Remote Sensing of Environment 294 (2023) 113596

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Marine big data-driven ensemble learning for estimating global phytoplankton group composition over two decades (1997–2020)

Yuan Zhang [a], Fang Shen [a,*], Xuerong Sun [b], Kun Tan [c]

[a] State Key Laboratory of Estuarine and Coastal Research, Center for Blue Carbon Science and Technology, East China Normal University, Shanghai, China
[b] Centre for Geography and Environmental Science, Department of Earth and Environmental Science, Faculty of Environment, Science and Economy, University of Exeter, Cornwall, United Kingdom
[c] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China

ARTICLE INFO

Edited by Dr. Menghua Wang

Keywords:
Phytoplankton group composition
HPLC pigments
Marine big data
Artificial intelligence
Ensemble learning

ABSTRACT

Accurate monitoring of the spatial-temporal distribution and variability of phytoplankton group (PG) composition is of vital importance in better understanding of marine ecosystem dynamics and biogeochemical cycles. While existing bio-optical algorithms provide valuable information, relying solely on satellite ocean color data remains insufficient to obtain high-precision retrieval of PG due to the intricate nature of the bio-optical signal and PG composition itself. An interdisciplinary approach combining advancements in machine learning with big data from ocean observations and simulations offers a promising avenue for more accurate quantification of PG composition. In this study, an ensemble learning approach, called the spatial-temporal-ecological ensemble (STEE) model, is developed to construct a robust prediction model for eight distinct phytoplankton groups (i.e., Diatoms, Dinoflagellates, Haptophytes, Pelagophytes, Cryptophytes, Green Algae, Prokaryotes, and Prochlorococcus). The proposed method introduces multiple data simultaneously: ocean color, physical oceanographic, biogeochemical, and spatial and temporal information. An ensemble strategy is applied to increase the performance of the model by merging three advanced machine-learning algorithms. The combined validation of multiple cross-validation (CV) strategies (i.e., standard, spatial block, and temporal block CVs) shows that the proposed STEE model has superior robustness and generalization ability. In addition, the analysis shows a high degree of concordance between the independent datasets and the modeled estimations for long-time series sites, indicating that the STEE model is capable of effectively monitoring long-term trends in phytoplankton group composition. Finally, the proposed model was utilized to retrieve global monthly phytoplankton group products (STEE-PG) over an extended period (September 1997 to May 2020), and comparisons demonstrated better rationality of spatio-temporal distribution than existing satellite-derived phytoplankton group products. Hence, this new model comprehensively integrates all kinds of observation data and yields long-term global PG products with high accuracy, which will enhance our understanding of the response of marine ecosystems to environmental and climate change.

1. Introduction

With escalating concerns about global climate change and carbon neutrality, phytoplankton, responsible for nearly half of the global net primary production (Field et al., 1998), has become prominent in earth system science. Through the photosynthesis-driven biological pump, phytoplankton plays a crucial role in regulating the carbon dioxide concentration in the atmosphere and the global carbon cycle (Gruber et al., 2019; Nair et al., 2008). Based on satellite products, clusters of any species and ecotype are usually referred to as phytoplankton groups (PG). Different PG have distinct biological functions, and fluctuations in their biomass and spatial distribution can directly or indirectly reflect trends in the marine environment and climate change (Poloczanska et al., 2013). Therefore, obtaining precise maps of PG distribution is essential for monitoring the health of aquatic ecosystems and investigating the effects of climate change on marine ecosystems (Bracher et al., 2017).

Given this importance, great efforts have been made to retrieve PGs. Abundance and spectral-based algorithms constitute the majority of available methods (Mouw et al., 2017; Xi et al., 2020), with the common

* Corresponding author.
E-mail address: fshen@sklec.ecnu.edu.cn (F. Shen).

https://doi.org/10.1016/j.rse.2023.113596
Received 15 July 2022; Received in revised form 13 April 2023; Accepted 17 April 2023
Available online 12 May 2023

denominator being ocean-color data. The abundance-based approach is univariate, i.e., it uses chlorophyll *a* (Chla) as the input data. This indirect approach relies on the observed pattern of variation in phytoplankton population structure with abundance, obtaining PG based on an empirical relationship linking in-situ diagnostic pigments to Chla (Hirata et al., 2011). This method is computationally simple and can be easily applied to Chla products from different sensors. Spectral-based approaches use the optical signatures of phytoplankton groups directly for their detection from space, relying on the spectral features of reflectance, absorbance, or backscatter spectra because of changes in phytoplankton composition. Powered by multivariate regression analysis models, spectral-based techniques map the spectral features of reflectance or absorbance measured by satellites to the PG (Sun et al., 2022; Werdell et al., 2014). Spectral data processing (or feature extraction) methods, such as derivative analysis (Alvain et al., 2005), differential spectra (Bracher et al., 2009; Losa et al., 2017; Sadeghi et al., 2012) and principal component analysis (Bracher et al., 2015), have also been introduced to extract the potentially relevant spectral features of target PGs. It has been demonstrated that abundance- and spectral-based algorithms can provide good predictability at lower levels of noise interference.

Although most retrieval algorithms perform well, ocean-color-based retrieval of PG remains limited. Abundance-based algorithms may fail when different phytoplankton types have similar chlorophyll levels and are less suitable for analysis at large temporal and spatial scales (Bracher et al., 2017). The following problems also challenge the spectral-based approach: (i) because of the overlap in pigment composition among PGs, the spectral fingerprints of some groups are too similar to discriminate (Sathyendranath et al., 2014). (ii) Except in highly turbid waters, the water signal comprises only a small fraction of photons that can reach a spaceborne sensor. Further attempts to decompose optical signals into correlated signals of various PGs are easily encountered by intrinsic constraints and interference from poor information content and noise signals. (iii) The variation in inherent optical properties generates ambiguity in different study areas and may introduce additional mistakes in categorization, particularly in Case-2 waters. High CDOM concentrations or enhanced reflectance because of benthic resuspension can also disrupt the optical algorithm, resulting in false positives (Nair et al., 2008).

Obtaining accurate information on the distribution and composition of PGs from satellite spectral data alone is challenging due to the complexity and resolution of the bio-optical signal and the complexity of the PGs composition itself. Hyperspectral remote sensing technology enhances practical information in the spectral dimension and has been used to improve the accuracy of phytoplankton species and groups retrieval (Dierssen et al., 2021; Oelker, 2021). The PhytoDOAS (Bracher et al., 2009; Sadeghi et al., 2012) method uses hyperspectral satellite data from the atmospheric sensor scanning imaging absorption spectrometer for atmospheric cartography (SCIAMACHY) and can quantitatively retrieve major PGs based on optical features. An alternative way to improve the accuracy of PG models is the ecologically based approach (Raitsos et al., 2008), which considers more environmental parameters in model development to gain information. This strategy is based on the rationale that phytoplankton growth is influenced by multiple environmental factors (Moore et al., 2013). Specifically, in long-term changes, such as global climate change, the marine environment can influence PG composition by altering water column stratification and macronutrient (i.e., nitrogen, phosphorus, and silicate) availability through ocean warming, acidification, changes in ocean circulation systems, and sea level rise (Henson et al., 2021; Holder and Gnanadesikan, 2021b; Longhurst et al., 1995). In particular, sea surface temperature (SST) influences PG composition (Sun et al., 2019; Ward, 2015) either directly (e.g., metabolism, Lopez-Urrutia and Moran (2015)) or indirectly (e.g., nutrients, Maranon et al. (2012)). Therefore, incorporating ecological variables or geographic knowledge with ocean-color data is expected to improve the performance of PG retrieval models. Encouragingly, constant marine data collection via various monitoring or simulation methods has resulted in a tremendous increase in data volume, and the era of big marine data is approaching (Guidi et al., 2020; Xi et al., 2021). Numerous disciplines and fields, such as biological oceanography, chemical oceanography, physical oceanography, and meteorology, have accumulated massive structured and unstructured datasets from various sensors, platforms, and even model simulation outputs, with immense application potential (Huang et al., 2015). In addition, considering that phytoplankton are simultaneously subjected to multiple stresses, artificial intelligence and machine learning have been introduced to capture their multivariate and non-linear relationships and to find patterns in complex ecological contexts (Zhou, 2020), which are not achievable using conventional methods. Several supervised learning methods, including neural networks (Flombaum et al., 2013), random forest regression (Stock and Subramaniam, 2020), and boosted regression trees (Busseni et al., 2020), have been successfully implemented.

Technological advances in machine learning and data processing methods, as well as improvements in the availability of large-scale ocean observational and modal data, offer new opportunities for the large-scale, long-term remote sensing of phytoplankton diversity. Raitsos et al. (2008) developed a holistic approach to discriminate different PGs using a probabilistic neural network that combines ecological knowledge with ocean-color parameters. Palacz et al. (2013) used artificial neural networks to estimate four plankton groups from satellite SST, wind speed, Chla, and mixed layer depth. Although the initial results of these approaches are encouraging, challenges remain. First, further research is required to demonstrate the applicability of ecological methods on a global scale. Second, there are still research gaps in generating highly accurate long-time-series PG products with the help of ecological methods. In addition, as the dimensionality of the data increases, the relationships between predictor variables and target parameters exhibit a high degree of nonlinearity, and these relationships may have potential cross-dimensional dependencies, such as temporal, spatial, and spectral dependencies. A single model may not be able to accurately construct complex mapping relationships between the environmental variables and PGs. This emphasizes the critical need to develop and construct higher precision and stability models.

Addressing these challenges will require developing new methods to effectively integrate existing marine environmental data into more advanced machine-learning frameworks. In this study, we sought to improve the accuracy and robustness of phytoplankton community retrieval by using advanced ensemble learning architectures supported by multi-source marine big data to achieve global-scale long-time-series PG mapping. In addition to ocean-color data, various oceanic environmental factors, such as chemical oceanography (e.g., nutrients) and physical oceanography (e.g., salinity), associated with PG distribution have been introduced. More geographical and temporal features were incorporated into the modeling procedure to identify the ecological niches where particular PGs may be found. The BorutaShap feature selection framework was adopted to determine the sensitive variables and minimize the probability of model overfitting. Three high-performance machine-learning algorithms were combined using the ridge regression ensemble to improve the data-mining capabilities, and thus, the accuracy of the estimations. By applying the proposed model to marine data, we reconstructed global monthly PG products over two decades (September 1997 to May 2020) and made comparisons with previous products.

The manuscript is organized as follows. Section 2 introduces the key information of the in-situ pigment dataset and the predictor data utilized in the study. The development of a spatial-temporal-ecological ensemble (STEE) model and a description of the analytical evaluation algorithms are also proposed in detail. Section 3 presents model validation and comparison. Section 4 discusses the potential and limitations of the proposed approach. Finally, Section 5 presents a summary.

2. Materials and methodology

2.1. Principle

The response of phytoplankton to environmental drivers is complicated and depends on a host of variables. Macronutrients, such as nitrogen, phosphorus, and silicate, together with physical factors, including mixed-layer depth, temperature, and wind stress, can significantly affect the spatial distribution and growth rate of phytoplankton (Holder and Gnanadesikan, 2021b). Therefore, we construct a multi-source data-driven PG retrieval model. Specifically, the spatial distribution of PG is modeled as a nonlinear mapping f_x of multiple environmental predictors (i.e., Bio-optical, Biogeochemistry, Physical, Meteorological, and distributed spatially and temporally), expressed as follows:

$$PG = f_x(\text{input predictors}) = f_x\left(p_{\text{Bio-optical}}, p_{\text{Biogeochemistry}}, p_{\text{Physical}}, p_{\text{Meteorological}}, p_{\text{Spatio-temporal}}\right) \quad (1)$$

where p represents the corresponding predictor variable. In this study, we construct f_x using ensemble machine learning. An overview of the proposed approach is shown in Fig. 1.

2.2. Datasets

2.2.1. In-situ HPLC pigment dataset

In this study, we utilized a compiled HPLC (High Performance Liquid Chromatography) pigment dataset collected from the global ocean between 1997 and 2020, which includes published datasets from various regions of the global ocean for model parameterization and long-time series datasets for independent validation (refer to Fig. 2 for details).

For global ocean datasets, Kramer and Siegel (2021) compiled a large dataset of in-situ HPLC phytoplankton pigment samples from 2000 to 2018. On this basis, we have added more open-source data to improve the spatial and temporal coverage of the global ocean dataset, including data from the PANGAEA Data Center (https://www.pangaea.de/), NASA SeaBASS archive (https://seabass.gsfc.nasa.gov/archive/), and Australian Ocean Data Network (http://portal.aodn.org.au/). More detailed information is provided in Supplementary Table S1.

As one of the largest shelf-edge seas in the world, the eastern China seas, including the Bohai Sea, Yellow Sea, East China Sea, and Changjiang River Estuary and its adjacent waters, have been hot spots for phytoplankton research. In this study, 405 samples from the surface ocean (0–3 m) collected from seven cruise campaigns between 2015 and 2020 in eastern China seas were utilized to enrich the existing global dataset (Fig. 2b). The procedure for sample collection and HPLC analysis has been described by Sun et al. (2022).

Long-term time-series data from static sites were independently utilized for model validation. We obtained independent time series of HPLC pigment datasets from six continuous observation sites, including Martha's Vineyard Coastal Observatory (MVCO), CArbon Retention In A Colored Ocean (CARIACO), the Plumes and Blooms program (PAB) in the Santa Barbara Channel, and three national reference stations in Australian coastal waters, Yongala, Port-Hacking (PH), and Maria Island (MI). More detailed information on these long-term series datasets is presented in Supplementary Table S2.

To control the quality of the pigment data, all in-situ samples were subjected to quality assurance procedures: (i) samples collected within the top water column (<10 m) were retained, (ii) samples with diagnostic pigment concentrations below 0.001 mg·m3 were rejected, and (iii) observations before 1997 were excluded. For duplicate samples collected or published synthetically, we calculated the average of the duplicate samples rather than making detailed distinctions.

2.2.2. Satellite imagery, reanalysis data, and spatio-temporal information

Multiple environmental factors data were collected as input predictors (Table 1). First, the merged SeaWiFS, MERIS, MODIS-Aqua, and VIIRS data of the Ocean-Color Climate Change Initiative (OC-CCI, version 5.0) from the European Space Agency were downloaded, with a spatial resolution of 4 km (Sathyendranath et al., 2019). The data variables used in this study included spectral remote sensing reflectance (R_{rs}), Chlorophyll-a concentration (Chla), particulate backscattering coefficient (b_{bp}), phytoplankton absorption coefficient (a_{ph}), diffuse attenuation coefficient at 490 nm (K_d490), and water class (water class memberships of each pixel to 14 optical water classes, Jackson et al. (2017)). Photosynthetically Available Radiation (PAR) data were obtained from the Copernicus Marine Environment Monitoring Service (CMEMS) GlobColour data archive (Team et al., 2017).

