Science of the Total Environment 502 (2015) 31–41

Contents lists available at ScienceDirect

Science of the Total Environment

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

Development of early-warning protocol for predicting *chlorophyll-a* concentration using machine learning models in freshwater and estuarine reservoirs, Korea

Yongeun Park [a], Kyung Hwa Cho [b], Jihwan Park [a], Sung Min Cha [c], Joon Ha Kim [a,*]

[a] School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Republic of Korea
[b] School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan 689-798, Republic of Korea
[c] Jeollanam-do Environmental Industries Promotion Institute, 650-94 Songgye-ro, Seongjeon-myeon, Gangjin-gun, Jeollanam-do, 527-811, Republic of Korea

HIGHLIGHTS

- Two machine learning models were used to predict *chlorophyll-a* concentration.
- Two models were trained and validated using a 7-year monitoring data.
- Sensitivity analysis determined the sensitivity of input variables for the models.
- The support vector machine was found as a reliable early-warning model.
- This study proposed a simple early warning protocol for managing algal blooms.

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 8 August 2014
Accepted 1 September 2014
Available online xxxx

Editor: Simon Pollard

Keywords:
Artificial neural network
Support vector machine
Early warning
Chlorophyll-a
Prediction accuracy
Sensitivity analysis

ABSTRACT

Chlorophyll-a (Chl-a) is a direct indicator used to evaluate the ecological state of a waterbody, such as algal blooms that degrade the water quality in lakes, reservoirs and estuaries. In this study, artificial neural network (ANN) and support vector machine (SVM) were used to predict Chl-a concentration for the early warning in the Juam Reservoir and Yeongsan Reservoir, which are located in an upstream region (freshwater reservoir) and downstream region (estuarine reservoir), respectively. Weekly water quality data and meteorological data for a 7-year period were used to train and validate both the ANN and SVM models. The Latin-hypercube one-factor-at-a-time (LH-OAT) method and a pattern search algorithm were applied to perform sensitivity analyses for the input variables and to optimize the parameters of the two models, respectively. Results revealed that the two models well-reproduced the temporal variation of Chl-a based on the weekly input variables. In particular, the SVM model showed better performance than the ANN model, displaying a higher prediction accuracy in the validation step. The Williams–Kloot test and sensitivity analysis demonstrated that the SVM model was superior for predicting Chl-a in terms of prediction accuracy and description of the cause-and-effect relationship between Chl-a concentration and environmental variables in both the Juam Reservoir and Yeongsan Reservoir. Furthermore, a 7-day interval was determined as an efficient early warning interval in the two reservoirs. As such, this study suggested an effective early-warning prediction method for Chl-a concentration and improved the eutrophication management scheme for reservoirs.

1. Introduction

Algal blooms commonly occur in receiving waterbodies, causing a potential deterioration of water quality, often resulting in problems such as depletion of oxygen, reduced water transparency, and decreased biodiversity in marine and freshwater environments (Hartnett and Nash, 2004). These problems subsequently pose serious direct and indirect threats to aquatic ecosystems and public health (Glasgow et al., 2004). To prevent severe occurences, efficient mitigation and management techniques should be developed by monitoring and modeling algal blooms.

One promising action could be reliable early-warning predictions of algal blooms by incorporating key environmental variables (e.g., temperature, light, and nutrients) (Anderson et al., 2001). Direct modeling of algal blooms, however, may be limited due to practical problems such as insufficient observations and the complex behavior of the algal community. Currently, the *chlorophyll-a* (Chl-a) concentration has been a useful indicator for measuring the abundance and

* Corresponding author. Tel.: +82 62 715 3277; fax: +82 62 715 2434.
E-mail address: joonkim@gist.ac.kr (J.H. Kim).

http://dx.doi.org/10.1016/j.scitotenv.2014.09.005
0048-9697/

variety of phytoplankton and/or algal biomass (Boyer et al., 2009). Because all photosynthetic algae include Chl-a, algal bloom can be easily predicted by investigation of Chl-a concentration in waterbodies.

Complex relationships between diverse environmental factors (e.g., temperature, light, nutrients) are involved in predicting Chl-a concentration (Lee et al., 2003). In order to reliably simulate the Chl-a concentration, models need to incorporate hydrological, geochemical, and ecological variables that impact algal growth. Several models have been used to predict Chl-a concentration, and can be categorized according to deterministic and stochastic approaches. Even though a process-based mathematical model has been widely implemented to predict the general ecological response of phytoplankton to several environmental factors (Thomann and Mueller, 1987), the physical dynamics of algal bloom phenomena are not understood well due to the uncertainty of kinetic rate coefficients for different species (Lee and Lee, 1995; Yabunaka et al., 1997). This fact limits the development of an appropriate formulation for simulating algal blooms and subsequently requires an alternative approach for modeling, such as the promotion of data-driven methodology (Lee et al., 2003). In this study, among current models, two data-driven approaches (i.e., stochastic approaches) are compared as an early-warning prediction model of Chl-a concentration in a lake system, through the evaluation of model performance.

Artificial neural network (ANN) and support vector machine (SVM) models are promising approaches used to reflect the nonlinearity between Chl-a concentration and environmental factors using stochastic error minimization approaches. In particular, ANN is a powerful pattern recognition approach that has been used in areas including business, industry, engineering, and science (Widrow et al., 1994) and has also been applied to predicting algal blooms (Lee et al., 2003). However, ANN has limitations in that empirical risk minimization (ERM) is only considered for minimizing the training error, not for generalizing its performance in the prediction step (Barton and Meckesheimer, 2006; Yuan and Wang, 2008). Recently, instead of ANN, SVM has been introduced as an alternative method for overcoming the intrinsic weaknesses of ANN modeling, while retaining all the advantages of ANN (Govindaraju, 2000). SVM maintains steady performance regardless of input dimensionality and correctly determines the global optimum during the regression process (Ren and Bai, 2010).

