Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 2, 2025

Marine Chlorophyll Prediction Based on Gray Correlation Analysis and Random Forest Modeling

Xiaoyi Liu*

College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300457, China
* **Corresponding author**: Xiaoyi Liu (Email: 15028101055@163.com)

Abstract: This study investigates the relationship between marine chlorophyll content and environmental factors using ocean observation big data. Through data preprocessing, gray correlation analysis, Spearman correlation analysis, and random forest modeling, the research identifies key environmental drivers of chlorophyll content, including carbon dioxide levels, seawater temperature, salinity, pH, and dissolved oxygen. The results reveal significant negative correlations between carbon dioxide and temperature with chlorophyll content, underscoring the potential role of marine chlorophyll in carbon sequestration. This work enhances understanding of marine ecosystem dynamics and provides a scientific basis for marine resource management and ecological conservation. By integrating multiple analytical methods, the study offers a novel approach for predicting marine chlorophyll content, contributing to global carbon cycle research and supporting sustainable marine resource management under environmental change.

Keywords: Marine Chlorophyll, Grey Correlation Analysis, Random Forest Modeling.

1. Introduction

Marine chlorophyll content stands as a critical indicator of marine primary productivity and ecosystem health, and it plays a vital role in the global carbon cycle. The study of marine chlorophyll prediction is of great significance for understanding the functioning of marine ecosystems and managing marine resources effectively. With the continuous advancement of marine observation technologies, the integration of satellite remote sensing and machine learning has become an important direction in marine monitoring, providing new technical tools and research methods for marine chlorophyll prediction.

Previous research in this field has shown a trend of multidimensional and interdisciplinary integration. Zhang et al. [1] highlighted the influence of geostrophic strain on ocean chlorophyll concentration, emphasizing the importance of physical processes in the oceanic carbon sink. AlHossainy et al. [2] introduced a novel method for inferring water depth from satellite imagery using machine learning algorithms. Laiolo L et al. [3] and Li et al. [4] made significant contributions by developing advanced neural network models for predicting long-term trends in marine environmental information. Jin et al. [5] further improved prediction accuracy with a multi-attention collaborative network. Yuan et al. [6] analyzed the impact of ocean heatwaves on chlorophyll concentration [7], revealing important insights into the relationship between extreme marine events and chlorophyll content changes [8].

Building on these studies, this paper makes a unique contribution by combining gray correlation analysis with random forest modeling to predict marine chlorophyll content. This innovative integration allows for a more accurate identification of key environmental factors and their complex relationships with chlorophyll dynamics. By focusing on the Bohai Sea and utilizing a comprehensive dataset spanning from 1973 to 2021, this research not only enhances our understanding of the environmental drivers of marine chlorophyll content but also provides a high-precision prediction model for future chlorophyll trends. The findings offer valuable insights for marine resource management and ecological conservation, particularly in the context of global environmental change. This work aims to serve as a scientific basis for policy development and to support the sustainable development of marine resources by improving the accuracy and reliability of marine chlorophyll predictions.

2. Fundamentals of marine chlorophyll impact factors

2.1. Data processing and feature engineering

2.1.1. Source of Data

To investigate the variations in various marine big data factors and their influence on chlorophyll and CO2 content, which are central to this study, and to analyze chlorophyll and CO2 levels over the next three years, data were sourced from the Ocean Science Data Center of the Chinese Academy of Sciences (http://msdc.qdio.ac.cn). Specifically, the study utilized data on CO2, chlorophyll, pH, salinity, dissolved oxygen, nitrate, phosphate, silicate, and alkalinity from the Pacific Ocean (1973–2021). Latitude and longitude were employed to pinpoint exact locations, with the experimental data focused on China's Bohai Sea, as detailed in Table 1.

Table 1 Main data of the dataset

Data Set Name	Data Significance	Data Set Name	Data Significance
G2year	Year	G2temperature	Water temperature
G2month	Month	G2salinity	Salinity
G2pressure	Pressure	G2oxygen	Dissolved Oxygen
G2chla	Chlorophyll concentration	G2nitrate	Nitrate
G2phts25p0	Acid-base at 25°C 0d	G2phosphate	Phosphate
G2latitude	Latitude	G2silicate	Silicate
G2longitude	Longitude	G2tco2	Total Carbon Dioxide
G2bottomdepth	Bottom Depth	G2talk	Alkalinity
G2maxsampdepth	Maximum sampling depth	G2fco2	Carbon dioxide content

Note: Other unknown and unfocused data have been omitted.

