Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data

Hongyan Xi[a,*], Svetlana N. Losa[a,b], Antoine Mangin[c], Mariana A. Soppa[a], Philippe Garnesson[c], Julien Demaria[c], Yangyang Liu[a,d], Odile Hembise Fanton d'Andon[c], Astrid Bracher[a,e]

[a] *Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*
[b] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia*
[c] *ACRI-ST, 06904 Sophia Antipolis Cedex, France*
[d] *Faculty of Biology and Chemistry, University of Bremen, Bremen, Germany*
[e] *Institute of Environmental Physics, University of Bremen, Bremen, Germany*

ARTICLE INFO

Edited by Menghua Wang

Keywords:
Retrieval algorithm
Chlorophyll *a*
Phytoplankton functional types
Empirical orthogonal functions
Remote sensing reflectance
HPLC pigments
CMEMS GlobColour
OLCI

ABSTRACT

This study presents an algorithm for globally retrieving chlorophyll *a* (Chl-*a*) concentrations of phytoplankton functional types (PFTs) from multi-sensor merged ocean color (OC) products or Sentinel-3A (S3) Ocean and Land Color Instrument (OLCI) data from the GlobColour archive in the frame of the Copernicus Marine Environmental Monitoring Service (CMEMS). The retrieved PFTs include diatoms, haptophytes, dinoflagellates, green algae and prokaryotic phytoplankton. A previously proposed method to retrieve various phytoplankton pigments, based on empirical orthogonal functions (EOF), is investigated and adapted to retrieve Chl-a concentrations of multiple PFTs using extensive global data sets of in situ pigment measurements and matchups with satellite OC products. The performance of the EOF-based approach is assessed and cross-validated statistically. The retrieved PFTs are compared with those derived from diagnostic pigment analysis (DPA) based on in situ pigment measurements. Results show that the approach predicts well Chl-*a* concentrations of most of the mentioned PFTs. The performance of the approach is, however, less accurate for prokaryotes, possibly due to their general low variability and small concentration range resulting in a weak signal which is extracted from the reflectance data and corresponding EOF modes. As a demonstration of the approach utilization, the EOF-based fitted models based on satellite reflectance products at nine bands are applied to the monthly GlobColour merged products. Climatological characteristics of the PFTs are also evaluated based on ten years of merged products (2002 – 2012) through inter-comparisons with other existing satellite derived products on phytoplankton composition including phytoplankton size class (PSC), SynSenPFT, OC-PFT and PHYSAT. Inter-comparisons indicate that most PFTs retrieved by our study agree well with previous corresponding PFT/PSC products, except that prokaryotes show higher Chl-*a* concentration in low latitudes. PFT dominance derived from our products is in general well consistent with the PHYSAT product. A preliminary experiment of the retrieval algorithm using eleven OLCI bands is applied to monthly OLCI products, showing comparable PFT distributions with those from the merged products, though the matchup data for OLCI are limited both in number and coverage. This study is to ultimately deliver satellite global PFT products for long-term continuous observation, which will be updated timely with upcoming OC data, for a comprehensive understanding of the variability of phytoplankton composition structure at a global or regional scale.

1. Introduction

Over the past decades, satellite ocean color (OC) remote sensing has been widely used for estimating chlorophyll *a* (Chl-*a*) concentration, which is often used as an indicator of phytoplankton biomass. Beyond that, extracting information on phytoplankton community structure, e.g., phytoplankton functional types (PFTs), size classes (PSCs) and taxonomic composition, has become a research topic of priority, as it plays an important role in understanding the marine food web and aids the modelling associated with climate change impacts on

[*] Corresponding author.
E-mail address: Hongyan.Xi@awi.de (H. Xi).

https://doi.org/10.1016/j.rse.2020.111704
Received 4 July 2019; Received in revised form 7 January 2020; Accepted 3 February 2020

biogeochemical and ecological cycling of oceans (e.g., Falkowski et al., 1998; Le Quéré et al., 2005; IPCC, 2013; Bracher et al., 2017). In addition, accurate estimation on phytoplankton diversity and group distribution provides valuable information on identifying blooms caused by specific toxic algae, i.e., harmful algal blooms such as cyanobacterial blooms and red tides (e.g., Craig et al., 2006; Hu et al., 2010; Wang et al., 2017). A PFT is usually defined as a homologous set of "organisms related through common biogeochemical processes" such as silicification, calcification, nitrogen fixation, or dimethyl sulfide production, but are not necessarily phylogenetically affiliated (Falkowski et al., 2003; Litchman et al., 2006; IOCCG, 2014). However, as many phytoplankton groups which can be detected by remote sensing are also functional types, (e.g., diatoms are silicifiers, some cyanobacteria are nitrogen fixers, and coccolithophorids are calcifiers) (Bracher et al., 2017), these satellite proxies have been named PFTs for brevity (e.g., Losa et al., 2017).

Satellite OC remote sensing enables observation of phytoplankton over large areas or even at global scale. With previous (e.g., Sea-Viewing Wide Field-of-View Sensor – SeaWiFS and MEdium Resolution Imaging Spectrometer – MERIS) and current available OC satellites Moderate Resolution Imaging Spectroradiometer (MODIS), Visible Infrared Imaging Radiometer Suite (VIIRS), and especially the newly launched OLCI onboard Sentinel-3A (in February 2016) and 3B (in April 2018), a vast amount of quality controlled OC data are collected, allowing us to contribute to developing and/or improving methods and the corresponding applications to satellite data for estimating biogeochemical parameters in terms of global observation. There is a clear need to implement a sound PFT retrieval algorithm to the recent OLCI data, as well as to previous and current satellite OC time series data such as CMEMS GlobColour merged products (ACRI-ST GlobColour Team et al., 2017).

Different bio-optical and ecological algorithms have been developed to identify PFTs and phytoplankton taxonomic composition at the ocean surface, mainly based on phytoplankton abundance and inherent/apparent optical properties. Abundance-based approaches seek to establish empirical relationships between the PFTs and phytoplankton abundance or biomass, such as Chl-*a* concentration that can be retrieved from satellites (e.g., Uitz et al., 2006; Brewin et al., 2010, 2015; Hirata et al., 2011). Ecological-based approaches incorporate additional environmental parameters to identify ecological niches where particular phytoplankton communities may be found (Raitsos et al., 2008; Palacz et al., 2013). Efforts have also been made to combine abundance and ecological-based approaches (e.g. Brewin et al., 2015; Ward, 2015). Spectral-based approaches are more direct as they target known optical signatures and use satellite observed spectra to extract the signatures of specific PFT (e.g., Ciotti and Bricaud, 2006; Devred et al., 2006; Alvain et al., 2005, 2008; Hirata et al., 2008; Bracher et al., 2009; Kostadinov et al., 2009; Werdell et al., 2014; Brewin et al., 2015; Correa-Ramirez et al., 2018). These methods are mainly based on radiative transfer or bio-optical models and generally require high computation performance and adaptations for specific sensors. More complete reviews of these approaches are well detailed by the works of the IOCCG (2014), Bracher et al. (2017), and Mouw et al. (2017).

In this study, we seek to establish an approach that uses satellite reflectance data which inherit the information of various phytoplankton pigments and, therefore, allows retrieving the Chl-*a* concentrations of multiple PFTs. We choose the empirical orthogonal function (EOF) analysis, also known as principal component analysis, as it has been previously used for predicting ocean color metrics and various phytoplankton pigment concentrations by assessing variance of structures in spectral remote sensing reflectance (R_{rs}) or water leaving radiance (e.g., Lubac and Loisel, 2007; Craig et al., 2012; Taylor et al., 2013; Bracher et al., 2015; Soja-Woźniak et al., 2017). The spectral data are subject to EOF analysis to reduce the high dimensionality of the data and derive the dominant signals (EOF modes) that best describe the variance within the data set. Studies also proved that the EOF analysis could provide reliable retrievals even with limited number of data points (Craig et al., 2012; Bracher et al., 2015). Another advantage is that the models exhibited negligible loss of skill when applied to data sets with a reduced spectral resolution, which enables the applicability to the previous or currently existing multispectral OC sensors and future hyperspectral satellite missions such as PACE (Gregg and Rousseaux, 2017), HyspIRI (Lee et al., 2015) and EnMAP (Guanter et al., 2015).

Given that the EOFs derived from in situ or satellite hyper-/multispectral R_{rs} data have provided reliable retrievals of the concentrations of Chl-*a* and different pigments/pigment groups (Taylor et al., 2013; Bracher et al., 2015), we intend to present an implementation of the method proposed in Bracher et al. (2015) to retrieve PFTs instead of pigments, and to up-scale the application from regional to global scale by constructing large in situ data sets and multi-sensor OC products. Therefore, with the use of extensive in situ phytoplankton pigment data sets, satellite OC products, and matchups between in situ and satellite data, we propose an EOF-based global PFT retrieval approach by linking the variances in R_{rs} spectral structures to different PFTs. In the present study, we aim firstly to establish the EOF fitted model based on the nearly globally covered matchups between the satellite R_{rs} and the PFT Chl-*a* concentrations derived from diagnostic pigment analysis (DPA) of in situ HPLC pigment data, and cross-validate the performance of the EOF-based algorithm statistically; secondly, to set up the PFT retrieval scheme based on the EOF modes obtained from the matchups for the implementation to satellite OC products; thirdly, to investigate and evaluate the climatological characteristics of the PFTs retrieved from merged OC products (2002–2012) through inter-comparisons with other existing PFT/PSC products at the same period, and finally, to explore the potential of applying the approach to OLCI products based on a prediction scheme using a much more limited number of matchups.

2. Data and methods

2.1. Data sets

2.1.1. In situ databases of phytoplankton pigments

2.1.1.1. Pigment Database I (1997–2012). A large data set of the quality controlled near surface (first 12 m) HPLC phytoplankton pigments built for the ESA SynSenPFT Project (Bracher et al., 2016) was used for the extraction of the collocated R_{rs} spectra from satellite data. This HPLC pigment data set includes >15,000 sets of phytoplankton pigment data spanning 25 years from 1988 to 2012 covering the global ocean, collected from SEABASS, MAREDAT, LTER, BATS, AESOP-CSIRO, LOV and also from our own data published at PANGAEA (see Table 1 in Losa et al., 2017). Since SeaWiFS as an earlier OC sensor was launched in 1997, a subset for the period of 1997–2012 including 11,977 sets of pigment data was taken as Pigment Database I and used for the extraction of the R_{rs} matchups from GlobColour merged products. Yearly coverage of this matchup database spans from 3.2% (the least data points for 2012) to 9.3% (the most for 2004). 24.1%, 17.4%, 21.1%, and 37.4% of the data were collected during March–May, June–August, September–November, and December–February, respectively. Fig. 1(A) shows the spatial distribution of all the data points in this database in which all pigments are included, but only total chlorophyll *a* concentration (TChl-*a*, sum of monovinyl chlorophyll *a*, divinyl chlorophyll *a*, chlorophyll *a* allomers, chlorophyll *a* epimers, and chlorophyllide *a*) is present in the figure.

2.1.1.2. Pigment Database II (2016–2018). A relatively smaller (n = 992) phytoplankton pigment database of quality controlled near surface HPLC pigments was also built for the OLCI matchups from 2016 to 2018, involving our recently published data sets of HPLC based phytoplankton pigment concentrations collected mainly in late spring and summer from five cruises – Heincke462 in the North Sea

Table 1
Numbers of available R_{rs} matchups (1 × 1 pixel and averaged by 3 × 3 pixels) with different band combinations from the CMEMS GlobColour merged products. Bold highlights the matchups used in the EOF based algorithm. SeaW = SeaWiFS (1997–2010), MO = MODIS (2002–present), ME = MERIS (2002–2012), V = VIIRS/Suomi-NPP (2012–present). Note that with SeaWiFS included merged products, the bands from SeaWiFS only contributed until December 2010. Waveband centers for the four sensors were listed in Table S1 in the supplementary document.

Sensors	No. of matchups		Available wavebands (nm)												No. of bands
	1 × 1	3 × 3	412	443	490	510	531	547	551	555	560	620	670	678	
SeaW	1223	609	×	×	×	×				×			×		6
SeaW/ME	381	125	×	×	×	×				×	×	×	×		8
SeaW/MO/ME	766	516	×	×	×		×	×		×			×	×	8
SeaW/MO/ME	394	265	×	×	×	×	×	×		×			×	×	9
MO + V	25	27	×	×	×		×	×	×	×			×	×	9
SeaW/MO/ME	183	63	×	×	×	×	×	×		×	×	×	×	×	11
MO/ME/V	3	2	×	×	×	×	×	×	×	×	×	×	×	×	12

(April–May 2016): https://doi.pangaea.de/10.1594/PANGAEA.899043 (Bracher and Wiegmann, 2019), PS99 in the North Sea and the Fram Strait Arctic (June–July 2016): https://doi.pangaea.de/10.1594/PANGAEA.905502 (PS99.1) and https://doi.pangaea.de/10.1594/PANGAEA.898102 (PS99.2) (Liu et al., 2019a, 2019c), PS103 in the Southern Ocean: https://doi.pangaea.de/10.1594/PANGAEA.898941 (Bracher, 2019) (December 2016–January 2017), PS107 in the Fram Strait Arctic (July–August 2017): https://doi.pangaea.de/10.1594/PANGAEA.898100 (Liu et al., 2019b), and PS113 in the trans-Atlantic Ocean (May–June 2018): https://doi.pangaea.de/10.1594/PANGAEA.911061 (Bracher et al., 2020). Fig. 1(B) shows the locations of the data points from Pigment Database II (including all the pigments but with only TChl-*a* concentration present in the figure), which covers a large range of latitudes but focuses on the Atlantic Ocean only (60°W–20°E).

2.1.2. Satellite ocean color data

Satellite normalized remote sensing reflectance (R_{rs}) Level-3 (L3) products from multiple sensors were obtained from the CMEMS GlobColour data archive (http://www.globcolour.info/). The R_{rs} products used for matchup analysis included daily R_{rs} L3 products with 4-km resolution at the bands from either individual sensors (SeaWiFS, MODIS, MERIS, and VIIRS onboard Suomi-NPP) or the merged products

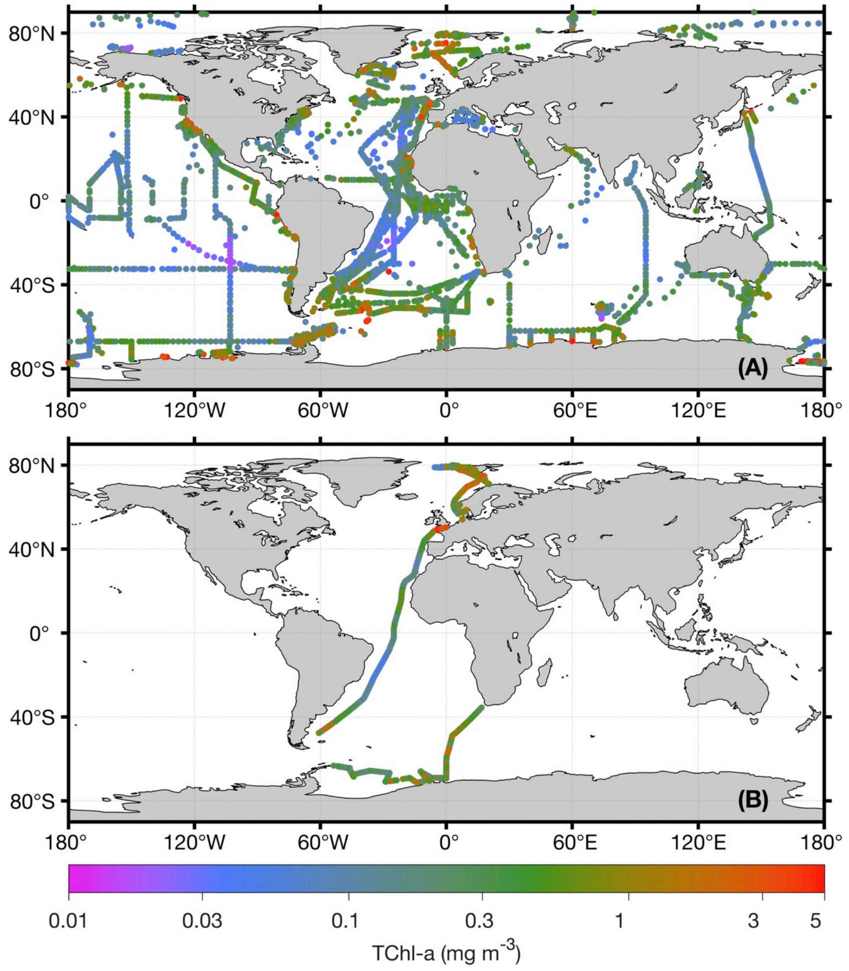

Fig. 1. Spatial distribution of the TChl-*a* concentration from the quality controlled in situ (A) Pigment Database I (1997–2012), and (B) Pigment Database II (2016–2018).

of two or more sensors. More details on the merged products are given in the GlobColour Product User Guide (ACRI-ST GlobColour Team et al., 2017). R_{rs} products from OLCI were not merged with any other sensor products and were therefore used separately for an OLCI only PFT retrieval scheme. Similar to the merged products, daily 4 km R_{rs} L3 products of OLCI were used for matchup extraction. In further application of the proposed approach to derive global long time series PFT products, monthly R_{rs} L3 products with 25 km spatial resolution from both, the merged products and OLCI data, were obtained for July 2002–April 2012 (time when SeaWiFS, MODIS and MERIS were in orbit, although SeaWiFS operation ended in late 2010 and then in late 2011 VIIRS was added), and April 2016–December 2018 (OLCI on Sentinel-3A in operation), respectively. In addition, the GlobColour merged ocean TChl-*a* monthly products with 25 km resolution in July 2002–April 2012 were also obtained for inter-comparison. The merged L3 TChl-*a* products were derived by a weighted average method (AVW) from single-sensor Level 2 chlorophyll products for case 1 waters (ACRI-ST GlobColour Team et al., 2017).

