


Technical Note

# Reconstruction of Sea Surface Chlorophyll-a Concentration in the Bohai and Yellow Seas Using LSTM Neural Network

Qing Xu <sup>1,2,3,\*</sup> , Guiying Yang <sup>1</sup>, Xiaobin Yin <sup>1,2,3</sup> and Tong Sun <sup>1</sup>

<sup>1</sup> College of Marine Technology, Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China; ygy1328@stu.ouc.edu.cn (G.Y.); yinxiaobin@ouc.edu.cn (X.Y.); suntong01116@stu.ouc.edu.cn (T.S.)

<sup>2</sup> Laboratory for Regional Oceanography and Numerical Modeling, Qingdao Marine Science and Technology Center, Qingdao 266100, China

<sup>3</sup> SANYA Oceanographic Institution, Ocean University of China, Sanya 572024, China

\* Correspondence: xuqing@ouc.edu.cn

**Abstract:** In order to improve the spatiotemporal coverage of satellite Chlorophyll-a (Chl-a) concentration products in marginal seas, a physically constrained deep learning model was established in this work to reconstruct sea surface Chl-a concentration in the Bohai and Yellow Seas using a Long Short-Term Memory (LSTM) neural network. Adopting the permutation feature importance method, time sequences of several geographical and physical variables, including longitude, latitude, time, sea surface temperature, salinity, sea level anomaly, wind field, etc., were selected and integrated to the reconstruction model as input parameters. Performance inter-comparisons between LSTM and other machine learning or deep learning models was conducted based on OC-CCI (Ocean Color Climate Change Initiative) Chl-a product. Compared with Gated Recurrent Unit, Random Forest, XGBoost, and Extra Trees models, the LSTM model exhibits the highest accuracy. The average unbiased percentage difference (UPD) of reconstructed Chl-a concentration is 11.7%, which is 2.9%, 7.6%, 10.6%, and 10.5% smaller than that of the other four models, respectively. Over the majority of the study area, the root mean square error is less than 0.05 mg/m<sup>3</sup> and the UPD is below 10%, indicating that the LSTM model has considerable potential in accurately reconstructing sea surface Chl-a concentrations in shallow waters.

**Keywords:** sea surface chlorophyll-a concentration; reconstruction; Long Short-Term Memory; deep learning



Academic Editor: Raphael Kudela

Received: 18 November 2024

Revised: 26 December 2024

Accepted: 31 December 2024

Published: 6 January 2025

**Citation:** Xu, Q.; Yang, G.; Yin, X.; Sun, T. Reconstruction of Sea Surface Chlorophyll-a Concentration in the Bohai and Yellow Seas Using LSTM Neural Network. *Remote Sens.* **2025**, *17*, 174. <https://doi.org/10.3390/rs17010174>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Chlorophyll-a (Chl-a) is the main indicator for evaluating the biomass of phytoplankton and measuring marine primary productivity [1,2]. The photosynthesis of marine phytoplankton not only provides a driving force for the material cycle of marine ecosystems but also has a profound impact on global climate change as a key link to the global carbon cycle [3]. Monitoring sea surface Chl-a concentration is of great significance for ecological protection and marine scientific research and applications [4,5].

Compared with traditional in situ measurements, spaceborne ocean color sensors, such as SeaWiFS (Sea-viewing Wide Field-of-view Sensor), MODIS (Moderate-resolution Imaging Spectroradiometer), COCTS (Chinese Ocean Color and Temperature Scanner), etc., regularly observe global Chl-a concentration with high spatial resolution. However, observations based on optical sensors are often affected by cloud cover, aerosols, and sun glint, resulting in a large amount of data gaps and severely reducing the spatiotemporal

continuity of the data [2]. This limitation has greatly weakened the application potential of water color remote sensing technology. Therefore, the effective reconstruction of missing sea surface Chl-a concentration using information of relevant marine environmental variables such as sea surface temperature (SST) or salinity (SSS) is of great significance for improving quantitative remote sensing observations and their applications.

Traditional Chl-a concentration reconstruction approaches include the optimal interpolation (OI), empirical orthogonal functions (EOF), data-interpolating EOF (DINEOF), or multivariate DINEOF [6–9]. When dealing with data loss issues, these techniques mainly estimate missing data by utilizing the values of surrounding pixels or rely on iteratively extracting spatiotemporal feature patterns from data time series. However, if there is too much missing data in the selected area (e.g., over 55%), it will be difficult to effectively utilize these algorithms for data reconstruction.

With the rapid development of artificial intelligence technology, machine learning or deep learning methods have also been used to fill gaps in satellite ocean color measurements [10–16]. Due to their powerful ability in data mining and capturing the complex nonlinear relationships between sea surface Chl-a and other environmental factors in spatial and temporal dimensions, these methods are highly suitable for solving data reconstruction problems with large amounts of missing data. Based on SST and sea surface height (SSH) data, Jouini et al. [17] used the Self Organizing Maps (SOM) method to reconstruct the missing data of daily sea surface Chl-a concentration in the Northwestern Atlantic Ocean. Xing et al. [18] applied the Extreme Gradient Boosting algorithm to reconstruct MODIS Chl-a products in data missing areas of the Northwest Pacific. Various environmental variables such as latitude, longitude, photosynthetically active radiation, and microwave sensor-observed significant wave height, sea surface wind field (SSW) and SSS were used as input factors. Compared with the original MODIS product, the coverage of reconstructed data was greatly improved, and there was no significant decrease in data quality. Based on the Data Interpolation Convolutional Autoencoder (DINCAE) method, Luo et al. [19] estimated sea surface Chl-a concentration in the Bohai and Yellow Seas. Compared with DINEOF, the DINCAE algorithm exhibits higher accuracy and efficiency in this region. Recently, a convolutional neural network (CNN) was used for global Chl-a data reconstruction from information of multi-physical variables [20], as it is capable of extracting complex variable features from both temporal and spatial domains. Ye et al. [21] developed a deep learning model, namely OI-SwinUnet, which combines the advantages of OI and SwinUnet to reconstruct daily MODIS Chl-a concentration data in the South China Sea.

Geographical factors (latitude and longitude of the location) and many oceanographic variables, such as SST (sea surface temperature), SSS (sea surface salinity), SSH (sea surface height), and SSW (sea surface wind), may affect the spatiotemporal patterns of sea surface Chl-a concentration. Which factors play a more important role in Chl-a concentration reconstruction in a certain region? This issue still needs further investigation. In addition, the link between ocean dynamics and phytoplankton may also change over time or show a phase difference. However, this has been overlooked in most studies.

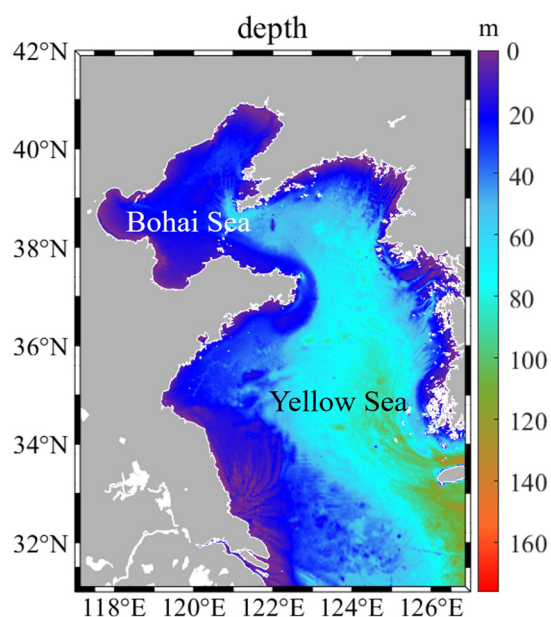
In this work, we aim to develop a physically constrained Long Short-Term Memory (LSTM) model for sea surface Chl-a concentration reconstruction in the Bohai and Yellow Seas. The main contributions of this study are as follows: (1) Different impacts of geographical and multiple physical variables on changes in Chl-a concentrations were quantitatively analyzed. (2) The proposed LSTM model with optimal input parameters shows great potential in improving both the accuracy and spatiotemporal coverage of satellite Chl-a concentration products in the study area.

