PREDICTION OF FRESHWATER HARMFUL ALGAL BLOOMS IN WESTERN LAKE ERIE USING ARTIFICIAL NEURAL NETWORK MODELING TECHNIQUES

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Haci Osman Guzel

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Natural Resource Management

December 2018

Fargo, North Dakota

North Dakota State University

Graduate School

Title

Prediction of Freshwater Harmful Algal Blooms in Western Lake Erie

Using Artificial Neural Network Modeling Techniques

By

Haci Osman Guzel

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Halis Simsek

Chair

Dr. Chiwon Lee

Dr. Kenneth Hellevang

Approved:

12/19/2018

Date

Shawn DeKeyser

Department Chair

ABSTRACT

Blue-green algae are a major environmental concern in freshwater produce toxins and cause a wide range of problems including oxygen depletion, fish kills, harm or death to other aquatic organisms, and subsequent habitat loss. Cyanobacteria are a type of blue-green algae that form harmful algal blooms (HABs) in water ecosystems. In this study, artificial intelligence techniques, in particular artificial neural networks, were developed to estimate blue-green algae fluorescence for the year-round data collected in 2016-17 from western Lake Erie, USA. Based on the lake's environmental conditions and available data, eight input parameters including phosphorous, nitrogen, chlorophyll-a, air temperature, water temperature, turbidity, wind speed, and pH were used to run the model. Five different learning algorithms were TESTED, and the Levenberg-Marquardt algorithm resulted in the highest R^2 values of 0.98 and 0.72 for eight, and three (phosphorous, nitrogen, and chlorophyll-a) input parameters, respectively. Eight input parameters produced the best estimation approach.

ACKNOWLEDGEMENTS

Foremost, I would first like to thank my thesis advisor Dr. Halis Simsek. The door to Dr. Simsek office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work but steered me in the right the direction whenever he thought I needed it.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Chiwon Lee, and Dr. Kenneth J. Hellevang for their encouragement and insightful comments.

My sincere thanks also go to Prof Senay Simsek who gave access to the laboratory and research facilities. I offer my endless thanks to Dr. Bilal Cemek from Ondokuz Mayis University in Turkey, for his precious support during my modeling study. I am profoundly grateful to my colleagues Yavuz Fatih Fidantemiz, Hakan Kadioglu and Ekrem Ergun who embellish my life in the United States.

I gratefully acknowledge the financial support provided by the Turkish Government, Ministry of Education (YLSY program) for my master's degree scholarship. Additionally, I would like to thank North Dakota Water Resource Research Institute (NDWRRI) for providing research funding for this study. The data for this study was collected from the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) (NOAA-GLERL, 2018). Last but not the least, I would like to thank my parents who support me in every part of my life.

DEDICATION

I dedicate this thesis to my wife Gulcin Guzel, and to my son Omer Halis Guzel.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN.....Artificial Neural Network

BP.....Back-Propagation

BR.....Bayesian Regularization

CFG.....Conjugate Gradient Function

GLERL.....Great Lakes Environmental Research Laboratory

LM.....Levenberg-Marguard

MAE.....Mean Absolute Error

MBE.....Mean Bias Error

MLP .....Multilayer Perceptron

NDAES .....North Dakota Agricultural Experiment Station

NDWRRI .....North Dakota Water Resource Research Institute

NOAA .....National Oceanic and Atmospheric Administration

RBP.....Resilient Back-Propagation

RMSE.....Root Means Square Error

SCG.....Scaled Conjugate Gradient

1. INTRODUCTION

Excess nutrients in freshwater environments stimulate blue-green algae, which rapidly increase and accumulate in lakes and rivers when the optimum environmental conditions are met. Cyanobacteria are a type of prokaryotic blue-green algae that can form harmful algal blooms (HABs) in water ecosystems and sometimes called CyanoHABs (O'Neil et al., 2012; Paerl et al., 2015). Some cyanobacteria genera, including *Microcystis spp., Planktothrix spp, Anabeana spp., Cylindrospermopsis spp., Aphanizomenon spp.,* and *Oscillatoria spp.*, in freshwater, are able to produce cyanobacterial metabolites and toxins (cyanotoxins). The cyanotoxins have been found to be causes of animal and human poisonings and may have lethal effects on aquatic organisms (Ferreira et al., 2001; Anderson et al, 2002; Mohamed and Shehri, 2010; O'Neil et al. 2012, Li et al., 2016).

The occurrence of HABs in freshwater increases the risk to human and animal health, reduces water transparency, creates oxygen-deprived aquatic zones, can cause taste and odor problems in drinking water, leads to death of plants and fishes, effects biodiversity, and decreases the recreational use of water (Carpenter et al., 1998; Smith, 1998; Hudnell et al., 2010). HABs are especially dangerous in a water body if the water is used as a municipal drinking water reservoir where possible cyanotoxins are piped into people's home and used for drinking, cooking, bathing, and other household chores.

Nitrogen (N) and phosphorus (P) are macro-nutrients required for growth by the photosynthetic cyanobacteria that make up HABs in freshwater ecosystem. Nutrient over-enrichment originated by human activity increases the HAB occurrence and can lead to eutrophication which has long been cited as a major cause of HABs. This abundance of nutrients has been linked to human activities, including agricultural and residential uses of fertilizer,

application of manure, discharge of municipal wastewaters, and inputs from industries (Anderson, 2009). Although P is a required macro nutrient in photosynthetic organisms' growth, it exists in small amounts in most freshwaters (Anderson et al., 2002). Some species of cyanobacteria are capable of providing their own N via N_2 fixation; therefore, P is the more limiting nutrient for controlling HABs. Besides nutrients, climatic factors also contribute to HABs. HAB proliferation was observed in regions where the temperature exceeds the optimal growth temperature, which is 25 °C (Paerl et al., 2011).

Intensification of HAB in freshwaters is not a simple process caused by a single event but rather multiple factors occurring simultaneously (Heisler et al., 2008). Innovative approaches are needed to prevent HAB occurrence, accumulation, and transport in freshwaters. The characteristics of freshwaters (lakes, rivers, streams, and reservoirs) are varied based on their hydrologic, geographic, climatic, morphologic, physical, chemical, geochemical, and biological features. Therefore, HAB control methods will be different in each water environment. For instance, controlling HABs in large water bodies are difficult, whereas control of HABs may be more manageable in small water environments such a waste ponds. External nutrient loading is usually the first target to control and prevent HABs in freshwaters even though limiting the nutrients might not be a solution in the near future (Hudnell et al., 2010).

