Check for updates

JGR Biogeosciences

RESEARCH ARTICLE
10.1029/2024JG008386

A Deep Learning Approach of Artificial Neural Network With Attention Mechanism to Predicting Marine Biogeochemistry Data

Mingzhi Liu[1,2], Yipeng Wang[1,2], Guoqiang Zhong[3], Yongxin Liu[1,2], Xiaoqing Liu[1,2], Jifan Shi[4], Yangli Che[1,2], and Rui Bao[1,2]

[1]Key Laboratory of Marine Chemistry Theory and Technology, Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ministry of Education/Institute for Advanced Ocean Studies, Ocean University of China, Qingdao, China, [2]Laboratory for Marine Geology, Qingdao Marine Science and Technology Center, Qingdao, China, [3]Department of Computer Science and Technology, Ocean University of China, Qingdao, China, [4]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, China

Abstract Predicting marine biogeochemical data is an effective method to solve the problem of marine data-scarcity and provides data support for fundamental research in marine science. Machine learning techniques are commonly used to improve the stability and accuracy of predicting biogeochemistry data. However, current methods based on Random Forest (RF) and Artificial Neural network (ANN) often struggle to effectively capture the intricate features of ocean data, resulting in suboptimal prediction accuracy. In this study, we develop a novel deep learning method called artificial neural network with attention mechanism (ANN-att) for predicting marine biogeochemistry data. We compare and evaluate the performance of RF, ANN, and ANN-att based on two widely used ocean data sets in marine biogeochemistry: GLODAP v2.2022 and MOSAIC 2.0. Our results show that the prediction accuracy of the ANN-att method is higher than other methods by 6% for GLODAP v2.2022 and 30% for MOSAIC v.2.0. Additionally, the prediction maps of surface ocean dissolved oxygen and $\Delta^{14}C$ in the West Pacific demonstrate that ANN-att has a significant advantage in predicting marine biogeochemistry data with stronger nonlinear characteristics.

Plain Language Summary The supplementation of marine biogeochemistry data enables a more comprehensive reflection of marine ecological and biogeochemical processes, thus supporting deeper scientific research. Many studies have used machine learning methods to supplement marine biogeochemistry data, but the prediction accuracy of existing methods still needs to be improved. In this study, we propose a new deep learning method for predicting the marine biogeochemistry data. This method can capture the important data features effectively, and in GLODAP v2.2022 and MOSAIC 2.0 data sets, compared with RF and ANN methods, our method has an average improvement of 6% and 30%, respectively. This method provides a new solution for the development of marine big data and biogeosciences.

Key Points:

- A new machine learning method called ANN-att has been developed to predict marine biogeochemistry data
- ANN-att method demonstrated high prediction accuracy across multiple ocean biogeochemistry data sets
- ANN-att method effectively captured the environmental factors influencing surface ocean dissolved oxygen and sedimentary organic carbon $\Delta^{14}C$ values in the Western Pacific Ocean

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

R. Bao,
baorui@ouc.edu.cn

Citation:

Liu, M., Wang, Y., Zhong, G., Liu, Y., Liu, X., Shi, J., et al. (2025). A deep learning approach of artificial neural network with attention mechanism to predicting marine biogeochemistry data. *Journal of Geophysical Research: Biogeosciences*, *130*, e2024JG008386. https://doi.org/10.1029/2024JG008386

Received 31 JUL 2024
Accepted 13 FEB 2025

Author Contributions:

Conceptualization: Rui Bao
Data curation: Mingzhi Liu
Formal analysis: Mingzhi Liu, Yipeng Wang, Rui Bao
Investigation: Mingzhi Liu
Methodology: Mingzhi Liu
Project administration: Rui Bao
Software: Mingzhi Liu
Supervision: Rui Bao
Validation: Mingzhi Liu, Rui Bao
Visualization: Mingzhi Liu, Rui Bao
Writing – original draft: Mingzhi Liu, Yipeng Wang, Guoqiang Zhong, Yongxin Liu, Xiaoqing Liu, Jifan Shi, Rui Bao
Writing – review & editing: Mingzhi Liu, Yipeng Wang, Yongxin Liu, Xiaoqing Liu, Jifan Shi, Yangli Che, Rui Bao

© 2025. American Geophysical Union. All Rights Reserved.

1. Introduction

Marine biogeochemistry data will help scientists gain a more comprehensive understanding of elemental changes in the marine environment and provide data support for the development of biogeosciences. For example, climate change and ecosystem stability in oceans can be predicted through the analysis of ocean hydrological (e.g., temperature and salinity) and biological data (Archer et al., 2004; Helm et al., 2010; Peralta & Morison, 2010). However, the measurement of some important marine biochemical data is often limited by technology and cost, making it difficult to obtain such data directly or within a short period of time, such as the radiocarbon (^{14}C) data. Therefore, supplementation of marine biochemical data is essential for data analysis and mining.

Big data prediction is one of the effective methods to supplement marine biogeochemistry data (Ahmad, 2019; Lou et al., 2023; Song et al., 2023), that is, predicting unknown data based on the relationship between known data. Big data prediction mainly contains statistical methods and machine learning methods. Statistical methods can usually construct simple linear relationships of marine biogeochemistry data (Egger et al., 2018), whereas the relationships between marine biochemical data are mostly nonlinear, the prediction results of this method have always large errors. Machine learning methods have been shown to be effective in predicting nonlinear biogeochemical relationships (Jordan & Mitchell, 2015; Lee et al., 2019; Chen et al., 2022), whereas random

forest (RF) (Atwood et al., 2020; Wu et al., 2024; Zhang et al., 2022) and artificial neural networks (ANN) (Agatonovic & Beresford, 2000; Liu & Luo, 2021) are widely applied in Earth sciences (Figure S1 in Supporting Information S1). The two methods can be used not only for predicting seawater but also for forecasting marine sediment biogeochemistry data. For example, Cui et al. (2023) used the RF method to predict global sea surface temperature data induced by tropical cyclones in the western North Pacific. Atwood et al. (2020) predicted global high-resolution marine sedimentary carbon reserves using the RF method. Paradis et al. (2024) used RF method to predict parameters such as total organic carbon (TOC), $\delta^{13}C$, and $\Delta^{14}C$ data of sediments in the western Pacific Ocean. Although the RF method is computationally efficient and runs quickly, it is prone to overfitting when the marine biogeochemistry data are concentrated in a specific region, which can lead to a decrease in accuracy (Song et al., 2023). Additionally, the findings of Paradis et al. (2024) demonstrate that less observational data (i.e., at the upper and lower ends of the variable ranges) can also result in reduced accuracy for the RF method. ANN (Tian et al., 2022), as another method to predict marine biogeochemistry data, can model nonlinear relations in arbitrary data. It is particularly effective in mitigating the influence of "abnormal" marine biogeochemistry data, which refers to data points that significantly deviate from expected trends due to factors such as measurement errors, environmental anomalies, or infrequent events that are not representative of typical marine conditions (Parameswaran et al., 2024). For example, Weber et al. (2019) predicted the global dispersed ocean-atmospheric methane fluxes, and their results showed that the prediction accuracy of ANN is basically comparable to that of RF. The prediction results of Liu et al. (2021) showed that the prediction accuracy of ANN is higher than that of RF in high concentration of particulate organic carbon (POC). Although ANN can model relationships between arbitrary data, it does not consider the impact of each input data on the output data. By understanding the influence of each input on the prediction, we can effectively leverage strong correlations while minimizing the impact of irrelevant factors, thus improving the prediction accuracy of the ANN.

Accurate prediction of marine biogeochemical data is crucial for understanding complex environmental processes. To further enhance prediction accuracy, there is an urgent need for the development of a new method capable of providing high accuracy for diverse data sets. Attention mechanism is a widely used data processing method in machine learning, applied across various machine learning tasks (Huang & Narayanan, 2017; Niu et al., 2021; Santana et al., 2022). The attention mechanism is inspired by the way humans process information (Liu et al., 2021). In cognitive science, attention is a selective process where the brain prioritizes certain stimuli or pieces of information based on their relevance, significance, or salience. This enables humans to focus on the most important aspects of a task improving efficiency and decision-making. Similarly, in machine learning, the attention mechanism allows a model to focus on the most relevant features of the input data while disregarding less important or redundant information. Recently, Ling et al. (2022) found that the attention mechanism is effective in selecting relevant information from the data, which in turn improves prediction accuracy. Attention mechanisms improve the accuracy and efficiency of the method by assigning learned weights to input features, selectively focusing on key parts of the data while reducing the influence of irrelevant features (Vaswani et al., 2017). This approach is also effective for predicting marine biogeochemistry parameters, where data sets are often characterized by heterogeneity and complex nonlinear relationships (Xie et al., 2022). For example, Zhi et al. (2024) elaborated that the attention mechanism can model complex relationships between data when predicting water quality. The dynamic selectivity of the attention mechanism would thus provide us with a new solution for high prediction accuracy for arbitrary data.