The rest of the paper is organized as follows. The datasets are introduced in Section 2. The LSTM model for reconstructing Chl-a concentration is described in Section 3. The model performance and error analysis are presented in Section 4. Section 5 is a summary.

## 2. Materials and Methods

### 2.1. Study Area

Figure 1 shows the spatial domain and water depth distribution of the study area. The Bohai Sea is a semi-enclosed inland sea of China and connected to the Yellow Sea through the Bohai Strait. Its average water depth of 18 m and unique geographical location lead to a long retention time (about 1.6 years) of suspended sediment concentration and pollutants in the area. The Yellow Sea is a marginal sea of the Northwestern Pacific Ocean and its water renewal cycle is relatively long. Rivers carry large amounts of sediment and organic matter into the Bohai and Yellow Seas, providing abundant freshwater resources and nutrients for the region, which in turn exacerbates eutrophication. The physical processes in the study area are mainly dominated by the Asian monsoon system. In winter, strong northwesterly winds and dry air prevail in the region. In summer, warm southeasterly winds associated with relatively high precipitation affect the area, while seasonal storm activity, including occasional typhoons, occurs in late summer and early fall.



**Figure 1.** Water depth (m) of the Bohai Sea and Yellow Seas.

### 2.2. Data

The concentration of sea surface Chl-a is influenced by various factors. Firstly, its variation is related to offshore distance and season, mainly reflected through latitude, longitude, and time (day of the year, DOY). Secondly, among the many factors that affect the growth of phytoplankton, atmospheric and marine environmental variables play a crucial role [22]. For example, there is a high correlation between SST and Chl-a concentration, and SST has a significant impact on the photosynthetic efficiency and growth rate of phytoplankton [23,24]. The wind-induced upwelling or cold eddy, which brings colder, nutrient rich waters to the surface, is usually accompanied by lower SST and SSH, which may also be related to the distribution of phytoplankton or Chl-a concentration [25]. River runoff can affect the nutrient content in the sea, indirectly affecting Chl-a concentration through changes in SSS [26].

Therefore, the initial input factors of the LSTM model constructed in this paper include eight physical variables and four geographical variables. The physical variables are SST, SST anomaly (SSTA), SSS, sea level anomaly (SLA), zonal wind speed (uwind), meridional wind speed (vwind), sea surface wind stress curl (SSWSC), and precipitation. The geographical variables are longitude, latitude,  $\sin(\theta)$  ( $\theta = 2\pi \times \text{DOY}/365.25$ ), and  $\cos(\theta)$ . The data used for training and validating the accuracy of the reconstruction model are from the OC-CCI (Ocean Color-Climate Change Initiative) product from 2013 to 2019.

### 2.2.1. OC-CCI Chl-a Product

The daily OC-CCI sea surface Chl-a concentration product (version 5) has a spatial resolution of 4 km [27]. In the processing of the product, atmospheric corrections were first applied to multi-band data from MODIS, SeaWiFS, VIIRS (Visible and Infrared Imaging Radiometer Suite), MERIS (Medium Resolution Imaging), and OLCI (Ocean Land Colour Instrument). These data were converted into values corresponding to the operating bands of MERIS, with center wavelengths of 412 nm, 443 nm, 490 nm, 510 nm, 560 nm, and 665 nm, respectively. The obtained remote sensing reflectance (Rrs) data of five sensors were then bias-corrected to MERIS level and merged. Finally, based on the fused Rrs data, the Chl-a concentration was calculated using an exponential algorithm and band ratio algorithm [28–30]. The OC-CCI Chl-a product has a high quality and has been widely used in marine research [31–33].

### 2.2.2. Satellite Data of Environment Variables

The daily SST data used in this study are from the GHR SST (Group for High-Resolution Sea Surface Temperature) L4 level product (Version 4.1) with a spatial resolution of 1 km (Figure 2a). The product was obtained by interpolating SST observations from various instruments, including AMSR-E (Advanced Microwave Scanning Radiometer-EOS), AMSR2 (Advanced Microwave Scanning Radiometer 2), WindSat microwave radiometer, MODIS onboard Aqua and Terra satellites, AVHRR (Advanced Very High Resolution Radiometer), and buoy measurements from iQuam. The SSTA can be calculated by subtracting the climatological mean from SST.

The daily SSS product (Version 3.21) is provided by the ESA-CCI (Climate Change Initiative), which merges observations from SMOS (Soil Moisture and Ocean Salinity), SMAP (the Soil Moisture Active Passive), Aquarius, etc.. The data has been spatially sampled on a 25 km EASE (Equal Area Scalable Earth) grid (Figure 2b). The product is in good agreement with in situ measurements of Argo, with a standard deviation of 0.15 psu [34].

The daily SLA data product was developed by AVISO (Archiving, Validation and Interpretation of Satellite Oceanographic) and distributed by Copernicus. The product merges multi-mission (Topex-Poseidon, Jason-1/2/3, Sentinel-6, CryoSat-2, ERS-1/2, etc.) satellite altimetry data with the optimal interpolation method. The L4 product is available from 1993 and has a spatial resolution of  $0.25^\circ \times 0.25^\circ$  (Figure 2c).

The six-hourly wind product (L4, Version 3.1) with spatial resolution of 25 km is from the Cross-Calibrated Multi-Platform (CCMP) (Figure 2d), which merges satellite observations of the wind field from scatterometers and microwave radiometers using the variational analysis data assimilation technique, including QuikScat and ASCAT-A/B, SSM/I (Special Sensor Microwave Imager), SSMIS (Special Sensor Microwave Imager Sounder), TMI (The Tropical Rainfall Measuring Mission Microwave Imager), GMI (Global Precipitation Measurement Microwave Imager), ASMR-E (Advanced Microwave Scanning Radiometer for EOS), AMSR2, and WindSat. The sea surface wind stress curl is then calculated with the following form [35]:

