Water Research 175 (2020) 115639

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Predicting cyanobacteria occurrence using climatological and environmental controls

Seungbeom Kim [a, b], Seokhyeon Kim [a, *], Rajeshwar Mehrotra [a], Ashish Sharma [a]

[a] School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, 2052, Australia
[b] K-water, Daejeon, 34350, Republic of Korea

ARTICLE INFO

Article history:
Received 5 August 2019
Received in revised form
14 February 2020
Accepted 20 February 2020
Available online 27 February 2020

Keywords:
Cyanobacterial bloom
River
Temperature
Total phosphorus
Velocity
Weighted function model

ABSTRACT

The occurrence of algal bloom results in deterioration of water quality, undesirable sights, tastes and odors, and the possibility of infections to humans and fatalities to livestock, wildlife and pets. Earlier studies have identified a range of factors including water temperature, flow, and nutrient concentrations that could affect cyanobacterial proliferation. Lack of enough data, independence in data across multiple sampling time steps, as well as the presence of more than one causative factors, each with different levels of influence on the response, has resulted in limited progress in the development of generalized prediction frameworks for cyanobacteria. In this study, a prediction model for cyanobacteria occurrences was developed using only three dominant environmental variables; water temperature, velocity and phosphorus concentration. These environmental variables were selected due to not only direct or joint contribution to algal bloom but also the ease of their availability either through direct measurements or as modelled responses in the river location of interest. In order to apply bacterial growth dynamic to the model, weight functions which quantify the importance assigned to the three variables depending on the cell number at the preceding time, were formulated. An extensive dataset spanning from 2013 to 2018 at 16 representative locations across the four major rivers in South Korea was used to develop and validate the model. Through cross-validation, this model was shown to have more than 75% forecasting accuracy despite the use of a relatively simple predictive algorithm. As the developed model makes use of commonly available environmental variables, it can easily be extended to locations across the country where very limited or no prior information about cyanobacteria bloom is available.

1. Introduction

The occurrence of cyanobacterial bloom is becoming a common water-related predicament (Brooks et al., 2016; Hallegraeff, 1993; Paerl and Huisman, 2008). Along with its unsightly color and unpleasant odor, cyanobacteria deteriorates water-quality and creates toxins that can harm humans and animals alike (Fleming et al., 2002). It also synthesizes added difficulty for water managers in deciding how much dam releases should occur in times of an active bloom event downstream. (Gilroy et al., 2000). An example of this comes from South Korea, where almost every summer, fast spreading algal blooms on major rivers have been recognized as a major concern for public health and drinking water (Cha et al., 2017; Kim, 2012; Park et al., 2017). Another example is the recent mass fish killing along the 40 km stretch of Darling river in Australia, offering an illustration of how unprecedented growth of cyanobacteria can impact our riverine ecosystems (Carman and Tomevska, 2019).

The increasing frequencies of cyanobacterial bloom occurrences across the world have prompted considerable research in identifying factors that may be responsible for their occurrences and proliferation. Richardson et al. (2018) showed the influences of three factors such as temperature, retention time and total phosphorus on the algal bloom in wide range of lakes in Europe. These factors can be broadly grouped into climatological, nutrient-induced and hydrological categories (Fornarelli et al., 2013; Paerl and Otten, 2013). It has been noted that some cyanobacteria need temperatures over 20 °C due to their competitiveness with eukaryotic phytoplankton. Moreover, above 25 °C, they are recorded to have a dominant position in comparison with diatom (Berg and Sutula, 2015; Davis et al., 2009). Additionally, Reynolds and Walsby (1975) suggested that a temperature range of 25–35 °C results in favorable conditions for cyanobacteria onset. Similarly,

* Corresponding author.
E-mail addresses: gogo1380@kwater.or.kr (S. Kim), seokhyeon.kim@unsw.edu.au (S. Kim), raj.mehrotra@unsw.edu.au (R. Mehrotra), a.sharma@unsw.edu.au (A. Sharma).

https://doi.org/10.1016/j.watres.2020.115639
0043-1354/

retention time or flow-velocity is also known to influence cyanobacteria occurrences. Low discharge or velocities hinder circulation and could trigger water stratification thereby creating a favorable environment for cyanobacteria and the buoyancy conditions needed for their growth (Beard et al., 1999; Brookes et al., 1999; Huber et al., 2012; Oliver, 1994). Finally, sustained nutrient influx through salinity, nitrogen and phosphorus are necessary for cyanobacterial growth. Phosphorus is known as a limiting nutrient and widely affects the total phytoplankton biomass in freshwater ecosystems (Elser et al., 1990; Schindler, 1974; Sommer, 1989; Sterner, 1994). In freshwater lake systems, phytoplankton biomass has been noted to be strongly dependent upon total phosphorus with a weaker relationship with total nitrogen (Guildford and Hecky, 2000). However, recent literature has revealed the impact of nitrogen on cyanobacterial growth in lakes (Gobler et al., 2016; Paerl et al., 2011). Conley et al. (2009) presented that regulating only phosphorus along with allowing nitrogen can decrease algal bloom in freshwater but cause eutrophication in estuaries. High phosphorus is generally considered as the main source of cyanobacteria in South Korea (Kim and Kang, 1993; Lee et al., 1998). Some studies (Schindler et al. 2008, 2012) have suggested that the presence of nitrogen-fixing cyanobacteria in water bodies control the concentration of nitrogen in water and constitute nitrogen as a less effective contributor of the growth of cyanobacteria.

Although much attention has been given to understand what drives algal bloom based on lots of specific studies, the development of simplified and generalizable forecasting models based on easily accessible variables has not been fully explored. Even under similar environmental conditions, the growth rate of cyanobacteria could differ depending on the existing levels of these conditions. According to the typical bacterial growth curve presented by Tortora et al. (2004), there exist four phases of growth: lag (slow or lack of growth); exponential (doubling cells); stationary (balance between growth and death of cells) and; death (net loss by death rate exceeding growth rate) phases.

Recently, forecasting models for algal bloom occurrences and counts have been developed based on Artificial Neural Networks (ANN) (Guzel, 2019; Sen et al., 2018; Srisuksomwong and Pekkoh, 2019). Such models can be susceptible to mimicking the training data characteristics resulting in over-fitting and hence producing inferior results with new data. This is especially the case when data length is limited as model complexity can be high with an ANN type approach. The physical models for water quality analysis such as CE QUAL W2 and Environmental Fluid Dynamics Code (EFDC) also require considerable parameter inputs or calibration data for simulating dynamics with adequacy (Gao and Li, 2014). McGillicuddy (2010) presents a nice overview of the strengths and limitations of commonly available conceptual to aggregated black-box models to predict harmful algal bloom. Even though plenty of models have been developed and proposed, a majority of them are physically based and site-specific and are not easy to be generalized.

