A machine learning algorithm to retrieve the red peak of phytoplankton absorption spectra from ocean-colour remote sensing

Remote Sensing Applications: Society and Environment 39 (2025) 101702

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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

A machine learning algorithm to retrieve the red peak of phytoplankton absorption spectra from ocean-colour remote sensing

Mohammad Ashphaq, Shovonlal Roy *

Department of Geography and Environmental Science, University of Reading, Whiteknights, Reading, RG6 6DR, UK

ARTICLE INFO

Keywords:
Machine learning
Phytoplankton absorption
Remote sensing reflectance
Ensemble models
Meta stacking

ABSTRACT

Light absorption by microscopic phytoplankton in marine ecosystems is a crucial process underpinning biological production and global biogeochemical cycles. Accurate estimation of phytoplankton absorption coefficients, an inherent optical property of ocean water, can improve remote sensing applications including spectral photosynthesis models and assessments of ocean health, biodiversity, and climate change impacts. However, considerable uncertainty exists in current satellite retrievals of phytoplankton absorption coefficients, particularly for $a_{ph}(676)$ - the phytoplankton absorption peak at red wavelengths near 676 nm - which is an input to several novel and advanced satellite algorithms. This uncertainty hinders operational use of algorithms for assessing phytoplankton physiology, size structure and oceanic carbon pools from space. We aimed to improve satellite-based estimation of a_{ph} (676) using advanced machine learning (ML) techniques. We compiled a comprehensive *in situ* dataset (n = 1576) of $a_{ph}(676)$ from published databases and matched with remote-sensing reflectance *Rrs* at six wavelengths (412, 443, 490, 510, 560, and 665 nm) from the Ocean Colour Climate Change Initiative. We extensively evaluated multiple base ML algorithms: Random Forest (RF), Gradient Boosting Machines, and Linear Regression; and implemented ensemble ML models: RF with Grid Search Cross-Validation, eXtreme Gradient Boosting Ensembled Model, Ensemble Forecast, Stacked Voting, Optimised Ensemble and Meta Stacking, integrating the base models through cross-validated hyperparameter tuning. Meta Stacking outperformed individual ML models in predictive accuracy across temporal resolutions, showing best results with daily composites. Our study addresses key limitations of previous models, including small training datasets, inconsistent performances, and lack of ensemble comparisons. We present a robust, extensively trained and validated ensemble ML model that significantly improves $a_{ph}(676)$ estimation and opens the possibility of routinely using in ocean colour remote sensing.

1. Introduction

Phytoplankton are microscopic, photosynthetic organisms essential to the marine food web, producing over 50 % of Earth's oxygen and regulating atmospheric CO_2 levels through absorption, making them vital indicators of ocean health, ecosystem changes, and climate dynamics (IOCCG, 2000; Machado et al., 2023; Cetinić et al., 2024). Phytoplankton cells contain pigments, especially

* Corresponding author.
E-mail address: shovonlal.roy@reading.ac.uk (S. Roy).

https://doi.org/10.1016/j.rsase.2025.101702
Received 24 April 2025; Received in revised form 19 July 2025; Accepted 23 August 2025
Available online 25 August 2025

chlorophyll (Chlor-a) that absorb light at specific wavelengths (Ciotti et al., 2002; Cleveland, 1995), enabling Ocean Colour satellites to detect their concentrations in the ocean (Huan et al., 2021; Mouw et al., 2017; Wang et al., 2021). The absorption coefficient represents the amount of light harvested by per milligram of phytoplankton Chlor-a and is an important quantify for remote sensing applications such as retrieval of cell size, pigment composition and photosynthesis models (Bricaud and Stramski, 1990; Bricaud et al., 1995). Phytoplankton absorption coefficient is an important quantity for understanding oceanic health and the impacts of climate change on oceanic ecosystems, as they are directly linked to the primary productivity of the ocean (Marra et al., 2007; Hirawake et al., 2011; Barnes et al., 2014; Silsbe et al., 2016). Nutrient stress and co-limitation of mineral availability can affect phytoplankton growth and alter pigment composition, thereby influencing phytoplankton absorption properties and in turn primary productivity (Robinson et al., 2017). The seasonal and spatial variability of phytoplankton absorption is vital for monitoring of ocean health indicators such as chlorophyll concentration and occurrence of certain harmful algal blooms (Shen et al., 2012; Wei et al., 2023; Xu et al., 2025). As climate change continues to be a driver in altering phytoplankton composition and nutrient regimes, integrating absorption coefficients into biogeochemical models enhances the predictive ability of the models to monitor and manage changes in marine ecosystems, sustaining conservation and initiate climate mitigation strategies (Patara et al., 2012; Paulsen et al., 2018).

Phytoplankton absorption coefficients a_{ph} (λ) can be estimated from laboratory measurements, field observations, as well as remote sensing, and numerical modelling (Pahlevan et al., 2021). Satellite remote sensing is now widely used for cost-effective global coverage and retrieval of phytoplankton physical properties (Churilova et al., 2019; Ciotti et al., 2002; Roelke et al., 1999), phytoplankton functional types (PFT) (Anderson, 2005; De Moraes Rudorff and Kampel, 2012; Roy et al., 2013) and phytoplankton size structure (PSC) (Kostadinov et al., 2023; Pérez et al., 2021; Roy et al., 2017). The retrieval algorithms use remote sensing wavelengths in the blue and red spectrum, leveraging the absorption properties of Chlor-a. The visible spectral range from 300 nm to 800 nm is essential for acquisition of solar energy and its conversion into chemical energy through primary production in the ocean. A large portion of this photosynthetically active radiation is utilised by marine phytoplankton through light-absorbing pigments, primarily chlorophyll-a, which shows strong absorption with a primary peak in the blue region near 440 nm and a secondary peak in the red region near 676 nm (Kirk, 1994). Most ocean colour sensors, such as NASA's MODIS and ESA's Sentinel-3 OLCI, also operate in this spectral window with multispectral channels designed to capture water leaving radiances (IOCCG, 2012). The reflectance values captured by these channels are then used to derive inherent optical properties in the ocean such as absorption coefficients (IOCCG, 2006). The peak in the absorption spectra of Chlor-a, the primary pigment in phytoplankton, at the 443 nm wavelength has been widely used (Cao et al., 2005; Carder et al., 1999; Hirata et al., 2008; Shang et al., 2011; Wang et al., 2008; Zheng and Stramski, 2013a, 2013b; Zheng et al., 2015) to estimate phytoplankton biomass and distribution from remote sensing. On the other hand, the secondary peak of Chlor-a near 675–676 nm wavelength, within the red spectrum, has proven crucial for detecting phytoplankton physiological properties (Allali et al., 1997; Cullen et al., 1997; Li et al., 2021; Meler et al., 2017; Seppälä et al., 2005; Shang et al., 2021; Sun et al., 2010; Zhang et al., 2010), and in particular, for obtaining advanced information on phytoplankton cell size, size spectrum and carbon content (Roy et al., 2011, 2013, 2017).

Various methodologies, categorised in analytic, semi-analytic, empirical, semi-empirical, quasi-empirical, LUT have been developed and applied for the estimation of a_{ph} (λ) from remote sensing data for ocean colour applications (Blondeau-Patissier et al., 2014; Huan et al., 2021; Pahlevan et al., 2021). More recently, Machine Learning-based approaches have been implemented due to their potential for enhancing the accuracy of a_{ph}(λ) prediction (Ahmed et al., 2017; Alam et al., 2024; Deng et al., 2019; Pahlevan et al., 2021). For example, Huan et al. (2021) evaluated pigment absorption at 670 nm for Case I and Case II waters demonstrating a higher predictive accuracy for Case I waters. Pahlevan et al., (2021) used a small data set (40 paired of Rrs and a_{ph} measurements) with HICO overpasses and mixture density networks (MDNs) reporting inconsistent error percentages and biases, and weak relationships across wavelengths. Further, Alam et al. (2024) used optimised ensemble ML models to estimate a_{ph}(λ) values from Rrs with 674 samples for a_{ph} at the 670 nm. None of the previous studies, however, explicitly retrieved phytoplankton absorption peak in the red wavelengths (i.e., at ~676 nm), which is a specific input to the advanced ocean-colour algorithms of our concern, particularly those for phytoplankton size spectrum and allometry-based carbon and nutritional values (Roy, 2018; Roy et al., 2013, 2017).

In this study, we utilise the Ocean Colour Climate Change Initiate dataset, which provides high-quality, consistent time series data spanning over two decades, to explicitly retrieve phytoplankton absorption at the red peak. In doing so, we address notable gaps in ML techniques applied by previous studies, by considerably increasing the sample size for ML training, extensively evaluating the performance of multiple ML methods, and ensuring optimal performance of ML algorithms under randomised and bootstrapped conditions. Our study, compiling a comprehensive dataset comprising 1576 samples, and exploring multiple base ML algorithms and subsequently creating an amalgamated hybrid ensemble model presents a robust model for phytoplankton absorption peak at the red bands, which would be readily applicable to ocean colour algorithms.

2. Data

2.1. In-situ database on phytoplankton absorption spectra

To compile a global database of *in situ* measurements of phytoplankton absorption spectra, a systematic search was made on PANGAEA data archive with keywords "phytoplankton", "a_{ph}", "Rrs 676", "a_{ph} (676)", spanning all published a_{ph} (676) datasets covering the available cruise missions. The datasets considered were the SeaWiFS Bio-optical Archive and Storage System (SEABASS) through NASA bio-Optical Marine Algorithm Dataset (NOMAD) database (https://seabass.gsfc.nasa.gov/wiki/NOMAD, Werdell and Bailey, 2005), Marine Optical Buoy (MOBY) (Brown et al., 2007; Brown et al., 2007), BOUSSOLE (Carr et al., 2006; Carr et al., 2006), and the Ocean Colour Climate Change Initiative (OC-CCI) validation dataset, which combined the first three datasets into a single

extended version spanning 1997 to 2021 (Valente et al., 2022). We sorted the datasets according to the availability of a_{ph} peak identifiable at 676 nm, along with associated Inherent Optical Properties and Diffuse Attenuation Coefficients across various wavelengths. This collated dataset before filtering consisted of 7425 points, with a_{ph} values ranging from 290 to 849, at intervals of 0.2 nm (Fig. 1). After filtering, we retained a subset comprised n = 1576 valid entries of a_{ph} (676) and associated variables.

2.2. Satellite ocean colour data

The methodology, described in detail in the following section, is applicable to ocean colour data from any satellite sensor. We have, however, chosen to use the ESA Ocean Colour Climate Change Initiative (OC-CCI) data because of its wide use, which was derived by merging ocean colour data from multiple sensors that were active in different or overlapping time scale, e.g. SeaWiFS, MERIS, MODIS-A, and VIIRS. The merging methodology included band-shifting, bias correction for remote sensing reflectance (Rrs) and implementation of various chlorophyll algorithms (Sathyendranath et al., 2019). We used these merged products as a good compromise between data accuracy and length of the time series covering the *in-situ* dataset. We acquired the Daily, 5Day, 8Day, & Monthly products of Rrs_412, Rrs_443, Rrs_490, Rrs_510, Rrs_560, and Rrs_665, from OC-CCI Version 6.0, 4-km (available at https://www.oceancolour.org/portal/) through the Composite Browser, OPeNDAP, Web GIS Portal, and FTP. Different spatial and temporal resolutions are critical for resolving oceanographic processes across scales. High temporal-resolution data (e.g. daily, 5-day) capture short-lived, localised events such as phytoplankton blooms. Low-resolution observations (e.g. monthly) are suited for analysing long-term, regional or basin-scale climate impacts. Integrating multi-resolution data (from daily to monthly) enables a deeper understanding of marine biophysical variability and the predictive model's relative performance across oceanic events.

Utilising Python scripts, data extraction from the server was automated, catering to various frequencies ranging from daily to monthly intervals. Sequential extraction across different time intervals produced Rrs products at Daily (448 data points), 5-days (591 data points), 8-days (62 data points) and Monthly (452 data points). In total, the extracted Rrs data corresponded to1576 a_{ph} (676) measurements, which on filtering for Chlor-a, & kd_490 availability resulted in 1553 matchups. The data descriptive summary of data is presented in Table 1.

3. Methodology

The major steps of the methodology (outlined in Fig. 2 and further discussed) involve (a) extraction of remotely sensed variables such as remote sensing reflectance (Rrs) from OC-CCI data archive at specific wavelengths (412–665 nm), matching up the collated phytoplankton absorption at 676 nm i.e. a_{ph} (676); (b) training possible forms of ML algorithms including hybrid ensemble techniques to model a_{ph} (676) and (c) evaluating and optimising the performance of the trained ML models to take it forward for application. Multiple statistical metrices, commonly used in the literature, are implemented to assess the ML model performance and to identify the overall best-performing model. The chosen model is then applied to satellite-derived reflectance to produce the predicted a_{ph} (676) maps.

3.1. Evaluating base ML algorithms

The base ML models used in our study have been extensively exploited in previous studies concluding, non-linear models, significantly outperform linear models in predicting wave runup due to their ability to model complex coastal dynamics (Durap, 2023);

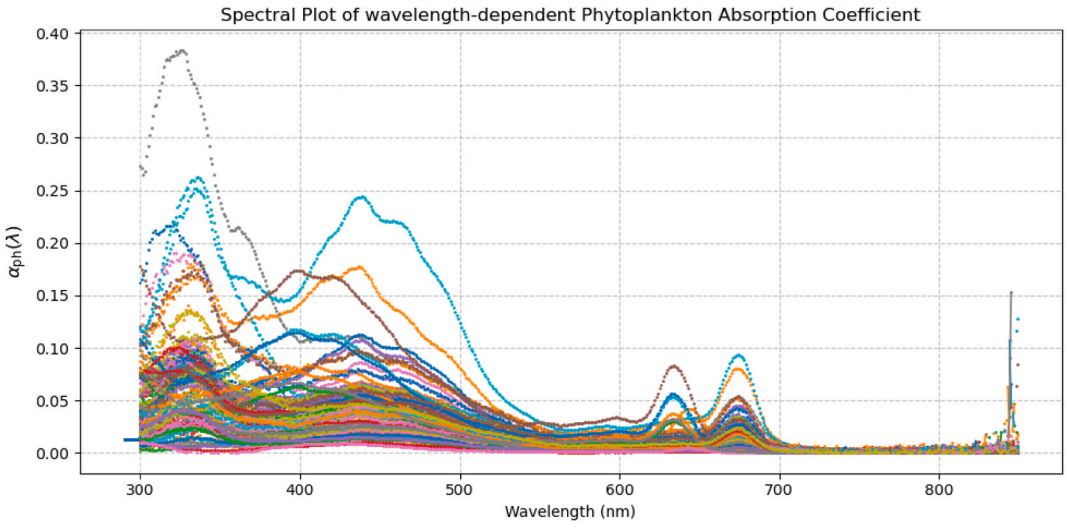


Fig. 1. Spectral plot of a_{ph} (λ) data from global dataset.