2.1.3. PFT retrieval input data

(A) PFT Chl-*a* concentrations derived from diagnostic pigment analysis (DPA)

Chl-*a* concentrations of PFTs were derived using an updated DPA method (Soppa et al., 2014; Losa et al., 2017). The DPA method was originally developed by Vidussi et al., 2001, adapted in Uitz et al. (2006) and further refined by Hirata et al. (2011) and Brewin et al. (2015). Basically, it relates the weighted sum of seven DPs (representative of individual PFTs) to TChl-*a* concentration, enabling us to determine the fraction of each PFT to the TChl-*a* thus to derive the PFT Chl-*a* concentrations. The partial coefficients of the DPs used in this study were derived from multiple linear regression using the data from the large global pigment data set as detailed in Table S1 of Supplementary Material in Losa et al. (2017) and were in good agreement with previous studies. The pigment concentrations of fucoxanthin, peridinin, 19′hexanoyloxy-fucoxanthin, 19′butanoyloxy-fucoxanthin, alloxanthin, chlorophyll *b*, zeaxanthin and divinyl chlorophyll *a* were used to derive the Chl-*a* concentrations of six PFTs in our study, that are, respectively, diatoms and dinoflagellates which are commonly considered as microphytoplankton, two types of nanophytoplankton – haptophytes and green algae (chlorophytes), and two picophytoplankton – prokaryotes, and *Prochlorococcus* which is a typical species of prokaryotes and commonly found in the subtropical region. PFT Chl-*a* concentrations <0.005 mg m^{-3} were excluded as such low values might contain much uncertainty. The rational for this threshold is that the surface Chl-*a* concentration encountered in the clearest ocean waters (South Pacific Gyre) was found to be in the range 0.01–0.02 mg m^{-3} (Morel et al., 2007). Therefore, values below 0.01 mg m^{-3} may be questionable. The corresponding PFT Chl-*a* concentration can be smaller. Considering the quality control on a large pigment data set as in Aiken et al. (2009), we chose the threshold of 0.005 mg m^{-3} for PFT Chl-*a* to minimize the influence of low accuracy in observations on the retrieval model, as it could bring much higher uncertainty to final prediction. The DPA derived PFT Chl-*a* concentrations for diatoms, haptophytes and prokaryotes from the pigment database I were published already in Losa et al. (2017) and are available from PANGAEA: https://doi.pangaea.de/10.1594/PANGAEA.875879 (Soppa et al., 2017).

(B) Matchups between in situ PFT and satellite R_{rs} data

Matchups to in situ PFT data were extracted from GlobColour global 4-km daily products for both merged and OLCI products. GlobColour "L3b" products with a sinusoidal projection were used so that each extracted pixel covers the same area. For each in situ measurement covered by a product, a matchup of 1 × 1 and 3 × 3 pixels around the in situ location was extracted. No specific quality filtering was applied at this stage because L3 products already exclude bad quality Level-2 pixels (ACRI-ST GlobColour Team et al., 2017). Averaged data based on 3 × 3 pixels were computed using the standard MERMAID tools (http://mermaid.acri.fr/) which follows the protocol from Bailey and Werdell (2006), in summary:

- only matchups containing at least 50% of valid pixels were kept;
- outlier pixels with (pixel value – median value) greater than ±1.5 ∗ standard deviation were removed;
- the matchups were removed if the coefficient of variation (CV) of the remaining pixels was higher than 0.15.

The same extraction and averaging protocol was used for merged and OLCI matchups. Based on the two HPLC pigment databases in Sect. 2.1.1, we have obtained the following matchups:

1) Matchups between daily merged Rrs products and in situ PFT data: the R_{rs} spectra at multispectral bands collocated with the PFT data derived from the Pigment Database I in Sect. 2.1.1 were extracted from the merged products (including SeaWiFS, MODIS, MERIS, VIIRS) from 1997 to 2012 archived in the GlobColour database. The extracted R_{rs} matchups included 1 × 1 pixel, and averaged R_{rs} by 3 × 3 pixels with the median and the standard deviation for each matchup. However, the same wavebands for R_{rs} data are not always available because different sensors have different spectral coverage at different periods (in addition to the exclusion of data with bad quality). Table 1 lists the numbers of matchups with different band combinations (from six to twelve bands) for R_{rs} matchups with 1 × 1 pixel and 3 × 3 pixels, respectively. Fig. 2 shows the corresponding geographical locations of 1 × 1 pixel matchups for R_{rs} at eight, nine and eleven bands, where the matchups were to some extent still globally distributed.
2) Matchups between daily OLCI R_{rs} and in situ PFT data: the Pigment Database II in Sect. 2.1.1 was used to derive the in situ PFT data and extract the corresponding OLCI R_{rs} matchups from 2016 to 2018. Table 2 lists the numbers of matchups with 10, 11 and 12 wavebands for R_{rs} data from S3A OLCI with 1 × 1 pixel and 3 × 3 pixels, respectively. Note that OLCI also includes the 709 nm and that OLCI itself does not have a band at 555 nm, but GlobColour database provides for MERIS and OLCI sensors the 555 nm through an inter-spectral conversion using:

$R_{rs}(555) = R_{rs}(560) * (1.02542–0.03757 * y - 0.00171 * y^2 + 0.0035 * y^3 + 0.00057 * y^4)$, where $y = \log10(\mathrm{CHL1})$ and CHL1 is the TChl-*a* concentration estimated by OC4 algorithm (ACRI-ST GlobColour Team et al., 2017). With this conversion, R_{rs} at 555 nm for OLCI were also included in our study.

2.2. Empirical orthogonal functions (EOF) based algorithm for PFT retrieval

2.2.1. EOF-based statistical approach

Following Bracher et al. (2015), each R_{rs} spectrum was firstly standardized by subtracting the mean spectral value and then divided by the spectral standard deviation (Taylor et al., 2013). The standardized data set of R_{rs}, denoted as matrix X (*M* observations × *N* wavelengths), was collocated to the respective DPA-based PFT data set C with *M* observations and 6 PFTs (*M* might be different for the six PFTs). As indicated in the model training box of Fig. 3, singular value decomposition (SVD) was applied to X for deriving the EOF modes:

$$X = U\Lambda V^T, \tag{1}$$

where matrix U (*M* × *N*) contains column vectors of scores associated with EOF modes, matrix V (*N* × *N*) contains the EOF loadings (spectral pattern), and Λ is an N × N matrix containing the singular values of X on the diagonal in decreasing order. For the PFT Chl-*a* prediction,

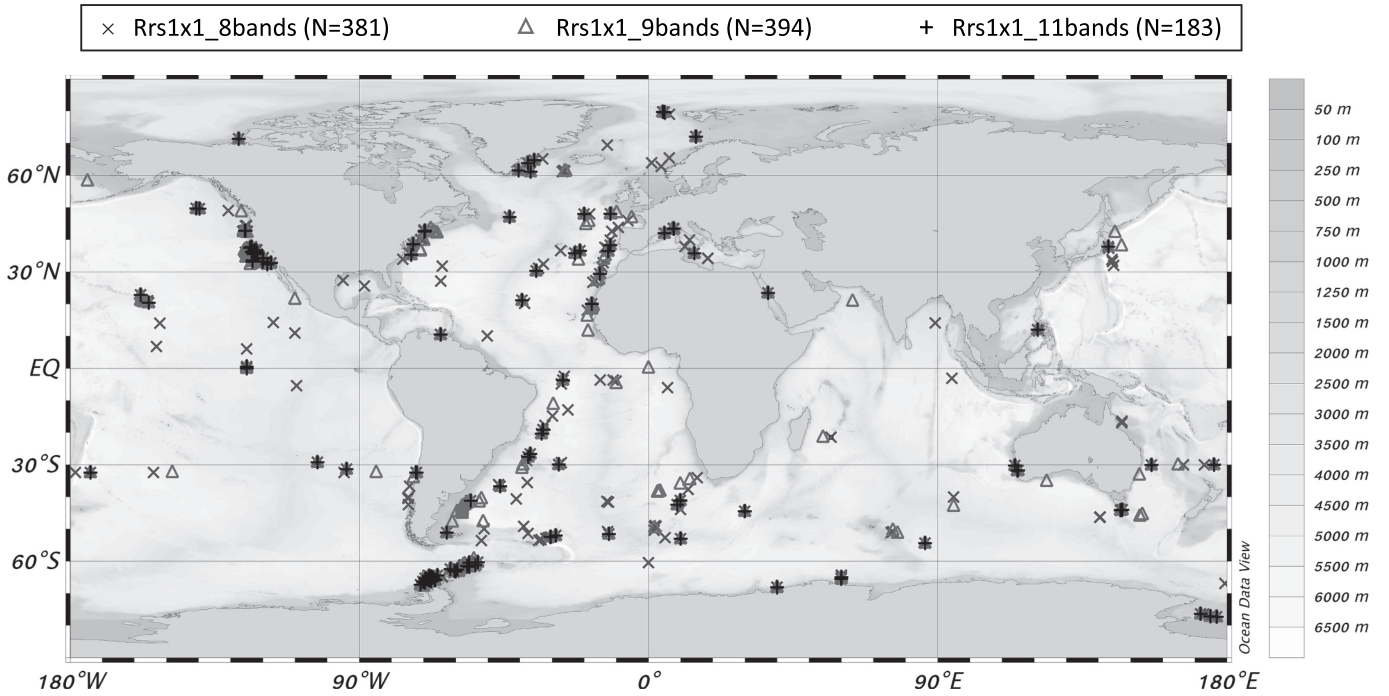

Fig. 2. Geographical locations of the single pixel matchups for merged R_{rs} at eight (in ×), nine (in △) and eleven bands (in +). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

generalized linear models (GLM) were created expressing the log-transformed Chl-*a* concentrations of each PFT, C_p, as a function of a subset of EOF scores (U). EOF modes with standard deviations (singular values from Λ) that are <0.0001 times the standard deviation of the first EOF mode were considered insignificant and thus omitted. The regression model for PFT prediction was expressed as:

$$\ln(C_p) = a_0 + a_1u_1 + a_2u_2 + \dots a_nu_n, \tag{2}$$

where $u_{1,2,\dots n}$ are the leading n EOFs from column vectors of U, a_0 is the intercept and $a_{1,2,\dots n}$ are the regression coefficients. In addition, a stepwise routine was applied to search for smaller regression models, i.e., less u variables, through minimization of the Akaike information criterion (AIC). The significance of included terms was defined by the change in AIC (ΔAIC) with each term's removal.

2.2.2. Model assessment

We consider the coefficient of determination (R^2), the slope (S) and the intercept (a) of the GLM regression, which are based on the log-scaled predicted ($\ln(C_p)$) against the log-scaled observed ($\ln(C_o)$) PFT Chl-*a* concentration data, while the root-mean-square difference (RMSD), the median percent difference (MDPD), and the bias are based on the non-log-transformed data. Model performance statistics are expressed as:

$$R^2 = \frac{\sum_{i=1}^{M}(\ln(C_{pi}) - \ln(\overline{C_{oi}}))^2}{\sum_{i=1}^{M}(\ln(C_{oi}) - \ln(\overline{C_{oi}}))^2}, \tag{3}$$

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{M}(C_{pi} - C_{oi})^2}{M}}, \tag{4}$$

$$\mathrm{MDPD} = \text{Median of}\left[\frac{|(C_{pi} - C_{oi})|}{C_{oi}} \times 100\right], i = 1, \dots\mathrm{M}, \tag{5}$$

$$\mathrm{bias} = \frac{100}{M}\sum_{i=1}^{M}\frac{(C_{pi} - C_{oi})}{C_{oi}}, \tag{6}$$

where M is the number of observations in C_o, and $\overline{C_{oi}}$ is the mean of the observations, i.e., $\overline{C_{oi}} = \frac{1}{M}\sum_{i=1}^{M} C_{oi}$.

To test the robustness of the fitted model, cross-validation of the model fitting was carried out, similar to the procedure performed in Bracher et al. (2015). The collocated data were randomly split into two subsets, in which 80% of the data was used for model fitting/training, which included X^{train} (standardized R_{rs} spectra) and C^{train} (PFT Chl-*a* concentrations), and the rest 20% was used for prediction validation including X^{val} and C^{val}. The procedure was run for 500 permutations to eliminate the model uncertainty produced based on a spatially or temporally biased data set. For each permutation, with Eqs. (1)–(2) and the stepwise routine, a regression model was fitted between $\ln(C^{train})$ and U^{train}. The standardized validation set X^{val} was then projected onto the EOF loadings V^{train} and the inverse of singular values $\Lambda^{train-1}$ to derive their EOF scores U^{val}:

$$U^{val} = X^{val} \bullet V^{train} \bullet \Lambda^{train-1} \tag{7}$$

Table 2
Numbers of available OLCI R_{rs} matchups with 10, 11 and 12 wavebands.

Number of OLCI matchups			OLCI central bands (nm)												No. of bands
1 × 1	3 × 3	3 × 3 all[a]	400	412	443	490	510	555	560	620	665	674	681	709	
115	33	924		×	×	×	×	×	×	×	×	×	×		10
115	33	924	×	×	×	×	×	×	×	×	×	×	×		11
86	25	749	×	×	×	×	×	×	×	×	×	×	×	×	12

[a] 3 × 3 all: all available pixels in the 3 × 3 square were selected, but only matchup data with more than five out of nine pixels available were used.

