https://doi.org/10.1016/j.jenvman.2021.112051

Machine Learning Based Marine Water Quality Prediction for Coastal Hydro-environment Management

Tianan Deng, Kwok-Wing Chau, Huan-Feng Duan*

Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR 999077

hf.duan@polyu.edu.hk

Abstract

During the past three decades, harmful algal blooms (HAB) events have been frequently observed in marine waters around many coastal cities in the world including Hong Kong. The increasing occurrence of HAB has caused acute influences and damages on water environment and marine aquaculture with millions of monetary losses. For example, the Tolo Harbour is one of the most affected areas in Hong Kong, where more than 30% HAB occurred. In order to forewarn the potential HAB incidents, the machine learning (ML) methods have been increasingly resorted in modelling and forecasting water quality issues. In this study, two different ML methods – artificial neural networks (ANN) and support vector machine (SVM) – are implemented and improved by introducing different hybrid learning algorithms for the simulations and comparative analysis of more than 30-year measured data, so as to accurately forecast algal growth and eutrophication in Tolo Harbour in Hong Kong. The application results show the good applicability and accuracy of these two ML methods for the predictions of both trend and magnitude of the algal growth. Specifically, the results reveal that ANN is preferable to achieve satisfactory results with quick response, while the SVM is suitable to accurately identify the optimal model but taking longer training time. Moreover, it is demonstrated that

the used ML methods could ensure robustness to learn complicated relationship between algal dynamics and different coastal environmental variables and thereby to identify significant variables accurately. The results analysis and discussion of this study also indicate the potentials and advantages of the applied ML models to provide useful information and implications for understanding the mechanism and process of HAB outbreak and evolution that is helpful to improving the water quality prediction for coastal hydro-environment management.

***Keywords**:* water quality; harmful algal blooms (HAB); machine learning (ML); coastal hydro-environment; marine environment

1. Introduction

With the increasing population growth and intensive agricultural and industrial activities since the last century, the eutrophic wastewaters discharged into coastal water bodies have greatly deteriorated the water quality as being a worldwide crisis on marine environment (Gill et al. 2018). Globally 415 regions were reported to have different forms of eutrophic symptoms according to an investigation conducted in 2008 (Selman et al. 2008). For example, the longest-lasting algal blooming (18 months) in the Eastern Florida Bay in 2005 (Glibert et al. 2009) and the largest water blooming from central California to Alaska in 2015 (McCabe et al. 2016; Michalak 2016). Meanwhile, the HAB have also been a major problem within the marginal sea between Asia continent and Pacific Ocean since the beginning of last century (Kim 1998; Li et al. 2004; Richlen et al. 2010; Al-Azri et al. 2014; Park et al. 2015). In particular, the annually recurrent HAB events last from early May to late June every year may affect up to 10,000 km^2 water area of the East China Sea (Yu et al. 2018).

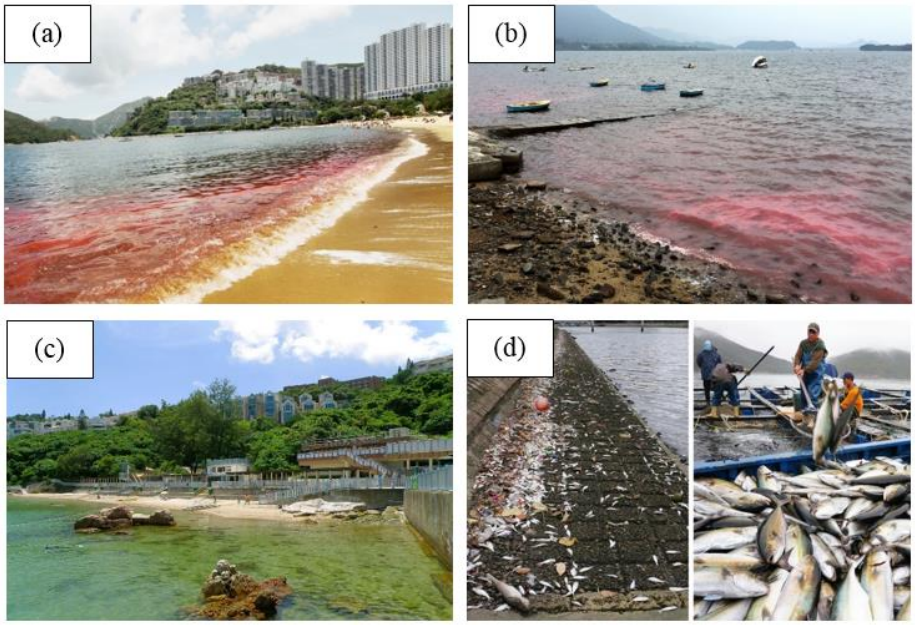


Figure 1: Typical HAB incidents in Hong Kong: (a) & (b) Water discoloration by HAB; (c) Recreational beach closed; (d) Fish kills (*Sources: newspapers and government websites*)

In Hong Kong, water quality degradation issues have been considered as one of the most serious threats on the coastal water ecosystem since 1980s, as typical examples shown in Figure 1. Hong Kong is a typical coastal city with the sea on its three sides where the marine water ecology may have significant impacts on the residential and environmental as well as economic development in that city. During the past decades, harmful algal blooms (HAB) events have frequently occurred in waters around Hong Kong. For example, in April 1998, the worst fish kills event in Hong Kong's history was attributed to the devastating algal growth with more than 3,000 tons fish death and over $ 40 million USD direct economic losses, which caused acute damages to both water ecology and aquaculture (Lee et al. 2003; Lu and Hodgkiss 2004; Muttil and Chau 2006; Selman et al. 2008).

In order to mitigate these potential damages and to improve the water quality condition, it is imperative to develop a usable model that can effectively predict the growth and evolution

process of the algal (including HAB), so as to allow the authority/administrator issue the early alert. Since 1980s, extensive process-based studies on predicting algal blooms have been carried out (Lu and Hodgkiss 2004; Lee et al. 2005; Yang et al. 2008; Xu et al. 2010; Yang et, al. 2019), in order to capture a deterministic relationship between growth dynamics of algal population and external environment variables. However, modelling dynamics of algal growth and evolution in a coastal water ecosystem remains challenging because the physical, chemical and biological processes involved are extremely complicated and more importantly, so that current theories and practice have not yet been well established by far (Xie et al. 2012; Yang et, al. 2019; de Oliveira et al. 2020).

Machine learning (ML) models can be important and useful complements and alternatives in HAB modelling and water quality prediction (Chau 2006). In principle, the ML models focus mainly on the relationship mapping between inputs and outputs of a system rather than complex process mechanisms. By learning from a large mass of historical data which has included the dynamic evolution process (e.g., coastal water and HAB growth), the highly nonlinear relationships can be accurately approximated with or without prior knowledge for the studied system. In this regard, there are different ML techniques have been successfully developed for algal prediction, including artificial neural networks (ANN) (Recknagel et al. 1997; Lee et al. 2003; Muttil and Chau 2007; Sivapragasam et al. 2010; Chang et al. 2017; Tian et al. 2017), genetic programming (GP) (Muttil and Chau 2006; Sivapragasam et al. 2010; Daghighi 2017), support vector machine (SVM) (Liu et al. 2009; Xie et al. 2012; Dai et, al. 2016; Mamun et al. 2020) and Random Forest (RF) (Segura et al. 2017; Zeng et al. 2017).

Amongst those ML techniques, ANN with error back-propagation (BP) algorithm is one of the widely used paradigms in water and environment field due to the rapid response and satisfactory modelling accuracy. However, one main defect of this gradient descent is attributed to the randomness of the initialization of parameters, which usually makes the model converge at a relatively slow speed or even trapped into a local optimum. In order to overcome such drawback, relevant optimization algorithms have been proposed and implemented in the ANN method in the literature, such as gradient descent method (GDM) (Rumelhart et al. 1985; Qian 1999; Lee et al. 2003; Muttil and Chau 2006), Levenberg-Marquardt algorithm (LM) (Levenberg 1944; Hagan and Menhaj 1994; Lourakis 2005; Gavin 2019), Genetic algorithm (GA) (Recknagel et al. 2002; Chau 2006; Ding et al. 2011; Mulia et al. 2013) and Particle Swarm Optimization (PSO) scheme (Kennedy and Eberhart 1995; Chau 2005a; Qi et al. 2018).

The SVM is another effective ML technique for non-linear classification and regression. Differently from the ANN, the SVM adopts the concept of structural risk minimization in which the learning strategy is aimed to minimize the regularized loss function. With the SVM, the generalization ability can be enhanced and the probability of overfitting can be reduced. The main tenet of SVM is to implicitly map a nonlinear problem from the original feature space into a higher or infinite dimensional space via the use of kernel functions where the original problem can be linearly described. From this perspective, the SVM is a promising forecasting paradigm that has been widely employed in many freshwater ecosystems.

Despite that many studies have been focused on the ML methods in different fields, there are so far very few researches on implementing and applying these ML methods (e.g., ANN

and SVM) for effective algal modelling and water quality prediction in marine systems (Li et al. 2014; Park et al. 2015). In this connection, this paper presents a further study on the coastal water quality prediction by using these two different ML methods (ANN and SVM), in order to establish a dynamic evolution relationship between the water quality consequence and various coastal system conditions and environmental factors. The marine water system of Tolo Harbour in Hong Kong is taken as example for the illustration and application of the developed method framework. Through the case study, the performances of these two different ML methods (ANN and SVM) are compared and discussed for coastal water quality prediction in terms of accuracy and efficiency. Furthermore, based on the developed models and obtained prediction relationships, the water quality results are analyzed and discussed for the influence and significance of different factors in the studied coastal system.

2. Study Area and Total Environment Conditions

Hong Kong is one of the worst regions suffered from HAB in the world (Lu and Hodgkiss 2004). Since records began in 1975, a total of 956 HAB incidents have been reported by 2019. Of these, 34.6% HAB events of Hong Kong occurred at Tolo Harbour and it is deemed as the most affected area in Hong Kong (AFCD, 2019). In this study, we select the field-measured water quality data over 30 years in Tolo Harbour for training both the ANN and SVM models.

2.1. Geographical Pattern of Tolo Harbour

Tolo Harbour, located between 22°24'N, 114°11'E and 22°31'N, 114°20'E, is an almost landlocked harbour of New Territories district, situated in the north-east of Hong Kong, connecting with open sea through the sole outlet Tolo channel (see Figure 2). The whole surface

area of Tolo Harbour was measured as 50 km^2 and average depth was about 12 m. The length from the inner harbour area to the only narrow exit to Mirs Bay as long as 16 km, which lead to a long water retention time. In addition, mixed semidiurnal tides of small height varied from 0.8 m to 2 m flushes this area with relatively slow velocity (Lee et al. 2003). Owing to its hydrological pattern, water movements of inner harbour zone influenced by tides are even more limited and the water column is often stratified. Thus, the water circulation of this area almost remains static or moves with a very slow pace, which impede the export of pollutants from inner zone and weaken the limited self-purification ability of Tolo Harbour. Due to the weak water circulation there, the harbingers of eutrophication were observed even before the commencements of modern exploitation (Chau 2007).