Our objective is to develop a new machine learning method capable of accurately supplementing marine biogeochemistry data with complex nonlinear relationships. Specifically, we propose an artificial neural network enhanced with an attention mechanism (ANN-att) to improve prediction accuracy and effectively capture the intricate patterns within the data. ANN-att uses the novel attention mechanism to improve the attention to important parameters and combines it with ANN methods. We compare and evaluate the performance of RF, ANN, and ANN-att using two widely used ocean data sets in marine biogeochemistry: Global Ocean Data Analysis Project (GLODAP v2.2022) (https://www.ncei.noaa.gov, Lauvset et al., 2022) and the Modern Ocean Sediment Archive and Inventory of Carbon (MOSAIC v.2.0) (https://sarah-paradis.github.io/MOSAIC, Paradis et al., 2023). Our experimental results demonstrate that the ANN-att significantly enhances the accuracy of predicting marine biogeochemistry data, providing a robust solution for handling complex data sets.

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. Materials and Methods

2.1. Data Description

The reasons for using the GLODAP v2.2022 and MOSAIC v.2.0 data sets are as follows:

1. These two data sets cover vast ocean and sediment areas, respectively, which provide a large number of data with diverse characteristics for machine learning.
2. Both data sets are open and continuously updated. Using these two data sets can provide experience and foundation for future research.

GLODAP v2.2022 provides a periodic compilation of seawater inorganic carbon chemistry parameters and related variables through chemical analysis of seawater samples. The data set collected measurements of nearly 1.4 million seawater samples from the global ocean collected from 1,085 cruises from 1,972 years to 2,021 years and contained 45 related variables. It includes 20,000 surface ocean samples (above 100 m) and 1.38 million water column samples with the water column data spanning depths from 100 m to 9,700 m. We excluded variables with more than 70% missing data and removed all missing data from the remaining data set. The abbreviations for parameters used in this study are referred to as Table S1 in Supporting Information S1. We considered that the performance of machine learning predictions is determined by both accuracy and generalization ability (the performance in regions with fewer observational data). Therefore, we did not use the entire GLODAP v2.2022 data set. Instead, we used a random selection approach to distribute the data across different areas, avoiding concentration in a single region, which could negatively affect the evaluation of the model's generalization. The Pacific region was selected due to its wide spatial heterogeneity and biogeochemical significance, including areas such as the equatorial upwelling zone and subtropical gyres, which represent a broad spectrum of marine biogeochemical variability.

MOSAIC v.2.0 is an organized and unified organic carbon database established to understand the temporal and spatial changes of organic carbon content, sources, and dynamics in marine sediments (Paradis et al., 2023). The main parameters contain latitude and longitude, water depth, sample depth, total organic carbon (TOC), total nitrogen (TN), organic nitrogen, carbon to nitrogen ratio (C/N), calcium carbonate, silicon dioxide, $\delta^{13}C$, and $\Delta^{14}C$. In MOSAIC v.2.0, the data for parameters such as organic nitrogen, calcium carbonate, and silicon dioxide contain a significant number of missing values, which are due to either unavailable measurements or the lack of relevant records. Therefore, these parameters will not be considered in this study. The parameters used in MOSAIC v.2.0 are shown in Table S2 of Supporting Information S1. We selected 232 sediment cores from the western Pacific Ocean along the continental margin all with sediment depths of less than 15 cm. These data simultaneously contain data for TOC, TN, C/N, $\delta^{13}C$, and $\Delta^{14}C$ and similarly divided into training sets, validation sets, and test sets in a ratio of 6:2:2.

2.2. Methods

2.2.1. Target and Input Variables

To comprehensively test the predictive performance of machine learning methods, we used multiple variables from the GLODAP v2.2022 and MOSAIC v2.0 data sets as target variables. The temperature/theta and salinity are critical factors influencing other environmental parameters. Our study focused on predicting more complex biogeochemical variables that are influenced by a range of factors beyond temperature and salinity, so we did not include them as target variables. The water depth/pressure is also not considered as a target variable in our study because of its limited significance in predicting biogeochemical data. When using the GLODAP v2.2022 data set for testing, we selected 14 variables, excluding latitude, longitude, temperature, theta, salinity, water depth and pressure, as target variables, whereas the remaining variables were used as input variables. For the MOSAIC v2.0 data set, we selected 5 variables, excluding latitude, water depth, and sediment sample depth, as target variables with the remaining variables used as input variables.

Machine learning methods utilize the training set to learn data features, the validation set to select optimal model parameters, and the test set for performance evaluation. To ensure that the method can learn sufficient features, the training set typically comprises a larger proportion of the data. If the validation and test sets are too small, it becomes challenging to accurately assess the model's performance. Therefore, we divided the data sets into training, validation, and test sets in a 6:2:2 ratio (Liu et al., 2023).

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.2.2. Random Forest (RF) and Artificial Neural Networks (ANN)

RF is a type of ensemble learning method in machine learning that is an optimized version of Bagging based on a tree model, where Bagging is a prominent ensemble method for training a single learner on an arbitrarily sampled subset of the original training data set (Ngo et al., 2022). It is effective in handling both classification and regression problems (Jafarzadeh et al., 2021). RF utilizes decision trees (Suthaharan, 2016) as its fundamental building blocks and makes decisions by aggregating the outputs of multiple independent decision trees through taking the mode or averaging. Compared to a single decision tree, RF demonstrates higher accuracy and generalization ability due to its use of the Bagging method to sample and construct multiple decision trees, thereby reducing sampling error and mitigating overfitting to some extent (Pothuganti, 2018). RF offers several advantages: first, it is straightforward to implement with low computational overhead, delivering robust performance and producing accurate results without the need for complex parameter tuning. Second, RF can handle high-dimensional input features without requiring dimensionality reduction and provide direct insight into the importance of each variable. Lastly, the training of RF method can be highly parallelized, making it advantageous in the era of big data.

ANN is an algorithmic method that mimics the functionality of a biological neural network, primarily utilized for feature learning and pattern recognition in data analysis (Abiodun et al., 2018). The fundamental principle of ANN involves emulating the interconnection and information transmission between neurons in the human brain. Multiple artificial neurons, also known as nodes, communicate with each other by exchanging signals through connection weights, enhancing the capacity to model nonlinear data through activation functions (Chen & Hu, 2017; Geiger et al., 2013). In instances where there is a significant disparity between the method output and the actual value, ANN adjusts the network weight using the back propagation algorithm (BP algorithm; LeCun et al., 1988; Cilimkovic, 2015) to minimize the error. The BP algorithm is iteratively applied until a predetermined number of training rounds or other stopping criteria are satisfied. ANN can find widespread application across various domains, boasting several advantages: (a) theoretically, ANN can model any intricate nonlinear data (Almeida, 2002). (b) ANN exhibits strong robustness and generalization (Chen & Hu, 2017); it can effectively model and predict even in the presence of noise in the original data. Despite its notable advantages, the limitations of ANN cannot be overlooked. First, due to its complex structure and numerous parameters, ANN necessitates a substantial amount of data for support. In scenarios with limited data, the method is prone to overfitting, resulting in a significant decrease in accuracy. Second, training the method is challenging, as the final outcome may converge to a local optimal solution rather than a global optimal solution, requiring continuous parameter adjustments and experimentation. This challenge is a key factor contributing to the lower prediction accuracy of ANN compared to RF. Although ANN can be applied to a wide range of data, achieving optimal results is challenging due to its inherent structure and other associated factors. Consequently, our research focuses on enhancing the predictive performance of ANN.

2.2.3. Artificial Neural Network With Attention Mechanism (ANN-att)

By dynamically assigning different weights to input data, the attention mechanism enables machine learning methods to focus on the most relevant features, thereby enhancing the expressiveness and generalization capabilities of method. Essentially, the attention mechanism functions as a weighting mechanism that creates new feature representations through the weighting of input data.

In this study, we propose an "ANN-att" method that integrates an attention mechanism with an artificial neural network (ANN), where the weights of the method are adjusted during training to prioritize important features and enhance prediction accuracy as illustrated in Figure 1. The method comprises three components: input, attention, and output. The input parameters are typically known and can have one-dimensional or multidimensional dimensions. The attention component selects the parameter information of method with thicker lines indicating a higher level of information selection. The output component presents the prediction results of the ANN-att method.