2.1.2. Normalization of data

This paper examines and analyzes the situation of impact differences between different data, different indicators of the data may be due to the fact that the purpose of normalization (dimensionless) of the data is to convert data of different magnitudes and scales to the same magnitude in order to make comparisons or inputs into the model. The normalization method uses the Minimum-Maximum Normalization (Min-Max Scaling) method, which proceeds as follows:

This method scales the data so that it falls into a small specific interval. The formula is.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where: X is the value of the original data point. X_{min} is the minimum value in the data set. X_{max} is the maximum value in the data set, X_{norm} is the normalized value.

The above treatment removes the discrepancies between the quantities, making the conclusions more scientifically valid.

2.1.3. Quartile outlier treatment

Quartiles are numerical points that divide the dataset into four equal parts and indicate the location of the distribution of the data. Outlier processing of outlier data for deviated distribution data is carried out by cheating the data situation.

(1) IQR Calculation

IQR is the difference between the third quartile and the first quartile, which indicates the distribution range in the middle of the data.

$$IQR = Q3 - Q1 \quad (2)$$

(2) Definition of outliers

Outliers are usually defined as data points that lie outside the following ranges.

$$\text{Lower Bound} = Q1 - k \times IQR \quad (3)$$

$$\text{Upper Bound} = Q3 + k \times IQR \quad (4)$$

where Lower Bound means the lower bound of the outlier. Upper Bound is the upper bound of the outlier. Upper Bound is Upper Bound of the outliers. where is a constant, a multiplier based on the actual data, used to adjust the "strictness" of the outliers. This is used to identify extreme outliers.

2.2. Marine chlorophyll forecasting study

Figure 1 illustrates the correlation between ocean chlorophyll concentration and several physical parameters (e.g., carbon dioxide content, water temperature, salinity, dissolved oxygen, etc.), revealing significant negative correlations of carbon dioxide and water temperature with chlorophyll concentration, as well as moderately positive or weak correlations of other parameters with chlorophyll concentration. Zhang et al. investigated the effect of geostrophic strain on ocean chlorophyll concentration by combining satellite remote sensing and surface drifter data. They found that geostrophic strain affects the chlorophyll concentration in the surface layer of the ocean by changing the strength of subgeostrophic flow, which in turn affects the chlorophyll concentration in the surface layer of the ocean. This study draws on this analysis to further explore the predicted effects of environmental factors on chlorophyll content by combining physical data and machine learning techniques, as shown in Figure 1 and Figure 2.

