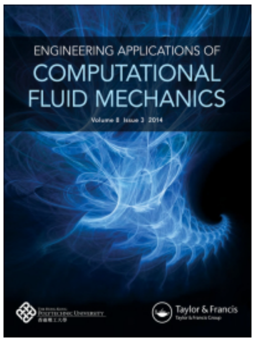


Taylor & Francis
Taylor & Francis Group

Engineering Applications of Computational Fluid Mechanics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcfm20

Spatiotemporal characterization and forecasting of coastal water quality in the semi-enclosed Tolo Harbour based on machine learning and EKC analysis

Tianan Deng, Huan-Feng Duan & Alireza Keramat

To cite this article: Tianan Deng, Huan-Feng Duan & Alireza Keramat (2022) Spatiotemporal characterization and forecasting of coastal water quality in the semi-enclosed Tolo Harbour based on machine learning and EKC analysis, Engineering Applications of Computational Fluid Mechanics, 16:1, 694-712, DOI: 10.1080/19942060.2022.2035257

To link to this article: https://doi.org/10.1080/19942060.2022.2035257

View supplementary material

Published online: 27 Feb 2022.

Submit your article to this journal

Article views: 727

View related articles

View Crossmark data

Citing articles: 1 View citing articles

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2022, VOL. 16, NO. 1, 694–712
https://doi.org/10.1080/19942060.2022.2035257

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Spatiotemporal characterization and forecasting of coastal water quality in the semi-enclosed Tolo Harbour based on machine learning and EKC analysis

Tianan Deng, Huan-Feng Duan and Alireza Keramat

Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR

ABSTRACT

Characterizing and forecasting coastal water quality and spatiotemporal evolution should be significant to coastal ecosystem management. However, high-quality modeling coastal water quality and their spatiotemporal evolutions is rather challenging due to complex dynamic mechanisms especially in a spatially and temporally heterogenous semi-enclosed bay. To this end, this study develops a framework incorporating machine learning (ML) algorithms and the Environmental Kuznets Curves (EKC) analysis to model, analyze and forecast the spatiotemporal variations of water quality indicators for different subzones and seasons in the semi-enclosed Tolo Harbour of Hong Kong. The application results indicate that the developed ML-based framework with an accuracy range of 0.672 ∼ 0.998 is well-suited in forecasting and understanding the coastal water evolution in a semi-enclosed harbour compared to conventional approach. Furthermore, the spatiotemporal characteristics of coastal water quality evolution in this semi-enclosed bay are analyzed and discussed for coastal hydro-environmental management. Moreover, the EKC analysis is also performed for determining the evolutions of essential water quality variables under 95% confidence interval of Hong Kong PCGDP projection and then implemented in the developed ML-based model for future prediction.

ARTICLE HISTORY
Received 21 September 2021
Accepted 22 January 2022

KEYWORDS
Coastal eutrophication; EKC analysis; machine learning; Tolo Harbour; spatiotemporal analysis

1. Introduction

In recent decades, many coastal ecosystems throughout the world have been facing the critical issue of hydro-environment degradation with different degrees owing to the rapid development of coastal cities and urbanization (Chau, 2007; Peng et al., 2012; Xu et al., 2010). The anthropogenic impacts such as coastal discharges of industrial waste and municipal sewage have significantly increased the nutritive contaminants and aggravated the water quality decline (Qiao et al., 2017). These environmental issues, along with other damaging impacts such as water temperature, acid–base, and meteorological conditions, usually stimulate the harmful algal bloom (HAB) events (also known as red tides), which is a typical eutrophic symptom frequently observed worldwide (Lee et al., 2003). This explosive growth and accumulation of algal species to an undesirable level in HAB events often result in a variety of adverse effects on coastal ecosystems, including water discoloration, transparency decrease, oxygen depletion, aquaculture fish kills, and beach closure that causes severe ecological and economic loss, especially in the semi-enclosed water bay areas (Deng et al., 2021; Guo et al., 2020).

The responses to nutrient enrichment of different ecosystems may vary considerably (Dai et al., 2016; Xu et al., 2010), and their leading causes are also very site-specific (Zhang & You, 2017). Particularly, semi-enclosed water bay areas with spatially and temporally heterogeneous conditions in different subzones usually respond differently to eutrophication susceptibility, which makes the simultaneously spatial and temporal analysis more complicated (Xu et al., 2010). Specifically, the pollutant matters easily get trapped in the water near the inner part due to slight tide flushing, longer water retention time, and weaker water exchange, making the areas more prone to eutrophication. On the contrary, the water near the outer open sea is often more affected by oceanic dynamics where frequent hydrodynamic exchanges with open seas impede the accumulation of nutrient substances (Li et al., 2014; Peng et al., 2012; Qiao et al., 2017). Therefore, in those semi-enclosed bay areas, spatial and temporal variations of hydrodynamic and hydro-environmental conditions can be very significant, even in a small-scale zone (Peng et al., 2012). Besides, various observations and practices have also demonstrated that it is essential to separately deduce the different causalities of eutrophica-

CONTACT Huan-Feng Duan hf.duan@polyu.edu.hk

Supplemental data for this article can be accessed here. https://doi.org/10.1080/19942060.2022.2035257

tion for different subzones and seasons to identify appropriate management strategies of water quality regulations for other portions over a heterogeneous ecosystem (Xu et al., 2004, 2010).

Recently, novel models incorporating machine learning (ML) algorithms demonstrate better performance in predicting eutrophic evolutions with higher accuracy than conventional mechanistic and regression models (Alizadeh et al., 2018; Li et al., 2014; Zeng et al., 2017). Among ML models, artificial neural network (ANN) is the most widely adopted algorithm in hydraulic and ecological modeling due to its superior self-adaptability, self-organization, and robustness (Deng et al., 2021; Hadjisolomou et al., 2021; Zeng et al., 2017). Nevertheless, a successful application of ANN model depends on a tremendous amount of training data and expertise in tuning various structural hyperparameters (Li et al., 2014; Xie et al., 2012). Random Forest (RF) is another robust machine learning algorithm that resorts to the concept of ensemble learning. Unlike other models trained individually, the ensemble learning model aggregates a group of simple models as a more powerful learner to potentially produce a vastly superior prediction performance than single models (Shin et al., 2020). In this regard, RF has also received broad interest in modeling and realizing environmental systems (Sattari et al., 2020; Zeng et al., 2017).

On the other hand, forecasting the future trend of water quality is still a main challenge of ML paradigm because the historical data used lacks the future information. Nevertheless, in ecological economics, some environmental quality indices are discovered to have significant associations with the economic development level of adjacent areas (He et al., 2014; Sarkodie & Strezov, 2018). This kind of association can be commonly concluded by the Environmental Kuznets Curve (EKC), which hypotheses a statistical relationship between environmental and socioeconomical indices (Diao et al., 2009). For example, Sarkodie and Strezov (2018) confirmed the validity of the EKC hypothesis in Australia, China, Ghana, and the USA; Chen et al. (2018) adopted EKC models to extrapolate the coastal environment evolutions with the economic developments in southeast China. In this case, the EKC model is expected to provide a possible solution to forecast the future eutrophic problems upon incorporating the ML models. However, the shapes of EKC curves are usually found to be very site-specific (Diao et al., 2009), and few studies disclose the EKC relationships along the Hong Kong coasts.

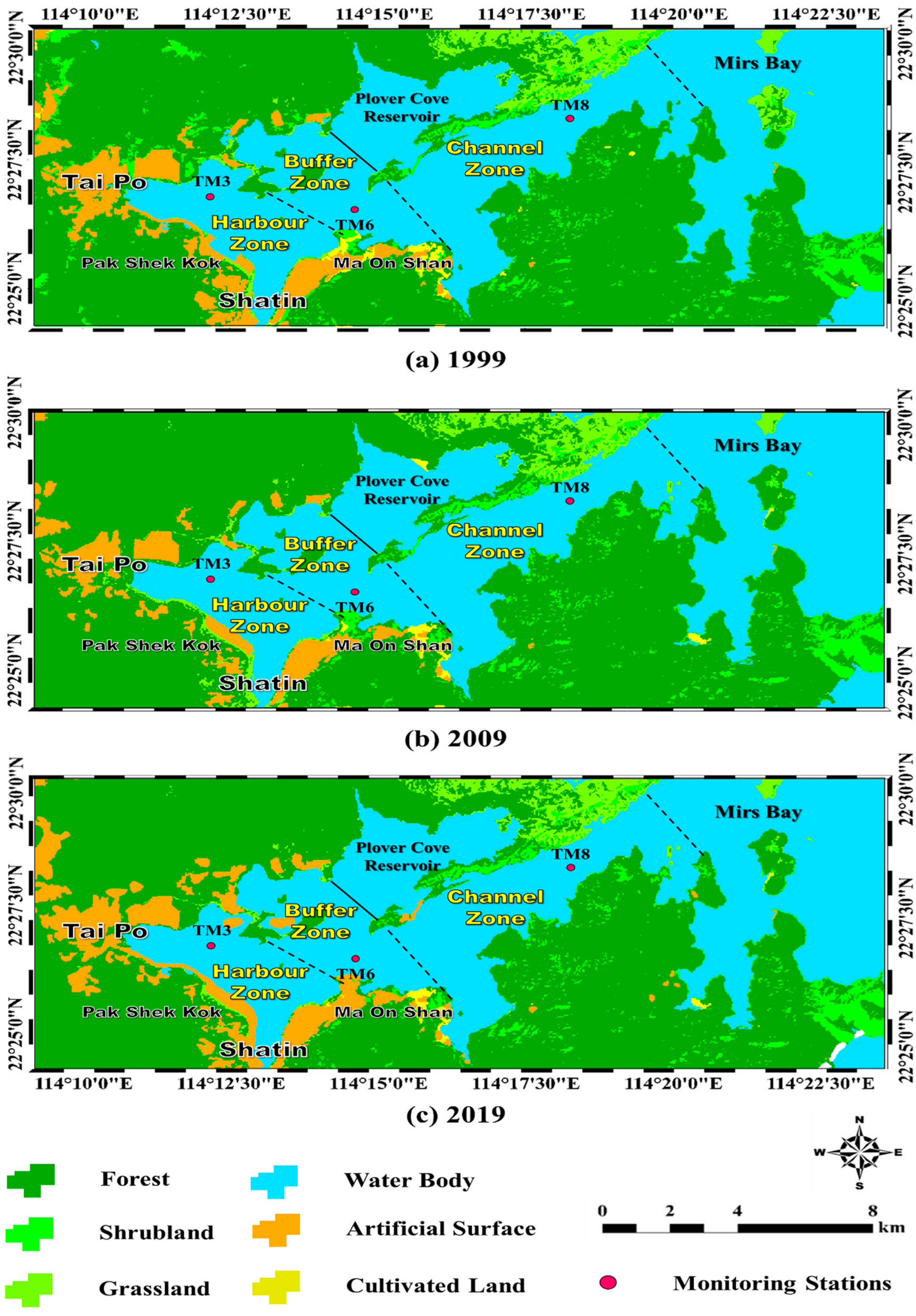
Among previous studies on ML-based eutrophic modeling especially in semi-enclosed bays, two major research gaps should be summarized as: (1) most researchers have focused on revealing the temporal variation of water quality of a single point, or as a whole of water bay, while the spatial differences are usually neglected (Li et al., 2014; Xie et al., 2012; Yu et al., 2021); (2) although a mass of ML-based analyses have demonstrated the validity of ML models in ecological modeling, most of them are only devoted to calibration and validation of subsistent data in essence (Hadjisolomou et al., 2021; Mamun et al., 2020; Shin et al., 2020) and seldom research aim to predict the future trend of eutrophic evolution owing to the uncertainty and complexity in environmental extrapolation.