Table 1
Descriptive summary of *in situ* and satellite match-up data.

	a_{ph} (676)	Rrs_412	Rrs_443	Rrs_490	Rrs_510	Rrs_560	Rrs_665	Chlor-a	kd_490
count	1553	1553	1553	1553	1553	1553	1553	1553	1553
mean	0.0196	0.0072	0.0071	0.0077	0.0075	0.0062	0.0090	0.9100	0.1039
std	0.0209	0.0057	0.0060	0.0074	0.0075	0.0079	0.0382	1.5460	0.1763
min	0.0002	0.0005	0.0002	0.0008	0.0011	0.0000	0.0000	0.0005	−0.0668
25 %	0.0060	0.0032	0.0031	0.0031	0.0030	0.0016	0.0002	0.0770	0.0050
50 %	0.0127	0.0048	0.0044	0.0042	0.0035	0.0023	0.0004	0.4629	0.0675
75 %	0.0260	0.0096	0.0100	0.0106	0.0112	0.0092	0.0020	0.9421	0.1184
max	0.2283	0.0589	0.0785	0.0852	0.0737	0.0463	0.6742	16.8658	2.2537

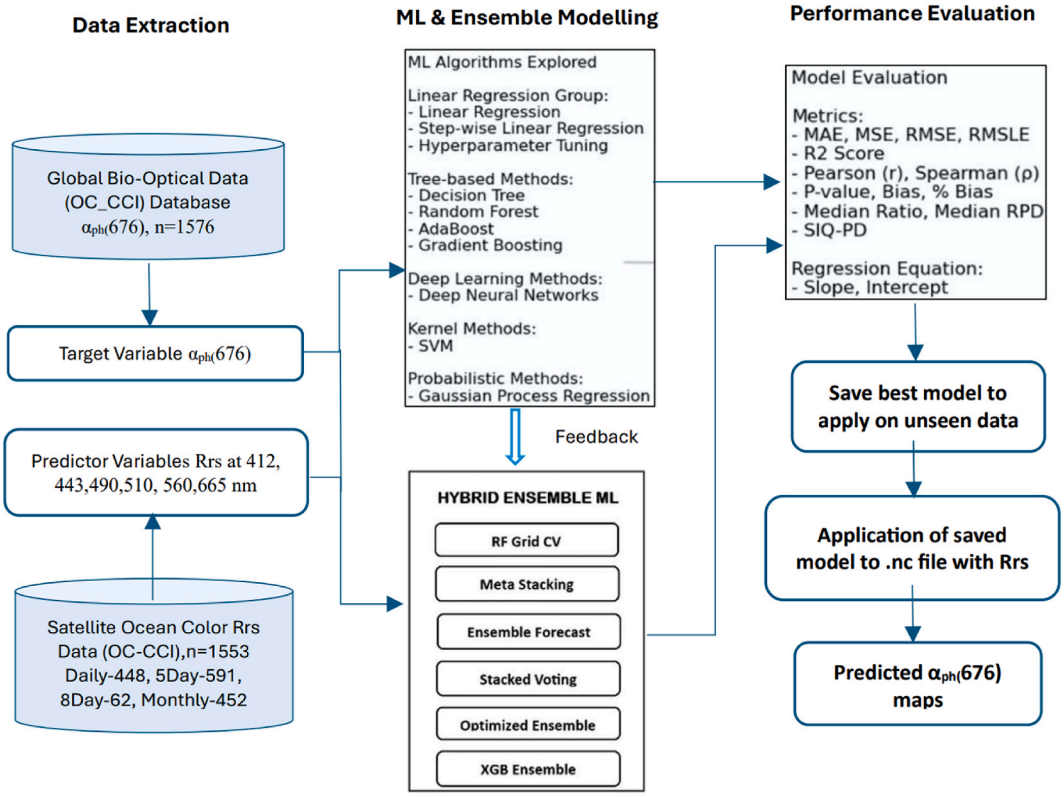


Fig. 2. Schematic diagram showing the major steps of our methodology involving data extraction, ML model development and evaluation of the model performance.

the Deep Learning models outperforms statistical models in predicting SST (Ali et al., 2021); the improved resolution and data inversion for enhanced carbon cycle analysis, nitrogen levels, and harmful algal blooms predictions based on classification techniques (Zhang et al., 2025); and also real-time applications like wave modelling and species distribution benefit from ML (Ahmad, 2019; Sadaiappan et al., 2023). Table 2 below summarises the advantages and limitations of ML techniques used as base models for advanced ensemble modelling. We first trained and validated a range of standard ML algorithms, described in the following, aimed at predicting a_{ph} values from Rrs values. Our methodology included adjustment of train-test ratios to discern the most effective model performance across different scenarios. The standard algorithms tested included (i) Linear Regression, (ii) Tree-based regression, (iii) Deep Learning method, (iv) Kernel Methods and (v) Probabilistic Methods.

In the *Linear Regression group*, *Linear Regression* stands as a foundational method for predicting continuous target variables, offering a straightforward approach to prediction. *Stepwise Linear Regression*, a variant, iteratively adjusts features based on their significance, refining the model for capturing relevant information. Additionally, *Linear Regression Hyperparameter Tuning* involves fine-tuning model parameters to optimise performance and enhance predictive accuracy. *Tree-based Methods group* includes *Decision Tree* employs a non-linear approach, partitioning data into subsets for predictions through binary decisions. *Random Forest* utilises ensemble techniques, constructing multiple decision trees and aggregating predictions to mitigate overfitting and enhance accuracy. *Ensemble methods*, *AdaBoost* and *Gradient Boosting*, combine weak learners, typically decision trees, to create robust predictive models, with Gradient Boosting sequentially improving upon errors. *Deep Neural Networks* utilise complex architectures with hidden layers, capable

Table 2
Advantages and limitations of ML techniques tested as base models.

Algorithm	Advantages	Limitations	Python	Sample Use Case
Linear Regression	Fast, baseline performance, may underfit complex data, Simple, interpretable, efficient for small linear datasets	Assumes linearity, sensitive to multicollinearity and outliers	scikit-learn, statsmodels	Baseline for temperature/ salinity modelling
Stepwise Linear Regression	Slightly better than basic linear regression, automatically selects significant variables	May overfit, computationally expensive for large data	statsmodels, custom code	Analysis of variable influence in ocean models
Linear Regression Hypertuning	Optimised parameters; accuracy via parameter tuning like regularisation	Still limited by linear assumptions	scikit-learn (Ridge, Lasso)	Improved linear models for salinity/temp trends
Decision Tree	Moderate, Easy to visualise, handles nonlinear data, no feature scaling	Overfits easily, unstable with small changes in data	scikit-learn	Marine habitat classification
Random Forest	High accuracy, robust, handles missing data, and nonlinearity, Reduces overfitting,	Less interpretable, slower training time	scikit-learn,	Species distribution modelling, SST prediction
AdaBoost	Boosts weak models, often outperforms single models	Sensitive to noisy data and outliers	scikit-learn	Marine species classification
Gradient Boosting	Very high accuracy, Captures complex patterns, tunable, supports regularisation	Computationally intensive, prone to overfitting if not tuned properly	xgboost, lightgbm, catboost,	Climate impact assessment, marine heatwave forecast
DNN (Deep Neural Network)	Handles high-dimensional and unstructured data well, high with enough data and tuning,	Requires large data, tuning, long training times	TensorFlow, Keras, PyTorch	Satellite image analysis, SST anomaly detection
SVM (Support Vector Machine)	Good with high-dimensional and smaller datasets with clear margins of separation	Not scalable to large datasets, hard to interpret	scikit-learn	classification, plankton data analysis
GPR (Gaussian Process Regression)	High for small datasets, very accurate, Provides uncertainty estimates	Very slow and memory-intensive for large datasets	scikit-learn, GPyTorch	Biogeochemical parameter modelling,

of learning intricate data patterns with high adaptability. Support *Vector Machine* is an established method for effectively classifying or regressing data by finding optimal hyperplanes. *Gaussian Process Regression* is a probabilistic method which employs Bayesian principles, modelling target variables as Gaussian processes and providing valuable uncertainty estimation in predictions (Ashphaq et al., 2024).

To ensure the ML model is robust and performs consistently under different sets of training and validation data, we used three train-test splits (50:50, 67:33, 80:20) to evaluate model performance. The model's performance and stability were assessed across the data splits, and the possibility of any overfitting or underfitting issues was investigated. We implemented a combination of 13 performance metrics, described below in Table 3, to test the algorithms' efficacy and performance. Sample sensitivity was analysed to verify the influence of changes in data splits on performance and generalisability. A log transformation was applied to normalise extreme values and reduce outlier impact on the dataset.

Table 3
Performance metrics used in this study.

Metric	Description	Desired Value	Units
Mean Absolute Error **(MAE)**	The average absolute difference between predicted and actual values.	Lower	Same as the target
Mean Squared Error **(MSE)**	The average of the squares of the errors between predicted and actual values.	Lower	Same as the target
Root Mean Squared Error **(RMSE)**	The square root of the MSE, representing the standard deviation of the residuals.	Lower	Same as the target
Root-Mean-Squared Log Error **(RMSLE)**	Similar to RMSE but calculated on the logarithm of the predicted and actual values. Useful for target variables with a large range.	Lower	Dimensionless (logarithmic scale)
R^2 Score	Measures proportion of the variance in the DV predictable from the IV	Higher (up to 1)	Dimensionless
Pearson Correlation Coefficient (r)	Measures the linear correlation between two variables, ranging from −1 to 1.	Close to 1 or -1	Dimensionless
Spearman's Correlation Coeff. (ρ)	Non-parametric measure of rank correlation, ranging from −1 to 1.	Close to 1 or -1	Dimensionless
P-value	The probability of observed results if the null hypothesis is true.	Lower (<0.05)	Dimensionless
Samples (n)	The number of samples used for testing.	Count	Dimensionless
Bias	The difference between the mean of predicted & mean of actual values.	Close to 0	Same as the target
% Bias	Bias expressed as a percentage of the mean of actual values.	Close to 0 %	Percentage
Median Ratio	The ratio of the median of predicted to the median of actual values.	Close to 1	Dimensionless
Median RPD (Relative Percent Difference)	The relative difference between predicted and actual values, as a % of the median of actual values.	Lower	Percentage
SIQ-PD (Scaled IQ Performance Descriptor)	Indicates model performance on a scaled IQ-like metric. Higher values suggest better performance.	Higher	Dimensionless (scaled score)

3.2. Hybrid ensemble modelling

The growing role of advanced machine learning approaches in environmental analysis, land-use changes, and thermal impacts have been demonstrated using the techniques like artificial neural networks (Zhang et al., 2021, 2024). These studies effectively model changes and forecast of land surface temperature, land use changes, and thermal impacts in urban settings by integrating spatial and temporal data to forecast urban heat patterns and carbon emissions (Zhang et al., 2023). Such predictive approaches are essential for understanding environmental dynamics to understand ocean health and climate change analogous to land-based systems. We then applied ensemble methods, which integrate multiple standard ML algorithms for enhancing the overall model performance. Ensemble methods synthesise predictions from multiple base models, such as decision trees, linear models, and others, to produce a final prediction that is typically more accurate and robust than that of any individual model. Analysis from previous steps highlighted the significant potential of machine learning models like Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR) in constructing a robust meta-learning framework based on ensemble techniques for the estimation of a_{ph} (676). In the following, we discuss various ensemble strategies adopted in our study: Meta Stacking, Ensemble Forecast, Stacked Voting, Optimised Ensemble, XGB Ensembled and hyper-tuned RF_Grid CV. These methods differ considerably in their composition, optimisation techniques, and applications.

3.2.1. Meta Stacking

Meta Stacking introduces complexity by leveraging a stacking ensemble approach. The base models RF and GBM are fine-tuned using Grid Search CV, and their predictions are combined by a meta-learner, a linear regression model. This method also includes bootstrapping with multiple iterations to ensure robust performance. The base models are individually optimised using GridSearchCV on resampled bootstrap samples of the training data. The optimised models generate predictions, which are aggregated into meta-features across multiple bootstrap iterations. After tuning, the base models are retrained on the entire training dataset with the best hyperparameters. For new data, predictions are made through a predict_stacked_model function, which first generates meta-features from the retrained base models and then applies the meta-model to produce the final prediction. The performance of this stacking ensemble is evaluated on a test set, providing a measure of accuracy as shown in flowchart Fig. 3. The stacking approach is ideal for complex tasks requiring high flexibility and advanced modelling techniques, albeit at the cost of higher training time and complexity.

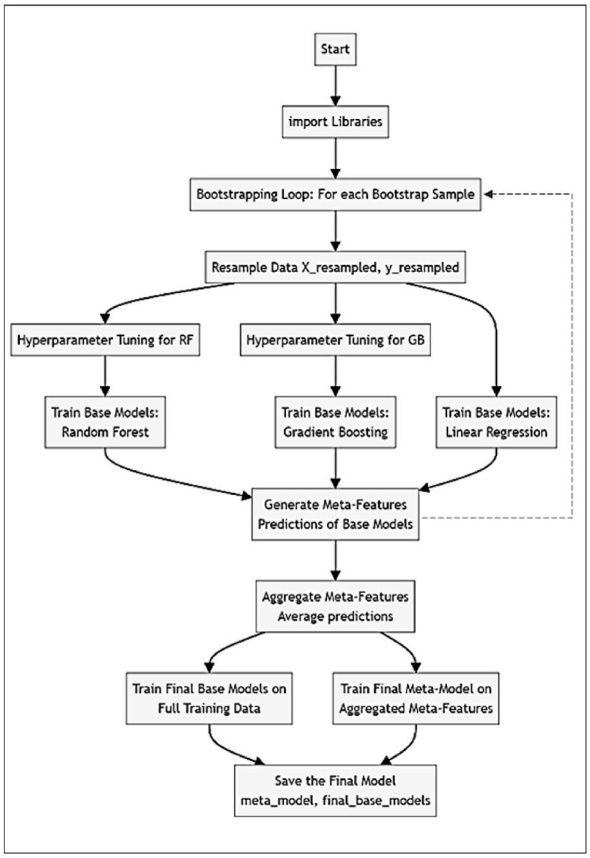


Fig. 3. Flowchart of meta stacking ensemble model.

3.2.2. Forecast ensemble

The Forecast Ensemble method integrates XGBoost, Gradient Boosting, Bagging, and Stacking, ensuring robustness and accuracy in predictions by averaging results across multiple iterations, making it ideal for demanding forecasting tasks despite high computational costs. The script demonstrates an advanced approach to create four forecasting function using Python: xgboost_forecast, gradient_boosting_forecast, bagging_forecast, and stacking_forecast. The first two functions leverage XGBoost and Gradient Boosting models, applying bootstrapping to produce reliable predictions. The bagging_forecast function combines outputs from both models through a bagging technique, averaging predictions to enhance accuracy. The stacking_forecast function further refines this by training a Linear Regression meta-model on the combined predictions of XGBoost and Gradient Boosting, potentially increasing forecasting accuracy. Hyperparameters such as n_estimators, control the number of boosting stages in both XGBoost and Gradient Boosting. The script emphasises model stability through bootstrapping and optimises accuracy using methods like grid search and cross-validation as shown in flowchart Fig. 4.

