missing values are usually linked to~~ Table 2), most likely due to cloud coverage or ~~masked data due to~~ coastal/ice ~~influence. However,~~presence/proximity. In-situ *psbO*-based observations also contained missing values ~~within the in-situ observation are~~ due to ~~a lack~~an absence of certain measurements ~~of~~at a given station. ~~And since D~~Since the in-situ dataset contains a low number of observations (~~211~~145 stations), every observation is valuable. ~~Therefore,~~In order to ~~tackle these~~overcome the several limitations~~, it is essential to use a non-linear multivariate regression~~ faced with this training dataset, we used the SOM algorithm that can deal with missing values and allow a robust generalization in the case of limited observations~~.~~ (Jouini et al., 2013).

Table ~~1.~~2. *Percentage of missing values within the initial database ~~(D).~~*

Formatted: Font color: Black

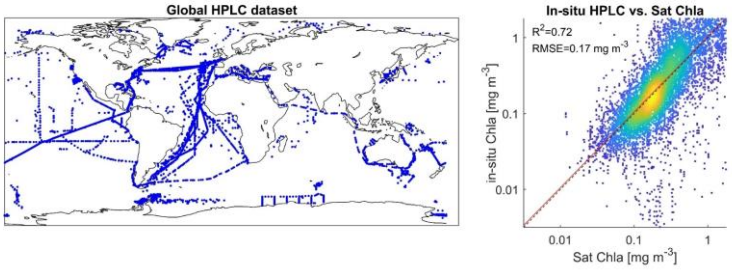
Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Tara Oceans	*psbO*		Sat										
D (~~211~~145 stations)	Relative cell abundance	Chla fraction per group	Chla	Rrs 412-645 nm	~~Rrs 443~~	~~Rrs 490~~	~~Rrs 555~~	SST	bbp	Kd490	NFLH	PAR
Percentage of missing values	~~31~~	7%	18 %	~~45~~	43-53%	~~43~~	~~43~~	30 %	55 %	53%	37%	14%

~~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).~~

~~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.~~



Split Cells
Formatted: Font: Not Italic
Formatted Table
Deleted Cells
Deleted Cells
Deleted Cells
Inserted Cells
Formatted: Font: Not Bold, Font color: Black
Deleted Cells
Deleted Cells
Inserted Cells
Formatted: Font: Bold, Font color: Black
Formatted: Space After: 0 pt, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Formatted: English (United States)
Formatted: Font color: Black
Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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.

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 regression 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.

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

Phytoplankton size class	Phytoplankton group	Diagnostic Pigment (DP)	Coefficients (α)*		
			Uitz et al., 2006	Soppa et al., 2014	Brewin et al., 2015

Formatted Table

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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 (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; $Chla=\sum \alpha DP$					

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.

3.1.3.2. Self-Organizing map applied to *Tara* Oceans *psbO* data (SOM-psbO):

3.1.1.3.2.1. Training General concept of the SOM psbO

The SOM algorithm (Kohonen, 2013) aims to cluster a is utilized for clustering multidimensional database (D) intodatabases by assigning them to classes represented by a fixed network of neurons (the SOM map). This network is used to defineknown as the SOM map. The SOM map consists of a rectangular grid of p x q neurons and defines a discrete distance between the neurons of the map, which present the shortest path between two neurons. Moreover, SOM enables, enabling the partitionpartitioning of D in which eachthe dataset. 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 bevector. Observations in the dataset are assigned to the closestnearest neuron.

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Normal, Space Before: 0 pt, Outline numbered + Level: 2 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.5" + Indent at: 1", Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font color: Black

Formatted: Normal, Outline numbered + Level: 3 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.75" + Indent at: 1.25", Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~in the sense of~~ based on the Euclidean Norm ~~(EN).~~. A ~~fundamental property~~key feature of ~~a~~ SOM is ~~the~~its ability to provide topological ordering ~~provided at the end of the clustering phase: two~~, where close neurons on the map ~~represent data that are close~~ correspond to similar observations 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~~are determined through a minimization process ~~depending~~of a cost function that depends on the distance between the neuron and its assigned observation.

SOMs have ~~frequently~~ been ~~used in the context of completing~~widely employed to complete missing data ~~(Jouini et al., 2013). Under these conditions, the distance between an observation c D and the neuron's vectors of the map is the Euclidean distance~~, utilizing the truncated distance (TD) that considers only the existing components ~~(the truncated distance or TD hereinafter). The use of the TD allows~~of the observation's vector, thus allowing for ~~considering the~~the integration of incomplete 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:~~

3.2.2. ~~a quantification error and a topographic error: the quantification~~Training phase

The implementation of the SOM methodology is summarized in Fig. 4 and 5. Briefly, we first split the *Tara* Oceans *psbO* datasets so as to obtain 80% of the data to train the SOM, and 20% of the data as a test set, the latter consisting of 30 observations with complete *psbO* information. We did this separately for D_{RCA} and D_{ChlF} sub-datasets so as to generate SOMRCA, which stands for the algorithm specialized in relative cell abundance estimation, and SOMChlF for the algorithm specialized in Chla fraction per phytoplankton group.

During the SOM training, different combinations of satellite variables were used to determine the best set of variables to estimate the 7 phytoplankton groups in terms of relative cell abundance and Chla fraction. For each combination of variables, we increased the number of neurons from 10 to 1000 neurons to determine the optimal size of the SOM. For each SOM obtained, we quantified quantization and topographic errors.

The quantization error represents the difference between an observation ~~D~~ and its closest neuron. This error is monitored during the training procedure ~~till~~until it reaches stability at a minimum value with increasing training epochs. This is where the training should stop to prevent overfitting. The quantization error is expressed as follows:

Formatted: Font: Times New Roman, 10 pt, Bold

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

$$qe = \frac{1}{n}\sum_{i=1}^{n} \|x(i) - wc(i)\|$$

where x(i) is the vector of an input observation i; wc(i) is the closest neuron's vector of sample x(i); n is the number of observations.

However, the topographic error is a representation of having, for each observation of the database, distant first and second best-matching neurons. ~~This~~ and is expressed as follows:

$$te = \frac{1}{n}\sum_{i=1}^{n} d(x(i))$$

where d(x(i)) = 1 if the first and second closest neuron to x(i) are not adjacent, else d(x(i)) = 0.

- Minimizing 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~~A one- leave-out ~~procedure, where each observation of D is used iteratively either as a test or for training. Therefore, at each~~ cross-validation procedure was performed to assign performance metrics (R^2 and RMSE) to help choose the best combination of SOM size and satellite variables. At each iteration of the cross-validation procedure, we ~~calculate~~chose randomly one observation as a test, whereas the other observations served to train the SOM with the given grid size. We calculated the closest neuron to the test observation based on ~~basis of~~ its satellite variables only and ~~associate~~associated these latter with the neuron's seven phytoplankton groups vector. When all the observations were used as a test, we ~~calculate~~calculated a mean R^2 and an RMSE, associated with the given size map, while comparing the estimated and ~~predicted~~observed 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.~~

~~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.~~

The ~~**Operational phase**~~best SOM configuration and variable combination are based on an optimum where *te*, *qe* and the RMSE are in low ranges while avoiding overfitting. The chosen SOM was tested using the 20% test set, providing independent performance metrics to evaluate the generalization of the chosen SOM. As a result, we present in the paper the performance metrics of the chosen SOM configuration based on the cross-validation procedure and the test set.

The optimal combination of satellite parameters for the SOMRCA and SOMChlF algorithms was determined to be Chla, SST, Rrs at four wavelengths (412, 443, 490, and 555 nm), bbp, and Kd490. The grid size for SOMRCA

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

was set at 242 neurons, while SOMChlF had a grid size of 222 neurons. This selection was based on several factors, including a high regression coefficient between estimated and observed phytoplankton values, low error values of quantization and topographic error, and a low global RMSE encompassing all phytoplankton groups. The choice of Rrs bands aligns with previous work conducted on the PHYSAT method by Alvain et al., (2005) and Ben Mustapha et al. (2013). The PHYSAT method utilizes reflectance anomalies in the same four selected bands to identify dominant phytoplankton functional types. It should be noted that the Rrs bands selected, including the additional 670 nm band, are commonly measured by all sensors used to build the Rrs product of Globcolour. This overlapping of different sensors enhances data availability and coverage, thus increasing the importance of these Rrs bands within the initial dataset. The inclusion of the Rrs at 670 nm did not significantly impact the performance of either SOMRCA or SOMChlF, primarily due to the open ocean nature of the dataset. In the clear open ocean, the information contained in the remote sensing reflectance (Rrs) bands beyond 555 nm is limited due to the strong absorption by water, as noted by Xi et al. (2015).

Through the iterative training process described above, the results show a significant increase in the general performance of the method when the number of neurons increases to a certain extent (Fig. 6). Using a number of neurons larger than the training dataset still allows a refined discretization. In this case, some neurons will capture a sample of the database, which permits to define a referent vector for these neurons. When the neuron did not capture any data observation, the discrete distance between the neighboring neurons was 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. However, the quantization error's lowest values above 350 neurons might indicate overfitting.

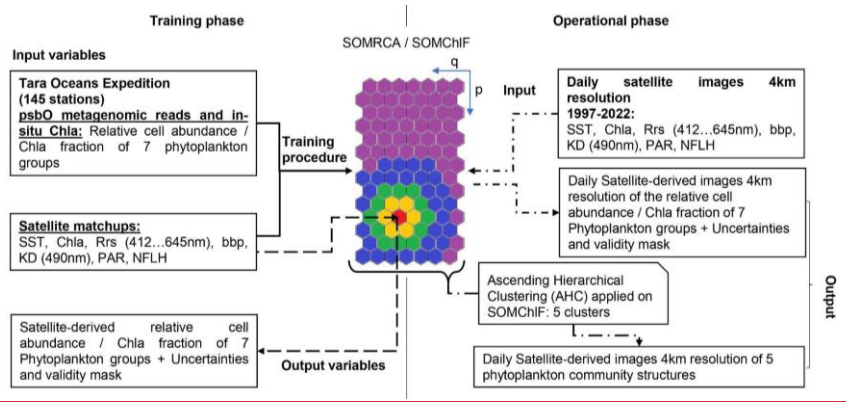


Figure 4. *General flowchart of the SOM methodology applied on both D_{RCA} and D_{ChlF} to construct SOMRCA and SOMChlF.*

