International Core Journal of Engineering Volume 11 Issue 5, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

Factors Influencing and Predicting Marine Chlorophyll Content

Jiaxin Tian[a], Shaosong Zhang[b]
College of science, North China University of Science and Technology, Tangshan, China
[a]2187635224@qq.com, [b]zss19831957585@126.com

Abstract

Marine chlorophyll content plays an important role in the global carbon cycle and climate change regulation. Based on machine learning and data mining techniques, this paper explores its key influencing factors and constructs a prediction model. Gray correlation analysis and Spearman's rank correlation coefficient method are used to screen out the main environmental variables, such as carbon dioxide concentration, pH, salinity, temperature and dissolved oxygen, and the Random Forest Model is used to analyze the importance of the characteristics and predict the future trend. The results showed that the chlorophyll content was significantly correlated with the environmental factors, especially affected by CO2 concentration. The results of the study can provide data support and scientific basis for the management of ocean carbon sinks and climate change response strategies.

Keywords

Marine Carbon Sinks; Chlorophyll Content; Machine Learning; Environmental Factors; Random Forests.

1. Introduction

Oceans, covering approximately 71% of the Earth's surface, play a crucial role in regulating global climate and supporting biodiversity. As a major component of the carbon cycle, the ocean functions as a carbon sink by absorbing atmospheric CO_2 through biological, chemical, and physical processes. Among these, phytoplankton photosynthesis, which is closely related to chlorophyll concentration, serves as a key mechanism for marine carbon sequestration. Understanding the factors influencing chlorophyll concentration is therefore essential for assessing the efficiency of marine carbon sinks and formulating climate mitigation strategies.

Previous studies have identified multiple environmental factors affecting chlorophyll concentration, including CO_2 concentration, pH, salinity, temperature, and dissolved oxygen levels. However, the complex, nonlinear interactions between these variables pose challenges for traditional statistical approaches. Recently, machine learning techniques have demonstrated strong potential in capturing these complex relationships and improving predictive accuracy. In particular, ensemble learning models such as Random Forest have been widely used for feature selection and forecasting in environmental and ecological studies.[1, 2]

This study employs a data-driven approach to analyze key environmental determinants of chlorophyll concentration and develop a predictive model. Using publicly available oceanographic datasets, we apply Grey Relational Analysis and Spearman correlation to identify the most influential factors. A Random Forest model is then constructed to assess feature importance and predict future trends in chlorophyll concentration. The findings of this study provide insights into the dynamics of marine carbon sinks and offer scientific support for climate change adaptation and ocean resource management.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

2. Data Processing and Feature Engineering

2.1 Data Sources

The dataset used in this study is obtained from the Chinese Academy of Sciences Marine Science Data Center. It includes oceanographic parameters such as chlorophyll concentration, CO_2 concentration, pH, salinity, temperature, dissolved oxygen, and various nutrient levels. The data spans from 1973 to 2021 and focuses on the Bohai Sea region. Latitude and longitude coordinates are used for spatial positioning, ensuring accurate regional analysis. The key variables in the dataset are summarized in Table 1.

Table 1. Summary of Key Variables in the Dataset

Variable Name	Description	Variable Name	Description
G2year	**Year**	**G2temperature**	**Temperature (°C)**
G2month	Month	G2salinity	Salinity
G2pressure	Pressure	G2oxygen	Dissolved oxygen
G2chla	Chlorophyll	G2nitrate	Nitrate content
G2phts25p0	pH (25°C)	G2phosphate	Phosphate content
G2latitude	Latitude	G2silicate	Silicate content
G2longitude	Longitude	G2tco2	Total CO_2 content
G2bottomdepth	Bottom depth	G2alk	Alkalinity

2.2 Data Preprocessing

To ensure the quality of the dataset, several preprocessing steps were performed:

1) Handling Missing Values: The dataset contains missing values, primarily represented by -9999. These were replaced with 0 to facilitate numerical processing.

2) Normalization: Different variables have varying magnitudes and units. To standardize the data, Min-Max Scaling was applied, normalizing each feature within the range of [0,1] using Equation (1):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is the original value, x_{min} and x_{max} are the minimum and maximum values of the variable, and x' is the normalized value.

3) Outlier Detection and Removal: Outliers were identified using the interquartile range (IQR) method. The IQR was computed as follows:

$$IQR = Q3 - Q1 \tag{2}$$

Data points falling outside the range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ were considered outliers and removed to ensure the robustness of the analysis.

