Ca-STANet: Spatio-Temporal Attention Network for Chlorophyll-a Prediction with Gap-Filled Remote Sensing Data

Min Ye, Bohan Li, Jie Nie, Yuntao Qian, *Senior Member, IEEE*, and Lie-Liang Yang, *Fellow, IEEE*

***Abstract*—Long-term chlorophyll-a (Chl-a) prediction has the potential to provide an early warning of red tide, and support fishery management and marine ecosystem health. The existing learning-based Chl-a prediction methods mostly predict a single point or multiple points with monitoring data. However, the monitoring data are subject to sparse sampling and difficult to be measured in a large-scale and synchronous way. Moreover, the advanced learning-based models for point Chl-a prediction, such as long short-term memory (LSTM) and convolutional neural network (CNN)-LSTM, are unable to fully mining the spatio-temporal correlation of Chl-a variations. Therefore, by using the satellite remote sensing data with extensive coverage, we design a framework, namely Ca-STANet, to simultaneously predict the Chl-a of all the locations in a large-scale area from the perspective of spatio-temporal field. Specifically in our method, the original data are firstly divided into multiple sub-regions to capture the spatial heterogeneity of large-scale area. Then, two modules are respectively operated to mine the spatial correlation and long-term dependency features. Finally, the outputs from the two modules are integrated by a fusion module to fully mine the spatio-temporal correlations, which are exploited to attain the final Chl-a prediction. In this paper, the proposed Ca-STANet is comprehensively evaluated and compared with the legacy methods based on the OC-CCI Chl-a 5.0 data of the Bohai Sea. The results demonstrate that the proposed Ca-STANet is highly effective for Chl-a prediction and achieves higher prediction accuracy than the baseline methods. Moreover, as the OC-CCI Chl-a 5.0 data have many missing areas, we introduce DINEOF method to fill the data gaps before using them for prediction.**

***Index Terms*—Chlorophyll-a (Chl-a), spatio-temporal prediction, deep learning (DL), remote sensing data, spatio-temporal attention, convolutional neural network.**

I. Introduction

In the ocean, the level of chlorophyll-a (Chl-a) has been demonstrated to be a key indicator of ecosystem changes, since it is used as a proxy for phytoplankton biomass [1], and also a paramount variable for studying the environmental effects arose from the ocean dynamic process [2]. Moreover, Chl-a has significant impact on the stability of the ocean ecosystem, the exchange of carbon dioxide flux across the air-sea interface and the distribution of marine aquatic resources [3–5]. Therefore, long-term and reliable Chl-a prediction is highly important for the research of global carbon cycle and the utilization of fishery resources, as well as for the timely warning of red tide disasters [6].

Generally, Chl-a prediction can be deemed as a problem of multivariate time-series forecasting, which exploits the historical Chl-a data and exogenous factors (e.g., temperature and PH) to predict the future Chl-a. To date, the classical approaches on Chl-a prediction can be roughly classified into two categories, namely the time-series analysis method and the physics-based method. Specifically, in [7], authors investigated the algal bloom prediction based on an autoregressive integrated moving average (ARIMA) model. However, the spatial-temporal evolution of Chl-a is in general nonlinear due to various factors, such as sea surface temperature, wind speed, and the light transmittance of seawater [8, 9]. As the existing time-series analysis methods are designed mainly for extracting inherent characteristics via linear evolution, they are incapable of providing the accurate prediction of Chl-a.

On the other side, the physics-based methods conduct Chl-a prediction mainly based on the ecological dynamics and flow-diffusion equation [10]. For instance, in [11], authors introduced the Earth system model to predict the seasonal and multi-year ocean chlorophyll. The National Aeronautics and Space Administration (NASA) established a global biogeochemical prediction model, which was used to predict the global ocean chlorophyll over a 9-month period [12]. To date, the physics-based methods have been comprehensively developed with the motivation to achieve high resolution and simultaneously deal with multi-physical processes. However, we should mention that these physics-based methods usually demand accurate knowledge about a large number of variables as well as their complicated relationships, making it hard to build a comprehensive model. Furthermore, the errors accumulated during prediction may hinder the model from reliable long-term prediction in practical implementation.

Recently, the learning-based methods for Chl-a prediction have captured a lot of research attention, as the result that the techniques for Chl-a data acquisition have been well developed and hence a lot of data are available. Initially, the learning-based methods introduced for Chl-a prediction include, e.g., random forest, support vector machine, artificial neural net-

This work was supported in part by the National Natural Science Foundation of China under Grant 62072418 and Grant 62172376, and in part by the Fundamental Research Funds for the Central Universities under Grant 202042008. The work of Min Ye was supported by the Chinese Scholarship Council (CSC) for her research at the School of Electronics and Computer Science, University of Southampton, U.K. L.-L. Yang would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council project EP/X01228X/1.(Corresponding authors: Jie Nie)

Min Ye, Jie Nie are with the Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: yemin@stu.ouc.edu.cn; niejie@ouc.edu.cn)

Yuntao Qian is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: ytqian@zju.edu.cn)

Bohan Li and Lie-liang Yang are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: {bl2n18,lly}@ecs.soton.ac.uk)

work (ANN) [13–17]. These methods usually have a relatively small number of hidden layers and can be classified as the shallow learning-based methods (SLBM) [18]. Furthermore, the ensemble learning-based models have been designed for Chl-a prediction. For instance, in [19], different wavelet-ANNs were integrated with the aid of the least square boosting ensemble and Bates-Granger techniques, which are capable of achieving more reliable Chl-a prediction than individual ANN. However, the above-mentioned methods are in general only feasible for handling the low dimensional data with simple nonlinear associations. They are not efficient for representing the dynamic evolution of Chl-a [20].

By contrast, the deep learning (DL) methods, which use multiple processing layers to learn the representations of data via multiple levels of abstraction, can exploit the extremely intricate functions of inputs [21]. Thus, they outperform the SLBM and have become the mainstream approaches for Chl-a prediction. For example, in [22], the recurrent neural network (RNN) was introduced for predicting the Chl-a in the Nakdong River, and demonstrated to provide a higher prediction accuracy than the SLBM. The long short-term memory (LSTM) model was introduced to predict Chl-a in [23–26]. As an enhanced RNN model, the LSTM with the long-term memory obtained by several control gates is feasible for solving the sequence modeling problems. To be a little more specific, the authors of [23] used the LSTM for the multi-step Chl-a prediction in Gongju, South Korea. The studies suggested that the LSTM is a high-efficiency model, which can exploit the temporal Chl-a variation. To improve the accuracy of prediction, the authors of [27] performed the Chl-a prediction at the Prespa lake using the hybrid convolutional neural network (CNN)-LSTM model. The studies revealed that the hybrid CNN-LSTM model outperforms the individual CNN or LSTM models in terms of the prediction accuracy.

The aforementioned learning-based methods mainly predict Chl-a at a single point or multiple points, as shown in Fig. 1. They are not suitable for the large-scale area Chl-a prediction, due to the fact that only a small number of points are involved in a model and each point is individually fed to the model for training. Thus, the spatio-temporal correlations within the whole area can not be fully exploited. To achieve efficient Chl-a prediction in large-scale areas, the spatio-temporal field, as illustrated in Fig. 1, combined with the DL methods are proposed for Chl-a prediction in this paper. Explicitly, the large-scale area Chl-a prediction requires continuous spatio-temporal grid data, while the on-site sampling and buoy-based data, which are difficult to be measured in a large-scale and synchronous way, can not satisfy this requirement. Therefore, in our studies, the satellite remote sensing data, which have the features of extensive coverage and the capability of realizing spatio-temporal continuous observation, will be used to construct the spatio-temporal grid data for Chl-a prediction.

However, in practice, the cloud obscuring measurements and the contamination of high sun glint may result in the loss of some satellite ocean color data of Chl-a. Hence, filling the remote sensing data gaps is indispensable for understanding the hydrodynamics and biophysical interactions of the ocean. In the open literature, various methods have been

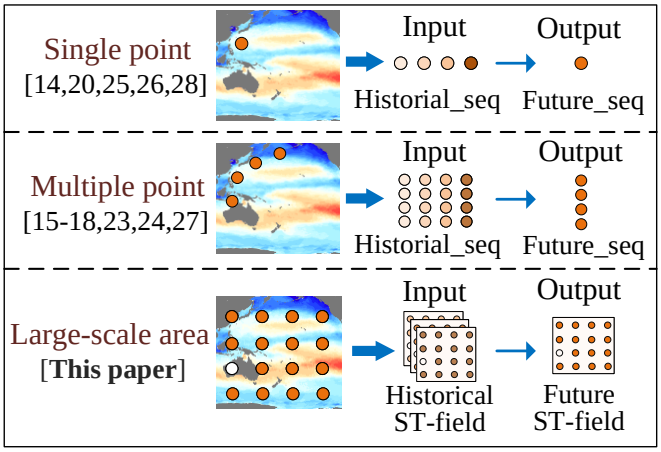

Fig. 1. Typical Chl-a prediction categories classified based on the learning-based methods. Historical_seq means the time-domain historical sequences, Future_seq means the future sequences, and ST-field means the spatio-temporal field.

introduced for recovering missing data, including the spline interpolation [28], optimal interpolation (OI) [29], Kriging interpolation [30], Data Interpolating Empirical Orthogonal Functions (DINEOF) [31], and the DL methods [32]. As demonstrated in [33], DINEOF can achieve similar filling results as the OI approach, but it is 30 times faster. Henn et al. [34] compared DINEOF with a temporal interpolation method and other three methods based on spatial interpolation. It was observed that DINEOF is the most accurate method for missing data recovery in large-scale areas. Finally, it is shown [32] that the CNN-relied DL approach can achieve better performance than the DINEOF method, however, it depends on a large amount of evenly distributed in situ data for training. In addition, the DINEOF method has been widely applied in reconstructing Chl-a data, such as Moderate Resolution Imaging Spectroradiometer (MODIS) and Sea-viewing Wide Field-of-view Sensor (SeaWiFS), in the Bohai and Yellow Seas [35–37]. Therefore, in this paper, the DINEOF method is introduced to fill the missing data in the Bohai Sea.