Formatted: Font: Not Bold

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

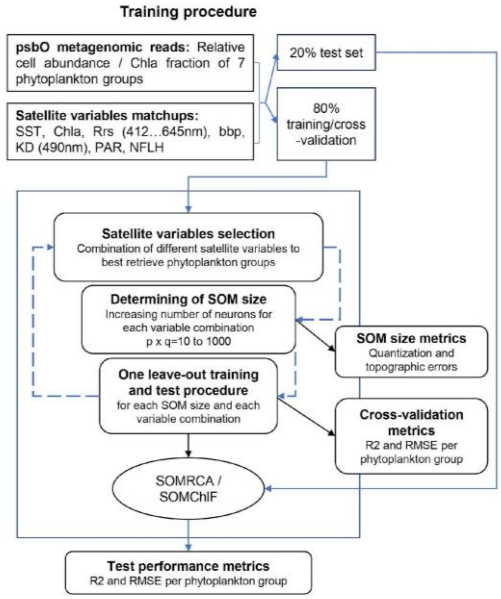
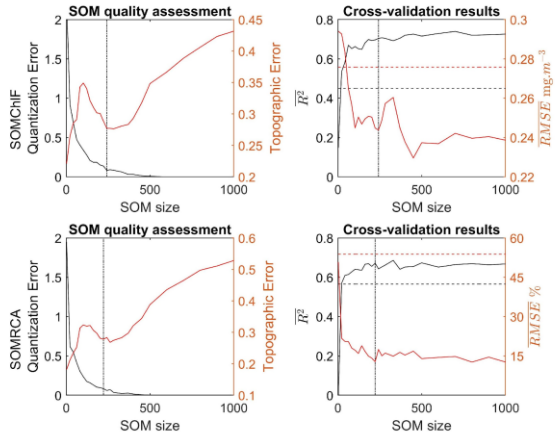


Figure 5. *Detailed training procedure of the SOM methodology applied on both DRCA and DChlF.*



~~3.1.3.~~ ~~***Figure 6.***~~

~~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.~~

Formatted: Font: Bold

Formatted: Font: Bold,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~In the second phase, which is~~*Quality assessment based on the quantization and topographic error related to the training of the SOMChlF and SOMRCA as a function of increasing SOM size (number of neurons) using* Chla, SST, Rrs at 4 wavelengths (412, 443, 490 and 555 nm), bbp and Kd490. *In parallel, the average regression coefficient and the root mean squared error as a function of increasing SOM size were calculated through a "one-leave out" cross-validation procedure. The dashed black and red lines correspond, respectively, to the* R^2 *and the RMSE using the "K-nearest neighbor" algorithm. Finally, the dotted lines correspond to the chosen SOM size for SOMChlF=242 neurons and SOMRCA=222 neurons.*

3.2.3. Operational phase

During the operational phase, we ~~estimate~~estimated the phytoplankton group variability using ~~different~~ the best combination of satellite ~~images~~parameters. The set of ~~ocean satellite observations~~parameters of a pixel ~~is~~was projected onto the SOM ~~psbO~~. In doing so, the ~~projected~~ parameters ~~are~~at each pixel were normalized ~~with the corresponding variances of D~~by the variance of that same parameter within the initial training dataset to maintain an equal weight among the parameters and ~~are~~were assigned with the closest best-matching neuron using the truncated distance. At the end of the assignment phase, each pixel ~~is~~was 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~~relative cell abundance~~, each Chla fraction of a neuron is divided by~~ in the ~~total~~case of SOMRCA, or Chla ~~value~~fraction in the case of ~~this same neuron.~~SOMChlF. Since the training was undergone ~~using~~for the whole phytoplankton ~~structure~~community 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~~groups at the same spatio-temporal resolution.

3.2.4. ~~However, since~~Masking and uncertainty evaluation

Given that our initial dataset ~~D~~ is of limited ~~to a short number of data,~~size, it is possible that it does not contain certain 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~~. In order to prevent abnormal predictions ~~to~~for cases ~~that are~~ not ~~seen within dataset D. A~~ observed in the initial dataset, we conducted a quality evaluation of the method's output. This evaluation involved quantifying a reliability index ~~was calculated between 1997 and 2021~~ by ~~testing~~comparing the set of satellite parameters' values ~~obtained for~~at a ~~given~~particular pixel ~~against~~with the values ~~in the original dataset; if~~ of the same parameters in the initial dataset. If a satellite variable's value ~~of a variable falls out of~~fell outside the ~~bond~~range defined within the initial dataset by the ~~+ 2 Std around the~~ mean value of the same variable's distribution~~, the variable is marked as~~ plus or minus two standard deviations, it was considered distant. This ~~is~~evaluation was performed ~~on~~for all ~~ten~~ satellite variables per pixel~~. The~~, and the reliability index ~~is calculated~~was determined by dividing the number of ~~distant~~accepted 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~~. A higher reliability index ~~indicate that~~ indicates greater reliability of the method ~~is reliable, and more care should be given to~~, while regions ~~where the~~with lower reliability index ~~is larger.~~ values require additional attention.

Formatted: Font: Times New Roman, 10 pt

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

In the context of the global ocean, numerous uncertainties are associated with satellite parameters and regions. The SOM algorithm is known to effectively reduce noise and mitigate the impact of uncertainties within the dataset (de Silva et al., 2013). However, the main source of uncertainty in the estimation process stems from selecting the best matching neuron. This involves finding and associating the closest neuron in the SOM with a new or unfamiliar observation, such as a satellite pixel. Due to the topology conservation, a pixel could be assigned to several close neurons, forming a neighborhood along a distance gradient. Consequently, a single satellite observation can represent various probabilities of phytoplankton group combinations.

To account for all uncertainties in the estimations, we opted to associate each pixel and phytoplankton group (based on relative cell abundance or Chla fraction) with a weighted standard deviation derived from the values of the ten closest neurons. The weights were determined by the distances between the first ten matching neurons and the pixel. This approach allowed us to incorporate uncertainties into the assignment process and provide a confidence measure for each pixel's assignment.

By considering both the reliability index and the weighted standard deviation, we could assess the influence of uncertainties in the satellite variables.

3.3. Characterisation of phytoplankton biomes

~~3.2.~~ To emphasize the predominant data structure learned by SOMChlF, the Ascending Hierarchical Clustering ~~applied on SOM-psbO~~

~~A hierarchical ascending clustering (HAC~~algorithm (AHC) 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.~~ to characterize phytoplankton biomes on the basis of their Chla fractions (a proxy of a phytoplankton group's biomass) and optical signature.

~~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 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.~~

The HAC is a bottom-up clustering algorithm. The HAC starts with 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

Formatted: Font: Not Bold,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

known a priori but at the end of the SOM+HAC procedure, several possibilities of a number of clusters to be taken into account were revealed. A compromise was made between the number of clusters we could explain from a physical point of view and the number of clusters for which we needed 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 SOMChlF was associated with a cluster. The association of several neurons in a cluster allows us to identify common phytoplankton community structures, and therefore characterize phytoplankton biomes. Upon applying SOMChlF as described in the operational phase section, each pixel of a satellite image could be associated with a cluster.

~~3.3.~~3.4. Evaluation of ~~the importance of~~ pigments to estimate phytoplankton groups ~~using random forest~~

Each ~~*psbO* derived~~ phytoplankton ~~group's~~group's *psbO* abundance was associated with its corresponding HPLC pigments measurement performed on the same *Tara* ~~Ocean's~~Oceans station. The ~~importance~~ability of pigments to predict phytoplankton groups was ~~computed~~evaluated using a ~~Bagged Random Forest~~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~~decision trees, each constituted of internal nodes and leaves. In the internal node, the selected feature (i.e.~~.~~.. pigment in this case) ~~is~~was 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~~Table 1 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~~SOMRCA and SOMChlF algorithms

~~Ten~~Cross-validation of different 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~~

Formatted: Font color: Black

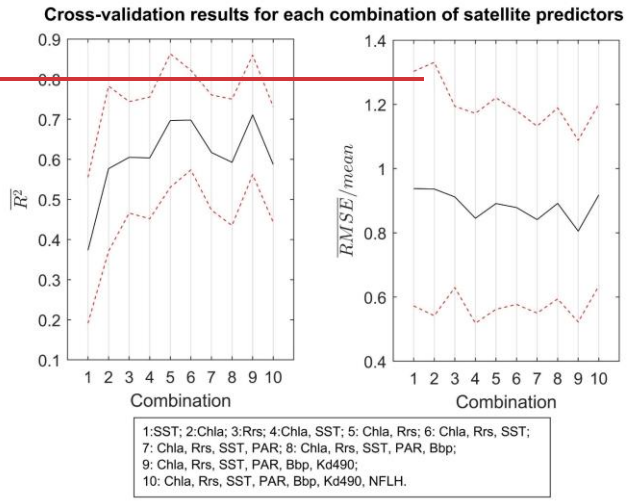
Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~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~~grid size revealed a performance of an average R^2 of 0.~~71 distributed between a maximal R^2= 0.8668~~ for ~~the green algae,~~SOMRCA and ~~a minimal R^2 =~~of 0.~~5674~~ for ~~the dinoflagellates and cryptophytes~~SOMChlF (Fig. ~~5~~7, 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~~83) and therefore a preservation of the initial phytoplankton quantity expressed in total Chla.



Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

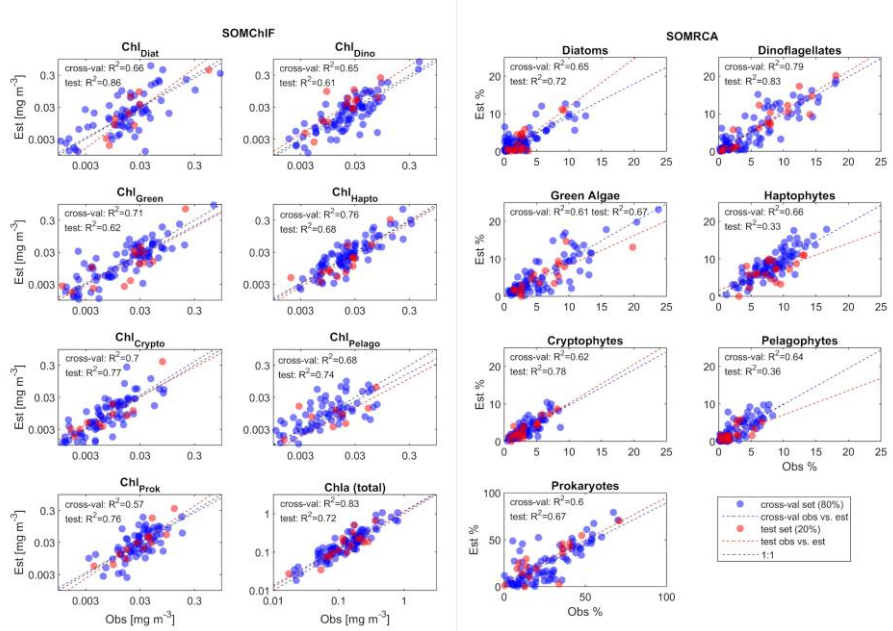
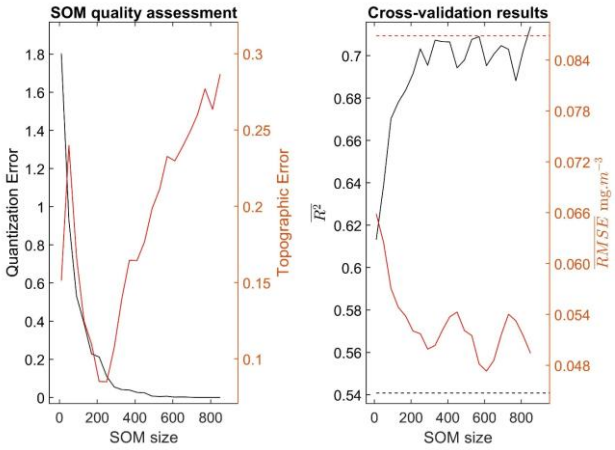