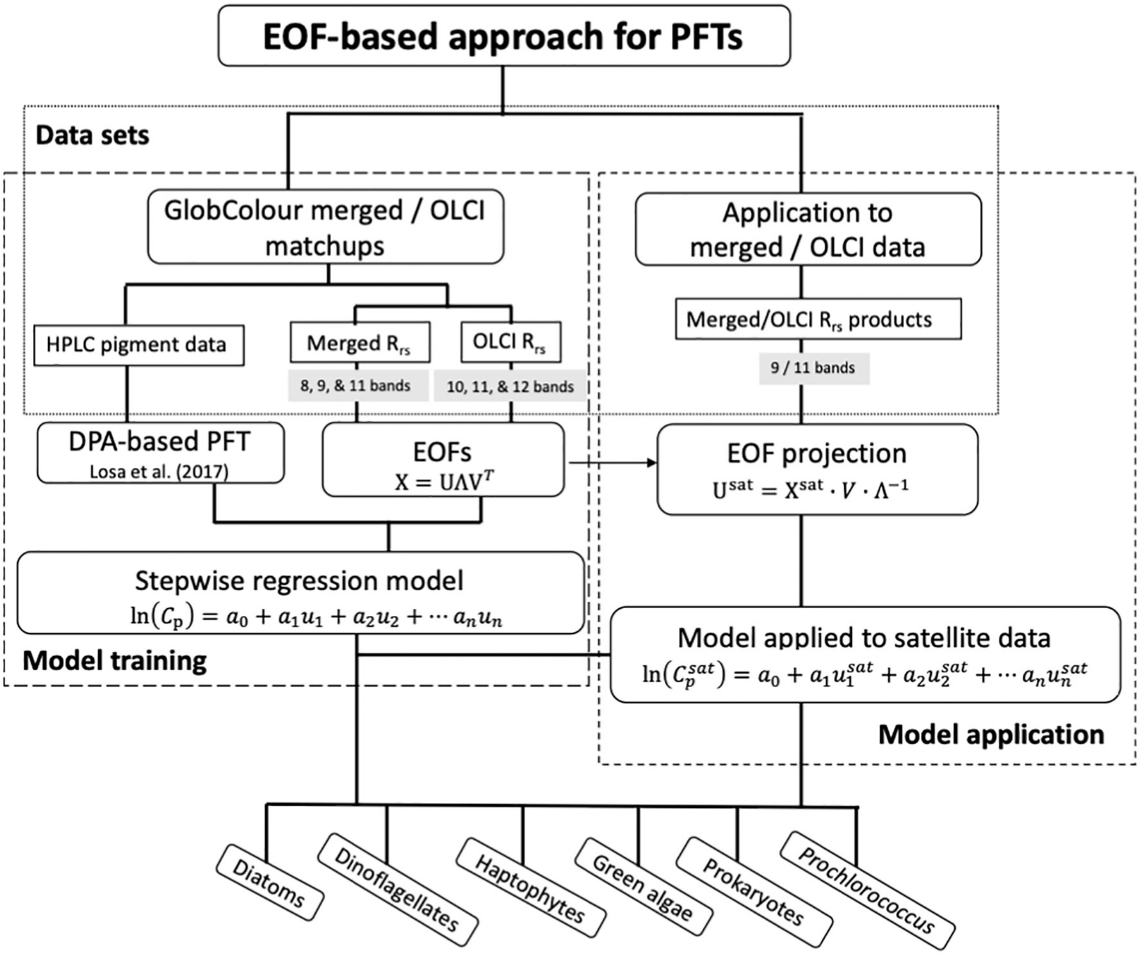
EOF-based approach for PFTs
Data sets
GlobColour merged / OLCI matchups
HPLC pigment data
Merged R_{rs}
OLCI R_{rs}
8, 9, & 11 bands
10, 11, & 12 bands
Application to merged / OLCI data
Merged/OLCI R_{rs} products
9 / 11 bands
DPA-based PFT
Losa et al. (2017)
EOFs
$X = U\Lambda V^T$
EOF projection
$U^{sat} = X^{sat} \cdot V \cdot \Lambda^{-1}$
Stepwise regression model
$\ln(C_p) = a_0 + a_1u_1 + a_2u_2 + \cdots a_nu_n$
Model training
Model applied to satellite data
$\ln(C_p^{sat}) = a_0 + a_1u_1^{sat} + a_2u_2^{sat} + \cdots a_nu_n^{sat}$
Model application
Diatoms
Dinoflagellates
Haptophytes
Green algae
Prokaryotes
Prochlorococcus

Fig. 3. Schematic flowchart of the EOF-based algorithm for predicting six PFTs with different input data sets. The left dashed-line box depicts the model training with the pigment-satellite matchup data and the right dashed-line box depicts the model application to satellite products. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Lastly, the PFT Chl-*a* concentrations for the validation data set ($C_p{}^{\mathrm{val}}$) were predicted using U^{val} of the selected EOF modes and the corresponding regression coefficients. The pairs of the observed and predicted PFT concentrations ($C_o{}^{\mathrm{val}}$ and $C_p{}^{\mathrm{val}}$) of the 500 permutations were recorded for model assessment.

For each permutation, the R^2 for cross validation based on $\ln(C_p{}^{\mathrm{val}})$ versus $\ln(C_o{}^{\mathrm{val}})$ is determined, and the mean value of the R^2 from all permutations (R^2cv) is calculated. Similarly, other statistical parameters for cross validation are determined as follows by taking the mean values of the parameters from all permutations:

$$R^2cv = \frac{\sum_{i=1}^{M} (\ln(c_{pi}^{val}) - \ln(\overline{c_{oi}^{val}}))^2}{\sum_{i=1}^{M} (\ln(c_{oi}^{val}) - \ln(\overline{c_{oi}^{val}}))^2} \quad (8)$$

$$\mathrm{RMSDcv} = \sqrt{\frac{\sum_{i=1}^{M} (C_{pi}^{val} - C_{oi}^{val})^2}{M}} \quad (9)$$

$$\mathrm{MDPDcv} = \text{Median of} \left[\frac{|(C_{pi}^{val} - C_{oi}^{val})|}{C_{pi}^{val}} \times 100\right], \text{ i} = 1, \ldots M \text{ (number of points for validation)} \quad (10)$$

2.2.3. PFT predictions from satellite data

As illustrated in Fig. 3 (model application part), we were able to apply the EOF analysis to satellite R_{rs} data listed in Sect. 2.1.2. Following Bracher et al. (2015), to predict PFTs globally using R_{rs} data from merged OC or OLCI products, for which we do not have corresponding pigment and PFT measurements, we projected standardized R_{rs} data from the satellite onto the EOF loadings (V) to derive a new set of EOF scores (U^{sat}), which was subsequently used for the prediction with the fitted model (see equations in model application of Fig. 3), where a_0 and $a_{1,2,\ldots n}$ were taken from the model developed with matchups from merged products or OLCI data as listed in Sect. 2.1.2.

2.3. PFT relative dominance

With the six retrieved PFTs in our study, we classified the relative PFT dominance in terms of Chl-*a* concentration on a global scale. The classification was performed simply based on the absolute values of the retrieved PFT Chl-*a* concentrations. For each set of the monthly PFT products, two steps were performed as follows. Step 1: the five PFTs – diatoms, dinoflagellates, haptophytes, green algae and prokaryotes – were compared pixelwise and the one with the highest Chl-*a* concentration was considered as the dominant PFT at this particular pixel. Since prokaryotes mainly contain *Prochlorococcus* and *Synechococcus*-like-cyanobacteria (SLC), Step 2 was performed to further assign the dominance of prokaryotes to either *Prochlorococcus*-dominated or SLC-dominated type. That is, for pixels where prokaryotes were the dominant group, we then compared the retrieved *Prochlorococcus* with prokaryotes – pixel with *Prochlorococcus* Chl-*a* concentration higher than 50% of that of the prokaryotes was defined as *Prochlorococcus* dominated, otherwise it was SLC dominated. With this straightforward classification we finally derived the dominance of diatoms,

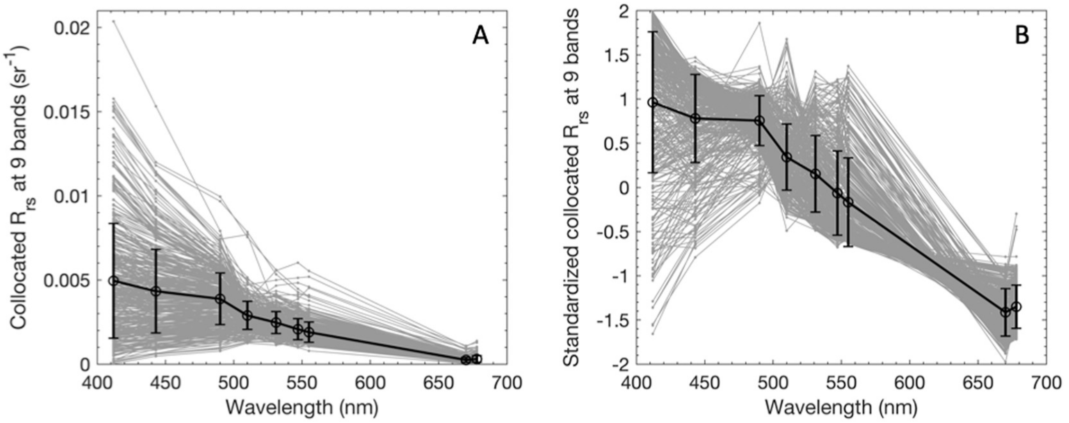

Fig. 4. (A) R_{rs} spectra at nine bands and (B) the corresponding standardized R_{rs} spectra from merged OC matchups at 1 × 1 pixel (in grey) with the mean spectra and standard deviation (black line with error bars).

dinoflagellates, haptophytes, green algae, *Prochlorococcus* and SLC from EOF-based PFTs.

3. Results and discussion

3.1. EOF analysis of Rrs data sets from GlobColour matchups

The matchups of satellite R_{rs} data sets highlighted in Table 1 with eight, nine and eleven bands (namely R_{rs_8}, R_{rs_9}, and R_{rs_11}) were taken as input data for the corresponding EOF analysis, respectively. The choice of the number of bands was based on previous positive experience with the eight MERIS bands (Bracher et al., 2015). In addition, it was tested if more spectral information would improve the retrieval results. As an example to illustrate satellite R_{rs} matchups, Fig. 4 shows the spectra of R_{rs_9} and the corresponding standardized spectra used in the EOF analysis. Most of the R_{rs} spectra presented quite typical spectral features of clean open ocean waters, i.e., high reflectance presented in blue band. However, our data set also contained cases of phytoplankton-rich waters with high reflectance in the green. With hyperspectral $R_{rs,}$ a few bio-optical features related to phytoplankton pigments and thus to PFTs can be caught only when they are prominent enough, such as phycocyanin (a marker pigment for cyanobacteria) which causes an obvious trough in 620–630 nm. While most spectral features in hyperspectral R_{rs} are often caused by a combined effect, e.g., both absorption and fluorescence peaks of phycoerythrin are located in green bands, where chlorophylls have the minimum absorption (Soja-Woźniak et al., 2017). With limited number of wavebands measured by multispectral sensors, it is even more challenging to identify directly the spectral features in terms of specific pigments of phytoplankton types.

As a statistical approach, EOF analysis on multispectral R_{rs} may not be able to catch the entire PFT absorption and scattering properties, but it provides information on to what extent the EOF modes (which have each their specific spectrum) are correlated to the PFTs. Following Sect. 2.2.1, the standardized R_{rs_8}, R_{rs_9}, and R_{rs_11} were decomposed by Eq. (1) into seven, eight, and ten EOF modes, respectively. As shown in Table 3, the first four modes already explain 99.51% to 99.71% of the total variance, with the first mode explaining 79.11%–82.51% of the total variance. Though previous studies (e.g., Craig et al., 2012; Bracher et al., 2015) have investigated the underlying bio-optical signature that the first several EOF modes may carry, it is still difficult to well define the distinct linkage between the EOF modes and the specific pigments or PFTs, as the significance level of the modes may change in different water types (Craig et al., 2012), and the PFT information cannot be the first-order reflected by the EOF modes derived from multispectral R_{rs} data. Nevertheless, a stepwise regression routine, via which the important modes to a certain PFT can be retained, was used to determine the PFT prediction models. Since the in situ PFT Chl-*a* concentrations derived from DPA are based on the marker pigments that were mostly identified in Bracher et al. (2015), we followed their study and included in the prediction model higher EOF modes. Though they contributed only a minute portion to the total R_{rs} variance, they might still inherit the optical signature by phytoplankton (partly group specific) pigments and therefore, be statistically significant for the prediction.

3.2. EOF-based algorithm for PFT retrievals

3.2.1. Stepwise regression procedure

As illustrated in Sect. 2.2.1, a stepwise routine was applied to determine the best EOF prediction model. The ΔAIC indicating the relative importance of the included terms (EOF modes) was presented in Table 4. For all three data sets, EOF-2 was the most important term in the respective models for TChl-*a* and Chl-*a* concentrations of most PFTs except for prokaryotes (also except for *Prochlorococcus* for R_{rs_11}). However, the second important EOF mode differed in PFT prediction models, and the total number of the EOF modes included in each model also varied. For instance, with data set R_{rs_9} only three EOFs were selected for *Prochlorococcus*, but all eight EOFs were included for haptophytes. It was also found that the most relevant EOF modes for prokaryotes and *Prochlorococcus* prediction were not fixed among the three R_{rs} data sets, indicating that the models are vulnerable and unstable, which was also reflected in their low performance (see Table 5 and Fig. 5). According to Bracher et al. (2015), EOF-2 is associated with Chl-*a*; the high importance of EOF-2 in the PFTs is likely due to the

Table 3
Percentage of total variance explained (%) by the decomposed EOF modes derived from three satellite matchup data sets R_{rs_8}, R_{rs_9}, and R_{rs_11} within the 1 × 1 pixel.

% of variance	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8	EOF-9	EOF-10
R_{rs_8} 1 × 1	82.51	14.78	2.14	0.28	0.18	0.08	0.02			
R_{rs_9} 1 × 1	79.11	17.75	2.03	0.79	0.22	0.06	0.03	0.01		
R_{rs_11} 1 × 1	79.28	17.60	1.76	0.87	0.25	0.13	0.05	0.05	0.01	0.01

Table 4
ΔAIC for the predictions of the TChl-*a* and six PFT Chl-*a* concentration by the EOF modes based on R_{rs} 1 × 1 matchups with eight, nine and eleven bands from merged OC products (R_{rs_8} 1 × 1, R_{rs_9} 1 × 1, and R_{rs_11} 1 × 1). Bold highlights the EOF mode with the highest ΔAIC for TChl-*a* and each derived PFT.

R_{rs_8} 1 × 1	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7
TChl-*a*	16.02	283.25	105.43	24.82			2.48
Diatom	8.16	130.24	90.83	10.53		0.89	
Haptophytes	42.34	214.50	4.57	24.04		1.45	
Prokaryotes	12.52						5.49
Dinoflagellates	5.69	122.46	54.56	0.41			
Green algae	1.14	92.25	8.05	1.49		9.25	
Prochlorococcus		7.29				6.87	0.73

R_{rs_9} 1 × 1	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8
TChl-*a*	38.27	416.17	109.26	58.11		3.13	10.07	
Diatom	20.05	217.09	80.52	30.17		9.43	7.07	1.14
Haptophytes	41.31	266.08	1.32	7.33	1.89	4.64	4.1	7.45
Prokaryotes	16.71		7.32	0.63	3.24	22.24	10.93	2.84
Dinoflagellates	4.85	177.95	27.59	24.62				7.14
Green algae		173.91	2.59		2.29	7.43	4.46	
Prochlorococcus		20.63				12.66	1.97	

R_{rs_11} 1 × 1	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8	EOF-9
TChl-*a*	13.34	181.37	48.59	6.66	1.94				
Diatom	7.86	105.23	44.49		0.32	3.41			
Haptophytes	25.35	123.10		0.58	0.82	0.55		6.38	1.32
Prokaryotes	9.45				3.15	6.86	0.55	4.52	
Dinoflagellates	10.32	86.57	8.95	5.10					2.03
Green algae		102.48			1.73		8.36	1.82	0.39
Prochlorococcus	9.30			0.06	0.65			10.87	

elevation of Chl-*a* concentration in most of the PFTs when TChl-*a* increases. Since prokaryotes and *Prochlorococcus* mainly dominate in oligotrophic regions with very low biomass concentration, they do not have a collinearity in their Chl-*a* concentration with TChl-*a* as most other PFTs. A similar statement was also given in Bracher et al. (2015) for predicting pigments.

3.2.2. Performance of retrieval models based on matchups of merged R_{rs} data sets

Satellite PFT Chl-*a* and TChl-*a* concentrations were predicted with the regression models built based on the EOF scores derived from the R_{rs} data sets and the in situ PFT Chl-*a* concentrations. Matchups at different band settings and pixel level (1 × 1, 3 × 3 pixels) were taken as input for comparison between the results from different band

Table 5
Statistics of regression models for TChl-*a* and six PFT Chl-*a* concentrations using EOF modes based on R_{rs} matchups R_{rs_8}, R_{rs_9}, and R_{rs_11} within 1 × 1 pixel from merged products. Cross-validation (cv) results are presented with 500 permutations for data splitting into 80% of the data used for training and 20% for validation. N = number of valid matchups for each parameter.