Development of a HAB early-warning system is highly dependent on reliable modeling methods that predict the HAB occurrence with high accuracy using current water and climate conditions and forecasts. Early warning systems provide practical guidance for water treatment plants about future lake contamination by cyanobacteria. In addition, early-warning systems provide critical knowledge for agencies, water utility managers and other stakeholders to prevent future hazards caused by algal toxin. In order to minimize the impact of HABs in aquatic systems,

the past and current situations and upcoming forecast should be evaluated using an appropriate model. Consideration of available data and sampling or scientific efforts are necessary for selecting the type of model to estimate HABs (in terms of blue-green algal fluorescence) in freshwater. The most common parameters used for modeling in rivers and streams are nutrient loading, water temperature, volumetric flow rate, water current and turbulence, water residence times, sunlight exposure, time, and intensity, quiescent or stagnant water, and depth of the water (deep or shallow).

Artificial intelligence techniques, in particular artificial neural network (ANN) techniques, have been extensively used in a variety of complex scientific and engineering problems to predict and classify environmental systems including system modeling, forecasting, hydrology, pattern recognition, sediment transport and accumulation, evaporation, evapotranspiration, rainfall, surface runoff, and watershed runoff (Holmberg et al., 2006; Paliwal and Kumar, 2009; Cobaner, 2011; Amiryousefi et al., 2011; Simsek et al., 2015). However, the application of these techniques to HAB estimation is very limited in the literature. Therefore, the objective of this study is to apply the ANN techniques, in particular multilayer perceptron (MLP) models to estimate blue-green algae in western Lake Erie, USA. MLP is a form of ANN modeling that consists of single-layer perceptron. External data in an MLP model is collected by the input layer which is known as the first layer. All existing datasets are randomly divided into a training (sample) and a testing (non-sampling) dataset. Back-propagation (BP) is accepted as a prevalent learning technique for MLP when obtained a training data set.

2. MATERIAL AND METHODS

2.1. Preliminary Study and Data Collection Strategy

A combination of eight input parameters, including phosphorus (μg/L), nitrogen (mg/L), chlorophyll-*a* (RFU), air temperature (℃), water temperature (℃), turbidity (NTU), wind speed (m/s), and pH were used in this study to estimate blue-green algae fluorescence in relative fluorescence units (RFU) in western Lake Erie, USA (Table 1). Blue-green algae fluorescence (a proxy for the occurrence of HABs and indicated by the HAB acronym in model output) may be used as a proxy for measure the cyanobacterial abundance of HAB that may turn toxic and is determined with a phycocyanin probe in the water or through satellite data. Optical phycocyanin sensors have provided early warnings of increased cyanobacteria abundance or elevated toxin concentrations (Brient et al., 2008; Marion et al., 2012; McQuaid et al., 2011) and have been used successfully in Lake Erie (Francy et al., 2016). Phycocyanin data from satellites are increasingly more accurate than chlorophyll-*a* data in the prediction of HABs (Yan et al., 2018).

All the input parameters were determined based on the lake's environmental conditions and the data availability. The data were collected real-time in the period of from June 30 to October 5 in 2016 and from May 1 to October 26 in 2017, by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) (NOAA-GLERL, 2018). The GLERL website runs a collaborative program, which uses data sharing to understand the environmental factors of HABs. To understand the long and short-term periodic changes in HAB occurrence, the data was collected using satellite images, remote sensing techniques, buoys, and an exhaustive observation and sample collection program in Lake Erie during the algal bloom season. The data was saved using the PostgreSQL database management system, which is a powerful, open source object-relational database system. To develop an MLP

model in this study, a large number of input data sets, which were about 13,300 data points from each parameter were processed to run the MLP model and the statistical analyses are presented as a supplementary document at Table S1. This table presents the distribution of the values of eight input parameters for training-only, testing-only and for all the data.

2.2. Model Development

The ANN model uses computer-based algorithms that can be trained to identify and classify complex patterns (Khan et al., 2001). The models have an input layer, hidden layer(s) and an output layer. All the computations are made in the hidden layers. Training, testing, and validation processes (machine learning systems) are used to confirm the models' performance (Takagi and Sugeno, 1985; Simsek, 2016). ANNs are classified according to the number of layers, nodes in each layer, and the way these nodes are connected to each other (Zhang et al., 1999). The network forms the model formula in the output layer, which is the last layer. Hidden layers are crucial for ANNs to define the complicated model data between the input and output layers. All the nodes in these layers are connected to each other from the lowest layer upwards (Zhang et al., 1999).

Completely connected, feed-forward BP neural network models were used in the ANN network with five different learning algorithms including Levenberg-Marquard (LM), Bayesian regularization (BR), conjugate gradient function (CGF), resilient back-propagation (RBP), and scaled conjugate gradient (SCG). A BP algorithm is a graphical approach that is used in ANNs to calculate a gradient of the error functions. A BP algorithm is commonly used to optimize the feed forward neural networks. A typical architecture of MLP structure is presented in Fig. 1.

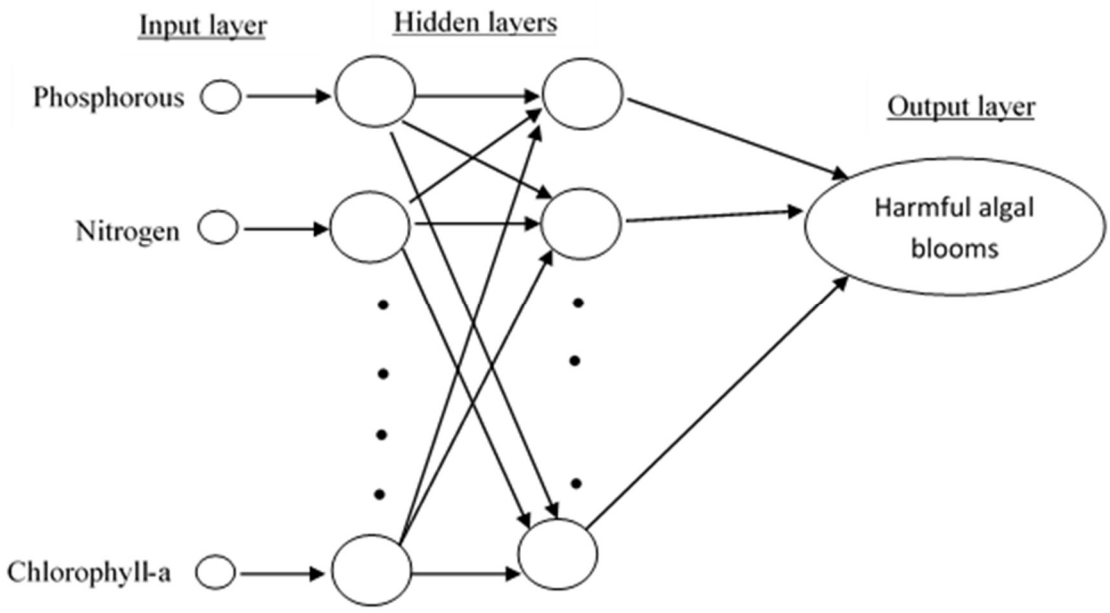

Figure 1. A schematic diagram of multilayer perceptron (MLP) structure with inputs, hidden layers and an output layer.

