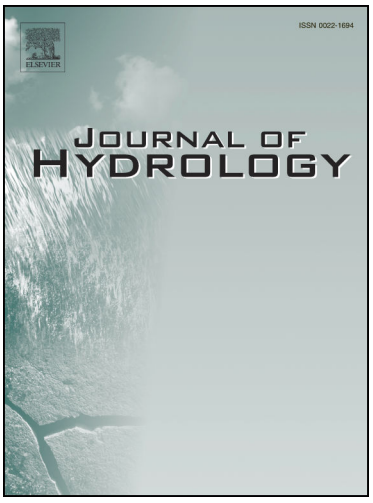
Journal Pre-proofs




Research papers

# Probabilistic forecasting of cyanobacterial concentration in riverine systems using environmental drivers

Seungbeom Kim, Raj Mehrotra, Seokhyeon Kim, Ashish Sharma

PII: S0022-1694(20)31087-8
DOI: https://doi.org/10.1016/j.jhydrol.2020.125626
Reference: HYDROL 125626

To appear in: *Journal of Hydrology*

Received Date: 23 July 2020
Revised Date: 6 October 2020
Accepted Date: 8 October 2020

Please cite this article as: Kim, S., Mehrotra, R., Kim, S., Sharma, A., Probabilistic forecasting of cyanobacterial concentration in riverine systems using environmental drivers, *Journal of Hydrology* (2020), doi: https://doi.org/10.1016/j.jhydrol.2020.125626

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier B.V.

Journal Pre-proofs

# Probabilistic forecasting of cyanobacterial concentration in riverine systems using environmental drivers

**Seungbeom Kim [a, b], Raj Mehrotra [a], Seokhyeon Kim [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

Corresponding author: Ashish Sharma (a.sharma@unsw.edu.au)

Journal Pre-proofs

## 1. Introduction

The manifestation of an active cyanobacterial bloom has been a critical issue in the operation and maintenance of water-systems worldwide (Brooks et al., 2016; Haakonsson et al., 2020; Hallegraeff, 1993; Paerl and Huisman, 2008). Cyanobacteria is believed to be one of the first organisms to spread across the earth (Herrero et al., 2008) and is known to recur when conditions are favorable (Anneville et al., 2015; Paerl and Ustach, 1982). In addition to its unattractive appearance and smelly odor, cyanobacterial bloom worsens quality of water and generates toxins which can negatively impact humans and animals (Fleming et al., 2002; Huisman et al., 2018). For instance, liver and kidney can be damaged by direct ingestion of water contaminated with cyanobacteria, an impact that occurs with humans and animals alike (Yunes, 2019). Unexpected outbreaks, rare number of events and site-specific characteristics present some of the obstacles in predicting algae accurately in advance, as a result of which water authorities have mainly adopted follow-up empirical approaches for prediction after the cyanobacterial bloom reaches a certain level. With the increasing frequency of cyanobacterial blooms around the world (Glibert et al., 2005), considerable research has gone into identifying drivers for both cyanobacteria outbreaks and propagation and formulating effective forecasting models. Although the specific conditions that lead to an algal bloom may be location or also event dependent, it has been established that algal blooms are dominantly affected by some common conditions characterizing the climatology (e.g. temperature and radiation), hydraulics (e.g. retention time and water velocity) and nutrient concentrations (e.g. nitrogen and phosphorus) of the water system (Fornarelli et al., 2013; O'Keeffe, 2019; Obenour et al., 2014; Paerl and Otten, 2013). Kim et al. (2020b) studied data from South Korea and concluded that out of dominant environmental variables, water temperature is the most important factor for algal bloom growth. This is in line with the

findings of Cha et al. (2017) who also found that temperature and retention time are the key factors for algal bloom formation in Korea.

The literature reports a variety of studies to develop forecasting models for cyanobacteria occurrences based on location specific process drivers. McGillicuddy (2010) analyzed various predictive models which loosely fall into the category of conceptual, empirical and numerical models. While numerical or dynamical models attempt to simulate the physics that leads to cyanobacterial evolution, conceptual models characterize the evolution through a simplified data driven statistical/stochastic model form that requires model parameters to be calibrated using observed data. Empirical models assume a causal relationship between the response and the process drivers, but often use the data alone for characterization of the modelled response. Empirical models based on the artificial neural network have been developed and applied to forecast cyanobacteria occurrence and growth conditional on environmental variables such as temperature, pH, phosphorus and nitrogen (Bowden et al., 2005a; Bowden et al., 2005b; Guzel, 2019; May et al., 2008; Pyo et al., 2020; Sen et al., 2018; Srisuksomwong and Pekkoh, 2019). To avoid the time lag problem of the availability of real time algal bloom information, Ibelings et al. (2003) used a modeling approach based on long-term weather forecasts to predict surface water bloom formation. They combined traditional numerical modeling based on differential equations with an expert system based on fuzzy logic. Zhang et al. (2013) coupled the near real time remotely sensed algal bloom information with process-based models to forecast bloom behavior over a period of days to weeks. Similarly, Cha et al. (2014) used a Bayesian hurdle Poisson model to identify the conditions that affect the abundance of cyanobacteria relative to other phytoplankton, and developed a model for cyanobacteria prediction. Zhao and Huang (2014) identified significant environmental factors influencing cyanobacterial bloom occurrence using two Probit models for short-term forecasts of bloom occurrence in the Hill Dagong water area of

Lake Tai in China. Lee and Lee (2018);Yi et al. (2018) applied artificial neural network-based machine learning techniques to predict chlorophyll-a as a surrogate of cyanobacteria concentration in rivers in Korea. Kim et al. (2020b) presented a binary forecasting model focusing on the occurrence or non-occurrence of cyanobacteria in rivers in Korea. They used three dominant environmental factors to define bacterial growth in a dynamic manner. Apart from these empirical predictive models, physical models, for example, Environmental Fluid Dynamics Code (EFDC) (a multidimensional water modelling system, including hydrodynamic, sediment-contaminant, and eutrophication components), and CE QUAL W2 (a water quality and hydrodynamic model for rivers, estuaries, lakes, reservoirs, and river basin systems), represent conventional physical/numerical models for water quality that have also been applied to this problem (He et al., 2011). Although such physical models are shown to capture well the observed response in their study regions well, they cannot be easily extended to other parts of the world because they require extensive data (El-Shafie et al., 2014) and fine-tuning/fitting of model parameters. Likewise, although the binary predictive models are simple to implement and can easily be extended to other regions, these can only predict bloom occurrence/nonoccurrence and are of limited assistance when the aim is to control bloom concentration through engineering controls such as upstream water releases.