As an extension to the previous study by authors (Deng et al., 2021), the main objective of this paper is to develop an ML-based framework for coastal water quality analysis in a semi-enclosed water bay in Hong Kong (i.e. Tolo Harbour) to (1) precisely characterize the spatiotemporal evolutions of objective parameters, (2) provide a practicable procedure for extrapolating the future trend of water quality indicators by combining a socioeconomic model. Unlike previous studies taking a semi-enclosed bay as a whole, we split the study area and period into different subzones and seasons for a more specific spatiotemporal analysis. To this end, we firstly investigate the effectiveness of the ANN and RF schemes compared with the MLR in eutrophication modeling for a spatially and temporally heterogeneous semi-enclosed ecosystem. Then, a seven-step framework for such modeling in a semi-enclosed water bay is proposed. For demonstrating the applicability of the developed framework, the eutrophication quantified by two typical response indicators, i.e. Chlorophyll-a concentration and transparency extent, is assessed, and the key contributors are interpreted over different seasons and subzones. After validating the effectiveness of the ML models for the water quality states, they are then extrapolated by innovatively incorporating EKC models using the projection of the Gross Domestic Products per capita (PCGDP) over the next decade in Hong Kong. On this basis, the well-trained ML models can predict the evolution of Chlorophyll and transparency in Tolo Harbour.

2. Material and methods

2.1. Study area

The Tolo Harbour, located in the north-eastern part of Hong Kong (see Figure 1), is an almost land-enclosed seawater body with only a narrow channel way out to Mirs Bay (Chau, 2007; Deng et al., 2021). The water area of Tolo Harbour is about 52 km^{2,} with an average length and width of 16 and 3 km, respectively (Sin & Chau, 1992). Figure 1 shows the change of geographical pattern and land-use pattern of Tolo Harbour over the past



(a) 1999

(b) 2009

(c) 2019

Figure 1. Long-term evolution (1999 ~ 2019) of geographical pattern and land-use around the semi-enclosed water bay of Tolo Harbour (Data Source: GLOBELAND30, available at http://www.globallandcover.com).

two decades. According to different geographic characters and hydrodynamic conditions, the whole west–east water body is generally divided into three subzones (i.e. Harbour, Buffer and Channel Zone) by dotted lines in Figure 1 (Muttil & Chau, 2007; Sin & Chau, 1992; Xu et al., 2004). Generally, the tidal current velocity of the inner part is only 0.01 ∼ 0.02 m/s, while that is about 0.2 ∼ 0.3 m/s at the outer channel part (Xu et al., 2010). Hence, the contaminant and nutriments trapped in the Harbour subzone are usually flushed out in a more extended period than the other two subzones (Sin & Chau, 1992).

The land-use type along the Tolo Harbour also shows significant heterogeneity. To be specific, the inner Harbour zone is almost surrounded by artificial surface including two new developed satellite towns, 'Tai Po' and 'Shatin' where the municipal and industrial waste from the local sewage discharge became the primary pollution source of the harbour zone (Deng et al., 2021; Muttil & Chau, 2007). There is also cultivated land producing agricultural waste along the south coast of the buffer zone, although artificial surfaces are gradually replacing it. In contrast, the land along the outer channel zone remains undeveloped state encircled with forest, shrubland, and grassland where the direct rainfall and runoff become the input of potential pollutants. Furthermore, the completion of reclamation projects at Tai Po (over 300 hectares), and Pak Shek Kok (about 74 hectares) in the 1990s ∼ 2000s resulted in the further decrease of the Tolo water area. After 2010, the new developments of Ma On Shan and new openings of Hong Kong Science Park (at Pak Shek Kok) increased industrial and residential land with a growing sewage discharge.

2.2. Modeling data and statistics

Monthly data of water quality and meteorology from 1999 to 2019 monitored by the Environmental Protection Department (EPD) (https://cd.epic.epd.gov.hk/) of Hong Kong and Hong Kong Observatory (HKO) (https://www.hko.gov.hk/) are used for modeling in this study. Three monitoring stations are selected to represent three subzones: TM3 (22°26′51″N, 114°12′10″E) for harbour zone, TM6 (22°26′38″N, 114°14′30″E) for buffer zone, and TM8 (22°28′24″N, 114°18′00″E) for channel zone. Based on the classification of trophic states in Table 1 suggested by NALMS (1990) and Sin and Chau (1992), Chlorophyll-a concentration (Chl-a, μg/L) and Secchi Disc Depth (SDD, m) are chosen as the response variables of eutrophication. Therein, Chl-a implies the algal biomass and SDD determines the water transparency. In addition, nine potential driven factors commonly used in previous eutrophication analyses are selected as modeling inputs (Mamun et al., 2020; Muttil & Chau, 2007; Xu et al., 2010), including Total Phosphorus (TP, mg/L), Total Nitrogen (TN, mg/L), Water Temperature (WT, °C), Suspended Solids (SS, mg/L), pH value, Dissolved Oxygen (DO, mg/L), 5-day Biochemical Oxygen Demand (BOD_5, mg/L), Rainfall (mm), Wind Speed (WS, m/s).

Table 1. Trophic state classification for Chl-a concentration and transparency.

Parameters	Oligotrophic	Mesotrophic	Eutrophic	Hypereutrophic
Chl-a	< 4 μg/L	4 ∼ 10 μg/L	10 ∼ 25 μg/L	> 25 μg/L
SDD	> 4 m	2 ∼ 4 m	1 ∼ 2 m	< 1 m

Figure 2 shows the variation pattern of the original annually average data span from 1999 to 2019 (i.e. 20 years' field data). Considering the climate feature of Hong Kong, all data for each subzone is classified into four seasons in this study: Spring (Mar – May), Summer (June – August), Autumn (September – November), and Winter (December – February). The spatial and temporal characteristics of environmental parameters and response variables are provided in Table S1 in the supplementary materials for the studied sea bay area.

According to the land-use heterogeneity and seasonal climate features, here we specify the spatial and temporal characteristics by separate modeling on different subzones and seasons.

2.3. Machine learning and regression methods

In this study, two different machine learning algorithms (i.e. artificial neural network and random forest) are applied for comparative investigation for Chl-a and transparency (SDD) prediction compared with a conventional statistical regressor (i.e. multiple linear regression).

2.3.1. Artificial neural network

Figure 3 shows the structure of the artificial neural network (ANN) adopted in this study. The typical architecture of ANN models consists of multiple functional layers, namely an input layer, one/multiple hidden layer(s), and an output layer, with interconnected artificial neurons. In practice, each neuron node accepts inputs from the precedent layer and generates output to the next layer (Cho et al., 2011; Deng et al., 2021; Lee, 2004) by conducting a simple nonlinear operation as:

$$a_j = f\left(\sum\nolimits_i w_{ji} a_i + b_j\right) \tag{1}$$

where a denotes the nodal data; w represents the weights along with interconnected nodes; b is the bias; f is the transfer function; the subscript i and j represent the node number of two adjacent layers.

This ANN model learns the implicit relationship by error back propagation algorithm that iteratively adjusts connecting weights w to minimize the global error (Gupta, 2013; Whittington & Bogacz, 2019). For a well-trained network, the relative importance of each input

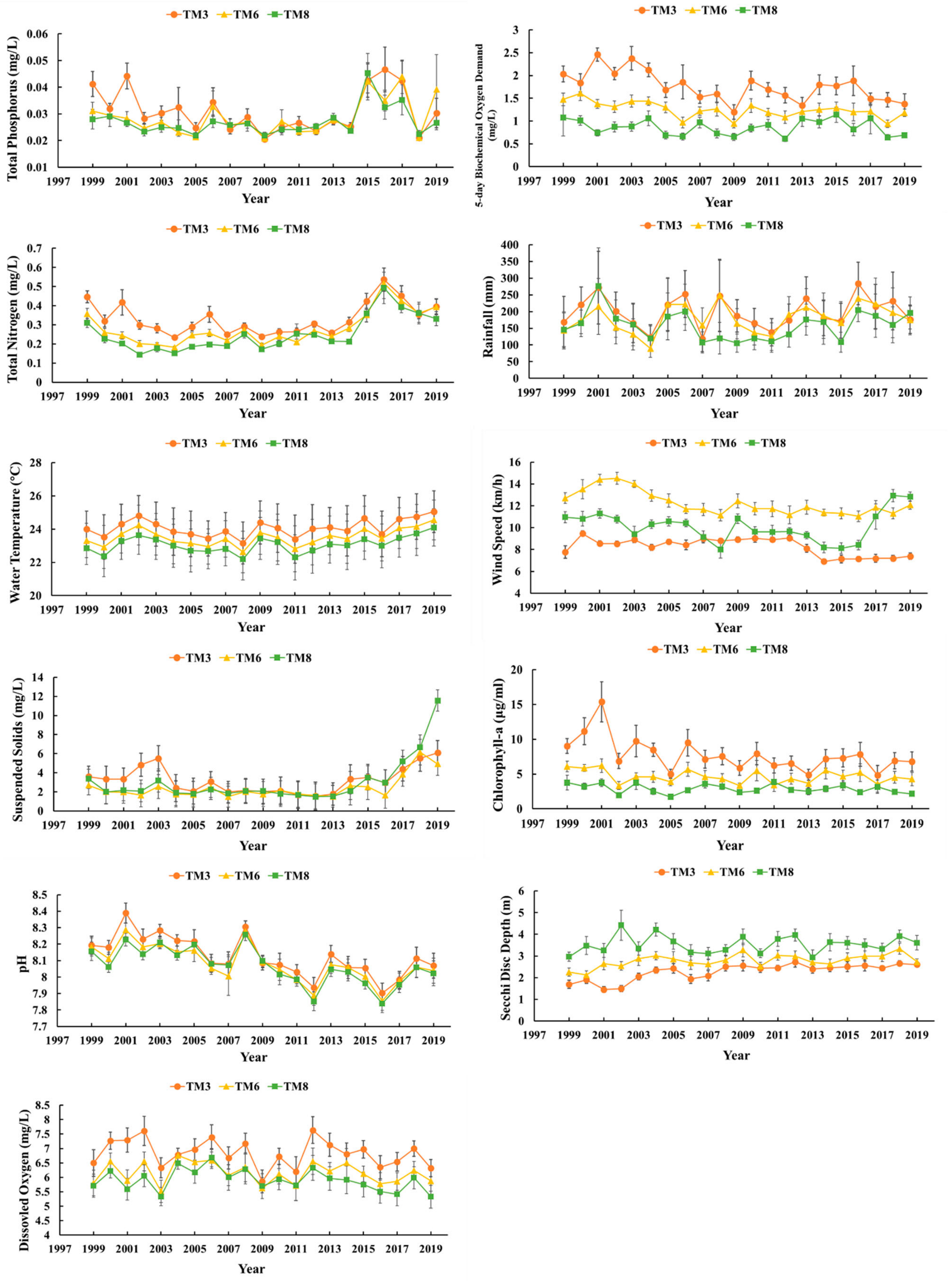


Figure 2. Mean annual variation of environmental variables at three stations (TM3, TM6 and TM8) from 1999 to 2019. (Bars = ±1 Standard Deviation, *n* = 21).