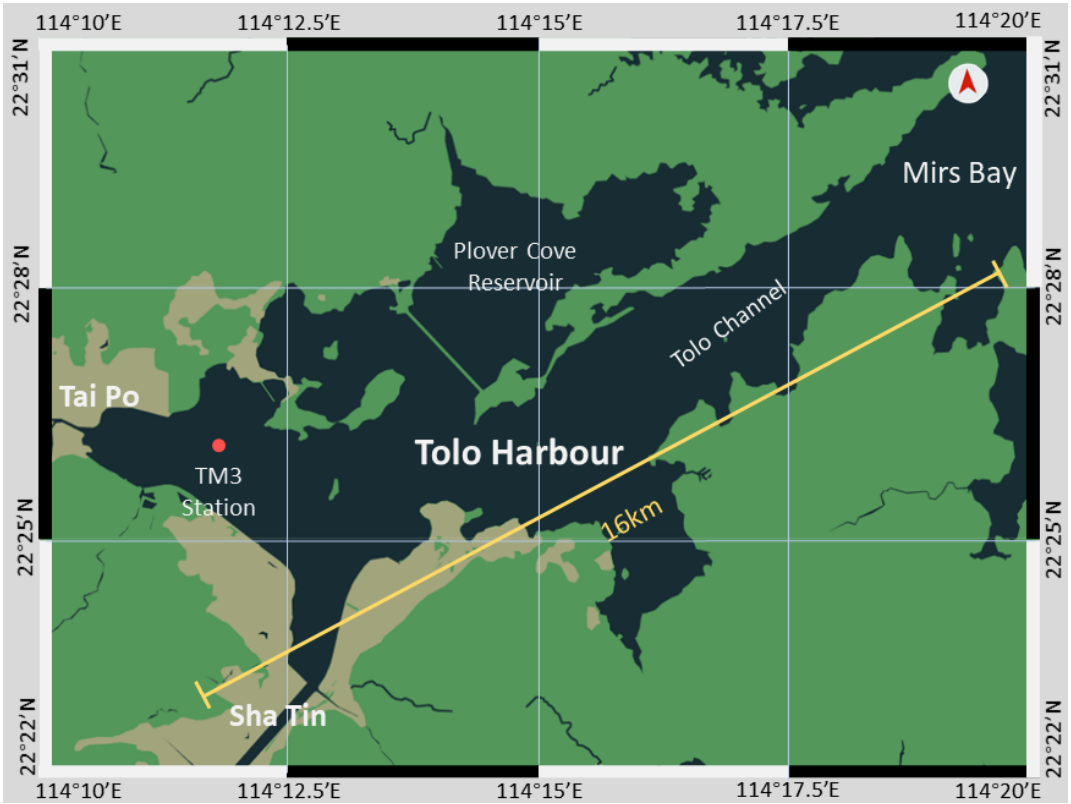


Figure 2: The map of Tolo Habour and sampling station location

2.2. Hydrosphere and Anthroposphere

Since 1970s, the heavy exploitations of Tolo Harbour started with the constructions of the Plover Cove reservoir. The new built reservoir cut off streams that directly flowed into Tolo Harbour before, resulting a significant reduction of freshwater runoff and great decrease of watershed area of Tolo Harbour (Xu et al. 2004a). Meanwhile, two new waterfront towns of Tai Po and Shatin were urbanized and industrialized. These excessive exploitations increased burden on the ecosystem of Tolo Harbour. In addition, a nearly doubled amount of population (from 0.5 million in 1986 to 0.9 million in 2001) in this area also caused a rise of wastewater discharge produced from both municipal and industrial activities (Xu et al. 2004b). Abundant nutrient elements such as nitrogen and phosphorus contained in sewage, especially nitrogen, phosphorus and other substances, were discharged into harbour zone, resulting in serious nutrient enrichment of Tolo Harbour. The waters in Tolo Harbour was thereby heavily eutrophicated and the deterioration was aggravated with years. In turn, the excessive pollutant load induced undesirable damages on the productive activities such as aquacultural fish deaths and harbour closures due to rapid phytoplankton accumulations. As water eutrophication led to successive HAB incidents, the aquatic ecology as well as aquaculture industry in Tolo Harbour hence suffered serious damages.

In order to control the pollution, the Hong Kong government scheduled Tolo Harbour as the first set of Water Control Zone (WCZ) in Hong Kong in 1982. Hereafter, two schemes namely the Tolo Harbour Action Plan (THAP) and Tolo Harbour Effluent Export Scheme (THEES) were also implemented in 1987 and 1995 respectively. After continuous efforts over

decades, the water environment in Tolo Harbour has been noticeably improved. At present, Tolo Harbour WCZ still maintains biweekly/monthly regular measurement of water quality, providing abundant historical data for water quality modelling researches.

3. Machine Learning Methods

In general, the procedure of machine learning modelling for prediction is composed of several key steps as follows. Firstly, the available data set will be split into training set, validation set and testing set respectively. After initial data preprocessing, a specific ML model is then selected, which will be trained and validated based on training set and validation set. Before to be tested with untrained data, the related hyper parameters will be tuned repeatedly until the preset training goal (precision) is met. Eventually, the testing set will be used to test the trained model and to evaluate the performance. For clarity, a flow chart of the ML modelling and application procedure is given in **Figure 3**.

In this study, to enhance the effectiveness of water quality (e.g., HAB) prediction for the studied case, two commonly used ML methods are improved, implemented and applied for this investigation, which are elaborated as follows. To be specific, all the algorithms and models involved in this study were implemented by in-house coding on the platform of MATLAB 2018a.

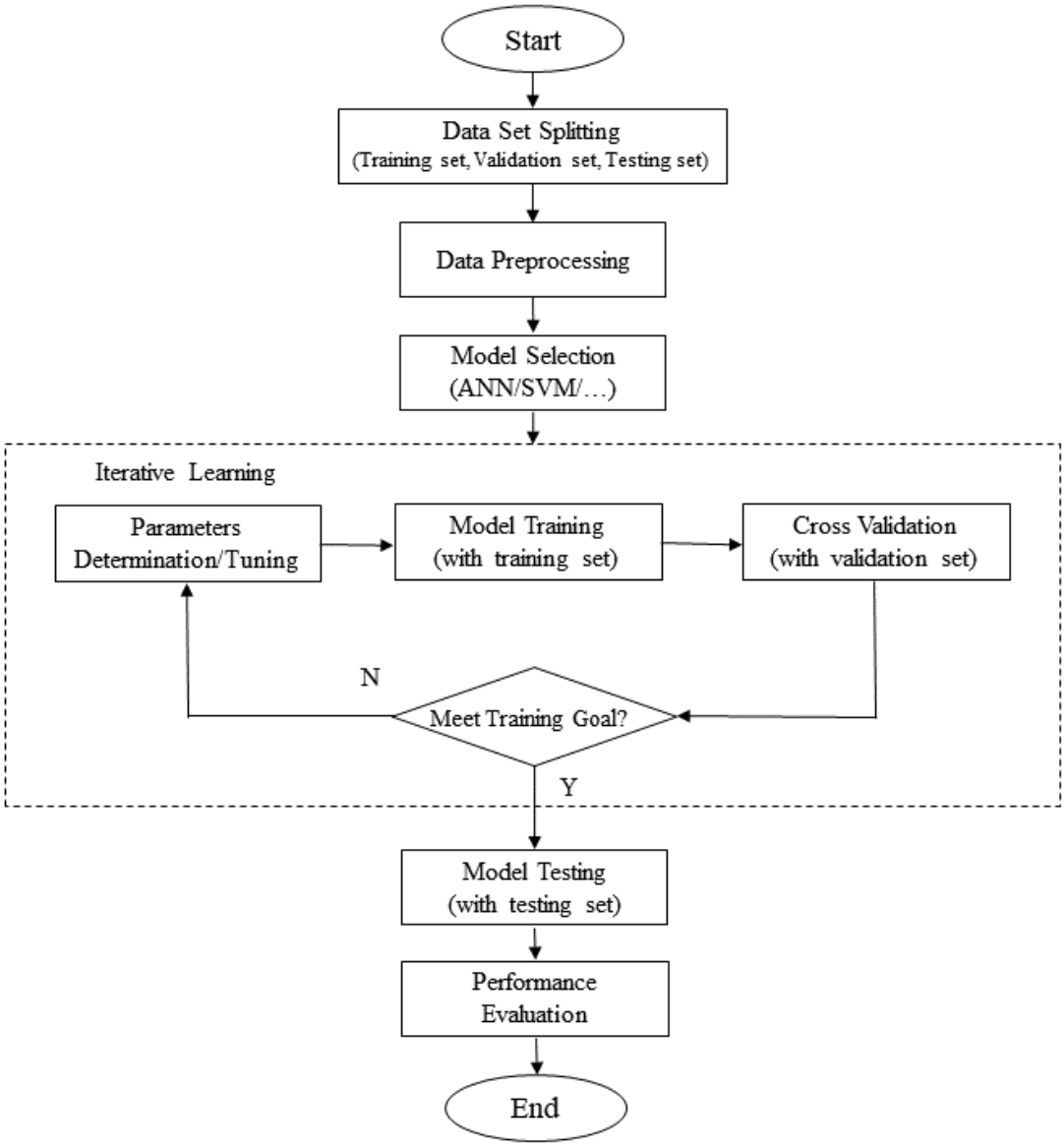


Figure 3 The flow chart of general procedure for machine learning modelling

3.1. ANN Framework and Improvement

ANN is a self-adaptive computational model that efficiently works on finding a mapping between input information and the desired output response. The principle of the classic ANN is shown in Figure 4. Due to the immediate modelling response, good fault tolerance and universality, ANN has been widely applied in non-linear simulations. Notwithstanding, there are several types of network structure in ANN, as mentioned, the BP network is the most used

ANN model so far and has successfully solved many problems in a number of fields. Typically, BP network is usually divided into three parts including the input layer, the output layer and one or more hidden layers. Each layer comprises numerous computing neurons which are highly interconnected with every neuron in the following layer and each neuron is a computing unit that conducts a nonlinear transfer operation following a linear summation.

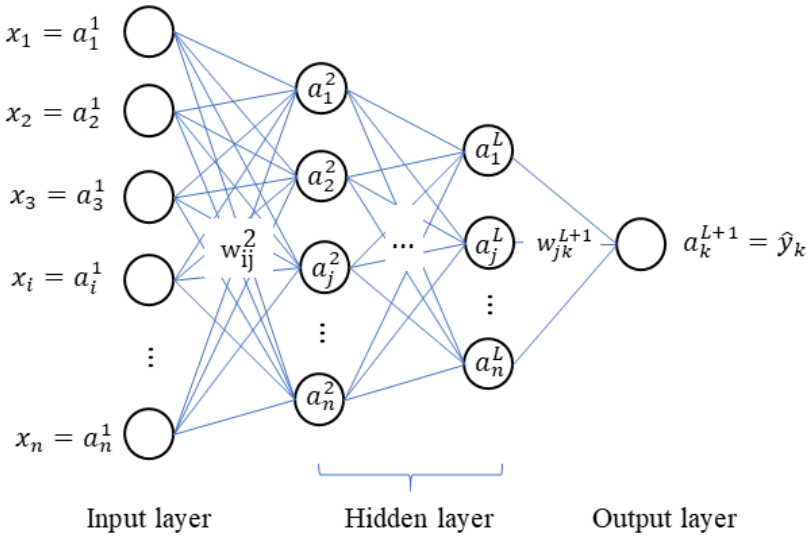


Figure 4: The framework and principle of classic ANN

The feedforward computational process can be described as Eq. (1):

$$a_j^{H+1} = f^{H+1}\left(b_j^{H+1} + \sum_{i=1}^{n} w_{ji}^{H+1} a_i^H\right) \quad (1)$$

$$\text{for } 1 \le H \le L,\ a_i^1 = x_I,\ a_k^{L+1} = \hat{y}_k$$

where the superscript H represents the number of layer, w_{ji} denotes the connecting weights between i^{th} neuron of H^{th} layer and j^{th} neuron of $(H+1)^{th}$ layer, which is usually expressed as a matrix form, a_i^H is the input of H^{th} layer and also the output of $(H+1)^{th}$, x_I and $\hat{y}_k$ respectively denote the initial input and predicted output of the entire network, n, b and f are the dimension of input vectors, the bias term and the nonlinear transfer function

that can be either sigmoid function, hyperbolic tangent function or rectified linear unit (ReLU) function.

After feedforward prediction, the mean squared error E_p (also called empirical error) is usually calculated to evaluate the performance of the network, which can be written as Eq. (2):

$$E_p = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}_k^i - y_o^i\right)^2 \quad (2)$$

where y_o is the desired output response and m is the size of training batch.

In practice, the modelling performance of ANN is largely dependent on both the learning algorithm and the initial weights (Sutskever et al. 2013). Gradient descent Eq. (3) is the most used algorithm that updates the randomly initialized weights incrementally towards the direction that E_p descents until certain conditions are met (e.g. the maximum iteration epoch, target accuracy or no significant changes between two iterations, etc.). Once training is completed, the weight matrix is fixed and the model can be used to predict untrained data.

$$w_{ji}(t+1) = w_{ji}(t) - \alpha\frac{\partial E_p}{\partial w_{ji}(t)} \quad (3)$$

where t denotes the epoch number, α is the learning rate to be preset.