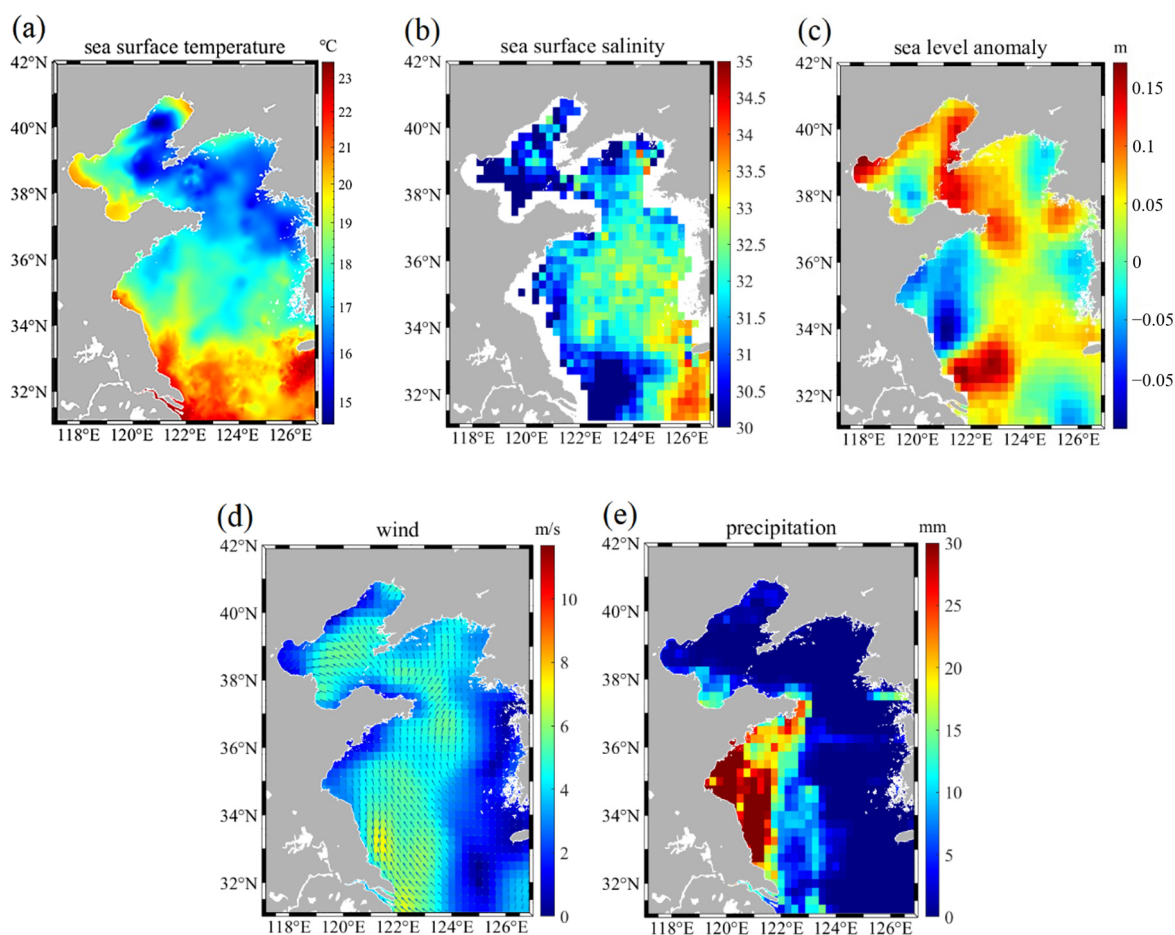
$$\text{curl}_z = \frac{\partial \tau_y}{\partial x} - \frac{\partial \tau_x}{\partial y} \quad (1)$$

$$\vec{\tau} = (\tau_x, \tau_y) = \rho_{air} C_D U_{10} \cdot (u, v) \quad (2)$$

where  $\vec{\tau}$  represents the sea surface wind stress and  $\tau_x$  and  $\tau_y$  are its zonal and meridional components, respectively;  $\rho_{air}$  is air density at the sea surface, which is generally  $1.225 \text{ kg/m}^3$ ;  $u$  and  $v$  are the zonal and meridional wind speed at a height of 10 m above the sea surface, respectively, and  $U_{10} = \sqrt{u^2 + v^2}$ ;  $C_D$  is the drag coefficient, which describes the frictional effect of wind on the sea surface and was calculated with the following form [36]:

$$C_D = \begin{cases} 1.2875 \times 10^{-3}, & U_{10} \leq 7.5 \text{ m/s} \\ (0.8 + 0.065U_{10}) \times 10^{-3}, & U_{10} > 7.5 \text{ m/s} \end{cases} \quad (3)$$

The TRMM (Tropical Rainfall Measuring Mission) daily product (3B42) with spatial resolution of 25 km was used to evaluate the effects of precipitation on Chl-a concentration patterns. The product merges TMI observations with AMSR-E, AMSU-B (Advanced Microwave Sounding Unit-B), and SSM/I data (Figure 2e).



**Figure 2.** Sea surface temperature (a), salinity (b), sea level anomaly (c), wind field (d), and precipitation (e) on 5 June, 2019, in the Bohai Sea and Yellow Seas.

### 2.3. Data Preprocessing

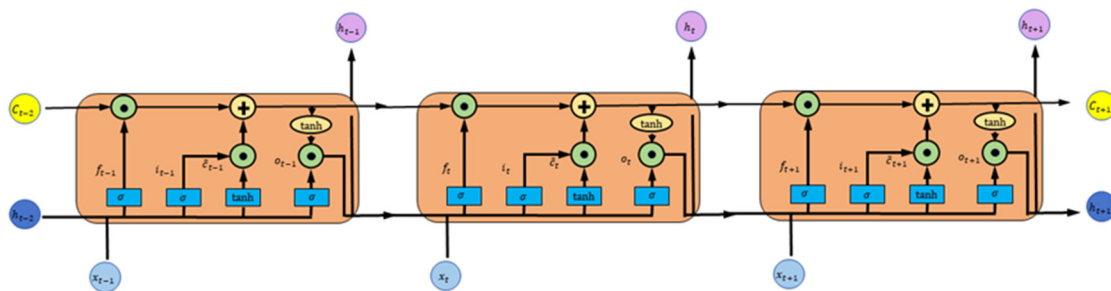
To maintain the consistency of the above datasets, the nearest neighbor interpolation algorithm was used to resample daily SST and Chl-a product to a  $25 \text{ km} \times 25 \text{ km}$  grid. Also, 6-hourly CCMP wind field data were averaged each day to obtain daily wind data. Then, a



dataset for model training and evaluation was generated by matching all environmental data with OC-CCI Chl-a concentration data from 2013 to 2019. In total, 489,316 data samples were obtained. Among them, 80% of the samples from 2013 to 2018 were randomly selected as the training set, the remaining 20% were used for validation, and data from 2019 constituted the testing set. To ensure the uniformity of Chl-a concentration distribution, a logarithmic transformation ( $\log_{10}$ ) was performed on the data.

#### 2.4. LSTM Model

LSTM is an optimized recurrent neural network (RNN), which has significant advantages in processing time series data. It is characterized by connections not only between layers but also between nodes within the hidden layers. This internal connection allows the model to capture and retain information from previous moments and pass it on to subsequent moments. As shown in Figure 3, LSTM introduces memory units in the hidden layer of RNN that can store information for a long time and adjusts the storage and updating of information in the memory units by adding gating units, effectively alleviating the problem of gradient vanishing and improving the performance of the network in processing long-time series data [37].



**Figure 3.** Structure of the LSTM neural network.

The processor of LSTM consists of three gates: input gate ( $i_t$ ), output gate ( $o_t$ ), and forget gate ( $f_t$ ). The data information is processed by these gates to obtain the neuron state at the current moment, which is subsequently transmitted along the time series and updated to record the current state when it reaches each neuron. The initial step of LSTM is to determine which information should be removed from the neuron state, which is determined by the sigmoid function of the forget gate [38]. The second step is to determine how much new information needs to be updated and saved. The sigmoid function decides which information needs to be updated, and tanh is responsible for generating new candidate variables. After the above processes, the neuron immediately updates its state at that moment. Finally, the output gate determines which neuron states are outputted based on sigmoid, and then these states are transformed through the tanh function to obtain a result between  $-1$  and  $1$ . The result is then multiplied by the output gate to obtain the output information.

In this study, to determine the optimal model input parameters, the permutation feature importance (PFI) method was used to analyze the contribution of different input variables. The impact of model hyperparameter settings on the reconstruction accuracy was also analyzed. Considering the possible lagged effects of various environmental factors on sea surface Chl-a concentration, parameter information from three consecutive days ( $t = 0, -1, -2$ ) was input into the model to reconstruct Chl-a concentration on a given date ( $t = 0$ ).