The BP network has a simple structure with a strong simulation capability and consists of two phases, which are a feed forward and backward phases. The feed forward phase sends external input information forward to the output node, and the second phase arranges to the connection strengths according to the discrepancy between the calculated and viewed information at the output unit (Cigizoglu and Alp, 2006; Goh, 1995). In BP neural networks, the mathematical relationships between the variables are not specified. Instead, they learn from the examples fed to them. Since there is no mathematical connection between the variables, BP neural networks learn from cases that they obtained.

The LM algorithm is a variation of Newton's method and derives from the error BP algorithm (Lourakis, 2005; Suratgar et al., 2007). The LM algorithm identifies the minimum function denoted as the sum of the squares of non-linear functions (Lourakis, 2005). Several approaches could be used in the LM method to accelerate the error BP algorithm, but most of these methods achieved minimally acceptable results in the literature. Even though LM has a high-speed

algorithm, it is not capable of minimizing error oscillation. Nevertheless, only the LM algorithm provides a fair exchange between the speed of the Newton algorithm and the determination of the steepest descent method (Suratgar et al., 2007).

Gradient-based learning methods are used as error reducing techniques to train BP nets (Bayati et al., 2009). BR is a mathematical technique that is improved to transform non-linear systems into ‘‘well posed’’ problems to minimize the potential for overfitting which causes a deficiency of generalization of the network (Saini, 2008).

RBP is a learning technique, which makes a direct adjustment of the weight step based on local gradient information. In RBP, it’s adaptation is not blurred by gradient behavior and it is almost 100 times faster than the simple BP technique because it depends on the sign of the derivative instead of the value of the derivative (Naoum et al., 2013; Saini, 2008). CGF, which uses orthogonal and linearly independent non-zero vectors, can be used as a method to reduce the network output error in conjugate directions (Man-Chung et al., 2000). SCG belongs to the class of conjugate gradient methods. SCG is faster than second order algorithms since it uses a step size scaling mechanism, which runs quickly for line-search per learning iteration (Orozco and García, 2003).

In order to explain the performance of training, testing and validation processes, some statistical calculations are necessary such as root mean square error (RMSE), mean absolute error (MAE), mean bias error (MBE), and coefficient of determination (R^2). The RMSE describes a short-term performance of a model by ensuring each unit compares to the real difference between the estimated value and the obtained value (Sanusi et al., 2013). The MBE describes the long-term behavior of a model, and at positive value indicates the average overestimate of the predicted value, whereas a negative value indicates the average underestimate of the predicted value

(Jacovides and Kontoyiannis, 1995; Sanusi et al., 2013). The RMSE value is expected to be as small as possible for a better result, similar to an MBE value (Sanusi et al., 2013). MBE, RMSE and MAE can be calculated using Eqs. 1, 2 and 3, respectively.

$$\text{MBE} = \frac{\sum_{i=1}^{n}(pi-ri)}{n} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(pi-ri)^2}{n}} \tag{2}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\text{pi}-\text{ri}|}{n} \tag{3}$$

Where, i is an index; p_i is the predicted value for i^{th} datum; r_i is the real value for i^{th} datum; and n is the observation number or sample size (Sanusi et al., 2013). The definition of the data set consists of a sequence of operations; the transmission functions are first assigned to a network layer to identify the input signals, and then the appropriate weight is calculated for the output signal. Logsig, tansig and purelin are the linear transfer functions that are used commonly in Matlab software. According to the ranges of these transfer functions, input and output data are normalized (Mohamed Ismail et al., 2012). The formula used for these three functions are presented in Eqs. 4, 5, and 6.

$$\text{Logsig}(\text{n}) = \frac{1}{1+\text{e}^{-\text{n}}} \tag{4}$$

$$\text{Tansig}(\text{n}) = \frac{2}{(1+\text{e}^{(-2\text{n})})-1} \tag{5}$$

$$\text{Purelin}(\text{n}) = \text{n} \tag{6}$$

Five different MLP models were designed to estimate blue-green algae accumulation based on fluorescence values (Table 1). The first four models contained air temperature and water

temperature data since these two parameters are important factors in the lake environment that promote algal growth (Fu et al., 2012; Wei et al., 2001).

Table 1. MLP models for estimation of blue-green algae fluorescence, a proxy for a harmful algal bloom (HAB).

Model	İnput								Output
	1	2	3	4	5	6	7	8	
MLP1	air temperature	water temperature	wind speed	pH	turbidity	chl-a	phosphorus	nitrogen	HAB
MLP2	air temperature	water temperature	wind speed	pH	turbidity				HAB
MLP3	air temperature	water temperature	wind speed	pH					HAB
MLP4	air temperature	water temperature	wind speed						HAB
MLP5	phosphorus	Nitrogen	chl-a						HAB

Note: The units are: phosphorus, micrograms per liter (µg/L), nitrogen milligrams per liter (mg/L), chl-a (chlorophyll-a, relative fluorescence units, RFU), air temperature (℃), water temperature (℃), turbidity nephelometric turbidity units (NTU), wind speed (m/s), and HAB stands for blue-green algae fluorescence (RFU). MLP: Multilayer perceptron.

Table 2. The network structure used in MLP models for both training and testing data sets.

Model	Network structure			
MLP1	8-10-1	8-12-1	8-15-1	8-10-15-1
MLP2	5-7-1	5-9-1	5-7-9-1	-
MLP3	4-5-1	4-7-1	4-5-7-1	-
MLP4	3-5-1	3-7-1	3-5-7-1	-
MLP5	3-5-1	3-7-1	3-5-7-1	-

Note: MLP represents the number of input parameters and the last number, which is 1, represents the output parameter. The other one or sometimes two numbers between first and second numbers: Multilayer perceptron. The table explains five different MLP models with their network structures. First number are the hidden layer structures.

All five MPL models were divided into their network structure as presented in Table 2. There were only 8, 5, 4, and 3 different inputs applied in this study. Some of the network structures

had one hidden layer, whereas others had more. Commonly, MLP models contain several layers of neurons in their network structure and each neuron receives input data. The input layer does not have any mission about calculation or computation in the neural structure, its role is transferring the input vector to the network vector. The input and output vectors in the system represent the inputs and the output of the MLP models and they can be represented as single vectors (Gardner and Dorling, 1998).