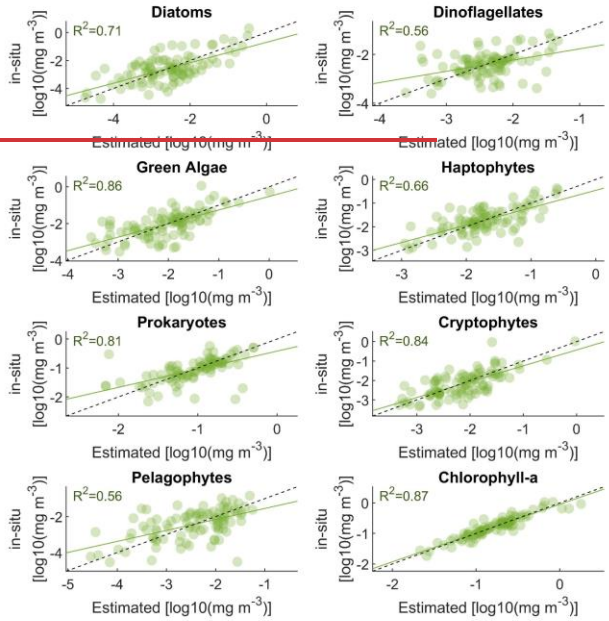
Figure *~~3: Cross~~7. Results of the two-step cross-validation ~~results expressed in terms of mean regression coefficient~~(blue) and ~~mean relative RMSE~~test (red) procedures for ~~each~~SOMChlF (left) and SOMRCA (right) with the chosen best* ***combination of satellite parameters and*** *~~for all the phytoplankton groups. The red dashed line d~~elimits 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.~~*



Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~a **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 correspondsgrid,~~ respectively ~~to the R^2 = 0.54 and the RMSE=0.029 mg m-3 using the "K nearest neighbor" algorithm.~~



~~, of 242 and 222 neurons. For the **Figure 5.** Results of the "one-leave out" cross-validation procedure. Each~~ , ~~each~~ observation ~~is~~, ~~among 80% of the initial data, was~~ used iteratively as a ~~train or~~training set ~~and~~ as a test set until all observations ~~serve~~served as tests. ~~(blue dots). This procedure was used to identify the best satellite combination and SOM grid size. Finally, the remaining 20% was used as a test to evaluate the generalization capacity of the SOM with the chosen configuration (red dots).~~

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

Phytoplankton group	~~Cross-validation~~ ~~Tara Oceans~~ SOMRCA Relative cell abundance (%)		~~HPLC-based Validation~~ ~~Global~~SOMChlF Phytoplankton chlorophyll-a fraction (mg m^{-3})	
	Cross-val	Test	Cross-val	Test

Formatted: Font: Not Bold

Formatted: Font: Bold

Formatted: Font: 9 pt, Not Bold, Font color: Black

Formatted: Widow/Orphan control, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font: 9 pt, Not Bold, Font color: Black

Formatted: Font: 9 pt, Not Bold, Font color: Black

Formatted: Widow/Orphan control, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Widow/Orphan control, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

	R^2	RMSE ~~[mg m^{-3}]~~(%)	~~MAE [mg m^{-3}]~~R^2	~~Rsp~~RMSE (%)	R^2	RMSE ~~[~~(mg m^{-3}~~]~~)	~~MAE [mg m^{-3}]~~ R^2	~~Rsp~~RMSE (mg m^{-3})
Diatoms	0.~~71~~65	~~0.074~~2.70	0.~~038~~72	~~0.76~~2.00	0.~~7~~866	0.~~29~~25	0.~~2~~086	0.~~85~~04
Dinoflagellates	0.~~56~~79	~~0.010~~5.45	0.~~006~~83	~~0.57~~6.15	0.~~5~~565	0.~~033~~09	0.~~015~~61	0.~~55~~04
Green Algae	0.~~86~~61	~~0.026~~5.64	0.~~031~~67	~~0.67~~4.32	0.~~5~~471	0.~~064~~03	0.~~059~~62	0.~~57~~01
Haptophytes	0.66	~~0.043~~3.42	0.~~023~~33	~~0.82~~3.04	0.~~5~~776	0.~~11~~07	0.~~085~~68	0.~~68~~01
Prokaryotes	0.~~81~~60	~~0.089~~19.27	0.~~050~~67	~~0.66~~20.53	0.~~2~~057	0.~~064~~05	0.~~1~~576	0.~~23~~09
Cryptophytes	0.~~84~~62	~~0.095~~1.98	0.~~021~~78	~~0.73~~1.98	0.~~5~~070	0.~~072~~03	0.~~031~~77	0.~~61~~01
Pelagophytes	0.~~56~~64	~~0.019~~2.60	0.~~010~~36	~~0.56~~1.96	0.4968	0.02	0.~~018~~74	0.~~58~~01
Chlorophyll-a	~~0.87~~				0.~~2~~3183	0.~~133~~23	0.~~8~~972	0.~~71~~31 ~~0.54~~ ~~0.45~~ ~~0.81~~

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~~group value distributions were compared between the ~~referent~~ vectors ~~(~~constituting the neurons~~)~~ of ~~the SOM~~ SOMRCA and ~~D.~~SOMChlF with D_{RCA} and D_{ChlF}, respectively. This analysis showed that the correlation coefficients were not altered within ~~the SOM, as well as~~SOMRCA nor SOMChlF, and neither was 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~~the initial dataset after the training procedure (Fig. ~~6~~A1 and A2).

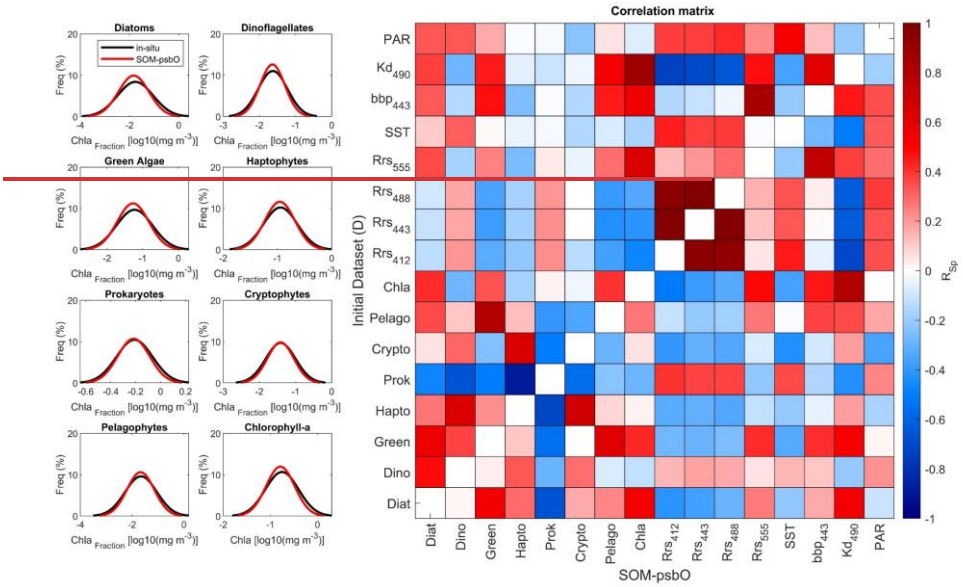


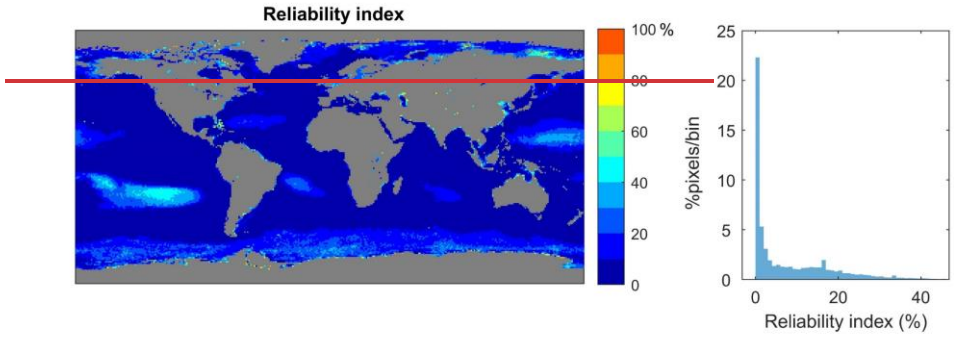
Figure 6. *Evaluation of the preservation of the initial dataset's characteristics; Left panel) distribution of the values of psbO-derived Chla fraction for each phytoplankton group and the total Chla in the initial dataset D and SOM neurons (n=270). Right panel) Spearman correlation coefficient matrix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

Formatted: Font: Bold

However, regardinggiven the limited size of the initial dataset, one should be cautious when applying SOM-psbOSOMRCA and SOMChlF to the global satellite data. Through must be done with caution. For each pixel and at each time step between 1997 to 2021, we performed 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.2.3 to illustrate the extentprovide a measure of the applicability of this method (Fig. 78). Regions of low confidence can be identified where more thanthe value of the reliability index does not exceed 40% of the pixels were masked throughout the time series between 1997 and 2021.. These regions are mainly shownfound in coastal and turbid waters, and inas well as the south pacificSouth Pacific Ocean gyre, and are characterized either by very high or very low Chla values. This result is expected sincebecause the SOM algorithm is not able tomainly adapted for case 1 waters and cannot extrapolate beyond the values' distribution of values in the initial dataset. Furthermore, moderate confidence regions can be defined in which around 2520% of the pixels fall out of the accepted bonds. And thesebounds, are highlighted by a reliability index under 80%. These regions are mainly concentrated infound at high latitudes, especially in the Southern Ocean, mainly due to the limited number of data points that sampled available samples in the area and the particular optical characteristics of that region (Mitchell et al., 1991).(Mitchell et al., 1991).