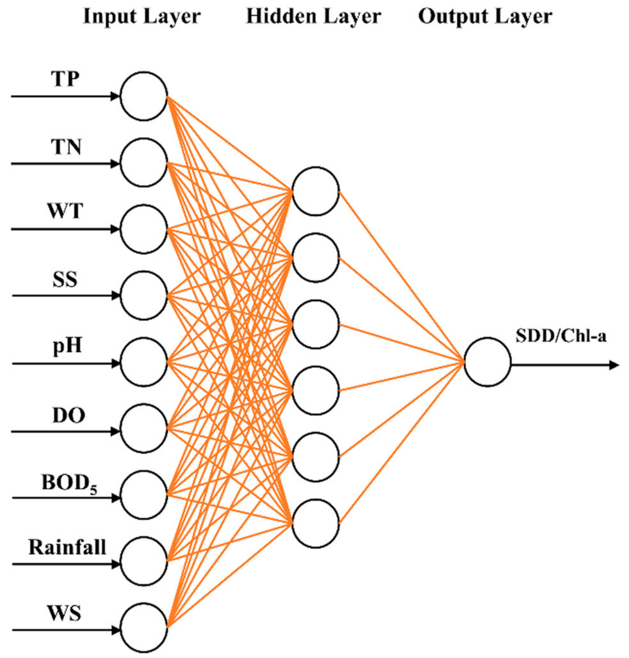


Figure 3. Schematic of artificial neural network (ANN) model and application principle.

can be calculated based on the ‘weight’ method as follows:

$$Q_{ih} = \frac{|w_{ih}|}{\sum_{i=1}^{ni} |w_{ih}|} \tag{2}$$

$$RI_i^{ANN} = \frac{\sum_{h=1}^{nh} Q_{ih}}{\sum_{h=1}^{nh} \sum_{i=1}^{ni} Q_{ih}} \times 100 \tag{3}$$

where w_{ih} is the weight connecting input and hidden layer; ni and nh are the numbers of input and hidden nodes. After comparing the effectiveness of different network structure, a single hidden-layered network is built with nine input nodes, six hidden nodes, and one output node. The Sigmoidal function is selected as the transfer function for each neuron.

2.3.2. Random forest

Differently from single machine learning algorithms, Random Forest (RF) is an ensemble learning method in which decision trees are incorporated as the basic units. Each decision tree is constructed on a subset randomly subsampled from the original dataset by the bagging approach (Kehoe et al., 2015; Sattari et al., 2020). The final output of an RF model is generated by aggregating (i.e. voting for classification and averaging for regression) all separate predictions of each tree in the forest (Zeng et al., 2017). Since the ensemble method combines many tree models, RF is believed to provide a better generalization performance and less possibility of overfitting (Zeng et al., 2017). For clarity, the schematic of the RF model is given in Figure 4.

Essentially, the generating process of decision trees depends on the selections and comparisons of different features. In this regard, RF evaluates which features (inputs) are comparatively important by the mean changes of out-of-bag error (errOOB) (Eq. (4) ~ (5)), which refers to the forecast error on unselected samples, with no additional manual extraction and judgment needed (Kehoe et al., 2015).

$$I^i = \sum_{j=1}^{t} \frac{(errOOB_2^i - errOOB_1^i)}{t} \tag{4}$$

$$RI_i^{RF} = \frac{I^i}{\sum_{i=1}^{n} |I^i|} \tag{5}$$

where $errOOB_1^i$ and $errOOB_2^i$ represent the out-of-bag error before and after random permutations of i^{th} feature; t is the maximum tree number in the forest; I^i denotes the importance of i^{th} feature.

A forest of 200 trees is built for an ensemble with the bagging approach in the present study. To assure the independent and identical distribution of sampling set, the data used for generating each tree is put back to the total dataset each time for possible reselection in the following steps (Kehoe et al., 2015). The whole generating process of different individual tree models is implemented in a parallel manner.

2.3.3. Multiple linear regression

Multiple linear regression (MLR) is a multivariate statistical technique that predicts the dependent variable using various independent variables. The purpose of applying MLR algorithm is to find the optimum relationship that fits the multiple observed data by determining the best combination of coefficients (Aiken et al., 2012; Cho et al., 2011). The general representation of MLR formula can be expressed as:

$$y_m = \sum_{i=1}^{n} \beta_i x_{mi} + \beta_0 + \epsilon \tag{6}$$

where x and y are independent variables (input) and dependent variable (output) respectively; β_0 is the y-interception and β_i is its regression coefficients; ϵ is the residual term;n and m denote the number of input variables and sampling size, respectively.

Theoretically, the individual contribution of each variable can be estimated by the absolute magnitude of the coefficient β_i. The relative importance values in MLR can be calculated by:

$$RI_i^{MLR} = \frac{\beta_i}{\sum_{i=1}^{n} |\beta_i|} \tag{7}$$

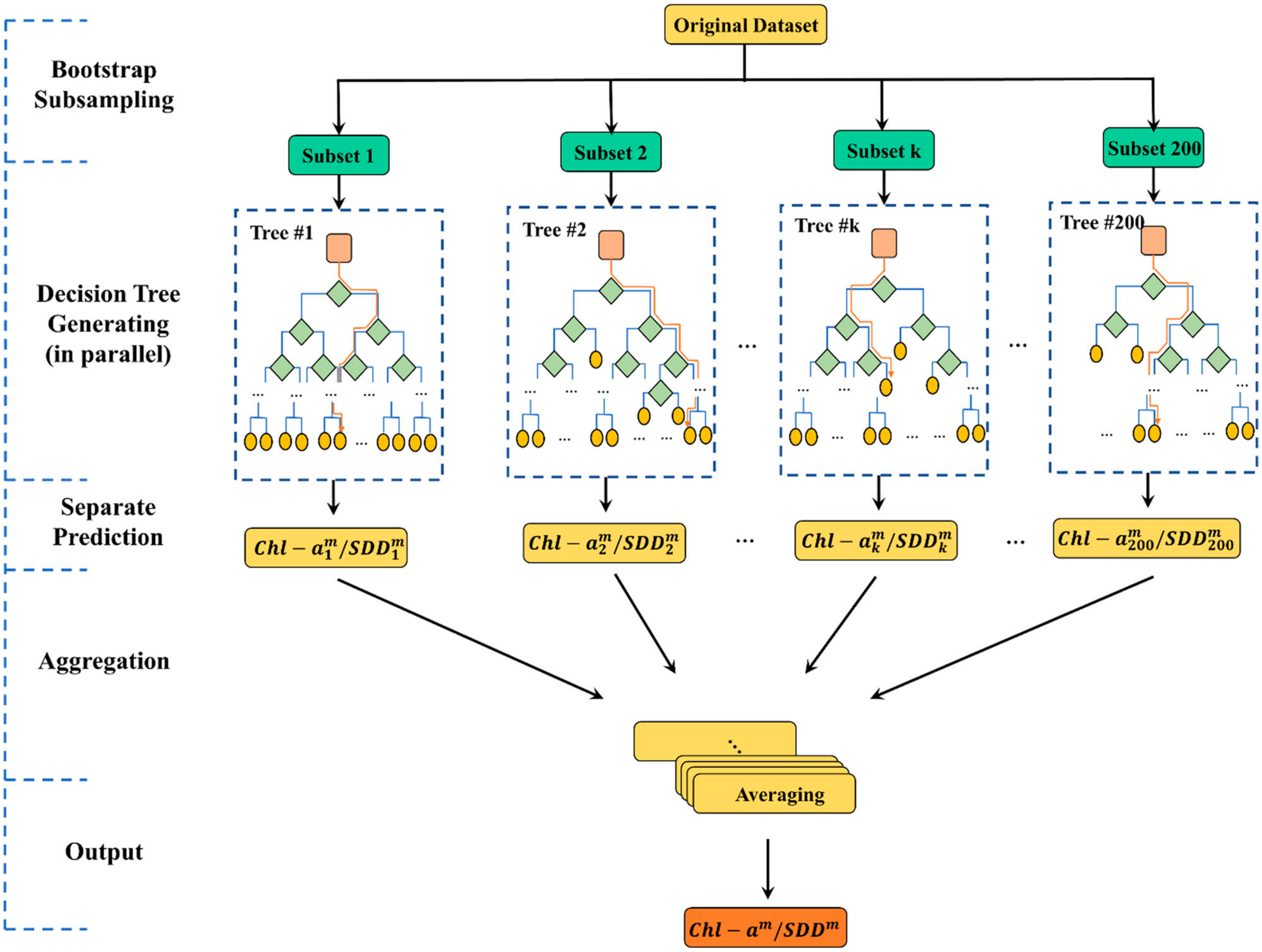


Figure 4. Schematic of random forest (RF) model and application principle.

2.4. EKC model and PCGDP projection

The Environmental Kuznets Curve (EKC) describes the potential relationship between ecological indices and Gross Domestic Products per capita (PCGDP) by building a multinomial curve model (Chen et al., 2018; Diao et al., 2009), which is generally written as the following equation:

$$y_t = \alpha_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + \beta_4 z_t \tag{8}$$

where y is the environmental estimator; x is the socioeconomic indicator represented by PCGDP in this study; z denotes the other factor contributing to the environmental degradation; α_0, β_1, β_2, β_3 and β_4 are the intercept and coefficients of the corresponding term to be determined; t denotes the time parameter. Note that all variables in the EKC analysis are annually averaged in this study.

Once the EKC relationships are determined, the future variation of the corresponding water quality indices can be forecasted based on the future projection of PCGDP. Since PCGDP is a nonstationary series, the Auto-Regressive Integrated Moving Average (ARIMA) method on first-order difference is used to project PCGDP in the next ten years with a 95% confidence level.

2.5. Model performance evaluation

In order to evaluate the modeling performance, three indicators are adopted to quantify the modeling accuracy, namely mean absolute error (MAE), root mean squared error (RMSE) and correlation coefficient (CC). To be specific, MAE can be mathematically calculated as Eq. (9), which gives the arithmetic average of global deviation between prediction and observation (Mamun et al., 2020; Xie et al., 2012).

