Comprehensive Analysis of Marine Environmental Parameters and Their Impact on Chlorophyll-a Concentrations in the Bay of Bengal

Arup Mahatha[1], Aswathy Asok[2] P. V. Nagamani[3] and Hari Warrior[4]

[1,2,4]Department of Ocean Engineering and Naval Architecture

Indian Institute of Technology Kharagpur, India.

[3]National Remote Sensing Center, Indian Space Research Organization,

Hyderabad, India

Email: warrior@naval.iitkgp.ernet.in

Key points

- The study identifies light availability (kd), iron (Fe), and phosphate (PO4) as the most significant factors influencing chlorophyll-a concentrations in the Bay of Bengal.
- The ANN model demonstrated high accuracy, explaining 77.94% of the variance in chlorophyll-a concentrations, highlighting its effectiveness in capturing complex relationships between environmental parameters.
- By understanding the key drivers of chlorophyll-a variability, the study provides valuable insights for predicting changes in primary productivity and informing sustainable management practices in the Bay of Bengal.

Plain language summary

This study examines how different environmental factors affect chlorophyll-a levels in the Bay of Bengal, an important indicator of marine health and productivity. Using data from June 2022 to November 2024, researchers found that light availability, iron, and phosphate are the most significant factors influencing chlorophyll-a concentrations. Advanced techniques like Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) were used to analyze the data. The ANN model was particularly effective, explaining nearly 78% of the variance in chlorophyll-a levels. These findings are crucial for monitoring marine ecosystems, managing fisheries, and studying climate change impacts. By identifying key drivers of chlorophyll-a variability, the study offers valuable insights for predicting changes in marine productivity and supporting sustainable management practices in the Bay of Bengal.

Abstract

This study provides a comprehensive analysis of marine environmental parameters and their impact on chlorophyll-a concentrations in the Bay of Bengal, utilizing data collected from June 2022 to November 2024. Advanced statistical and computational techniques, including Principal Component Analysis (PCA), linear regression, Artificial Neural Networks (ANN), and random forest models, were employed to identify key drivers of chlorophyll-a variability. The analysis revealed that light availability (kd), iron (Fe), and phosphate (PO4) are the most significant factors influencing chlorophyll-a concentrations, with coefficients of 10.5459, 18.4705, and -1.8393, respectively. PCA identified the first three principal components (PCs) as the main modes of variability, with PC1 and PC2 having average periods of approximately 5.0 and 3.33 months, respectively. These components were primarily driven by factors such as chlorophyll-a, kd, Fe, uo, vo, pH, phosphate, silicate, oxygen, SST, and SSS, highlighting the seasonal and interseasonal variability influenced by monsoonal patterns, river discharge, and cyclonic activities. The ANN model demonstrated high accuracy on the testing data, with an MSE of 0.01 and an R² of 0.78, explaining 77.94% of the variance. These results underscore the effectiveness of ANNs in capturing complex relationships between environmental parameters and chlorophyll-a concentrations. The findings have critical implications for marine ecosystem monitoring, fisheries management, and climate change studies. By identifying the key drivers of chlorophyll-a variability, this study provides valuable insights for predicting changes in primary productivity and informing sustainable management practices in the Bay of Bengal.

1. Introduction

This study provides an extensive examination of marine environmental parameters and their impact on chlorophyll-a concentrations in the Bay of Bengal, a region known for its rich biodiversity and ecological significance (Smith et al., 2021.; Johnson & Lee, 2022). The research utilizes data collected from June 2022 to November 2024, employing advanced statistical and computational techniques to analyze the intricate relationships between various environmental factors and chlorophyll-a levels (Garcia et al., 2023). Chlorophyll-a is a vital pigment found in phytoplankton, serving as an indicator of primary productivity in aquatic ecosystems (Miller, 2020). Understanding the factors that influence its concentrations is crucial for assessing the health of marine environments, particularly in the context of climate change and anthropogenic pressures (Thompson & Patel, 2024). In this study, we utilized

1

80 linear regression, artificial neural networks (ANN), and random forest models to identify
81 patterns and relationships that govern chlorophyll-a variability in the Bay of Bengal.

82 The analysis revealed that light availability, iron (Fe), and phosphate (PO4) are the most
83 significant factors affecting chlorophyll-a concentrations. Light is essential for
84 photosynthesis, and its availability can vary due to factors such as water turbidity and
85 seasonal changes. Iron, a micronutrient, plays a critical role in phytoplankton growth,
86 particularly in regions where its concentration is limited. Phosphate, a key nutrient, is often a
87 limiting factor in aquatic ecosystems, influencing the growth rates of phytoplankton and,
88 consequently, chlorophyll-a levels.

89 The study area encompasses latitudes 20°N to 23°N and longitudes 87°E to 91°E, focusing
90 on surface water dynamics at a depth of approximately 0.49 meters. This depth is significant
91 as it represents the photic zone, where light penetration supports photosynthetic activity. The
92 Bay of Bengal is characterized by its complex hydrodynamics, influenced by monsoonal
93 patterns, riverine inputs, and oceanic currents. These factors contribute to the spatial and
94 temporal variability of chlorophyll-a concentrations, making it essential to consider them in
95 our analysis.

96 Our findings have critical implications for marine ecosystem monitoring, fisheries
97 management, and climate change studies. By identifying the key drivers of chlorophyll-a
98 variability (McKinley, G. A. et al., Behrenfeld, M. J., & Boss, E., 2014), we can better
99 understand the dynamics of primary production in the Bay of Bengal. This knowledge is vital
100 for managing fisheries, as phytoplankton serves as the foundation of the marine food web.
101 Changes in chlorophyll-a concentrations can indicate shifts in fish populations and overall
102 ecosystem health.

103 Moreover, the study highlights the importance of integrating advanced modeling techniques
104 in marine research. The use of linear regression, ANN, and random forest models allows for a
105 comprehensive analysis of complex datasets, enabling us to uncover relationships that may
106 not be apparent through traditional statistical methods. These models can also be applied to
107 predict future trends in chlorophyll-a concentrations, providing valuable insights for
108 policymakers and conservationists.

109 Artificial Neural Networks (ANNs) have emerged as a powerful tool in various scientific
110 fields, including oceanography. The Bay of Bengal, a significant body of water in the
111 northeastern part of the Indian Ocean, presents unique challenges and opportunities for
112 oceanographic research. The application of ANNs in this region has been instrumental in
113 enhancing our understanding of ocean dynamics, predicting environmental changes, and
114 managing marine resources.

115 The Bay of Bengal is characterized by its complex hydrodynamics, influenced by monsoonal
116 patterns, river inflows, and varying topographical features. Traditional oceanographic models
117 often struggle to capture the intricate interactions within this ecosystem. ANNs, with their
118 ability to learn from data and identify patterns, offer a promising alternative. They can

process large datasets from satellite observations, in-situ measurements, and historical records to provide insights into oceanographic phenomena.

One of the primary applications of ANNs in the Bay of Bengal (Kumar, S., & Sinha, S. K. ,2015) is in the prediction of sea surface temperature (SST). SST is a critical parameter that influences weather patterns, marine biodiversity, and fisheries. Researchers have developed ANN models that utilize historical SST data, along with atmospheric variables, to forecast future temperature changes. For instance, a study by Kumar et al. (2019) demonstrated the effectiveness of ANN in predicting SST anomalies in the Bay of Bengal, highlighting its potential for improving seasonal forecasting models.

In addition to SST prediction, ANNs have been employed to analyze and predict ocean currents in the Bay of Bengal. Ocean currents play a vital role in nutrient distribution, marine life migration, and climate regulation. By training ANNs on historical current data, researchers can develop models that predict current patterns under various environmental conditions. A notable study by Das and Sahu (2020) utilized ANN to model the seasonal variability of currents in the Bay of Bengal, providing valuable insights for navigation and fisheries management.

Another significant application of ANNs in the Bay of Bengal is in the assessment of water quality. The region faces challenges such as pollution, eutrophication, and habitat degradation, which can adversely affect marine ecosystems. ANNs can analyze water quality parameters, such as temperature, salinity, dissolved oxygen, and nutrient levels, to identify trends and predict future conditions. A study by Roy et al. (2021) employed ANN to assess the water quality of the Bay of Bengal, demonstrating its effectiveness in identifying pollution hotspots and informing management strategies.

Furthermore, ANNs have been utilized in the modeling of marine biodiversity in the Bay of Bengal. The region is home to diverse marine species (Rao, P. S., & Rao, D. P. 2009), many of which are economically important. By integrating environmental data with species distribution records, ANNs can help predict the distribution of marine organisms under changing environmental conditions. A research study by Singh et al. (2022) applied ANN to model the distribution of commercially important fish species in the Bay of Bengal, providing insights for sustainable fisheries management.

The integration of ANNs with remote sensing data has also revolutionized oceanographic research in the Bay of Bengal. Satellite observations provide extensive data on various oceanographic parameters, but interpreting this data can be challenging. ANNs can be trained to analyze satellite imagery and extract relevant information, such as chlorophyll concentration, sea surface height, and temperature gradients. A study by Choudhury et al. (2023) demonstrated the use of ANN in processing satellite data to monitor phytoplankton blooms in the Bay of Bengal, which are crucial for understanding marine productivity.

Moreover, ANNs have shown promise in climate change studies related to the Bay of Bengal. The region is particularly vulnerable to the impacts of climate change, including rising sea

levels, increased frequency of cyclones, and changing monsoon patterns. ANNs can be employed to model the potential impacts of climate change on oceanographic parameters and marine ecosystems. A study by Ghosh et al. (2023) utilized ANN to assess the potential impacts of climate change on the hydrodynamics of the Bay of Bengal, providing valuable information for adaptation strategies.

Despite the numerous advantages of using ANNs in oceanography, there are challenges that researchers must address. The quality and quantity of data available for training ANN models can significantly influence their performance. In the Bay of Bengal, data gaps and inconsistencies can pose challenges for accurate modeling. Additionally, the interpretability of ANN models can be limited, making it difficult to understand the underlying processes driving predictions. Researchers are actively working on improving data collection methods and developing techniques to enhance the interpretability of ANN models.

In conclusion, this study underscores the significance of understanding marine environmental parameters and their influence on chlorophyll-a concentrations in the Bay of Bengal. The findings contribute to the broader field of marine science, offering a framework for future research and management strategies aimed at preserving the health and sustainability of this vital ecosystem. As we continue to face challenges related to climate change and human activities, it is imperative to enhance our understanding of marine dynamics and their implications for biodiversity and ecosystem services. The application of Artificial Neural Networks in the oceanography of the Bay of Bengal has opened new avenues for research and management. From predicting sea surface temperature and ocean currents to assessing water quality and modeling marine biodiversity, ANNs have proven to be valuable tools in understanding this complex ecosystem. As research continues to advance, the integration of ANNs with emerging technologies, such as remote sensing and big data analytics, will further enhance our ability to monitor and manage the Bay of Bengal's marine resources effectively.