Keeping these limitations in mind, the present study proposes a probabilistic forecasting model for cyanobacteria concentration to assist with risk-based management of the bloom by water authorities who can modulate pertinent environmental control variables. The proposed model not only forecasts the occurrence probability of cyanobacteria, but also provides quantitative forecasts using only a few environmental variables. Being simple and effective, the model can be used for assessing proactive measures for the cyanobacterial bloom control by evaluating the changes in the probability distribution by adjusting environmental factors. The model proposed here provides a one week ahead probabilistic forecast of cyanobacteria

using the current values of cyanobacteria concentration and identified environmental predictors. The model provides a one week ahead probabilistic forecast of cyanobacteria using the current values of cyanobacteria concentration and identified environmental predictors.

This paper is organized as follows. In Section 2, the datasets and the proposed probabilistic modelling framework are described. Section 3 explains and discusses the results obtained by using the proposed model. Finally, in Section 4, conclusions of this study are presented.

## 2. Data and method

### 2.1 Study area and Data

This study uses data from the four major rivers of South Korea to illustrate the proposed modelling approach. As shown in Figure 1a, the study area is located in the southern part of the Korean Peninsula, latitude: 33°N–39°N and longitude: 124°E–130°E. There are four distinct seasons, spring, summer, autumn and winter, in South Korea belonging to East Asian monsoonal region. As a result of its geographical setting, the region experiences a temperate climate (Savada and Shaw, 1997). The country has four major rivers; Han, Nakdong, Geum, and Yeongsan, which most of the Korean population and industries rely on as major water sources. The overall catchment area of four rivers which cover almost 63 % of South Korea is about 63,016km$^2$ (South Korea Ministry of Environment, 2003), with environmental and climatological conditions not significantly different to each other.

The selection of environmental variables is crucial and sensitive in our modelling strategy. In an earlier study, Kim et al. (2020b) found four environmental variables: $T$ (water temperature); $P$ (total phosphorous); $N$ (total Nitrogen); and $V$ (flow velocity) as significant predictors of cyanobacteria occurrence in South Korean rivers. This study utilizes weekly water quality data for six years from January 2013 to December 2018, obtained from the

Water Environment Information System (http://water.nier.go.kr/) run by the Ministry of Environment of South Korea (South Korea Ministry of Environment, 2012). In detail, the datasets used in this study consist of observed measurements of water temperature (*T*, °C), total phosphorus (*P*, mg/L), total nitrogen (*N*, mg/L), cyanobacterial cell count (*C*, total number of cells/mL) and water flow rate ($m^3$/sec) at 16 stations of these rivers (Figure 1a). The minimum distance between two stations is 20 km, and the environmental factors like velocity, nutrient vary over the locations because of diverse elevations, cross sections, velocities and inflows from adjoining locations. According to the types of data released in South Korea, the number of cyanobacteria cells denotes the overall cell counts (sum of cell counts) of the four toxic cyanobacteria species, *Anabaena, Aphanizomenon, Microcystis* and *Oscillatoria,* which are categorized into cyanobacterial toxins (Bartram and Chorus, 1999) creating a negative impact on the liver or nervous system according to World Health Organization (2001). Nevertheless, it should be noted that every cyanobacteria species acts differently and prefers different environmental conditions. The measured water flow rate ($m^3$/sec) is converted to water velocity (m/sec) using HEC-RAS model (Hydrologic Engineering Center - River Analysis System) (Brunner, 1995), which requires regular-interval cross-section, roughness of each section and water flow rate to estimate mean cross-sectional velocity at measurement points. Physical data such as cross-sections and roughness are collected from the national river management plan of 2015 (Han River Flood Control Office, 2018) updated every five years for all Korean major rivers by the South Korea government. Mean daily flow rate of every measuring spots are collected from the *MyWater*, managed by K-water, State-owned company in South Korea (https://www.water.or.kr/).

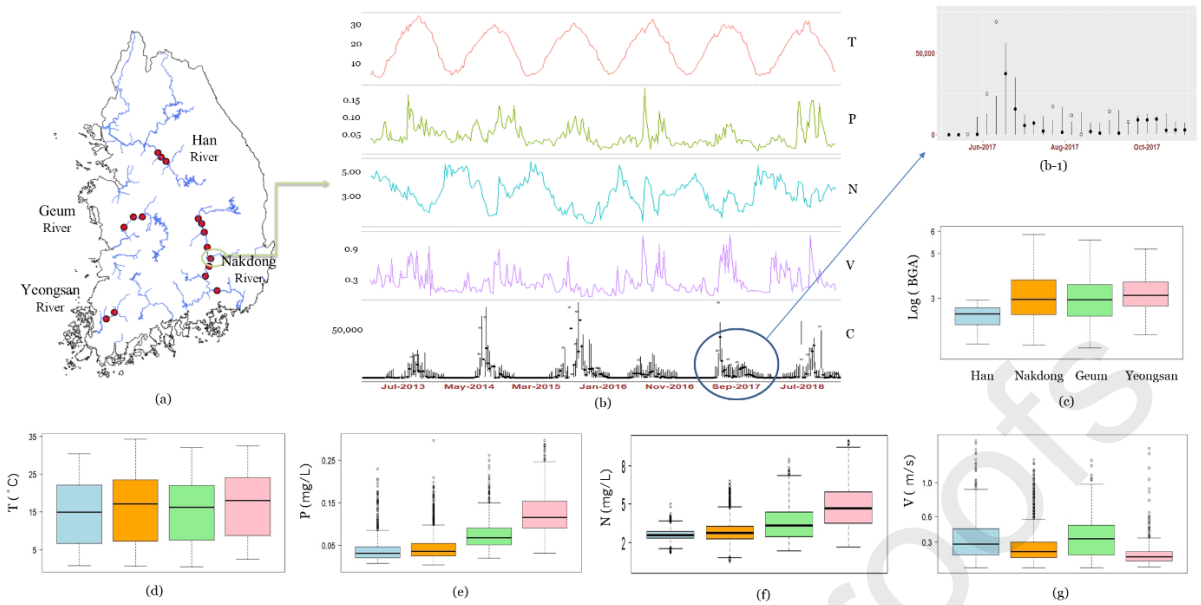



Figure 1. Study area and data summary. (a) Location map of 16 stations along the four major rivers; Han, Nakdong, Geum and Yeongsan, (b) Example of weekly time series for *T, P, N, V* and *C* at Gangjung station in Nakdong river. (b-1) the vertical line at each time step is the range between 1st and 3rd quantile of the model forecasts, ● is the case when the observed value belongs to the range, and in other case, it is marked as ○. (c)-(g) Box plots using 6-year weekly *C, T, P, N* and *V* at the 16 stations. *T* = water temperature; *P* = total phosphorous; *N*= total Nitrogen, *V* = flow velocity; *C* = number of cyanobacterial cells.