This study utilized several reanalysis products as ancillary predictors, including biogeochemical hindcasts, physical reanalysis, and meteorological data. Specifically, we obtained NC (nitrate concentration), PC (phosphate concentration), SC (silicate concentration), and DO (dissolved oxygen) for biogeochemical variables from the global biogeochemical multi-year hindcast products (identified as GLOBAL_MULTIYEAR_BGC_001_029) of the Copernicus Marine Environment Monitoring Service (CMEMS; http://marine.copernicus.eu/). For physical variables, we used the sea surface temperature (SST) products produced by the ESA SST CCI project (Merchant et al., 2019), and the SSS (sea surface salinity), SSH (sea surface height), UML (upper mixed layer depth), and OCV (ocean current velocity) data from the CMEMS Global Ocean Physics Reanalysis products (identified as GLOBAL_MULTIYEAR_PHY_001_030). Additionally, we obtained meteorological data from the CMEMS, including sea surface wind speed, the west-to-east component, and the south-to-north component of the wind vector (identified as WIND_GLO_WIND_L4_REP_OBSERVATIONS_012_006).

Spatial and temporal environmental heterogeneity, that is, spatial autocorrelation and patchiness, as well as temporal variation, are among the essential characteristics of phytoplankton. To enhance the description of phytoplankton variation, we incorporated both temporal and spatial information into the modeling process. The spatial properties can be described by latitude, longitude, and haversine distance to the coast. The time terms used in the model include (i) Year (without transformation), (ii) C_{mon} (i.e., month, using cosine transformation) and (iii) N_{mon}, which is the number of whole months from September 1997 to the target time.

2.2.3. Data processing and match-ups selection

The ocean-color imagery and reanalysis products described above have a variety of gridding, time scales, and spatial resolutions. In this study, other raster data were resampled to adjust the spatial and temporal resolutions using ocean-color data as a benchmark. Values for the no-data regions were calculated using inverse distance weighting based on surrounding pixel values. After interpolation, smoothing iterations were performed using 3 × 3 average filters on interpolated pixels to eliminate artifacts. The median imputation method was used to fill gaps after interpolation. Finally, we normalized every piece of the modeled data. Specifically, we subtract the minimum value from each entry and divide the result by the range, where the range is the difference between the maximum and minimum values. This process is implemented with the "preprocessing.normalize" function in Scikit-Learn (Pedregosa et al., 2011), a machine learning library based on the python language. In addition, for non-normally distributed products (i.e., Chla, MLT, and K_d490), we performed a logarithmic transformation at a base of 10. The above operations were executed using open-source geospatial data abstraction library (GDAL) libraries. (van der Walt et al., 2011);

We matched the in-situ pigment data with the variables from the corresponding products in time (with a 1-day window) and space (in 3 × 3-pixel boxes with the closest latitude and longitude). The results are shown in Fig. 3. For the matched in-situ data, we performed the Diagnostic Pigment Analysis (DPA) program to determine the Chla concentration of PG. The DPA is a rapid and accurate method for determining phytoplankton abundance in marine environments(Uitz et al., 2006;

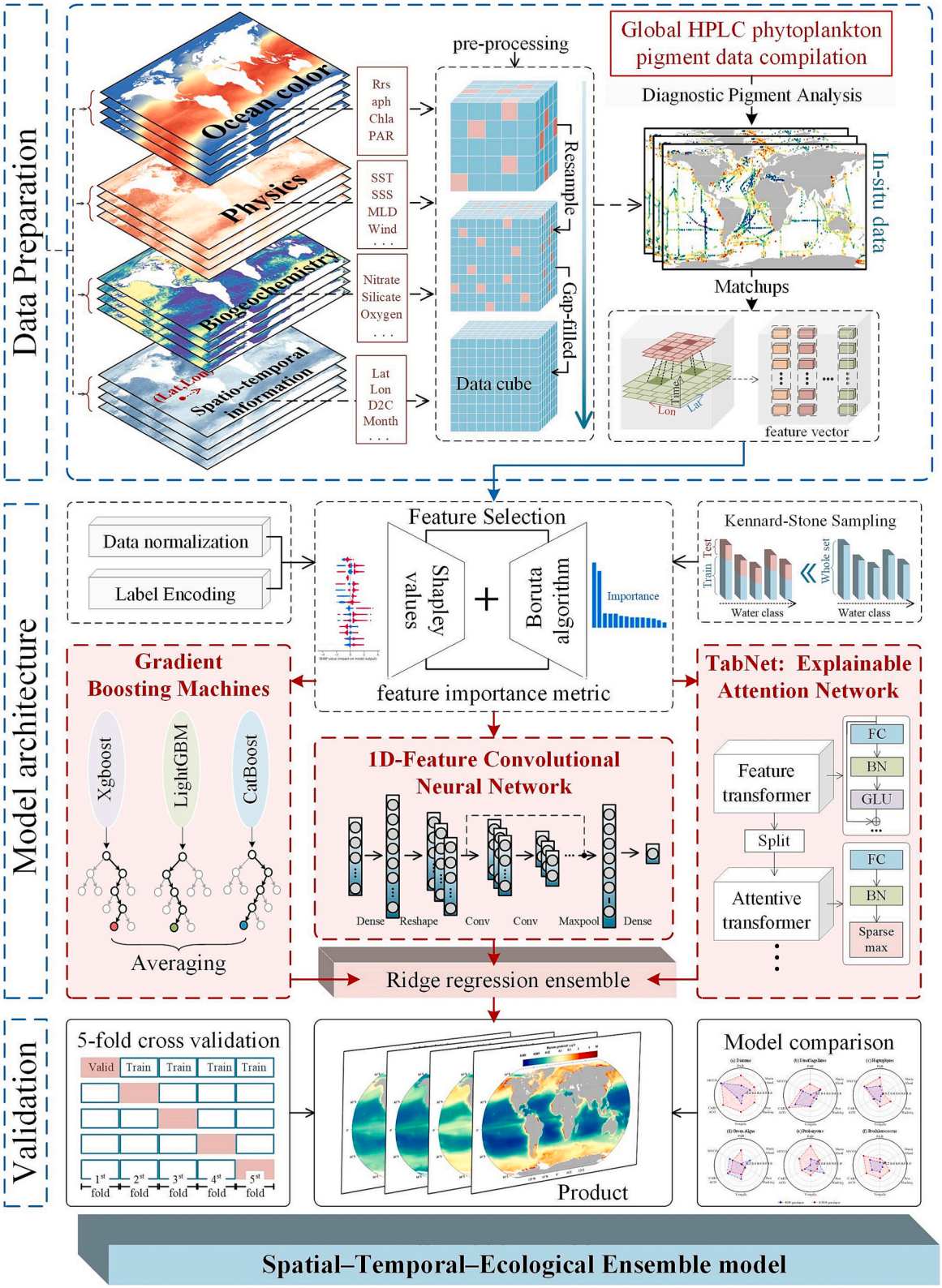

Fig. 1. Schematic flow of the methodological approach in this study. A global field HPLC pigment dataset, including samples from all the major marine regions, is compiled. Simultaneous datasets on physical oceanography, chemical oceanography, meteorology, and bio-optical are also collected as input variables for the regression model. Sensitive variables are selected based on the BorutaShap method to provide a basis for further analysis and modeling. Subsequently, a multi-model ensemble learning approach, named as the Spatial–Temporal–Ecological Ensemble model (STEE) model, is implemented based on three machine learning techniques to deal with complex supervised regression problems in multi-source data, resulting in the construction of a robust PG estimation model. Finally, global monthly estimation products are generated for eight phytoplankton groups (i.e., Diatoms, Dinoflagellates, Haptophytes, Pelagophytes, Cryptophytes, Green Algae, Prokaryotes, and Prochlorococcus) from September 1997 to May 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

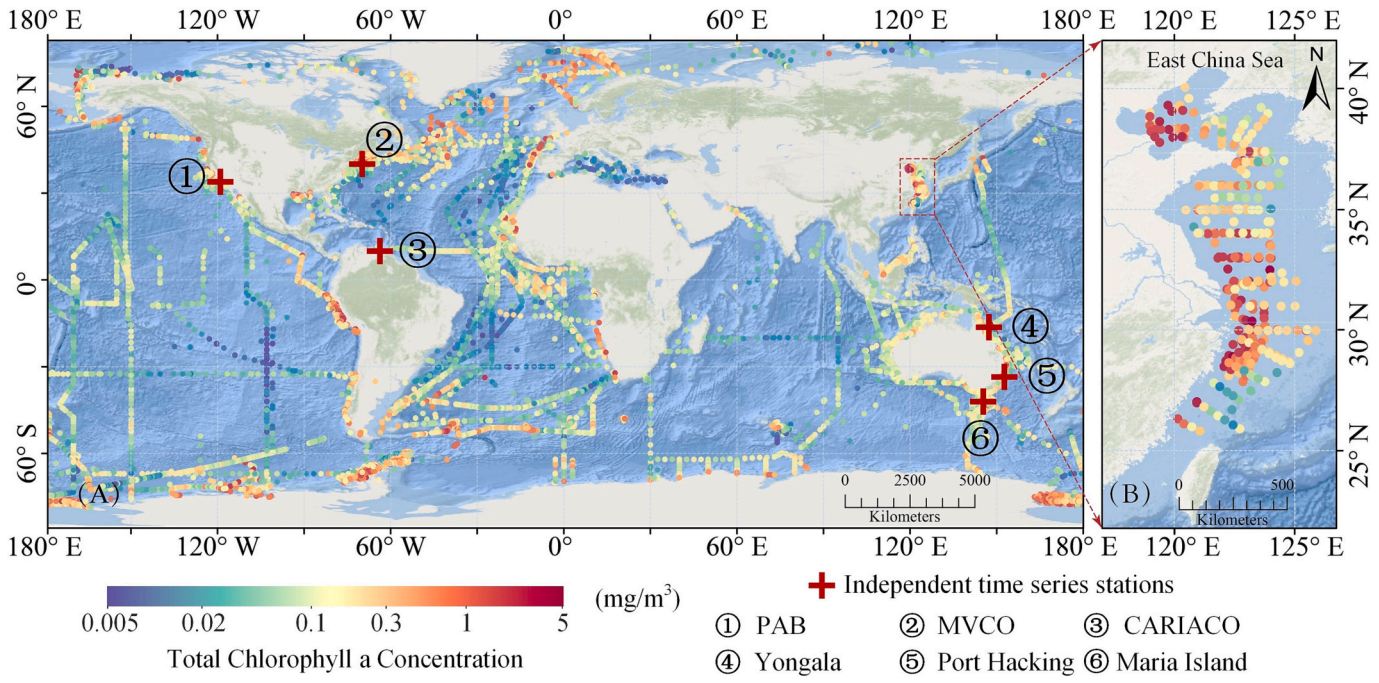

Fig. 2. Spatial distribution of in-situ HPLC pigment datasets in (a) global surface ocean, and (b) eastern China seas. The crosses and numbers in (a) marks represent the location of the six independent long-tern time series stations.

Table 1
Predictors and corresponding data products.

Dataset	Abbreviation	Definition	Resolution	Reference/DOI
Bio-optical data	$R_{412–670}$	Remote sensing reflectance at 412, 443, 490, 510,555 and 670 nm	4 km, Daily, 1997.9–2020.5	Sathyendranath et al. (2019)
	$a_{412–670}$	QAA absorption due to phytoplankton at 412, 443, 490, 510,555 and 670 nm		
	$b_{412–670}$	QAA backscatter due to particulate matter at 412, 443, 490, 510,555 and 670 nm		
	K_d490	diffuse attenuation coefficient at 490 nm		
	Chla	Chlorophyll-a concentration		
	WC	memberships of each pixel to 14 optical water classes		
Biogeochemistry data	PAR	Photosynthetically Available Radiation	1/4°, Daily, 1997.9–2020.5	Team et al. (2017)
	NC	Nitrate concentration		https://doi.org/10.48670/moi-00019
	PC	Phosphate concentration		
	SC	Silicate concentration		
	DO	Dissolved oxygen		
Ocean Physical data	SST	sea surface temperature	1/20°, Daily, 1997.9–2020.5	Merchant et al. (2019)
	SSS	sea surface salinity	1/12°, Daily, 1997.9–2020.5	https://doi.org/10.48670/moi-00021
	UML	Upper Mixed Layer depth		
	EOV	Eastward ocean current velocity		
	NOV	Northward ocean current velocity		
Meteorological data	WS	sea surface wind speed	1/4°, Daily, 1997.9–2020.5	https://doi.org/10.48670/moi-00185
	EW	West to East component of wind-to vector		
	NW	North component of the wind-to vector		
Spatio-temporal information	sLat	Sine of latitude	–	–
	sLon	Sine of longitude		
	D2C	Haversine distance to coast		
	Year	Year		
	Cmon	months of the year, converted using cosine		
	Nmon	Number of months since September 1997		

Vidussi et al., 2001). The DPA weights utilized in this study were referenced from Losa et al. (2017), obtained from the global ocean using multiple regression analysis. The concentrations of eight pigments, i.e., fucoxanthin (Fuco), peridinin (Peri), 19′-hexanoyloxyfucoxanthin (Hex), 19′-butanoyloxyfucoxanthin (But), alloxanthin (Allo), chlorophyll *b* (Chlb), zeaxanthin (Zea) and divinyl chlorophyll *a* (DVChla) were used to determine the Chla concentrations of eight PG (Diatoms, Dinoflagellates, Haptophytes, Pelagophytes, Cryptophytes, Green Algae, Prokaryotes, and Prochlorococcus). Fig. S1 in the Supplementary material presents the histograms of the Chla concentrations of the eight PG at log-10 scale, along with the statistics.

2.3. Model development

In this study, a multimodel ensemble learning approach, called the STEE, is developed to deal with complex supervised regression issues in

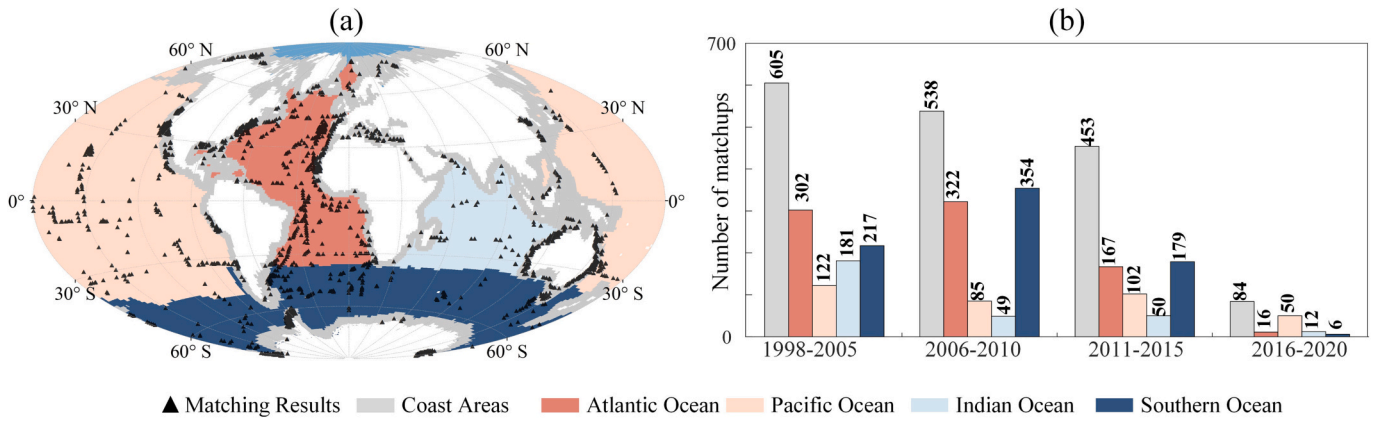

Fig. 3. Geographical location (a) and the number of matchups in different regions (b) between in-situ pigments data and ecological factors data. The boundaries of the ocean basin and coastal regions are provided by the RECCAP2 project (see https://reccap2-ocean.github.io/regions/).

multi-source data and to construct a robust PG prediction model. We designed an ensemble learning framework that wraps the PG retrieval model between the inputs, which are environmental variables, and the outputs that represent different PGs, as shown in Fig. S2.

