A Cluster-Stacking-Based Approach to Forecasting Seasonal Chlorophyll-a Concentration in Coastal Waters

WEIJIA JIA, JIE CHENG, AND HONGZHI HU

1 School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264200, China
2 Weihai Marine and Fishery Monitoring and Hazard Mitigation Center, Weihai 264200, China

Corresponding author: Jie Cheng (chjie@sdu.edu.cn)

ABSTRACT Prediction of Chlorophyll-a (Chl-a) concentration is significant for marine ecology and environmental protection. This paper presents an integrated approach to forecast seasonal Chl-a concentration in coastal waters. Before modeling, feature construction procedures, such as simplification, combination, and normalization, are conducted to identify the potentially significant features. The feature extraction method based on Random Forest (RF) and eXtreme Gradient BOOSTing (XGBoost) is applied to select relevant variables. Then, we propose a Cluster-stacking-based approach which includes a station-oriented clustering model and a stacking-based regression model. The former model is used to divide the observation stations into several groups, thus partitions the study region into several sub-regions and the study dataset into several subsets according to the corresponding stations. In each subset, single regression models including K-Nearest Neighbor (KNN), Support Vector Regression (SVR), Multi-Layer Perceptron regression (MLP) and XGBoost are established in level 0 space and integrated by RF in level 1 space via stacked generalization. We compare the performance of the Cluster-stacking model with that of Cluster-KNN, Cluster-SVR, Cluster-MLP, Cluster-XGBoost and the regression stacking model without cluster. The model evaluation shows that the Cluster-stacking-based approach outperforms others in forecasting Chl-a concentration with a coefficient of determination \( R^2 \) of 0.848 and a mean absolute error (MAE) of 0.665 ug/l.

INDEX TERMS Chlorophyll-a concentration forecast, cluster-stacking-based approach, regression stacking, coastal waters.

I. INTRODUCTION

Since phytoplankton is a fundamental component of oceanic food webs [1], Chl-a concentration which is considered as an indicator of phytoplankton biomass reflects the primary productivity of waters. The changes of Chl-a concentration reflect the dynamic changes of water physical and chemical properties [2]. Therefore, the prediction of Chl-a concentration is of great significance for marine ecology and environmental protection [3], such as preventing the red tides and improving water quality. Hence, an effective method to forecast Chl-a concentration is urgently needed to achieve an early warning.

Currently, researches on Chl-a concentration forecasting are commonly based on data-driven methods. Previous studies [2], [4] have shown that Chl-a concentration is related to a variety of different water environmental factors, such as temperature, nutrients, and human activities. The data-driven approach is to study the regularity of Chl-a concentration and other environmental factors from observed data. Zeng et al. [5] applied an artificial neural network model and a multi-varieties regression model with weekly collected data to predict Chl-a concentration in urban lakes. Liu et al. [4] used a multivariate statistical approach to model the influence of 18 chemistry variables on Chl-a in a lake with monthly collected data. Malek et al. [6] accessed four predictive ecological models, including fuzzy logic, recurrent neural network, hybrid evolutionary algorithm, and multiple linear regressors to forecast Chl-a concentration with bi-weekly collected data in a lake. However, the above studies did not achieve “real” predictions. That is, they just explored the relationship between Chl-a concentration and other water...
quality environmental variables in the same period but did not make future predictions, thus they could not give an early warning.

To forecast days-ahead Chl-a concentration and give an early warning, some studies predicted Chl-a concentration of the next periods based on historical data of related factors. For example, Wang et al. [7] designed a Chl-a predicting model with dynamic neural network models in a simulated indoor environment. The prediction performance of the proposed models was examined from 1-hour ahead to 35-hour ahead. Park et al. [8] gave a 7-day early warning prediction of Chl-a concentration in reservoirs using weekly water quality observation data. Their method is based on artificial neural networks and support vector machines. Besides, many other machine learning approaches such as backpropagation artificial neural network, support vector regression and random forests have also been applied to forecasting Chl-a concentration in lakes and rivers [9], [10].

This paper focuses on seasonal Chl-a concentration in coastal waters which is different from indoor environments, reservoirs, rivers or lakes. Due to the high cost of the data collection, the acquisition of monitoring data is limited [11]. As Huang [12] estimated, long-term and large-scale analysis can describe the developing trend of Chl-a concentration better than small range data. Besides, since Chl-a biomass in ocean surface is regulated by a complex interaction of physiological, oceanographic and ecological factors [13], to get an accurate prediction, we applied as many relevant factors as possible in experimental data. Based on these datasets, we propose an effective hybrid approach to give a one-season-ahead prediction of Chl-a concentration from four-season historical observation data.

Hybrid models have been applied in many industries. Most of these models can be regarded as a combination of different data preprocessing, modeling and optimization techniques [14]. For instance, to forecast the exchange rate, Shen et al. [15] proposed an improved deep belief network (DBN) combined with continuous restricted Boltzmann machines, which updated the classical DBN to fit continuous data. In the field of air pollution prediction, Zhu et al. [16] proposed a hybrid model based on complementary ensemble empirical mode decomposition, particle swarm optimization and gravitational search algorithm, support vector regression, generalized regression neural network and grey correlation analysis, to forecast daily PM$_{2.5}$ concentration. Bi et al. [17] presented an integrated forecasting method equipped with noise filtering and data frequency representation to predict the amount of workload in future time slots. Although these methods cannot be directly applied to the prediction of Chl-a concentration, they provided approaches to integrating different models and avoiding the insufficiency of single ones.

The Chl-a concentration forecasting approach presented in this paper, the Cluster-stacking-based approach, combines the advantages of both station-oriented clustering and stacking-based regression to improve forecasting performance. The major contributions of this work are as follows.