Our study aims to develop an approach for predicting the onset and end of cyanobacteria occurrence as a function of environmental variables as well as pre-existing concentration as an additional indicator. The bloom occurrence here is defined based on cyanobacteria cell count exceeding a specified threshold. The choice of a binary predictive model is also based on operational considerations, as a confident prediction can help trigger alerts and activate damage control measures by the water resources agencies in charge. Here, simplicity and computational effectiveness of the model are achieved by adopting simple algorithms based on easily accessible inputs and bacterial growth models. We verify the developed model over multiple representative locations across four major rivers in South Korea.

This paper is structured as follows. Section 2 briefly presents the relevant datasets used and explains the proposed predictive modelling framework. In Section 3, results for the proposed model through cross-validation are presented. Finally, conclusions are presented in Section 4.

2. Data and method

2.1. Study area and data

Our study focuses on the major river system in South Korea and uses long-term, comprehensive, high-quality data across the river system that has been meticulously collected for multiple years by the Ministry of Environment, South Korea (South Korea Ministry of Environment, 2009). As presented in Fig. 1(a), the study region covers the southern part of the Korean Peninsula, latitude: 33°N - 39°N and longitude: 124°E – 130°E. The area has a temperate climate with four distinct seasons as a part of the East Asian monsoonal region (Savada and Shaw, 1997). There are four major rivers in South Korea; Han, Nakdong, Geum, and Yeongsan, all of which represent major water sources for most of the population and industries. The total basin area of these rivers is about 63,016 km^2 which occupies almost 63% of South Korea (South Korea Ministry of Environment, 2003).

Six-year weekly water quality data from January 2013 to December 2018 are used in this study. These datasets are mostly sourced from the Water Environment Information System (http://water.nier.go.kr/) operated by Ministry of Environment of South Korea (South Korea Ministry of Environment, 2012). The datasets include weekly measurements of water temperature (T, °C), total phosphorus (P, mg/L), Number of cyanobacteria cells (C, total number of cells/mL) and discharge rate (m^3/sec) from 16 stations along the four major rivers. The frequency of routine observation of the data is once a week but twice a week for serious algal bloom events. The cyanobacterial cell numbers in the above datasets are collected as the sum of cell counts for harmful cyanobacteria species such as *Microcystis, Anabaena, Aphanizomenon* and *Oscillatoria*. Those four species are also classified as cyanobacterial toxins affecting the liver or nervous system (World Health Organization, 2001). The 16 stations presented in Fig. 1(a) have been operating since 2012 for monitoring algal bloom and extracting water quality samples. One of the reasons for continuous measurement of this data is due to repeating occurrences of cyanobacteria every summer in the country especially in recent years as temperatures rise across the world (Srivastava et al., 2015).

As there are no velocity measurements in the datasets, the flow velocity was indirectly obtained by using the HEC-RAS model (Hydrologic Engineering Center - River Analysis System). The HEC-RAS model makes use of cross-section, roughness and discharge information to derive average cross-sectional velocity at defined measurement points (Brunner, 1995). The cross-section and roughness data were obtained from the national river basic management plan of 2015 (Han River Flood Control Office, 2018), which has been published for the rivers every five years by the South Korea government. Daily discharges of every measuring point were derived from the database, *MyWater*, operated by K-water of South Korea (https://www.water.or.kr/).

2.2. Methodology

2.2.1. Climatological and environmental factors governing cyanobacteria growth

As mentioned earlier, it is difficult to attribute the occurrence of cyanobacteria to one specific factor as there exist combinations of factors that trigger and sustain cyanobacteria growth. Based on a