3. RESULT AND DISCUSSION

Blue-green algae fluorescence was estimated using five different MLP models (Table 1) with various network structures (Table 2) in each model and only the best estimation models were presented in this study. The lake parameters for the MLP models were selected based on the lake's environmental, ecological, and climatic conditions. Among all the MLP models, the highest R^2 values (≤0.98) for both training and testing data sets were obtained by MLP1 model, which used eight input parameters to stimulate HAB occurrence in Lake Erie as shown in the Fig. 2a and b and in Table 3. The best learning algorithm was LM and the best network structure was 8-10-15-1 for the eight input parameters. The best ANN transfer functions of tansig-tansig-purelin for both training and testing data sets were also observed in 8-10-15-1 network structures. The detail training and testing results for eight input parameters for MLP1 models are presented at Table S2 and S3. In general, good performance was achieved as indicated by small values of RMSE, MBE, and MAE as well as large values of R^2 (Jacovides and Kontoyiannis, 1995). These results showed that we were able to forecast blue-green algae fluorescence with MLP models, which could lead to early mitigation and thus reduce human health risks and ecological effects of toxic algae.

Among all the MLP models, the highest coefficient of determination values (≤0.98) for both training and testing data sets were obtained at MLP1 model, which used eight input parameters to stimulate HAB occurrence in Lake Erie as shown in the Figure 2a and b and in the Table 3. The best learning algorithm was LM and the best network structure was 8-10-15-1 for the eight input parameters. The best ANN transfer functions of tansig-tansig-purelin for both training and testing data sets were also observed in 8-10-15-1 network structures. The detailed of training and testing results for eight input parameters for MLP1 models are presented at Table S2 and S3. In general, good performance was achieved as indicated by small values of RMSE, MBE, and

MAE as well as large values of R^2 (Jacovides and Kontoyiannis, 1995). These results showed that we were able to forecast blue-green algae fluorescence with MLP models, which could lead to early mitigation and thus reduce human health risks and ecological effects of toxic algae.

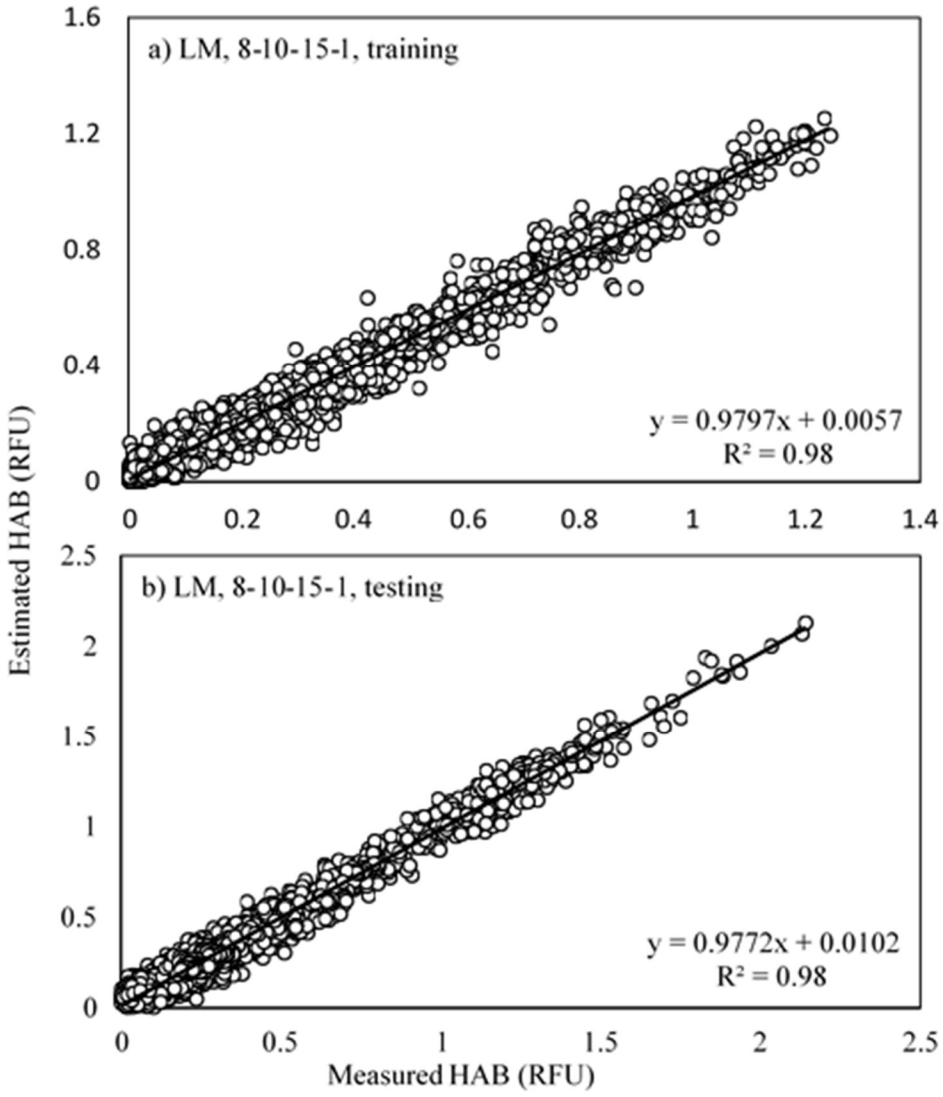

Figure 2. MLP1 model for measured and estimated blue-green algae fluorescence, as a surrogate for harmful algal blooms (HAB) (a) training and (b) testing data sets (MLP: Multilayer perceptron, LM: Levenberg-Marquardt).

Table 3. The summary of MLP1, MLP2 and MLP3 model statistics for training and testing data set.

`

Network structure	Activation	Learning algorithm	Training				Testing			
			MBE	MAE	RMSE	R^2	MBE	MAE	RMSE	R^2
8-10-15-1	**Tansig-tansing-purelin**	**LM**	**0 0.02**	**0.03**	**0.98**	**0**	**0.04**	**0.05**	**0.98**	
8-10-15-1	Tansig-logsig-purelin	BR	0	0.02	0.03	0.98	0	0.04	0.05	0.98
8/15/2001	Tansig-purelin	CGF	0	0.05	0.07	0.91	-0.01	0.08	0.11	0.91
8/12/2001	Logsig-purelin	RP	0	0.06	0.07	0.91	-0.01	0.08	0.11	0.9
8/15/2001	Logsig-purelin	SCG	0	0.05	0.07	0.92	-0.01	0.07	0.1	0.91