3.2.3. Stacked Voting Ensemble

The Stacked Voting Ensemble pipeline is an advanced system designed for optimised regression modeling, leveraging ensemble learning techniques like stacking and voting. It begins with selecting diverse base models—RF, GB, SVR, and XGB—each chosen for its

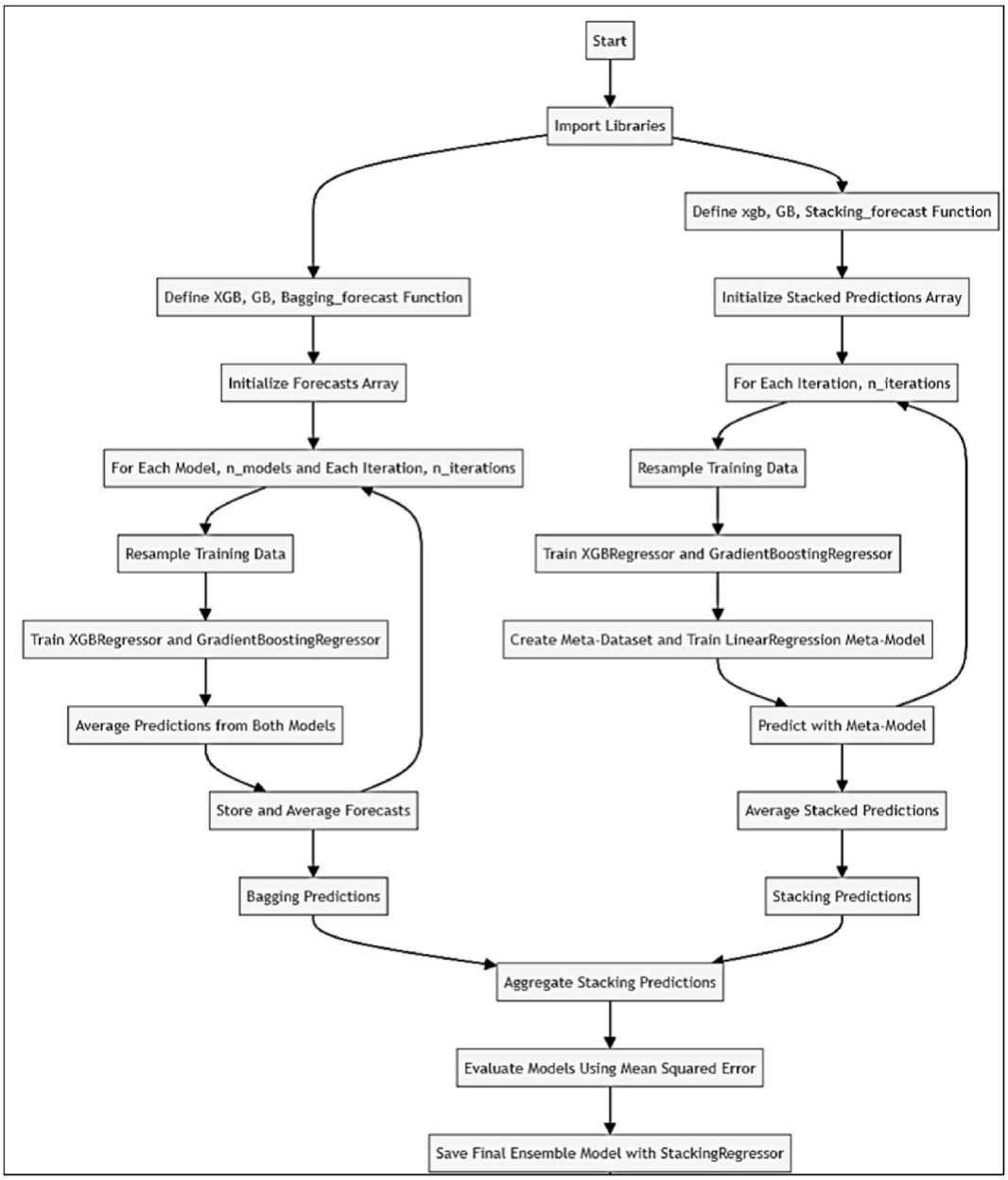


Fig. 4. Flowchart of forecast ensemble model.

ability to handle different aspects of the data. To maximise each model's performance, GridSearchCV is utilised for hyperparameter tuning, systematically exploring parameter combinations such as estimators, depth, learning rates, and regularisation. After identifying the best configurations, the models are combined in a Stacking Regressor framework, where their predictions are aggregated using a Voting Regressor as the meta-learner. This method enhances predictive accuracy by leveraging the strengths of the different models. Bootstrapping is applied to further improve robustness, reducing variance by averaging predictions from multiple resampled training sets. The final model's performance is assessed, ensuring it is both accurate and stable across various data subsets as shown in flowchart Fig. 5. This sophisticated pipeline, although computationally intensive, is designed to achieve high accuracy and reliability in regression tasks by combining model diversity and optimal tuning.

3.2.4. XGB ensemble

This ensemble technique utilises 'EnsembleWrapper' class which is a utility for managing collections of XGB models, facilitating their saving, loading, and prediction processes. Upon initialisation, it can be provided with lists of models and filenames or create empty lists if none are supplied. It saves models to files and metadata to a JSON file for future reference. For loading, it reads the metadata to reconstruct models and prepare them for predictions. During prediction, unseen data is converted to XGB's DMatrix format, predictions from each model are averaged, and the final output is evaluated using Root Mean Squared Error (RMSE).

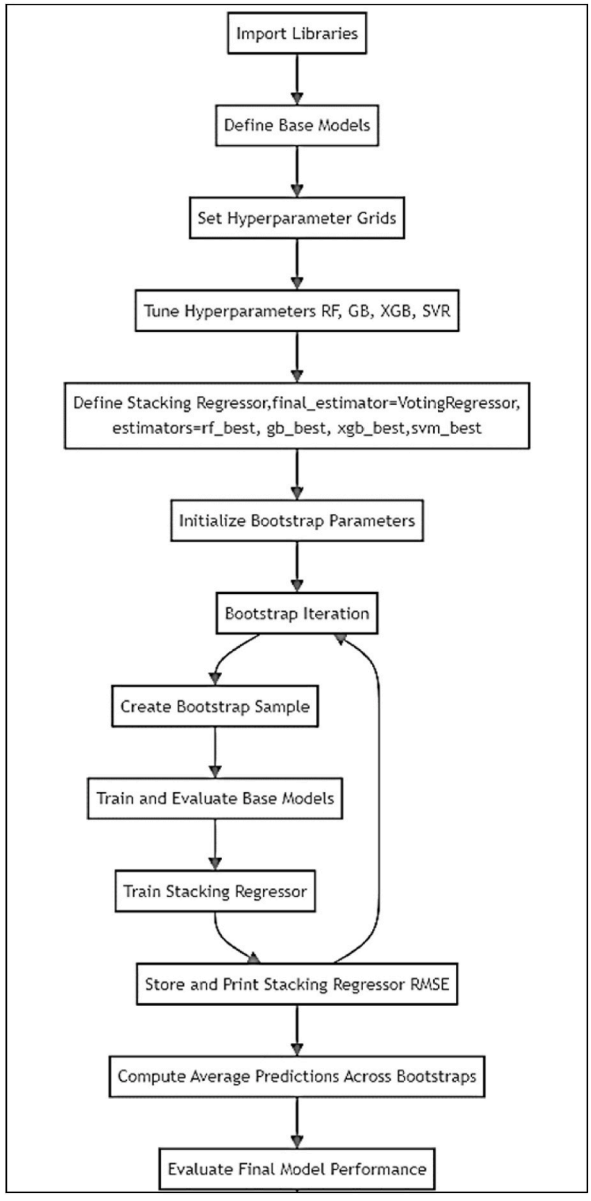


Fig. 5. Flowchart of stacked voting ensemble model.

Optimisation functions are critical for tuning XGB models' hyperparameters. The AHA Optimisation function adjusts the learning rate (eta) and L2 regularisation (lambda) iteratively, selecting the parameter set with the lowest RMSE. BWO Optimisation uses evolutionary strategies, starting with random parameter sets, and refines them through procreation and mutation, retaining the best-performing sets. AOA Optimisation employs arithmetic adjustments to fine-tune parameters based on a mathematical model, selecting the configuration with the best RMSE. FHO Optimisation simulates natural selection, refining parameters based on performance to select the optimal set. Initial hyperparameters for training include setting the objective to 'reg' for regression, 'eval_metric' to 'rmse', with an initial learning rate (eta) of 0.1, maximum tree depth of 6, subsample and colsample_bytree both at 0.8, and an L2 regularisation term (lambda) of 1.0. Optimisation functions adjust these parameters, and the refined parameters are used to train

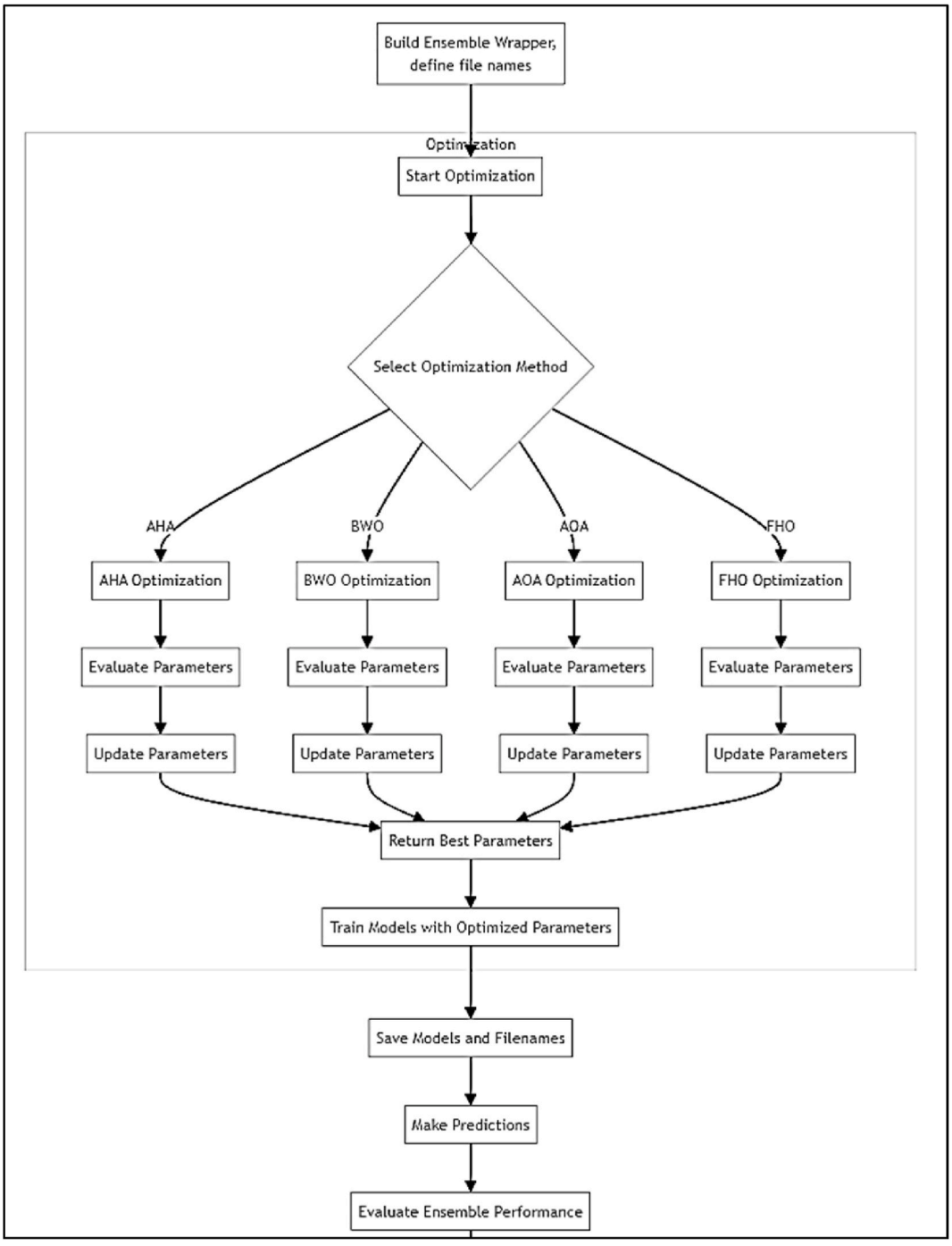


Fig. 6. Flowchart of xgb ensemble model.

models with the xgb.train() function. Models are then saved, and an EnsembleWrapper instance is created for managing these models. Predictions are made by the ensemble, averaged, and assessed for accuracy. This systematic workflow shown in flowchart Fig. 6 ensures effective model optimisation, management, and evaluation.

3.2.5. Optimised Ensemble

This ensemble begins with configuring and training individual models—Random Forest, Gradient Boosting, Extra Trees, AdaBoost, and Bagging, using predefined hyperparameters such as the number of estimators and a fixed random seed to ensure reproducibility. Hyperparameter tuning, though performed randomly in this case, involves adjusting parameters like the number of trees or boosting stages. This random tuning introduces variability into the performance metrics, enabling a diverse exploration of potential configurations and affecting the MSE scores. This process helps to evaluate how well each model generalises to unseen data, revealing their respective strengths and weaknesses. In addition to individual models, the code utilises a VotingRegressor to combine the predictions from the trained models. By averaging predictions from Random Forest, Gradient Boosting, Extra Trees, AdaBoost, and Bagging, the Voting Regressor aims to improve overall predictive accuracy and stability. The effectiveness of this ensemble method is evaluated through its MSE, demonstrating the advantages of aggregating multiple regression techniques. Finally, the code introduces an Optimised Ensemble, which integrates six base models—Random Forest (RF), Gradient Boosting Machine (GBM), Extra Trees Regressor (ETR), AdaBoost (ADA), Bagging, and Voting. This ensemble leverages stacking and voting mechanisms, along with hyperparameter optimisation using Grid Search CV, to enhance modelling flexibility and scalability (Alam et al., 2024). Despite its complexity and higher resource demands, this approach is designed for advanced tasks that benefit from the combined strengths of multiple algorithms.

3.2.6. RF_GridCV

This code is only focused on optimising a RandomForestRegressor model using GridSearchCV for hyperparameter tuning. Key hyperparameters include n_estimators (number of trees), max_features (features per split), max_depth (tree depth), min_samples_split (samples required to split a node), min_samples_leaf (samples per leaf), and bootstrap (sampling method, default enabled). GridSearchCV performs a comprehensive search across these hyperparameters using n-fold cross-validation to find the optimal configuration. After identifying the best parameters, the model is trained, and its performance is evaluated using Mean Squared Error (MSE) on a testing dataset. It's effective for moderate datasets due to its straightforward nature, balancing relatively high training time with quick prediction speed and good scalability.

3.3. Comparison to conventional models for estimating phytoplankton absorption at 676

We compared our ML model output with the following algorithms that have been previously used to retrieve phytoplankton absorption at 676 nm.

3.3.1. Carder et al., 1999

Developed by Carder et al. (1999), this semi analytical algorithm estimates water optical parameters using remote sensing reflectance (Rrs) values at 412 nm, 443 nm, 490 nm, and 560 nm. It calculates particulate and total backscattering coefficients, and absorption coefficients for phytoplankton and dissolved organic matter. Initial estimates for the parameters a675val (absorption at 675 nm) and adg400 (backscattering at 400 nm) are refined by solving nonlinear equations that relate observed Rrs ratios to model predictions. The function ultimately returns the estimated phytoplankton absorption coefficient at 675 nm, reflecting phytoplankton's contribution to water absorption.

3.3.2. Empirical method

An empirical method used by Roy et al. (2017) to estimate phytoplankton absorption at 676 from remote sensing data uses a simple formulation utilising phytoplankton absorption at 443 and 510. a_{ph} (676) is estimated as a product of a_{ph} (443) is raised to the power of 0.8478, and a_{ph} (510) is raised to the power of 0.2674.

3.4. Evaluation criteria for ensemble models

Evaluating machine learning (ML) approaches involves a comprehensive assessment of several essential and additional criteria to ensure effectiveness and applicability. Essential criteria include usability, which addresses the ease of implementation and user interface; applicability, ensuring the ML method aligns with the specific problem and data characteristics; and ease of application, reflecting the practical simplicity of deploying the model. Replicability is crucial for consistent performance, while time of execution evaluates the efficiency of training and prediction processes. Model diversity and ensemble integration are important for capturing various data aspects and improving performance through combined models. Additional factors encompass scalability, which examines the model's capacity to handle growing datasets without performance degradation; hyperparameter complexity, focusing on the ease of tuning; and interpretability, which is vital for understanding model predictions and making data-driven decisions. Data requirements and robustness assess the model's efficiency with different data volumes and its resilience to noise. Computational resource requirements consider the necessary hardware and software, and flexibility evaluates the model's adaptability to changing data and tasks. These criteria collectively provide a nuanced understanding of an ML approach's strengths, limitations, and suitability for specific applications.

