Applied Water Science (2019) 9:164
https://doi.org/10.1007/s13201-019-1044-3

ORIGINAL ARTICLE

Application of artificial intelligence to estimate phycocyanin pigment concentration using water quality data: a comparative study

Salim Heddam[1] · Hadi Sanikhani[2] · Ozgur Kisi[3]

Received: 28 October 2017 / Accepted: 23 September 2019 / Published online: 30 September 2019

Abstract

In the present investigation, the usefulness and capabilities of four artificial intelligence (AI) models, namely feedforward neural networks (FFNNs), gene expression programming (GEP), adaptive neuro-fuzzy inference system with grid partition (ANFIS-GP) and adaptive neuro-fuzzy inference system with subtractive clustering (ANFIS-SC), were investigated in an attempt to evaluate their predictive ability of the phycocyanin pigment concentration (PC) using data from two stations operated by the United States Geological Survey (USGS). Four water quality parameters, namely temperature, pH, specific conductance and dissolved oxygen, were utilized for PC concentration estimation. The four models were evaluated using root mean square errors (RMSEs), mean absolute errors (MAEs) and correlation coefficient (*R*). The results showed that the ANFIS-SC provided more accurate predictions in comparison with ANFIS-GP, GEP and FFNN for both stations. For USGS 06892350 station, the *R*, RMSE and MAE values in the test phase for ANFIS-SC were 0.955, 0.205 µg/L and 0.148 µg/L, respectively. Similarly, for USGS 14211720 station, the *R*, RMSE and MAE values in the test phase for ANFIS-SC, respectively, were 0.950, 0.050 µg/L and 0.031 µg/L. Also, using several combinations of the input variables, the results showed that the ANFIS-SC having only temperature and pH as inputs provided good accuracy, with *R*, RMSE and MAE values in the test phase, respectively, equal to 0.917, 0.275 µg/L and 0.200 µg/L for USGS 06892350 station. This study proved that artificial intelligence models are good and powerful tools for predicting PC concentration using only water quality variables as predictors.

Keywords Modeling · Phycocyanin concentration · Feedforward neural networks · Gene expression programming · Adaptive neuro-fuzzy inference system · Grid partition · Subtractive clustering

Introduction

Nowadays, cyanobacterial harmful algal bloom (HAB) has become a serious problem, contributes seriously to the degradation of the drinking water quality and affects human health and the aquatic life with long-lasting effects (Sivapragasam et al. 2010), including bad odors and tastes, reduction in water clarity and oxygen depletion (hypoxia or anoxia) during bloom decay (Sharaf et al. 2019). Monitoring cyanobacteria also known as blue–green algae (CBG) is of great importance for freshwater ecosystems; however, it has been very difficult over the years to ensure effective and adequate monitoring of cyanobacteria in freshwater (Backer 2002). Traditional methods used for monitoring cyanobacteria are mainly based on: (i) standard methods of chlorophyll-a determination, (ii) cell counting and (iii) direct in situ measurement of cyanotoxin (Kong et al. 2014). However, it is reported that fluorescence is a fast, real-time

✉ Salim Heddam
heddamsalim@yahoo.fr

Hadi Sanikhani
hsanikhani12@gmail.com

Ozgur Kisi
ozgur.kisi@iliauni.edu.ge

[1] Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, Hydraulics Division, Agronomy Department, Faculty of Science, University 20 Août 1955, Route El Hadaik, BP 26, Skikda, Algeria

[2] Water Sciences and Engineering Department, Agriculture Faculty, University of Kurdistan, Sanandaj, Iran

[3] School of Technology, Ilia State University, 0162 Tbilisi, Georgia

monitoring method to measure the concentration of phytoplankton in natural water bodies (Xiaoling et al. 2019). One of the most accessory pigment characteristics of cyanobacteria is certainly phycocyanin pigment concentration (PC), and it is considered as the main light-harvesting pigment in cyanobacteria (Simis et al. 2012). PC is more suitable for monitoring cyanobacterial blooms and toxic cyanobacteria and is a functional protein found in cyanobacteria with high intracellular variability (Yan et al. 2018). PC plays an imperative role in the energy transfer cascade by funneling the light energy toward reaction center of the photosystems (Patel et al. 2018). According to Kuo et al. (2018), cyanobacterial blooms are strongly associated with phycocyanin concentrations.

According to Gregor et al. (2007), when PC is excited by light around 590–630 nm with a maximum of 620 nm (Mishra et al. 2009), it emits red light with a maximum at 650 nm. Two methodologies were employed for assessing PC: (i) models prediction of PC utilizing satellite remotely detected data and (ii) laboratory analysis and directly in situ measurement utilizing sensors. In addition, McQuaid et al. (2011) have demonstrated that PC has the property of being soluble in water and strongly fluorescent and consequently the quantitatively detection of PC based on portable instruments is possible. However, measuring PC cannot be easily accomplished and there is no standard measurement technique (Tebbs et al. 2013). Assuming that the traditional method used for quantifying the PC is based upon laboratory analysis that is costly and time-consuming (Le et al. 2011; Kong et al. 2014; Song et al. 2013a, b), a wide variety of alternative approaches based on remote sensing have been proposed and tested to estimate PC as function of reflectance measurement at different wavelengths. In this context, depending on the magnitude of the reflectance trough around 620 nm, three different algorithms are available (Le et al. 2011): (i) semi-baseline (Dekker 1993), (ii) a single reflectance band ratio (Schalles and Yacobi 2000) and (iii) a nested band ratio semi-analytical algorithms (Simis et al. 2005). PC estimation utilizing remotely detected data has been extensively examined by the researchers (Simis et al. 2005; Li et al. 2010; Le et al. 2011).