The model accuracy was evaluated using root mean square error (RMSE), unbiased percentage difference (UPD), and  $\log_{10}$  logarithmic form of correlation coefficient ( $R(\log)$ ), which take the following forms [5,39]:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{\text{REC}_i} - y_{\text{OBS}_i})^2} \quad (4)$$

$$\text{UPD} = \frac{1}{N} \sum_{i=1}^N \left( \left| \frac{y_{\text{REC}_i} - y_{\text{OBS}_i}}{y_{\text{OBS}_i} + y_{\text{REC}_i}} \right| \right) \times 200\% \quad (5)$$

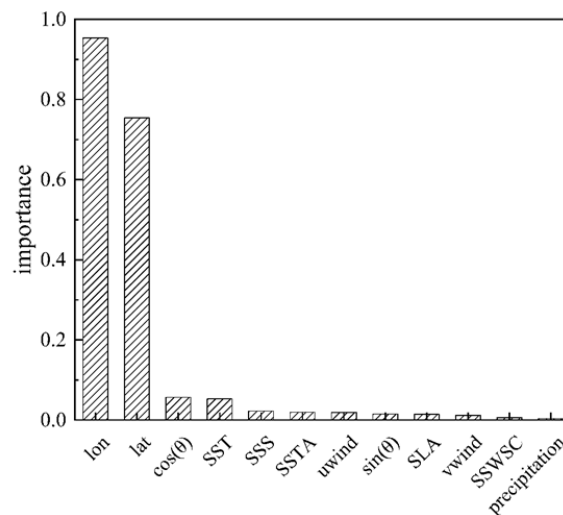
$$\text{R}(\log) = \frac{\sum_{i=1}^N \left[ \log_{10}(y_{\text{REC}_i}) - \overline{\log_{10}(y_{\text{REC}})} \right] \left[ \log_{10}(y_{\text{OBS}_i}) - \overline{\log_{10}(y_{\text{OBS}})} \right]}{\sqrt{\sum_{i=1}^N \left[ \log_{10}(y_{\text{REC}_i}) - \overline{\log_{10}(y_{\text{REC}})} \right]^2 \sum_{i=1}^N \left[ \log_{10}(y_{\text{OBS}_i}) - \overline{\log_{10}(y_{\text{OBS}})} \right]^2}} \quad (6)$$

where  $y_{\text{REC}}$  and  $y_{\text{OBS}}$  represent the reconstructed sea surface Chl-a concentration and true values, respectively;  $\log_{10}(y_{\text{RES}})$  and  $\log_{10}(y_{\text{OBS}})$  represent the average value of  $y_{\text{REC}}$  and  $y_{\text{OBS}}$  after  $\log_{10}$  transformation;  $N$  represents the total number of data samples. These evaluation indicators can reveal the reconstruction ability of the LSTM model from multiple dimensions and also provide theoretical support for further optimization of the model.

### 3. Results

#### 3.1. Contribution of the Environmental Variables to Chl-a Concentration Reconstruction

To explore the contribution of various geographical and physical variables to the LSTM-based Chl-a reconstruction model, the PFI method was used to analyze the relative importance of different input parameters. The initial width of the model is set as 64 and the number of hidden layers is 3. From Figure 4, it is found that the geographic location (latitude and longitude) are the most important factors, followed by  $\cos(\theta)$  and SST. We further explored the optimal factors for Chl-a reconstruction through comparative experiments. Table 1 shows the statistical error of the LSTM model when the input parameters are the top 2 to 12 factors sorted by the PFI method. One can see that the model demonstrates the highest accuracy when all 12 factors were used as input.



**Figure 4.** Importance of different environmental factors based on PFI analysis.

#### 3.2. Overall Model Performance

In order to further determine the optimal hyperparameters of the LSTM model with all 12 environmental factors as input, we adopted the Bayesian optimization algorithm [2]. Taking into account the Bayesian optimization results and model training time, the optimal hyperparameter configuration for the model was ultimately determined as follows: model

width of 128, initial learning rate of 0.001, hidden layers of 3, optimizer of “Nadam”, and activation function of “tanh”.

**Table 1.** Accuracy of the LSTM model with different input parameters. Factor numbers 1 to 12 represent the importance ranking from high to low as shown in Figure 4. The model width and number of hidden layers are 64 and 3, respectively.

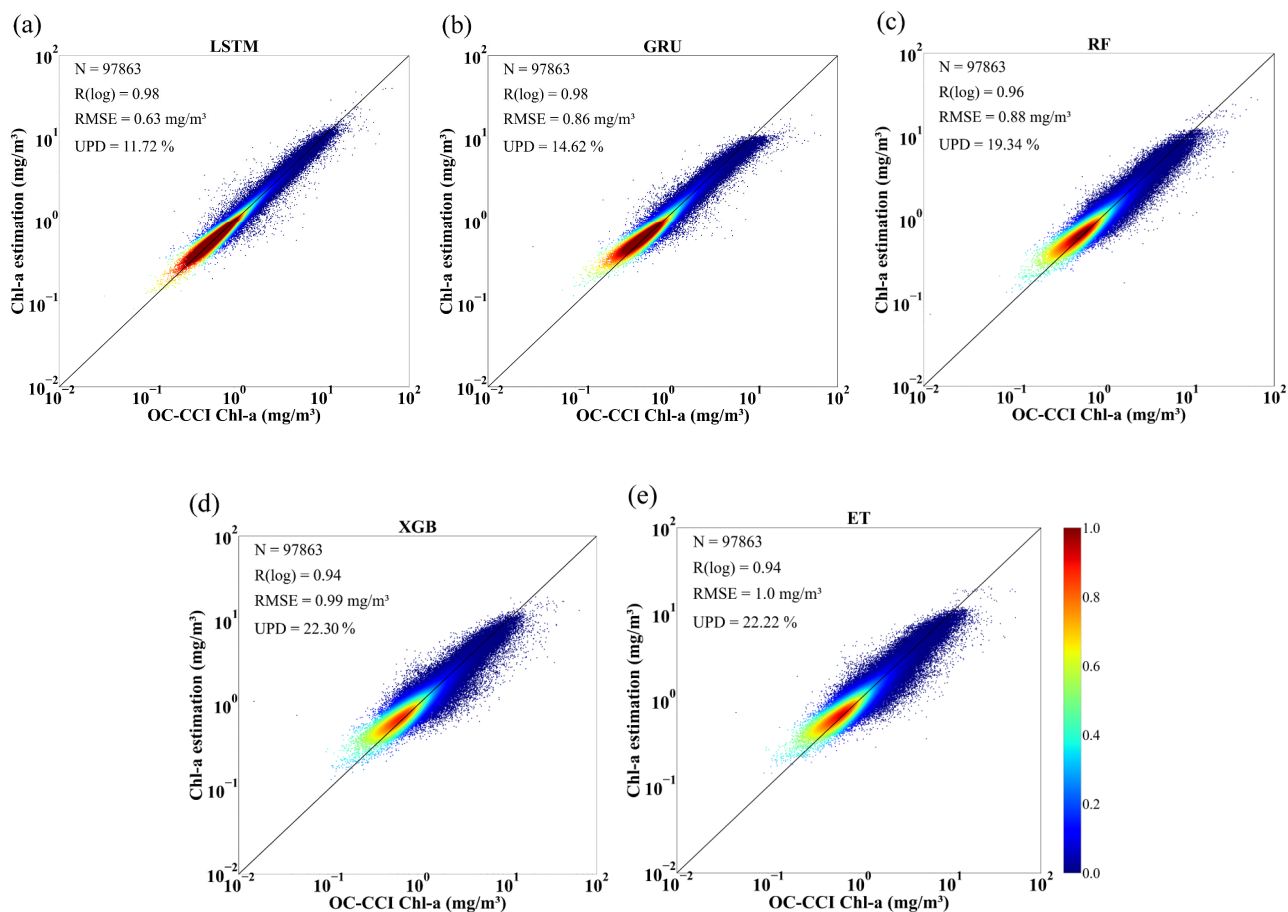
Model ID	Input	R(log)	RMSE (mg/m <sup>3</sup> )	UPD (%)
LSTM_1	Factor 1—2	0.84	1.36	37.45
LSTM_2	Factor 1—3	0.93	1.04	23.62
LSTM_3	Factor 1—4	0.94	1.00	21.44
LSTM_4	Factor 1—5	0.95	0.96	20.26
LSTM_5	Factor 1—6	0.95	0.95	19.76
LSTM_6	Factor 1—7	0.96	0.91	18.47
LSTM_7	Factor 1—8	0.96	0.91	18.19
LSTM_8	Factor 1—9	0.96	0.91	18.14
LSTM_9	Factor 1—10	0.96	0.89	17.59
LSTM_10	Factor 1—11	0.96	0.89	17.25
LSTM_11	Factor 1—12	0.97	0.88	17.24

The performance of the LSTM model based on this configuration was evaluated and compared with four machine learning algorithms, namely the Gated Recurrent Unit (GRU), Random Forest (RF), XGBoost (XGB), and Extra Trees models (ET). GRU is a simplification of LSTM [40] which integrates the input gate and forget gate of LSTM into an update gate and thus only contains an update gate and reset gate. Both ET and RF are essentially decision tree sets based on ensemble learning.