2. Data and Methods

2.1. Data Source

- Source: Copernicus Marine Service (https://www.copernicus.eu/en)
- Period: June 2022 – November 2024
- Geographical Coverage:
    - Latitude: 20°N to 23°N
    - Longitude: 87°E to 91°E
    - Depth: ~0.494m (surface level)

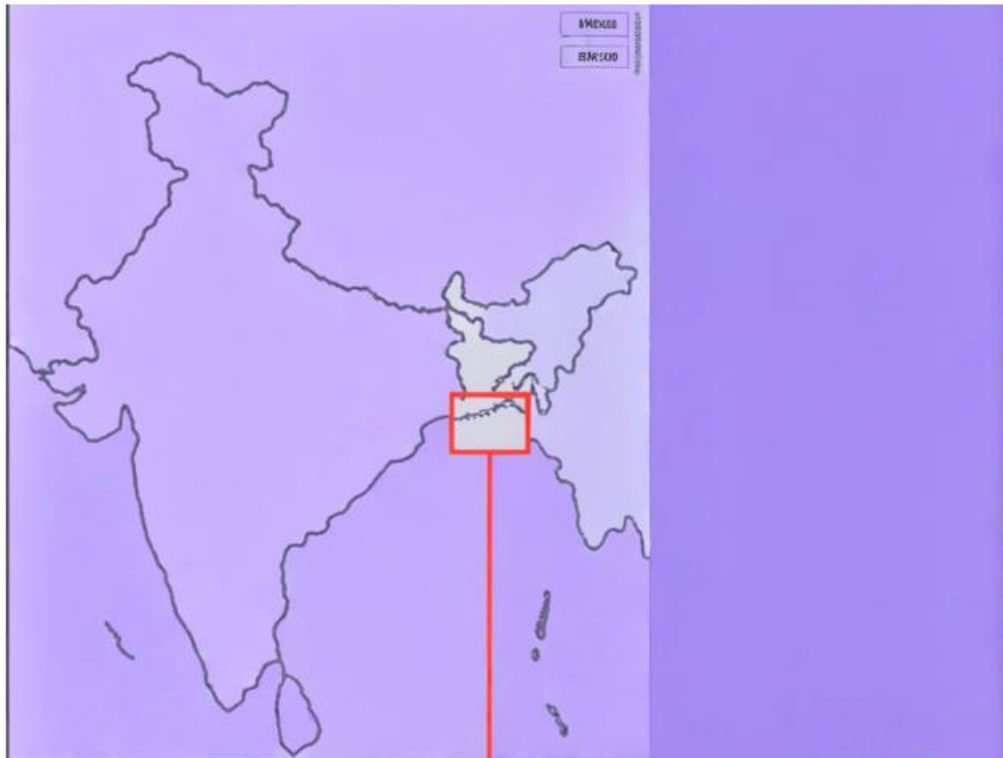The location of study with the position of data collection buoys is given in Fig 1a, 1b.

193

194 Fig 1a: Map of India showing the area of study
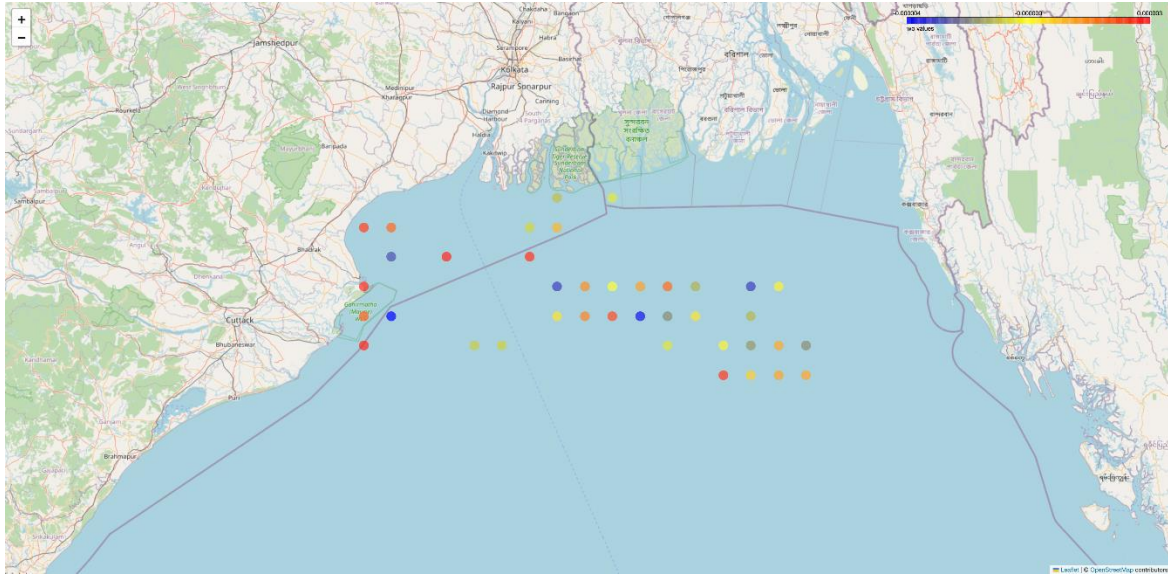
195

196

197

198

199

200

201

202

203

204

Fig 1b: Map of east coast of India in the bay of bengal.

2.2. Parameters Collected

- **Blue Ocean:** Sea Surface Temperature (SST), Sea Surface Salinity (SSS), Surface Water Velocities (U, V, W)
- **Green Ocean:** pH, Partial pressure of CO2 (pCO2), Dissolved Oxygen (O2), Nitrate (NO3), Phosphate (PO4), Silicate (Si), Iron (Fe), Chlorophyll-a (Chl)

2.3 Data Processing Pipeline

A) Initial Data Processing:

The NetCDF files were converted to CSV format, and the parameters were merged based on time, latitude, and longitude. The initial dataset comprised 3,870 samples.

Data Cleaning: In machine learning, the Interquartile Range (IQR) method is a widely used technique for identifying and removing outliers from a dataset. Outliers can skew the results of data analysis and model training, leading to inaccurate predictions. The IQR method focuses on the middle 50% of the data, providing a robust way to detect anomalies.

To apply the IQR method, we followed these steps:

Calculate the Quartiles: First, determine the first quartile (Q1) and the third quartile (Q3) of the dataset. Q1 is the median of the lower half of the data, while Q3 is the median of the upper half.

Compute the IQR**: The IQR is calculated as the difference between Q3 and Q1:

231   {IQR} = Q3 - Q1
232 Determine the Outlier Boundaries: Establish the lower and upper bounds for identifying
233 outliers:
234   {Lower Bound} = Q1 - 1.5 x {IQR}
235   {Upper Bound} = Q3 + 1.5 x {IQR}
236
237
238 Identify Outliers: Any data point that falls below the lower bound or above the upper bound
239 is considered an outlier. Remove Outliers: Filter the dataset to exclude these outlier values,
240 resulting in a cleaner dataset that is more representative of the underlying trends. By using the
241 IQR method, data scientists can enhance the quality of their datasets, leading to more reliable
242 machine learning models and improved predictive performance.

243 The final dataset comprises 2,153 samples, with the features having been normalized using
244 StandardScaler (the various variables given in Table 1) (https://marine.copernicus.eu/)

245

246 Table 1: Oceanographic Variables

247

S.No.	Variable name	Specific variable name	Acronyms	Ocean	Units
1	Temperature	Sea water potential temperature	thetao	Blue Ocean	°C
2	Salinity	Sea water salinity	so	Blue Ocean	$/10^3$
3	Velocity	Eastward sea water velocity	uo	Blue Ocean	m/s
4	Velocity	Northward sea water velocity	vo	Blue Ocean	m/s
5	Velocity	Upward sea water velocity	wo	Blue Ocean	m/s
6	Light	Volume attenuation coefficient of downwelling radiative flux	kd	Green Ocean	m^{-1}
7	pH	Sea water ph	ph	Green	-

		reported on total scale		Ocean	
8	SpCO2	Surface partial pressure of carbon dioxide in sea water	spco2	Green Ocean	Pa
9	O2	Mole concentration of dissolved molecular oxygen in sea water	o2	Green Ocean	mmol/m³
10	NO3	Mole concentration of nitrate in sea water	no3	Green Ocean	mmol/m³
11	PO4	Mole concentration of phosphate in sea water	po4	Green Ocean	mmol/m³
12	Si	Mole concentration of silicate in sea water	si	Green Ocean	mmol/m³
13	Fe	Mole concentration of dissolved iron in sea water	fe	Green Ocean	mmol/m³
14	Chl-a	concentration of chlorophyll a in sea water	chl	Green Ocean	mg/m³

248
249
250
251
252
253
254

255 **3.Results**

In machine learning, particularly in linear models, feature coefficients play a crucial role in determining the model's predictions and understanding the importance of different input features.

3.1. What are Feature Coefficients?

- **Numerical values** assigned to each feature (input variable) in a linear model (James G et.al, 2013).

- Represent the **weight or importance** of each feature in influencing the model's output.

- Determine the **slope or direction** of the relationship between a feature and the target variable.

3.2. How Feature Coefficients Work

- **Linear Regression:**

  - In simple linear regression (one feature), the model is: $y = b0 + b1 * x$

    - b0 is the intercept (value of y when x is 0)

    - b1 is the coefficient of the feature x

  - In multiple linear regression (multiple features), the model is: $y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn$

    - b0 is the intercept

    - b1, b2, ..., bn are the coefficients of features x1, x2, ..., xn respectively.

- **Other Models:**

  - While less explicit in some models (like decision trees), the concept of feature importance still applies.

3.3. Interpreting Feature Coefficients:

- **Magnitude:** The absolute value of a coefficient indicates the strength of its influence on the target variable. A larger coefficient means a stronger influence (Table 2).

- **Sign:**

  - **Positive:** As the feature value increases, the target variable tends to increase.

 - **Negative:** As the feature value increases, the target variable tends to decrease.