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right



Uncertainty values reached 20% relative cell abundance for SOMRCA and 0.15 mg m^{-3} of Chla SOMChlF, respectively, and in each case displayed regional patterns. Generally, uncertainty values followed the concentration gradient in Chla fraction and cell abundance per group. High latitudes exhibited the highest uncertainties for diatoms, green algae, and haptophyte relative cell abundances, while the Southern Ocean showed the highest uncertainties for prokaryotic cell abundance. The elevated uncertainty in prokaryotes within the Southern Ocean can be attributed to the limited sampling in this area, resulting in greater dissimilarity between satellite data in this region and the data sampled in the initial dataset, corroborating the findings of the reliability index. This is also consistent with the very low abundance of cyanobacteria in the area (Flombaum et al 2013), for which model uncertainty may be higher.

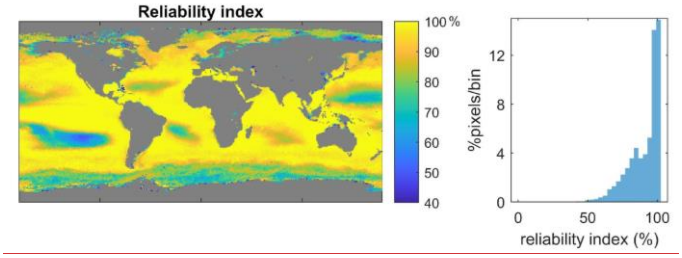


Figure ~~*7*~~*8. Applicability of the satellite psbO-based method~~*; Geographical*~~*. The geographical (left) and values distribution (right) of the reliability index were calculated between 1997 and 2021 by testing the set of* ~~*values obtained for*~~*satellite parameters at a given pixel against the values in the original dataset (D).*

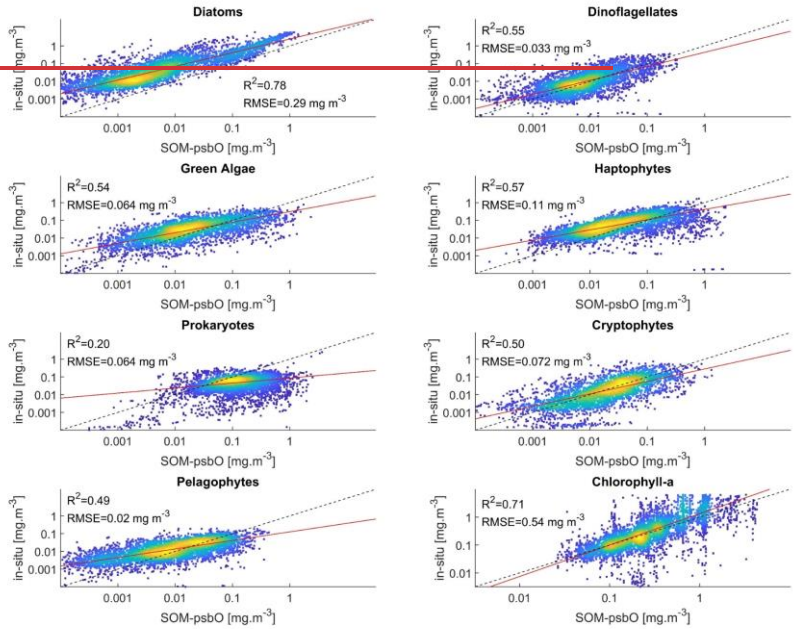
~~4.1.1.~~4.2. ~~Independent validation using~~ Comparison with global HPLC pigment dataset

The global in-situ HPLC dataset was then used to estimate Chla fractions for each phytoplankton group~~.~~ using the diagnostic pigment approach (DPA). This dataset was compared to ~~its matching phytoplankton group's~~the Chla fraction matching each phytoplankton group that was estimated ~~using SOM-psbO and satellite data~~ by SOMChlF (Fig. ~~8~~9). Evaluating the sum of Chla fractions and comparing it with in-situ Chla can be considered as a baseline

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

evaluation of this method. This comparison ~~shows~~showed a satisfying correspondence ~~scoring an~~score of R^2=0.~~7 with an RMSE of 0.17 mg m-3. A relatively~~72. Relatively good correspondence is noted for ~~the~~ diatoms and haptophytes, showing an R^2=0.~~7864 between in-situ and ~~SOM-psbO's~~SOMChlF for diatoms ~~Chla fraction.~~and 0.65 for haptophytes. Moderate correspondence ~~is noted~~was found for ~~dinoflagellates,~~ green algae, ~~haptophytes,~~ cryptophytes, and pelagophytes, with an R^2 ranging between 0.~~57~~43 and 0.~~49. The prokaryotes~~39. Prokaryotes and dinoflagellates had the lowest correspondence between both outputs. The comparison between DPA-based phytoplankton groups and ~~SOM-psbO's~~SOMChlF 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~~. For example, the ~~cross-validation performances of SOM-psbO. Diatom~~diatom Chla fraction~~,~~ is well captured by ~~this~~the latter, and the values agree with ~~the ones~~those estimated using HPLC observations~~,~~; however, we ~~note~~noted a major ~~over-estimation~~overestimation within the HPLC DPA method. For prokaryotes, ~~the satisfying cross-validation performances of SOM-psbO lead~~this comparison leads us to say that the use of zeaxanthin as an indicator of ~~cyanobacteria's abundance~~the cyanobacterial contribution to Chla may not be entirely representative of this group.



The permutation-based importance analysis using Random Forest, performed on the in-situ *Tara* Oceans *psbO* and HPLC measurements, emphasizes the necessity of a multivariate approach for predicting phytoplankton community structure based on pigments (see Fig. 10). Notably, the diagnostic pigments mentioned in Table 1 exhibited dominant importance in determining the relative abundance of their respective assigned phytoplankton

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

groups. For instance, peridinin represented dinoflagellates, Chlorophyll-b characterized green algae, and zeaxanthin indicated prokaryotes (Table 1). These pigments demonstrated the highest importance for their respective groups, as illustrated in Fig. 10, accompanied by a positive Spearman correlation. However, individually, these pigments accounted for less than 25% of the variance in their respective groups. Conversely, in the case of cryptophytes, diatoms, and haptophytes, no pigment stood out in terms of importance, and the observed correlations were related to co-variation between pigments (e.g., Chlb and Fuco in diatoms), possibly influenced by Chla variability. Therefore, the variability within each group is best explained not by a single diagnostic pigment, but rather by the overall pigment composition. It is crucial to consider how natural variability can influence the interpretation of pigment composition in relation to phytoplankton community structure. Pigment ratios not only vary with phytoplankton composition but also reflect the diverse strategies employed by different phytoplankton types to acclimate to environmental factors such as light, temperature, nutrients, and other variables.

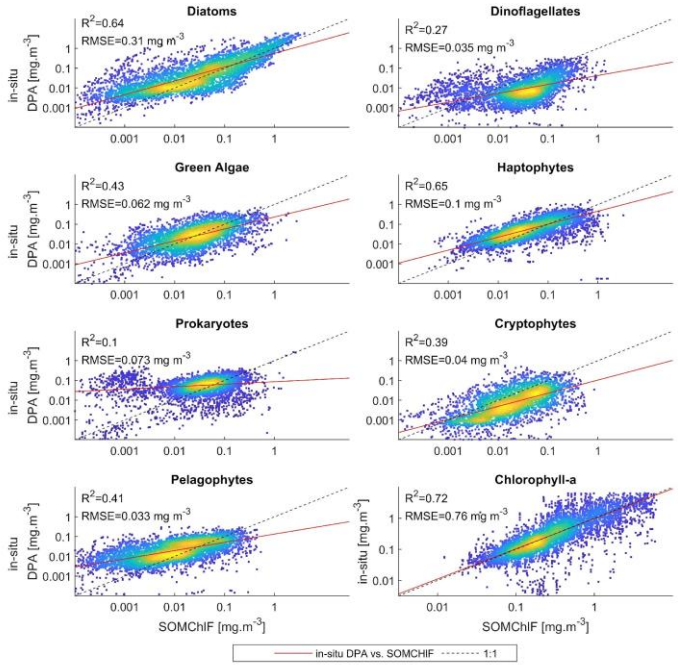


***Figure 8**: Evaluation of SOM-psbO method while comparing**9**: Comparison between the outputs to of SOMChlF with the DPA approach applied on an in-situ global HPLC based phytoplankton group measurements (DPA approach). dataset.*

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

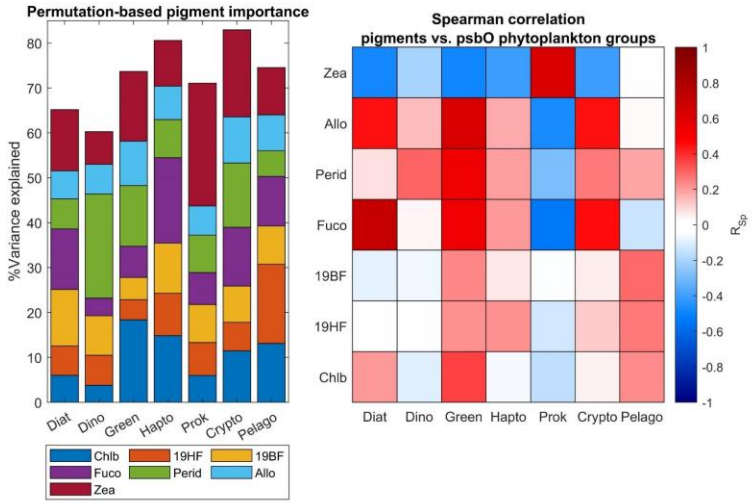


Figure 10. *Evaluation of secondary pigment weighting for the estimation of different phytoplankton groups. The left panel represents the percentage of variance of each phytoplankton group explained by a set of frequently used phytoplankton secondary pigments. This analysis has been done using a random forest algorithm applied to the in-situ Tara Oceans psbO and HPLC datasets. A Spearman correlation coefficient has been calculated between each pigment and the phytoplankton groups (right panel).*

Formatted: Strikethrough

~~4.2.~~4.3. Global patterns of satellite-derived phytoplankton groups ~~using SOM-psbO~~

~~Using~~We then applied our method to Glocolour satellite data~~, we generated~~ to generate a daily database ~~of the relative abundance of the seven focus phytoplankton groups~~spanning from 1997 to 2021~~. From~~ , capturing the relative cell abundance and Chla fraction of seven phytoplankton groups of interest. Fig. 1~~0~~1 presents the annual patterns of relative cell abundance and Chla fraction for each phytoplankton group, derived from this satellite~~-derived dataset, we computed the seasonal relative abundance patterns for each phytoplankton group (Fig. 9).~~ dataset.