2.3 Feature Engineering

Feature engineering was performed to reduce dimensionality and enhance the model's predictive power:

1) Grey Relational Analysis (GRA): GRA was used to assess the relationship between chlorophyll concentration and other environmental variables. The grey correlation coefficient for each feature was calculated using Equation (2):

$$\xi_i(k) = \frac{\min\Delta + \rho\max\Delta}{\Delta_i(k) + \rho\max\Delta} \tag{3}$$

where $\Delta_i(k)$ represents the absolute difference between the reference and comparative sequences, ρ is the distinguishing coefficient (typically 0.5), and $\xi_i(k)$ is the grey relational coefficient. The grey relational degree for each feature is summarized in Table 2.

Table 2. Grey Relational Degree of Features

Variable	Grey Relational Degree	Variable	Grey Relational Degree
G2tco2	0.811	G2temperature	0.654
G2nitrate	0.721	G2phts25p0	0.672
G2latitude	0.678	G2salinity	0.478

Based on the results, high-correlation features such as CO_2 concentration, temperature, and nitrate content were retained for further analysis.

2) Feature Selection Using Random Forest: The importance of each feature in predicting chlorophyll concentration was determined using a Random Forest model. The feature importance was computed using Equation (3):

$$I_{RF}(X_i) = \sum_{t=1}^{T} \sum_{s\in S_t, v(s)=X_i} \Delta I(s,t) \tag{4}$$

where $I_{RF}\left(X_i\right)$ represents the importance of feature X_i, T is the total number of trees, S_t is the set of nodes in tree t, and $I(s,t)$ is the impurity decrease at node sss. The results of the feature importance analysis are illustrated in Figure 1.

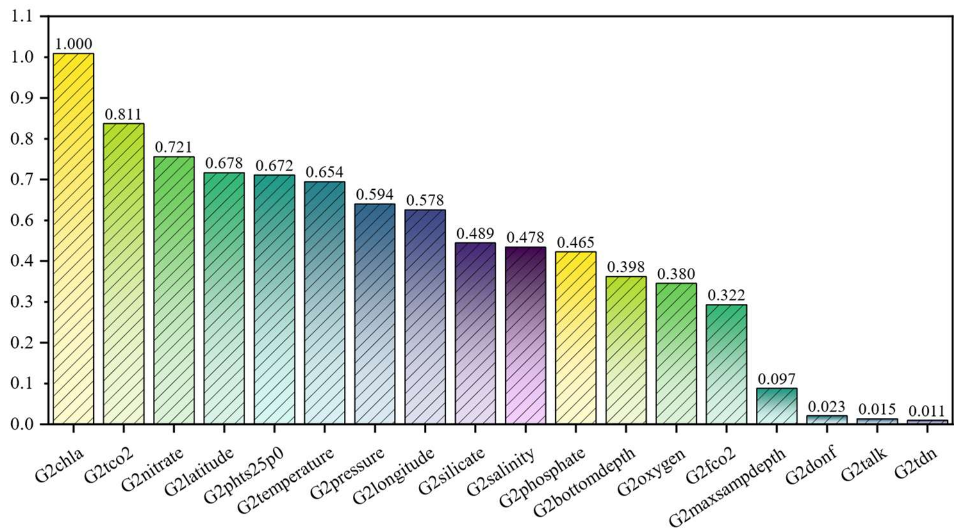

Figure 1. Feature Importance Analysis Using Random Forest

The results indicate that dissolved oxygen, nitrate concentration, and total CO_2 content are the most influential factors affecting chlorophyll concentration. These selected features were used in the subsequent predictive modeling.

3. Analysis of Key Influencing Factors

Understanding the key factors influencing chlorophyll concentration is crucial for evaluating its role in marine carbon sequestration. This section examines the statistical properties of the dataset, determines correlations between variables using Spearman's rank correlation, and applies a Random Forest model to assess feature importance.

3.1 Normality and Continuity Tests

Before performing correlation analysis, it is essential to evaluate the distribution and continuity of the data. The Jarque-Bera (JB) test was applied to assess normality, while a difference test was used to determine data continuity.

1) Continuity Test: To verify whether the data follows a continuous trend, the first-order difference was computed:

$$\Delta x_i = x_i - x_{i-1} \tag{5}$$

where x_i represents the value of a given variable at time i. If the differences remain relatively uniform, the data is considered continuous; otherwise, discontinuities may exist.