	N	MDPD (%)	RMSD (mg m^{-3})	R^2	MDPDcv (%)	RMSDcv (mg m^{-3})	R^2cv
R_{rs_8} 1 × 1							
TChl-*a*	381	40.66	1.38	0.72	40.97	1.40	0.71
Diatoms	286	80.28	1.25	0.59	81.56	1.27	0.58
Haptophytes	366	57.16	0.30	0.58	57.97	0.30	0.54
Prokaryotes	348	62.32	0.15	0.05	62.95	0.14	0.04
Dinoflagellates	258	59.14	0.91	0.56	60.52	0.64	0.54
Green algae	239	60.51	0.12	0.50	61.81	0.12	0.47
Prochlorococcus	139	41.92	0.03	0.13	42.77	0.03	0.08
R_{rs_9} 1 × 1							
TChl-*a*	394	37.41	1.24	0.76	37.08	1.27	0.75
Diatoms	306	73.70	1.21	0.65	74.74	1.29	0.63
Haptophytes	387	47.16	0.22	0.64	48.62	0.24	0.61
Prokaryotes	367	53.70	0.13	0.15	55.08	0.13	0.11
Dinoflagellates	272	55.32	0.93	0.62	57.29	0.72	0.59
Green algae	262	55.81	0.11	0.51	56.26	0.11	0.48
Prochlorococcus	142	39.65	0.02	0.24	42.68	0.02	0.18
R_{rs_11} 1 × 1							
TChl-*a*	183	38.15	1.42	0.75	40.20	1.43	0.73
Diatoms	148	75.56	1.26	0.68	77.42	1.28	0.64
Haptophytes	179	53.04	0.28	0.61	55.84	0.29	0.54
Prokaryotes	171	61.41	0.17	0.13	62.61	0.16	0.08
Dinoflagellates	132	64.32	1.20	0.56	66.75	0.83	0.51
Green algae	116	54.52	0.12	0.60	58.60	0.13	0.48
Prochlorococcus	52	41.83	0.02	0.35	50.60	0.03	0.14

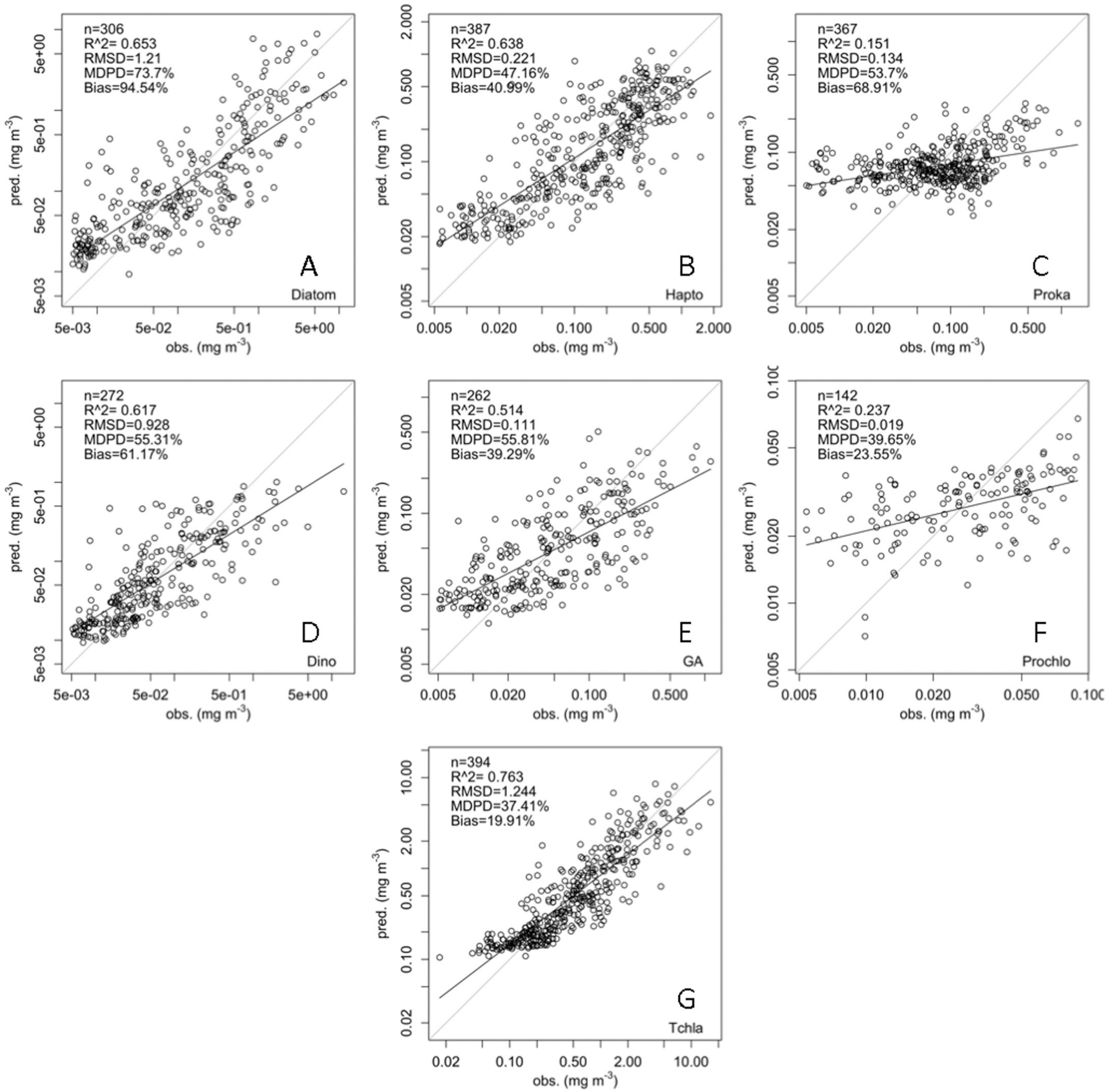

Fig. 5. Regressions between observed (x-axis, obs.) and predicted (y-axis, pred.) Chl-*a* concentrations of (A) diatoms, (B) haptophytes, (C) prokaryotes, (D) dinoflagellates, (E) green algae, (F) *Prochlorococcus*, and (G) TChl-*a* using EOF modes derived from merged R_{rs} products at 9 bands (1 × 1 pixel).

numbers, pixels and data points. Prediction model performances of using R_{rs} data sets with 1 × 1 and 3 × 3 matchups were statistically similar. Therefore, here we only presented and discussed in detail the results of the 1 × 1 pixel matchups, as there were more collocated data which should provide more robust predictions (statistics based on R_{rs} 3 × 3 data sets are presented in Table S2 in the supplementary document). The prediction models developed from the 1 × 1 collocated R_{rs} data sets were also later applied to the satellite products.

Statistics of the EOF-based regression models are listed in Table 5 for different R_{rs} data sets (R_{rs_8}, R_{rs_9} and R_{rs_11}). The predicted PFT Chl-*a* concentrations display slight differences between different band settings of the input R_{rs}. With all three data sets, the predicted and observed (based on in situ data) TChl-*a* and Chl-*a* concentrations for diatoms, haptophytes, dinoflagellates and green algae are well correlated, with $R^2 \geq 0.50$ and R^2cv ≥ 0.47. TChl-*a* has the highest correlation ($R^2 \geq 0.72$), while Prokaryotes and *Prochlorococcus* have the weakest correlation between the predicted and observed concentrations but are generally better correlated using data set R_{rs_9} compared to the other two data sets. The MDPD are lowest for TChl-*a* and *Prochlorococcus* (< 42%) and low for haptophytes, dinoflagellates, green algae and prokaryotes (< 60% for data set R_{rs}_9). The highest MDPD was found for diatoms (< 80%). The MDPDcv of all cases are slightly higher but still comparable with the MDPD, indicating that the prediction models are stabilized. R_{rs_9} presents an overall lowest MDPD among the three data sets. RMSD values were calculated in non-log transformed manner, and thus vary depending on the corresponding

Chl-*a* concentration ranges of individual PFTs. TChl-*a* has the highest RMSD as it is the indicator of all phytoplankton biomass, whereas Chl-*a* of *Prochlorococcus* which is always low in concentration has the lowest RMSD. Among the three data sets, the lowest RMSD are found for R_{rs_9}. Hence, we conclude that the EOF-based models with R_{rs} at nine bands (see Table 1) perform best and slightly better than those with eleven bands, while the weakest are the models based on eight bands. This to some extent indicates that the performance of prediction models is not only subject to the number of bands (i.e., the more bands the better), but also to the number of matchups (with R_{rs_11} the least).

As a summary, Fig. 5 shows the observed against the predicted TChl-*a* and Chl-*a* concentrations for the six PFTs by the EOF-based method using R_{rs_9}. Corresponding to the statistics in Table 5, TChl-*a* and Chl-*a* of diatoms, haptophytes, dinoflagellates, and green algae which have relatively larger ranges in magnitude show relatively good predictions, with regression lines close to the 1:1 reference line and lower intercepts. Prokaryotes and *Prochlorococcus* are of weaker correlations with slopes much lower than 1 and higher intercepts, mainly due to their low concentrations, the narrow range of the variation, as well as the low variability in the concentrations especially for prokaryotes that could not be well interpreted by the EOF modes. Slopes of all regression lines <1 indicate that the models to some extent overestimate the variables in low concentrations and underestimate them in higher concentrations. Slopes of <1 were also shown in Bracher et al. (2015) for all the predictions of pigments and pigment composition, though in their study the prediction performance for some important pigments was statistically better compared to our prediction of PFT Chl-*a* concentration. Among the well predicted pigments in Bracher et al. (2015), zeaxanthin, typically used as a marker pigment for prokaryotes, showed the lowest correlation but reasonable MDPD, which corresponds to our lower R^2 values for prokaryotic phytoplankton. It is worth investigating further the prediction models and perform certain tuning procedure through mathematical methods to reduce these over- or underestimations, especially for picophytoplankton which are usually very low in concentration.

The cross-validation procedure effectively examined the robustness of the prediction models. The statistical parameters for cross-validation (averaged for all 500 permutations with 20% data for prediction) were nearly or as equivalently good as the statistics for the model trained with the whole data set (Table 5). This suggests that the number of data points (matchups) is adequate for a robust model establishment. In fact, in our study there were 52–394 data points for all matchups with different band settings, which is much higher than that was suggested to be necessary for robust model development by Craig et al. (2012) (15 points at a seasonal cycle) and Bracher et al. (2015) (50 points). However, since their studies were rather regional while we are focusing on the global scale, a higher number of points is expected in our study to enable a comprehensive coverage of the global ocean water types. From Table 5 one can see that the statistics of the cross validation are much worse than the original statistics for the green algae and *Prochlorococcus* Chl-*a* predictions using the data set R_{rs_11}, for which less available matchups were obtained. Therefore, though lower R^2 and higher MDPD were obtained with the data set R_{rs_9}, for these two PFTs, the cross-validation showed better results than that from the data set R_{rs_11}, convincing us the nine-band setting of the R_{rs} to be optimal for PFT model applications to satellite products without in situ matchups.

To better understand the performance of the EOF-based algorithm, Fig. S1 in the supplementary document shows the uncertainty for different ocean biomes in the algorithm derived Chl-*a* concentrations of the six PFTs using GlobColour merged R_{rs} at nine bands (global projection of the uncertainty is detailed in the supplementary document). Diatoms show underestimation in coastal regions (mean deviation of -0.11 mg m^{-3} in this biome), slight underestimation in high latitudes and near the equator (~-0.02 mg m^{-3}), and very slight overestimation in the subtropical regions ($\sim$0.013 mg m^{-3}). Haptophytes, dinoflagellates, and green algae present similar uncertainty distributions, i.e., overestimation in higher than 40°N and subtropical regions and underestimation near the equator and in the Southern Ocean, but with different amplitudes. Both prokaryotes and *Prochlorococcus* show distinct overestimation in the central part of the oligotrophic gyres (0.026 and 0.014 mg m^{-3}, respectively) but underestimation in the surrounding areas of the gyres (-0.06 and $-$ 0.012 mg m^{-3}, respectively).

3.2.3. Application to merged products for global PFT retrieval

Given that the EOF-based PFT models based on the matchups of merged R_{rs} at nine bands show the best performance, we applied these models (based on the full data set fit) to the merged R_{rs} global products at the same nine bands for the period of 2002–2012. Selection criterion of the nine bands from merged R_{rs} products is detailed in Sect. 3.2.2 of the supplementary document. The numerical matrices and regression coefficients determined by Eqs. (1) and (2) used for the model implementation to the merged R_{rs} products at nine bands are also explained and provided in Tables S3 and S4 in the supplementary document.

Fig. 6 illustrates the global mean distribution Chl-*a* concentration of each PFT, based on the monthly PFT products derived from the merged R_{rs} products with 25 km resolution from 2002 to 2012. Diatom Chl-*a* concentrations are generally higher in high latitudes, marginal seas and coastal upwelling regions but are much lower in the tropical regions and extremely low in the subtropical gyres. The typical diatom abundant regions are higher than 40°N (North Atlantic, Bering Sea and Labrador Sea up to the Arctic Ocean), the Patagonian upwelling and most part of the Southern Ocean. The average Chl-*a* concentration of diatoms over the globe is ~0.08 mg m^{-3}. Chl-*a* concentration of dinoflagellates is low nearly over the whole globe (~0.02 mg m^{-3}) but higher in the Arctic Ocean and Patagonian upwelling. Haptophytes with a global average Chl-*a* of 0.09 mg m^{-3} follow in distributions of the diatoms but have more spread regions of high Chl-*a* in the high latitudes, waters near the coasts, and equatorial regions (such as the west coast of Africa). Chl-*a* concentration of green algae (global average of 0.03 mg m^{-3}) is found typically higher in the Arctic and the near coast oceans around the southern part of South America. Prokaryotes and *Prochlorococcus* show distinctly different distribution features from the other four PFTs. Prokaryotes with a global average Chl-*a* concentration of 0.07 mg m^{-3} are much more abundant in the subtropical regions but also substantially contribute (~5–30% of TChl-*a*) in the Arctic Ocean. Waters such as the Baltic Sea, the east coast of China, and the west coast of Africa (around 5°S and 10–20°N) show very prominent abundance of prokaryotes. *Prochlorococcus* are generally very low on a global scale (global average 0.03 mg m^{-3}), especially in high latitude waters (not really detectable), slightly higher in subtropical regions and apparently abundant in some parts of the west coast of Africa similar to prokaryotes. Distribution of *Prochlorococcus* is supported by previous findings (Flombaum et al., 2013). Their quantitative model based on a large number of observations well defined the assessment of the *Prochlorococcus* abundance and the results match well our retrievals. In general the global average Chl-*a* concentrations of the PFTs retrieved from our study are consistent with those from Hirata et al. (2011), except that prokaryotes Chl-*a* is higher (0.07 mg m^{-3} in our study versus 0.04 mg m^{-3} from Hirata et al., 2011), mainly due to our elevated Chl-*a* prediction in the subtropics for prokaryotes. To illustrate the changes in the PFT Chl-*a* distribution with seasons, the monthly climatological products of each PFT are provided in Figs. S2-S7 in the supplementary document. For instance, diatom blooms are mainly detected during early summer in the Southern Ocean (December–January) and in the subarctic and Arctic waters (May–June). Haptophytes show similar seasonal changes in high latitudes as diatoms, but highly increase during the summer season in the equatorial Atlantic. A strong prokaryotic enhancement is also found during July–August at the west coast of South Africa.

Distribution of TChl-*a* retrieved by the EOF-based algorithm is

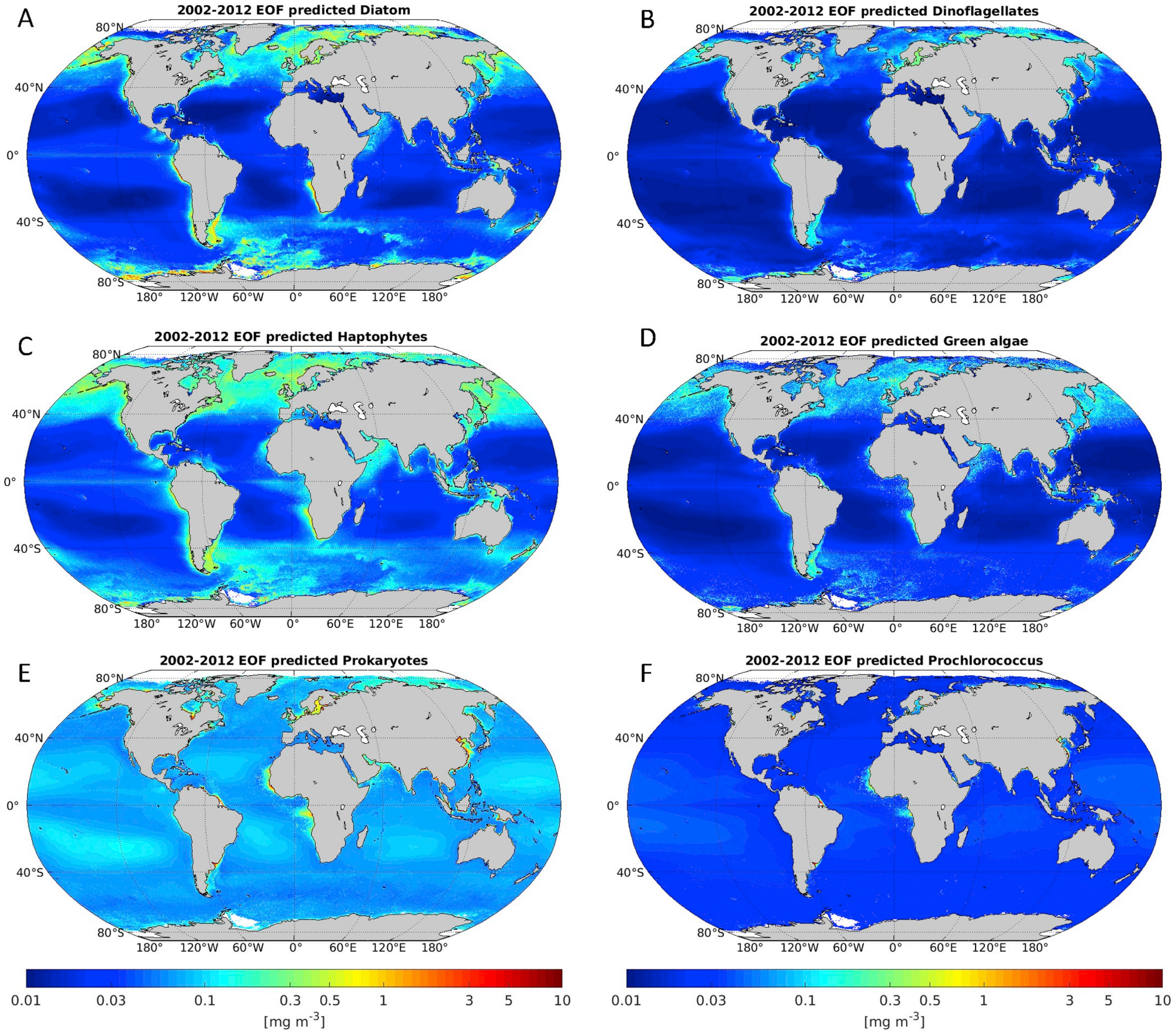
A
2002-2012 EOF predicted Diatom
B
2002-2012 EOF predicted Dinoflagellates
C
2002-2012 EOF predicted Haptophytes
D
2002-2012 EOF predicted Green algae
E
2002-2012 EOF predicted Prokaryotes
F
2002-2012 EOF predicted Prochlorococcus
80°N
40°N
0°
40°S
80°S
180° 120°W 60°W 0° 60°E 120°E 180°
0.01 0.03 0.1 0.3 0.5 1 3 5 10
[mg m-3]