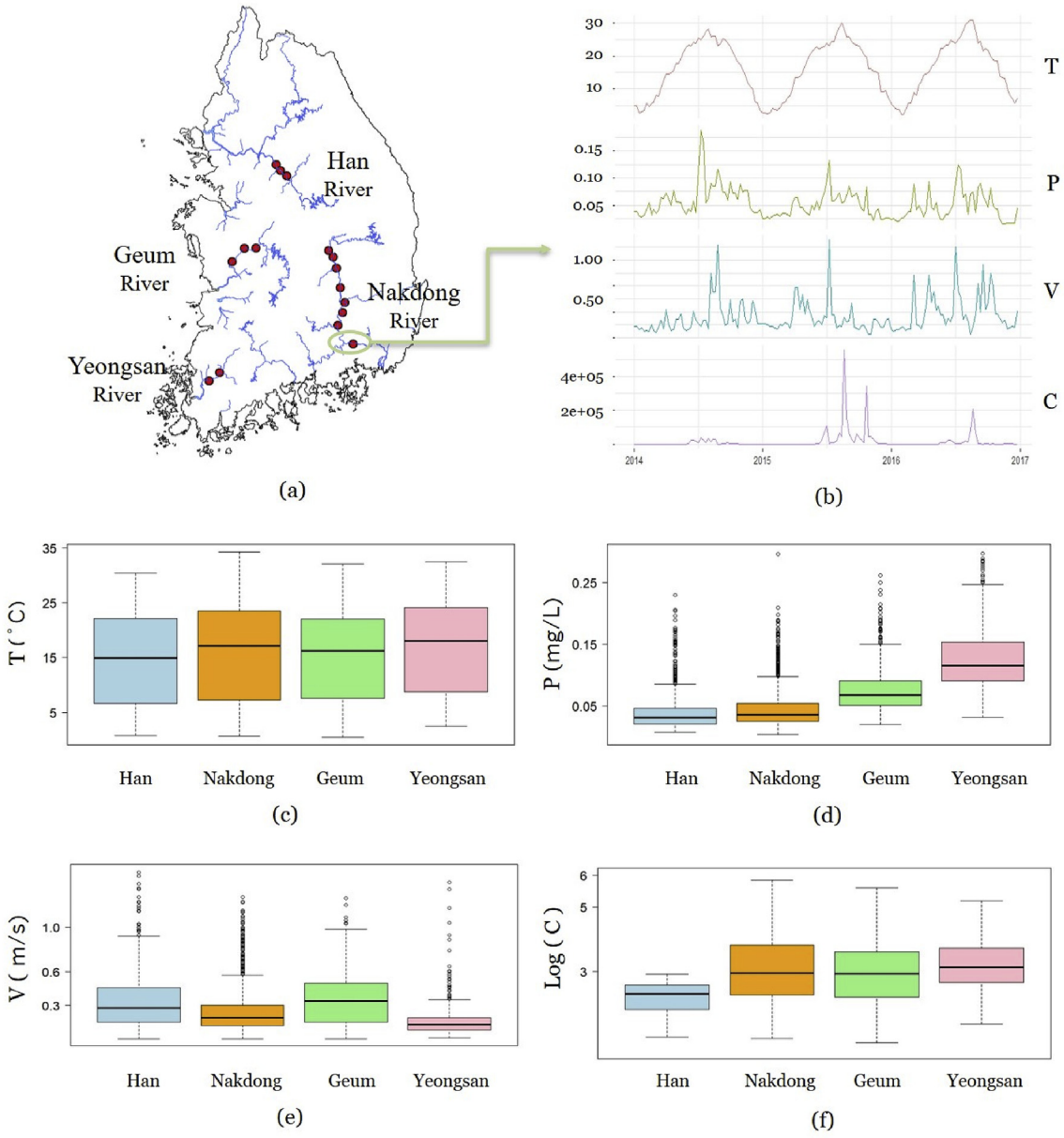

Fig. 1. Overview of the study area. (a) Locations of the four major rivers (Han, Nakdong, Geum and Yeongsan) in South Korea and the 16 data measurement stations used for this study, (b) time series of *T*, *P*, *V* and cyanobacteria at an example location (Haman station in Nakdong river). (c)–(f) Box plots of weekly *T*, *P*, *V* and log (*C*) of the four major rivers obtained from the 16 stations. T = water temperature (°C); P = total phosphorus (mg/L); V = flow velocity (m/s); C = number of cyanobacterial cells/mL.

review of pertinent literature, in this study, three factors; water temperature (T), total phosphorus (P) and flow velocity (V), were considered as the dominant environmental variables leading to the occurrence and growth of cyanobacteria. Although combination of excess nitrogen and phosphorus can worsen cyanobacterial concentration in rivers, phosphorus is considered as the main cause of cyanobacteria in South Korea (Kim and Kang, 1993; Lee et al., 1998) and surplus nitrogen is known to have limited influence on the cyanobacterial formulation (Schindler et al. 2008, 2012). Hence, in order to place more emphasis on the single nutrient, this study focuses on P as a representative nutrient factor.

As shown in the time series of T, P, V and C at one of the representative stations in Nakdong River in Fig. 1(b), definite patterns in the environmental variables and cyanobacteria growth are noticeable. T and P are relatively higher at the time of cyanobacteria events whereas V is lower. In addition to this, it is quite clear from Fig. 1(c)–(f) that these variables have positive/negative correlations with cyanobacteria count. Notwithstanding these observations, it is not straightforward to derive a direct relationship between the cyanobacteria and the environmental variables due to low Pearson correlation coefficients (0.21, 0.03 and −0.06 for T, P and V respectively) and their joint dependence. Realizing this, we refrain from a prediction of numeric cell counts and focus on developing a binary prediction model for cyanobacteria occurrence (or non-occurrence).

The predictive model for the cyanobacterial bloom occurrence can be stated as:

$$[\mathrm{O}|(T,P,V),\boldsymbol{\theta}] = \left\{ \begin{matrix} 0 \\ 1 \end{matrix} \right\} \tag{1}$$

where the binary response variable (i.e. O) is defined as,

$$O = \begin{cases} 1 & \text{if } C \geq C^* \\ 0 & \text{else,} \end{cases}$$

C^* representing a pre-specified threshold of cyanobacteria cell concentrations C (to account for the uncertainty in the cell count measurements), and is adopted equal to 1000 cells/mL following the South Korean Government's algal alert system standards (South Korea Ministry of Environment, 2009; Srivastava et al., 2015). Additionally, (T, P, V) represent temperature (°C), total phosphorus concentration (mg/L) and flow velocity (m/s) respectively, with the vector $\boldsymbol{\theta}$ denoting model parameters to be ascertained.

As the response for the model in Equation (1) is binary, various formulations of the model can be explored. A commonly used form involves specifying Equation (1) through a logistic regression relationship (Kleinbaum et al., 2002) using continually varying representations for predictors (T,P,V). An even simpler model uses binary predictors defined based on thresholds that become the parameters $\boldsymbol{\theta}$ to be estimated. Our proposed model predicts the occurrence of cyanobacteria if T, P and V values are simultaneously above or below the critical thresholds, defined as $\boldsymbol{\theta} = [\tilde{\theta}^T, \tilde{\theta}^P, \tilde{\theta}^V]$. In other words, if at a given time step T and P are higher than the thresholds $\tilde{\theta}^T$, $\tilde{\theta}^P$, respectively, and V is lower than the threshold, $\tilde{\theta}^V$, then the model predicts the occurrence of cyanobacteria at that time step. However, during our study we found that these critical thresholds are not directly applicable because of the persistence characteristic of cyanobacterial bloom. The cyanobacterial bloom tends to persist for some time after it reaches the exponential phase of its typical growth curve even under not-so favorable environmental conditions (Tortora et al., 2004). Following this, we considered that these thresholds are not static but a function of the recent past concentration of cyanobacteria. For this, we propose the following formulation to update the thresholds of T, P and V conditional on the cyanobacteria cell count at the preceding time step:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_t^T \\ \theta_t^P \\ \theta_t^V \end{bmatrix} = \begin{bmatrix} \tilde{\theta}^T \tilde{\theta}^P \tilde{\theta}^V \end{bmatrix} \text{if } C_{t-1} = 0$$

$$= \begin{bmatrix} \tilde{\theta}^T - s^T f(C_{t-1})\tilde{\theta}^P - s^P f(C_{t-1})\tilde{\theta}^V + s^V f(C_{t-1}) \end{bmatrix} \text{else,} \quad (2)$$

where the vector, $[\theta_t^T \quad \theta_t^P \quad \theta_t^V]'$, denotes the updated thresholds of T, P and V at the given time step (t) conditioned on the cyanobacteria cell count at the preceding time step (i.e. C_{t-1}). s^T, s^P and s^V denote the sample standard deviations of T, P and V, respectively, and $f(C_{t-1})$ is a common scaling factor (positive real number) estimated by a weight function using cyanobacteria concentration at previous time step, C_{t-1}. It should be noted that the $C^* = 1000$ in Equation (1) is to define the occurrence of a cyanobacteria bloom operationally, which should be altered by users as per local guidelines that may be in place. Also, as implied in Equation (2), only non-zero values of C_{t-1} are used to modify the dynamic thresholds for T, P and V. These modified thresholds then decide whether a bloom has occurred or not for the time step in question. With these modified threshold values (lower $(-)$ for T and P; higher $(+)$ for V), it is possible to incorporate the persistence nature of cyanobacterial bloom. In this study, three forms of weight function were tested and compared. These are:

$$\text{Sigmoid: } f(C_{t-1}) = \frac{\alpha}{1 + e^{-\beta(C_{t-1} - \gamma)}} \quad (3)$$

$$\text{Linear: } f(C_{t-1}) = \begin{cases} \alpha C_{t-1} & \text{if } C_{t-1} < \beta \\ \alpha\beta & \text{if } C_{t-1} \geq \beta \end{cases} \quad (4)$$

$$\text{Exponential: } f(C_{t-1}) = \alpha\left(1 - e^{-\beta C_{t-1}}\right) \quad (5)$$

Here, α, β and γ represent model parameters, to be ascertained along with the three thresholds $(\tilde{\theta}^T, \tilde{\theta}^P, \tilde{\theta}^V)$.