(1) It proposes a feature extraction procedure for coastal water quality environmental factors including simplification, combination, normalization, and importance analysis.

(2) It proposes a station-oriented clustering model to group the water quality observation stations. By means of clustering, the comprehensive influences of biological, physical and chemical reactions in waters are reflected to that of geographical differences.

(3) It presents a regression-stacking model to deal with regressors, thus the forecast of Chl-a concentration can avoid the insufficiency of single models and achieve an acceptable performance.

(4) It proposes an integrated approach to forecasting Chl-a concentration in coastal waters that includes the station-oriented clustering and the stacking-based regression. The extensive experimental results show that the Cluster-stacking-based approach achieves better performance than single models with or without clustering.

The rest of this paper is organized as follows. Section II describes the study region, data source, and data preprocessing. Particularly, data preprocessing includes simplification, combination, normalization, and feature importance analysis. Section III details the Cluster-stacking-based approach which includes a station-oriented clustering model and a stacking-based regression model. The former model is applied to group the observation stations and the latter one is used to predict Chl-a concentration in each group. Section IV discusses the experimental results. Section V concludes the paper and suggests the directions for future works.

II. STUDY AREA AND DATA PREPROCESSING

A. STUDY AREA AND DATA SOURCE

The study area of this paper is Weihai sea area (latitude 36°30’N to 37°45’N and longitude 121°15’E to 122°50’E), which is located in the eastern part of Shandong Peninsula, China [18]. Weihai coastal region has many industrial enterprises and it is also a significant region in terms of population, agriculture and marketing systems [19]. In this paper, we collected an amount of historical water quality monitoring data, including space-time factors, biological factors, chemical factors, physical factors, meteorological factors, hydrological factors, and some other kinds of water quality factors, from the 1st season of 2015 to the 4th season of 2018.

Chl-a distribution in the Weihai coastal area has prominent regional characteristics [20]. Fig.1 illustrates the general distribution of monitoring stations used in this study. As can be seen from Fig.1, the distribution of Chl-a near shore is higher compared with that far shore, particularly in Rushan. And the Chl-a concentration is similar in a relatively large area of seas, such as north of Huancui, east of Rongcheng, south of Wendeng, and south of Rushan.

Apart from spatial distribution characteristics, Weihai coastal region also has temporal distribution characteristics of seasonal average Chl-a concentration. As shown in Fig.2, Chl-a concentration in winter (the 4th season) is lower than that in spring, summer, and autumn. The possible causes
FIGURE 1. Map of the study area. Green points denote observation stations and the sizes of them show the average value of Chl-a concentration during the study period. The yellow area and blue area represent land and ocean respectively.

FIGURE 2. Seasonal Chl-a concentration changes over time.

of this phenomenon are the influences of sea wind, cold water mass and atmospheric pressure in different seasons. The spatial and temporal characteristics of Weihai coastal waters make the prediction of Chl-a concentration possible.

TABLE 1 lists the observed water indicators used in this research. Since the frequency of sampling is not fixed, usually one to three times a season, we then give the average value of each indicator seasonally to ensure the coherence of the dataset. The column Mean and Skew show the mean value and skewness coefficient of each indicator respectively.

Water indicators in this study can be categorized into spatial, temporal, biological, chemical, physical, meteorological, hydrological, and anthropic factors. For clarity, we locate the spatial-temporal feature by five parameters including longitude (x), latitude (y), sampling depth (Depth), year (Year), and season (Season) of the record date. The biological indicator, Chl-a, is the main object of the study. Physical and chemical indicators listed on the table (NO2-N, NO3-N, NH4-N, DIN, DIP, TN, TP, Si, pH, Sal, DO, COD, Petro, SS, T, Tr) have all been proved to be associated with water quality [4], [8]–[10], [21]. Apart from common physicochemical factors in related studies, some meteorological and hydrological characteristics such as seasonal average wind speed (WindS), wind direction (WindD), watercolor (WaterC), weather (weather), sea condition (SeaC) were also contained. To be specific, WindS and WindD are essential indicators for marine phytoplankton presence-absence prediction during calm weather [22]. The amount of dissolved and particulate material influences WaterC [23]. weather is of four types, cloudy, fog, rainy, and sunny. According to the sea surface conditions, the shape of wave crests, the degree of ruptures, and the number of foam waves, SeaC is divided into ten levels (0-9).

A multitude of anthropic influences has significantly altered the environmental conditions and the diversity of marine biological communities [24]. However, understanding and predicting the combined impacts of single and multiple stressors are particularly challenging [21]. In this research, impacts of human were concluded as four indicators, including the population of the nearest town or street from the observation station (NP), function name of the location of the observation station (FN) and type of the place where observation stations are located (FL) and the type of the nearest sewage (ST). Six types of FL are agriculture, conservation, industrial, harbor, recreation, and reservation. Five types of ST are the land stream, rural drain, sea farming, town drain, and others.

B. DATA PREPROCESSING

Since nonlinear functions can describe most of the relationships between predictors and target values better than simple linear regressors [14], feature construction and feature extraction are needed before modeling.
In this section, three different feature construction procedures, including simplification, combination, and normalization, are used to identify potentially significant features. Then, we applied a feature importance analysis method based on RF and XGBoost for feature extraction.

