Estimation of Chlorophyll in the North Indian Ocean Using Machine Learning Approach

Deepanshu Malik

Estimation of Chlorophyll in the North Indian Ocean Using Machine Learning Approach

Thesis Submitted to the
Indian Institute of Technology Bhubaneswar
in Partial Fulfilment for the Award of the Degree

of

Master of Technology

in

Climate Science and Technology

by

Deepanshu Malik

Under the guidance of

Dr. Sourav Sil and Dr. Manmeet Singh

SCHOOL OF EARTH, OCEAN, AND CLIMATE SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR

(April, 2023)

APPROVAL OF THE VIVA-VOCE BOARD

27/04/2023

Certified that the thesis entitled "**Estimation of Chlorophyll in the North Indian Ocean Using Machine Learning Approach**", submitted by **Deepanshu Malik (Roll No: 21CL06009)** to the Indian Institute of Technology Bhubaneswar, for the award of the degree of Master of Technology in Climate Science and Technology, has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

(Supervisor I)	(Supervisor II)
(Internal Examiner – I)	(Internal Examiner - II)
(Chairman)	(External Examiner)

भारतीय प्रौद्योगिकी संस्थान भुवनेश्वर

Indian Institute of Technology Bhubaneswar

पृथ्वी, महासागर एवं जलवायु विज्ञान विद्यापीठ

School of Earth, Ocean and Climate Sciences

Certificate

27/04/2023

This is to certify that the thesis entitled "**Estimation of Chlorophyll in the North Indian Ocean Using Machine Learning Approach",** submitted by **Deepanshu Malik (Roll No.: 21CL06009)** to the Indian Institute of Technology, Bhubaneswar, is a record of bona fide research work under our supervision and I consider it worthy of consideration for the award of the degree of Master of Technology in Climate Science and Technology.

Dr. Sourav Sil
(Supervisor I)

Dr. Manmeet Singh
(Supervisor II)

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to Dr. Sourav Sil (Associate Professor, School of Earth, Ocean, and Climate Science, Indian Institute of Technology, Bhubaneswar) and Dr. Manmeet Singh (Scientist-D, Indian Institute of Tropical Meteorology, Pune), my supervisors, for their continuous guidance and support throughout the research, review work and during writing this report.

I would like to thank Rahul Deogharia and Shouvik Dey along with other research scholars of our school, for their valuable input which has helped refine this work.

I also would take this opportunity to thank IIT Bhubaneswar for providing me with the resources required and all the eminent faculties of School of Earth, Ocean and Climate Science, whose inputs and constant monitoring of work progress have helped invaluably in the completion of this work.

I would also like to thank IITM Pune, for allowing me to use HPC resources and guidance in the final stages of this project.

Lastly, I express my gratitude to my family whose constant support has been crucial in academics and beyond.

DECLARATION

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor(s).
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in writing the thesis.
d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving the required details in the references.

Signature of the Student
(Deepanshu Malik)

Abstract

The amount of chlorophyll in the water is a crucial variable in studying marine environments. The satellite data provides details about the Earth's surface. However, data from the BGC-Argo indicate that the subsurface in the North Indian Ocean is where chlorophyll concentrations are highest. However, North Indian Ocean (NIO) BGC-Argo are quite rare and scattered. To estimate and rebuild the 3D chlorophyll structure in the North Indian Ocean, we employ a novel approach based on satellite and biogeochemical Argo (BGC-Argo) data. Along with well-known models like the Random Forest (RF) and Neural Networks (NN), the Transformer model from 'Embeddings for Numerical Features in Tabular Deep Learning' is implemented. This technique uses near-surface and vertical features to infer the Chla vertical distribution. Near-surface features from satellite products, vertical physical properties from temperature and salinity profiles from BGC-Argo floats, and temporal and spatial features (day of the year, longitude, and latitude) are the three types of input variables. Data from BGC-Argo floats measuring Chla and temperature-salinity parameters from 2011 to 2021 is used to train and test models, together with data from synchronous satellite-derived products. The correlation coefficient between the predicted Chla values and the validation dataset is 0.94, and the root-mean-square error is 0.012. When compared to the random forest, neural network, and NEMO numerical model, the Transformer model retrieves a more accurate and reliable vertical Chla profile in the NIO. Obtaining the 3D Chla structure at high vertical resolution is a primary use for this cphl product. This will aid in the precise quantification of NIO phytoplankton productivity and carbon fluxes. We anticipate this method can be used to create long-time series products, to comprehend the 3D Chla's variability in potential climate change scenarios.

Contents

List of Figures

Chapter 1

Introduction

Oceans cover three-quarters of the Earth's surface, and they produce about 80 percent of the world's oxygen. By affecting the amount of carbon dioxide (CO2) in the air, the ocean is an important part of how our planet's climate works. As the ocean's biological carbon pump removes atmospheric CO2, the photosynthetic generation of organic carbon by marine phytoplankton is essential [2] [4]. Also, phytoplankton is the first link in marine food chains, and the amount of photosynthesis is a big factor in how many fish there are. A better mechanistic understanding of how phytoplankton behaves is important for solving problems that are important to society, like climate change or managing fisheries. Climate change, which affects the surface (e.g., warming of the oceans), and Physical and chemical forcing, such as ocean acidification, have a big effect on how much phytoplankton photosynthesis happens, on the communities of phytoplankton, and on marine ecosystems as a whole. Therefore, a thorough understanding of ocean carbon sink requires the study and scientific evaluation of phytoplankton abundance in the euphotic layer where phytoplankton lives and grows [58][79]. For a long time, the majority of in situ measurements used to characterize upper ocean biogeochemical processes were obtained by ship-based observations, resulting in serious under-sampling and associated observational gaps. However, most recent studies have utilized satellite data to concentrate on near-surface water [8][22][27][30]. The variability of phytoplankton biomass below the surface is not captured by satellite-derived Chla, which only accounts for one-fifth of the total Chl a concentration in the euphotic layer [19][65]. Therefore, seasonal processes and patterns of phytoplankton in the near-surface layer do not accurately reflect variation in the entire upper-water column [29][79][89], particularly in areas with a subsurface chlorophyll maximum layer [19][71]. For instance, the southern Bay of Bengal (BoB) was reported to have an SCML [43][75], which contributed to column-integrated productivity [34][40], with magnitude often comparable to the highly productive Arabian Sea [54][44][79][89]. Predictions of the ocean carbon sink can be improved by expanding phytoplankton biomass assessments to encompass the entire euphotic layer, which can extend to depths of 0-150 m in the North Indian Ocean (NIO, 40°E to 100°E, -10°N to +10°N; the rectangle in Figure 1.1).

In marine ecological and biogeochemical research [56][9][16], the concentration of chlorophyll a (Chla; units: mg/m3) has attracted a significant amount of attention due to its widespread use as a proxy for phytoplankton biomass. Recent years have seen an increase in the use of satellite observations for estimating long-term trends

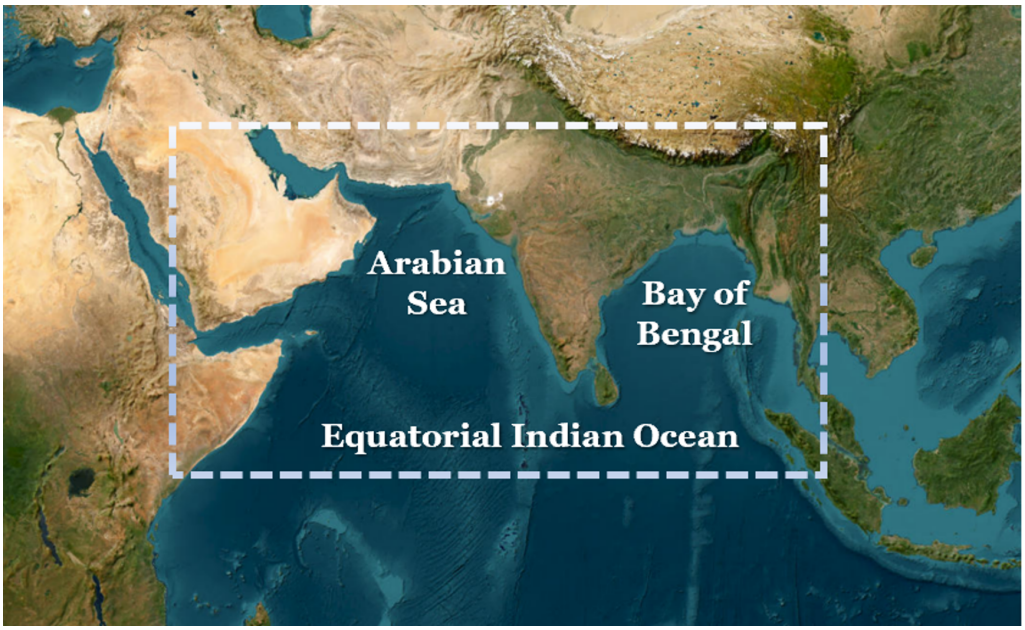

Figure 1.1: Geographical representation of North Indian Ocean (NIO) consisting of Arabian Sea (AS), Equatorial Indian Ocean (EIO) and Bay of Bengal (BOB).

in Chla at both the basin and the global scale [14][62]. In the past four decades, satellite measurements have suggested that worldwide Chla has been dropping, particularly in the subtropical gyres, and oligotrophic portions of all oceans have been expanding [56][9][16]. However, satellite-derived Chla only accounts for one-fifth of the total Chla content beneath the euphotic layer [65]. As a result, satellite-derived Chla fails to reflect the variability of phytoplankton biomass below the surface [19]. The number of concurrent in-situ observations of Chla at various depths has significantly risen thanks to the advent of oceanic autonomous observation platforms, in particular, biogeochemical Argo (BGC-Argo) profiling floats [9][12], large number of vertical profiles of physical (including pressure, temperature, and salinity) and bio-optical properties acquired by BGC-Argo floats offer a 3-Dimensional (3D) view of the ocean's interior in the upper 2000 m layer. At present, more than 9,000 temperature-salinity and Chla profiles have been obtained in the NIO. These vertical profiles combined with ocean satellite products provide an invaluable opportunity for the retrieval of the vertical Chla profiles.
The availability of light and nutrients, which are controlled primarily by physical dynamic processes linked with the monsoon-induced seasonal reversal current, upwelling, and vertical movement of pycnocline in the NIO [10][9], has a significant impact on the vertical pattern of Chla. In general, this pattern is determined by the availability of light and nutrients. Therefore, it is feasible to extrapolate surface-layer Chla to the entire water column by determining the relationship between Chla and the many physical properties of the water column that are known. Empirical statistical methods and ML techniques are just two examples of the models that have been created to construct vertical Chla profiles. The first study to employ near-surface estimates to investigate Chla vertical distribution established a statistical connection between near-surface and euphotic Chla. By contrasting the two Chla levels, this correlation was discovered. In order to infer the near-surface Chla, vertical distribution of Chla, column-integrated phytoplankton biomass, and phy-

toplankton community composition, [80] suggested a statistical approach based on the High-Performance Liquid Chromatography (HPLC) pigment database. Recent advances in artificial intelligence (AI) have led to the widespread adoption of data-driven machine learning (ML) techniques in the field of environmental remote sensing. This is because ML techniques offer many benefits over traditional approaches when modeling nonlinear relationships [68][57][92][91][23].

[68] used an ANN technique to develop a 3D particle backscattering coefficient for the oceans of the world. Satellite-derived surface outputs are combined with temperature and salinity profiles in this technique. In addition, the ANN method was used to rebuild the 3-D Chla field in the Mediterranean Sea using surface satellite products[65]. In this example, we opted for a method that makes use of the regression principle to either predict or categorize the value of an output variable. RF is an excellent machine learning technique that has found a lot of applications in the environmental and earth sciences[59]. The hierarchical tree structure used to make regression decisions is more transparent and straightforward to read than ANN [15][59]. Since the generalization error converges as the number of trees increases, the RF also avoids overfitting the data [59]. This prevents the RF from skewing the results in any way. The RF may also determine the relative importance of the numerous input variables by computing the Shapley additive explanations values (SHAP;[42]). To better understand the reaction of each input variable to the RF-Model and pick the best input variables to enhance the RF-Model's prediction performance [50], the SHAP is extremely useful for multi-source and high-dimensional datasets.

Learning-to-rank, click-through rate prediction, credit scoring, and a plethora of other industrial ML applications all involve tabular problems with data characterized by a set of heterogeneous attributes. Although deep learning models have recently taken over the ML literature, practitioners still frequently turn to "old-school" decision tree ensembles (like as GBDT) when faced with tabular challenges. The answer to the question "tabular DL or GBDT?" is still up in the air, as it was only later that multiple studies offered deep models that challenge the supremacy of GBDT in the tabular domain [3][26][73][26]. Pretraining deep model's parameters with a well-designed objective give them a possible performance edge over GBDT. To fine-tune for subsequent jobs, these pretrained parameters are a better starting point than random. Pretraining has been proven to be essential for state-of-the-art performance in the computer vision and NLP domains [28][21], hence it has become the de facto norm in these areas. However, this consensus has yet to be reached for tabular problems and recommended techniques for tabular pretraining have yet to be determined. In particular, unlike pretraining in vision or NLP problems, for which massive "extra" data is available on the Internet, pretraining for tabular issues is often performed directly on the downstream target datasets. Although many previous publications [7][20][78][90] have focused on pretraining tabular DL models, it is difficult to draw solid conclusions about the usefulness of pretraining in tabular DL from the literature due to the wide variety of experimental setups. While this shows that pretraining is effective, the performance of supervised baselines is somewhat constrained because certain evaluation algorithms presume the unlabeled data is copious yet employ a small fraction of labels from each dataset during fine-tuning for evaluation target datasets, as opposed to the vast amounts of "extra" data available on the Internet for pretraining in vision or NLP challenges. Although

many previous publications have focused on pretraining tabular DL models, it is difficult to draw solid conclusions about the usefulness of pretraining in tabular DL from the literature due to the wide variety of experimental setups. While this shows that pretraining is effective, the performance of supervised baselines is somewhat constrained because certain evaluation algorithms presume the unlabeled data is copious yet employ a small fraction of labels from each dataset during fine tuning for evaluation.

Using information gathered from ocean satellites and BGC-Argo floats, this study aims to create a novel Transformer-based approach for projecting the three-dimensional distribution of Chla in the NIO. For the rest of the thesis, we will refer to the process used to obtain the Chla 3D structure from a combination of satellite-derived products and BGC-Argo float data as the Transformer-method for the transformer deep learning model. The projected chla profiles are then used to calculate the vertical phytoplankton biomass in the NIO zone. Data from BGC-Argo floating observations, including 9,358 vertical profiles of pressure, temperature, salinity, and Chla, were combined with matching satellite-derived products to build the transformer model (Figure 2.1). Since the NIO database includes a wide range of hydrological and biological parameters, as well as physical processes like mixing and stratification, our model may be applied to the NIO.

Chapter 2

Data Used

2.1 Training Data

2.1.1 BGC-Argo Float Data

During the time span of January 2014 to December 2021, observations of Chla profiles were collected from 49 biogeochemical Argo (BGC-Argo) floats (see table no: 2.1 for WMO no.). These floats were mostly dispersed across three of the most important parts of the North Indian Ocean (40°-100°E, 10°S-30°N): the Arabian Sea (AS), the Bay of Bengal (BoB), and the Equatorial Indian Ocean (EIO)(see fig. 2.1 a). There are a total of 9385 profiles derived from official BGC-Argo floats that have a quality control flag of "1" (good) in each variable (i.e., pressure, temperature, salinity, and Chl a). These profiles can be accessed through the CORIOLIS Global Data Assembly Center.(ftp://ftp.ifremer.fr/ifremer/argo). In the upper 200 meters, the depth resolution is around 2 meters, and the profiling frequency for each float is once per day. Optical sensors that measure fluorometry-based Chla (Ex: 470 nm, 695 nm) are included in the BGC-Argo floats' ECO-Triplet or MCOMS, which are both equipped with the BGC-Argo floats [85][18]. Conductivity-Temperature-Depth (CTD) sensors paired with (Seabird) SBE41CP were used in all BGC-Argo floats to record pressure, temperature, and salinity profiles. The float data processing protocols are described in detail by [11]. These protocols include the following steps: conversion of numerical counts into Chl a physical unit; visual quality control of the profiles and potential probe corrections; removal of out-of-range values; dark-offset correction for Chl a; and correction for non-photochemical quenching for Chl a [11][18] Cornec et al. A correction factor of two was added in order to account for the fact that the Seabird-Wetlabs fluorometers overestimated the Chl a concentration [60]. In addition, the data coverage that is available assures that the Chl a variability that is shown by the BGC-Argo is mostly impacted by physical processes and nutrient dynamics [24][83].The integrated Chla (a proxy of phytoplankton biomass,[84]) change rate is defined as a deviation from the annual mean of 2015–2019 a total of 9,738 profiles flagged as 'good data' in each variable (i.e., pressure, temperature, and salinity, and Chla) were collected from all 49 BGC-Argo floats.

Table 2.1: BGC Argo WMO No.