Simis et al. (2005) introduced a basic optical model-based reflectance band ratio algorithm, for modeling PC of highly eutrophic Loosdrecht and Ijsselmeer lakes, Netherlands. They have used band settings of the MEdium Resolution Imaging Spectrometer (MERIS), and they have found a very high coefficient of determination (R^2) equal to 0.94 between measured PC and predicted PC by the proposed algorithm, with measured specific absorption coefficients at 620 nm called a_{pc}*(620). Using hyperspectral airborne imaging spectrometer for applications (AISA) imagery from central Indiana, USA, Li et al. (2010) built up a model that linked spectral indices, called (x) to the measured PC, called (y). The authors have tested four different univariate regressions: (i) linear, (ii) exponential, (iii) power and (iv) polynomial. As a result of the study, they have demonstrated that PC concentration correlated best with the reflectance trough 628 nm (R_{628}), via an exponential relation, with an R^2 equal to 0.80 and root mean square error (RMSE) equal to 25.52 (μg L^{-1}). Le et al. (2011) compared two semi-analytical algorithms for modeling PC of Lake Taihu, China, including highly turbid water. The two algorithms are: the semi-analytical four-band algorithm already suggested by Le et al. (2009) and the nested band ratio algorithm; the two models are based upon hyperspectral reflectance measurements. The authors have obtained the following results: (i) the nested band ratio algorithm for PC modeling has provided an R^2 equal to 0.68 and a very high RMSE equal to 10.43 mg/m^{-3} and (ii) the semi-analytical four-band algorithm produced good predictions as compared to the first algorithm with an R^2 equal to 0.86 and a very low RMSE value equal to 4.83 mg/m^{-3}. Song et al. (2012) proposed a new model called genetic algorithm partial least squares (GA-PLS) for PC retrieval. The model was compared to three-band algorithm (TBM), and the two were applied together in the three reservoirs, Eagle Creek, Morse and Geist reservoirs, in the Indianapolis, Indiana, USA. The authors used hyperspectral data obtained from in situ and airborne image. As a result of the study, both GA-PLS and TBA provided good accuracy, and the GA-PLS model is more accurate than the TBA. Song et al. (2013a) used data from five drinking water sources in South Australia and central Indiana, USA, for developing models using in situ hyperspectral data. The authors compared four types of algorithms, namely (i) TBM three-band, (ii) OBR optimal band ratio, (iii) SM05 Simis et al. (2005) band ratio and (iv) SY00 Schalles and Yacobi (2000) models. As a result, the four models yielded an R^2 in the validation phase equal to 0.95, 0.94, 0.94 and 0.12 for TBM, OBR, SM05 and SY00, respectively, and the TBM model was the best among the all others. In another study, Song et al. (2013b) compared three different models for estimating PC in the Eagle Creek reservoir, Indianapolis, Indiana, USA. The three models were: (i) three-band, (ii) two-band and (iii) optimal band models. Utilizing simulated MEdium Resolution Imaging Spectrometer (MERIS) and Hyperion spectra pooled datasets, the three models yielded an R^2 equal to 0.68, 0.64 and 0.74 for three-band, two-band and optimal band models, respectively. Li et al. (2012) introduced a semi-analytical method called TBBA to estimate PC using as input the absorption coefficients at 624 nm (A_{PC} (624)). The algorithm combines both three-band indices and the baseline algorithm. The investigation was conducted using data from in three reservoirs: Eagle Creek Reservoir (ECR), Geist Reservoir (GR) and Morse Reservoir (MR), at central Indiana, USA. Compared with the baseline and three-band algorithms, the TBBA provided better PC estimates with R^2 equal to 0.86.

Obviously, predicting PC concentration using remote sensing is broadly discussed in the literature and much effort has been devoted in this subject. Although the aforementioned models are robust enough, the proposition of a new kind of models is most welcome. Artificial intelligence (AI) techniques have been successfully applied in many areas of scientific researches; however, few studies have reported an application of the AI for predicting PC concentration. Sun et al. (2012) modeled PC by support vector machines (SVMs) and linear regression model utilizing band ratios as inputs. The authors have used three different reflectance forms, namely single-band, band ratio and three-band combination, and they have chosen three lakes in China as cases studies: Lake Taihu, Lake Chaohu and Lake Dianchi. To demonstrate the ability of the proposed SVM model, the authors have compared the results obtained with previous proposed algorithms, which are: (i) the baseline algorithm, (ii) the linear algorithm using band ratio, (iii) the quadratic algorithm using band ratio, (iv) the three-band combination algorithm and (v) the semi-analytical algorithm. As a result of the study, the low RMSE was found to be 38.4 (mg m^{-3}), obtained from SVM model. Song et al. (2014) developed and compared three different models: (i) a partial least squares-artificial neural network (PLS-ANN) model, (ii) artificial neural network (ANN) and (iii) three-band model (TBM). The three models used the remote sensing reflectance spectra (R_{rs}) as input to predict the PC concentration as output. The three models were applied using data from central Indiana, USA, and South Australia. The results obtained showed that the PLS-ANN was the best, followed by TBM and the ANN ranked in the last place. Although the two studies applied AI techniques for predicting PC, they are based on the integration of the remote sensing reflectance band ratio as inputs. Recently, Heddam (2016a) proposed a new kind of models based on ANN paradigm for predicting PC utilizing water quality data as input to the model. Four water quality parameters were measured at 15-min interval of time, namely water temperature (TE), pH, specific conductance (SC) and dissolved oxygen (DO), measured at the lower Charles River Buoy, USA. The author has demonstrated that the multilayer perceptron neural network (MLPNN) satisfactorily predicted the PC with high accuracy and a coefficient of correlation equal to 0.975 in the test phase.

Therefore, the main contributions of this study are the proposition of a new kind of models based on AI for predicting PC concentration. We develop and apply four models, namely (i) feedforward neural networks (FFNNs), (ii) gene expression programming (GEP), (iii) adaptive neuro-fuzzy inference system with grid partition (ANFIS-GP) and (vi) adaptive neuro-fuzzy inference system with subtractive clustering (ANFIS-SC), for predicting PC using data from two stations operated by the United States Geological Survey (USGS).

Materials and methods

Feedforward neural network

Artificial neural network (ANN) is a nonlinear model inspired from the behavior of the biological neuron. ANN is arranged in different layers, and their functioning is mainly based on the adaptation of the parameters through a learning process, generally the backpropagation algorithm (Haykin 1999). The most common architecture of ANN is the feedforward neural network (FFNN), selected in the present study. FFNN is composed of three layers: one input layer with four inputs, one hidden layer of neurons with sigmoid activation function and one output layer consisting of only one neuron corresponding to the PC. FFNN is a universal approximator (Hornik 1991; Hornik et al. 1989). The structure of the FFNN developed is shown in Fig. 1. The general equations of the FFNN from the input layer to the output layer can be presented as:

$$Y = f_2\left[\sum_{j=1}^{n} w_{jk}\left(f_1\left(\sum_{j=1}^{n} x_i w_{ij} + \delta_j\right)\right) + \delta_0\right] \quad (1)$$

where x_i is the input variable, w_{ij} weight between the input i and the hidden neuron j and δ_j is the bias of the hidden neuron j. w_{jk} indicates the connection weight between the neuron j in hidden layer and the neuron k in the output layer, and δ_0 denotes the bias of the neuron k in the output layer. f_2

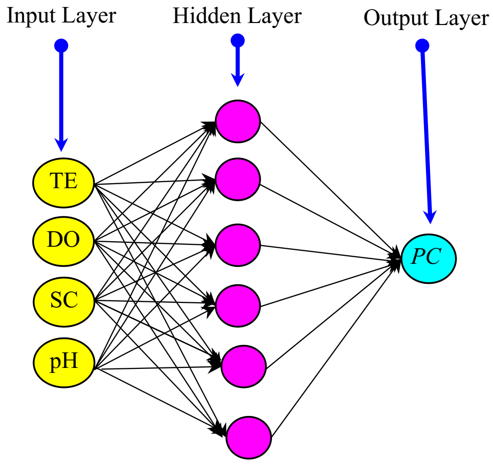

Fig. 1 Architecture of FFNN with four input variables used for modeling PC concentration

is the linear activation function, and f_1 the sigmoid activation function, expressed by Eq. (2).

$$f_1(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Adaptive neuro-fuzzy inference system