1) SIMPLIFICATION

We use the Individual Water Quality Index (IWQI) class [25] to evaluate the grades of single factor concentrations. Classification criteria used in this study are listed in APPENDIX and the calculation formula for IWQI is described as (1).

\[
IWQI = \frac{(IWQI_{UI} - IWQI_{LI})(C_i - C_{LI})}{C_{UI} - C_{LI}} + IWQI_{LI} \tag{1}
\]

Given a water quality observation factor \(i\), \(IWQI_i\) is the desired \(IWQI\) of \(i\). \(C_i\) is the concentration of \(i\). \(C_{UI}\) is the upper bound of factor concentrations of a specific water quality grade while \(C_{LI}\) is the lower bound. \(IWQI_{UI}\) and \(IWQI_{LI}\) are the corresponding individual water quality index values of \(C_{UI}\) and \(C_{LI}\), respectively. Then, we simplified the numerical \(IWQI\) of different factors into four rates as \(IWQICOD\), \(IWQIDIP\), \(IWQIDIN\), and \(IWQIPetro\). Based on these, \(IWQIMax\) is constructed, which is the max value of the four grades described above. Simplification factors are shown in TABLE 2.

2) COMBINATION

There are many parameters to assess water quality according to the national and international criteria and standards. It has been proved that nutrient concentrations and the ratio of them affect Chl-a concentration primarily [26]. Thus, various ratios of different types of nutrients were constructed to improve the precision of model forecasting. In our study, the combination method is described as (2),

\[
a/b = \frac{\text{concentration(a)}}{\text{concentration(b)}} \tag{2}
\]

in which, \(a\) and \(b\) are two different nutrient concentration factors and \(a/b\) is the constructed feature represents a nutrient ratio. TABLE 2 also shows the constructed indicators.

3) NORMALIZATION

As can be seen in TABLE 1 and TABLE 2, different coastal waters observation features usually have various dimensions.
and different order of magnitude. If the analysis is performed directly with original index values, the role of the higher-value indicators in the comprehensive analysis would be highlighted and the effect of lower-level indicators would be relatively weakened. Therefore, to ensure the reliability of forecasting, the original data needs to be normalized. We applied the min-max normalization method [27] described as formula (3) to deal with all features, including origin features and constructed features.

$$x^* = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$  (3)

In (3), $x^*$ is the normalized value, $x$ is the original value, $x_{\text{min}}$ and $x_{\text{max}}$ are respectively the minimum and maximum value of the corresponding feature.

Fig. 3 compares data distributions measured before and after min-max transformation. Taking COD as an example, Fig. 3 shows two histograms of data distributions with a kernel density estimate and maximum likelihood Gaussian distribution fitting curve.

### 4) IMPORTANCE ANALYSIS

To select relevant variables, we acquire the importance scores of each indicator by two tree-based models. Let $x = (x_1, x_2, \ldots, x_n)$ be a vector of water quality parameters, where $x_i$ describes the samples of water quality observation factor $i$, $i \in [1, n]$. Applying XGBoost [28] and RF [29] models on $x$, we can get the importance scores of each indicator, as $XG_i$ and $RF_i$. For each of these indicators, its importance score is the average of the scores acquired from the above two models [14]. The calculation method is described as (4),

$$IM_i = \frac{1}{2}(XG_i + RF_i)$$  (4)

where $XG_i$ and $RF_i$ are the importance scores of water quality observation factor $i$ calculated by XGBoost and RF, respectively. $IM_i$ is the importance score of factor $i$ used in this study, which is the mean value of $XG_i$ and $RF_i$. Then we dropped the least essential indicators to increase the computation speed and generalization performance.

The two base models, XGBoost and RF, generate feature importance scores in different ways. For the XGBoost model, it generates importance score $XG_i$ based on the frequency at which $x_i$ is used to segment the study data during node segmentation of trees established [30]. RF calculates the importance score $RF_i$ by determining the average contribution of $x_i$ on each tree [31]. Calculating the average value of

| Indicators | Factors | Type* | Description | Mean | Skew |
|------------|---------|-------|-------------|------|------|
| Simplification | IWQCOD | N | COD rate calculated by Single Factor Index Method | 0.48 | 2.16 |
| IWQDIP | N | DIP rate calculated by Single Factor Index Method | 0.53 | 3.68 |
| IWQDIN | N | DIN rate calculated by Single Factor Index Method | 0.51 | 3.49 |
| IWQPN | N | Petroleum rate calculated by Single Factor Index Method | 0.49 | 1.98 |
| IWQMax | N | Max value of IWQCOD, IWQDIP, IWQDIN and IWQPN | 1.11 | 3.63 |
| Combination | NO3/NO2 | N | NO3 concentration divided by NO2 concentration | 38.22 | 6.70 |
| NO3/NH4 | N | NO3 concentration divided by NH4 concentration | 6.77 | 8.42 |
| NO2/NH4 | N | NO2 concentration divided by NH4 concentration | 0.28 | 15.43 |
| TN/TP | N | TN divided by TP | 27.89 | 23.61 |
| N/P | N | DIN concentration divided by DIP concentration | 32.46 | 9.83 |
| Si/N | N | Si concentration divided by DIN concentration | 1.68 | 5.06 |
| Si/P | N | Si concentration divided by DIP concentration | 24.34 | 2.98 |

* An $N$ denotes a numerical variable in the column of Type.

| TABLE 2. Description of simplification and constructed factors. |
$X_G$ and $RF_i$ can reduce the influence of the difference caused by different importance analysis methods.

### III. METHODOLOGY

The proposed Cluster-stacking-based approach includes two parts, a station-oriented clustering, and a stacking-based regression.

#### A. STATION-ORIENTED CLUSTERING MODEL

Cluster technique can help to reveal hidden information in a large dataset, hence considering clustering results as input variables for a regression model can improve model performance [32]. In addition, since Chl-a in sea surface has geographical and dynamic features [33], it is necessary to segment the study region by grouping the stations. So, we first propose a station-oriented clustering model to cluster observation stations into several groups, by which, the study region can be partitioned into several sub-regions. The procedure is described as Algorithm 1.