For inputs with M parameters, denoted as F^M, the workflow of the method is as follows: the input features are fully connected with the neurons in the first layer (blue circle, Figure 1), and its mathematical expression is

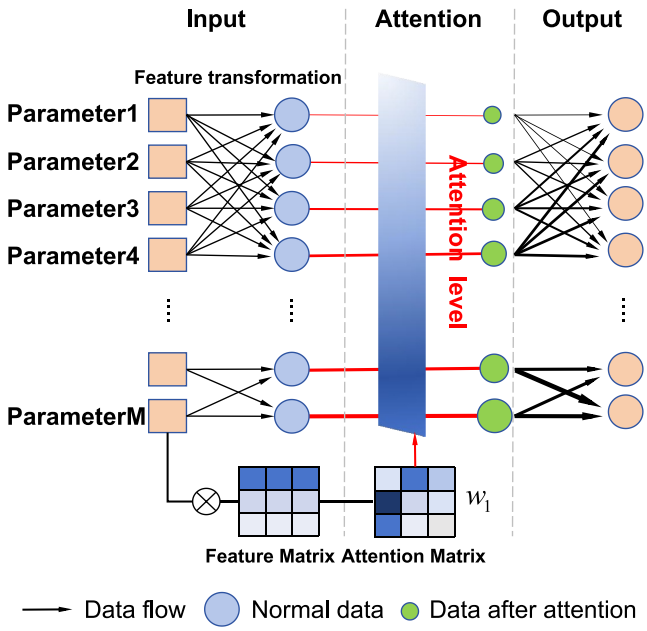

Figure 1. The framework of the ANN-att method.

$$H_1 = F^M \otimes W^1_{MN} \tag{1}$$

where W^1_{MN} represents the weights of the first-layer neurons with a parameter size of $M \times N$, $\otimes$ denotes the symbol of matrix product operation, and H_1 is the output of the first layer with a dimension corresponding to N, the number of neurons in the first layer. In our paper, when the training data set is GLODAP v2.2022, M equals 20; when the data set is MOSAIC v2.0, M equals to 8. The output results are further selected by the attention mechanism.

$$\mathrm{AH}_1 = H_1 \cdot w_1 \tag{2}$$

where w_1 is the attention matrix formed by combining input parameters with the parameter matrix, and its value ranges from 0 to 1. $\cdot$ is the matrix inner product operator. AH_1 is the output of attention mechanism, and its dimension is N. The method computes the k-layer output, AH_k, by iteratively using the output of the previous layer as the input for the next layer of neurons.

$$AH_k = H_{k-1} \cdot w_k \tag{3}$$

where $k \geq 2$. The prediction result of the output part is

$$Y = AH_k \otimes w^N \tag{4}$$

where w^N is the parameter of the output layer and Y is the prediction result. During the aforementioned process, the method performs an operation; however, the predicted outcome may deviate significantly from the actual value (due to parameter randomization). To minimize the disparity between the predicted and actual values, the method undergoes optimization through a gradient descent algorithm. After a certain number of repetitions of the aforementioned operations, the disparity reaches a minimum, and the method at this point is considered the trained method. The trained method can then be utilized for making predictions. The parameters required for the method training process are shown in Text S1 and Figure S2 of Supporting Information S1.

2.2.4. Training and Vailding the Methods

The predictive results of RF, ANN, and ANN-att are all determined by their respective parameters (Figures S3–S5 in Supporting Information S1). The RF method establishes the relationship between inputs and outputs through the number and depth of the "trees," whereas ANN and ANN-att do so via the parameters of the hidden layers (Rafter et al., 2019). To find the optimal parameters that bring the predictions closer to the target results, the training set was used to train all three models, employing an Adam-based parameter search method. During training, the validation set was used to assess the generalization of each method (to ensure that the models were not solely effective on the training set and could perform on other data) and to avoid overfitting (indicated by a gradual decline in prediction accuracy on the validation set). After training on the training set and validating with the validation set, the optimal parameter combinations for the current data set were determined for all three methods. We then used a test set, which none of the methods had previously "learned" to validate the models and assess the accuracy of their predictions. In this study, we used the coefficient of determination (R^2) to evaluate the predictive accuracy of all three methods during both the training and validation phases. R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

where y_i represents the true values of the target variable, $\bar{y}$ denotes the mean of all target variables, $\hat{y}$ refers to the predicted results from different methods, and n signifies the number of predictions. The closer the R^2 value is to 1, the more accurate the predictions are.

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

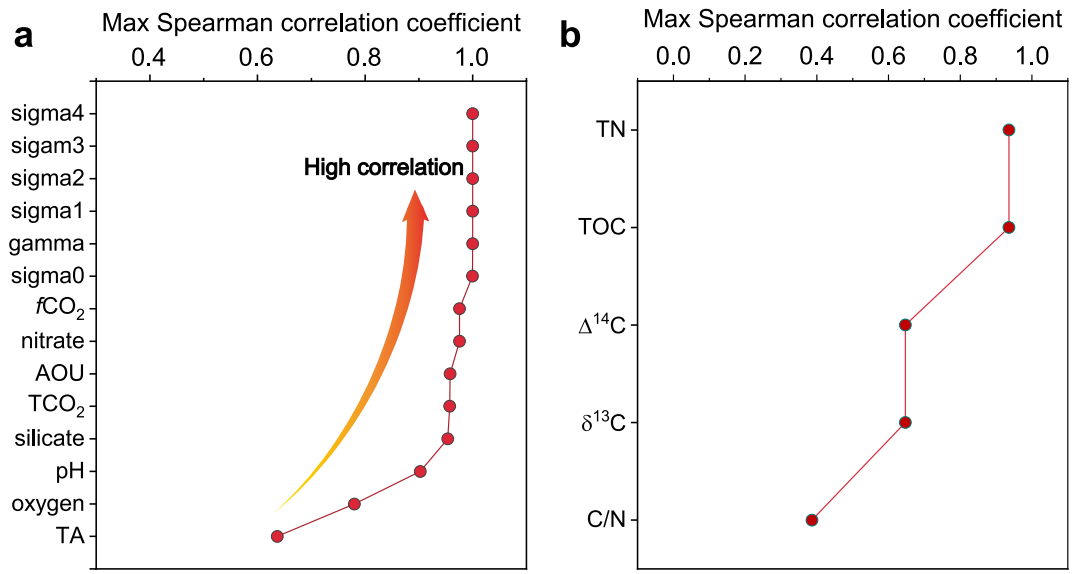

Figure 2. Maximum Spearman correlation corresponding to different variables. (a) Maximum Spearman correlation for each variable in the GLODAP v2022 data set. (b) Maximum Spearman correlation for each variable in the MOSAIC v.2.
0 data set.

2.2.5. Spearman Correlation Analysis

Spearman correlation analysis is a data analysis method based on statistical data level, which is used to detect the linear relationship between two variables (Ali Abd Al-Hameed, 2022; Winter et al., 2016; Xiao et al., 2016). Spearman correlation coefficient (*Spcc*, Stirnimann et al., 2024) is an important index in Spearman analysis, which can be used to measure the strength of linear relationships between data A and B, and its calculation is based on formula 6.

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

where d represents the difference in the ranks of the data A and B when calculating the Spearman correlation coefficient, and n denotes the total number of observed samples in the data sets.

3. Results

3.1. Maximum Spearman Correlation of Variables

Figure 2 presents the maximum Spearman correlation coefficients between different variables in the GLODAP v2022 and MOSAIC v2.0 data sets. In the GLODAP v2022 data set, oxygen and TA show the lowest maximum Spearman correlation coefficient ($Spcc = 0.75$), whereas sigma3 and sigma4 ($Spcc = 0.99$) have the highest correlations (Figure 2a). This indicates that oxygen and TA exhibit the strongest nonlinear relationships with other parameters. In the MOSAIC v2.0 data set, TOC and TN have the highest Spearman correlation coefficient ($Spcc = 0.935$), whereas C/N shows the lowest correlation. For machine learning methods, predicting variables with stronger nonlinear relationships tends to be more challenging compared to those with weaker correlations, which makes it harder to achieve higher prediction accuracy.

3.2. Methods Evaluation for Various Target Variables

To evaluate the performance of different methods in predicting marine biogeochemistry data, we tested three methods using the GLODAP v2.2022 and MOSAIC v2.0 data sets in both marine water column and sediment environments, respectively. A common method to evaluate the predictive performance of a method is to assess the deviation between the predicted results and the actual values, such as the coefficient of determination (R^2; Tao et al., 2023; C. Wang et al., 2024, Z. Wang et al., 2024).