First, the input variables were selected based on the BorutaShap algorithm to increase the prediction accuracy. To enhance the model diversity, the STEE model is implemented by integrating three powerful machine-learning regression techniques: Gradient Boosting Machine (GBM), One-dimensional Convolutional Neural Network (1d-CNN), and Attentive Interpretable Tabular Learning neural network (TabNet). Each sub-model obtains the input data and generates an independent prediction output. Finally, ridge regression was introduced as a combinatorial approach to form ensemble predictions that maximize robustness while minimizing the possibility of overfitting.

2.3.1. Sensitive variables selection

Feature selection is a vital step in machine learning model development that involves identifying the most significant subset of input variables for prediction or decision-making. By doing so, it can boost the model's performance, interpretability, and efficiency while reducing the risk of overfitting. In this study, the term "feature" pertains to the environmental variables used as inputs for the STEE model. The applicability of the proposed STEE model as a data-driven algorithm is heavily dependent on the selection of input parameters. Therefore, screening important variables instead of all variables is required to improve model efficiency and reduce overfitting. We used the BorutaShap feature selection method (Keany, 2020) to select sensitive variables. BorutaShap combines the Boruta feature selection algorithm (Kursa and Rudnicki, 2010) with the SHAP (SHapley Additive exPlanations, Lundberg and Lee (2017)) technique, aiming to find a minimal optimal feature set rather than all the features relevant to the target variable. In the BorutaShap algorithm, the z-score value is used to characterize the importance of the variables; the calculation process is detailed in (Keany, 2020). The BorutaShap algorithm is independent of the dataset size because it uses the tree structure to compute a global feature ranking, making it much faster than SHAP when dealing with larger datasets.

The random forest model (Breiman, 2001) served as the foundation for the BorutaShap algorithm in this study. Features with an average z-core value in the top 80% were selected as sensitive variables after running 1000 iterations of the BorutaShap algorithm, and the rest are discarded. Note that we performed feature selection for each cross-validation iteration. After completing the cross-validation evaluation, we retrained the proposed STEE model with all the data (a combination of training and testing datasets) for the final deployment and generated global PG products. Fig. S3 shows the feature selection results of the final deployed model. Given that the main objective of this study was to construct a predictive model, the response mechanisms between each ecological variable and phytoplankton distribution were not investigated in detail.

2.3.2. Individual base model

Each machine learning model has distinct perceptual abilities and is designed to extract distinct data features. It has been demonstrated that multi-model averaging of ensemble members can produce more accurate and reliable forecasts than a single model (Pena and van den Dool, 2008). Therefore, given the characteristics of the data utilized in this study, three heterogeneous machine-learning algorithms were applied as the base regression models, which are explained in detail below.

2.3.2.1. Gradient boosting machine. The Gradient Boosting Machine (GBM), comprised of multiple weak learners, can better reduce overfitting issues by constructing solutions in a stagewise manner over many boosting iterations and has become the go-to algorithm for training on structural data. The term "multiple weak learners" is used to describe a set of models that have limited predictive accuracy when used in isolation. By aggregating the predictions of multiple weak learners, the resulting model can attain higher accuracy than any of its individual constituents. Recently, a series of advanced GBM algorithms have been extended, and three of them with widely recognized performance are eXtreme Gradient Boosting (XGBoost, Chen and Guestrin (2016)), Light Gradient Boosting Machine (LightGBM, Ke et al. (2017)), and Categorical Boosting machine (CatBoost, Prokhorenkova et al. (2018)). In this study, the following hyperparameters were tuned for three GBM algorithms: (i) For XGBoost, the learning rate is set at 0.01, the maximum depth of the tree is set at 8, the minimum loss reduction is set at 0, and the subsampling rate is set at 0.5; (ii) For LightGBM, the learning rate is set at 0.01, the maximum number of leaves of the tree is set at 5, and the boosting type is set as Gradient Boosting Decision Tree (GBDT); (iii) For CatBoost, the boosting type is set as GBDT, the learning rate is set at 0.01, and the maximum depth of the tree is set at 10. The other model parameters followed default settings. After completing the training, the prediction results of the three models were averaged to maximize the data-mining capability.

2.3.2.2. One-dimensional convolutional neural network. Convolutional Neural Networks (CNN, Lecun et al. (1998)) have been one of the most potent developments in artificial intelligence in recent decades. With limited data, the One-dimensional Convolutional Neural Network (1d-CNN) may be a promising approach for data-mining one-dimensional signals (Malek et al., 2018; Núñez et al., 2022). It combines feature extraction, transformation, and data fusion in a single framework. Because neurons are sparsely connected with tied weights, the 1d-CNN can process significant inputs with excellent computational efficiency

compared to conventional fully connected multilayer perceptron networks. In this study, we constructed a 1d-CNN, as shown in Fig. S4 in Supplementary material. First, the feature dimension was increased through a fully connected layer in the architecture. Next, the features are extracted in several 1D-Conv layers with a shortcut-like connection. Finally, the extracted features predict the targets through a fully connected layer after flattening. We choose the mean absolute error as the loss function for the regression.

2.3.2.3. Attentive interpretable tabular learning neural network. The Attentive Interpretable Tabular Learning neural network (TabNet) is an advanced general-purpose deep neural network architecture for tabular learning (Arik and Pfister, 2021). The model combines the advantages of deep neural networks and tree models. The TabNet model selects a meaningful subset of features based on a sequential attention mechanism to enhance performance and interpretability. In this study, the Optuna framework (Akiba et al., 2019) was used to optimize the model hyperparameters. Following the hyperparameter optimization, we trained TabNet for 1000 iterations to reach the optimum.

2.3.3. Ridge regression ensemble

Although a multimodel ensemble strategy can effectively incorporate the benefits of various models and produce a more effective learner, it also increases the risk of overfitting. To achieve an ideal balance between model performance and the risk of overfitting, ridge regression is used to produce an optimal predictive model from the base models. Ridge regression is multiple linear regression technique that reduces the risk of overfitting (Pena and van den Dool, 2008). Specifically, ridge regression adds a regularization term to the loss function L of multiple linear regression, which is expressed as the L2 norm of the weight vector, ω, multiplied by the regularization coefficient λ:

$$L = \min_{\omega}\left[\|y - X\omega\|_2^{\,2} + \lambda\|\omega\|_2^{\,2}\right] \tag{2}$$

where $X = (x_1, x_{2,\ldots}, x_K)$ are the independent predictions of the K submodels, and y represents the true (or observed) value. Minimization of L leads to the weight vector ω, which is calculated as follows:

$$\omega = \left(X^{\mathrm{T}}X + \lambda I\right)^{-1}X^{\mathrm{T}}y \tag{3}$$

where I denotes the identity matrix. In this study, a grid search is conducted to determine the optimal value of λ between 0.001 and 1. It should be noted that we implemented the ridge regression algorithm using the "sklearn.linear_model.Ridge" function from the open source machine learning library Scikit-Learn (Pedregosa et al., 2011), where λ is the hyperparameter, and used grid search for optimization.

2.4. Accuracy assessment

2.4.1. Regression evaluation metrics

The coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and symmetric mean absolute percentage error (sMAPE) were utilized to quantify the performance of the model, according to:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}[p_i - \widehat{p}_i]^2}{\sum_{i=1}^{n}[p_i - \overline{p}]^2} \tag{4}$$

$$\mathrm{RMSE} = \left[\frac{1}{N}\sum_{i=1}^{N}(p_i - \widehat{p}_i)^2\right]^{1/2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|p_i - \widehat{p}_i| \tag{6}$$

$$\mathrm{sMAPE} = \frac{100}{N}\sum_{i=1}^{N}\frac{|\widehat{p}_i - p_i|}{(\widehat{p}_i - p_i)/2} \tag{7}$$

where p_i and $\widehat{p}_i$ are the log10-scaled observed and estimated Chla concentrations of each PG for sample i, N is the number of observations, $\overline{p}$ is the log10-scaled mean of the observed values.

2.4.2. Cross-validation approach

The 5-fold cross-validation (CV) method is a widely used validation tool that repeats the validation process five times, with four-fifths of the samples selected for training modeling and the remaining one-fifth for validation in each validation process to ensure that all samples were trained and validated. In many remote sensing applications, data segmentation is based on simple random selection. However, based on Tobler's first law of geography (Tobler, 1970), which states that "everything is related to everything else, but near things are more related than distant things", spatial data violate the assumption of independence required for many traditional statistical tests (Meyer and Pebesma, 2022; Ploton et al., 2020). This phenomenon, known as spatial autocorrelation, causes potential leakage of information from training to validation folds in the standard CV setting (Roberts et al., 2017; Stock, 2022), which may produce optimistically biased prediction performance estimates for spatial models and must be accounted for in machine-learning approaches (Stock and Subramaniam, 2022). To address this issue, a promising approach is to partition data into discrete "blocks" according to either time or space, which enables the creation of independent training and validation folds using spatial or temporal blocking. This method of block CV generates error estimates that provide a more accurate representation of the model's intended application in areas or periods where in situ data is not available. Consequently, we have adopted three distinct strategies: standard 5-fold CV, spatial block 5-fold CV, and temporal block 5-fold CV, which are elaborated upon below.

(1) Standard 5-fold CV

This study introduced information on the 14 optical water classes from the OC-CCI into the standard CV process (Fig. S5 in Supplementary material). Specifically, the random partitioning process of the datasets was modified as follows: the in-situ data were first grouped by optical water class, and then the training and testing datasets of each group were partitioned in a 4:1 ratio using the Kennard-Stone algorithm. Finally, the training and test datasets for all the groups were pooled.

(2) Spatial block 5-fold CV

We designed two zoning methods: (i) by ocean basin (Fig. S6a). We divided the global ocean into 11 spatial blocks based on the ocean area mask into 11 spatial blocks. (ii) Hexagonal gridding (Fig. S6b). A hexagonal grid was created at 20° horizontal and vertical intervals, and regions without sampling points were removed for 180 hexagonal regions. We arranged the in-situ samples in the order of the regions. In each group of five regions, samples from four regions were used to train the STEE model, and the rest were used for testing for a total of five iterations (Fig. S6c).

(3) Temporal block 5-fold CV

We placed the in-situ samples in chronological order and then divided them into groups of years. Four group samples were used to train the STEE model every five neighboring years, and the rest were used for testing. Five iterations were carried out, where the testing year changed in each iteration, as shown in Fig. S7.

2.4.3. Other models

Models from previous studies were included for comparison in this study, including abundance-based, spectral-based, and three other machine-learning models. The abundance-based approach is a univariate method that only uses chlorophyll as input. In our implementation of this approach, we used chlorophyll products from the OC-CCI as the sole input, but employed a polynomial regression technique that effectively reduced model complexity while maintaining predictive power. In

contrast, the spectral-based method is a multivariate approach that typically uses satellite observations of $R_{rs}(\lambda)$ or a derived spectral feature as input. For this study, we used R_{rs} data at wavelengths of 412, 443, 490, 510, 555, and 670 nm as inputs, without performing any spectral feature transformation, and fed them directly into the random forest model.

Moreover, three other typical machine-learning models were introduced as the ecological approach: i.e., decision trees (DT), multilayer perceptron (MLP), and support vector machines (SVM). These three models have been successfully applied in previous studies (Hu et al., 2018; Palacz et al., 2013; Stock and Subramaniam, 2020). The above three contrastive machine-learning models used the same input variables as the proposed STEE model in the specific implementation. All the models followed a consistent CV process to ensure the validity of the final accuracy comparison.

To further evaluate the proposed STEE-based PG products, we downloaded the global PG products from CMEMS (identified as OCEANCOLOUR_GLO_BGC_L4_MY_009_104) for comparison. These PG products are based on the empirical orthogonal function (EOF) approach (Xi et al., 2020), which uses dimensionality reduction to describe the dominant signal of structural variance in the spectral data, followed by parametric regression methods to construct the statistical PG model. EOF-based products are derived from multi-sensor merged ocean-color products or Sentinel-3A Ocean and Land Color Instrument data from the CMEMS, which are different from our products. Both the proposed STEE-based and EOF-based products have a resolution of 4 km. However, it should be noted that the EOF-based products only include six PGs.

3. Results

3.1. Model validation

3.1.1. Standard 5-fold CV

The Standard 5-fold CV procedure is used to test the performance of the proposed STEE-PG model. In general, the estimates of PG Chla concentration were highly consistent with in-situ measurements (Fig. 4). For all eight PGs, the determination coefficients (R^2) were higher than 0.6, demonstrating the superior capacity of the proposed model for data mining. Among the eight PG, diatoms exhibited the highest prediction accuracy with an R^2 value of 0.88. Prochlorococcus had the lowest level of accuracy, with an R^2 value of 0.61. The accuracies of the other estimated PG were between these values. The performance of the STEE model on different PG appears to be influenced by the concentration. The lower the concentration, the more difficult it is to estimate accurately. Through comparison, we found that all three machine-learning methods used in this study have superior data-mining capabilities, with the GBM model contributing the best modeling performance (see Fig. S8 in Supplementary Material). Moreover, after ridge regression merging, the prediction accuracy of the integrated model was further improved, which proves the effectiveness of the multimodel ensemble strategy proposed in this study. However, it should be noted that the slope of the regression line between the observed and estimated values was less than one, and the intercept was greater than zero. This phenomenon indicates that the suggested STEE-PG model may overestimate the Chla concentrations of PG at low phytoplankton biomass and underestimate those at high phytoplankton biomass.

3.1.2. Spatial and temporal block 5-fold CV

We employed spatial and temporal block cross-validation techniques to comprehensively evaluate the predictive capability of the proposed STEE model in regions or time periods lacking in situ data. The results of the temporal block 5-fold CV (Fig. 5) indicated that the model achieved R^2 values >0.5 for Diatoms, Dinoflagellates, Haptophytes, Green Algae, and Prokaryotes. However, the accuracy of the model for Prochlorococcus was relatively lower, with an R^2 value of 0.38, suggesting the need for further refinement of the model to improve its predictive performance for Prochlorococcus.

