Master Thesis

Marine Big Data-driven Machine Learning Based Monitoring of Phytoplankton Groups in the Arctic Ocean

Programme:
M.Sc. Space Sciences and Technologies

Submitted by:
Alfredo J. Bellido Rosas
6236057

Supervisors:
Dr. Hongyan Xi
Prof. Dr. Astrid Bracher

Faculty of Physics and Electrical Engineering
University of Bremen
Bremen, Germany

Alfred Wegener Institut
Helmholtz Zentrum für Polar und Meeresforschung
Bremerhaven, Germany

September 10, 2024

Contents

Abstract

The Arctic Ocean is experiencing rapid and significant changes due to climate warming, profoundly impacting its physical and biological systems. Phytoplankton, as primary producers, play a crucial role in marine ecosystems and biogeochemical cycles. Monitoring their distribution and abundance is essential for understanding the health of the Arctic marine environment. This study focuses on developing an ensemble machine learning model to predict concentrations of Total Chlorophyll-a (TChl-a) and various Phytoplankton Functional Types (PFTs) in the Arctic Ocean, leveraging data from satellite observations and in-situ measurements.

The ensemble model combines Gradient Boosting Machine (GBM), Fully Connected Neural Network (FCNN), Random Forest Regression (RFR), and Support Vector Machine (SVM) through a Ridge Regression Ensemble approach. The model was trained by using satellite data and model simulations outputs from Copernicus Marine Service (CMEMS) that were matched with in situ data collected during 1997-2020 and validated using in-situ measurements from the PS131 expedition [1]. The model demonstrates strong predictive capabilities, particularly for Diatoms and TChl-a, which are crucial for understanding primary production and nutrient dynamics in the Arctic.

Results indicate that the ensemble model performs well in capturing the spatial and temporal distribution of TChl-a and PFTs. The model's robust performance during the training phase and its ability to generalise to the validation dataset, regardless of its higher variability respect to the training dataset, underscore its potential for large-scale ecological monitoring.

The creation of Arctic maps for PFTs and TChl-a provided valuable insights into the spatial distribution of these variables. In the maps created, higher concentrations of Diatoms were observed near coastal areas, aligning with known nutrient-rich environments such as river outflows and upwelling zones, particularly along the coasts of northern Europe and northern Asia. Green Algae showed a patchy distribution influenced by localised environmental factors, such as variations in light availability and nutrient inputs from specific sources like the Barents Sea and Laptev Sea. Haptophytes exhibited specialised niches in cooler waters, reflecting their ecological roles in regions like the Kara Sea, where lower temperatures and nutrient availability favor their growth. Dinoflagellates were distributed along various coastal regions in northern Europe and northern Asia without a specific area of high concentration, suggesting their adaptability to a range of environmental conditions.

In general, the model effectively identified areas of high phytoplankton activity, which are essential for understanding the Arctic marine food web and biogeochemical cycles. Hence, this study demonstrates the potential of using machine learning models for predicting phytoplankton dynamics in the Arctic, offering a robust tool for monitoring and managing marine ecosystems.

1. Introduction

The Arctic Ocean is undergoing rapid and significant changes due to climate warming, impacting its physical and biological systems. Phytoplankton plays different roles in the dynamics of the ocean systems. They are the primary producers in the marine ecosystem, forming the base of the food web and supporting a diverse array of marine organisms, from zooplankton to large mammals and birds [2]. The significant reduction in sea-ice extent and the resulting increase in open-water habitat have led to changes in phytoplankton bloom dynamics, including earlier blooms and later termination of it [3]. Phytoplankton dynamics influence the entire marine food web, affecting species interactions, population structures, and the overall health of marine ecosystems. Changes in phytoplankton composition and abundance can cascade through the food web, impacting the productivity and survival of higher trophic levels [4].

Phytoplankton is also responsible of capturing the carbon through the biological pump, transferring carbon from the surface to the deep ocean and sediments. They absorb carbon dioxide during photosynthesis, and when they die, they sink to the ocean floor, capturing carbon and helping to regulate atmospheric CO_2 levels. While these processes are common for all phytoplankton, some species have specific chemical requirements due to their distinct physiological functions, therefore playing different roles in the ocean biogeochemical cycles. Some phytoplankton such as dinoflagellates and prymnesiophytes (haptophytes) influence the Earth's climate by affecting ocean albedo and through the production of dimethyl sulfide (DMS), which impacts cloud formation and climate regulation. Changes in phytoplankton dynamics can alter these feedback mechanisms, affecting regional and global climate patterns, like a reduction in sea-ice cover increases the amount of sunlight absorbed by the ocean, potentially enhancing phytoplankton growth and altering the production of DMS, which in turn influences cloud cover and climate [2]. Other phytoplankton species as diatoms use Si to form their silica cell walls. Prokaryotes, particularly cyanobacteria, are important for their role in nitrogen fixation, converting atmospheric nitrogen into forms usable by other organisms. This process is crucial in oligotrophic (nutrient-poor) regions of the ocean where nitrogen is a limiting nutrient [5] [6]. These functional differences have led to phytoplankton to be classified into Phytoplankton Functional Types (PFT) [6]. In order to quantify the contributions of these PFT, accurate monitoring is crucial for understanding the marine ecosystems and global biogeochemical cycles. Essential to this monitoring is the measurement of chlorophyll-a, a pigment found in all phytoplankton that plays a key role in photosynthesis.

The concentration of chlorophyll-a (Chl-a) in the ocean is directly related to the amount of phytoplankton present, which respond rapidly to changes in environmental conditions. As consequence of the dependency of nutrients and light, the concentration of Chl-a in the Arctic Ocean is seasonal (Fig. 1.1), since the availability of these are driven by the sea-ice cover dynamics, the seasonal sun-light and the changes in the mixed layer. It was observed that in the last past two decades the annual NPP (Net Primary Production) has increased [7]. The NPP is the difference between the total amount of carbon fixed through photosynthesis and the amount of carbon lost through plant respiration, which is a crucial measure of ecosystem productivity. This increment of NPP is linked to the sea-ice decline, and consequently to the emergence of earlier peaks in the year of Chl-a concentration in the Arctic. Therefore, monitoring spatial-temporal distribution and variability of Chl-a and PFT in the area has crucial importance in better understanding of marine ecosystem dynamics and biogeochemical cycles, in order to be able to separate potential long-term climate signals from natural variability in the short term [2].

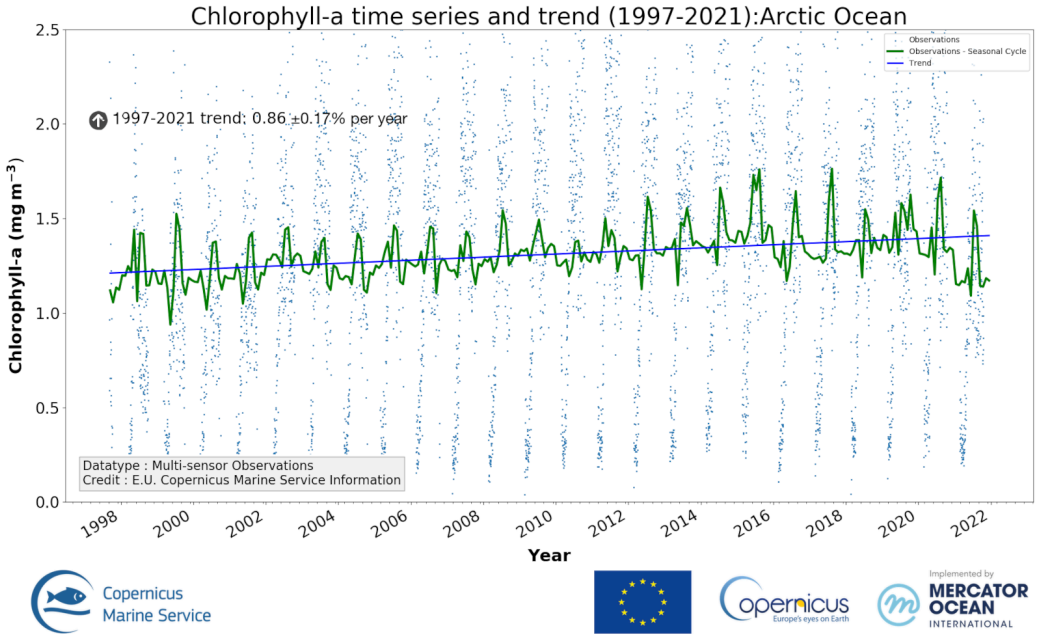

Figure 1.1: Arctic Ocean time series and trend (1997-2021) of satellite chlorophyll, based on CMEMS product OCEANCOLOUR_ARC_BGC_L3_MY_009_123. The daily regional average (weighted by pixel area) time series is shown in blue, with the de-seasonalized time series in green and the linear trend in blue. [7]

Several algorithms have been developed to estimate PFTs from ocean color data, each with unique strengths and limitations. Traditional bio-optical algorithms, such as abundance-based models, rely on empirical relationships between Chl-a concentrations and specific pigment markers to estimate PFTs [6]. However, these models often face challenges in capturing the intricate bio-optical signals and overlapping spectral signatures of different phytoplankton groups, especially in regions with complex environmental conditions like the Arctic [8].

Advanced machine learning approaches, such as the Spatial-Temporal-Ecological Ensemble (STEE) model, have demonstrated improved performance by combining multiple data sources and leveraging spatio-temporal patterns to predict PFT distributions globally [9]. The STEE model, for example, integrates diverse ecological, temporal, and spatial data to enhance its predictive capabilities across various marine environments, providing a useful framework for PFT prediction. Inspired by such methodologies, this study develops an ensemble machine learning model specifically tailored to the Arctic region, addressing its unique ecological dynamics and the need for higher-resolution predictions.

An interdisciplinary method that combines the use of machine learning with extensive marine data from ocean observations and simulation outputs (Figure 1.2), specifically focused on the sub-arctic and arctic regions (above 50° latitude), could provide improved quantification of PFTs and Chl-a concentrations, thereby compensating for the lack of in-situ data in this region. This study developed an ensemble machine learning model designed to produce robust predictions for five distinct PFTs: Diatoms, Dinoflagellates (Dino), Haptophytes (Hapto), Prokaryotic phytoplankton (Proka), and Green Algae (GA) and Total Chlorophyll-a (TChl-a). The model leverages innovative machine learning techniques adapted to meet the specific requirements of the Arctic's unique environmental conditions.

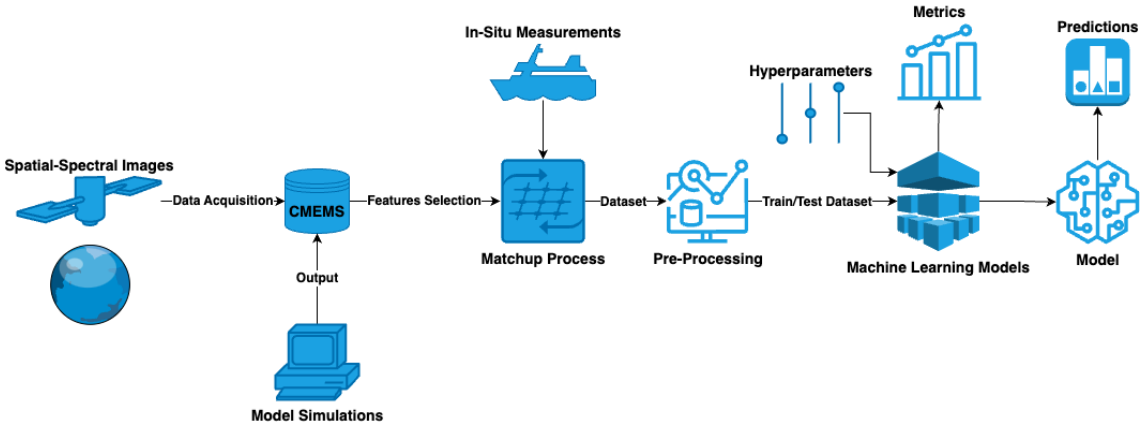

Figure 1.2: Flowchart - Ensemble Model Developing Steps.

2. Data and Methods

2.1 In-situ Dataset

The in-situ PFT measurements used in this study were derived using an updated diagnostic pigment analysis (DPA) method [10] with retuned coefficients [11]. The values of retuned DPA weighting coefficients for PFT Chla determination are: 1.56 for fucoxanthin, 1.53 for peridinin, 0.89 for 19'-hexanoyloxyfucoxanthin, 0.44 for 19'-butanoyloxyfucoxanthin, 1.94 for alloxanthin, 2.63 for total chlorophyll b, and 0.99 for zeaxanthin. The coefficient retuning was based on an updated global HPLC pigment data base for the open ocean (water depth > 200 m), which was compiled based on the previously published data sets spanning from 1988 to 2012 [12], with updates [13] [11], by adding other newly available HPLC pigment data collected between 2012 and 2018 mainly from SeaBASS, PANGAEA, British Oceanographic Data Centre (BODC), and Australian Open Access to Ocean Data (AODN). The complete data set composes a large amount of quality controlled in situ measurements of major pigments based on HPLC collected from various expeditions across the Atlantic Ocean spanning from 71°S to 84°N. This complete global data set where the Arctic data were extracted, covers the years from 1997 to 2020. The location threshold set for this study for the in-situ measurements is above 50°N (Fig. 2.1).

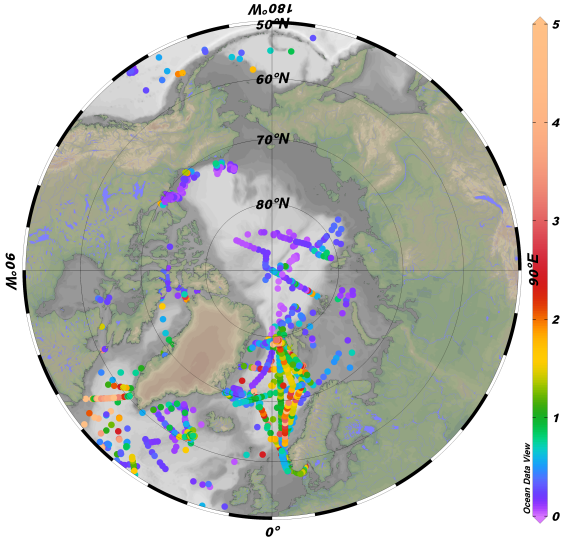

Figure 2.1: In-situ Arctic measurements locations — Total chlorophyll-a (mg/m^3).