**Algorithm 1** The Station-Oriented Clustering

**Input:** $D = \{s_1, s_2, \ldots, s_q\}$, $M = \{m_1, m_2, \ldots, m_p\}$.
**Output:** $C = \{C_1, C_2, \ldots, C_{k_{best}}\}$.

1. Let $k = \{1, 2, \ldots, \lceil \sqrt{q} \rceil\}$.
2. For $m_i \in M$, $k_j \in k$:
   - $m_i(k_j, D)$.
3. $\text{Score}_{CH_i} \leftarrow CH(m_i, k_j)$.
4. $\text{Score}_{CH} = \max \{\text{Score}_{CH_i}\}_{p \times \lceil \sqrt{q} \rceil}$.
5. $m_{best}, k_{best} \leftarrow \text{optimal} \text{Score}_{CH}$.
6. $C \leftarrow m_{best}(k_{best}, D)$.

The input of Algorithm 1 contains a water quality observation dataset $D = \{s_1, s_2, \ldots, s_q\}$ and a set of clustering methods $M = \{m_1, m_2, \ldots, m_p\}$, $p \geq 1$, where $s_i, 1 \leq i \leq q$, represents a time series data of seasonal average Chl-a concentration of a water quality observation station, and $q$ is the scale of historical data used in this study; $m_i \in M$ represents a specific clustering method, and $p$ is the total number of clustering methods used in this study. The output is a cluster set $C = \{C_1, C_2, \ldots, C_{k_{best}}\}$, where $C_i \in C$ contains several water quality observation stations, and $k_{best}$ is the optimal number of clusters. An appropriate number of clusters is one of the most influential factors on the results of clustering models. Thus we use Algorithm 1 to select not only the optimum clustering method $m_{best}$ but also the optimum number of clusters $k_{best}$.

We first let $k = \{1, 2, \ldots, \lceil \sqrt{q} \rceil\}$ (Step 1). For $k_j \in k$ and clustering method $m_i \in M$ (Step 2): Train the clustering model by method $m_i$ in total $k_j$ clusters based on dataset $D$ (Step 3). Then calculate the Calinski-Harabasz score $\text{Score}_{CH_i}$ [34] of the model via (5).

$$
\text{Score}_{CH_i} = \frac{tr(B_{m_i, k_j})}{tr(W_{m_i, k_j})} \frac{q - k_j}{k_j - 1}
$$

where $k_j$ is the amount of clusters, $B_{m_i, k_j}$ is the covariance matrix between categories, $W_{m_i, k_j}$ is the covariance matrix of intra-category data and $tr$ is the trace of a matrix. A high CH score means a small covariance of data within categories and large covariance between them. The higher the value of $\text{Score}_{CH}$, the better the performance of a clustering model will be (Step 4).

Let $\{\text{Score}_{CH_i}\}_{p \times \lceil \sqrt{q} \rceil}$ as a set of CH score got through the loop. Then compare the elements in it (Step 5) and determine the parameters $m_{best}$ and $k_{best}$ corresponding to the optimal clustering model, which is the model with maximum $\text{Score}_{CH}$ (Step 6). Divide observation stations into $k_{best}$ number of clusters by the model $m_{best}$ as $C = \{C_1, C_2, \ldots, C_{k_{best}}\}$ (Step 7).

In this paper, we set $p = 3$ and $M = \{m_1, m_2, m_3\}$ where $m_1$ to $m_3$ are respectively set to K-means Cluster, Agglomerative Cluster, and Gaussian Mixture Cluster. K-means Cluster is a distance-based clustering method [35]; the Agglomerative Cluster is a hierarchical clustering method [36], and the Gaussian Mixture Cluster is based on Gaussian mixture method [37]. Each of the three approaches has its advantages for water quality observation data.

K-means model is applicable to cope with the feature of a large amount of water quality observation data because of its fast convergence with high efficiency. Compared with K-means, the Gaussian mixture cluster is more general and can form clusters of different sizes and shapes, which is conducive to clusters with a different number of observation stations. The main advantage of hierarchical clustering is that it can find the hierarchical relationship between clusters, which is beneficial to analyze observation stations with geographical proximity. According to the clusters of observation stations, the sampling dataset we used is thus divided into several smaller sub-datasets, which lays a foundation for the establishment of the regression model.

#### B. STACKING-BASED REGRESSION MODEL

To avoid the insufficiency of single regression model, stacking strategy that combines multiple regressors or classifiers for predictions has been proved helpful [38]. Hence, we propose a stacking-based regression model to forecast seasonal average Chl-a concentration from a water quality observation dataset. The principle of the integrated approach is to combine the predictions of several base learners to improve the robustness and generality of a single learner [39]. The stacking-based regression model can be concluded as Algorithm 2.

In Algorithm 2, the input contains a water quality observation dataset $D = (x, y)$, in which $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m})$ represents a sample of water quality observation data including historical records of 4 seasons, $x_{i,p}, 1 \leq p \leq m$, represents the value of a water quality factor, and $m$ is the total number of water quality factors used in this study; $y_i$ is the observed value of Chl-a concentration corresponding to $x_i$; $n$ denotes the scale of the dataset. In addition, the stacking-based regression model is...
composed of two levels. The input also contains a set of basic regression learning machines, where \( L_j, 1 \leq j \leq k \) and a new one \( L_{new} \) are applied in level-0 and level-1, respectively, \( k \) is the number of basic regression learning machines in level-0. The output is the predicted value of one-season-ahead Chl-a concentration \( \hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) \), where \( \hat{y}_i \) represents the predicted value of \( y_i \).