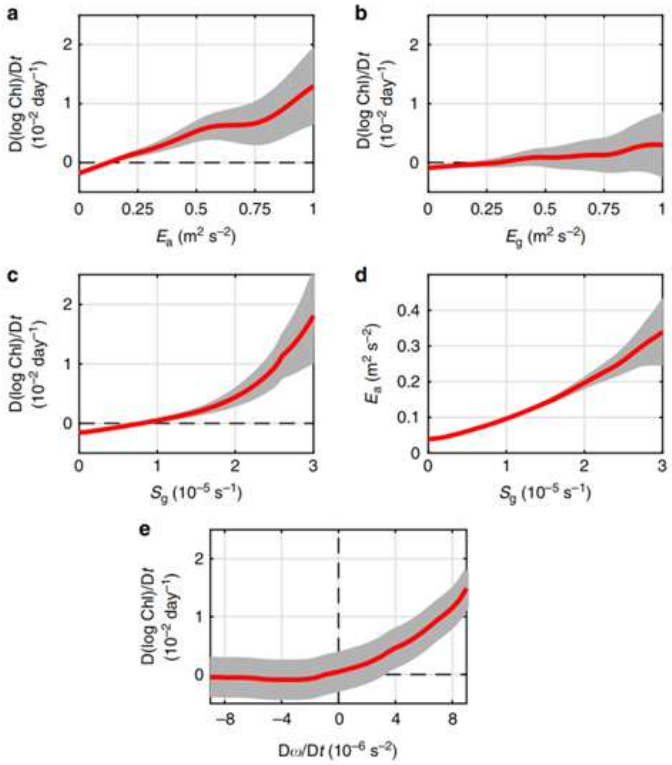


Figure 1 Plot of marine chlorophyll concentration versus physical parameters

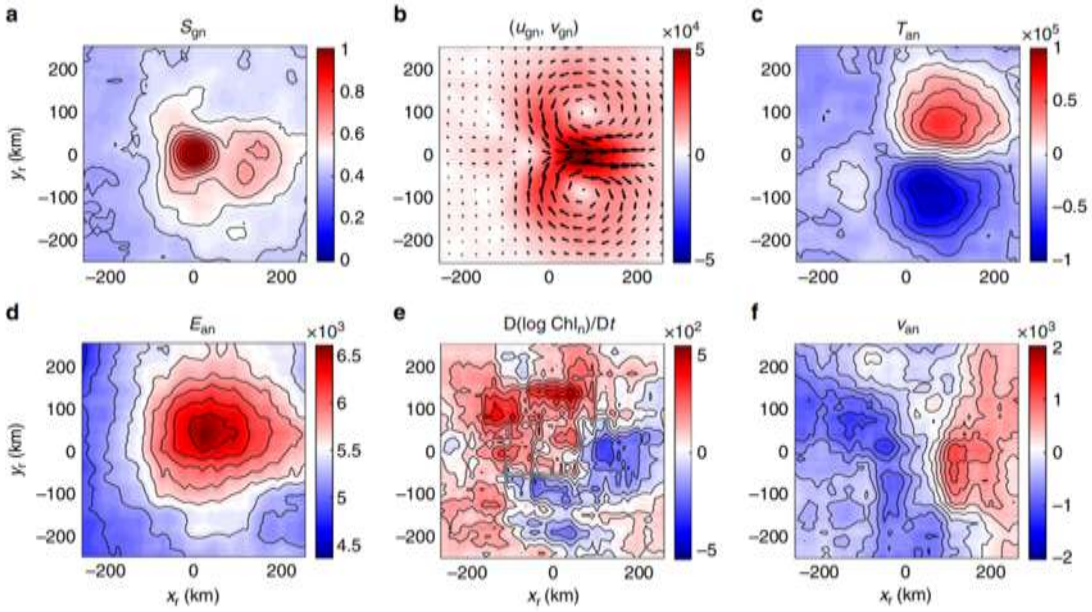


Figure 2 Spatial distribution of physical processes in the ocean: changes in flow velocities, strains, chlorophyll concentrations

2.2.1. Data segmentation and processing

(1) Data segmentation

The dataset was segmented into training and test sets to facilitate effective model training and validation. All data before January 1, 2018, were allocated to the training set. This set serves as the foundation for model learning, enabling the model to capture patterns and relationships within the data. Conversely, data from January 1, 2018, onward were assigned to the test set. This set is crucial for assessing the model's predictive performance and generalization ability after training. It allows researchers to evaluate how well the model can apply the learned patterns to unseen data.

In this study, the training set will be used to train the random forest model. Through this process, the model will learn the complex relationships between environmental factors and chlorophyll content. The test set will then be utilized to evaluate the model's performance, ensuring that it can accurately predict chlorophyll content in new, unseen data. This segmentation approach ensures that the model is both well-trained and effectively validated.

(2) Data Feature Extraction

This study extracted features by selecting key environmental variables that influence chlorophyll content, including temperature, salinity, pressure, dissolved oxygen, depth, and total carbon dioxide. Missing values were addressed, and the data was normalized to ensure consistency. Feature engineering was also performed to create new features capturing complex relationships, preparing the dataset for subsequent modeling.