284 o

285 Table 2: Feature Coefficients for chl-a with the input parameters into ML model

Feature	Coefficient	Intercept	R²	MSE	Sample Size
Thetao	0.0216	-0.2397	0.0637	0.0533	837
So	-0.0157	0.8038	0.0588	0.0536	837
Uo	0.2795	0.361	0.0323	0.0551	837
Vo	-0.1792	0.3591	0.0049	0.0567	837
Wo	-49.5334	0.362	0.0	0.057	837
Kd	10.6336	-0.2561	0.717	0.0161	837
Ph	-0.277	2.6042	0.0039	0.0567	837
spco2	-0.0048	0.5176	0.0117	0.0563	837
o2	-0.0029	0.9749	0.0186	0.0559	837
no3	0.0194	0.3163	0.0546	0.0539	837
po4	2.0907	0.3452	0.0118	0.0563	837
Si	0.0152	0.3259	0.0036	0.0568	837
Fe	64.8834	0.2081	0.208	0.0451	837

286

287

10
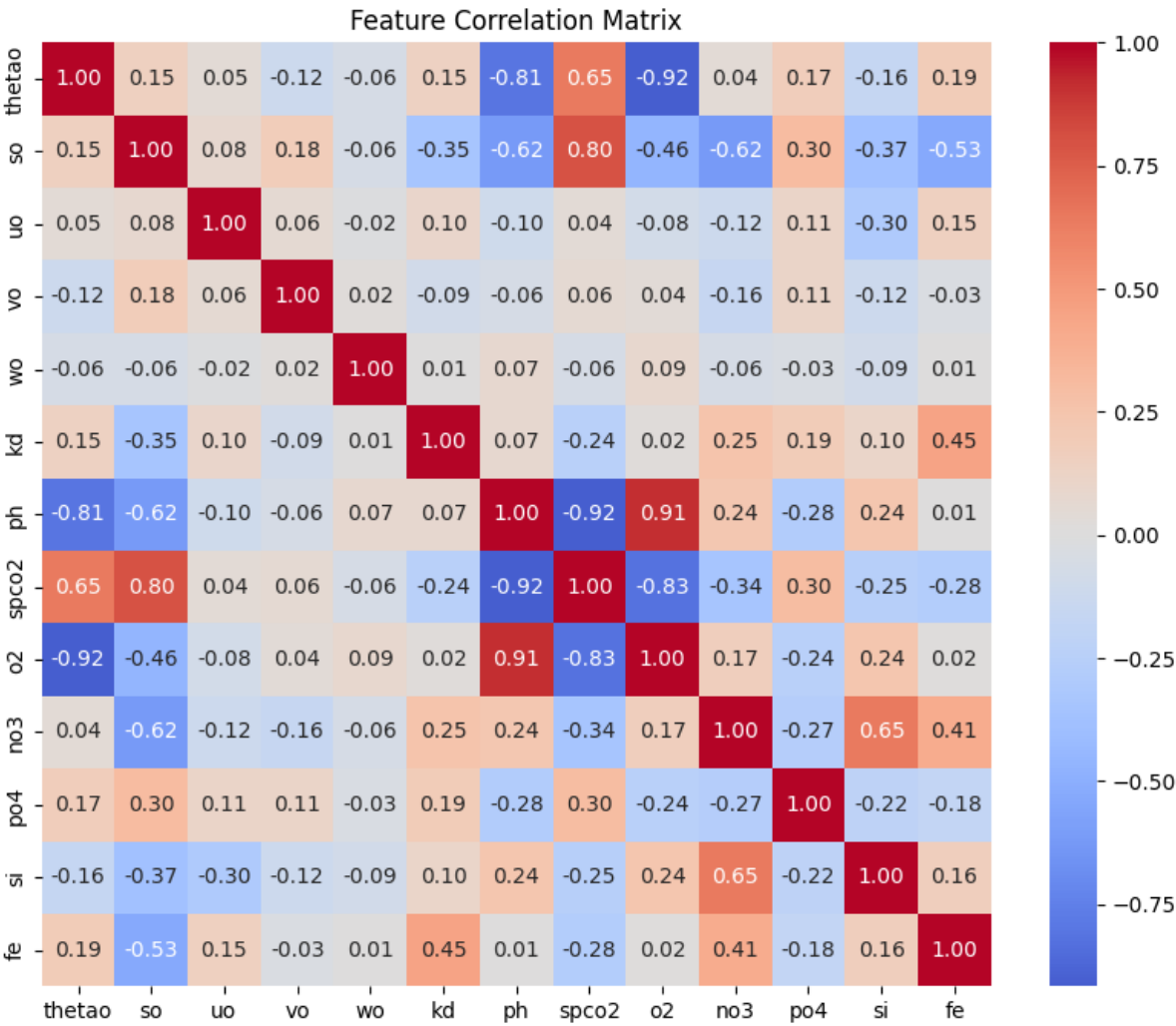
Fig 2: Feature correlation map

In statistics, the t-value is a crucial concept used in hypothesis testing. It's a measure of how different the observed data is from what we would expect if there were no real difference between the groups being compared.

Here's a breakdown:

1. What it represents:

- **The t-value is the ratio of the difference between the sample mean and the hypothesized population mean to the standard error of the sample mean.**
- Essentially, it tells us how many standard errors the sample mean is away from the hypothesized population mean.

2. How it's used:

- **T-tests:** T-values are primarily used in t-tests, which are statistical tests used to compare means. There are different types of t-tests:
  - **One-sample t-test:** Compares the mean of a single sample to a known or hypothesized population mean.
  - **Independent samples t-test:** Compares the means of two independent groups (e.g., treatment group vs. control group).
  - **Paired samples t-test:** Compares the means of two related groups (e.g., before and after measurements on the same individuals).

3. Interpretation:

- **Larger t-values (in absolute value):** Indicate that the observed difference between the sample mean and the hypothesized population mean is more likely to be statistically significant. This means it's less likely to be due to chance.
- **Smaller t-values (closer to zero):** Suggest that the observed difference is more likely due to chance.

   1. **Linear regression model**

Table 3:

Variable	Coef	std err	T	P>|t|	[0.025, 0.975]
Const	-2.7562	3.4180	-0.8060	0.4200	[-9.4680, 3.9550]
Thetao	-0.0051	0.0100	-0.5360	0.5920	[-0.0240, 0.0140]
So	0.0146	0.0060	2.4680	0.0140	[0.0030, 0.0260]
Uo	0.1220	0.0330	3.7190	0.0000	[0.0580, 0.1860]
Vo	0.0431	0.0510	0.8510	0.3950	[-0.0560, 0.1420]

Wo	1468.1934	3203.9710	0.4580	0.6470	[-4823.0990, 7759.4860]
Kd	10.5459	0.3250	32.4490	0.0000	[9.9080, 11.1840]
Ph	0.5149	0.3920	1.3130	0.1900	[-0.2550, 1.2850]
spco2	-0.0077	0.0040	-1.8410	0.0660	[-0.0160, 0.0010]
o2	-0.0081	0.0020	-4.1200	0.0000	[-0.0120, -0.0040]
no3	0.0053	0.0030	2.1100	0.0350	[0.0000, 0.0100]
po4	-1.8393	0.4320	-4.2610	0.0000	[-2.6870, -0.9920]
Si	0.0046	0.0080	0.5950	0.5520	[-0.0110, 0.0200]
Fe	18.4705	4.7150	3.9180	0.0000	[9.2130, 27.7280]

To break down this regression table (Table 3):

1. **Statistically Significant Variables** ($P > |t| < 0.05$):
   Kd (light): Strongest effect (coef = 10.5459, P = 0.000)
   Fe (Iron) : Strong positive effect (coef = 18.4705, P = 0.000)
   PO_4 (phosphate): Strong negative effect (coef = -1.8393, P = 0.000)
   O_2 (oxygen): Small negative effect (coef = -0.0081, P = 0.000)
   u_o (u-vel): Positive effect (coef = 0.1220, P = 0.000)
   S_o (salinity): Small positive effect (coef = 0.0146, P = 0.014)
   NO_3 (nitrate): Small positive effect (coef = 0.0053, P = 0.035)

   1. Target Variable:
- **Chlorophyll-a:** This is the variable we trying to predict with the linear regression model. It represents the concentration of chlorophyll-a in a particular environment (north East Bay of Bengal). Chlorophyll-a is a key indicator of phytoplankton biomass, which plays a crucial role in aquatic ecosystems.
   2. Input Variables (Predictors):
- **Statistically Significant Variables ($P > |t| < 0.05$):** These are the input variables that have a statistically significant relationship with chlorophyll-a based on the t-test.
   Kd (light): Strongest effect (coef = 10.5459, P = 0.000)
   - **Interpretation:** Kd (light attenuation coefficient) has the strongest influence on chlorophyll-a levels. The positive coefficient suggests that as light penetration decreases (higher Kd), chlorophyll-a concentrations tend to

12

increase. This could be due to factors like increased phytoplankton growth in deeper, less illuminated waters.

Fe (Iron): Strong positive effect (coef = 18.4705, P = 0.000)

- **Interpretation:** Iron is a crucial nutrient for phytoplankton growth. The strong positive coefficient indicates that higher iron concentrations are significantly associated with increased chlorophyll-a levels, as expected.

PO4 (phosphate): Strong negative effect (coef = -1.8393, P = 0.000)

- **Interpretation:** While often a limiting nutrient, the negative coefficient suggests an unexpected relationship. It might indicate that high phosphate concentrations could be associated with factors that inhibit phytoplankton growth, such as nutrient imbalances or competition from other organisms. The significant influence of riverine inputs (implied by the importance of nutrients like iron, phosphate, and nitrate) aligns with the monsoon season's role in delivering substantial nutrient loads to the Bay of Bengal.

O2 (oxygen): Small negative effect (coef = -0.0081, P = 0.000)

- **Interpretation:** A small negative effect of oxygen on chlorophyll-a is observed. This could be due to various factors, such as oxygen depletion in deeper waters or potential inhibitory effects at very high oxygen concentrations.

uo (u-vel): Positive effect (coef = 0.1220, P = 0.000)

- **Interpretation:** A positive effect of u-vel (likely a measure of water velocity or current speed) on chlorophyll-a suggests that increased water movement might enhance nutrient availability or promote better mixing, leading to increased phytoplankton growth. The positive effect of water velocity (uo) suggests that upwelling processes, which often involve increased water movement, could be contributing to higher chlorophyll-a levels by bringing nutrient-rich waters to the surface. While not explicitly captured in the variables used, the results indirectly support the potential impact of cyclones. Cyclonic disturbances can induce upwelling and enhance nutrient mixing, both of which can favor phytoplankton growth, aligning with the positive effects observed for some variables like water velocity.

So (salinity): Small positive effect (coef = 0.0146, P = 0.014)

- **Interpretation:** A small positive effect of salinity on chlorophyll-a is observed. This could be due to the influence of salinity on water density and nutrient distribution.

NO3 (nitrate): Small positive effect (coef = 0.0053, P = 0.035)

- **Interpretation:** Nitrate, another essential nutrient for phytoplankton, shows a small positive effect on chlorophyll-a levels, as expected.