**Algorithm 2 The Stacking-Based Regression**

| Input: \( D = (x, y), L = \{L_1, L_2, \ldots, L_k, L_{new}\} \) |
| Output: \( \hat{y} \) |
|---|
| 1. \( \{D_{train}, D_{test}\} \leftarrow k\text{-fold} (D) \) |
| 2. \( \{D_1, D_2, \ldots, D_k\} \leftarrow k\text{-fold} D_{train} \) |
| 3. For \( L_j, 1 \leq j \leq k \): |
| 4. \( D_{j} = D - D_j \) |
| 5. \( L_j \leftarrow \text{train} (L_j, D_{j}) \) |
| 6. \( \hat{y}_{j} = \text{average}(\hat{y}(1_{train}), \hat{y}(2_{train}), \ldots, \hat{y}(k_{train})) \) |
| 7. \( \hat{y}_{train} = \hat{y}(1_{train}), \hat{y}(2_{train}), \ldots, \hat{y}(k_{train}) \) |
| 8. \( \hat{y}_{test} = \text{average}(\hat{y}(1_{test}), \hat{y}(2_{test}), \ldots, \hat{y}(k_{test})) \) |
| 9. \( \hat{y} \leftarrow L_{new} (D_{new}) \) |

First, we divide the original dataset \( D \) into \( D_{train} (80\%) \) and \( D_{test} (20\%) \) randomly (Step 1), apply \( k\)-fold cross-validation to \( D_{train} \) and divide it into \( k \) subsets as \( \{D_1, D_2, \ldots, D_k\} \) (Step 2). For each basic learning machine \( L_j \) in level-0 (Step 3): Let dataset \( D_j, 1 \leq j \leq k \) as a validation set, and \( D_{j} = D - D_j \) as a training set in level-0 (Step 4). Then we train the regression model \( L_j \) on dataset \( D_{j} \) (Step 5), and apply \( L_j \) to forecast the samples in \( D_j \), recording the forecasting result as \( \hat{y}_{j} \). At the meantime, apply \( L_j \) to forecast the samples in \( D_{test} \), recording the forecasting result as \( \hat{y}_{test} \) (Step 6). Reconstruct the forecasting results as \( \hat{y}_{train} \) and \( \hat{y}_{test} \) and let it be a new training set and testing set for layer-1 respectively (Step 7). That is, \( D_{new} = (\hat{y}_{train}, \hat{y}_{test}) \) is the dataset used in level-1 space (Step 8). Applying a new base learning machine \( L_{new} \) on \( D_{new} \) in level-1 space, we can then obtain the forecasting value \( \hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) \) of Chl-a concentration (Step 9).

The stacking-based regression model combines KNN, MLP, SVR, XGBoost, and RF. In this paper, \( k\)-fold is set to 4, and the regression models above are respectively used as \( L_1, L_2, L_3, L_4 \) and \( L_{new} \). The stacking modeling process is shown in Fig.4. Within the five models, KNN is a representative of “lazy learning”. The complexity of the K-nearest neighbor regression is low [40], which is applicable to the water quality observation data whose volume increases year by year. MLP is based on the present understanding of the biological nervous system. It is a massive parallel system composed of many processing elements connected by links of variable weights [41] and it has been used widely for water quality factor forecasting [42], [43]. SVR is a model based on a Support Vector Machine to solve regression problems [44]. Regress analysis of SVR is performed via introducing a loss function to project the data into a higher-dimensional feature space for linear regression through a nonlinear mapping. As compared to the MLP model normally with multiple local minima, the SVR gives a unique solution resulting from the convex nature of the optimality problem [45]. XGBoost is an integrated algorithm based on classification and regression tree [28]. As a massively parallel and distributed machine learning solution, XGBoost runs faster for model exploration. It aims to prevent over-fitting and optimize the computation resources, which makes the model performs well when applied to the water quality observation dataset with many features. Furthermore, an early stopping strategy is adopted for MLP and XGBoost to prevent overfitting. RF is a bagging algorithm based on decision trees and introduces random attribute selection in the training process of a decision tree [29]. A random forest can describe both linear and nonlinear relationships, without any additional assumptions concerning either the independent or the dependent variables [46]. This enables us to make Chl-a concentration prediction without knowing the correlation between water quality variables.

### C. cluster-stacking-based approach
Based on the models above, we give the Cluster-stacking-based approach for the Chl-a concentration prediction in coastal waters, which is illustrated as Algorithm 3. It is an integration of the station-oriented clustering model and the stacking-based regression model, for the usage of how to transform data into an appropriate input matrix, and how to combine Algorithm 1 and Algorithm 2 to forecast Chl-a concentration in coastal waters.

**Algorithm 3 The Chl-a Concentration Prediction Based on Cluster-Stacking Model**

| Input: \( D^{ini} = (x^{ini}, y^{ini}) \) |
| Output: \( \hat{y} \) |
|---|
| 1. \( D^{sta} \leftarrow \text{convert}^1 D^{ini} \) |
| 2. \( C \leftarrow \text{Algorithm1} D^{sta} \) |
| 3. \( D^{pre} \leftarrow \text{convert}^2 D^{sta} \) |
| 4. \( \{D_1^{pre}, D_2^{pre}, \ldots, D_k^{pre}\} \leftarrow C = \{C_1, C_2, \ldots, C_k\} \) |
| 5. For each \( D_i^{pre} \) in \( D^{pre} \): |
| 6. \( \hat{y}(D_i^{pre}) \leftarrow \text{Algorithm2} D_i^{pre} \) |
| 7. \( \hat{y} = (\hat{y}(D_1^{pre}), \hat{y}(D_2^{pre}), \ldots, \hat{y}(D_k^{pre})) \) |