The above process is illustrated in Fig. 2(b)–(e) using cyanobacteria occurrence predictions at Gongju station in Geum river, as an example. The second panel in the figure presents the time series of recorded cyanobacteria cell numbers (C) with solid line and threshold cell number ($C^* = 1000$) with dotted line. In the 3rd to 5th panels from the top, $\boldsymbol{\theta}$ is determined, for example of temperature, either $\tilde{\theta}^T$ or $\tilde{\theta}^T - s^T f(C_{t-1})$ using Equation (2) based on cyanobacteria cell counts of preceding time steps. The environmental variable values at each time step are compared with the thresholds θ_t^T, θ_t^P and θ_t^V to see whether there is a cyanobacteria occurrence or not. Consider the first case in May 2014 where a cyanobacteria occurrence is noted. Here, a triggering condition did not occur as T_t is lower than θ_t^T even though P_t and V_t meet the specified criteria. Accordingly, this prediction is regarded as a correct rejection. Next, there was a 2-week long cyanobacteria occurrence in July 2014 and the triggering conditions were correctly predicted (i.e. hit). Lastly, the cyanobacteria occurrence disappeared in August 2014 mainly resulting from increased V_t at that time and was also correctly predicted by the model (i.e. correct rejection).

2.2.2. Model calibration using maximum likelihood

As shown earlier, the proposed model first adjusts the threshold $\boldsymbol{\theta}$ values at the current time step as θ_t^T, θ_t^P and θ_t^V on the basis of cyanobacteria cell count at the preceding time step and thereafter compares them with the observed environmental variables T_t, P_t and V_t. A 'Yes' or 'No' alert is predicted depending upon whether these values are below/above the thresholds.

A maximum likelihood estimate of the model parameters $\boldsymbol{\theta} = (\tilde{\theta}^T, \tilde{\theta}^P, \tilde{\theta}^V)$ and function parameters (α, β, γ) given n observations can be obtained as;

$$L\left(\tilde{\theta}^T, \tilde{\theta}^P, \tilde{\theta}^V, \alpha, \beta, \gamma; Data\right) = \sum_{i=1}^{n} \{O_i \log(p) + (1 - O_i)\log(1 - p)\} \quad (6)$$

where p represents the probability of joint exceedance of the adopted thresholds, or:

$p = Prob[T \geq \theta^T, P \geq \theta^P, V \leq \theta^V]$, and L () the log-likelihood.

As the modelled outcomes are binary, the above maximum likelihood estimate can be expressed using the form of a 2 × 2 contingency table by considering the four scalar attributes, hits, false alarms, misses and correct rejections (Wilks, 2011) as presented in Fig. 2(a). The maximum likelihood estimation of the model parameters would be analogous to that obtained through maximization of the probability of correct hits and rejections, referred to as "proportion correct" or PC:

(a)

2×2 contingency analysis		Observation	
		Yes $(C_t > C^*)$	No $(C_t \le C^*)$
Prediction	Yes $T_t \ge \Theta^T_t$ & $P_t \ge \Theta^P_t$ & $V_t \le \Theta^V_t$	Hits (a)	False alarms (b)
	No (else)	Misses (c)	Correct rejections (d)

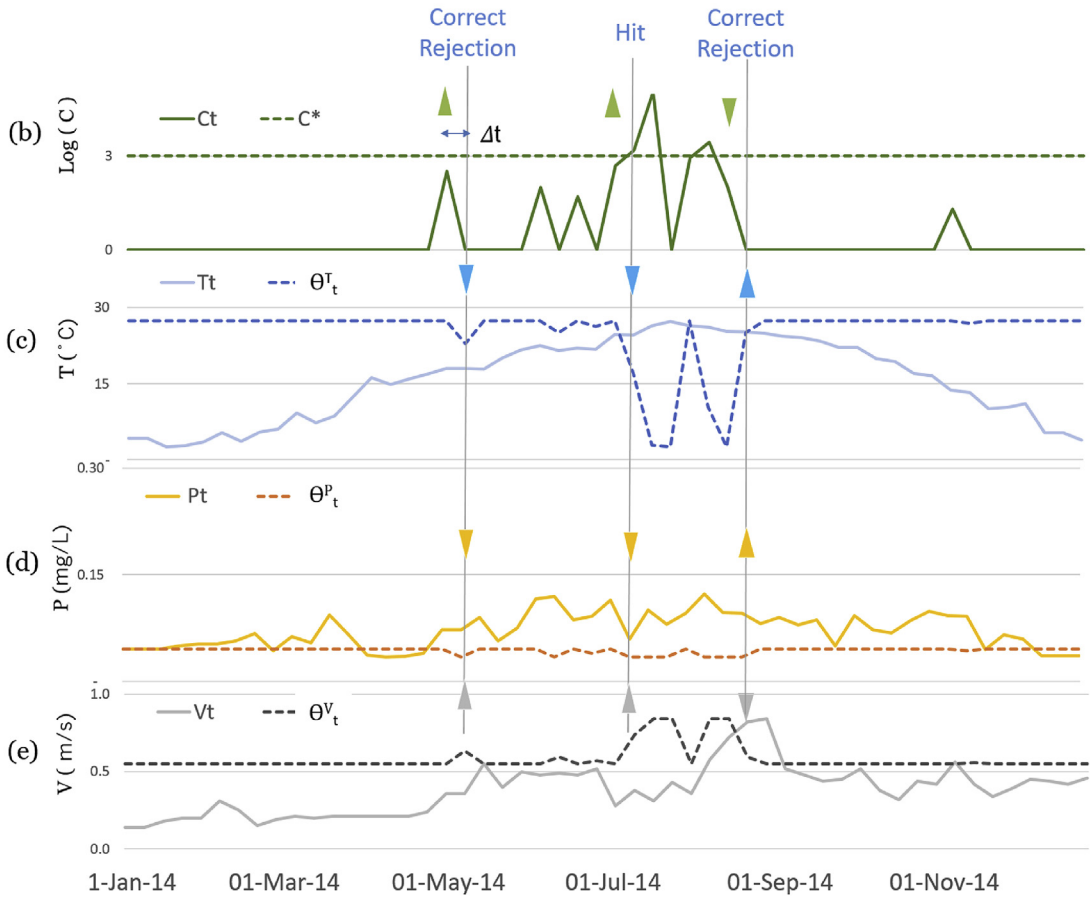

Fig. 2. Example showing observed and modelled occurrences of cyanobacteria at Gongju station in Geum river. (a) Components of contingency table applied, and time series of (b) C = number of cyanobacterial cells/mL except for 0 (▲increase, ▼decrease), (c) T = water temperature (°C), (d) P = total phosphorus (mg/L) and (e) V = water velocity (m/s).