High-performance liquid chromatography (HPLC) is a widely utilised method for estimating phytoplankton community structure and obtaining information about PFTs in ocean samples. HPLC measures the concentrations of various phytoplankton pigments in water samples, with some pigments serving as chemotaxonomic markers for specific phytoplankton groups. This allows researchers to revise several major groups of phytoplankton on a global scale, such as cyanobacteria, diatoms/dinoflagellates, haptophytes, and green algae. Despite its strengths, HPLC has limitations, such as the variable occurrence and plasticity of pigments across species, groups, strains, and environmental conditions. Nevertheless, HPLC remains a crucial method for characterising

phytoplankton community structure, particularly in the context of ocean colour remote sensing and the long-term monitoring of marine ecosystems [5].

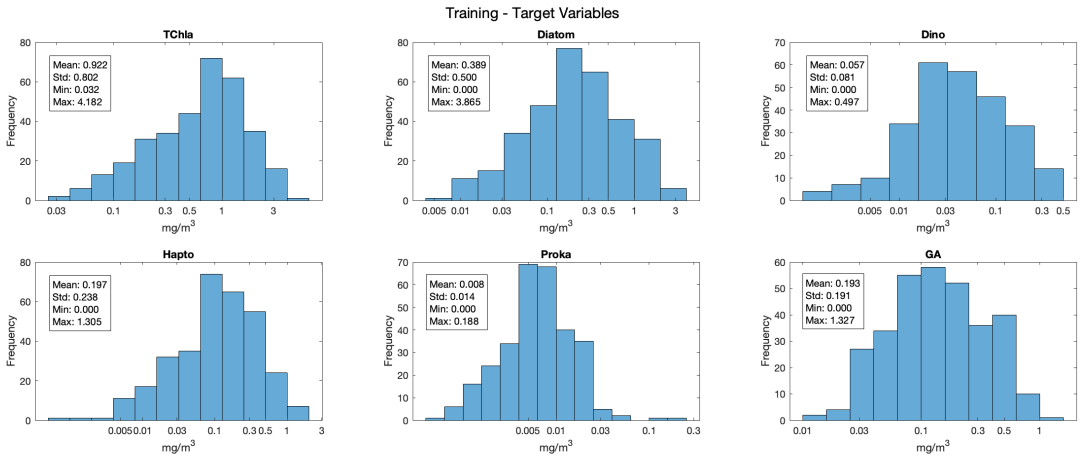

Figure 2.2: Histogram Statistics - Training Target Variables — In-situ measurements.

For this study, a subset of Chl-a concentration of five PFTs and TChl-a of the in-situ measurements was selected to train the ensemble model as target variables. The statistical summary of these target variables provides valuable insights into their distribution and variability (Fig. 2.2). The mean TChl-a concentration was 0.922 mg/m^3, with a median of 0.739 mg/m^3, indicating a skewed distribution towards higher values. The standard deviation of 0.802 mg/m^3 suggests considerable variability, with values ranging from 0.032 mg/m^3 to 4.182 mg/m^3. This variability is crucial for training the model to accurately predict TChl-a under different conditions.

Diatoms showed a mean concentration of 0.389 mg/m^3 and a median of 0.213 mg/m^3, with a higher standard deviation of 0.500 mg/m^3, reflecting their significant presence and variability in the Arctic Ocean. Dino, with a mean concentration of 0.057 mg/m^3 and a median of 0.026 mg/m^3, exhibited lower abundance but also had notable variability as indicated by the standard deviation of 0.081 mg/m^3. Hapto and GA had mean concentrations of 0.197 mg/m^3 and 0.193 mg/m^3 respectively, both with considerable variability as shown by their standard deviations. Proka had the lowest mean concentration at 0.008 mg/m^3, with a median of 0.006 mg/m^3 and a standard deviation of 0.014 mg/m^3, indicating their sparse distribution and reduced presence in this region.

2.2 CMEMS Dataset

In this study, datasets from Copernicus Marine Service (CMEMS), which include different remote sensing and simulation output products were utilised (Table 2.1), with parameters or features such as Chl-a concentration, sea surface temperature, nutrients, optical features and other biogeochemical variables (Fig. 2.3). These datasets span multiple years (1997-2020) and cover above the 50°N, providing the basis for the ensemble model development.

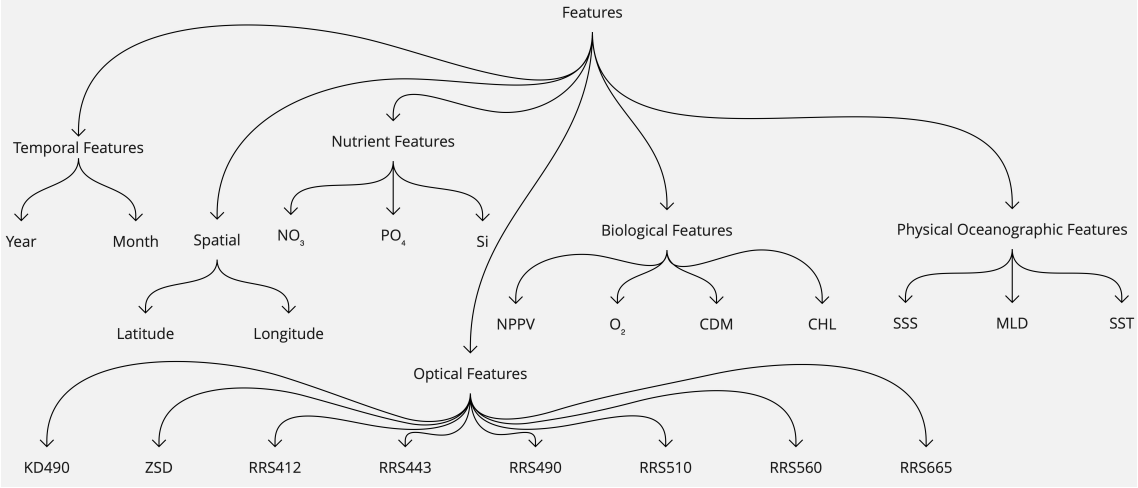

Figure 2.3: Selected Features for the Ensemble Model.

Table 2.1: Copernicus Marine Service Datasets - Daily Resolution.

Dataset ID	**Variables**	**Horizontal Resolution**	**Source**
cmems_mod_glo_bgc_my_0.25_P1D-m	NO_3, PO_4, Si, $NPPV$, O_2	$\sim 25km \times 25km$	Model
c3s_obs-oc_glo_bgc-plankton_my_l3-multi-4km_P1D	CHL	$4km \times 4km$	Observations
c3s_obs-oc_glo_bgc-reflectance_my_l3-multi-4km_P1D	$RRS412$, $RRS443$, $RRS490$, $RRS510$, $RRS560$, $RRS665$	$4km \times 4km$	Observations
cmems_obs-oc_glo_bgc-transp_my_l3-multi-4km_P1D	$KD490$, ZSD	$4km \times 4km$	Observations
cmems_obs-oc_glo_bgc-optics_my_l3-multi-4km_P1D	CDM	$4km \times 4km$	Observations
cmems_mod_glo_phy_my_0.083deg_P1D-m	SSS, MLD	$\sim 10km \times 10km$	Model
METOFFICE-GLO-SST-L4-REP-OBS-SST	SST	$\sim 5km \times 5km$	Observations

Figure 2.3 shows how the selected features are grouped according to their nature for better understanding. The table 2.1 shows the datasets ID of CMEMS from where the features were taken, together with their original resolution and if they are either simulation model outputs or observations by satellite sensors. In a machine learning models, features are the individual measurable properties or characteristics of the phenomenon being observed. They serve as the input variables that the model uses to make predictions. Features directly impact the accuracy and performance of machine learning models, therefore choosing the right set of features is critical for building effective models. A brief description of the selected features is summarised as below:

Latitude and longitude, the geographical coordinates of a sample location, are crucial for

predicting PFT groups because different phytoplankton thrive in distinct geographical regions. Variations in climate, nutrient availability, and water conditions across latitudinal and longitudinal gradients significantly influence the distribution of phytoplankton groups [14].

Temporal factors such as year and month provide essential context for understanding seasonal and inter-annual variations in PFT distribution. Seasonal changes impact light availability, temperature, and nutrient cycling, which in turn influence phytoplankton growth patterns [15].

Nitrate (NO_3) concentration is a critical nutrient parameter that affects phytoplankton growth. Different PFT groups have varying nitrate requirements and uptake mechanisms. Nitrate availability can thus shape the composition of phytoplankton communities [16].

Phosphate (PO_4) is another vital nutrient for phytoplankton, and its concentration in water influences phytoplankton diversity and productivity. Phosphate limitation can restrict the growth of certain PFT groups, favouring those that can efficiently utilise low phosphate levels [17].

Silicate (Si) concentration is particularly important for diatoms, which require silicate for their frustules. The availability of silicate can thus directly affect the abundance and distribution of diatoms relative to other PFT groups [18].

Net primary productivity of vegetation (NPPv) is an indicator of the overall productivity of the phytoplankton community. High NPPv values often correlate with blooms of certain phytoplankton groups, reflecting their capacity for rapid growth under favorable conditions [19].

Oxygen (O_2) levels can influence phytoplankton metabolism and the composition of PFT groups. Oxygen concentration is related to both photosynthesis and respiration processes in marine environments [20].

Colored Dissolved Organic Matter (CDM) and Chlorophyll-a (CHL) concentrations provide insights into the presence and activity of phytoplankton. CDM can affect light penetration in water, while chlorophyll-a is a direct proxy for phytoplankton biomass. The concentration of CDM can have a significant effect on biological activity in aquatic systems. CDM diminishes light intensity as it penetrates water. Very high concentrations of CDM can have a limiting effect on photosynthesis and inhibit the growth of phytoplankton [21, 22].

The diffuse attenuation coefficient at 490 nm (KD490) and Secchi disk depth (ZSD) are measures of water clarity and light penetration. The presence of phytoplankton contributes to the light attenuation in the upper ocean layers, affecting the value of KD490. The ZSD is a simple measure of water transparency or turbidity, determined by lowering a white disk into the water until it disappears from view. ZSD is inversely related to KD490, higher KD490 values (greater light attenuation) correspond to shallower Secchi depths (lower transparency) [21].

Remote sensing reflectance at various wavelengths (RRS412, RRS443, RRS490, RRS510, RRS560, RRS665) These are remote sensing reflectance values at different wavelengths (measured in nanometers). They indicate how much light is being reflected by the water at these specific wavelengths. Changes in these values can be indicative of different types and concentrations of substances in the water, including phytoplankton [23].

Sea surface salinity (SSS) and mixed layer depth (MLD) are important physical parameters that affect phytoplankton distribution. Salinity can influence phytoplankton physiology and community structure, lower sea surface salinity from ice melt creates favorable conditions for phytoplankton blooms in polar regions while mixed layer depth affects nutrient availability and light conditions [24].

Sea surface temperature (SST) is a key determinant of phytoplankton growth rates and species composition. Temperature influences metabolic rates and the stratification of the water column, which in turn affects nutrient availability [25].

2.3 Matchup Extraction

To ensure the accuracy and reliability of the machine learning model for predicting the target variables (PFTs and TChl-a), matching satellite-derived and model output data with in-situ measurements is a critical step. This process involves several key stages, each essential for aligning and validating the datasets.

The first step involves selecting relevant datasets from the CMEMS. To access these datasets, an FTP connection is established with the CMEMS server, allowing retrieval of data files stored

in a structured directory system organised by year and month. The downloading process involves navigating through these directories and fetching files for each day within the specified date range, ensuring comprehensive temporal coverage.

The critical task of aligning dataset points with corresponding in-situ measurements, known as the matchup process, ensures that model/observed-derived features accurately reflect real-world conditions. Basically, matching in-situ measurements with dataset points that are geographically and temporally close.

This process begins by reading the in-situ data from a CSV file, containing information about the location (latitude and longitude) and time of each measurement. For each in-situ data point, the corresponding dataset file is identified based on the date. The dataset, stored in NetCDF format, is then accessed. To find the dataset points nearest to the in-situ measurement locations, the latitude and longitude from the in-situ data are compared with those in the dataset points. The nearest data points are identified by calculating the minimum Euclidean distance between the in-situ location and the grid points in the dataset points. This precise matching ensures that the data points used is as close as possible to the actual measurement location.

Once the nearest dataset points are identified, the values of the relevant features are extracted. To ensure robustness, a 3x3 grid of surrounding points is considered, accounting for spatial variability. The values from these grid points are averaged, provided they meet criteria such as a minimum number of valid data points and a low coefficient of variation. This 3x3 matrix approach helps to smooth out any local anomalies and provides a more representative average value for the area surrounding the in-situ measurement, thereby reducing noise and improving data reliability.

For example, the Euclidean distance d between an in-situ point (lat_{IS}, lon_{IS}) and a dataset grid point (lat_{sat}, lon_{sat}) is calculated as:

$$d = \sqrt{(lat_{IS} - lat_{cmems})^2 + (lon_{IS} - lon_{cmems})^2}$$

The nearest grid point minimises this distance. The coefficient of variation (CV) is calculated to ensure data consistency:

$$CV = \frac{\sigma}{\mu}$$

where σ is the standard deviation and μ is the mean of the values within the 3x3 grid. If $CV < 0.25$, the values are considered consistent and averaged.

The matched data with a total of 334 points including the collocated predictor variables with the in situ measurements, are compiled into a structured format that will be used for the training of the model.