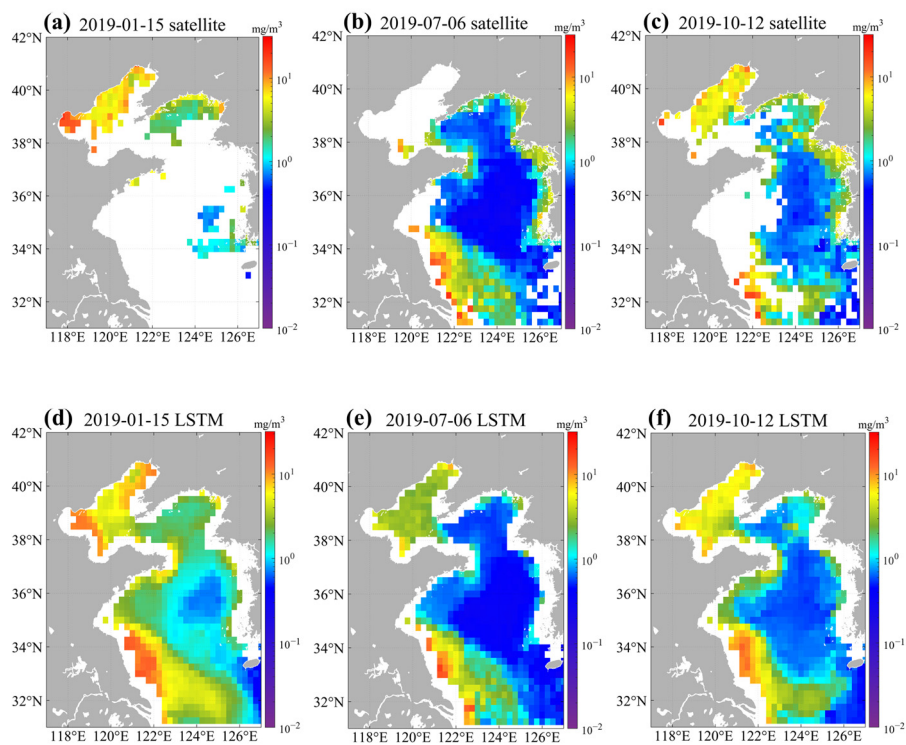
The scatter probability density between the reconstruction results of each model and the OC-CCI Chl-a product is shown in Figure 5. It is obvious that the time-series-based LSTM model exhibits the highest accuracy, with an UPD of 11.72%, RMSE of 0.63 mg/m<sup>3</sup>, and R(log) of 0.98. Compared with the GRU, RF, XGB, and ET models, the UPD is 2.9%, 7.6%, 10.6%, and 10.5% smaller, respectively. In Figure 5a, it can be found that LSTM reconstructed results are evenly and tightly distributed on both sides of the diagonal. Except for this model, other models generally overestimate Chl-a concentration when the value is less than 0.2 mg/m<sup>3</sup>. The accuracy of the GRU model is slightly lower than that of the LSTM model, but the training speed is faster. In general, the LSTM model performs the best and is more suitable for filling gaps of Chl-a concentration data in the Bohai and Yellow Seas.

Figures 6 and 7 show the spatial distribution of LSTM and GRU model reconstruction results on different dates in 2019. In Figure 6, there are extensive data gaps in the satellite product, while very few data are missing in Figure 7 during cloud-free conditions. Due to the limited spatial coverage of SSS data used in this study, there is a lack of reconstructed Chl-a concentration values in coastal regions. From Figure 6, it can be seen that the overall coverage of Chl-a concentration has been significantly improved. The Chl-a pattern reconstructed by LSTM shows good consistency with the OC-CCI product. The reconstructed results clearly reflect the decreasing trend of Chl-a concentration with increasing offshore distance. The model also captures the seasonal variation of Chl-a concentration, with higher values in winter and relatively lower values in summer. This may be mainly due to the fact that the rapid growth of phytoplankton in spring consumed a large amount of nutrients and limited its growth in summer. After entering autumn, due to the increase in wind speed, the enhanced mixing of seawater brought the nutrients from the bottom layer to the surface, thereby gradually increasing the concentration of Chl-a at the sea surface.