MLP2

Network structure	Activation	Learning algorithm	Training				Testing			
			MBE	MAE	RMSE	R^2	MBE	MAE	RMSE	R^2
5-7-9-1	**Logsig-tansig-purelin**	**LM**	**-0.01**	**0.06**	**0.09**	**0.88**	**0.01**	**0.08**	**0.12**	**0.89**
5-7-9-1	Tansig-logsig-purelin	BR	0	0.06	0.08	0.88	0.01	0.08	0.12	0.89
5/9/2001	Logsig-purelin	CGF	-0.02	0.09	0.14	0.71	0.05	0.13	0.18	0.81
5-7-9-1	Logsig-tansig-purelin	RP	-0.02	0.08	0.12	0.79	0.05	0.13	0.19	0.77
5/7/2001	Tansig-purelin	SCG	-0.02	0.09	0.13	0.74	0.04	0.13	0.19	0.79

MLP3

Network structure	Activation	Learning algorithm	Training				Testing			
			MBE	MAE	RMSE	R2	MBE	MAE	RMSE	R2
4-5-7-1	**Logsig-tansig-purelin**	**LM**	**-0.01**	**0.09**	**0.12**	**0.76**	**0.02**	**0.1**	**0.15**	**0.82**
4-5-7-1	Logsig-tansig-purelin	BR	-0.01	0.09	0.12	0.75	0.03	0.12	0.17	0.78
4-5-7-1	Logsig-tansig-purelin	CGF	-0.02	0.11	0.16	0.61	0.04	0.15	0.21	0.68
4-5-7-1	Tansig-tansing-purelin	RP	-0.01	0.1	0.15	0.62	0.03	0.14	0.2	0.72
4-5-7-1	Logsig-tansig-purelin	SCG	-0.01	0.11	0.15	0.65	0.03	0.14	0.2	0.72

Note: [LM, Levenberg-Marquard; BR, Bayesian regularization; CGF conjugate gradient function; RP, resilient backpropogation; SCG, scaled conjugate gradient; MBE, mean bias error; MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination]. MLP: Multilayer perceptron. Bold numbers were selected as the best results and their figures were presented in this study. Bold numbers were selected as the best results and their figures were presented in this study.

Even though eight input structures showed the best estimation of blue-green algae fluorescence, this model might not be feasible in real-world applications since it will be time consuming and costly to obtain all eight parameters. Hence, three, four, and five input parameters were tested as well in this study.

Table 3 shows the best estimation of MLP modeling results for five and four input parameters (MLP2 and MLP3) using five different transfer functions. Five different learning algorithms were applied, and the best estimations of blue-green algae fluorescence were obtained at LM and BR algorithm with 0.89 R^2 values in both algorithms. Only the selected algorithms were presented in Table 3 and Figure 3 for both training and testing data sets. LM algorithm was one of the fastest medium-sized feedforward algorithms with a set of simple interconnected units (neurons or nodes) (Karul et al., 2000). Five input parameters produced little better estimation with R^2 values of 0.88 for training data sets compared to four input parameters which produced 0.76 R^2 values for training data sets.

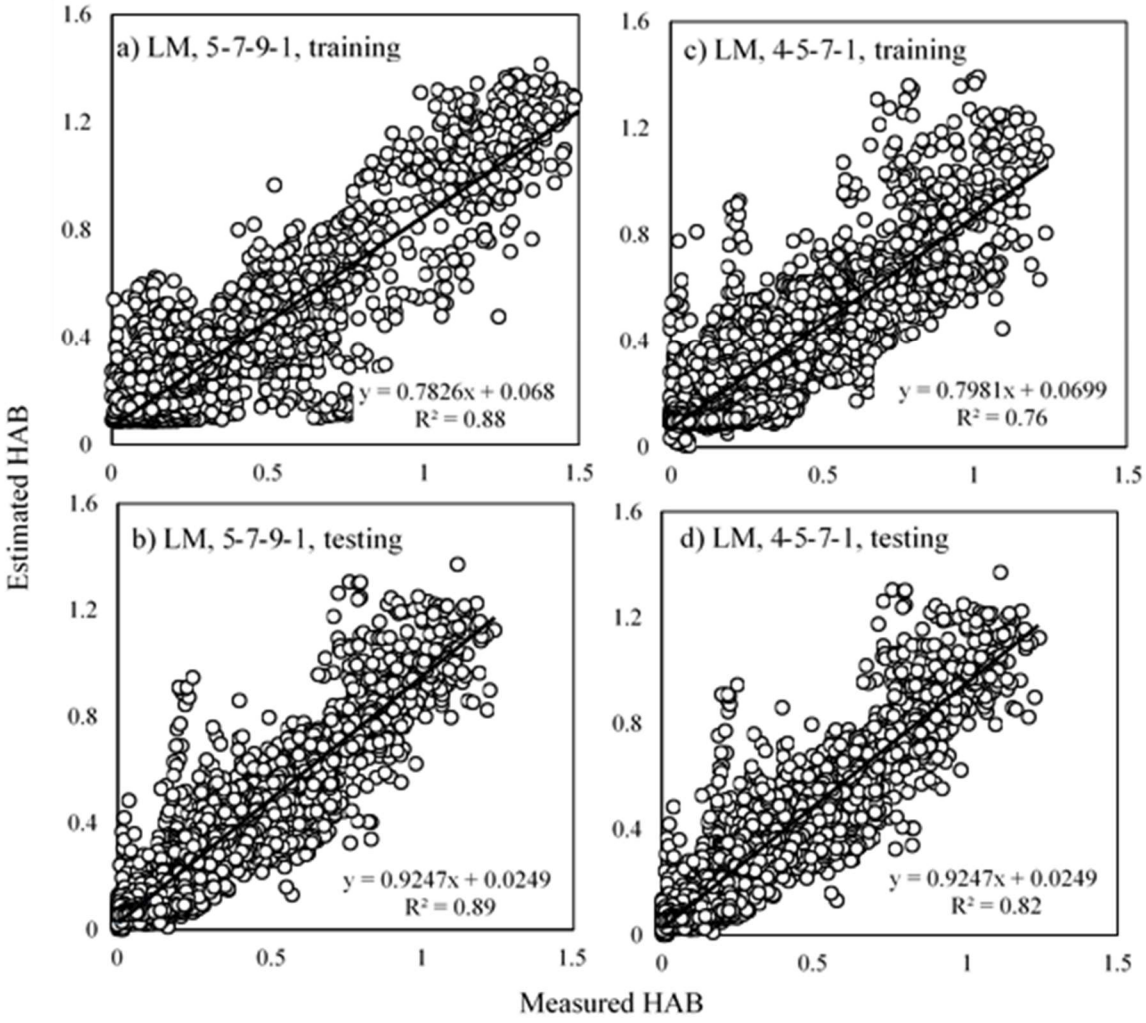

Figure 3. MLP2 model (a) training and (b) testing data sets; MLP3 model (c) training and (d) testing data sets (MLP: Multilayer perceptron, LM: Levenberg-Marquardt).