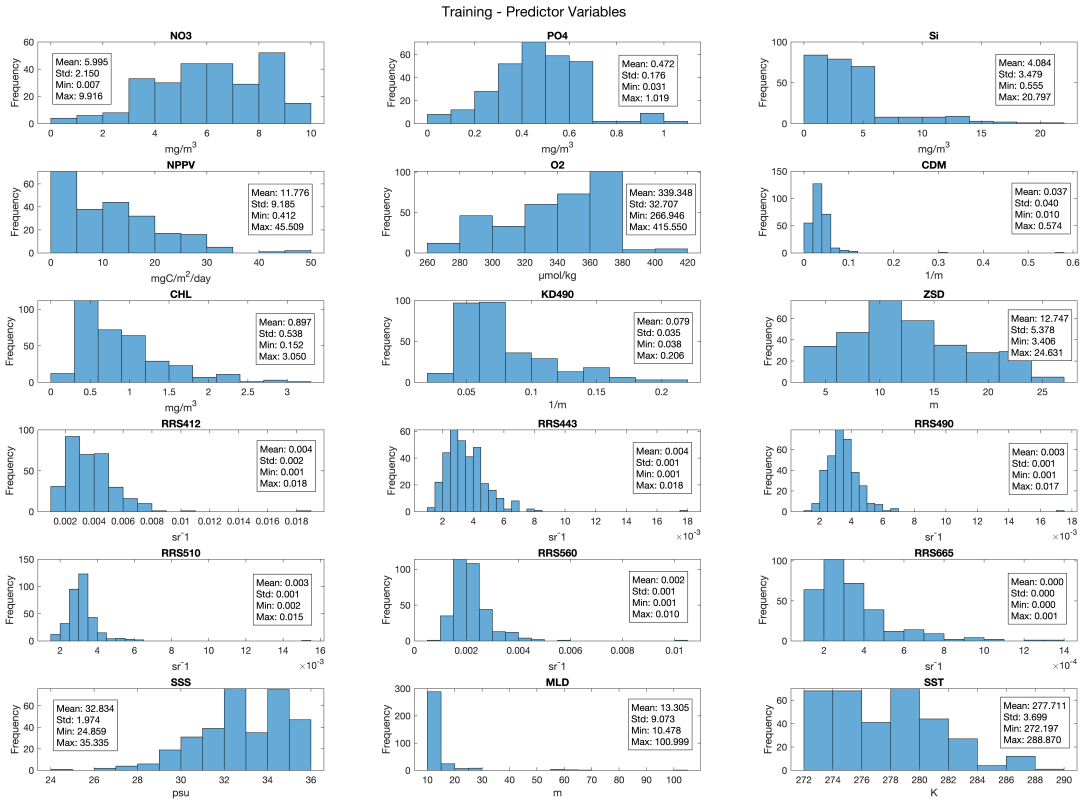

Figure 2.4: Histogram and Statistics - Training Predictor Variables from CMEMS Dataset after Matchup Extraction.

The descriptive statistics of these predictor variables in the matchup dataset, illustrated in Figure 2.4, provide a comprehensive understanding of their distributions and variabilities, which are crucial for model training. For instance, nitrate (NO_3) concentrations, with a mean of 5.995 and a standard deviation of 2.150, exhibit significant variability, indicating the diverse nutrient conditions that phytoplankton experience in the Arctic Ocean. Phosphate (PO_4) and silicate (Si) show similar patterns of variability, essential for understanding the nutrient limitations and requirements of different PFTs.

Net primary productivity of vegetation (NPPv), with a mean of 11.776 and a high standard deviation of 9.185, reflects the dynamic productivity levels in the Arctic, influenced by various environmental factors. Oxygen (O_2) levels, relatively stable with a mean of 339.348 and a lower standard deviation, provide a consistent measure of the marine environment's health and its influence on phytoplankton metabolism.

The optical properties, such as CDM, CHL, KD490, and ZSD, highlight the variability in water clarity and phytoplankton biomass. For example, the mean CHL concentration is 0.897 mg/m^3, with a standard deviation of 0.538, indicating a broad range of phytoplankton biomass across different regions and conditions in the Arctic.

Remote sensing reflectance values at various wavelengths (RRS412, RRS443, RRS490, RRS510, RRS560, RRS665) provide detailed insights into the optical characteristics of the water, essential for detecting phytoplankton presence and concentration. The relatively low mean values and standard deviations reflect the subtle variations in water color that can be linked to phytoplankton and other substances.

Sea surface salinity (SSS) and mixed layer depth (MLD) are critical physical parameters. SSS,

with a mean of 32.834 psu, shows moderate variability, influencing phytoplankton physiology and distribution. MLD, with a mean of 13.305 m and a higher standard deviation, reflects the dynamic nature of water column mixing, affecting nutrient availability and light conditions for phytoplankton growth.

Sea surface temperature (SST), with a mean of 277.711 K and a standard deviation of 3.699, underscores the thermal conditions that influence phytoplankton metabolic rates and species composition. The variability in SST is crucial for understanding seasonal and inter-annual changes in phytoplankton dynamics.

2.4 Pre-Processing

Preparing the dataset for a machine learning model is a crucial step that ensures the reliability and accuracy of the predictions. The data preparation process involves a series of techniques, each designed to handle different aspects of the dataset. This section provides a brief explanation of these techniques, focusing on the logic behind each step.

The initial stage involves loading two datasets, the first one obtained during the matchup extraction (Section 2.3) which includes the predictor variables, and the second one with the in-situ measurement, which are the so-called target variables (TChl-a and PFT). These datasets are read from CSV files into structured tables, allowing for efficient data manipulation.

One of the primary challenges in working with real-world datasets is handling missing values (NaN values). In this process, missing values in the predictor variables dataset due to the matchup extractions are addressed using K-Nearest Neighbours (KNN) imputation. KNN imputation predicts missing values based on the values of the nearest neighbours in the dataset. Mathematically, this involves identifying the k-nearest samples in the feature space and imputing the missing value with the mean or median of these neighbours. For example, if we have a missing value in a feature, KNN imputation will find the 20 closest samples (neighbours) in the dataset and calculate the mean of their values to replace the missing one. This method is effective because it preserves the local structure of the data and leverages the information from similar instances to fill in the gaps.

After addressing missing values, the process continues with handling zeros in the in-situ dataset. In this case the target variables dataset. Zeros can be problematic because they might represent missing data or values below the detection limit rather than actual zero measurements. To mitigate this issue, all zeros are replaced with a small positive value ($0.001 mg/m^3$). This substitution prevents distortions in the dataset that could arise from treating zeros as true values and ensures that logarithmic transformations applied later do not encounter undefined values.

Logarithmic transformation is applied to the target dataset (Figure 2.2), and to CHL in the predictor dataset (Figure 2.4). Log transformation is used to stabilise the variance, normalise the distribution, and reduce the skewness of the data. By transforming the data to a logarithmic scale, extreme values are compressed, and the overall distribution becomes more symmetrical. This is beneficial for two reasons, one is that the Chl-a concentration of PFTs, thus the TChl-a have a Gaussian distribution in the logarithmic scale (Figure 2.2), and the second reason is that machine learning algorithms assume already normally distributed input data.

Normalisation of the predictor data is performed to ensure that all features contribute equally to the model. Normalisation involves scaling the features to a standard range, typically between 0 and 1, or by transforming them to have a mean of zero and a standard deviation of one (standardisation). For example, normalisation can be expressed as:

$$\text{normalised value} = \frac{\text{original value} - \min(\text{feature})}{\max(\text{feature}) - \min(\text{feature})}$$

This scales the values to a range of [0, 1]. Standardisation can be expressed as:

$$\text{standardised value} = \frac{\text{original value} - \mu}{\sigma}$$

where μ is the mean of the feature, and σ is the standard deviation.

Once normalisation is complete, the target dataset is also normalised using a similar approach to ensure compatibility with the predictors. The normalised predictors and targets are then combined into a single dataset, ready for partitioning into training and testing sets.

The final step involves splitting the dataset into training and testing sets based on a specified proportion. This is achieved using a cross-validation partitioning method, specifically the holdout approach. A random partition of the data is created, holding out a specified proportion for testing, while the remaining data is used for training the model. This approach ensures that the model's performance can be evaluated on unseen data, providing an estimate of its generalisation ability. Mathematically, if the total number of samples is N and the test proportion is p, the training set will contain $(1-p)N$ samples, and the test set will contain pN samples.

In summary, the data preparation process involves loading the datasets, handling missing values with KNN imputation, addressing zeros, applying logarithmic transformation, normalising the data, and splitting it into training and testing sets. Each step is carefully designed to ensure that the data is clean, consistent, and suitable for training a robust machine learning model.

2.5 Model Selection

To accurately predict the concentrations of TChl-a and the different PFTs, a diverse ensemble of machine learning models was selected and trained using the prepared datasets. The selected models include the Gradient Boosting Machine (GBM), Fully Connected Neural Network (FCNN), Random Forest Regression (RFR), and Support Vector Machine (SVM). Each of these models was trained individually to capture different aspects of the data and leverage their unique strengths. After training, a final ensemble model was constructed using Ridge Regression Ensemble to combine the predictions from all four models. This ensemble approach aims to enhance the overall prediction accuracy and robustness by integrating the strengths of multiple learning algorithms.

2.5.1 Gradient Boosting Machine (GBM)

GBM is a machine learning algorithm that is optimised for regression and classification tasks. The essence of GBM lies in its ensemble approach, which sequentially builds an ensemble of weak learners, typically decision trees. Unlike traditional ensemble methods that average the predictions of individual models, GBM constructs each new model to correct the errors made by its predecessors, enhancing the overall accuracy iteratively.

The foundation of GBM is the boosting technique, which improves model performance by combining multiple simple models. In this sequential process, each model is trained to minimise the errors of the combined ensemble from the previous iterations. This iterative refinement is guided by gradient descent, a powerful optimisation method that minimises the chosen loss function. The loss function quantifies the difference between the predicted and actual values, guiding the model in learning from its mistakes.

Mathematically, GBM starts with an initial model, $f_0(x)$, which is typically a simple constant value that minimises the loss function over the training data. In each subsequent iteration m, a new model $h_m(x)$ is added to the ensemble. This new model is trained to predict the negative gradient of the loss function with respect to the current ensemble's predictions. Formally, the update at each iteration can be represented as:

$$f_m(x) = f_{m-1}(x) + \rho_m h_m(x)$$

where ρ_m is the learning rate, a parameter that controls the contribution of each new model to the ensemble. The learning rate is crucial as it balances the model's convergence speed and its ability to generalise to new data. A smaller learning rate usually requires more iterations but can lead to better generalisation, while a larger learning rate may converge faster but risk overfitting.

The core optimisation in GBM involves minimising the loss function $L(y, f(x))$, where y represents the true values, and $f(x)$ is the model's prediction. The negative gradient of the loss function, $-\frac{\partial L(y,f(x))}{\partial f(x)}$, indicates the direction in which the model should adjust its predictions to reduce the error. Each new model in the ensemble is trained to approximate this gradient, effectively learning the direction and magnitude of necessary adjustments.

An important aspect of GBM is the choice of the base learner. Decision trees are commonly used due to their flexibility and ability to handle non-linear relationships. Each tree in GBM is typically shallow, often referred to as a "stump" when it has only one split. Shallow trees ensure that each new model introduces only minor corrections, preventing overfitting and maintaining the model's generalisation ability.

To further enhance performance and prevent overfitting, GBM employs regularisation techniques such as shrinkage and subsampling. Shrinkage, implemented through the learning rate ρ, reduces the influence of each added model, promoting gradual improvements and robustness. Subsampling, on the other hand, involves training each new model on a random subset of the data, introducing diversity and reducing variance.

In practical applications, GBM has demonstrated significant success across various domains. Its flexibility in handling different types of loss functions makes it adaptable to a wide range of problems, such as predicting Phytoplankton Functional Types (PFTs) using tabular data. [26]

2.5.2 Fully Connected Neural Network (FCNN)

FCNN is a fundamental type of artificial neural network in which each neurone in one layer is connected to every neurone in the subsequent layer. This architecture enables the network to learn complex patterns and representations by allowing the neurones to interact freely across layers. FCNNs are widely used for various tasks, including regression, classification, and pattern recognition, due to their flexibility and powerful learning capabilities.

The core of an FCNN is its layered structure, typically consisting of an input layer, one or more hidden layers, and an output layer. Each layer is composed of neurones, or nodes, which perform computations and pass their outputs to the next layer. The connections between neurones are weighted, and these weights are the parameters learned during the training process. The input layer receives the raw data, which is then transformed and propagated through the network.

Mathematically, the output of a neurone j in layer l can be described by the following equation:

$$a_j^l = \phi\left(\sum_{i=1}^{n} w_{ij}^l a_i^{l-1} + b_j^l\right)$$

where a_i^{l-1} represents the activations from the previous layer, w_{ij}^l are the weights connecting neurone i in layer $l-1$ to neurone j in layer l, b_j^l is the bias term for neurone j in layer l, and ϕ is the activation function. The activation function introduces non-linearity into the network, allowing it to model complex relationships. Common activation functions include the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU).

Training an FCNN involves adjusting the weights and biases to minimise a loss function, which measures the difference between the predicted outputs and the actual targets. This optimisation is typically performed using backpropagation, an efficient algorithm that computes the gradient of the loss function with respect to each weight by applying the chain rule of calculus. The gradients

indicate the direction in which the weights should be adjusted to reduce the loss.

The backpropagation algorithm involves two main steps: forward propagation and backward propagation. During forward propagation, the input data is passed through the network, and the activations of each neurone are computed. The final activations at the output layer represent the network's predictions. The loss function $L(y, \hat{y})$, where y is the true value and $\hat{y}$ is the predicted value, is then calculated. In the backward propagation step, the gradients of the loss function with respect to each weight are computed. These gradients are used to update the weights using an optimisation algorithm, such as stochastic gradient descent (SGD):

$$w_{ij}^{l} = w_{ij}^{l} - \eta \frac{\partial L}{\partial w_{ij}^{l}}$$

where η is the learning rate, a hyperparameter that controls the size of the weight updates.

FCNNs are highly flexible and can be tailored to specific tasks by adjusting their architecture and hyperparameters. The number of hidden layers, the number of neurones per layer, and the choice of activation function can all be tuned to optimise performance. Regularisation techniques, such as dropout and weight decay, are often applied to prevent overfitting. Dropout randomly disables a fraction of neurones during training, forcing the network to learn more robust features, while weight decay penalises large weights, promoting simpler models. [27]

2.5.3 Random Forest Regression (RFR)

RFR is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees. This technique combines the strengths of decision trees while mitigating their limitations, such as overfitting. Random Forests are renowned for their robustness, accuracy, and ability to handle high-dimensional data and complex interactions among features.

At the core of RFR is the concept of ensemble learning, which involves aggregating the predictions of several base models to improve overall performance. In a Random Forest, each base model is a decision tree. Decision trees split the data at various points based on feature values to minimise a loss function, such as mean squared error (MSE) for regression tasks. The tree structure allows for capturing non-linear relationships and interactions between features.

The construction of a Random Forest involves the following key steps. First, a large number of decision trees are created using bootstrapped samples of the training data. Bootstrapping, also known as bagging, involves randomly sampling the training data with replacement, resulting in different subsets for training each tree. This introduces diversity among the trees, as each tree is trained on a slightly different dataset.

Mathematically, for a dataset D with N samples, a bootstrap sample D_b is created by sampling N times with replacement from D. This means that some samples may be repeated in D_b, while others may be omitted. Each decision tree is then trained on its respective bootstrap sample.

During the construction of each tree, a random subset of features is selected at each split point. This process, known as feature bagging, ensures that the trees are diverse not only in their training data but also in the features they consider for splitting. The number of features selected at each split is typically a user-defined parameter, often set to the square root of the total number of features.