However, the basic gradient descent algorithm is usually not fully ideal in nonlinear problem solving since it converges at a low speed, produces unstable results and is easily trapped into the local optimum etc. Aimed at these deficiencies, a number of optimized learning algorithms are proposed. The following optimized algorithms are implemented in the ANN framework so as to enhance the prediction effectiveness in this study:

(1) Gradient Descent with Momentum (GDM): GDM introduces the concept of inertia (a momentum term) in weight update process which considers both current gradient and

gradient change of previous steps (Qian 1999). It is the simplest way to avoid oscillations especially near local minimums and speed up convergence rate.

(2) Levenberg-Marquardt algorithm (LM): It combines Gradient Descent with Gauss-Newton algorithm by introducing a damping term (Gavin 2019). LM remarkably accelerates the convergence speed since it considers both first-order derivatives (gradient) and second derivatives (Hessian matrix).

(3) Genetic Algorithm (GA): Unlike gradient-based optimizations above, GA is a population-based optimization that determines the optimal weight matrix (solution) by promoting explorations. (Ghaffari et al. 2006). GA searches the global optimal solution by a group of potential solutions and their offspring. The evolutionary manipulations such as reproduction, selection, crossover and mutation will iterate repeatedly until the optimal one is found (Recknagel et al. 2002; Chau 2006). Actually, GA can be described as a global optimization algorithm that does not dependent on the initial values and gradient information (Mirzazadeh et al. 2008).

(4) Particle Swarm Optimization (PSO): PSO is another promising populated evolutionary algorithm that mimics the social behaviors of gregarious animals to search the optimal solution by cooperation and competition (Chau 2005a). In PSO, each potential solution flies to the current optimum based on both the swarm best position and individual best position (Chau 2005b). Since the complex evolutionary operators such as crossover and mutation in GA are not involved, the computational cost of PSO is much inexpensive and still can accomplished satisfactory results in

many cases.

To sum up, gradient-based optimizations (i.e. GDM and LM) are good at local convergence but they are prone to find a local optimum while population-based optimizations (i.e. GA and PSO) are robust to search for best region in the whole solution space but are inefficient in fine-tuning local search especially within the near-optima region (Ghaffari et al. 2006). Some scholars have proposed a hybrid ANN models integrating different population-based and gradient-based algorithms to make full use of advantages on them and obtained better performances than using either one exclusively (Chau 2005b). These integrated models are adopted in Section 4 with four candidate models developed.

3.2. SVM Framework and Implementation

SVM is another promising machine learning algorithm (as depicted in Figure 5), which has been successfully applied to classification as well as regression problems. The preliminary goal of SVM is to determine the optimal nonlinear relation $f(x)$ between input and output by mapping feature vectors from original space to a high dimensional space where the relation can be linearly described. Assuming the training data set is Eq. (4):

$$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m) \quad (4)$$

The input vectors mapped to high dimensional space are described as Eq. (5):

$$\Phi(x) = \left(\Phi(x_1), \Phi(x_2), \dots, \Phi(x_m)\right) \quad (5)$$

Therefore, the high dimensional target function can be written as Eq. (6):

$$f(x) = \omega^T \cdot \Phi(x) + b \quad (6)$$

where $\Phi(x)$ denotes the high dimensional mapping on input vector from the original space x.

ω and b are parameters to be estimated by learning. Y_m is the label of m^{th} input vector.

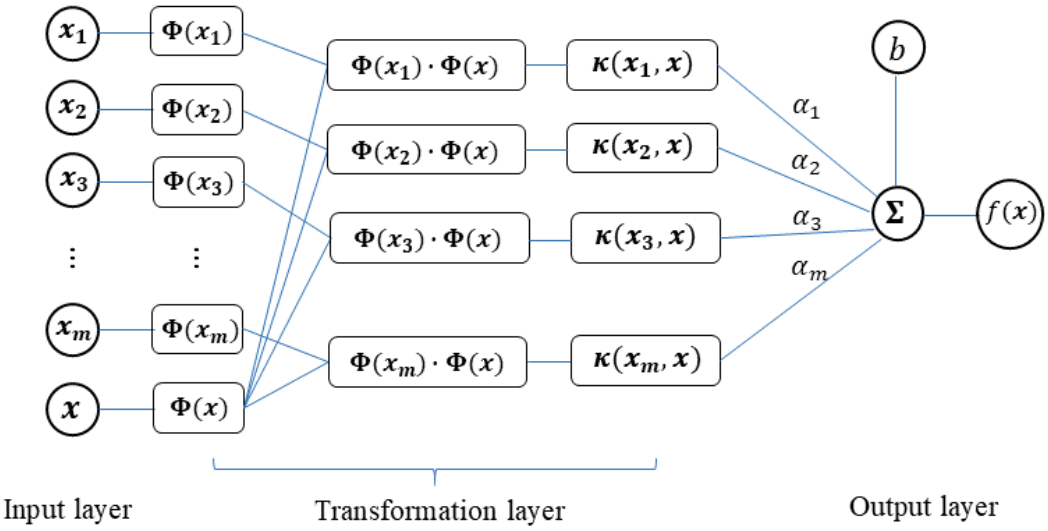


Figure 5: The framework and principle of the SVM algorithm

Based on the principle of structural risk minimization, the learning strategy of SVM is to minimize the upper limit of structural error rather than empirical error adopted by other machine learning techniques like ANN. Mathematically, the objective function of SVM for regression (SVR) can be expressed as Eq. (7):

$$\min_{\omega,b,\xi_I,\hat{\xi}_i} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)$$
$$\text{s.t.}\begin{cases} f(x_i) - y_i \le \epsilon + \xi_i \\ y_i - f(x_i) \le \epsilon + \hat{\xi}_i \\ \xi_i \ge 0, \hat{\xi}_i \ge 0, i = 1,2,\dots,m \end{cases} \quad (7)$$

where $\|\omega\|$ is a regularized penalty term which assures the flatness of function; In SVM model, an ϵ-insensitive zone is introduced which ignores the errors less than ϵ; ξ_I and $\hat{\xi}_I$ are nonnegative slackness variables which measure the deviation between actual value and the boundary of the insensitive zone; C is a constant trade-off the empirical error and the flatness.

Obviously, the original problem of SVR Eq. (7) is a convex quadratic optimization

problem that assures the solution unique and global optimal (Hsu et al. 2003; Xie et al. 2012; Lou et al. 2017). By introducing the Lagrange multipliers, Eq. (7) can be equivalently transformed as the its dual expression Eq. (8):

$$\max_{\alpha,\hat{\alpha}} \sum_{i=1}^{m} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)\Phi(x_i)\Phi(x_j) \quad s.t\begin{cases}\sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i) = 0 \\ 0 \le \alpha_i, \hat{\alpha}_i \le C\end{cases} \tag{8}$$

where $\alpha_I, \hat{\alpha}_I$ are Lagrange multipliers; $\Phi(x_i)^T\Phi(x_j)$ involves the inner product of high dimensional vectors which may lead to geometrically increase of computational complexity. In order to simplify the computational complexity, the kernel tricks are usually adopted. The kernel tricks $\Phi(x_i)^T\Phi(x_j) = \kappa(x_I, x_j)$ are normally used to represent the inner product of two high-dimensional vectors by inner product of two low-dimensional vectors. Finally, the high dimensional decision function with kernel functions can be expressed as Eq. (9):

$$f(x) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)\,\kappa(x_I, x) + b \tag{9}$$

By using kernel functions, all calculations in SVM can be implemented in the original input domain without complex high dimensional computations. The commonly used kernel functions are as Eq. (10):

$$\kappa(x_I, x_j) = \begin{cases}(x_i^T \cdot x_j + 1)^d & \text{Polynomial} \\ \exp\left(-\gamma\|x_i - x_j\|^2\right) & \text{Gaussian} \\ \tanh(\gamma x_i^T \cdot x_j + r) & \text{Sigmoidal}\end{cases} \tag{10}$$

Different kernel functions map the input vectors into a higher dimensional space in different ways. The choice of kernel function in SVM directly affects the number of parameters as well as the computational complexity. Moreover, parameters in SVM such as γ, C and ϵ should also be selected carefully because the predictive performances of models vary

295 significantly under different combinations of preset parameters (Hsu et al. 2003). Therefore,
296 the selection of appropriate kernel function and associated parameters is a critical procedure in
297 building efficient SVM models.

298 *3.3* ***Interpretative Methods of Important Variables***

299 In order to understand the underlying mechanism of algal growth dynamics, identifying
300 and interpreting the important environment variables is an essential step after the modeling and
301 prediction. A number of explanatory methods to interpret the importance of environmental
302 variables in ecological machine learning models have been proposed and discussed by many
303 researchers (Gevrey et al. 2003; Olden et al. 2004). Amongst them, the 'weight' methods that
304 explain the factor importance by looking at the magnitude of connecting weights and the
305 'stepwise' methods that rank the importance by continuously adding or deleting individual
306 variable changing the model error largest are commonly used. Moreover, these methods are
307 considered as effective sensitivity analysis techniques for machine learning methods that are
308 conducive to interpret the relative importance of input variables (Gevrey et al. 2003; Lee et al.
309 2003; Chau et al. 2007; Foo et al. 2016). In this study, the forward stepwise method and the
310 simplified 'weight' method suggested by Gevrey et al. (2003) are adopted to rank the
311 contributions and quantify the relative importance of each environmental variable.

312 In the first step of forward stepwise process, each out of the eight variables is trained as
313 input by individual model, so that the respective variable of smallest RMSE is ranked the most
314 significant. Then this determined variable is combined with each of the other seven variables
315 respectively to form seven models results. This process is repeated, and each step ranked one

of the remaining variable, until all variables are achieved. As a result, the order of integration of the input variables in the network is the order of the importance of their contributions (Gevrey et al. 2003; Olden et al. 2004).

In the 'weight' method, the connecting weights between the input layer and hidden layer represent the importance of each input. The variable with a higher RI is supposed to have more contribution (that is, more important). This indicator can be defined by Eq. (11):

$$Q_{ih} = \frac{|w_{ih}|}{\sum_{i=1}^{ni}|w_{ih}|},\ RI(\%)_i = \frac{\sum_{h=1}^{nh} Q_{ih}}{\sum_{h=1}^{nh}\sum_{i=1}^{ni} Q_{ih}} \times 100 \quad (11)$$

where w_{ih} is the weight connecting input and hidden neuron; ni and nh are number of input and hidden neuron.

4. Application Procedure for Water Quality Prediction

4.1. Data Preparation

4.1.1 Dataset Selection and Division

The water quality data in Tolo Harbour is biweekly/monthly monitored by the Environment Protection Department (EPD) of Hong Kong. The weakest flushed monitoring station TM3 at 22°27'N, 114°12'E (Figure 2) is selected as the sampling point so that the hydrodynamic effects can be separated (Lee et al. 2003). In this study, the 30-year water quality data from 1988 to 2018 are used for modelling. Since the raw data are measured biweekly or monthly, we applied linear interpolation to obtain daily values. Therefore, totally 11293 interpolated daily samples are obtained where the first 9000 samples (from 1988 to 2012) are selected as training set and the remaining 2293 untrained data (from 2012 to 2018) are used as

testing set. The training set is originally fed into both ANN and SVM for model training, while the retained testing set is used to test the capability of the model to predict the output for those new samples that were not contained in the training set, which is also termed as generalization performance (Xie et al. 2012). Given that the 5-fold cross validation method is adopted for model validation in this study, the folded validation set is randomly partitioned from training set into 5 equal sized subsets and then used for model validation before the testing stage.

4.1.2 Input Variables and Time Lags

Similar to previous modelling and filed studies in Tolo Harbour (Lee et al. 2003; Muttil and Chau 2007; Li et al. 2014), the following water quality indicators are taken as model inputs, including total inorganic nitrogen (TIN, mg/L), phosphorus (PO_4, mg/L), Chlorophyll-a (Chl-a, μg/L), dissolved oxygen (DO, mg/L), water temperature (℃), and the secchi-disc depth (SDD, m) which measures the light intensity. In addition, some studies also suggest that 5-day biological oxygen demand (BOD5, which measures the organic pollutants) and acid-base conditions (pH) of water are directly influence to algal growth but usually were ignored in previous researches on HAB issues of Tolo Harbour. As complement, we also take the variables of BOD5 (mg/L) and pH as consideration in this study. All the field measured data are measured at surface, middle and bottom of water column and then depth-averaged for analysis. The modelling output should be an indicator that represents the magnitude of algal reasonably. Chlorophyll-a is one of the important components of algal cells which is a commonly used estimator to reflect the algal abundance in HAB studies (Li et al. 2004; de Oliveira et al. 2020). Therefore, the concentration of Chl-a at time t is selected as model output.