Fuzzy inference system (FIS) is used to create nonlinear models, linking a set of inputs to an output, generally achieved in three important processes: (i) selection of membership function, (ii) applying fuzzy set operation and (iii) elaboration of the rules base (Kotti et al. 2016). These types of models use the fuzzy numbers, while the models based on statistical regression are based on the error term (Kitsikoudis et al. 2016). Adaptive neuro-fuzzy inference system (ANFIS) was first suggested by Jang (1993). ANFIS combines the learning abilities of ANN and the fuzzy logic concept (Jang 1993). ANFIS is a MLPNN based on fuzzy inference system (FIS), where each node applies a particular function on incoming signals (Jang 1993). As illustrated in Fig. 2, the ANFIS is composed of exactly six layers: (i) input layer, (ii) fuzzification layer, (iii) rules layer, (iv) normalization layer, (v) defuzzification layer and (vi) summation (output or decision) layer. In the ANFIS structure, there are only two adaptive layers, namely the fuzzification layer and the defuzzification layer. In the fuzzification layer, two modifiable parameters ($\{\sigma_i, c_i\}$), which are identified with the input membership functions, exist, while in the defuzzification layer there are three adjustable parameters ($\{p_i, q_i, r_i\}$) (Jang 1993). ANFIS utilizes a hybrid learning algorithm composed of the gradient descent for the premise parameters (nonlinear) parameters and the least square estimate (LSE) for the linear (consequent) parameters. The learning process is achieved into two phases: forward and backward passes. Simply assume that we have a FIS having two inputs, x and y, and one output z.

Assume that the rule base includes two fuzzy if–then rules (Takagi and Sugeno type):

$$\text{Rule 1} = \text{If } \left(x \text{ is } A_1\right) \text{ and } \left(y \text{ is } B_1\right) \text{ Then } \left(f_1 = p_1x + q_1y + r_1\right) \tag{3}$$

$$\text{Rule 2} = \text{If } \left(x \text{ is } A_2\right) \text{ and } \left(y \text{ is } B_2\right) \text{ Then } \left(f_2 = p_2x + q_2y + r_2\right) \tag{4}$$

where x and y denote the inputs, A_i and B_i indicate the fuzzy sets, f_i are the outputs within the fuzzy region indicated by the fuzzy rule and p_i, q_i and r_i show the design parameters that are identified in the training phase. The ANFIS structure to actualize these two rules is shown in Fig. 2, in which a circle demonstrates a fixed node, whereas a square shows an adaptive node.

Layer 1: the input layer that only fixes the input variable of the system.

Layer 2: the fuzzification layer. Every node i in this layer is a square node with a node function:

$$\mathrm{O}_i^1 = \mu_{A_i}(x), \quad i = 1, 2, \tag{5}$$

$$\mathrm{O}_i^1 = \mu_{B_{i-2}}(y), \quad i = 3, 4 \tag{6}$$

where x (or y) is the input to node i, A_i (or B_{i-2}) is the linguistic label (small, large, etc.) associated with this node function and $\mu_{\mathrm{A}_i}(x)$ and $\mu_{\mathrm{B}_{i-2}}(y)$ can adopt any fuzzy

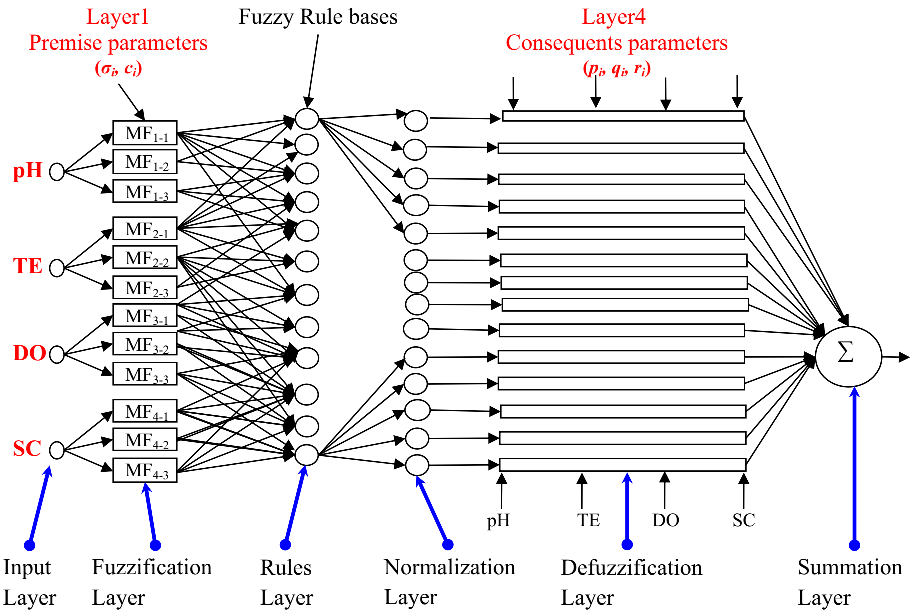

Fig. 2 Architecture of ANFIS with four input variables used for modeling PC concentration

membership function. Assuming a Gaussian function as a membership function, A_i can be computed as

$$\mu_{A_i}(x) = \exp\left\lfloor -0.5 \times \left\{\left(x - c_i\right)/\sigma_i\right\}^2 \right\rfloor, \tag{7}$$

where (σ_i, c_i) denote parameter sets. Parameters in this layer are called as premise parameters.

Layer 3: the rules layer. Each node i in this layer is a fixed node. These nodes multiply the incoming signals and outputs the product.

$$\mathrm{O}_i^2 = w_i = \mu_{A_i}\mu_{B_i}, \quad i = 1, 2, \tag{8}$$

The output signal w_i indicates the firing strength of a rule. The node numbers in this layer are equal to the number of fuzzy rules in the FIS.

Layer 4: the defuzzification layer. In this layer, the nodes are adaptive. Each node's output of this layer is the product of the normalized firing strength and a first-order polynomial. Thus, this layer's outputs are expressed as

$$\mathrm{O}_i^3 = \bar{w}_i = \left(w_i/\left(w_1 + w_2\right)\right), \quad i = 1, 2, \tag{9}$$

Outputs of this layer are named as normalized firing strengths.

Layer 5: the defuzzification layer. In this layer, the nodes are adaptive nodes. The output of each node in this layer is simply the product of the normalized firing strength and a first-order polynomial (for a first-order Sugeno model). Thus, this layer's outputs are expressed as

$$\mathrm{O}_i^4 = \bar{w}_i f_i = \bar{w}_i\left(p_i x + q_i y + r_i\right), \quad i = 1, 2 \tag{10}$$

where $\bar{w}_i$ is the output of Layer 3 and ({p_i, q_i, r_i}) denotes the parameter set of this node. This layer's parameters will be called as consequent parameters.

Layer 6: the summation (output or decision) layer. This layer's node is a fixed node labeled Σ, which calculates the overall output as the sum of all incoming signals, i.e.,

$$\mathrm{O}_i^5 = \sum_{i=1} \bar{w}_i f_i = \left(\sum_{i=1} w_i f_i/\left(w_1 + w_2\right)\right). \tag{11}$$

Explicitly, this layer sums the node's output of the previous layer to calculate the whole network's output.

ANFIS uses two different identification approaches: the grid partition (GP) and the subtractive clustering (SC) (Sylaios et al. 2008). A detail of the methods is reported in the following.