The decision tree construction can be mathematically described as follows. Let Θ denote the set of parameters that define a particular tree, including the splits and the values at the nodes. Each tree $T(\Theta_b)$ in the forest is trained on its bootstrap sample D_b and uses a subset of features at each split to minimise the loss function. For regression, the prediction of each tree for a given input x is the average of the target values in the leaf node where x falls.

Once all trees are trained, the Random Forest makes a prediction by averaging the predictions of the individual trees. If there are M trees in the forest, the prediction for an input x is given by:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x)$$

where $T_m(x)$ is the prediction of the m-th tree. This averaging process reduces variance and improves the overall accuracy of the model.

RFR offers several advantages. The aggregation of multiple trees reduces the risk of overfitting, which is a common problem with individual decision trees. Additionally, the use of bootstrapping and feature bagging introduces randomness, leading to a more robust and stable model. Random Forests can also handle large datasets with higher dimensionality and provide estimates of feature importance, helping to identify the most influential features in the prediction task. [28]

2.5.4 Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm widely used for classification and regression tasks. SVM is designed to find the optimal hyperplane that separates data points of different classes with the maximum margin. In the context of regression, SVM is adapted to Support Vector Regression (SVR), which aims to find a function that deviates from the actual target values by a margin of tolerance.

The fundamental concept behind SVM is the idea of finding a hyperplane in an n-dimensional space (where n is the number of features) that distinctly classifies the data points. The optimal hyperplane is the one that maximises the margin, which is the distance between the hyperplane and the nearest data points from either class, known as support vectors. The maximum margin criterion ensures that the classifier is robust and has good generalisation capabilities.

For a given dataset with features $\mathbf{x}_i$ and target values y_i, where $i = 1, 2, \ldots, N$, the goal of SVM is to solve the following optimisation problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

subject to the constraints:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

where $\mathbf{w}$ is the weight vector, b is the bias term, and ξ_i are slack variables introduced to handle misclassifications. The term $\|\mathbf{w}\|^2/2$ represents the margin, and the constraints ensure that the data points are correctly classified with a margin of at least 1, allowing for some misclassification controlled by the slack variables.

In the case of Support Vector Regression (SVR), the objective is to find a function $f(\mathbf{x})$ that approximates the target values y_i within a margin of tolerance ϵ. The optimisation problem for SVR is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

subject to the constraints:

$$\begin{aligned} &y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \leq \epsilon + \xi_i \\ &(\mathbf{w} \cdot \mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0, \quad \forall i \end{aligned}$$

where C is a regularisation parameter that controls the trade-off between maximising the margin and minimising the training error, and ξ_i, ξ_i^* are slack variables that allow deviations from the

margin ϵ.

A key feature of SVM is the use of kernel functions, which enable the algorithm to handle non-linear relationships by mapping the input features into a higher-dimensional space where a linear hyperplane can effectively separate the data points. Commonly used kernel functions include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel. The kernel trick allows SVM to perform complex transformations without explicitly computing the coordinates in the higher-dimensional space, thus maintaining computational efficiency.

The choice of kernel and its parameters significantly influences the performance of the SVM model. For example, the RBF kernel, defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

introduces a parameter γ that controls the width of the Gaussian function, effectively determining the influence of a single training example.

SVM is particularly effective for high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. Its robustness to overfitting, especially in high-dimensional spaces, makes it a suitable choice for complex datasets. [29]

2.5.5 Ridge Regression Ensemble (RRE)

RRE is a powerful technique used to combine the predictions of multiple machine learning models, improving the overall accuracy and robustness of the final prediction. This method leverages the principles of ridge regression, a type of linear regression that includes a regularisation term to prevent overfitting and ensure stability in the presence of multicollinearity.

The core idea of ridge regression is to introduce a penalty on the size of the coefficients, which helps in shrinking them towards zero and thereby reducing the model complexity. Mathematically, the ridge regression model aims to minimise the following objective function:

$$\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} w_j^2 \right\}$$

where $\mathbf{w}$ is the vector of coefficients, $\mathbf{x}_i$ represents the input features, y_i is the target value, N is the number of samples, p is the number of features, and λ is the regularisation parameter. The term $\lambda \sum_{j=1}^{p} w_j^2$ is the ridge penalty, which discourages large coefficients and thereby prevents overfitting.

In the context of ensemble learning, RRE can be used to combine the outputs of several base models. The ensemble model predicts the target value as a weighted sum of the predictions from the individual models. Suppose there are M base models, each providing a prediction $f_m(x)$ for an input x. The ensemble prediction $\hat{y}$ can be expressed as:

$$\hat{y} = \sum_{m=1}^{M} w_m f_m(x)$$

where w_m are the weights assigned to the predictions of the base models. The goal is to find the optimal weights that minimise the prediction error on the training data while also applying the ridge penalty to avoid overfitting.

The optimisation problem for the ridge regression ensemble can be formulated as:

$$\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} w_m f_m(x_i) \right)^2 + \lambda \sum_{m=1}^{M} w_m^2 \right\}$$

This formulation ensures that the ensemble model not only fits the training data well but also maintains generalisability to new data.

RRE is particularly effective when combining models that are diverse in their predictions. By weighting the predictions of different models, the ensemble can capture various aspects of the data that individual models might miss. This diversity is crucial in reducing the overall variance and improving the stability of the predictions.

In this study, RRE is used to combine the outputs of four different models: GBM, FCNN, RFR, and SVM. Each of these models brings unique strengths to the ensemble. [30]

2.6 Accuracy Assessment

2.6.1 Performance Metrics

To evaluate the performance of the machine learning models used in predicting Chl-a concentration of PFTs, four key metrics are employed: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), the coefficient of determination (R^2), and the Median Absolute Percentage Error (MdAPE). Each metric provides unique insights into the model's accuracy and reliability.

Root Mean Squared Error (RMSE)

The RMSE is a widely used metric for measuring the differences between predicted and actual values. It is defined as the square root of the average of the squared differences between the predicted values ($\hat{y}_i$) and the actual values (y_i):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where N is the number of observations. RMSE gives a sense of how concentrated the data is around the line of best fit. It is particularly useful because it penalises larger errors more than smaller ones, due to the squaring of the differences. Lower RMSE values indicate a better fit of the model to the data. This metric is sensitive to outliers, which can disproportionately affect the result, making it a comprehensive measure of model performance.

Mean Absolute Error (MAE)

The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as the average of the absolute differences between the predicted values and the actual values:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

MAE provides a straightforward interpretation of the prediction error, as it represents the average absolute deviation of the predictions from the actual values. Unlike RMSE, MAE does not penalise larger errors more than smaller ones, providing a more balanced view of the model's prediction accuracy. Lower MAE values indicate better predictive performance. This metric is less sensitive to outliers compared to RMSE, making it a useful measure for understanding the typical size of the errors.

Coefficient of Determination (R^2)

The coefficient of determination, R^2, is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values. R^2 values range from 0 to 1, with higher values indicating a better fit. An R^2 value of 1 signifies that the model perfectly explains the variance in the data, while an R^2 value of 0 indicates that the model does not explain any of the variance. R^2 provides an intuitive measure of the model's explanatory power, showing how well the predictions match the observed data.

Median Absolute Percentage Error (MdAPE)

The MdAPE is a robust measure of prediction accuracy that is less sensitive to outliers compared to RMSE and MAE. It is defined as the median of the absolute percentage errors between the predicted values and the actual values:

$$\text{MdAPE} = \text{median}\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\right)$$

MdAPE provides an easy-to-understand percentage error metric, representing the typical prediction error in percentage terms. Lower MdAPE values indicate better predictive accuracy. This metric is particularly useful for comparing model performance across datasets with different scales and is less affected by extreme values than the mean absolute percentage error (MAPE).

Interpretation of Metrics

These performance metrics together offer a comprehensive view of the model's accuracy and reliability. RMSE provides insight into the magnitude of the prediction errors, heavily penalising larger deviations, and is useful for understanding the overall fit of the model. MAE offers a straightforward measure of average error, giving a balanced view of prediction accuracy without disproportionately weighting larger errors. R^2 complements these metrics by indicating the proportion of variance explained by the model, providing a clear measure of its explanatory power. MdAPE offers a robust percentage-based error metric that is less sensitive to outliers and provides an intuitive interpretation of typical prediction error.

In the context of predicting Phytoplankton Functional Types (PFTs), these metrics help in evaluating how well the models capture the underlying patterns in the data. Low RMSE and MAE values, along with high R^2 values and low MdAPE values, indicate that the model performs well, accurately predicting the PFTs based on the given features. By using these metrics, the effectiveness and robustness of the machine learning models can be quantitatively assessed, ensuring reliable and accurate predictions.

2.6.2 Cross-validation approach

Cross-validation is a fundamental technique in machine learning used to assess the generalisability and robustness of a model. In this study, a 5-fold cross-validation approach is employed, which involves training the model five times with different holdout sets and then averaging the results to obtain a comprehensive evaluation. This method splits the dataset into a training set and a test set, typically using 70% of the data for training and 30% for testing. This process is repeated five times, each time with a different random split, ensuring that the model's performance is evaluated on multiple data splits.

One primary advantage of cross-validation is its ability to reduce overfitting by training and validating the model on different subsets of data, thereby providing a more reliable estimate of the model's ability to generalise. This method ensures every data point is used for both training and validation, maximising the efficient use of available data, which is especially beneficial in fields with limited data availability, like environmental science. Additionally, cross-validation is essential for model selection and hyperparameter tuning, helping identify the best-performing model configuration. By using cross-validation to test and tune hyperparameters, the chosen model and its parameters are better suited to the data, leading to improved predictive performance. The variability in performance metrics across iterations also offers insights into the model's stability and robustness, highlighting areas for potential improvement.

2.7 Validation and PFT mapping

2.7.1 In-situ Data

The validation of the ensemble model developed in this study was conducted using an independent in situ PFT data set collected from PS131 expedition. The PS131 expedition, also known as ATWAICE, was carried out from June 27, 2022, to August 17, 2022, aboard the German research vessel Polarstern. The primary objectives of the expedition included investigating ocean-ice-air interactions, studying the impacts of climate change on the Arctic environment, and collecting data on various oceanographic and biological parameters [1].

The in-situ measurements were collected from June 29, 2022, to August 12, 2022, covering a period of 45 days (Figure 2.5 and Figure 2.6). Using the temporal range of this data, corresponding predictor variables were extracted from the CMEMS dataset for the same 45-day period. The CMEMS dataset included the same parameters as in the predictor dataset in the training phase (Table 2.2). But unlike the training phase, where only matching points based on in-situ locations and dates were used, the validation involved downloading all the data points of the Arctic over the 45-day period.

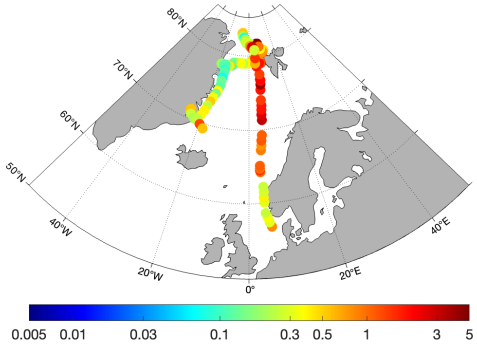

Figure 2.5: PS131 in-situ measurements location - TChl-a (mg/m^3) – 45 days.

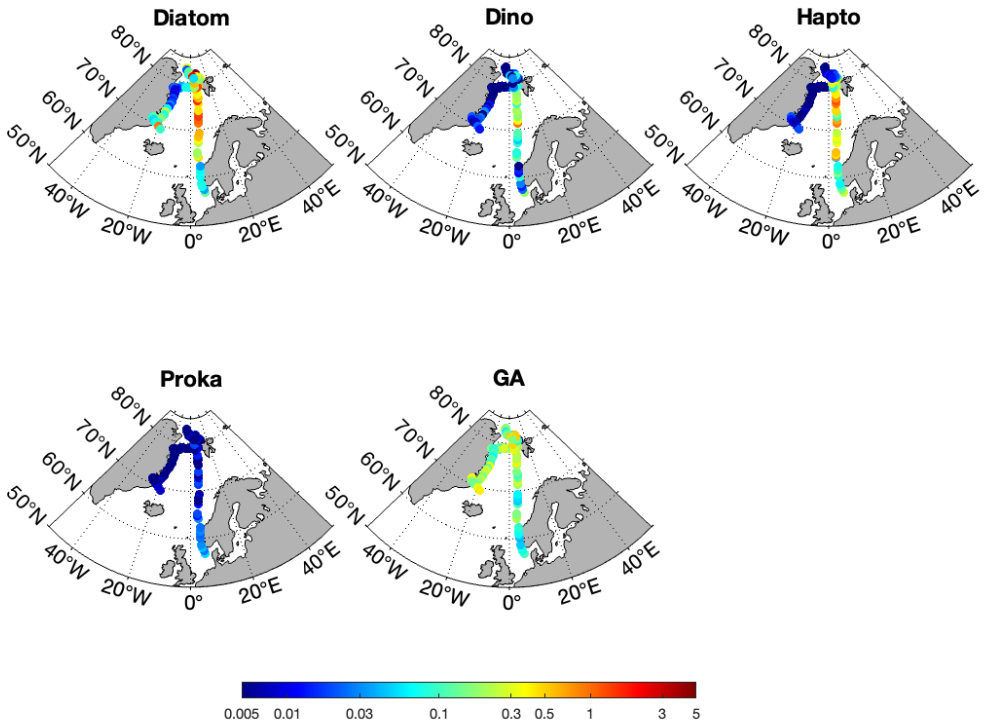

Figure 2.6: PS131 in-situ measurements location - Chl-a PFT (mg/m^3) – 45 days.