In this present study, a 1-week prediction of the algal blooms in Tolo Harbour is set as the modelling target based on the consideration of the ecological process and sampling frequency (Lee et al. 2003). However, the reoccurrence of algal blooms explosion in Tolo Harbour has been observed with a periodic cycle of 1-2 week. Lee et al. (2003) confirmed this phenomenon with the observation based on the continuous telemetric techniques, which suggested a significant self-correlation of algal dynamics up to time lag of around 2 weeks. To this end, the lag times of 7-13 days are introduced for lead-time prediction to identify the significant input variables (Muttil and Chau 2006). In other word, 8 environmental variables with 7 lag times (i.e. t-13, t-12 …, t-7) are chosen as 56 input variables of model (t denotes the time to predict).

4.1.3 Normalization

Considering that the values of the eight environmental variables are not of the same magnitude, all variables are normalized as Eq. (12) to ensure that the data used for modeling is homogeneous so that models converge effectively.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (12)$$

where x and x' denote the original and normalized value respectively; x_{max} and x_{max} denote the respective maximum and minimum value of each variable.

4.2. Model Determinations

The modelling performance of ANN is directly affected by network structure hence an important procedure in ANN model construction is to select the network structure and its configuration which makes the model reach the optimal performance. Since a single hidden layered model is well enough to approximate any continuous function at an arbitrary precision

(Cybenko 1989), in this study, the frame of the BP neural network with an input layer, a hidden layer and an output layer is selected.

In input layer, there should be 56 input nodes, corresponding to 56 variables determined in the previous subsection, while the output layer should have only one node which produces the predicted Chl-a concentration. However, there is no deterministic principle for selecting the number of hidden nodes. The trial and error method is conducted to determine the optimal hidden nodes. By building 11 networks and varying the number of hidden neurons between 3 to 13, the best performed network model is found with 5 hidden nodes. The sigmoidal function Eq. (13) is used as transfer function between input layer and hidden layer as well as the transfer function between hidden layer and output layer to assure the best performance.

$$f(x) = \frac{1}{1+e^{-x}} \tag{13}$$

As mentioned before, in order to make full use of advantages of both gradient-based and population-based algorithms, four ANN models with integrated learning algorithms (i.e. GDM-GA, GDM-PSO, LM-GA and LM-PSO) are compared in this study. In these hybrid learning process, GA or PSO are employed for global search and then GDM or LM are used for fast local convergence. Before modelling, the parameters of each algorithm are claimed as follows. The learning rate and training epochs are determined as 0.01 and 1000 respectively, the momentum of GDM is selected as 0.9, the probability of crossover and mutation of GA are 0.7 and 0.01 respectively, the particle velocity of PSO are ranged from -2.0 to 2.0 and the particle position are ranged from -1.5 to 1.5. For both GA and PSO, each generation consists of 30 individuals and the operations will terminate at the maximum generation of 50. The main

parameters of the ANN models are listed in Table 1.

Compared with ANN, the model structure of SVM requires many fewer parameters to be specified. As mentioned previously, the modelling performance of SVM largely hinge on the applied kernel function and predefined parameters. Empirically, the Gaussian kernel function (RBF function) should be the first choice since it can ideally handle nonlinear problems by relatively simpler calculations and fewer parameters (Hsu et al. 2003). It is also widely employed in many cases of HAB events forecasting with satisfactory performances (Xie et al. 2012; Park et al. 2015). Thus, the Gaussian kernel function is selected herein to nonlinearly map feature vectors. The main parameters of SVM model are summarized in Table 2.

Table 1 Key parameters of ANN models

Input Nodes	56	Hidden Layer	1
Hidden Nodes	5	Output Node	1
Transfer Functions	Sigmoid/Sigmoid	Training Algorithms	GDM/LM/PSO/GA
Gradient Descent with Momentum (GDM):			
Learning Rate	0.01	Training Goal	0.001
Learning Epochs	1000	Momentum Term	0.9
Levenberg-Marquardt algorithm (LM):			
Learning Rate	0.01	Training Goal	0.001
Learning Epochs	1000		
Genetic Algorithm (GA):			
Maximum Generation	50	Population size	30
Crossover rate	0.7	Mutation rate	0.01
Generation gap	0.95	Chromone length	20
Crossover strategy	Single-point	Selection Strategy	Roulette wheel
Particle Swarm Optimization (PSO):			
Maximum Generation	50	Population size	30
Particle velocity	-2.0~2.0	Particle position space	-1.5~1.5

410 **Table 2** Key parameters of SVM model

Kernel Function	Gaussian Function
Gaussian Function Parameter	0.25
Insensitive Factor	0.1
Loss Function	8
Cross validation	5-fold
Grid-search space	2^{-4}~2^{4}

411

412 Before applying Gaussian kernel functions in SVM, there are only three parameters to be
413 determined: (1) The constant C, which penalize the outliers. (2) The insensitive parameter ϵ,
414 which controls the error tolerance of insensitive zone. (3) The Gaussian parameter γ.
415 Theoretically, inappropriate choices of these values may induce overfitting or underfitting. In
416 order to determine the optimal combination of three parameters, the cross-validation and grid-
417 search method recommended by Hsu et al. (2003) is conducted. A 5-fold cross-validation is
418 carried out and the parameters with minimum cross-validation error are picked to train the
419 model. In present case, the ideal parameters are found as $C = 8;\ \epsilon = 0.1;\ \gamma = 0.25$.

420 ***4.3. Performance Indicators***

421 In order to quantitatively describe the modelling performance, we select the root-mean-
422 square-error (RMSE) Eq. (14) to evaluate measure the deviation between the predicted value
423 and the measured value and use the correlation coefficient (CC) Eq. (15) to measure the
424 goodness of fit. Generally, a model with smaller RMSE is considered to have less modelling
425 error while a model with CC closer to 1 is considered to have a better positive correlation.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{m}} \quad (14)$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{\hat{y}}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{m}(\hat{y}_i - \bar{\hat{y}}_i)^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2}} \quad (15)$$

where y_I and $\hat{y}_I$ denote actual value and predicted value respectively and the variables with a capped bar represents the average value; m is the number of samples.

In this study, both the modelling accuracy and its generalization ability are considered in performance evaluation. After the model is trained, the training set will be re-input into the model to check the accuracy of the model and then the untrained testing set will be used to test the generalization ability to process new data. In addition, the training time (T) is also employed to reflect the computing cost.

To sum up, the application principle and procedure of the ML-based water quality prediction developed in this study starts with the model description and variable selection. Prepared data set with 8 selected environmental variables then will be fed into two different ML models (i.e. ANN and SVM) and relevant parameters will be dynamic adjusted repeatedly. After predicting and results analysis, the performances of each model will be evaluated and compared so that the best adopted model can be selected. For clarity, the integral process of the developed ML-based scheme is presented in Figure 6.

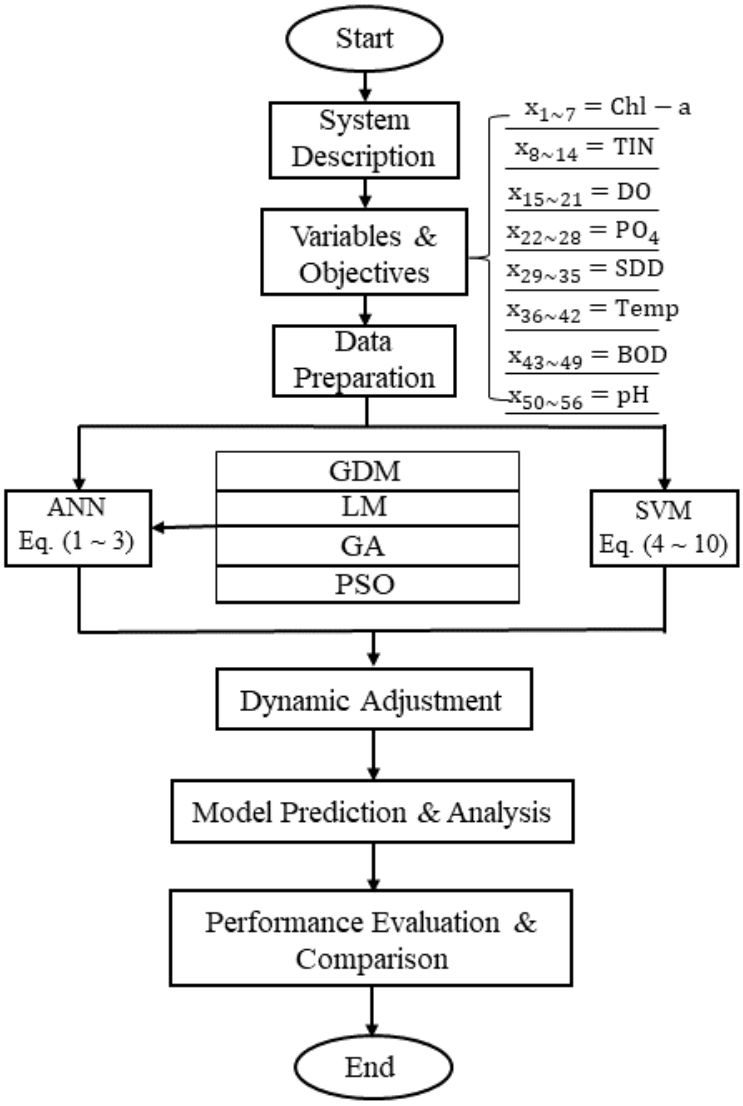


Figure 6: The application principle and procedure of ML-based water quality prediction

5. Results and Discussion

5.1. Comparison of Predicting Performances

In this section, the full models trained with all mentioned 8 environmental variables are established based on both ANN and SVM techniques as given in Figure 6. Table 3 lists the modelling results evaluated by error, correlation and training time. In terms of ANN models, predicting performances of four learning algorithms are compared and the results of

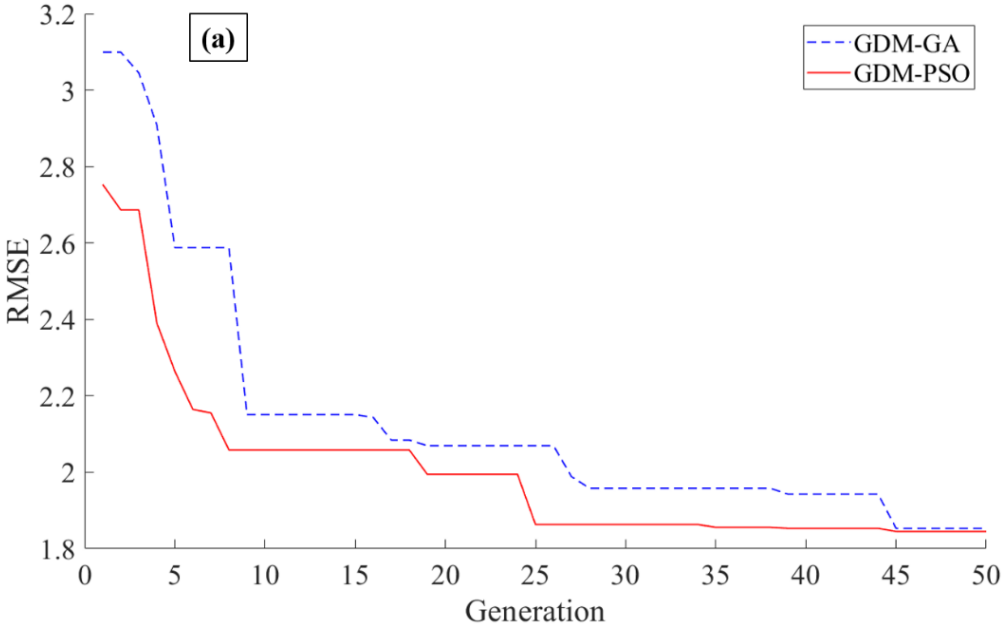
evolutionary process and water quality prediction are shown in Figures 7 and 8, respectively. Overall, four algorithms all showed good predicting ability to accurately capture both the growth trend and magnitude of Chl-a concentration both in training and testing set (Figure 8), which means that the four models are effectively established and there is no overfitting problem. Based on the results of modelling error and correlation, the global optimal solution was found by PSO with fewer generation steps than GA as indicated in Table 3 and Figure 7. In this case, it reveals that PSO may have better search efficiency on global optimization.