3.5. Application of saved model (.pkl file) to raster data (.nc) file

The Python code is a developed to apply a saved ML model to generate predictions and visualise the results on a geographical map. The process involves importing critical libraries such as *xarray* for managing and processing multi-dimensional data arrays, *numpy* for numerical operations, *joblib* for loading pre-trained machine learning models, and *matplotlib.pyplot* for creating visualisations. The NetCDF file is accessed using xarray's open_dataset function, enabling convenient handling of the dataset. Six specific reflectance bands (Rrs_412, Rrs_443, Rrs_490, Rrs_510, Rrs_560, and Rrs_665) are identified and extracted from the dataset, representing reflectance at different wavelengths. These bands are combined into a single 3D NumPy array, where the third dimension corresponds to the different wavelengths. The data is then reshaped from the 3D array into a 2D format where each row represents a pixel, and each column represents a wavelength band. A mask is created to filter out invalid data points, removing rows with Nan, zero, or negative values, ensuring that only valid data is used for further analysis. The script loads a pre-trained stacking model from a *.joblib* file, which is applied to valid data points to predict the a_{ph} (676) parameter. These predictions are mapped back onto the original geographical grid, and a new array is initialised to store the predicted values, initially filled with NaN to indicate missing data. The predicted data is then visualised using matplotlib.pyplot, where the predictions are plotted on a geographical map with a color-coded scale representing a_{ph} (676) values.

4. Results and discussion

4.1. Performance of the base ML algorithms

The base ML models tested are listed in Table 4, which include Linear Regression, Stepwise Linear Regression, Linear Regression Hyper tuning, Decision Tree, Random Forest, AdaBoost, Gradient Boost, Deep Neural Network (DNN), Support Vector Machine (SVM), Gaussian Process Regression (GPR). Each ML model is evaluated using independent sets training and testing data, drawn randomly from the *in-situ* datasets with three different training: testing split ratios: 50:50, 67:33, and 80:20, and their corresponding performances are analysed using the validation metrics (Table 4).

Intercomparison of the algorithms' performance (Table 4) shows that the Linear Regression model, although produces a generally consistent performance across different training: testing splits, resulting in relatively low MAE (0.23), MSE (0.09), and RMSE (0.29), its predictive power is R^2 Score ranging from 0.52 to 0.55. The Stepwise Linear Regression model demonstrates performance like that of the Linear Regression. The decision Tree, AdaBoost, Gradient Boost, DNN, SVM, and GPR models exhibit moderate performance with metrics comparable to each other. The Random Forest (RF) model, however, outperforms all other models with the lowest MAE (0.21), MSE (0.076), and RMSE (0.09), and the higher R^2 score (Table 4).

4.2. Performance of hybrid ensemble models

We have attempted to retrieve all match ups corresponding to the daily data. But, to maximize the number of match ups, when there were gaps in the daily data, we used the overlapping 5-day, 8-day and monthly satellite images. It should be noted that the accuracy of the daily matchups should be higher than any of the three temporal resolutions. As a compromise between the sample size and resolution we have merged all the matchups into the final validation dataset. We have further tested each of the ML model's performance across all the subsets (i.e., daily, 5-day, 8-day, monthly and merged) of the validation dataset, which are described below.

For daily predictions, Meta Stacking demonstrates the highest R^2 value (0.702), the best slope of regression (0.78), high correlation coefficients (Pearson r 0.84503, and Spearman's ρ 0.81101), and relatively low RMSE (0.2414), indicating high predictive accuracy and strong validation performance (Table 4). Forecast model and Optimised Ensemble show comparable R^2 values but slightly lower slopes of regression (0.71 and 0.70, respectively), suggesting their slightly inferior performance, compared to Meta Stacking. Carder method shows the highest RMSE (1.957) and lowest slope (0.01) among all the models, indicating its significantly inferior performance compared with the ensemble ML models (see Fig. 7).

The performances are generally consistent across all other temporal scales. For example, in 5-day predictions, Meta Stacking again stands out with the highest R^2 (0.47), lowest RMSE (0.3391) and comparable slope of regression, suggesting its better performance across other metrics compared to its peers. However, in this case Optimised Ensemble and XGB Ensembled models' performance is also strong with relatively high R^2 values and good fit metrics, placing them in the just below Meta Stacking. Carder model sustains it poor performance, but the empirical model shows better slope with high RMSE, suggesting less prediction accuracy (Table 5).

For the 8-day, the performance of XGB Ensembled is closely comparable with Meta stacking indicated by R^2 (0.725 and 0.7247), RMSE (0.248 and 0.2266) and slope (0.70 and 0.71), with a minor edge for Meta stacking due to RMSE and slope (Table 5). However, in the monthly predictions, XGB shows better performance over the other models with lower RMSE (0.329) and higher R^2 (0.591), although the slope is slightly lower than Meta stacking (0.67 vs 0.71).

When combining all four temporal datasets, Meta Stacking shows superior performance with the highest R^2 (0.5849), highest slope (0.57) and the lowest RMSE (0.3003), suggesting it better overall performance across the metrics, and outperforming all other models. Across all temporal scales, Meta Stacking performs significantly better than the Empirical algorithm and Carder et al. (1999) algorithm (Fig. 6).

Overall, across all temporal scales, Meta Stacking stands out with superior variance explanation, linear and rank correlation, accuracy (Table 5). RF_Grid CV and Optimised Ensemble are notably strong with robust predictive accuracy. Forecast Ensemble also performs generally well balancing accuracy and correlation. In contrast, Stacked Voting Ensemble and XGB Ensembled show higher

Table-4
The summary of results for all executed algorithm for Daily matchup with log-transformed data (n-448).

METHOD	TRAIN: TEST	MAE	MSE	RMSE	RMSLE	R^2	Pearson r	Spearman (ρ)	P	(n)	Bias	% Bias	Median Ratio	Median RPD	SIQ-PD	Reg Equation
LINEAR REGRESSION	50:50	0.2367	0.0899	0.2998	0.1052	0.5266	0.7257	0.76397	0	224	0.0809	4.3186	1.05421	10.76449	18.2657	y = 0.93x + 0.06
	67:33	0.2327	0.0921	0.3035	0.10725	0.5456	0.73866	0.79797	0	148	0.9261	4.9262	1.06205	9.88157	19.3098	y = 0.98x + −0.06
	80:20	0.2301	0.0865	0.2942	0.10348	0.5549	0.74494	0.76898	0	90	0.0639	3.35238	1.05444	9.63197	16.9819	y = 0.96x + 0.02
STEP-WISE LINEAR REGRESSION	50:50	0.2409	0.0916	0.3026	0.10706	0.5177	0.71953	0.75032	0	224	0.0769	4.10101	1.06248	10.88031	18.5965	y = 0.94x + 0.04
	67:33	0.2652	0.1522	0.3902	0.12309	0.2488	0.49887	0.77403	0	148	0.1147	6.10419	1.05895	11.26833	22.4904	y = 0.66x + 0.56
	80:20	0.2257	0.0865	0.2941	0.10404	0.5553	0.74521	0.76562	0	90	0.0556	2.91392	1.04338	9.19655	17.0964	y = 0.98x −0.03
LINEAR REGRESSION HYPERTUNING	50:50	0.2367	0.0899	0.2998	0.10523	0.5266	0.7257	0.76397	0	224	0.0809	4.3186	1.05421	10.76449	18.2657	y = 0.93x + 0.06
	67:33	0.2327	0.0921	0.3035	0.10725	0.5456	0.73866	0.79797	0	148	0.0926	4.9262	1.06205	9.88157	19.3098	y = 0.98x −0.06
	80:20	0.2301	0.0865	0.2942	0.10348	0.5549	0.74494	0.76898	0	90	0.0639	3.35238	1.05444	9.63197	16.9819	y = 0.96x + 0.02
DECISION TREE	50:50	0.2776	0.1381	0.3716	0.12719	0.2728	0.52236	0.64951	0	224	0.0294	1.56779	1.01623	11.32348	21.2703	y = 0.64x + 0.66
	67:33	0.3054	0.1635	0.4044	0.13937	0.1933	0.43971	0.58871	0	148	0.0822	4.37382	1.04772	12.63165	24.9663	y = 0.62x + 0.67
	80:20	0.2660	0.1396	0.3736	0.1221	0.2825	0.53151	0.67446	0	90	0.0384	2.01432	1.02686	10.45904	19.9361	y = 0.63x + 0.67
RANDOM FOREST	50:50	0.2298	0.0881	0.2969	0.10402	0.5359	0.73205	0.76741	0	224	0.0641	3.42126	1.04641	9.64474	18.7048	y = 0.92x + 0.10
	67:33	0.2127	0.0768	0.2771	0.09868	**0.6210**	0.78807	0.79771	0	148	0.070	3.75121	1.05101	8.98946	18.06746	y = 1.02x −0.10
	80:20	0.2285	0.0920	0.3034	0.10311	0.5268	0.72582	0.75284	0	90	0.0643	3.37249	1.05554	10.77715	16.8982	y = 0.85x + 0.23
ADABOOST	50:50	0.2719	0.1121	0.3348	0.11855	0.4097	0.64014	0.73011	0	224	0.1102	5.87852	1.07859	12.84391	21.9427	y = 1.01x −0.13
	67:33	0.2722	0.1096	0.3310	0.11919	0.4593	0.67777	0.75971	0	148	0.1121	5.96417	1.0912	12.15978	22.5196	y = 1.17x −0.45
	80:20	0.2817	0.1164	0.3412	0.11978	0.4014	0.63358	0.71186	0	90	0.1108	5.80976	1.09613	12.68709	21.1081	y = 1.01x −0.13
GRADIENT BOOST	50:50	0.2401	0.0922	0.3037	0.10595	0.5141	0.71703	0.7447	0	224	0.0651	3.47467	1.04641	10.937	18.57462	y = 0.88x + 0.17

(*continued on next page*)

Table-4 (*continued*)

METHOD	TRAIN: TEST	MAE	MSE	RMSE	RMSLE	R^2	Pearson r	Spearman (ρ)	P	(n)	Bias	% Bias	Median Ratio	Median RPD	SIQ-PD	Reg Equation
	67:33	0.2266	0.0836	0.2892	0.10254	0.5873	0.76641	0.772	0	148	0.0475	2.5312	1.04617	10.751	17.86402	y = 1.00x - 0.05
	80:20	0.2211	0.0829	0.2879	0.09866	0.5737	0.75747	0.76246	0	90	0.0474	2.48585	1.05451	10.079	15.91661	y = 0.90x + 0.14
DNN	50:50	0.2465	0.1270	0.3564	0.13475	0.3311	0.57541	0.72533	0	224	0.0401	2.13901	1.04349	10.266	19.88814	y = 0.73x + 0.49
	67:33	0.2538	0.1477	0.3844	0.14803	0.2711	0.52073	0.71722	0	148	0.0296	1.57767	1.03895	10.342	21.65022	y = 0.69x + 0.56
	80:20	0.2409	0.0940	0.3067	0.10637	0.5164	0.71863	0.73047	0	90	0.0367	1.92633	1.04987	10.507	17.1662	y = 0.92x + 0.12
SVM	50:50	0.2239	0.0862	0.2937	0.10269	0.5456	0.7387	0.76291	0	224	0.0401	2.1427	1.0408	9.9131	17.41604	y = 0.95x + 0.05
	67:33	0.2190	0.0858	0.2930	0.10326	0.5764	0.75924	0.78418	0	148	0.0378	2.01325	1.0489	9.3935	18.16648	y = 1.07x −0.16
	80:20	0.2382	0.0976	0.3124	0.10685	0.4981	0.70583	0.70997	0	90	0.0145	0.76028	1.04134	10.515	16.76476	y = 0.91x + 0.16
GPR	50:50	1.6346	3.2007	1.7890	0.97929	0.0026	0.05178	0.0474	0.48	224	−1.611	−85.952	0	100	92.6825	y = 0.03x + 1.87
	67:33	1.6316	3.2395	1.7998	0.97927	0.0013	−0.0364	−0.0587	0.47	148	−1.596	−84.930	0	100	92.0269	y = −0.02x+ 1.89
	80:20	1.0750	4.9394	2.2224	nan	0.0091	0.0957	0.3329	0.0	90	−0.181	−9.4970	1.01588	20.561	124.633	y = 0.02x + 1.88

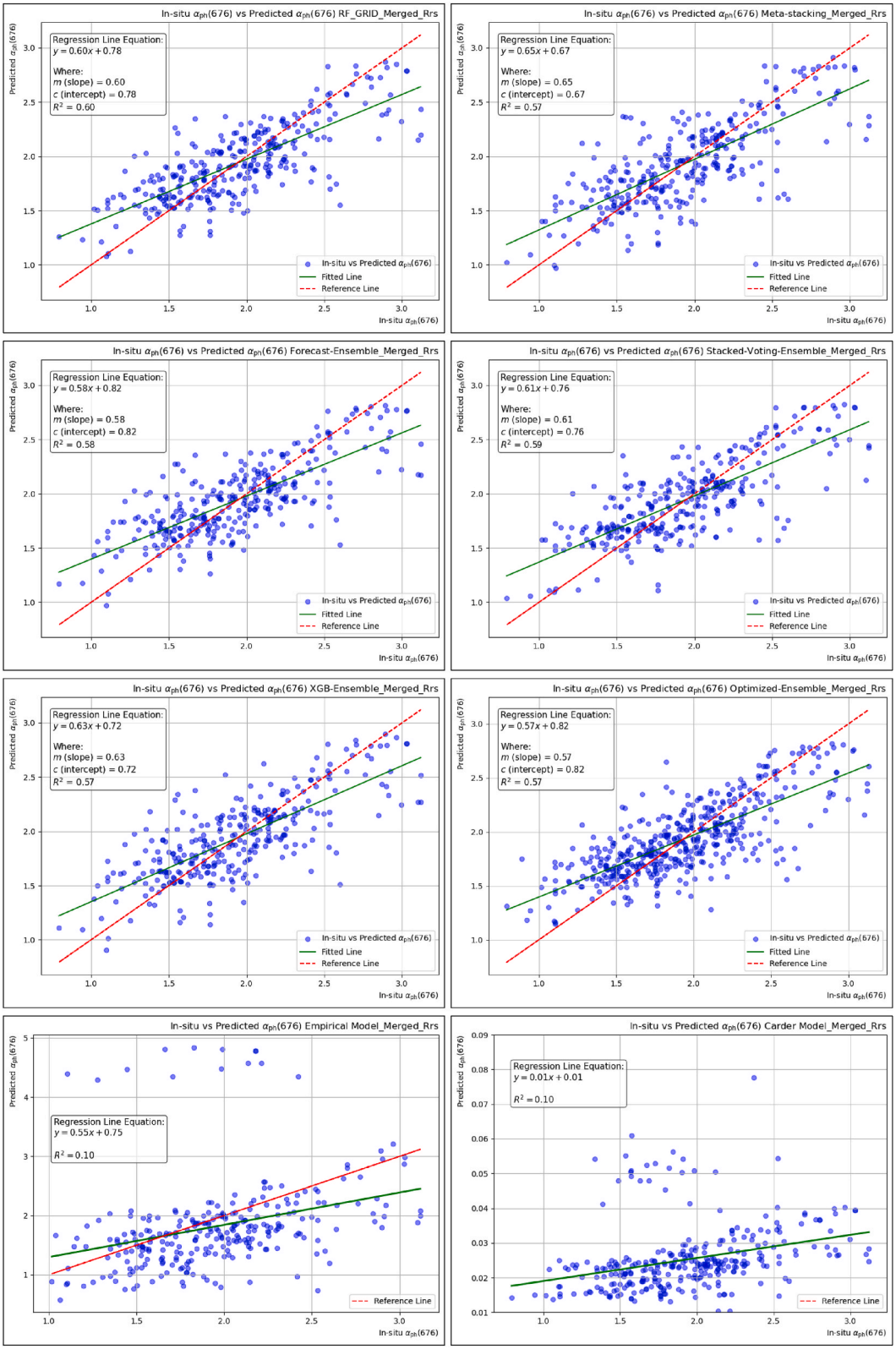


Fig. 7. Scatterplots and regression results for ensemble models validation with the merged Rrs data. Results are shown for RF-Grid, Meta stacking, forecast ensemble, Stacked-voting ensemble, XGB ensemble and Optimised ensemble, along with empirical model and Carder models.

Table-5
The summary of results for all Ensemble Model for all match-up.