2.7.2 CMEMS Products

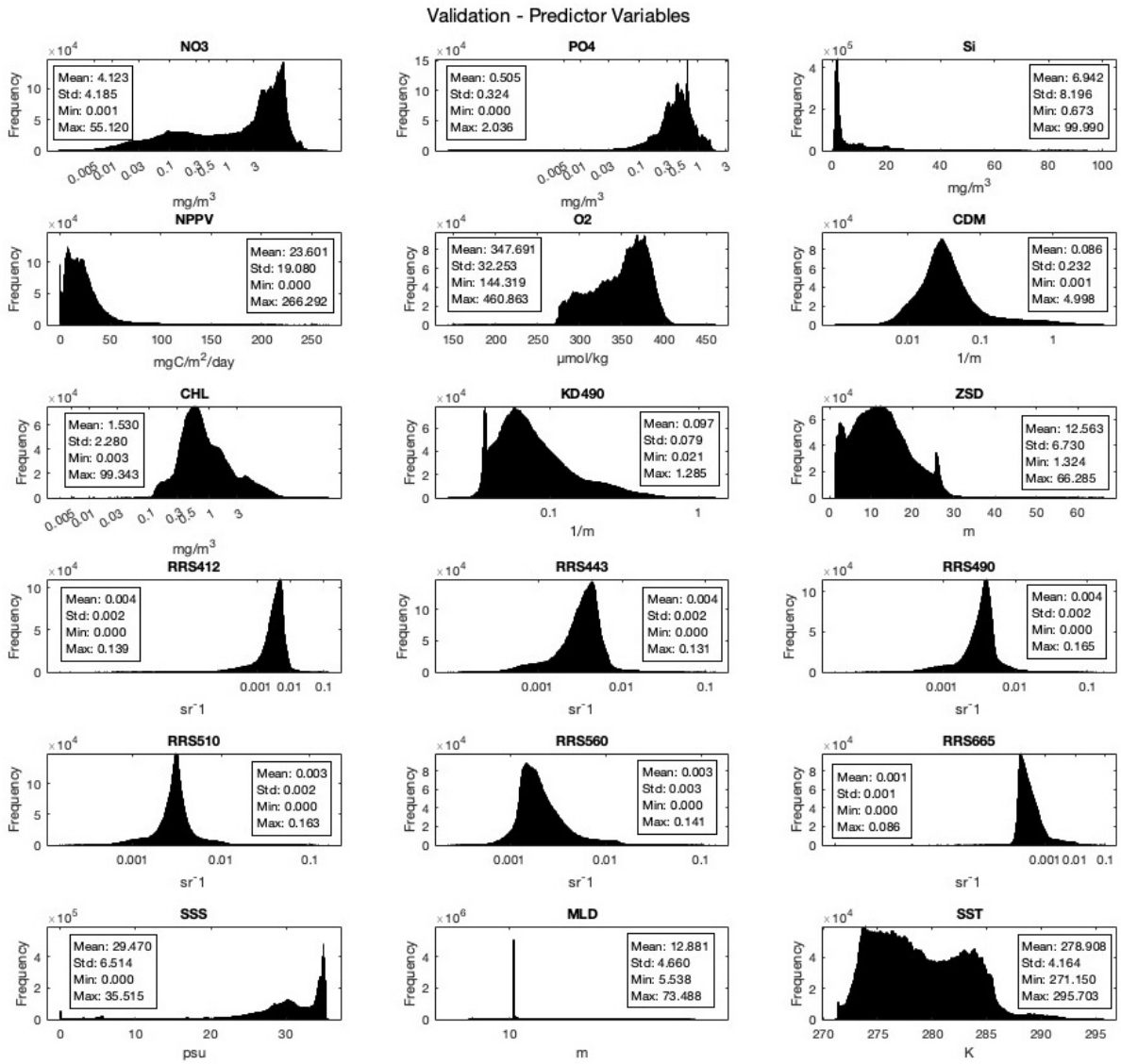

Figure 2.7: Histogram Descriptive Statistics - Validation Predictor Variables from CMEMS Dataset – 45 days.

The descriptive statistics of the validation predictor variables dataset, illustrated in Figure 2.7, obtained from the CMEMS products listed in the table 2.2, provide critical insights into the environmental conditions during the validation period. For instance, nitrate (NO_3) concentrations in the validation dataset have a mean of 4.123 with a higher standard deviation of 4.185 compared to the training dataset, indicating more variability and a broader range of nutrient conditions. The maximum value for nitrate in the validation set is significantly higher, suggesting the presence of regions with extremely high nutrient inputs.

Table 2.2: Validation - Copernicus Marine Service Datasets - Daily Resolution.

Dataset ID	**Variables**	**Horizontal Resolution**	**Source**
cmems_mod_glo_bgc-nut_anfc_0.25_P1D-m	NO_3, PO_4, Si	$\sim 25km \times 25km$	Model
cmems_mod_glo_bgc-bio_anfc_0.25_P1D-m	$NPPV$, O_2	$\sim 25km \times 25km$	Model
c3s_obs-oc_glo_bgc-plankton_my_l3-multi-4km_P1D	CHL	$4km \times 4km$	Observations
c3s_obs-oc_glo_bgc-reflectance_my_l3-multi-4km_P1D	$RRS412$, $RRS443$, $RRS490$, $RRS510$, $RRS560$, $RRS665$	$4km \times 4km$	Observations
cmems_obs-oc_glo_bgc-transp_my_l4-gapfree-multi-4km_P1D	$KD490$, ZSD	$4km \times 4km$	Observations
cmems_obs-oc_glo_bgc-optics_my_l3-multi-4km_P1D	CDM	$4km \times 4km$	Observations
cmems_mod_glo_phy-so_my_0.083deg_P1D-m	SSS	$\sim 10km \times 10km$	Model
cmems_mod_glo_phy_anfc_0.083deg_P1D-m	MLD	$\sim 10km \times 10km$	Model
METOFFICE-GLO-SST-L4-REP-OBS-SST	SST	$\sim 5km \times 5km$	Observations

Phosphate (PO_4) concentrations in the validation dataset show a mean of 0.505, slightly higher than in the training dataset, with a higher variability indicated by the standard deviation of 0.324. This suggests varying nutrient dynamics during the validation period. Silicate (Si) concentrations exhibit a mean of 6.942 and a high standard deviation of 8.196, highlighting substantial variability and the presence of areas with high diatom growth potential.

Net primary productivity of vegetation (NPPv) has a mean of 23.601 with a high standard deviation of 19.080, reflecting dynamic productivity levels and substantial variability in phytoplankton growth conditions. Oxygen (O_2) levels, with a mean of 347.691 and a standard deviation of 32.253, remain relatively stable, providing a consistent measure of the marine environment's health during the validation period.

The optical properties, such as CDM, CHL, KD490, and ZSD, highlight variations in water clarity and phytoplankton biomass. The mean CHL concentration in the validation dataset is 1.530 mg/m^3, with a higher standard deviation of 2.280, indicating significant fluctuations in phytoplankton biomass. The higher mean and variability compared to the training dataset suggest more dynamic phytoplankton activity during the validation period.

Remote sensing reflectance values at various wavelengths (RRS412, RRS443, RRS490, RRS510, RRS560, RRS665) in the validation dataset show slightly higher means and variability, reflecting changes in water optical properties and possibly different phytoplankton compositions and concentrations.

Sea surface salinity (SSS) has a mean of 29.470 with a higher standard deviation of 6.514, indicating substantial changes in salinity levels, possibly due to melting ice or freshwater inputs. The mixed layer depth (MLD) shows a mean of 12.881 with a standard deviation of 4.660, suggesting variations in water column mixing and nutrient availability. Sea surface temperature (SST), with a mean of 278.908 K and a higher standard deviation of 4.164, underscores the thermal variability

influencing phytoplankton growth rates and species composition.

These descriptive statistics of the validation predictor variables highlight the diverse and dynamic environmental conditions during the validation period. Comparing these with the training dataset reveals important differences, such as higher variability in nutrients and optical properties, which can affect the performance of the ensemble model. The robust preprocessing and alignment of these variables ensure that the model's predictions are based on accurate and consistent data, enhancing the reliability of the validation results.

2.7.3 Pre-Processing

Since the products including all predictor variables acquired from CMEMS files had different resolutions, interpolation was necessary to align the data spatially. This interpolation ensured that all predictor variables were available at the same spatial resolution, providing consistent input for the model. Then, common pixels between the CMEMS data were identified using the remote sensing reflectance at 665 nm (RRS665) as reference, as it had the fewest points. This step was crucial for ensuring that the spatial points from different datasets corresponded correctly and all predictor files contained the same number of location points corresponding to the identified common pixels.

Interpolation

This section involved the interpolation of geospatial data from various sources with different spatial resolutions (Table 2.2) to a unified 4km grid, subsequently storing the interpolated data in new files. The process began by reading the latitude and longitude data of the target 4km grid from a 4km resolution reference file, followed by setting up grids for the variables to be interpolated from their respective original resolutions. For each variable group such as salinity, layer thickness, nutrients, bio, and surface temperature dimensions and variables were defined in new files to store the interpolated data.

The interpolation process for each variable involved creating a mesh grid based on the original resolution, defining the necessary dimensions and variables in the target file, and iterating over the time steps to read slices of the original data. These data slices were then interpolated to the 4km grid using linear interpolation methods and subsequently written to the new file. This approach ensures that all data, regardless of its original resolution, could be compared or combined on a consistent 4km grid, thus facilitating further analysis or visualisation.

Common Pixels

To prepare the dataset for analysis, geospatial data from various CMEMS products was read and filtered based on a reference variable. The reference variable, "RRS665", was selected due to its relevance in the dataset and was used to determine valid entries. The latitude and longitude values associated with "RRS665" were used to create a mesh grid representing the geospatial coordinates. A logical mask was generated to identify valid (non-NaN) entries in the "RRS665" data, which served as a basis for filtering out invalid (NaN) values from other variables in subsequent steps.

The next step involved defining a list of predictor variables from the different CMEMS products, each associated with specific filenames and variable names. The filtering process applied the non-NaN mask to each predictor variable to ensure that only valid data points were retained. The corresponding data was then read from each file, and non-NaN entries were preserved. These filtered values were combined to create a comprehensive dataset, and any remaining rows containing NaN values were removed to ensure completeness. The consolidated data, including latitude, longitude, and all predictor variables without NaN values, was then organised into a single table, ready for further analysis or visualisation.

After these preprocessing steps, the dataset was prepared similarly to the training phase, including normalisation and transformation processes. The predictor variables were then input into the trained ensemble model to generate daily predictions. These daily predictions were averaged

to create a comprehensive map of the Arctic for Chl-a PFTs and TChl-a over the entire period.

2.7.4 PFT Mapping

To generate maps of the Chl-a PFTs and TChl-a, the daily distributions were predicted using the ensemble model, and these predictions were averaged over a specified number of days to produce the resulting maps. The process began with defining the predictor and target variables, followed by loading the trained models and the coefficients of the ridge regression ensemble model. An initial storage framework was established for daily data predictions and their corresponding coordinates.

For each day within the defined period, data from CSV files was loaded and preprocessed. Multiple machine learning models (GBM, FCNN, RFR, SVM) were applied to generate predictions for each PFT. These predictions were then combined using an ensemble method to enhance robustness. The predictions and their associated geospatial coordinates were stored for each day to maintain continuity over the period.

Subsequently, reference coordinates and the accumulated predictions were loaded, and storage was initialised to accumulate the log-transformed values of the predictions. For each PFT, the predicted values were log-transformed and accumulated over the days according to their corresponding coordinates. A nearest neighbour search was employed to match the prediction coordinates with the reference grid. The mean log-transformed values for each grid point were then calculated, with any grid points lacking data marked as missing. Finally, the averaged predictions were plotted using an Arctic projection to create the maps.

2.7.5 PFT Validation using in-situ Matchups

The goal of the validation was to assess the performance of the model by comparing its predictions with the in-situ measurements. This involved matching the predicted data with the in-situ measurements based on location (latitude and longitude) and date and applying the same metrics as during the training phase.

3. Results and Discussions

3.1 Training Phase Performance

To analyse the performance of the Ridge Regression Ensemble model, it is crucial to delve into the specific metrics for each PFT and TChl-a. This detailed analysis provides insights into the strengths and weaknesses of the ensemble approach and highlights areas for potential improvement.

Analysis of Individual Models - Table 3.1

GBM demonstrated notable results in predicting the concentration of various PFTs and TChl-a. Specifically, GBM achieved the highest R^2 values for Hapto and TChl-a, indicating a strong predictive capability for these categories. However, its performance was less effective for Proka and Diatoms, suggesting that the model struggled to accurately capture the dynamics of these PFTs in the Arctic Ocean. The relatively higher RMSE and MAE values for these less effective categories indicate larger prediction errors. This can be partly attributed to the lower mean and higher variability of these PFTs in the training dataset, as indicated by the descriptive statistics: Diatoms (mean: 0.389 mg/m^3, std: 0.500 mg/m^3) and Prokaryotes (mean: 0.008 mg/m^3, std: 0.014 mg/m^3).

FCNN, known for its versatility and power in various applications, showed varying degrees of success in this context. The highest R^2 values were observed for TChl-a and Hapto, indicating reasonable effectiveness in these areas. However, the performance significantly dropped for Proka and Diatoms, highlighting potential limitations in the model's ability to generalise across different PFTs. The increased RMSE and MAE values for these PFTs suggest that the FCNN had difficulty in making precise predictions, possibly due to the low concentrations and high variability of these groups in the dataset.

SVM presented a balanced performance across different PFTs. It achieved its best results for Hapto and TChl-a, with moderate R^2 values indicating satisfactory predictive accuracy. Nonetheless, the values for Diatoms and Proka were relatively lower, indicating moderate predictive accuracy. The relatively low RMSE and MAE values for these PFTs suggest that while the SVM model was generally effective, there was still room for improvement. The variability in the predictor variables, such as nutrients and temperature, likely influenced these results.

RFR exhibited robust performance, particularly for Hapto and TChl-a. This model demonstrated its capability to handle non-linear relationships and interactions between features. However, similar to other models, the R^2 values for Diatoms and Proka indicated room for improvement. The relatively low RMSE and MAE values for these PFTs suggest that RFR was effective in making precise predictions but struggled with certain PFTs due to their lower abundance or complex ecological roles.

Ridge Regression Ensemble Analysis - Table 3.2

The Ridge Regression Ensemble model, which combines the strengths of the individual models, showed the highest overall performance. For Diatoms, the ensemble model achieved an R^2 of 0.814, indicating a strong predictive capability, likely due to the high abundance and well-defined seasonal cycles of Diatoms in the Arctic. The relatively low RMSE and MAE values suggest that the model is able to make precise predictions with minimal error. However, it is crucial to consider the MdAPE value as well, which provides a robust measure of prediction accuracy less sensitive to

outliers.

For Dino, the ensemble model achieved an R^2 of 0.612. While the R^2 value is lower than that for Diatoms, the relatively low RMSE and MAE values might indicate that the variance in the actual measurements of Dino is low, leading to smaller error margins. This could suggest that the model's predictions are not as challenged by the data variability, potentially masking underlying prediction difficulties. The MdAPE value here helps confirm the model's reliability by providing an intuitive percentage-based error metric.