Additionally, Supplementary Fig. S9 and Fig. S10 present the results for the two spatial blocks of 5-fold CV, showing that both spatial CV strategies displayed similar outcomes. Diatoms maintained high modeling accuracy, with R^2 values >0.7, which demonstrates the STEE

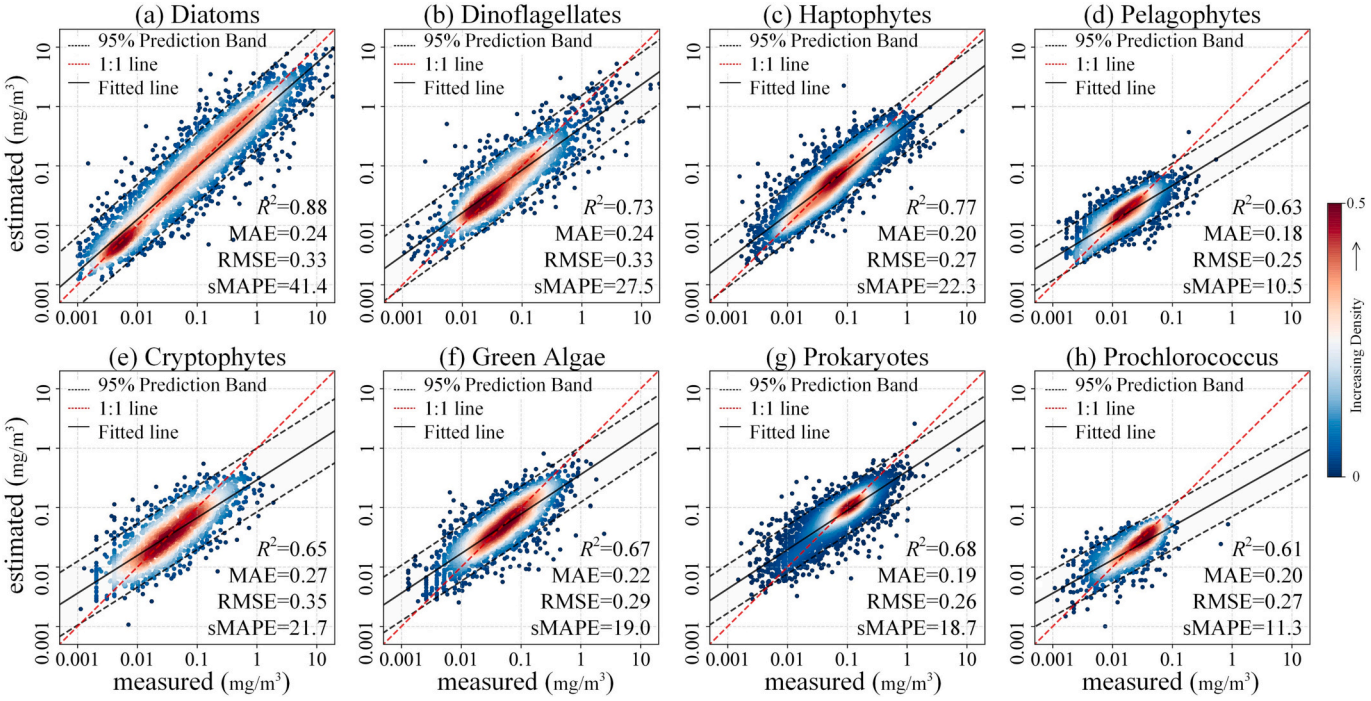

Fig. 4. Scatter diagrams (based on standard 5-fold CV procedure) of the predicted vs. measured Chla concentrations of (a) Diatoms, (b) Dinoflagellates, (c) Haptophytes, (d) Pelagophytes, (e) Cryptophytes, (f) Green Algae, (g) Prokaryotes and (h) Prochlorococcus, along with the evaluation metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

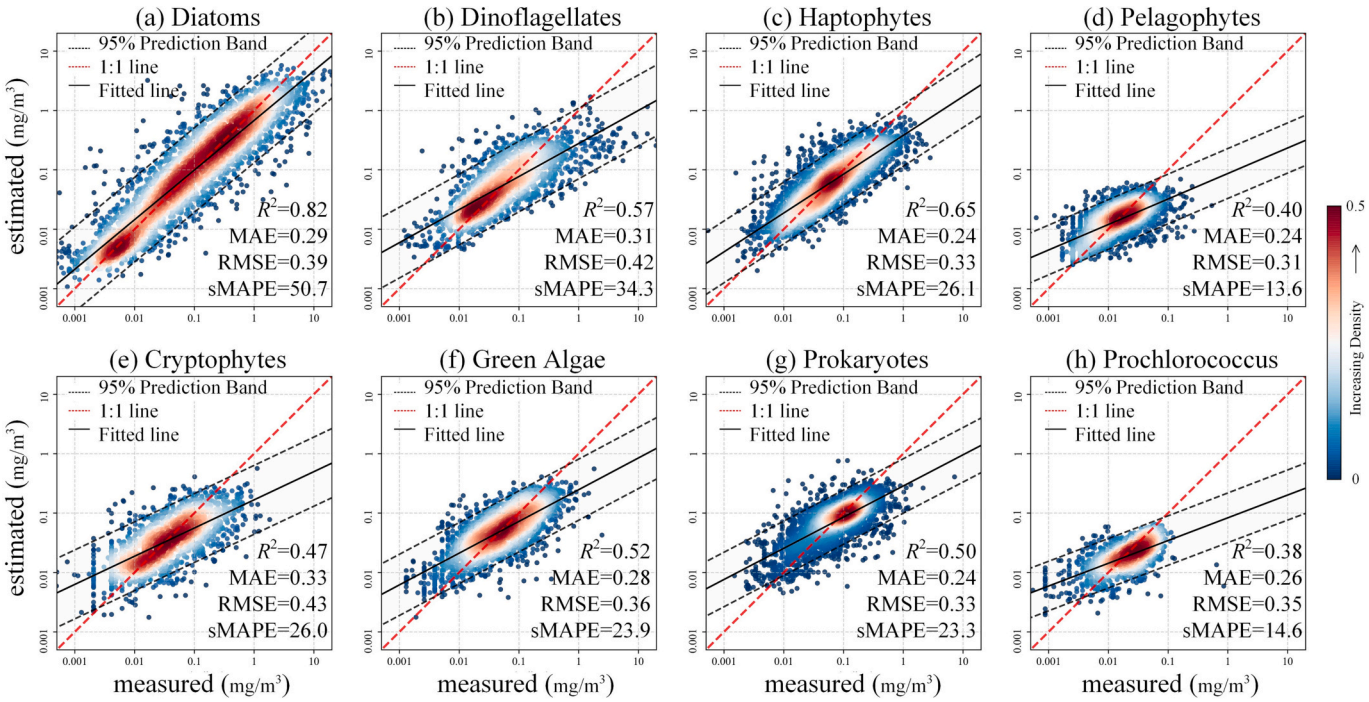

Fig. 5. Scatter diagrams (temporal blocks CV procedure) of the predicted vs. measured Chl-a concentrations of (a) Diatoms, (b) Dinoflagellates, (c) Haptophytes, (d) Pelagophytes, (e) Cryptophytes, (f) Green Algae, (g) Prokaryotes and (h) Prochlorococcus, along with the evaluation metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model's good extrapolation capability at the spatial scale. In contrast, Prochlorococcus and Pelagophytes exhibited significant performance degradation, with R^2 values below 0.4.

Fig. 6 illustrates the accuracy of the model under different cross-validation strategies. The results indicate that the STEE model exhibited a reduction in accuracy under spatial and temporal CV strategies when compared to standard CV. Nevertheless, in general, the model was still able to show good extrapolation in regions or time periods without in situ data, with particularly promising outcomes for Diatoms, Dinoflagellates, and Haptophytes.

3.2. Assessment and applicability

3.2.1. Global ocean scale

We compared the performance in estimating the PG between our proposed STEE-PG model and five models from previous studies (see Section 2.4.3). The standard CV procedure was first applied to compare the performances of the models. A rose diagram (Fig. 7) and normalized Taylor diagrams (Fig. 8) were utilized to provide a more comprehensive and visualized result for the comparison. The evaluation metrics for each model are presented in Table 2.

As demonstrated in the normalized Taylor diagrams (Fig. 8), the points of the STEE-PG model (i.e., red circles) are considerably closer to the reference point (i.e., blue circles represent the in-situ data), indicating that the STEE model outperforms the other models in terms of predictive capacity. The performance of each model on diatoms was relatively good, as shown in Table 2. In contrast, the estimation of Prochlorococcus was the most difficult, with R^2 <0.3 for all the comparison models except the proposed STEE model, which achieved an R^2 value of 0.6. The overall performance of the MLP and DT models is comparable, but both are marginally worse than that of the SVR model. Machine-learning-based ecological methods consistently outperform

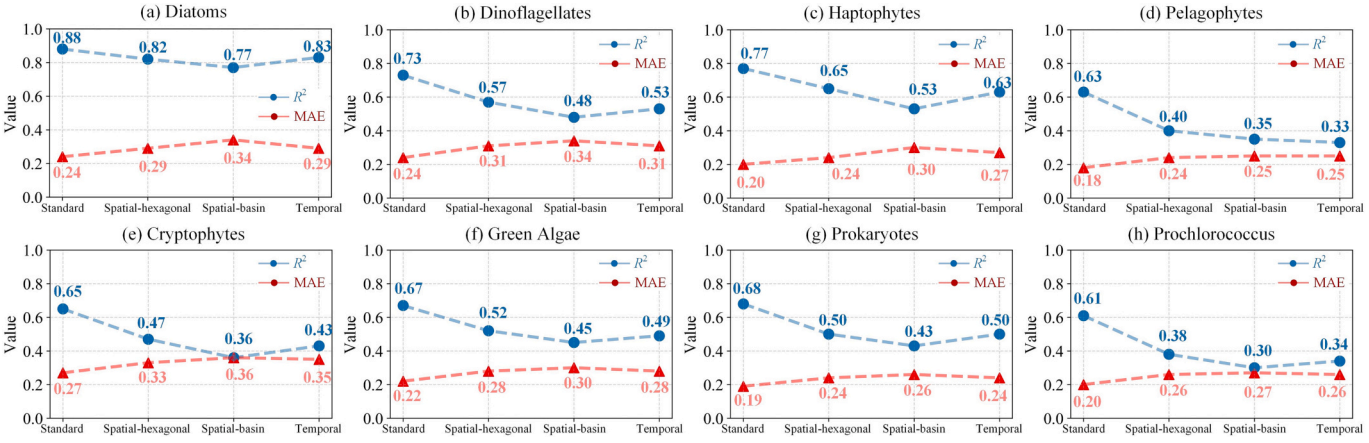

Fig. 6. Comparison of the results obtained using different CV methods, including standard CV, spatial block CV based on ocean basin, spatial block CV based on hexagonal grid, and temporal block CV. For further details on these methods, please refer to Section 2.4.2.

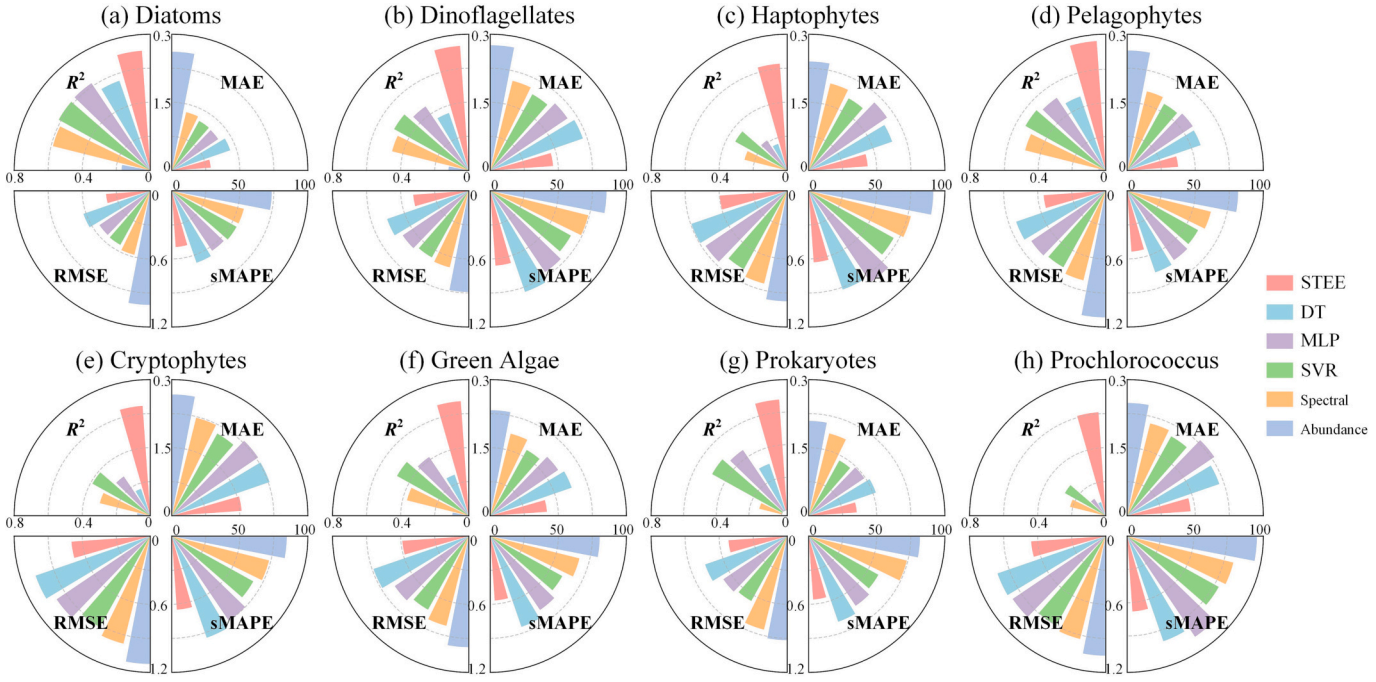

Fig. 7. Model performance metrics (R^2, MAE, RMSE, and sMAPE) of STEE model and five models for eight PG. Note that the evaluation metrics here are calculated based on the standard 5-fold CV procedure.