The input is the initial version of the water quality observation dataset \( D^{ini} = (x^{ini}, y^{ini}) \). In \( D^{ini} \), \( x^{ini} = (x_1^{ini}, x_2^{ini}, \ldots, x_n^{ini}) \) and \( y^{ini} = (y_1^{ini}, y_2^{ini}, \ldots, y_n^{ini}) \), where \( x_{i,p}^{ini} \) represents a sample of water quality observation data including all observed factors apart from Chl-a, \( x_{i,p}^{ini}, 1 \leq p \leq m \), represents the value of a water
In particular, the dataset $D$ concentration with (Step 6). Obtain the 1-season ahead forecasting value $\hat{D}$ and apply it on based on the corresponding dataset $D$. Divide historical data used in this study. Then let $q$ a water quality observation station, and a time series data of seasonal average Chl-a concentration of tors used in this study; quality factor, and $m$ represents the total number of water quality factors used in this study. $y_{ini}$ represents the corresponding Chl-a concentration with $x_{ini}$, and $n$ denotes the scale of the dataset. In particular, the dataset $D^{ini}$ here is preprocessed already. The output $\hat{y}$ is the forecasting value of Chl-a concentration.

We first convert $D^{ini}$ into $D^{pre}$ by grouping the dataset $D^{ini}$ according to the station number, making it applicable to Algorithm 1 (Step 1). In $D^{sta} = \{s_1, s_2, \ldots, s_q\}$, $s_i$ represents a time series data of seasonal average Chl-a concentration of a water quality observation station, and $q$ is the scale of historical data used in this study. Then let $D = D^{sta}$ as an input, we can get the cluster set $C = \{C_1, C_2, \ldots, C_K\}$ through Algorithm 1 (Step 2). Convert $D^{ini}$ into $D^{pre} = (\hat{x}^{pre}, \hat{y}^{pre})$ by combining each sample with four-season historical records which have the same station number, making it applicable to Algorithm 2 (Step 3). In $D^{pre}$, $x = (x_1^{pre}, x_2^{pre}, \ldots, x_n^{pre})$, $y = (y_1^{pre}, y_2^{pre}, \ldots, y_n^{pre})$, where $x_i^{pre} = (x_{i,1}^{pre}, x_{i,2}^{pre}, \ldots, x_{i,m}^{pre})$ represents a sample of water quality observation data including historical records of 4 seasons, $x_{i,p}^{pre}, 1 \leq p \leq m$, represents the value of a water quality factor, and $m$ is the total number of water quality factors used in this study; $y_{ini}$ represents the corresponding Chl-a concentration with $x_{ini}$. Divide $D^{pre}$ into several sub-datasets $\{D_1^{pre}, D_2^{pre}, \ldots, D_K^{pre}\}$ according to $C = \{C_1, C_2, \ldots, C_K\}$ (Step 4). For dataset $D_l^{pre}$ in $D^{pre}$ (Step 5): Train the stacking regression model based on the corresponding dataset $D_l^{pre}$ through Algorithm 2 and apply it on $D_l^{pre}$. Record the forecasting value as $\hat{y}(D_l^{pre})$ (Step 6). Obtain the 1-season ahead forecasting value $\hat{y}$ of Chl-a concentration (Step 7). The schematic of the proposed methodology is illustrated in Fig.5.

IV. RESULTS AND DISCUSSION

We wrote Python code for the integration of the clustering and stacking approach. To complete the seasonal Chl-a concentration forecasting, we set the input variables as the historical records of four seasons before and the output variable as the one-season-ahead seasonal average Chl-a concentration. All historical records were handled with simplification, combination, and normalization. Python packages including sklearn, XGBoost, pandas, numpy, matplotlib and seaborn were used for these methods. Meanwhile, we set all the parameters not mentioned explicitly in this paper by the default value in the above packages.

A. FEATURE IMPORTANCE ANALYSIS

Fig.6 ranks the top 20 most important indicators obtained from RF (Fig.6(a)) and XGBoost (Fig.6(b)), combined with the average score of these two methods in Fig.6(c). According to the analysis for the survey data of the current season, the significant variables for forecasting one-season-ahead Chl-a concentration is the average sea condition in the last one season ($L1_{SeaC}$), followed by average Chl-a concentration in the same season of last year ($L4_{Chl-a}$), and the average total nitrogen concentration in last one season ($L1_{TN}$).

As the most important indicator for one-season-ahead Chl-a forecasting, $SeaC$ reflects the sea surface conditions, the shape of wave crests and their degree of bursting, and the amount of foam waves, which is closely related to the Chl-a concentration. As for the Chl-a concentration ($L4_{Chl-a}$), the possible reason is the seasonal difference in growth mechanism of phytoplankton. In spring, seawater phytoplankton is in the period of proliferation, while phytoplankton is in decline in autumn because their growth and reproduction ability are diminished. $N$, $P$, $Si$ or the combination factors of them are listed at the top 5 critical factors, such as $L1_{TN}$ and $L4_{Si/P}$. Land source input is the primary source of nutrients in the coastal waters and the main influencing factor of phytoplankton activities. When phytoplankton has steady growth and reproduction ability, they need to absorb nutrients, so $N$, $P$, and $Si$ become three of the significant influencing factors. This importance analysis score list also explains the importance of nutrients to plant growth.

As spatial indicators, longitude and latitude of the observation stations are both listed in the top 20, which verifies...
the necessity of our clustering steps. Besides, many types of chemical and physical indicators or the combination of them are on the list. However, $pH$ is not an important indicator because of its slight change. In Fig.6, some meteorological indicators, hydrological indicators, and indicators related to human impacts are also listed, which indicates the importance of these indicators in the forecasting of Chl-a concentration. It should be noted that constructed factors are shown in none of these three figures. But it doesn’t mean that they are not essential. Simplification factors are not listed in the figures but also in top 50%. Categorical variables are not on the list because they are encoded in One-Hot, which decomposes a single indicator into multiple indicators. Hence, these constructed factors are all taken into consideration in this study.