~~In terms of~~Regarding relative cell abundance, ~~we could distinguish two~~the prokaryotes stand out as a dominant group. This group largely ~~dominant groups with antagonist spatial distributions: haptophytes and prokaryotes. This latter dominates largely in~~ dominated tropical regions ~~all year round, reaching,~~ with a relative abundance of ~~70~~up to 80% in subtropical gyres. ~~Such~~ Haptophytes, green algae, and diatoms exhibited higher abundance in mid and high latitudes as well as the equatorial region, showing a maximum relative abundance of 30%. The remaining three phytoplankton groups displayed relative abundances that barely exceeded 10% of the total phytoplankton community. Pelagophytes and dinoflagellates were primarily observed in mid and subtropical latitudes, while cryptophytes were found in coastal areas and high latitudes.

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Examination of how each phytoplankton group contributed to total Chla revealed that diatoms had a significant contribution at high latitudes and equatorial regions. Prokaryotes, on the other hand, had an overall low to moderate contribution to total Chla.

Qualitatively, the information captured by SOMChlF was clustered into five groups, each characterized by a distinct remote sensing reflectance spectrum that corresponded to the phytoplankton community structure (Fig. 1~~1~~2). To illustrate the link between each group's contribution to total Chla concentration and relative cell abundance, we depicted the latter while evaluating the pixel's assigned relative abundance values for each of the five clusters. This approach revealed that three out of the five clusters are dominated by prokaryotes in terms of cell abundance (C1, C2, and C3). However, based on their relative contribution to Chla, C1 was found to be dominated by prokaryotes and dinoflagellates, C2 exhibited a mixed composition, C3, and C4 represented diatoms and other eukaryotes, whereas C5 was predominantly composed of diatoms. The shift from relative cell abundance to size-integrated relative Chla fraction illustrates how cell size influences Chla contribution and variability.

Each cluster is characterized by a specific optical signature in terms of Rrs spectra. The Rrs values per wavelength were normalized based on their corresponding variance, enabling intercomparison regardless of magnitude. For instance, C1, which exhibits higher reflectance in the blue wavelength, represents clear, oligotrophic waters. In such environments ~~suffer from ultra-oligotrophy. In conditions where~~ with low nutrients and high surface stratification~~ prevail~~, picophytoplankton groups ~~such as~~like cyanobacteria ~~strive with~~thrive due to their high ~~biovolume~~surface-to-size ratio ~~making them adequate to dominate such regions (ref). 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.~~(Chisholm, 1992; Raven, 1998). C2 represents normalized Rrs spectra with insignificant differences between normalized bands, suggesting an average state where the phytoplankton community appears mixed. In C3 and C4, we observed an increase in normalized Rrs values in the green compared to the blue wavebands, indicating higher Chla in these environments. Given that C3 and C4 are located in high-latitude regions with ample nutrient resources and exceptional seasonal variability of light intensity, larger cell-sized phytoplankton groups, including diatoms, are favored, leading to increased biomass and Chla contribution (Brun et al. 2015). C5, with the greatest difference between Rrs in the blue and green, represents eutrophic waters, known for their high productivity and diatom-dominated blooms (Brun et al. 2015).

~~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~~

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~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~~global distributions of these clusters ~~on a global scale, we can define,~~ several biomes~~, citing:~~ can be defined. C1 is centered in subtropical gyres, C2 is found in transitional zones such as mid-latitude regions ~~as well as~~and the equatorial region, ~~C4~~C3 is ~~shown on~~observed in the ~~edge of the subtropical Antarctic front, C5 for Polar and~~ Southern Ocean, C4 corresponds to high ~~latitudes~~-latitude regions, and ~~C6 for~~C5 is prevalent in coastal and eutrophic waters.

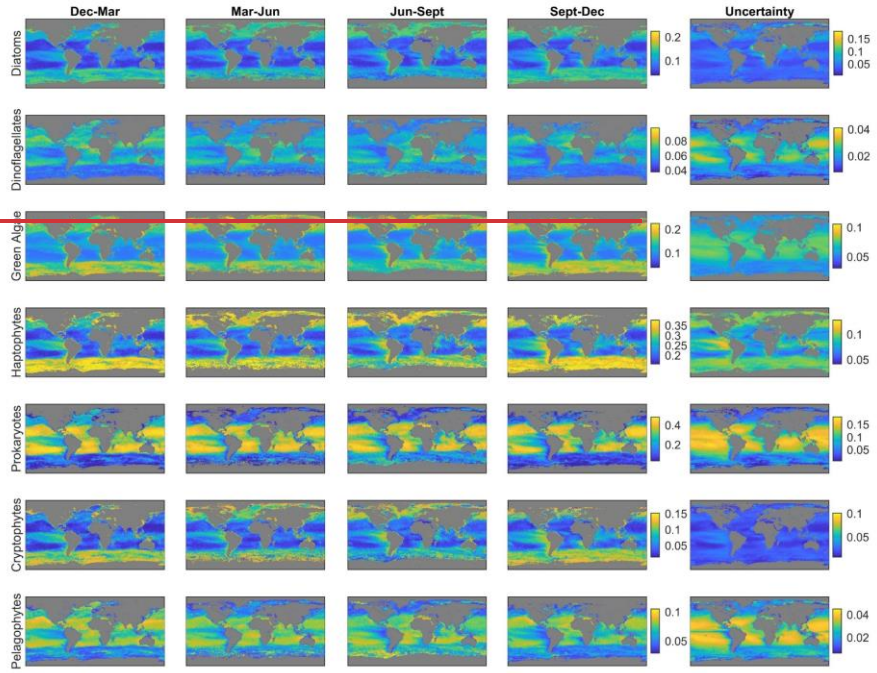
Different temporal variability is evident for each cluster across different latitudinal bands. In northern high latitudes, an increase in C5 indicates maximal productivity occurring in that region around May. At mid-latitudes, the winter maximum is marked by an increase in C5 and C4 clusters. A secondary, less pronounced peak can be observed in autumn, attributed to the break in the thermocline and remineralization processes. During summer, C1 dominates the mid-latitude regions. In tropical regions, C1 is predominant, with a cyclic increase of C2 suggesting coastal influences, likely due to the proximity of C2 to nutrient-rich zones like upwelling systems. In contrast to northern high latitudes, the Southern Ocean exhibits a different temporal variability. The presence of prokaryotes is signified by C1 in this region, whereas C3 dominates during the bloom season in January. This analysis confirms the Antarctic nature of C3 in contrast to C4, highlighting differences in water types between the two regions based on phytoplankton community structure and satellite data.

This ~~structuration~~division into parallel and transitional biomes ~~supports~~underscores the ~~important effect~~significant influence of ~~the~~ latitudinal physical gradients ~~on the structuration of the phytoplankton community (add Ref) such as~~ , including light availability and temperature~~.~~, on the structuring of the phytoplankton community in terms of types and size. These findings align with previous global phytoplankton studies conducted in situ (Ibarbalz et al., 2019; Sommeria-Klein et al., 2021) as well as satellite estimates (Alvain et al., 2006; Hirata et al., 2011; Ben Mustapha et al., 2013; El Hourany et al., 2019a; Xi et al., 2020).

~~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., 2006; Hirata et al., 2011; Ben Mustapha et al., 2013; El Hourany et al., 2019a; Xi et al., 2020).~~

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right



Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

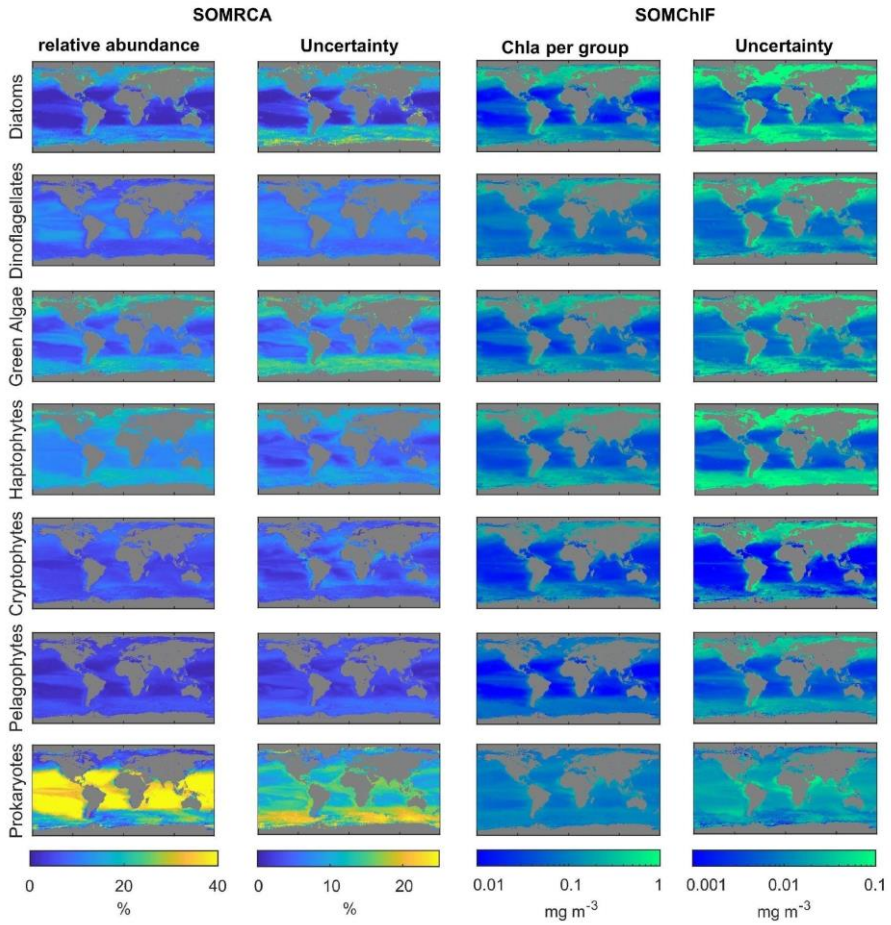


Figure ~~***9. Mean monthly***~~***11.*** *Annual composites of the* ***relative abundances*** *and Chl fractions of the seven* ~~*satellite*~~ *psbO-derived phytoplankton groups based on satellite data (compiled using data from 1997-2021). The* ~~*uncertainty*~~*uncertainties related to each group* ~~*represents the relative standard deviation between all estimated values at different initialization of the SOM and at each "one-leave-out" cross-validation iteration*~~*and each method is because of their different possible combinations through the weighted standard deviations, as described in Section 3.2.3. We not that the scales for uncertainty are smaller than those in the abundance and Chla columns*.