Having the spatio-temporal relied data prepared, the next step is to introduce a predictor, which can efficiently use the data to make a reliable Chl-a prediction in large-scale areas. To this regard, we should note that the above-mentioned advanced LSTM and CNN-LSTM models for point-source Chl-a prediction are unable to fully mine the spatio-temporal correlation of the Chl-a variations. By contrast, the DL-assisted spatio-temporal methods have received a lot of research attention for prediction in large-scale areas. In [38], the authors proposed a multiscale CNN network for forecasting the sea surface temperature (SST) in the eastern equatorial Pacific Ocean. This kind of methods [38–40] developed based on CNN can effectively improve the mining ability of spatial correlation within the whole area through the multiple convolution and downsampling operations. Moreover, the authors of [41] proposed a multilayer fusion RNN, which uses a cell to fuse the global and local spatio-temporal features, to predict the sea surface height anomaly (SSHA) areas. It was shown that the RNN assisted methods [41–43], such as Convlstm and

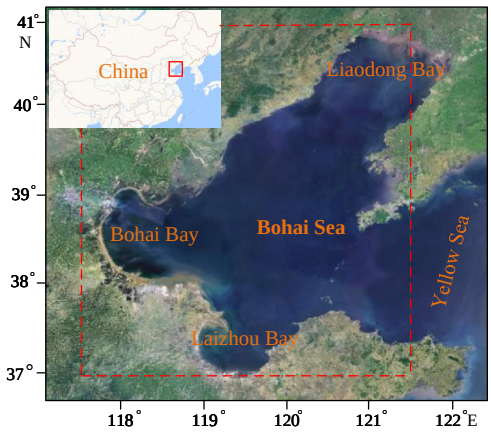

Fig. 2. Area of Bohai Sea studied (in the red rectangle).

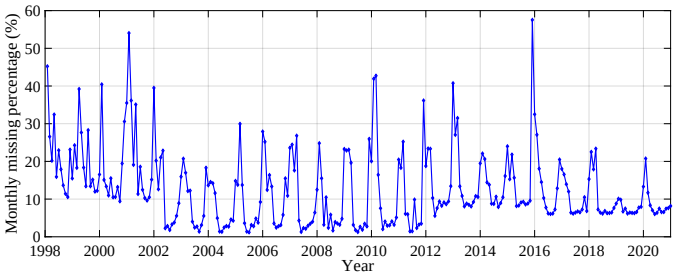

Fig. 3. Time series showing the area-averaged missing rate of the monthly Chl-a data for the Bohai sea.

MLFrnn, are capable of extracting the information of spatio-temporal evolutions from neighboring grids.

However, the above-mentioned DL-assisted spatio-temporal prediction methods cannot fully exploit the evolution features in the respective domains of space and time. More recently, the authors of [44] proposed a deep generative model with the dual spatial and temporal discriminator for precipitation nowcasting, showing the improved forecast quality and forecast consistency. Inspired by the work in [44] and combined with the spatio-temporal characteristics of Chl-a evolution, in this paper, we design a framework to simultaneously predict the Chl-a of all the locations in a large-scale area by a spatio-temporal feature fusion network, which is termed as the Ca-STANet. In our method, the original data are first divided into multiple sub-regions to capture the spatial heterogeneity in large-scale area prediction. Then, for each sub-region, two modules, namely the spatial attention feature extraction (SAFE) and temporal attention feature extraction (TAFE), are constructed to enhance the feature extraction of the spatial variation and temporal dependency. Moreover, each of the sub-regions is trained independently so as to avoid the possible catastrophic forgetting caused by a single model based training. In our double-module based training, the SAFE module introduces a spatial attention mechanism based on the sub-region's inputs and global inputs, aiming at extracting the features of the spatial variation within a sub-region, but at the same time, exploiting the global dynamic correlation between sub-regions. By contrast, the TAFE module focuses on extracting the features in the time-domain to capture the long-term dependency existing in the data. After that, the outputs from the two modules are integrated by a fusion module, where the spatio-temporal correlation is fully mined for the Chl-a prediction. Finally, the predicted results from sub-regions are combined to provide the Chl-a prediction of the whole large-scale area.

The main contributions of this paper are summarized as follows:

1) Different from the existing point source Chl-a prediction, the Chl-a prediction in large-scale areas is investigated with the aid of the DL models operated in spatio-temporal field.

2) The DINEOF method is introduced to fill the 5-day and monthly missing Chl-a data obtained by the remote sensing OC-CCI 5.0 in Bohai sea. The impact of the gap-filled data in winter on the Chl-a prediction is analyzed.

3) A Ca-STANet is proposed to provide the Chl-a prediction in large-scale areas. In our Ca-STANet, an original large-scale area is first divided into multiple sub-regions to capture the spatial heterogeneity. Then, the SAFE and TAFE modules are constructed to enhance the feature extraction of spatial variation and temporal dependency. Finally, a fusion module is used to mine the spatio-temporal correlation from the outputs provided by the two modules.

4) The effectiveness of the proposed method is validated by comparing our method with the state-of-the-art prediction methods, demonstrating that our method is capable of efficiently mining the spatio-temporal correlations and hence, providing higher accuracy of Chl-a prediction.

The remainder of the paper is organized as follows. Section II presents the study area and data sets. Section III details the scheme for preparing the Chl-a data and the proposed Ca-STANet framework for Chl-a prediction. Section IV reports the experimental results and provides our analysis. Section V further analyzes the prediction results of winter data and their influence on the model. Finally, in Section VI, we summarize the research observations and discuss the possible future research issues.

II. Data

A. Study Area

Bohai Sea is located in the north of China (bounded by 37°-41°N, 117.5°-121.5°E), as shown in Fig.2. It is a shallow shelf sea, with an average water depth of about 18 m and a total area of about 77,000 km^2. Under the influence of human activities in Bohai Rim area, the eutrophication of seawater has become serious and the phytoplankton biomass has increased rapidly. The distribution of Chl-a shows a pattern of high concentration in the shallow waters near land, and the Chl-a gradually decreases when moving away from the shores [45].

B. Data Sets

In this study, the 5-day and monthly surface Chl-a 5.0 data of Bohai sea with a spatial resolution of 4 km were downloaded from the Ocean Colour Climate Change Initiative (OC-CCI), which is available at http://www.esa-oceancolour-cci.org. The Chl-a 5.0 dataset is obtained by merging the data from Medium Resolution Imaging Spectrometer (MERIS),

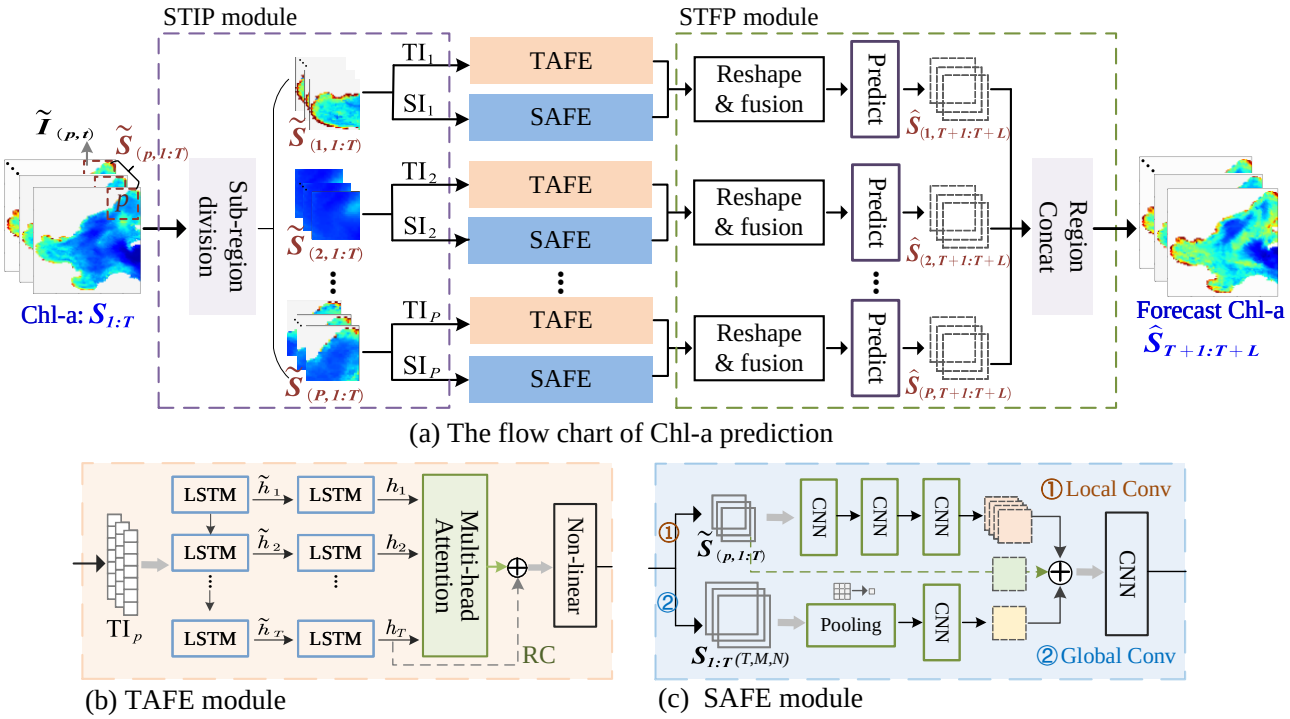

Fig. 4. Overall network structure of Ca-STANet, including mainly the STIP module, TAFE module, SAFE module, and STFP module.

MODIS, SeaWiFS, and Visible Infrared Imaging Radiometer Suite (VIIRS) data. In this paper, the Chl-a dataset covers the period from 1998 to 2020.

III. Methods

A. Data Pre-processing and Gap Filling

The loss of Chl-a data due to clouds and sun glint are different in different time periods. It is found that during the 23 years considered, the mean missing rate of 5-day and monthly Chl-a data are 45.8% and 12.6%, respectively. For example, Fig. 3 shows the missing percentage of monthly Chl-a data in the Bohai Sea, which is uneven in both time and space. In particular, sea ice and heavy clouds contribute to the high loss rate of Chl-a data in winter. In order to avoid the loss of important local information and achieve more accurate prediction, the missing data should be appropriately reconstructed.