2) Normality Test (JB Test): The JB test evaluates whether the dataset follows a normal distribution by considering skewness and kurtosis:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4} \right) \tag{6}$$

where:

n is the sample size,

S is the skewness coefficient, computed as

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma} \right)^3 \tag{7}$$

K is the kurtosis coefficient, calculated as

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma} \right)^4 \tag{8}$$

If the JB statistic is significantly high (p-value < 0.05), the null hypothesis of normality is rejected.

Volume 11 Issue 5, 2025
DOI: 10.6919/ICJE.202505_11(5).0019

Table 3. Normality and Continuity Test Results

Variable	JB Statistic	p-value	Continuity
G2latitude	572.57	0	No
G2longitude	452.34	0	No
G2chla	45.237	0	No
G2tco2	76.732	0	No
G2temperature	78.691	0	No
G2salinity	478.79	0	No

The results indicate that none of the variables follow a normal distribution or exhibit strict continuity. As a result, non-parametric methods such as Spearman's rank correlation were chosen for further analysis.

3.2 Spearman's Rank Correlation Analysis

Given the non-normality of the data, Spearman's rank correlation coefficient was used to quantify the monotonic relationship between chlorophyll concentration and environmental factors. Spearman's correlation is computed as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{9}$$

where d_i represents the difference between the ranks of the paired observations and n is the sample size.

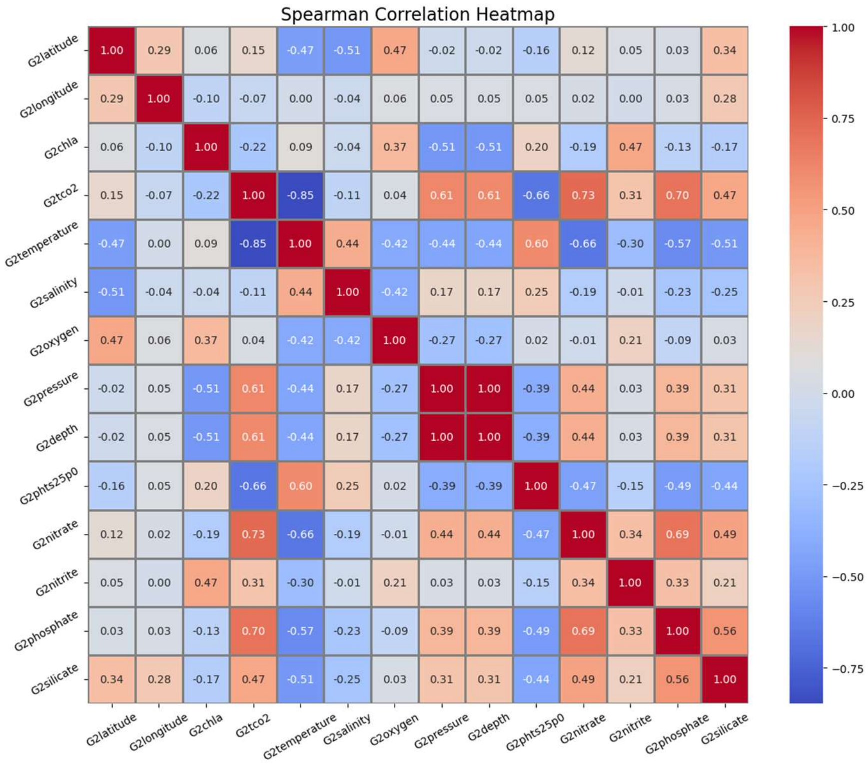

Figure 2. Heatmap of Spearman's Correlation Coefficients

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

Key Findings:

Total CO_2 concentration (G2tco2) shows a strong negative correlation (-0.51) with chlorophyll concentration, suggesting that increased CO_2 may inhibit phytoplankton growth due to ocean acidification.

Temperature (G2temperature) is negatively correlated (-0.51) with chlorophyll concentration, indicating that warming conditions may reduce marine primary productivity.

Dissolved oxygen (G2oxygen) exhibits a moderate positive correlation (0.37) with chlorophyll concentration, likely due to photosynthesis.

These results suggest that ocean acidification and warming trends may negatively impact phytoplankton populations, thereby reducing marine carbon sequestration efficiency.

3.3 Feature Importance Analysis Using Random Forest

A Random Forest (RF) regression model was employed to evaluate the relative importance of environmental factors affecting chlorophyll concentration. Feature importance is measured by the reduction in mean squared error (MSE) when a variable is used for node splitting:

$$I_{RF}(X_i) = \sum_{t=1}^{T} \sum_{s \in S_t, v(s)=X_i} \Delta I(s,t) \tag{10}$$

where:

$I_{RF}\left(X_i\right)$ is the importance of feature X_i ,

T is the number of decision trees,

S_t is the set of nodes in tree t,

$\Delta I(s,t)$ is the impurity reduction at node s .