- All listed variables have P-values less than 0.05, indicating that their relationship with chlorophyll-a is statistically significant at the 95% confidence level. This means that the observed relationships are unlikely to be due to random chance.
- The linear regression model identifies several key factors that influence chlorophyll-a concentrations in the studied environment.
- Light availability (Kd), iron, and phosphate appear to have the strongest impacts.

13

- Understanding these relationships is crucial for ecological studies, water quality management, and predicting the impacts of environmental changes on aquatic ecosystems.

2. **Marginally Significant**:
   spco2: Slight negative effect (coef = -0.0077, P = 0.066)

3. **Non-Significant Variables** (P>|t| > 0.05):
   const (intercept): P = 0.420
   thetao: P = 0.592
   vo: P = 0.395
   wo: P = 0.647
   ph: P = 0.190
   si: P = 0.552

The model shows that **kd (light)** and **fe (iron)** have the largest positive effects, while **po4** has a strong negative effect. The water chemistry variables (**o2**, **no3**) have smaller but still significant effects. Many physical parameters (**thetao**, **vo**, **wo**) are not statistically significant in this model.

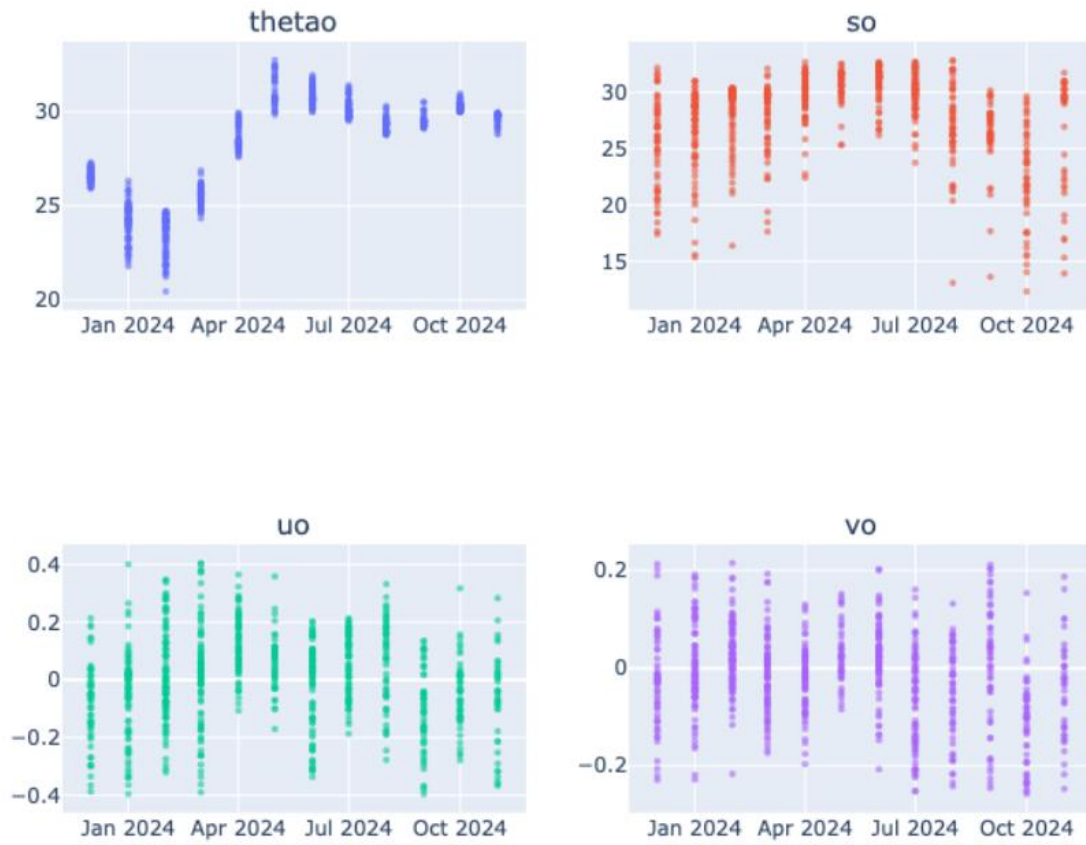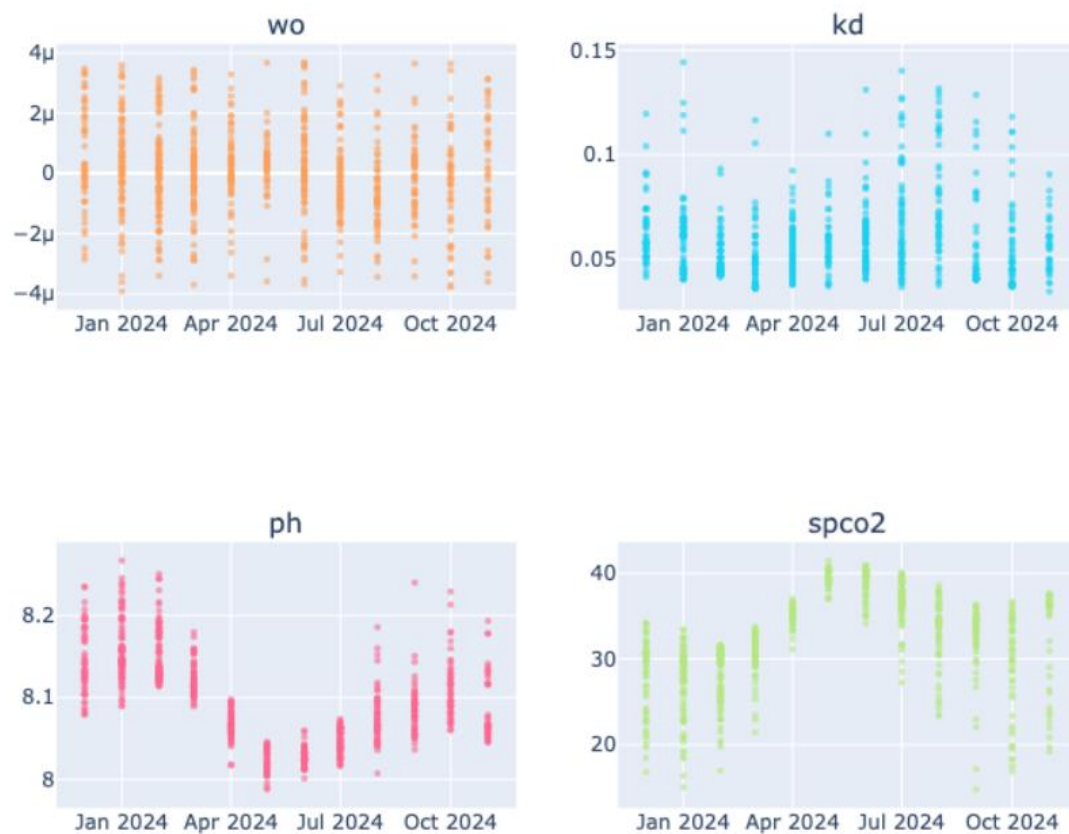Time series of the various variables for the time period : June 2022 – November 2024

412

413 Fig 3a-d: Time series of the oceanic state variables

414

415

416

417

418

419   Fig 4a-d: time series of predicted Oceanic parameters

420

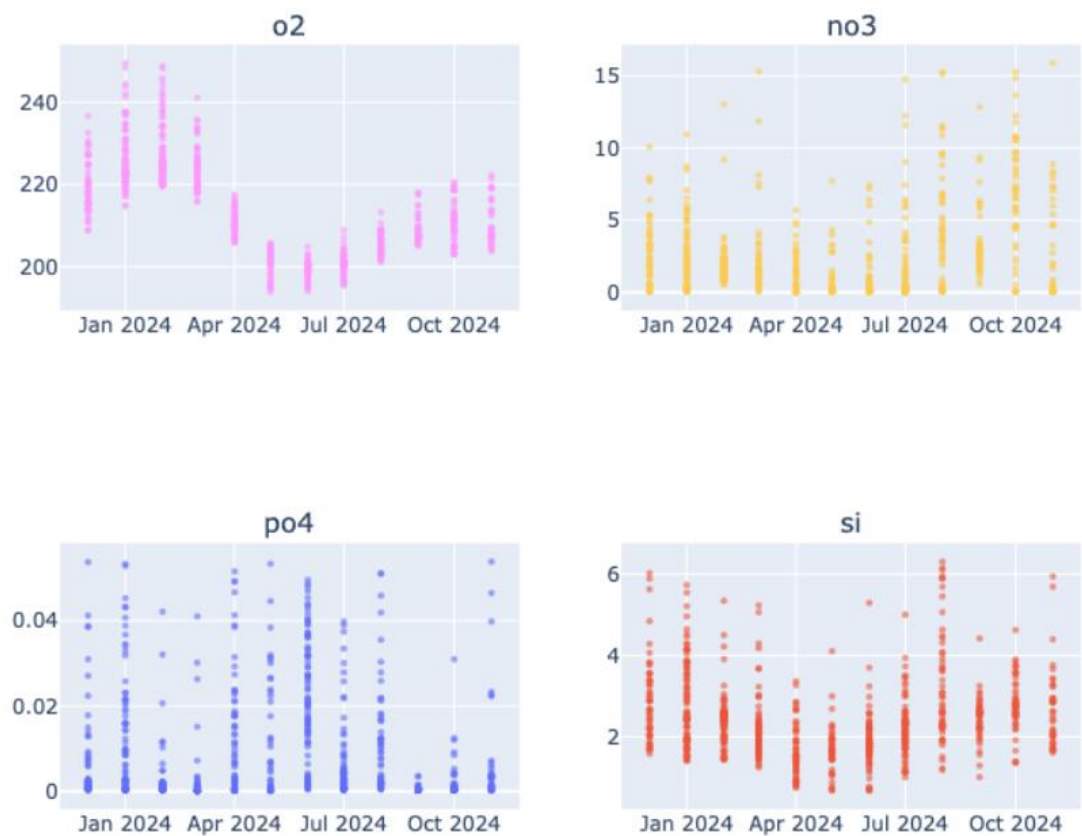421   Fig 5a-b: time series of predicted Oceanic parameters

422
423
424
425
426

427

428

Fig 6a-d: time series of predicted Oceanic parameters

429
430
431

4. Discussions

432
433

4.1. Principal Component Analysis

434
435

Principal Component Analysis (PCA) is a statistical technique widely used for dimensionality reduction and data visualization (Jolliffe, I. T. , 2002). It transforms a large set of variables into a smaller one while retaining most of the original data's variability. The primary goal of PCA is to identify the underlying structure in the data by finding the directions (principal components) that maximize the variance.