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

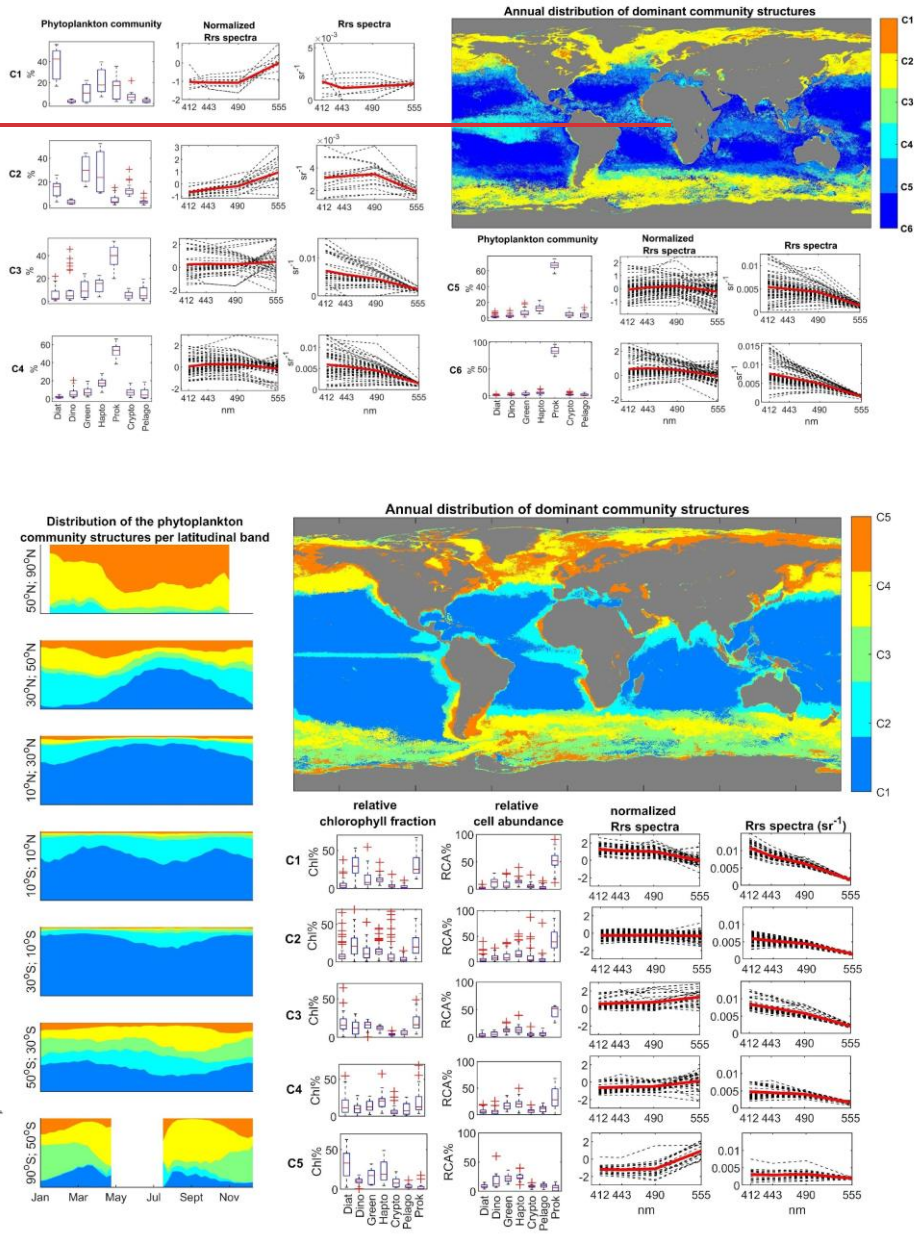


Figure ~~10~~12*: Satellite-derived biomes of* ~~phytoplanktonic~~*phytoplankton communities, obtained by unsupervised clustering (Hierarchical clustering)* ~~*on the SOM's referent vectors. The normalized*~~ *of SOMChlF neurons. Normalized relative cell abundances and original Rrs* ~~*spectrum was*~~*spectra were also derived* ~~*to characterize*~~

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

~~*each cluster's optical signature*~~. *The global map shows the most frequent community structure recorded during the 1997-2021 period.* *A spatio-temporal analysis was conducted to highlight latitudinal patterns.*

~~4.3.~~4.4. Intercomparison of satellite-derived phytoplankton ~~groups~~group products

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

~~Some~~13). However, some differences between the estimated quantities of Chla phytoplankton groups ~~fraction are~~can be noted~~:~~. For diatoms, the outputs based on El Hourany et al., 2019 and SOMChlF exhibit higher Chla ~~Diat~~ values, ~~and the ones~~while those based on Xi et al., 2020 show low values 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, the three products show matching latitudinal variability, with only minor discrepancies in values at high and subtropical latitudes. For prokaryotes, the outputs of Xi et al., ~~SOM-psbO records~~2020 show higher ~~concentrations than the other two products which are relatively matching~~estimates, particularly~~. For prokaryotes, the outputs of Xi et al., 2020 show higher estimation, accentuated~~ near the ~~arctic~~Arctic and ~~the~~ equatorial ~~region. And last, as~~regions. Lastly, for ~~the~~ dinoflagellates, the SOM-Pigments method ~~shows~~yielded lower Chla values ~~of Chla Dino~~, especially in subtropical gyres, whereas SOMChlF showed the highest Chla estimates for this taxonomic group.

Addressing the differences between the outputs referring to the same phytoplankton group is not a straightforward task. Two methods are based on the DPA approach, ~~the latter is not trivial due to~~which displays uncertainties related to the choice of pigments to delimit certain groups. ~~Indeed~~For instance, several studies showed that the DPA approach tends to overestimate ~~some groups such as~~ diatoms ~~(Brewin et al., 2014; Chase et al., 2020).~~(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~~, which allows the development of robust algorithms. On the other hand, the method described in this paper and the generated outputs are based for the first time on a complete and harmonized database ~~on the~~of phytoplankton taxonomic community structure on a global scale; an approach that provides an unbiased picture of phytoplankton cell abundances. ~~But~~ However,the major limitation of this approach at this time is the low number of observations from which ~~makes~~ the ~~global generalization of such a method a major challenge~~metric has been derived.

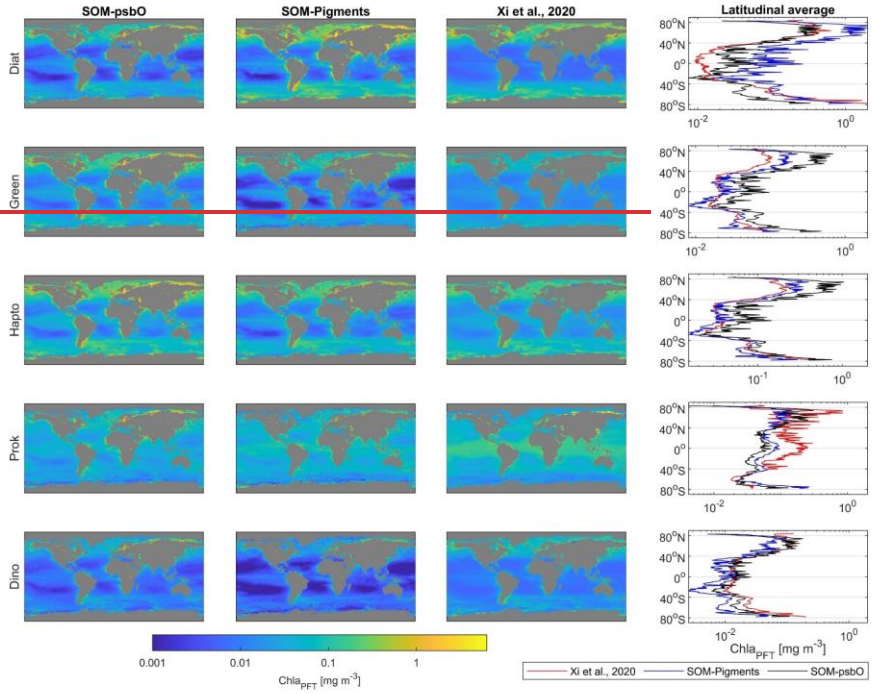
~~The random forest analysis, performed using the in-situ *Tara* Oceans' *psbO* and HPLC measurements, highlights the need for a multivariate approach to predict the phytoplankton community structure from pigments (Fig. 12). Therefore, the variability of each group is best explained not only by one diagnostic 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.~~

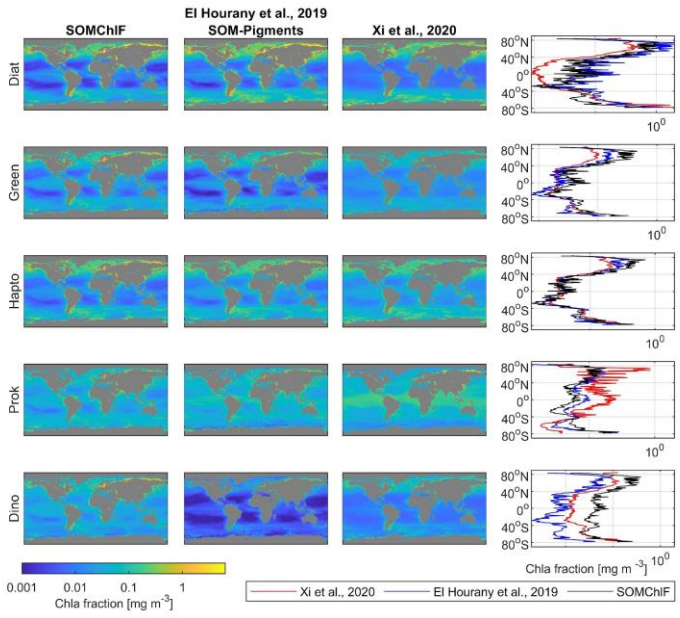
Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right





Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

***Figure 1113**: intercomparison of five satellite-derived phytoplankton group Chla fractions based on ~~SOM-psbO~~SOMChlF, SOM-Pigments (El hourany et al., 2019), and Xi et al., 2020 algorithms. The average per latitude of each Chla fraction is calculated to reveal latitudinal patterns~~. (right panels).*