## 2.2 Methodology

Cyanobacteria evolution proceeds through interaction with a variety of variables that constantly change in time and space making it difficult to physically model cyanobacterial occurrence and growth. However, one can simplify the relationship by casting this evolution in a probabilistic setting. Markov models are frequently used to synthesize the evidence available for any process, finding uses in nearly every discipline as a result. They assume that the random variable exhibits a finite order of dependence on its past occurrences, an assumption that is often adapted by invoking exogenous variables that modulate the Markovian process. Following this, we propose a Markov model of cyanobacteria growth which takes into account cyanobacteria concentration and associated exogenous factors. The major factors that influence the growth of cyanobacteria are temperature, velocity and the

nutrient composition of the suspending medium. Transition probabilities of a Markov model express the likelihood of cyanobacteria growth to change from one state to another.

The model described here is intended to provide a weekly probabilistic forecast of the cyanobacteria occurrence and concentration based on the cyanobacteria count and values of environmental variables at the previous time step. Thus, the model can be considered as a short-term probabilistic forecasting model. The model ascertains the probability distribution of the one-timestep-ahead cyanobacterial occurrence or concentration characterizing the random variable as a Markov chain and expressing its probability distribution using Kernel Density Estimation (KDE). The key reasoning for adopting Markovian dependence originates from the bacterial growth mechanism which imparts cyanobacterial growth as a clear function of its preceding state (Kim et al., 2020a; Kim et al., 2020b; Tortora et al., 2004). The kernel density estimation (KDE) procedure is a non-parametric means to define the probability density using a finite data sample, used here to express the conditional probability distribution of Cyanobacterial concentration. In the approach outlined below, cyanobacteria occurrences and counts are generated separately at each time step. The probability of cyanobacteria occurrences in occurrence model is generated using a 1$^{st}$ order Markov model conditional on the previous time step values of exogenous variables (environmental variables in our case here) and cell count of cyanobacteria. On occasions where the occurrence model predicts an event, the cell count or cyanobacterial concentration is ascertained using KDE, again conditional on previous time step values of exogenous predictors and cyanobacteria concentration. The details of the occurrence and count stages of the model are presented next.

### 2.2.1 Cyanobacteria occurrence model

Hereinafter, we denote a cyanobacterial occurrence/amount at time $t$ as $C_t$ and at the $q$th time step prior to the current as $C_{t-q}$. It should be noted that the same notation ($C_t$) is used both for cyanobacterial occurrence and amount and should be taken in the context of the issue being

addressed. Also, both single-variables and parameters are defined as non-bold while both multi-variable vectors and matrices are presented as bold characters or symbols.

The cyanobacteria forecasting model can be viewed as the conditional prediction of $C_t|\mathbf{Z}_{t-1}$ where $\mathbf{Z}_{t-1}$ serves as a vector of variables at a given location at time *t-1* and includes exogenous environmental control variables, $\mathbf{X}_{t-1}$ along with cyanobacterial count/occurrence, $C_{t-1}$ at the preceding time step. The parameters (in case of a binary response, the transition probabilities) of a first-order Markov model are defined by $P(C_t|C_{t-1})$ when $\mathbf{Z}_{t-1}$ consists of $C_{t-1}$ alone. However, if the conditioning vector $\mathbf{Z}_{t-1}$ includes additional predictors $\mathbf{X}_{t-1}$, these transition probabilities would modify as $P(C_t|C_{t-1},\mathbf{X}_{t-1})$. To estimate $P(C_t|C_{t-1},\mathbf{X}_{t-1})$, the following parameterization is applied (Mehrotra and Sharma, 2007):

$$
\begin{aligned}
P(C_t = 1 \,|\, C_{t-1} = i,\mathbf{X}_{t-1}) &= \frac{P(C_t = 1,C_{t-1} = i,\mathbf{X}_{t-1})}{P(C_{t-1} = i,\mathbf{X}_{t-1})} \\
&= \frac{f(\mathbf{X}_{t-1}|C_t = 1,C_{t-1} = i) \times P(C_t = 1,C_{t-1} = i)}{f(\mathbf{X}_{t-1}|C_{t-1} = i) \times P(C_{t-1} = i)} \\
&= \frac{P(C_t = 1,C_{t-1} = i)}{P(C_{t-1} = i)} \times \frac{f(\mathbf{X}_{t-1}|C_t = 1,C_{t-1} = i)}{f(\mathbf{X}_{t-1}|C_{t-1} = i)} \\
&= \frac{P(C_t = 1,C_{t-1} = i)}{P(C_{t-1} = i)} \times \frac{f(\mathbf{X}_{t-1}|C_t = 1,C_{t-1} = i)}{[f(\mathbf{X}_{t-1} \,|\, C_t = 1, C_{t-1} = i)P(C_t = 1 \,|\, C_{t-1} = i)] + [f(\mathbf{X}_{t-1} \,|\, C_t = 0,C_{t-1} = i)P(C_t = 1 \,|\, C_{t-1} = i)]}
\end{aligned}
\tag{1}
$$

The first term of Eq. (1), $P(C_t|C_{t-1})$, represents the standard transition probabilities of a first-order Markov model. The second term presents the additional effects introduced by the predictor set $\mathbf{X}_{t-1}$ in the conditioning vector $\mathbf{Z}_{t-1}$. In the most simplified form, $\mathbf{X}_{t-1}$ is assumed to be consisting of environmental variables which are normally distributed. Under this assumption, the associated conditional probability density $f(\mathbf{X}_{t-1}|C_t = 1,C_{t-1} = i)$ can be approximated as a multivariate normal. As a result, $P(C_t|C_{t-1},\mathbf{X}_{t-1})$ can be simplified as

$$= P_{1i}\frac{\frac{1}{\det(\mathbf{V}_{1,i})^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{1,i})\mathbf{V}_{1,i}^{-1}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{1,i})'\}}{\left[\frac{1}{\det(\boldsymbol{V}_{1,i})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{1,i})\mathbf{V}_{1,i}^{-1}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{1,i})'\right\}P_{1i}\right]+\left[\frac{1}{\det(\mathbf{V}_{0,i})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{0,i})\mathbf{V}_{0,i}^{-1}(\mathbf{X}_{t-1}-\boldsymbol{\mu}_{0,i})'\right\}P_{1i}\right]} \quad (2)$$

where $\boldsymbol{\mu}_{1,i}$ is the conditional mean vector of X when $C_t = 1, C_{t-1} = i$ and $\boldsymbol{\mu}_{0,i}$ when $C_{t-1} = i$ and $C_t = 0$. Likewise, $\mathbf{V}_{1,i}$ and $\mathbf{V}_{0,i}$ are the corresponding variance–covariance matrixes. $P_{1i}$ represents the transition probabilities of a first-order Markov model $P(C_t = 1|C_{t-1} = i)$ and det() represents the determinant operation.