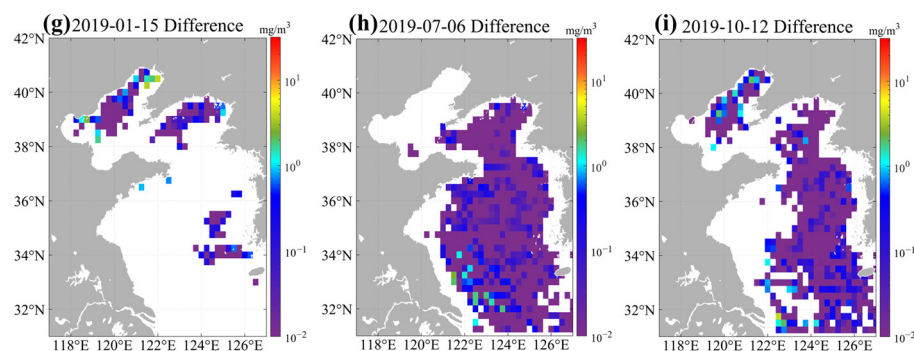




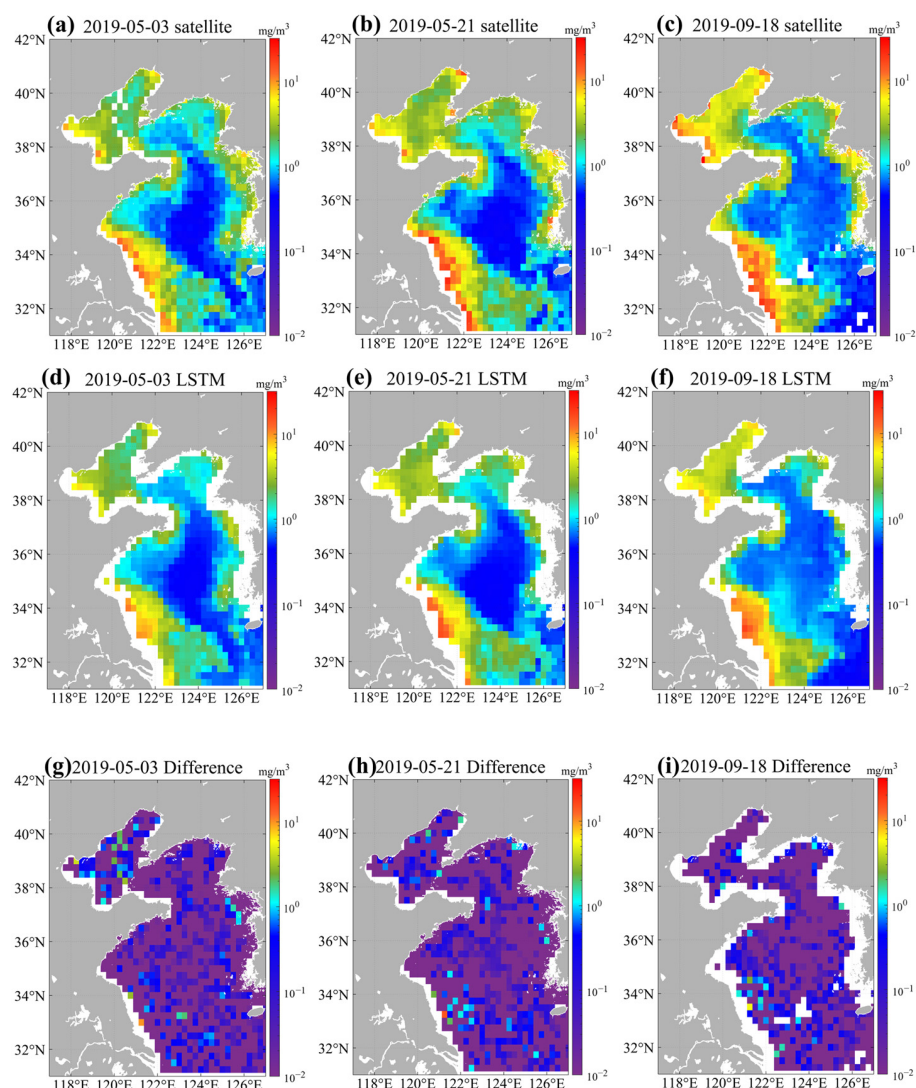
**Figure 5.** Comparison of sea surface Chl-a concentration reconstructed by different models with the OC-CCI product on the testing dataset: (a) LSTM; (b) GRU; (c) RF; (d) XGB; (e) ET. N denotes the total number of data samples.



**Figure 6.** Cont.



**Figure 6.** Spatial distribution of daily OC-CCI product with extensive data gaps (a–c), LSTM reconstructed Chl-a concentration (d–f) in 2019, and the difference between them (g–i).



**Figure 7.** Same as Figure 6 but for daily OC-CCI product during cloud free conditions.

#### 4. Discussion

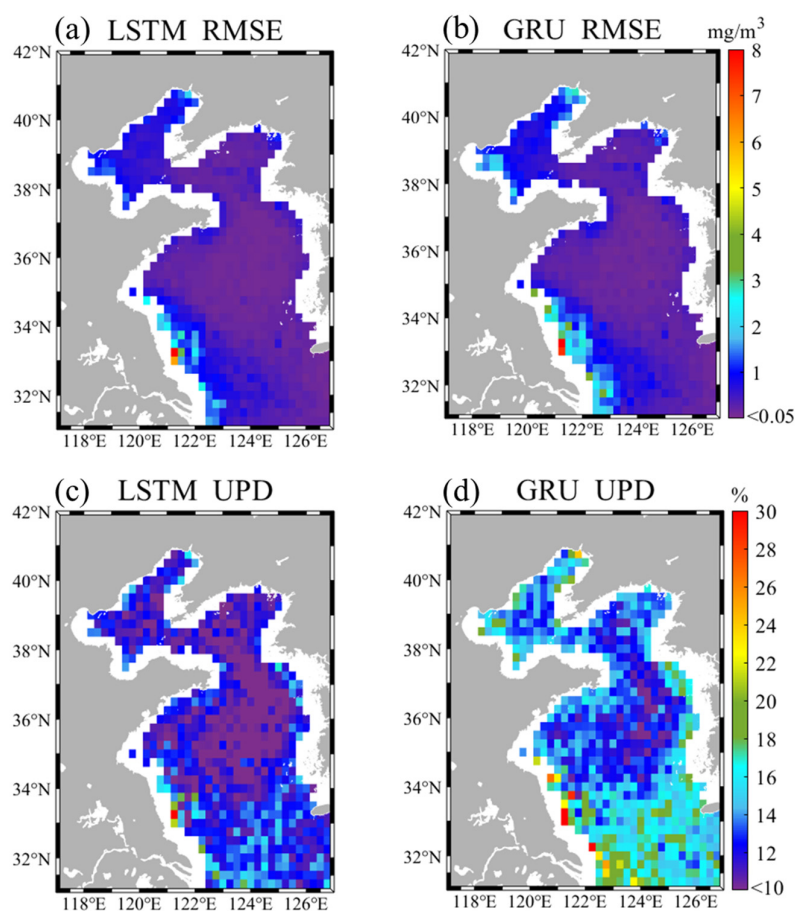
In the above study, sea surface Chl-a concentrations on a given date ( $t = 0$ ) were reconstructed using 12 environmental variables from the previous two days and this given date ( $t = -2, -1, 0$ ) as model inputs. To what extent does the historical information of input factors affect the performance of the LSTM model? Comparative experiments were conducted using 1-day ( $t = 0$ ) and 2-day ( $t = -1, 0$ ) input windows, respectively. As shown in Table 2, compared with the model with only 1 day of input in Experiment 1, the

additional input from the previous day in Experiment 2 significantly improves the model performance, with a 4.53% reduction in UPD. The use of a 3-day input window further helps to reduce UPD and RMSE by 2.49% and 0.14 mg/m<sup>3</sup>, respectively. This verifies the contribution of time series of input data in reconstructing Chl-a concentration.

**Table 2.** Accuracy of the LSTM model with different time window of input data.

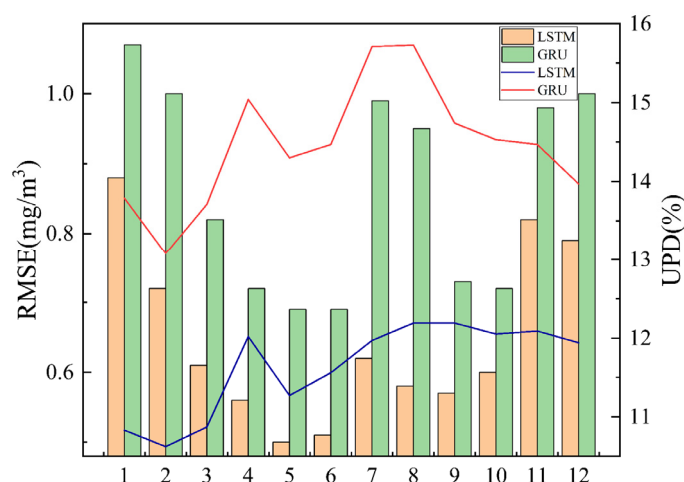
Experiment ID	Time Window of Input Data	RMSE (mg/m <sup>3</sup> )	UPD (%)
1	1 day (t = 0)	0.95	18.74
2	2 days (t = −1, 0)	0.77	14.21
3	3 days (t = −2, −1, 0)	0.63	11.72

To further evaluate the performance of optimal LSTM and GRU models in different regions of the Bohai and Yellow Seas, spatiotemporal analysis of reconstruction errors was conducted. Figure 8 shows the average spatial distributions of RMSE and UPD in 2019. The overall performance of the LSTM model is better than that of the GRU model, and its advantages in the nearshore area are more significant. The error decreases with the increase in offshore distance. The RMSE is less than 0.05 mg/m<sup>3</sup> in most areas but is larger in coastal areas of Jiangsu Province and the Yangtze River estuary, where the turbidity of the waters is extremely high. The UPD of the two models is small in the waters north of 34°N. In particular, the LSTM-derived UPD is less than 10% in most regions. Comparing Figure 8c,d, it is found that the UPD of the LSTM model is about 3% smaller than that of the GRU model on average, and the difference is even more than 5% in some areas of the South Yellow Sea.