Table 3 Performance indicators of different ANN algorithms

	Training Set		Testing Set		Training
	RMSE	CC	RMSE	CC	Time(s)
GDM-GA	3.750	0.798	1.853	0.863	3.85
GDM-PSO	3.943	0.770	1.845	0.803	3.39
LM-GA	1.717	0.961	0.769	0.976	2.12
LM-PSO	1.615	0.965	0.765	0.972	2.04
SVM	1.243	0.980	0.660	0.984	61.19

On the other hand, by comparing results of two different gradient-based algorithms in Table 3 and Figure 7, it can be clearly seen that outputs predicted by LM has a better agreement with observations than that predicted by GDM. The models using the GDM algorithm showed larger errors and is more prone to phase mismatch as given in Figure 8(b). Furthermore, the convergence rate of LM is also considered to be superior with nearly a half computing time of GDM. Comprehensively, the network using LM-PSO algorithm for training is considered as a better performed model because of the higher accuracy and efficiency in both training (with RMSE of 1.615 and CC of 0.965) and testing sets (with RMSE of 0.765 and CC of 0.972).

Therefore, from this comparative study, LM-PSO algorithm is retained in the ANN model for further analysis.



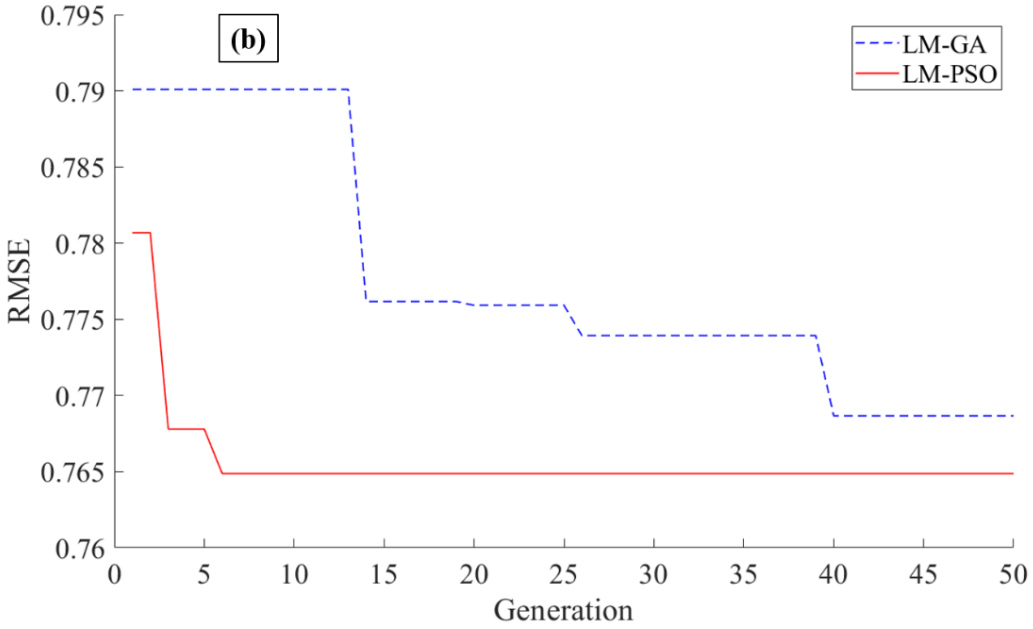


Figure 7: Evolutionary process of four different hybrid algorithms: (a) GDM with GA and PSO; (b) LM with GA and PSO

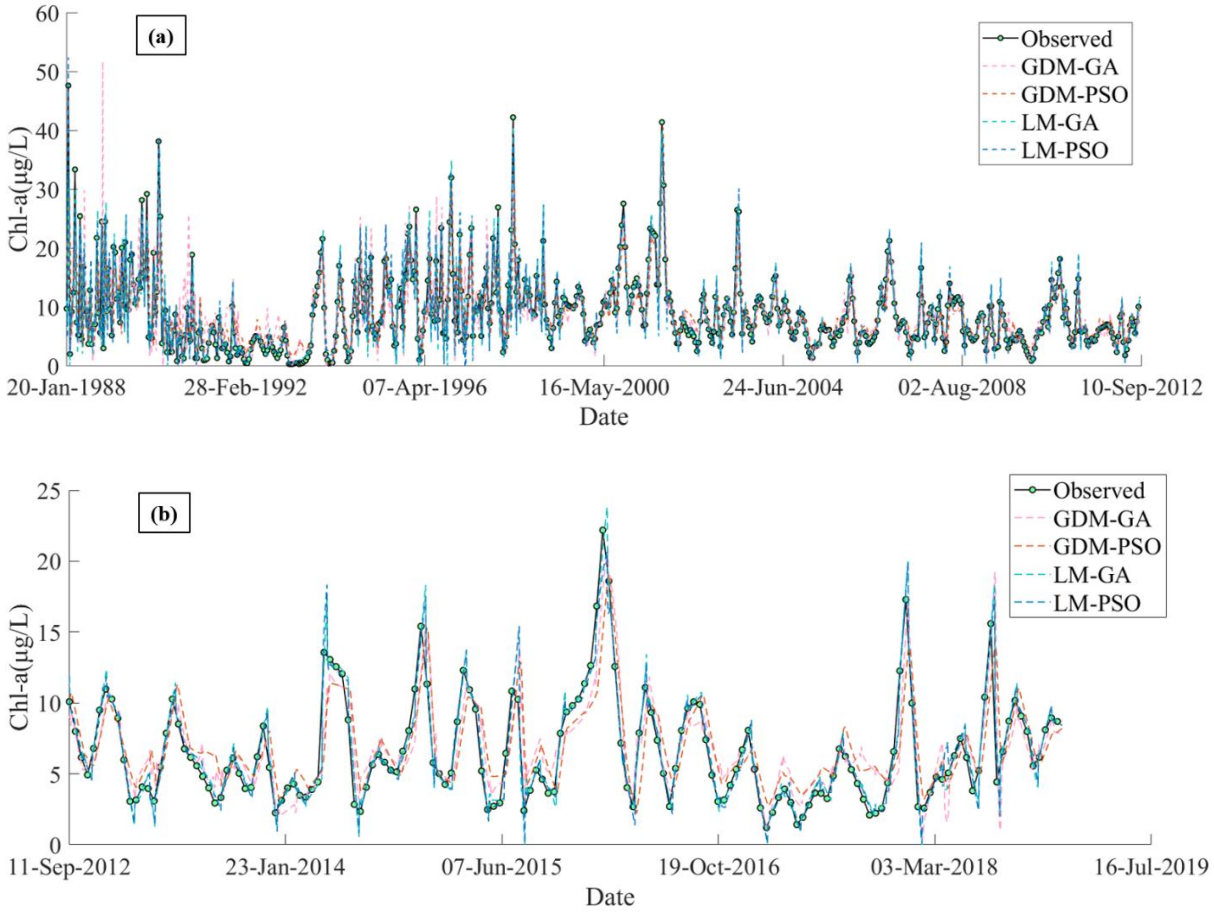


Figure 8: Results Comparison of four hybrid ANN models: **(a)** training set **(b)** testing set

In terms of RMSE and CC, Table 3 also shows that the SVM model trained with the same training set data performed even slightly better than the best ANN model (i.e. LM-PSO above). The comparisons of these two ML methods are further depicted in Figures 9 and 10 for the results of water quality prediction and performance, respectively. By using SVM technique, the highest correlation coefficient was achieved in both training set (CC=0.980) and testing set (CC=0.984) as well as the lowest RMSE (1.243 for training set and 0.660 for testing set respectively) (see Figure 10). Meanwhile, it is revealed in Figure 9 that the SVM can handle better the nonlinear relationship between water quality variables and chlorophyll concentration than the ANN models. However, it should be noted that it takes much longer to train the SVM

(~60s) than the ANN (2~4 s) as the SVM takes a quadratic programming with time complexity of $O(m^3)$ (m is the number of examples).

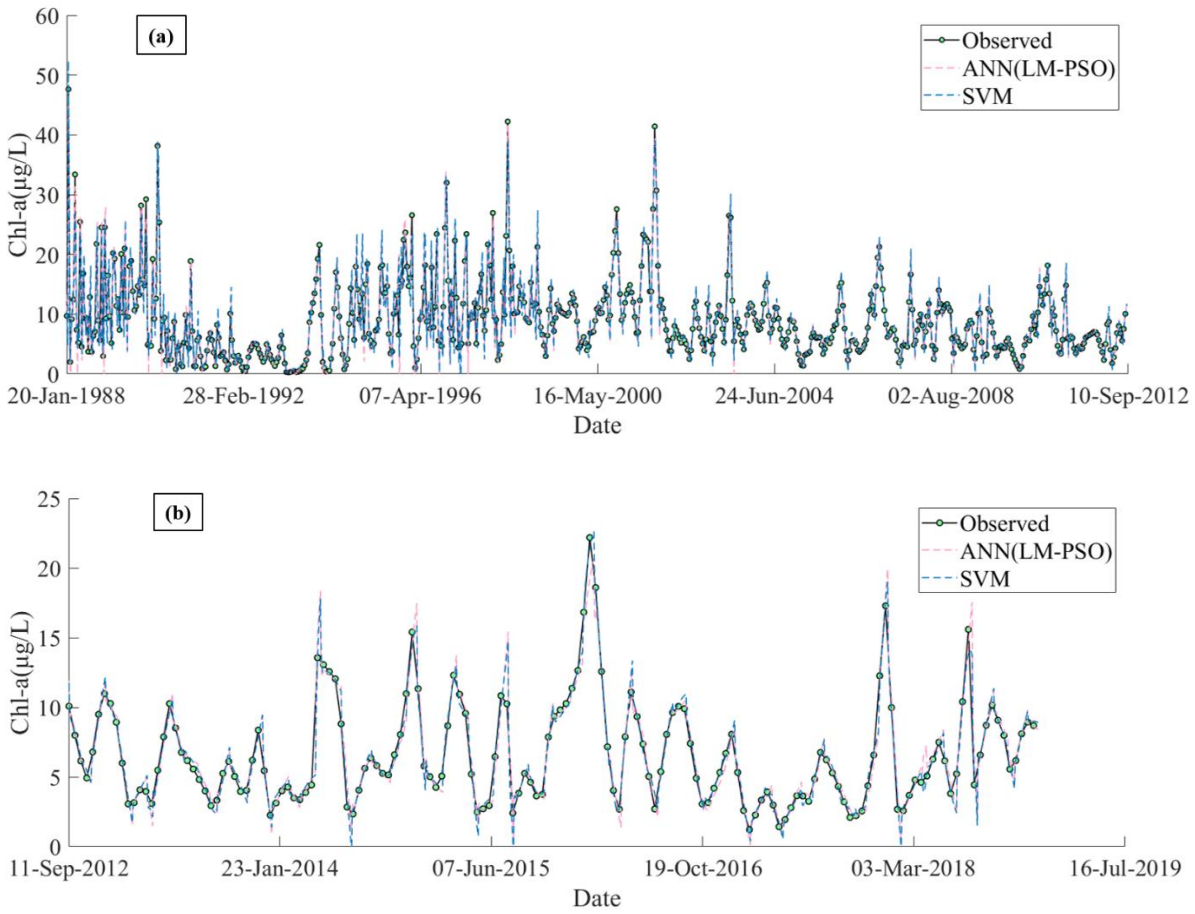
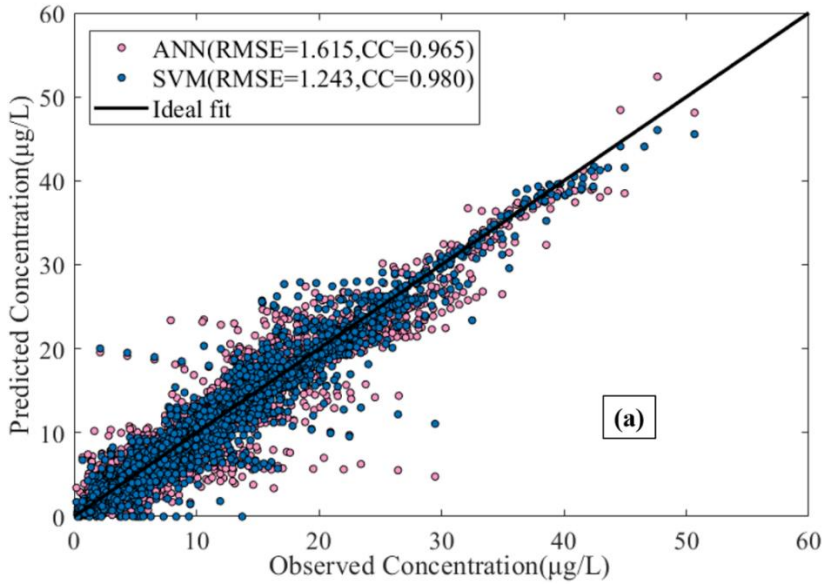
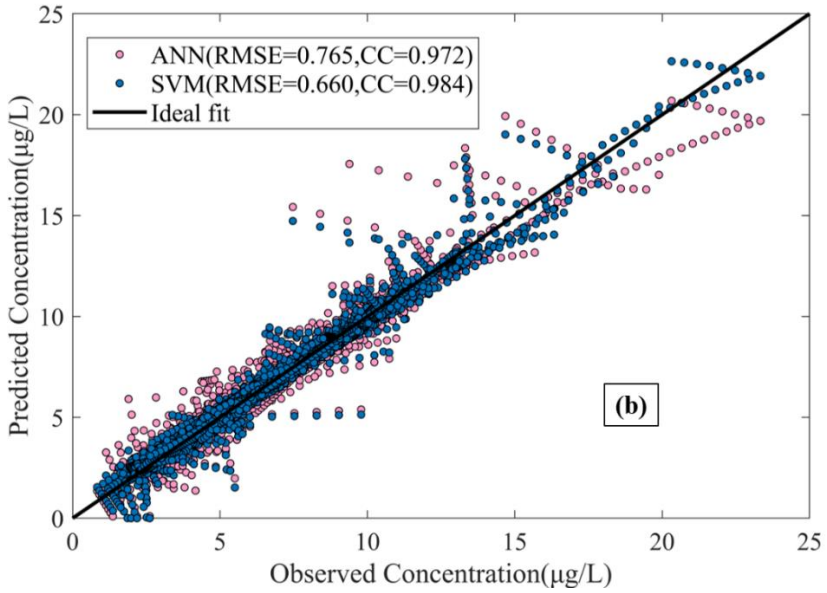