ML/Ensemble Method	Rank	MAE	MSE	RMSE	RMSLE	R2	Pearson r	Spearman's (ρ)	p	n	Bias	% Bias	Median Ratio	Median RPD	SIQ-PD	Regression Equation
Daily_Rrs (OC-CCI)																
RF_Grid CV		0.179	0.059	0.2433	0.084	0.69	0.831	0.799	0	90	−0.008	−0.412	1.009	7.063	12.962	y = 0.71x + 0.54
Meta_Stacking	1	0.1818	0.0583	0.2414	0.0829	0.702	0.8379	0.8099	0	90	0.0056	0.2904	0.9947	6.8749	13.154	y = 0.78x + 0.43
Forecast Ensemble	2	0.184	0.05531	0.235	0.0817	0.701	0.842	0.824	0	90	−0.0003	−0.016	1.022	7.701	12.75	y = 0.71x + 0.56
Stacked-Voting Ensemble		0.197	0.0631	0.251	0.0871	0.672	0.82	0.787	0	90	−0.0053	−0.277	1.019	8.589	13.578	y = 0.73x + 0.53
XGB Ensembled		0.194	0.0635	0.252	0.0861	0.676	0.822	0.797	0	90	−0.0056	−0.288	1.011	7.958	13.219	y = 0.76x + 0.46
Optimised Ensemble	3	0.18	0.0557	0.236	0.0818	0.70	0.841	0.821	0	90	0.002	0.104	1.017	7.029	13.08	y = 0.70x + 0.59
Empirical Model (SR)		0.411	0.258	0.508	0.193	0.391	0.626	0.561	0	88	−0.316	−16.468	0.824	20.092	25.946	y = 0.69x + 0.29
Carder et al., 1999 **Model**		1.909	3.8306	1.957	1.0518	0.425	0.652	0.616	0	90	−1.9088	−98.788	0.012	98.775	98.766	y = 0.01x + 0.01
5Days_Rrs (OC-CCI)																
RF_Grid CV		0.24	0.12	0.3465	0.12	0.447	0.668	0.636	0	119	−0.0322	−1.617	1.006	7.172	23.319	y = 0.47x + 1.02
Meta_Stacking	1	0.2374	0.115	0.3391	0.1179	**0.47**	0.6856	0.6567	0	119	−0.0338	−1.7013	1.0072	7.8798	23.1747	y = 0.46x + 1.04
Forecast Ensemble		0.245	0.12766	0.357	0.1226	0.409	0.639	0.648	0	119	−0.021	−1.054	1.009	8.061	25.188	y = 0.43x + 1.11
Stacked-Voting Ensemble		0.249	0.1285	0.358	0.124	0.413	0.642	0.63	0	119	−0.0284	−1.429	0.996	8.583	24.595	y = 0.47x + 1.03
XGB Ensembled		0.233	0.1222	0.35	0.1212	0.445	0.667	0.664	0	119	−0.0337	−1.692	1	7.483	24.544	y = 0.51x + 0.94
Optimised Ensemble	2	0.238	0.121	0.348	0.1206	0.441	0.664	0.657	0	119	−0.0162	−0.815	1.015	6.869	24.404	y = 0.48x + 1.02
Empirical Model (SR)		0.66735	1.12177	1.05914	0.28843	0.08626	0.2937	0.46111	0	114	0.1394	7.13442	0.93994	18.542	55.82783	y = 0.74x + 0.64
Carder et al., 1999 **Model**		1.963	4.0652	2.016	1.0682	0.088	0.296	0.377	0	119	−1.9629	−98.676	0.013	98.666	98.624	y = 0.01x + 0.01
8Days_Rrs (OC-CCI)																
RF_Grid CV		0.22	0.076	0.2751	0.094	0.636	0.797	0.751	0.0031	13	−0.1116	−5.595	0.99	8.588	13.766	y = 0.61x + 0.67
Meta_Stacking	1	0.1775	0.0514	0.2266	0.0791	0.7247	0.8513	0.8232	0.0005	13	−0.0606	−3.0379	0.9738	6.728	12.167	y = 0.71x + 0.51
Forecast Ensemble		0.223	0.06806	0.261	0.0895	0.666	0.816	0.79	0.0013	13	−0.0919	−4.609	0.946	10.824	13.505	y = 0.58x + 0.74
Stacked-Voting Ensemble		0.218	0.0759	0.275	0.0913	0.632	0.795	0.794	0.0012	13	−0.104	−5.214	0.97	10.716	12.995	y = 0.56x + 0.77

(*continued on next page*)

Table-5 (*continued*)

ML/Ensemble Method	Rank	MAE	MSE	RMSE	RMSLE	R2	Pearson r	Spearman's (ρ)	p	n	Bias	% Bias	Median Ratio	Median RPD	SIQ-PD	Regression Equation
XGB Ensembled	2	0.192	0.0613	0.248	0.0841	0.725	0.852	0.713	0.0063	13	−0.117	−5.869	0.978	8.581	11.811	y = 0.70x + 0.47
Optimised Ensemble		0.204	0.068	0.261	0.0879	0.689	0.83	0.691	0.009	13	−0.1172	−5.878	0.989	8.472	12.388	y = 0.65x + 0.58
Empirical Model (SR)		0.257	0.1592	0.399	0.1789	0.526	0.725	0.771	0.0055	11	−0.1939	−10.24	0.943	7.539	22.697	y = 0.99x + −0.18
Carder et al., 1999 **Model**		1.969	4.0445	2.011	1.0703	0.668	0.817	0.845	0.0003	13	−1.9687	−98.722	0.013	98.672	98.72	y = 0.01x + −0.00
Monthly_Rrs (OC-CCI)																
RF_Grid CV		0.268	0.114	0.3374	0.118	0.559	0.748	0.731	0	91	0.0087	0.442	1.011	11.823	20.172	y = 0.65x + 0.70
Meta_Stacking		0.2726	0.1204	0.3469	0.12	0.5598	0.7482	0.7096	0	91	0.0046	0.2319	0.9991	10.7459	20.0012	y = 0.71x + 0.57
Forecast Ensemble		0.266	0.11243	0.335	0.1181	0.555	0.745	0.709	0	91	0.0009	0.046	1.009	11.623	20.374	y = 0.61x + 0.76
Stacked-Voting Ensemble		0.267	0.1137	0.337	0.1185	0.548	0.74	0.721	0	91	−0.0067	−0.339	0.988	12.383	20.602	y = 0.60x + 0.79
XGB Ensembled	1	0.255	0.1081	0.329	0.116	0.581	0.762	0.723	0	91	0.0206	1.046	1.011	10.419	19.745	y = 0.67x + 0.67
Optimised Ensemble	2	0.264	0.1081	0.329	0.1159	0.57	0.755	0.729	0	91	0.0049	0.248	1.005	11.573	20.124	y = 0.62x + 0.76
Empirical Model (SR)		0.462	0.2954	0.544	0.1831	0.071	0.266	0.266	0.0308	66	−0.3014	−14.213	0.818	19.94	26.65	y = 0.14x + 1.53
Carder et al., 1999 **Model**		1.951	4.0542	2.013	1.067	0.363	0.603	0.597	0	91	−1.9511	−98.876	0.011	98.93	98.822	y = 0.00x + 0.01
Merged_Rrs (OC-CCI)																
RF_Grid CV		0.227	0.094	0.3065	0.108	0.567	0.753	0.73	0	313	−0.0148	−0.751	1.004	8.979	19.809	y = 0.56x + 0.84
Meta_Stacking	1	0.2196	0.0902	0.3003	0.1057	**0.5849**	0.7648	0.7367	0	313	−0.0141	−0.7169	0.9984	8.3846	19.5087	y = 0.57x + 0.83
Forecast Ensemble	3	0.22512	0.09199	0.30329	0.10644	0.578	0.75979	0.7401	0	313	−0.01825	−0.92737	1.00802	8.43004	19.73532	y = 0.56x + 0.84
Stacked-Voting Ensemble		0.231	0.0969	0.311	0.1098	0.554	0.744	0.716	0	313	−0.0147	−0.745	0.999	9.443	20.197	y = 0.55x + 0.87
XGB Ensembled	2	0.223	0.0929	0.305	0.1067	0.577	0.76	0.737	0	313	−0.0178	−0.906	1.004	8.387	19.26	y = 0.63x + 0.71
Optimised Ensemble		0.222	0.0915	0.303	0.1066	0.578	0.76	0.735	0	313	−0.0017	−0.087	1.01	8.067	19.98	y = 0.56x + 0.87
Empirical Model (SR)		0.52186	0.61606	0.78489	0.23442	0.10028	0.31668	0.48563	0	279	−0.12187	−6.15136	0.86921	19.46401	40.9166	y = 0.57x + 0.74
Carder et al., 1999 **Model**		1.944	3.9937	1.998	1.0632	0.176	0.42	0.513	0	313	−1.9442	−98.768	0.012	98.77	98.726	y = 0.01x + 0.01

error metrics, suggesting generally less accurate predictions. All ensemble ML models reliability in predictions is generally higher than the Empirical algorithm and Carder et al. (1999) algorithms (Table 5).

4.3. Synthesis of the models' structure and performance

In evaluation of ensemble models, Meta Stacking and Optimised Ensemble stand out for their sophisticated ensemble methodology. While Meta Stacking employs a pure stacking approach, Optimised Ensemble combines stacking with voting to improve performance (Table 6). Meta_Stacking demonstrates the most consistent and superior performance across temporal data of Rrs, achieving the lowest MAE (0.1775–0.2374) and highest R^2 (up to 0.725), particularly in 8-day and 5-day datasets. It's architecture (Table 6) building on stacking Random Forest, Gradient Boosting Machine, and Logistic Regression with Grid CV optimisation and bootstrapping, yields strong predictive power despite higher complexity and moderate prediction time.

Both RF_Grid CV and Meta Stacking utilise Grid search cross validation (Grid CV) for robust optimisation, whereas Optimised Ensemble and XGB Ensembled incorporate advanced techniques like metaheuristic optimisation (AHA, BWO, AOA, and FHO). All models use hyperparameter tuning, however, Meta Stacking, Stacked Voting, and Optimised Ensemble explicitly leverage this process, which is crucial for improving performance. Additionally, Meta Stacking, Ensemble Forecast, Stacked Voting, and Optimised Ensemble models incorporate bootstrapping to enhance model robustness. XGB Ensembled, performs better in monthly Rrs with MAE = 0.255, RMSE = 0.329, and R^2 = 0.581, and also shows strong performance in 8-day (MAE = 0.192, R^2 = 0.725), highlighting the value of precision-tuned XGB frameworks. The Optimised Ensemble, with its broader model diversity and Grid CV, maintains top three MAE across all resolutions, peaking in daily (MAE = 0.18, R^2 = 0.70) and merged Rrs (MAE = 0.222, R^2 = 0.578), demonstrating robust generalisability. Among ensemble methods, Meta Stacking and Stacked Voting are particularly effective, though Meta Stacking's pure stacking approach may offer a slight edge. RF_Grid CV, a non-ensemble baseline, has moderate complexity, making it easier to manage compared to the higher complexity of other models. However, RF_Grid CV shows moderate metrics, with daily R^2 = 0.69, 5-day MAE = 0.24, and monthly MAE = 0.268.

In terms of scalability, all ensemble models perform moderately, with Meta Stacking and Optimised Ensemble having a slight edge. For our application, Meta Stacking and Optimised Ensemble would be preferable, while RF_Grid CV remains a further possibility for this application. Overall, Meta Stacking combined with the base models and effective optimisation methodology emerges as the top performer due to its relatively better validation metrics (e.g. R^2, Pearson r, and Spearman's ρ).

4.4. Applications: spatial maps of a_{ph} (676) and uncertainty

We have applied the Meta Stacking algorithm to raster data from the OC-CCI archive to generate spatial maps of a_{ph} (676). These maps represent two seasons (Fig. 8), January and August in 2023, showing the seasonal variation in spatial distribution of a_{ph} (676). In the Northern Hemisphere, above 40°, a_{ph} (676) values are noticeably higher in August compared to January, reflecting increased concentration of chlorophyll during the summer months. Conversely, in the Southern Ocean, a_{ph} (676) values are higher in January, corresponding with the austral summer, and indicating elevated concentrations of phytoplankton. These observed patterns are consistent with the seasonal phytoplankton bloom dynamics, such as diatom blooms, in southern hemisphere and northern hemisphere, suggesting the algorithm's ability to capture seasonal trends of phytoplankton absorption from remote sensing.

The uncertainty levels in algorithm prediction vary spatially, depending on the input *Rrs* values. A geographical residual plot of the training and test data (Fig. 9) suggests that, except at very latitudes in both the Northern and Southern Hemisphere, residuals generally remain below 35 %, indicating a reasonable level of prediction uncertainty. It is noteworthy that, due to unavailability of *in situ* data,

Table 6
Comparison of evaluation criteria for ensemble models.

Criteria	RF_Grid CV	Meta Stacking	Ensemble Forecast	Stacked Voting	Optimised Ensemble	XGB Ensembled
Model Type	RF	Stacking Ensemble	XGB, GB, Bagging, Stacking	Stacking + Voting	Stacking + Voting	XGB Ensemble
Base Models	RF	RF, GBM, LR	XGB, GB, Bagging, Stacking	RF, GBM, XGB, SVR, Voting	RF, GBM, ETR, ADA, Bagging, Voting	XGB Models
Optimisation	Grid CV	Grid CV	Bootstrapping	Grid CV	Grid CV	AHA, BWO, AOA, FHO
Hyperparameter Tuning	Yes	Yes	Implicit	Yes	Yes	Yes
Bootstrapping	Yes	Yes	Yes	Yes	Yes	No
Ensemble Method	NO	Yes (Stacking)	XGB, GB, Bagging, Stacking	Yes (Stacking + Voting)	Yes (Stacking + Voting)	Yes
Training Time	High	High	Moderate	High	High	High
Prediction Time	Fast	Moderate	Moderate	Moderate	Moderate	Moderate
Complexity	Moderate	High	High	High	High	High
Scalability	Good	Moderate to Good	Moderate to Good	Moderate to Good	Moderate to Good	Moderate to Good
Application	Versatile	Complex Tasks	Robust Performance	Advanced Modelling	Advanced Modelling	High Accuracy Tasks

the residuals could not be estimated across all oceanic regimes.

5. Conclusions

The primary objective this study was to enhance the predictive accuracy of remote sensing estimates of phytoplankton absorption peak at the red band, i.e., $a_{ph}(676)$, which is a critical input for several remote sensing algorithms used to retrieve phytoplankton size classes, as well as carbon and nutritional content (Roy et al., 2013, 2017; Roy, 2018). We presented a new machine learning (ML) algorithm using ocean colour satellite data from OC-CCI, developed through extensive training and validation of various ML model formulations. To obtain a robust ML model, we adopted a rigorous approach by compiling a comprehensive *in situ* training dataset of $a_{ph}(676)$ and matched it with remote-sensing reflectance at six wavelengths in the visible range. We then extensively evaluated a range of base ML algorithms, e.g., Random Forest (RF), Gradient Boosting Machines, and Linear Regression; and further implemented advanced ensemble ML models such as RF with Grid Search Cross-Validation, eXtreme Gradient Boosting Ensembled Model, Ensemble Forecast, Stacked Voting, Optimised Ensemble, and Meta Stacking, by integrating the base models. The best-performing model was identified by evaluating its performance against the large *in situ* $a_{ph}(676)$ database compiled in this study.