**Figure 8.** Distribution of Chl-a concentration reconstruction errors based on LSTM (left) and GRU (right) models: (a,b) RMSE; (c,d) UPD.

Figure 9 shows the monthly averaged errors of the LSTM and GRU models. In most months, the LSTM model-derived RMSE is below  $0.6 \text{ mg/m}^3$ , while that of the GRU model is always above  $0.7 \text{ mg/m}^3$  and even higher than  $1.0 \text{ mg/m}^3$ . For the LSTM model, the RMSE is smaller in summer, with the minimum value less than  $0.5 \text{ mg/m}^3$  in May, and larger in winter when the sea surface Chl-a concentration is higher. The seasonal trend of UPD obtained by LSTM and GRU is similar, but most of the time, the UPD of the LSTM model is 3% lower than that of the GRU model. This further verifies the conclusion that the LSTM model performs better than the GRU model in the study area.



**Figure 9.** Monthly averaged reconstruction errors of sea surface Chl-a concentration based on LSTM and GRU models. The histogram is RMSE and the curve is UPD.

## 5. Summary

Considering the impacts of various environmental variables on sea surface Chl-a concentration in the Bohai and Yellow Seas, a physically constrained LSTM model was developed to reconstruct Chl-a concentration from continuous observations of these variables, filling the data gap in satellite ocean color products. Through the permutation feature importance analysis, the optimal input factors were determined, including eight oceanic and atmospheric variables, including SST, SSTA, SSS, SLA, meridional wind speed, zonal wind speed, wind stress curl, and precipitation for three consecutive days, and four geographic variables, such as longitude, latitude,  $\sin(\theta)$ , and  $\cos(\theta)$ . Compared with the OC-CCI product, the  $R(\log)$ , RMSE, and UPD of the LSTM model are 0.98,  $0.63 \text{ mg/m}^3$ , and 11.72%, respectively. The UPD is 2.9%, 7.6%, 10.6%, and 10.5% smaller than that of the GRU, RF, XGB, and ET models, respectively. In most regions, the RMSE of the LSTM model is less than  $0.05 \text{ mg/m}^3$  and the UPD is less than 10%, which fully demonstrates the effectiveness of the model in reconstructing sea surface Chl-a concentration. The outcomes provide robust methodological and data support for marine ecological environment monitoring and protection, fisheries resource assessment, and global climate change research.

In this study, in order to maintain consistency between different datasets used for model training and testing, the spatial resolution of satellite observations of oceanographic variables was selected to be 25 km. But this does not mean that the proposed LSTM model can only reconstruct Chl-a concentrations with a spatial resolution of 25 km. In the near future, by collecting satellite observations with higher spatial resolution, such as the 2 km resolution SWOT (Surface Water and Ocean Topography) SLA product [41], the model can be further optimized and validated through more in situ measurements to meet the needs of both coastal and deep-water applications.



**Author Contributions:** Conceptualization, Q.X.; methodology, G.Y.; software, G.Y. and X.Y.; validation, G.Y. and T.S.; formal analysis, G.Y.; investigation, Q.X.; resources, Q.X.; data curation, G.Y. and T.S.; writing—original draft preparation, G.Y.; writing—review and editing, Q.X.; visualization, G.Y. and T.S.; supervision, Q.X. and X.Y.; project administration, Q.X.; funding acquisition, Q.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant numbers T2261149752, 42476172, and Hainan Province Science and Technology Special Fund, grant number SOLZSKY2024005.

**Data Availability Statement:** Data are contained within the article.

**Acknowledgments:** The OC-CCI data are available at <http://www.esa-oceancolour-cci.org/> (accessed on 1 May 2022). The SST data are available at <https://catalog.data.gov/dataset/ghrsst-global-1-km-sea-surface-temperature-g1sst-global-0-01-degree-2010-2017-daily> (accessed on 1 May 2022). The SSS data are available at <https://climate.esa.int/en/odp/#/project/sea-surface-salinity> (accessed on 1 May 2022). The SLA data are available at <https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-sea-level-global> (accessed on 1 May 2022). The sea surface wind data are available at <https://www.remss.com/measurements/ccmp> (accessed on 1 May 2022). The precipitation data are available at [https://disc.gsfc.nasa.gov/datasets/TRMM\\_3B42\\_Daily\\_7/summary?keywords=TRMM](https://disc.gsfc.nasa.gov/datasets/TRMM_3B42_Daily_7/summary?keywords=TRMM) (accessed on 1 May 2022).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Bojinski, S.; Verstraete, M.; Peterson, T.C.; Richter, C.; Simmons, A.; Zemp, M. The concept of essential climate variables in support of climate research, applications, and policy. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 1431–1443. [\[CrossRef\]](#)
2. Yang, G.; Ye, X.; Xu, Q.; Yin, X.; Xu, S. Sea surface chlorophyll-a concentration retrieval from hy-1c satellite data based on residual network. *Remote Sens.* **2023**, *15*, 3696. [\[CrossRef\]](#)
3. Hammond, M.L.; Beaulieu, C.; Henson, S.A.; Sahu, S.K. Regional surface chlorophyll trends and uncertainties in the global ocean. *Sci. Rep.* **2020**, *10*, 15273. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Hu, C.; Feng, L.; Guan, Q. A machine learning approach to estimate surface chlorophyll a concentrations in global oceans from satellite measurements. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4590–4607. [\[CrossRef\]](#)
5. Ye, X.; Liu, J.; Lin, M.; Ding, J.; Zou, B.; Song, Q. Global ocean chlorophyll-a concentrations derived from COCTS onboard the HY-1C satellite and their preliminary evaluation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9914–9926. [\[CrossRef\]](#)
6. Jayaram, C.; Pavan Kumar, J.; Udaya Bhaskar, T.; Bhavani, I.; Prasad Rao, T.; Nagamani, P. Reconstruction of gap-free OCM-2 chlorophyll-a concentration using DINEOF. *J. Indian Soc. Remote Sens.* **2021**, *49*, 1419–1425. [\[CrossRef\]](#)
7. Ma, C.; Zhao, J.; Ai, B.; Sun, S. Two-decade variability of sea surface temperature and chlorophyll-a in the northern South China Sea as revealed by reconstructed cloud-free satellite data. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9033–9046. [\[CrossRef\]](#)
8. Wang, Y.; Gao, Z.; Liu, D. Multivariate DINEOF reconstruction for creating long-term cloud-free chlorophyll-a data records from SeaWiFS and MODIS: A case study in Bohai and Yellow Seas, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1383–1395. [\[CrossRef\]](#)
9. Wang, Y.; Liu, D. Reconstruction of satellite chlorophyll-a data using a modified DINEOF method: A case study in the Bohai and Yellow seas, China. *Int. J. Remote Sens.* **2014**, *35*, 204–217. [\[CrossRef\]](#)
10. Barth, A.; Alvera-Azcárate, A.; Troupin, C.; Beckers, J.-M.; Van der Zande, D. Reconstruction of missing data in satellite images of the Southern North Sea using a convolutional neural network (DINCAE). In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 7493–7496.
11. Chen, S.; Hu, C.; Barnes, B.B.; Xie, Y.; Lin, G.; Qiu, Z. Improving ocean color data coverage through machine learning. *Remote Sens. Environ.* **2019**, *222*, 286–302. [\[CrossRef\]](#)
12. Han, Z.; He, Y.; Liu, G.; Perrie, W. Application of dincae to reconstruct the gaps in chlorophyll-a satellite observations in the south China sea and west Philippine sea. *Remote Sens.* **2020**, *12*, 480. [\[CrossRef\]](#)
13. Hu, Q.; Chen, X.; Bai, Y.; He, X.; Li, T.; Pan, D. Reconstruction of 3-D ocean chlorophyll a structure in the northern Indian ocean using satellite and BGC-argo data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 4200513. [\[CrossRef\]](#)