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{9}$$

Theoretically, the model with better performance has a lower MAE value. The estimator RMSE is presented in Eq. (10), which calculates the squared root of the average

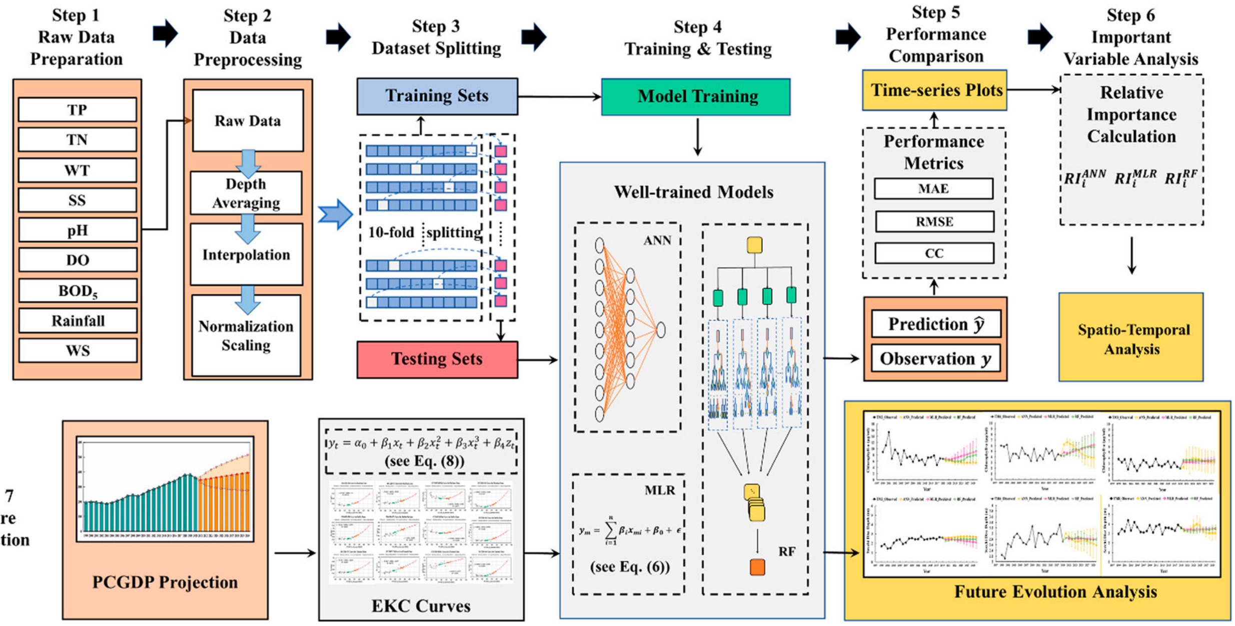


Figure 5. Flowchart of the seven-step coastal water quality analysis framework based on ANN, RF and MLR models.

residual between predicted and actual values (Shin et al., 2020; Zhang & You, 2017).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{m}} \tag{10}$$

The model with smaller RMSE can be considered to have better accuracy. The correlativity between predicted data and actual measured data can be statistically evaluated by the Correlation Coefficient (CC). As a rule, a model with a CC closer to 1 indicates a better goodness-of-fit (Deng et al., 2021). The following equation can calculate CC values:

$$\text{CC} = \frac{\sum_{i=1}^{m} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2}} \tag{11}$$

where m represents the number of the sample size used for modeling; y_i and $\hat{y}_i$ denote the real observation and its estimated value, respectively; parameters capped with a bar refer to the arithmetic averaging.

2.6. Workflow for coastal eutrophication analysis

This study aims to use different machine learning/regression models to simulate and characterize the spatial and temporal evolutions of Chl-a concentration and transparency (SDD). It enables the interpretation of the essential variables over different seasons (time domain) and different subzones (space domain) and estimates the future change of eutrophication in the Tolo Harbour. To this end, seven key steps are outlined in Figure 5 and elaborated as follows.

- Step 1: Raw data preparation – Collecting raw data of variables potentially related to the eutrophication issues over the study period.
- Step 2: Data preprocessing – The raw data measured at different depths is depth-averaged firstly; then the monthly data is interpolated into daily data to increase the number of samples; lastly, the information is scaled to the range of $-1 \sim 1$ to ensure the different variables are at the same order of magnitude.
- Step 3: Dataset Splitting – In this study, a 10-fold splitting method is adopted to produce 10 pairs of training/testing sets prepared for 10-run modeling.
- Step 4: Model training and testing – Three different models are trained separately using a set of training data while the corresponding testing dataset is used to test the accuracy of the trained model.
- Step 5: Performance Comparison – model performances are evaluated based on three metrics, i.e. MAE, RMSE, and CC. The time-series plots of both observation and prediction results are plotted for visual comparison. Considering the 10-fold cross-training adopted in this study, all these performance indicators used in this study are averaged with 10 runs.
- Step 6: Importance Variable Analysis – Relative Importance (RI) values are calculated and evaluated for different seasonal and spatial scenarios for comparison.
- Step 7: EKC model development – To build desired models for each critical variable and estimate future evolutions based on the next 10-year PCGDP projection. Accordingly, the future variation trends of Chl-a and SDD can be estimated by well-trained ML models.

All these algorithms and models (ANN, RF, and MLR) are implemented through in-house coding in the MATLAB program platform to achieve better application flexibility.

3. Results and analysis

3.1. Spatial characteristics of Chl-a prediction in different zones

Figure 6 presents the 20-year Chl-a evolution at the three inspected locations (as shown in Figure 1) compared with the predictions based on MLR, ANN, and RF models. These time series plots show that the highest Chl-a concentration level with a maximum of 42.40 μg/ml is found in the Harbour zone, fluctuating more violently than the buffer and channel sections. For inspection, the trophic status of the water body in Tolo Harbour classified in Table 1 is used herein for illustration. During the early periods from 1999 to 2003, the trophic status at TM3 was almost mesotrophic and eutrophic with even three hypereutrophic events occurred, in line with the relatively high annual levels of TP, TN DO, and BOD_5 in those years (Figure 2). Afterwards, the average concentration has been maintained between the eutrophic and mesotrophic state, though oligotrophic conditions have been frequently observed, especially after 2011. In the buffer zone, the trophic condition was almost oligotrophic to mesotrophic, but increasing cases of eutrophication have been observed recently. Channel zone is the most deficient productivity area with a low Chl-a level and maintains almost oligotrophic over the whole study period.

In Table 2, three descriptors, namely MAE, RMSE and CC, are adopted to quantify the performances of different models. Overall, the modeling performances of training set are slightly better than the results of the testing set without overfitting problems for all methods. Specifically, RF model provides a better agreement (CC over 0.99) than ANN and MLR models in simulating Chl-a growth. Meanwhile, the MLR presents higher training errors at TM3 station ($RMSE = 3.526$, $MAE = 2.611$) than ANN and RF. Regarding spatial comparison, the ANN algorithm obtains the best testing performance in terms of correlation at TM3 ($CC = 0.824$) in comparison to TM6 and TM8. As a result, the relatively low MAE and RMSE but high CC values demonstrate the favorable predictive availability of ANN and RF, whereas the MLR performs worst with the lowest goodness-of-fit.

Figure 8(a) exhibits the relative importance (RI) values of different external influence variables in terms of Chl-a prediction in various spatial domains of Tolo Harbour. In general, the better the model performs, the more reliable its interpretation of essential variables is. Based on the RI values, the BOD_5, which represents organic matter content, is recognized as the foremost important variable in all three inspected regions of Tolo Harbour. Moreover, the Chl-a concentration in the Harbour subzone is also largely influenced by nutrients (TN and TP) and SS, while the impacts of them may decrease adjacent to the buffer region.

3.2. Temporal characteristics of Chl-a prediction in different seasons

The variation patterns of Chl-a results over the study period are exhibited in Figure 7, indicating seasonal changes from spring to winter. In both summer and autumn periods, relatively high concentration levels in Chl-a have been observed in Tolo Harbour. Remarkably, the hypereutrophic cases have appeared in the summer and autumn periods of 2000, 2001, and 2003. On the contrary, the trophic status of Tolo Harbour remains weakest in wintertime, during which the oligotrophic condition is more likely to occur. However, it also shows the frequent winter eutrophications since 2016 but recovered by 2019. The performances of different models are elaborated in Table 3 in supplementary material, revealing that the lowest modeling errors (MAE and RMSE) are found in winter while the highest ones are found in spring. Furthermore, the CC values by RF are found to be much higher than ANN and MLR during all seasons in both training and testing stages.

The RI indexes for Chl-a prediction over the four seasons are presented in Figure 8(b). During all seasons, BOD_5 is ranked as the foremost decisive contributor to algal growth. In addition to BOD_5, SS is also suggested as another salient factor during spring and winter periods as the concentration varies largely. However, the relative significances of other attributes are nearly equivalent in the summer and autumn periods. The modeling results also disclose that TN is the dominant nutrient limiting algal growth in different seasons in Tolo Harbour.

3.3. Spatial characteristics of transparency prediction in different zones

As shown in Figure S1 (in supplementary), the transparency of Tolo Harbour shows a spatial gradient since the average secchi disc depth decreases from the outer channel part towards the inner harbour zone. Notably, from 1999 to 2003, the harbour section has suffered from immense turbidity with SDD of almost less than 2 m, which was then maintained as the mesotrophic state (2 ∼ 4 m) after a significant rise during the period from 2003 to 2006. By contrast, the buffer zone water is more transparent whose trophic status is almost mesotrophic