S.No	Arabian Sea	Eq Indian Ocean	Bay of Bengal
1	2902091	2902088	2902087
2	2902092	2902156	2902086
3	2902118	2902179	2902113
4	2902120	2902215	2902114
5	2902123	2902216	2902158
6	2902124	2902238	2902160
7	2902174	2902239	2902161
8	2902175	2902240	2902189
9	2902199	2902241	2902193
10	2902202	2902242	2902195
11	2902204	2902243	2902196
12	2902205	2902244	2902217
13	2902209	2902245	2902264
14	2902210		2902294
15	2902211		
16	2902263		
17	2902270		
18	2902271		
19	2902272		
20	2902273		
21	2902274		
22	2902275		
23	2902276		
24	2902277		
25	5903586		

After that, a 3-point moving median filter was applied to each profile within 0-200 meters in order to remove the layer containing the spikes. De-spiking the profiles resulted in a decrease in the median value of all Chla observations, which went from 0.195 mg/m3 to 0.157 mg/m3, and an increase in the interquartile range, which went from 2 mg/m3 to 0.164 mg/m3. The temporal resolution of the profiling data for each float was 5 to 10 days, and the vertical resolution of all of the float profiles was nearly 1 to 5 m in the depth layer of the top 200 meters. At the same time, make use of the data for the pressure, temperature, and salinity profiles. In this study, we particularly train and evaluate the model utilizing depth, temperature, and salinity data received from BGC-Argo observations. Chla data was also included in the training process.

The data obtained by BGC Argo stations are unevenly scattered across the entirety of the north Indian Ocean. Because the majority of the locations are located in the open ocean at depths of the bathymetry of more than 2,000 meters, the primary emphasis of our study is on the determination of chlorophyll levels in the open ocean. In this case, we do not take into account the coastal effect(see Figure 2.1b).

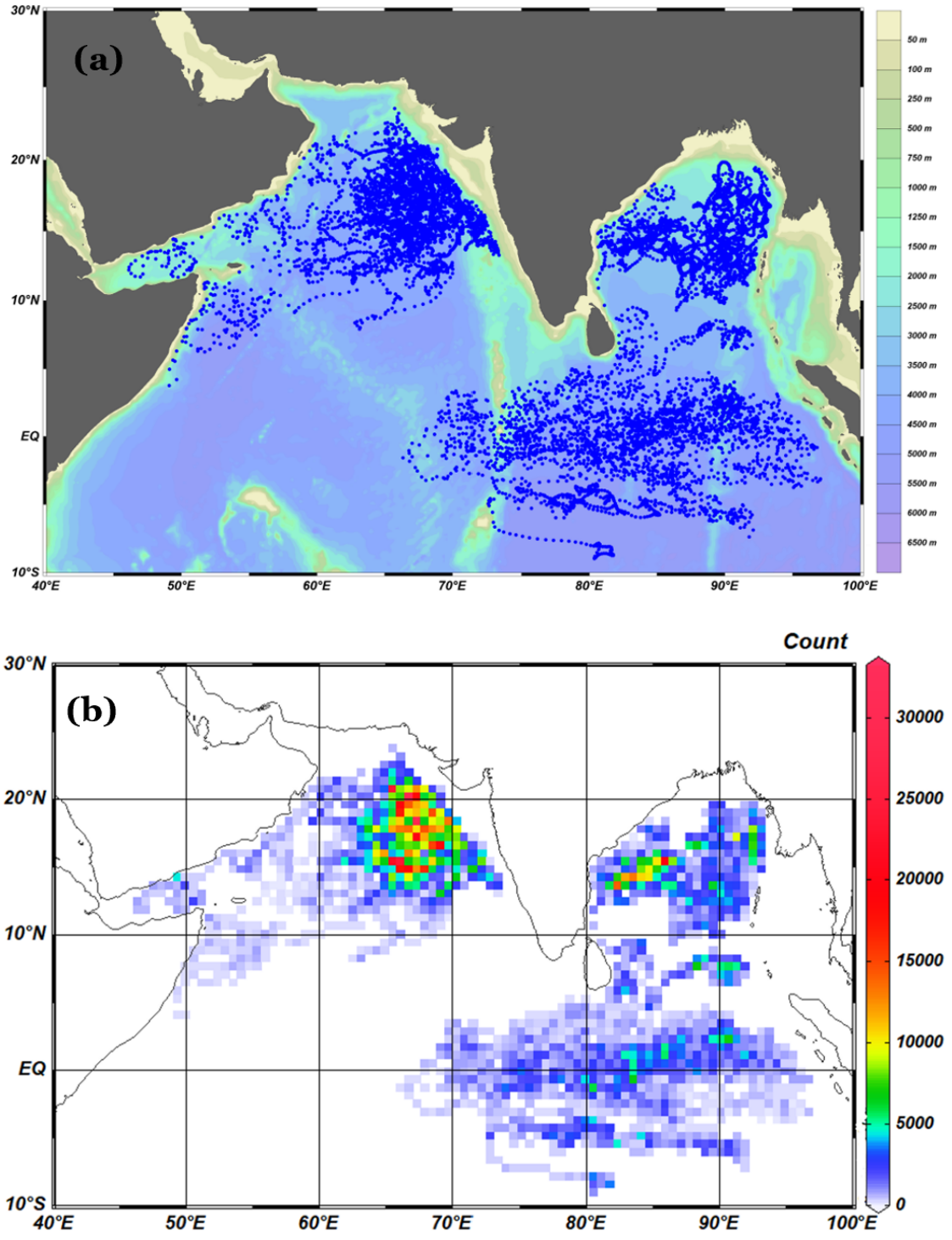

Figure 2.1: (a) Geographic distribution of 9,385 profile stations in the northern Indian Ocean (NIO). For each station (blue dots), the data contains vertical profiles from BGC-Argo (including pressure, temperature, salinity, and Chla) and matched satellite products.(b) Density distribution over NIO

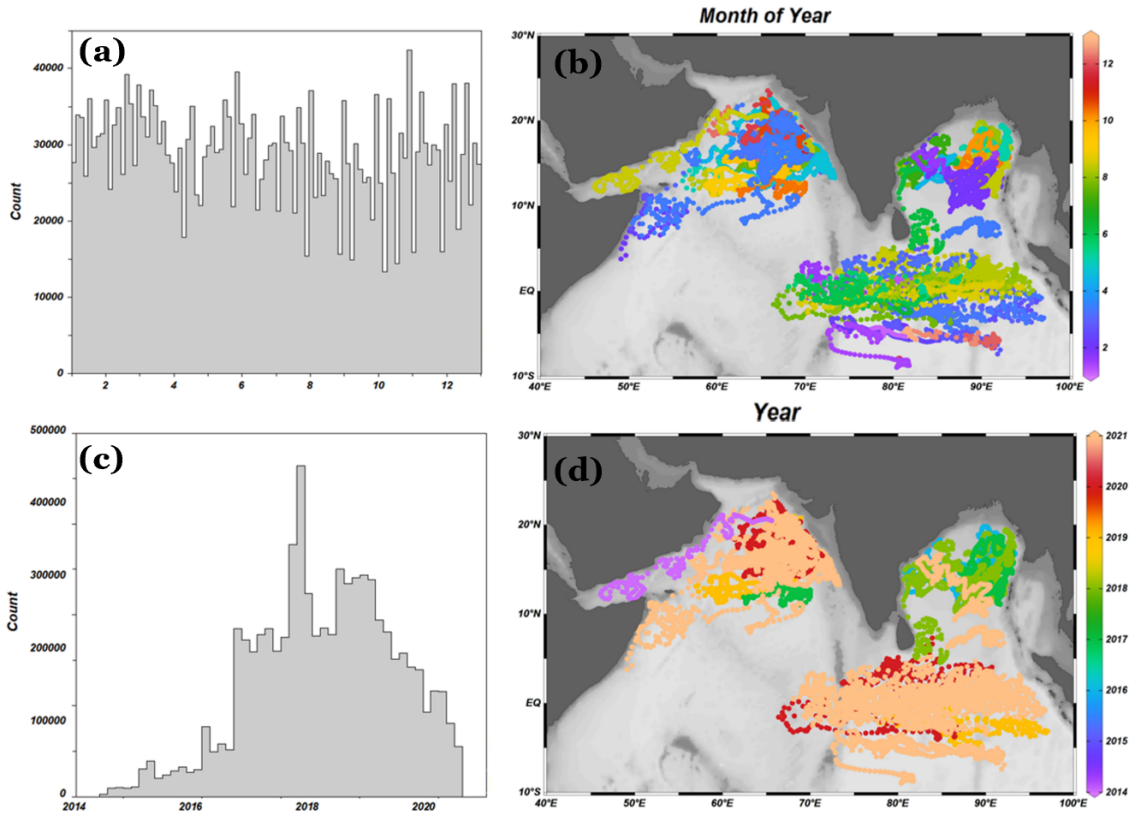

Figure 2.2: Temporal distribution of the number of float profiles was available as a function of years and months, with the color representing different years and months respectively. (b) and (d) are the temporal isosurface representation of data distribution of months and years respectively using different colors. (a) and (c)are the histograms of the temporal distribution of data for months and years respectively.

2.1.2 Satellite Data

Satellites with color-sensitive ocean sensors, like the Visible Infrared Imaging Radiometer Suite (VIIRS) and the Moderate Resolution Imaging Spectroradiometer (MODIS), can see how the ocean surface reflects light at different wavelengths. Based on existing bio-optical relationships, these measurements can then be used to figure out the concentration of chlorophyll. Remote sensing products, which are gridded datasets that give estimates of chlorophyll concentration at regular intervals and spatial resolutions, are usually how satellite data on ocean surface chlorophyll concentration is gathered. The scientific community can freely access these products, which are frequently produced by space agencies like NASA and ESA. For this investigation, a lot of satellite data from different sources with a resolution of one day was used. Detailed information on each satellite product is as follows:

1. Sea Surface Temperature (SST): The Level 3 Standard Mapped Image (SMI) products (contains a Plate Carrée, the pixel-registered grid of floating-point numbers (or scaled integer representations of the values) for a single geophysical parameter during 2011–2021) were downloaded from NASA's ocean color database (https://oceancolor.gsfc.nasa.gov/). Level 3 products are derived geophysical variables that have been projected onto a well-defined spatial grid over a well-defined time period. Aqua MODIS measures SST by detecting the thermal radiation emit-

ted by the ocean surface in the infrared portion of the electromagnetic spectrum. The sensor measures the brightness temperature, which is then converted into SST using algorithms that take into account various factors such as atmospheric conditions, sensor calibration, and oceanographic conditions. [32] has shown that Chl-a concentrations are highly correlated with sea surface temperature.

2. Photosynthetically Active Radiation (PAR): Daily PAR data were downloaded from the same source as of SST. Imaging Spectrometer (MODIS) Aqua on 4 km $\times$ 4 km grid during 2011-2021 was downloaded from NASA's ocean color website (https://oceancolor.gsfc.nasa.gov/). Photosynthetically Active Radiation (PAR) refers to the portion of the electromagnetic radiation spectrum that is absorbed by plants for photosynthesis. It consists of wavelengths between 400 and 700 nanometers (nm), which correspond to the visible light spectrum that humans can perceive. PAR is a critical component of light that is required for photosynthesis, the process by which plants use light energy to synthesize carbohydrates and produce oxygen. PAR plays a crucial role in plant growth and development, as it provides the energy needed for photosynthesis, which is essential for plant metabolism and the production of organic matter. PAR is typically measured in units of energy per unit of time, such as micromoles per square meter per second (mol/m^2/s), and is used as a standard measure to quantify the amount of light available for photosynthesis.

3. Sea Surface Chlorophyll (Chla): Chla data was constructed by merging measurements from four different sensors: MODIS, sea-viewing wide field-of-view sensor (SeaWiFS), medium resolution imaging spectrometer (MERIS), and visible infrared imaging radiometer suite (VIIRS). These merging data with a spatial resolution of 4 km $\times$ 4 km were obtained from the Ocean Colour-Climate Change Initiative website (OC-CCI; https://climate.esa.int/en/projects/) [66]. Sea surface chlorophyll refers to the concentration of chlorophyll-a, a pigment found in marine phytoplankton, at the surface of the ocean. Chlorophyll-a is an essential component of photosynthesis, the process by which plants and algae use sunlight to convert carbon dioxide and nutrients into organic matter and oxygen. Phytoplankton are microscopic, single-celled plants that float near the ocean's surface and form the base of the marine food chain. They are responsible for producing a significant portion of the world's oxygen and play a critical role in regulating global carbon cycles. Phytoplankton, including diatoms, dinoflagellates, and coccolithophores, contain chlorophyll-a, which allows them to capture sunlight and convert it into chemical energy through photosynthesis. The concentration of chlorophyll-a in the sea surface is an important indicator of phytoplankton biomass and productivity. Satellites and other remote sensing tools are used to measure sea surface chlorophyll levels, which are typically expressed in units of milligrams of chlorophyll-a per cubic meter of seawater (mg/m^3). Higher chlorophyll-a concentrations indicate higher phytoplankton abundance and productivity, while lower concentrations may indicate lower productivity or nutrient limitation. Sea surface chlorophyll levels are influenced by various factors, including sunlight availability, nutrient availability (such as nitrogen, phosphorus, and iron), water temperature, and ocean currents. Changes in sea surface chlorophyll concentrations can have significant ecological impacts on marine ecosystems, as they affect the availability of food for higher trophic levels, including fish, marine mammals, and seabirds. Additionally, sea surface chlorophyll levels are used in oceanographic research and ecosystem management to monitor changes in ocean productivity, track oceanographic processes, and study the impacts of climate

change on marine ecosystems.

4. Remote Sensing Reflectance (Rrs): Rrs at 412, 443, 490, 560, and 665 nm wavelengths were also taken from (OC-CCI; https://climate.esa.int/en/projects/) [66]. The remote-sensing reflectance R rs is defined as :-

$$R_{rs}(\theta, \phi, \lambda) = \frac{L_w(\theta, \phi, \lambda)}{E_d(\lambda)}$$

Here θ and ϕ specify the polar and azimuthal directions, respectively, in some convenient coordinate system, and λ is the wavelength. $L_w(\theta, \phi, \lambda)$ is the water-leaving spectral radiance in direction (θ, ϕ), that is, the radiance heading upward just above the sea surface that originated from underwater light, which was transmitted upward through the sea surface into the direction (θ, ϕ). $E_d(\lambda)$ is the downwelling spectral plane irradiance incident onto the sea surface. The measurements implicit in above equation are generally made within a few meters of the sea surface and at wavelengths from the near ultraviolet to the near-infrared, e.g., for λ from 350 to 800 nm [48]. Rrs(412) is often used as a reference wavelength in remote sensing studies due to its sensitivity to changes in water constituents, such as chlorophyll-a, suspended particles, and colored dissolved organic matter, which are important indicators of water quality and ecosystem health [38].

5. Diffuse attenuation coefficient at 490 nm (Kd490): kd490 was also downloaded from OCCCI website [38]. For many oceanographic studies, the diffuse attenuation coefficient, Kd(λ) (where λ is the light wavelength in free space), of the spectral solar downward irradiance, $E_d(\lambda)$, plays a critical role. These studies include the heat transfer in the upper ocean [17][39][49][93], photosynthesis and other biological processes in the water column [45][46][52][67], and turbidity of oceanic and coastal waters [31][36]. K(λ) is an apparent optical property [38], so it varies to some extent with solar zenith angle, sky and surface conditions, as well as with depth even within the well-mixed water column.

6. Sea Level Anomaly (SLA): Daily multi-satellite merged SLA on 0.25° × 0.25° grid during 2011-2021 was provided by the Copernicus Marine Environment Monitoring Service (CMEMS; http://marine.copernicus.eu/). The products/datasets are extended in time at least once a year so that these datasets cover 1993 up to five to twelve months before present. This dataset benefit from the highest quality altimeter measurements and geophysical corrections, and are produced with a unique system (no changes in time in the set up, extended in time using the same system) to minimize the risk of quality loss or spurious signals appearing in time.

7. Sea Surface Wind (SSW): Daily gridded cross-calibrated multiplatform (CCMP) version 2.0 sea-surface wind (SSW, including both u and u components) with a spatial resolution of 0.25° × 0.25° for the period 2011-2021 is available from the Remote Sensing Systems (RSS, www.remss.com). The Cross-Calibrated Multi-Platform (CCMP) is a gridded Level 4 (L4) product that provides vector wind over the world's oceans. CCMP is a combination of ocean surface (10m) wind retrievals from multiple types of satellite microwave sensors and a background field from reanalysis. The resulting product is a spatially complete dataset available every six hours that remains closely tied to the satellite retrievals where they are available and closely collocated in time and space. Where satellite retrievals are not available, CCMP is statistically consistent with satellite winds. Creating a L4 product using this method of combining satellite and reanalysis data ensures a smooth transition

in the wind field between regions with and without satellite retrievals.

In the present study, the large database of the model consists of a total of 18 variables, 17 of which come from satellite and BGC-Argo. These variables include the day of the year, longitude, latitude, Rrs at 412, 443, 490, 560, and 665 nm, SST, Kd490, PAR, SLA, wind (u and v components), temperature, salinity, and depth. These variables were used to train and evaluate the model. The vertical Chla profiles in the resulting matchup database serve as the response variable. These profiles were measured using data from the ocean color satellite and the BGC-Argo float. Then, each profile that was obtained from the BGC-Argo observations that were discussed earlier was matched up with concurrent satellite-derived products (such as Rrs, SST, Kd490, PAR, SLA, wind, and Chla) by using the closest pixel from daily satellite image composites that had a resolution of 4 kilometers. Before beginning to train the model, the near-surface Chla readings obtained from the ocean color satellite and the BGC-Argo float were compared to ensure that they were consistent. The scatter of near-surface Chla values is centered around the 1:1 line and has a strong linear relationship with high R2 values of 0.85 (Figure 2.3). Additionally, this scatter has a strong linear relationship. This suggests that the pattern and amount of surface Chla in the research area as seen from the satellite were in good agreement with the observations made by BGC-Argo. The time-series validation of satellite surface chlorophyll with Bgc Argo surface chlorophyll (the average value of 0-10 m) is also done on a Bgc Argo float (WMO no. 2902158) by averaging along latitude and longitude as shown in (Figure 2.3).