In Eq. (2), $\boldsymbol{\mu}_0$, $\mathbf{V}_0$, $\boldsymbol{\mu}_1$, $\mathbf{V}_1$ and $p_{1i}$ are the parameters of the modified Markov model for $P(C_t|C_{t-1}, \mathbf{X}_{t-1})$ These can be derived by forming different conditional subsets of the given cyanobacterial series and by calculating the conditional probabilities, means, variances and covariances. For example, $\boldsymbol{\mu}_{1,i}$ is the mean of a subset of data formed by pooling out weeks when there is Cyanobacteria occurrence noted ($C_t$>1000). For the cases, where the assumption of a multivariate normal is violated, conditional probability densities $f(\mathbf{X}_{t-1}|C_t, C_{t-1})$ and $f(\mathbf{X}_{t-1}|C_{t-1})$ of Eq. (1) may be estimated either using nonparametric alternatives, for instance, kernel density estimation or adopting more appropriate probability distributions. For this study, the assumption of a multivariate normal distribution is adopted.

### 2.2.2 Cyanobacterial count model

The number of cyanobacterial count should be predicted for each time step by the occurrence model as an event. The cyanobacterial count model is based on the kernel density procedure defined in previous studies (Harrold et al., 2003; Mehrotra and Sharma, 2007; Sharma, 2000; Sharma and O'Neill, 2002; Sharma et al., 1997). The model structure description described here is adopted from the above publications and is reproduced here for the sake of convenience of the readers. The model is mentioned hereafter as the Kernel Density Estimation (KDE) model.

Similar to the occurrence model, the KDE predicts the number of cyanobacterial counts conditional on the values of cyanobacteria count and the environmental variables at the preceding time step. Let the cyanobacterial count at time $t$ be $C_t$, the conditioning vector be $\mathbf{X}_t$ consisting of $q$ predictor variables. The conditional density estimates for week $t$ for a multivariate Gaussian kernel can be written as:

$$f(C_t \mid \mathbf{X}_{t-1}) = \sum_{i=1}^{N} \frac{1}{(2\pi\lambda^2 S')^{\frac{1}{2}}} w_i exp\left(-\frac{(C_t - b_i)^2}{2\lambda^2 S'}\right) \quad (3)$$

where $f(C_t|\mathbf{X}_{t-1})$ is the calculated conditional multivariate probability density for week $t$, which is defined as a weighted sum of $N$ Gaussian density functions with $b_i$ (mean) and $\lambda^2 S'$ (covariance). Here, $N$ is the total number of observations and $S'$ is a spread measure of the conditional density, calculated as

$$S' = S_{RR} - \mathbf{S}_{XR}^T \mathbf{S}_{XX}^{-1} \mathbf{S}_{XR} \quad (4)$$

the covariance of $[C_t, \mathbf{X}_{t-1}]$ is defined as:

$$\mathrm{Cov}[C_t, \mathbf{X}_{t-1}] = \begin{bmatrix} S_{RR} & \mathbf{S}_{XR}^T \\ \mathbf{S}_{XR} & \mathbf{S}_{XX} \end{bmatrix} \quad (5)$$

The contributed portion of each kernel in building the conditional probability density is assigned weight $w_i$ and is expressed as

$$\omega_i = \frac{exp\left(-\frac{1}{2\lambda^2}\{[\mathbf{X_{t-1}} - \mathbf{X_i}]\boldsymbol{\Psi}\}^T [\mathbf{S}_{XX}]^{-1}\{[\mathbf{X_{t-1}} - \mathbf{X_i}]\boldsymbol{\Psi}\}\right)}{\sum_{i=1}^{N}\left(-\frac{1}{2\lambda^2}\{[\mathbf{X_{t-1}} - \mathbf{X_i}]\boldsymbol{\Psi}\}^T [\mathbf{S}_{XX}]^{-1}\{[\mathbf{X_{t-1}} - \mathbf{X_i}]\boldsymbol{\Psi}\}\right)}, \quad (6)$$

where $\boldsymbol{\psi}$ is the diagonal matrix of influence weights (Mehrotra and Sharma, 2006). It incorporates the relative influence of each predictor in forming the conditional probability

density. $\lambda$, a kernel bandwidth, is a spread measure and, $b_i$ is the conditional mean related to each kernel, described as:

$$b_i = C_i + [\mathbf{S}_{XR}]^T [\mathbf{S}_{XX}]^{-1} \{[\mathbf{X}_{t-1} - \mathbf{X}_i]\mathbf{\Psi}\} \quad (7)$$

Basic principles of Eq. (3-7) and detailed discussions about the kernel density procedure are described in Mehrotra and Sharma (2007b) and Sharma and O'Neill (2002).

2.2.3 Predictor sets with combinations of environmental variables

In a so-called leave-one-out validation framework, we develop/calibrate the model using data for 15 stations and test/validate the developed model on the left out station. This process is repeated 16 times, by rotating the left out station to obtain the cross-validated predictions at all 16 stations. In addition, at each time step 1,000 predictions are made to form a probabilistic forecast at that time step.

In order to select an optimal set of predictors, we form multiple predictor sets using all possible combinations of these four environmental variables (e.g. M1 = TPV, M2 = TV,…, and M15 = P) as shown in Figure 3, and use them to both occurrence and count models to predict the cyanobacteria occurrence probability and concentration and evaluate the results.