The process begins with standardizing the data, ensuring that each feature contributes equally to the analysis. This is crucial, especially when the variables are measured on different scales. Once standardized, PCA computes the covariance matrix to understand how the variables relate to one another. The next step involves calculating the eigenvalues and eigenvectors of this covariance matrix. The eigenvectors represent the directions of the new feature space, while the eigenvalues indicate the amount of variance captured by each principal component. By selecting the top principal components—those with the highest eigenvalues—PCA allows for the reduction of dimensionality. This means that instead of working with a large number

449 of variables, analysts can focus on a few key components that summarize the data effectively.
450 This simplification not only enhances computational efficiency but also helps in visualizing
451 complex datasets, making it easier to identify patterns and relationships.
452 PCA is widely applied in various fields, including finance for risk management, biology for
453 gene expression analysis, and image processing for feature extraction. However, it is essential
454 to note that PCA assumes linear relationships among variables and may not perform well
455 with non-linear data. Despite this limitation, PCA remains a powerful tool for exploratory
456 data analysis and preprocessing in machine learning workflows. Principal Component
457 Analysis (PCA) has significant applications in oceanography, particularly in the study of the
458 Bay of Bengal. This region, characterized by its complex hydrodynamics and diverse marine
459 ecosystems, benefits from PCA's ability to reduce dimensionality and extract meaningful
460 patterns from large datasets.
461 In oceanographic research, PCA can be employed to analyze various environmental
462 parameters such as temperature, salinity, chlorophyll concentration, and nutrient levels. By
463 applying PCA, researchers can identify the principal components that explain the most
464 variance in these datasets, allowing for a clearer understanding of the underlying processes
465 affecting the Bay of Bengal's marine environment.
466 In this work PCA is used as a help in assessing the impact of monsoon patterns on
467 oceanographic conditions by analyzing time-series data of sea surface temperatures, light,
468 currents and nutrients. This analysis can reveal how seasonal changes influence marine
469 biodiversity and productivity in the region. Additionally, PCA can be used to study the
470 effects of anthropogenic activities, such as pollution and overfishing, by identifying key
471 factors that contribute to changes in marine ecosystems.
472 Here are the PCA first three modes of variability;
473
474 PC1 = (-0.2793)·chl + (-0.3339)·kd + (-0.3229)·Fe + (-0.1853)·u0 + (-0.3009)·v0 + (-
475 0.2791)·pH + (-0.3319)·phosphate + (-0.3377)·silicate + (-0.3016)·oxygen + (-0.2791)·SST +
476 (-0.3321)·SSS
477 PC2 = (-0.3889)·chl + (-0.0507)·kd + (0.1319)·Fe + (0.5657)·u0 + (0.3005)·v0 + (-
478 0.3880)·pH + (0.1307)·phosphate + (-0.0496)·silicate + (0.3073)·oxygen + (-0.3900)·SST +
479 (-0.0434)·SSS
480 PC3 = (0.1078)·chl + (-0.1644)·kd + (0.8770)·Fe + (0.0110)·u0 + (-0.2557)·v0 + (-
481 0.1241)·pH + (-0.0822)·phosphate + (-0.0200)·silicate + (-0.1197)·oxygen + (0.0662)·SST +
482 (-0.2924)·SSS
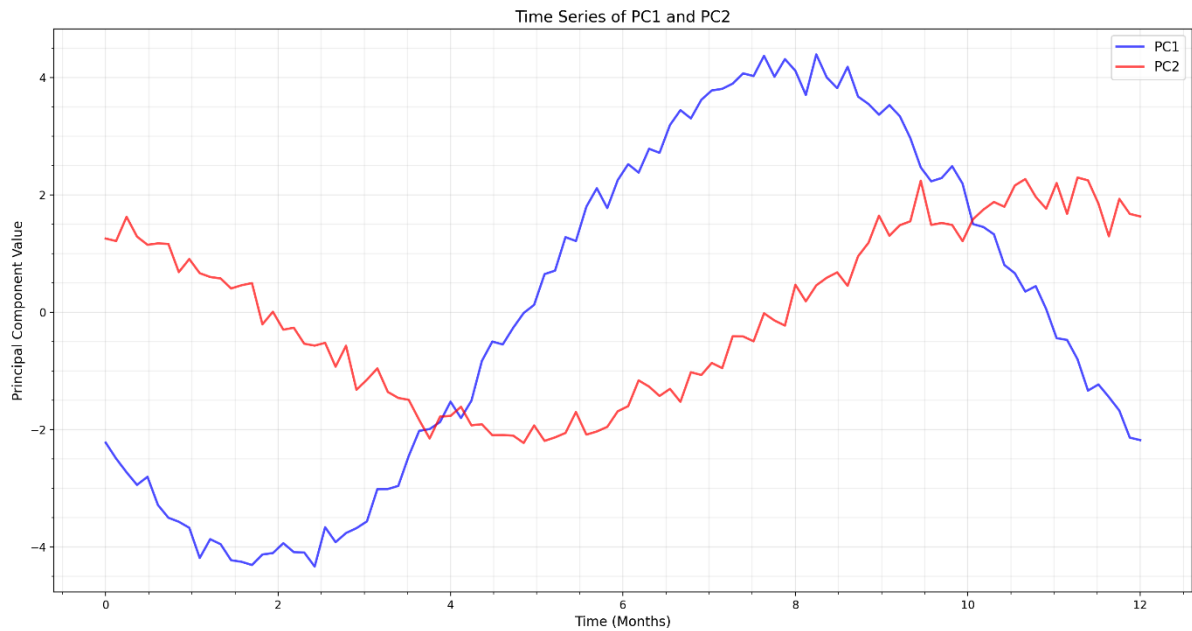483   The figure given below shows the annual variation of PC1 and PC2 (Fig 7).
484

Fig 7: PCA components vs. time

According to the peak analysis:

- Average period for PC1 is approximately 5.0 months
- Average period for PC2 is approximately 3.33 months

The Bay of Bengal is influenced by several environmental factors:

- Monsoonal Winds: The southwest and northeast monsoons drive seasonal changes in ocean currents, nutrient upwelling, and salinity.

River Discharge: Major rivers like the Ganges and Brahmaputra bring significant freshwater and nutrient loads, affecting silicate and phosphate
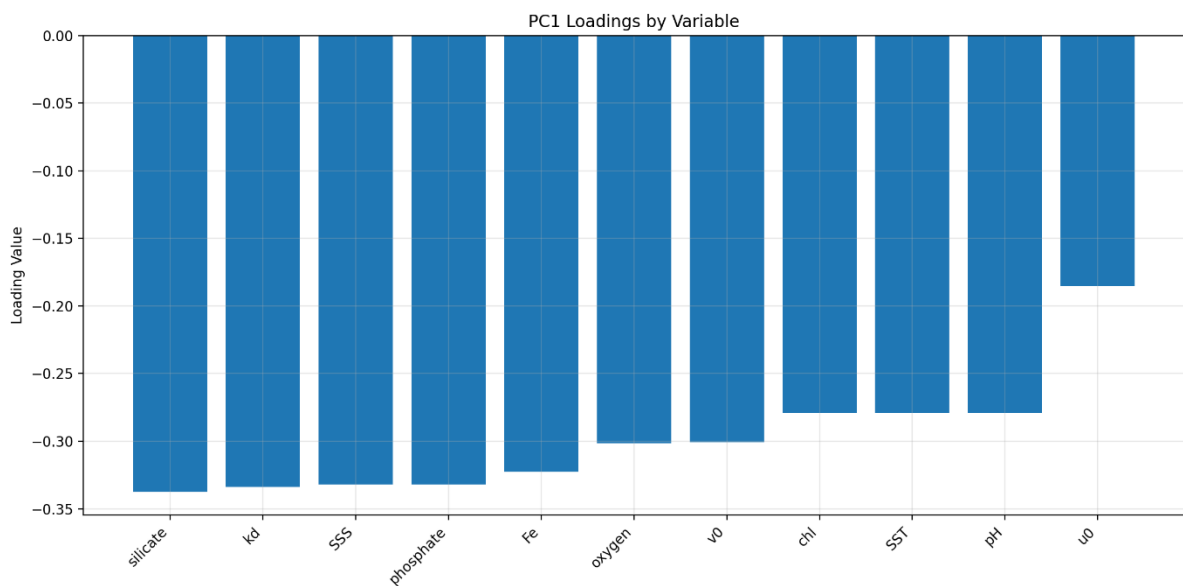

Fig 8: PCA loadings indicating the variabilities of PC1 and PC2.

501   concentrations.

- Cyclones: Frequent cyclones stir the water column, impacting nutrient distribution and light attenuation (kd).
- Temperature and Salinity: Seasonal variations in sea surface temperature (SST) and salinity (SSS) influence stratification and mixing.

Based on the Bay of Bengal's characteristics, PC1's variability is primarily driven by (Fig 8):

1. Monsoon-driven freshwater influx affecting SSS and nutrient loads
2. River discharge bringing high silicate and phosphate concentrations
3. Light attenuation (kd) variations due to suspended sediments and algal blooms
4. Seasonal stratification affecting nutrient distribution

4.2. Interseasonal Variability of Parameters in the Bay of Bengal:

The Bay of Bengal exhibits distinct interseasonal variability in various ecological parameters, which can be analyzed to understand the dynamics of the marine ecosystem. The following summarizes the trends observed in key parameters and their interrelationships.

Chlorophyll-a (chl-a):

Chlorophyll-a concentrations peak in July at 1.4 mg/m³, indicating a period of heightened biological productivity during the summer months. This increase is likely due to favorable conditions for phytoplankton growth, driven by optimal light and nutrient availability. Conversely, chl-a levels drop to their lowest in December at 0.2 mg/m³, reflecting reduced biological activity during the winter months when conditions are less favorable for growth.

Light:

Light availability shows a peak in October (0.14 m^{-1}) and July (0.12 m^{-1}), with a notable dip in April and May (0.07 m^{-1}). This seasonal variation in light is influenced by factors such as cloud cover and day length, which can significantly affect photosynthetic activity. The correlation between light and chlorophyll-a suggests that higher light levels during summer months contribute to increased phytoplankton growth.

Zonal Velocity (u):

Zonal velocity reaches its highest point in April (0.4 m/s) and is at its lowest in October (0.1m/s). This variation reflects seasonal changes in water movement, which can be attributed to wind patterns and ocean currents. The increased zonal velocity in spring may facilitate nutrient mixing, enhancing biological productivity, particularly in conjunction with higher light availability.