Table 2.3: Details of Input Variables

Type of component	Variable
Spatial	Latitude (deg N)
Spatial	Longitude (deg E)
Temporal	Day of year (radian)
Vertical	Pressure(dbar)
Vertical	Temperature(deg C)
Vertical	Salinity(siemens/m)
Surface	Sea Surface Chlorophyll (Chla) (mg/m^3)
Surface	Rrs @412, 443, 490, 560, and 665 nm(J/m^2)
Surface	Attenuation Coeff @490 nm(dB/m)
Surface	Sea surface wind (both u and v) (m/s)
Surface	sea surface Temperature (K)
Surface	Sea Level Anomaly(m)
Surface	Photosynthetically Active Radiation (PAR) (YPF)
Target	Chlorophyll-a conc(mg/m^3)

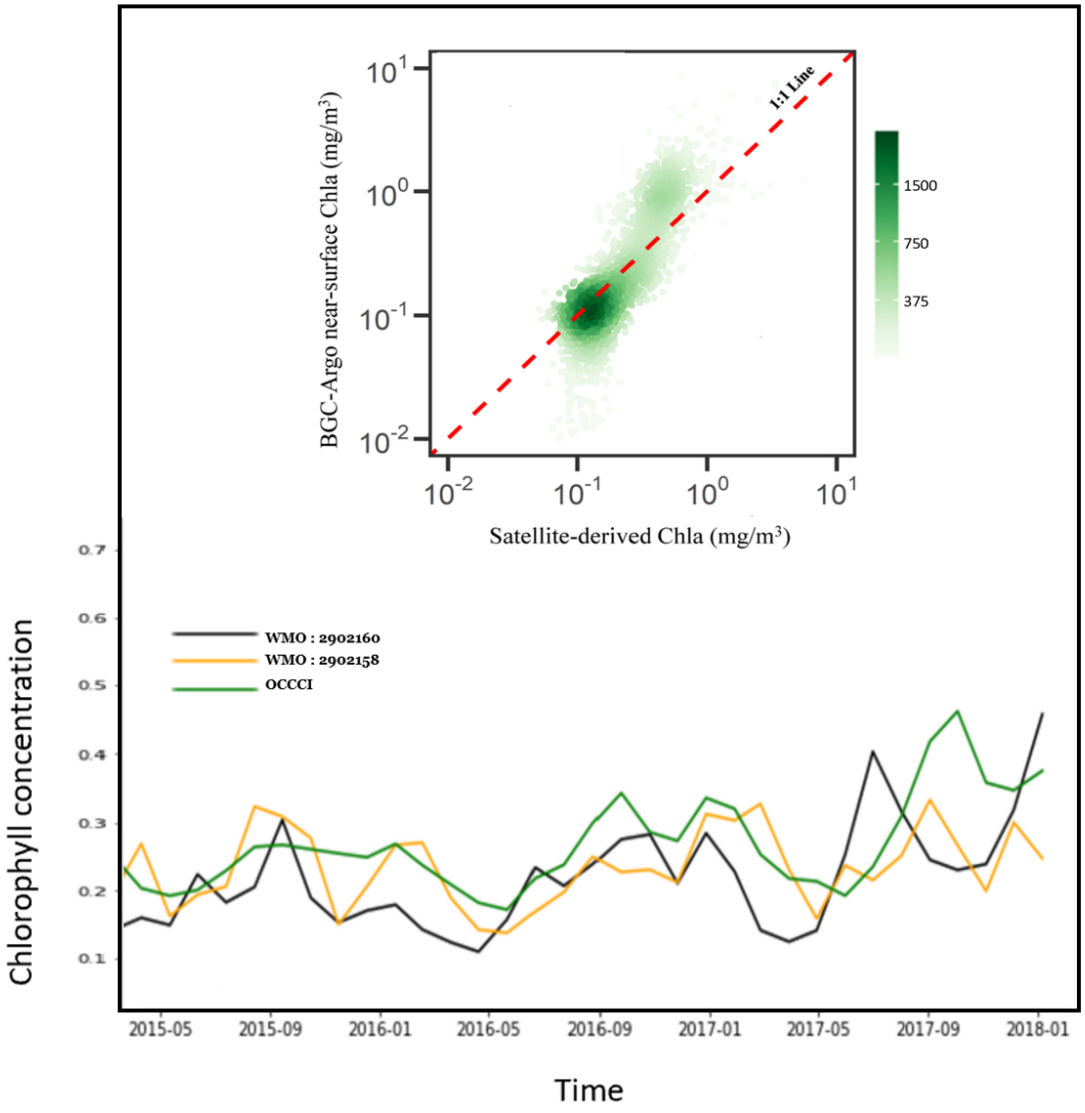

Figure 2.3: Comparison of near-surface Chla from ocean color satellite and BGC-Argo floats using scatter plot of the satellite Chla (on the x-axis) vs. the BGC-Argo surface Chla (on the y-axis); surface Chla derived from BGC-Argo is calculated as the average value of 0-10 m. Also, time-series comparison of near-surface Chla from OCCCI and BGC-Argo float ((WMO:2902158 and 2902160)) data from May 2015 to January 2018

2.2 Deployment Data

Many data gaps were present in the daily satellite products because of serious potential contamination in the NIO, hence monthly data was used in this study for the deployment phase of the trained RF-Model. Moreover, both the daily and monthly satellite data came from the same remote sensing sensor, thus they may be used interchangeably in the model. The CMEMS global ocean ensemble reanalysis, which incorporates satellite and in-situ observations, yields a dataset with a spatial resolution of 1/4° * 1/4°and 32 vertical levels from 0 to 200 m (GLOBAL REANALYSIS PHY 001 031, https://doi.org/10.48670/moi-00024; including, monthly temperature and salinity). A numerical ocean model that is bound by the data assimilation of satellite and in situ observations is used to construct global ocean reanalyses, which are homogenous, 3D gridded descriptions of the physical condition of the ocean that cover several decades. These reanalyses are constructed to be as realistic as feasible and to be in agreement with the physics of the model. They are built to be as near as possible to the observations. Estimation of uncertainties or error bars in the ocean state is made possible by the use of the multi-model ensemble technique. The ensemble mean may even provide for certain regions and/or periods a more reliable estimate than any individual reanalysis product. Consistency checks were performed on the BGC-Argo float and reanalysis dataset's vertical temperature and salinity profiles before to releasing the trained transformer model. The reanalysis data showed a vertical temperature and salinity distribution that was quite similar to that predicted by BGC-Argo. Furthermore, in monthly climatology from 2014-2019, there was strong agreement between BGC-Argo and reanalysis data (Figure 2.4). Therefore, the monthly 3D Chla product in the NIO, with a spatial resolution of 1/4° * 1/4° and 32 vertical levels spanning from 0 to 200 m, may be reliably generated using the trained model in conjunction with the monthly reanalysis and satellite data.

2.3 Validation Data

In order to validate the monthly 3D Chla product produced by the trained model was evaluated using observation and operational global 3D Chla datasets. First, for surface chlorophyll validation we used monthly OCCCI surface chlorophyll data, which was constructed by merging measurements from four different sensors: MODIS, sea-viewing wide field-of-view sensor (SeaWiFS), medium-resolution imaging spectrometer (MERIS), and visible infrared imaging radiometer suite (VIIRS). These merging data with a spatial resolution of 4 km × 4 km were obtained from the Ocean Colour-Climate Change Initiative website (OC-CCI; https://climate.esa.int/en/projects/) [66]. Then, for vertical chlorophyll validation, we used Bgc-Argo dataset downloaded from euro argo fleet monitoring website. Additionally, we used the CMEMS global biogeochemical multi-year hindcast GLOBAL_REANALYSIS_BIO_001_029 (https://resources.marine.copernicus.eu/) generated by the Nucleus for European Modelling of the Ocean (NEMO), Pelagic Interaction Scheme for Carbon and Ecosystem Studies (PISCES; https://www.nemo-ocean.eu/). It provides monthly 3D biogeochemical fields of Chla, nitrate, phosphate, silicate, oxygen, and net primary production for the period 1993-2019 with 1/4° × 1/4°mhorizontal resolution and 75 vertical levels from the surface to 5500 m depth.

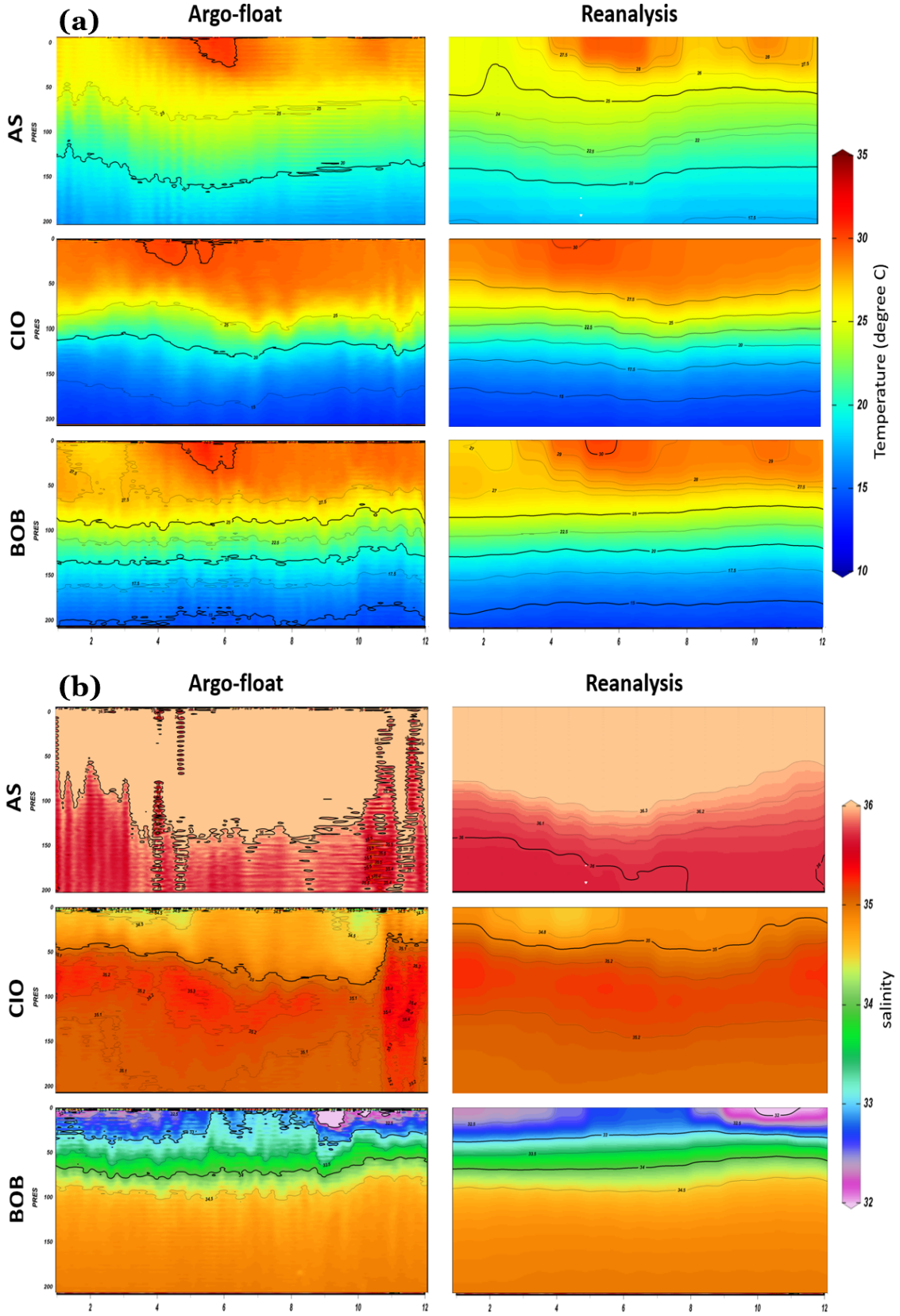

Figure 2.4: Comparison of the monthly climatology vertical profile of (a) temperature and (b) salinity obtained from BGC-Argo float and reanalysis data at the depth range of 0-200 m from 2014 to 2019.

Chapter 3

Methodology

3.1 Data Preprocessing

After deducting the number of dark counts from the raw signal, the chlorophyll a fluorescence was then converted into the chlorophyll a concentration using the calibration coefficients that were provided by the manufacturer [13]. The real-time dedicated quality control procedure identified the occurrence of negative spikes, adjusted chlorophyll a concentration profiles for cases of nonzero values at depth, and verified the range of measured values according to the technical specifications provided by the manufacturer [13]. This procedure was carried out in accordance with the procedures that are described in the [69]. In addition, the concentration of chlorophyll a seemed to increase at depths where it should have been zero in certain bioregions, such as subtropical gyres and the Black Sea, despite the fact that this should not be the case. This behavior was attributed to the influence of fluorescence that originated from nonalgal materials, according to reference proctor2010new. Consequently, corrections were made to profiles in accordance with the [86]. In the end, in accordance with the advice provided by [60] for Chla measurements from WET Labs ECO fluorometers, the calibrated quality-controlled Chla values were divided by a correction factor of 2 in order to account for the error. A global comparison of paired high-performance liquid chromatography (HPLC) and in situ fluorescence Chla data was used to derive the correction factor. This factor was then confirmed by optical proxies of Chla, such as light absorption line height [61] or in situ radiometry [87]. [60] presents an in-depth analysis of the regional variability of this average correction factor, in addition to the probable uncertainties associated with these factors. In order to determine the sensitivity of the bbp-to-Chla relationship to the factor that was used to adjust the fluorescence-based Chla values, we carried out an investigation. For the purpose of this study, a total of 9,385 profiles were acquired from each of the 49 BGC-Argo floats between the years 2014 and 2021. You can get more information on the detailed processing of the BGC-Argo float data at this website: (https://biogeochemical-argo.org/).
Despiking was performed on each profile within 0-200 meters using a 3-point moving median filter in order to get rid of the spikes layer. The suppression of unwanted noise can be accomplished with the use of a nonlinear signal processing technique called median filtering. In 1971, [76] proposed using it as a tool for conducting time series analysis; during the years after then, it has also found application in the field of image processing. The process of median filtering involves letting a window move

over the points of a picture (or sequence) and then replacing the value at the center of the window with the median of the original values that are contained within the window. This process is repeated until the desired result is achieved. This results in an output picture (sequence) that is typically smoother than the one that was inputted originally. Utilizing a linear low-pass filter is the traditional method of smoothing, and in many instances, this is the method that is best suited for the job. However, there are circumstances in which median filtering is preferable, and two of its primary benefits are as follows: I) The median filtering method maintains the sharpness of the edges, in contrast to the linear low-pass filtering method which blurs the edges. II) When it comes to smoothing down spiky noise, median filters are among the most effective tools [33]. The median value of all Chla observations dropped from 0.192 mg/m3 to 0.151 mg/m3 after the profiles were de-spiked and spiked, and the interquartile range shrunk from 0.367 mg/m3 to 0.172 mg/m3.
Next, Regridding satellite data from multiple sources is a crucial step in integrating and analyzing remote sensing data. Satellites capture data at various spatial resolutions and grids, depending on the sensor and orbit. To compare and combine data from different sources, it is necessary to regrid them onto a common grid with a consistent spatial resolution. This process involves interpolating or resampling the data to match the desired grid system. Regridding is necessary to ensure that the data aligns properly and can be accurately compared, analyzed, and integrated into models. It requires careful consideration of factors such as data resolution, spatial extent, and projection systems to ensure that the resulting regridded data accurately represents the underlying phenomena and enables meaningful analysis and interpretation. The resolution of the satellite data we have ranges from 4 km to 25 km. Python's package xESMF: Universal Regridder for Geospatial Data (uses ESMF/ESMPy as backend and can regrid across general curvilinear grids with all ESMF regridding methods, including bilinear, conservative, and closest neighbor) is used to interpolate them into a common grid.
Then, each profile obtained from BGC-Argo observations mentioned above was matched up with concurrent satellite-derived products (including Rrs, SST, Kd490, PAR, SLA, wind, and Chla) using the closest pixel from daily satellite image composites with a 4 km resolution. By Calculating the distances between the extracted pixel values and the locations of your in-situ points we assign the value of the closest pixel to the in-situ point based on latitude, longitude and time.
Standardization, also known as feature scaling or normalization, is a crucial step in the data preprocessing phase of machine learning. It involves transforming the features of a dataset to a common scale or range in order to ensure that they are on the same level of magnitude. This is done to avoid any bias towards features with higher values and to prevent numerical instability during model training. Z-score normalization, Min-max normalization, and Log transformation are the commonly used feature scaling methods. In response to our problem, we have used Min-max normalization method that scales data to a specific range, typically [0, 1], by linearly transforming the data.

The formula for Min-max scaling is:

$$X\_scaled = \frac{X - min}{max - min}$$

where 'x' is the original value of the feature, 'min' is the minimum value of the feature, and 'max' is the maximum value of the feature. The Min-max scaling

bounds the data to a specific range, which can be useful in cases where data within a certain range is desired, such as in neural networks with activation functions that are sensitive to the input data range. It is a simple and intuitive method that preserves the relative relationships between data points.