To quantitatively evaluate the performances of these predictor sets, we apply a 2×2 contingency table (Wilks, 2011), on the weekly forecasts (both occurrence and count) of our prediction model (M = 1 or 0) to assess if it matches with the observations (O = 1 or 0). Therefore, it is composed of four scalars: hit (M = 1 and O = 1, denoted as “a”); false alarm (M = 1 and O = 0, denoted as “b”); miss (M = 0 and O = 1, denoted as “c”); and correct rejection (M = 0 and O = 0, denoted as “d”). While the occurrence calculation is performed based on the overall observed dataset, count concentration is calculated on the reduced datasets where occurrence of cyanobacteria is noted (i.e. O = 1). For the occurrence assessment in Figure 2a, for example, a hit (a) is identified if the probability of occurrence

model at a given time step is greater than 50% (out of 1000 forecasts), i.e. M = 1; and the observed cyanobacterial cell count also exceeds the threshold, i.e. O = 1. In the count assessment in Figure 2b, a hit (a) is captured when the observed cyanobacterial cell count is within the interquartile range of forecasting distribution (formed from the 1000 forecasts) at a certain time step. Here, the threshold is defined as 1000 cells/mL based on South Korean Government's algal alert system in place since 1997 (Srivastava et al., 2015).

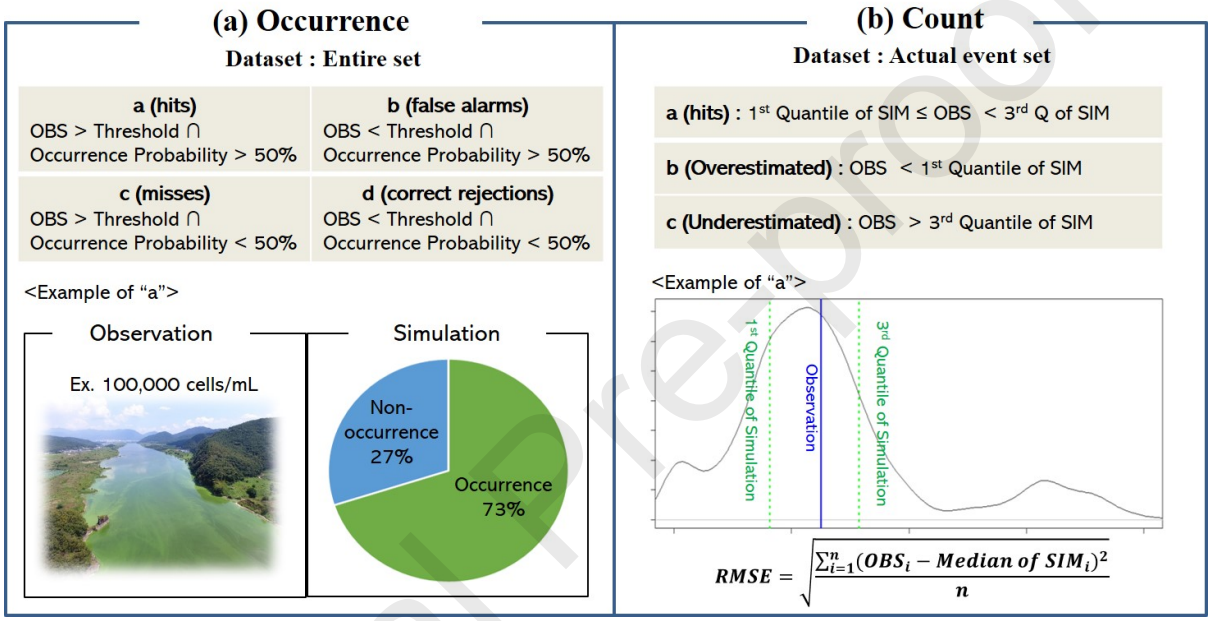



Figure 2. Assessment criteria of both occurrence and count of cyanobacteria, OBS means observation and SIM stands for simulation.

We adopt two scores to evaluate the performance of cyanobacteria occurrences forecasts. These are the Proportional Correct (PC) and Threat score (TS). PC in Equation (8) is the most direct and intuitive measure of the accuracy of a forecast for discrete events. Although, PC is straightforward to calculate, for rarer events like cyanobacterial bloom occurrences in our case (about 18.6 % event rate; 932 out of 4,992 datasets), the score is significantly influenced by the large number of non-events and inflates the model performance. Therefore, TS is

selected to complement the limitation of PC since it does not consider the correct rejections (d) in the calculations (equation 9). Perfect score of both metrics is 1 and the worst value is 0.

$$\mathrm{PC} = \frac{a + d}{a + b + c + d} \tag{8}$$

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

For the count assessment, we use TS only as there is no "d" element in count assessment and therefore PC can't be calculated. In addition to this, for the count assessment, we also calculate Root Mean Square Error (RMSE) as an additional measure of the differences between the observed and forecast cell counts, and defined it as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i - S_{Med,\,i})^2}{n}} \tag{10}$$

Where, $O_i$ is the observed cyanobacterial cell count and $S_{Med}$ is the median value of model forecast at time step *i*.

### 2.2.4 Sensitivity analysis with controllable predictors

In addition to estimating PC, TS and RMSE, a sensitivity analysis is also conducted to assess the influence of individual environmental variables on the probability of cyanobacteria occurrence. A predictor set consisting of temperature and velocity variables is selected for this sensitivity analysis since these variables could be controlled by external operations. Generally, temperature of water at higher elevation is lower than at low elevations. Similarly, water temperature of upstream dam is lower than that of downstream due to its geological location. By releasing water from the upstream dam and by operating the weir-gate

instantaneously, temperature and velocity can be varied in a controlled manner to possibly create unfavorable conditions for the growth of cyanobacteria.

To understand better how the occurrence model reacts when these environmental variables change, three cases are assessed. In the first two cases, the changes in the occurrence probability are noted when temperature and velocity are reduced by a unit standard deviation ($1\sigma$ = 8.4 ºC, 0.25m/s) separately. Even though dropping water temperature by 8.4 ºC does not seem to be realistic, data does exhibit these variations and it would help to visualize easily how the model reacts when a variable is modified.

In the last case, these two variables are reduced together by one standard deviation. Results from these and other assessments are presented next.