Our evaluation demonstrated that Meta Stacking ensemble learning was the most effective algorithm in terms of predictive

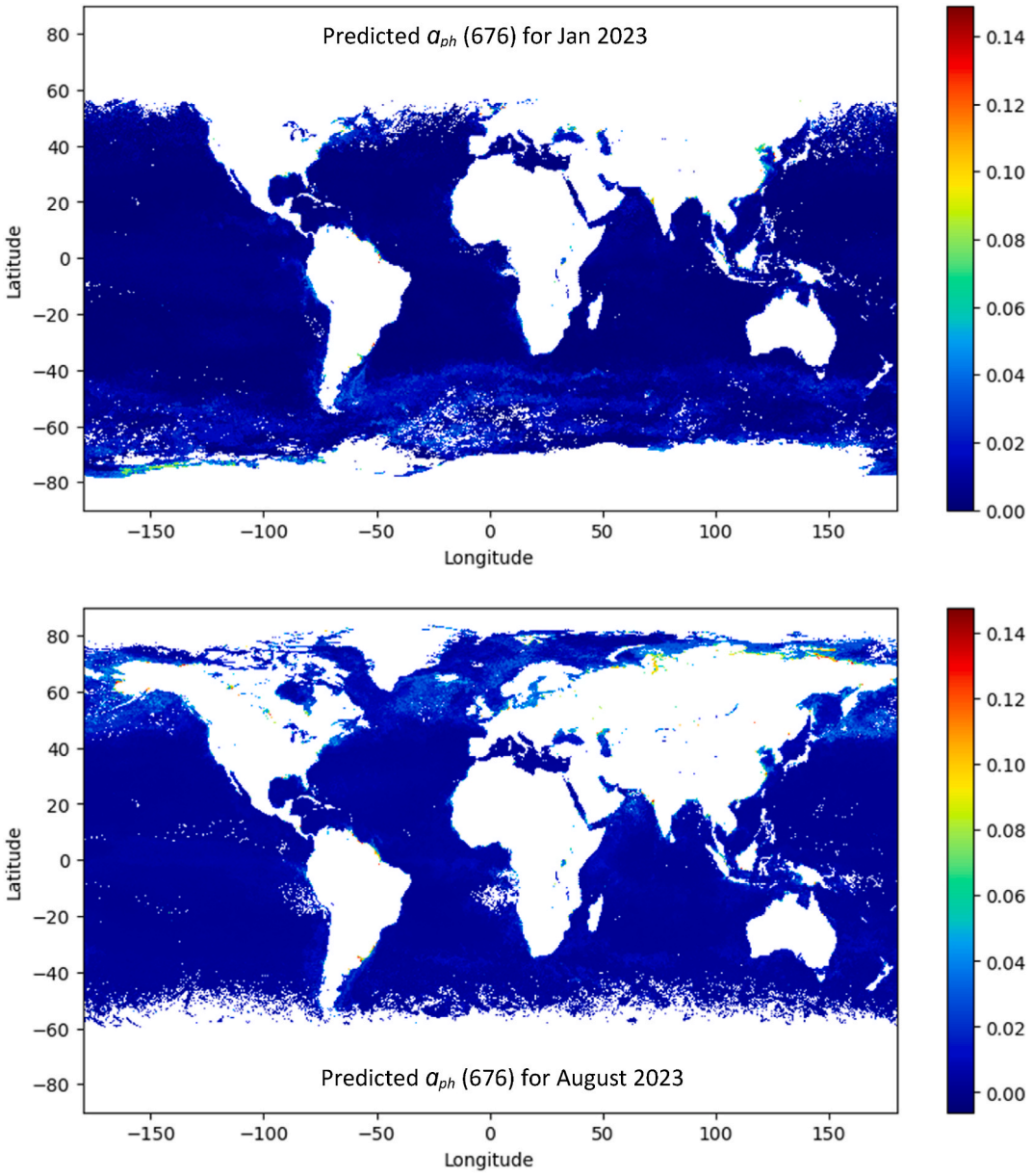


Fig. 8. Application of the model to raster data from OC-CCI, resulting in a_{ph} (676) prediction maps.

accuracy and ability to perform well with various temporal data resolutions of ocean colour data. Our analysis suggests that the choice of ML model and temporal resolution of satellite data are crucial for accurately estimating phytoplankton absorption from satellite remote sensing. Meta Stacking as an algorithm may be particularly effective for $a_{ph}(676)$ prediction due to its robust combination of diverse base models and optimisation techniques, especially when paired with daily data from satellites for higher accuracy.

Our study addresses key limitations identified in the literature on estimating $a_{ph}(676)$ from *Rrs* by developing an ensemble machine learning model. We addressed the challenges such as small sample sizes of the training $a_{ph}(676)$ data, inconsistent error percentages in the previously developed ML models for phytoplankton absorption, weak relationships across wavelengths, lack of baseline performance comparisons, and absence of evaluations comparing ensemble methods (e.g., Alam et al., 2024; Pahlevan et al., 2021). By compiling an extensive *in situ* $a_{ph}(676)$ dataset, the largest till date, and implementing more advanced ML techniques such as hyper parameter tuning, our study ensures the robustness and generalisability of the developed ML model. Furthermore, we conducted comparative evaluation of different ML algorithms evaluations through baseline performance metrics to identify the most effective approach for estimating $a_{ph}(676)$ values. Our study thus confronts prevailing limitations in estimating $a_{ph}(676)$ values by systematically optimising ensemble machine learning models.

The ML model performance across satellite matchups obtained on various temporal resolutions (daily, 5-day, 8-day, monthly, and merged Rrs datasets) indicates that finer temporal granularity improves the predictive accuracy of the model. Consistent with our understanding, ML models trained on higher-resolution inputs (daily, 8-day) yielded lower errors and higher R^2 values, reflecting better apprehension of short-term variability and seasonal patterns in ocean-colour biogeochemical properties. In contrast, coarser resolutions (e.g. monthly) increased uncertainty, leading to underfitting and reduced model responsiveness. By tackling the challenge of obtaining longitudinal ocean colour satellite data, our research advances remote sensing application of phytoplankton absorption for wider ecological research. Future studies may focus on exploring multi-resolution training, temporal embeddings, and dynamic ensemble weighting to enhance robustness and generalisation of our ML model across datasets.

The ML model that we have developed for retrieving phytoplankton absorption can potentially support policy-relevant studies by enhancing the accuracy of satellite-derived advanced biogeochemical products. By improving the estimates of a key inherent optical variable i.e., $a_{ph}(676)$, our model can help advance the accuracy and reliability of satellite retrieval algorithms for large-scale environmental assessments critical for ecosystem management and policy. More specifically, improved satellite-based estimation of $a_{ph}(676)$ using our approach can refine advanced algorithms for deriving phytoplankton size classes and phytoplankton carbon, for which $a_{ph}(676)$ is the key input (e.g., Roy et al., 2013, 2017; Roy, 2018). Accurate estimation of phytoplankton carbon from space is particularly important because it serves as a key component of oceanic and aquatic carbon budgets (Falkowski et al., 1998; Field et al., 1998). These estimates are increasingly sought after by the global scientific and policy community for better quantifying carbon fluxes and stocks in marine ecosystems (CEOS, 2014). So, our ML model outputs can contribute to global carbon and climate models and can inform climate change assessments and mitigation strategies, such as those presented in IPCC reports (Calvin et al., 2023).

Despite our efforts, the available *in situ* dataset of $a_{ph}(676)$ that we have compiled and used for training the ML model, is mainly restricted to the Atlantic Ocean and parts of the Pacific Ocean, leaving several major oceanic regimes underrepresented, for example, the Indian Ocean, Southern Ocean, and much of the Pacific. Any future sampling efforts, expanding *in situ* observations to encompass these diverse and ecologically distinct regions will be crucial for further improving the robustness and spatial coverage of the training data. A more globally distributed training dataset would enhance the generalisability of the ML model and may reduce regional biases. If an expanded dataset becomes available in the future, re-training our proposed ML model will be necessary to incorporate the new data and improve the overall accuracy of the model prediction. Furthermore, for specific applications at regional scales such as monitoring harmful algal blooms or assessing phytoplankton community structure in coastal zones, it may be useful to develop regionally trained ML models. These localised ML models could better capture unique bio-optical characteristics and ecological

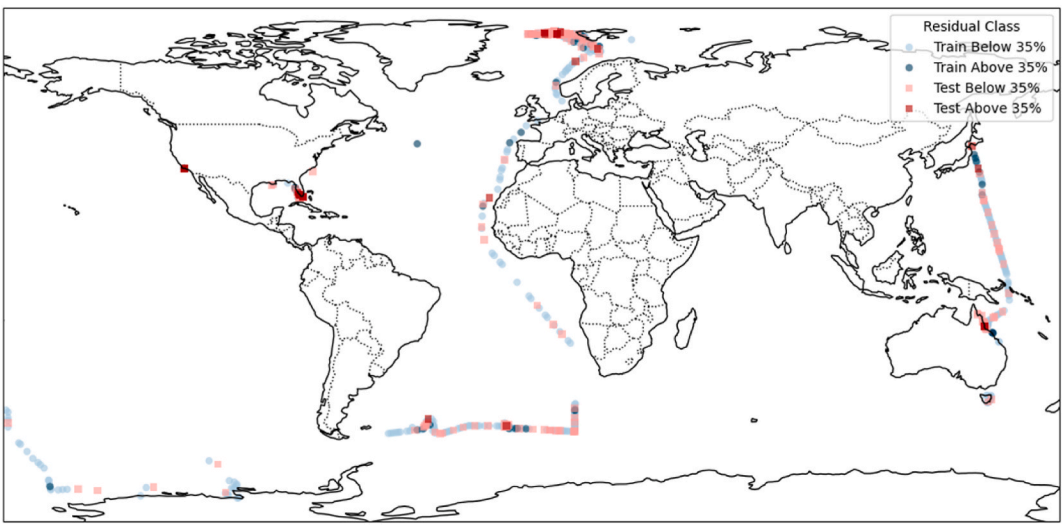


Fig. 9. A geographical map of predicted residuals for a_{ph} (676), shown at the locations of available *in situ* data.

dynamics and potentially improve the reliability and relevance of satellite-based predictions of phytoplankton absorption in management or conservation efforts.

CRediT authorship contribution statement

Mohammad Ashphaq: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Shovonlal Roy:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Funding acquisition, Conceptualization.

Ethical statement

We declare that:

- all authors have agreed to submit this manuscript to Remote Sensing Applications: Society and Environment
- this work is not under consideration for publication elsewhere and it will not be submitted elsewhere before a final decision is made.
- the written work is entirely original, and any work and/or text from others has been appropriately cited or quoted.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used AI-assisted technologies in order to get help with language corrections. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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.

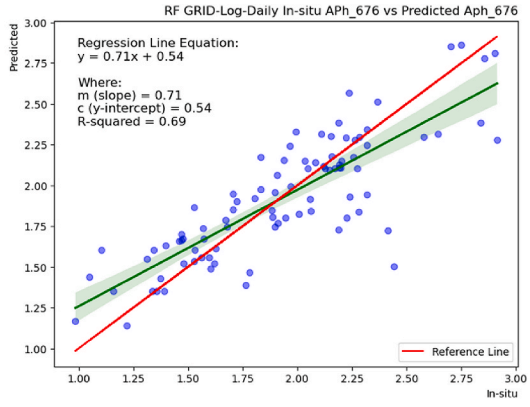
Acknowledgement

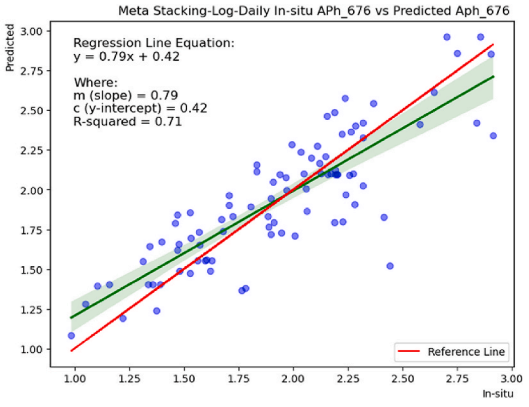
This research was funded by a Research Endowment Trust Fund (RETF) grant from the University of Reading. We thank the investigators who collected the *in situ* data used in this study for making their datasets freely available through various data repositories. We also acknowledge the Ocean Colour Climate Change Initiative (OC-CCI) team for providing free access to the satellite data via their online portal.

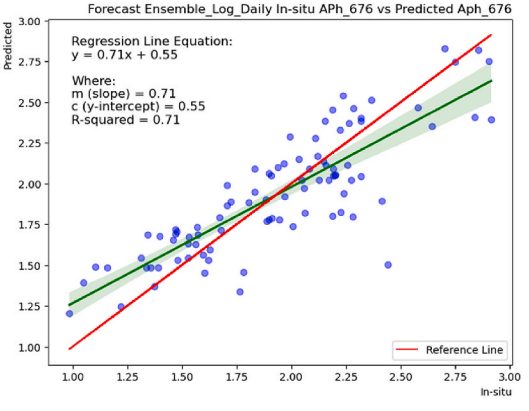
Appendix-1. Abbreviations Used in This Study

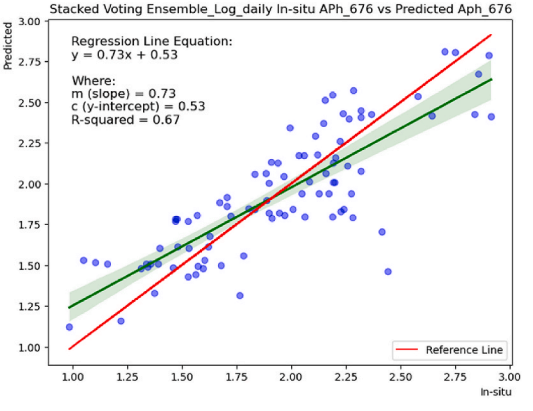
Abbreviation	Meaning
AHA	Adaptive Hyperparameter Algorithm (used for optimising hyperparameters)
BWO	Black Widow Optimisation (a metaheuristic optimisation algorithm)
AOA	Artificial Owl Algorithm (an optimisation algorithm based on owl behaviour)
FHO	Firefly Optimisation (an optimisation algorithm inspired by firefly behaviour)
RMSE	Root Mean Squared Error (a metric for evaluating model performance)
XGBoost	Extreme Gradient Boosting (a machine learning algorithm for regression and classification)
SVR	Support Vector Regression (a regression algorithm based on Support Vector Machines)
RF	Random Forest (a type of ensemble learning method using multiple decision trees)
GB	Gradient Boosting (a boosting algorithm that builds models sequentially)
XGB	XGBoost (an optimised version of Gradient Boosting)
DMatrix	A data structure used by XGBoost for optimised training and prediction
CV	Cross-Validation (a technique for assessing model performance by splitting data into training and validation sets)
GridSearchCV	A method for hyperparameter tuning that performs an exhaustive search over specified parameter values
SVR	Support Vector Regression (a regression technique using support vector machines)
n_jobs	Number of CPU cores to use during computation (in contexts like GridSearchCV)
Bootstrap	A statistical method for resampling with replacement to estimate the distribution of a statistic
Model	A trained machine learning algorithm used for making predictions based on input data
Ensemble	A method combining multiple models to improve performance (e.g., VotingRegressor, StackingRegressor)

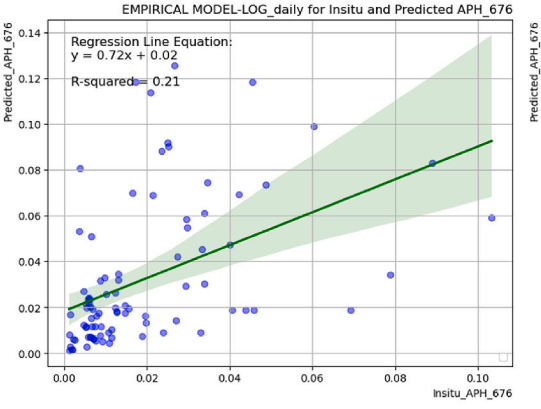
Appendix 2. Scatterplot for Daily_Rrs

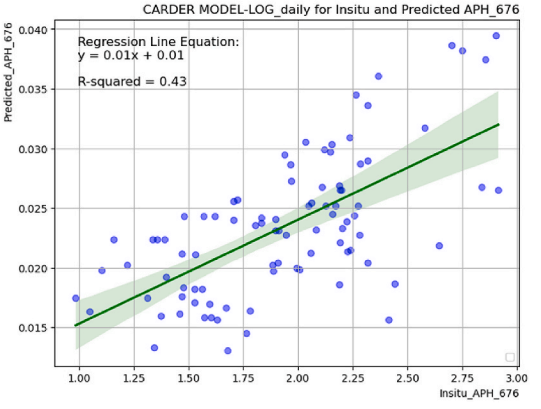

RF GRID-Log-Daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.71x + 0.54
Where:
m (slope) = 0.71
c (y-intercept) = 0.54
R-squared = 0.69
Reference Line
Predicted
In-situ