3.2 Machine Learning Framework

Machine learning is a field of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable computers to learn from and make predictions or decisions based on data. In the context of climate science, machine learning can help analyze large datasets and extract valuable insights for understanding and predicting climate patterns, impacts, and phenomena. Ocean science generates massive amounts of data from various sources such as satellites, floats, and buoys. These datasets are often complex, and multidimensional, and require advanced analytical techniques to extract meaningful information. Machine learning can help analyze large amounts of ocean data to identify patterns, trends, and anomalies. For example, it can be used to analyze historical data to identify long-term climate patterns, such as changes in sea surface temperature, ocean currents, or sea level rise, which can help in understanding climate change impacts. Also, one of the potential applications of machine learning is the reconstruction of 3d chlorophyll a concentration product using satellite and irregular in-situ observations. In this study, we utilized a robust machine-learning framework to analyze a huge observational dataset and construct a 3d chlorophyll repository in NIO. The framework I employed leveraged the power of popular Python libraries such as scikit-learn (sklearn) and TensorFlow to build and train machine learning models. Sklearn provided a comprehensive suite of tools for data preprocessing, feature selection, and model evaluation. I used its rich collection of algorithms, including decision trees, support vector machines, and random forests, to develop predictive models that could capture the non-linear relationships and interactions among the various climate variables. Sklearn's intuitive APIs and built-in functions for cross-validation and hyperparameter tuning allowed me to optimize the performance of the models and ensure their reliability. On the other hand, TensorFlow proved to be a powerful deep learning framework that enabled me to build complex neural network architectures for tasks such as image and time series data analysis. Its flexible and scalable design allowed me to experiment with different deep learning models, including convolutional neural networks (CNNs) and sequential models (RNNs, transformers), to extract meaningful features from the ocean data and make accurate predictions. TensorFlow's extensive support for GPU acceleration also facilitated efficient training of large-scale models, enabling me to handle the massive datasets typically encountered in ocean science research. Overall, the machine learning framework I employed in my research paper using Python, sklearn, and TensorFlow provided a robust and efficient toolkit for analyzing climate data and uncovering hidden patterns and insights. The combination of these powerful libraries empowered me to develop sophisticated models that could contribute to our understanding of the chlorophyll variability in NIO.

3.2.1 Multivariate Linear regression (MLR)

MLR is a common form of regression analysis. Multiple linear regression attempts to explain the relationship between one dependent variable and two or more independent variables by ftting a linear Eq. It has been widely used for climate studies for downscaling and impact analysis. In general, MLR can be mathematically written as:-

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_2 + \epsilon$$

where y is the dependent variable, x_i are independent variables, β_i are parameters, ϵ is the error. In this study, the ordinary linear least squares (LLS) regression which minimizes the residual sum of squares between the observed values and the ensemble values was used. This was implemented using 'sklearn.linear_model' module in python.

3.2.2 Support Vector regression (SVR)

SVM is based on Vapnik–Chervonenkis (VC) theory and the rule of structural risk minimization [81]. SVM is used for various climate change and hydrological applications [64][88]. Support Vector Regression (SVR) is the SVM that elucidates nonlinear regression problems by mapping the low-dimensional data to a high-dimensional feature space using kernel functions. Mathematically, SVR model can be represented as follows:-

$$y = \sum_{i=1}^{n} (\alpha_i - \hat{\alpha}) Kernel(x_i, x) + b$$

here Kernel(x_i, x) represents the kernel function used; α_i and $\hat{\alpha}$ denote the Lagrange multipliers; x_i denote the vectors; x represents the independent vector; b represents the bias parameter. SVR uses a symmetrical loss function, which equally penalizes high and low misestimates. Using Vapnik's Open image in new window -insensitive approach, a fexible tube of minimal radius is formed symmetrically around the estimated function, such that the absolute values of errors less than a certain threshold Open image in new window are ignored both above and below the estimate. In this manner, points outside the tube are penalized, but those within the tube, either above or below the function, receive no penalty. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy[5]. Estimations that used the polynomial kernel function performed better than the estimations that used other kernel functions. As a result, in this study, the polynomial kernel function was used in the same way that[64] did. The choice of hyperparameters plays a great role in machine learning methods. In the current study, Bayesian hyperparameter optimization was used to determine the hyperparameters for all machine learning algorithms. The "optuna" package in Python was used to implement. Optuna is an automatic hyperparameter optimization software framework, particularly designed for machine learning. It features an imperative, define-by-run style user API. The important hyperparameters optimized in SVR are C, kernel function, and epsilon.

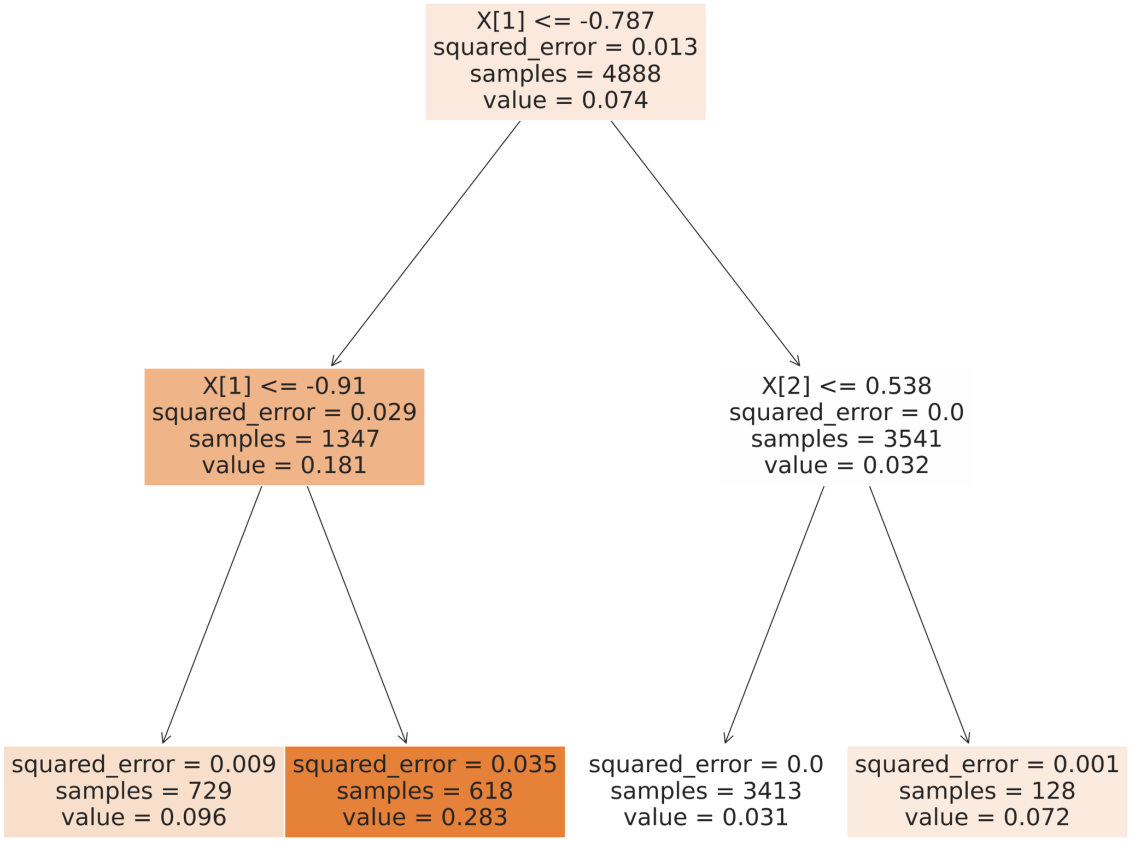

Figure 3.1: Simplified representation of Random Forest method

3.2.3 Random Forest regression (RF)

In order to make an estimate of Chla's three-dimensional structure, an initialized RF-Model was given training with the parameters 300 decision trees and a maximum depth of 16. Figure 4 provides a synopsis of the primary inputs to the RF-Model as well as the associated processing flow. The predictors are the 17 input variables, which may be broken down into the following categories: day of the year, longitude, latitude, Rrs at 412, 443, 490, 560, and 665 nm, SST, Kd490, PAR, SLA, wind (u and v components), temperature, salinity, MLD, and depth. These variables can be further subdivided into the following categories: temperature, salinity, and MLD. (1) the temporal aspect, which is represented by the day of the year. Because the variation in the temporal variable (day of the year) indicates the seasonal cycles, it is translated into radians using the equation, which is as follows: is the day of the year when expressed in radian units, and the coefficient 182.625 stands for half of the total number of days in a year (365.25). (2) the spatial component, which is represented by longitude and latitude; (3) the sea-surface component, which is represented by Rrs at 412, 443, 490, 560, and 665 nm, SST, Kd490, PAR, SLA, and wind (u and v components); and (4) the vertical component, which is represented by temperature, salinity, MLD, and depth. Training and validation datasets are used to fine-tune and test the RF-Model's hyperparameters as it learns. The effectiveness of the trained RF is also assessed. Decision Forests (DF) are a family of ML techniques that may rival the speed and accuracy of neural networks when working with tabular data. The supervised ML method known as the RF-describe produces a large regression tree that may be utilized to describe the complex nonlin-

ear relationship with a high degree of accuracy. The RF model is an instance of an ensemble method for machine learning. Combining statistical learning theory with classification or regression techniques, RF is offered [15]. Over-fitting is avoided, and input variables of varying types are calibrated, thanks to the multiple classification and regression decision tree (CART) built into the method. This algorithm produces a large number of unrelated trees, from which it draws a conclusion based on the features of nonparametric statistical regression and chance. A decision tree has three levels of nodes: a root, a child, and a child's child. A judgment level is represented by a leaf node, and a judgment rule is stored in a sub-node. The method yields an average prediction based on all trees. Internally, RF is OOB-scored for cross-validation purposes.

3.2.4 Artificial Neural Network (ANN)

[47] introduced the idea of ANNs in 1943, but it wasn't until 1986 when [63] developed the backpropagation algorithm, that ANNs really took off. In the field of artificial intelligence, ANNs are commonly used. (AI) is the research and development of techniques to enable machines to learn independently. Machine Learning Course at Stanford University states that the ultimate aim of IA is to create machines that are as intelligent as the human brain. In an effort to mimic the human brain, Neural Networks (NNs) seek to model both the neurons themselves (represented by nodes) and the connections between them (represented by transfer functions). NNs are constructed from layers, which include the input layer, which receives the known data, the hidden layers, which execute the actual processing, and the output layer, which receives the desired value. Each layer is made up of a different type of unit (neurons or nodes), and these layers communicate with one another via a transfer function. In ANNs, the output of layer i-1 feeds into layer i. The input layer's values are multiplied by an initial weight and summed before being delivered to the output layer via a sequence of functions, where the known data and bias term have been inputted.ANNs are the most prevalent method for developing nonlinear regression models among the many machine-learning techniques [72]. Training an ANN needs selections including the network structure (i.e. the number of hidden layers and nodes per layer), proper initialization of the weights, learning rate, and training algorithm. In this work, the input layer was physical parameters, and the output layer was CPHL. We optimized a two-layer back propagation neural network (BPNN) with tan-sigmoid ($i.e. f_x = 21 + e2x1$) hidden neurons and log sigmoid ($i.e. g_x = 11 + ex$) output neurons using the Levenberg–Marquardt algorithm. Meanwhile, a cross-validation procedure was employed to set the number of nodes per layer

3.2.5 Transformer (Trans-q-lr)

There has been a plethora of deep models for tabular data presented in recent years [37][53][3][74][6][29][73]. However, these models are not consistently superior to decision tree ensembles such as GBDT (Gradient Boosting Decision Tree) in systematic evaluations. In addition, follow-up research has shown that the proposed

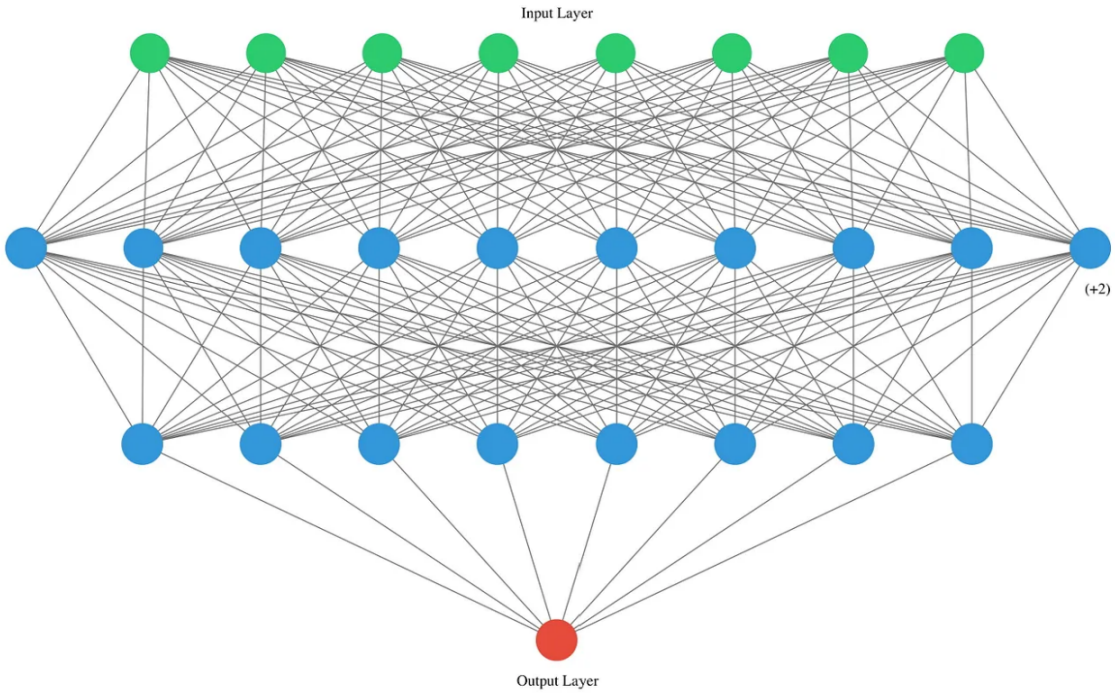

Figure 3.2: Simplified representation of Neural Networks

complicated architectures are not superior to well-adjusted basic models like MLP and ResNet. In contrast to the aforementioned literature, we do not plan to use a unique backbone design in this investigation. Instead, we focus on finer-grained approaches to dealing with numerical features; these developments may be utilized in tandem with any model, whether it's a basic MLP or something more complex like the Transformer model of the future. There have been a plethora of deep models for tabular data proposed in recent years [37][53][3][74][6][29][73]. Ensembles of decision trees, such as GBDT (Gradient Boosting Decision Tree) [26] [55], are often the first-choice in different ML competitions [26][70]. The proposed complex architectures have also been demonstrated to be inferior to well-tuned basic models, such as MLP and ResNet, in a number of recent publications [35]. In contrast to the existing literature, our goal here is not to suggest a fundamentally different infrastructure design. Our work instead centers on more precise methods for dealing with numerical features, and it is designed to be easily integrated with a wide variety of models, from classic MLPs to the most cutting-edge Transformer-like variants. The general framework for what we call "embeddings for numerical features", shows their usage for MLP-like architectures, and describes the main building blocks used in the experimental comparison from the work by [25]. For implementation, We use California Housing(CA) dataset from the previous work on tabular DL and Kaggle competitions. Preliminary data preprocessing is known to be crucial for the optimization of tabular DL models. For this dataset, we use the quantile transformation from the Scikit-learn library [51]. We also apply standardization to regression targets for all algorithms. Then for tuning on dataset, we carefully tune each model's hyperparameters. The best hyperparameters are the ones that perform best on the validation set, so the test set is never used for tuning. For most algorithms, we use the Optuna library [1] to run Bayesian optimization (the Tree-Structured Parzen

Estimator algorithm), which is reported to be superior to random search [77]. For each tuned configuration, we run 15 experiments with different random seeds and report the average performance on the test set. Then, we obtain three ensembles by splitting the 15 single models into three disjoint groups of equal size and averaging predictions of single models within each group. The implementations of the Transformer backbone was taken from [26]. We minimize mean squared error for regression problems. We use the AdamW optimizer [41]. We do not apply learning rate schedules. For CA dataset, we use a predefined batch size. We continue training until there is patience + 1 consecutive epoch without improvements on the validation set; we set patience = 16 for the transformer model. When differentiable components are presented, we tune the output dimensions of the corresponding linear layers. Hyperparameters that define the PLE-representations (Piecewise Linear Encoding) are tuned and they are the same for all features. For quantile-based bins, we tune the number of quantiles. For embeddings based on the Periodic module, we tune σ (it is the same for all features). we use the combination of backbone and embedding i.e. the "Backbone-Embedding" pattern to name the models, where "Backbone" denotes the backbone (e.g. Transformer) and "Embedding" denotes the embedding type (e.g. ReLU, Linear, PLEq). The implementation of Transformer-Q-LR model was done by using a docker container setup developed at IITM Pune.

3.2.6 Optuna Hyperparameter Tuning

As we know GridSearchCV and RandomizedSearchCV work well when we have a small number of parameters to be optimized. So when parameters get increased these optimization techniques become computationally expensive. So to optimize in a better way when a number of parameters are large we have OPTUNA. A number of algorithms are provided by the well-known Python package Optuna for determininging the best hyperparameters. The process of choosing the appropriate hyperparameters for a machine learning model is known as hyperparameter optimisation. Hyperparameters are the parametersters that must be defined before the model is trained because they cannot be learned from the data. Optuna effectively searches the hyperparameter space and minimizes the objective function via Bayesian optimization. Using earlier assessments, it builds a model of the objective function and offers fresh sets of hyperparameters based on model predictions. You must build an objective function that accepts hyperparameters as input and outputs a scalar value that represents the model's performance in order to use Optuna for hyperparameter optimization. Following an iterative evaluation of the objective function with various hyperparameter values, Optuna updates the model and suggestsgests further hyperparameters for evaluation based on the findings.
Optuna makes it simple to improve the hyperparameters of multiple machine learning models by offering a number of built-in integration with well-known machine learning libraries, including Scikit-learn, PyTorch, and TensorFlow. Furthermore, it offers distributed computation, allowing for effective hyperparameter optimization on substantial datasets or intricate models. It uses the Bayesian optimization technique.