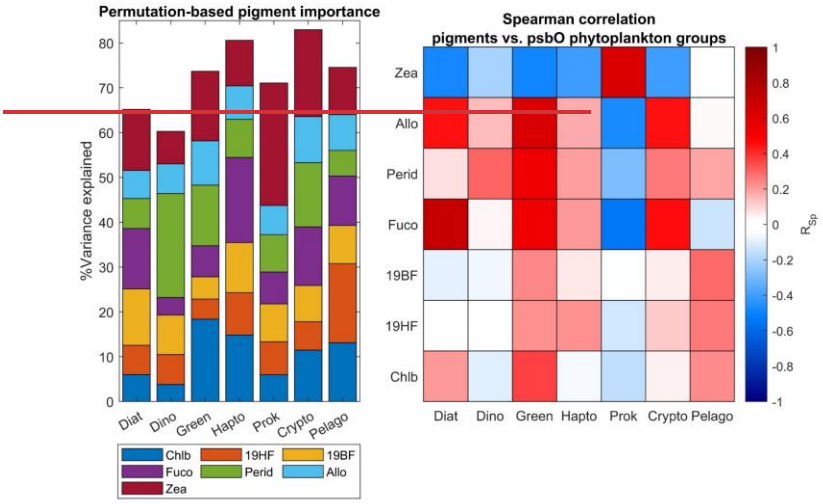


***Figure 12.** Evaluation of secondary pigments' importance regarding the estimation of phytoplankton groups. The left panel represents the percentage of the variance of each phytoplankton group explained by a set of frequently used phytoplankton secondary pigments. This analysis has been done using 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).*

5. Conclusion:

~~We found~~By employing an alternative approach utilizing in-situ metagenomic observations, a ~~remarkable congruence between~~reliable ocean color algorithm for detecting phytoplankton groups was developed in this work. This achievement is noteworthy considering the ~~data derived from satellites and omics, despite the relatively small number~~limited availability of omics data~~.~~ used in our analysis. The ~~link has been~~successful implementation was made possible ~~due to~~ by leveraging machine learning techniques and ~~the preservation of~~preserving the data structure using Self-Organizing ~~maps~~Maps. The methodology ~~showed satisfying performances to provide~~ demonstrated satisfactory performance in producing robust estimates ~~of~~for the seven major phytoplankton groups, albeit with ~~few~~some limitations ~~regarding the~~ in terms of global generalization ~~of the method. The size of the training database is essential to provide a straightforward easy generalizable method. However, in this case,~~due to the limited availability of data. For instance, it is important to exercise caution when interpreting estimates ~~in~~for regions ~~like~~such as the subtropical gyres ~~should be interpreted with caution.~~ As DNA sequencing costs continue

Formatted: Font: Not Bold, Italic

Formatted: Strikethrough

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

to decrease and new expeditions generate molecular data from undersampled ocean regions, we expect the training datasets to increase rapidly in future years, ~~and thus~~which should further increase the accuracy of our method. Furthermore, this study ~~presented~~presents a ~~valuable~~new global dataset of the relative ~~quantities of phytoplankton groups based on a new molecular method, leading to unique~~cell abundances of the seven phytoplankton groups and their contributions to total Chla. These two types of information carry different implications. Chla serves as a biomass proxy, which is crucial for energy and matter fluxes in various ecological and biogeochemical processes. On the other hand, cell abundance represents species abundance for unicellular organisms, providing insights into community assembly processes.

This dataset opens up possibilities for inter-~~comparison~~comparisons with ~~the~~existing approaches, such as DPA-based ~~approach used on~~methods using in-situ and satellite data ~~to identify phytoplankton groups. Nevertheless, these datasets show different yet~~ . The results provide coherent yet distinct information ~~on the~~about phytoplankton~~, valuable for the~~ communities, contributing to a better understanding of ~~the community structure~~their composition. While our focus was on seven broad phytoplankton groups, it is worth mentioning that the deep taxonomic resolution achievable through molecular methods allows for species-level monitoring, which can be an interesting avenue for future implementation.

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 needs to be validated with in-situ observations as well. ~~Few steps away from~~Of further interest is the impending 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.

Data and Code availability. ~~psbO dataset: https://www.ebi.ac.uk/biostudies/studies/S-BSST761~~*psbO* dataset: https://www.ebi.ac.uk/biostudies/studies/S-BSST761; Globcolour dataset: ~~https://www.globcolour.info/, https://hermes.acri.fr/~~https://www.globcolour.info/, https://hermes.acri.fr/. SST CCI dataset: ~~https://data.marine.copernicus.eu/product/SST_GLO_SST_L4_REP_OBSERVATIONS_010_024/description.~~https://data.marine.copernicus.eu/product/SST_GLO_SST_L4_REP_OBSERVATIONS_010_024/description. Global HPLC pigment dataset: MAREDAT, POLERSTERN data, Labrador Sea expeditions data, and *Tara* Oceans Expedition data, all available on ~~https://pangaea.de/~~https://pangaea.de/, GeP&Co database (accessed at ~~http://www.obs-vlfr.fr/proof/php/x_datalist.php?xxop=gepco&xxcamp=gepco~~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/)~~.https://seabass.gsfc.nasa.gov/). Following best practices, the ~~SOM-psbO~~SOMChlF and SOMRCA will be deposited into a public domain repository accessible upon publication alongside with the datasets generated in this study. Prerequisite software library SOM Toolbox 2.0 for Matlab is

Formatted: Font: Italic

Formatted: Font color: Black

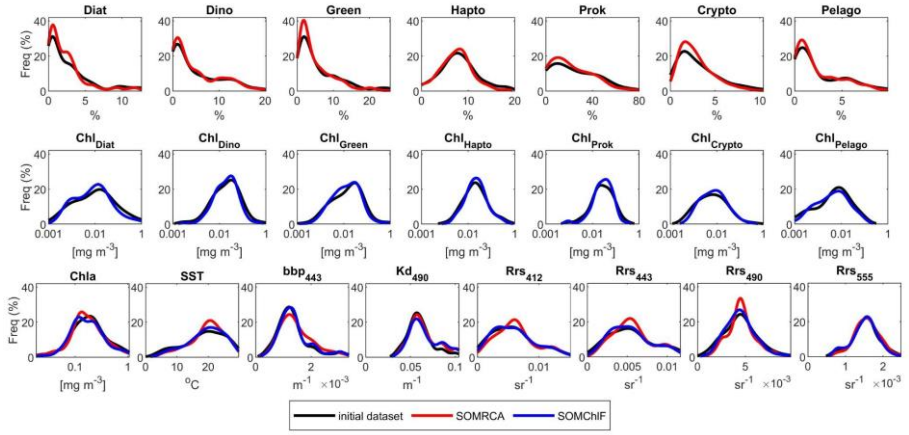
Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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.~~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~~JJPK. Formal analysis, RE, ~~JPK~~JJPK, ML, CB. Investigation, RE, ~~JPK~~JJPK, LZ, HL, ML, CB. Resources, ML, CB. Data curation, ~~JPK~~JJPK, RE. Writing-original draft preparation, RE, Writing-review and editing, RE, ~~JPK~~JJPK, 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.

Acknowledgments. 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~~Grant agreement No. 835067), the Horizon Europe projects 'Marco-Bolo' (Grant Agreement No. 101082021) and 'BlueRemediomics' (Grant Agreement No. 101082304), 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.



Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Figure A1. *Evaluation of the preservation of the initial dataset's characteristics; distribution of the values within the initial dataset (DRCA and DChlF) and the SOMRCA and SOMChlF neurons.*

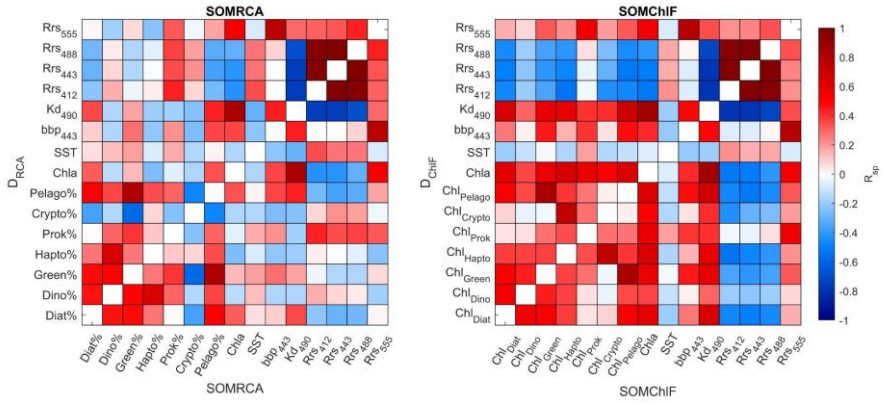


Figure A2. *Evaluation of the preservation of the initial dataset's characteristics; Spearman correlation coefficient matrix comparing intra-correlations in the initial datasets DRCA and DChlF and the SOMRCA and SOMChlF neurons respectively.*

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. Res. Part II Top. Stud. Oceanogr., 56, 899–917, https://doi.org/10.1016/j.dsr2.2008.09.017, 2009.

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

Alvain, S., Moulin, C., Dandonneau, Y., Loisel, H., and Bréon, F. M.: A species-dependent bio-optical model of case I waters for global ocean color processing, Deep. Res. Part I Oceanogr. Res. Pap., 53, 917–925, https://doi.org/10.1016/j.dsr.2006.01.011, 2006.

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 Biogeochem. Cycles, 22, 1–15, https://doi.org/10.1029/2007GB003154, 2008.

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, Biogeosciences, 751–764 pp., 2009.

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

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

Brewin, R. J. W., 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

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Sens. Environ., 168, 437–450, https://doi.org/10.1016/J.RSE.2015.07.004, 2015.

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, Limnol. Oceanogr. Methods, 18, 570–584, https://doi.org/10.1002/LOM3.10385, 2020.

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, Front. Mar. Sci., 4, 126, https://doi.org/10.3389/fmars.2017.00126, 2017.

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 Res. Part II Top. Stud. Oceanogr., 51, 303–318, https://doi.org/10.1016/j.dsr2.2003.07.018, 2004.

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

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 Discuss., 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, 13 May 2009.

Gieskes, W. W. C. and Kraay, G. W.: Dominance of Cryptophyceae during the phytoplankton spring bloom in the central North Sea detected by HPLC analysis of pigments, Mar. Biol., 75, 179–185, https://doi.org/10.1007/BF00406000, 1983.

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, Limnol. Oceanogr., 54, 1951–1963, https://doi.org/10.4319/LO.2009.54.6.1951, 2009.

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

Henson, S. A., Cael, B. B., Allen, S. R., and Dutkiewicz, S.: Future phytoplankton diversity in a changing climate, Nat. Commun. 2021 121, 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 (Cop.)., 24, 251–256, https://doi.org/10.1034/J.1600-0587.2001.240302.X, 2001.

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

Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W. 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. M., 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 Res. Part II Top. Stud. Oceanogr., 53, 459–512,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

https://doi.org/10.1016/J.DSR2.2006.01.025, 2006.