## 3. Results and discussion

### 3.1 Performance evaluation

The modeled results and those obtained from the sensitivity analysis of parameters are presented in Figure 3a and in Table S1 of the Supporting Information (hereafter the prefix "S" for figures and tables indicate those in the Supporting Information). In addition, Figure S2 shows time series of predicted and observed cyanobacteria concentration over three representative stations, Gongju, Seungchon and Gangjung. In Figure 3a, the subscripts "o" and "c" in the description of PC and TS scores, respectively, represent the occurrence and count models. All results represent validation performance using the leave-one-out cross-validation procedure outlined before. Overall, all datasets exhibit equally good performances except a few instances where these scores are low. For example, all datasets achieve a PC score of 0.86 at a minimum. However, as mentioned in Section 2.2.3, many non-bloom events (about 81.3% of non-event) result in high score of correct rejections (d) (i.e. non-bloom in both observation and model) leading to an inflated PC score. As TS ignores the

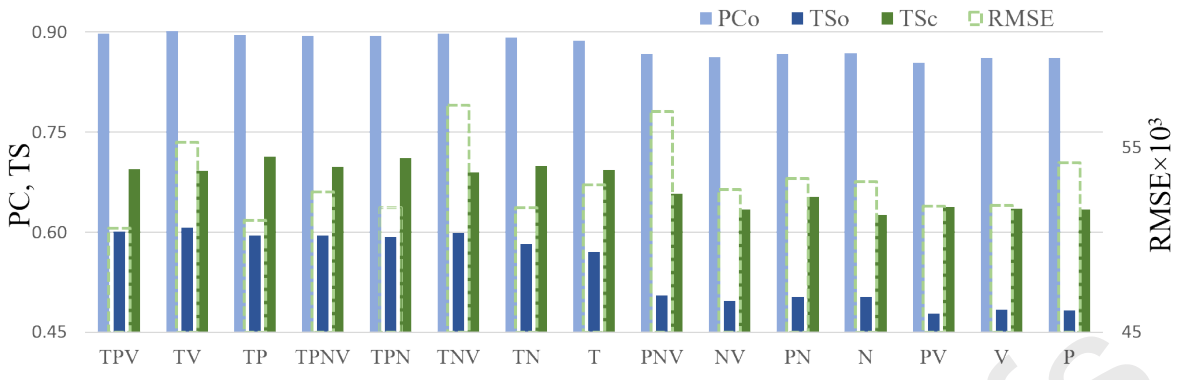
correct rejection term, the score was lower than PC with an averaged value close to 0.50 except for three datasets: *P, V, NV* and *PV* (Figure 3).

When comparing the performances of the four single-variable datasets (i.e. *T, N, V* or *P*), water temperature (*T*) comes up as the most dominant variable amongst the four for both occurrence and count models. It shows comparable performance against datasets consisting of two or more variables (Figure 3). This finding highlights the importance of water temperature in the formulation of any predictive model (Cha et al., 2017; Kim et al., 2020b).

The performance of datasets, *T* and *TN*; *P* and *PV*; *N* and *PN*, are almost identical, probably because of the high joint dependence of the two variables considered (Pearson correlation coefficients are -0.45 between *T* and *N*, 0.32 between *P* and *N*, and 0.31 between *V* with *P*, respectably).

Finally, *TPV* produces the best results whereas both *TPN* and *TPNV* less perform. Following these results, the use of either *TPV, TN, or TV* dataset is likely to be recommended depending on the availability of data.

Overall, by including probabilistic results, this model enables users to make a risk-based decision. In addition, the model can also be viewed as an improved water quality model because it not only provides probabilistic output (occurrence probability) but also a deterministic count of cyanobacteria (median value of count) that physical models normally aim to simulate. While a physical model requires considerable amount of time for calibration and running with large data requirements to specify current boundary and initial conditions, the proposed model does a similar job with a few environmental variables and with a reasonable accuracy. As a result of the model performance, water authorities in the study area can benefit from various aspects such as simple and reliable forecasting, evaluation of proactive measures and most influential environmental factors.




(a)

| TPV model | Overall | $C_{t-1} < 1{,}000$ | $C_{t-1} > 1{,}000$ | $C_{t-1} < C_t$ | $C_{t-1} > C_t$ |
|---|---|---|---|---|---|
| $TS_o$ | 0.60 | 0.30 | 0.77 | 0.66 | 0.53 |
| $TS_c$ | 0.70 | 0.71 | 0.67 | 0.54 | 0.75 |

(b)

Figure 3. (a) Overall performance of the combination of environmental input variables, (b) performance of the TPV dataset under different cyanobacteria concentration (in cells/mL) threshold conditions for the preceding time step. Note TS represents the threat score, with subscripts 'o' and 'c' denoting occurrence and counts respectively.

### 3.2 TPV dataset performance conditional to preceding cyanobacterial count

In this section, we evaluate the predictive performance of the forecasting model under varying cyanobacteria concentration conditions. To assess the stability of the approach, we select the *TPV* dataset because it provides a more consistent performance across various conditions compared to other groups of datasets. We divide the whole data into subsets based on cyanobacteria cell counts for the preceding time step. For example, $C_{t-1} < 1000$ in Figure 3b means the subset when the cyanobacterial bloom did not happen at the preceding time step. Likewise, $C_{t-1} < C_t$ denotes the subset when the current cyanobacterial cell count is greater than the one in the preceding time step. Since calculating TS does not include correct

rejections (d) occupying a high portion of the forecasting cases, we only consider TS to provide an indication of model behavior under varying conditions.

Overall, the occurrence model performs well when $C_{t-1}$ >1,000 cells/mL because it is based on first-order Markovian dependence with that threshold being used. On the other hand, a poor result is obtained when $C_{t-1}$ < 1,000 cells/mL. This is because of a relatively large number of false alarms (b) in calculating TS (Equation 9). Specifically, about 80 % of total observed cell counts in this group showed greater than zero but less than 1000 cells/mL because of which TS is likely to be higher if the threshold (1,000 cells/mL) is lowered. In the count model, the *TPV* exhibits a good performance with fewer deviations across the cases.

### 3.3 Influence of controlled environmental variables on the model outcome

In order to understand the influence of the externally controlled environmental variables on the probability of cyanobacteria occurrence, sensitivity analysis is performed using the TV data alone. This follows the reasoning that it is possible to control water temperature and velocity to some extent by releasing the water from upstream storages where feasible. The change in probability resulting from changing the T, V and both together as boxplots is presented in Figure S1.

By dropping temperature by one standard deviation ($1\sigma$ = 8.4°C), occurrence probability decreases by about 19 % (median value). On the other hand, occurrence probability is reduced by about 7 % (median value) by increasing the velocity by one standard deviation. When changes in these two variables are applied together, the occurrence probability drops by about 24 %. These results suggest that temperature is more effective than velocity and varying both variables together is much more effective in controlling the cyanobacteria occurrence.

### 3.4 Caveats and follow-up studies

This study used four easily assessable environmental variables; water temperature (*T*), total phosphorus (*P*), total nitrogen (*N*) and flow velocity (*V*). Our results show that all of them influence the occurrence and concentration of cyanobacterial blooms to a varying degree. However, other factors such as turbidity, pH, salinity, irradiance, electric conductivity, can also influence the cyanobacteria occurrence. These variables are not considered in the present study for the sake of simplicity and unavailability of data. Remote sensing information can also be used as a surrogate to supplement the ground information needed to study the cyanobacteria occurrence and concentration. For example, Teta et al. (2017) showed that remote sensing can be applied to identify initial cyanobacterial bloom by combining with aerial and in-situ data.