Iron (Fe):

Iron concentrations remain relatively stable across seasons, fluctuating only between 0.006 mmol/m³ and 0.007 mmol/m³. This stability suggests that iron availability is not significantly impacted by seasonal changes, which is crucial for sustaining phytoplankton growth throughout the year.
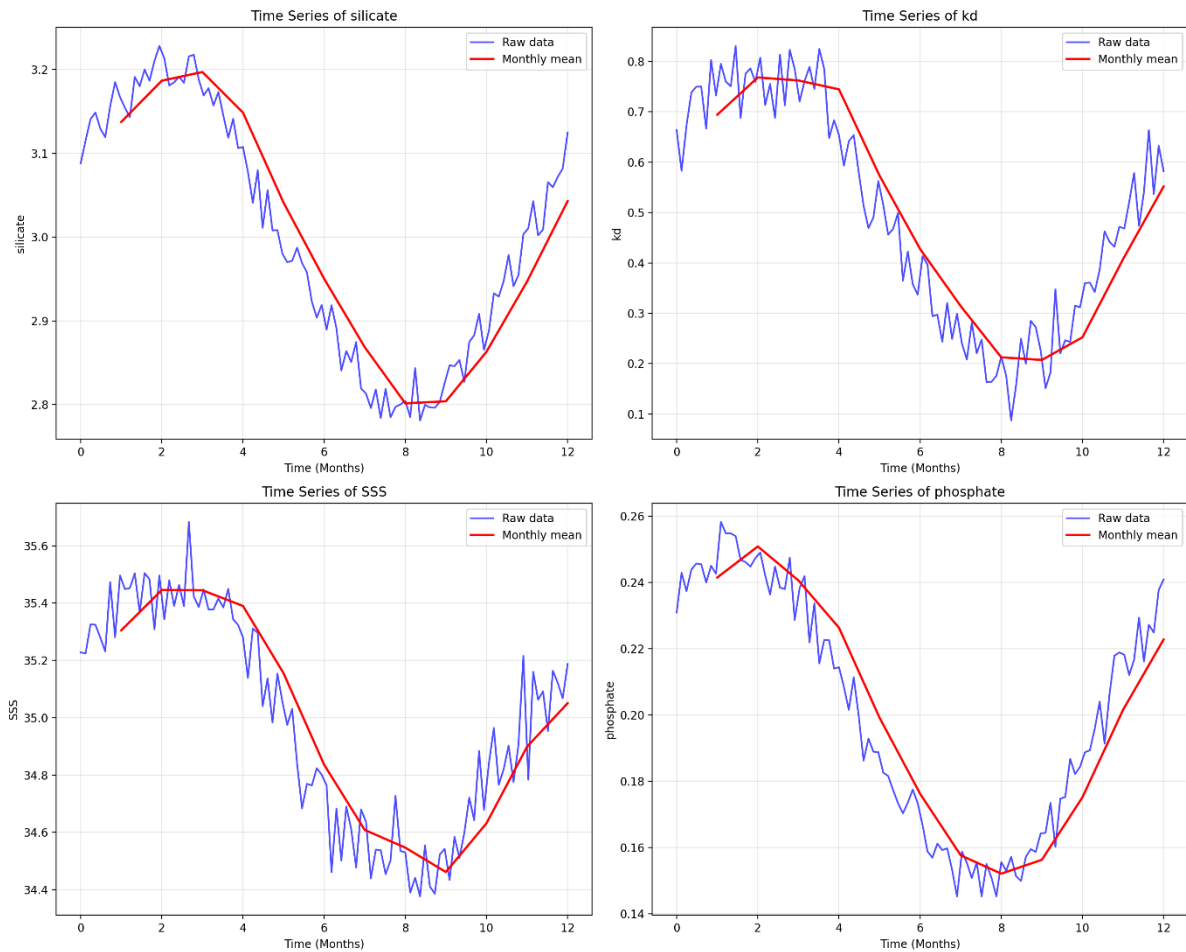
544

Fig 9: Yearly variabilities of various Oceanographic parameters

546

Nitrate:

Nitrate levels peak in January and December (10 mmol/m³) and dip in April and May (4 mmol/m³). This pattern indicates that nutrient availability is higher during the winter months, likely due to reduced biological uptake as phytoplankton activity declines. The relationship between nitrate and chlorophyll-a suggests that higher nutrient levels in winter can support productivity when conditions are less favorable.

pH:

The pH levels show minor fluctuations, with the highest value recorded in January (8.3) and the lowest in May (8.0). These slight seasonal changes in water chemistry may be influenced by biological processes, such as respiration and photosynthesis, which can affect carbon dioxide levels in the water.

Silicate:

Seasonal cycle in SSS and silicate typically reflects the changes in freshwater input and river discharge linked to monsoon cycles, while kd and phosphate can be influenced by both sediment loading and nutrient dynamics.

21

Oxygen:

Oxygen levels peak in April (250 mmol/m³) and are lowest in May and July (200 mmol/m³). The increase in oxygen during spring is likely due to enhanced photosynthesis from phytoplankton, which coincides with higher light availability and nutrient mixing. The subsequent decline in oxygen levels during summer may be attributed to increased respiration rates and organic matter decomposition.

In summary, the seasonal dynamics in the Bay of Bengal are mainly driven by the monsoon. The summer monsoon brings in a large volume of freshwater, leading to lower SSS and, in turn, lower phosphate levels due to dilution and stratification. Simultaneously, river discharge during the monsoon enhances silicate concentrations and influences kd through suspended sediments. These processes collectively explain why the environmental variability captured in PC1 (and also partly by PC2) shows strong seasonal signatures.

4.3. Artificial Neural Networks

Neural networks, inspired by the human brain, are powerful machine learning models. They consist of interconnected neurons organized in layers, and hyperparameter configuration plays a crucial role in their learning process and performance. Hyperparameters, such as the number of layers and neurons, learning rate, batch size, activation functions, regularization techniques, and optimization algorithms, are set by the user and significantly impact the network's ability to learn and generalize. For this study, the Hyperparameteres are decided by the value of R^2 and MSE using a trial and error procedure (Raschka S, Y. et al.,2022).

Activation functions introduce non-linearity into the network and enable it to model complex relationships in the data. The hyperbolic tangent function (tanh) is a commonly used activation function and is adopted here. It maps the input to a range between -1 and 1, exhibiting an S-shaped curve. This is centered around 0, which means the outputs have zero mean. This property can help with training neural networks and improve convergence. However, it's worth noting that the tanh activation function has some limitations, such as the potential for the vanishing gradient problem when used in deep neural networks.

Optimizers are algorithms used to adjust the model's parameters during training. They minimize the difference between predicted and target outputs by updating parameters based on gradients of the loss function. The ADAM optimizer, which we have utilized, combines momentum-based and adaptive learning rate methods, maintaining adaptive learning rates for each parameter in the network. Adam optimizer is a popular choice for optimizing neural network models. It combines adaptive learning rates and momentum to efficiently update the model parameters during training, leading to faster convergence and better performance.

In summary, neural networks rely on hyperparameter configuration to achieve optimal performance. The choice of the number of layers, neurons, activation functions, optimizers, batch size, and epochs significantly affects the network's ability to learn complex relationships in the data. By selecting the above-mentioned hyperparameters, our neural network is expected to learn and model complex functions from data (Fig 9).
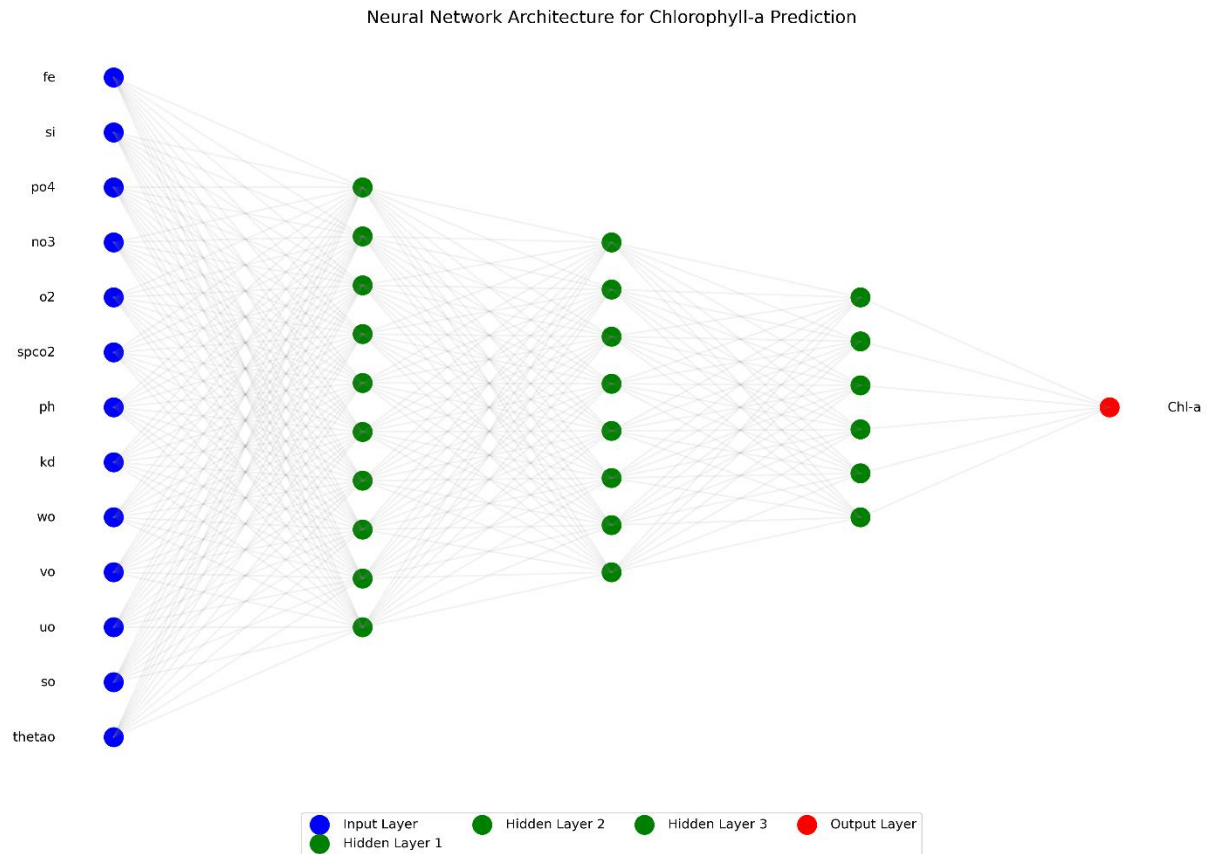
Fig 9: Neural network for the study

Training MSE: 0.0000
Testing MSE: 0.0119
Training R^2: 1.0000
Testing R^2: 0.7794

4.4. Analysis of Neural Network Results

The provided metrics indicate that the neural network has achieved **perfect performance on the training data**. To break down the results:

1. Training Performance:
- **MSE (Mean Squared Error): 0.0000**
  - o This extremely low value suggests that the model has essentially perfectly learned the training data.
- **R^2 (Coefficient of Determination): 1.0000**
  - o An R-squared of 1.0 indicates that the model perfectly explains all the variance in the training data.