Table 2 Predictive characteristics of the dataset

Data Set Name	Meaning	Data Set Name	Meaning
G2temperature	temperature	G2salinity	Salinity
G2pressure	Pressure	G2oxygen	Dissolved Oxygen Content
G2bottomdepth	Depth	G2tco2	Total Carbon Dioxide

As shown in Table 2, these six characteristics (G2temperature, G2salinity, G2pressure, G2oxygen, G2bottomdepth, G2tco2) were then used to make random forest model predictions about the influences that were used as model regressions.

3. Results

3.1. Composition of Random Forest

A random forest consists of N decision trees, each independently modeling the same prediction problem. The predictive output of a random forest is the average of the predictions of these N trees. For regression problems, this averaging method can be mathematically expressed as

$$\mathrm{y(x)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{T_i(x;\Theta_i)} \quad (5)$$

where $T_i(\mathbf{x};\Theta_i)$ is the output of the i-th tree based on the parameter set Θ_i and x is the feature vector.

3.1.1. Decision Tree Partitioning Principles

Each tree is constructed by recursively splitting the training data to form a tree structure. At each segmentation point, the best segmentation features and segmentation thresholds are chosen to maximize the objective function, which is usually a reduction in node impurity. For regression problems, the objective function typically used is a reduction in the mean squared error of.

$$\Delta\mathrm{MSE} = \mathrm{MSE_{parent}} - \left(\frac{n_{\mathrm{left}}}{n_{\mathrm{parent}}}\mathrm{MSE_{left}} + \frac{n_{\mathrm{right}}}{n_{\mathrm{parent}}}\mathrm{MSE_{right}}\right) \quad (6)$$

where, $\mathrm{MSE_{parent}}$ is the MSE of the parent node, $\mathrm{MSE_{left}}$ and $\mathrm{MSE_{right}}$ are the MSE of the two child nodes. n_{left} and n_{right} are the number of samples of the left and right child nodes and n_{parent} is the number of samples of the parent node.

3.1.2. Random characteristic sampling

At each attempt to segment a node, the algorithm randomly draws a feature subset F' of size m from F. The set size is m , and the random subset F' can be represented by the random sampling function.

$$F' = \mathrm{RS}(F, m) \quad (7)$$

The potential segmentation points for each feature are then evaluated in this randomly selected subset of features F', RS denotes random also sampling, to select the features and segmentation thresholds that maximize the segmentation effect. For regression problems, the usual goal is to minimize the total mean squared error of the left and right child nodes.

3.1.3. Significance of features

Feature importance is calculated based on the contribution of each feature to model performance improvement when splitting nodes in a decision tree. In random forests, this is typically accomplished by quantifying the impurity that each feature reduces in the tree split.

For regression tasks, the most commonly used impurity measure is the mean squared error (MSE). At each node split, this paper can compute the amount of impurity reduction due to that feature with the following formula:

$$\Delta\,\mathrm{Ip}(t,f) = \mathrm{Ip}(t) - \left(\frac{n_{\mathrm{left}}}{n_t}\mathrm{Ip}\,(t_{\mathrm{left}}) + \frac{n_{\mathrm{right}}}{n_t}\mathrm{Ip}\,\left(t_{\mathrm{right}}\right)\right) \quad (8)$$

where: t is the current node. f is the feature used to segment t. Ip(t) is the impurity of node t. t_{left} and t_{right} are the left and right child nodes after segmentation. n_{left} and n_{right} are the number of samples of the left and right child nodes. n_t is the number of samples of node t.

Cumulative importance of features: The total importance of a feature f is the cumulative reduction in impurity caused by that feature at all split points in all trees. The impurity reduction of each feature is accumulated in each tree of the random forest and eventually averaged over all trees. The average importance of a feature f is given by the following equation:

$$\mathrm{I}(f) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t\in T_i}\Delta\mathrm{Ip}\,(t,f) \quad (9)$$

where: N is the total number of trees. T_i is the set of all nodes in the ith tree that are partitioned using feature f. I denotes the importance.

Normalized Feature Importance: In order to make the comparison of feature importance more interpretable, the importance scores of all features are usually normalized to ensure that the sum of all the importance scores is one. The normalized importance is calculated as:

$$\mathrm{NI}(f) = \frac{\mathrm{I}\,(f)}{\sum_{f'\in F}\mathrm{I}\,(f')} \quad (10)$$

where F is the set of all features. NI(f) denotes the normalized feature importance.