In this study, we assumed that the environmental conditions leading to cyanobacterial growth do not change across space within the study area. This assumption is based on the relatively small spatial extent of the study area (around 100,000 $km^2$) where the environmental and climatological conditions do not vary significantly in space. However, it should be noted that the conditions of environmental variables that trigger algal bloom might differ depending on the regions or event since each location or outbreak has a unique and native environmental control. If enough data is available, this framework should ideally be applied separately to each location, especially so if climatic and land-use conditions are markedly different from each other.

The weekly time step applied in the model framework follows the weekly data sampling interval adopted by South Korea (South Korea Ministry of Environment, 2012). Using data at finer temporal resolutions is expected to further improve the model performance. The frequency of sampling is dictated by a number of factors including indented use, the cost of monitoring, the season and the growth rate of the cyanobacteria. The Guidance manual of Global Water Research Coalition (Newcombe, 2012) recommends that sampling for high

risk/high security supplies (i.e. drinking supplies) should occur on at least a weekly basis and probably twice-weekly when cyanobacterial count of > 2,000 cells/mL is reached. For supplies where the public health risk is deemed to be low (i.e. low cell counts in non-supply reservoirs), fortnightly sampling may be adequate. As duration of cyanobacterial blooms is usually greater than a month (in temperate zones Cyanobacteria can last 2 to 4 months, whereas in tropical and subtropical regions they can continue all year round), a weekly sampling interval is expected to provide a good idea of the Cyanobacteria growth behavior (Sivonen and Jones, 1999).

## 4. Conclusion

This paper demonstrated the applicability of a probabilistic short-term forecasting model for cyanobacterial bloom occurrence and count. The model works in two parts: prediction of cyanobacteria occurrence and simulation of cyanobacteria counts following a positive simulation of occurrence. The occurrence model simulates cyanobacterial bloom using a first-order Markov model conditional on environmental variables. Cyanobacterial counts are simulated based on a first order conditional kernel density-based model. Results of a sensitivity analysis conducted using various combinations of four environmental control variables, suggest *TPV, TN* and *TV* as sets of optimal environmental variables for prediction of cyanobacteria over the study region.

A minimum PC score of 0.86 and conditional TS of 0.5 capturing rare occurrence events and stable results across various conditions and stations consolidate the performance of the proposed model. The novel elements of this research include the idea of modelling the cyanobacteria growth rate conditional on the previous state of cyanobacteria occurrence and environmental variables under a Markovian assumption. Water temperature is found to be the most dominant variable in the cyanobacteria predictions followed by the nutrients such as N

and P. The results of sensitivity analysis can help in identifying appropriate proactive measure required to control the cyanobacteria occurrence.

Although, the model is developed and applied using the data for South Korea, the logic adopted and the model structure followed is quite generic and it can easily be extended to other areas depending on the availability of environmental data. Care should, however, be taken to assess whether the same environmental control variables, or variables indicative of localized conditions, climate and land use are to be used.

**CRediT authorship contribution statement**

**Seungbeom Kim:** Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, Funding acquisition. **Raj Mehrotra:** Conceptualization, Methodology, Software, Resources, Writing - Review & Editing, Supervision. **Seokhyeon Kim:** Methodology, Investigation, Data Curation, Writing - Review & Editing, Project administration, Supervision. **Ashish Sharma:** Conceptualization, Methodology, Writing - Review & Editing, Supervision.

**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 appreciate K-water for providing sponsorship of the first author and supporting 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 phosphorous, total nitrogen, 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/) (language: Korean).

Journal Pre-proofs

**References**

Anneville, O., Domaizon, I., Kerimoglu, O., Rimet, F., Jacquet, S., 2015. Blue-green algae in a "Greenhouse Century"? New insights from field data on climate change impacts on cyanobacteria abundance. Ecosystems, 18(3): 441-458.

Bartram, J., Chorus, I., 1999. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management. CRC Press.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005a. Input determination for neural network models in water resources applications. Part 1—background and methodology. Journal of Hydrology, 301(1-4): 75-92.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2005b. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. Journal of Hydrology, 301(1-4): 93-107.

Brooks, B.W. et al., 2016. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? Environmental toxicology and chemistry, 35(1): 6-13.

Brunner, G.W., 1995. HEC-RAS River Analysis System. Hydraulic Reference Manual. Version 1.0, HYDROLOGIC ENGINEERING CENTER DAVIS CA.

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 research, 124: 11-19.

Cha, Y., Park, S.S., Kim, K., Byeon, M., Stow, C.A., 2014. Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. Water Resour Res, 50(3): 2518-2532.



Journal Pre-proofs

El-Shafie, A., Najah, A., Alsulami, H.M., Jahanbani, H., 2014. Optimized neural network prediction model for potential evapotranspiration utilizing ensemble procedure. Water Resour Manage, 28(4): 947-967.

Fleming, L.E. et al., 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 Resources Research, 49(6): 3626-3641.

Glibert, P.M., Anderson, D.M., Gentien, P., Granéli, E., Sellner, K.G., 2005. The global, complex phenomena of harmful algal blooms.

Guzel, H.O., 2019. Prediction of Freshwater Harmful Algal Blooms in Western Lake Erie Using Artificial Neural Network Modeling Techniques.

Haakonsson, S. et al., 2020. Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-Gamma model. Water Research: 115710.

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,.

Harrold, T.I., Sharma, A., Sheather, S.J., 2003. A nonparametric model for stochastic generation of daily rainfall occurrence. Water Resour Res, 39(10).

He, G. et al., 2011. Application of a three-dimensional eutrophication model for the Beijing Guanting Reservoir, China. Ecological Modelling, 222(8): 1491-1501.

Herrero, A., Flores, E., Flores, F.G., 2008. The Cyanobacteria: Molecular Biology, Genomics, and Evolution. Caister Academic Press.

Huisman, J. et al., 2018. Cyanobacterial blooms. Nature Reviews Microbiology, 16(8): 471-483.

Ibelings, B.W., Vonk, M., Los, H.F., van der Molen, D.T., Mooij, W.M., 2003. Fuzzy modeling of cyanobacterial surface waterblooms: validation with NOAA-AVHRR satellite images. Ecological Applications, 13(5): 1456-1472.