Fig. 6. Ten-year mean distribution (July 2002–April 2012) of the PFT Chl-*a* concentration for (A) diatoms, (B) dinoflagellates, (C) haptophytes, (D) green algae, (E) prokaryotes, and (F) *Prochlorococcus* retrieved by EOF-based algorithm from merged monthly R_{rs} products at nine bands.

presented in comparison to the GlobColour merged ocean chlorophyll products (mean over all years in Fig. 7, and climatological monthly mean in Figs. S8–S9). The ten-year mean of our EOF-based predicted TChl-*a* is generally in good agreement considering the distribution patterns with the standard products, though it is clearly seen that the EOF-based TChl-*a* shows higher/lower values in the subtropical gyres/coastal waters than the standard products. This was however expected, as the EOF-based retrieval models based on matchups already showed an over-/under-estimation for lower/higher values for all the retrieved variables/PFTs, as illustrated in Fig. 5. This flattening effect of the prediction is most prominent in prokaryotes and *Prochlorococcus*, of which the EOF-based models present the weakest correlation. An accurate retrieval of prokaryotic phytoplankton or its corresponding marker pigments (zeaxanthin, divinyl Chl-*a*) has always been a challenge so far (e.g., Bracher et al., 2015; Losa et al., 2017), as the picophytoplankton Chl-*a* concentrations are usually globally very low, even when dominating in oligotrophic oceans. This results in a narrow variation range and low variability in their concentrations compared to other PFTs, and also in a weak imprint on the spectral shape which are limited for the detection via the spectral analysis. An exception is that in the Baltic Sea prokaryotes can have high Chl-*a* concentrations especially during blooms. This is also reflected in our retrievals, though there are no matchups available included in the EOF analysis.

3.3. Evaluation of the EOF-based PFT products

3.3.1. Inter-comparison with other PFT/PSC products

To evaluate our retrieval algorithm, the derived Chl-*a* concentrations of diatoms, haptophytes and prokaryotes were compared with SynSenPFT Chl-*a* of diatoms, coccolithophores and cyanobacteria (Losa et al., 2017) and Chl-*a* of three PSCs (micro- >20 μm, nano- 2–20 μm, and picophytoplankton <2 μm, Sieburth et al., 1978) obtained with the PSC model of Brewin et al. (2010, 2015). Both SynSenPFT and PSC products developed within the frame of the SynSenPFT project (Losa et al., 2017) were available globally at 4 km daily resolution over the period from 2002 to 2012. Prior to the inter-comparison, both products

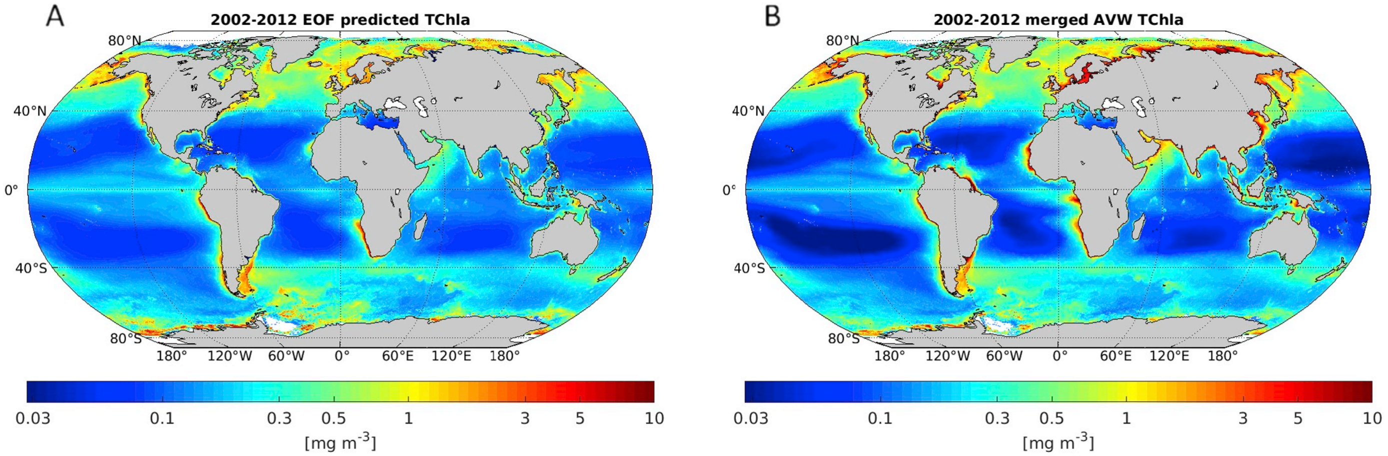

Fig. 7. Ten-year mean distribution (July 2002–April 2012) of (A) TChl-*a* concentration retrieved by EOF-based algorithm from merged monthly R_{rs} products at nine bands and (B) GlobColour AVW merged TChl-*a* concentration based on open ocean L2 chlorophyll products from SeaWIFS, MODIS and MERIS sensors.

were binned to monthly averages and re-gridded to 25 km resolution, to be consistent with our EOF-based PFT products. For simplification, in the following text the SynSenPFT derived Chl-*a* concentrations of diatoms, coccolithophores, and cyanobacteria are denoted as dia-SynSenPFT, coc-SynSenPFT, and cya-SynSenPFT, respectively; the Chl-*a* concentrations of micro-, nano- and picophytoplankton derived from PSC model are denoted as c-micro, c-nano, and c-pico, respectively. The EOF-based Chl-*a* products of the other two PFTs, green algae and *Prochlorococcus* were compared to those derived by OC-PFT method proposed by Hirata et al. (2011) using GlobColour AVW merged TChl-*a* monthly 25-km products as input for the same period (2002–2012). Dinoflagellates were not considered for comparison as the OC-PFT derived dinoflagellates showed very poor validation result (Hirata et al., 2011). It is noteworthy that OC-PFT also allows the retrieval of Chl-*a* concentrations of diatoms, haptophytes and prokaryotes, but as they are intrinsic in the SynSenPFT products (Losa et al., 2017) they were not included as separate products for the inter-comparison.

Following Losa et al. (2017), the time-latitude Hovmöller diagrams were generated covering the monthly means from 2002 to 2012 of the different PFT/PSC products. Since globally the Chl-*a* concentration is typically log-normally distributed (Campbell, 1995), all averaging was done in logarithmic space and then back-transformed to the original scale. The Hovmöller diagrams are presented in Figs. 8–11, where the left side of each subplot shows the monthly variation during the ten-year period (2002–2012), and the right side shows the climatological annual cycle. Since different studies tend to provide different retrieval information in terms of phytoplankton composition, the optimal way for the inter-comparison is to select the variables carrying the most similar PFT information, but one has to keep in mind that the products compared here are not always representing exactly the same quantities.

Diatoms derived from our study (Fig. 8A) and dia-SynSenPFT (Fig. 8B) show similar distributions with both lowest diatom Chl-*a* concentration in the subtropical regions especially in the gyres and higher concentration in high latitudes. Compared to dia-SynSenPFT, the EOF-based diatoms show generally lower Chl-*a* in the polar and tropical regions, however they indicate the same blooming periods for diatoms in May–June in the Arctic and December–January in the Southern Ocean. Dia-SynSenPFT presented distinct higher Chl-*a* from 10°S to 10°N during December to February 2005–2006, 2007–2008 and 2010–2011 than other years, whereas the change between the years is not evident in either our results or the c-micro products (Fig. 8C). Since microphytoplankton contain not only diatoms but also other micro-size phytoplankton such as dinoflagellates, the sum of EOF-based diatoms and dinoflagellates was also shown (Fig. 8D), presenting similar seasonal variation to c-micro but higher/lower Chl-*a* in the gyres/high latitudes.

Before comparing the EOF-based haptophytes to other products, it should be noted that coccolithophores are a main contributing PFT to haptophytes, while haptophytes are a part of nanophytoplankton, with the latter containing also *Phaeocystis*, cryptophytes, and a few other groups. Haptophytes derived from our study (Fig. 9A) are well consistent with coc-SynSenPFT (Fig. 9B), although again our retrievals show a relatively mild pattern with lower Chl-*a* in high latitudes and the 10°S-10°N equator belt. Chl-*a* concentration of coc-SynSenPFT from 10°N to 40°N during the summer time is significantly higher, but this pattern is not found in either our products or c-nano. Our haptophytes present similar distribution with c-nano (Fig. 9c) but lower Chl-*a* in the high latitudes and equatorial regions as expected. The climatological annual cycles of both are in very good agreement in the Southern Ocean, while in the Arctic c-nano shows much Chl-*a* enhancement in May–July. In addition, c-nano spreads more to the north until 25°N from the equator. However, caution should be taken since our DPA derived haptophytes contain only their nanophytoplankton fraction while their picophytoplankton fraction is neglected, whereas Brewin et al. (2015) consider part of the haptophytes in the picophytoplankton group when TChl-*a* is below 0.08 mg m^{-3}.

The overall Chl-*a* concentration of our EOF-based prokaryotes (Fig. 10A) is generally low (0.03–0.20 mg m^{-3}), but higher concentrations are found in the subtropical regions, only slightly lower than the maxima in the Arctic and in the Southern Ocean from 70°S to 80°S during the summer. On the contrary, both distributions of cya-SynSenPFT (Fig. 10B) and c-pico (Fig. 10C) show the lowest Chl-*a* in the gyres. Similar seasonality (with little changes) between the cya-SynSenPFT and c-pico is observed at the mid- to high latitudes, while the EOF-based prokaryotes show slightly lower Chl-*a* maxima as well as a different seasonal change in the Arctic, which have a clear elevation in Chl-*a* from spring to summer. It is noteworthy that the cyanobacteria derived from SynSenPFT include all the prokaryotic phytoplankton (Losa et al., 2017) which should thus be the same product as our EOF retrieved prokaryotes. The product of c-pico from Brewin et al. (2015) contains not only prokaryotes but also other picoeukaryotic phytoplankton (green algae and pico-sized haptophytes), therefore we also presented in Fig. 10D the sum of the prokaryotes and green algae Chl-*a* from our study, which shows much higher Chl-*a* concentration in general compared to c-pico, simply due to the predictions of high Chl-*a* of prokaryotes in the subtropical regions. Nevertheless, the high prokaryotes Chl-*a* concentrations in the subtropical regions are not only shown in our study, but are also found in the cyanobacteria simulated by NASA Ocean Biogeochemical Model (NOBM), which is a global biogeochemical model with coupled circulation and radiative models (Gregg, 2002; Gregg and Casey, 2007, figure not shown here but can be found in Losa et al., 2017). However, our prokaryotic phytoplankton

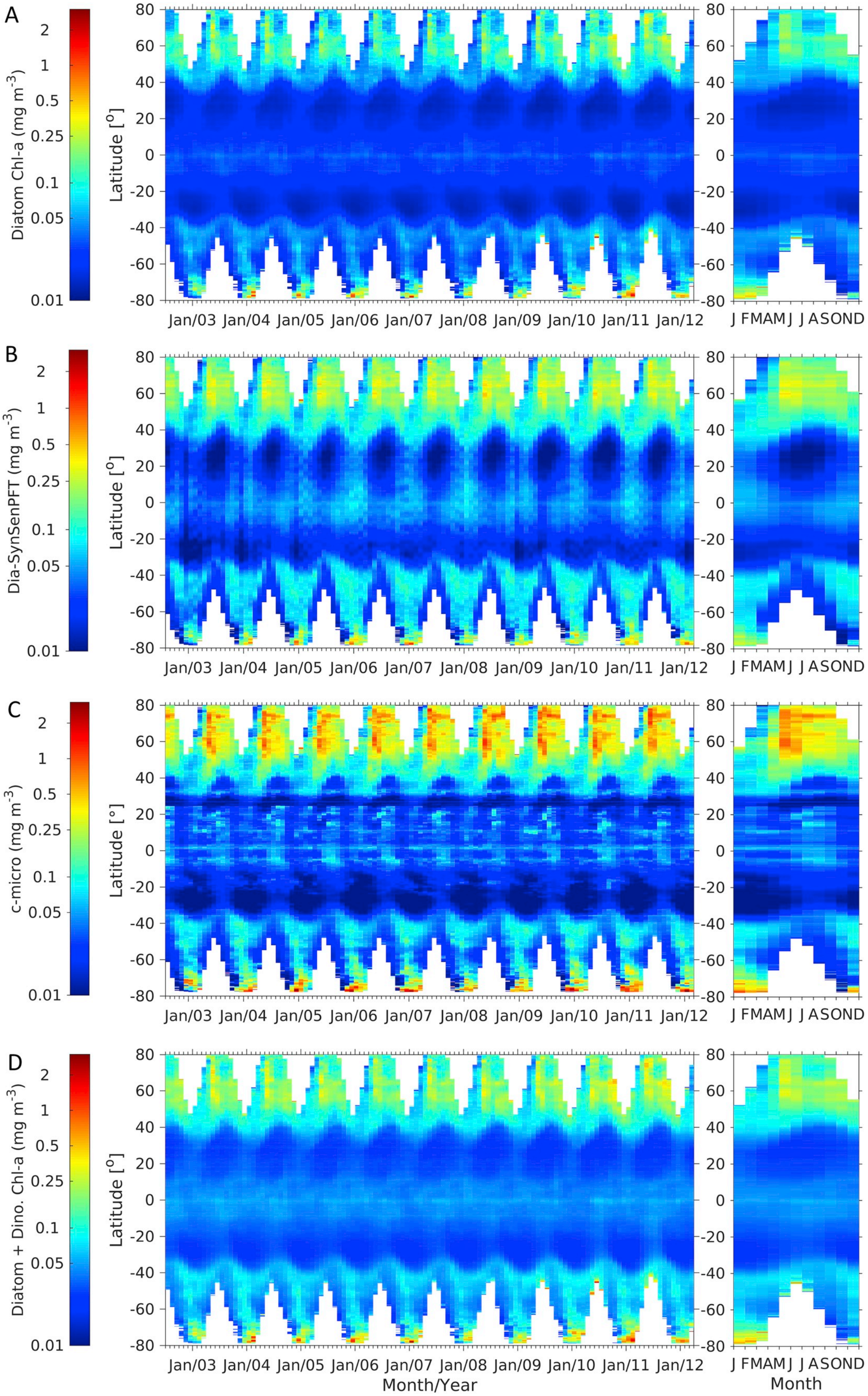

Fig. 8. Hovmöller diagrams of Chl-*a* concentrations of (A) diatoms derived from our study, (B) dia-SynSenPFT (Losa et al., 2017), (C) c-micro derived from PSC method (Brewin et al., 2015), and (D) sum of diatoms and dinoflagellates (Diatom + Dino.) from our study.