$$PC = \frac{a + d}{a + b + c + d} \tag{7}$$

where a = hits, b = false alarms, c = misses and d = correct rejections. The difficulty in using the maximum likelihood formulation above lies in the fact that the proportion of non-occurrences is markedly higher than the proportion of occurrences. As a result, the above maximization ends up maximizing 'd' as it dominates the PC score. This can result in parameters that will correctly predict the non-occurrences but not the occurrences that are of greater importance.

2.2.3. Modified parameter estimation approach

In order to overcome the limitation of parameter estimation approach outlined above and to enforce a greater importance to actual occurrences, various modifications to the Maximum Likelihood estimator in Equation (7) were investigated. After assessing alternate performance measures, the Peirce Skill Score (PSS, Equation (8)) was selected as the objective function for calibrating the parameters. PSS can be calculated as the difference between the hit rate (H = a/(a+c)) and the false alarm rate (F = b/(b + d)); that is, PSS = H − F:

$$PSS = \frac{ad - bc}{(a + c)(b + d)} \tag{8}$$

Using PSS as the basis for parameter estimation, perfect forecasts provide a score of one (as b = c = 0 or alternatively, H = 1 and F = 0), while random forecasts give negative scores. The advantage PSS offers over the Maximum Likelihood estimator in Equation (7) is that the model is not discouraged from forecasting rare occurrences (Wilks, 2011; Kim et al., 2018), which is highly relevant to the cyanobacteria prediction problem at hand.

There are two parameters and three critical thresholds of environmental variables to be ascertained depending on the weight functions used in Equations (3)–(5). For γ in the sigmoid model presented in Equation (3), the median value of cyanobacteria cell numbers in the training datasets was used for simplification. In order to identify these parameters and critical thresholds, the Shuffled Complex Evolution (SCE) optimization approach from R package *Soilhyp* was applied. SCE has been widely used for a broad class of hydrologic and environmental optimization problems (Duan et al., 1993; Vrugt et al., 2003).

2.2.4. Assessment of model sensitivity and cross-validation

Besides PSS, the threat score (TS) (Equation (9)) was also ascertained on the basis of PSS optimized parameters for an independent assessment of the applicability of the three weight functions as well as the sensitivity of each variable. Wilks (2011) suggested that TS is particularly useful when the event to be forecasted occurs less frequently than its nonoccurrence and is computed as;

$$\mathrm{TS} = \frac{a}{a+b+c} \tag{9}$$

In addition to the original TS, four conditional TS scores were also evaluated. The first (TS_1) calculates TS for events which do not have a preceding cyanobacterial event. In other words, it judges the model's forecasting accuracy at the time of abrupt cyanobacterial occurrence. The other three scores (TS_2 to TS_4) measure TS for the cases where temperature and phosphorus are less than their median values ($T_t < T_{50\%}$ and $P_t < P_{50\%}$), and velocity is greater than the median velocity ($V_t > V_{50\%}$). These conditional TS scores can be interpreted as sensitivity estimates of the model performance to exceedances of each environmental variable above (or below) its median.

Entire dataset was first used to get a stable estimate of model parameters. Following this, further checks on these calibrated variables and their applicability were conducted using a 4-fold cross validation wherein 12 out of 16 stations were randomly selected as training datasets and the remaining four stations were used as test datasets. This validation process was repeated 50 times. The calibrated parameters and calculated skill scores are presented in the next section.

3. Results and discussion

3.1. Calibration results

Using the entire dataset, the PSS, parameters and critical thresholds for each environmental variable were ascertained. The PSS for the three weight functions range from 0.79 to 0.80 (Table 1). The critical thresholds of environmental variables are also within physically acceptable ranges. For instance, $\tilde{\theta}^T$ is in line with the favorable temperature range of 25–35 °C for cyanobacteria occurrences suggested by Reynolds and Walsby (1975). Accordingly, this threshold can explain non-occurrence of cyanobacteria in Han river in Fig. 1(c). These results also support the finding of Huang et al. (2008) that the change of the flow velocity in a range of less than 0.4 m/s would accelerate the growth of algae and the occurrence of bloom in river-type reservoir. Similarly, the critical velocity about 0.05 m/s which suppress the development of stratification suggested by Mitrovic et al. (2003) can be included in this range of velocity.

Table 1
Model calibration results for three weight functions.

Function Type	PSS	$\tilde{\theta}^T$ (°C)	$\tilde{\theta}^P$ (mg/L)	$\tilde{\theta}^V$ (m/s)
Sigmoid	0.79	26.97	0.04	0.63
Linear	0.79	26.40	0.04	0.55
Exponential	0.80	27.75	0.03	0.61

3.2. Cross-validation results

In order to verify the calibration result and assess the future applicability of the model, the 4-fold cross-validation was implemented and results are presented in Fig. 3.

Fig. 3(a) presents box plots of the distribution of PSS for the three weight functions. It is clear from the figure that the all PSS and threshold values are similar to those obtained from the calibration results (Table 1). The linear weight function provides the most robust range of the environmental variables (Fig. 3(b)–(d)) with relatively narrower interquartile ranges.

With the conditional TS results which were implemented for assessing the model reliability in infrequent situations, all three weight functions predicted over 40% accuracy even under unusual events described in Fig. 3(e). The exponential model generally shows higher TSs than others except for TS_1 which is slightly lower than those of the sigmoid model. Therefore, it is recommended to use the linear or exponential weight function for the cyanobacteria prediction while dealing with the similar environmental conditions.

Furthermore, it was apparent from the averaged values of conditional TS_2 to TS_4 (Fig. 3(e)) that the most dominant factor for cyanobacteria occurrence in the study area was temperature (TS_2) followed by velocity (TS_4) and phosphorus (TS_3). This is because the variable showing high dependence on cyanobacteria even in relatively low apparent conditions can be interpreted as having a lower correlation with cyanobacteria. In detail, TS_2 dataset includes the cases where temperature is less than the median value of the temperature (temperature being the most dominant variable, resulting TS_2 is much smaller than normal TS). Contrasting to TS_2, TS_3 includes the cases where total phosphorus is less than the median value of total phosphorus (phosphorus being the least dominant variable, resulting in TS_3 is slightly smaller than the normal TS). TS_4 results fall in between these two categories. Following these results, the gap between the normal TS and conditional TSs can be interpreted as the relative contribution of each environmental variable towards cyanobacteria predictive model formulation.