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

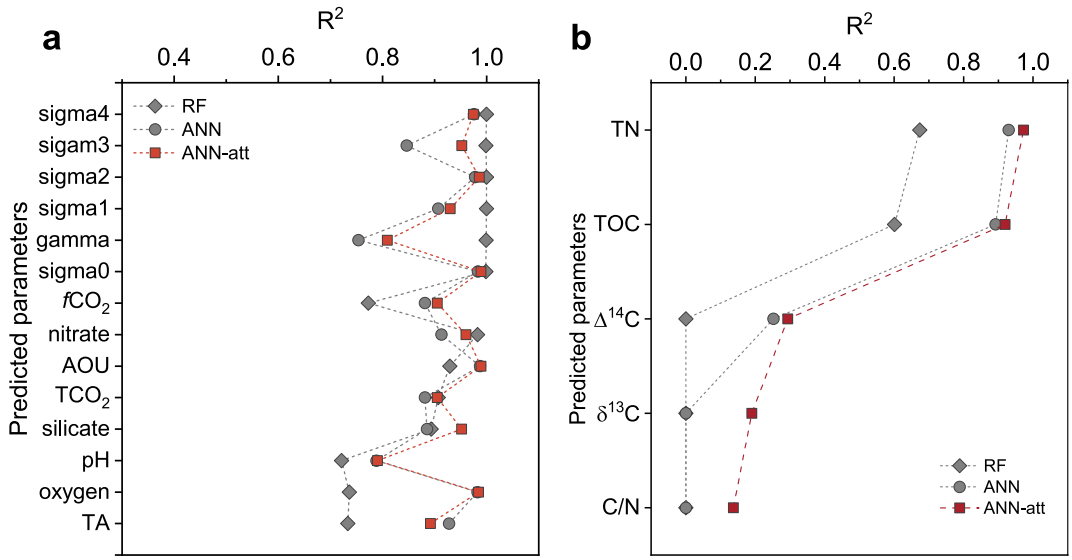

Figure 3. The prediction results of RF, ANN, ANN-att. (a, b) show the R^2 evaluation results for the three methods in the GLODAP v2022 data set and the MOSAIC v.2.0 data set, respectively.

We evaluate the predictive performance of ANN-att, ANN, and RF on marine water column environments using R^2 based on the selected and filtered GLODAP v2022 data set (n = 3,200). Our results show that the R^2 of ANN-att improved by 2.4% and 2.3% compared with ANN and RF, respectively, indicating that the predictive results of ANN-att are more accurate when the prediction targets are total alkalinity (TA), oxygen, nitrate, fCO_2, AOU, TCO_2, pH, and silicate. The R^2 of ANN-att decreased by 0.2% compared with RF when the prediction targets are gamma, sigma0, sigma1, sigma2, sigma3, and sigma4 (Figure 3a). Several factors contribute to the lower prediction accuracy of ANN-att compared to RF. We have provided a detailed discussion of these factors in Section 4.1. After evaluating the predictive capabilities of the three methods on a water column environments, we further assessed their performance on a marine sediment environments data set (n = 162) based on the MOSAIC v.2.0 data set. Our results indicate that the R^2 of ANN-att improved by 20.9% and 9.8% compared with ANN and RF on the MOSAIC v.2.0 data set, respectively (Figure 3b). We further evaluated the prediction capabilities of ANN-att by mapping the prediction results. The detailed results of evaluation for mapping were shown in Text. S2, Figures S6 and S7 of Supporting Information S1. In summary, these results suggest that ANN-att achieved the highest accuracy in predicting marine biogeochemistry data.

3.3. Predicted Values of Global Surface Ocean (<10 m) Dissolved Oxygen and Marginal Sea Sediment $\Delta^{14}C$ in Western Pacific Ocean

The above experiments show that ANN-att has significant advantages in predicting marine biogeochemistry data. To evaluate the generalization ability of ANN-att in predicting marine chemical data, particularly its accuracy in regions with limited observational data, we selected dissolved oxygen and $\Delta^{14}C$, both of which exhibit stronger nonlinear characteristics (Figure 2), from the GLODAP v2022 data set (n = 3,200) and the MOSAIC v2.0 data set (n = 232) as the prediction targets, respectively. Specifically, due to the limitations in the data ranges within these data sets, we could only select a subset of the original input variables for the prediction.

When predicting dissolved oxygen, we selected surface ocean (<10 m) dissolved oxygen as the target variable and latitude, longitude, salinity, and temperature, pH, nitrate and silicate as input variables, as these factors are known to be related to surface dissolved oxygen (Shao et al., 2023; C. Wang et al., 2024, Z. Wang et al., 2024). In the ANN-att method, the predicted results show a gradual increase in dissolved oxygen from high-temperature regions to low-temperature regions (Figure S8 in Supporting Information S1) with the highest prediction accuracy (R^2 = 0.38, Figure S9a in Supporting Information S1). This result once again demonstrates the obvious advantage of ANN-att in predicting the variables of nonlinear relationships. In the RF method, the predicted dissolved oxygen does not align with the "temperature-dissolved oxygen" theory (Figure 4b), particularly in the Atlantic region, and exhibits the lowest accuracy (R^2 = 0.05, Figure S9b in Supporting Information S1). In the ANN

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

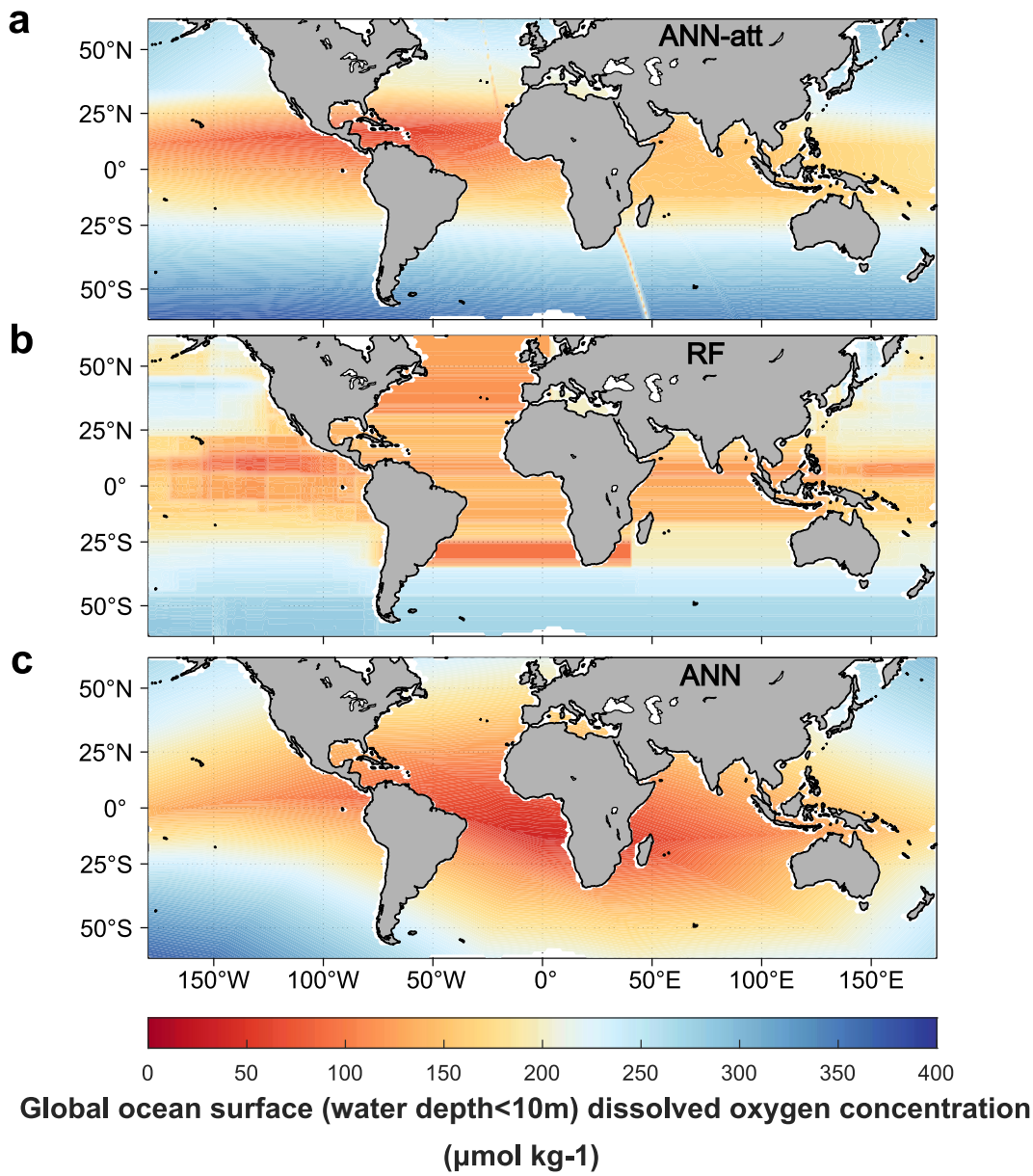

Figure 4. Global ocean surface (water depth<10 m) dissolved oxygen concentration (μmol kg^{-1}): (a–c) represent the mapping plots of dissolved oxygen predictions from ANN-att, RF, and ANN methods, respectively. The maps consist of grids with a resolution of 1° × 1°, where each grid represents a predicted dissolved oxygen value.

method, the predicted dissolved oxygen aligns with the "temperature-dissolved oxygen" theory (Figure 4c), but it underestimates oxygen concentration compared to ANN-att. All three methods significantly underestimate dissolved oxygen, and overall prediction accuracy remains low (Figure S9c in Supporting Information S1).