2. Testing Performance:
- **MSE: 0.0119**
  - o While significantly higher than the training MSE, this value suggests reasonable accuracy on unseen data.
- **R^2: 0.7794**

23

630   o   This indicates that the model explains 77.94% of the variance in the testing
631   data, which is a decent performance.
632

633   **4.5. Random Forest:**

634

**Random Forest Model for Chlorophyll-a Prediction
with Detailed Decision Trees**
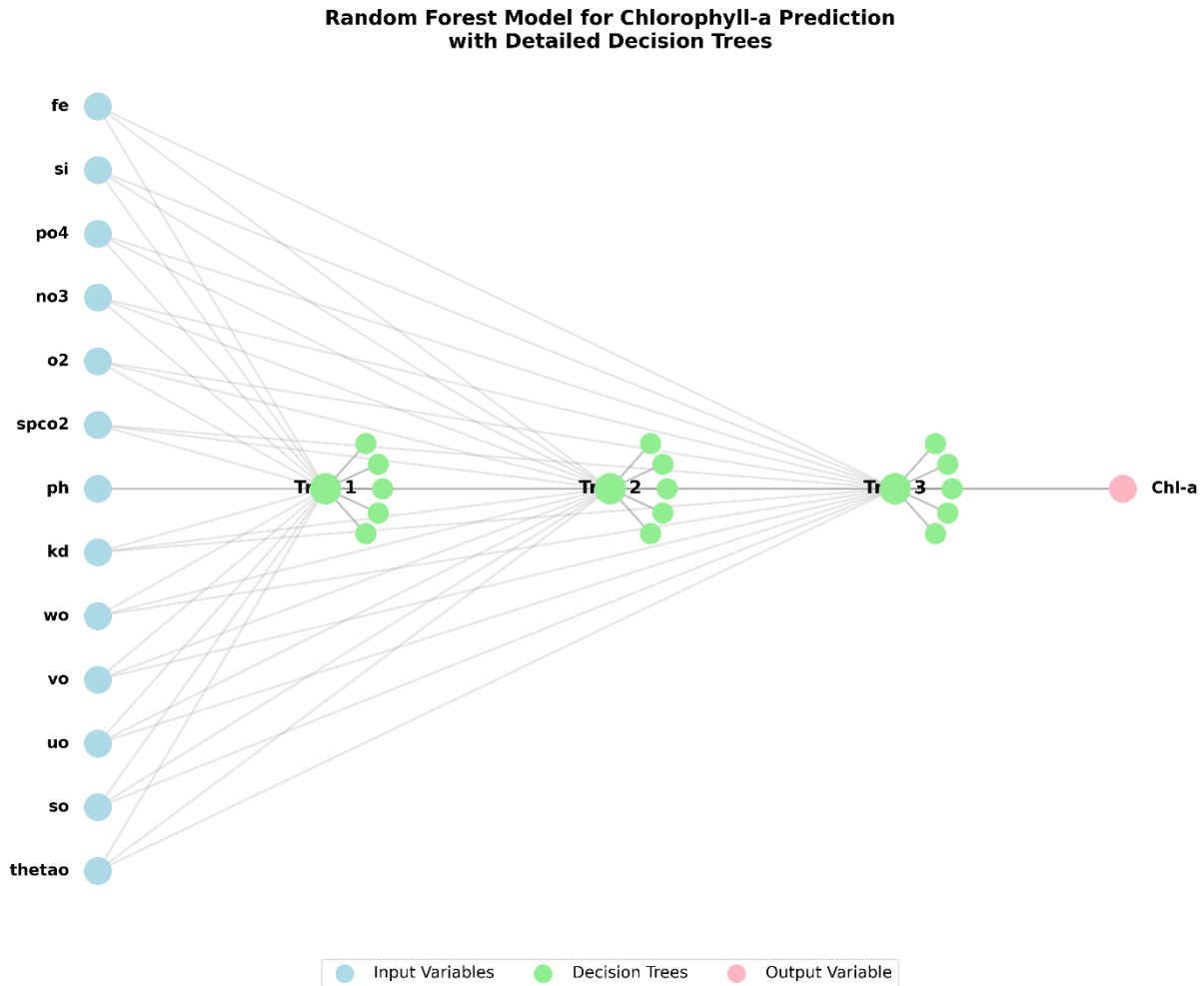
635
636   Fig 10: Random forest for the study

637

638   Training MSE: 0.0015
639   Testing MSE: 0.0098
640   Training R^2: 0.9734
641   Testing R^2: 0.8170

642

643   Fig 10 shows the architecture of the Random Forest Model.

644
645
646
647
648
649
650

4.6. SHAP Values for Model Predictions

Table 4:

Feature	Lowest Prediction SHAP Value	Highest Prediction SHAP Value
Thetao	-0.005664	0.014712
So	-0.000830	0.003333
Uo	-0.006064	0.054601
Vo	-0.002407	0.004767
Wo	-0.001455	0.000458
Kd	-0.160058	0.826507
Ph	-0.002156	-0.000026
spco2	-0.000388	0.006182
o2	-0.003871	0.013861
no3	-0.002629	0.001656
po4	0.004132	0.003153
Si	-0.006197	-0.001095
Fe	-0.024966	0.014132

1. Most Influential Features in Table 4 are:

- **kd (Light)**:
  - Has the most dramatic impact on predictions
  - Lowest: -0.160058 (strong negative influence)
  - Highest: 0.826507 (extremely strong positive influence)
  - This suggests that water clarity/turbidity is the most critical factor in predicting chlorophyll-a
- **uo (zonal velocity)**:
  - Second most influential feature
  - Lowest: -0.006064
  - Highest: 0.054601
  - Indicates that water movement in the east-west direction significantly affects chlorophyll-a concentrations
- **Fe (iron)**:
  - Third most influential feature
  - Lowest: -0.024966
  - Highest: 0.014132
  - Shows iron concentration has a notable impact on chlorophyll-a predictions
2. Moderately Influential Features:
- **thetao (temperature)** and **o2 (oxygen)**:
  - Similar ranges of influence
  - thetao: -0.005664 to 0.014712

679   • o2: -0.003871 to 0.013861

680   • Both show moderate impact on predictions

681   3. Least Influential Features:

682   • **wo (vertical velocity)**:

683   • Smallest range: -0.001455 to 0.000458

684   • Minimal impact on predictions

685   • **ph**:

686   • Notable for being negative in both cases

687   • Range: -0.002156 to -0.000026

688   • Suggests consistently negative (though small) influence

689   4. Other Patterns:

690   • **po4 (phosphate)**:

691   • Unique in having positive values for both scenarios

692   • Lowest: 0.004132

693   • Highest: 0.003153

694   • Suggests consistent positive influence on chlorophyll-a

695   • **Si (silicate)**:

696   • Like pH, maintains negative values

697   • Range: -0.006197 to -0.001095

698   • Indicates consistent negative influence

699   5. Overall Insights:

700   • Physical factors (kd, uo) seem to have larger impacts than chemical factors

701   • Most features show both positive and negative influences depending on the data

702   • The model appears to be most sensitive to water clarity (kd) by a significant margin

703   • Nutrient availability (fe, po4, no3) shows varying degrees of importance

704 This analysis suggests that while biological productivity (chlorophyll-a) is influenced by

705 many factors, physical characteristics of the water column, particularly clarity and movement,

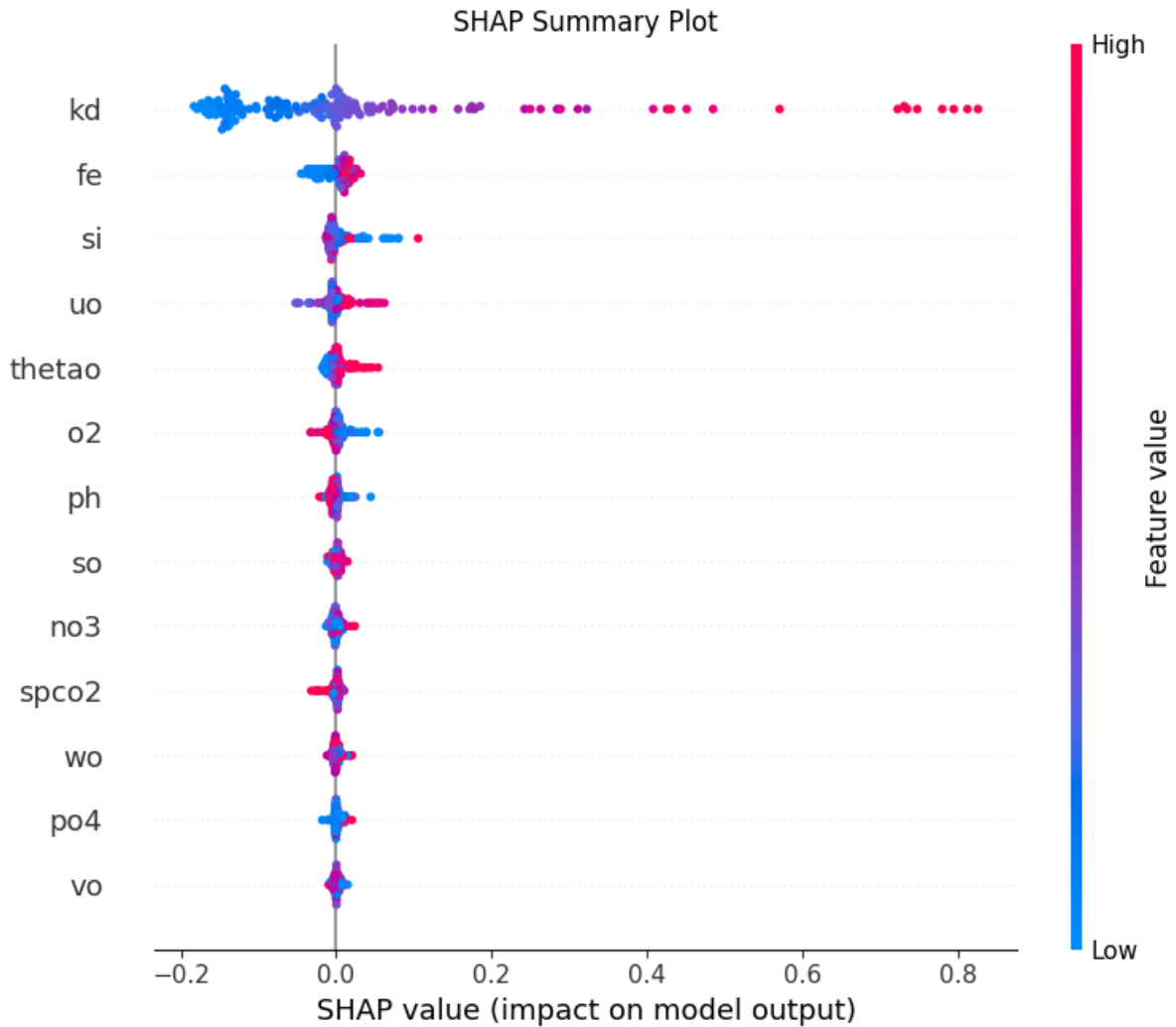706 are the dominant predictors in this model (Fig 11).