3.3. Maximum likelihood results

The usefulness of adopting a new basis for parameter estimation in PSS, and not using the maximum likelihood-based estimate which would have amounted to maximizing PC (Equation (8)) was also assessed through the cross-validation described in Section 3.2.

Based on this, the maximized PC was found to be as high as 0.91 ± 0.02 and the normal TS (Equation (9)) was almost identical to the PSS-derived value reported in the previous sub-section. However, the TS_2 to TS_4 conditioned by relatively implausible *T*, *P* and *V* conditions for cyanobacteria occurrences (Section 2.2.4) were found to be slightly lower than those from the PSS based cross-validation results. Furthermore, TS_1 was unquestionably lower than the PSS-derived result, which were 0.13 ± 0.08 for the PC case and 0.69 ± 0.02 for the PSS case (Fig. 3(e)). Namely, the model performance when calibrated by maximizing PC tends to sharply drop for cases that do not have a preceding cyanobacteria event. Accordingly, the use of a metric such as the PSS is recommended for the specification of rare occurrence variables such as cyanobacteria.

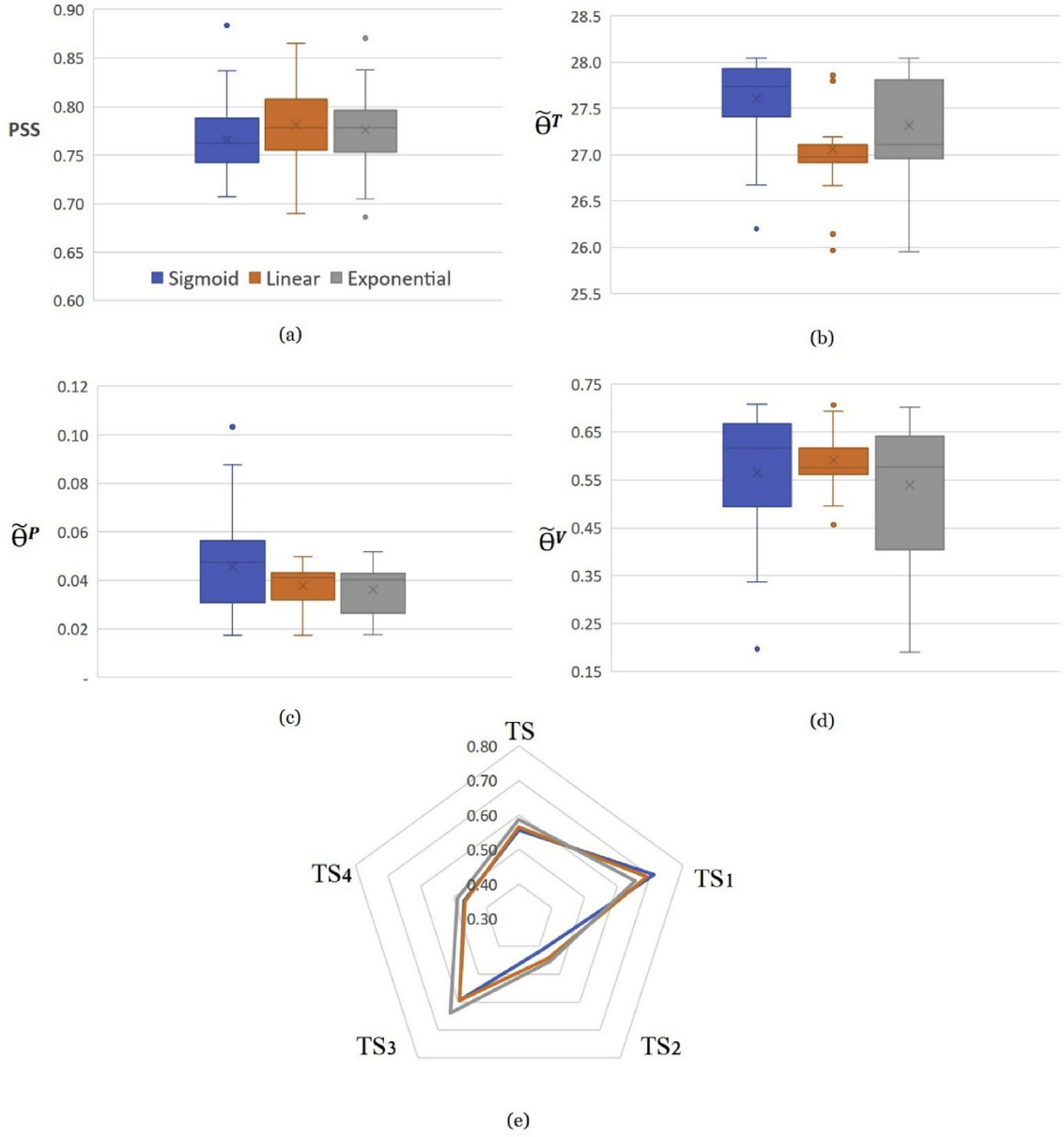

Fig. 3. Cross validation results by three weight functions. Box plots of (a) PSS and (b)–(d) thresholds of three environmental variables (i.e. water temperature, total phosphorus and water velocity), and (e) radar chart for average values of conditional TSs.

3.4. Caveats and follow-up studies

This study focused on the three dominant environmental variables of cyanobacterial growth; water temperature (*T*), total phosphorus (*P*) and flow velocity (*V*). Although the stable range of each predictor threshold in Fig. 3 suggest that climatological conditions over study area are not significantly different to each other, there could be additional environmental and climatological conditions, that are important in cyanobacterial growth depending on the location being modelled. Similarly, notwithstanding an obvious strong relationship of cyanobacteria with the three variables used in the study, other factors are also known to affect the appearance of cyanobacteria, such as pH, nitrogen, salinity, turbidity, irradiance, electric conductivity, etc. Furthermore, results of conditional TSs indicate that exceptional changes in one of the variables can influence algal bloom occurrence. In this study, we assumed phosphorus as an indicator of the nutrients in the water. In order to verify if nitrogen is a better choice than phosphorus, PSS and TS were estimated first using total nitrogen (*N*) instead of total phosphorus (*P*) and both variables together, thereafter. The model forecasting accuracy was found to be slightly reduced (statistically insignificant) in comparison to the base case. It can thus be interpreted that total phosphorus is marginally a better choice than total nitrogen or the combination together. Otherwise also, its relative impact on cyanobacterial growth is much smaller than water temperature and flow velocity (*V*). Likewise, if all relevant factors are available, it would be meaningful to rank their relative contributions, which is a limitation of most data collection systems as measurements are often not complete.