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), J. Geophys. Res. Ocean., https://doi.org/10.1029/2018JC014450, 2019a.

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, J. Geophys. Res. Ocean., 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, J. Geophys. Res. Ocean., 126, e2020JC016808, https://doi.org/10.1029/2020JC016808, 2021.

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 Biogeochem. 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, Nat. 2004 4296994, 429, 863–867, https://doi.org/10.1038/nature02593, 2004.

Jeffrey, S. W.: Algal Pigment Systems, in: Primary Productivity in the Sea, Springer US, Boston, MA, 33–58, https://doi.org/10.1007/978-1-4684-3890-1_3, 1980.

Jeffrey, S. W. and Hallegraeff, G. M.: Chlorophyllase distribution in ten classes of phytoplankton: a problem for chlorophyll analysis, https://doi.org/10.2307/24825001, 1987.

Jouini, M., Lévy, M., Crépon, M., and Thiria, S.: Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method, Remote Sens. Environ., 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.

Loreau, M.: Biodiversity and ecosystem functioning: A mechanistic model, Proc. Natl. Acad. Sci., 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, Nat. 2001 4126842, 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 Syst. Sci. Data, 4, 47–73, https://doi.org/10.5194/essd-4-47-2012, 2012.

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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

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 Sens. Environ., https://doi.org/10.1016/j.rse.2013.08.046, 2013.

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, J. Geophys. Res., 103, 24937, 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), Appl. Opt., 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., 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. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., 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. W., Stuart, S. N., Turak, E., Walpole, M., and Wegmann, M.: Essential biodiversity variables, Science (80-.)., 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, Sci. Data, 2, https://doi.org/10.1038/sdata.2015.23, 2015.

Pierella Karlusich, J. J., Ibarbalz, F. M., and Bowler, C.: Phytoplankton in the Tara Ocean, https://doi.org/10.1146/annurev-marine-010419-010706, 3 January 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, Mol. Ecol. Resour., 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.

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, Glob. Chang. Biol., 0, 051013014052005-???, https://doi.org/10.1111/j.1365-2486.2005.1004.x, 2005.

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, 45 pp., 2014.

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

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

Righetti, D., Vogt, M., Gruber, N., Psomas, A., and Zimmermann, N. E.: Global pattern of phytoplankton diversity driven by temperature and environmental variability, Sci. Adv., 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, Glob. Ecol. Biogeogr., 24, 527–538, https://doi.org/10.1111/GEB.12274, 2015.

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

Sarzeaud, O. and Stephan, Y.: Data interpolation using Kohonen networks, Proc. Int. Jt. Conf. 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, IOCCG; 15., edited by: S. Sathyendranath and V. Stuart, International Ocean-Colour Coordinating Group, Dartmouth, Nova Scotia, B2Y 4A2, Canada., 156 pp., 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, 2826–2832, https://doi.org/10.1109/IJCNN.2009.5178623, 2009.

Smith, V. H.: Microbial diversity–productivity relationships in aquatic ecosystems, FEMS Microbiol. Ecol., 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 (80-.)., 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 Sens., 6, 10089–10106, https://doi.org/10.3390/rs61010089, 2014.

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, C08005, 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, J. Geophys. Res. Ocean., 106, 19939–19956, 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 Sens. Environ., 98, 122–140, https://doi.org/10.1016/j.rse.2005.07.001, 2005.

Wright, S. W. and Jeffrey, S. W.: Fucoxanthin pigment markers of marine phytoplankton analysed by HPLC and HPTLC, https://doi.org/10.2307/24825629, 1987.

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 Sens. Environ., 240,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

111704, https://doi.org/10.1016/J.RSE.2020.111704, 2020.

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. Res. Part II Top. Stud. Oceanogr., 56, 899–917, https://doi.org/10.1016/j.dsr2.2008.09.017, 2009.

Alvain, S., Moulin, C., Dandonneau, Y., and Bréon, F. M.: Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery, Deep Sea Res. Part I Oceanogr. Res. Pap., 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 Biogeochem. Cycles, 22, 1–15, https://doi.org/10.1029/2007GB003154, 2008.

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, Biogeosciences, 751–764 pp., 2009.

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

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

Brewin, R. J. W., 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 Sens. Environ., 168, 437–450, https://doi.org/10.1016/J.RSE.2015.07.004, 2015.

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

Brun, P., Vogt, M., Payne, M.R., Gruber, N., O'brien, C.J., Buitenhuis, E.T., Le Quéré, C., Leblanc, K. and Luo, Y.W.,: Ecological niches of open ocean phytoplankton taxa. Limnol. Oceanogr. 60 (3): 1020–38, https://doi.org/10.1002/lno.10074, 2015

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, Limnol. Oceanogr. 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.

da Silva, L.E.B., Costa, J.A.F. (2013). Clustering, Noise Reduction and Visualization Using Features Extracted from the Self-Organizing Map. In: , et al. Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41278-3_30

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, Front. Mar. Sci., 4, 126, https://doi.org/10.3389/fmars.2017.00126, 2017.

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 Res. Part II Top. Stud. Oceanogr., 51, 303–318, https://doi.org/10.1016/j.dsr2.2003.07.018, 2004.

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

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

609-2020, 2020.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., ... & Martiny, A. C. (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.

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 Discuss., 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, 13 May 2009.

Gieskes, W. W. C. and Kraay, G. W.: Dominance of Cryptophyceae during the phytoplankton spring bloom in the central North Sea detected by HPLC analysis of pigments, Mar. Biol., 75, 179–185, https://doi.org/10.1007/BF00406000, 1983.

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, Limnol. Oceanogr., 54, 1951–1963, https://doi.org/10.4319/LO.2009.54.6.1951, 2009.

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

Henson, S. A., Cael, B. B., Allen, S. R., and Dutkiewicz, S.: Future phytoplankton diversity in a changing climate, Nat. Commun. 2021 121, 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 (Cop.)., 24, 251–256, https://doi.org/10.1034/J.1600-0587.2001.240302.X, 2001.

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

Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W. 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. M., 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 Res. Part II Top. Stud. Oceanogr., 53, 459–512, https://doi.org/10.1016/J.DSR2.2006.01.025, 2006.

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), J. Geophys. Res. Ocean., https://doi.org/10.1029/2018JC014450, 2019a.

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, J. Geophys. Res. Ocean., 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, J. Geophys. Res. Ocean., 126, e2020JC016808, https://doi.org/10.1029/2020JC016808, 2021.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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 Biogeochem. 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, Nat. 2004 4296994, 429, 863–867, https://doi.org/10.1038/nature02593, 2004.

Jeffrey, S. W.: Algal Pigment Systems, in: Primary Productivity in the Sea, Springer US, Boston, MA, 33–58, https://doi.org/10.1007/978-1-4684-3890-1_3, 1980.

Jeffrey, S. W. and Hallegraeff, G. M.: Chlorophyllase distribution in ten classes of phytoplankton: a problem for chlorophyll analysis, https://doi.org/10.2307/24825001, 1987.

Jouini, M., Lévy, M., Crépon, M., and Thiria, S.: Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method, Remote Sens. Environ., 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.

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 Syst. Sci. Data, 4, 47–73, https://doi.org/10.5194/essd-4-47-2012, 2012.

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

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 Sens. Environ., https://doi.org/10.1016/j.rse.2013.08.046, 2013.

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, J. Geophys. Res., 103, 24937, 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), Appl. Opt., 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., 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

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., 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. W., Stuart, S. N., Turak, E., Walpole, M., and Wegmann, M.: Essential biodiversity variables, Science (80-.)., 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, Sci. Data, 2, https://doi.org/10.1038/sdata.2015.23, 2015.

Pierella Karlusich, J. J., Ibarbalz, F. M., and Bowler, C.: Phytoplankton in the *Tara* Ocean, Ann Rev Mar Sci, 12, 233-265.https://doi.org/10.1146/annurev-marine-010419-010706, 3 January 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, Mol. Ecol. Resour., 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.

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, Glob. Chang. Biol., 0, 051013014052005-???, https://doi.org/10.1111/j.1365-2486.2005.1004.x, 2005.

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, 45 pp., 2014.

Richardson, A. J., Risien, C., and Shillington, F. A.: Using self-organizing maps to identify patterns in satellite imagery, Prog. Oceanogr., 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, Sci. Adv., 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, Glob. Ecol. Biogeogr., 24, 527–538, https://doi.org/10.1111/GEB.12274, 2015.

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

de Salas, M. F., Eriksen, R., Davidson, A. T., and Wright, S. W.: Protistan communities in the Australian sector of the Sub-Antarctic Zone during SAZ-Sense, Deep. Res. Part II Top. Stud. Oceanogr., 58, 2135–2149, https://doi.org/10.1016/j.dsr2.2011.05.032, 2011.

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.,

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right

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, IOCCG; 15., edited by: S. Sathyendranath and V. Stuart, International Ocean-Colour Coordinating Group, Dartmouth, Nova Scotia, B2Y 4A2, Canada., 156 pp., 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, 2826–2832, https://doi.org/10.1109/IJCNN.2009.5178623, 2009.

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

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 Sens., 6, 10089–10106, https://doi.org/10.3390/rs61010089, 2014.

Tara Ocean Foundation; Tara Oceans; European Molecular Biology Laboratory (EMBL); European Marine Biological Resource Centre - European Research Infrastructure Consortium (EMBRC-ERIC). Priorities for ocean microbiome research. Nat Microbiol. 2022 Jul;7(7):937-947. doi: 10.1038/s41564-022-01145-5. Epub 2022 Jun 30. PMID: 35773399.

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, C08005, 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, J. Geophys. Res. Ocean., 106, 19939–19956, 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 Sens. Environ., 98, 122–140, https://doi.org/10.1016/j.rse.2005.07.001, 2005.

Wright, S. W. and Jeffrey, S. W.: Fucoxanthin pigment markers of marine phytoplankton analysed by HPLC and HPTLC, https://doi.org/10.2307/24825629, 1987.

Wright, S. W., van den Enden, R. L., Pearce, I., Davidson, A. T., Scott, F. J., and Westwood, K. J.: Phytoplankton community structure and stocks in the Southern Ocean (30-80°E) determined by CHEMTAX analysis of HPLC pigment signatures, Deep. Res. Part II Top. Stud. Oceanogr., 57, 758–778, https://doi.org/10.1016/j.dsr2.2009.06.015, 2010.

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 Sens. Environ., 240, 111704, https://doi.org/10.1016/J.RSE.2020.111704, 2020.

Formatted: Font color: Black

Formatted: Normal, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border), Tab stops: 3.25", Centered + 6.5", Right