Figure 9: Results comparison of the ANN and SVM methods: **(a)** training set **(b)** testing set





494

495 **Figure 10:** Prediction performance comparison of the ANN and SVM methods: **(a)** training
496 set; **(b)** testing set

497 ***5.2. Variable Importance Analysis***

498 The identification of the significant factors that strongly influence algal dynamics is
499 imperative to understand the causality and mechanism of algal blooms events, which are also
500 beneficial in early warning and precautionary measures implement. Two different methods
501 mentioned in Section 3.3, namely 'stepwise' method and 'weight' method, are conducted in
502 this study.

503 Tables 4 and 5 list the best input combination of the model with lowest RMSE in each
504 step for the 'stepwise' methods of both ANN and SVM respectively. Noted that the
505 redundancies, noise and irrelevant components are likely to be introduced by adding variables,
506 while the performance of the model does not increase monotonically with the increase of
507 variables as shown in Figure 11. According to the results of 'stepwise' method in Tables 4 and
508 5, both ANN and SVM method suggest Chl-a concentration as the most significant factor

contributing to algal growth in the studied area. Other variables including BOD5, TIN, DO and pH are also considered to have relatively higher contributions to the algal growth, which are ranked 2-4 in the forward selection process. It is also noting that with the increase of the number of variables, the training time of SVM may greatly exceed that of ANN, but with limited improvement on the model performance (Figure 11). More details on obtaining the final results of Table 4 and Table 5 may refer to the supplementary materials of this paper.

Table 4 Results and Performance of the 'stepwise' method for the ANN method (all the variables with 7-13 lagged days)

Step	Best Combination of Inputs	Training Set		Testing Set		Time
		RMSE	CC	RMSE	CC	(s)
1	Chl-a	1.835	0.955	0.799	0.977	1.50
2	Chl-a, BOD5	1.775	0.958	0.822	0.975	1.64
3	Chl-a, BOD5, TIN	1.783	0.957	0.839	0.974	1.57
4	Chl-a, BOD5, TIN, DO	1.751	0.959	0.821	0.974	1.64
5	Chl-a, BOD5, TIN, DO, pH	1.712	0.961	0.772	0.977	1.61
6	Chl-a, BOD5, TIN, DO, pH, PO4	1.774	0.958	0.825	0.974	1.65
7	Chl-a, BOD5, TIN, DO, pH, PO4, SDD	1.696	0.961	0.839	0.974	1.89
8	Chl-a, BOD5, TIN, DO, pH, PO4, SDD, Temp	1.615	0.965	0.765	0.972	2.04

Table 5 Results and Performance of the 'stepwise' method for the SVM method (All the variables with 7-13 lagged days)

Step	Best Combination of Inputs	Training Set		Testing Set		Time
		RMSE	CC	RMSE	CC	(s)
1	Chl-a	2.093	0.945	0.799	0.977	9.62
2	Chl-a, BOD5	1.713	0.962	0.746	0.979	25.12
3	Chl-a, BOD5, TIN	1.814	0.958	0.773	0.978	28.83
4	Chl-a, BOD5, TIN, DO	1.474	0.971	0.751	0.979	42.24
5	Chl-a, BOD5, TIN, DO, pH	0.778	0.992	0.597	0.987	42.77
6	Chl-a, BOD5, TIN, DO, pH, PO_4	0.806	0.991	0.588	0.987	50.43
7	Chl-a, BOD5, TIN, DO, pH, PO_4, SDD	0.717	0.993	0.556	0.988	52.21
8	Chl-a, BOD5, TIN, DO, pH, PO_4, SDD, Temp	1.243	0.980	0.660	0.984	61.19

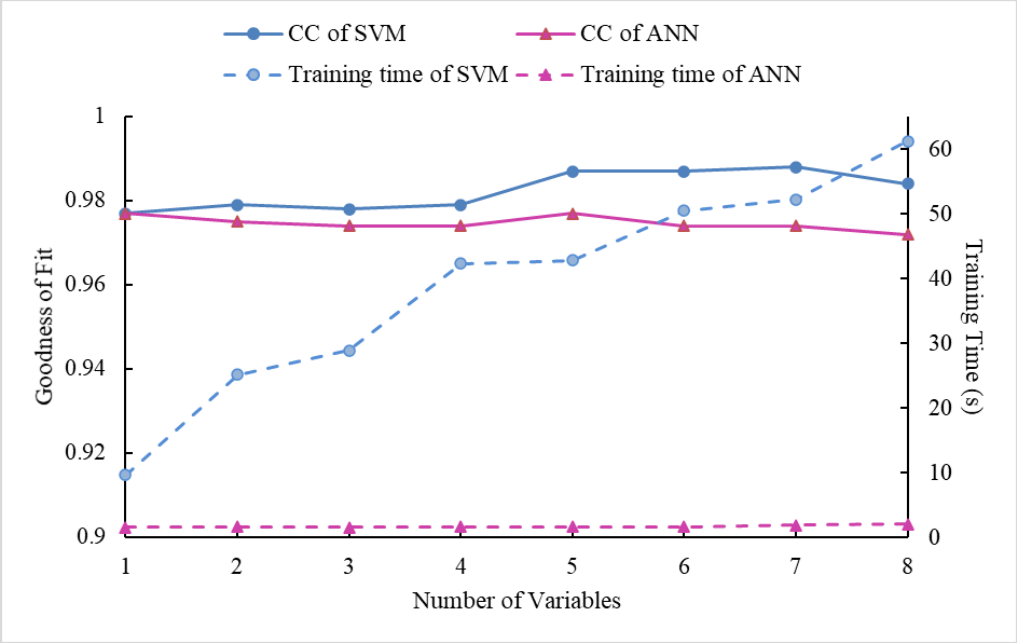


Figure 11: Performance comparison based on 'stepwise' method (The CC indicator is for testing set; training time is for training set)

In addition, the suggested 'weight' method (Gevrey et al. 2003) is also conducted to quantify the relative importance (RI) for each variable. The results of RI values for each input variable are shown in Table 6 and Figure 12. Specifically, the values that are larger than the overall average (1/56=1.78%) are deemed to be relatively more significant and thus shaded in blue in the table. In the analysis of 'weight' method, all variables with lag times of (t-7) and (t-13) in Table 6 indicate potentially high relation with current Chl-a concentration (output) (as visualized in Figure 12). Furthermore, considering all time lags of a variable as a whole, the summation of the RI for each environmental variable can be calculated by Eq. (16) and is plotted in Figure 13.

$$S_i = \sum_{i=1}^{nT} RI(\%)_i \quad (16)$$

Not surprisingly, the 'weight' method also suggests that the factor of Chl-a being the most

significant variable with the sum total RI of 22.64% (see Figure 13).

By comparison, the result of the "weight" method is overall consistent with that of the former "stepwise" method, which also confirms the applicability and accuracy of the ML methods for the water quality prediction proposed in this study.

Table 6 The RI results by the 'weight' method

Variable	RI of Each Input Variable (%)							Sum
	t-7	t-8	t-9	t-10	t-11	t-12	t-13	
Chl-a	8.28	6.69	1.61	0.51	1.24	1.97	2.34	22.64
TIN	2.83	1.44	2.54	2.94	1.03	1.20	3.17	15.15
DO	2.70	1.05	1.81	0.31	1.54	1.01	1.19	9.60
PO_4	2.28	1.61	1.56	1.15	0.72	1.20	0.97	9.48
SDD	1.90	0.49	1.58	0.80	0.49	0.75	2.18	8.18
Temp	2.30	0.68	0.62	1.02	0.96	0.65	1.07	7.30
BOD5	5.30	1.90	1.01	1.08	0.94	1.77	3.90	15.90
pH	3.82	1.00	0.64	0.87	1.92	1.55	1.94	11.74
							Total	100