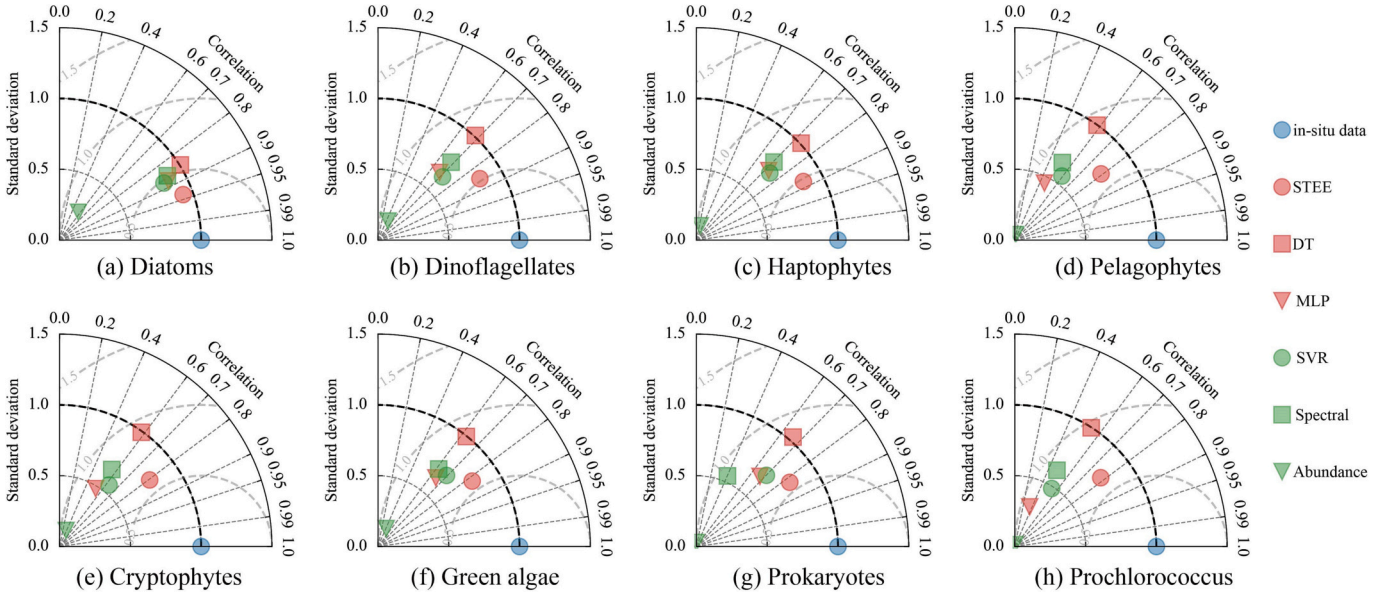

Fig. 8. Taylor diagrams (based on standard 5-fold CV procedure) for comparing the performance between STEE-PG and five models. The radial dimension represents the model standard deviations normalized by the observations. Correlation coefficients are represented in angular coordinates, whereas the arcs show the RMSE. The blue circle represents the referring point of each PG. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

spectral-based and abundance-based models in terms of prediction accuracy. This is attributable to the ecological approach, which introduces more valid environmental variables, and the superior data-mining capabilities of machine-learning models. Although the spectral-based model performed relatively well for Diatoms, Dinoflagellates, and Pelagophytes ($R^2 > 0.4$), it showed poor performance for other PGs. The abundance-based model performed the worst among all models, where the simple polynomial regression utilized to implement the model could be the reason for the low performance. In general, the performance of the proposed STEE model surpasses that of the other models with higher R^2 and lower MAE, RMSE, and sMAPE for all eight PGs.

Fig. 9 illustrates the Taylor diagrams for the time-block 5-fold CV, while Supplementary Fig. S11 and S12 display the spatial block 5-fold CV. The evaluation metrics for each model can be found in Table S3, S4 and S5. According to the Taylor diagram, the results of the proposed STEE model were consistently the closest to the reference point, regardless of the CV method. While the performance of the STEE model decreased under the block CV-based strategy, the proposed model still exhibited the best prediction accuracy compared to the other models, as shown in Tables S3, S4, and S5. This indicates the clear superiority of the

Table 2
Comparison of Model performance metrics (R^2, MAE, RMSE, and sMAPE, based on standard 5-fold CV procedure) between STEE-PG model and other models (see Section 2.4.3 for details) in eight PG. For each PG, the evaluation metrics with higher performance are shown in bold.

PG	Metrics	Model					
		STEE	DT	MLP	SVR	Spectral	Abundance
Diatoms	R^2	**0.88**	0.70	0.77	0.77	0.74	0.21
	MAE	**0.24**	0.37	0.34	0.34	0.36	0.71
	RMSE	**0.33**	0.52	0.45	0.45	0.48	0.84
	sMAPE	**41.38**	55.84	53.00	54.43	54.55	73.48
Dinoflagellates	R^2	**0.73**	0.36	0.45	0.50	0.46	0.12
	MAE	**0.24**	0.37	0.35	0.33	0.35	0.47
	RMSE	**0.33**	0.51	0.47	0.45	0.46	0.59
	sMAPE	**27.53**	39.22	35.99	33.86	37.09	42.69
Haptophytes	R^2	**0.77**	0.46	0.52	0.54	0.49	0.04
	MAE	**0.20**	0.30	0.30	0.29	0.31	0.45
	RMSE	**0.27**	0.42	0.40	0.38	0.41	0.56
	sMAPE	**22.34**	31.64	30.65	29.76	31.67	40.74
Pelagophytes	R^2	**0.63**	0.17	0.21	0.35	0.26	0.01
	MAE	**0.18**	0.27	0.29	0.25	0.27	0.33
	RMSE	**0.25**	0.37	0.36	0.33	0.35	0.41
	sMAPE	**10.52**	15.33	16.20	14.27	15.46	18.30
Cryptophytes	R^2	**0.65**	0.17	0.28	0.39	0.31	0.08
	MAE	**0.27**	0.39	0.39	0.35	0.38	0.46
	RMSE	**0.35**	0.53	0.50	0.46	0.49	0.56
	sMAPE	**21.65**	31.55	29.95	27.38	29.46	33.77
Green algae	R^2	**0.67**	0.25	0.42	0.48	0.37	0.10
	MAE	**0.22**	0.33	0.31	0.29	0.32	0.40
	RMSE	**0.29**	0.44	0.39	0.37	0.41	0.49
	sMAPE	**18.95**	28.18	26.14	24.21	27.00	32.16
Prokaryotes	R^2	**0.68**	0.30	0.45	0.49	0.14	0.00
	MAE	**0.19**	0.27	0.26	0.24	0.32	0.36
	RMSE	**0.26**	0.38	0.34	0.33	0.42	0.46
	sMAPE	**18.65**	26.58	24.66	23.40	29.70	32.67
Prochlorococcus	R^2	**0.61**	0.08	0.12	0.27	0.22	0.00
	MAE	**0.20**	0.30	0.32	0.28	0.29	0.34
	RMSE	**0.27**	0.42	0.41	0.37	0.39	0.44
	sMAPE	**11.34**	16.69	18.22	15.72	16.43	19.48

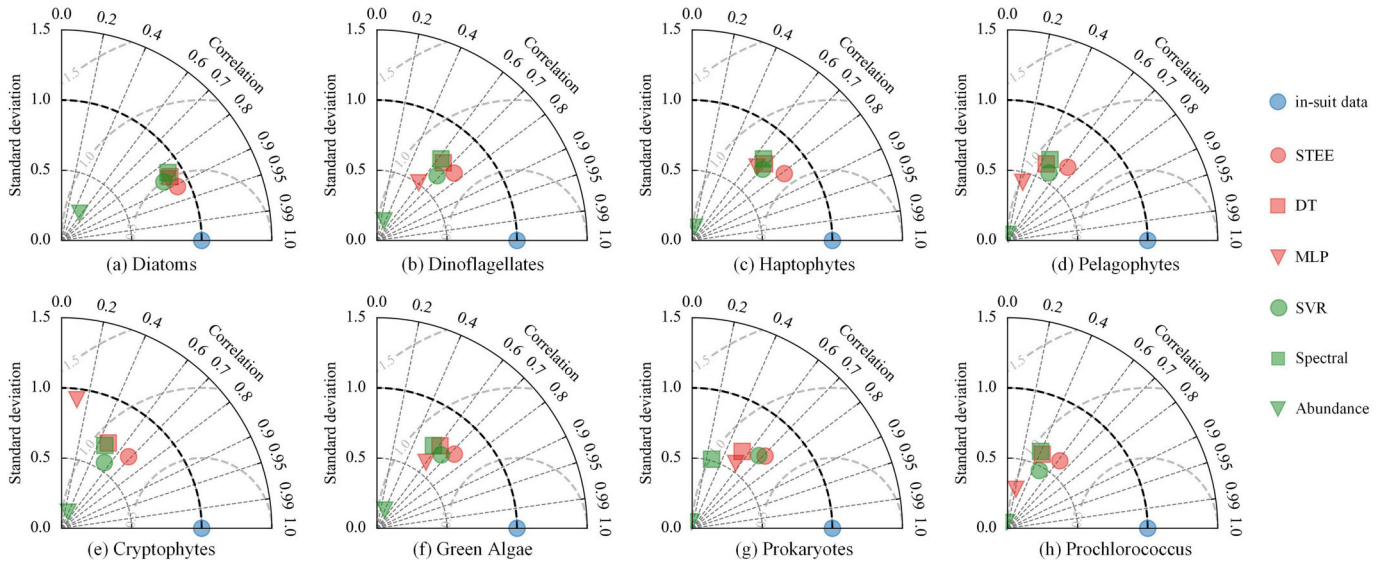

Fig. 9. Taylor diagram of temporal block CV for comparing the performance of STEE-PG with five other models. The radial dimension represents the model standard deviations normalized by the observations. Correlation coefficients are represented in angular coordinates, whereas the arcs show the RMSE. The blue circle represents the referring point of each PG. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

STEE model. These CV strategies complement each other, demonstrating the validity and effectiveness of the proposed STEE model.

3.2.2. *Ocean basins scale*

We further examined the effectiveness of the proposed STEE model in various ocean basins (i.e., the Atlantic, Pacific, Indian, and Southern Oceans) and coastal areas, where the boundaries of the ocean basin are referred to as the RECCAP2 project (Fig. 3). The results in Fig. 10 show that the STEE model outperformed the other five models for all PG in both oceanic and coastal regions. Similar to the results for the global ocean, the abundance-based model is less effective in accurately estimating PG. However, the three different machine-learning approaches (SVR, MLP, and DT) outperformed the spectral-based and abundance-based methods. The proposed STEE model showed consistent and

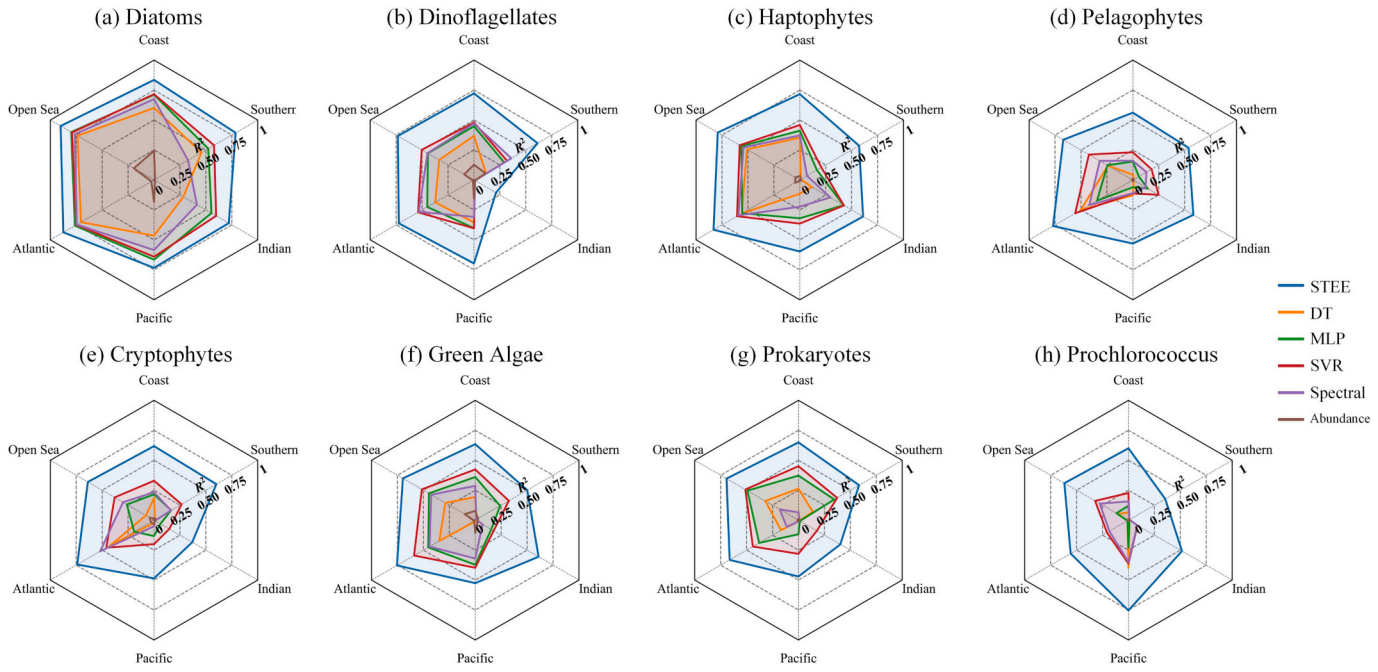

Fig. 10. The performance of each model in different regions. The radial dimension represents R^2. Note that the evaluation metrics here are calculated based on the standard 5-fold CV procedure.

satisfactory prediction accuracy across different regions and PG. However, in the Indian Ocean, all six models exhibited low prediction accuracies for dinoflagellates, which could be attributed to the imbalanced distribution of the in-situ datasets (i.e., Indian Ocean samples are few, see Fig. 3b). More detailed evaluation metrics of the Fig. 10 are shown in Table S6 in the Supplementary Material.

3.2.3. Optically complex waters

To explore the application potential of the proposed STEE model, a comparison experiment was conducted to verify its validity in optically complex waters. Specifically, instead of using global data, we extracted samples from optically complex waters to train the model. Cross-validation was used to compare the performance of the proposed STEE model with those of other models in optically complex waters. The optical classification of a pixel by OC-CCI products indicates, to some extent, whether the pixel is likely to belong to Case-1 or Case-2 waters. As a rule of thumb, higher-numbered categories are more likely to belong to Case-2 waters with high scattering and are mostly located near coastal areas.

First, we extracted samples belonging to optical water classes 12, 13, and 14. The location distribution and satellite spectra of the sampled points are shown in Fig. S30 in the Supplementary Information. Subsequently, using these sampling points as typical representatives of complex water bodies, we compared the prediction accuracy of the proposed model with that of five other models (detailed in Section 2.4.3) using the CV method. Because of the small number of samples, we only performed standard random CV and did not use spatial and temporal block cross-validation strategies. The results are presented in Fig. S31 and normalized Taylor diagrams (Fig. S32) were used to provide more comprehensive and visualized results for the model comparison. The results show that the proposed STEE model has better prediction accuracy, even for optically complex water bodies. The standard Taylor diagram shows that the STEE model is closer to the reference point, indicating a better prediction of the model. The specific model accuracy evaluation metrics are listed in Table S7, which further illustrates that the STEE model achieves the best model accuracy. Therefore, the proposed STEE model outperforms other methods in optically complex waters. The introduction of ecological parameters is effective in improving the PG model. Further combination with the powerful non-linear modeling capability of machine learning can significantly improve the accuracy of the PG models.

3.2.4. Long-time series observations

Using independent long-term in-situ observations from six global sites (Fig. 2a), we validated and compared the performance of the two global PG products, that is, the STEE-PG and EOF-PG products (see Section 2.4.3). Note that we did not use other competitive models for time-series comparisons. The EOF-PG product has been well validated; therefore, the comparison of this product is a better demonstration of the validity of the proposed STEE model. Fig. S14-S21 in the Supplementary Material compares the field data, STEE-PG, and EOF-PG products for eight PGs at six long-term series sites. Generally, good agreement was observed between the STEE-PG, EOF-PG, and field data.