Therefore, the DINEOF method [31] is introduced to first reconstruct the missing data. In principle, DINEOF uses an empirical orthogonal function (EOF) for reconstructing the missing data in a geophysical data set, via deciphering the dominant variability modes within the data [46]. The advantages of this method include that: a) it does not require the priori information about the de-correlation scale, and b) it can exploit the multiple data types with inherent correlation to increase data coverage.

The DINEOF method can be described by the following four steps:

1) Initial data processing: Consider an initial matrix $\bar{\boldsymbol{X}}$, where the rows denote the number of spatial points and the columns denote the length of time. A data matrix $\boldsymbol{X}$ is then obtained by subtracting the mean value from $\bar{\boldsymbol{X}}$, while setting the initially missing data to zero.

2) Iterative data replacement: An EOF decomposition is performed on $\boldsymbol{X}$, based on which the values in $\boldsymbol{X}$ are updated using the following equation:

$$\boldsymbol{X}_{i,j} = \sum_{k=1}^{K} \rho_k \left(\boldsymbol{u}_k\right)_i \left(\boldsymbol{v}_k^T\right)_j \tag{1}$$

where i and j are the spatial and temporal indexes of the missing data in $\boldsymbol{X}$, K represents the number of EOF modes, $\boldsymbol{u}_k$ and $\boldsymbol{v}_k^T$ are the spatial and temporal functions of the corresponding EOF mode, and ρ_k is the corresponding singular value. The above EOF decomposition and data updating of $\boldsymbol{X}$ are repeated until convergence is achieved. In our study, the convergence is assumed to be achieved, when the maximum error between the input $\boldsymbol{X}_{i,j}$ values and their reconstruction ones reaches a Lanczos convergence threshold of 10^{-8}.

3) Finding the optimal number of EOF modes: Step 2) is repeated with $K = 1, 2..., K_{max}$ and the correspondingly reconstructed matrices $\boldsymbol{X}_K$ are obtained. Then, the optimal number of EOF modes (denoted as K_{opt}) is obtained through a cross-validation technique [47]. Specifically, K_{opt} is chosen as the K value that minimizes the error between the data set left for validation and their correspondingly reconstructed values.

4) Gap-filling of missing data: Once K_{opt} is obtained, Steps 1) and 2) are operated again on the basis of K_{opt} and $\bar{\boldsymbol{X}}$ to obtain a $\boldsymbol{X}$, where $\bar{\boldsymbol{X}}$ also includes the data previously set aside for cross-validation. Finally, the mean value of $\bar{\boldsymbol{X}}$ is added back to $\boldsymbol{X}$ to obtain the finally reconstructed gap-free data.

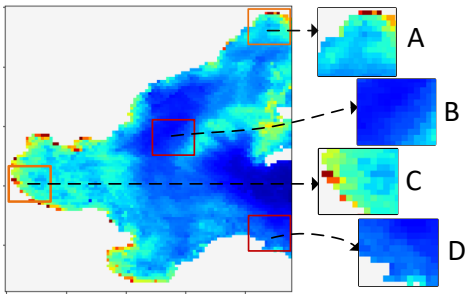

Fig. 5. Example of different levels of spatial heterogeneity throughout the containing sub-regions.

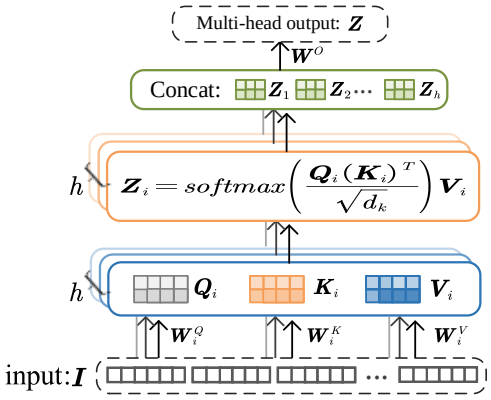

Fig. 6. Structure of multi-head attention mechanism.

B. Formulation of the Chl-a Prediction Problem

We assume a spatial region represented by an $M \times N$ grid, which consists of M rows and N columns. Let $\boldsymbol{S}_{1:T} = (\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_T)$ be the dynamic Chl-a sequence of length T, where $\boldsymbol{I}_t \in \mathbb{R}^{M \times N}$ contains the Chl-a values of the t-th frame. Hence, $\boldsymbol{S}_{1:T}$ is a 3-dimensional vector of size (T, M, N). The goal of Chl-a prediction is to predict a future Chl-a sequence up to L steps subsequent to the historical satellite remote sensing sequence $\boldsymbol{S}_{1:T}$, which can be formulated as:

$$\widehat{\boldsymbol{S}}_{T+1:T+L} = \arg\max_{\boldsymbol{S}_{T+1:T+L}} p\left(\boldsymbol{S}_{T+1:T+L} \mid \boldsymbol{S}_{1:T}\right) \quad (2)$$

where $\widehat{\boldsymbol{S}}_{T+1:T+L} = \left(\hat{\boldsymbol{I}}_{T+1}, \ldots, \hat{\boldsymbol{I}}_{T+L}\right)$ represents the predicted Chl-a sequence of length L.

C. Structure of Prediction Model

This section describes in detail the proposed Ca-STANet, which realizes reliable long-term Chl-a prediction by making efficient use of the spatio-temporal correlation features existing in the satellite sensed Chl-a data. The overall structure of Ca-STANet is shown in Fig. 4(a). The inputs to the Ca-STANet are the historical data $\boldsymbol{S}_{1:T}$, as previously defined. As shown in Fig. 4(a), $\boldsymbol{S}_{1:T}$ is firstly divided into sub-regions to tackle the spatial heterogeneity in the spatio-temporal input preprocessing (STIP) module. Then, for each sub-region, the SAFE module and TAFE module are respectively operated to extract the features of spatial variation and temporal dependency. After that, a fusion module is used to mine the spatio-temporal correlations for the Chl-a prediction, which is referred as the spatio-temporal fusion prediction (STFP) module. Finally, the results from multiple sub-regions are concatenated to provide the Chl-a prediction of the whole region considered. Below we analyze these modules one-by-one in detail.

1) STIP module: The objective of this module is to subdivide the original area and generate the required inputs for the TAFE and SAFE modules. From the Chl-a data of Bohai sea, we noticed that the spatial relationship within each sub-region may vary greatly. As illustrated in Fig. 5, the point Chl-a values in the sub-regions B and D of the offshore water far away from land are similar and have a low level of fluctuation. They show a high degree of positive correlation between nearby points and even between the distant points. By contrast, in the sub-regions A and C located at the shallow waters near land, the point Chl-a values are usually different and vary significantly in different seasons. These different correlation relationships existing in the various sub-regions imply the phenomenon of spatial heterogeneity, which is defined as the uneven distribution of a trait, event, or relationship over a region [48]. However, the spatial heterogeneity of sub-regions may be overlooked, if the data of a whole region are used as the inputs. To solve this problem, in our method, each of the sub-regions is trained independently, so as to capture the detailed spatial heterogeneity, while also avoiding the occurrence of the possible catastrophic forgetting caused by using a single model based training for all sub-regions.

In detail, each $\boldsymbol{I}_t$ of the original Chl-a inputs, which provides the data of a whole area, is divided into P sub-regions. Considering the rectangular type of inputs of CNN layers in the SAFE module and the computational efficiency, the area of Bohai sea is evenly segmented into P sub-regions. The data of these sub-regions are denoted as $\widetilde{\boldsymbol{I}}_{(p,t)}$ with $p = 1, \cdots, P$, and we have $\boldsymbol{I}_t = \cup_{p=1}^{P} \widetilde{\boldsymbol{I}}_{(p,t)}$. For one sub-region p, the historical sequences are now expressed as $\widetilde{\boldsymbol{S}}_{(p,1:T)} = \left(\widetilde{\boldsymbol{I}}_{(p,1)}, \widetilde{\boldsymbol{I}}_{(p,2)}, \ldots, \widetilde{\boldsymbol{I}}_{(p,T)}\right)$, as shown in left-most box of Fig. 4(a). The historical sequence $\widetilde{\boldsymbol{S}}_{(p,1:T)}$ and the global historical sequence $\boldsymbol{S}_{1:T}$ are encapsulated as the inputs to the SAFE module, which are expressed as SI_p in Fig. 4(a). On the other side, an $\widetilde{\boldsymbol{S}}_{(p,1:T)}$ is converted to the 2-dimensional vector of size $(T, (M \times N)/P)$, which produces the inputs to the TAFE module, shown as TI_p in Fig. 4(a). Let us now detail the operations of the TAFE and SAFE modules.

2) TAFE Module: The purpose of the TAFE module is mainly to capture the long-term dependency of sub-regions and generate temporal features for the following STFP module. As shown in Fig. 4(b), the TAFE module consists of the LSTM layers, a multi-head attention layer, and a "Non-linear" block. It is well recognized that the LSTM architecture is feasible for capturing the long-term dependency in a sequential pattern. Hence, in this study, we use LSTM as a building block for the TAFE module. Specifically, in the TAFE module, the hidden vectors $\boldsymbol{H}_T = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T\}$ are first obtained with the aid of a two-layer LSTM network, as shown in Fig. 4(b). Then, the multi-head attention layer is executed to learn the internal characteristics and time-dependence between different $\boldsymbol{h}_t$.

After this learning, the weights for different $\boldsymbol{h}_t$ are reassigned to make the network focus on the relatively important time features. This is achieved via the residual connection (RC) [49] and the output of the multi-head attention layer shown in Fig. 4(b). Finally, we express the output of the TAFE module as $\boldsymbol{T}_{\text{tem}}$, which is the output of the "Non-linear" block. Note that, the "Non-linear" block includes the fully connected layers associated with the rectified linear unit (ReLU) function [50].