**Numbers in blue are for IR > 1.78% (i.e., overall average)*

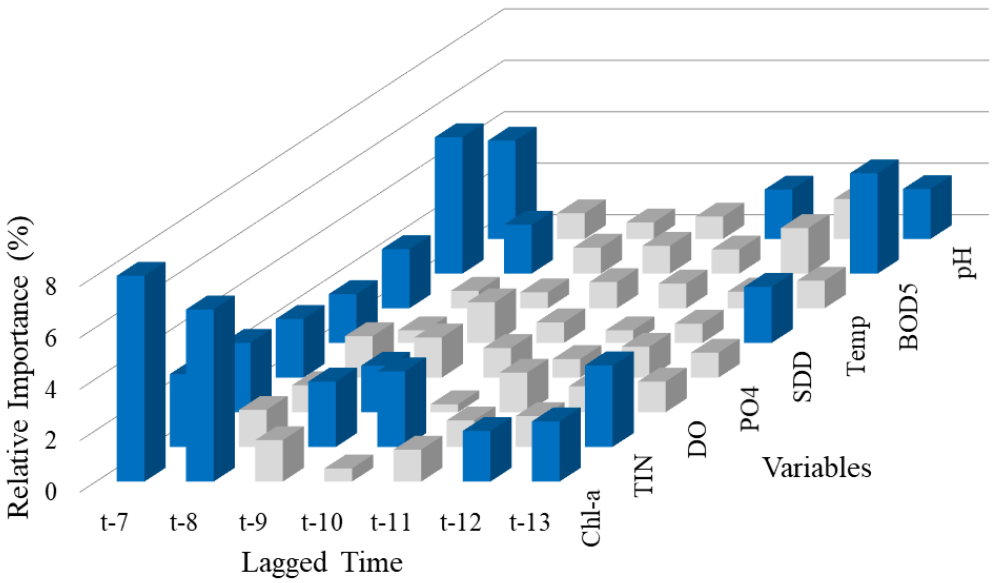


Figure 12: The RI value of each influence variable

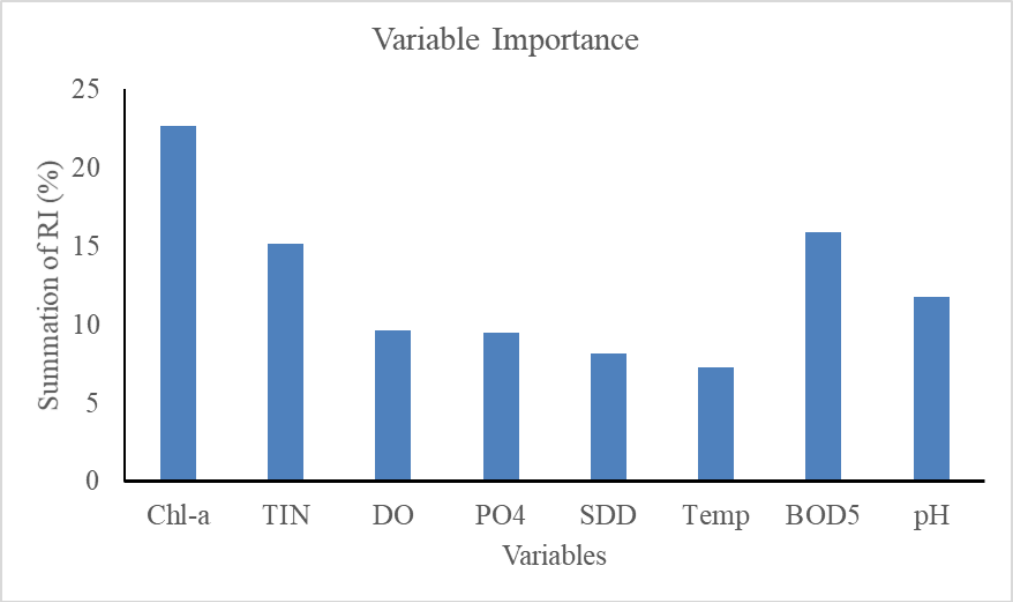


Figure 13: The summation of the RI for each environmental variable (S_i) by considering all time lags of a variable as a whole

5.3. Environmental Interpretation

Based on the above results and analysis by the proposed ML methods, the time-lagged Chl-a concentration is considered to be the most significant variable contributed to the current algae abundance. In other word, the upcoming algal bloom events are strongly related to Chl-a concentration with 1-2 weeks ahead, which indicates the occurrence of HAB in Tolo Harbour with a cycle of 1-2 weeks. This auto-regressive characteristic of algal growth dynamics is also observed and concluded by other scholars (Lee et al. 2005; Muttil and Lee 2005; Muttil and Chau 2006). After comparing with three differently flushed stations, Muttil and Lee (2005) confirmed the phenomenon was related to the tidal flushing conditions. The sampling station TM3 is located at a side cove in Tolo Harbour. Due to the very limited tidal flushing, the water circulation is extremely weak (with average current velocity of 0.04m/s) and water residence time is relatively long (about 28 days). Hence, it is justifiable that the time-series reoccurrence

phenomenon is very obvious in this semi-closed coastal water due to the inertial process of the physical system (Muttil and Chau 2007).

The BOD5 is ranked as the secondary important variable by 'stepwise' method of both ANN and SVM and also obtained a relatively high RI (15.90%) in the 'weight' method. Biologically, the indicator of BOD5 represents the amount of oxygen demanded by decomposing micro-organisms to break down organic substance over five days, which is usually adopted as the indicator of the degree of organic pollution in water. It has been determined as a main factor contributing to eutrophication (Chen et al. 2002; Solanki et al. 2010) and a linear and positive relations between BOD5 and Chl-a are observed based on numerical models (Xu and Xu 2015). In Tolo Harbour, there was also consistent observations that the highest level of BOD5 was detected during the period of the worst eutrophication (Li et al. 2004; Xu et al. 2004b).

The ambient nutrient variable TIN is ranked as the third significant variable in all analyses, while the other nutrient variable PO4 had a less important contribution. In general, the concentration of nutrient elements in water, such as nitrogen (N) and phosphorus (P), promote the growth of aquatic phytoplankton. It is interpretable that the growth and reproduction of algal are directly dependent on nutrient supply. (Xu et al. 2004a; Chau 2007; Xu et al. 2010; Davidson et al. 2012). The municipal and industrial wastewater containing heavy nitrogen and phosphorus load stimulate the growth of algal biomass in response to the increase in nutrient load (Chang et al. 2017). Conversely, the exhaustion of essential nutrient limits development of algal flora. However, unlike freshwater in riverine or reservoir, the nitrogen is more likely

582 to be the limiting factor in many coastal eutrophic water systems rather than phosphorus (Elser
583 et al. 2007; Davidson et al. 2012; Paerl et al. 2014; Park et al. 2015). In Hong Kong, the Tolo
584 Harbour was also classified as a nitrogen limiting water system especially during the period of
585 frequent HAB (Xu et al. 2010). Therefore, it is reasonable that the RI of TIN is much higher
586 than PO4, which also means that nitrogen plays a much more important role in the growth of
587 algae than phosphorus in Tolo Harbor.

588 The dissolved oxygen (DO) and pH are ranked as the fourth and fifth significant variables,
589 respectively. Theocratically, DO is necessary for the aquatic organisms in terms of respiration
590 and other important biochemical reaction. When a large number of algae accumulate rapidly,
591 the dissolved oxygen in the water will be depleted, resulting in hypoxia issues. In Tolo Harbour,
592 a negative correlation between the dissolved oxygen and algal abundance was also observed
593 and verified based on both statistical data and three-dimensional numerical eutrophication
594 models (Lee et al. 2005; Chau 2007). The pH value is considered as another plant growth
595 limiting factor which directly affect the absorption of nutrient solution (Khan and Ansari 2005).
596 The formation of Chl-a is also limited by acid environment, while alkaline environment with
597 high pH value was demonstrated to promote the growth of algal and often results in bloom
598 (George and Heaney 1978; Wei et al. 2001; Yang et al. 2008).

599 In this study, the water temperature and SDD are suggested relatively insignificant to the
600 algal dynamics in Tolo Harbour. SDD is a measure of light penetration into water body which
601 is usually used to water transparency. Theoretically, SDD is mainly influenced by light intensity.
602 However, light was considered to rarely limit the growth of phytoplankton in inner Tolo (Xu et

603 al. 2010). Some scholars also pointed out that water temperature was also an important factor
604 to promote algal growth (Park et al. 2015; Michalak 2016), but the annual temperature variation
605 is quite small (less than 1℃ per month) in Tolo Harbour due to the tropical climate, hence it
606 is also reasonable that the water temperature has limit impacts on algal growth (Flewelling et
607 al. 2005; Muttil and Chau 2006; Cressey 2017).

608 In conclusion, the current chlorophyll concentration has a predictive effect on the
609 occurrence of HAB events in the upcoming week. In the process of preventing and controlling
610 HAB events in Tolo Harbour, though it is also important to focus on DO and pH, the load of
611 organic pollutants and nitrogen should be reduced at priority, while SDD and water temperature
612 are not the key points in water quality restoration. To some extent, these machine learning
613 models are promising to provide environment management department with some useful
614 information to understand the insight of the algal growth in Tolo Harbour so that suitable
615 strategies can be made to restore eutrophication and mitigate the harmful bloom impacts.

616 ***5.4. Long-term Change of Water Quality in Tolo Harbour***

617 With intensified anthropogenic exploitations since 1970s, the ecosystem of Tolo Harbour
618 began to degrade. In early 1980s, increased population migrated to Tai Po and Shatin, two new
619 towns along the Tolo Harbour, where industrial areas were also developed. The water
620 environment degraded sharply due to excessive domestic and industrial sewage discharging
621 into Tolo Harbour nearby. Averagely, during the decade of 1980s, the daily BOD and TIN loads
622 caused by sewage discharge to Tolo Harbour were recorded as high as 14,000 and 6,000kg/day
623 respectively (Xu et al. 2004a), which stimulated the development and production of algal

species. In late 1980s, harmful algal blooms or red tides events began to occur in Tolo Harbour frequently with the worst situation of 43 incidents in 1988 alone. In order to mitigate environmental pollutions, the Hong Kong authority announced Tolo Harbour Action Plan (THAP) including a series of schemes, such as livestock waste control, effluent diversion schemes and sewage treatment works with the target level of BOD and TIN daily discharge decrease to 5,000 kg/day and 600kg/day. After such scheme implementation, the water quality has been improved gradually, with the average annual HAB incidents in Tolo Harbour decreased from 16 during 1986~1996 to only 5 during 2008~2018. Specifically, at the most affected TM3 station, the mean concentrations of BOD5 and TIN were declined by 37.9% and 61.9% respectively, and the average annual Chl-a now is almost less than 10 (μg/L) (EPD, 2019).

The successful restoration of Tolo Harbor shows that reduction of BOD and TIN load as the THAP's primary targets has a very obvious effect on reducing red tide and water bloom. It is noteworthy that the similar enlightenments can be obtained from the ANN and SVM models developed in this study. In reality, each water ecosystem has its own individuality hence it is hard to fully grasp a causative pattern of algal developments. Before the complicated relationship between algal and environmental variables is well-understood, machine learning models seem to be good supplements to understanding the complex process. Although machine learning models are regarded as a 'black box' model, the case study of Tolo Harbour confirms that the results and interpretations can play a significant role in restoring water degradation.

6. Conclusions

In this study, two machine learning (ML) models namely ANN and SVM are implemented

and applied to model and predict the algal growth trend and magnitude in Tolo Harbour by training with 30-year monitored data. In general, both ANN and SVM could provide very satisfactory results. During the model training stage of the ANN, four hybrid learning algorithms are implemented and compared for their performance in improving the water quality prediction. In terms of accuracy and generalization, LM-PSO algorithm is proved to be the best predictive performance over other ANN models. In addition, the performance of SVM is better than all ANN models in terms of water quality prediction results, but with lower computational efficiency due to the inclusion of the nonlinear relationships among variables and outputs.

Based on the application results and analysis, it is demonstrated that the upcoming algal bloom events are strongly related to Chl-a concentration with 1-2 weeks ahead of the time, which indicates the auto-regressive characteristics of algal dynamics. The analysis results also reveal that the variables of BOD, TIN, DO, PO4 and pH can be key variables contributing to abundance of blooms in Tolo Harbour during past three decades. This is evidenced by the practice that the occurrence of HAB events has been noticeably decreased after the long-term efforts to reduce BOD and nutrient load in this studied area. This result is also consistent with the recommendation from the ML methods in this study, which confirms the usefulness of the interpretations of important variables by these methods in restoring water degradation.

Finally, the results and findings of this study also suggest that the ML methods can provide supplementary information for the understanding of the complicated algal behavior and eutrophication mechanisms as well as appropriate suggestions on water quality prediction and improvement for total coastal hydro-environmental management.

Acknowledgement

This work was supported by the research project from the Hong Kong Polytechnic University and the Hong Kong Research Grants Council (no. 15200719 and no. 15201017).

References

AFCD (Agricultural Fisheries and Conservation Department), 2019. Hong Kong Red Tide Information Network. https://www.afcd.gov.hk/english/fisheries/hkredtide/redtide.html

Al-Azri, A. R., Piontkovski, S. A., Al-Hashmi, K. A., Goes, J. I., Gomes, H. d. R. and Glibert, P. M. (2014). Mesoscale and nutrient conditions associated with the massive 2008 Cochlodinium polykrikoides bloom in the Sea of Oman/Arabian Gulf. Estuaries and Coasts **37**(2): 325-338.

Chang, N. B., Bai, K. and Chen, C. F. (2017). Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management. Journal of Environmental Management, **201**: 227-240.

Chau, K. W. (2005a). Algal bloom prediction with particle swarm optimization algorithm. In International Conference on Computational and Information Science (pp. 645-650).

Chau, K. (2005b). A split-step PSO algorithm in prediction of water quality pollution. In International Symposium on Neural Networks (pp. 1034-1039).

Chau, K. (2006). A review on the integration of artificial intelligence into coastal modeling. Journal of Environmental Management, **80**(1), 47-57.

Chau, K. (2007). Integrated water quality management in Tolo Harbour, Hong Kong: a case study. Journal of Cleaner Production **15**(16): 1568-1572.