Hapto exhibited an R^2 of 0.745. The high R^2 value reflects the model's strong ability to predict the concentrations of Hapto. The relatively low RMSE and MAE values further emphasise the accuracy of the model in predicting this PFT, likely due to their moderate abundance and distinct physiological characteristics. The MdAPE value corroborates this accuracy, indicating robust performance.

Proka had an R^2 of 0.641. Although the R^2 value is lower compared to other PFTs, the exceptionally low RMSE and MAE values suggest that the model might be overfitting to specific characteristics of the data, or that the variance in Proka' measurements is low. This indicates that while predictions appear precise, they may not be robust. Therefore, a focus on R^2 and MdAPE values provides a better understanding of the model's generalisation capabilities and avoids misleading interpretations based on RMSE and MAE alone.

GA showed an R^2 of 0.696. These metrics suggest a moderate level of predictive accuracy, with the model performing reasonably well. The lower abundance of GA compared to Diatoms and Hapto likely contributed to the lower R^2 value, but the relatively low RMSE and MAE values still indicate that the model can make accurate predictions. The MdAPE value here again provides a more reliable measure of prediction accuracy.

For TChl-a, the ensemble model achieved an R^2 of 0.809. The high R^2 value reflects the model's robust performance in predicting overall chlorophyll concentrations, which is crucial for understanding primary production in the Arctic Ocean. The RMSE and MAE values indicate a strong predictive accuracy, essential for comprehensive ecosystem monitoring. The MdAPE value further supports this accuracy, ensuring that the model's predictions are reliable.

Our model performance is also compared to the previous global PFT retrieval models developed based on larger global data sets, such as those by Xi et al. [13] and the STEE model [9]. Xi et al. (2021) employed an EOF-based approach for global PFT retrievals and achieved good performance metrics, with R^2 values of 0.82 for Total Chlorophyll-a (TChl-a), 0.71 for Diatoms, and 0.53 for Green Algae (GA), among other PFTs. However, this EOF-based approach may not be optimally suited for Arctic-specific conditions due to its reliance on global data sets that may not capture the unique environmental characteristics of the Arctic, such as the presence of sea ice, strong seasonal light variations, and specific nutrient dynamics.

In contrast, our machine learning-based approach demonstrates several advantages over the EOF-based method of Xi et al. (2021) when applied to the Arctic region. Our ensemble model, specifically trained on Arctic data, achieved an R^2 value of 0.809 for TChl-a, 0.814 for Diatoms, and 0.696 for GA. These results indicate that our model is able to account for the localised environmental factors and data sparsity characteristic of the Arctic, providing improved retrievals of PFT concentrations in this challenging region.

When compared with the STEE model, which achieved R^2 values higher than 0.6 for all eight PFTs globally (with a maximum R^2 of 0.88 for Diatoms), our ensemble model's performance is slightly downgraded but still comparable, especially given the more limited data availability and the specific regional focus of the Arctic Ocean. For example, our model achieved an R^2 of 0.814 for Diatoms, close to the global STEE model's performance, and comparable R^2 values for other PFTs such as Haptophytes and Prokaryotes (0.745 and 0.641, respectively). This demonstrates that, despite the reduced amount of data, the regional training tailored to the Arctic Ocean context

Table 3.1: Training Performance Metrics - Individual Models.

PFT	RMSE	MAE	MdAPE	R^2
Gradient Boosting Machine				
Diatom	0.357	0.203	40.433	0.586
Dino	0.076	0.038	43.881	0.535
Hapto	0.199	0.109	39.115	0.770
Proka	0.0114	0.005	32.970	0.499
GA	0.179	0.108	37.089	0.524
TChl-a	0.402	0.242	19.901	0.791
Fully Connected Neural Network				
Diatom	0.411	0.233	46.755	0.505
Dino	0.091	0.048	67.527	0.446
Hapto	0.264	0.154	63.129	0.604
Proka	0.013	0.005	49.317	0.266
GA	0.205	0.122	42.623	0.427
TChl-a	0.470	0.294	31.424	0.774
Support Vector Machine				
Diatom	0.344	0.193	36.034	0.570
Dino	0.083	0.043	59.492	0.504
Hapto	0.205	0.117	37.903	0.716
Proka	0.008	0.005	41.796	0.462
GA	0.143	0.083	28.427	0.469
TChl-a	0.414	0.259	24.202	0.783
Random Forest Regression				
Diatom	0.302	0.169	32.468	0.612
Dino	0.071	0.035	49.301	0.546
Hapto	0.180	0.094	29.969	0.789
Proka	0.011	0.004	34.999	0.496
GA	0.152	0.086	29.329	0.561
TChl-a	0.399	0.249	22.870	0.773

is still valid and capable of providing meaningful insights into phytoplankton dynamics.

Overall, while our model's performance is slightly lower than the global models, the results highlight the benefits of using a regionally focused machine learning approach to enhance the understanding of PFT distributions in the Arctic, thereby underscoring its utility for ecological monitoring in areas with unique environmental conditions and limited data.

Table 3.2: Validation Performance Metrics - Ensemble Model.

Ridge Regression Ensemble				
PFT	**RMSE**	**MAE**	**MdAPE**	R^2
Diatom	0.292	0.164	32.605	0.814
Dino	0.075	0.0372	47.380	0.612
Hapto	0.171	0.094	28.080	0.745
Proka	0.011	0.004	26.120	0.641
GA	0.132	0.079	30.302	0.696
TChl-a	0.380	0.224	19.568	0.809

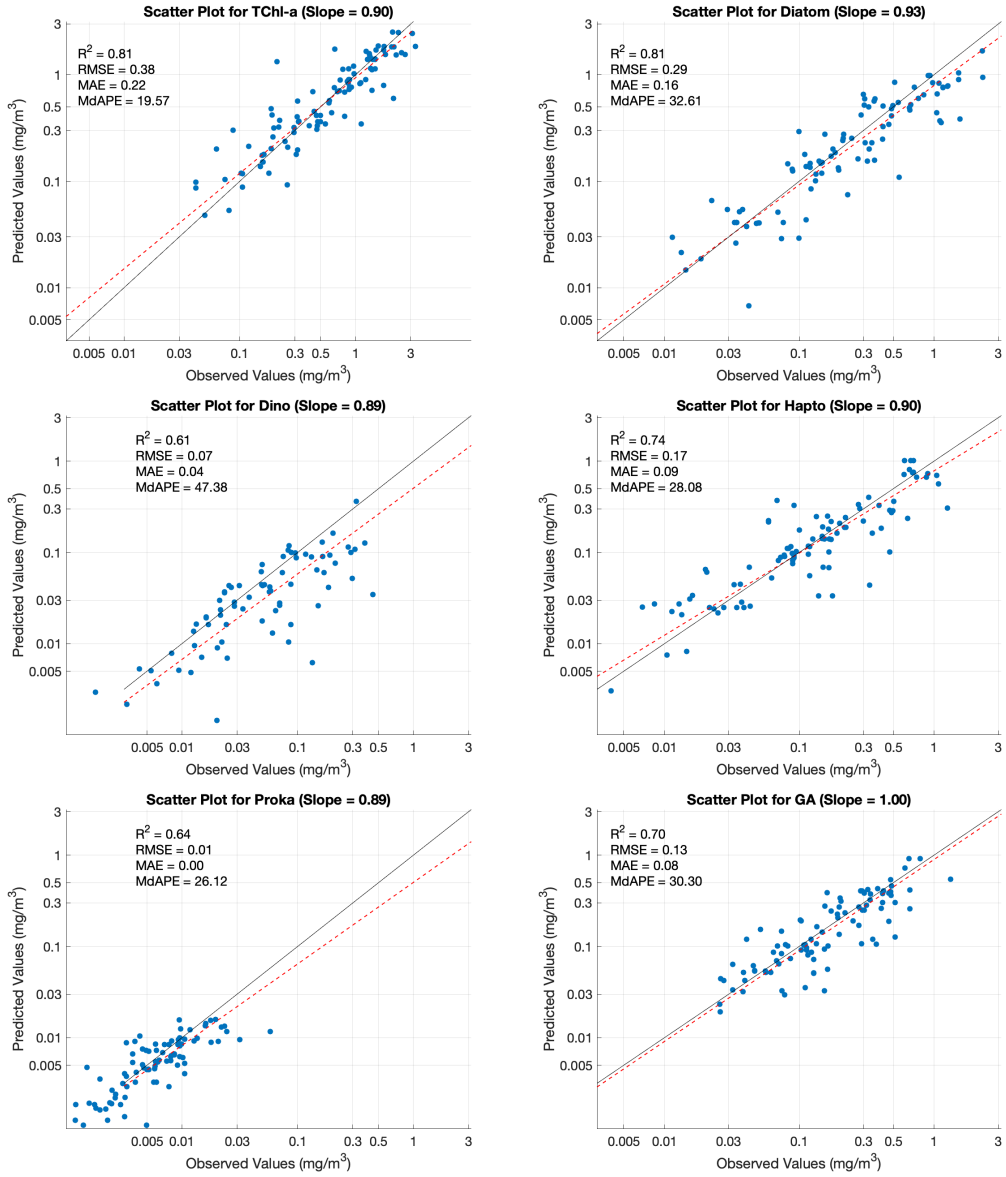

Figure 3.1: Scatterplots of predicted vs observed TChla and PFT Chla values (Black line 1:1 fit).

3.2 Validation Analysis

To validate the ensemble model, a series of systematic steps were undertaken to ensure the accuracy of the predictions. Initially, predictor variables (Fig. 2.7) parameters were prepared. These predictors were used to generate daily predictions for 45 days using pre-trained models, GBM, FCNN, RFR, and SVM. These individual model predictions were then combined using a RRE approach using the pre-loaded coefficients from the training phase.

The validation process included matching these predictions with in-situ measurements collected during the PS131 expedition. A threshold of 0.04 degrees, corresponding to the spatial resolution of the prepared CMEMS products, was applied to ensure that only predictions within this radius from the in-situ data points were considered. This step was crucial for ensuring spatial relevance and accuracy. For each in-situ point, the nearby predicted values were averaged, allowing for a meaningful comparison. The final step involved calculating performance metrics as for the training performance such as RMSE, MAE, and MdAPE to quantitatively assess the model's prediction accuracy. These metrics provided insights into the model's performance, highlighting both its strengths and areas for potential improvement, and make easier the comparison with the training phase performance and other validated models.

To assess the performance of the ensemble model, it is essential to compare the validation in-situ dataset with the training in-situ dataset.

3.2.1 Comparison of Training and Validation Datasets

The descriptive statistics indicate notable differences between the training and validation datasets, which could impact the model's predictive capabilities. For instance, the mean and maximum values of TChl-a are higher in the validation dataset compared to the training dataset, suggesting more variable environmental conditions during the validation period. This variability could pose a challenge for the model, potentially affecting its accuracy.

Comparison:

- Diatoms also exhibit higher mean and maximum values in the validation dataset, indicating a more substantial presence and variability during the validation period. This increased variability in Diatom concentrations could impact the model's performance, as it may need to account for a broader range of values.
- Dino show a slight decrease in mean concentration in the validation dataset, but with a few higher outliers. This could suggest sporadic blooms or specific conditions favouring Dinoflagellates during the validation period, which the model needs to handle effectively.
- Hapto and Proka present lower mean concentrations in the validation dataset compared to the training dataset. The lower abundance of these PFTs in the validation dataset could potentially simplify the model's predictions but also require it to be sensitive to smaller concentrations.
- GA exhibit a higher mean concentration in the validation dataset, which could reflect environmental conditions that favour their growth during the validation period. The model's ability to accurately predict GA concentrations under these conditions is crucial for its overall performance.

Overall, the differences between the training and validation datasets highlight the importance of ensuring that the model is robust and capable of generalising across different environmental conditions. The variability in the validation dataset, particularly the higher values for some PFTs, poses a significant challenge that the model needs to overcome to demonstrate its effectiveness.

3.2.2 Validation Performance

To assess the performance of the Ridge Regression Ensemble model, it is essential to examine the metrics obtained during the validation phase and compare them with the training performance metrics, as shown in Table 3.3 and the scatter plots in the figure 3.2, in which can be seen the latitude of the matching points. Additionally, considering the differences in the descriptive statistics of the validation and training in-situ datasets, it is expected that the model will face challenges due to these variations.

Diatoms: The RRE model achieved an R^2 of 0.463 for Diatoms during the validation phase, a significant decrease from the training phase R^2 of 0.814. This drop can be attributed to the higher mean (1.080 mg/m^3) and increased variability (standard deviation of 2.212 mg/m^3) of Diatoms in the validation dataset compared to the training dataset (mean of 0.389 mg/m^3 and standard deviation of 0.500 mg/m^3). The higher RMSE and MAE values during validation further indicate increased prediction errors, reflecting the model's struggle with the more diverse validation data. Such a drop is within the expected range, considering the significant difference in the data distributions.

Dino: For Dino, the R^2 during validation is 0.332, lower than the training phase R^2 of 0.612. The validation dataset had a lower mean (0.036 mg/m^3) and a similar standard deviation (0.078 mg/m^3) compared to the training dataset (mean of 0.057 mg/m^3 and standard deviation of 0.081 mg/m^3). The increased RMSE and MAE values indicate that the model's predictions for Dino are less accurate during validation, possibly due to the sporadic higher values observed in the validation dataset. This drop in performance is typical for models when dealing with less abundant and more variable datasets.

Hapto: The model's R^2 for Hapto during validation is 0.304, a decrease from the training phase R^2 of 0.745. The validation dataset showed a lower mean (0.120 mg/m^3) and a similar standard deviation (0.215 mg/m^3) compared to the training dataset (mean of 0.197 mg/m^3 and standard deviation of 0.238 mg/m^3). The higher RMSE and MAE values during validation suggest that the lower mean concentration and variability in the validation dataset posed a challenge for the model, leading to decreased performance. This drop in performance is expected due to the variability in the validation dataset.

Proka: The R^2 for Proka during validation is 0.483, which is relatively close to the training phase R^2 of 0.641. The validation dataset had a lower mean (0.004 mg/m^3) and a similar standard deviation (0.008 mg/m^3) compared to the training dataset (mean of 0.008 mg/m^3 and standard deviation of 0.014 mg/m^3). The consistent RMSE and MAE values suggest that the model maintained its predictive accuracy for Proka despite the lower mean concentration and reduced variability in the validation dataset. The small drop in performance indicates the model's robustness for this PFT, likely due to the low abundance of Prokaryotes in both datasets, which resulted in less variability to influence the model's performance.