B. STATION-ORIENTED CLUSTERING RESULTS

We tried to determine the appropriate architectures for each clustering model until $Score_{CH}$ reached the optimum case. Fig.7 shows the comparison of three different clustering
models, k-means cluster, Agglomerative cluster and Gaussian Mixture cluster. In total, different clustering models were evaluated to find the optimal one. We can see from this figure that stations in the same cluster are basically very close in location, further confirming that the cluster idea is practical.

It should be noted that stations of the third type of sub-region (blue circle points) are partitioned into 2 parts, as well as the first type (red star points). The possible reason is that the east side of Rongcheng and Southside of Wendeng are both containing bays, which influence the Chl-a concentration in current, wind, and other factors. And both of the two parts are open sea area may be the possible reason for the situation in the third type of sub-region.

C. MODEL EVALUATION

To verify the effectiveness of the Cluster-stacking approach in Chl-a concentration prediction, we compared the performance of the Cluster-stacking model with that of other 5 models, which are Cluster-KNN, Cluster-MLP, Cluster-SVR, Cluster-XGB, and the regression stacking model without cluster. We divided the available dataset, both observed and constructed features, into two sub-sets for training (80%) and testing (20%), and tried to determine the best parameters for each model until they reached the optimum case of $R^2$ and MAE. The mathematical equations of these statistical indicators are described as (6) and (7).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$  \hspace{1cm} (6)

$$MAE(y, \hat{y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|$$  \hspace{1cm} (7)

In (6) and (7), $\hat{y}_i$ denotes the predicted value of the $i$th observation sample, $y_i$ is the real value of the $i$th observation sample, $\bar{y}_i$ is the average value of all observation samples, and $n$ is the number of the observation samples. The closer $R^2$ is to 1 and the lower MAE values are, the better the model performance will be.

The evaluation results of the models are graphically compared in Fig.9 in the form of scatter plots. The scatter plots can be partitioned into 6 groups corresponding to the 6 models. In each group, the left scatter plot and the right one shows the results of a model on the training set and testing set, respectively. The horizontal and vertical axis respectively express...
the observed value and predicted value. The black line represents $y = x$. The red line is the linear regression fitting of the observed value and predicted value, i.e., when a point falls on the black line, the prediction is exact. The closer the red line is to the dark line, the more accurate the prediction is.

TABLE 3 and TABLE 4 list the training and testing results of the 6 models in 7 types of sub-region and the whole region, concerning the evaluation of MAE and $R^2$. The results in the column of whole region is the integration of forecasting results in all the sub-regions. In these tables, the bold values show the optimal one. As can be seen, Cluster-stacking model shows the best performance with the maximum $R^2$ and minimum MAE in most of the sub-regions, both in training and testing phases.

According to the test results of Fig.9, TABLE 3, and TABLE 4, combined with the station-oriented cluster results, the performance of the regression stacking model with k-means clustering is better than those of any other methods in whole region, with $R^2$ of 0.848 and MAE of 0.665 ug/l in the testing phase.
In the training phase, each model in this study can fit the data reasonably well. For all types of sub-regions and the whole region, the Cluster-stacking-based model in the training phase performs well with $R^2$ higher than 0.95 and MAE less than 0.5 ug/l. Only MLP obtained a little better regression in sub-region of type 2, which can be negligible. However, as for in testing phase, not all models perform well. MLP performs better in $R^2$ than other methods in sub-region of type 2, 4, and 7; yet when concerning MAE, KNN performs best in sub-region of type 3 and 4, MLP performs best in sub-region of type 2 and 7. SVR has stable performance in all sub-regions. In total, when referring to the whole region, the Cluster-stacking model shows the best overall performance with the highest $R^2$ and lowest MAE.
V. CONCLUSION
This paper is based on the historical water quality observation data collected from 2015 to 2018 in Weihai coastal waters. The following conclusions can be drawn from the study.

First, we propose a set of feature extraction methods for water quality factors including simplification, combination, normalization and importance analysis, and a feature selection method based on RF and XGBoost to select relevant variables of Chl-a concentration. The feature importance analysis shows that sea condition in the last season (L1_TN) plays an essential role in forecasting seasonal average Chl-a concentration, followed by average Chl-a concentration in the same season of the previous year (L4_Chl-a) and average value of total nitrogen concentration in the last one season (L1_TN).

Second, we propose a station-oriented clustering model to group the observation stations. Due to the spatial characteristic of Chl-a concentration in coastal waters, station clustering has been a crucial step of Chl-a prediction. By means of clustering, the dataset of waters can be partitioned into several subsets, thus the influence of biological, physical and chemical reactions in waters can be simplified to that of geographical differences. In our study, we have tested many cluster models, such as k-means cluster, Agglomerative cluster and Gaussian Mixture cluster, with a great number of parameter-sets, and find that k-means model performs the best. Since the clustering has shown its importance in Chl-a concentration prediction, it can also be treated as a useful step to forecasting many other kinds of water quality factors. Furthermore, it can provide a meaningful idea in water regions partition.