More details related to the multi-head attention layer, defined as an ensemble of h different self-attention blocks, can be found in Fig. 6. Specifically, the input matrix $\boldsymbol{I}$ is first mapped to the h different subspaces through linear transformation as [51]:

$$\begin{aligned}[\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_h] &= \left[\boldsymbol{I}\boldsymbol{W}_1^Q,\ldots,\boldsymbol{I}\boldsymbol{W}_h^Q\right]\\ [\boldsymbol{K}_1,\ldots,\boldsymbol{K}_h] &= \left[\boldsymbol{I}\boldsymbol{W}_1^K,\ldots,\boldsymbol{I}\boldsymbol{W}_h^K\right]\\ [\boldsymbol{V}_1,\ldots,\boldsymbol{V}_h] &= \left[\boldsymbol{I}\boldsymbol{W}_1^V,\ldots,\boldsymbol{I}\boldsymbol{W}_h^V\right]\end{aligned} \tag{3}$$

where $\boldsymbol{Q}_i, \boldsymbol{K}_i$ and $\boldsymbol{V}_i$, $i \in [1,h]$, are the query, key, and value matrices of each subspace, $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V$, $i \in [1,h]$, are the learnable conversion matrices.

Then, as shown in Fig. 6, the attention value of each subspace is computed as:

$$\boldsymbol{Z}_i = \text{softmax}\left(\frac{\boldsymbol{Q}_i(\boldsymbol{K}_i)^T}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{4}$$

where $\boldsymbol{Z}_i$ is the attention matrix of the i-th subspace, and $\sqrt{d_k}$ is applied to change the attention matrix to follow the standard normal distribution so as to achieve the gradient stability. As shown in Fig. 6, $\boldsymbol{Z}_i$ is obtained by multiplying $\boldsymbol{V}_i$ and the attention coefficient that is generated by the softmax. The inputs $\boldsymbol{I}$ are given by the hidden vectors in $\boldsymbol{H}_T$, as previously defined. In this way, the short and long-term memory features generated by LSTM are further processed in the multi-head attention layer.

Finally, the attention matrices $\boldsymbol{Z}_i$ are concatenated and projected to produce the multi-head output as:

$$\boldsymbol{Z} = \text{Concat}\left(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_h\right)\boldsymbol{W}^O \tag{5}$$

where $\boldsymbol{W}^O$ is the learnable weight matrix and Concat$(\cdot)$ is the concatenation operation.

Note that in our work, we set $h = 2$ and $\sqrt{d_k}$ =32.

3) SAFE Module: The function of this module is mainly to capture the spatial correlation existing in a sub-region and between sub-regions and to generate spatial features for the following STFP module. Our SAFE module is designed to include both the 'Local Conv' and 'Global Conv' as noted in Fig. 4(c). The 'Local Conv' pays attention on extracting the features of the spatial variation within a sub-region, while the 'Global Conv' emphasizes on deriving the global spatial correlation between sub-regions. Furthermore, as the inputs to the SAFE module represent the historically evolving data, the SAFE module can also mine the spatio-temporal correlation of the input data of different sub-regions.

To be more specific, as shown in Fig. 4(c), the $\widetilde{\boldsymbol{S}}_{(p,1:T)}$ of SI_p, as previously defined in the STIP module as seen in Fig. 4(a), are input to the 'Local Conv' to extract the spatial correlation features within the sub-regions. The 'Local Conv' consists of three cascaded CNN layers, each of which is followed by a ReLU function. The convolution kernel used in the first CNN layer has the size of 5×5, while that used in the following two CNN layers has the size of 3×3. The number of convolution kernels of these three cascaded CNN layers are 20, 20, and 30, respectively. By contrast, as the 'Global Conv' sub-module is introduced to extract the Chl-a variation between different sub-regions, the data input to the sub-module are only the $\boldsymbol{S}_{1:T}$. The functions of this sub-module are achieved by one downsampling layer and one CNN layer. The downsampling layer uses the AvgPool to compress information and the convolution kernel size of the CNN layer is 3×3. Additionally, considering that the input Chl-a data at the present time step has a higher correlation with the future Chl-a data, the input Chl-a data at the present time step, i.e., $\widetilde{\boldsymbol{I}}_{(p,T)}$ in $\widetilde{\boldsymbol{S}}_{(p,1:T)}$, shown as the green feature map in Fig. 4(c), is directly linked to the last CNN layer along with the features obtained from both the 'Local Conv' (shown by the orange feature map in Fig. 4(c)) and the 'Global Conv' (shown by the yellow feature map in Fig. 4(c)). The last CNN layer seen in Fig. 4(c) applies a kernel with the size of 3×3. This CNN layer accomplishes the feature extraction of spatial correlation and prediction, yielding the output expressed as $\boldsymbol{S}_{\text{spa}}$.

4) STFP Module: Finally, the STFP module as shown in Fig. 4(a) is designed to incorporate the outputs $\boldsymbol{T}_{\text{tem}}$ by TAFE with the outputs $\boldsymbol{S}_{\text{spa}}$ of SAFE to mine the overall spatio-temporal correlation, and provide the Chl-a prediction for the whole lead time series. In detail, in the STFP module shown in Fig. 4(a), for each sub-region relied model, both $\boldsymbol{T}_{\text{tem}}$ and $\boldsymbol{S}_{\text{spa}}$ are input to the "Reshape&fusion" block to integrate the spatio-temporal features. In each of these blocks, $\boldsymbol{T}_{\text{tem}}$ with the size of $(C, (M \times N)/P)$ is first structured to have the same size as $\boldsymbol{S}_{\text{spa}}$, which is $(C, \frac{M}{\sqrt{P}}, \frac{N}{\sqrt{P}})$, where C represents the number of channels. Then, the two feature maps are concatenated to form a new feature map of size $(2C, \frac{M}{\sqrt{P}}, \frac{N}{\sqrt{P}})$, which is sent to the " Predict" block.

The "Predict" block is functioned to predict the future Chl-a values of each sub-region via the association mining of the spatio-temporal fusion features. More specifically, the high-dimensional association mining of the spatio-temporal evolution features is achieved by the multiple cascaded CNN layers, which have the 3×3 sized convolution kernels. The prediction results with the size of $(L, \frac{M}{\sqrt{P}}, \frac{N}{\sqrt{P}})$, namely $\widehat{\boldsymbol{S}}_{(p,T+1:T+L)}$ shown in Fig. 4(a), are the outputs of the CNN layers. Finally, the Chl-a predictions of multiple sub-regions are joined in the "Region Concat" block to complete the future Chl-a prediction in the whole area considered. For clarity, the implementation procedures of our Ca-STANet for the Chl-a prediction are summarized in Algorithm 1.

IV. Experiments and Results

A. Gap Filling for 5-day and Monthly Chl-a Data

Before gap filling, the parts with more than 90% of missing data in the 5-day and monthly Chl-a data are filtered to ensure the accuracy of later reconstruction. Furthermore, in our study, 5% of the filtered data are randomly selected to form the cross-validation set so as to evaluate the performance of DINEOF.

Algorithm 1: Ca-STANet Algorithm for Chl-a prediction

Input: Historical Chl-a sequence $\boldsymbol{S}_{1:T}$, No. of sub-regions P, maximum epoch γ, learning rate λ, and prediction step size L.

Output: Future Chl-a sequence $\widehat{\boldsymbol{S}}_{T+1:T+L}$.

Step 1: ***Spatio-temporal input preprocessing (STIP)***

Construct the Chl-a sequence $\widetilde{\boldsymbol{S}}_{(p,1:T)}$, $p = 1, \cdots, P$ via sub-region division;

Construct the inputs TI_p and SI_p for each sub-region.

Step 2: ***Training Ca-STANet***

for $p = 1, ..., P$, **do**

  Initialize TAFE, SAFE and STFP modules;

  while *training epoch is less than* γ, **do**

    1) Generate temporal features $\boldsymbol{T}_{\text{tem}}$ using TAFE module;

    2) Generate spatial features $\boldsymbol{S}_{\text{spa}}$ using SAFE module;

    3) Generate spatio-temporal correlations by fusing $\boldsymbol{T}_{\text{tem}}$ and $\boldsymbol{S}_{\text{spa}}$ using STFP module;

    4) Predict $\widehat{\boldsymbol{S}}_{(p,T+1:T+L)}$ and update modules' parameters using gradient descent.

  end

end

Output $\widehat{\boldsymbol{S}}_{T+1:T+L}$ via concatenating $\widehat{\boldsymbol{S}}_{(p,T+1:T+L)}$, $p = 1, \cdots, P$.

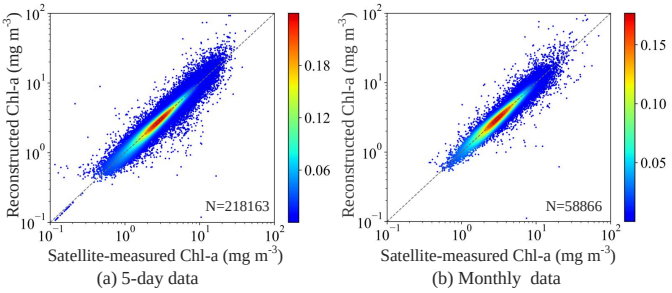

Fig. 7. Scatterplots of satellite-measured Chl-a data versus reconstructed Chl-a data: (a) 5-day Chl-a data and (b) monthly Chl-a data.

After cross-validation, this 5% of data are then included to form the gap-free data for the following Chl-a prediction.