Grid partitioning

The grid partition method (GP) separates the data into rectangular subspaces depending on the pre-defined membership functions' number and types (Sylaios et al. 2008). Using GP method, network partitioning is uniformly utilized and with initialization (Rad et al. 2015). The major drawback of the ANFIS-GP is the so-called the curse of dimensions, which implies that the number of fuzzy rules exponentially increases when there is an increment in the number of input variables (Wei et al. 2007; Noori et al. 2009). According to the study of Jang (2016) and Jang et al. (1997), the number of input variables must be small and < 6 to apply GP. For example, in the case of building a model with high number of inputs (e.g., 10) and if it is necessary to select much membership functions (MF_s) for each input, for example, three MF_s for each input, the number of rules will be: ($3^{10} = 2187$) rules, and the calculation and optimization of this model are a difficult task, rather impossible with the actual computer machines. In the current study, modeling PC concentration was achieved using four input variables and therefore applying an ANFIS-GP model is feasible. Using ANFIS-GP, the total number of model parameters that need to be optimized is computed as follows (Heddam 2014):

Using GP method in ANFIS, the total number of modifiable parameters (Ψ) is computed as:

$$\Psi = \beta + \delta \tag{12}$$

where β is the premise parameters' number and δ consequent parameters' number, and β and δ are computed as:

$$\beta = N_I \times N_{\mathrm{MF}s} \times N_{\mathrm{MP}} \tag{13}$$

$$\delta = N_{\mathrm{FR}} \times \left(N_\mathrm{I} + N_\mathrm{O}\right) \tag{14}$$

$$N_{\mathrm{FR}} = \left(N_{\mathrm{MFs}}\right)^{N_I} \tag{15}$$

where N_I is the input variable number, N_{MFs} MF number of each input and N_{MP} the number of modifiable parameters for each MF, for example, for Gaussian membership function ($N_{\mathrm{MP}} = 2$), N_{FR} numbers of fuzzy rules that will be produced by all inputs and N_O system output which is equal to one (in his study, PC concentration).

Subtractive clustering

Subtractive clustering (SC) is utilized to avoid the problem of curse of dimensionality encountered when using the GP method. SC leads to a reduction in the high number of fuzzy rules and generates significantly smaller rule base depending only on one parameter: the so-called cluster radius (Vasileva-Stojanovska et al. 2015). The influential radius is very essential for calculating the number of clusters. By choosing a smaller radius, too many smaller clusters are obtained in the data space and more rules are required and vice versa

(Kisi and Zounemat-Kermani 2014). SC is a modified version of the original mountain clustering approach (Yager and Filev 1994) suggested by Chiu (1994). The SC approach is utilized to decide the number of antecedent MFs and rules by taking into consideration every cluster center (D_i) as a fuzzy rule. In this method, each data point of a set of N data points $\{x_{1\ldots}\ x_N\}$ in a p-dimensional space is considered as the cluster centers' candidate (Wei et al. 2007). Then, the density measure at data point x_i can be expressed as (Aqil et al. 2007):

$$D_i = \sum_{j=1}^{N} \exp\left(- \frac{\left\| x_i - x_j \right\|^2}{\left(r_a/2 \right)^2} \right) \tag{16}$$

where r_a = a positive constant named cluster radius. A data point is marked as a cluster center when more data points are closer to it. Accordingly, the data point (x_1^*) with highest density measure (D_1^*) is considered as the first cluster center (Wei et al. 2007). Now removing the impact of the first cluster center, the density measure of all other data points is recalculated as:

$$D_i = D_i - D_i^* \cdot \mu\left(x_i^* \right) \tag{17}$$

$$\mu\left(x_i^* \right) = \exp\left(- \frac{\left\| x_i - x_j \right\|^2}{\left(r_b/2 \right)^2} \right) \tag{18}$$

where r_b ($r_b > r_a$) = a positive constant that yields a measurable reduction in density measures of neighborhood data points to avoid closely spaced cluster centers (Chiu 1994). Using ANFIS-SC, the total number of model parameters that need to be optimized is computed as follows (Heddam 2014):

With SC partition approach for the ANFIS model, the number of modifiable parameters (Φ) can be computed as:

$$\varphi = \alpha + \lambda \tag{19}$$

where α is the premise parameters' number and λ the consequent parameters' number, and α and λ are computed as:

$$\alpha = N_I \times N_{\mathrm{MFs}} \times N_{\mathrm{MP}} \tag{20}$$

$$\lambda = N_{\mathrm{FR}} \times \left(N_{\mathrm{I}} + N_{\mathrm{O}} \right) \tag{21}$$

$$N_{\mathrm{FR}} = N_{\mathrm{C}} = N_{\mathrm{MFs}} \tag{22}$$

From the above equations, it can be seen that, when fuzzy systems are designed utilizing SC approach, every cluster corresponds to a fuzzy rule. At that point, the total number of modifiable parameters is equivalent to the quantity of premise parameters in addition to the number of consequent parameters.

Gene expression programming

Gene expression programming (GEP) was introduced by Ferreira in 1999 (Ferreira 2001). This paradigm has some similarity with genetic algorithm (GA) and genetic programming (GP). In GEP similar to GA, linear and chromosomes with fixed length are used. Furthermore, in GEP similar to pars tree of GP, ramified structure is applied. GEP can be used successfully in the following situations: (i) identifying the internal relation of dependent variables is very complex, (ii) finding the size and shape of final variable is complex, (iii) common methods cannot represent the analytical solution for a given problem, (iv) an approximate solution is appropriate, (v) every small improvement in performance is measured routinely and highly valuable and (vi) the amount of data that should be evaluated and classified by computers are huge (Banzhaf et al. 1998). Some preliminary steps before implement of GEP should be considered as follows: (1) select the terminals set (i.e., problem variables and fixed stochastic numbers), (2) select the function set that required for mathematical formula creation, (3) choose the appropriate fitness function for evaluating the fitness of formulas, (4) determine the parameters that control the model evolve (i.e., population size, probability of genetic operators) and (5) determine a criterion for end of program and represent the results of model. In this study for modeling the phycocyanin pigment concentration (PC) using GEP method, various steps were considered. In the first step, the suitable fitness function was selected. In this research, root relative squared error (RRSE) was chosen as fitness function. In the second step, the input variables (i.e., pH, TE, SC and DO) and functions set were selected. In the third step, chromosomal architecture (i.e., in this study the head length and number of genes were 8 and 3, respectively) was determined. In the fourth step, linkage function for creating link between sub-expression trees was selected. Finally in the fifth step, genetic operators and theirs rates were determined. The genetic operators and theirs values are presented in Table 1. In this study for implementation of GEP, GeneXpro Tools

Table 1 Genetic operators and their values utilized in this study for GEP model

Genetic operator	Rate	Genetic operator	Rate
Number of chromosomes	30	One-point recombination rate	0.3
Head size	8	Two-point recombination rate	0.3
Number of genes	3	Gene recombination rate	0.1
Mutation rate	0.044	Gene transposition rate	0.1
Inversion rate	0.1	IS transposition rate	0.1
RIS transposition	0.1		

was utilized. More details about GEP model can be found in Ferreira (2006).