Table 4. Feature Importance Ranking (Random Forest)

Feature	Importance Score
G2oxygen	0.271
G2nitrate	0.219
G2depth	0.181
G2salinity	0.132
G2pressure	0.101
G2tco2	0.096

Key Findings:

Dissolved oxygen is the most important predictor of chlorophyll concentration, reinforcing the role of photosynthesis.

Nutrient availability (nitrate) also significantly influences phytoplankton growth, confirming its role in primary productivity.

CO_2 concentration, while negatively correlated with chlorophyll, has relatively lower feature importance, suggesting that its effects may be indirect or modulated by other variables.

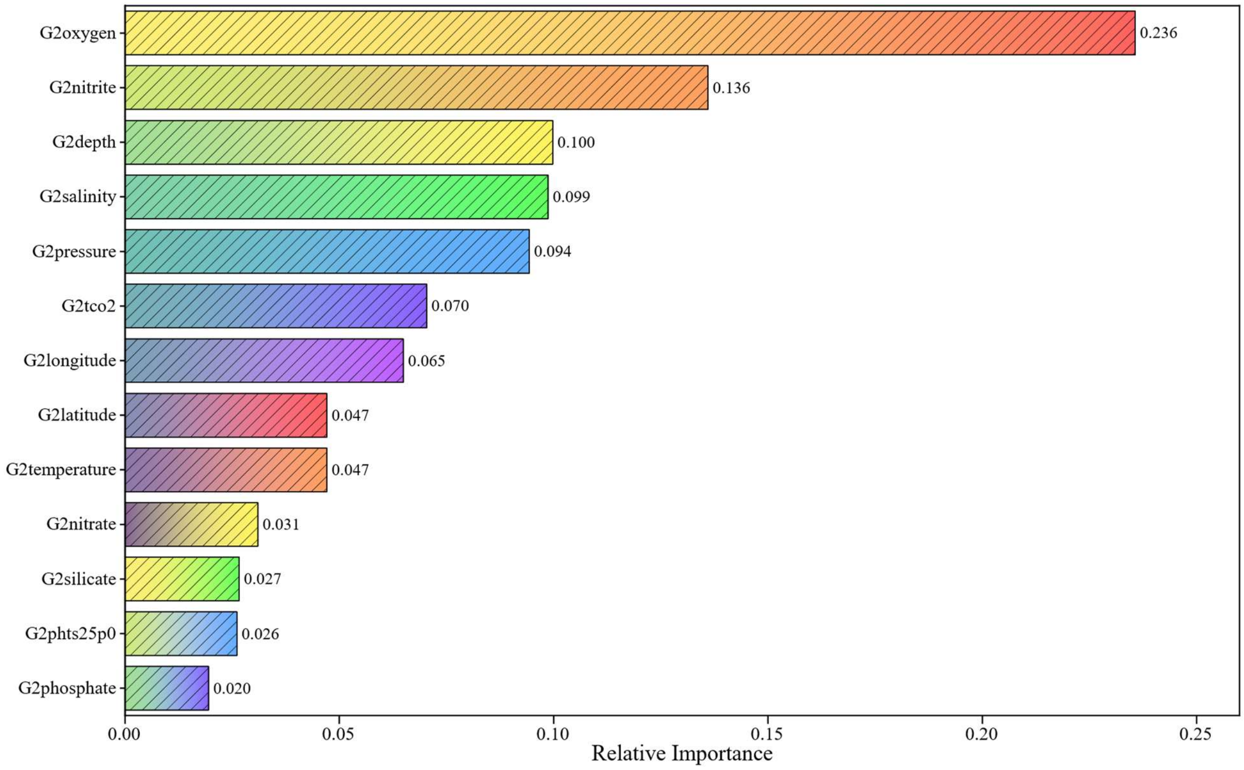

Figure 3. Feature Importance Visualization (Random Forest)

Summary

The dataset does not follow a normal distribution or exhibit strict continuity, necessitating the use of non-parametric correlation analysis.

Spearman's correlation analysis confirms significant relationships between chlorophyll concentration and CO_2, temperature, and dissolved oxygen levels.

Random Forest feature importance analysis highlights dissolved oxygen and nitrate as the most influential factors in predicting chlorophyll concentration.