To evaluate the accuracy of the reconstructed Chl-a data, the cross-validation set of data after the DINEOF process are plotted against the original satellite-measured data, as shown in Fig. 7, for both the 5-day and monthly Chl-a data. Here, the density of scatters is expressed by a kernel density estimation using the Gaussian kernel. The satellite-measured 5-day and monthly Chl-a data vary in the range of 0.11-93.08 ($mg \cdot m^{-3}$) and 0.11-92.47 ($mg \cdot m^{-3}$), respectively. It can be seen from Fig. 7 that the data scatter around the 1:1 line in both Fig. 7(a) and Fig. 7(b), with the root mean square error (RMSE) of 0.88 and 1.25 ($mg \cdot m^{-3}$), respectively. Besides, the coefficient of determination (R^2) [52] is 0.89 for the 5-day data and 0.83 for the monthly data. Furthermore, if we normalize the distribution using a natural logarithm transformation, the RMSE is 0.06 and 0.07, respectively, while R^2 is 0.93 for the 5-day data and 0.92 for the monthly data. Based on these results, we can

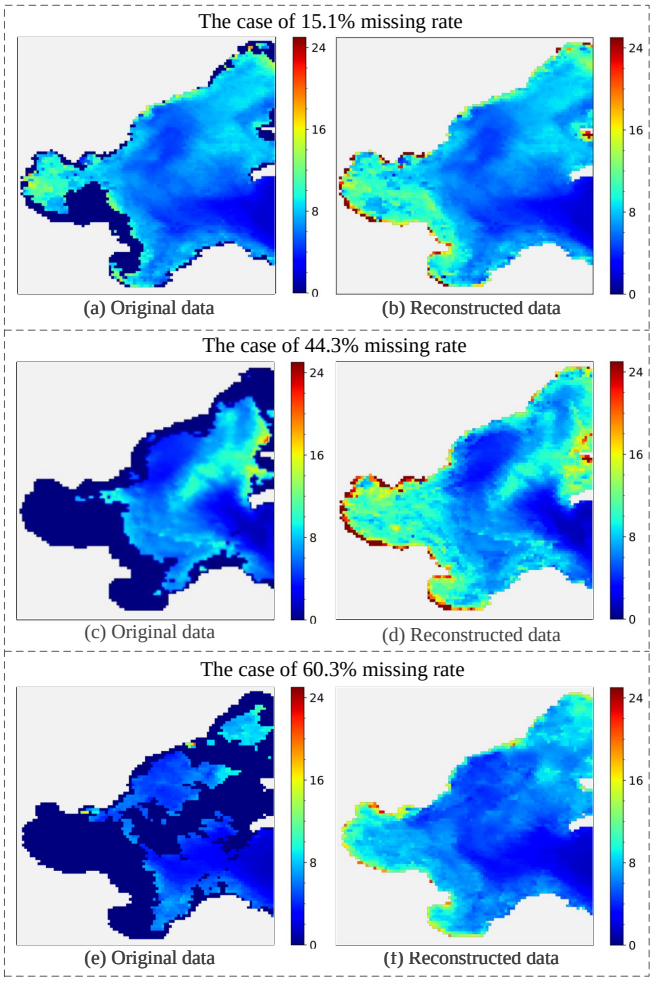

Fig. 8. Original monthly (a, c) and 5-day (e) Chl-a data, and their reconstructions (b, d and f).

confident that DINEOF performs well in the reconstruction of the Chl-a data, which can be further exploited to analyze the spatial and temporal patterns in the Chl-a data of Bohai sea.

The original and reconstructed Chl-a data for three scenarios with different missing rates are shown in Fig. 8. The gray background represents the land or sea areas not involved in the reconstruction. The dark blue areas in Fig. 8(a), 8(c) and 8(e) are the missing parts needing reconstruction. The data missing rates for these scenarios are 15.1%, 44.3% and 60.3%, respectively. On the right-side, Fig. 8(b), 8(d) and 8(f) give the correspondingly reconstructed Chl-a data. From Fig. 8, it is clear that under various missing rates, the spatial distribution of Chl-a can be reconstructed with high accuracy.

B. Experiment Settings for Prediction

Next, we show the performance of Chl-a prediction. Our forecasting algorithm is operated on a single NVIDIA 2080TI GPU. During training, the mean square error (MSE) as the loss function and the adaptive moment estimation (Adam) as the optimizer are applied. After reconstructing the corresponding Chl-a data, we select 70% of the data as the training samples, while the remaining 30% data are used as the testing data.

TABLE I
SUB-REGION EXPERIMRNT WITH MONTHLY DATA

Methods	Metrics	Prediction Months		
		1-step	3-step	6-step
whole-area	RMSE	1.91	2.39	2.41
	MAE	1.16	1.63	1.60
	PCorr	0.880	0.839	0.840
4 sub-regions	RMSE	1.83	2.15	2.17
	MAE	1.09	1.38	1.40
	PCorr	0.881	0.848	0.847
9 sub-regions	RMSE	1.82	2.12	2.11
	MAE	1.09	1.32	1.33
	PCorr	0.879	0.850	0.849
16 sub-regions	RMSE	1.85	2.17	2.13
	MAE	1.09	1.33	1.29
	PCorr	0.874	0.848	0.851

Furthermore, the training and testing samples are chosen in the order of time rather than randomly, where the testing samples are the future samples relative to the training samples. The learning rate is set to 0.001 for the monthly data and 0.0001 for the 5-day data. The size of the mini-batch is set to 8, and the model is trained by 40 epochs. Inspired by [39], which used the sea surface temperature based spatio-temporal sequences of three consecutive months to predict the ENSO, our model uses the Chl-a data of three consecutive steps as the inputs and the direct strategy to perform the multistep-ahead forecasting. More specifically, for the 5-day Chl-a data, the input data are constructed in the format: [1485, 3, 96, 96], where the first number is the number of the Chl-a samples, the next number is the length T of the historical Chl-a sequence, and the last two numbers are the height M and width N of the input data, respectively. Similarly, for the monthly Chl-a data, the input data are constructed in the format: [273, 3, 96, 96].

1) Evaluation Metrics: RMSE, mean absolute error (MAE) and Pearson correlation coefficient (PCorr) are used to assess the performance of the Ca-STANet, which are evaluated from the formulas of:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(p_i - o_i\right)^2}{n}} \tag{6}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|p_i - o_i|}{n} \tag{7}$$

$$\text{PCorr} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \sum(o_i - \bar{o})^2}} \tag{8}$$

where p_i is the predicted Chl-a value and o_i is the Chl-a value sensed by satellite, whose averages are denoted by $\bar{p}$ and $\bar{o}$, respectively, while n is the total number of testing samples.

2) Baseline Models: The following learning-based methods are selected as the baselines for comparison with our model, which are:

(a) CNN [53]: Note that the CNN model is slightly modified by removing the fully-connected layer and a max-pooling layer, when the Chl-a in large-scale area is predicted in our study.
(b) LSTM-S [54]: LSTM-S model uses all the points in a field as one sample. However, in our study of simultaneously considering a large number of points, the model easily leads to memory overflow. To solve this problem, the original area is divided into two sub-areas for training.
(c) CNN-LSTM [27]: The CNN-LSTM, consisting of three convolutional layers, a flattening layer, and an LSTM layer, was proposed to predict two water quality variables, i.e., dissolved oxygen and Chl-a.
(d) ConvLSTM [43]: ConvLSTM was proposed to overcome the drawbacks of LSTM in handling the spatio-temporal data. The ConvLSTM-assisted Chl-a prediction model is the same as that in [43].

C. Setting of Parameter P

In this subsection, the influence of the number of sub-regions P on the prediction performance is studied. It can be understood that using the whole-area sequences as inputs can also achieve good performance for the large-area Chl-a prediction, if there are enough training samples available. However, when only thousands of 5-day or monthly data are available for Chl-a prediction, the performance achieved is generally limited, which can be improved by dividing the whole region into sub-regions. Table I shows the performance results of the monthly prediction, when different numbers of sub-regions are employed. It can be seen that the performance of using 4, 9 or 16 sub-regions is superior to that of using the whole-area, i.e, 1 region. Furthermore, it is observed that there exists an appropriate number of sub-regions that results in the best performance. As seen in Table I, for the cases considered, this is given by the case with 9 sub-regions. In this case, the Ca-STANet is capable of making a relatively clear distinction between the sub-regions in the shallow waters with high fluctuation and those far away from land with a low variation. By contrast, if there are too many tiny sub-regions, some of the neighborhood spatial information may be missed by the Ca-STANet, yielding the degraded prediction accuracy. Furthermore, a bigger number of sub-regions results in the expansion of model parameters, which in turn results in bigger complexity for the model training. Therefore, in our following studies, we set P to 9.

D. Experimental Results and Analysis

Now we demonstrate and compare the Chl-a prediction results by different methods. During the experiments, all the baseline models are trained using the same hyperparameters and prediction strategy, as mentioned during the Ca-STANet training in Section IV-B.

Table II shows the RMSE, MAE and PCorr values obtained by the different prediction algorithms. From Table II, we can see that for the various steps of predictions based on the 5-day data and monthly data, CNN achieves the poorest performance, as it does not exploit the chronological correlation of the input data. LSTM-S performs slightly better than CNN, especially for the long-term prediction. This is because LSTM-S has the ability to mine the long-term dependency of the Chl-a data by its recurrent network structure and gating mechanisms. CNN-LSTM and ConvLSTM, which make use of the spatio-temporal correlation of input data, outperform both the CNN, which only mines the spatial correlation, and

TABLE II
PREDICTION RESULTS ON THE CHL-A DATA SETS OF BOHAI SEA

Models	Metrics	Prediction 5-days				Prediction Months		
		1-step	3-step	6-step	9-step	1-step	3-step	6-step
CNN	RMSE	1.86	1.96	2.11	2.18	2.31	2.69	2.84
	MAE	1.33	1.41	1.56	1.64	1.52	1.81	1.90
	PCorr	0.848	0.822	0.798	0.787	0.870	0.799	0.770
LSTM-S	RMSE	1.85	1.89	2.00	2.06	2.67	2.62	2.59
	MAE	1.22	1.24	1.32	1.38	1.74	1.68	1.67
	PCorr	0.828	0.826	0.820	0.809	0.829	0.812	0.814
CNN-LSTM	RMSE	1.74	1.74	1.80	2.02	2.11	2.47	2.44
	MAE	1.25	1.26	1.31	1.58	1.41	1.72	1.69
	PCorr	0.836	0.826	0.824	0.809	0.856	0.827	0.827
ConvLSTM	RMSE	**1.46**	1.62	1.83	2.01	1.90	2.55	2.76
	MAE	**0.98**	1.12	1.33	1.52	1.14	1.72	1.91
	PCorr	0.857	0.833	0.807	0.787	0.870	0.795	0.775
Ca-STANet	RMSE	1.47	**1.58**	**1.73**	**1.85**	**1.82**	**2.12**	**2.11**
	MAE	1.00	**1.11**	**1.27**	**1.39**	**1.09**	**1.32**	**1.33**
	PCorr	**0.866**	**0.849**	**0.834**	**0.826**	**0.879**	**0.850**	**0.849**

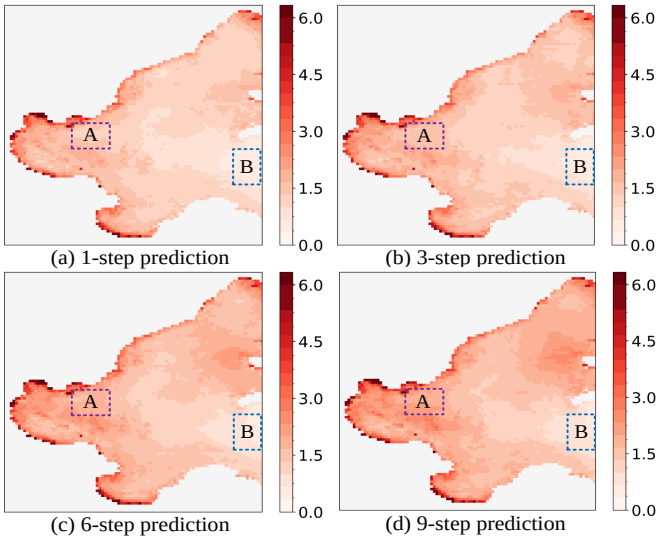

Fig. 9. Spatial distributions of RMSE calculated from all testing samples of the 5-day Chl-a data, where one prediction step represents a 5-day interval. Specifically, area A is close to the shore, and area B is far away from the shore and in deep offshore waters.