Two sets of three input parameters were designed (MLP4 and MLP5) to determine the best blue-green algae fluorescence estimation although the network structures and transfer algorithms used are the same (Table 4). In MLP4, the input factors are air temperature, water temperature and wind speed while in MLP5, the input factors are phosphorus, nitrogen and chl-*a*. In both models, the training and testing modeling results for the LM and BR learning algorithms are similar and only the LM algorithm was presented in Fig. 4 for both MLP4 and MLP5 models. However, the input parameters for MLP5 (nutrients and chlorophyll concentrations) are vital since nitrogen and

phosphorous are essential nutrient sources for HAB formation and they are essential to the productivity of HABs in aquatic ecosystem. Optimal amounts of nutrients are important to support aquatic life; however, in high concentrations they can be detrimental. This is supported in the research where natural and/or anthropogenic nutrient over enrichment of a water body increased algal abundance (Paerl and Huisman, 2009). Abundance of cyanobacteria, chlorophytes, and cryptophytes increased after nutrient addition to (Lake Taihu, China); whereas diatoms showed a slower abundance response than the other algal groups (Paerl et al., 2015).

Optimal amounts of nutrients are important to support aquatic life; however, in high concentrations they can be detrimental. This is supported in the research where natural and/or anthropogenic nutrient over enrichment of water body promotes proliferation of HABs (Paerl and Huisman, 2009); different type of HABs including cyanobacteria, chlorophytes, and cryptophytes grew well under nutrient addition to a lake (Lake Taihu, China); whereas diatoms were moderately stimulated by the nutrient loading (Paerl et al., 2015).

Table 4. The summary of MLP4 and MLP5 model statistics for training and testing data set.

MLP4

Network structure	Activation	Learning algorithm	Training				Testing			
			MBE	MAE	RMSE	R2	MBE	MAE	RMSE	R2
3-5-7-1	**Tansig-logsig-purelin**	**LM**	**0.01**	**0.09**	**0.13**	**0.7**	**-0.02**	**0.12**	**0.16**	**0.81**
3-5-7-1	Tansig-logsig-purelin	BR	0.01	0.09	0.13	0.71	-0.01	0.12	0.16	0.81
3-7-1	Tansig-purelin	CGF	-0.02	0.13	0.18	0.73	-0.02	0.13	0.18	0.73
3-5-1	Tansig-purelin	RP	-0.02	0.14	0.19	0.73	-0.02	0.14	0.19	0.73
3-5-1	Tansig- purelin	SCG	-0.02	0.13	0.17	0.76	-0.02	0.13	0.17	0.76

MLP5

Network structure	Activation	Learning algorithm	Training				Testing			
			MBE	MAE	RMSE	R^2	MBE	MAE	RMSE	R^2
3-5-7-1	**Tansig-tansing-purelin**	**LM**	**-0.01**	**0.13**	**0.18**	**0.46**	**0.03**	**0.12**	**0.19**	**0.72**
3-5-7-1	Tansig-logsig-purelin	BR	-0.01	0.13	0.18	0.45	0.03	0.12	0.19	0.72
3-5-7-1	Tansig-tansing-purelin	CGF	-0.02	0.15	0.21	0.29	0.04	0.17	0.24	0.57
3-7-1	Logsig-purelin	RP	-0.02	0.14	0.2	0.37	0.04	0.14	0.21	0.67
3-5-7-1	Tansig-tansing-purelin	SCG	-0.02	0.15	0.21	0.31	0.04	0.17	0.24	0.6

Note: [LM, Levenberg-Marquard; BR, Bayesian regularization; CGF conjugate gradient function; RP, resilient backpropogation; SCG, scaled conjugate gradient; MBE, mean bias error; MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination]. MLP: Multilayer perceptron. Bold numbers were selected as the best results and their figures were presented in this study.

The coefficient of determination (R^2) values of the MLP5 model (nutrient and chlorophyll inputs) were low (0.46 and 0.72) in training and testing data sets, respectively. Overall, the amount of phosphorous in the lake was low, in microgram per liter level. Out of 13,300 data points for phosphorous (concentrations), the values of about 900 data points were less than 1.0 µg/L, (Fig. S1). About 7,100 phosphorous data points were under 10.0 µg/L and there were only 450 data points in between 100 and 146 µg/L. Similarly, nitrogen values were also low in the lake, however at least they were at the mg/L level. The distribution of nitrogen data was as follows; the nitrogen concentration of about 4,880 data points were under 0.5 mg/L, about 2120 data points were in between 0.5 and 1.0 mg/L, and about 6,300 data points were in between 1.0 and 4.8 mg/L. Since blue-green fluorescence and nutrient concentration in the lake has a negative relationship, it would be expected to measure low phosphorous and nitrogen concentrations in the lake when blue-green algae fluorescence is high. When the data were analyzed based on the summer season, the

concentrations of phosphorous and nitrogen parameters decreased through the end of the summer (September and October) in both years (2016 and 2017) even though the concentrations of these two nutrients fluctuate during the beginning and middle of the summer. A better understanding of lake-wide nutrient input and its utilization by HABs or other organisms could be determined to create more accurate prediction results.