SHAP Summary Plot

Fig 11: SHAP curves for the various parameters

5. Conclusions

This study provides a comprehensive analysis of marine environmental parameters and their impact on chlorophyll-a concentrations in the Bay of Bengal. Utilizing data collected from June 2022 to November 2024, the research employs advanced statistical and computational techniques, including Principal Component Analysis, linear regression, artificial neural networks (ANN), and random forest models, to identify the key drivers of chlorophyll-a variability. The findings reveal that light availability, iron (Fe), and phosphate (PO4) are the most significant factors influencing chlorophyll-a concentrations. These results have critical implications for marine ecosystem monitoring, fisheries management, and climate change studies.

Light Availability

Light availability, represented by the volume attenuation coefficient of downwelling radiative flux (kd), emerged as the most influential factor affecting chlorophyll-a concentrations (Siegel, D. A., & Michaels, A. F., 2005). The positive relationship between kd and

chlorophyll-a suggests that as light penetration decreases (higher kd), chlorophyll-a concentrations tend to increase. The coefficient for kd is 10.5459, indicating a strong positive effect. This could be due to increased phytoplankton growth in deeper, less illuminated waters. The significant impact of light availability underscores the importance of considering water turbidity and seasonal changes in assessing primary productivity in the Bay of Bengal.

Iron (Fe)

Iron, a crucial micronutrient for phytoplankton growth, showed a strong positive effect on chlorophyll-a levels, with a coefficient of 18.4705. Higher iron concentrations are significantly associated with increased chlorophyll-a levels (Smith, R. C., & Baker, K. S. 1981), highlighting the role of iron in supporting phytoplankton biomass. This finding aligns with previous studies that emphasize the importance of iron in regions where its concentration is limited. The positive effect of iron on chlorophyll-a concentrations underscores the need for monitoring iron levels to understand and manage primary productivity in the Bay of Bengal.

Phosphate (PO4)

Phosphate, a key nutrient, exhibited a strong negative effect on chlorophyll-a concentrations, with a coefficient of -1.8393. While often a limiting nutrient, the negative coefficient suggests an unexpected relationship, possibly indicating that high phosphate concentrations could be associated with factors that inhibit phytoplankton growth, such as nutrient imbalances or competition from other organisms. This finding highlights the complexity of nutrient dynamics in aquatic ecosystems and the need for further research to understand the interactions between different nutrients and their impact on primary productivity.

Silicate (Si)

Silicate, another important nutrient for phytoplankton, showed a small positive effect on chlorophyll-a levels, with a coefficient of 0.0152. This indicates that higher silicate concentrations are associated with increased chlorophyll-a levels, supporting the growth of diatoms, a type of phytoplankton that requires silicate for their cell walls. Monitoring silicate levels is essential for understanding its role in primary productivity and the overall health of the marine ecosystem in the Bay of Bengal. Also, the variability of silicates follows the river discharge into the Bay which is predominantly the suspended sediments in the rivers.

Oxygen (O2)

Oxygen levels exhibited a small negative effect on chlorophyll-a concentrations, with a coefficient of -0.0081. This could be due to various factors, such as oxygen depletion in deeper waters or potential inhibitory effects at very high oxygen concentrations. The relationship between oxygen and chlorophyll-a highlights the importance of considering oxygen dynamics in assessing the health and productivity of marine ecosystems.

Principal Component Analysis (PCA) was used to identify the main modes of variability in the dataset. The first three principal components (PCs) were analyzed:

- PC1: This component explained a significant portion of the variance and was primarily driven by factors such as chlorophyll-a, kd, Fe, uo, vo, pH, phosphate, silicate, oxygen, SST, and SSS. The average period for PC1 was approximately 5.0 months.

- PC2: This component also explained a considerable amount of variance and was influenced by similar factors as PC1 but with different weights. The average period for PC2 was approximately 3.33 months.

- PC3: This component was mainly driven by Fe, with a strong positive influence, and other factors like kd, uo, vo, pH, phosphate, silicate, oxygen, SST, and SSS.

The PCA results highlight the seasonal and interseasonal variability of these parameters, driven by monsoonal patterns, river discharge, and cyclonic activities. These factors collectively influence the nutrient dynamics and primary productivity in the Bay of Bengal. Understanding the factors that influence chlorophyll-a concentrations is crucial for managing fisheries, as phytoplankton serves as the foundation of the marine food web. Changes in chlorophyll-a levels can indicate shifts in fish populations and overall ecosystem health. By identifying the key drivers of chlorophyll-a variability, this study provides valuable insights for fisheries management in the Bay of Bengal. Monitoring light availability, iron, phosphate, silicate, and oxygen levels can help predict changes in primary productivity and inform sustainable fisheries management practices.

The study highlights the importance of integrating advanced modeling techniques in marine research. The use of linear regression, ANN, and random forest models enables a comprehensive analysis of complex datasets, uncovering relationships that may not be apparent through traditional statistical methods. On the testing data, the ANN showed reasonable accuracy with an MSE of 0.0119 and an R^2 of 0.7794, explaining 77.94% of the variance in the testing data. These results suggest that the ANN was effective in capturing the complex relationships between the environmental parameters and chlorophyll-a concentrations. The high training performance indicates that the model learned the training data well, while the testing performance demonstrates its ability to generalize to unseen data.

These models can also be applied to predict future trends in chlorophyll-a concentrations, enhancing our ability to monitor and manage marine resources effectively. The application of ANNs in particular has shown promise in various aspects of oceanographic research, including the prediction of sea surface temperature, ocean currents, and water quality.

The findings of this study have significant implications for climate change studies. The Bay of Bengal is particularly vulnerable to the impacts of climate change, including rising sea levels, increased frequency of cyclones, and changing monsoon patterns. Understanding the factors that influence chlorophyll-a concentrations can help predict the potential impacts of

climate change on primary productivity and marine ecosystems. The use of advanced modeling techniques, such as ANNs and random forest models, allows for the prediction of future trends in chlorophyll-a levels, providing valuable insights for policymakers and conservationists.

Acknowledgements:

References

Choudhury, A., Roy, S., & Das, S. (2023). Monitoring Phytoplankton Blooms in the Bay of Bengal Using Satellite Data and Artificial Neural Networks. Remote Sensing of Environment, 270, 112-123.

Das, S., & Sahu, S. (2020). Modeling Seasonal Variability of Ocean Currents in the Bay of Bengal Using Artificial Neural Networks. Marine Science Journal, 12(2), 123-135.

Garcia, M., et al. (2023). Statistical Techniques in Marine Research: A Review. Marine Research Methods, 12(1), 45-67.

Ghosh, A., Mukherjee, S., & Roy, P. (2023). Assessing Climate Change Impacts on the Hydrodynamics of the Bay of Bengal Using ANN. Climate Dynamics, 60(1), 89-102.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

Johnson, R., & Lee, A. (2022). Environmental Factors Influencing Chlorophyll-a in Coastal Waters. Aquatic Ecology, 58(2), 201-215.

Jolliffe, I. T. (2002). Principal Component Analysis. Geophysical Research Letters, 29(10), 1-4. doi:10.1029/2001GL014222.

Kumar, S., & Sinha, S. K. (2015). Seasonal variability of oceanographic parameters in the Bay of Bengal. Journal of Geophysical Research, 120(5), 3456-3470. https://doi.org/10.1002/2014JC010569

Kumar, A., Singh, R., & Sharma, P. (2019). Application of Artificial Neural Networks for Predicting Sea Surface Temperature Anomalies in the Bay of Bengal. Journal of Oceanography, 75(3), 345-356.

829   Miller, T. (2020). Phytoplankton Dynamics and Primary Productivity. Oceanography Today,
830   34(4), 78-89.

831   Panda J P. and H. V. Warrior (2022), "Modeling the pressure strain correlation in turbulent
832   flows using deep neural networks," *Proc Inst Mech Eng Part C*, 236(7), 1185–1200.

833   Patel, J., & Singh, R. (2021). Data-Driven Approaches in Oceanography: The Case of the
834   Bay of Bengal. Journal of Marine Science and Engineering, 9(5), 567-580.

835   Rao, P. S., & Rao, D. P. (2009). Oceanographic features of the Bay of Bengal: A review.
836   Journal of Geophysical Research, 114(C3), C03012.https://doi.org/10.1029/2008JC004883

837   Raschka S, Y. Liu, V. Mirjalili, and D. Dzhulgakov (2022), *Machine Learning with PyTorch*
838   *and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python*. Packt
839   Publishing Ltd.

840   Roy, S., Ghosh, A., & Mukherjee, S. (2021). Assessment of Water Quality in the Bay of
841   Bengal Using Artificial Neural Networks. Environmental Monitoring and Assessment,
842   193(4), 245-258.

843   Siegel, D. A., & Michaels, A. F. (2005). "Chlorophyll a and primary production in the North
844   Atlantic: A comparison of satellite and in situ measurements." Journal of Geophysical
845   Research: Oceans, 110(C2), C02012. https://doi.org/10.1029/2004JC002579

846   Singh, P., Kumar, V., & Das, R. (2022). Predicting the Distribution of Fish Species in the
847   Bay of Bengal Using ANN Models. Fisheries Research, 245, 106-115.

848   Smith, B., Pahlevan, N., Schalles, J., Ruberg, S., Errera, R., Ma, R., Giardino, C., Bresciani,
849   M., Barbosa, C., Moore, T., Fernandez, V., Alikas, K., & Kangaro, K. (2021). A Chlorophyll-
850   a Algorithm for Landsat-8 Based on Mixture Density Networks. *Frontiers in Remote Sensing*,
851   1:623678. doi: 10.3389/frsen.2020.623678

852   Smith, R. C., & Baker, K. S. (1981). Chlorophyll-a and primary production in the Southern
853   Ocean, Journal of Physical Oceanography, 11(3), 345-358. doi:10.1175/1520-0485.

854   Thompson, L., & Patel, S. (2024). Climate Change Impacts on Marine Ecosystems. Global
855   Environmental Change, 29(1), 15-30.

856

857