When predicting $\Delta^{14}C$, we chose the western Pacific Ocean as the target region for prediction due to the dispersed distribution of $\Delta^{14}C$ of organic carbon in marine sediment within the MOSAIC v.2.0 data set (Figure 5a). Due to the limited availability of $\delta^{13}C$ and TN data in the South China Sea, we used latitude, longitude, depth, and TOC as input variables. The prediction results from ANN-att indicate that the distribution of $\Delta^{14}C$ values aligns closely with depth (Figure 5b), and the predicted values differ significantly from observed $\Delta^{14}C$ data ($R^2 = 0$, Figure S10a in Supporting Information S1). The predictions of RF show that $\Delta^{14}C$ values are more negative in the Yellow River and Yangtze River shelf areas (Figure 5b; −600‰ to −450‰) and are the closest to observed $\Delta^{14}C$ values ($R^2 = 0.568$, Figure S7b in Supporting Information S1). However, the RF predictions in the South China Sea region

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

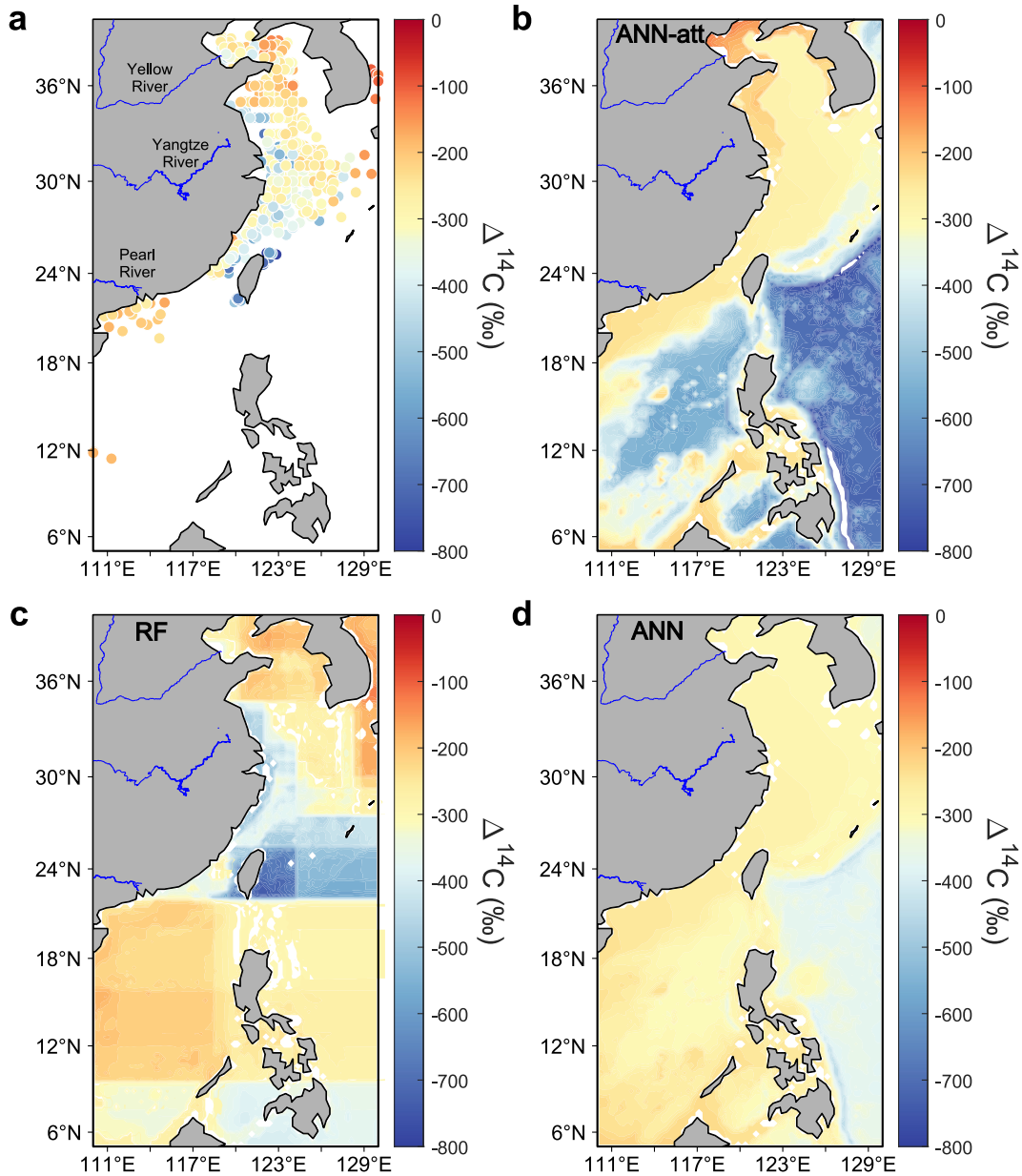

Figure 5. The values of $\Delta^{14}C$ (‰) of OC in the western Pacific Ocean: (a) shows the distribution of observed $\Delta^{14}C$; (b–d) represent the mapping plots of $\Delta^{14}C$ predictions from ANN-att, RF, and ANN, respectively. The maps consist of grids with a resolution of 0.25° × 0.25°, where each grid represents a predicted $\Delta^{14}C$ value.

appear very coarse. ANN predictions suggest that $\Delta^{14}C$ values in the Bohai Sea, Yellow Sea, and East China Sea generally range between −400‰ and −300‰, whereas in the South China Sea, values range from −450‰ to −350‰ (Figure 5c). Similarly, the ANN predictions also show significant discrepancies compared to observed $\Delta^{14}C$ data ($R^2 = 0.056$, Figure S10c in Supporting Information S1). All three methods demonstrate considerable differences between predictions and actual observations in regions with sparse data (Paradis et al., 2024).

4. Discussions

4.1. Which Methods Are More Accurate?

Our preliminary experiments validated that ANN-att possesses strong predictive capabilities. However, when the target variables were gamma, sigma0, sigma1, sigma2, sigma3, and sigma4, the prediction accuracy of ANN-att

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

was lower than that of RF. This indicates that the high predictive accuracy of ANN-att has certain limitations. Based on the maximum Spearman correlation results, we found that gamma, sigma0, sigma1, sigma2, sigma3, and sigma4 can be calculated from temperature and salinity (Yu et al., 2022), meaning these variables exhibit weaker nonlinear characteristics (Figure 2a). In contrast, variables such as TA, oxygen, nitrate, fCO_2, AOU, TCO_2, pH, and silicate demonstrate stronger nonlinear characteristics due to the influence of various oceanic factors. For instance, C. Wang et al. (2024), Z. Wang (2024) found that oxygen is influenced by ocean currents and biological factors and shows significant spatial and temporal variability. This suggests that the advantage of ANN-att diminishes when predicting variables with weaker linear relationships. We hypothesize that the attention mechanism, while providing enhanced data analysis capabilities, may also complicate the analysis of simpler data, a phenomenon we refer to as "overcomplicating simple problems."

To investigate whether this phenomenon is widespread, we applied binning to the raw data (Rafter et al., 2019) and selected the portion with the highest data (depth <10 m, $n = 210$) as the new training data sets. Binning creates new data features to validate the models and is commonly used in oceanographic data analysis. The results confirmed that the "overcomplicating simple problems" led to lower prediction accuracy of ANN-att compared to RF and ANN when predicting variables with stronger linear relationships (Figure S11 in Supporting Information S1). When testing with the MOSAIC v2.0 data set, as the linearity of the target parameters increased, the prediction results of ANN and ANN-att became increasingly similar, indicating that the "overcomplicating simple problems" effect persists regardless of the data environment.

In summary, ANN-att performs better when predicting data with stronger nonlinear relationships followed by ANN. When predicting variables with weaker nonlinear relationships, RF and ANN show superior performance.