In this paper, a detailed assessment for two stochastic models is performed in order to evaluate the early warning predictability of Chl-a concentration through the prediction accuracy and an analysis of the relationship between input and output. Although a straightforward comparison (e.g., determining model performance using prediction accuracy) between ANN and SVM has been presented (Balabin and Lomakina, 2011; Behzad et al., 2009; Chen et al., 2005), there has yet to be a comprehensive analysis in terms of the application of parameter optimization and sensitivity analysis. Therefore, to investigate the model performances in predicting the Chl-a concentration, ANN and SVM models were set up in the Juam Reservoir (freshwater reservoir) and Yeongsan Reservoir (estuarine reservoir) in the southwestern part of Korea, which have distinct site-specific characteristics. The objectives of this study are: 1) to develop a reliable model for early warning prediction of Chl-a using ANN and SVM by optimizing key model parameters, 2) to evaluate model-specific features based on model performance in terms of a statistical evaluation and sensitivity analysis in response to different input variables, and 3) to propose a simple early warning protocol for managing algal blooms using Chl-a concentration as a key decision-supporting system.

2. Materials and methods

2.1. Site description and data acquisition

The Juam Reservoir (JAR) and Yeongsan Reservoir (YSR) are located in the southwestern region of Korea (see Fig. 1). The JAR, surrounded by mountainous valleys at approximately 400 m altitude above sea level, is a freshwater lake that flows into the Seomjin River. This lake is the most important freshwater resource in the region (e.g., Gwangju, Naju, Mokpo) and supplies 25 million m^3/day for drinking water. It is approximately 40 km long, and has a surface area of 33 km^2, average depth of 14 m (maximum depth of 47 m), and basin area of 1010 km^2 (Shin et al., 2000). In contrast, the YSR is an estuarine reservoir built in 1981 by damming the downstream end of the Yeongsan River, which supplies agricultural water and prevents flooding in the surrounding regions; it is 23.5 km from the Mongtan Bridge to the Yeongsan Estuarine Dam, and has a surface area of 34.6 km^2, average depth of 10 m (maximum depth of 21 m), and a basin area of 3468 km^2 (Lee et al., 2009).

Water quality in the JAR is maintained in relatively good conditions compared to other major reservoirs in South Korea, in terms of nutrients and Chl-a concentration (Jones et al., 2003). Because the JAR basin is dominantly composed (80%) of forested area (8% agricultural area, 0.6% urban area, etc.), nutrients released from the basin were relatively insignificant (Jones et al., 2006). Water quality in the YSR, however, has been drawing researchers' attention in recent years because of the aggravated aqua-ecological state caused by its structural deficiency (i.e., estuarine dam) and the pollutant load from the Yeongsan Watershed. As the dam is constructed at the outlet of the Yeongsan River, natural water circulation has been inhibited, resulting in a degradation of water quality due to the anoxic and hypoxic conditions in the bottom layer (Lee et al., 2009, 2010). Furthermore, there are numerous point- and non-point sources discharging into the YSR from the Yeongsan Watershed, causing eutrophication (Ki et al., 2007; Cho et al., 2009a).

The water quality data in the two reservoirs were monitored by the Yeongsan River Environmental Research Center (near dike dams in both the JAR and YSR; see Fig. 1). Surface water samples were taken at weekly intervals over a 7-year period (from 2006 to 2012). Water quality observations in this study include *chlorophyll-a* (Chl-a), phosphate phosphorus (PO_4-P), ammonium nitrogen (NH_3-N), nitrate nitrogen (NO_3-N), and water temperature. Daily meteorological data were monitored by the Korea Meteorological Administration at local weather stations. In addition, solar radiation and wind speed were used as data in this study. The five water quality and two meteorological data were used as the inputs and outputs for the two stochastic models. The descriptive statistics for the data are shown in Table 1, including the number of data, minimum, maximum, mean, and standard deviation. As mentioned above, a wider range of water quality parameters was observed in the YSR, compared to the JAR.

2.2. Theoretical background of applied stochastic models

2.2.1. Artificial neural network

ANN is a useful method for classifying patterns of multi-variable datasets and modeling complex environmental processes (Cho et al., 2011). The structure of ANN consists of two or more layers of nodes, including an input layer, hidden layer, and output layer that are connected by links having varying weights. The nodal data are multiplied by the weights to compute the signal strength, and are then transferred to the next node in the network; the input layer nodes accept the input vectors and forward the signals to the next layer according to the connection. This process is continued until the signals reach the output layer. Note that a back propagation learning algorithm was applied in this study to minimize the objective function. ANN consists of three layers having p input nodes ($x^i_1, x^i_2, \ldots, x^i_{p=8}$), q hidden nodes ($H^i_1, H^i_2, \ldots, H^i_q$), and one output node ($g^i_1$) (see Fig. S1 in the Supplementary Material). These structural components were applied to develop ANN, and followed a generalized mathematical expression (Norgaard et al., 2000). Hidden node outputs (H^i_q) were determined using Eq. (1) based on a transfer function (f_1) associated with the input elements (x^i), weight (w^h_{pq}), and bias (b^q_1), and the final network output (g^i_1) was then calculated from the hidden node output (H^i_q) using the transfer function (f_2) having a connection weight (w^o_{q1}) and bias (b^1_2).