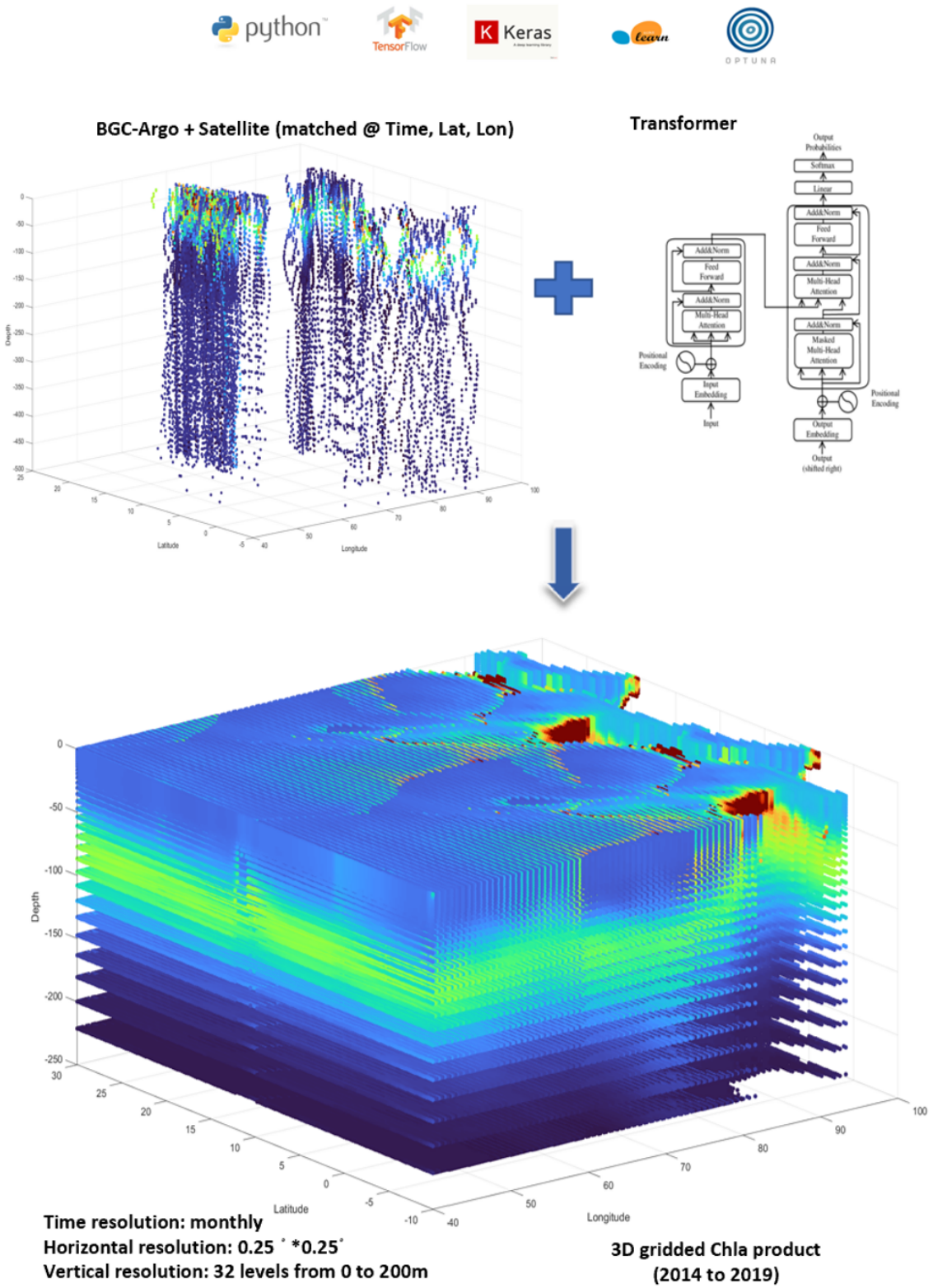

Figure 3.3: Overview of the Transformer model-based method that the 3D structure of Chla inferred from merged satellite-derived products and BGC-Argo float data.

1. Build a surrogate probability model of the objective function.
2. Find the hyperparameters that perform best on the surrogate.
3. Apply these hyperparameters to the true objective function.
4. Update the surrogate model incorporating the new results.
5. Repeat steps 2-4 until max iterations or time is reached.

Surrogate probability model means it takes the parameter which we have given and prepares a probability model over it. The probability model means a function that it prepares by doing a hit and trial and it selects a build-up function and then it applies those hyperparameters which we have given to it over the surrogate probability model and the best-performing hyperparameters it will select on that surrogate probability model and apply the same probability to the true objective fun that has to maximize. It uses the terms study and trial as follows: Study: optimizationtion based on an objective function Trial: a single execution of the objective function.

```python
import optuna
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

def objective(trial):
    # Define the search space for the hyperparameters
    criterion = trial.suggest_categorical("criterion", ['squared_error']) #, 'absolute_error', 'poisson'
    max_features = trial.suggest_categorical("max_features", ['sqrt', 'log2', 'auto'])
    max_depth = trial.suggest_int("max_depth", 2, 32, log=True)
    n_estimators = trial.suggest_int("n_estimators", 100,500)
    min_samples_split = trial.suggest_int("min_samples_split", 2,20)
    min_samples_leaf = trial.suggest_int("min_samples_leaf", 1,10)

    # Create a random forest regressor model with the hyperparameters
    model = RandomForestRegressor(criterion =criterion,
            max_depth=max_depth,
            n_estimators=n_estimators,
            max_features = max_features,
            min_samples_split = min_samples_split,
            min_samples_leaf = min_samples_leaf)

    # Train the model
    model.fit(train_x, train_y)

    # Predict on the test set
    y_pred = model.predict(test_x)

    # Calculate the mean squared error on the test set
    mse = mean_squared_error(test_y, y_pred)

    return mse

# Define the study to search for the best hyperparameters
study = optuna.create_study(direction='minimize')
study.optimize(objective, n_trials=100)

# Print the best hyperparameters found
print(f'Best parameters: {study.best_params}')

```

Figure 3.4: Demonstration of Optuna implementation on Random Forest Regression model

3.3 Evaluation Metrics

The chlorophyll observation and prediction values are split 80:20 across a training and test set. Root-Mean-Square Error (RMSE) or Root-Mean-Square Deviation (RMSD), determination coefficient (R2), and correlation coefficient (R) were used to evaluate performance on test data. During model training, the weight values that cause the loss to be minimized are identified. The Mean Squared Error(MSE) is one of the simplest and most common loss functions. To determine the MSE, we square the disparity between the model's predictions and the observed data and then take an average over the entire dataset. The formula for the MSE is:-

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

RMSE measures the Euclidean distance between predicted values and actual values, it provides the measure that how far the prediction is from the actual or true values. It is calculated by finding the square of the residuals (i.e. the differences between the predicted and true values) at every data point, then the average of the residual is computed, and finally, the square root of the average is calculated to obtain the RMSE. The formula for RMSE is given below. Root mean square error can be expressed as:-

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

The coefficient of determination (R^2) is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable when predicting the outcome of a given event. In other words, this coefficient, more commonly known as R-squared (or R^2), assesses how strong the linear relationship is between two variables the determination coefficient (R^2) is defined as:-

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

Chapter 4

Results and Discussion

4.1 Develop and validate machine learning models to estimate chlorophyll profiles

4.1.1 Model performance

In this section, we will test how well our model performs on a subset of our data and report our findings. In the early stages of this study, it is crucial to show that the machine learning approaches for chlorophyll estimation are feasible by using a wide range of ocean data. Using temperature, salinity, pressure, latitude, and longitude as input factors, we have begun working with BGC-Argo float data in the Bay of Bengal region to estimate chlorophyll. All model's performances were analyzed with the validation data and compared using R^2 and RMSE values as measures. Table 4.1 displays the comparisons between the R^2 and RMSE values for estimated (using the RF, SVR, or NN model) and measured (using the cphl) values for each station. By and large, the RF model outperforms its closest competitors—MLR, SVR, and ANN.

Table 4.1: Performance of machine learning models

S.No	Model	R^2	RMSE
1	Multivariate L.R.	0.38	0.094
2	Support Vector R.	0.52	0.089
3	Neural Network	0.57	0.083
4	Random Forest	0.69	0.02

4.1.2 Prediction accuracy of cphl vertical profiles

All four machine learning methods were used to predict the cphl vertical profile in the BoB region from 2014 to 2019. Figure 4.1 depicts this comparison at a given time step (i.e., April 10, 2016) by using color to denote the various models under consideration. When compared to other models, the RF model provides the best fit because it accounts for all types of seasonal and interannual variability. Figure 4.1 shows that RF has a Pearson's correlation coefficient of more than 0.85 for predicting deep chlorophyll maxima (DCM) in the upper ocean's top 200 m layer, with a significance level of more than 99.9 percent.

The effectiveness of several machine learning models can be compared by plotting a scatter diagram against the actual data. The y-axis of the scatter plot shows the

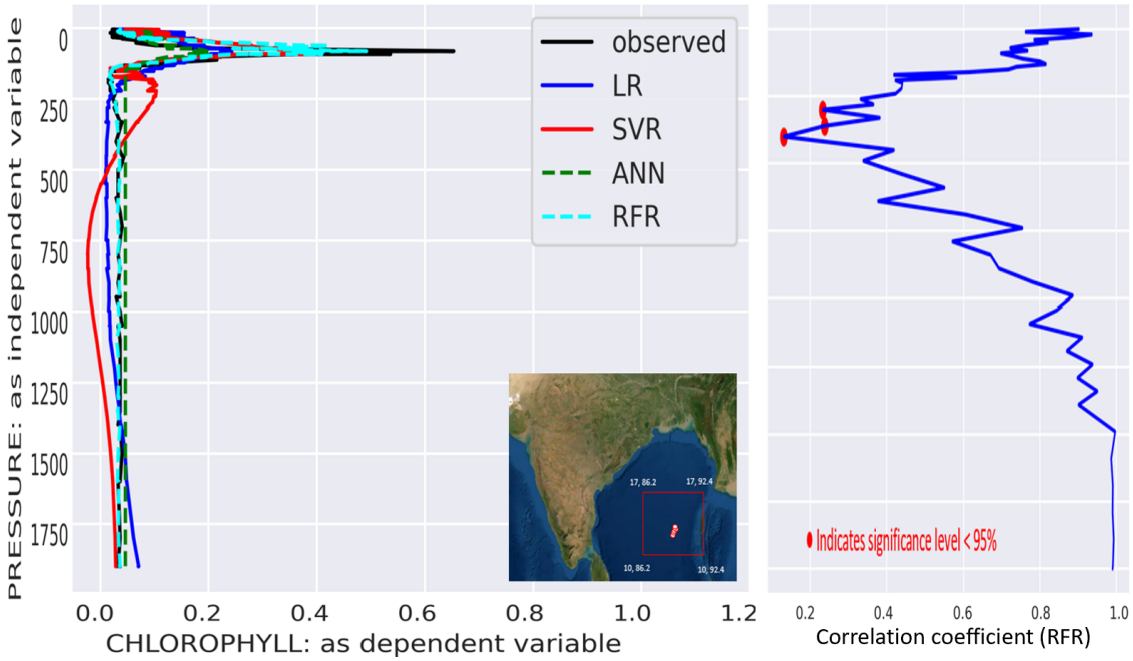

Figure 4.1: vertical profiles from different models as mentioned in color label along with the vertical profile of correlation and insignificant points as red dots

model predictions, while the x-axis shows the actual values. The position of each data point on the scatter plot indicates the accuracy of the prediction provided by the corresponding model.

A perfect match between the projected and actual values would be shown by the close proximity of the data points to a diagonal line. The model's predictions may be off and the possibility for improvement exists if the data points are dispersed around the plot. Scatter plots are useful for visually comparing the results of different machine learning models and understanding how well they can predict the true values. Figure 4.2 shows that the RF model has better predictive abilities, as the green spots are the least dispersed.

4.1.3 Prediction of DCM and Surface chlorophyll

In this section, we have presented the results in the form of visual plots, specifically time series plots of the vertical profile of chlorophyll in the Bay of Bengal spanning the years 2015 to 2017 and 2015 alone as shown in figure 4.3. These plots serve as a visual representation of the data and provide a clear and concise way to compare the skills of best-performing RF model predictions with observations from Bgc-Argo data. By comparing the machine learning model predictions with the observed data from ARGO floats, I am able to assess the skill of the model in capturing the temporal variations in chlorophyll concentration.

Figure 4.3 clearly illustrates the similarities and differences between the model

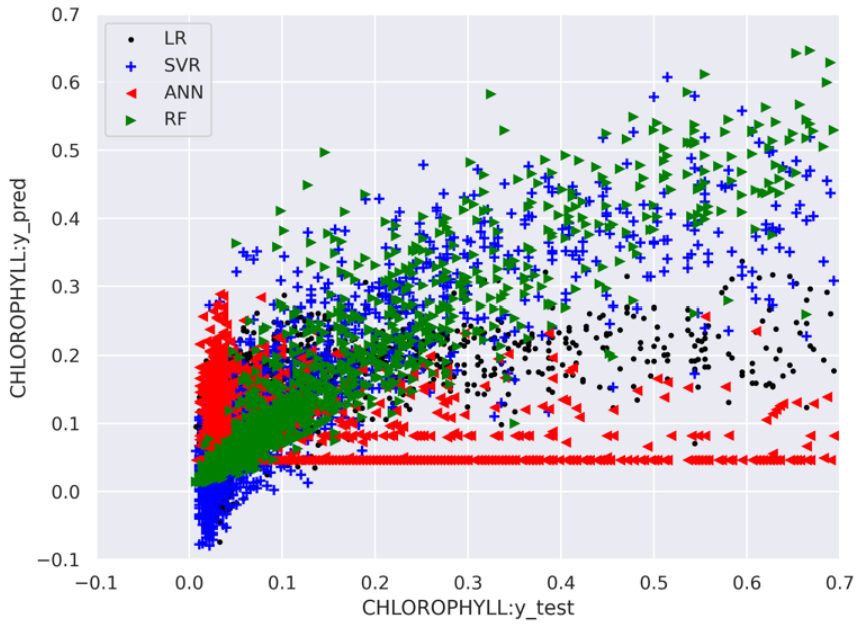

Figure 4.2: scatter plot of chla from test dataset and chla from predictions of different models

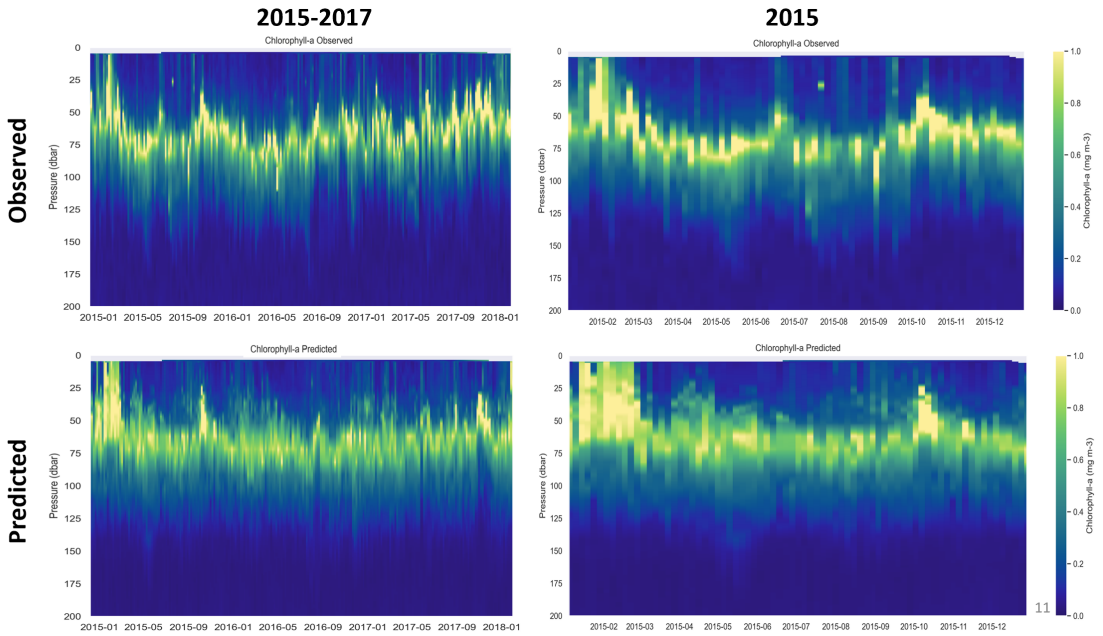

Figure 4.3: visual representation of RF model performance with observation in vertical chla time series plot from 2015 to 2017 (left) and 2015 (right)

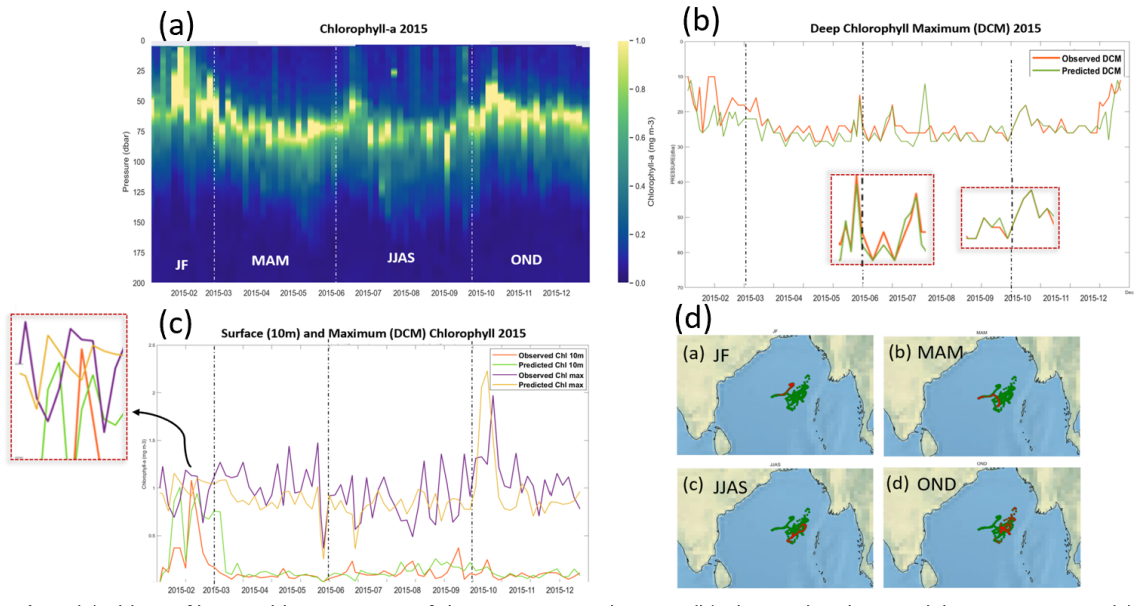

Figure 4.4: (a) Chla profile monthly time series of the year 2015 in the BOB. (b) observed and RF model DCM time series (c) observed and RF model surface and max. chla timeseries. (d) Seasonal profile locations

predictions and the observed data, allowing for a comprehensive evaluation of the model's performance. Figure 4.3 shows a time series plot of the RF-predicted chla versus the observed chla values. Figure 4.4(a) is a seasonal separation line superimposed on a time series of the BGC Argo measured chla profile. Figure 4.4(b) displays the deep chlorophyll maxima (DCM), while Figure 4.4(C) depicts the maximum and surface time series.