GA: The model achieved an R^2 of 0.563 for Green Algae during validation, slightly lower than the training phase R^2 of 0.696. The validation dataset showed a higher mean (0.253 mg/m^3) and a similar standard deviation (0.239 mg/m^3) compared to the training dataset (mean of 0.193 mg/m^3 and standard deviation of 0.191 mg/m^3). The slightly higher RMSE and MAE values during validation indicate a marginal decrease in predictive accuracy, likely due to the increased mean concentration of GA in the validation dataset. This performance drop is within the expected range.

TChl-a: The R^2 for TChl-a during validation is 0.744, compared to the training phase R^2 of 0.809. The validation dataset had a higher mean (1.533 mg/m^3) and increased variability (standard deviation of 2.363 mg/m^3) compared to the training dataset (mean of 0.922 mg/m^3 and standard deviation of 0.802 mg/m^3). The higher RMSE and MAE values during validation indicate less accurate predictions, reflecting the model's struggle with the more variable validation data. However, the relatively high R^2 value indicates that the model performed well despite the increased variability, which is a positive outcome.

Expected Performance Drop: Performance metrics can vary significantly between training and validation datasets, which could lead to the expected drops in performance metrics like R^2, depending on the complexity and variability of the data [31]. In this study, the drop in performance is within this expected range, indicating that the model generalises reasonably well despite the challenges posed by the validation dataset.

Table 3.3: Validation Performance Metrics - Ensemble Model.

Ridge Regression Ensemble				
PFT	**RMSE**	**MAE**	**MdAPE**	R^2
Diatom	0.268	0.178	48.512	0.463
Dino	0.048	0.037	53.532	0.332
Hapto	0.311	0.248	73.477	0.304
Proka	0.011	0.009	23.787	0.483
GA	0.062	0.042	18.613	0.563
TChl-a	0.380	0.289	36.233	0.744

Overall, the performance metrics indicate that the Ridge Regression Ensemble model experienced a decrease in predictive accuracy during the validation phase compared to the training phase. This is expected due to the increased variability and higher mean concentrations in the validation dataset compared to the training dataset. The model's performance reflects the challenges of generalising across different datasets and highlights the importance of considering dataset variability and ensuring the model's robustness across varying environmental conditions.

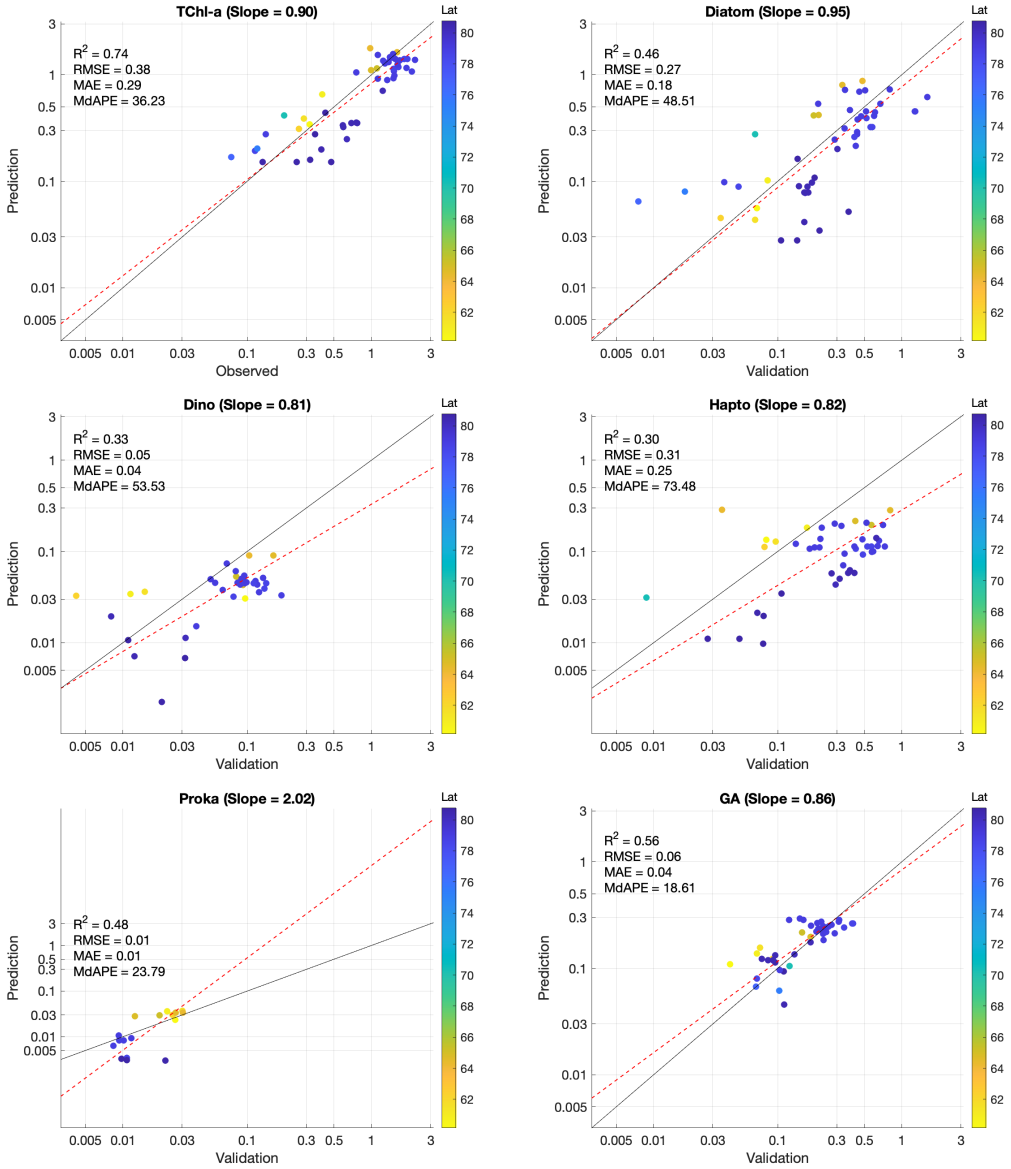

Figure 3.2: Scatterplots showing Predicted vs. Observed Values using the independent validation data set from PS131 (mg/m^3) (Black line 1:1 fit).

3.3 Mapping of the Arctic PFTs

In order to map the PFT distribution for the Arctic Ocean, daily predicted values for each PFT were stored along with their corresponding geographic coordinates. This allowed for the accumulation of predictions across the 45-day period. The next step involved calculating the mean log-transformed values of these predictions. This transformation was performed to mitigate the effect of outliers and stabilise the variance. The log-transformed values were then averaged over the entire period, providing a comprehensive representation of the predicted TChl-a and PFT concentrations as can be seen in Figure 3.3 and Figure 3.4, respectively.

The distribution of TChl-a and various PFTs in the Arctic can be influenced by multiple environmental factors, including nutrient availability, light conditions, and temperature. Generally, we expect to find higher concentrations of Diatoms near coastal areas where nutrient input is higher due to upwelling and river discharges. Diatoms thrive in nutrient-rich waters, making coastal zones ideal habitats. Conversely, open waters, particularly in higher latitudes, might show lower

concentrations of Diatoms.

GA are typically found in both coastal and open waters but are more sensitive to changes in light and nutrient conditions. Their distribution can be quite patchy, reflecting localised variations in these factors. The maps show varying concentrations of GA across different regions, influenced by local environmental conditions. Specifically, GA are more present along the coasts of northern Europe and northern Asia, particularly in the Barents Sea, Kara Sea, Laptev Sea, and parts of the East Siberian Sea. These areas likely offer the optimal light and nutrient conditions that Green Algae need.

Hapto tend to have more specialised niches and are more common in cooler waters, playing a significant role in biogeochemical cycling. The maps reflect these ecological preferences, with Hapto showing the highest concentrations around Bugrino in Russia, where cooler temperatures and favourable nutrient conditions support their growth. Like GA, they are also distributed along the coasts of northern Europe (Norway and Sweden) and northern Asia.

Dino are present along the coasts of northern Europe and northern Asia but do not exhibit a specific area with particularly high values. This more generalised distribution suggests that Dino may adapt to a broader range of environmental conditions within the Arctic region, which might include both stratified waters and varying nutrient levels.

Proka, which include cyanobacteria, are generally found in lower concentrations in the Arctic due environmental factors such as temperature and nutrient availability [32]. However, they play crucial roles in nutrient cycling and primary production, especially in areas with lower competition from other phytoplankton groups. The maps indicate that Proka have a slightly higher presence along the coasts of Norway, potentially due to specific local conditions such as nutrient availability and light.

The TChl-a map provides an overarching view of change to phytoplankton abundance in the Arctic, integrating the contributions of all PFTs. Higher TChl-a concentrations are expected in regions with favourable growth conditions for phytoplankton, such as nutrient-rich coastal waters and areas with optimal light and temperature conditions. These maps serve as a valuable tool for assessing the accuracy of the model predictions by comparing them with known ecological patterns and validating against in-situ measurements (Figure 2.5 and Figure 2.6).

In summary, the creation of these maps and their analysis provides a deeper understanding of the spatial distribution and ecological dynamics of PFTs and TChl-a in the Arctic. The machine learning ensemble model effectively captures the distribution of PFTs, revealing plausible patterns that align with known ecological behaviours. The results are comparable to global model retrievals, demonstrating the model's potential to improve TChl-a and Chl-a PFT estimations specifically for the Arctic region. The validation against in-situ data and consideration of environmental factors confirm the model's strengths while also identifying areas where further refinements could enhance its performance.

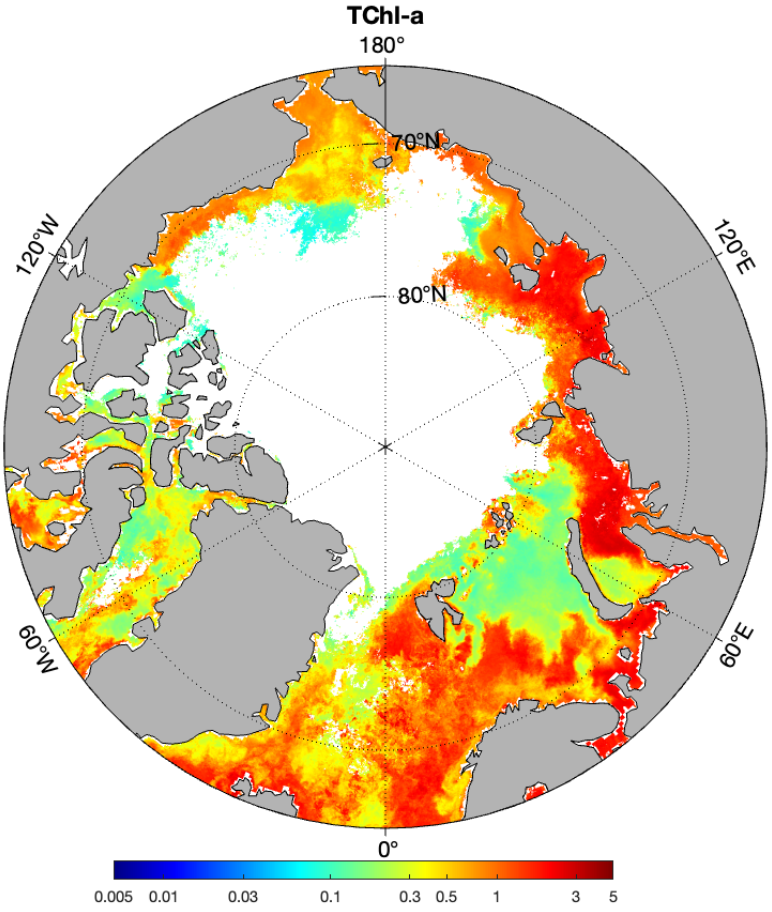

Figure 3.3: Mean TChla distribution in the Arctic Ocean (mg/m^3) during 29 June - 12 August 2022.

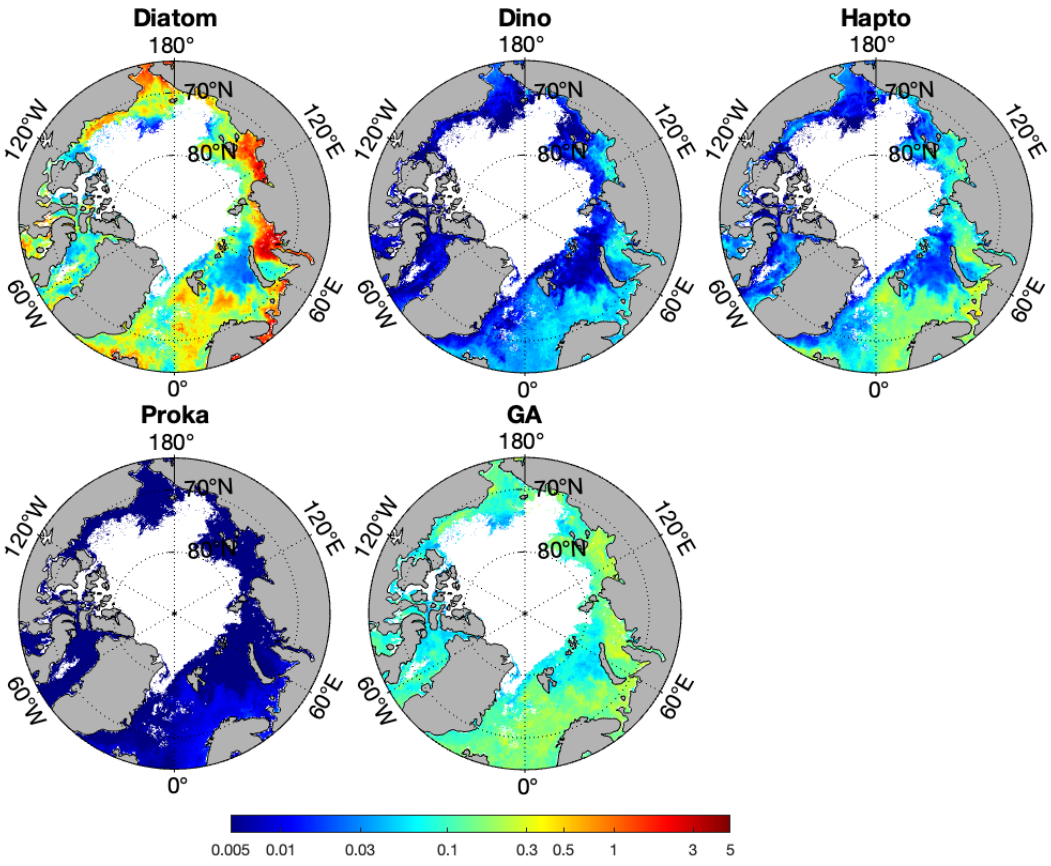

Figure 3.4: Mean Chl-a PFT distribution in the Arctic Ocean (mg/m^3) during 29 June - 12 August 2022.