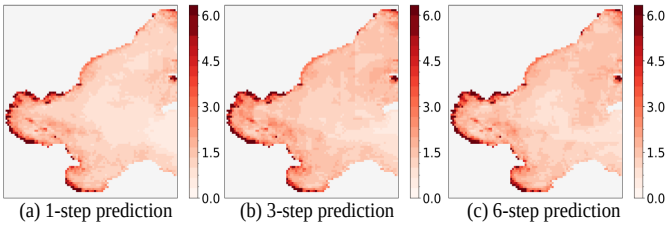

Fig. 10. Spatial distributions of RMSE calculated from all testing samples of the monthly Chl-a data, where one prediction step represents one month.

the LSTM-S, which focuses only on the time-dependency. As shown in Table II, for the relatively short term prediction, ConvLSTM outperforms CNN-LSTM. In particular, for the 1-step prediction, ConvLSTM attains the best RMSE and MAE performance among the five algorithms considered. By contrast, CNN-LSTM may be superior to ConvLSTM in medium and relatively long-term prediction, especially in the monthly data case. The possible reason behind is that the structure of ConvLSTM is relatively complex, making it prone to the inadequate training for the small sample size, resulting in unstable performance for multi-step prediction.

From Table II we can explicitly see that our proposed model outperforms all the other considered models and attains the best performance in nearly all the cases. Specifically, for the monthly data, it outperforms the ConvLSTM by 0.08, 0.43 and 0.65 in terms of RMSE, when the 1-step, 3-step and 6-step predictions are implemented. In terms of MAE, our method outperforms ConvLSTM by 0.05, 0.40 and 0.58, also when these monthly predictions are considered. In terms of the PCorr, the result achieved by Ca-STANet is higher by 0.009, 0.055, 0.074 than that obtained by ConvLSTM for the 1-step, 3-step and 6-step predictions. For the 5-day data, our Ca-STANet outperforms all the other considered models in terms of RMSE and MAE performance, and has the capability to mine the correlation existing in the data.

In general, Table II shows that the prediction error of all the 5 models becomes larger when the step-size of prediction becomes bigger, but among them our model increases at the slowest pace. This implies that introducing the TAFF module enables Ca-STANet to efficiently extract the feature of the long-term dependency that exists in the Chl-a data sequences. As shown in Table II, our Ca-STANet significantly outperforms ConvLSTM for the monthly prediction relying on the small size data samples. The reason behind is that, by dividing the data of Bohai sea into sub-regions according to the spatial heterogeneity, our method is capable of fully exploiting the characteristics of each sub-region in the case of a relatively small number of samples. On the other side, our method can efficiently fuse the regional features and global features to provide the reliable and robust long-term prediction. Owing to the above-mentioned experimental results, our approach is in general capable of outperforming the legacy schemes for the Chl-a prediction.

Figs. 9 and 10 depict the RMSE spatial distributions obtained by the Ca-STANet for the 5-day and monthly data, respectively, which were calculated grid-by-grid from the prediction errors of all the samples during the testing period.

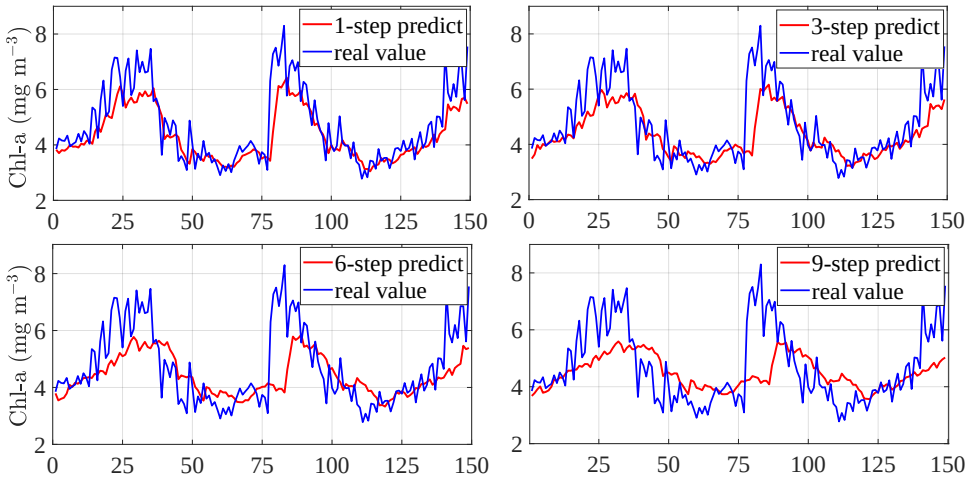

Fig. 11. Temporal trend of the real and predicted Chl-a values for the different step sizes, when the prediction step size is 5-day.

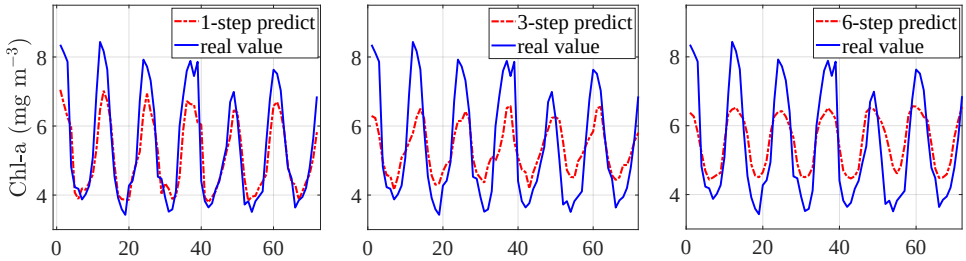

Fig. 12. Temporal trend of the real and predicted Chl-a values for different step sizes, when the prediction step size is one month.

Explicitly, the RMSE in the coastal waters is higher than that in the other offshore waters for both data sets. This is because in the coastal waters, there are large Chl-a spatial gradient, fast temporal variation and outbreak of red tides. From Fig. 9, we observe that as the step size of prediction increases, the RMSE increases quickly at the locations where Chl-a changes fast on the 5-day basis. By contrast, the RMSE is relatively stable at the locations in the deep offshore waters in the Bohai sea. For instance, the RMSES are 1.41, 1.55, 1.84 and 2.09, respectively, when 1-step, 3-step, 6-step and 9-step predictions are executed within area A (bounded by 38.50°-39.00°N, 118.50°-119.13°E, in Fig. 9) that is close to the shoreside. While for the area B (bounded by 38.00°-38.63°N, 121.00°-121.50°E, in Fig. 9) in the deep offshore waters, the corresponding RMSE values are 0.47, 0.54, 0.61 and 0.67, respectively. Furthermore, it is shown that for the 1-step to 9-step prediction, the proportion of the grid points with their RMSE below 2 ($mg \cdot m^{-3}$) is 92.4%, 89.9%, 83.2% and 75.0%, respectively. A similar comparison in the context of the monthly Chl-a data can be obtained based on Fig. 10. From these results, it can be implied that the proposed model enables a stable prediction performance for most of the areas in the Bohai sea.

The temporal trends of the real and predicted Chl-a values for the 5-day and monthly data are present in Fig. 11 and Fig. 12, respectively, which were calculated sample-by-sample from the forecasting results. Specifically, Fig. 11 compares the predicted Chl-a values with the real values on the 5-day intervals from 2018 to 2020. It can be observed that the predicted results by the Ca-STANet fit well with the real Chl-a data. Apparently, as the number of prediction steps increases, the prediction accuracy decreases. Similarly, Fig. 12 compares the predicted Chl-a values with the real Chl-a values on the monthly data basis between January 2014 to December 2020. As shown by the results, the real Chl-a exhibits clearly a monthly periodic characteristics. Our proposed method can efficiently extract this periodicity during the training procedures, resulting in a good periodic fitting with the real Chl-a data for all the prediction steps considered. Furthermore, with the increase of the prediction step size from 1 month to 6 months, error does not increase significantly. The reason behind this is that the monthly data has a seasonal cycle, and our proposed method can improve the accuracy of long-term prediction. In general, the results of Fig. 11 and Fig. 12 explain that the proposed Ca-STANet is capable of efficiently making use of the temporal correlation for improving the accuracy of Chl-a prediction. However, due to the complexity and chaos of the climate variability, the proposed model underestimates the Chl-a peaks in the winter but overestimates the Chl-a valleys in the summer, especially when the number of prediction steps increases.