Third, to forecasting seasonal average Chl-a concentration in coastal waters, we present an integrated approach that includes station-oriented clustering and stacking-based regression. The main idea of the station-oriented clustering is to partition the whole water region into several sub-regions which has similar regularity of Chl-a concentration evolution. In each sub-region, the stacking-based regression is applied to forecast Chl-a concentration and avoid the insufficiency of single models. The evaluation demonstrates that the Cluster-stacking-based method outperforms the approaches of clustering with single regression models such as Cluster-KNN, Cluster-SVR, Cluster-MLP, Cluster-XGBoost, as well as the regression stacking model without cluster.

The evaluation demonstrates the feasibility of the Cluster-stacking-based approach in forecasting Chl-a concentration in coastal waters, but we noticed that there is still a gap between prediction and observation data, so it deserves more efforts to explore the features and refine the model. Besides, how to apply the approach proposed to a broader range of waters will also be our future work.

APPENDIX

TABLE 5 shows the classification criteria used in this study, which are excerpted from the Water Quality Index of Weihai [25].

| TABLE 5. IWQI classification and corresponding concentration limits of different factors. |
|-----------------------------------------------|---|---|---|---|
| | I | II | III | IV |
| COD (mg/l) | 2 | 3 | 4 | 5 |
| DIN (mg/l) | 0.2 | 0.3 | 0.4 | 0.5 |
| DIP (mg/l) | 0.015 | 0.03 | / | 0.045 |
| Petro (mg/l) | 0.05 | / | 0.3 | 0.5 |

REFERENCES
[1] National Research Council and Committee on Earth Studies, Issues in the Integration of Research and Operational Satellite Systems for Climate Research. Washington, DC, USA: National Academies Press, 2000.
[2] W. Ying, J. Hong, X. Zhang, Y. Wang and C. Cheng, “Temporal-spatial variations of chlorophyll-a in the adjacent sea area of Yangtze river estuary,” in Proc. Int. Conf. Geoinfor., 2010, pp. 1–5.
[3] G. Xu, P. Li, K. Lu, Z. Tantai, J. Zhang, Z. Ren, X. Wang, K. Yu, P. Shi, and Y. Cheng, “Seasonal changes in water quality and its main influencing factors in the dan river basin,” Catena, vol. 173, pp. 131–140, Feb. 2019, doi: 10.1016/j.catena.2018.10.014.
[4] Y. Liu, H. Guo, and P. Yang, “Exploring the influence of lake water chemistry on chlorophyll a: A multivariate statistical model analysis,” Ecological Model., vol. 221, no. 4, pp. 681–688, Feb. 2010, doi: 10.1016/j.ecolmodel.2009.03.010.
[5] Y. Zeng, Z. Yang, and J. Liu, “Prediction of the concentration of chlorophyll-alpha for Luhua urban lakes in Beijing City,” J. Environ. Sci.-China, vol. 18, no. 4, pp. 827–831, 2006.
[6] S. Malek, S. M. Syed Ahmad, S. K. K. Singh, P. Milow, and A. Saleh, “Assessment of predictive models for chlorophyll—a concentration of a tropical lake,” BMC Bioinf., vol. 12, no. S13, p. S12, Dec. 2011, doi: 10.1186/1471-2105-12-S13-S12.
[7] H. Wang, X. Yan, H. Chen, C. Chen, and M. Guo, “Chlorophyll-a predicting model based on dynamic neural network,” Appl. Artif. Intell., vol. 29, no. 10, pp. 962–978, 2015, doi: 10.1080/08839514.2015.1097142.
[8] 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, South Korea,” Sci. Total Environ., vol. 502, pp. 31–41, 2015, doi: 10.1016/j.scitotenv.2014.09.005.
[9] Y. Xu, C. Ma, Q. Liu, B. Xi, G. Qian, D. Zhang, and S. Huo, “Method to predict key factors affecting lake eutrophication—A new approach based on support vector regression model,” Int. Biodeterioration Biodegradation, vol. 102, pp. 308–315, Aug. 2015, doi: 10.1016/j.ibiod.2015.02.013.
[10] B. Li, G. Yang, R. Ren, G. Hörmann, J. Huang, N. Frohler, 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, Dec. 2017, doi: 10.1016/j.ecolind.2017.07.033.
[11] Y. Fu, S. Xu, and J. Liu, “Temporal-spatial variations and developing trends of Chlorophyll–A in the bohai sea, China,” Estuarine, Coastal Shelf Sci., vol. 173, pp. 49–56, May 2016, doi: 10.1016/j.ecss.2016.02.016.
[12] J. Huang, J. Gao, and Y. Zhang, “Eutrophication prediction using a Markov chain model: Application to lakes in the yangtze river basin, China,” Environ. Model. Assessment, vol. 21, no. 2, pp. 233–246, Apr. 2016, doi: 10.1007/s10666-015-9472-4.
[13] P. A. Gilibrand, B. Siemering, P. I. Miller, and K. Davidson, “Individual-based modelling of the development and transport of a karelia mikimotoi Bloom on the north-west European continental shelf,” Harmful Algae, vol. 53, pp. 118–134, Mar. 2016, doi: 10.1016/j.hal.2015.11.011.
[14] B. Zhai and J. Chen, “Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China,” Sci. Total Environ., vol. 635, pp. 644–658, Sep. 2018, doi: 10.1016/j.scitotenv.2018.04.040.
[15] F. Shen, J. Chao, and J. Zhao, “Forecasting exchange rate using deep belief networks and conjugate gradient method,” Neurocomputing, vol. 167, pp. 243–253, 2015, doi: 10.1016/j.neucom.2015.04.071.
[16] S. Zhu, X. Lian, L. Wei, J. Che, X. Shen, L. Yang, X. Qiu, X. Liu, W. Gao, X. Ren, and J. Li, “PM2.5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors,” Atmos. Environ., vol. 183, pp. 20–32, Jun. 2018, doi: 10.1016/j.atmosenv.2018.04.004.