4. Conclusions and Outlook

4.1 Conclusions

This study focused on the development and validation of an ensemble machine learning model to predict the concentrations of TChl-a and Chl-a of various PFTs in the Arctic Ocean. Given the significant changes in the Arctic region due to climate change, understanding the dynamics of phytoplankton is crucial for monitoring marine ecosystems and biogeochemical cycles.

The ensemble model combined the strengths of GBM, FCNN, RFR, and SVM through a Ridge Regression Ensemble approach. This method demonstrated promising results during the training phase, particularly for Diatoms and TChl-a, with R^2 values of 0.814 and 0.809, respectively. However, performance varied across different PFTs, reflecting the complexity and variability of phytoplankton dynamics in the Arctic.

Validation of the model using in-situ measurements from the PS131 expedition demonstrated the applicability of the machine learning ensemble approach, showing an overall good agreement between the model predictions and the observed data. This positive validation supports the model's capability to capture key patterns and trends in PFTs and TChl-a concentrations in the Arctic Ocean. However, the validation performance metrics also revealed some challenges, including a drop in predictive accuracy compared to the training phase. This decrease in performance can be attributed to the inherent differences between the training and validation datasets, highlighting the importance of accounting for environmental variability and further underscoring the need for robust model generalisation to handle diverse Arctic conditions effectively.

The creation of Arctic maps for PFT Chl-a and TChl-a over a 45-day period provided valuable insights into the spatial distribution of these variables. The maps revealed higher concentrations of Diatoms near coastal areas, reflecting nutrient-rich conditions, while GA showed a patchy distribution influenced by localised environmental factors. Hapto exhibited specialised niches in cooler regions, such as around Bugrino, where favourable nutrient conditions also support their growth. Dino were distributed along various coastal regions without a specific area of high concentration, possibly due to their ability to thrive in diverse environmental conditions. Proka, found in lower concentrations, highlighted the harsh conditions of the Arctic.

Overall, this study demonstrates the potential of using machine learning models for predicting phytoplankton dynamics in the Arctic. The ensemble model showed promise in capturing the complex patterns of PFT distribution and TChl-a concentration. The insights gained from this study can inform future research and monitoring efforts in the Arctic region.

4.2 Outlook

Future work should focus on several key areas to improve the accuracy and reliability of phytoplankton predictions in the Arctic Ocean:

1. **Enhanced Data Collection**: Increasing the spatial and temporal resolution of in-situ measurements will provide a more comprehensive dataset for model training and validation. Collaborative efforts to collect more data during different seasons and across various locations in the Arctic will enhance the model's robustness.

2. **Incorporation of Additional Variables**: Integrating more environmental variables, such as light availability, ice cover, and ocean currents, can improve the model's ability to capture the factors influencing phytoplankton dynamics. These variables can provide a more holistic view of the conditions affecting phytoplankton growth and distribution.

3. **Advanced Model Architectures**: Exploring another advanced machine learning architectures, such as deep learning models and hybrid approaches, and try with other different combinations for the ensemble model, could possibly give a better predictive performance.

4. **Long-term Monitoring and Prediction**: Developing models capable of long-term predictions will be crucial for understanding the impacts of climate change on Arctic phytoplankton. These models should be designed to account for long-term trends and shifts in environmental conditions.

5. **Interdisciplinary Approaches**: Collaborating with a larger group of experts in oceanography, climatology, and ecology will provide a more comprehensive understanding of the Arctic ecosystem, and ideas for additional variables or different in-between methods. Interdisciplinary approaches can integrate various data sources and perspectives, leading to more accurate and meaningful predictions.

6. **Policy and Conservation Implications**: The insights gained from phytoplankton predictions can inform policy decisions and conservation efforts in the Arctic. Understanding phytoplankton dynamics is crucial for managing fisheries, protecting marine biodiversity, and mitigating the impacts of climate change.

In conclusion, this study lays the groundwork for future research on phytoplankton dynamics in the Arctic Ocean where observations are reliable. By addressing the challenges identified and exploring new methodologies, we have shown much potential of improving the retrieval capability of the PFTs in this particular region, which help enhance our understanding of this critical component of the marine ecosystem and its response to a rapidly changing environment.

Acknowledgments

I would like to express my deepest gratitude to my professor, Astrid Bracher, for giving me the incredible opportunity to join the Phytooptics Group at the Alfred Wegener Institute in Bremerhaven. Her support and encouragement have been invaluable throughout this journey.

I am especially grateful to my supervisor, Hongyan Xi, for her unwavering guidance and mentorship. She entrusted me with this project and provided continuous support, offering ideas and solutions for the various challenges and uncertainties I encountered along the way. Her expertise and dedication were instrumental in the completion of this work. This research was made possible with the financial support from the AWI Innovation Fund through the project "ML-Phyto" led by Hongyan Xi.

I also extend my thanks to my colleague, Ehsan Mehdipour, for his helpful advices and support throughout this project.

I am thankful to all the scientists and crew who were involved in the in situ HPLC pigment data collection and analyses for making their data publicly available. In situ data from the AWI conducted expeditions PS74, PS76, PS78, PS80, PS85, PS93.2, PS99, PS106, PS107, PS121, PS122, MSM93, and PS131 were specially acknowledged. Thanks to the Copernicus Marine Service for all the products from satellite observations and model simulations.

Thank you all for making this journey an enriching and memorable experience.

Alfredo J. Bellido Rosas

This study has been conducted using EU Copernicus Marine Service information:

- https://doi.org/10.48670/moi-00019
- https://doi.org/10.48670/moi-00282
- https://doi.org/10.48670/moi-00280
- https://doi.org/10.48670/moi-00021
- https://doi.org/10.48670/moi-00168
- https://doi.org/10.48670/moi-00281
- https://doi.org/10.48670/moi-00165
- https://doi.org/10.48670/moi-00016
- https://doi.org/10.48670/moi-00015

Bibliography

[1] "Alfred-Wegener-Institut Helmholtz Zentrum fuer Polar und Meeresforschung (2017). Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute. Journal of large-scale research facilities, 3, A119." [Online]. Available: http://dx.doi.org/10.17815/jlsrf-3-163

[2] M. Ardyna and K. R. Arrigo, "Phytoplankton dynamics in a changing arctic ocean," *Nature Climate Change*, vol. 10, no. 10, pp. 892–903, 2020.

[3] P. Assmy, M. Fernández-Méndez, P. Duarte, A. Meyer, A. Randelhoff, C. J. Mundy, L. M. Olsen, H. M. Kauko, A. Bailey, M. Chierici *et al.*, "Leads in arctic pack ice enable early phytoplankton blooms below snow-covered sea ice," *Scientific reports*, vol. 7, no. 1, p. 40850, 2017.

[4] H. C. Kang, H. J. Jeong, J. H. Ok, A. S. Lim, K. Lee, J. H. You, S. A. Park, S. H. Eom, S. Y. Lee, K. H. Lee *et al.*, "Food web structure for high carbon retention in marine plankton communities," *Science advances*, vol. 9, no. 50, p. eadk0842, 2023.

[5] S. Sathyendranath, T. Platt, G. Zibordi, M. Babin, E. Boss, C. Klaas, H. Konno, Y. Liu, S. R. Signorini, Z.-P. Lee, Y. Huot, C. Mouw, L. Brown, and A. Mannino, *Phytoplankton Functional Types from Space*. Springer International Publishing, 2014.

[6] T. Hirata, N. Hardman-Mountford, R. Brewin, J. Aiken, R. Barlow, K. Suzuki, T. Isada, E. Howell, T. Hashioka, M. Noguchi-Aita *et al.*, "Synoptic relationships between surface chlorophyll-a and diagnostic pigments specific to phytoplankton functional types," *Biogeosciences*, vol. 8, no. 2, pp. 311–327, 2011.

[7] "Arctic ocean chlorophyll-a time series and trend from observations reprocessing, E.U. Copernicus Marine Service Information (CMEMS), DOI:10.48670/moi-00188, (Accessed on 02.06.2024)."

[8] A. Bracher, H. A. Bouman, R. J. Brewin, A. Bricaud, V. Brotas, A. M. Ciotti, L. Clementson, E. Devred, A. Di Cicco, S. Dutkiewicz *et al.*, "Obtaining phytoplankton diversity from ocean color: a scientific roadmap for future development," *Frontiers in Marine Science*, vol. 4, p. 55, 2017.

[9] Y. Zhang, F. Shen, X. Sun, and K. Tan, "Marine big data-driven ensemble learning for estimating global phytoplankton group composition over two decades (1997–2020)," *Remote Sensing of Environment*, vol. 294, p. 113596, 2023.

[10] M. A. Soppa, T. Hirata, B. Silva, T. Dinter, I. Peeken, S. Wiegmann, and A. Bracher, "Global retrieval of diatom abundance based on phytoplankton pigments and satellite data," *Remote Sensing*, vol. 6, no. 10, pp. 10 089–10 106, 2014.

[11] L. Alvarado, M. Soppa, P. Gege, S. Losa, I. Droescher, H. Xi, and A. Bracher, "Retrievals of the main phytoplankton groups at lake constance using olci, desis, and evaluated with field observations," 2022.

[12] S. N. Losa, M. A. Soppa, T. Dinter, A. Wolanin, R. J. Brewin, A. Bricaud, J. Oelker, I. Peeken, B. Gentili, V. Rozanov *et al.*, "Synergistic exploitation of hyper-and multi-spectral precursor sentinel measurements to determine phytoplankton functional types (synsenpft)," *Frontiers in Marine Science*, vol. 4, p. 203, 2017.

[13] H. Xi, S. N. Losa, A. Mangin, P. Garnesson, M. Bretagnon, J. Demaria, M. A. Soppa, O. Hembise Fanton d'Andon, and A. Bracher, "Global chlorophyll a concentrations of phytoplankton functional types with detailed uncertainty assessment using multisensor ocean color and sea surface temperature satellite products," *Journal of Geophysical Research: Oceans*, vol. 126, no. 5, p. e2020JC017127, 2021.

[14] M. J. Follows, S. Dutkiewicz, S. Grant, and S. W. Chisholm, "Emergent biogeography of microbial communities in a model ocean," *Science*, vol. 315, no. 5820, pp. 1843–1846, 2007.

[15] M. J. Behrenfeld, R. T. O'Malley, D. A. Siegel, C. R. McClain, J. L. Sarmiento, G. C. Feldman, A. J. Milligan, P. G. Falkowski, R. M. Letelier, and E. S. Boss, "Climate-driven trends in contemporary ocean productivity," *Nature*, vol. 444, no. 7120, pp. 752–755, 2006.

[16] R. C. Dugdale and F. P. Wilkerson, "Nitrate uptake rates in a coastal upwelling regime: A comparison of pn-specific, absolute, and chl a-specific rates," *Limnology and Oceanography*, vol. 43, no. 5, pp. 1081–1093, 1998.

[17] T. Tyrrell, "The relative influences of nitrogen and phosphorus on oceanic primary production," *Nature*, vol. 400, no. 6744, pp. 525–531, 1999.

[18] O. Ragueneau, P. Tréguer, A. Leynaert, J.-M. André, C. Pierre, and M. Panouse, "Biogeochemical transformations of inorganic nutrients in the mixing zone between the rhone river and the mediterranean sea: Implications for nutrient budgets in coastal zones," *Biogeochemistry*, vol. 50, no. 1, pp. 117–144, 2000.

[19] M. J. Behrenfeld and E. S. Boss, "Student's tutorial on bloom hypotheses in the context of phytoplankton annual cycles," *Global change biology*, vol. 24, no. 1, pp. 55–77, 2018.

[20] D. Breitburg, L. A. Levin, and v. n. p. y. p. Oschlies, Andreas and Grégoire, Marilaure and Chavez, Francisco P and Conley, Daniel J and Garçon, Veronique and Gilbert, Denis and Gutiérrez, Dimitri and Isensee, Kirsten and others, journal=Science, "Declining oxygen in the global ocean and coastal waters."

[21] M. Babin, A. Morel, V. Fournier-Sicre, F. Fell, and D. Stramski, "Light scattering properties of marine particles in coastal and open ocean waters as related to the particle mass concentration," *Limnology and oceanography*, vol. 48, no. 2, pp. 843–859, 2003.

[22] J. E. Cloern, S. Q. Foster, and A. E. Kleckner, "Primary production in the world's estuarine-coastal ecosystems," *Biogeosciences*.

[23] K. Xue, R. Ma, D. Wang, and M. Shen, "Optical classification of the remote sensing reflectance and its application in deriving the specific phytoplankton absorption in optically complex lakes," *Remote Sensing*, vol. 11, no. 2, p. 184, 2019.

[24] A. Zampollo, T. Cornulier, R. O'Hara Murray, J. F. Tweddle, J. Dunning, and B. E. Scott, "The bottom mixed layer depth as an indicator of subsurface chlorophyll a distribution," *Biogeosciences*, vol. 20, no. 16, pp. 3593–3611, 2023.

[25] C. Ji, Y. Zhang, Q. Cheng, J. Tsou, T. Jiang, and X. San Liang, "Evaluating the impact of sea surface temperature (sst) on spatial distribution of chlorophyll-a concentration in the east china sea," *International journal of applied earth observation and geoinformation*, vol. 68, pp. 252–261, 2018.

[26] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[27] Y. Zhang, J. Lee, M. Wainwright, and M. I. Jordan, "On the learnability of fully-connected neural networks," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 83–91. [Online]. Available: https://proceedings.mlr.press/v54/zhang17a.html

[28] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24–31, 2016.

[29] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*. Elsevier, 2020, pp. 101–121.

[30] T. C. Carneiro, P. A. Rocha, P. C. Carvalho, and L. M. Fernández-Ramírez, "Ridge regression ensemble of machine learning models applied to solar and wind forecasting in brazil and spain," *Applied Energy*, vol. 314, p. 118936, 2022.

[31] M. A. Lones, "How to avoid machine learning pitfalls: a guide for academic researchers," *arXiv preprint arXiv:2108.02497*, 2021.

[32] M. Royo-Llonch, P. Sánchez, C. Ruiz-González, G. Salazar, C. Pedrós-Alió, K. Labadie, L. Paoli, T. O. Coordinators, S. Chaffron, D. Eveillard *et al.*, "Ecogenomics of key prokaryotes in the arctic ocean," *bioRxiv*, pp. 2020–06, 2020.