687 Chen, X., Li, Y. and Li, Z. (2002). Spatio-temporal distribution of Chlorophyll-a concentration
688 in Hong Kong's coastal waters. Acta Geographica Sinica: 422-428 (In Chinese).
689 Cressey, D. (2017). Climate change is making algal blooms worse. Nature (London).
690 Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics
691 of Control, Signals and Systems **2**(4): 303-314.
692 Daghighi, A. (2017). Harmful algae bloom prediction model for western lake erie using
693 stepwise multiple regression and genetic programming. ETD Archive. 964
694 Dai, C., Tan, Q., Lu, W. T., Liu, Y., & Guo, H. C. (2016). Identification of optimal water transfer
695 schemes for restoration of a eutrophic lake: An integrated simulation-optimization method.
696 Ecological Engineering, **95:** 409-421.
697 Davidson, K., Gowen, R. J., Tett, P., Bresnan, E., Harrison, P. J., McKinney, A., Milligan, S.,
698 Mills, D. K., Silke, J. and Crooks, A.-M. (2012). Harmful algal blooms: how strong is the
699 evidence that nutrient ratios and forms influence their occurrence? Estuarine, Coastal and
700 Shelf Science **115**: 399-413.
701 de Oliveira, T. F., de Sousa Brandao, I. L., Mannaerts, C. M., Hauser-Davis, R. A., de Oliveira,
702 A. A. F., Saraiva, A. C. F., de Oliveira, M. A. and Ishihara, J. H. (2020). Using
703 hydrodynamic and water quality variables to assess eutrophication in a tropical
704 hydroelectric reservoir. Journal of Environmental Management, **256**: 109932.
705 Ding, S., Su, C. and Yu, J. (2011). An optimizing BP neural network algorithm based on genetic
706 algorithm. Artificial Intelligence Review **36**(2): 153-162.
707 EPD (Environment Protection Department), 2019. Marine Water Quality Data.

https://cd.epic.epd.gov.hk/EPICRIVER/marine/?lang=en

Elser, J. J., Bracken, M. E., Cleland, E. E., Gruner, D. S., Harpole, W. S., Hillebrand, H., Ngai, J. T., Seabloom, E. W., Shurin, J. B. and Smith, J. E. (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecology Letters **10**(12): 1135-1142.

Flewelling, L. J., Naar, J. P., Abbott, J. P., Baden, D. G., Barros, N. B., Bossart, G. D., Bottein, M.-Y. D., Hammond, D. G., Haubold, E. M. and Heil, C. A. (2005). Red tides and marine mammal mortalities. Nature **435**(7043): 755-756.

Foo, Y. W., Goh, C., and Li, Y. (2016). Machine learning with sensitivity analysis to determine key factors contributing to energy consumption in cloud data centers. In 2016 International Conference on Cloud Computing Research and Innovations (ICCCRI) (pp. 107-113). IEEE.

Gavin, H. P. (2019). The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. Department of Civil and Environmental Engineering, Duke University http://people. duke. edu/~ hpgavin/ce281/lm. pdf: 1-19.

George, D. and Heaney, S. (1978). Factors influencing the spatial distribution of phytoplankton in a small productive lake. The Journal of Ecology: 133-155.

Gevrey, M., Dimopoulos, I. and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling **160**(3): 249-264.

Ghaffari, A., Abdollahi, H., Khoshayand, M., Bozchalooi, I. S., Dadgar, A. and Rafiee-Tehrani,

729 M. (2006). Performance comparison of neural network training algorithms in modeling of
730 bimodal drug delivery. International Journal of Pharmaceutics **327**(1-2): 126-138.
731 Gill, D., Rowe, M. and Joshi, S. J. (2018). Fishing in greener waters: understanding the impact
732 of harmful algal blooms on Lake Erie anglers and the potential for adoption of a forecast
733 model. Journal of Environmental Management, **227**: 248-255.
734 Glibert, P., Heil, C., Rudnick, D., Madden, C., Boyer, J. and Kelly, S. (2009). Florida Bay:
735 Status, trends, new blooms, recurrent problems. Contributions in Marine Science **38**: 5-
736 17.
737 Hagan, M. T. and Menhaj, M. B. (1994). Training feedforward networks with the Marquardt
738 algorithm. IEEE Transactions on Neural Networks **5**(6): 989-993.
739 Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2003). A practical guide to support vector
740 classification, Taipei.
741 Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings
742 of ICNN'95-International Conference on Neural Networks (Vol. 4, pp. 1942-1948). IEEE.
743 Khan, F. A. and Ansari, A. A. (2005). Eutrophication: an ecological vision. The botanical
744 review **71**(4): 449-482.
745 Kim, H. (1998). Cochlodinium polykrikoides blooms in Korean coastal waters and their
746 mitigation. Harmful Algae: 227-228.
747 Lee, J. H., Huang, Y., Dickman, M. and Jayawardena, A. W. (2003). Neural network modelling
748 of coastal algal blooms. Ecological Modelling **159**(2-3): 179-201.
749 Lee, J. H. W., Hodgkiss, I. J., Wong, K. and Lam, I. (2005). Real time observations of coastal

algal blooms by an early warning system. Estuarine, Coastal and Shelf Science **65**(1-2): 172-190.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics **2**(2): 164-168.

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., Chen, X., Wai, O. W. and 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.

Liu, Z., Wang, X., Cui, L., Lian, X. and Xu, J. (2009). Research on water bloom prediction based on least squares support vector machine. 2009 WRI World Congress on Computer Science and Information Engineering (Vol. 5, pp. 764-768). IEEE.

Lou, I., Xie, Z., Ung, W. K. and Mok, K. M. (2017). Integrating Support Vector Regression with Particle Swarm Optimization for numerical modeling for algal blooms of freshwater. Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs:125-141.

Lourakis, M. I. (2005). A brief description of the Levenberg-Marquardt algorithm implemented by levmar. Foundation of Research and Technology **4**(1): 1-6.

Lu, S. and Hodgkiss, I. (2004). Harmful algal bloom causative collected from Hong Kong waters. Asian Pacific Phycology in the 21st Century: Prospects and Challenges: (pp. 231-238)

771 Mamun, M., Kim, J.-J., Alam, M. A. and An, K.-G. (2020). Prediction of algal chlorophyll-a
772 and water clarity in monsoon-region reservoir using machine learning approaches. Water
773 **12**(1): 30.
774 McCabe, R. M., Hickey, B. M., Kudela, R. M., Lefebvre, K. A., Adams, N. G., Bill, B. D.,
775 Gulland, F. M., Thomson, R. E., Cochlan, W. P. and Trainer, V. L. (2016). An
776 unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions.
777 Geophysical Research Letters **43**(19): 10,366-310,376.
778 Michalak, A. M. (2016). Study role of climate change in extreme threats to water quality.
779 Nature **535**(7612): 349-350.
780 Mirzazadeh, T., Mohammadi, F., Soltanieh, M. and Joudaki, E. (2008). Optimization of caustic
781 current efficiency in a zero-gap advanced chlor-alkali cell with application of genetic
782 algorithm assisted by artificial neural networks. Chemical Engineering Journal **140**(1-3):
783 157-164.
784 Mulia, I. E., Tay, H., Roopsekhar, K. and Tkalich, P. (2013). Hybrid ANN–GA model for
785 predicting turbidity and chlorophyll-a concentrations. Journal of Hydro-Environment
786 Research **7**(4): 279-299.
787 Muttil, N. and Chau, K.-W. (2006). Neural network and genetic programming for modelling
788 coastal algal blooms. International Journal of Environment and Pollution **28**(3-4): 223-
789 238.
790 Muttil, N. and Chau, K.-W. (2007). Machine-learning paradigms for selecting ecologically
791 significant input variables. Engineering Applications of Artificial Intelligence **20**(6): 735-

744.

Muttil, N. and Lee, J. H. (2005). Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling **189**(3-4): 363-376.

Olden, J. D., Joy, M. K. and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling **178**(3-4): 389-397.

Paerl, H. W., Gardner, W. S., McCarthy, M. J., Peierls, B. L. and Wilhelm, S. W. (2014). Algal blooms: noteworthy nitrogen. Science **346**(6206): 175.

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

Qi, C., Fourie, A. and Chen, Q. (2018). Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. Construction and Building Materials **159**: 473-478.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. Neural Netw **12**(1): 145-151.

Recknagel, F., Bobbin, J., Whigham, P. and Wilson, H. (2002). Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. Journal of Hydroinformatics **4**(2): 125-133.

Recknagel, F., French, M., Harkonen, P. and Yabunaka, K.-I. (1997). Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling **96**(1-3): 11-

813 28.

814 Richlen, M. L., Morton, S. L., Jamali, E. A., Rajan, A. and Anderson, D. M. (2010). The
815 catastrophic 2008–2009 red tide in the Arabian Gulf region, with observations on the
816 identification and phylogeny of the fish-killing dinoflagellate Cochlodinium
817 polykrikoides. Harmful Algae **9**(2): 163-172.

818 Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by
819 error propagation (No. ICS-8506). California University San Diego La Jolla Institute for
820 Cognitive Science.

821 Segura, A., Piccini, C., Nogueira, L., Alcántara, I., Calliari, D. and Kruk, C. (2017). Increased
822 sampled volume improves Microcystis aeruginosa complex (MAC) colonies detection
823 and prediction using Random Forests. Ecological Indicators **79**: 347-354.

824 Selman, M., Greenhalgh, S., Diaz, R. and Sugg, Z. (2008). Eutrophication and hypoxia in
825 coastal areas: a global assessment of the state of knowledge. World Resources Institute
826 **284**: 1-6.

827 Sivapragasam, C., Muttil, N., Muthukumar, S. and Arun, V. (2010). Prediction of algal blooms
828 using genetic programming. Marine Pollution Bulletin **60**(10): 1849-1855.

829 Solanki, V. R., Hussain, M. M. and Raja, S. S. (2010). Water quality assessment of Lake Pandu
830 Bodhan, Andhra Pradesh State, India. Environmental Monitoring and Assessment **163**(1-
831 4): 411-419.

832 Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization
833 and momentum in deep learning. In International Conference on Machine Learning (pp.

1139-1147).

Tian, W., Liao, Z. and Zhang, J. (2017). An optimization of artificial neural network model for predicting chlorophyll dynamics. Ecological Modelling **364**: 42-52.

Wei, B., Sugiura, N. and Maekawa, T. (2001). Use of artificial neural network in the prediction of algal blooms. Water Research **35**(8): 2022-2028.

Xie, Z., Lou, I., Ung, W. K. and Mok, K. M. (2012). Freshwater algal bloom prediction by support vector machine in macau storage reservoirs. Mathematical Problems in Engineering **2012**.

Xu, F.-l., Lam, K., Dawson, R., Tao, S. and Chen, Y. (2004b). Long-term temporal-spatial dynamics of marine coastal water quality in the Tolo Harbor, Hong Kong, China. Journal of Environmental Sciences **16**(1): 161-166.

Xu, F., Lam, K., Zhao, Z., Zhan, W., Chen, Y. D. and Tao, S. (2004a). Marine coastal ecosystem health assessment: a case study of the Tolo Harbour, Hong Kong, China. Ecological Modelling **173**(4): 355-370.

Xu, J., Yin, K., Liu, H., Lee, J. H., Anderson, D. M., Ho, A. Y. T. and 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.

Xu, Z. and 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): 4171.

Yang, X.-e., Wu, X., Hao, H.-l. and He, Z.-l. (2008). Mechanisms and assessment of water

855 eutrophication. Journal of Zhejiang University Science B **9**(3): 197-209.
856 Yang, Q., Liu, G., Hao, Y., Zhang, L., Giannetti, B. F., Wang, J., & Casazza, M. (2019). Donor-
857 side evaluation of coastal and marine ecosystem services. Water Research, 166, 115028.
858 Yu, R.-C., Lü, S.-H. and Liang, Y.-B. (2018). Harmful algal blooms in the coastal waters of
859 China. Global Ecology and Oceanography of Harmful Algal Blooms (pp.309-316)
860 Zeng, Q., Liu, Y., Zhao, H., Sun, M. and Li, X. (2017). Comparison of models for predicting
861 the changes in phytoplankton community composition in the receiving water system of an
862 inter-basin water transfer project. Environmental Pollution **223**: 676-684.
863