4.2. Limitations of the Methods in Application

The evaluation results indicate that ANN-att, RF, and ANN performed poorly in predicting dissolved oxygen and $\Delta^{14}C$. Although all three methods underestimated dissolved oxygen values, both ANN-att and ANN successfully captured the relationship between temperature and oxygen fluctuations. This suggests that the poor prediction results may be influenced by the data range. To further investigate, we used all Pacific data as training data and re-predicted global surface dissolved oxygen. The results showed a significant improvement in prediction accuracy across all methods (Figure 6, Figure S12 in Supporting Information S1) with predictions better reflecting actual dissolved oxygen variations. Additionally, prediction accuracy in the Indian Ocean was higher than in the Atlantic. This can be attributed to the greater generalizability of the expanded Pacific data sets and the similarity between surface dissolved oxygen in the Pacific and Indian Oceans. In contrast, the unique characteristics of the Atlantic, such as high-salinity regions at higher latitudes, could not be captured by models trained solely on Pacific data. Furthermore, as the expanded dissolved oxygen data set exhibited stronger linearity ($Spcc = 0.985$), the ANN-att method, which incorporates attention mechanisms, showed lower prediction accuracy compared to RF and ANN.

The concentration of $\Delta^{14}C$ in sediments is influenced by multiple factors (C. Wang et al., 2024), and relying solely on depth and TOC may not yield accurate predictions. We included additional variables (Table S3 in Supporting Information S1) that could potentially influence $\Delta^{14}C$ and re-predicted $\Delta^{14}C$ in marginal sea sediments. The new data demonstrated stronger nonlinearity for $\Delta^{14}C$ ($Spcc = 0.40$). The updated results revealed a significant improvement in prediction accuracy across all three methods (Figure 7), with ANN-att achieving the highest accuracy in the training data ($R^2 = 0.893$, Figure S13a in Supporting Information S1) consistent with the findings of Bao et al. (2016). Furthermore, the ANN-att has a more refined spatial distribution of $\Delta^{14}C$. In contrast, RF exhibits a "block-like" spatial distribution of $\Delta^{14}C$ with minimal spatial variation within each block. This pattern is associated with the classification-based approach of RF, which results in predictions being heavily influenced by dominant factors. The ANN model, due to its lower prediction accuracy, shows a significant deviation from the actual spatial distribution of $\Delta^{14}C$ (Bao et al., 2016). In the region east of Taiwan (approximately between 121°E to 130°E longitude and 0 to 24°N latitude, where data is scarce), the predicted $\Delta^{14}C$ values by ANN-att ranged from −400‰ to −200‰ (Figure 7b), RF from −500‰ to −400‰ (Figure 7c), and ANN from −200‰ to 0‰ (Figure 7d). The significant discrepancies in prediction results among the three methods in this region are likely due to limitations in data generalizability. The training data of $\Delta^{14}C$ used in this study primarily represent shallow water environments (0–200 m), whereas the region east of Taiwan consists mostly of deep water (average depth around 4,000 m). Organic carbon burial mechanisms may differ between shallow and deep

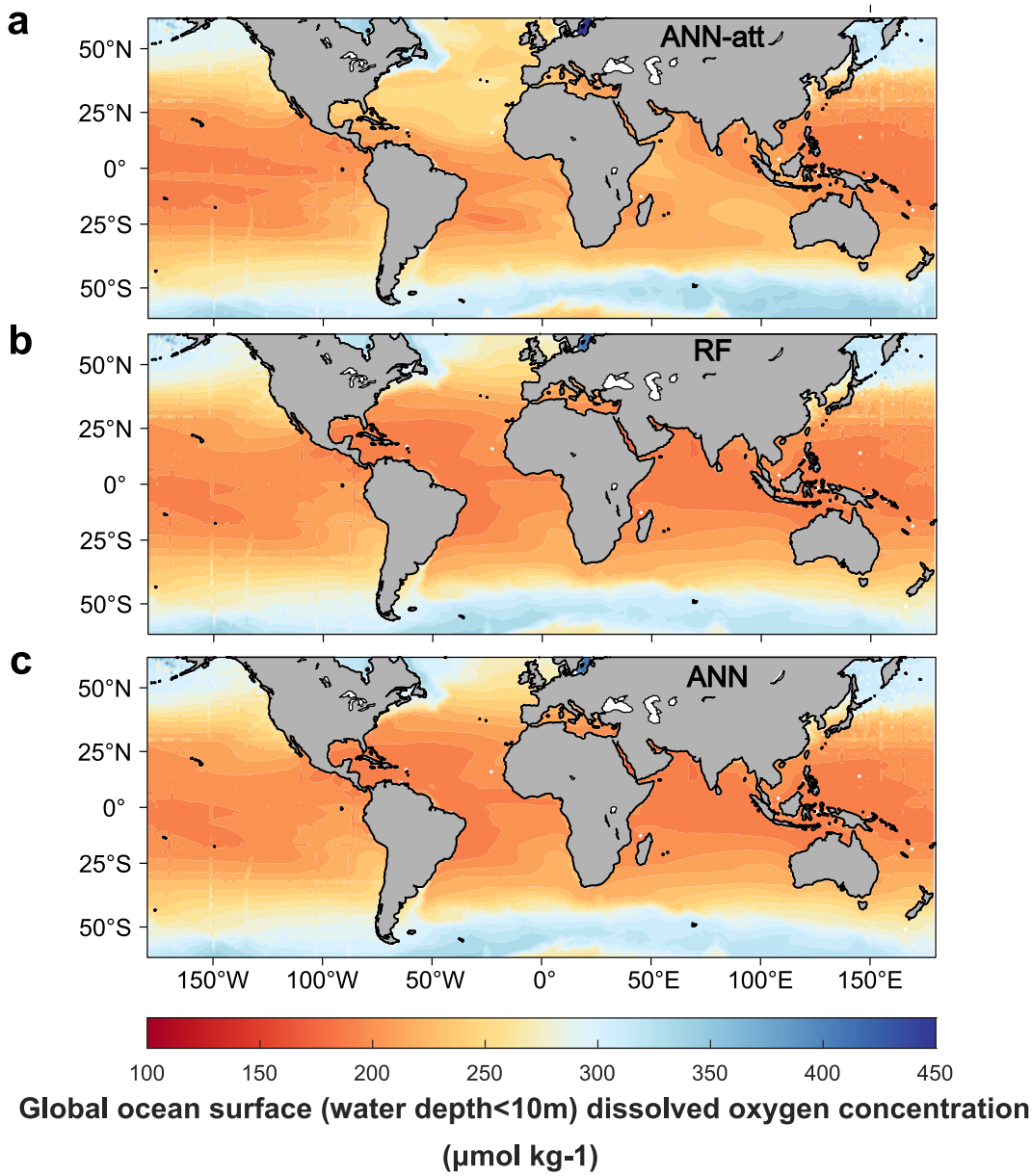

Figure 6. Global ocean surface (water depth<10 m) dissolved oxygen concentration (μmol kg^{-1}) based on the entire Pacific data set: (a–c) represent the mapping plots of dissolved oxygen predictions from ANN-att, RF, and ANN, respectively. The maps consist of grids with a resolution of 1° × 1°, where each grid represents a predicted dissolved oxygen value.

waters. By incorporating additional relevant factors, such as terrestrial inputs and hydrodynamic processes, we anticipate achieving more reliable predictions in this area.

The above analysis further demonstrates that ANN-att has a clear advantage when predicting data with strong nonlinear relationships. Additionally, the distribution patterns and correlations of the input data play a crucial role in influencing prediction accuracy. Input data with stronger generalization properties can effectively improve the overall prediction accuracy.

5. Conclusion

This study proposes a novel deep learning method, ANN-att, to improve the accuracy of predicting marine biogeochemistry data. The results show that the prediction of ANN-att improved the R^2 by 2.4% and 2.3%