Based on the data type published in South Korea and the cyanobacteria toxic classification from World Health Organization, the sum of toxic cyanobacteria species including *Microcystis, Anabaena,*

Aphanizomenon and *Oscillatoria* was classified as cyanobacteria in this study. However, it should be recognized that every cyanobacteria species acts somewhat differently and favors different environmental conditions.

A follow-up to this study would be to develop a probabilistic forecasting model that provides cyanobacteria predictions in a probabilistic manner thereby allowing water managers to make decisions based on the tolerable risks allowed in the river system of interest. Furthermore, expansion of the model to suppress or delay the dominance of cyanobacteria, that is, to reduce the probability of occurrence of the events, could be another way forward. In situations where there is no or insufficient data, it would be valuable to utilize remote sensing data of the three environmental factors as a surrogate (O'Reilly et al., 2015). The United States' National Oceanic and Atmospheric Administration (NOAA) has developed the National Phytoplankton Monitoring Program, which was first developed in coastal areas for monitoring marine harmful algal blooms (HABs). More recently, this program which used satellite imagery to identify surface cyanobacteria events has begun to concentrate on freshwater HABs (Brooks et al., 2016). As another example to detect early algal bloom by satellite image, Teta et al. (2017) showed that remote sensing data combined with aerial and in-situ data was an efficient method to detect early blooms of cyanobacteria. Investigation of these added data resources to establish a globally applicable predictive model would be of interest.

4. Conclusion

We proposed a simple binary model for predicting cyanobacteria blooms in rivers, conditioned on the three environmental variables. These variables were identified through literature review and are easily measurable at all locations.

The novel elements of this research were to propose a prediction model of cyanobacteria occurrence conditional on the preceding state of cyanobacteria concentration, and, a procedure for parameter estimation that provides equal consideration to the cases of a very few samples of cyanobacteria occurrence as compared to the non-occurrence cases. Based on the preceding time step cyanobacteria concentration, thresholds of the variables at the current time step were transformed by using three types of weight function (i.e. sigmoid, linear and exponential) in which the linear and exponential function were found to be performing better.

More than 0.75 PSS and 0.4 conditional TS scores capturing rare combinations of occurrence of the three environmental variables, gave us confidence in the accuracy of the proposed model. Furthermore, results based on conditional TS showed that the most dominant factor for cyanobacteria in this study area was temperature followed by velocity and phosphorus, providing avenues for an extension to other regions and even catchments with limited ground observations.

Declaration of competing interest

□ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank K-water (doc#: 6673) for providing sponsorship to the first author and sharing the data and information of the cyanobacteria measuring system currently in place. We are grateful to the contributors to the datasets used in this study. The weekly environmental data including water temperature, total phosphorus, water velocity and cyanobacteria cell numbers for the four major rivers in South Korea are freely available from the Water Environment Information System operated by Ministry of Environment in South Korea (http://water.nier.go.kr/).

References

Beard, S., Handley, B., Hayes, P., Walsby, A., 1999. The diversity of gas vesicle genes in Planktothrix rubescens from Lake Zürich. Microbiology 145 (10), 2757–2768.
Berg, M., Sutula, M., 2015. Factors Affecting the Growth of Cyanobacteria with Special Emphasis on the Sacramento-San Joaquin Delta. Southern California Coastal Water Research Project Technical Report 869.
Brookes, J.D., Ganf, G.G., Green, D., Whittington, J., 1999. The influence of light and nutrients on buoyancy, filament aggregation and flotation of Anabaena circinalis. J. Plankton Res. 21 (2).
Brooks, B.W., Lazorchak, J.M., Howard, M.D., Johnson, M.V.V., Morton, S.L., Perkins, D.A., Reavie, E.D., Scott, G.I., Smith, S.A., Steevens, J.A., 2016. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? Environ. Toxicol. Chem. 35 (1), 6–13.
Brunner, G.W., 1995. HEC-RAS River Analysis System. Hydraulic Reference Manual. Version 1.0. HYDROLOGIC ENGINEERING CENTER DAVIS CA.
Carman, R., Tomevska, S., 2019. A Million Fish Dead in 'distressing' Outback Algal Bloom at Menindee.
Cha, Y., Cho, K.H., Lee, H., Kang, T., Kim, J.H., 2017. The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers. Water Res. 124, 11–19.
Conley, D.J., Paerl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C., Likens, G.E., 2009. Controlling Eutrophication: Nitrogen and Phosphorus. American Association for the Advancement of Science.
Davis, T.W., Berry, D.L., Boyer, G.L., Gobler, C.J., 2009. The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of Microcystis during cyanobacteria blooms. Harmful Algae 8 (5), 715–725.
Duan, Q., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theor. Appl. 76 (3), 501–521.
Elser, J.J., Marzolf, E.R., Goldman, C.R., 1990. Phosphorus and nitrogen limitation of phytoplankton growth in the freshwaters of North America: a review and critique of experimental enrichments. Can. J. Fish. Aquat. Sci. 47 (7), 1468–1477.
Fleming, L.E., Rivero, C., Burns, J., Williams, C., Bean, J.A., Shea, K.A., Stinn, J., 2002. Blue green algal (cyanobacterial) toxins, surface drinking water, and liver cancer in Florida. Harmful Algae 1 (2), 157–168.
Fornarelli, R., Galelli, S., Castelletti, A., Antenucci, J.P., Marti, C.L., 2013. An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers. Water Resour. Res. 49 (6), 3626–3641.
Gao, L., Li, D., 2014. A review of hydrological/water-quality models. Frontiers of Agricultural Science and Engineering 1 (4), 267.
Gilroy, D.J., Kauffman, K.W., Hall, R.A., Huang, X., Chu, F.S., 2000. Assessing potential health risks from microcystin toxins in blue-green algae dietary supplements. Environ. Health Perspect. 108 (5), 435.
Gobler, C.J., Burkholder, J.M., Davis, T.W., Harke, M.J., Johengen, T., Stow, C.A., Van de Waal, D.B., 2016. The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. Harmful Algae 54, 87–97.
Guildford, S.J., Hecky, R.E., 2000. Total nitrogen, total phosphorus, and nutrient limitation in lakes and oceans: is there a common relationship? Limnol. Oceanogr. 45 (6), 1213–1223.
Guzel, H.O., 2019. Prediction of Freshwater Harmful Algal Blooms in Western Lake Erie Using Artificial Neural Network Modeling Techniques.
Hallegraeff, G.M., 1993. A review of harmful algal blooms and their apparent global increase. Phycologia 32 (2), 79–99.
Han River Flood Control Office, 2018. River Management Information System, Han River Flood Control Office. Han River Flood Control Office.
Huang, Y.l., Liu, D.f., Chen, M.X., 2008. Simulation of Algae Bloom under Different Flow Velocity, vol. 19 (10).
Huber, V., Wagner, C., Gerten, D., Adrian, R., 2012. To bloom or not to bloom: contrasting responses of cyanobacteria to recent heat waves explained by critical thresholds of abiotic drivers. Oecologia 169 (1), 245–256.
Kim, E.-H., Kang, S.-K., 1993. The effect of heavy metal ions on the growth of Microcystis aeruginosa. Journal of Korean Society on Water Environment 9 (3), 139–200.
Kim, Y., 2012. Algae Threatens Water Supply. Korea Herald, Korea.
Kim, S., Paik, K., Johnson, F., Sharma, A., 2018. Building a Flood-Warning Framework for Ungauged Locations Using Low Resolution, Open-Access Remotely Sensed Surface Soil Moisture, Precipitation, Soil, and Topographic Information. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (2), 375–387. https://doi.org/10.1109/JSTARS.2018.2790409, 17545550. In this issue.
Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M., 2002. Logistic Regression. Springer.
Lee, T.-G., Park, S.-W., Yu, T.-S., Kim, J., 1998. The growth and coagulation characteristics of Microcystis aeruginosa during water treatment processes. Journal Korea Technological Society of Water and Waste Water Treatment 6 (1).
McGillicuddy Jr., D., 2010. Models of harmful algal blooms: conceptual, empirical, and numerical approaches. J. Mar. Syst.: journal of the European Association of Marine Sciences and Techniques 83 (3–4), 105.