The time series plots of the vertical profile of chlorophyll concentration in the Bay of Bengal from 2015 to 2017 provide a visual representation of the temporal dynamics of phytoplankton in the region. They serve as a powerful tool for comparing the accuracy of machine learning model predictions with observed data and help to strengthen the validity and robustness of my research findings.

4.2 Reconstruction of the 3D structure of chlorophyll in the North Indian Ocean

4.2.1 Model Performance

In this research project, several machine learning models were employed, including the transformer model, random forest, and neural networks, to predict the outcomes of the data analyzed. The results show that the transformer model outperformed the other models, achieving an R-squared value of 0.94, indicating that it explained 94% of the variability observed in the data as shown in Figure 4.5. The random forest model also performed well with an R-squared value of 0.92, indicating it explained 92% of the variability observed in the data. In contrast, the neural network model performed relatively poorly, achieving an R-squared value of 0.831. These findings suggest that the transformer and random forest models are appropriate for this

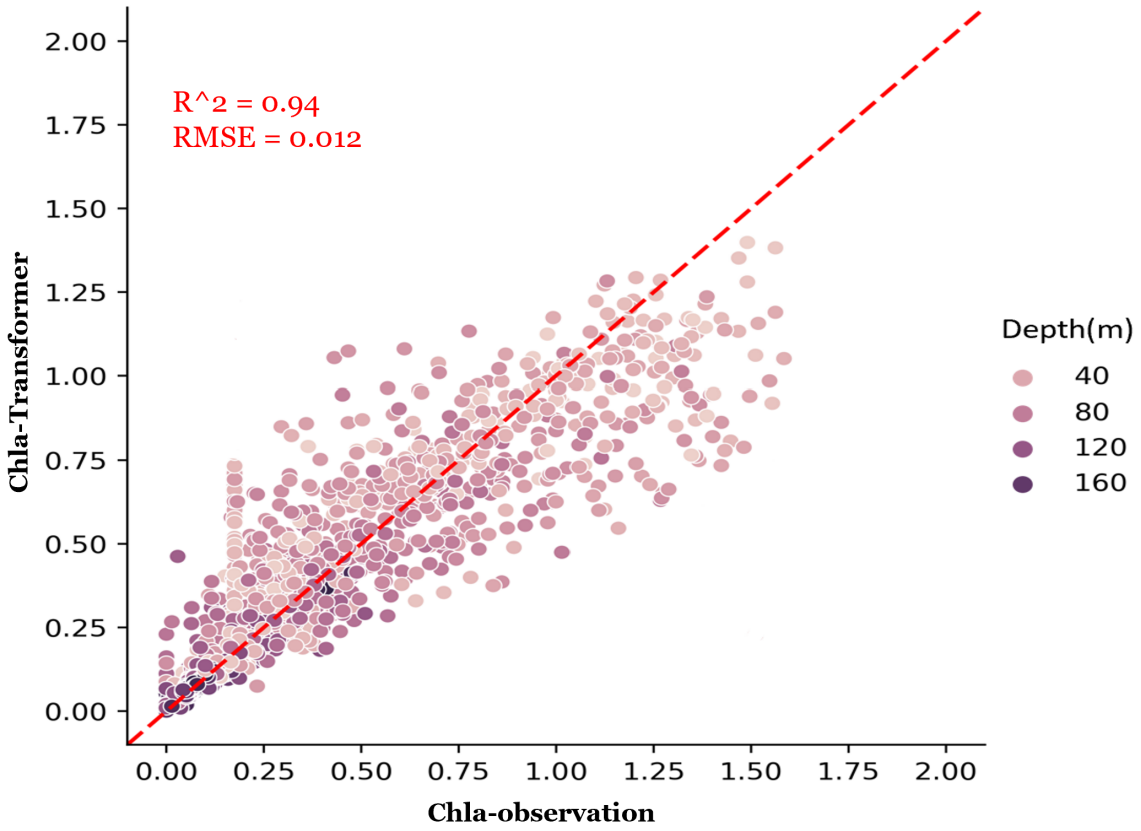

Figure 4.5: scatter plot of observed chlorophyll on x-axis with the transformer model predicted chlorophyll on y-axis

particular research problem, as they can effectively capture the complexity of the data and explain most of the variation. It is important to note that the neural network model may have the potential for improvement by tweaking its parameters, but in this study, it was not as effective as the other models. These results have implications for future research in this area, as well as practical applications for using machine learning models in similar contexts. Overall, the results highlight the importance of carefully selecting and testing multiple models to determine the most effective approach for a given research problem.

4.2.2 Sensitivity study

In this part, we also saw how changing one or more of the input variables affected the performance of the Transformer model and the RF-Model (Figure 4.6). The surface variable Kd490 and the vertical variable temperature were chosen as proxies. The accuracy of both the Transformer model and the RF-Model declined when Kd490 (Temperature) was eliminated, with the RMSE increasing from 0.012 to 0.017 (0.021) and 0.0178 to 0.019 (0.026) (Figure 4.6). Also assessed was the impact of both horizontal (depth, temperature, and salinity) and vertical (surface) variables (Rrs at 412, 443, 490, 560, and 665 nm, Kd490, PAR, SLA, u, and v) on these models precision. The accuracy was drastically diminished after the surface/vertical

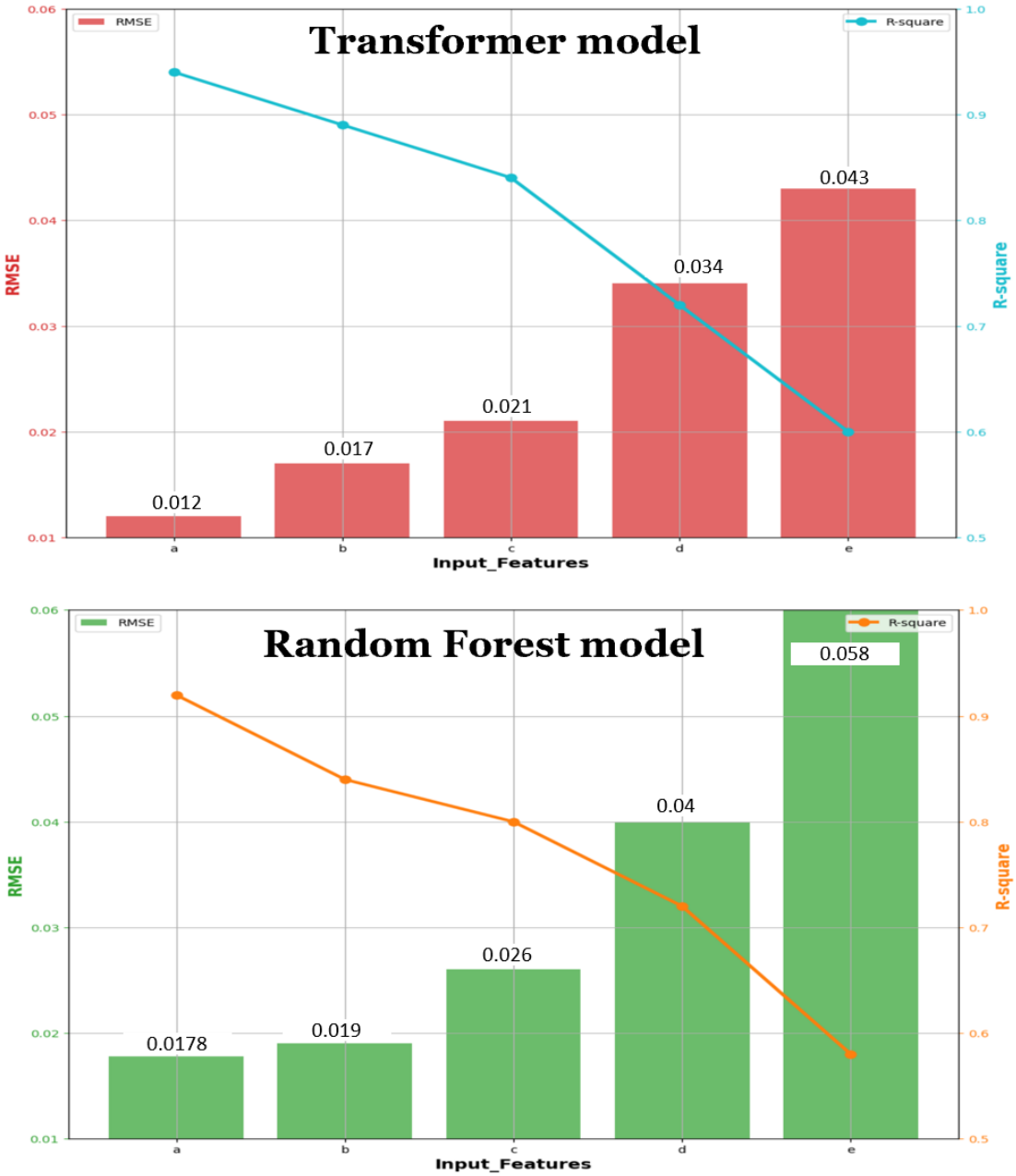

Figure 4.6: Transformer and RF-Model sensitivity changes for removing single and multiple input variables, respectively

variables were removed, where R2 dropped to 0.94 from 0.6 for the transformer model and R2 dropped to 0.92 from 0.58 for the RF model. At the same time, RMSE rose from 0.012 to 0.043 for the transformer model. Furthermore, we discovered that the multiple vertical variables contribute more to the performance of these models than the surface variables, suggesting that the vertical physical properties of the water column should be taken into account when reconstructing the vertical Chla variables.

4.2.3 Validation of surface and vertical chlorophyll

1) Validation of Surface Chla: A comparative study was carried out between the Transformer chla profile product, the RF Chla profile product, the NN chla profile product, observation data (from satellites and BGC-Argo), and global ocean 3D Chla products generated by the NEMO model. This study was carried out with the purpose of further validating the temporal and spatial accuracy of our Chla profile products. In Figure 4.7, the OC-CCI [Figure 4.7(a-d)], Transformer Model [Figure 4.7(e-h)], RF Model [Figure 4.7(i-l)], NN Model [Figure 4.7(m-p)], and NEMO Model [Figure 4.7(q-t)] are each used to represent the seasonal variation of surface Chla in the NIO, respectively. We evaluate the surface RF Chla values by contrasting them with the satellite data and global ocean 3D Chla products that correspond (using the NEMO-Model). When it comes to seasonal cycles, the geographic patterns and magnitude of the surface Chla values that were derived from the Transformer model and the RF-Model show a good agreement with those that were seen by ocean color satellites composite. This is the case regardless of whatever model was used to obtain the values. When summer and autumn come around, the quantity and spatial extent of phytoplankton blooms in the western border and central AS that are generated from the NN and NEMO techniques are, to some degree, underestimated. On the other hand, the Transformer model does a good job of representing the blooms that have actually been observed. The Transformer-model and the RF-model are able to capture phytoplankton blooms in the northern AS pretty effectively over the winter and spring seasons, however the NN Model and the NEMO Model are unable to do so. In a similar fashion, low Chla levels (less than 0.25 mg/m3) with minimal seasonal fluctuation were found to be disclosed by the Transformer Model in the equatorial band and BoB. These findings may be found in the matching ocean color satellite measurements. In addition, in order to evaluate how well the transformer model can predict the future, further research is conducted on 2 typical places that experience large Chla changes: eastern Sri Lanka, and the western Asian subcontinent. These regions are depicted in Figure 4.9 and Figure 4.10, respectively. Also, we have done time series analyses for the entire decade of 2014-2019 (not shown in figure), and we discovered that the seasonal and interannual changes of surface Chla portrayed from the Transformer Model are highly compatible with satellite data. In fact, they are more consistent with satellite observations than the seasonal and interannual variations depicted from the RF, NN, and NEMO-Model.

2). Validation of Vertical Chla Pattern: In order to acquire fully depth-resolved vertical profiles of Chla in the 0-200 m layer, we performed a data-interpolating variational analysis (DIVA) interpolation between each of the Chla profile values that were produced from the BGC-Argo, Transformer-model, RF-Model, NN-Model, and NEMO-Model (Figure 4.8). This allowed us to obtain fully depth-resolved vertical

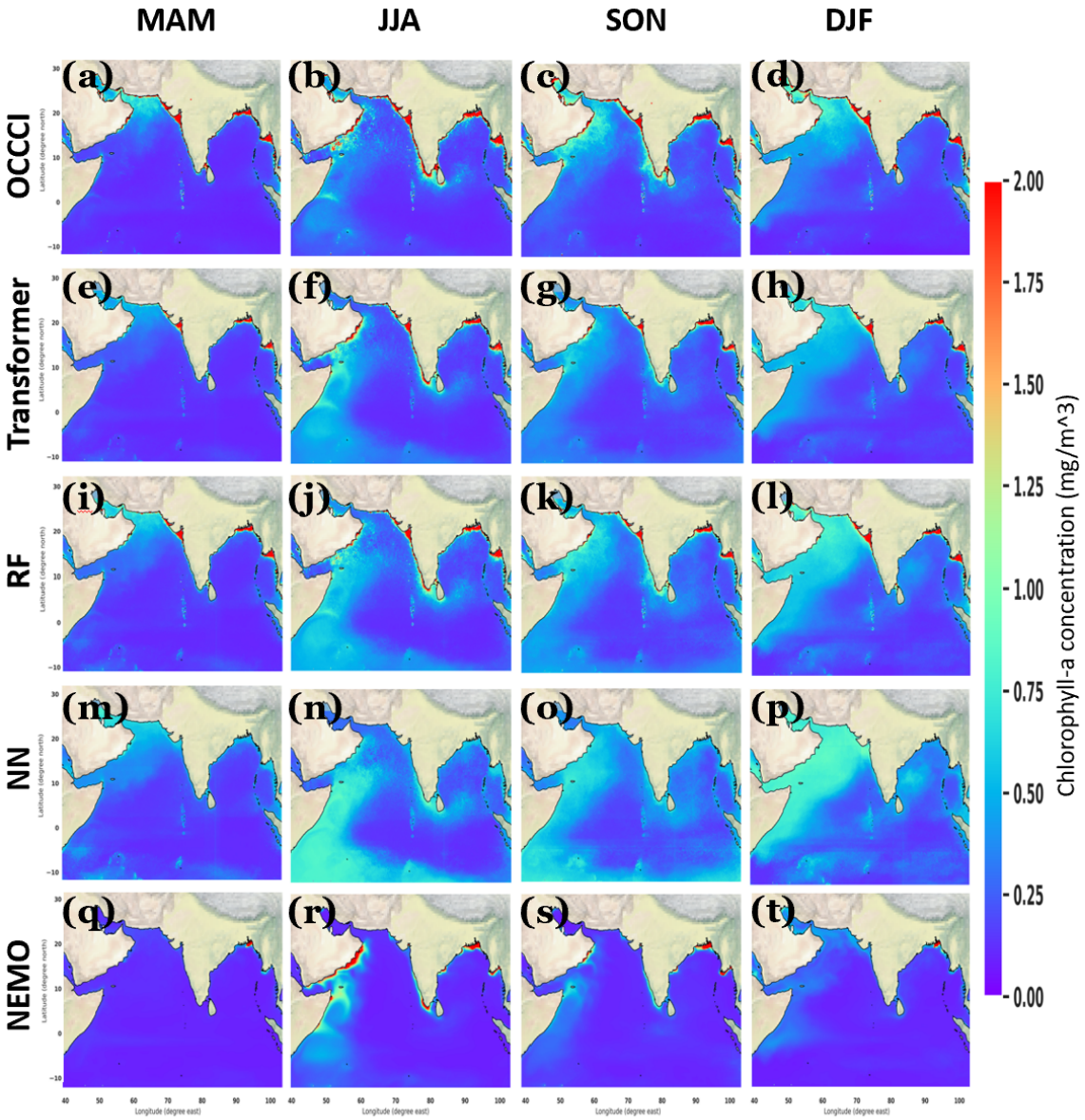

Figure 4.7: Seasonal variability of surface Chla in the northern Indian Ocean from OC-CCI (a-d), Transformer-Model(e-h), RF-Model(i-l), NN-Model(m-p), and NEMO-Model(q-t) for spring season i.e. MAM (a,e,i,m,q), summer season i.e JJA (b,f,j,n,r), autumn season i.e SON (c,g,k,o,s) and winter season i.e DJF (d,h,l,p,t)