The findings provide a basis for predictive modeling and future trend analysis, helping to assess the impact of environmental changes on marine carbon sequestration.

4. Prediction Model for Chlorophyll Concentration

To forecast chlorophyll concentration and assess its potential role in marine carbon sequestration, a Random Forest regression model was employed. This section details the dataset partitioning, model construction, and evaluation of predictive performance.

4.1 Data Partitioning and Processing

To ensure robust model training and evaluation, the dataset was split into a training set (pre-2018 data) and a test set (2018 onwards). A rolling window technique was applied to smooth short-term fluctuations, ensuring reliable predictions.

Data Splitting Strategy

Training Set: Data before January 1, 2018, used to train the model.

Test Set: Data from January 1, 2018, onwards, used for validation and performance assessment.

To mitigate missing values, a 7-day rolling window average was computed:

DOI: 10.6919/ICJE.202505_11(5).0019

$$x_t^{'} = \frac{1}{7}\sum_{i=t-6}^{t} x_i \tag{11}$$

where $x_t^{'}$ represents the smoothed value at time t.

Feature Selection for Prediction

Based on correlation analysis and Random Forest feature importance ranking, six key predictors were retained (Table 5).

Table 5. Selected Features for Prediction

Feature	Description
G2temperature	Temperature (°C)
G2salinity	Salinity
G2pressure	Pressure
G2oxygen	Dissolved oxygen
G2bottomdepth	Depth
G2tco2	Total CO_2 content

4.2 Random Forest Model Construction and Prediction

A Random Forest (RF) regression model was implemented due to its robustness in handling non-linear relationships and multi-variable interactions. The model was trained using the selected features to predict chlorophyll concentration.

Prediction Function of Random Forest

Given a feature vector XXX, the RF model's prediction is computed as the average output of multiple decision trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \tag{12}$$

where:

$\hat{y}$ is the predicted chlorophyll concentration,

T is the total number of trees,

$f_t(X)$ is the prediction from the t-th tree.

Model Training Parameters

The RF model was optimized using the following parameters:

Number of trees: 200

Maximum depth: 10

Minimum samples per leaf: 5

Feature selection method: Mean decrease in impurity

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

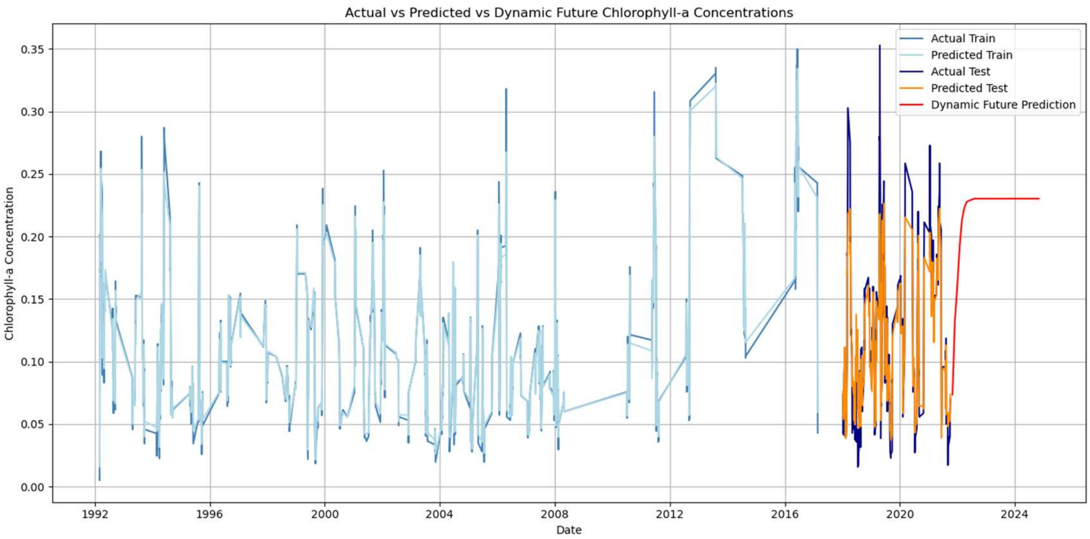

Figure 4. Chlorophyll Concentration Prediction Results

4.3 Model Evaluation and Performance Analysis

To assess the predictive performance of the Random Forest model, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score were calculated.