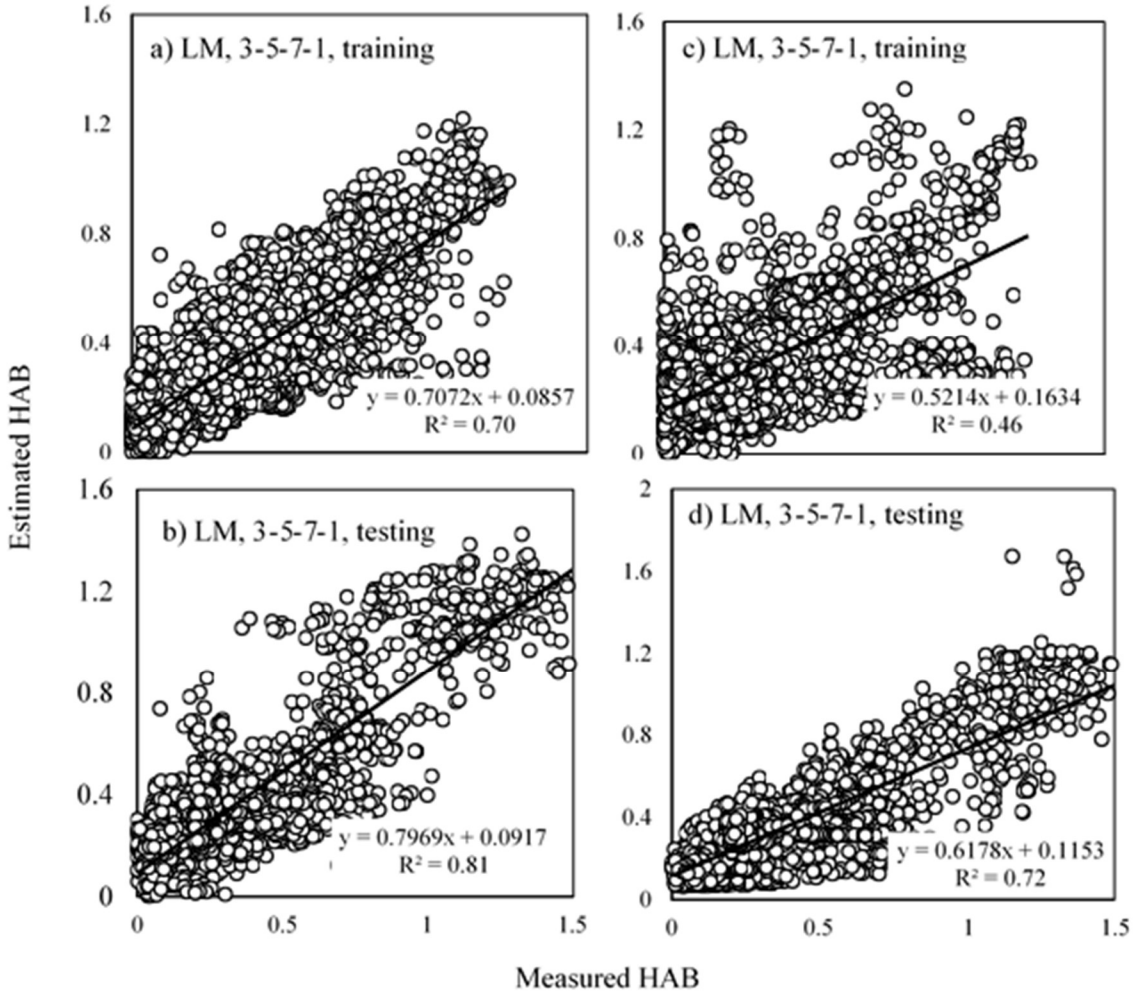

Figure 4. MLP4 model (a) training and (b) testing; MLP5 model (c) training and (d) testing data sets estimating of blue-green algae fluorescence as a surrogate for harmful algal blooms (HAB) (MLP: Multilayer perceptron, LM: Levenberg-Marquardt).

4. CONCLUSIONS

In this study, five different Multilayer perceptron (MLP) models with five different activation functions including LM, BR, CGF, RP, and SCG were developed to estimate blue-green algae fluorescence as a surrogate for the occurrence of HABs in western Lake Erie. The best estimation of blue-green algae fluorescence was achieved using eight different input parameters, which were phosphorus, nitrogen, chlorophyll-*a*, air temperature, water temperature, turbidity, wind speed, and pH (MLP1 model). Two of the models, MLP3 and MLP4 proved that the blue-green algae occurrence in the lake could be predicted quickly and cost effectively with simple field measurements of air and water temperature, wind speed and pH (MLP4 only). Therefore, using only these two models could help to create an early warning system to indicate the likelihood of a HAB more efficiently and cost effectively than MLP1.

Phosphorus, chlorophyll-*a*, and nitrogen input parameters provided weak correlation (MLP5) even though, nutrients and algal proliferation tend to correlate in many systems (Carpenter et al., 1998). However, having more than 2 years' data might give better estimation using nutrient parameters. Overall, the ANN modeling approach described here proved that, developing and implementing MLP models to provide accurate forecasting of blue-green algae fluorescence depends on appropriate and representative data measurements in the lake environment. Determining physical, ecological, biological and chemical parameters of the lake would improve the forecast capability of the model. Harmful algal blooms are a growing concern for lake management and estimating blue-green algae fluorescence as a surrogate for the occurrence of HABs can be a key planning element for lake environment and hydrological studies; use of neuro computing techniques offer new opportunities for rapid estimation of HABs in freshwaters.

REFERENCES

Amiryousefi, M.R., Mohebbi, M., Khodaiyan, F., Asadi, S., 2011. An empowered adaptive neuro-fuzzy inference system using self-organizing map clustering to predict mass transfer kinetics in deep-fat frying of ostrich meat plates. Comput. Electron. Agric. 76, 89–95. https://doi.org/10.1016/j.compag.2011.01.008.

Anderson D.M., Glibert PM, Burkholder JM., 2002. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25(4b):562–84.

Anderson D.M., 2009. Approaches to monitoring, control and management of harmful algal blooms (HABs). Ocean Coastal Manage. 52 (7), 342–347.

Bayati, A.Y. Al, Sulaiman, N.A., Sadiq, G.W., 2009. A modified conjugate gradient formula for back propagation neural network algorithm 5, 849–856.

Brient, L., Lengronne, M., Bertrand, E., Rolland, D., Sipel, A., Steinmann, D., Baudin, I., Legeas, M., Le Rouzic, B., Bormans, M., 2008. A phycocyanin probe as a tool for monitoring cyanobacteria in freshwater bodies. J. Environ. Monit. 10, 248–255. https://doi.org/10.1039/b714238b.

Carpenter, S.R., Caraco, N.F., Correll, D.L., W.Howarth, R., Sharpley, A.N., Smith, V.H., 1998. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. 8, 559–568. https://doi.org/10.1890/1051-0761(1998)008[0559:NPOSWW]2.0.CO;2.

Cigizoglu, H.K., Alp, M., 2006. Generalized regression neural network in modelling river sediment yield. Adv. Eng. Softw. 37, 63–68. https://doi.org/10.1016/j.advengsoft.2005.05.002.

Cobaner, M., 2011. Evapotranspiration estimation by two different neuro-fuzzy inference systems. J. Hydrol. 398, 292–302. https://doi.org/10.1016/j.jhydrol.2010.12.030.

Ferreira, F.M.B., Soler, J.M.F., Fidalgo, M.L., Fernández-Vila, P., 2001. PSP toxins from Aphanizomenon flos-aquae (cyanobacteria) collected in the Crestuma-Lever reservoir (Douro river, northern Portugal). Toxicon 39, 757–761. https://doi.org/10.1016/S0041-0101(00)00114-8.

Francy, D.S., Brady, A.M.G., Ecker, C.D., Graham, J.L., Stelzer, E.A., Struffolino, P., Dwyer, D.F., Loftin, K.A., 2016. Estimating microcystin levels at recreational sites in western Lake Erie and Ohio. Harmful Algae 58, 23–34. https://doi.org/10.1016/j.hal.2016.07.003

Fu, F.X., Tatters, A.O., Hutchins, D.A., 2012. Global change and the future of harmful algal blooms in the ocean. Mar. Ecol. Prog. Ser. 470, 207–233. https://doi.org/10.3354/meps10047.

Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron): a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9, 143–151. https://doi.org/10.1016/0954-1810(94)00011-S

Heisler, J., Glibert, P.M., Burkholder, J.M., Anderson, D.M., Cochlan, W., Dennison, W.C., Dortch, Q., Gobler, C.J., Heil, C.A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H.G., Sellner, K., Stockwell, D.A., Stoecker, D.K., Suddleson, M., 2008. Eutrophication and harmful algal blooms: A scientific consensus. Harmful Algae 8, 3–13. https://doi.org/10.1016/j.hal.2008.08.006.

Holmberg, M., Forsius, M., Starr, M., Huttunen, M., 2006. An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. Ecol. Modell. 195, 51–60. https://doi.org/10.1016/j.ecolmodel.2005.11.009.

Hudnell, H.K., Jones, C., Labisi, B., Lucero, V., Hill, D.R., Eilers, J., 2010. Freshwater harmful algal bloom (FHAB) suppression with solar powered circulation (SPC). Harmful Algae 9, 208–217. https://doi.org/10.1016/j.hal.2009.10.003.

Jacovides, C.P., Kontoyiannis, H., 1995. Statistical procedures for the evaluation of evapotranspiration computing models. Agric. Water Manag. 27, 365–371. https://doi.org/10.1016/0378-3774(95)01152-9.

Karul, C., Soyupak, S., Çilesiz, A.F., Akbay, N., Germen, E., 2000. Case studies on the use of neural networks in eutrophication modeling. Ecol. Modell. 134, 145–152. https://doi.org/10.1016/S0304-3800(00)00360-4.

Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673–9. https://doi.org/10.1038/89044.

Li, J., Zhang, J., Huang, W., Kong, F., Li, Y., Xi, M., Zheng, Z., 2016. Comparative bioavailability of ammonium, nitrate, nitrite and urea to typically harmful cyanobacterium Microcystis aeruginosa. Mar. Pollut. Bull. 110, 93–98. https://doi.org/10.1016/j.marpolbul.2016.06.077.

Lourakis, M.I., 2005. A brief description of the Levenberg-Marquardt algorithm implemened by levmar. Matrix 3, 2. https://doi.org/10.1016/j.ijinfomgt.2009.10.001

Man-Chung, C., Chi-Cheong, W., Chi-Chung, L.A.M., 2000. Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization 61.

Marion, J.W., Lee, J., Wilkins, J.R., Lemeshow, S., Lee, C., Waletzko, E.J., Buckley, T.J., 2012. In vivo phycocyanin flourometry as a potential rapid screening tool for predicting elevated microcystin concentrations at eutrophic lakes. Environ. Sci. Technol. 46, 4523–4531. https://doi.org/10.1021/es203962u.

McQuaid, N., Zamyadi, A., Prévost, M., Bird, D.F., Dorner, S., 2011. Use of in vivo phycocyanin fluorescence to monitor potential microcystin-producing cyanobacterial biovolume in a drinking water source. J. Environ. Monit. 13, 455–463. https://doi.org/10.1039/c0em00163e.

Mohamed Ismail, H., Ng, H.K., Queck, C.W., Gan, S., 2012. Artificial neural networks modelling of engine-out responses for a light-duty diesel engine fueled with biodiesel blends. Appl. Energy 92, 769–777. https://doi.org/10.1016/j.apenergy.2011.08.027

Mohamed, Z.A., Al Shehri, A.M., 2010. Microcystin production in epiphytic cyanobacteria on submerged macrophytes. Toxicon 55, 1346–1352. https://doi.org/10.1016/j.toxicon.2010.02.007.

Naoum, R.S., Abid, N.A., Al-Sultani, Z.N., 2013. An enhanced resilient backpropagation artificial neural network for intrusion detection system. IJCSNS Int. J. Comput. Sci. Netw. Secur. 13, 98–104.

O'Neil, J.M., Davis, T.W., Burford, M.A., Gobler, C.J., 2012. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. Harmful Algae 14, 313–334. https://doi.org/10.1016/j.hal.2011.10.027.

Orozco, J., García, C.A.R., 2003. Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. Eur. Symp. Artif. Neural Networks, Bruges 349–354.
Paerl, H.W., Xu, H., Hall, N.S., Rossignol, K.L., Joyner, A.R., Zhu, G., Qin, B., 2015. Nutrient limitation dynamics examined on a multi-annual scale in Lake Taihu, China: Implications for controlling eutrophication and harmful algal blooms. J. Freshw. Ecol. 30, 5–24. https://doi.org/10.1080/02705060.2014.994047.
Paliwal, M., Kumar, U.A., 2009. Neural networks and statistical techniques: A review of applications. Expert Syst. Appl. 36, 2–17. https://doi.org/10.1016/j.eswa.2007.10.005
Saini, L.M., 2008. Peak load forecasting using Bayesian regularization, Resilient and adaptive backpropagation learning based artificial neural networks. Electr. Power Syst. Res. 78, 1302–1310. https://doi.org/10.1016/j.epsr.2007.11.003.
Sanusi, Y.K., Abisoye, S.G., Abiodun, A.O., 2013. Application of artificial neural networks to predict daily solar radiation in Sokoto. Int. J. Curr. Eng. Technol. 3 (2), 647–652.
Simsek, H., 2016. Mathematical modeling of wastewater-derived biodegradable dissolved organic nitrogen. Environ. Technol. (United Kingdom) 37, 2879–2889. https://doi.org/10.1080/09593330.2016.1167964.
Simsek, H., Cemek, B., Odabas, M.S., Rahman, S., 2015. Estimation of nutrient concentrations in runoff from beef cattle feedlot using adaptive neuro-fuzzy inference systems. Neural Netw. World 25. https://doi.org/10.14311/NNW.2015.25.025.
Smith, V.H., Tilman, G.D., Nekola, J.C., 1998. Eutrophication: Impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environ. Pollut. 100, 179–196. https://doi.org/10.1016/S0269-7491(99)00091-3.
Suratgar, A.A., Tavakoli, M.B., Hoseinabadi, A., 2007. Modified Levenberg–Marquardt Method for Neural Networks Training. World Acad. Sci. Eng. Technol. 1, 1745–1747.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. Syst. Man Cybern. IEEE Trans. SMC-15, 116–132. https://doi.org/10.1109/TSMC.1985.6313399.
Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Res. 35, 2022–8. https://doi.org/10.1016/S0043-1354(00)00464-4
Yan, Y., Bao, Z., Shao, J., 2018. Phycocyanin concentration retrieval in inland waters: A comparative review of the remote sensing techniques and algorithms. J. Great Lakes Res. https://doi.org/10.1016/j.jglr.2018.05.004.
Zhang, G., Hu, M.Y., Patuwo, B.E., Indro, D.C., 1999. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. Eur. J. Oper. Res. 116, 16–32. https://doi.org/10.1016/S0377-2217(98)00051-4.