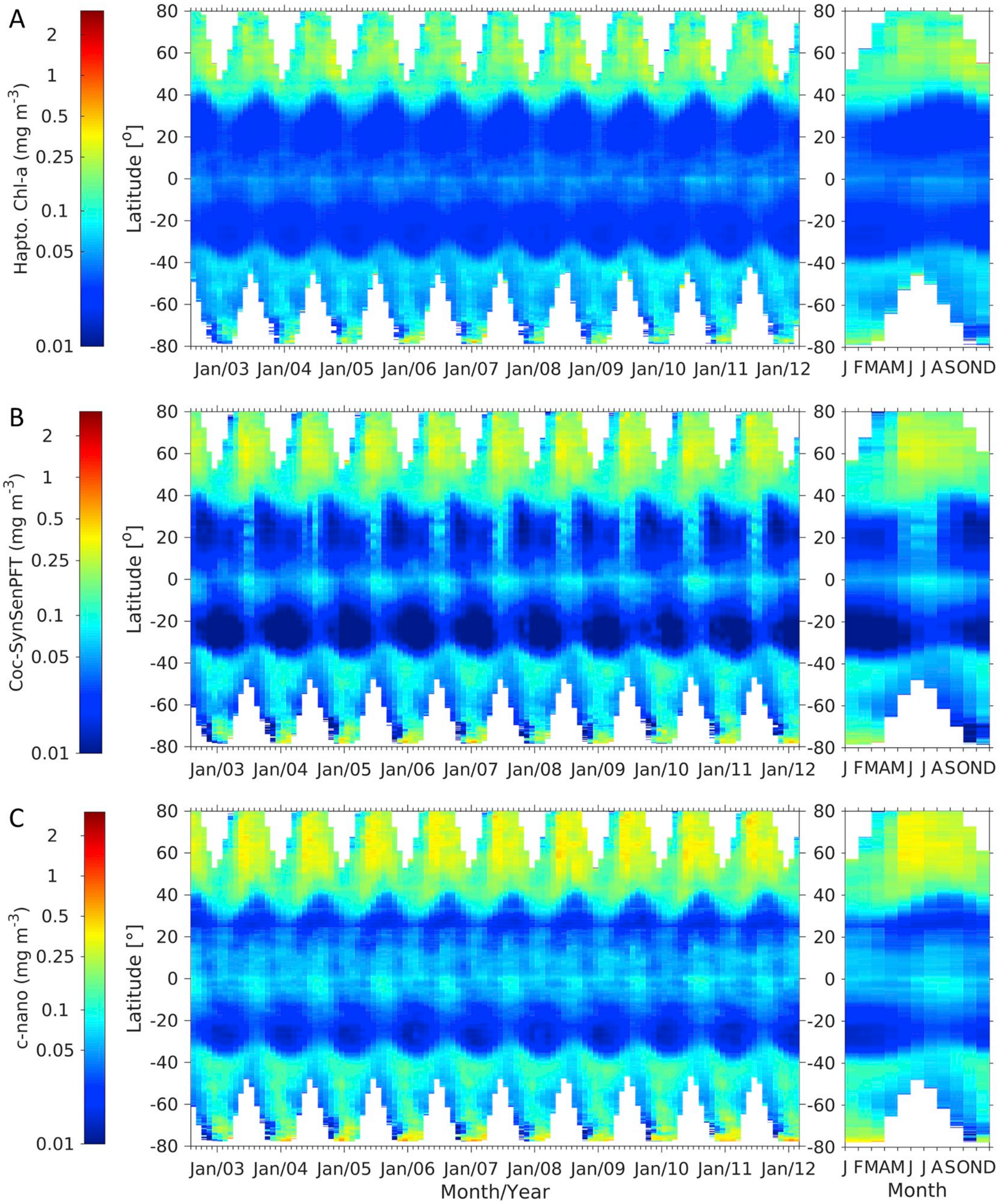

Fig. 9. Hovmöller diagrams of Chl-*a* concentrations of (A) haptophytes derived from our study, (B) coc-SynSenPFT (Losa et al., 2017), and (C) c-nano derived from PSC method (Brewin et al., 2015).

retrieval performance still needs to be further improved by potentially scaling the low concentration range or using non-linear prediction models.

Hovmöller diagrams of green algae and *Prochlorococcus* Chl-*a* concentrations derived by our study (Fig. 11A–B) are presented in comparison with those from OC-PFT (Hirata et al., 2011, Fig. 11D–E). Green algae in both products show distinct seasonality but Chl-*a* concentrations of green algae from our study are generally lower than those from OC-PFT (except for the subtropical regions), especially in the Arctic where OC-PFT shows enhanced green algae from late spring to early winter, whereas the EOF-based green algae show the lowest Chl-*a* during summer and increase in autumn to winter. *Prochlorococcus* Chl-*a* is generally very low (< 0.1 mg m^{-3}) for both products with quite different patterns presented. The EOF-based *Prochlorococcus* Chl-*a* concentrations are higher in mid- to low latitudes but lower in polar regions, corresponding to previous findings by Flombaum et al. (2013), while the OC-PFT *Prochlorococcus* shows higher Chl-*a* in the Southern Ocean which is outside the known distribution range and likely caused by undersampling of the in situ data (Hirata et al., 2011). Dinoflagellates show similar distribution with diatoms but with much lower Chl-*a* concentration, which is almost neglectable in subtropical regions and only higher than 0.05 mg m^{-3} in higher than 40°N with clear seasonality observed (Fig. 11C). However, an equivalent product is still necessary for dinoflagellates evaluation.

3.3.2. PFT Chl-a dominance comparison with PHYSAT products

We compared the PFT Chl-*a* dominance derived from our study for the period of 2002–2012 to the PHYSAT product from 1997 to 2006 (Alvain et al., 2008) which empirically relates the radiance anomaly to specific dominant phytoplankton groups. It is worth noting that the periods of the two compared products do not coincide, because we could only obtain the 12-month PHYSAT climatology data from 1997 to

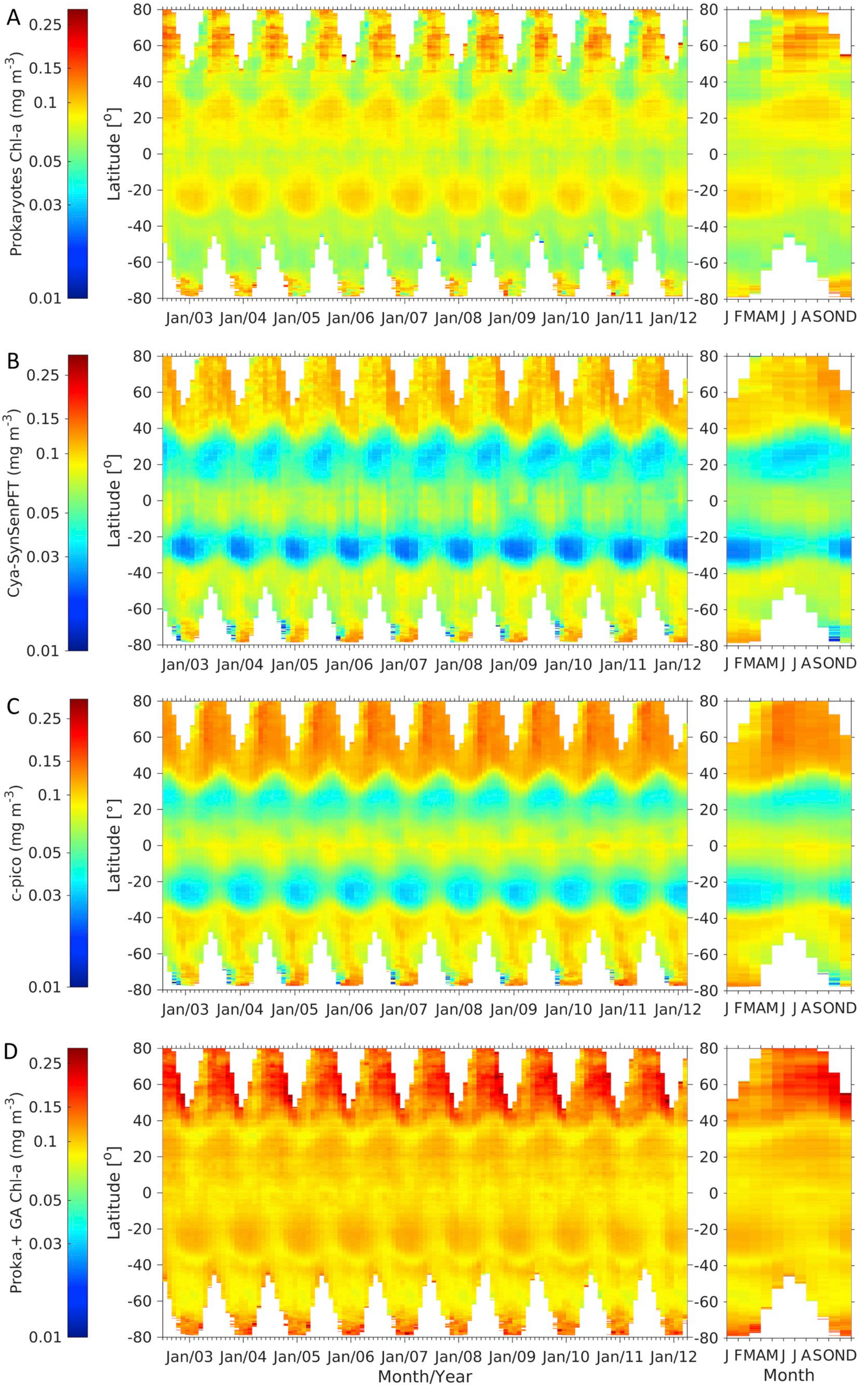

Fig. 10. Hovmöller diagrams of Chl-*a* concentrations of (A) prokaryotes derived from our study, (B) cya-SynSenPFT (Losa et al., 2017), (C) c-pico derived from PSC method (Brewin et al., 2015), and (D) sum of prokaryotes and green algae (Proka. + GA) from our study. Note that the color scale is different from Figs. 8–9.

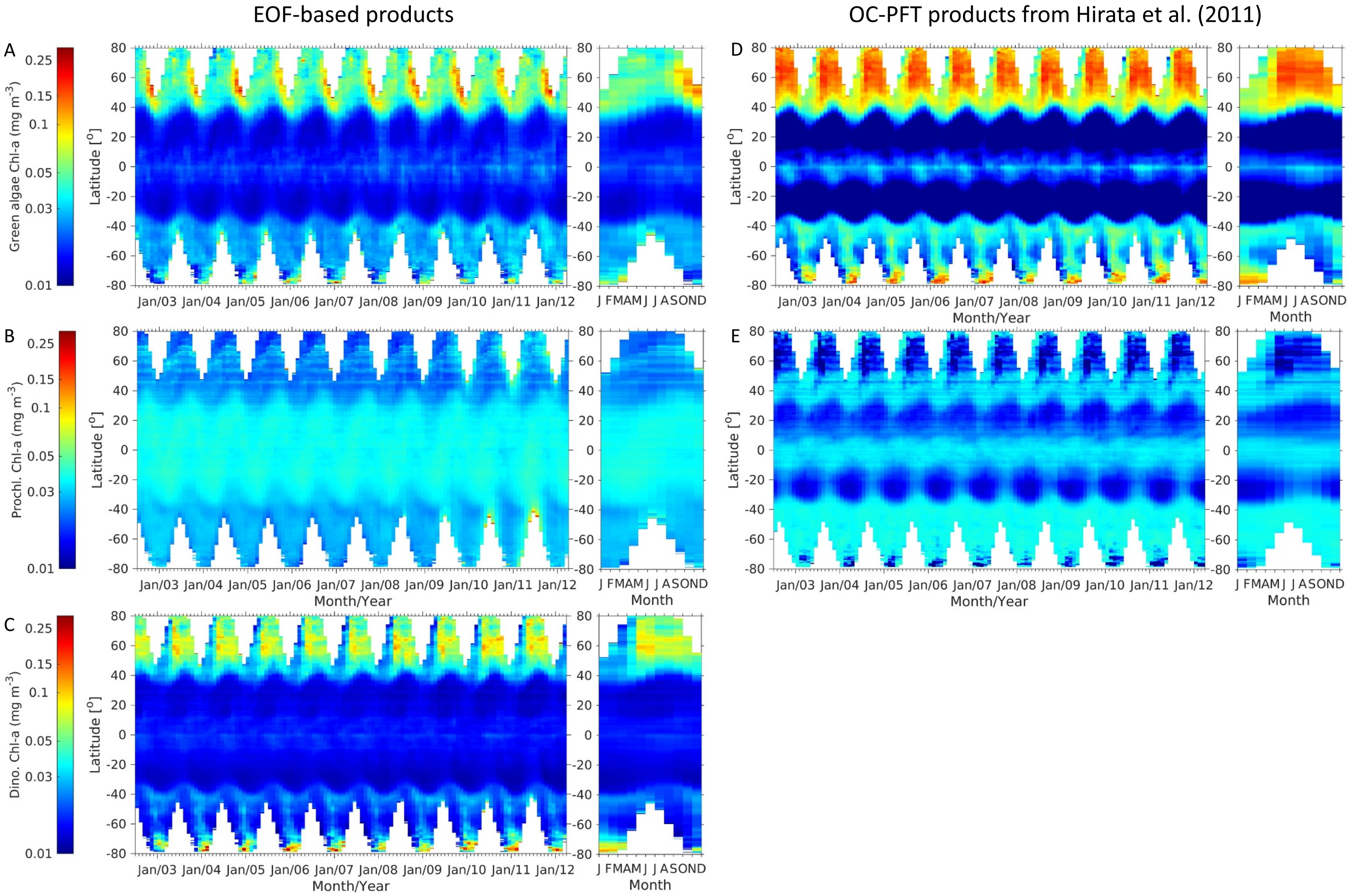

Fig. 11. Hovmöller diagrams of Chl-*a* concentrations of (A) green algae, (B) *Prochlorococcus*, and (C) dinoflagellates derived from our study, with the former two in comparison to (D) green algae and (E) *Prochlorococcus* from OC-PFT (Hirata et al., 2011). Note that the color scale is the same as Fig. 10.

2006 due to our limited access to the PHYSAT product. Diatoms, haptophytes, *Prochlorococcus* and SLC were included but the dominance of *Phaeocystis*-like group derived by PHYSAT was not available in our products. Distributions of dominant PFTs extracted from our products for four representative months (Fig. 12) generally present haptophytes and diatoms dominating in high latitudes and *Prochlorococcus* and SLC dominating in lower latitudes, which is well consistent with PHYSAT products. A more detailed comparison is described as follows.

In the high latitudes of the north hemisphere, our classified haptophytes dominance in January spreads a smaller range (30°N–50°N) than that from PHYSAT. In April, both products show haptophytes dominating above 30°N while diatoms are dominating in coastal areas only in our product. In July the two products show similar identification results with diatoms dominating in some parts of the North Sea, Norwegian Sea, Bering Sea and the Arctic waters, while haptophytes are still the major dominant PFT. Similar distribution is found for October as well despite of our product showing more diatoms in nearshore waters.

In the mid- to low latitudes, haptophytes and diatoms mainly dominate in coastal waters. Our product also shows dominance of haptophytes in the equatorial waters especially in the Pacific Ocean nearly for all seasons, which is barely presented in PHYSAT products. *Prochlorococcus* and SLC are two largely dominant groups in the mid- to low latitudes (Zubkhov et al., 1998). *Prochlorococcus* dominance is hardly found in 20°N–35°N from our product, whereas it is prominent in PHYSAT product especially in the north Pacific and north Atlantic gyres for all seasons. *Prochlorococcus* is found dominating in the low latitudes approximately between 15°S–15°N and SLC mainly dominates in the south Pacific gyre in both products. In the central to south Atlantic and Indian Ocean (equator to 40°S), our product shows SLC dominance in most of the regions in January and April, which decreases and is taken over by *Prochlorococcus* and haptophytes in July. However, PHYSAT products present dominance of both *Prochlorococcus* and SLC in this region in January and April, which is then gradually taken over by haptophytes in July with *Prochlorococcus* only dominating in the gyres. In the southern Pacific Ocean near 40°S both products show mainly *Prochlorococcus* dominating for nearly all seasons.

In the high latitudes of the south hemisphere (40°S–80°S), our product shows that *Prochlorococcus* and SLC spread more to the Southern Ocean especially from the south Pacific Ocean. In January, diatoms dominance of our product is found in Patagonian coastal waters and the south part of the Southern Ocean, while PHYSAT shows extensive diatoms dominance in 40°S–80°S with haptophytes and *Phaeocystis* also detected. For the other seasons, our product presents a smaller coverage of haptophytes dominance compared to PHYSAT products, and that diatoms are always dominant in Patagonian coastal waters and the west coast of South Africa.