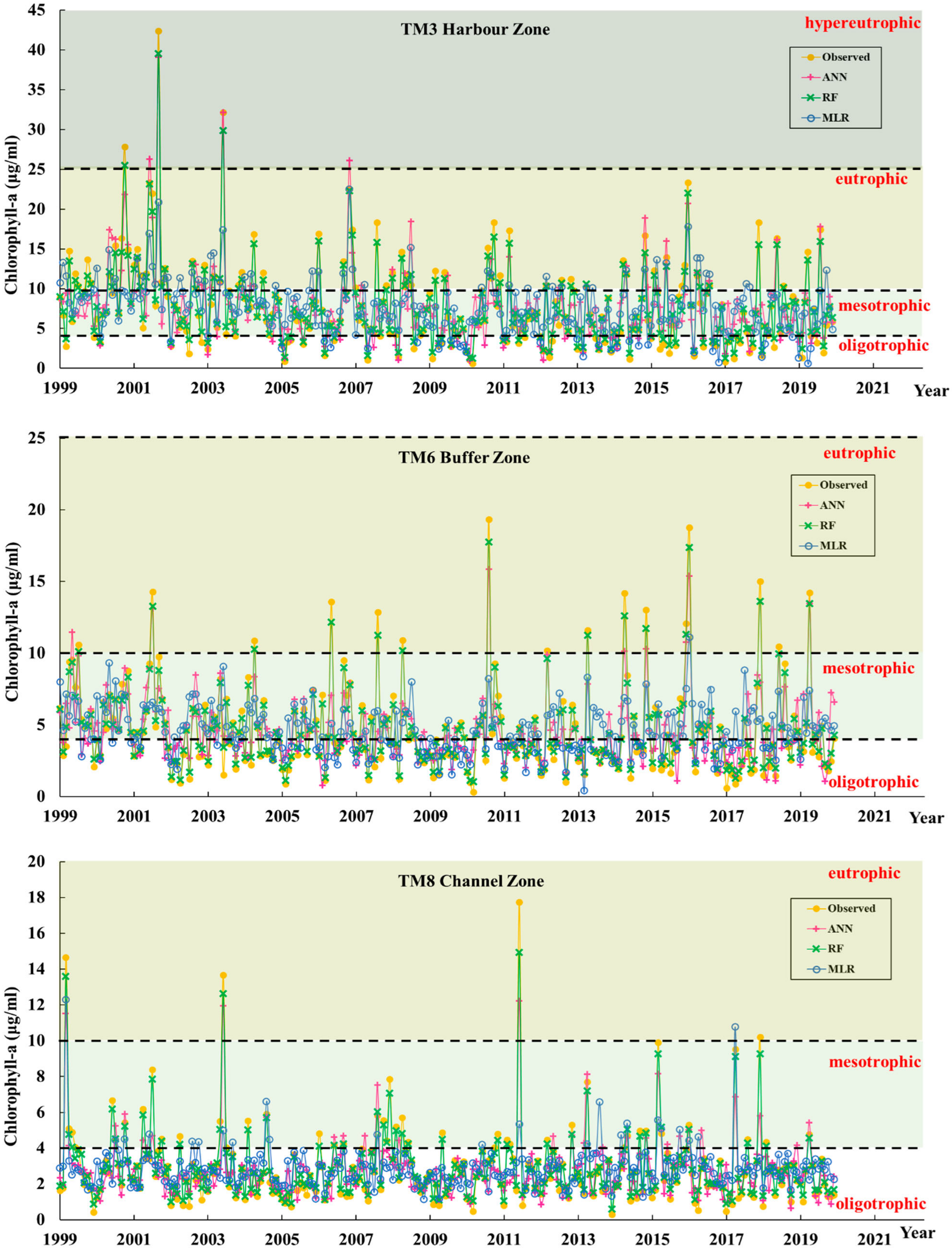


Figure 6. Comparison of 20-year spatial variation of observed and predicted Chl-a concentration results based on ANN, RF and MLR models at TM3, TM6 and TM8.

Table 2. Modeling performances of Chl-a prediction at harbour, buffer and channel subzones using different methods, represented by MAE, RMSE and CC.

	TM3 (Harbour Zone)			TM6 (Buffer Zone)			TM8 (Channel Zone)		
Performance Estimators	ANN	RF	MLR	ANN	RF	MLR	ANN	RF	MLR
				Training Set					
MAE	1.948	0.231	2.611	1.954	0.125	1.592	0.792	0.089	0.999
RMSE	2.546	0.374	3.526	2.573	0.204	2.137	1.055	0.165	1.433
CC	0.833	0.998	0.644	0.829	0.998	0.557	0.795	0.997	0.568
				Testing Set					
MAE	1.987	0.411	2.616	1.975	0.224	1.594	0.805	0.158	1.001
RMSE	2.603	0.655	3.533	2.613	0.357	2.140	1.088	0.296	1.432
CC	0.824	0.992	0.641	0.822	0.994	0.554	0.780	0.990	0.561

Table 3. Modeling performances of Chl-a prediction in spring, summer, autumn and winter using different methods, represented by MAE, RMSE and CC.

				Spring		
	Training Set			**Testing Set**		
Performance Estimators	**ANN**	**RF**	**MLR**	**ANN**	**RF**	**MLR**
MAE	1.325	0.173	1.869	1.365	0.308	1.872
RMSE	1.798	0.306	2.574	1.855	0.531	2.575
CC	0.855	0.997	0.671	0.845	0.991	0.670
				Summer		
	Training Set			**Testing Set**		
Performance Estimators	**ANN**	**RF**	**MLR**	**ANN**	**RF**	**MLR**
MAE	1.191	0.179	1.883	1.219	0.316	1.887
RMSE	1.659	0.323	2.665	1.700	0.552	2.671
CC	0.908	0.997	0.740	0.902	0.992	0.738
				Autumn		
	Training Set			**Testing Set**		
Performance Estimators	**ANN**	**RF**	**MLR**	**ANN**	**RF**	**MLR**
MAE	1.121	0.149	1.723	1.150	0.265	1.727
RMSE	1.517	0.314	2.591	1.579	0.530	2.592
CC	0.930	0.998	0.779	0.923	0.993	0.778
				Winter		
	Training Set			**Testing Set**		
Performance Estimators	**ANN**	**RF**	**MLR**	**ANN**	**RF**	**MLR**
MAE	0.946	0.142	1.447	0.971	0.252	1.450
RMSE	1.290	0.272	2.014	1.335	0.466	2.018
CC	0.925	0.997	0.806	0.920	0.992	0.806

with very few eutrophic events, and no hypereutrophic issue occurs. The channel part follows a similar pattern but with an even deeper average depth and more frequent oligotrophic cases.

The modeling performances are listed in Table S2. The training CC values do not display any difference by RF over different regions (CC = 0.997), while the best testing CC (0.992) can be observed in harbour part. Moreover, ANN performs better at TM3 with the lowest MAE and RMSE among all these three stations. Moreover, the relatively low CC values (less than 0.5) indicate that MLR performs unsatisfactorily, especially in the channel subzone with the CC even less than 0.3.

According to modeling results of the two ML methods (ANN and RF) (Figure S3(a)), SS and TN can be identified as the key factors to water turbidity (i.e. low transparency) for harbour subzone. However, it is also noted that both the nutrients (i.e. TN and TP) and organic pollutants present even higher contributions than SS from the buffer zone to the channel zone. Moreover, the precipitation also contributes important influences on transparency in the channel zone compared to the other two subzones.

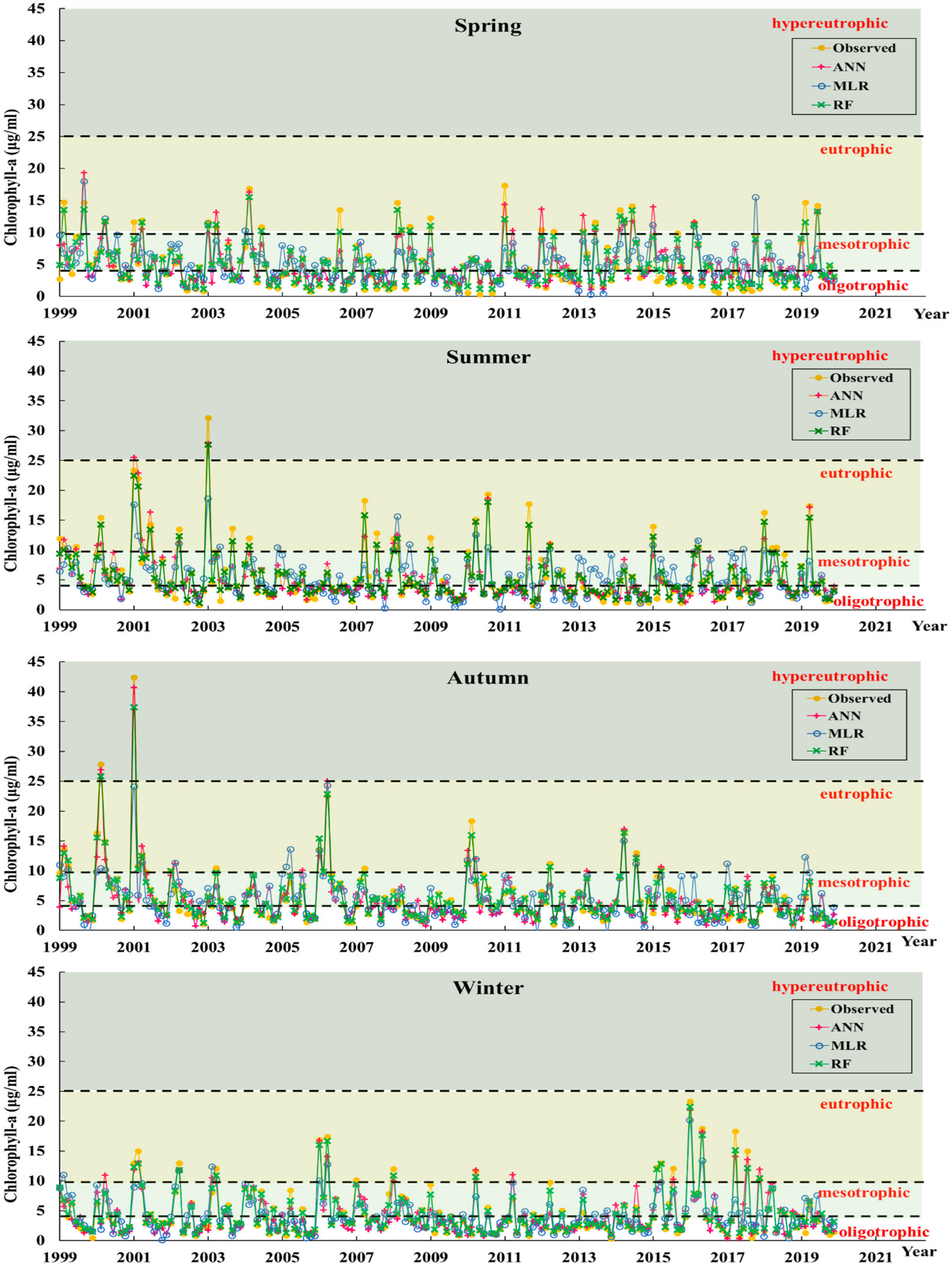
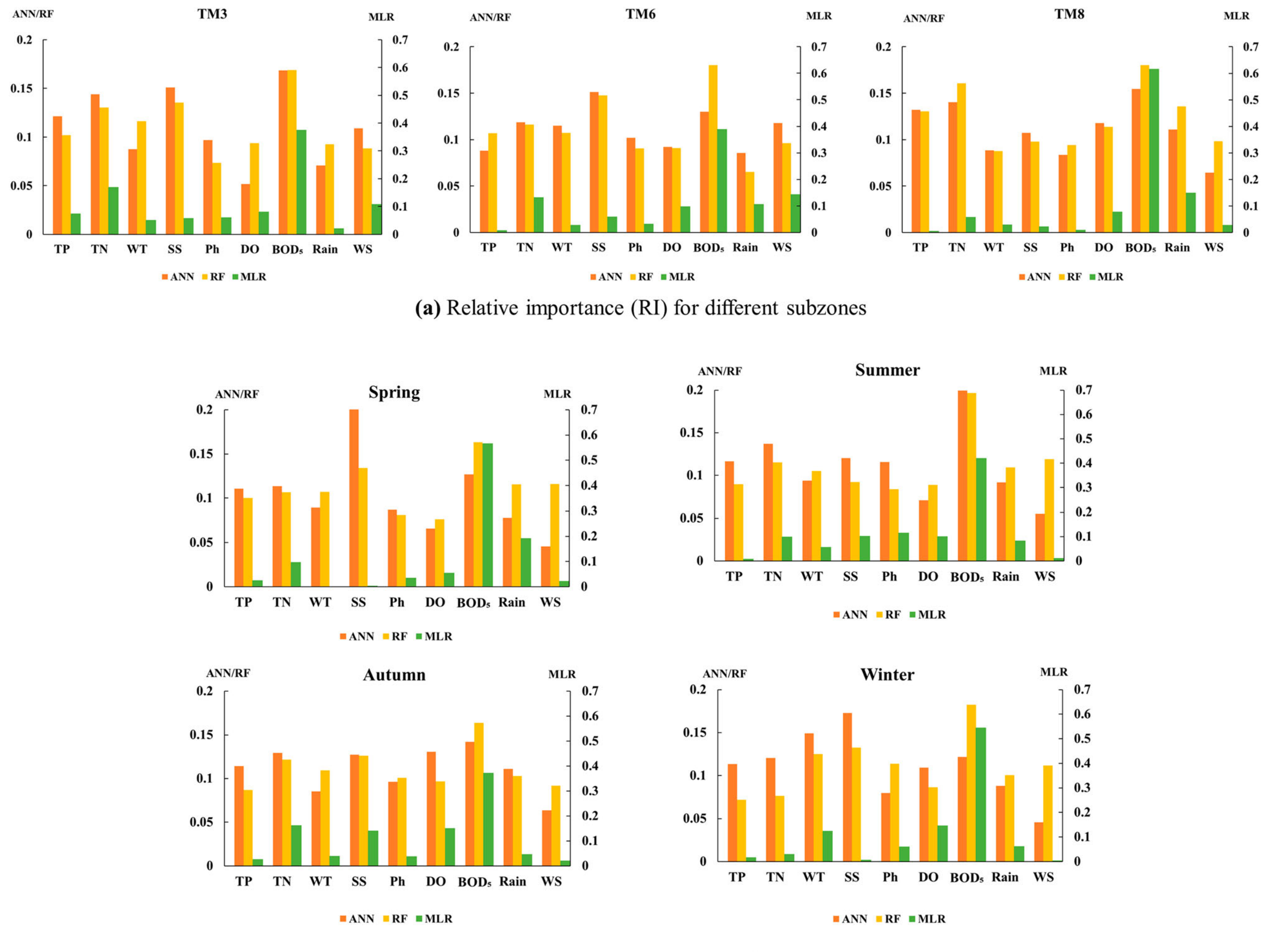