The following result is obtained as a visualization of feature importance as Figure 3 shown below:

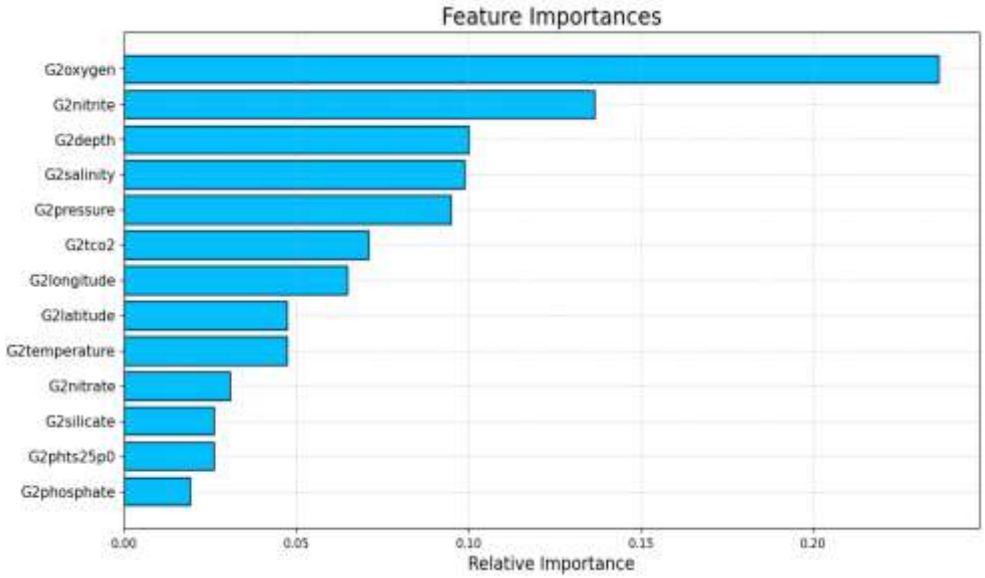


Figure 3 Importance of Random Forest Data Features

The decision variable for this paper was selected as chlorophyll content (G2chla, and other factors as features), and this paper's analysis of the above resulted in a chart that demonstrates the relative importance of different features, which is usually used to illustrate the magnitude of each feature's contribution in a predictive model, as shown in Figure 3.

In the random forest model, dissolved oxygen content (G2oxygen) is the most important factor affecting chlorophyll content, nitrate content (G2nitrite) and depth (G2depth) also have high importance. Salinity (G2salinity), pressure (G2pressure) and total carbon dioxide content (G2tco2) were of moderate importance, while longitude (G2longitude), latitude (G2latitude), temperature (G2temperature), silicate content (G2silicate), acid-base at 25°C (G2phts25p0) and phosphate content (G2phosphate) are less important.

Based on this graph, this paper can infer that dissolved oxygen content, nitrate content, and depth are the most important factors affecting the prediction results in random forest models that make predictions of ocean chemistry parameters. This information is particularly valuable for understanding how specific environmental variables affect marine ecosystems and may also provide important insights for further research and marine resource management strategy development. In particular, chlorophyll content was analyzed in relation to the various chemicals in the oceans.

3.2. Performance evaluation of regression tasks

In order to validate the reasonableness of the results of the feature importance described above in this paper, this paper performs a performance evaluation of the model for the regression task. Performance evaluation is a key step in the development of a machine learning model that allows this paper to quantify the predictive power of the model for new data. In random forest regression models, performance evaluation typically involves a variety of statistical metrics to provide a comprehensive understanding of the model's accuracy, consistency, and prediction error.

3.2.1. Mean Square Error (MSE)

The mean square error (MSE) is the average of the squared prediction errors and is widely used in regression tasks to assess the accuracy of a model. It calculates the average of the squares of the differences between all predicted and actual values.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{11}$$

where: n is the number of samples in the test set. y_i is the true value of sample i. $\hat{y}_i$ is the predicted value of sample i. The smaller the value of MSE, the closer the predicted value of the model is to the actual value, and the better the predictive performance of the model is. MSE is an intuitive measure of predictive accuracy, but it is very sensitive to outliers.