Case studies

In the present study, historical PC concentration and four water quality data from January 1, 2015, to December 31, 2015, were utilized for developing the AI models; data can be obtained from the United States Geological Survey (USGS) Web site: http://or.water.usgs.gov. The data from two water quality stations, namely USGS 06892350 (latitude 38°59′00″, longitude 94°57′52″ NAD27) and USGS 14211720 (latitude 45°31′03″, longitude 122°40′09″ NAD83), were used in this study. The water quality data consisted of measured water temperature (TE, °C), dissolved oxygen (DO mg/L), pH (Std. unit), specific conductance (SC, µS/cm) and PC (µg/L). For USGS 06892350 station, data were measured at 15-min interval of time, while for USGS 14211720 station the data were measured at 30-min interval of time. The dataset selected had a total of 18,139 patterns for USGS 06892350 station and 17,195 for USGS 14211720. Table 2 represents the statistic parameters of water quality variables for the two stations. In the table, the terms X_{mean}, $X_{\max}$, $X_{\min}$, S_x, C_v and R indicate the mean, maximum, minimum, standard deviation, variation coefficient and the coefficient of correlation between the variable and the PC, respectively. The correlations between the water quality variables and PC are generally higher in station 06892350 than in station 14211720, except DO having the lowest correlation with PC in station 06892350. Coefficients of correlation are given in Table 3. The dataset is separated into three subsets (Table 4): (i) a training subset, (ii) a validation subset and (iii) a test subset, with a ratio of 60%, 20% and 20%, respectively. We have tested different train–test–validation splitting strategies, by changing the training ration from 20, 30, 40 and 60%. The best accuracy was obtained using 60% of the data for training.

Table 3 Pearson's correlation coefficients between and among physical water quality parameters and PC concentration

	TE (°C)	SC (µS/cm)	pH (–)	DO (mg/L)	PC (µg/L)
USGS ID 06892350					
TE (°C)	1.000				
SC (µS/cm)	−0.057	1.000			
pH (–)	0.192	0.764	1.000		
DO (mg/L)	−0.584	0.266	0.351	1.000	
PC (µg/L)	0.428	0.569	0.710	−0.025	1.000
USGS ID 14211720					
TE (°C)	1.000				
SC (µS/cm)	0.728	1.000			
pH (–)	0.595	0.512	1.000		
DO (mg/L)	−0.938	−0.826	−0.545	1.000	
PC (µg/L)	−0.234	−0.391	0.231	0.273	1.000

In the present study, before applying the three models, all the four input variables and the PC were normalized to contain the same scale with mean equal to 0 and standard deviation equal to 1, utilizing the Z-score by Eq. (23). Using the Z-score method, the performance of the developed models has been substantially improved (Olden et al. 2004; Heddam 2016b, c).

$$x_{n_{i,k}} = \frac{x_{i,k} - m_k}{S_{dk}} \tag{23}$$

where $x_{ni,\,k}$ denotes the normalized value of the k variable (input or output) for every sample i. $x_{i},\,_{k}$ is the original value of the k variable. m_k and S_{dk} are the mean value and standard deviation of the variable k, respectively.

Table 2 Statistical parameters of dataset

Station	Dataset	Unit	X_{mean}	$X_{\max}$	$X_{\min}$	S_x	C_v	CC
USGS ID 06892350	TE	°C	17.466	32.400	1.200	7.190	0.412	0.428
	SC	µS/cm	681.926	990.00	213.00	189.01	0.277	0.569
	pH	–	8.518	9.300	7.400	0.375	0.044	0.710
	DO	mg/L	10.222	17.100	5.700	2.117	0.207	−0.025
	PC	µg/L	1.141	3.460	0.010	0.677	0.593	1.000
USGS ID 14211720	TE	°C	14.874	26.900	4.800	6.142	0.413	−0.234
	SC	µS/cm	86.303	134.00	57.000	10.882	0.126	−0.391
	pH	–	7.312	8.600	6.900	0.221	0.030	0.231
	DO	mg/L	10.631	14.600	5.700	1.857	0.175	0.273
	PC	µg/L	0.167	1.000	0.000	0.157	0.944	1.000

X_{mean}, mean; $X_{\max}$, maximum; $X_{\min}$, minimum; S_x, standard deviation; C_v, coefficient of variation; CC, the correlation coefficient with PC; PC, phycocyanin; SC, specific conductance; DO, dissolved oxygen; µg/L, microgram per liter; µS/cm, microsiemens per centimeter; mg/L, milligram per liter

Table 4 Summary description of dataset

Description	USGS 06892350	USGS 14211720
Year	2015	2015
Begin date	01/01/2015	01/01/2015
End date	31/12/2015	31/12/2015
Total pattern	35,040**	17,520*
Incomplete pattern	16,901	00325
Finale pattern	18,139	17,195
Training	10,885	10,317
Validation	3627	03439
Test	3627	03439

*17,520: data measured at 30-min interval of time, thus: 365 days × 24 × 2 = 17,520 patterns; 35,040: data measured at 15-min interval of time, thus: 365 days × 24 × 4 = 35,040 patterns

Table 5 Structure of the developed models

Models	Input combination
M1	pH, TE, SC and DO
M2	pH, TE and SC
M3	pH, SC and DO
M4	pH and SC
M5	pH and TE
M6	TE and SC

Application and results

In the current study, an attempt is made to estimate PC concentration using water quality variables as inputs. Several combinations of the water quality variables were selected, and in total, six scenarios were compared (Table 5), and those are: (i) TE, pH, SC and DO; (ii) TE, pH and SC; (iii) DO, pH and SC; (iv) pH and SC; (v) TE and pH; and (vi) TE and SC. The selection of the six combinations is mainly based on the correlation coefficient. In this study, three performance indices were utilized to evaluate the developed models. These three indices are: the coefficient of correlation (*R*), the root mean squared error (RMSE) and the mean absolute error (MAE), calculated as follows:

$$R = \left[\frac{\frac{1}{\mathrm{N}}\sum\left(\mathrm{O}_i - \mathrm{O}_m\right)\left(\mathrm{P}_i - \mathrm{P}_m\right)}{\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{n}\left(\mathrm{O}_i - \mathrm{O}_m\right)^2}\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{n}\left(\mathrm{P}_i - \mathrm{P}_m\right)^2}}\right] \quad (24)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left(\mathrm{O}_i - \mathrm{P}_i\right)^2} \quad (25)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{O}_i - \mathrm{P}_i\right| \quad (26)$$

where N denotes the number of data points, O_i is the measured value and P_i is the corresponding model output (prediction). O_m and P_m indicate the average values of O_i and P_i, respectively.

Predicting PC at USGS 06892350 station

In this section, GEP, FFNN, ANFIS_SC and ANFIS_GP were developed and compared to estimate PC concentration using four water quality variables. Results obtained in the training, validation and test stages are given in Table 6. According to Table 6, the four models achieved good accuracy with high R and low RMSE and MAE values. Table 6 clearly shows that the four models yield different accuracies for different input combinations. In the training stage as given in Table 6, the *R* values, respectively, range from 0.869 to 0.946, 0.872 to 0.946, 0.893 to 0.955 and 0.870 to 0.940 for the FFNN, ANFIS_GP, ANFIS_SC and GEP, highlighting high level of accuracy. In addition, the RMSE values, respectively, range 0.219–0.335, 0.222–0.334, 0.203–0.307 and 0.231–0.334 µg/L for the FFNN, ANFIS_GP, ANFIS_SC and GEP.