Kim, K.B., Jung, M.-K., Tsang, Y.F., Kwon, H.-H., 2020a. Stochastic modeling of chlorophyll-a for probabilistic assessment and monitoring of algae blooms in the lower Nakdong River, South Korea. Journal of Hazardous Materials: 123066.

Kim, S., Kim, S., Mehrotra, R., Sharma, A., 2020b. Predicting cyanobacteria occurrence using climatological and environmental controls. Water Research, 175: 115639. DOI:https://doi.org/10.1016/j.watres.2020.115639

Lee, S., Lee, D., 2018. Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. International journal of environmental research and public health, 15(7): 1322.

May, R.J., Dandy, G.C., Maier, H.R., Nixon, J.B., 2008. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environmental Modelling & Software, 23(10-11): 1289-1299.

McGillicuddy, D., Jr, 2010. Models of harmful algal blooms: conceptual, empirical, and numerical approaches. Journal of marine systems: journal of the European Association of Marine Sciences and Techniques, 83(3-4): 105.

Mehrotra, R., Sharma, A., 2006. Conditional resampling of hydrologic time series using multiple predictor variables: A k-nearest neighbour approach. Adv Water Resour, 29(7): 987-999. DOI:10.1016/j.advwatres.2005.08.007

Mehrotra, R., Sharma, A., 2007. A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low-frequency variability. Journal of Hydrology, 335(1-2): 180-193.

Newcombe, G., 2012. International guidance manual for the management of toxic cyanobacteria. IWA Publishing.

O'Keeffe, J., 2019. Cyanobacteria and Drinking Water: Occurrence, Risks, Management and Knowledge Gaps for Public Health. National Collaborating Centre for Environmental Health.

Obenour, D.R., Gronewold, A.D., Stow, C.A., Scavia, D., 2014. Using a B ayesian hierarchical model to improve L ake E rie cyanobacteria bloom forecasts. Water Resour Res, 50(10): 7847-7860.

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. Microbial ecology, 65(4): 995-1010.

Paerl, H.W., Ustach, J.F., 1982. Blue-green algal scums: An explanation for their occurrence during freshwater blooms 1. Limnology and oceanography, 27(2): 212-217.

Pyo, J. et al., 2020. An Integrative Remote Sensing Application of Stacked Autoencoder for Atmospheric Correction and Cyanobacteria Estimation Using Hyperspectral Imagery. Remote Sensing, 12(7): 1073.

Savada, A.M., Shaw, W., 1997. South Korea: A country study, 550. Diane Publishing.

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.

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—A nonparametric probabilistic forecast model. J Hydrol, 239(1-4): 249-258.

Sharma, A., O'Neill, R., 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water Resour Res, 38(7): 5-1-5-10.

Sharma, A., Tarboton, D.G., Lall, U., 1997. Streamflow simulation: A nonparametric approach. Water resources research, 33(2): 291-308.

Sivonen, K., Jones, G., 1999. Cyanobacterial toxins. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management, 1: 43-112.

South Korea Ministry of Environment, 2003. Water Resources Management Informatoin System (WAMIS). Republic of Korea Ministry of Environment.

South Korea Ministry of Environment, 2012. Water Environment Information System. Republic of Korea Ministry of Environment, Republic of Korea (South Korea), pp. Real-time water quality (river, lake), water level and precipitation over major points in 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 research international, 2015.

Teta, R. et al., 2017. Cyanobacteria as indicators of water quality in Campania coasts, Italy: a monitoring strategy combining remote/proximal sensing and in situ data. Environmental Research Letters, 12(2): 024001.

Tortora, G.J., Funke, B.R., Case, C.L., Johnson, T.R., 2004. Microbiology: an introduction, 9. Benjamin Cummings San Francisco, CA.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences, 100. Academic press.

World Health Organization, 2001. Water-related diseases.

Yi, H.-S., Park, S., An, K.-G., Kwak, K.-C., 2018. Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. International journal of environmental research and public health, 15(10): 2078.

Yunes, J.S., 2019. Cyanobacterial Toxins, Cyanobacteria. Elsevier, pp. 443-458.

Zhang, H. et al., 2013. An improved ecological model and software for short-term algal bloom forecasting. Environ Modell Softw, 48: 152-162.

Zhao, L., Huang, W., 2014. Models for identifying significant environmental factors associated with cyanobacterial bloom occurrence and for predicting cyanobacterial blooms. Journal of Great Lakes Research, 40(2): 265-273. DOI:https://doi.org/10.1016/j.jglr.2014.02.011

**Abstract**

Toxic cyanobacteria blooms such as *Anabaena, Aphanizomenon, Microcystis* and *Oscillatoria* are of critical concern for public health and environmental system globally. An algal bloom is largely influenced by factors that jointly characterize the climatology (e.g., water temperature), hydraulics (e.g., water velocity) and nutrient concentrations (e.g., phosphorus and nitrogen). While a wide range of efforts has been made to predict a cyanobacterial bloom, there is still a need for computational tools to characterize the bloom concentration effectively. Here, we present a short-term cyanobacteria forecasting model that not only predicts the occurrences of algal bloom but also provides their concentration conditional on the selected dominant environmental variables. The prediction model operates in two stages. In the first stage, cyanobacterial occurrences are predicted using a first-order Markov model conditioned on a few selected environmental variables. On occasions where a cyanobacterial occurrence is predicted, the second stage predicts cyanobacterial cell counts again conditional on the selected environmental variables. In an application using data for four major rivers in South Korea, a minimum Threat Score of 0.56 (56% forecasting accuracy) with a single environmental variable, temperature, is attained. This simple model provides one week ahead probabilistic prediction of cyanobacteria occurrence and cell concentration making it easier to prioritize proactive measures based on the probability changes caused by relevant changes in the conditioning environmental variables.

Journal Pre-proofs

**Keywords:** Cyanobacterial bloom, River, Probabilistic model, Markov Chain, Stochastic sampling, Kernel density estimation

**Conflict of 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.

**CRediT authorship contribution statement**

**Seungbeom Kim:** Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, Funding acquisition. **Raj Mehrotra:** Conceptualization, Methodology, Software, Resources, Writing - Review & Editing, Supervision. **Seokhyeon Kim:** Methodology, Investigation, Data Curation, Writing - Review & Editing, Project administration, Supervision. **Ashish Sharma:** Conceptualization, Methodology, Writing - Review & Editing, Supervision.

**Highlights**

- A probabilistic model for short-term forecasts of riverine algal blooms is developed
- Model is based on a first-order Markov chain and Kernel Density Estimation
- Model is tested over the four major rivers in South Korea
- Water temperature is found to be the most dominant variable in the forecasting setup

Journal Pre-proofs