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

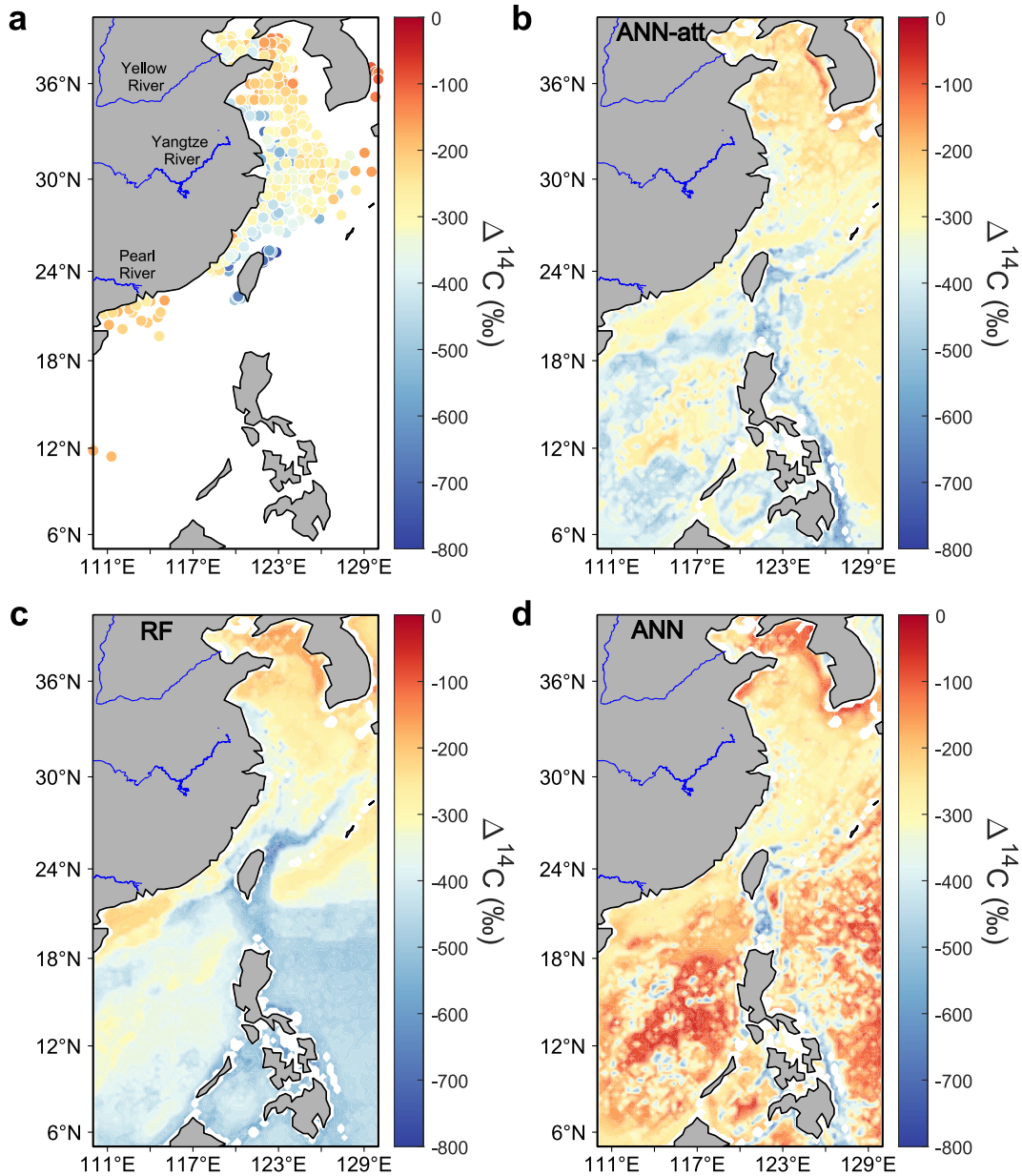

Figure 7. The values of $\Delta^{14}C$ (‰) of OC in the western Pacific Ocean based on more input variables (pH, chlorophyll, salinity, etc.) compared to Figure 5: (a) shows the distribution of observed $\Delta^{14}C$ values; (b–d) represent the mapping plots of $\Delta^{14}C$ predictions from ANN-att, RF, and ANN methods, respectively. The maps consist of grids with a resolution of 0.25° × 0.25°, where each grid represents a predicted $\Delta^{14}C$ value.

compared to RF and ANN in GLODAP v2022 (n = 3,200) and increased by 20.9% and 9.8% in MOSAIC v.2.0 (n = 232). The experimental results demonstrate that the ANN-att method significantly improves prediction accuracy in both marine water and sediment environments. Notably, when the target variable exhibits strong nonlinear characteristics, the prediction accuracy of ANN-att increases more substantially, which may be influenced by the "overcomplicating simple problems" effect introduced by the attention mechanism. By predicting global surface ocean dissolved oxygen concentration and $\Delta^{14}C$ data in sediments, we further found that the distribution patterns and correlations of the input data play a crucial role in influencing prediction accuracy. Input data with stronger generalization capabilities can effectively enhance overall prediction performance. Additionally, incorporating existing biogeochemical knowledge has been shown to improve prediction accuracy.

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In future work, we plan to explore combining artificial intelligence methods with established biogeoscience approaches to further increase the reliability of predictions.

Conflict of Interest

We comment that when using data generated by models, it is essential to adhere to academic standards and ethical guidelines, ensuring proper citation of original data and relevant literature. This not only respects intellectual property rights but also enhances the transparency and credibility of research.

Data Availability Statement

Data and code of the paper have been upload to a Zenodo repository [https://doi.org/10.5281/zenodo.14015992].

Acknowledgments
This work is funded by the National Natural Science Foundation of China (42450135 and 42450192), the Fundamental Research Funds for the Central Universities (202042010), the Taishan Young Scholars (tsqn202103030), and Shandong Province Natural Science Foundation (ZR2021JQ12).

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of Artificial Neural Network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1

Ahmad, H. (2019). Machine learning applications in oceanography. *Aquatic Research*, *2*(3), 161–169. https://doi.org/10.3153/ar19014

Ali Abd Al-Hameed, K. (2022). Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, *13*(1), 3249–3255.

Almeida, J. S. (2002). Predictive non-linear modeling of complex data by artificial neural networks. *Current Opinion in Biotechnology*, *13*(1), 72–76. https://doi.org/10.1016/s0958-1669(02)00288-4

Archer, D., Martin, P., Buffett, B., Brovkin, V., Rahmstorf, S., & Ganopolski, A. (2004). The importance of ocean temperature to global biogeochemistry. *Earth and Planetary Science Letters*, *222*(2), 333–348. https://doi.org/10.1016/s0012-821x(04)00183-9

Atwood, T. B., Witt, A., Mayorga, J., Hammill, E., & Sala, E. (2020). Global patterns in marine sediment carbon stocks. *Frontiers in Marine Science*, *7*, 165. https://doi.org/10.3389/fmars.2020.00165

Bao, R., McIntyre, C., Zhao, M., Zhu, C., Kao, S.-J., & Eglinton, T. I. (2016). Widespread dispersal and aging of organic carbon in shallow marginal seas. *Geology*, *44*(10), 791–794. https://doi.org/10.1130/g37948.1

Chen, G., Cheng, Q., Lyons, T. W., Shen, J., Agterberg, F., Huang, N., & Zhao, M. (2022). Reconstructing Earth's atmospheric oxygenation history using machine learning. *Nature Communications*, *13*(1), 5862. https://doi.org/10.1038/s41467-022-33388-5

Chen, S., & Hu, C. (2017). Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. *Remote Sensing of Environment*, *201*, 115–132. https://doi.org/10.1016/j.rse.2017.09.004

Cilimkovic, M. (2015). Neural networks and back propagation algorithm. *Institute of Technology Blanchardstown, Blanchardstown Road North*, *15*(1).

Cui, H., Tang, D., Mei, W., Liu, H., Sui, Y., & Gu, X. (2023). Predicting tropical cyclone-induced sea surface temperature responses using machine learning. *Geophysical Research Letters*, *50*(18), e2023GL104171. https://doi.org/10.1029/2023gl104171

Egger, M., Riedinger, N., Mogollón, J. M., & Jørgensen, B. B. (2018). Global diffusive fluxes of methane in marine sediments. *Nature Geoscience*, *11*(6), 421–425. https://doi.org/10.1038/s41561-018-0122-8

Geiger, E. F., Grossi, M. D., Trembanis, A. C., Kohut, J. T., & Oliver, M. J. (2013). Satellite-derived coastal ocean and estuarine salinity in the Mid-Atlantic. *Continental Shelf Research*, *63*, S235–S242. https://doi.org/10.1016/j.csr.2011.12.001

Helm, K. P., Bindoff, N. L., & Church, J. A. (2010). Changes in the global hydrological-cycle inferred from ocean salinity. *Geophysical Research Letters*, *37*(18). https://doi.org/10.1029/2010gl044222

Huang, C. W., & Narayanan, S. S. (2017). Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 583–588).

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation. *Remote Sensing*, *13*(21), 4405. https://doi.org/10.3390/rs13214405

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Lauvset, S. K., Lange, N., Tanhua, T., Bittig, H. C., Olsen, A., Kozyr, A., et al. (2022). GLODAPv2. 2022: The latest version of the global interior ocean biogeochemistry data product (version 2.2022) [Dataset]. *National Centers for Environmental Information*. https://doi.org/10.25921/1f4w-0t92

LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988). A theoretical framework for back-propagation. *Proceedings of the 1988 connectionist models summer school*, *1*, 21–28.

Lee, T. R., Wood, W. T., & Phrampus, B. J. (2019). A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Global Biogeochemical Cycles*, *33*(1), 37–46. https://doi.org/10.1029/2018gb005992

Ling, F., Luo, J. J., Li, Y., Tang, T., Bai, L., Ouyang, W., & Yamagata, T. (2022). Multi-task machine learning improves multi-seasonal prediction of the Indian Ocean Dipole. *Nature Communications*, *13*(1), 7681. https://doi.org/10.1038/s41467-022-35412-0

Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., et al. (2021). Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods. *Remote Sensing of Environment*, *256*, 112316. https://doi.org/10.1016/j.rse.2021.112316

Liu, J. W., & Luo, X. L. (2021). Research progress in attention mechanism in deep learning. *Chinese Journal of Engineering*, *43*(11), 1499–1511.