Fig. 11 depicts the correlation analysis of the two products at six long-time series sites. Given the irregular frequency of the in-situ sampling, we calculated the monthly average of the in-situ observations before evaluating the correlation coefficients. Similar to Fig. 4, the STEE model has the most accurate predictions for diatoms, with correlation coefficients of over 0.6 at all six sites. Except for the PAB site, the correlation coefficients for Green Algae were <0.5 at all the other five sites, indicating that the STEE model has poor generalization performance for Green Algae. For Diatoms, Dinoflagellates, and Haptophytes, the correlation coefficients between the STEE-based products and in-situ measurements were higher than those derived from the EOF models. For Prokaryotes, the prediction accuracy of the STEE model was significantly greater than that of the EOF model, except for the Port-Hacking site. This is because that the number of matches of EOF-PG products obtained at the Port-Hacking site is less than those of STEE-PG products, leading to deviations from the actual situation. Overall, the estimates of PG from the STEE model have high correlation coefficients with independent in-situ measurements, suggesting that the STEE model has the potential to estimate PG in a long time series with high accuracy.

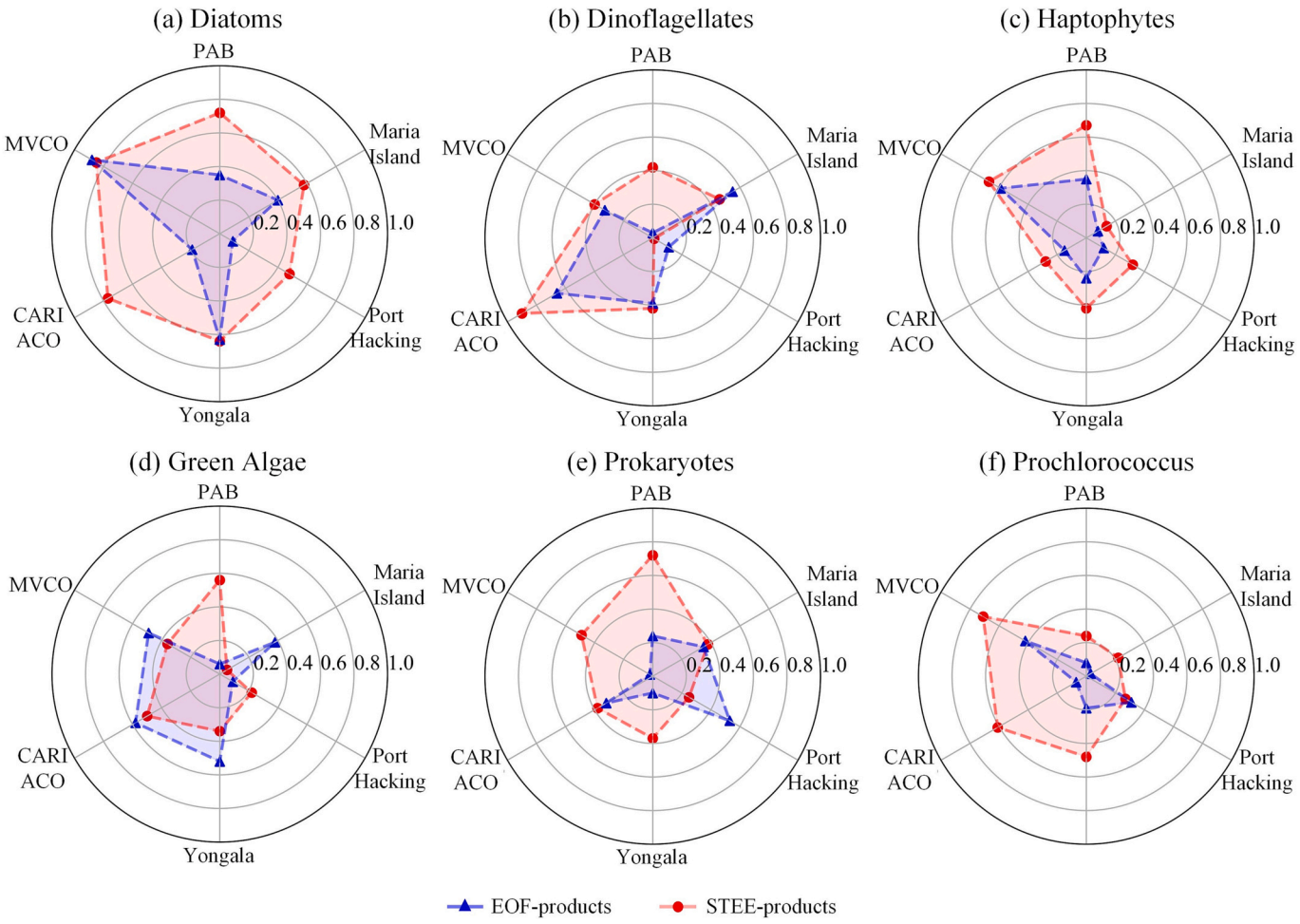

Fig. 11. Performance comparison of STEE-PG and EOF-PG products at six independent time series stations.

3.3. Global products of PG composition

We retrieved monthly global satellite products of PG Chla concentration composition from September 1997 to May 2020 with a resolution of 4 km by applying the proposed STEE model to monthly imagery, reanalysis datasets, and spatio-temporal information given in Section 2.2. After calculating the mean value of each pixel from all monthly products, Fig. 12 shows that each PG exhibits a distinctive spatial distribution pattern. In general, the geographic pattern of the eight PGs calculated using the proposed STEE model is consistent with current knowledge. Diatoms, for example, represent an important component of phytoplankton biomass at high latitudes (>60°) and coastal waters with higher nutrient supply and turbulent conditions, such as those found in the eastern China Sea, the Bering Sea, and the Southern Ocean. The Chla of diatoms is substantially lower in the Pacific and Atlantic Ocean gyres than that of other PGs. Dinoflagellates have a spatial distribution pattern similar to diatoms; however, dinoflagellates have lower Chla values than diatoms. Higher Chla values for Haptophytes, Pelagophytes, and Green Algae were observed in middle-latitude regions and the Eastern Equatorial Pacific, whereas lower values were found in gyres and at higher latitudes. Cryptophytes are more uniformly distributed in the global ocean than the other seven PG, with slightly larger concentrations observed in nearshore waters and medium latitudes in the Northern Hemisphere. The distribution patterns of Prochlorococcus and Prokaryotes differed significantly from those of the other PG. They are most abundant at low latitudes, such as the warm euphotic zone of tropical and subtropical oligotrophic oceans. To illustrate the seasonal variation in the Chla distribution for different PGs, the monthly climatological products for eight PGs are provided in Fig. S22-S29 in the Supplementary Material.

4. Discussion

4.1. Different PG products

To demonstrate the consistency and discrepancy between the proposed STEE-PG and EOF models (Xi et al., 2020), comparisons of the global monthly products derived from the two models from 2003 to 2019 are shown in Fig. 13 and Fig. S13. In general, STEE-based products reveal a more stable pattern than EOF-based products over a long time series. Taking diatoms as an example, the EOF-based products (blue triangles, Fig. 13 a1) exhibit a more pronounced oscillation phenomenon than those from the STEE model (red circles, Fig. 13 a1). Before 2012, Diatoms Chla from the EOF model had greater differences between high latitudes (~60°) and low latitudes (0–30°) (Fig. 13 a2), and a significant decline in Chla was observed after 2012, where differences in Chla between high and low latitudes diminished. From April 2016 onwards, the Dinoflagellates products from the EOF model experienced a notable increase (Fig. 13 b1), which was caused by a significant increase at high latitudes (40–60°, Fig. 13 b2), and a decrease in Chla in low-latitude regions was observed in the EOF-derived products. Similarly, owing to the decline of Chla at most latitudes, the Prokaryotes products of the EOF model exhibited a rapid decrease after 2016 (Fig. 13 c1), which is consistent with the STEE-derived products.

Inconsistencies in the OC data from different sensors may cause unreasonable mutations in EOF-PG products. In addition, the retrieval

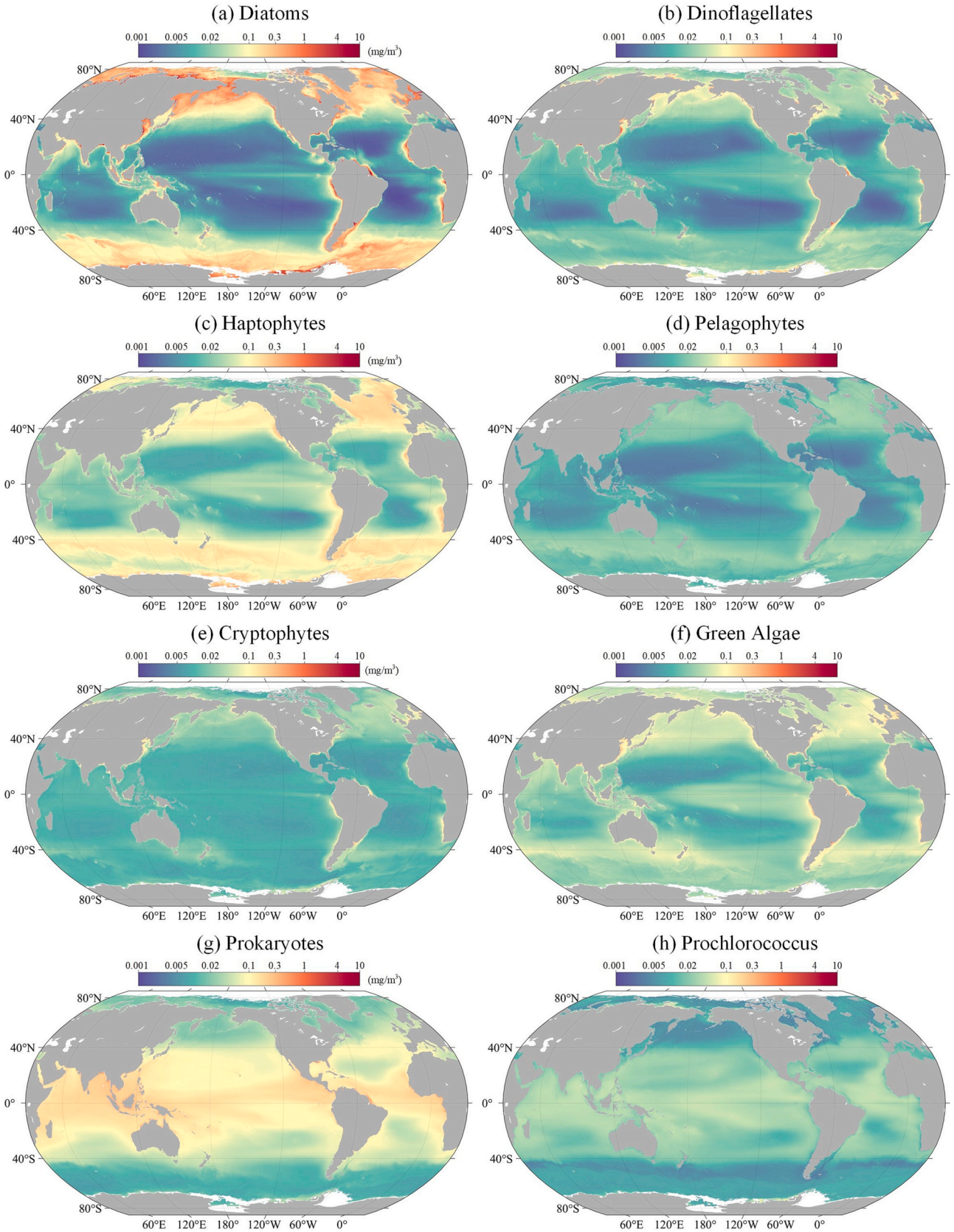

Fig. 12. The global mean distribution (September 1997–May 2020) of the Chla concentration for (a) Diatoms, (b) Dinoflagellates, (c) Haptophytes, (d) Pelagophytes, (e) Cryptophytes, (f) Green Algae, (g) Prokaryotes and (h) Prochlorococcus. The grey areas represent lands. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

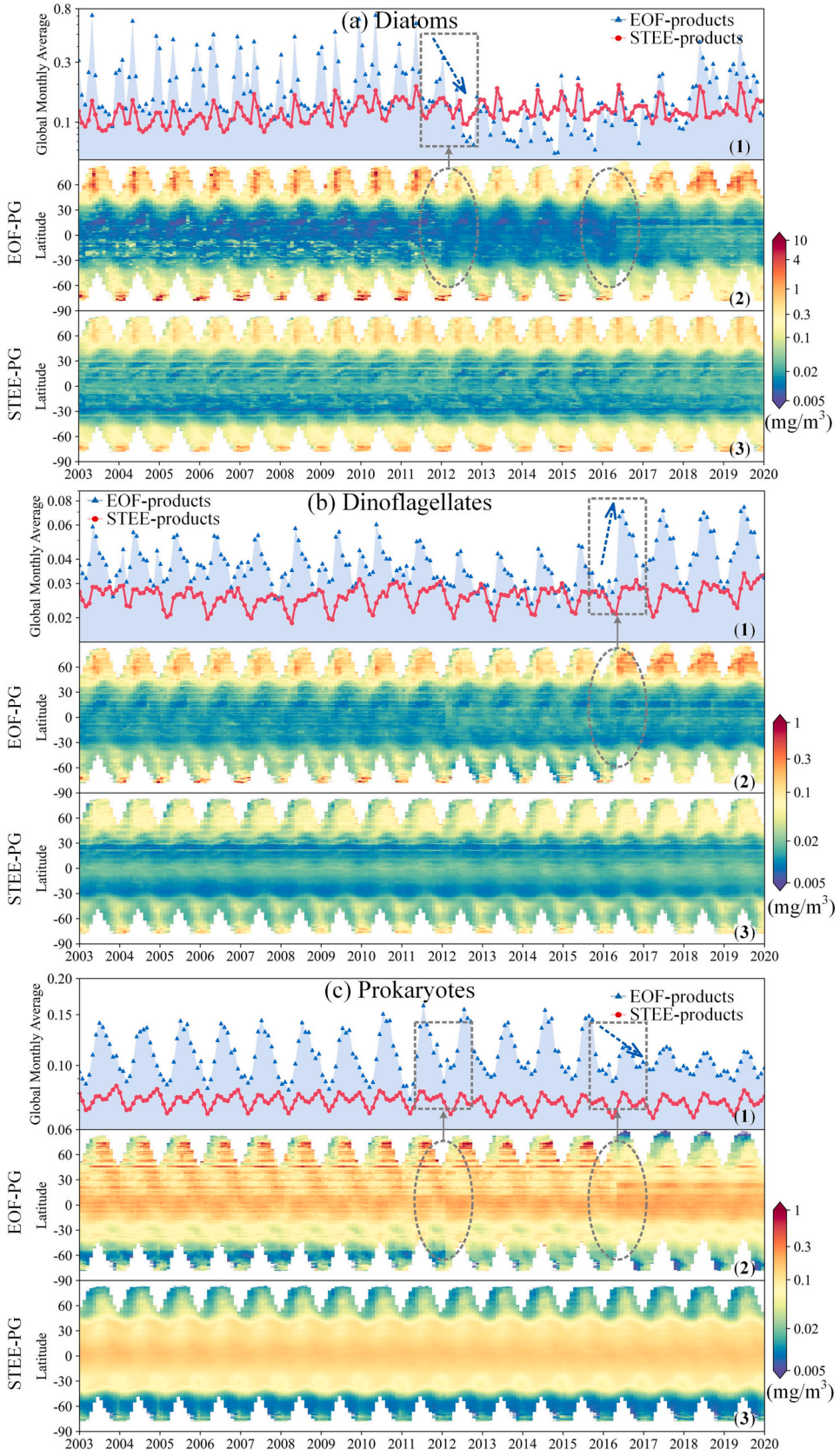

Fig. 13. Comparison of the two PG products from 2003 to 2020 of (a) Diatoms, (b) Dinoflagellates, and (c) Prokaryotes. The first row of each sub-figure shows the global monthly average variation of the product, the second and the third rows are Hovmöller diagrams of the EOF and STEE-derived products, respectively.

equations were derived based on matchups from 2002 to 2012. The EOF-PG product was produced based on GlobColour OC data (http://www.globcolour.info/), which was merged from multiple sources, including SeaWiFS, MODIS-Aqua, MERIS, VIIRS-NPP, and Sentinel-3A OLCI data. The starting dates of the VIIRS-NPP and OLCI data are 2012-01-02 and 2016-04-25, respectively, which coincide with the mutation points of the products.