Overall, besides some different distribution in diatoms dominance between PHYSAT and our products, the other main difference exists in the dominance of *Prochlorococcus* and SLC distributed in the central oceans, likely attributed to the low retrieval performance of prokaryotes. One should also keep in mind that PHYSAT product presents the climatology of 1997–2006 while ours for a more recent period (2002–2012) as explained in the beginning of this section. Recent long-term observations in the Arctic have shown a shift in phytoplankton composition from diatom-dominated to haptophyte-dominated and an

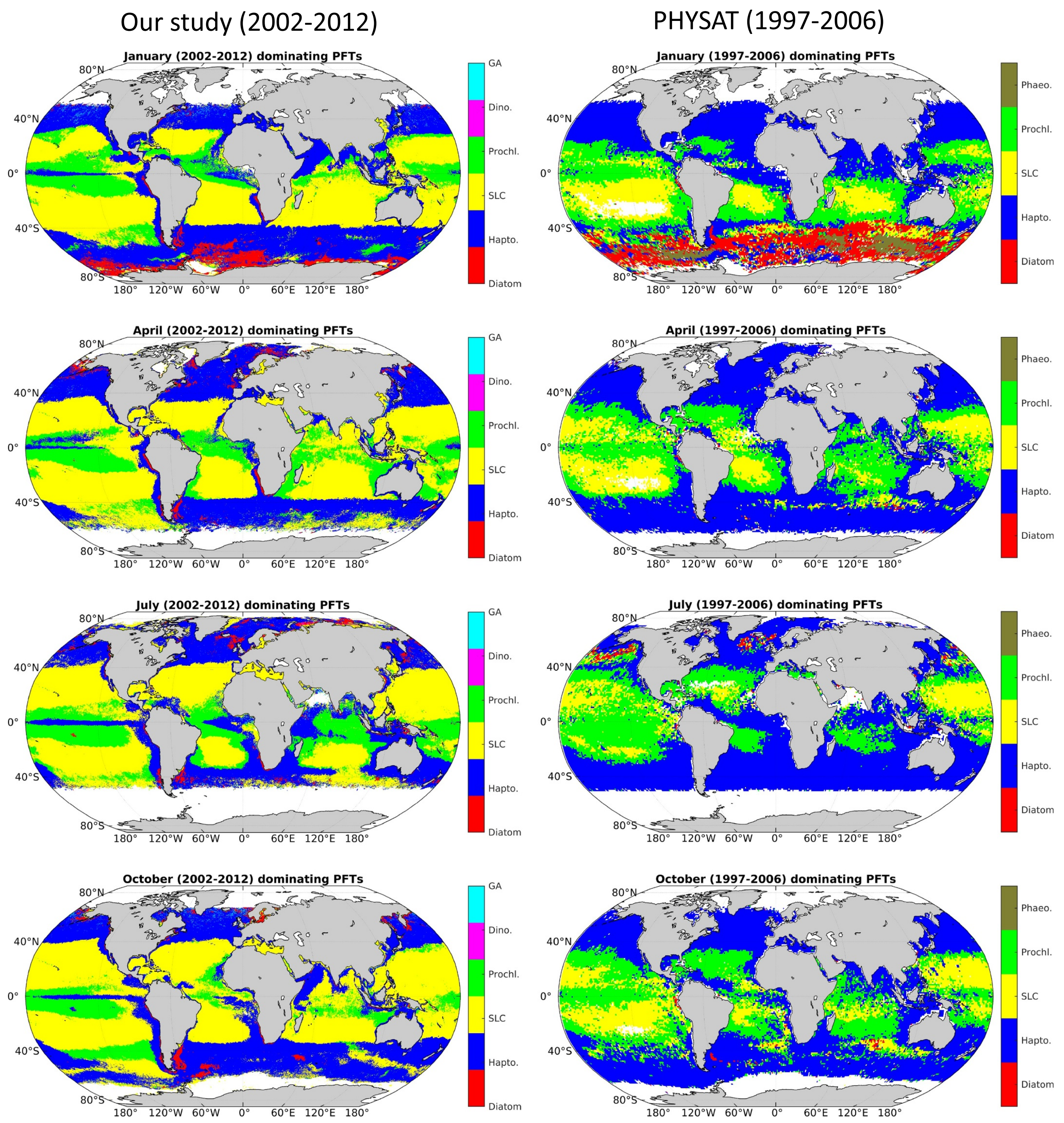
Our study (2002-2012)
PHYSAT (1997-2006)
January (2002-2012) dominating PFTs
January (1997-2006) dominating PFTs
April (2002-2012) dominating PFTs
April (1997-2006) dominating PFTs
July (2002-2012) dominating PFTs
July (1997-2006) dominating PFTs
October (2002-2012) dominating PFTs
October (1997-2006) dominating PFTs
80°N
40°N
0°
40°S
80°S
180°
120°W
60°W
0°
60°E
120°E
180°
GA
Dino.
Prochl.
SLC
Hapto.
Diatom
Phaeo.

Fig. 12. PFT Chl-*a* dominance extracted from the EOF retrieved PFTs (2002–2012, left panel) versus the results derived by PHYSAT (1997–2006, Alvain et al., 2008, right panel) for representative months. Note that the classified dominant PFTs are not all the same as the PHYSAT product, i.e., dinoflagellates and green algae dominance were included in our product while *Phaeocystis* dominance was included in PHYSAT. Blank areas indicate no available data.

enhancement of prokaryotic phytoplankton indicating that smaller phytoplankton species appear more often in the high latitudes (Nöthig et al., 2015), which is also presented in our product. In general, the overall high consistency between the two products confirms that realistic information on the PFT dominance can be extracted from the EOF-based PFT products.

3.4. Potential application to Sentinel-3A OLCI products

3.4.1. EOF-based PFT algorithm based on collocated OLCI R_{rs} and in situ HPLC pigments

Based on the R_{rs} matchups extracted specifically from the OLCI products (listed in Table 2), the EOF-based algorithm for OLCI application was built using the matchups of R_{rs} at 10 bands, 11 bands and 12 bands (denoted as OLCI R_{rs_10}, OLCI R_{rs_11}, and OLCI R_{rs_12}, respectively) at 1 × 1 pixel and corresponding DPA-derived PFTs. The matchups for 3 × 3 averaged data were not used due to low number of points. As shown in Fig. 3, the same procedure was applied to OLCI matchups regarding EOF analysis and regression model establishment. Similar to Sects. 3.1 and 3.2.1, EOF analysis was performed and the contribution to total variance of each important EOF mode was provided in Table S5 (supplementary document), showing EOF-1 takes >85% and the first four EOF modes contribute >99.9% of the total spectral variance. Table S6 in supplementary document presents the stepwise routine generated ΔAIC showing the importance of the EOF modes. Different from that based on merged matchups (Table 4), EOF-3 and EOF-2 for most PFTs both have high ΔAIC scores, indicating both are relatively important. Statistical results of the prediction performance provided in Table S7 (supplementary document) shows little differences between using different band numbers for the input R_{rs} data sets. For all PFTs the predictions are comparable to that gained using merged matchups in Sect. 3.2.2, however low number of matchups led to weaker cross validation statistics. As an example, Fig. S10 (supplementary document) shows the comparison between the predicted and observed PFTs using OLCI R_{rs_11} at 1 × 1 pixel. In general, good predictions are achieved with OLCI R_{rs_11} for diatoms, haptophytes, dinoflagellates, green algae and *Prochlorococcus*, especially the predictions of the latter two PFTs were obviously better than those from merged matchups. However, performances for TChl-*a* and diatom Chl-*a* prediction are a bit downgraded with OLCI data compared to merged matchups, possibly due to the low quality of corrected R_{rs} at blue bands for OLCI. This needs to be further investigated as it does not apply to other PFTs. As the retrieval approach is also an empirical method based on regressions, other factors such as the lower number of matchup points andvariation range of input data do also have impacts on the OLCI model performance. Prokaryotes prediction still has the least good performance. Good performance for *Prochlorococcus* estimation is achieved but the robustness could be weak due to little number of matchups (only 17–22 points for 1 × 1 pixel).

3.4.2. Test output of global PFTs retrieved from S3A OLCI products

The EOF fitted models based on OLCI R_{rs_11} at 1 × 1 pixel were selected and applied to the OLCI R_{rs} L3 monthly products with 25 km spatial resolution. Fig. 13 shows the mean distribution of each PFT Chl-*a* concentration derived from OLCI products during April 2016–December 2018. Compared to the PFTs derived from merged products, diatoms derived from OLCI are also well represented for coastal regions and show similar distribution in polar regions, but have higher Chl-*a* in the gyres and lower at the Equator. Haptophytes, dinoflagellates and green algae show nearly identical distributions with those from merged products, suggesting that the fitted models of these PFTs are well defined for both satellite products. Prokaryotes present elevated Chl-*a* concentration in the oligotrophic regions, however, opposed to that derived from the merged products, low abundance of prokaryotes is detected in coastal waters and high latitudes especially in the Arctic. *Prochlorococcus* shows much spatial variation with higher Chl-*a* concentration in the Arctic Ocean, north and central Atlantic Ocean, central Pacific Ocean, most coastal waters, and scattered regions in the Southern Ocean. This is apparently inconsistent with the consensus that *Prochlorococcus* is hardly detectable in most of these regions. This misinterpretation might be attributed to the ill-defined prediction model due to low number of valid matchups for *Prochlorococcus*.

It is noteworthy that the PFTs are retrieved for different periods between OLCI (2016–2018) and merged products (2002–2012), and that the matchups extracted from OLCI data are not adequate for a global coverage. Relatively weak performance for prokaryotes and *Prochlorococcus* retrieval lies in both applications of OLCI and merged products, suggesting again that improvement on their models is necessary to achieve more reliable retrievals.

4. Conclusion and outlook

An EOF-based global retrieval algorithm for quantifying multiple PFTs was developed using collocated satellite R_{rs} data and DPA derived PFT Chl-*a* concentrations from in situ pigment data. R_{rs} matchups with different band numbers extracted from the GlobColour SeaWiFS/MODID/MERIS merged products were used to assess and compare the performance of corresponding EOF fitted models in predicting PFTs. The models developed using R_{rs} data set with nine bands slightly outperformed those using the other data sets. The retrieval skills for six PFTs (diatoms, dinoflagellates, haptophytes, green algae, prokaryotes and *Prochlorococcus*) were investigated and cross-validated via a bootstrapping method. Satisfactory retrievals were achieved for diatoms, dinoflagellates, haptophytes and green algae, while the correlation generated by the EOF-based models for prokaryotic phytoplankton was relatively weak, resulting in less accurate retrievals for prokaryotes and *Prochlorococcus*. Global PFT retrievals over a ten-year period (2002–2012) were obtained based on the EOF-based models using merged R_{rs} L3 products at nine bands, showing plausible distributions for most of the investigated PFTs in the open ocean.

Evaluations on the EOF-based PFT products were carried out through inter-comparisons with SynSenPFT, PSC and OC-PFT products. Time-latitude Hovmöller diagrams covering monthly means of 2002–2012 showed generally good agreement between our EOF-based PFTs and other PFT/PSC products, despite that prokaryotic phytoplankton showed higher Chl-*a* concentrations in the subtropical gyres, which needs to be further validated. Dominance of PFTs derived from the EOF-based PFT products was also in high agreement with PHYSAT-products. Implementation of EOF models to OLCI products showed potential for a continuous observation, though differences for certain PFTs appeared comparing to the PFT products derived from merged products, likely due to lower number and limited coverage of matchups.

Different from abundance-based PFT algorithms, the proposed retrieval algorithm directly uses the reflectance data from satellites, thus can avoid uncertainty generated in the chlorophyll products, and links the variation of satellite reflectance spectra via PFT specific EOF based regression models. In addition, the retrieval algorithm is still an empirical approach, which is subject to the input data sets for training with regard to the number of observations, range of data variation and homogeneity of the data distribution in space and time. The apparent over−/underestimation feature in the regression models should also be further investigated. Nevertheless, this study showed the high potential of the EOF-based algorithm for quantitatively retrieving PFTs globally using satellite reflectance products from different sensors, which was not adequately reported in previous studies. Future efforts will be put in improving the current algorithm especially for prokaryotes prediction, such as applying proper scaling to the data sets and using non-linear fitting models. Further work, which is ongoing, is focusing on updating and extending the global in situ pigment data sets (especially from 2012 to present). The updated data sets will be used for EOF re-training with hopefully globally distributed matchups for OLCI (2016–present) and MODIS/VIIRS (2012–present) merged products, to fill the gap between MERIS and OLCI (2012–2016), and ultimately enable a continuous PFT observation from multi-sensor data. The updated in situ data sets will also be used for a thorough validation of the satellite retrieved PFTs.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2020.111704.

Author contributions

Conceptualization: H.X. and A.B.; *Methodology*: A.B. and H.X.; *Writing – Original Draft*: H.X., *Writing – Review & Editing*: H.X., A.B., M.S., and Y.L.; *Validation*: H.X., A.B., S.L., and M.S.; *Formal analysis*: H.X.; *Investigation*: H.X., A.B., S.L., A.M., and P.G.; *Resources*: A.M., P.G., J.D., O.H.A., A.B., Y.L., M.S., and S.L.; *Visualization*: H.X. and S.L.; *Supervision*: A.B., M.A., and O.H.A.; *Project administration*: A.B. and H.X.

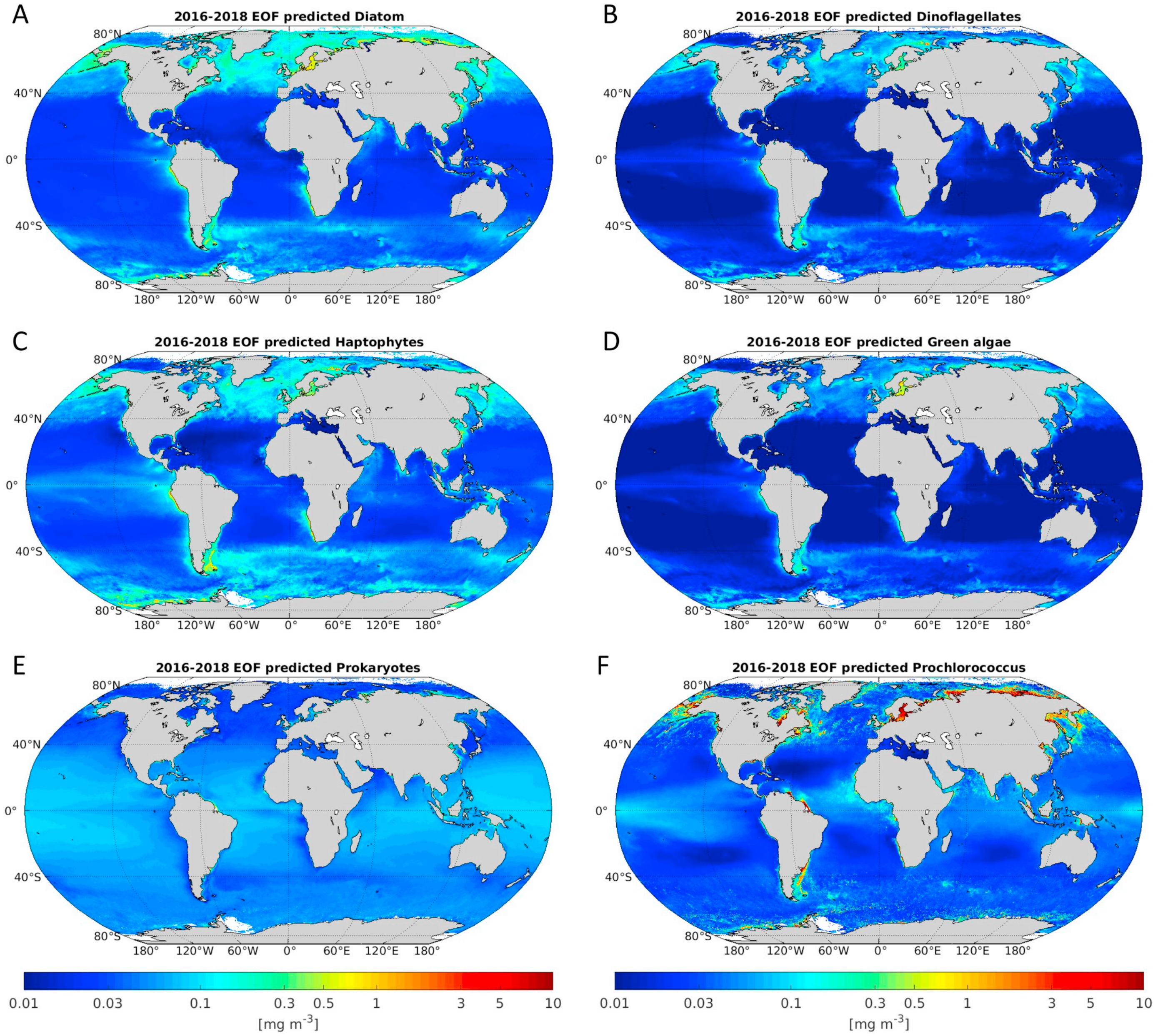
A
2016-2018 EOF predicted Diatom
B
2016-2018 EOF predicted Dinoflagellates
C
2016-2018 EOF predicted Haptophytes
D
2016-2018 EOF predicted Green algae
E
2016-2018 EOF predicted Prokaryotes
F
2016-2018 EOF predicted Prochlorococcus
80°N
40°N
0°
40°S
80°S
180° 120°W 60°W 0° 60°E 120°E 180°
0.01 0.03 0.1 0.3 0.5 1 3 5 10
[mg m-3]

Fig. 13. Yearly mean distribution (April 2016–December 2018) of the PFT Chl-*a* concentrations for (A) diatoms, (B) dinoflagellates, (C) haptophytes, (D) green algae, (E) prokaryotes, and (F) *Prochlorococcus* retrieved by the EOF-based algorithm from OLCI monthly R_{rs} products at 11 bands.