E. Ablation Study

Next, we carry out an extensive ablation study to demonstrate the effectiveness of the SAFE module and the TAFE module in our Ca-STANet. To serve the purpose, four models are considered, including: 1) **Model_base**, which only uses three layers of the CNN with the original inputs instead of the

TABLE III
ABLATION EXPERIMRNT WITH 5-DAY DATA

Method	RMSE(mg · m^{-3})			
	1-step	3-step	6-step	9-step
Model_base	1.86	1.96	2.11	2.18
Model_S	1.69	1.83	1.98	2.10
Model_ST(w/oAtt)	1.46	1.58	1.74	1.87
Model_ST	1.47	1.58	1.73	1.85
Method	MAE(mg · m^{-3})			
	1-step	3-step	6-step	9-step
Model_base	1.33	1.41	1.56	1.64
Model_S	1.20	1.34	1.49	1.61
Model_ST (w/oAtt)	0.99	1.11	1.28	1.41
Model_ST	1.00	1.11	1.27	1.39
Method	PCorr			
	1-step	3-step	6-step	9-step
Model_base	0.848	0.822	0.798	0.787
Model_S	0.860	0.833	0.812	0.801
Model_ST (w/oAtt)	0.866	0.850	0.833	0.825
Model_ST	0.866	0.849	0.834	0.826

TABLE IV
ABLATION EXPERIMRNT WITH MONTHLY DATA

Method	RMSE(mg · m^{-3})		
	1-step	3-step	6-step
Model_base	2.31	2.69	2.84
Model_S	2.05	2.25	2.22
Model_ST (w/oAtt)	1.89	2.22	2.23
Model_ST	1.82	2.12	2.11
Method	MAE(mg · m^{-3})		
	1-step	3-step	6-step
Model_base	1.52	1.81	1.90
Model_S	1.32	1.44	1.37
Model_ST (w/oAtt)	1.11	1.41	1.34
Model_ST	1.09	1.33	1.30
Method	PCorr		
	1-step	3-step	6-step
Model_base	0.870	0.799	0.770
Model_S	0.878	0.840	0.839
Model_ST (w/oAtt)	0.879	0.846	0.843
Model_ST	0.879	0.850	0.849

outputs of the STIP module; 2) **Model_S**, which includes only the SAFE module; 3) **Model_ST(w/oAtt)**, which includes the SAFE module and the TAFE module but without the multi-head attention layer; 4) **Model_ST**, which includes both the SAFE module and the TAFE module with the multi-head attention layer.

Experimental results are shown in Table III and Table IV for the 5-day and monthly data, respectively, where one step represents a 5-day interval in Table III but one month in Table IV. As shown in the tables, by using the Model_S, the RMSE and MAE for both the 5-day and monthly data are reduced in comparison with that of the Model_base. Specifically, the MAE of Chl-a prediction is decreased by 0.2, 0.36 and 0.53, when 1-step, 3-step and 6-step prediction are respectively carried out on the monthly data. The Model_S can also improve the PCorr effect. As shown in Table III, PCorr is increased by 0.006, 0.016, 0.022, and 0.025, respectively, when the 1-step, 3-step, 6-step and 9-step Chl-a predictions are applied. The results demonstrate that the SAFE module is capable of mining the spatial correlations existing in both the sub-regions and the global region, yielding an improved prediction performance.

From Table III and Table IV, we can explicitly see that Model_ST (w/oAtt) outperforms Model_S in both the 5-day and monthly Chl-a prediction. This implies that the introduction of a time branch for extracting the temporal features is important for improving the prediction accuracy. Furthermore, Model_ST, which adds an extra multi-head attention layer on Model_ST (w/oAtt), performs better than Model_ST (w/oAtt) in most cases. This advantage is obtained because the added attention layer can implement the differential modeling for each output vector based on its different correlation attention scores within all inputs, which enhances the mining of the long-term dependencies of the entire historical Chl-a inputs.

Consequently, it can be seen from Table III and Table IV that the Model_ST in general achieves the best RMSE and MAE performance, and yields the highest PCorr values. Therefore, by introducing the TAFE and SAFE modules to the Ca-STANet, the prediction performance can be improved, especially for the long-term prediction, owing to that it can simultaneously extract the temporal and spatial features for the fusion prediction. This implies that the correlations in both the temporal domain and spatial domain should be exploited to increase the prediction accuracy.

TABLE V
THE INFLUENCE OF WINTER GAP-FILLED DATA ON PREDICTION IN SPRING, SUMMER AND AUTUMN

Data	Metrics	Prediction Months		
		1-step	3-step	6-step
All-data	RMSE	1.68	1.94	2.08
	MAE	0.95	1.20	1.31
Except-winter	RMSE	1.66	1.82	1.96
	MAE	0.94	1.14	1.24

V. Analysis of Results and Influence of Winter Data

Sea-ice and heavy cloud cause a high loss of Chl-a data of Bohai sea in winter. Hence, the winter Chl-a data, especially in the ice-covered area, are mainly gap-filled data obtained by the DINEOF method. Therefore, it is important to investigate the impact of the winter gap-filled data on the Chl-a prediction in spring, summer, and autumn, as well as on the accuracy of Chl-a prediction in winter. To serve this purpose, we set up the following experiments. Firstly, we only consider the Chl-a prediction in spring, summer, and autumn to analyze the impact of the winter gaped-filled data on the performance of the Ca-STANet. As shown in Table V, 'All-data' means that all four seasons' data are used for training, whereas 'Except-winter' means that only the data in spring, summer, and autumn seasons are used for training. Table V explicitly shows that the corresponding results obtained by including and without including the training data of winter season agree well with each other and have only small differences for the monthly Chl-a prediction in spring, summer, and autumn.

Secondly, to investigate the prediction accuracy of winter season data, some grid points with the true satellite-measured Chl-a values at different times are chosen to verify the performance of the Ca-STANet. The settings and results are shown in

TABLE VI
PREDICTION RESULTS FOR SOME GRID POINTS WITH TRUE CHL-A VALUES IN WINTER

Grid	Dataset	Miss_rate	Steps	True_v	P_gap	P_w/ogap
G1	monthly	0.4%	1-step	5.37	4.90(Δ-0.45)	4.85(Δ-0.52)
G2	monthly	23.9%	3-step	11.18	10.17(Δ-1.01)	6.04(Δ-5.14)
G3	monthly	55.1%	1-step	7.48	9.84(Δ+2.36)	0.86(Δ-6.62)
G4	5-day	23.8%	1-step	3.37	3.29(Δ-0.08)	3.57(Δ+0.20)
G5	5-day	33.8%	3-step	1.95	1.67(Δ-0.28)	0.98(Δ-0.97)
G6	5-day	79.7%	1-step	7.75	8.90(Δ+1.15)	1.77(Δ-5.98)

Table VI, 'Miss_rate' represents the miss rate of the grid points in the monthly or 5-day Chl-a dataset, 'True_v' represents the true value of the grid point at a certain time, 'P_gap' represents the predicted Chl-a value for the certain time by using the gap-filled data as inputs, 'P_w/ogap' represents the predicted Chl-a value for the certain time by using the original incomplete data as inputs, and finally Δ represents the difference between the predicted value and the true Chl-a value. From Table VI, we can clearly see that the differences between 'P_gap' and 'True_v' are relatively small for the six grid points with different miss rates and Chl-a concentrations. By contrast, the differences between 'P_w/ogap' and 'True_v' are typically several times bigger than the corresponding differences between 'P_gap' and 'True_v'. This is especially the case, when the 'Miss_rate' is high. Hence, even though the DINEOF algorithm fills the missing data of the Bohai Sea in the winter without considering in detail the ice-covered areas, the prediction results obtained by Ca-STANet based on the gap-filled data have much higher accuracy than that obtained based on the original incomplete data. Therefore, we can conclude that the gap-filled data by the DINEOF algorithm are effective for Ca-STANet to achieve good performance.

VI. Conclusion

Accurate Chl-a prediction is paramount for providing an early warning of red tide and keeping the marine ecosystem healthy. To the best of our knowledge, most existing learning-based Chl-a prediction approaches predict the Chl-a of a single point or multiple points in a small region, based mainly on the monitoring data. By contrast, the long-term Chl-a prediction in the relatively large-scale area is still in its infancy. In this paper, from the perspective of the spatio-temporal field, the remote sensing data of OC-CCI were used to predict the Chl-a in the whole Bohai sea. First, the DINEOF method was introduced to fill the gap in the OC-CCI Chl-a data between 1980 and 2020. For prediction, a new framework, namely the Ca-STANet, was designed to predict the Chl-a in a large-scale area by exploiting both the spatial and temporal correlations. In our method, the original input data were divided into multiple sub-regions not only to tackle the spatial heterogeneity of large-scale areas but also to avoid the occurrence of catastrophic forgetting. A TAFE module was designed to derive the important time features and capture the long-term dependency. A SAFE module was proposed to extract the spatial evolution features in the sub-regions, and at the same time mine the dynamic correlation of global spatial neighborhood. Owing to these well-designed modules, Ca-STANet is capable of capturing the spatio-temporal correlations to achieve highly reliable Chl-a prediction. The experimental results demonstrated that the Ca-STANet achieves better performance than the state-of-the-art methods in both the 5-day and monthly Chl-a prediction. Moreover, Ca-STANet can even perform efficiently with a small number of data samples, such as the monthly Chl-a data.

In this paper, the missed Chl-a data in the Bohai sea are reconstructed by the DINEOF without considering in detail sea-ice coverage. In the future, we will make an endeavor to study the high-quality Chl-a dataset reconstruction by considering the dynamic sea-ice mask and investigate its impact on our Ca-STANet and other Chl-a prediction algorithms. Moreover, in this paper, the introduced Chl-a prediction depends only on Chl-a data. In practice, however, Chl-a may be affected by the other hydrological and meteorological conditions, as well as the land-sourced pollutant emission. Therefore, in our future work, we will endeavor to further improve our Ca-STANet by exploiting the influence of these factors.

Acknowledgments

The authors would like to give special thanks to the European Space Agency (ESA) for providing the chlorophyll-a (Chl-a) concentration data (https://esa-oceancolour-cci.org/).

References

[1] M. J. Behrenfeld and P. G. Falkowski, "Photosynthetic rates derived from satellite-based chlorophyll concentration," *Limnology and oceanography*, vol. 42, no. 1, pp. 1–20, 1997.