Finally, as given in Table 6, MAE values range 0.164–0.256, 0.167–0.258, 0.145–0.229 and 0.176–0.257 µg/L for the FFNN, ANFIS_GP, ANFIS_SC and GEP, respectively. According to Table 6, the M1 combination with TE, pH, SC and DO yielded the highest efficiency than all the others for the all four developed models, while the M4 combination with TE and SC yielded the lowest accuracy in comparison with the all other four developed models. In the training stage, the ANFIS_SC M1 model is the best among the four developed models, with an *R* equal to 0.955, RMSE equal to 0.203 µg/L and MAE equal to 0.145 µg/L, followed by FFNN and ANFIS_GP that lead almost the same accuracy regarding the three performances indices, and the GEP took in the third place with an *R* equal to 0.940, RMSE equal to 0.231 µg/L and MAE equal to 0.176 µg/L. From the six input combinations proposed, when the four AI models have included only two inputs, M5 combination with pH and TE is always the best, and ANFIS_SC M5 model performed the best with an *R* equal to 0.916, RMSE equal to 0.275 µg/L and MAE equal to 0.197 µg/L.

In the validation phase, as given in Table 6, the M1 combination is always the best for the four developed models. The FFNN, ANFIS_GP, ANFIS_SC and GEP M1 models

Table 6 Performances of the FFNN, ANFIS_SC, ANFIS_GP and GEP models in different phases for USGS 06892350 station

Models	Model	Training			Validation			Test		
		R	RMSE	MAE	*R*	RMSE	MAE	*R*	RMSE	MAE
FFNN	M1	0.946	0.219	0.164	0.945	0.221	0.165	0.945	0.222	0.166
	M2	0.933	0.243	0.180	0.933	0.243	0.181	0.930	0.249	0.186
	M3	0.913	0.277	0.205	0.909	0.281	0.208	0.913	0.276	0.204
	M4	0.869	0.335	0.256	0.865	0.338	0.259	0.864	0.341	0.258
	M5	0.905	0.289	0.212	0.907	0.284	0.213	0.901	0.294	0.217
	M6	0.878	0.324	0.238	0.877	0.324	0.239	0.871	0.333	0.243
ANFIS_GP	M1	0.946	0.222	0.167	0.946	0.220	0.166	0.947	0.221	0.166
	M2	0.940	0.233	0.170	0.941	0.233	0.168	0.940	0.235	0.170
	M3	0.921	0.266	0.195	0.921	0.266	0.195	0.924	0.263	0.191
	M4	0.872	0.334	0.258	0.873	0.334	0.258	0.875	0.333	0.255
	M5	0.902	0.294	0.221	0.910	0.285	0.217	0.904	0.294	0.221
	M6	0.882	0.322	0.240	0.884	0.320	0.238	0.879	0.327	0.242
ANFIS_SC	M1	0.955	0.203	0.145	0.955	0.202	0.146	0.955	0.205	0.148
	M2	0.955	0.203	0.143	0.949	0.217	0.151	0.953	0.209	0.149
	M3	0.930	0.251	0.176	0.928	0.256	0.181	0.930	0.253	0.179
	M4	0.893	0.307	0.229	0.878	0.326	0.243	0.888	0.316	0.234
	M5	0.916	0.275	0.197	0.921	0.268	0.195	0.917	0.275	0.200
	M6	0.906	0.290	0.205	0.899	0.301	0.217	0.899	0.302	0.214
GEP	M1	0.940	0.231	0.176	0.936	0.237	0.179	0.936	0.238	0.181
	M2	0.926	0.256	0.193	0.923	0.259	0.195	0.923	0.261	0.195
	M3	0.903	0.291	0.219	0.898	0.296	0.223	0.904	0.290	0.218
	M4	0.870	0.334	0.257	0.866	0.337	0.259	0.869	0.335	0.256
	M5	0.905	0.288	0.209	0.909	0.281	0.208	0.903	0.291	0.213
	M6	0.873	0.331	0.242	0.872	0.331	0.242	0.863	0.342	0.249

used for predicting PC concentration yielded *R* values of 0.945, 0.946, 0.955 and 0.936, respectively, and RMSE values of 0.221, 0.220, 0.202 and 0.237 µg/L, respectively. Finally, the four models yielded MAE values of 0.165, 0.166, 0.146 and 0.179 µg/L, respectively. Similar to the results obtained in the training stage, in the validation stage ANFIS_SC M1 is always the best, followed by FFNN, and ANFIS_GP took in the third place. ANFIS_SC M1 yielded an R equal to 0.945, RMSE equal to 0.221 µg/L and MAE equal to 0.165 µg/L. Using only two input variables (pH and TE), ANFIS_SC M5 model is the best among all the others. According to Table 6, in the test stage ANFIS_SC M1 is the best model and performs superior to the FFNN, ANFIS_GP and GEP in all combinations. In the test phase as given in Table 6, the ANFIS_SC M1 improved the FFNN, ANFIS_GP and GEP M1 models of about 7.57%, 7.23% and 13.86% and 10.84%, 10.84% and 18.23% decrement in RMSE and MAE, respectively. Additionally, results were improved with respect to R statistics in the test stage by approximately 1.0%, 0.8% and 1.9%, respectively.

The cluster radius was calculated as 0.10 by trial and error. The optimal cluster number was found to be 40, and consequently, the ANFIS_SC M1 model having four input variables has a total of 40 fuzzy rules. The detailed description of the two ANFIS model parameters is reported in Table 7. As can be clearly seen from the table that the ANFIS_SC has much more parameters than the ANFIS_GP model. In Table 8, we report the testing results, different functions set and linkage function for developing GEP models. GEP model provided the best accuracy with F5 operators and addition linking function. The equation of the GEP M1 model for PC concentration using TE, pH, SC and DO as inputs is given by:

$$PC = \left(\exp\left(\sin\left(\arctan\left(\frac{3.9(SC - TE)}{TE^2}\right)\right)\right)\right) + \frac{\cos(0.8pH)}{\log\left(\exp\left(\frac{pH}{DO}\right)\right)} + \cos(pH + \arctan(1.4 + TE - DO)) \tag{27}$$

Table 7 Total number of parameters for the two ANFIS models developed for USGS 06892350 station

Station	Designation	Models	
		ANFIS_SC	ANFIS_GP
	Number of linear parameters	200	80
	Number of nonlinear parameters	320	16
	Total number of parameters	520	96
	Number of fuzzy rules	40	16

Table 8 Testing results of different functions set and linkage function for developing GEP for USGS 06892350 station

Definition		RMSE
F1	$\{+, -, \times, \div\}$	0.245
F2	$\{+, -, \times, \div, \ln, e^x\}$	0.241
F3	$\left\{+, -, \times, \div, \sqrt{\ }, \sqrt[3]{\ }, x^2, x^3\right\}$	0.241
F4	$\left\{+, -, \times, \div, \ln, e^x, \sqrt{\ }, \sqrt[3]{\ }, x^2, x^3\right\}$	0.239
F5	$\left\{+, -, \times, \div, \ln, e^x, \sqrt{\ }, \sqrt[3]{\ }, x^2, x^3, \sin x, \cos x, \mathrm{arctg} x\right\}$	0.237
Linking functions		
Addition		0.237
Multiplication		0.240
Subtraction		0.242
Division		0.245

Figures 3, 4, 5 and 6 illustrate scatter plots of the computed versus measured PC for FFNN, ANFIS_GP, ANFIS_SC and GEP model M1, in the training, validation, test and all data. Comparison of the figures apparently indicates that the ANFIS_SC model M1 provides less scattered estimates with a fit line equation closer to the exact line and a higher R^2 value than those of the other models.