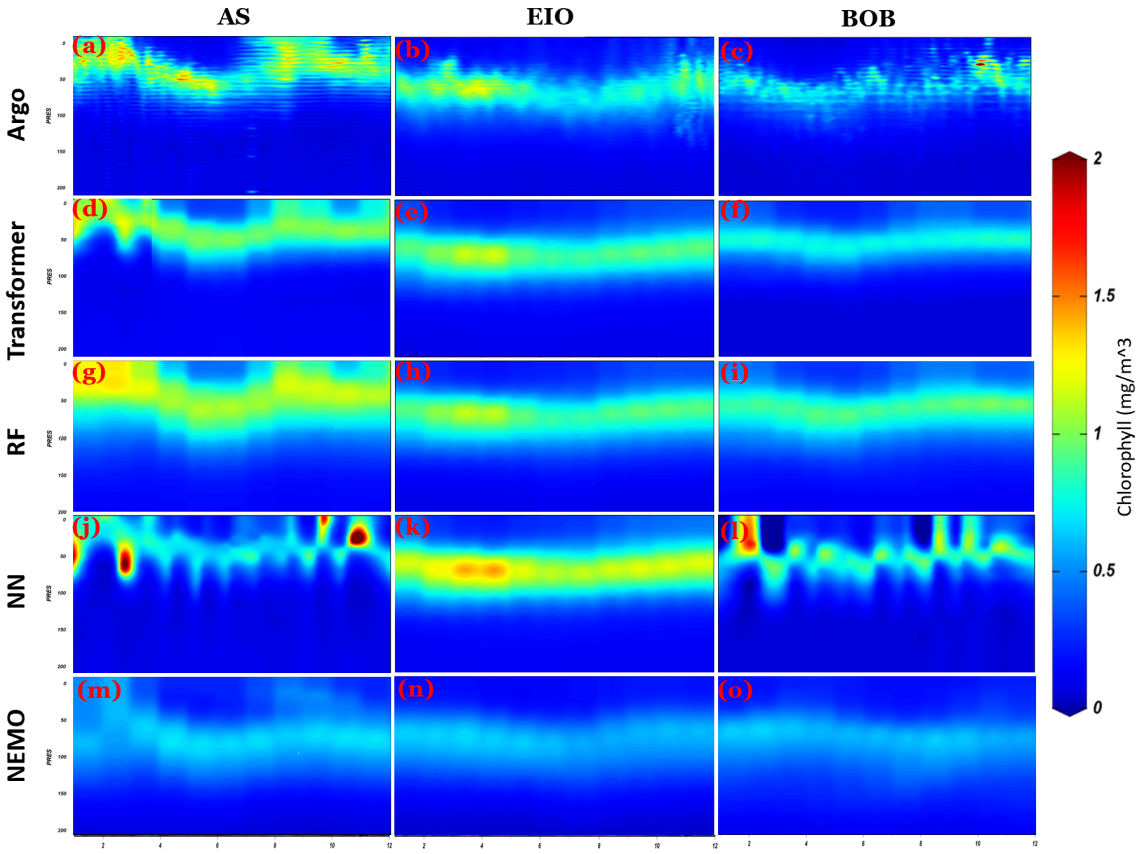

Figure 4.8: Monthly climatology of vertical Chla in the Arabian Sea(AS) (c,f,i,l,o), the central tropical Indian Ocean (CIO) (b,e,h,k,m), and the Bay of Bengal (BOB) (a,d,g,j,m) from BGC-Argo (a-c), Transformer-Model (d-f), RF-Model (g-i), NN-Model (j-l), and NEMO-Model(m-o), respectively.

profiles of Chla. The resulting seasonal cycle is then compared to its float equivalent for each of the three typical locations (AS, CIO, and BoB; see Figure 4.8). The Transformer and RF model's predictions of the vertical patterns of Chla are in agreement with the patterns that the BGC-Argo saw in the 0-200 m layer. In the AS, phytoplankton blooms and the DCM co-exist in the seasonal cycle. Transformer-Chla reproduces the phytoplankton blooms at a depth of 0-50 meters during the summer and winter, while the DCM is present during the months of March to June and September to November. In every season, the values of Chla that were received from the CIO agreed with the data from the BGC-Argo. The exceptionally high (low) Chla values at 50-100 m in March-April (June-July) are caught by the Transformer and RF-Model to a reasonable degree of accuracy. With a significant DCM present throughout the year, the RF-Chla readings in the BoB are in agreement with the BGC-Argo float observations. When compared to the BGC-Argo data, the Chla profile datasets that were produced from the NN and NEMO approaches still display non-negligible gaps in temporal and spatial scales. On the other hand, the vertical Chla profile that was retrieved from the Transformer-model was found to be more precise and resilient. Further validation of the surface and vertical retrieved-Chla profile was carried out in the western tropical Indian Ocean close to the Somali coast.

4.2.4 Capturing the biological response on the southern tip of India and around Somali coast

In response to the summer monsoon, it is widely accepted that the western part of the Arabian Sea has an exceptionally high level of biological output. The entrainment of nutrients into the mixed layer can be attributed to upwelling, which can be driven by alongshore winds or by Ekman pumping in the open ocean. As a result of coastal upwelling, phytoplankton blooms of a smaller intensity are known to occur along the west coast of Arabian Sea during the summer monsoon season (see Figure 4.10).

Also, during the peak of the summer monsoon, the oceanic region that surrounds the southern tip of India (Figure 4.9) is home to a very active and dynamic physical environment. The southwest monsoon current (SMC) may be found to the south of India and moves in an easterly direction. It receives its nutrients from two separate flows in the Arabian Sea: one moving westward from the south-central region, and the other moving southeastward from the southeast region. The SMC makes a sharp curve to the northeast and eventually empties into the Bay of Bengal just to the east of Sri Lanka. In addition, sections of the SMC can be found in the northern and southeastern parts of the Bay of Bengal. During March, there is some amount of chla in Palk Bay and the rest of the region is almost devoid of any phytoplankton. Chla patches higher than 1 mg/m3 appear along the southern coast of Sri Lanka during April. By the month of May, there is high chla along the Indian coast as well as along the southern coast of Sri Lanka. The chla bloom is fully developed during June–July-August and there is a high chla in the Gulf of Mannar, Palk Bay, and along the southern coast of Sri Lanka [82]. The chla in the upper layer of tropical oceans is, in general, limited by the availability of nutrients. Therefore, oceanic processes that can bring nutrients into the euphotic zone are of prime importance. Nutrients can be brought in by coastal upwelling driven by alongshore winds, open ocean upwelling driven by Ekman pumping, entrainment due to wind stirring at the base of the mixed layer, and by horizontal advection due to ocean currents.

Satellite observations showed that phytoplankton blooms occurred in summer and winter along the coast of Somalia, and the intensity of phytoplankton blooms in summer is stronger than that in winter. Here, phytoplankton blooms at the surface during winter and summer monsoons were well characterized by the OCCCI composite and Transformer-Model [Fig. 12(b) and (c)]. Compared with NEMO-Model, the Transformer-Model depicted a similar seasonal variation pattern near Somali coast. Taken together, the Transformer-Model has a better performance to retrieve 3D Chla structure in the NIO.

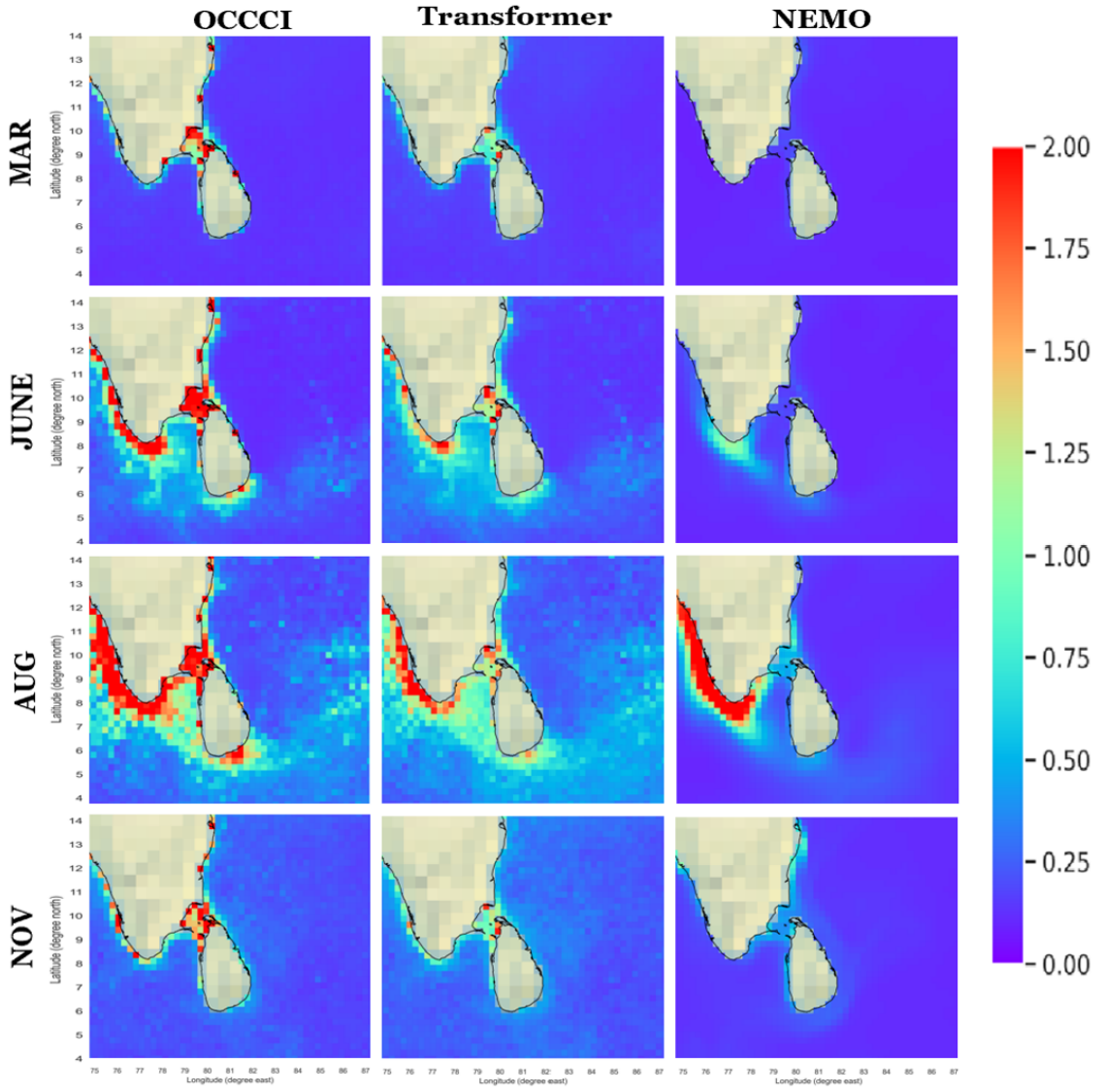

Figure 4.9: Monthly surface chla comparison of Transformer model with OCCCI composite and NEMO model along the southern tip of India.

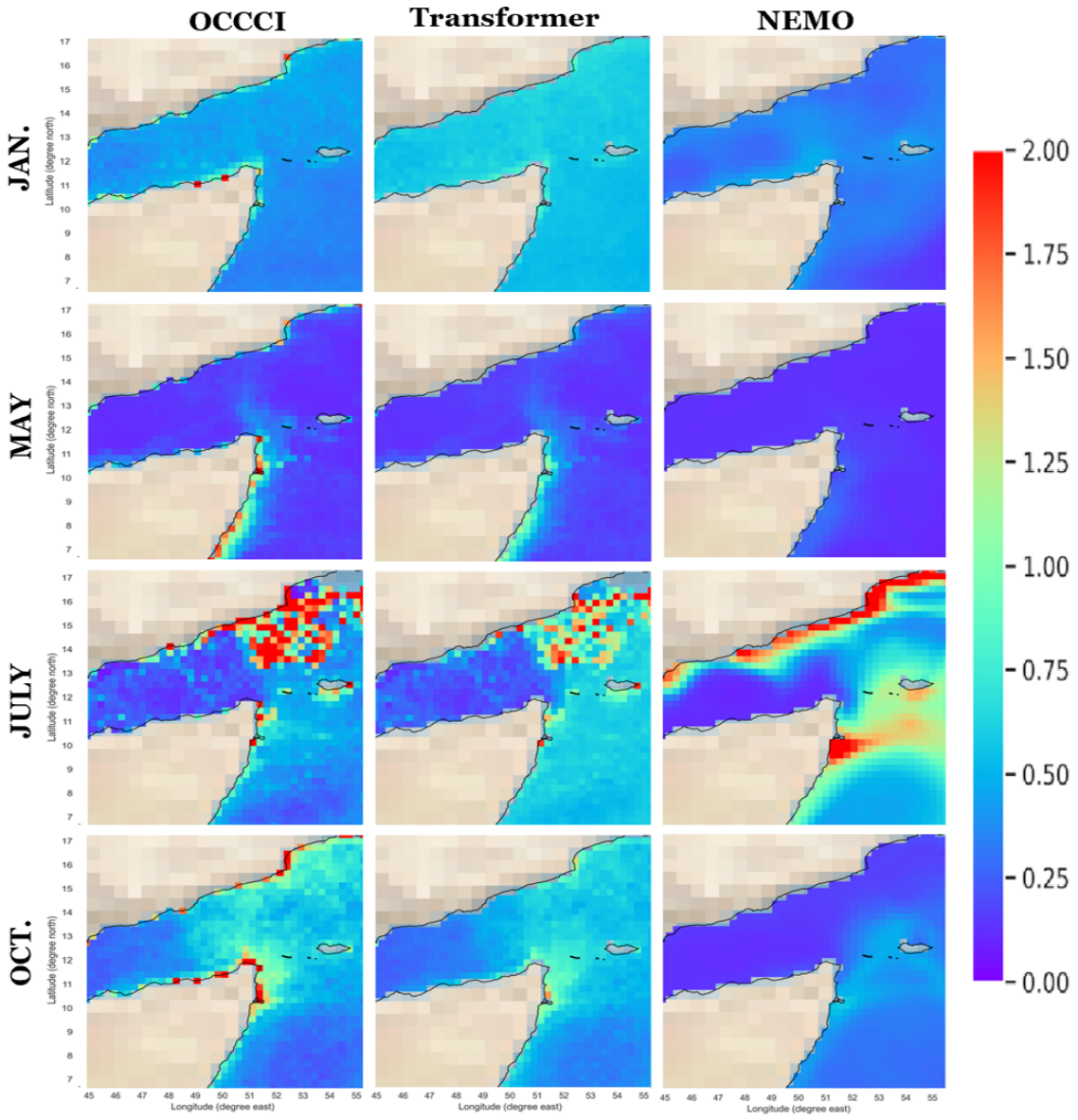

Figure 4.10: Monthly surface chla comparison of Transformer model with OCCCI composite and NEMO model along the coast of Somalia

Chapter 5

Conclusions

The vertical distribution of Chla was predicted using a transformer-based technique that integrated satellite-derived products with BGC-Argo float profiles. This research is a crucial resource for understanding the lateral distribution of Chla in the NIO. The proposed transformer based method, uses day, longitude, and latitude in addition to the surface variables provided by satellites and the vertical variables, such as temperature and salinity measured by BGC-Argo floats, to infer the vertical distribution of Chla. The Transformer model is meant to be adaptable to most open-ocean habitats in the NIO because it is trained on a huge database representing the different tropic conditions existing in the three typical regions (AS, CIO, and BoB). Because of this, estimated Chla can be relied upon as a model used for oceanographic research; for example, the climatological data generated by the Transformer model can be utilized as an initialization or validation dataset to enhance the precision of biogeochemical models. The creation of long-term series products to illustrate 3D Chla fluctuation under future climate change scenarios is of special interest. The ability to obtain vertically resolved 3D models of phytoplankton structures and biomass is an important outcome of this research. Also in this study, we provide a unique data-driven approach for the investigation of the presence and variability of phytoplankton blooms in the oceanic regions of the North Indian Ocean. Better characterization and quantification of NIO carbon fluxes is a crucial step in this direction. The vertical structure of chla through time has been suggested to be studied using a method that is more effective than physics-based models.

5.1 Limitations of the study

Every research has limitations and ours is no exception. Here are some possible limitations of our study: 1. Limited spatial coverage: Although we were able to obtain a considerable amount of data from the open ocean, the study focused mainly on the North Indian Ocean. Therefore, the results obtained may not be applicable to other regions.

2. Less reliable in coastal water: Despite the use of BGC-Argo data, the lack of in situ data from the coastal regions may limit the accuracy of the study's results. This is because the in situ data could have provided a more accurate representation of the chlorophyll concentration in the coastal waters.

3. Data quality: The accuracy of the study's results is heavily dependent on the

quality of the data used. The data used may be subject to errors and uncertainties that could affect the reliability of the study's findings.

4. Limited temporal coverage: Although the study reconstructed the 3D chlorophyll structure over the North Indian Ocean from 2014 to 2019, this is a relatively short time period. A longer temporal coverage could have provided more insights into the long-term variability of chlorophyll concentration in the region.

5. Methodological limitations: The use of machine learning methods to estimate chlorophyll concentration may have introduced biases and uncertainties into the study's findings. Furthermore, the method may not be applicable to other regions with different physical and biological characteristics.

It is important to acknowledge the limitations of this research. By doing so, we can provide a more accurate assessment of the applicability and reliability of our results, as well as suggest potential areas for improvement and future research.