Figure 7. The 20-year seasonal variation of the observed and predicted Chl-a results based on ANN, RF and MLR models in different seasons.



(a) Relative importance (RI) for different subzones

(b) Relative importance (RI) for different seasons

Figure 8. Relative importance (RI) of exploratory variables in Chl-a prediction (a) for different subzones; (b) for different seasons based on ANN, RF and MLR models.

3.4. Temporal characteristics of transparency prediction in different seasons

The time series plots of both measured and predicted transparency results for Tolo Harbour over different seasons are presented in Figure S2. It is noticeable from the results that the secchi disc depth fluctuates much violently with higher standard deviations (Table S1 in the supplementary material) in the winter and spring periods, indicating a larger variation in transparency over these two seasons. In fact, more oligotrophic states are observed in winter while hypereutrophic conditions more likely occurred in summer in this bay area, especially before 2007. After a pronounced rise of transparency from 2007 to 2009, Tolo Harbour hardly shows hypereutrophic states and mostly maintains mesotrophic conditions during summer periods. Thereafter, three severe eutrophic events occurred in the spring of 2013, followed by frequent oligotrophic conditions after 2016 as the improved trophic states. Similarly, the performances of different models are elaborated in Table S3 in the supplementary material. The results indicate that the most significant modeling errors and lowest CC value imply that ANN and RF models have the worst modeling performance in winter among all seasons. Moreover, the low CC values (approximately 0.5) of MLR indicates its invalidity in transparency prediction.

From the relative importance analysis (Figure S3(b)), nutrient contents (TN and TP) and organic pollutants (BOD5) are suggested as the dominating contributors to transparency during the summer and autumn periods. In contrast, the influence of SS on water turbidity dominates over the TN, TP, and BOD**5**, especially in spring and winter. Moreover, the physical factors including WT, WS, and Rainfall have not shown an evident seasonal variation of relative importance on transparency.

3.5. Prediction and analysis of water quality evolution

The EKC model is a commonly used statistical method that builds multinomial curves to estimate the change of environmental variables with PCGDP (Chen et al., 2018; Sarkodie & Strezov, 2018). According to the previous analysis, EKC curves are plotted in Figure 9 for four relatively essential variables (i.e. TN, TP, BOD_5, and SS) at three different stations in Tolo Harbour. These curves show that all four selected variables almost have a U-shaped quadratic relationship with economic growth except for the monotonically linear increases of TN at buffer zone and BOD5 at channel zone.

Once the EKC relationships are built, the future variations of water quality can be estimated using the projection of PCGDP data. *2020 Gross Domestic Product* published by the Census and Statistics Department (CSD) of the Hong Kong government (https://www.censtatd.gov.hk/en/) reviews PCGDP data of Hong Kong over the past two decades. Based on these data, the future growth of Hong Kong PCGDP in the next decade is projected by the ARIMA method under the assumption of current strategies and policies in operation (Figure 10). With a 95% confidence level, the baseline, upper, and lower confidence limit line represent three different economic developing patterns over the following ten years. According to these three patterns, the corresponding water quality evolutions have been predicted under low, basic, and high economic growth (Figure 9). Hereafter, they are input into the well-trained ML models to expect the future variation of Chl-a and SDD.

Figure 11 shows the trends as well as varying ranges of future Chl-a and SDD, which were predicted based on three models in the next decade. Comparing to Chl-a, the extrapolating forecast of SDD performs smaller ranges, indicating a more stable evolution in oncoming years. Nevertheless, Chl-a has wider variation ranges and remarkable oscillated developments, which demonstrated a more complex pattern of blossom mechanism, especially at the buffer zone where has more comprehensive effects of industrial, agricultural, and residential pollutants.

Specifically, the future predictions of growth trend and range by MLR method almost increase linearly with time, simply neglecting possible nonlinear oscillation, which is reasonable to assume that MLR is too simplistic in extrapolation predicting. The ANN model performs more conservatively with a larger potential range of variation at the regions away from the inner coast (i.e. buffer and channel section) due to more sensitive responses on different PCGDP growth cases. This high predictive sensitivity can be attributed to the suspicion of overfitting when extrapolating predictions using networks trained with past data. In contrast, RF performs specifically with a more reasonable variation, providing a reliable extended forecast.

4. Discussion and implications

The modeling performances shows that different models have the different effectiveness in Chl-a and transparency prediction. Specifically, ANN and RF perform better than MLR due to their excellent abilities to handle nonlinear problems (Mamun et al., 2020; Xie et al., 2012). Moreover, ensemble learning model (RF) even provides a more accurate performance than ANN by aggregating sub-models (Shin et al., 2020; Zeng et al., 2017).

The water quality in Tolo Harbour has been observed to show a consistent decrease from less developed channel segment to the densely populated inner region (Deng et al., 2021; Wong et al., 2018; Xu et al., 2010). The ML models have also correctly displayed this variation trend. Specifically, the most severe region with frequent eutrophic events and lower transparency is found to be the inner harbour zone with a direct inflow of coastal wastes. In addition, the study indicates a relatively weak variation in seasonal patterns in the water quality of Tolo Harbour (Xu et al., 2010), except the Chlorophyll-a concentration showed slightly higher in summer and autumn periods than others.

Although nitrogen and phosphorous in the water body are typically considered as the principal factors of Chl-a growth (Deng et al., 2021; Mamun et al., 2020), the higher RI values of BOD_5 indicate that organic pollutant contributes more substantially in the Tolo Harbour. This finding is consistent with previous studies for the Tolo Harbour (Li et al., 2004). Furthermore, a strong relationship between biochemical oxygen demand and eutrophication has also been confirmed by Xu and Xu (2015). This result is reasonably related to sewage diversion plans launched by the Hong Kong government in the late 1990s, which reduced 90% discharge loading and controlled TN and TP within an acceptable range (Xu et al., 2010). After that, the critical limiting factor to algal growth could not be attributed to nutrient availability (Harrison et al., 2010; Nurohman et al., 2021). In addition, BOD_5 and SS are also found as another significant stimulus to algal growth in Tolo Harbour, especially in lower velocity regions (i.e. inner harbour and buffer zone) and high variation seasons (i.e. spring and winter). This finding also concurs with the conclusions of relevant research (Gorokhova et al., 2020; Nurohman et al., 2021).

Water transparency is an integrated indicator related to concentrations of inorganic suspended solids, phytoplankton, and dissolved organic matter in the water, so that could reflect the turbidity, watercolor and abundance of algae (Ao et al., 2018; Mamun et al., 2020). In Tolo Harbour, three variables are found as key factors of water transparency: SS, $BOD_{5,}$ and TN. Wherein, the SS is found to be the leading driven factor near the