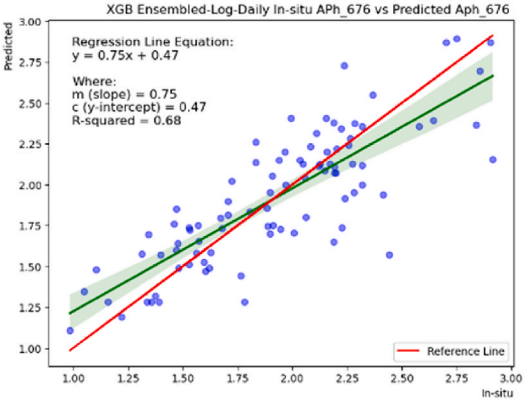

Meta Stacking-Log-Daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.79x + 0.42
Where:
m (slope) = 0.79
c (y-intercept) = 0.42
R-squared = 0.71
Reference Line
Predicted
In-situ

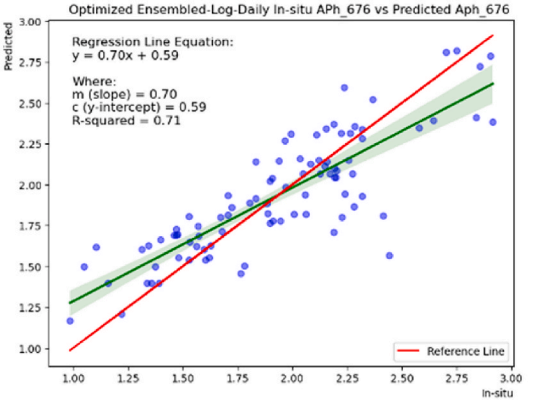

Forecast Ensemble_Log_Daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.71x + 0.55
Where:
m (slope) = 0.71
c (y-intercept) = 0.55
R-squared = 0.71
Reference Line
Predicted
In-situ


Stacked Voting Ensemble_Log_daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.73x + 0.53
Where:
m (slope) = 0.73
c (y-intercept) = 0.53
R-squared = 0.67
Reference Line
Predicted
In-situ

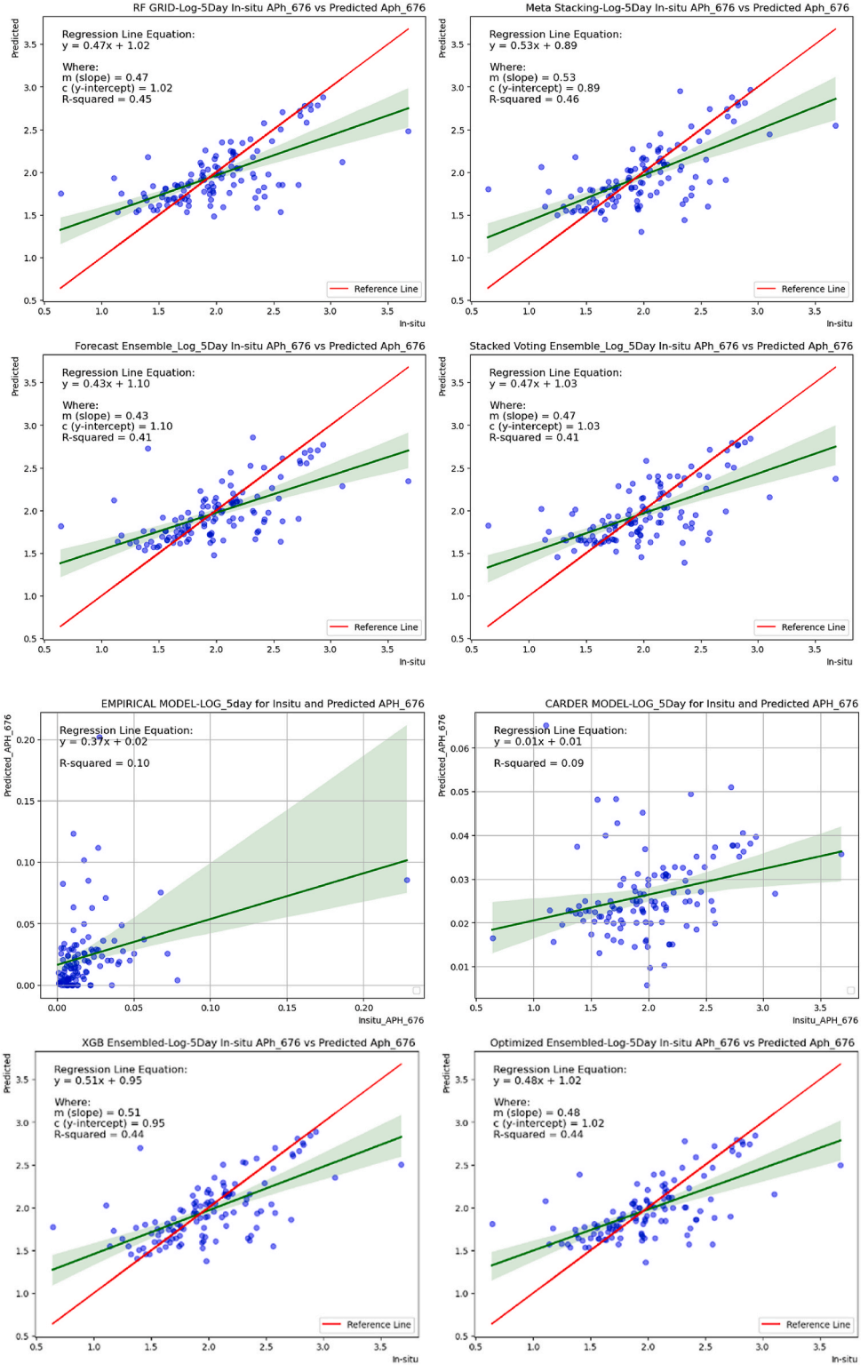

EMPIRICAL MODEL-LOG_daily for Insitu and Predicted APH_676
Regression Line Equation:
y = 0.72x + 0.02
R-squared = 0.21
Predicted_APH_676
Insitu_APH_676


CARDER MODEL-LOG_daily for Insitu and Predicted APH_676
Regression Line Equation:
y = 0.01x + 0.01
R-squared = 0.43
Predicted_APH_676
Insitu_APH_676

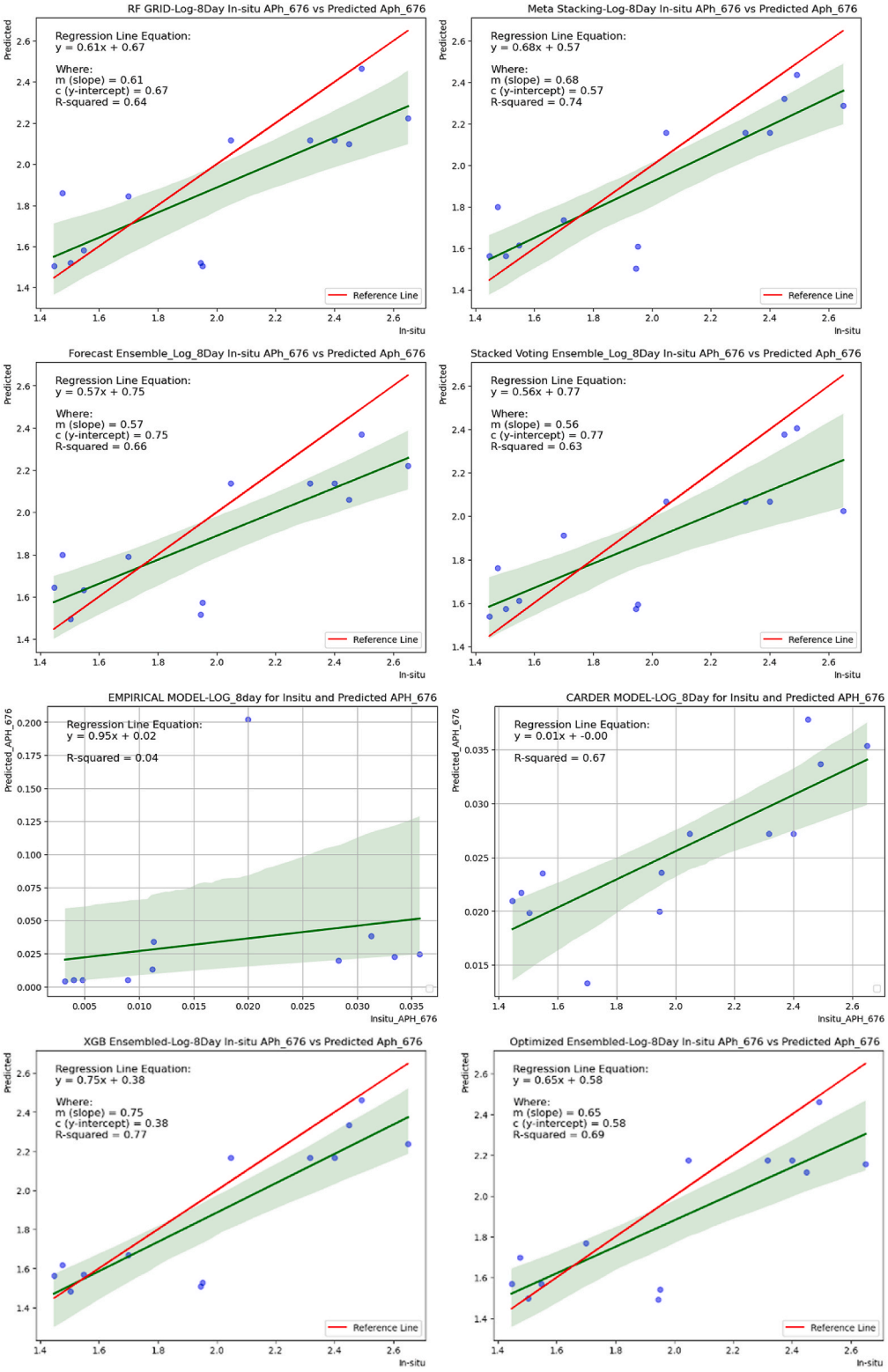

XGB Ensembled-Log-Daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.75x + 0.47
Where:
m (slope) = 0.75
c (y-intercept) = 0.47
R-squared = 0.68
Reference Line
Predicted
In-situ


Optimized Ensembled-Log-Daily In-situ APh_676 vs Predicted Aph_676
Regression Line Equation:
y = 0.70x + 0.59
Where:
m (slope) = 0.70
c (y-intercept) = 0.59
R-squared = 0.71
Reference Line
Predicted
In-situ

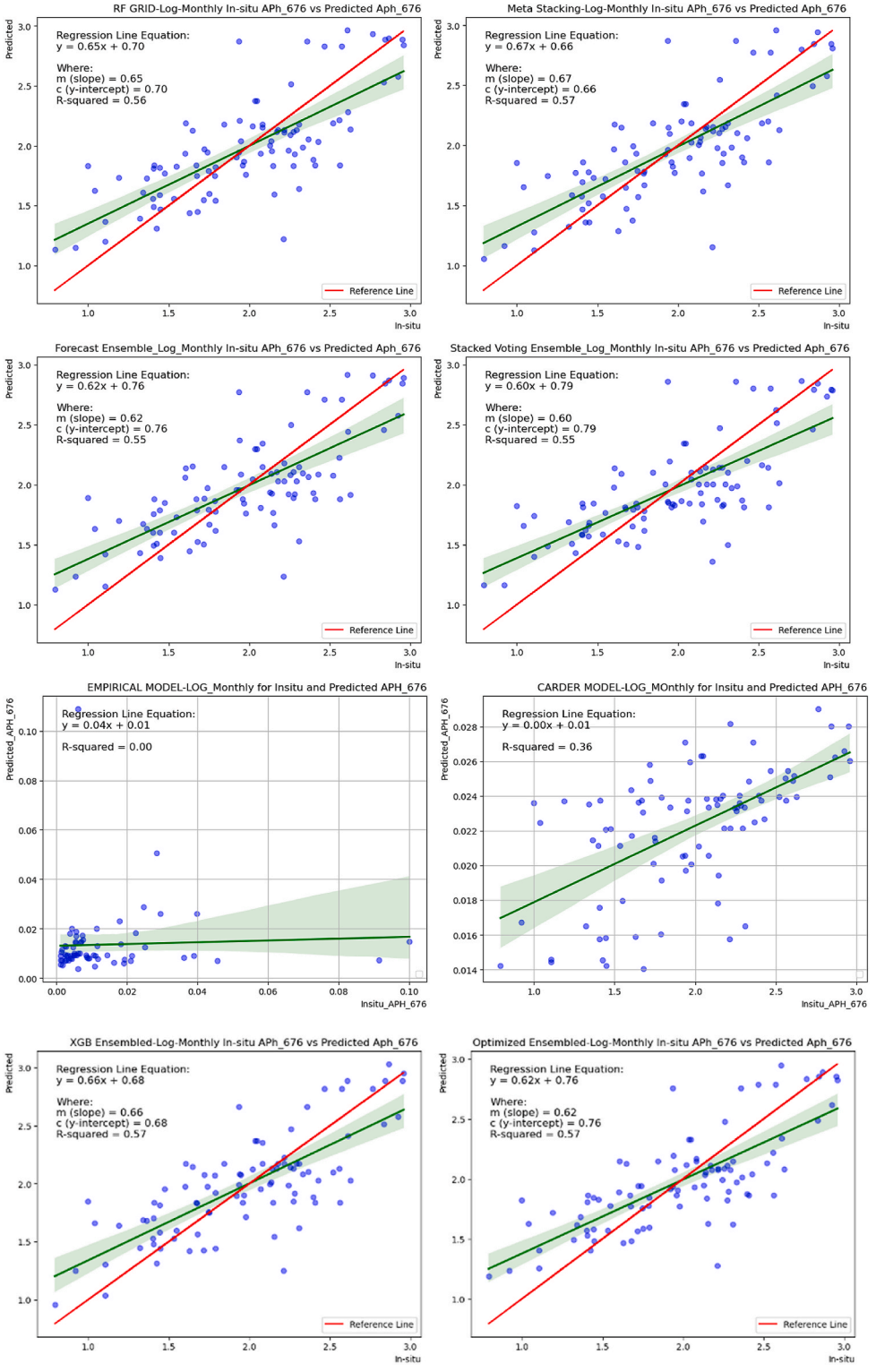
Appendix 3. Scatterplot for 5Day_Rrs



Appendix 4. Scatterplot for 8Day_Rrs



Appendix 5. Scatterplot for Monthly_Rrs



Data availability

Data will be made available on request.

References

Ahmad, H., 2019. Machine learning applications in oceanography. Aqu. Res. 161–169. https://doi.org/10.3153/AR19014.

Ahmed, S., El-Habashi, A., Lovko, V., 2017. In: (Will) Hou, W., Arnone, R.A. (Eds.), Neural Network Retrievals of Phytoplankton Absorption and Karenia brevis Harmful Algal Blooms in the West Florida Shelf, p. 101860L. https://doi.org/10.1117/12.2261848.

Alam, MdS., Tiwari, S.P., Rahman, S.M., 2024. Optimized ensemble machine learning models for predicting phytoplankton absorption coefficients. IEEE Access 12, 5760–5769. https://doi.org/10.1109/ACCESS.2024.3350328.