The differences in the Hovmoller latitude-averaged Chla of the PG derived from the two models are shown in Fig. 14. Overall, the trend consistency between the two products was good, although there were significant differences between Prokaryotes and Prochlorococcus. According to Xi et al. (2020), the modeling accuracy of EOF-PG on Prokaryotes and Prochlorococcus is lower, which may be the reason for the more significant difference between the two products. In general, the EOF-based products had slightly higher values of Chla than those from STEE-based products, especially at higher latitudes.

The comparison of products from two different models demonstrates that the use of bio-optical properties and a set of static model parameters alone makes the model more susceptible to spectral variability and noise, which may lead to unreasonable product variations. In contrast, the proposed STEE model is more robust, and its derived PG products have better stability and spatiotemporal consistency.

4.2. Assessment strategies

The cross-validation method is a commonly used approach for evaluating models in PG modeling and other remote sensing inversion fields. Nevertheless, recent studies have revealed the inadequacies of using random selection of test data as it does not guarantee independence from the training data. To overcome this challenge, it is necessary to incorporate temporally and spatially separated block cross-validation strategies. These techniques can help mitigate the effects of spatial autocorrelation present in remotely sensed data and reduce overly optimistic estimates of model accuracy by testing blocks of spatial and temporal data that are not used in the training process.

In the present study, we employed multiple complementary cross-validation (CV) techniques, including standard CV, spatial block CV, and temporal block CV. As illustrated in Fig. S6 and S7, the block CV methods reduce the optimistic bias in standard CV by introducing spatial or temporal separation, which allows us to estimate the predictive accuracy of the STEE model in regions and years without field data. This is particularly important for global PG mapping, where reliable in situ data may be scarce. While the use of block CV methods increases the estimation error of the STEE model, it still outperforms other competing models, as evidenced in Table S3, S4, and S5. It is worth noting that the choice of dividing spatial or temporal blocks can be subjective and may affect the block CV errors. Therefore, the use of multiple complementary methods is recommended to obtain a more reliable assessment of the model's predictive performance in real-world scenarios.

More refined local assessments are one of the next research priorities. With increasing field sampling data, we will further validate the accuracy in local areas using independent validation data in a subsequent study. In addition, in future studies, we will evaluate the pixel-by-pixel uncertainty of PG composition retrievals by considering the uncertainty of the input data and model parameters in combination with uncertainty propagation methods, which will provide reliable uncertainty for estimating PG composition products and enable us to better understand the quality of the products in time and space.

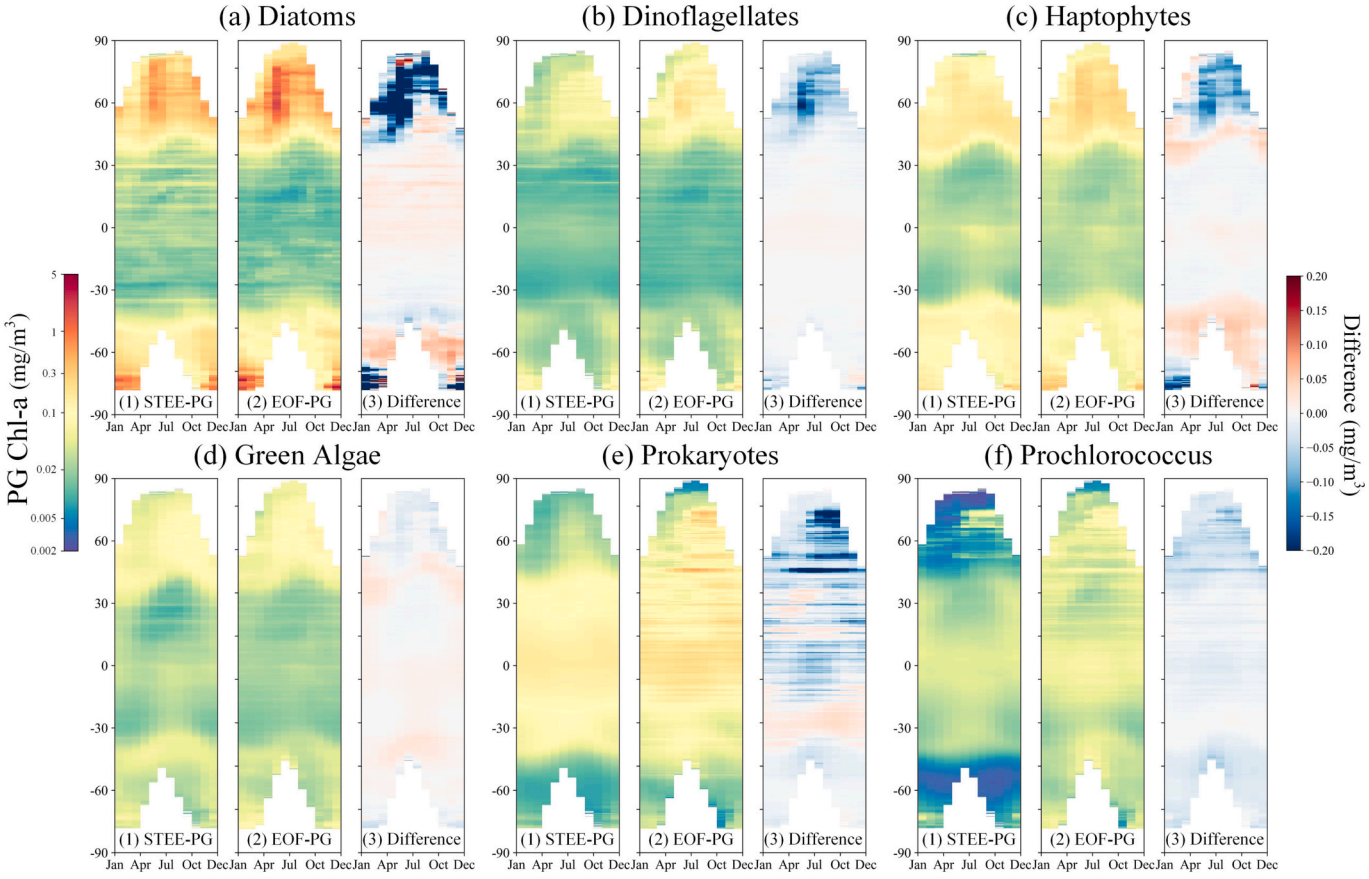

Fig. 14. The time-latitude Hovmöller diagrams of the annual climatological cycle for STEE-based and EOF-based products from 2003 to 2019. The colormaps on the left and right represent the Chla and difference of Chla, respectively.

4.3. Potential and limitations

It is common knowledge that marine science is entering the era of big data, characterized by an explosion of various in situ observations, quantitative remote sensing products, reanalysis data, and supporting calibration and validation data. The accompanying large volume of optical, physical, chemical, meteorological, and other marine environmental data continues to increase in spatial coverage, time, and quality (Huang et al., 2015). However, these precious ancillary data have not yet taken full advantage of phytoplankton remote sensing. This study suggests that the combination of multi-source data including satellite data and ocean environment data by machine learning techniques can liberate researchers from the challenge of PG subtle spectral discrimination, and provide a flexible way to obtain PG distributions in global scale. The big data-driven modeling paradigm also offers new ideas for subsequent spectral and hyperspectral based remote sensing studies of phytoplankton. We believe that the proposed method has broad application prospects in future research, including and not limited to (i) more accurate and refined carbon estimation of phytoplankton, (ii) research on the response mechanism of PGs to environmental change, and (iii) prediction of the future global distribution pattern of PGs under climate change.

Although the research results show that the STEE model has good accuracy in estimating PG composition, there is still room for further improvement. In this study, the slopes of the fitted line of the STEE model between the measured and estimated values were typically <1 (Fig. 4), indicating that there were underestimates at high values and overestimates at low values. This phenomenon has also occurred in several previous PG models (Xi et al., 2021; Xi et al., 2020), where the uneven distribution of samples in the field data is probably the major reason. It is worth mentioning that all the PG (Fig. S1 in the supplementary) typically have long-tailed or log-normal distributions in global oceans, meaning that most of the samples are scattered in the low-value region. Most previous studies have log-transformed the in-situ HPLC pigment data during the analysis (Brewin et al., 2010; Kramer and Siegel, 2019; Ward, 2015). However, this log-transformation strategy cannot fully eliminate the data imbalance and may jeopardize the generalizability of the model. Therefore, subsequent studies must explore new label transformation methods or loss functions to reduce the impact of unbalanced regression. (Liu et al., 2020; Yang et al., 2021).

The proposed STEE model implicitly captures interactions between environmental factors and other poorly understood biogeochemical factors. However, the STEE model has black-box properties and lacks a mechanistic basis for phytoplankton distribution. The development of interpretable artificial intelligence has the potential to gain mechanistic insights into complex machine-learning models (McGovern et al., 2019; Reichstein et al., 2019), creating the possibility of opening the black box. One effective way to do this is to highlight the most important variables in the input space that help the model make a specific prediction (Toms et al., 2020). For example, by combining random forest-based feature importance measures and partial dependence plots to identify drivers of phytoplankton abundance (Rivero-Calle et al. (2015)), or by using neural network ensembles to model the interactions between predictors and their effects on phytoplankton biomass (Holder and Gnanadesikan (2021a)). In addition, Monte Carlo estimates (Sobol, 2001) and also Jacobian matrix (Maddy and Boukabara, 2021) also have promising applications in examining the sensitivity of the results to perturbations in the inputs. This will be further explored in future applied research on the proposed STEE model.

In addition, the proposed model has the following limitations. First, the model is highly dependent on the input data; uncertainties can be raised by the error of in-situ measurements and HPLC experimental data from different regions. Because of the large number of sources in the dataset used in this study (Table 1), different laboratory processing methods and criteria are involved in obtaining pigments, which may lead to uncertainties. Similar concerns exist for other input datasets (i.e., physical, chemical, and bio-optical dataset). Therefore, it is difficult to obtain detailed estimates of the uncertainty for each subdataset. Second, the weights used in the DPA could be another source of uncertainty. A quality-controlled HPLC dataset can be used with data-driven statistical methods to characterize phytoplankton communities reasonably. In this study, we used the weight values from Losa et al. (2017), which were calculated based on a global dataset. However, previous studies have demonstrated that because of phytoplankton composition, the relationships between each PG and total Chla have regional differences, as reflected in weight differences (Bracher et al., 2017; Mouw et al., 2017). Based on the results of the model developed by Kramer and Siegel (2019), only four PGs could be detected in the surface ocean, whereas many more groups require further analysis at a local scale. Therefore, as DPA is applied to future studies of ocean ecosystems, satellite algorithm development, and ecosystem models, its inherent biases and uncertainties must be considered, and the weights must be constantly revisited at local scales. Thirdly, the STEE model incorporates spatial information (i.e. geographic, latitude and longitude) and timestamps directly into the model construction. However, it needs to be considered that the relationships between geographic location, season, and phytoplankton composition may break under climate change, thus compromising the generalization ability of the model. Therefore, incorporating additional biogeochemical knowledge as mechanistic constraints in future research and product application could be considered. Furthermore, developing more reasonable spatiotemporal coding methods is necessary to avoid overfitting of the spatial structure. Due to the susceptibility of ocean-color remote sensing to cloud coverage, a large number of invalid values in the data could hinder global assessments. In this circumstance, robust cloud-filling methods are required to fill the gap in satellite products. These issues require further investigation in future research.

5. Conclusions

With the advent of the marine big data era, the diversity and range of large environmental variable datasets and technological advances in machine learning have provided an unprecedented opportunity to quantify the composition and distribution of phytoplankton groups. Based on the ensemble machine-learning strategy and multi-source data integration, this study established a high-precision global PG estimation model. To the best of our knowledge, this study is the first global-scale attempt to apply data mining and machine learning to improve the long-term series retrieval of PG.

The main contributions of this paper are as follows:

(i) The present study simultaneously included dozens of environmental variables, such as ocean color, physical ocean, biogeochemical, and spatial and temporal information variables as predictors, thereby considerably increasing the predictive potential of the model.

(ii) This study presents a new STEE model that combines three state-of-the-art machine learning methods (GBM, 1d-CNN, and TabNet) using an ensemble strategy to enhance the model's robustness. Multiple cross-validation techniques were employed to evaluate the performance of the proposed framework, and the results demonstrate its effectiveness in achieving accurate predictions. The standard cross-validation analysis showed strong correlation (R^2 values >0.6) between the model and all eight phytoplankton groups. Furthermore, despite a decrease in prediction accuracy for some phytoplankton groups under the block CV strategy (e.g., R^2 values <0.4 for Prochlorococcus and Pelagophytes), the STEE model outperformed the other models in block CV, indicating its robustness and considerable extrapolation ability on spatial and temporal scales.

(iii) The proposed STEE model was applied to map the monthly global products for the eight PGs. The STEE-PG product has better stability and spatiotemporal consistency than the published EOF-PG product.

We conclude that ecological approaches based on multi-source data

integration and artificial intelligence provide new insights into the ecological modeling of phytoplankton and could be powerful tools for advancing phytoplankton ecosystem observations. Our approach will likely enable us to address the mechanism by which environmental factors affect the distribution of phytoplankton and investigate the response of phytoplankton to climate change in future research.

Author contributions

Conceptualization – Project Administration: Fang Shen;
Methodology: Yuan Zhang, Fang Shen, and Kun Tan;
Writing – Original: Yuan Zhang;
Writing – Review & Editing: Yuan Zhang, Fang Shen and Xuerong Sun.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was funded by the National Natural Science Foundation of China (Nos. 42076187 and 42271348). This study is also a part of the research objectives of the IMBeR OC-PC working group and is supported by the ESA-FE EO-WPI project. Xuerong Sun is supported by a UKRI Future Leader Fellowship (MR/V022792/1). This study utilizes various in situ observations and databases, and we express gratitude to the many scientists and crew involved with collecting and processing the data, which have been madefreely and publicly available. All data used in the study are properly cited and referenced We are extremely grateful to three anonymous reviewers and Prof. Frédéric Mélin for their insightful and constructive comments and suggestions, which help us improve the quality of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2023.113596.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. In: Optuna: A Next-generation Hyperparameter Optimization Framework. Kdd'19: Proceedings of the 25th Acm Sigkdd International Conferencce on Knowledge Discovery and Data Mining, pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.