Data availability

All data used in this study were obtained via the links provided in Sect. 2.1 and in Losa et al. (2017). Numerical matrices and regression coefficients regarding the EOF-based algorithm are provided in the supplementary document.

Declaration of competing interest

None.

Acknowledgement

This work is supported by a collaborative project between ACRI-ST and Phytooptics team at Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, OLCI-PFT (ACRI-AWI Offer #209-180104). The contribution of Svetlana N. Losa is partly made in the framework of the state assignment of the Federal Agency for Scientific Organizations (FASO) Russia (theme 0149-2019-0015). We thank Marc Taylor for the R script on the EOF method, Sonja Wiegmann for her dedication in measuring and processing the HPLC data for our own cruises, and all the previous and current Phytooptics team members who participated in the past cruises for data collection and analysis. We are thankful to all the scientists and crew involved in the global HPLC data collection and analyses for providing their pigment data. We thank Robert Brewin for the PSC products and Séverine Alvain for making the PHYSAT data available to us. Thanks also to NASA, ESA and EUMETSAT for the SeaWiFS, MODIS, VIIRS, MERIS, and OLCI data, and specially the CMEMS GlobColour program for providing the OLCI and merged ocean color L3 products. Marine Bretagnon is also acknowledged for providing useful suggestions during the OLCI-PFT project discussions. We finally thank the three reviewers for providing valuable comments that substantially improved the manuscript.

References

ACRI-ST GlobColour Team, Mangin, A., Fanton d'Andon, O., 2017. GlobColour Product User Guide, GC-UM-ACR-PUG-01, Version 4.1. (Sophia-Antipolis).

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

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

Alvain, S., Moulin, C., Dandonneau, Y., Loisel, H., 2008. Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: a satellite view. Glob. Biogeochem. Cycles. https://doi.org/10.1029/2007GB003154.

Bailey, S.W., Werdell, P.J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2006.01.015.

Bracher, A., 2019. Phytoplankton pigment concentrations in the Southern Ocean during RV POLARSTERN cruise PS103 in Dec 2016 to Jan 2017. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEAhttps://doi.pangaea.de/10.1594/PANGAEA.898941.

Bracher, A., Wiegmann, S., 2019. Phytoplankton pigment concentrations in the North Sea and Sogne Fjord from 29 April to 7 May 2016 during RV HEINCKE cruise HE462. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEAhttps://doi.pangaea.de/10.1594/PANGAEA.899043.

Bracher, A., Vountas, M., Dinter, T., Burrows, J.P., Röttgers, R., Peeken, I., 2009. Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. Biogeosciences. https://doi.org/10.5194/bg-6-751-2009.

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

Bracher, A., Soppa, M., Losa, S., Dinter, T., Wolanin, A., Brewin, R., Bricaud, A., 2016. Final report. version 1.2, 30 Nov 2016, electronic version: SEOM-SynSenPFT-FR-D4.3_v1.2.pdf.

Bracher, A., Bouman, H.A., Brewin, R.J.W., Bricaud, A., Brotas, V., Ciotti, A.M., Clementson, L., Devred, E., Di Cicco, A., Dutkiewicz, S., Hardman-Mountford, N.J., Hickman, A.E., Hieronymi, M., Hirata, T., Losa, S.N., Mouw, C.B., Organelli, E., Raitsos, D.E., Uitz, J., Vogt, M., Wolanin, A., 2017. Obtaining phytoplankton diversity from ocean color: a scientific roadmap for future development. Front. Mar. Sci. 4, 1–15. https://doi.org/10.3389/fmars.2017.00055.

Bracher, Astrid, Wiegmann, Sonja, Xi, Hongyan, 2020. Phytoplankton pigment concentration and phytoplankton groups measured on water samples obtained during POLARSTERN cruise PS113 in the Atlantic Ocean. PANGAEA. https://doi.org/10.1594/PANGAEA.911061.

Brewin, R.J.W., Sathyendranath, S., Hirata, T., Lavender, S.J., Barciela, R.M., Hardman-Mountford, N.J., 2010. A three-component model of phytoplankton size class for the Atlantic Ocean. Ecol. Model. https://doi.org/10.1016/j.ecolmodel.2010.02.014.

Brewin, R.J.W., Sathyendranath, S., Jackson, T., Barlow, R., Brotas, V., Airs, R., Lamont, T., 2015. Influence of light in the mixed-layer on the parameters of a three-component model of phytoplankton size class. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2015.07.004.

Campbell, J.W., 1995. The lognormal distribution as a model for bio-optical variability in the sea. J. Geophys. Res. https://doi.org/10.1029/95JC00458.

Ciotti, A.M., Bricaud, A., 2006. Retrievals of a size parameter for phytoplankton and spectral light absorption by colored detrital matter from water-leaving radiances at SeaWiFS channels in a continental shelf region off Brazil. Limnol. Oceanogr. Methods 4, 237–253. https://doi.org/10.4319/lom.2006.4.237.

Correa-Ramirez, M., Morales, C.E., Letelier, R., Anabalon, V., Hormazabal, S., 2018. Improving the remote sensing of phytoplankton functional types (PFT) using empirical orthogonal functions: a case study in a coastal upwelling region. Remote Sens. 10 (4), 498. https://doi.org/10.3390/rs10040498.

Craig, S.E., Jones, C.T., Li, W.K.W., Lazin, G., Horne, E., Caverhill, C., Cullen, J.J., 2012. Deriving optical metrics of coastal phytoplankton biomass from ocean colour. Remote Sens. Environ. 119, 72–83. https://doi.org/10.1016/j.rse.2011.12.007.

Craig, S.E., Lohrenz, S.E., Lee, Z., Mahoney, K.L., Kirkpatrick, G.J., Schofield, O.M., Steward, R.G., 2006. Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, Karenia brevis. Appl. Opt. 45, 5414–5425. https://doi.org/10.1364/AO.45.005414.

Devred, E., Sathyendranath, S., Stuart, V., Maass, H., Ulloa, O., Platt, T., 2006. A two-component model of phytoplankton absorption in the open ocean: theory and applications. J. Geophys. Res. Ocean. https://doi.org/10.1029/2005JC002880.

Falkowski, P.G., Barber, R.T., Smetacek, V., 1998. Biogeochemical controls and feedbacks on ocean primary production. Science 281, 200–206. https://doi.org/10.1126/science.281.5374.200.

Falkowski, P.G., Laws, E.A., Barber, R.T., Murray, J.W., 2003. Phytoplankton and their role in primary, new, and export production. In: Fasham, M.J.R. (Ed.), Ocean Biogeochemistry: The Role of the Ocean Carbon Cycle in Global Change. Springer, pp. 99–121.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., Veneziano, D., Vera, C.S., Vrugt, J.A., Martiny, A.C., 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. Proc. Natl. Acad. Sci. 110, 9824–9829. https://doi.org/10.1073/pnas.1307701110.

Gregg, W.W., 2002. A coupled ocean-atmosphere radiative model for global ocean biogeochemical models. In: Suarez, M. (Ed.), NASA Global Modeling and Assimilation Series. 22. NASA Technical Memorandum 2002–104606, Greenbelt, MD, pp. 33.

Gregg, W.W., Casey, N.W., 2007. Modeling coccolithophores in the global oceans. Deep. Res. Part II Top. Stud. Oceanogr. https://doi.org/10.1016/j.dsr2.2006.12.007.

Gregg, W.W., Rousseaux, C.S., 2017. Simulating PACE Global Ocean Radiances. Front. Mar. Sci. 4, 1–19. https://doi.org/10.3389/fmars.2017.00060.

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., Straif, C., Fischer, S., Schrader, S., Storch, T., Heiden, U., Mueller, A., Bachmann, M., Mühle, H., Müller, R., Habermeyer, M., Ohndorf, A., Hill, J., Buddenbaum, H., Hostert, P., Van Der Linden, S., Leitão, P.J., Rabe, A., Doerffer, R., Krasemann, H., Xi, H., Mauser, W., Hank, T., Locherer, M., Rast, M., Staenz, K., Sang, B., 2015. The EnMAP spaceborne imaging spectroscopy mission for earth observation. Remote Sens. 7, 8830–8857. https://doi.org/10.3390/rs70708830.

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

Hirata, T., Hardman-Mountford, N.J., Brewin, R.J.W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., Yamanaka, Y., 2011. 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.

Hu, C., Cannizzaro, J., Carder, K.L., Muller-Karger, F.E., Hardy, R., 2010. Remote detection of Trichodesmium blooms in optically complex coastal waters: examples with MODIS full-spectral data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2010.04.011.

IOCCG, 2014. Phytoplankton functional types from space. In: Sathyendranath, S., Stuart, V. (Eds.), Reports of the International Ocean Color Coordinating Group No. 15. IOCCG, Dartmouth, NS.

IPCC, 2013. Climate change 2013: the physical science basis. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M.M.B., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK; New York, NY, USA, pp. 1535.

Kostadinov, T.S., Siegel, D.A., Maritorena, S., 2009. Retrieval of the particle size distribution from satellite ocean color observations. J. Geophys. Res. Ocean. https://doi.org/10.1029/2009JC005303.

Le Quéré, C., Harrison, S.P., Prentice, C.I., Buitenhuis, E.T., Aumonts, O., Bopp, L., et al., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob. Chang. Biol. 11, 2016–2040. https://doi.org/10.1111/j.1365-2486.2005.1004.x.

Lee, C.M., Cable, M.L., Hook, S.J., Green, R.O., Ustin, S.L., Mandl, D.J., Middleton, E.M., 2015. An introduction to the NASA Hyperspectral InfraRed Imager (HyspIRI) mission and preparatory activities. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2015.06.012.

Litchman, E., Klausmeier, C.A., Miller, J.R., Schofield, O.M., Falkowski, P.G., 2006. Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. Biogeosciences 3, 585–606. https://doi.org/10.5194/bg-3-585-2006.

Liu, Y., Boss, E., Chase, A., Xi, H., Zhang, X., Röttgers, R., Pan, Y., Bracher, A., 2019a. Phytoplankton Pigment Concentration Estimated From Underway AC-S Particulate Absorption Data During POLARSTERN Cruise PS99.2. PANGAEAhttps://doi.org/10.1594/PANGAEA.898102.

Liu, Y., Boss, E., Chase, A., Xi, H., Zhang, X., Röttgers, R., Pan, Y., Bracher, A., 2019b. Phytoplankton Pigment Concentration Estimated From Underway AC-S Particulate Absorption Data During POLARSTERN Cruise PS107. PANGAEAhttps://doi.org/10.1594/PANGAEA.898100.

Liu, Y., Hellmann, S., Wiegmann, S., Bracher, A., 2019c. Phytoplankton Pigment Concentrations Measured by HPLC During POLARSTERN Cruise PS99.1. PANGAEA. https://doi.org/10.1594/PANGAEA.905502.

Losa, S.N., Soppa, M.A., Dinter, T., Wolanin, A., Brewin, R.J.W., Bricaud, A., Oelker, J., Peeken, I., Gentili, B., Rozanov, V., Bracher, A., 2017. Synergistic exploitation of hyper- and multi-spectral precursor sentinel measurements to determine phytoplankton functional types (SynSenPFT). Front. Mar. Sci. 4, 1–22. https://doi.org/10.3389/fmars.2017.00203.

Lubac, B., Loisel, H., 2007. Variability and classification of remote sensing reflectance spectra in the eastern English Channel and southern North Sea. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2007.02.012.

Morel, A., Gentili, B., Claustre, H., Babin, M., Bricaud, A., Ras, J., Tieche, F., 2007. Optical properties of the "clearest" natural waters. Limnol. Oceanogr. 52 (1), 217–229.

Mouw, C.B., Hardman-Mountford, N.J., Alvain, S., Bracher, A., Brewin, R.J.W., Bricaud, A., Ciotti, A.M., Devred, E., Fujiwara, A., Hirata, T., Hirawake, T., Kostadinov, T.S., Roy, S., Uitz, J., 2017. A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the Global Ocean. Front. Mar. Sci. 4. https://doi.org/10.3389/fmars.2017.00041.

Nöthig, E.M., Bracher, A., Engel, A., Metfies, K., Niehoff, B., Peeken, I., Bauerfeind, E., Cherkasheva, A., Gäbler-Schwarz, S., Hardge, K., Kilias, E., Kraft, A., Kidane, Y.M., Lalande, C., Piontek, J., Thomisch, K., Wurst, M., 2015. Summertime plankton ecology in Fram Strait - a compilation of long-and short-term observations. Polar Res. https://doi.org/10.3402/polar.v34.23349.

Palacz, A.P., St. John, M.A., Brewin, R.J.W., Hirata, T., Gregg, W.W., 2013. Distribution of phytoplankton functional types in high-nitrate, low-chlorophyll waters in a new diagnostic ecological indicator model. Biogeosciences. https://doi.org/10.5194/bg-10-7553-2013.

Raitsos, D.E., Lavender, S.J., Maravelias, C.D., Haralabous, J., Richardson, A.J., Reid, P.C., 2008. Identifying four phytoplankton functional types from space: an ecological approach. Limnol. Oceanogr. https://doi.org/10.4319/lo.2008.53.2.0605.

Sieburth, J.M., Smetacek, V., Lenz, J., 1978. Pelagic ecosystem structure: heterotrophic compartments of the plankton and their relationship to plankton size fractions. Limnol. Oceanogr. 23, 1256–1263. https://doi.org/10.4319/lo.1978.23.6.1256.

Soja-Woźniak, M., Craig, S.E., Kratzer, S., Wojtasiewicz, B., Darecki, M., Jones, C.T., 2017. A novel statistical approach for ocean colour estimation of inherent optical properties and cyanobacteria abundance in optically complex waters. Remote Sens. 9, 1–22. https://doi.org/10.3390/rs9040343.

Soppa, M.A., Hirata, T., Silva, B., Dinter, T., Peeken, I., Wiegmann, S., Bracher, A., 2014. Global retrieval of diatom abundance based on phytoplankton pigments and satellite data. Remote Sens. 6, 10089–10106. https://doi.org/10.3390/rs61010089.

Soppa, M.A., Peeken, I., Bracher, A., 2017. Global Chlorophyll “a” Concentrations for Diatoms, Haptophytes and Prokaryotes Obtained with the Diagnostic Pigment Analysis of HPLC Data Compiled From Several Databases and Individual Cruises. PANGAEAhttps://doi.org/10.1594/PANGAEA.875879.

Taylor, B.B., Taylor, M.H., Dinter, T., Bracher, A., 2013. Estimation of relative phycoerythrin concentrations from hyperspectral underwater radiance measurements - a statistical approach. J. Geophys. Res. Ocean. 118, 2948–2960. https://doi.org/10.1002/jgrc.20201.

Uitz, J., Claustre, H., Morel, A., Hooker, S.B., 2006. Vertical distribution of phytoplankton communities in open ocean: an assessment based on surface chlorophyll. J. Geophys. Res. Ocean. 111. https://doi.org/10.1029/2005JC003207.

Vidussi, F., Claustre, H., Manca, B.B., Luchetta, A., Marty, J.-C., 2001. Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter. J. Geophys. Res. Ocean. https://doi.org/10.1029/1999JC000308.

Wang, G., Lee, Z., Mouw, C., 2017. Multi-spectral remote sensing of phytoplankton pigment absorption properties in cyanobacteria bloom waters: a regional example in the Western Basin of Lake Erie. Remote Sens. https://doi.org/10.3390/rs9121309.

Ward, B.A., 2015. Temperature-correlated changes in phytoplankton community structure are restricted to polar waters. PLoS One. https://doi.org/10.1371/journal.pone.0135581.

Werdell, P.J., Roesler, C.S., Goes, J.I., 2014. Discrimination of phytoplankton functional groups using an ocean reflectance inversion model. Appl. Opt. 53, 4833. https://doi.org/10.1364/ao.53.004833.

Zubkhov, M.V., Sleigh, M.A., Tarran, G.A., Burkhill, P.H., Leakey, R.J., Raymond, J.G., 1998. Picoplanktonic community structure on an Atlantic transect from 50°N to 50°S. Deep-Sea Res. I 45, 1339–1355. https://doi.org/10.1016/S0967-0637(98)00015-6.