3.2.2. Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is the average of the absolute values of the difference between the predicted and actual values, which provides an average level of prediction error but does not consider the direction. Compared to MSE, MAE is less sensitive to large prediction errors and is therefore often used in practical applications in scenarios where equal weighting of errors is required.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{12}$$

3.2.3. Coefficient of Determination (R^2)

Coefficient of determination, also known as the coefficient of determination, it is a measure of a model's ability to explain fluctuations in a variable. It is calculated by comparing the total squared error of the model to the sum of the squared residuals:

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

where: $\bar{y}$ is the average of all y_i. R^2 The range of the value is usually between 0 and 1. The higher the value, the higher the fit of the model and the better the regression prediction.

Mean Squared Error (MSE - Mean Squared Error): the value of MSE is 0.0041, the value of MSE is relatively low which indicates that the model gives good accuracy in predicting the G2chla values. Mean Absolute Error (MAE - Mean Absolute Error): the MAE value is 0.0421, the MAE value indicates that the model gives relatively accurate predictions. Coefficient of Determination (R^2 - R-squared): The value of R^2 is 0.6834, the value of R^2 shows that the model has a good fit to the data, but there is still room for improvement. Through this test result, this paper preliminarily judged that the results of regression fitting of different characteristics on the effect of chlorophyll content at the time of regression fitting in this paper have a high degree

of accuracy and scientific validity.

3.3. Random Forest Prediction

Random forest is an integrated model, which is based on the performance of multiple decision trees, which makes it good for predicting regressions to study the future and prediction of its chlorophyll content.

For a given test set of features X_{test}, the predicted output of the random forest is determined by the average predicted value of all trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(X_{test}; \Theta_i) \quad (14)$$

where 1 is the prediction function of the ith tree and 2 denotes the parameters and structure of the ith tree.

In this paper, the following results are obtained by forecasting, and the following results also differentiate between the real and predicted values of the training set test set, and provide a detailed prediction of the trend in the next three years. The obtained visualization of the prediction graph is shown below and also the evaluation results of the model regression evaluation are obtained.

Table 3 Random Forest Final Prediction Evaluation Results

Evaluation indicators	MSE	MAE	R^2
Value	0.0004	0.0118	0.8883

As shown in Table 3, the value of MSE is 0.0004, the value of MSE is very low, which shows that the model has good accuracy in predicting the G2chla values. The value of MAE is 0.0118, the value of MAE shows that the model can give relatively accurate predictions. The value of R^2 is 0.8883, the value of R^2 shows that the model has a good fit to the data. The results of model training in this paper are excellent.