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

Arik, S.O., Pfister, T., 2021. In: TabNet: Attentive Interpretable Tabular Learning. Thirty-Fifth Aaai Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence and the Eleventh Symposium on Educational Advances in Artificial Intelligence, 35, pp. 6679–6687, 10.48550/arXiv.1908.07442.

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 https://doi.org/10.3389/fmars.2017.00055.

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

Bracher, A., Vountas, M., Dinter, T., Burrows, J.P., Rottgers, R., Peeken, I., 2009. Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. Biogeosciences 6, 751–764. https://doi.org/10.5194/bg-6-751-2009.

Breiman, L., 2001. Random forests. Machine Learn. 45, 5–32. https://doi.org/10.1023/a:1010933404324.

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. 221, 1472–1483. https://doi.org/10.1016/j.ecolmodel.2010.02.014.

Busseni, G., Caputi, L., Piredda, R., Fremont, P., Mele, B.H., Campese, L., Scalco, E., de Vargas, C., Bowler, C., d'Ovidio, F., Zingone, A., d'Alcala, M.R., Iudicone, D., 2020. Large scale patterns of marine diatom richness: drivers and trends in a changing ocean. Glob. Ecol. Biogeogr. 29, 1915–1928. https://doi.org/10.1111/geb.13161.

Chen, T.Q., Guestrin, C., 2016. In: XGBoost: A Scalable Tree Boosting System. Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Dierssen, H.M., Ackleson, S.G., Joyce, K.E., Hestir, E.L., Castagna, A., Lavender, S., McManus, M.A., 2021. Living up to the hype of hyperspectral aquatic remote sensing: science, resources and outlook. Front.Environ.Sci. 9 https://doi.org/10.3389/fenvs.2021.649528.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–240. https://doi.org/10.1126/science.281.5374.237.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K., 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. U. S. A. 110, 9824–9829. https://doi.org/10.1073/pnas.1307701110.

Gruber, N., Clement, D., Carter, B.R., Feely, R.A., van Heuven, S., Hoppema, M., Ishii, M., Key, R.M., Kozyr, A., Lauvset, S.K., Lo Monaco, C., Mathis, J.T., Murata, A., Olsen, A., Perez, F.F., Sabine, C.L., Tanhua, T., Wanninkhof, R., 2019. The oceanic sink for anthropogenic CO2 from 1994 to 2007. Science 363, 1193–1199. https://doi.org/10.1126/science.aau5153.

Guidi, L., Guerra, A.F., Canchaya, C., Curry, E., Foglini, F., Irisson, J.O., Malde, K., Marshall, C.T., Obst, M., Ribeiro, R.P., 2020. Big data in marine science. European Marine Board. https://doi.org/10.5281/zenodo.3755793.

Henson, S.A., Cael, B.B., Allen, S.R., Dutkiewicz, S., 2021. Future phytoplankton diversity in a changing climate. Nat. Commun. 12, 5372. https://doi.org/10.1038/s41467-021-25699-w.

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.

Holder, C., Gnanadesikan, A., 2021a. Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? - a proof-of-concept study. Biogeosciences 18, 1941–1970. https://doi.org/10.5194/bg-18-1941-2021.

Holder, C., Gnanadesikan, A., 2021b. Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – a proof-of-concept study. Biogeosciences 18, 1941–1970. https://doi.org/10.5194/bg-18-1941-2021.

Hu, S.B., Liu, H.Z., Zhao, W.J., Shi, T.Z., Hu, Z.W., Li, Q.Q., Wu, G.F., 2018. Comparison of machine learning techniques in inferring phytoplankton size classes. Remote Sens. 10 https://doi.org/10.3390/rs10030191.

Huang, D.M., Zhao, D.F., Wei, L.F., Wang, Z.H., Du, Y.L., 2015. Modeling and analysis in marine Big Data: advances and challenges. Math. Probl. Eng. 2015 https://doi.org/10.1155/2015/384742.

Jackson, T., Sathyendranath, S., Melin, F., 2017. An improved optical classification scheme for the Ocean Colour Essential Climate variable and its applications. Remote Sens. Environ. 203, 152–161. https://doi.org/10.1016/j.rse.2017.03.036.

Ke, G.L., Meng, Q., Finley, T., Wang, T.F., Chen, W., Ma, W.D., Ye, Q.W., Liu, T.Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30 (Nips 2017), 30.

Keany, E., 2020. BorutaShap: A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values, 10.5281/zenodo.4247618.

Kramer, S.J., Siegel, D.A., 2019. How can phytoplankton pigments be best used to characterize Surface Ocean phytoplankton groups for ocean color remote sensing Algorithms? J. Geophys. Res. Oceans 124, 7557–7574. https://doi.org/10.1029/2019JC015604.

Kramer, S.J., Siegel, D.A., 2021. Global HPLC phytoplankton pigment data compilation, Version 2. In: PANGAEA.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta Package. J. Stat. Softw. 36, 1–13. https://doi.org/10.18637/jss.v036.i11.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324 doi: Doi.

Liu, Z., Wei, P., Jiang, J., Cao, W., Bian, J., Chang, Y., 2020. MESA: boost ensemble imbalanced learning with meta-sampler. Adv. Neural Inf. Proces. Syst. 33, 14463–14474. https://doi.org/10.48550/arXiv.2010.08830.

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

Lopez-Urrutia, A., Moran, X.A.G., 2015. Temperature affects the size-structure of phytoplankton communities in the ocean. Limnol. Oceanogr. 60, 733–738. https://doi.org/10.1002/lno.10049.

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 https://doi.org/10.3389/fmars.2017.00203.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inform. Process. Syst. 30 (Nips 2017), 30, 10.48550/arXiv.1705.07874.

Maddy, E.S., Boukabara, S.A., 2021. MIIDAPS-AI: an explainable machine-learning algorithm for infrared and microwave remote sensing and data assimilation preprocessing-application to LEO and GEO sensors. Ieee J.Select.Top.Appl.Earth Observ.Remote Sens. 14, 8566–8576. https://doi.org/10.1109/Jstars.2021.3104389.

Malek, S., Melgani, F., Bazi, Y., 2018. One-dimensional convolutional neural networks for spectroscopic signal regression. J. Chemom. 32 https://doi.org/10.1002/cem.2977.

Maranon, E., Cermeno, P., Latasa, M., Tadonleke, R.D., 2012. Temperature, resources, and phytoplankton size structure in the ocean. Limnol. Oceanogr. 57, 1266–1278. https://doi.org/10.4319/lo.2012.57.5.1266.

McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T., 2019. Making the Black Box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100, 2175–2199. https://doi.org/10.1175/Bams-D-18-0195.1.

Merchant, C.J., Embury, O., Bulgin, C.E., Block, T., Corlett, G.K., Fiedler, E., Good, S.A., Mittaz, J., Rayner, N.A., Berry, D., Eastwood, S., Taylor, M., Tsushima, Y., Waterfall, A., Wilson, R., Donlon, C., 2019. Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci.Data 6, 223. https://doi.org/10.1038/s41597-019-0236-x.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13 https://doi.org/10.1038/s41467-022-29838-9.

Moore, C.M., Mills, M.M., Arrigo, K.R., Berman-Frank, I., Bopp, L., Boyd, P.W., Galbraith, E.D., Geider, R.J., Guieu, C., Jaccard, S.L., Jickells, T.D., La Roche, J., Lenton, T.M., Mahowald, N.M., Maranon, E., Marinov, I., Moore, J.K., Nakatsuka, T., Oschlies, A., Saito, M.A., Thingstad, T.F., Tsuda, A., Ulloa, O., 2013. Processes and patterns of oceanic nutrient limitation. Nat. Geosci. 6, 701–710. https://doi.org/10.1038/Ngeo1765.

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.

Nair, A., Sathyendranath, S., Platt, T., Morales, J., Stuart, V., Forget, M.H., Devred, E., Bouman, H., 2008. Remote sensing of phytoplankton functional types. Remote Sens. Environ. 112, 3366–3375. https://doi.org/10.1016/j.rse.2008.01.021.

Núñez, J., Catalán, P.A., Valle, C., Zamora, N., Valderrama, A., 2022. Discriminating the occurrence of inundation in tsunami early warning with one-dimensional convolutional neural networks. Sci. Rep. 12, 1–20. https://doi.org/10.1038/s41598-022-13788-9.

Oelker, J., 2021. Suitability of atmospheric satellite sensors for ocean color applications. In: Universität Bremen. https://doi.org/10.26092/elib/1100.

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 10, 7553–7574. https://doi.org/10.5194/bg-10-7553-2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pena, M., van den Dool, H., 2008. Consolidation of multimodel forecasts by ridge regression: application to Pacific Sea surface temperature. J. Clim. 21, 6521–6538. https://doi.org/10.1175/2008jcli2226.1.

Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pelissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 https://doi.org/10.1038/s41467-020-18321-y.

Poloczanska, E.S., Brown, C.J., Sydeman, W.J., Kiessling, W., Schoeman, D.S., Moore, P. J., Brander, K., Bruno, J.F., Buckley, L.B., Burrows, M.T., Duarte, C.M., Halpern, B.S., Holding, J., Kappel, C.V., O'Connor, M.I., Pandolfi, J.M., Parmesan, C., Schwing, F., Thompson, S.A., Richardson, A.J., 2013. Global imprint of climate change on marine life. Nat. Clim. Chang. 3, 919–925. https://doi.org/10.1038/Nclimate1958.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inform. Process. Syst. 31 (Nips 2018), 31, 10.48550/arXiv.1706.09516.

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. 53, 605–613. https://doi.org/10.4319/lo.2008.53.2.0605.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., <check>Prabhat, check, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Rivero-Calle, S., Gnanadesikan, A., Del Castillo, C.E., Balch, W.M., Guikema, S.D., 2015. Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO2. Science 350, 1533–1537. https://doi.org/10.1126/science.aaa8026.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schroder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881.

Sadeghi, A., Dinter, T., Vountas, M., Taylor, B.B., Altenburg-Soppa, M., Peeken, I., Bracher, A., 2012. Improvement to the PhytoDOAS method for identification of coccolithophores using hyper-spectral satellite data. Ocean Sci. 8, 1055–1070. https://doi.org/10.5194/os-8-1055-2012.

Sathyendranath, S., Aiken, J., Alvain, S., Barlow, R., Bouman, H., Bracher, A., Brewin, R., Bricaud, A., Brown, C., Ciotti, A., 2014. In: Phytoplankton functional types from Space. (Reports of the International Ocean-Colour Coordinating Group (IOCCG) 15). International Ocean-Colour Coordinating Group, pp. 1–156, 10.25607/OBP-106.

Sathyendranath, S., Brewin, R.J.W., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Cipollini, P., Couto, A.B., Dingle, J., Doerffer, R., Donlon, C., Dowell, M., Farman, A., Grant, M., Groom, S., Horseman, A., Jackson, T., Krasemann, H., Lavender, S., Martinez-Vicente, V., Mazeran, C., Melin, F., Moore, T.S., Muller, D., Regner, P., Roy, S., Steele, C.J., Steinmetz, F., Swinton, J., Taberner, M., Thompson, A., Valente, A., Zuhlke, M., Brando, V.E., Feng, H., Feldman, G., Franz, B.A., Frouin, R., Gould, R.W., Hooker, S.B., Kahru, M., Kratzer, S., Mitchell, B.G., Muller-Karger, F.E., Sosik, H.M., Voss, K.J., Werdell, J., Platt, T., 2019. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (OC-CCI). Sensors (Basel) 19. https://doi.org/10.3390/s19194285.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. 55, 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6.

Stock, A., 2022. Spatiotemporal distribution of labeled data can bias the validation and selection of supervised learning algorithms: a marine remote sensing example. ISPRS J. Photogramm. Remote Sens. 187, 46–60. https://doi.org/10.1016/j.isprsjprs.2022.02.023.

Stock, A., Subramaniam, A., 2020. Accuracy of empirical satellite algorithms for mapping phytoplankton diagnostic pigments in the Open Ocean: a supervised learning perspective. Front. Mar. Sci. 7 https://doi.org/10.3389/fmars.2020.00599.

Stock, A., Subramaniam, A., 2022. Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing. Giscie.Remote Sens. 59, 1281–1300. https://doi.org/10.1080/15481603.2022.2107113.

Sun, X.R., Shen, F., Brewin, R.J.W., Li, M.Y., Zhu, Q., 2022. Light absorption spectra of naturally mixed phytoplankton assemblages for retrieval of phytoplankton group composition in coastal oceans. Limnol. Oceanogr. 67, 946–961. https://doi.org/10.1002/lno.12047.

Sun, X.R., Shen, F., Brewin, R.J.W., Liu, D.Y., Tang, R.G., 2019. Twenty-year variations in satellite-derived chlorophyll-a and phytoplankton size in the Bohai Sea and Yellow Sea. J. Geophys. Res. Oceans 124, 8887–8912. https://doi.org/10.1029/2019jc015552.

Team, A.-S.G., Mangin, A., d'Andon, O.F., 2017. GlobColour Product User Guide, GC-UM-ACR-PUG-01, Version 4.1.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240. https://doi.org/10.2307/143141.

Toms, B.A., Barnes, E.A., Ebert-Uphoff, I., 2020. Physically interpretable neural networks for the geosciences: applications to earth system variability. J.Adv.Model.Earth Syst. 12 https://doi.org/10.1029/2019MS002002.

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. Oceans 111. https://doi.org/10.1029/2005jc003207.

van der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The NumPy Array: a structure for efficient numerical computation. Comput.Sci.Eng. 13, 22–30. https://doi.org/10.1109/mcse.2011.37.

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. Oceans 106, 19939–19956. https://doi.org/10.1029/1999jc000308.

Ward, B.A., 2015. Temperature-correlated changes in phytoplankton community structure are restricted to polar waters. Plos One 10, e0135581. 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, 4833–4849. https://doi.org/10.1364/AO.53.004833.

Xi, H.Y., Losa, S.N., Mangin, A., Garnesson, P., Bretagnon, M., Demaria, J., Soppa, M.A., D'Andon, O.H.F., Bracher, A., 2021. Global chlorophyll a concentrations of phytoplankton functional types with detailed uncertainty assessment using Multisensor Ocean color and sea surface temperature satellite products. J. Geophys. Res. Oceans 126. https://doi.org/10.1029/2020JC017127.

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

Yang, Y.Z., Zha, K.W., Chen, Y.C., Wang, H., Katabi, D., 2021. In: Delving into deep imbalanced regression.International Conference on Machine Learning, 139, p. 139. https://doi.org/10.48550/arXiv.2102.09554.

Zhou, X.J., 2020. Application of deep learning in ocean Big Data mining. J. Coast. Res. 614–617 https://doi.org/10.2112/Si106-139.1.