5.2 Future scope

This research has laid the foundation for further investigation of the ocean ecosystem in the North Indian Ocean such as climatological data produced by the Transformer model can be considered as an initialization or validation dataset to improve the accuracy of biogeochemical models. In particular, the development of long-time series products to depict the variability of 3D Chla in future climate change scenarios. Deploying more floats will increase the measurement density of Chla profiles in the world's oceans, which will assist expand the training dataset and enhance the precision of our machine learning-based model as the worldwide BGC-Argo project progresses. Therefore, it is important to pay more attention to the evolution of the BGC-Argo database and the quality of profile data, as well as the retrieval of the vertical bio-optical characteristics of seawater. Other bio-optical parameters (such as nitrate and particle backscatter coefficient, POC) recorded by BGC-Argo floats and satellite-derived products may be inferred in the future using similar methods. Also, this research can extend to studying the relationship between chlorophyll and primary productivity, carbon fluxes, and ocean acidification. You may also investigate the impact of climate change on chlorophyll concentration and oceanic processes. Additionally, incorporating coastal data could provide a more generalized model and further improve the accuracy of your model. Finally, this research has significant implications for the management and conservation of marine resources in the region, which could be explored in future studies.

Bibliography

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] M Anju, MG Sreeush, V Valsala, BR Smitha, Faseela Hamza, G Bharathi, and CV Naidu. Understanding the role of nutrient limitation on plankton biomass over arabian sea via 1-d coupled biogeochemical model and bio-argo observations. *Journal of Geophysical Research: Oceans*, 125(6):e2019JC015502, 2020.

[3] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[4] Lionel A Arteaga, Emmanuel Boss, Michael J Behrenfeld, Toby K Westberry, and Jorge L Sarmiento. Seasonal modulation of phytoplankton biomass in the southern ocean. *Nature Communications*, 11(1):5364, 2020.

[5] Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. Support vector regression. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pages 67–80, 2015.

[6] Sarkhan Badirli, Xuanqing Liu, Zhengming Xing, Avradeep Bhowmik, Khoa Doan, and Sathiya S Keerthi. Gradient boosting neural networks: Grownet. *arXiv preprint arXiv:2002.07971*, 2020.

[7] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

[8] Michael J Behrenfeld. Climate-mediated dance of the plankton. *Nature Climate Change*, 4(10):880–887, 2014.

[9] Michael J Behrenfeld, Robert T O'Malley, Emmanuel S Boss, Toby K Westberry, Jason R Graff, Kimberly H Halsey, Allen J Milligan, David A Siegel, and Matthew B Brown. Revaluating ocean warming impacts on global phytoplankton. *Nature Climate Change*, 6(3):323–330, 2016.

[10] Michael J Behrenfeld, Robert T O'Malley, David A Siegel, Charles R McClain, Jorge L Sarmiento, Gene C Feldman, Allen J Milligan, Paul G Falkowski, Ricardo M Letelier, and Emmanuel S Boss. Climate-driven trends in contemporary ocean productivity. *Nature*, 444(7120):752–755, 2006.

[11] Marco Bellacicco, Marin Cornec, E Organelli, RJW Brewin, Griet Neukermans, G Volpe, Marie Barbieux, A Poteau, C Schmechtig, F d'Ortenzio, et al. Global variability of optical backscattering by non-algal particles from a biogeochemical-argo data set. *Geophysical Research Letters*, 46(16):9767–9776, 2019.

[12] Henry C Bittig, Tanya L Maurer, Joshua N Plant, Catherine Schmechtig, Annie PS Wong, Hervé Claustre, Thomas W Trull, TVS Udaya Bhaskar, Emmanuel Boss, Giorgio Dall'Olmo, et al. A bgc-argo guide: Planning, deployment, data handling and usage. *Frontiers in Marine Science*, 6:502, 2019.

[13] EB Boss and N Haëntjens. Primer regarding measurements of chlorophyll fluorescence and the backscattering coefficient with wetlabs flbb on profiling floats. 2016.

[14] Daniel G Boyce, Marlon R Lewis, and Boris Worm. Global phytoplankton decline over the past century. *Nature*, 466(7306):591–596, 2010.

[15] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[16] Fei Chai, Kenneth S Johnson, Hervé Claustre, Xiaogang Xing, Yuntao Wang, Emmanuel Boss, Stephen Riser, Katja Fennel, Oscar Schofield, and Adrienne Sutton. Monitoring ocean biogeochemistry with autonomous platforms. *Nature Reviews Earth & Environment*, 1(6):315–326, 2020.

[17] Grace C Chang and Tommy D Dickey. Coastal ocean optical influences on solar transmission and radiant heating rate. *Journal of Geophysical Research: Oceans*, 109(C1), 2004.

[18] Marin Cornec, Hervé Claustre, Alexandre Mignot, Lionel Guidi, Leo Lacour, A Poteau, F d'Ortenzio, Bernard Gentili, and Catherine Schmechtig. Deep chlorophyll maxima in the global ocean: Occurrences, drivers and characteristics. *Global Biogeochemical Cycles*, 35(4):e2020GB006759, 2021.

[19] John J Cullen. Subsurface chlorophyll maximum layers: enduring enigma or mystery solved? *Annual Review of Marine Science*, 7:207–239, 2015.

[20] Sajad Darabi, Shayan Fazeli, Ali Pazoki, Sriram Sankararaman, and Majid Sarrafzadeh. Contrastive mixup: Self-and semi-supervised learning for tabular domain. *arXiv preprint arXiv:2108.12296*, 2021.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] WL Emkey and C Jack. Analysis and evaluation of graded-index fiber lenses. *Journal of Lightwave Technology*, 5(9):1156–1164, 1987.

[23] Madeline C Evans and Christopher S Ruf. Toward the detection and imaging of ocean microplastics with a spaceborne radar. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–9, 2021.

[24] MS Girishkumar, M Ravichandran, and V Pant. Observed chlorophyll-a bloom in the southern bay of bengal during winter 2006–2007. *International Journal of Remote Sensing*, 33(4):1264–1275, 2012.

[25] Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 35:24991–25004, 2022.

[26] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[27] Watson W Gregg, Margarita E Conkright, Paul Ginoux, John E O'Reilly, and Nancy W Casey. Ocean primary production and climate: Global decadal changes. *Geophysical Research Letters*, 30(15), 2003.

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[29] Qiwei Hu, Xiaoyan Chen, Xianqiang He, Yan Bai, Fang Gong, Qiankun Zhu, and Delu Pan. Effect of el niño-related warming on phytoplankton's vertical distribution in the arabian sea. *Journal of Geophysical Research: Oceans*, 126(11):e2021JC017882, 2021.

[30] Qiwei Hu, Xiaoyan Chen, Wanyi Huang, and Fenghua Zhou. Phytoplankton bloom triggered by eddy-wind interaction in the upwelling region east of hainan island. *Journal of Marine Systems*, 214:103470, 2021.

[31] Nils Gunnar Jerlov. *Marine optics*. Elsevier, 1976.

[32] Chenxu Ji, Yuanzhi Zhang, Qiuming Cheng, JinYeu Tsou, Tingchen 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*, 68:252–261, 2018.

[33] BI Justusson. Median filtering: Statistical properties. *Two-Dimensional Digital Signal Prcessing II: Transforms and Median Filters*, pages 161–196, 2006.

[34] R Jyothibabu, NV Madhu, PA Maheswaran, KV Jayalakshmy, KKC Nair, and CT Achuthankutty. Seasonal variation of microzooplankton (20–200 μm) and its possible implications on the vertical carbon flux in the western bay of bengal. *Continental Shelf Research*, 28(6):737–755, 2008.

[35] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941, 2021.

[36] John TO Kirk. *Light and photosynthesis in aquatic ecosystems*. Cambridge university press, 1994.

[37] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

[38] Zhongping Lee, Kendall L Carder, Curtis D Mobley, Robert G Steward, and Jennifer S Patch. Hyperspectral remote sensing for shallow waters. i. a semi-analytical model. *Applied optics*, 37(27):6329–6338, 1998.

[39] Marion R Lewis, Mary-Elena Carr, Gene C Feldman, Wayne Esaias, and Chuck McClain. Influence of penetrating solar radiation on the heat budget of the equatorial pacific ocean. *Nature*, 347(6293):543–545, 1990.

[40] Gang Li, Qiang Lin, Guangyan Ni, Pingping Shen, Yanzhi Fan, Liangmin Huang, and Yehui Tan. Vertical patterns of early summer chlorophyll a concentration in the indian ocean with special reference to the variation of deep chlorophyll maximum. *Journal of Marine Biology*, 2012, 2012.

[41] I Loshchilov and F Hutter. Decoupled weight decay regularization,[in] 7th international conference on learning representations (iclr). *New Orleans, LA, USA, May*, (6-9):2019, 2019.

[42] SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, J Himmelfarb, N Bansal, and SI Lee. From local explanations to global understanding with explainable ai for trees. nat mach intell. 2020; 2: 56–67.

[43] NV Madhu, R Jyothibabu, PA Maheswaran, Vijay John Gerson, TC Gopalakrishnan, and KKC Nair. Lack of seasonality in phytoplankton standing stock (chlorophyll a) and production in the western bay of bengal. *Continental Shelf Research*, 26(16):1868–1883, 2006.

[44] John Marra and Richard T Barber. Primary productivity in the arabian sea: A synthesis of jgofs data. *Progress in Oceanography*, 65(2-4):159–175, 2005.

[45] John Marra, Christopher Langdon, and Carol A Knudson. Primary production, water column changes, and the demise of a phaeocystis bloom at the marine light-mixed layers site (59° n, 21° w) in the northeast atlantic ocean. *Journal of Geophysical Research: Oceans*, 100(C4):6633–6643, 1995.

[46] Charles R McClain, Kevin Arrigo, King-Sheng Tai, and Daniella Turk. Observations and simulations of physical and biological processes at ocean weather station p, 1951–1980. *Journal of Geophysical Research: Oceans*, 101(C2):3697–3713, 1996.

[47] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[48] Curtis D Mobley. Estimation of the remote-sensing reflectance from above-surface measurements. *Applied optics*, 38(36):7442–7455, 1999.

[49] André Morel and David Antoine. Heating rate within the upper ocean in relation to its bio–optical state. *Journal of physical oceanography*, 24(7):1652–1665, 1994.

[50] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

[51] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[52] Trevor Platt, Shubha Sathyendranath, Carla M Caverhill, and Marlon R Lewis. Ocean primary production and available light: further algorithms for remote sensing. *Deep Sea Research Part A. Oceanographic Research Papers*, 35(6):855–879, 1988.

[53] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

[54] S Prasanna Kumar, PM Muraleedharan, TG Prasad, Mangesh Gauns, N Ramaiah, SN De Souza, Smeeta Sardesai, and M Madhupratap. Why is the bay of bengal less productive during summer monsoon compared to the arabian sea? *Geophysical Research Letters*, 29(24):88–1, 2002.

[55] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[56] Marie-Fanny Racault, Shubha Sathyendranath, Robert JW Brewin, Dionysios E Raitsos, Thomas Jackson, and Trevor Platt. Impact of el niño variability on oceanic phytoplankton. *Frontiers in Marine Science*, 4:133, 2017.

[57] M Reichstein, G Camps-Valls, B Stevens, M Jung, J Denzler, and N Carvalhais. the national energy research supercomputing center in lawrence berkeley national laboratory, berkeley, ca, usa: Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204.

[58] Katherine Richardson and Jørgen Bendtsen. Vertical distribution of phytoplankton and primary production in relation to nutricline depth in the open ocean. *Marine Ecology Progress Series*, 620:33–46, 2019.

[59] V Rodriguez-Galiano, M Sanchez-Castillo, M Chica-Olmo, and MJOGR Chica-Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818, 2015.

[60] Collin Roesler, Julia Uitz, Hervé Claustre, Emmanuel Boss, Xiaogang Xing, Emanuele Organelli, Nathan Briggs, Annick Bricaud, Catherine Schmechtig, Antoine Poteau, et al. Recommendations for obtaining unbiased chlorophyll estimates from in situ chlorophyll fluorometers: A global analysis of wet labs eco sensors. *Limnology and Oceanography: Methods*, 15(6):572–585, 2017.

[61] Collin S Roesler and Andrew H Barnard. Optical proxy for phytoplankton biomass in the absence of photophysiology: Rethinking the absorption line height. *Methods in Oceanography*, 7:79–94, 2013.

[62] Mathew Koll Roxy, Aditi Modi, Raghu Murtugudde, Vinu Valsala, Swapna Panickal, S Prasanna Kumar, M Ravichandran, Marcello Vichi, and Marina Lévy. A reduction in marine primary productivity driven by rapid warming over the tropical indian ocean. *Geophysical Research Letters*, 43(2):826–833, 2016.

[63] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[64] DA Sachindra, Khandakar Ahmed, Md Mamunur Rashid, S Shahid, and BJC Perera. Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, 212:240–258, 2018.

[65] Michela Sammartino, Salvatore Marullo, Rosalia Santoleri, and Michele Scardi. Modelling the vertical distribution of phytoplankton biomass in the mediterranean sea from satellite data: A neural network approach. *Remote Sensing*, 10(10):1666, 2018.

[66] Shubha Sathyendranath, Robert JW Brewin, Carsten Brockmann, Vanda Brotas, Ben Calton, Andrei Chuprin, Paolo Cipollini, André B Couto, James Dingle, Roland Doerffer, et al. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (oc-cci). *Sensors*, 19(19):4285, 2019.

[67] Shubha Sathyendranath, Trevor Platt, Carla M Caverhill, Roderick E Warnock, and Marlon R Lewis. Remote sensing of oceanic primary production: computations using a spectral model. *Deep Sea Research Part A. Oceanographic Research Papers*, 36(3):431–453, 1989.

[68] R Sauzède, Hervé Claustre, J Uitz, C Jamet, G Dall'Olmo, F d'Ortenzio, B Gentili, A Poteau, and C Schmechtig. A neural network-based method for merging ocean color and argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient. *Journal of Geophysical Research: Oceans*, 121(4):2552–2571, 2016.

[69] Catherine Schmechtig, Herve Claustre, Antoine Poteau, and Fabrizio D'Ortenzio. Bio-argo quality control manual for chlorophyll-a concentration. version 1.0, december 17th 2014. 2014.

[70] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[71] Greg M Silsbe and Sairah Y Malkin. Where light and nutrients collide: The global distribution and activity of subsurface chlorophyll maximum layers. *Aquatic microbial ecology and biogeochemistry: A dual perspective*, pages 141–152, 2016.

[72] Haykin Simon. Neural networks: A comprehensive foundatio, printice-hall. *Englewood Cliffs, NJ*, 1999.

[73] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[74] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1161–1170, 2019.

[75] Venugopal Thushara, Puthenveettil Narayana Menon Vinayachandran, Adrian J Matthews, Benjamin GM Webber, and Bastien Y Queste. Vertical distribution of chlorophyll in dynamically distinct¡? xmltex\break?¿ regions of the southern bay of bengal. *Biogeosciences*, 16(7):1447–1468, 2019.

[76] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[77] Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, pages 3–26. PMLR, 2021.

[78] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

[79] J Uitz, H Claustre, and A Morel. Hooker s. b., 2006. vertical distribution of phytoplankton communities in open ocean: an assessment based on surface chlorophyll. *Journal of Geophysical Research*, 111(10.1029).

[80] Julia Uitz, Hervé Claustre, André Morel, and Stanford B Hooker. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research: Oceans*, 111(C8), 2006.

[81] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[82] PN Vinayachandran, P Chauhan, M Mohan, and S Nayak. Biological response of the sea around sri lanka to summer monsoon. *Geophysical Research Letters*, 31(1), 2004.

[83] PN Vinayachandran, Yukio Masumoto, Tetsuya Mikawa, and Toshio Yamagata. Intrusion of the southwest monsoon current into the bay of bengal. *Journal of Geophysical Research: Oceans*, 104(C5):11077–11085, 1999.

[84] T Westberry, MJ Behrenfeld, and DA Siegel. i boss, e.(2008). carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles*, 22.

[85] Xiaogang Xing, Emmanuel Boss, Shuangling Chen, and Fei Chai. Seasonal and daily-scale photoacclimation modulating the phytoplankton chlorophyll-carbon coupling relationship in the mid-latitude northwest pacific. *Journal of Geophysical Research: Oceans*, 126(10):e2021JC017717, 2021.

[86] Xiaogang Xing, Hervé Claustre, Emmanuel Boss, Collin Roesler, Emanuele Organelli, Antoine Poteau, Marie Barbieux, and Fabrizio d'Ortenzio. Correction of profiles of in-situ chlorophyll fluorometry for the contribution of fluorescence originating from non-algal matter. *Limnology and Oceanography: Methods*, 15(1):80–93, 2017.

[87] Xiaogang Xing, Hervé Claustre, Julia Uitz, Alexandre Mignot, Antoine Poteau, and Haili Wang. Seasonal variations of bio-optical properties and their interrelationships observed by b io-a rgo floats in the subpolar n orth a tlantic. *Journal of Geophysical Research: Oceans*, 119(10):7372–7388, 2014.

[88] Ren Xu, Nengcheng Chen, Yumin Chen, and Zeqiang Chen. Downscaling and projection of multi-cmip5 precipitation using machine learning methods in the upper han river basin. *Advances in Meteorology*, 2020:1–17, 2020.

[89] Yi Xu, Ying Wu, Huiwu Wang, Zhenqiu Zhang, Jian Li, and Jing Zhang. Seasonal and interannual variabilities of chlorophyll across the eastern equatorial indian ocean and bay of bengal. *Progress in Oceanography*, 198:102661, 2021.

[90] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

[91] Xiaolei Yu, Shuangling Chen, and Fei Chai. Remote estimation of sea surface nitrate in the california current system from satellite ocean color measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021.

[92] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.

[93] J Ronald V Zaneveld, James C Kitchen, and Hasong Pak. The influence of optical water type on the heating rate of a constant depth mixed layer. *Journal of Geophysical Research: Oceans*, 86(C7):6426–6428, 1981.