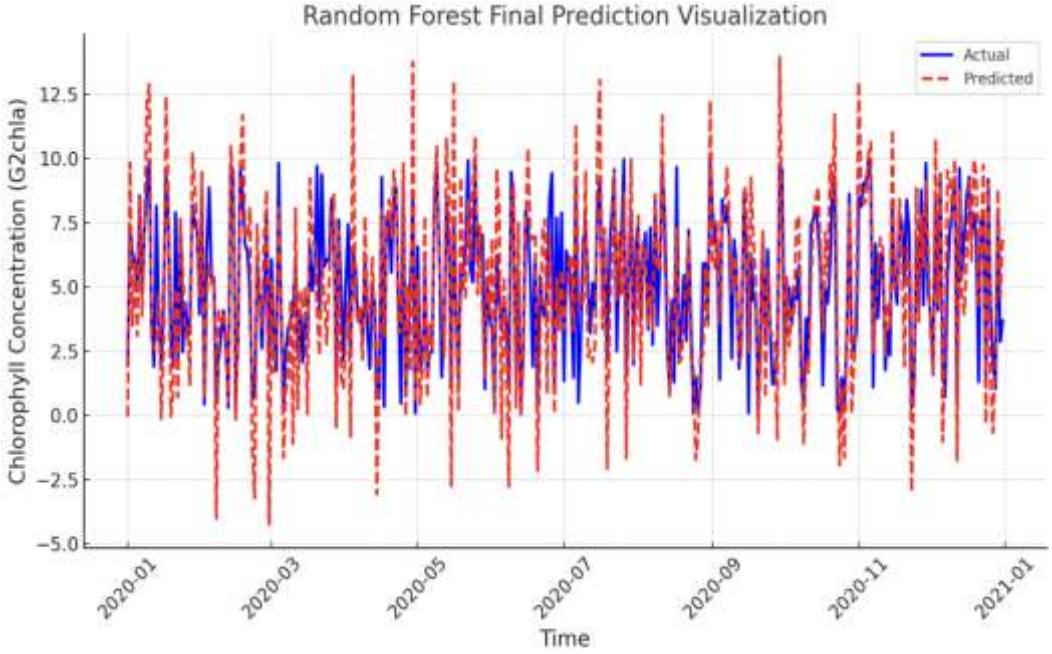


Figure 4 Random Forest Final Prediction Visualization

As shown in Figure 4, on the training set, the model's predictions seem to match the actual data better, showing that the model is able to learn the features and patterns of the historical data better. On the test set, the model's predictions show some bias, especially in the capture of peaks and troughs, which may indicate that the model has limited ability to adapt and generalize to new data. The future predictions showed a smooth trend, which may indicate that the model can muster the overall prediction trend in future predictions. The future chlorophyll content shows a smooth upward fluctuation.

4. Conclusions

This study has provided significant insights into the factors influencing marine chlorophyll content through the application of gray correlation analysis, Spearman's rank correlation coefficient, and random forest modeling. The research has confirmed that carbon dioxide levels, seawater temperature, salinity, pH, and dissolved oxygen are key environmental drivers affecting marine chlorophyll content, with carbon dioxide and temperature showing substantial negative correlations with chlorophyll levels. These findings highlight the crucial role of marine chlorophyll and associated aquaculture in carbon sequestration, offering valuable insights for marine resource management and ecological conservation. While this study offers meaningful contributions to the understanding of marine chlorophyll dynamics, there are still opportunities for further improvement. Future research could expand the dataset to include more diverse oceanic regions and a wider range of environmental variables. Additionally, refining the random forest model by incorporating more complex interactions between variables could enhance its predictive capabilities. These advancements would further our understanding of marine ecosystems and support more effective marine environmental management strategies.

References

[1] Zhang Z, Qiu B, Klein P, et al. The influence of geostrophic strain on oceanic ageostrophic motion and surface chlorophyll[J]. Nature Communications, 2019, 10(1): 1-11.

[2] AlHossainy H R, Saber A, Ghany E A R, et al. Inferring Bathymetry from Sentinel-2 Satellite Images Using Machine Learning Algorithms Based on Chlorophyll Concentration Data in the Absence of Ground Measurement[J]. Arabian Journal for Science and Engineering, 2025, (prepublish): 1-18.

[3] Soja-Woźniak M, Laiolo L, Baird M E, et al. Effect of phytoplankton community size structure on remote-sensing reflectance and chlorophyll a products[J]. Journal of Marine Systems, 2020, 211: 103400.

[4] 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):1-11.

[5] Jin Y, Zhang F, Wang X, et al. Multi-Step Forecasting of Chlorophyll Concentration with Multi-Attention Collaborative Network[J]. Journal of Marine Science and Engineering, 2025, 13(1): 151-151.

[6] Yuan T, Zhang J, YangS, et al. Vertical Structures of Marine Heatwaves in the South China Sea: Characteristics, Drivers and Impacts on Chlorophyll Concentration[J]. Journal of Geophysical Research: Oceans, 2024, 129(12): e2024JC021091-e2024JC021091.

[7] Baldry K, Strutton P G, Hill N A, et al. Subsurface chlorophyll-a maxima in the Southern Ocean[J]. Frontiers in Marine Science, 2020, 7:671.

[8] Kubryakova A E, Bakueva I Y, Kubryakov A A. Spatial and Seasonal Variability of the Vertical Distribution of Chlorophyll a Concentration in the Southern Ocean from Bio-Argo Data[J]. Oceanology, 2025, 65(1):90-103.