Ali, A., et al., 2021. Marine data prediction: an evaluation of machine learning, deep learning, and statistical predictive models. Comput. Intell. Neurosci. 2021 (1). https://doi.org/10.1155/2021/8551167.

Allali, K., Bricaud, A., Claustre, H., 1997. Spatial variations in the chlorophyll-specific absorption coefficients of phytoplankton and photosynthetically active pigments in the equatorial Pacific. J. Geophys. Res.: Oceans 102 (C6), 12413–12423. https://doi.org/10.1029/97JC00380.

Anderson, T.R., 2005. Plankton functional type modelling: running before we can walk? J. Plankton Res. 27, 1073–1081.

Ashphaq, M., Srivastava, P.K., Mitra, D., 2024. Satellite-derived bathymetry in dynamic coastal geomorphological environments through machine learning algorithms. Earth Space Sci. 11 (7). https://doi.org/10.1029/2024EA003554.

Barnes, M., et al., 2014. Absorption-based algorithm of primary production for total and size-fractionated phytoplankton in coastal waters. Mar. Ecol. Prog. Ser. 504, 73–89. https://doi.org/10.3354/meps10751.

Blondeau-Patissier, D., et al., 2014. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. Prog. Oceanogr. 123, 123–144. https://doi.org/10.1016/j.pocean.2013.12.008.

Bricaud, A., et al., 1995. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: analysis and parameterization. J. Geophys. Res.: Oceans 100 (C7), 13321–13332. https://doi.org/10.1029/95JC00463.

Bricaud, A., Stramski, D., 1990. Spectral absorption coefficients of living phytoplankton and nonalgal biogenous matter: a comparison between the Peru upwelling areaand the Sargasso Sea. Limnol. Oceanogr. 35 (3), 562–582. https://doi.org/10.4319/lo.1990.35.3.0562.

Brown, S.W., et al., 2007. In: Meynart, R., et al. (Eds.), The Marine Optical Buoy (MOBY) Radiometric Calibration and Uncertainty Budget for Ocean Color Satellite Sensor Vicarious Calibration. https://doi.org/10.1117/12.737400, 67441M.

Calvin, K., et al., 2023. *IPCC, 2023:* climate change 2023: synthesis report. In: Lee, H., Romero, J. (Eds.), Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team. IPCC, Geneva, Switzerland. https://doi.org/10.59327/IPCC/AR6-9789291691647.

Cao, W., et al., 2005. Spectral absorption coefficient of phytoplankton and its relation to chlorophyll a and remote sensing reflectance in coastal waters of southern China. Prog. Nat. Sci. 15 (4), 342–350. https://doi.org/10.1080/10020070512331342210.

Carder, K.L., et al., 1999. Semianalytic moderate-resolution imaging spectrometer algorithms for chlorophyll a and absorption with bio-optical domains based on nitrate-depletion temperatures. J. Geophys. Res.-Oceans 104 (C3), 5403–5421.

Carr, M.-E., et al., 2006. A comparison of global estimates of marine primary production from ocean color. Deep Sea Res. Part II Top. Stud. Oceanogr. 53 (5–7), 741–770. https://doi.org/10.1016/j.dsr2.2006.01.028.

CEOS, 2014. CEOS Strategy for Carbon Observations from Space.

Cetinić, I., et al., 2024. Phytoplankton composition from sPACE: requirements, opportunities, and challenges. Rem. Sens. Environ. 302, 113964. https://doi.org/10.1016/j.rse.2023.113964.

Churilova, T., et al., 2019. Phytoplankton light absorption in the deep chlorophyll maximum layer of the black sea. Europ. J. Rem. Sens. 52 (Suppl. 1), 123–136. https://doi.org/10.1080/22797254.2018.1533389.

Ciotti, Á.M., Lewis, M.R., Cullen, J.J., 2002. Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient. Limnol. Oceanogr. 47 (2), 404–417. https://doi.org/10.4319/lo.2002.47.2.0404.

Cleveland, J.S., 1995. Regional models for phytoplankton absorption as a function of chlorophyll *a* concentration. J. Geophys. Res.: Oceans 100 (C7), 13333–13344. https://doi.org/10.1029/95JC00532.

Cullen, J.J., et al., 1997. Optical detection and assessment of algal blooms. Limnol. Oceanogr. 42 (5), 1223–1239. http://www.jstor.org/stable/2839014.

Deng, L., et al., 2019. Retrieving phytoplankton size class from the absorption coefficient and chlorophyll A concentration based on support vector machine. Remote Sens. 11 (9), 1054. https://doi.org/10.3390/rs11091054.

Durap, A., 2023. A comparative analysis of machine learning algorithms for predicting wave runup. Anthro. Coasts 6 (1), 17. https://doi.org/10.1007/s44218-023-00033-7.

Falkowski, P.G., Barber, R.T., Smetacek, V., 1998. Biogeochemical controls and feedbacks on ocean primary production. Science 281 (5374), 200–206. https://doi.org/10.1126/science.281.5374.200.

Field, Michael J., Randerson, James T., Falkowski Paul, G., C, B.B., 1998. Primary production of the biosphere: integrating terrestrial and Oceanic components. Amer. Assoc. Adv. Sci. (AAAS) 281 (5374), 237–240. https://doi.org/10.1126/science.281.5374.237.

Hirata, T., et al., 2008. An absorption model to determine phytoplankton size classes from satellite ocean colour. Rem. Sens. Environ. 112 (6), 3153–3159. https://doi.org/10.1016/j.rse.2008.03.011.

Hirawake, T., et al., 2011. A phytoplankton absorption-based primary productivity model for remote sensing in the Southern Ocean. Polar Biol. 34 (2), 291–302. https://doi.org/10.1007/s00300-010-0949-y.

Huan, Y., et al., 2021. Phytoplankton "Missing" absorption in marine waters: a novel pigment compensation model for the packaging effect. J. Geophys. Res.: Oceans 126 (1). https://doi.org/10.1029/2020JC016458.

IOCCG, 2000. Remote sensing of ocean colour in coastal, and other optically-complex, waters. In: Dartmouth, S. Sathyendranath, NS (Eds.), Reports of the International Ocean-Colour Coordination Group.

IOCCG, 2006. Remote sensing of inherent optical properties: fundamentals, tests of algorithms, and applications. In: Lee, Z.P. (Ed.), Reports of the International Ocean-Colour Coordination Group. Dartmouth, Canada.

IOCCG, 2012. Mission Requirements for Future Ocean-Colour Sensors. Canada, Dartmouth.

Kirk, J.T.O., 1994. Light and Photosynthesis in Aquatic Ecosystems. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511623370.

Kostadinov, T.S., et al., 2023. Ocean color algorithm for the retrieval of the particle size distribution and carbon-based phytoplankton size classes using a two-component coated-sphere backscattering model. Ocean Sci. 19 (3), 703–727. https://doi.org/10.5194/os-19-703-2023.

Li, Y., et al., 2021. Research trends in the remote sensing of phytoplankton blooms: results from bibliometrics. Remote Sens. 13 (21), 4414. https://doi.org/10.3390/rs13214414.

Machado, K.B., et al., 2023. Systematic mapping of phytoplankton literature about global climate change: revealing temporal trends in research. Hydrobiologia 850 (1), 167–182. https://doi.org/10.1007/s10750-022-05052-y.

Marra, J., Trees, C.C., O'Reilly, J.E., 2007. Phytoplankton pigment absorption: a strong predictor of primary productivity in the surface ocean. Deep Sea Res. Oceanogr. Res. Pap. 54 (2), 155–163. https://doi.org/10.1016/j.dsr.2006.12.001.

Meler, J., et al., 2017. Light absorption by phytoplankton in the southern Baltic and Pomeranian lakes: mathematical expressions for remote sensing applications. Oceanologia 59 (3), 195–212. https://doi.org/10.1016/j.oceano.2017.03.010.

De Moraes Rudorff, N., Kampel, M., 2012. Orbital remote sensing of phytoplankton functional types: a new review. Int. J. Rem. Sens. 33 (6), 1967–1990. https://doi.org/10.1080/01431161.2011.601343.
Mouw, C.B., et al., 2017. A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global Ocean. Front. Mar. Sci. 4 (41). https://doi.org/10.3389/fmars.2017.00041.
Pahlevan, N., et al., 2021. Hyperspectral retrievals of phytoplankton absorption and chlorophyll-a in inland and nearshore coastal waters. Rem. Sens. Environ. 253, 112200. https://doi.org/10.1016/j.rse.2020.112200.
Patara, L., et al., 2012. Global response to solar radiation absorbed by phytoplankton in a coupled climate model. Clim. Dyn. 39 (7–8), 1951–1968. https://doi.org/10.1007/s00382-012-1300-9.
Paulsen, H., et al., 2018. Light absorption by marine Cyanobacteria affects tropical climate mean state and variability. Earth Syst. Dyn. 9 (4), 1283–1300. https://doi.org/10.5194/esd-9-1283-2018.
Pérez, G.L., et al., 2021. Variability of phytoplankton light absorption in stratified waters of the NW Mediterranean Sea: the interplay between pigment composition and the packaging effect. Deep Sea Res. Oceanogr. Res. Pap. 169, 103460. https://doi.org/10.1016/j.dsr.2020.103460.
Robinson, C.M., et al., 2017. Phytoplankton absorption predicts patterns in primary productivity in Australian coastal shelf waters. Estuar. Coast Shelf Sci. 192, 1–16. https://doi.org/10.1016/j.ecss.2017.04.012.
Roelke, D.L., Kennedy, C.D., Weidemann, A.D., 1999. Use of discriminant and fourth-derivative analyses with high-resolution absorption spectra for phytoplankton research: limitations at varied signal-to-noise ratio and spectral resolution. Gulf Mex. Sci. 17 (2). https://doi.org/10.18785/goms.1702.02.
Roy, S., et al., 2013. The global distribution of phytoplankton size spectrum and size classes from their light-absorption spectra derived from satellite data. Rem. Sens. Environ. 139, 185–197. https://doi.org/10.1016/j.rse.2013.08.004.
Roy, S., 2018. Distributions of phytoplankton carbohydrate, protein and lipid in the world oceans from satellite ocean colour. ISME J. 12 (6), 1457–1472. https://doi.org/10.1038/s41396-018-0054-8.
Roy, S., Sathyendranath, S., Platt, T., 2011. Retrieval of phytoplankton size from bio-optical measurements: theory and applications. J. R. Soc. Interface 8 (58), 650–660.
Roy, S., Sathyendranath, S., Platt, T., 2017. Size-partitioned phytoplankton carbon and carbon-to-chlorophyll ratio from ocean-colour by an absorption-based bio-optical algorithm. Rem. Sens. Environ. 194, 177–189. https://doi.org/10.1016/j.rse.2017.02.015.
Sadaiappan, B., et al., 2023. Applications of machine learning in chemical and biological oceanography. ACS Omega 8 (18), 15831–15853. https://doi.org/10.1021/acsomega.2c06441.
Sathyendranath, S., et al., 2019. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (OC-CCI). Sensors 19 (19), 4285. https://doi.org/10.3390/s19194285.
Seppälä, J., Ylöstalo, P., Kuosa, H., 2005. Spectral absorption and fluorescence characteristics of phytoplankton in different size fractions across a salinity gradient in the Baltic sea. Int. J. Rem. Sens. 26 (2), 387–414. https://doi.org/10.1080/01431160410001723682.
Shang, S., et al., 2011. MODIS observed phytoplankton dynamics in the Taiwan strait: an absorption-based analysis. Biogeosciences 8 (4), 841–850. https://doi.org/10.5194/bg-8-841-2011.
Shang, Y., et al., 2021. Variations in the light absorption coefficients of phytoplankton, non-algal particles and dissolved organic matter in reservoirs across China. Environ. Res. 201, 111579. https://doi.org/10.1016/j.envres.2021.111579.
Shen, L., Xu, H., Guo, X., 2012. Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework. Sensors 12 (6), 7778–7803. https://doi.org/10.3390/s120607778.
Silsbe, G.M., et al., 2016. The CAFE model: a net production model for global ocean phytoplankton. Glob. Biogeochem. Cycles 30 (12), 1756–1777. https://doi.org/10.1002/2016GB005521.
Sun, D., et al., 2010. Partitioning particulate scattering and absorption into contributions of phytoplankton and non-algal particles in winter in Lake Taihu (China). Hydrobiologia 644 (1), 337–349. https://doi.org/10.1007/s10750-010-0198-7.
Valente, A., et al., 2022. A compilation of global bio-optical in situ data for ocean colour satellite applications – version three. Earth Syst. Sci. Data 14 (12), 5737–5770. https://doi.org/10.5194/essd-14-5737-2022.
Wang, G., et al., 2008. Partitioning particulate absorption coefficient into contributions of phytoplankton and nonalgal particles: a case study in the northern South China Sea. Estuar. Coast Shelf Sci. 78 (3), 513–520. https://doi.org/10.1016/j.ecss.2008.01.013.
Wang, Guifen, et al., 2021. Estimation of phytoplankton pigment concentration in the South China Sea from hyperspectral absorption data. Acta Opt. Sin. 41 (6), 0601002. https://doi.org/10.3788/AOS202141.0601002.
Wei, J., et al., 2023. Chlorophyll-specific absorption coefficient of phytoplankton in world oceans: seasonal and regional variability. Remote Sens. 15 (9), 2423. https://doi.org/10.3390/rs15092423.
Werdell, P.J., Bailey, S.W., 2005. An improved bio-optical data set for ocean color algorithm development and satellite data product validation. Remote Sens. Environ. 98 (1), 122–140. https://doi.org/10.1016/j.rse.2005.07.001.
Xu, X., et al., 2025. A new algorithm based on the phytoplankton absorption coefficient for red tide monitoring in the East China Sea via a geostationary ocean color imager (GOCI). Remote Sens. 17 (5), 750. https://doi.org/10.3390/rs17050750.
Zhang, M., et al., 2021. Simulating the relationship between land use/cover change and urban thermal environment using machine learning algorithms in wuhan city, China. Land 11 (1), 14. https://doi.org/10.3390/land11010014.
Zhang, M., et al., 2023. Impact of urban expansion on land surface temperature and carbon emissions using machine learning algorithms in wuhan, China. Urban Clim. 47, 101347. https://doi.org/10.1016/j.uclim.2022.101347.
Zhang, M., et al., 2024. Predicting the impacts of urban development on urban thermal environment using machine learning algorithms in Nanjing, China. J. Environ. Manag. 356, 120560. https://doi.org/10.1016/j.jenvman.2024.120560.
Zhang, Y., et al., 2010. Seasonal-spatial variation and remote sensing of phytoplankton absorption in Lake Taihu, a large eutrophic and shallow lake in China. J. Plankton Res. 32 (7), 1023–1037. https://doi.org/10.1093/plankt/fbq039.
Zhang, Z., et al., 2025. A review of machine learning applications in ocean color remote sensing. Remote Sens. 17 (10), 1776. https://doi.org/10.3390/rs17101776.
Zheng, G., Stramski, D., 2013a. A model based on stacked-constraints approach for partitioning the light absorption coefficient of seawater into phytoplankton and non-phytoplankton components. J. Geophys. Res.: Oceans 118 (4), 2155–2174. https://doi.org/10.1002/jgrc.20115.
Zheng, G., Stramski, D., 2013b. A model for partitioning the light absorption coefficient of suspended marine particles into phytoplankton and nonalgal components. J. Geophys. Res.: Oceans 118 (6), 2977–2991. https://doi.org/10.1002/jgrc.20206.
Zheng, G., Stramski, D., DiGiacomo, P.M., 2015. A model for partitioning the light absorption coefficient of natural waters into phytoplankton, nonalgal particulate, and colored dissolved organic components: a case study for the <scp>C</scp> hesapeake <scp>B</scp> ay. J. Geophys. Res.: Oceans 120 (4), 2601–2621. https://doi.org/10.1002/2014JC010604.