Mitrovic, S., Oliver, R., Rees, C., Bowling, L., Buckney, R., 2003. Critical flow velocities for the growth and dominance of Anabaena circinalis in some turbid freshwater rivers. Freshw. Biol. 48 (1), 164–174.
O'Reilly, C.M., Sharma, S., Gray, D.K., Hampton, S.E., Read, J.S., Rowley, R.J., Schneider, P., Lenters, J.D., McIntyre, P.B., Kraemer, B.M., 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys. Res. Lett. 42 (24), 10,773–710, 781.
Oliver, R.L., 1994. Floating and sinking in gas-vacuolate cyanobacteria1 30 (2), 161–173.
Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320 (5872), 57–58.
Paerl, H.W., Otten, T.G., 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb. Ecol. 65 (4), 995–1010.
Paerl, H.W., Xu, H., McCarthy, M.J., Zhu, G., Qin, B., Li, Y., Gardner, W.S., 2011. Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): the need for a dual nutrient (N & P) management strategy. Water Res. 45 (5), 1973–1983.
Park, Y., Pyo, J., Kwon, Y.S., Cha, Y., Lee, H., Kang, T., Cho, K.H., 2017. Evaluating physico-chemical influences on cyanobacterial blooms using hyperspectral images in inland water, Korea. Water Res. 126, 319–328.
Reynolds, C., Walsby, A., 1975. Water-blooms. Biological reviews 50 (4), 437–481.
Richardson, J., Miller, C., Maberly, S.C., Taylor, P., Globevnik, L., Hunter, P., Jeppesen, E., Mischke, U., Moe, S.J., Pasztaleniec, A., 2018. Effects of multiple stressors on cyanobacteria abundance vary with lake type. Global Change Biol. 24 (11), 5044–5055.
Savada, A.M., Shaw, W., 1997. South Korea: A Country Study. Diane Publishing.
Schindler, D.W., 1974. Eutrophication and recovery in experimental lakes: implications for lake management. Science 184 (4139), 897–899.
Schindler, D.W., Hecky, R., Findlay, D., Stainton, M., Parker, B., Paterson, M., Beaty, K., Lyng, M., Kasian, S., 2008. Eutrophication of lakes cannot be controlled by reducing nitrogen input: results of a 37-year whole-ecosystem experiment. Proc. Natl. Acad. Sci. Unit. States Am. 105 (32), 11254–11258.
Schindler, D.W., Hecky, R.E., McCullough, G.K., 2012. The rapid eutrophication of Lake Winnipeg: greening under global change. J. Great Lake. Res. 38, 6–13.
Sen, S., Nandi, S., Dutta, S., 2018. Application of RSM and ANN for optimization and modeling of biosorption of chromium (VI) using cyanobacterial biomass. Applied Water Science 8 (5), 148.
Sommer, U., 1989. Nutrient status and nutrient competition of phytoplankton in a shallow, hypertrophic lake. Limnol. Oceanogr. 34 (7), 1162–1173.
South Korea Ministry of Environment, 2003. Water Resources Management Information System (WAMIS). Republic of Korea Ministry of Environment.
South Korea Ministry of Environment, 2009. Water Environment Information System, Republic of Korea Ministry of Environment. Republic of Korea (South Korea).
South Korea Ministry of Environment, 2012. Water environment information system. In: Real-time Water Quality (River, Lake), Water Level and Precipitation over Major Points in South Korea. Republic of Korea Ministry of Environment, Republic of Korea (South Korea).
Srisuksomwong, P., Pekkoh, J., 2019. Artificial neural network model to prediction of eutrophication and Microcystis aeruginosa bloom in maekuang reservoir, Chiangmai, Thailand. Numerical Computations: Theory and Algorithms NUMTA 2019, 235.
Srivastava, A., Ahn, C.-Y., Asthana, R.K., Lee, H.-G., Oh, H.-M., 2015. Status, alert system, and prediction of cyanobacterial bloom in South Korea. BioMed Res. Int. 2015 (584696), 8.
Sterner, R.W., 1994. Seasonal and spatial patterns in macro-and micronutrient limitation in Joe Pool Lake, Texas. Limnol. Oceanogr. 39 (3), 535–550.
Teta, R., Romano, V., Della Sala, G., Picchio, S., De Sterlich, C., Mangoni, A., Di Tullio, G., Costantino, V., Lega, M., 2017. Cyanobacteria as indicators of water quality in Campania coasts, Italy: a monitoring strategy combining remote/proximal sensing and in situ data. Environ. Res. Lett. 12 (2), 024001.
Tortora, G.J., Funke, B.R., Case, C.L., Johnson, T.R., 2004. Microbiology: an Introduction. Benjamin Cummings, San Francisco, CA.
Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour. Res. 39 (8).
Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic press.
World Health Organization, 2001. Water-related Diseases.