[2] M. Lévy, P. J. Franks, and K. S. Smith, "The role of submesoscale currents in structuring marine ecosystems," *Nature communications*, vol. 9, no. 1, pp. 1–16, 2018.

[3] A. R. Fay and G. A. McKinley, "Correlations of surface ocean pCO2 to satellite chlorophyll on monthly to interannual timescales," *Global Biogeochemical Cycles*, vol. 31, no. 3, pp. 436–455, 2017.

[4] E. Zohdi and M. Abbaspour, "Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction," *International Journal of Environmental Science and Technology*, vol. 16, no. 3, pp. 1789–1806, 2019.

[5] R. Hidayat, M. Zainuddin, A. R. S. Putri *et al.*, "Skipjack tuna (katsuwonus pelamis) catches in relation to chlorophyll-a front in Bone Gulf during the southeast monsoon," *Aquaculture, Aquarium, Conservation & Legislation*, vol. 12, no. 1, pp. 209–218, 2019.

[6] C. A. Stock, J. G. John, R. R. Rykaczewski, R. G. Asch, W. W. Cheung, J. P. Dunne, K. D. Friedland, V. W. Lam, J. L. Sarmiento, and R. A. Watson, "Reconciling fisheries catch and ocean productivity," *Proceedings of the National Academy of Sciences*, vol. 114, no. 8, pp. E1441–E1449, 2017.

[7] Q. Chen, T. Guan, L. Yun, R. Li, and F. Recknagel, "Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials," *Harmful Algae*, vol. 43, pp. 58–65, 2015.

[8] M. Chen, J. Li, X. Dai, Y. Sun, and F. Chen, "Effect of phosphorus and temperature on chlorophyll a contents and cell sizes of scenedesmus obliquus and microcystis aeruginosa," *Limnology*, vol. 12, no. 2, pp. 187–192, 2011.

[9] E. Chassot, S. Bonhommeau, N. K. Dulvy, F. Mélin, R. Watson, D. Gascuel, and O. Le Pape, "Global marine primary production constrains fisheries catches," *Ecology letters*, vol. 13, no. 4, pp. 495–505, 2010.

[10] S. E. Jørgensen, H. Mejer, and M. Friis, "Examination of a lake model," *Ecological Modelling*, vol. 4, no. 2-3, pp. 253–278, 1978.

[11] J.-Y. Park, C. A. Stock, J. P. Dunne, X. Yang, and A. Rosati, "Seasonal to multiannual marine ecosystem prediction with a global Earth system model," *Science*, vol. 365, no. 6450, pp. 284–288, 2019.

[12] C. S. Rousseaux, W. W. Gregg, and L. Ott, "Assessing the skills of a seasonal forecast of chlorophyll in the global pelagic oceans," *Remote Sensing*, vol. 13, no. 6, p. 1051, 2021.

[13] T. Rajaee and A. Boroumand, "Forecasting of chlorophyll-a concentrations in South San Francisco Bay using five different models," *Applied Ocean Research*, vol. 53, pp. 208–217, 2015.

[14] W. Tian, Z. Liao, and J. Zhang, "An optimization of artificial neural network model for predicting chlorophyll dynamics," *Ecological Modelling*, vol. 364, pp. 42–52, 2017.

[15] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea," *Science of the Total Environment*, vol. 502, pp. 31–41, 2015.

[16] B. Li, G. Yang, R. Wan, G. Hörmann, J. Huang, N. Fohrer, and L. Zhang, "Combining multivariate statistical techniques and random forests model to assess and diagnose the trophic status of Poyang Lake in China," *Ecological Indicators*, vol. 83, pp. 74–83, 2017.

[17] H. Yajima and J. Derot, "Application of the random forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases," *Journal of Hydroinformatics*, vol. 20, no. 1, pp. 206–220, 2018.

[18] H. Yang, W. Zhong, Y. Ma, H. Geng, R. Chen, W. Chen, and B. Yu, "Vlsi mask optimization: From shallow to deep learning," *Integration*, vol. 77, pp. 96–103, 2021.

[19] S. Shamshirband, E. Jafari Nodoushan, J. E. Adolf, A. Abdul Manaf, A. Mosavi, and K.-w. Chau, "Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 91–101, 2019.

[20] J. Yu and X. Yan, "Whole process monitoring based on unstable neuron output information in hidden layers of deep belief network," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3998–4007, 2019.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] Y. Shin, T. Kim, S. Hong, S. Lee, E. Lee, S. Hong, C. Lee, T. Kim, M. S. Park, J. Park *et al.*, "Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods," *Water*, vol. 12, no. 6, p. 1822, 2020.

[23] H. Cho, U.-J. Choi, and H. Park, "Deep learning application to time-series prediction of daily chlorophyll-a concentration," *WIT Trans. Ecol. Environ*, vol. 215, pp. 157–163, 2018.

[24] N. A. P. Rostam, N. H. A. H. Malim, R. Abdullah, A. L. Ahmad, B. S. Ooi, and D. J. C. Chan, "A complete proposed framework for coastal water quality monitoring system with algae predictive model," *IEEE Access*, vol. 9, pp. 108 249–108 265, 2021.

[25] X. He, S. Shi, X. Geng, L. Xu, and X. Zhang, "Spatial-temporal attention network for multistep-ahead forecasting of chlorophyll," *Applied Intelligence*, vol. 51, no. 7, pp. 4381–4393, 2021.

[26] L. Zheng, H. Wang, C. Liu, S. Zhang, A. Ding, E. Xie, J. Li, and S. Wang, "Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models," *Journal of Environmental Management*, vol. 295, p. 113060, 2021.

[27] R. Barzegar, M. T. Aalami, and J. Adamowski, "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model," *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 2, pp. 415–433, 2020.

[28] R. Everson, P. Cornillon, L. Sirovich, and A. Webber, "An empirical eigenfunction analysis of sea surface temperatures in the western north atlantic," *Journal of Physical Oceanography*, vol. 27, no. 3, pp. 468–479, 1997.

[29] R. He, R. H. Weisberg, H. Zhang, F. E. Muller-Karger, and R. W. Helber, "A cloud-free, satellite-derived, sea surface temperature analysis for the west florida shelf," *Geophysical Research Letters*, vol. 30, no. 15, 2003.

[30] T. Wu and Y. Li, "Spatial interpolation of temperature in the united states using residual kriging," *Applied Geography*, vol. 44, pp. 112–120, 2013.

[31] A. Alvera-Azcárate, A. Barth, M. Rixen, and J.-M. Beckers, "Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature," *Ocean Modelling*, vol. 9, no. 4, pp. 325–346, 2005.

[32] A. Barth, A. Alvera Azcárate, M. Licer, and J.-M. Beckers, "A convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations (dincae)," in *EGU General Assembly Conference Abstracts*, 2020, p. 9414.

[33] T. N. Miles, R. He, and M. Li, "Characterizing the south atlantic bight seasonal variability and cold-water event in 2003 using a daily cloud-free sst and chlorophyll analysis," *Geophysical research letters*, vol. 36, no. 2, 2009.

[34] B. Henn, M. S. Raleigh, A. Fisher, and J. D. Lundquist, "A comparison of methods for filling gaps in hourly near-surface air temperature data," *Journal of Hydrometeorology*, vol. 14, no. 3, pp. 929–945, 2013.

[35] Y. Wang and D. Liu, "Reconstruction of satellite chlorophyll-a data using a modified dineof method: a case study in the bohai and yellow seas, china," *International journal of remote sensing*, vol. 35, no. 1, pp. 204–217, 2014.

[36] Y. Wang, Z. Gao, and D. Liu, "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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 5, pp. 1383–1395, 2019.

[37] J. Guo, J. Lu, Y. Zhang, C. Zhou, S. Zhang, D. Wang, and X. Lv, "Variability of chlorophyll-a and secchi disk depth (1997–2019) in the bohai sea based on monthly cloud-free satellite data reconstructions," *Remote Sensing*, vol. 14, no. 3, p. 639, 2022.

[38] G. Zheng, X. Li, R.-H. Zhang, and B. Liu, "Purely satellite data–driven deep learning forecast of complicated tropical instability waves," *Science advances*, vol. 6, no. 29, p. eaba1482, 2020.

[39] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, "Deep learning for multi-year ENSO forecasts," *Nature*, vol. 573, no. 7775, pp. 568–572, 2019.

[40] Y. Zhou, K. Chen, and X. Li, "Dual-branch neural network for sea fog detection in geostationary ocean color imager," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[41] Y. Zhou, C. Lu, K. Chen, and X. Li, "Multilayer fusion recurrent neural network for sea surface height anomaly field prediction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[42] G. Xu, D. Xian, P. Fournier-Viger, X. Li, Y. Ye, and X. Hu, "Amconvgru: a spatio-temporal model for typhoon path prediction," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5905–5921, 2022.

[43] C. Xiao, N. Chen, C. Hu, K. Wang, Z. Xu, Y. Cai, L. Xu, Z. Chen, and J. Gong, "A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data," *Environmental Modelling & Software*, vol. 120, p. 104502, 2019.

[44] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge *et al.*, "Skilful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.

[45] N. Zhao, G. Zhang, S. Zhang, Y. Bai, S. Ali, and J. Zhang, "Temporal-spatial distribution of chlorophyll-a and impacts of environmental factors in the Bohai Sea and Yellow Sea," *IEEE Access*, vol. 7, pp. 160 947–160 960, 2019.

[46] A. Alvera-Azcárate, A. Barth, G. Parard, and J.-M. Beckers, "Analysis of SMOS sea surface salinity data using DINEOF," *Remote sensing of environment*, vol. 180, pp. 137–145, 2016.

[47] M. Rixen, J.-M. Beckers, J.-M. Brankart, and P. Brasseur, "A numerically efficient data analysis method with error map generation," *Ocean Modelling*, vol. 2, no. 1-2, pp. 45–60, 2000.

[48] L. Anselin, "Thirty years of spatial econometrics," *Papers in regional science*, vol. 89, no. 1, pp. 3–25, 2010.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[52] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.

[53] J.-H. Choi, J. Kim, J. Won, and O. Min, "Modelling chlorophyll-a concentration using deep neural networks considering extreme data imbalance and skewness," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2019, pp. 631–634.

[54] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.