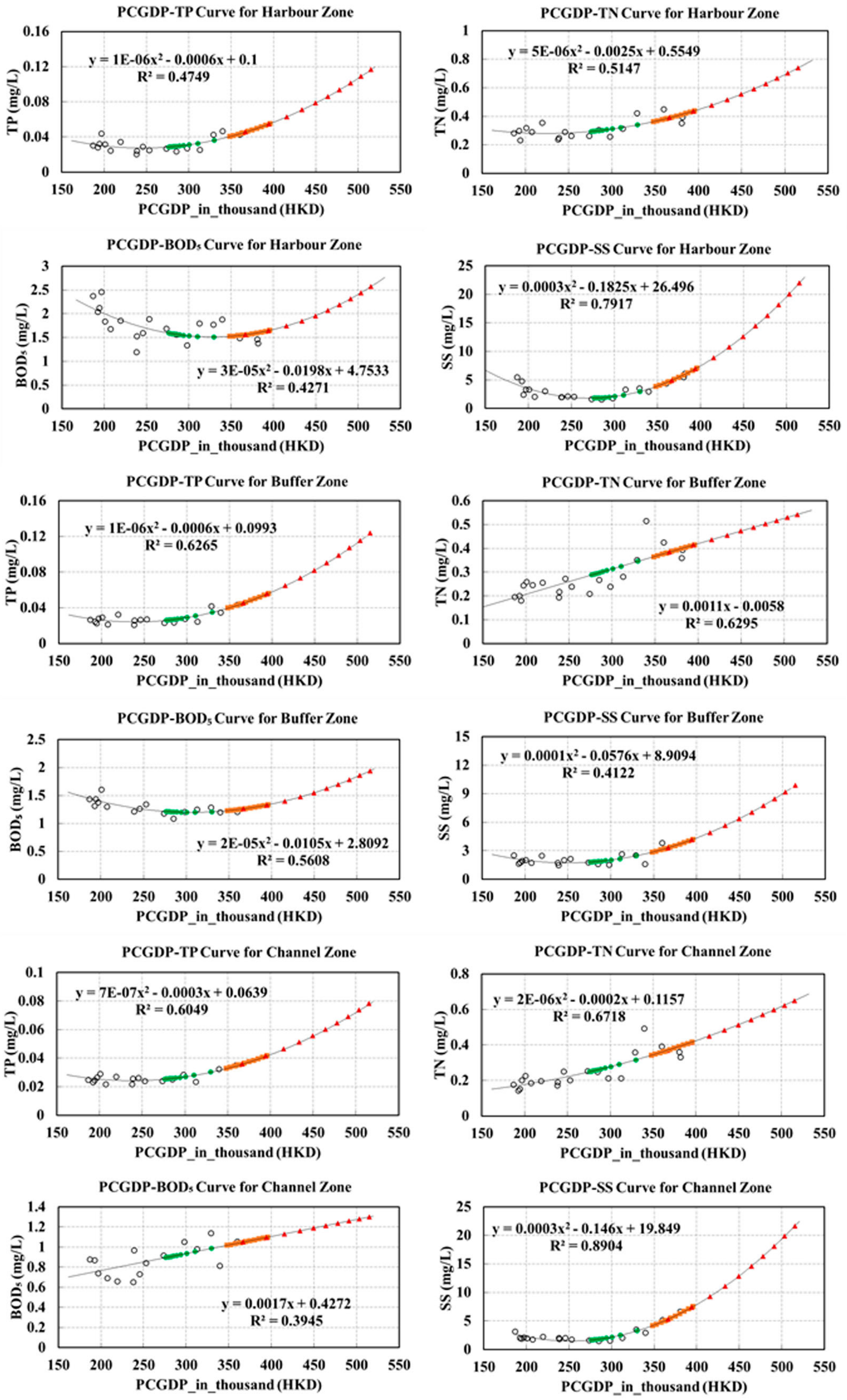


Figure 9. EKC relationships between PCGDP and four variables; represents the observed data, the predicted value under low growth case of PCGDP, and the predictions under basic and high growth cases respectively.

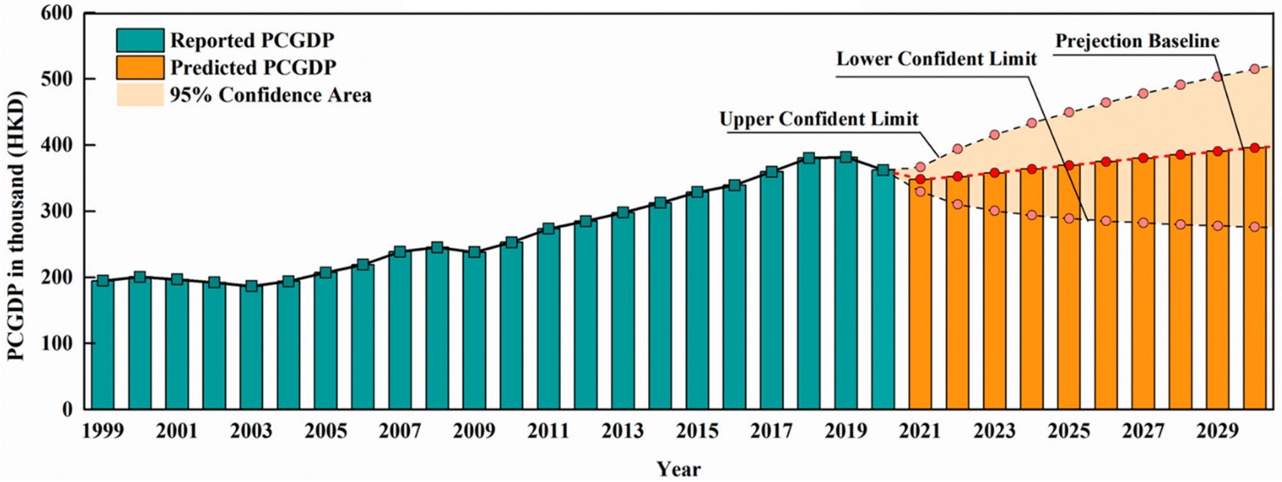


Figure 10. PCGDP Projection of Hong Kong in the next 10 years by ARIMA method with 95% confidence level.

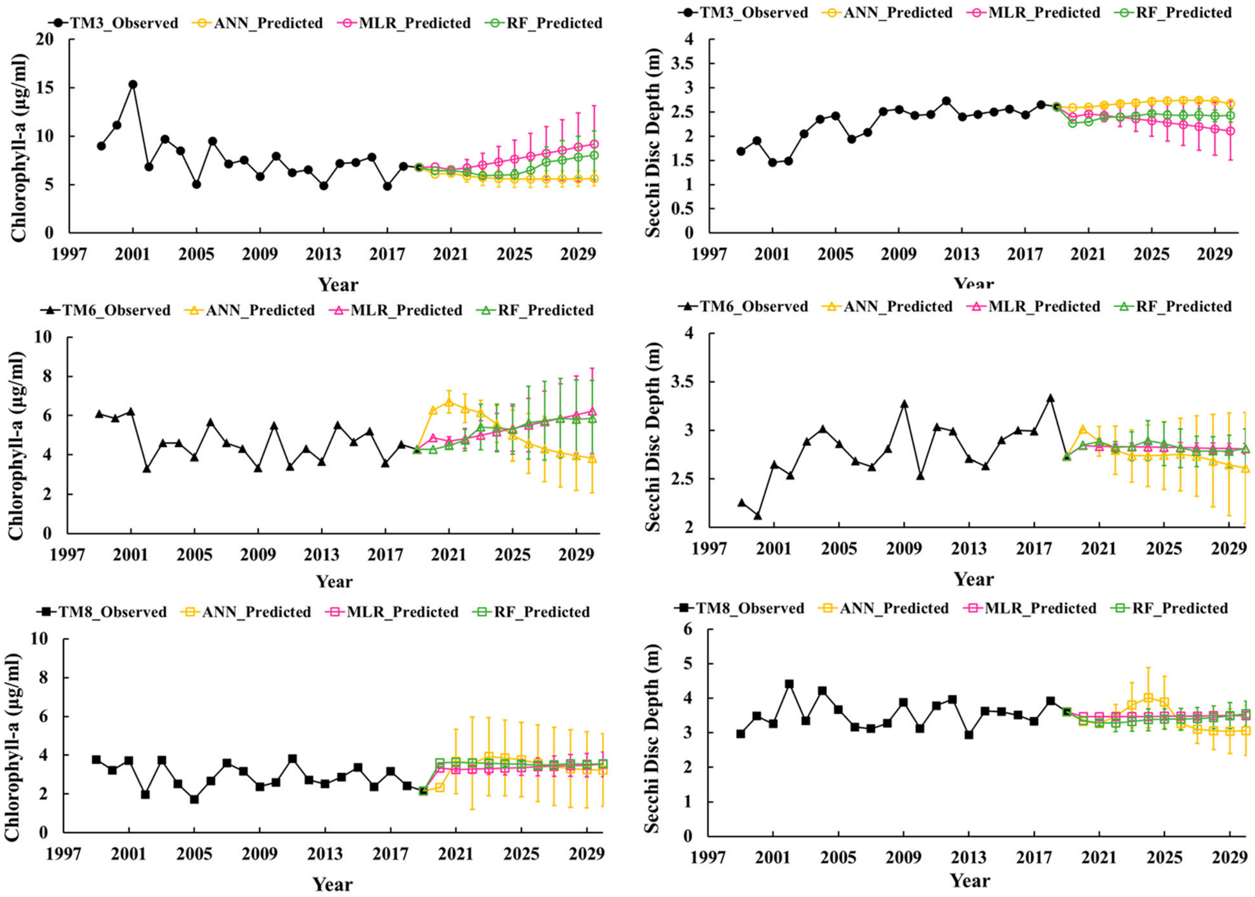


Figure 11. Future prediction of Chl-a and SDD evolutions in the next 10 years, the variation ranges are capped with the maximum and minimum values based on high and low growth PCGDP case.

inner harbour zone. The sudden increase of SS in recent years (Figure 2) should be worthy of the attention, which could be a forewarning of eutrophic issues. This could be attributed to the large-scale constructions due to rapid exploitation along Tolo Harbour in recent decade, including the town expansions of Shatin, Taipo, and Ma On Shan as well as the new developments of Hong Kong Science Park. On the other hand, TN and BOD_5 also have critical contributions to transparency. Nevertheless, the similar impacts of the two on algal growth and water turbidity evidence that these two symptoms of eutrophication highly correlate in the Tolo Harbour. This finding has also been confirmed in recent studies (Hadjisolomou et al., 2021; Mamun et al., 2020). Furthermore, the lower importance of weather conditions, such as WT, WS, and Rainfall, also corroborates that the deterioration should be blamed on anthropogenic activities rather than natural evolution.

From a long-term view of future prediction, the water quality will remain more stable near the channel part (TM8) under current policies and strategies. In contrast, the water deterioration (Chl-a increases and SDD decreases) at TM3 is likely to recur as the economy develops within the next decade but not as severe as 20 years ago. Regarding the buffer region, more attention should be given to regulating and controlling policies owing to the complicated dependency between socioeconomic development and water quality evolution.

In this study, the developed methodology and application results should be the preliminary work of coastal hydro-environment management and is of great significance for water environment protection. Specifically, the more specific spatiotemporal analysis than previous studies characterize different key factors to eutrophication over different subzones and seasons, which are expected to provide more tailored management measures to particular subzones and seasons on coastal eutrophication problems. Furthermore, the well-trained framework incorporating EKC model is able to forecast the future eutrophic trend in the following years, which provides practitioners with a possible way to take the precautionary measures in advance.

5. Summary and conclusion

In conclusion, this study proposes a machine learning framework (using ANN and RF models) to characterize the spatial and temporal variations of eutrophication indicators (i.e. Chl-a concentration and transparency) in a semi-enclosed water bay (Tolo Harbour) in Hong Kong. Furthermore, the developed framework provides a practicable way to forecast future eutrophicate evolutions in the coming years once incorporating the EKC models.

The modeling results indicate that the developed framework had an excellent performance in coastal water quality modeling. The spatiotemporal patterns of key contributors to eutrophication are also rationally interpreted based on this framework. Specifically, the relative importance analysis reveals that the water quality deterioration in the Tolo Harbour is mainly attributed to nutrients, organic pollutants, and suspended solids. Accordingly, their treatment deserves high priority when launching strategies for alleviating eutrophication in this sea bay area in the future. Especially, the sudden increase of nutrients and suspended solids monitored in Tolo Harbour should be worthy of the attention.

According to the future analysis, the eutrophicate condition will remain almost at a constant level in the channel zone. In the inner Harbour zone, although the water deterioration may rebound in oncoming years, the situation will not be as strict as two decades ago. The results also imply that this method still has uncertainties in detailed predicting, especially at the TM6 station, due to more comprehensive anthropogenic impacts along the coasts, which need more inspections and modifications when implementing regulations in the future.