Predicting PC at USGS 14211720 station

The main purpose of this section is the comparison of the accuracy of the four AI models developed for predicting PC concentration using data from USGS 14211720 station. The statistics indices of performance are listed in Table 9. Firstly, from the results given in Table 9, the superiority of the ANFIS_SC model can be clearly seen, in all training, validation and test phases. Secondly, in either case, when comparing the six developed combinations (M1 to

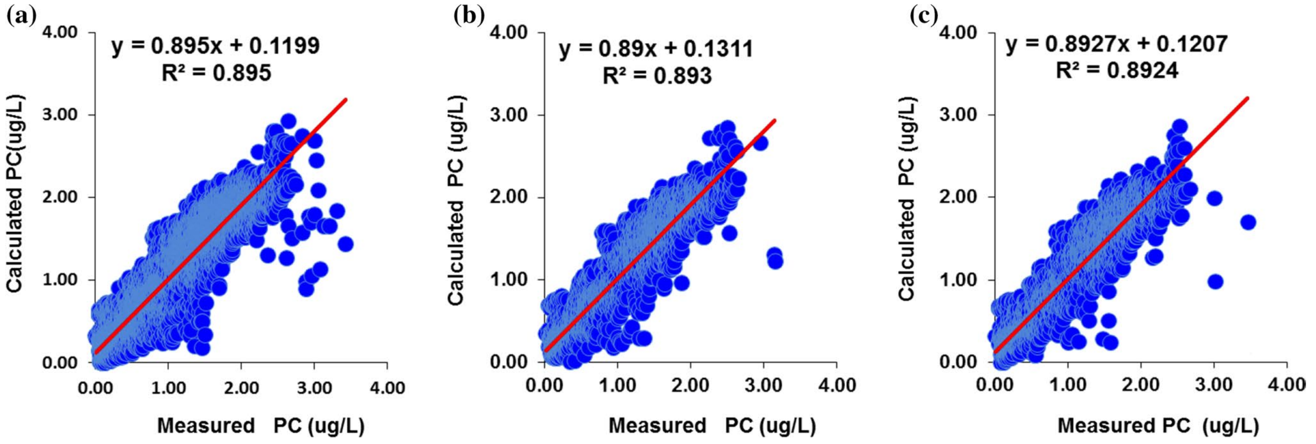

Fig. 3 Scatterplots of estimated versus measured values of PC for FFNN (M1) model: **a** training, **b** validation and **c** test data—station ID: 06892350

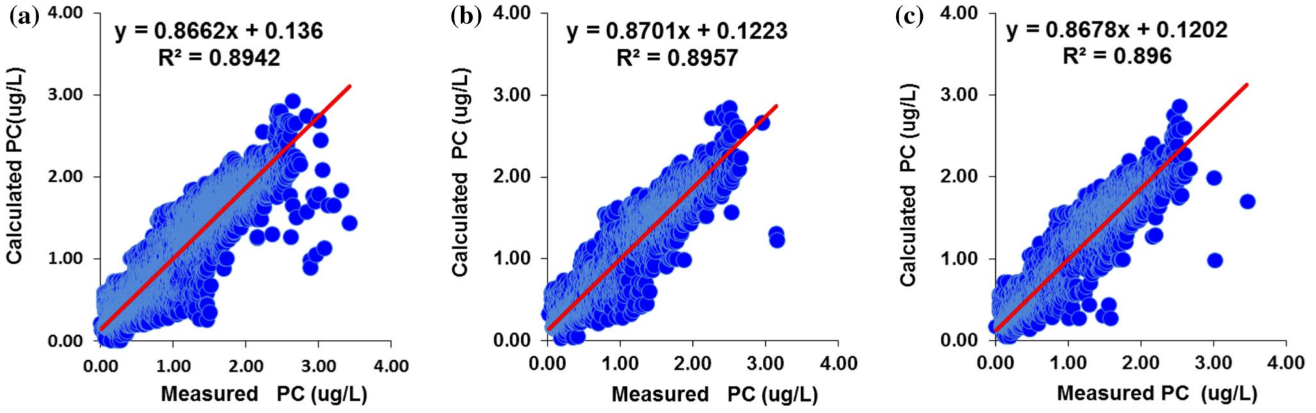

Fig. 4 Scatterplots of estimated versus measured values of PC for ANFIS_GP (M1) model: **a** training, **b** validation and **c** test data—Station ID: 06892350

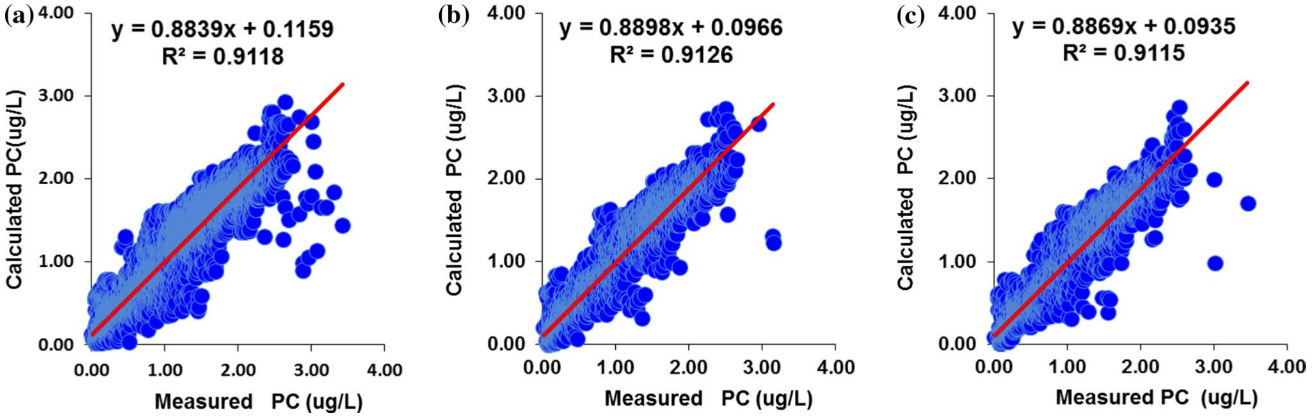

Fig. 5 Scatterplots of estimated versus measured values of PC for ANFIS_SC (M1) model: **a** training, **b** validation and **c** test data—Station ID: 06892350