14. Krasnopolsky, V.; Nadiga, S.; Mehra, A.; Bayler, E.; Behringer, D. Neural networks technique for filling gaps in satellite measurements: Application to ocean color observations. *Comput. Intell. Neurosci.* **2016**, *2016*, 6156513. [[CrossRef](#)] [[PubMed](#)]
15. Martinez, E.; Gorgues, T.; Lengaigne, M.; Fontana, C.; Sauzède, R.; Menkes, C.; Uitz, J.; Di Lorenzo, E.; Fablet, R. Reconstructing global chlorophyll-a variations using a non-linear statistical approach. *Front. Mar. Sci.* **2020**, *7*, 464.
16. Park, J.; Kim, H.-C.; Bae, D.; Jo, Y.-H. Data reconstruction for remotely sensed chlorophyll-a concentration in the ross sea using ensemble-based machine learning. *Remote Sens.* **2020**, *12*, 1898. [[CrossRef](#)]
17. Jouini, M.; Lévy, M.; Crépon, M.; Thiria, S. Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method. *Remote Sens. Environ.* **2013**, *131*, 232–246. [[CrossRef](#)]
18. Xing, M.; Yao, F.; Zhang, J.; Meng, X.; Jiang, L.; Bao, Y. Data reconstruction of daily MODIS chlorophyll-a concentration and spatio-temporal variations in the Northwestern Pacific. *Sci. Total Environ.* **2022**, *843*, 156981. [[CrossRef](#)]
19. Luo, X.; Song, J.; Guo, J.; Fu, Y.; Wang, L.; Cai, Y. Reconstruction of chlorophyll-a satellite data in Bohai and Yellow Sea based on DINCAE method. *Int. J. Remote Sens.* **2022**, *43*, 3336–3358. [[CrossRef](#)]
20. Roussillon, J.; Fablet, R.; Gorgues, T.; Drumetz, L.; Littaye, J.; Martinez, E. A Multi-Mode Convolutional Neural Network to reconstruct satellite-derived chlorophyll-a time series in the global ocean from physical drivers. *Front. Mar. Sci.* **2023**, *10*, 1077623. [[CrossRef](#)]
21. Ye, H.; Yang, C.; Dong, Y.; Tang, S.; Chen, C. A daily reconstructed chlorophyll-a dataset in South China Sea from MODIS using OI-SwinUnet. *Earth Syst. Sci. Data Discuss.* **2024**, *2023*, 3125–3147. [[CrossRef](#)]
22. Dai, Y.; Yang, S.; Zhao, D.; Hu, C.; Xu, W.; Anderson, D.M.; Li, Y.; Song, X.-P.; Boyce, D.G.; Gibson, L. Coastal phytoplankton blooms expand and intensify in the 21st century. *Nature* **2023**, *615*, 280–284. [[CrossRef](#)] [[PubMed](#)]
23. Ji, C.; Zhang, Y.; Cheng, Q.; Tsou, J.; Jiang, T.; San Liang, X. Evaluating the impact of sea surface temperature (SST) on spatial distribution of chlorophyll-a concentration in the East China Sea. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *68*, 252–261. [[CrossRef](#)]
24. Krug, L.A.; Platt, T.; Sathyendranath, S.; Barbosa, A.B. Unravelling region-specific environmental drivers of phytoplankton across a complex marine domain (off SW Iberia). *Remote Sens. Environ.* **2017**, *203*, 162–184. [[CrossRef](#)]
25. Gupta, G.; Sudheesh, V.; Sudharma, K.; Saravanane, N.; Dhanya, V.; Dhanya, K.; Lakshmi, G.; Sudhakar, M.; Naqvi, S. Evolution to decay of upwelling and associated biogeochemistry over the southeastern Arabian Sea shelf. *J. Geophys. Res. Biogeosciences* **2016**, *121*, 159–175. [[CrossRef](#)]
26. Jo, Y.-H.; Kim, D.-W.; Kim, H. Chlorophyll concentration derived from microwave remote sensing measurements using artificial neural network algorithm. *J. Mar. Sci. Technol.* **2018**, *26*, 10.
27. Sathyendranath, S.; Jackson, T.; Grant, M.; Brewin, R.; Brotas, V.; Brockmann, C. *ESA Ocean Colour Climate Change Initiative (Ocean\_Colour\_cci): Version 5.0 Data*; NERC EDS Centre for Environmental Data Analysis: Didcot, UK, 2021.
28. Gordon, H.R.; Morel, A.Y. *Remote Assessment of Ocean Color for Interpretation of Satellite Visible Imagery: A Review*; Springer-Verlag: New York, NY, USA; Berlin/Heidelberg, Germany; Tokyo, Japan, 1983; 114p.
29. Hu, C.; Lee, Z.; Franz, B. Chlorophyll algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *J. Geophys. Res. Ocean.* **2012**, *117*, C01011. [[CrossRef](#)]
30. O'Reilly, J.E.; Maritorena, S.; Mitchell, B.G.; Siegel, D.A.; Carder, K.L.; Garver, S.A.; Kahru, M.; McClain, C. Ocean color chlorophyll algorithms for SeaWiFS. *J. Geophys. Res. Ocean.* **1998**, *103*, 24937–24953. [[CrossRef](#)]
31. Geng, B.; Xiu, P.; Shu, C.; Zhang, W.Z.; Chai, F.; Li, S.; Wang, D. Evaluating the roles of wind-and buoyancy flux-induced mixing on phytoplankton dynamics in the northern and central South China Sea. *J. Geophys. Res. Ocean.* **2019**, *124*, 680–702. [[CrossRef](#)]
32. Keerthi, M.G.; Prend, C.J.; Aumont, O.; Lévy, M. Annual variations in phytoplankton biomass driven by small-scale physical processes. *Nat. Geosci.* **2022**, *15*, 1027–1033. [[CrossRef](#)]
33. Ma, W.; Xiu, P.; Chai, F.; Li, H. Seasonal variability of the carbon export in the central South China Sea. *Ocean Dyn.* **2019**, *69*, 955–966. [[CrossRef](#)]
34. Boutin, J.; Reul, N.; Köhler, J.; Martin, A.; Catany, R.; Guimbard, S.; Rouffi, F.; Vergely, J.-L.; Arias, M.; Chakroun, M. Satellite-based sea surface salinity designed for ocean and climate studies. *J. Geophys. Res. Ocean.* **2021**, *126*, e2021JC017676. [[CrossRef](#)]
35. Large, W.G.; Pond, S. Open Ocean momentum flux measurements in moderate to strong winds. *J. Phys. Oceanogr.* **1981**, *11*, 324–336. [[CrossRef](#)]
36. Wu, J. Wind-stress coefficients over sea surface from breeze to hurricane. *J. Geophys. Res. Ocean.* **1982**, *87*, 9704–9706. [[CrossRef](#)]
37. Hochreiter, S.; Schmidhuber, J. *Long Short-Term Memory*; Neural Computation MIT-Press: Cambridge, MA, USA, 1997.
38. Ergen, T.; Kozat, S.S. Unsupervised anomaly detection with LSTM neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 3127–3141. [[CrossRef](#)]
39. Hooker, S.B.; Lazin, G.; Zibordi, G.; McLean, S. An evaluation of above- and in-water methods for determining water-leaving radiances. *J. Atmos. Ocean. Technol.* **2002**, *19*, 486–515. [[CrossRef](#)]

40. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
41. AVISO/DUACS. SWOT Level-3 KaRIn Low Rate SSH Expert (v1.0). CNES. 2024. Available online: <https://doi.org/10.24400/527896/A01-2023.018> (accessed on 16 November 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.