Liu, M., Liu, G., & Sun, L. (2023). Spatial-temporal dependence and similarity aware traffic flow forecasting. *Information Sciences*, *625*, 81–96. https://doi.org/10.1016/j.ins.2022.12.107

Lou, R., Lv, Z., Dang, S., Su, T., & Li, X. (2023). Application of machine learning in ocean data. *Multimedia Systems*, *29*(3), 1815–1824. https://doi.org/10.1007/s00530-020-00733-x

Ngo, G., Beard, R., & Rohitash, C. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*, *510*, 1–14. https://doi.org/10.1016/j.neucom.2022.08.055

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

Paradis, S., Diesing, M., Gies, H., Haghipour, N., Narman, L., Magill, C., et al. (2024). Unraveling environmental forces shaping surface sediment geochemical "isodrapes" in the East Asian marginal seas. *Global biogeochemistry Cycles*, *38*(4), e2023GB007839. https://doi.org/10.1029/2023gb007839

Paradis, S., Nakajima, K., Van der Voort, T. S., Gies, H., Wildberger, A., Blattmann, T., et al. (2023). The Modern Ocean Sediment Archive and Inventory of Carbon (MOSAIC) (version 2.0) [Dataset]. *Mosaic*, *15*(9), 4105–4125. https://doi.org/10.5168/mosaic019.1

Parameswaran, N., González, E., Burwicz-Galerne, E., Braack, M., & Wallmann, K. (2024). NN-TOC v1: Global prediction of total organic carbon in marine sediments using deep neural networks [preprint]. EGUsphere. https://doi.org/10.5194/egusphere-2024-1360

Peralta-Ferriz, C., & Morison, J. (2010). Understanding the annual cycle of the Arctic Ocean bottom pressure. *Geophysical Research Letters*, *37*(10). https://doi.org/10.1029/2010gl042827

Pothuganti, S. (2018). Review on over-fitting and under-fitting problems in machine learning and solutions. International journal of advanced research in electrical. *Electronics and Instrumentation Engineering*, *7*(9), 3692–3695.

Rafter, P. A., Bagnell, A., Marconi, D., & DeVries, T. (2019). Global trends in marine nitrate N isotopes from observations and a neural network-based climatology. *Biogeosciences*, *16*(13), 2617–2633. https://doi.org/10.5194/bg-16-2617-2019

SantanaColombini, C. E. L. A., & Colombini, E. L. (2022). Attention, please! A survey of neural attention models in deep learning. *Artificial Intelligence Review*, *55*(8), 6037–6124. https://doi.org/10.1007/s10462-022-10148-x

Shao, J., Huang, S., Chen, Y., Qi, J., Wang, Y., Wu, S., et al. (2023). Satellite-based global sea surface oxygen mapping and interpretation with spatiotemporal machine learning. *Environmental Science & Technology*, *58*(1), 498–509. https://doi.org/10.1021/acs.est.3c08833

Song, T., Pang, C., Hou, B., Xu, G., Xue, J., Sun, H., & Meng, F. (2023). A review of artificial intelligence in marine science. *Frontiers in Earth Science*, *11*, 1090185. https://doi.org/10.3389/feart.2023.1090185

Stirnimann, L., Bornman, T. G., Forrer, H. J., Mirkin, J., Ryan-Keogh, T. J., Flynn, R. F., et al. (2024). A circum-Antarctic plankton isoscape: Carbon export potential across the summertime Southern Ocean. *Global Biogeochemical Cycles*, *38*(1), e2023GB007808. https://doi.org/10.1029/2023gb007808

Suthaharan, S. (2016). Decision tree learning. In *Machine learning models and algorithms for big data classification: Thinking with examples for effective learning* (pp. 237–269).

Tao, F., Huang, Y., Hungate, B. A., Manzoni, S., Frey, S. D., Schmidt, M. W., et al. (2023). Microbial carbon use efficiency promotes global soil carbon storage. *Nature*, *618*(7967), 981–985. https://doi.org/10.1038/s41586-023-06042-3

Tian, T., Cheng, L., Wang, G., Abraham, J., Wei, W., Ren, S., et al. (2022). Reconstructing ocean subsurface salinity at high resolution using a machine learning approach. *Earth System Science Data*, *14*(11), 5037–5060. https://doi.org/10.5194/essd-14-5037-2022

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, C., Qiu, Y., Hao, Z., Wang, J., Zhang, C., Middelburg, J. J., et al. (2024). Global patterns of organic carbon transfer and accumulation across the land–ocean continuum constrained by radiocarbon data. *Nature Geoscience*, *17*(8), 1–9. https://doi.org/10.1038/s41561-024-01476-4

Wang, Z., Xue, C., & Bao, P. (2024). A reconstructing model based on time–space–depth partitioning for Global Ocean dissolved oxygen concentration. *Remote Sensing*, *16*(2), 228. https://doi.org/10.3390/rs16020228

Weber, T., Wiseman, N. A., & Kock, A. (2019). Global ocean methane emissions dominated by shallow coastal waters. *Nature Communications*, *10*(1), 4584. https://doi.org/10.1038/s41467-019-12541-7

Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273–290. https://doi.org/10.1037/met0000079

Wu, H., Wang, L., Ling, X., Cui, L., Sun, R., & Jiang, N. (2024). Spatiotemporal reconstruction of global ocean surface pCO_2 based on optimized random forest. *Science of the Total Environment*, *912*, 169209. https://doi.org/10.1016/j.scitotenv.2023.169209

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, *28*(14), 3866–3878. https://doi.org/10.1002/cpe.3745

Xie, W., Kimura, M., Takaki, K., Asada, Y., Iida, T., & Jia, X. (2022). Interpretable framework of physics-guided neural network with attention mechanism: Simulating paddy field water temperature variations. *Water Resources Research*, *58*(5), e2021WR030493. https://doi.org/10.1029/2021wr030493

Yu, H., Zheng, J., & Lin, Q. (2022). Strength prediction of seawater sea sand concrete based on artificial neural network in python. *Materials Research Express*, *9*(3), 035201. https://doi.org/10.1088/2053-1591/ac5957

Zhang, Y., Liu, X., Lei, L., & Liu, L. (2022). Estimating global anthropogenic CO_2 gridded emissions using a data-driven stacked random forest regression model. *Remote Sensing*, *14*(16), 3899. https://doi.org/10.3390/rs14163899

Zhi, W., Appling, A. P., Golden, H. E., Podgorski, J., & Li, L. (2024). Deep learning for water quality. *Nature Water*, *2*(3), 228–241. https://doi.org/10.1038/s44221-024-00202-z

References From the Supporting Information

Assis, J., Fernández Bejarano, S. J., Salazar, V. W., Schepers, L., Gouvêa, L., Fragkopoulou, E., et al. (2024). Bio-ORACLE v3.0. Pushing marine data layers to the CMIP6 Earth system models of climate change research (Version 3.0) [Dataset]. *Easier access to scientific data*. https://erddap.bio-oracle.org/erddap/index.html

Bradley, J. A., Hülse, D., LaRowe, D. E., & Arndt, S. (2022). DomHu/ClosingTheBioCPump_OMEN: InitialPublication [Dataset]. *Zenodo*. https://zenodo.org/badge/latestdoi/566835035

Hendricks, S. (2024). ESA Cryo-TEMPO-Northern hemisphere land/ocean flag and distance to coast at resolution of 250 m. (Version 1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.12156765

Straume, E. O., Gaina, C., Medvedev, S., Hochmuth, K., Gohl, K., Whittaker, J. M., et al. (2019). GlobSed: Updated total sediment thickness in the world's oceans [Dataset]. *Geochemistry, Geophysics, Geosystems*, *20*(4), 1756–1772. https://doi.org/10.1029/2018gc008115

Schwestermann, T., Huang, J., Konzett, J., Kioka, A., Wefer, G., Ikehara, K., et al. (2020). Multivariate statistical and multiproxy constraints on earthquake-triggered sediment remobilization processes in the central Japan Trench. Geochemistry. *Geophysics, Geosystems*, *21*(6), e2019GC008861. https://doi.org/10.1029/2019gc008861

Shaik, I., Yadav, S., Krishna, G., Mahesh, P., Nagamani, P. V., Begum, S. K., et al. (2024). Global surface ocean total alkalinity estimation: The machine learning approach. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5. https://doi.org/10.1109/lgrs.2024.3398048

21698961, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JG008386 by Ocean University Of China, Wiley Online Library on [19/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License