From the application results and analysis of this study, the proposed ML-based framework is proven to be well-suited and effective to model and understand the spatiotemporal evaluation of coastal water quality in a semi-enclosed seawater bay under various heterogeneous conditions. Furthermore, it can provide helpful tools and methods for coastal hydro-environmental forecasting. However, some limitations of this work should also be noted: (1) the proposed ML-based framework is highly data-depended, whose performances directly rely on the reliability of the data used; (2) the ML-models are very site-specific, which means a new training process must be conducted before each application; (3) the projected socioeconomic statistics and extrapolated water quality data based on different assumptions may be very different with reality. To overcome these limitations, this proposed method could be improved in the future studies by higher resolution data and still needs modifications by successive water quality data monitoring.

Acknowledgments

The authors would like to thank the Research Grants Council, University Grants Committee of Hong Kong for the financial supports [Grant Number: 15200719] and Grant [Grant Number: 15201017]. The relevant Departments of Hong Kong Government (EPD, HKO, CSD) are appreciated for the available datasets of water quality, meteorology and economic statistics in the studied region.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors would like to thank the Research Grants Council, University Grants Committee of Hong Kong for the financial supports [Grant Number: 15200719] and Grant [Grant Number: 15201017].

ORCID

Tianan Deng http://orcid.org/0000-0001-7230-0561
Huan-Feng Duan http://orcid.org/0000-0002-9200-904X

Alireza Keramat http://orcid.org/0000-0002-6280-4931

References

Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2012). Multiple linear regression. In David B. Baker (Ed.), *Handbook of psychology* (2nd ed., pp. 511–542). Jone Wiley & Sons, Inc.

Alizadeh, M. J., Kavianpour, M. R., Danesh, M., Adolf, J., Shamshirband, S., & Chau, K. W. (2018). Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 810–823. https://doi.org/10.1080/19942060.2018.1528480

Ao, D., Luo, L., Dzakpasu, M., Chen, R., Xue, T., & Wang, X. C. (2018). Replenishment of landscape water with reclaimed water: Optimization of supply scheme using transparency as an indicator. *Ecological Indicators*, *88*, 503–511. https://doi.org/10.1016/j.ecolind.2018.01.007

Chau, K. W. (2007). Integrated water quality management in Tolo Harbour, Hong Kong: A case study. *Journal of Cleaner Production*, *15*(16), 1568–1572. https://doi.org/10.1016/j.jclepro.2006.07.047

Chen, K., Liu, Y., Huang, D., Ke, H., Chen, H., Zhang, S., Yang, S., & Cai, M. (2018). Anthropogenic activities and coastal environmental quality: A regional quantitative analysis in southeast China with management implications. *Environmental Science and Pollution Research*, *25*(4), 3093–3107. https://doi.org/10.1007/s11356-017-9147-6

Cho, K. H., Sthiannopkao, S., Pachepsky, Y. A., Kim, K. W., & Kim, J. H. (2011). Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. *Water Research*, *45*(17), 5535–5544. https://doi.org/10.1016/j.watres.2011.08.010

Dai, C., Tan, Q., Lu, W. T., Liu, Y., & Guo, H. C. (2016). Identification of optimal water transfer schemes for restoration of a eutrophic lake: An integrated simulation-optimization method. *Ecological Engineering*, *95*, 409–421. https://doi.org/10.1016/j.ecoleng.2016.06.080

Deng, T., Chau, K. W., & Duan, H. F. (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, *284*, 112051. https://doi.org/10.1016/j.jenvman.2021.112051

Diao, X. D., Zeng, S. X., Tam, C. M., & Tam, V. W. (2009). EKC analysis for studying economic growth and environmental quality: A case study in China. *Journal of Cleaner Production*, *17*(5), 541–548. https://doi.org/10.1016/j.jclepro.2008.09.007

Gorokhova, E., Ek, K., & Reichelt, S. (2020). Algal growth at environmentally relevant concentrations of suspended solids: Implications for microplastic hazard assessment. *Frontiers in Environmental Science*, *8*, 223. https://doi.org/10.3389/fenvs.2020.551075

Guo, J., Dong, Y., & Lee, J. H. (2020). A real time data driven algal bloom risk forecast system for mariculture management. *Marine Pollution Bulletin*, *161*, 111731. https://doi.org/10.1016/j.marpolbul.2020.111731

Gupta, N. (2013). Artificial neural network. *Network and Complex Systems*, *3*(1), 24–28. https://iiste.org/Journals/index.php/NCS/article/view/6063.

Hadjisolomou, E., Stefanidis, K., Herodotou, H., Michaelides, M., Papatheodorou, G., & Papastergiadou, E. (2021). Modelling freshwater eutrophication with limited limnological data using artificial neural networks. *Water*, *13*(11), 1590. https://doi.org/10.3390/w13111590

Harrison, P. J., Xu, J., Yin, K., Lee, H. W. J., Anderson, D. M., Liu, H., & Ho, A. (2010). Algal blooms and red tides in Hong Kong: Who, when, where and why. In *Proceedings of 13th international conference on harmful algae* (p. 49).

He, Q., Bertness, M. D., Bruno, J. F., Li, B., Chen, G., Coverdale, T. C., Altieri, A. H., Bai, J., Sun, T., Pennings, S. C., Liu, J., Ehrlich, P. R., & Cui, B. (2014). Economic development and coastal ecosystem change in China. *Scientific Reports*, *4*(1), 1–9. https://doi.org/10.1038/srep05995

Kehoe, M. J., Chun, K. P., & Baulch, H. M. (2015). Who smells? Forecasting taste and odor in a drinking water reservoir. *Environmental Science & Technology*, *49*(18), 10984–10992. https://doi.org/10.1021/acs.est.5b00979

Lee, J. H., Huang, Y., Dickman, M., & Jayawardena, A. W. (2003). Neural network modelling of coastal algal blooms. *Ecological Modelling*, *159*(2-3), 179–201. https://doi.org/10.1016/S0304-3800(02)00281-8

Lee, T. L. (2004). Back-propagation neural network for long-term tidal predictions. *Ocean Engineering*, *31*(2), 225–238. https://doi.org/10.1016/S0029-8018(03)00115-X

Li, X., Yu, J., Jia, Z., & Song, J. (2014). Harmful algal blooms prediction with machine learning models in Tolo Harbour. In *2014 International conference on smart computing* (pp. 245–250). IEEE.

Li, Y. S., Chen, X., Wai, O. W., & King, B. (2004). Study on the dynamics of algal bloom and its influence factors in Tolo Harbour, Hong Kong. *Water Environment Research*, *76*(7), 2643–2654. https://doi.org/10.1002/j.1554-7531.2004.tb00226.x

Mamun, M., Kim, J. J., Alam, M. A., & An, K. G. (2020). Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. *Water*, *12*(1), 30. https://doi.org/10.3390/w12010030

Muttil, N., & Chau, K. W. (2007). Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of AI*, *20*(6), 735–744. https://doi.org/10.1016/j.engappai.2006.11.016

North American Lake Management Society (NALMS). (1990). The lake and reservoir restoration guidance manual.

Nurohman, H., Lee, T. G., & Kwon, H. J. (2021, June). Water quality investigation of the main river in Daegu, South Korea. In *IOP conference series: Earth and environmental science* (Vol. 789, No. 1, p. 012042). IOP Publishing.

Peng, S., Qin, X., Shi, H., Zhou, R., Dai, M., & Ding, D. (2012). Distribution and controlling factors of phytoplankton assemblages in a semi-enclosed bay during spring and summer. *Marine Pollution Bulletin*, *64*(5), 941–948. https://doi.org/10.1016/j.marpolbul.2012.03.004

Qiao, Y., Feng, J., Cui, S., & Zhu, L. (2017). Long-term changes in nutrients, chlorophyll-a and their relationships in a semi-enclosed eutrophic ecosystem, Bohai Bay, China. *Marine Pollution Bulletin*, *117*(1-2), 222–228. https://doi.org/10.1016/j.marpolbul.2017.02.002

Sarkodie, S. A., & Strezov, V. (2018). Empirical study of the environmental Kuznets curve and environmental sustainability curve hypothesis for Australia, China, Ghana and USA. *Journal of Cleaner Production*, *201*, 98–110. https://doi.org/10.1016/j.jclepro.2018.08.039

Sattari, M. T., Falsafian, K., Irvem, A., & Qasem, S. N. (2020). Potential of kernel and tree-based machine-learning models for estimating missing data of rainfall. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 1078–1094. https://doi.org/10.1080/19942060.2020.1803971

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., Lee, C., Kim, T., Park, M. S., Park, J., & Heo, T. Y. (2020). Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods. *Water*, *12*(6), 1822. https://doi.org/10.3390/w12061822

Sin, Y. S., & Chau, K. W. (1992). Eutrophication studies on Tolo Harbour, Hong Kong. *Water Science and Technology*, *26*(9-11), 2551–2554. https://doi.org/10.2166/wst.1992.0785

Whittington, J. C., & Bogacz, R. (2019). Theories of error backpropagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

Wong, K. T., Chui, A. P. Y., Lam, E. K. Y., & Ang Jr, P. (2018). A 30-year monitoring of changes in coral community structure following anthropogenic disturbances in Tolo Harbour and Channel, Hong Kong. *Marine Pollution Bulletin*, *133*, 900–910. https://doi.org/10.1016/j.marpolbul.2018.06.049

Xie, Z., Lou, I., Ung, W. K., & Mok, K. M. (2012). Freshwater algal bloom prediction by support vector machine in macau storage reservoirs. *Mathematical Problems in Engineering*, *2012*(1). https://doi.org/10.1155/2012/397473

Xu, F. L., Lam, K. C., Zhao, Z. Y., Zhan, W., Chen, Y. D., & Tao, S. (2004). Marine coastal ecosystem health assessment: A case study of the Tolo Harbour, Hong Kong, China. *Ecological Modelling*, *173*(4), 355–370. https://doi.org/10.1016/j.ecolmodel.2003.07.010

Xu, J., Yin, K., Liu, H., Lee, J. H., Anderson, D. M., Ho, A. Y. T., & Harrison, P. J. (2010). A comparison of eutrophication impacts in two harbours in Hong Kong with different hydrodynamics. *Journal of Marine Systems*, *83*(3-4), 276–286. https://doi.org/10.1016/j.jmarsys.2010.04.002

Xu, Z., & Xu, Y. J. (2015). Rapid field estimation of biochemical oxygen demand in a subtropical eutrophic urban lake with chlorophyll a fluorescence. *Environmental Monitoring and Assessment*, *187*(1), 1–14. https://doi.org/10.1007/s10661-014-4171-1

Yu, P., Gao, R., Zhang, D., & Liu, Z. P. (2021). Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, *123*, 107334. https://doi.org/10.1016/j.ecolind.2020.107334

Zeng, Q., Liu, Y., Zhao, H., Sun, M., & Li, X. (2017). Comparison of models for predicting the changes in phytoplankton community composition in the receiving water system of an inter-basin water transfer project. *Environmental Pollution*, *223*, 676–684. https://doi.org/10.1016/j.envpol.2017.02.001

Zhang, C. X., & You, X. Y. (2017). Application of EFDC model to grading the eutrophic state of reservoir: Case study in Tianjin Erwangzhuang Reservoir, China. *Engineering Applications of Computational Fluid Mechanics*, *11*(1), 111–126. https://doi.org/10.1080/19942060.2016.1249411