Evaluation Metrics

1) Mean Squared Error (MSE): Measures average squared deviations between predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{13}$$

2) Mean Absolute Error (MAE): Computes the absolute differences between predicted and observed values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{14}$$

3) R^2 Score (Coefficient of Determination): Evaluates model goodness-of-fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{15}$$

Table 6. Model Performance Metrics

Metric	MSE	MAE	R^2
Value	0.0004	0.0118	0.8883

Key Findings

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

The low MSE (0.0004) and MAE (0.0118) indicate high prediction accuracy.

The high R^2 value (0.8883) suggests that the model explains 88.8% of the variance in chlorophyll concentration, confirming its reliability.

Future chlorophyll concentration is projected to increase steadily with minor fluctuations, likely due to ecological and climate-driven factors.

Summary

The dataset was split into training (pre-2018) and test sets (2018+), with rolling window smoothing applied for noise reduction.

A Random Forest regression model was built using six key environmental factors, providing robust predictions.

Evaluation metrics confirmed high accuracy and reliability (R2=0.8883), supporting the feasibility of machine learning-based chlorophyll concentration forecasting.

The model predicts a gradual increase in chlorophyll concentration, offering valuable insights for marine carbon sink management and climate policy development.

5. Conclusion and Recommendations

5.1 Conclusion

This study systematically analyzed the key environmental factors influencing chlorophyll concentration in the ocean and developed a predictive model using machine learning techniques. The major findings are summarized as follows:

1) Key Environmental Influences:

Through Grey Relational Analysis and Spearman's Rank Correlation, total CO_2 concentration, temperature, salinity, pH, and dissolved oxygen were identified as significant factors affecting chlorophyll concentration.

A strong negative correlation was observed between chlorophyll concentration and CO_2 concentration, suggesting potential impacts of ocean acidification on primary productivity.

2) Feature Importance and Prediction:

Random Forest analysis revealed that dissolved oxygen and nitrate concentration were the most critical variables influencing chlorophyll concentration.

The predictive model demonstrated high accuracy (MSE = 0.0004, MAE = 0.0118, R2 = 0.8883), confirming its reliability in forecasting chlorophyll trends.

Future projections suggest a gradual increase in chlorophyll concentration, potentially influenced by environmental and anthropogenic factors.

3) Implications for Marine Carbon Sequestration:

Given the strong correlation between chlorophyll and CO_2 levels, enhancing phytoplankton productivity through sustainable marine resource management could contribute to increased carbon sequestration.

The findings provide scientific support for climate mitigation policies aimed at optimizing oceanic carbon sinks.

5.2 Recommendations

1) Enhanced Monitoring and Data Integration

Establish high-frequency monitoring stations in key oceanic regions to track variations in chlorophyll and CO_2 levels.[3,4]

Utilize satellite-based remote sensing and AI-driven data analytics for real-time assessment of oceanic primary productivity.

2) Policy and Management Strategies

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0019

Implement marine conservation policies that protect phytoplankton-rich ecosystems, such as reducing nutrient pollution and regulating fishing activities.

Incorporate oceanic carbon sequestration into global carbon markets, providing economic incentives for conservation efforts.

3) Future Research Directions

Extend the machine learning framework to incorporate deep learning and ensemble models for improved chlorophyll prediction.

Investigate regional variations in phytoplankton responses to climate change, considering interactions between multiple oceanographic parameters.

By integrating machine learning with oceanographic data analysis, this study provides a scientific basis for marine ecosystem management and climate policy formulation. Future work should focus on refining predictive models and exploring adaptive strategies to enhance the role of oceans in global carbon sequestration.[5]

References

[1] Orr J C, Fabry V J, Aumont O, et al. Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms [J]. Nature, 2005, 437(7059): 681-686.

[2] Ryther J H ,Yentsch C S. The estimation of phytoplankton production in the ocean from chlorophyll and light data. Limnology and Oceanography,1957,2(3) : 281-286.

[3] Jiang Y, Zhang T, Gou Y, et al. High-resolution temperature and salinity model analysis using support vector regression[J]. Journal of Ambient Intelligence and Humanized Computing, 2018:1-9.

[4] Gou Y, Liu J, Zhang T. KNN regression model-based refinement of thermohaline data[C]//Proceedings of the Thirteenth ACM International Conference on Underwater Networks & Systems. 2018:1-8.

[5] Li Z, Wang J, Cao D, et al. Investigating Neural Activation Effects on Deep Belief Echo-State Networks for Prediction Toward Smart Ocean Environment Monitoring[J]. Arabian Journal for Science and Engineering, 2021, 46(4):3913-3923.