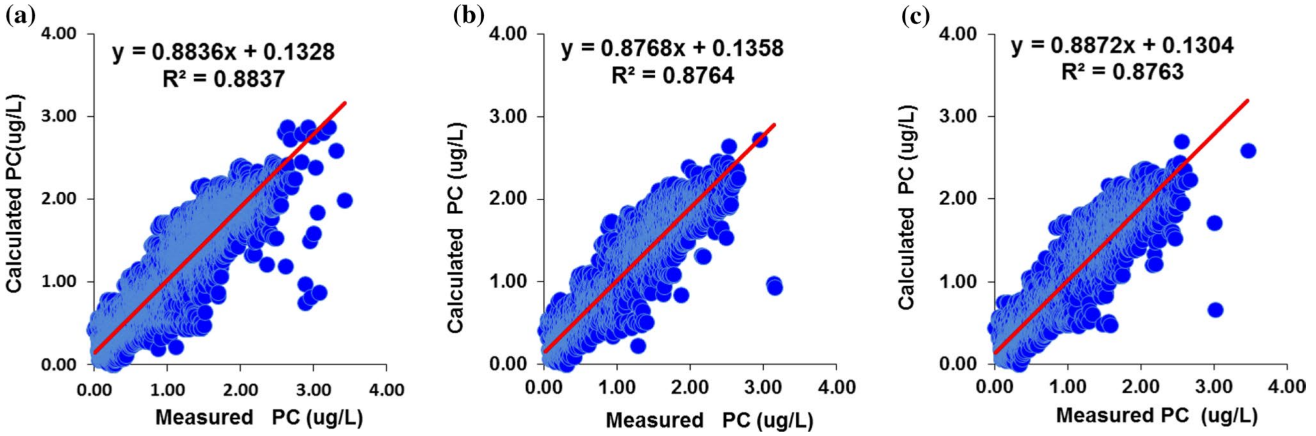

Fig. 6 Scatterplots of estimated versus measured values of PC for GEP (M1) model: **a** training, **b** validation and **c** test data—Station ID: 06892350

M6), ANFIS_SC is always the best among all the others. Thirdly, contrary to the results obtained in USGS 06892350 station, where the M5 combination was the best when only two input variables were included (TE and pH), herein for USGS 14211720 station, M5 combination is the worst with the lowest R and highest RMSE and MAE values. This is certainly due to the fact that the pH has a high coefficient of correlation with PC concentration in the USGS 06892350 station (0.710) and low in the other station (0.231). For the other three models, FFNN, ANFIS_GP and GEP, as given in Table 9, the three models gave relatively the similar results, especially for the M1 combination. In the training phase, ANFIS_SC M1 is the best model with *R*, RMSE and MAE values equal to 0.949, 0.049 µg/L and 0.031 µg/L, respectively. Comparing the ANFIS_SC with the FFNN, ANFIS_GP and GEP, ANFIS_SC has reduced RMSE by 12.50%, 14.03% and 12.50%, and MAE by 20.51%, 22.50% and 16.21% and improved the *R* by 1.5%, 1.8% and 1.5%, respectively.

In the validation stage as given in Table 9, ANFIS_SC M1 is always the best with *R*, RMSE and MAE values equal to 0.95, 0.049 µg/L and 0.032 µg/L, respectively. Comparing the ANFIS_SC with the FFNN, ANFIS_GP and GEP, ANFIS_SC has reduced RMSE by 12.50%, 15.51% and 12.50% and MAE by 17.94%, 21.95% and 15.78% and improved the R by 1.5%, 1.8% and 1.6%, respectively. In the test stage, the ANFIS_SC performed the best with the M1 combination in light of the results obtained in the training and validations phases. The corresponding *R*, RMSE and MAE values were 0.95, 0.050 µg/L and 0.031 µg/L. It is obvious from Table 9 that the ANFIS_SC M1 yields the best performances among the M1 to M6 input combinations. The detailed description of the two ANFIS models parameters is reported in Table 10. Similar to the previous application, here also the ANFIS_SC seems to be more complicated and

Table 9 Performances of the FFNN, ANFIS_SC, ANFIS_GP and GEP models in different phases for USGS 14211720 station

Models	Model	Training			Validation			Test		
		R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE
FFNN	M1	0.934	0.056	0.039	0.935	0.056	0.039	0.931	0.058	0.039
	M2	0.927	0.059	0.039	0.931	0.058	0.039	0.927	0.060	0.040
	M3	0.913	0.064	0.046	0.917	0.063	0.046	0.914	0.064	0.046
	M4	0.812	0.091	0.069	0.821	0.091	0.068	0.808	0.093	0.070
	M5	0.709	0.110	0.071	0.716	0.111	0.073	0.715	0.111	0.072
	M6	0.858	0.080	0.052	0.864	0.081	0.053	0.863	0.082	0.052
ANFIS_GP	M1	0.931	0.057	0.040	0.932	0.058	0.041	0.926	0.060	0.042
	M2	0.927	0.059	0.039	0.923	0.061	0.039	0.916	0.064	0.041
	M3	0.908	0.066	0.048	0.900	0.069	0.049	0.903	0.068	0.049
	M4	0.802	0.093	0.071	0.808	0.093	0.071	0.797	0.096	0.071
	M5	0.658	0.118	0.081	0.646	0.121	0.083	0.661	0.119	0.082
	M6	0.800	0.094	0.064	0.803	0.095	0.065	0.800	0.095	0.064
ANFIS_SC	M1	0.949	0.049	0.031	0.950	0.049	0.032	0.950	0.050	0.031
	M2	0.943	0.052	0.033	0.943	0.053	0.035	0.942	0.053	0.035
	M3	0.934	0.056	0.038	0.934	0.057	0.039	0.930	0.058	0.039
	M4	0.844	0.084	0.059	0.847	0.084	0.058	0.835	0.086	0.058
	M5	0.751	0.103	0.065	0.751	0.105	0.068	0.746	0.105	0.067
	M6	0.867	0.078	0.049	0.867	0.079	0.050	0.863	0.080	0.049
GEP	M1	0.934	0.056	0.037	0.934	0.056	0.038	0.935	0.056	0.038
	M2	0.915	0.063	0.043	0.917	0.063	0.043	0.914	0.064	0.043
	M3	0.901	0.068	0.052	0.901	0.069	0.053	0.903	0.068	0.052
	M4	0.834	0.086	0.064	0.836	0.087	0.065	0.825	0.087	0.065
	M5	0.712	0.109	0.073	0.711	0.111	0.074	0.719	0.110	0.073
	M6	0.824	0.088	0.059	0.827	0.089	0.060	0.829	0.089	0.060

Table 10 Total number of parameters for the two ANFIS models developed for USGS 14211720 station

Station	Designation	Models	
		ANFIS_SC	ANFIS_GP
	Number of linear parameters	155	80
	Number of nonlinear parameters	248	16
	Total number of parameters	403	96
	Number of fuzzy rules	31	16

has much more parameters than the ANFIS_GP. In Table 11, we report the testing results, different functions set and linkage function for developing GEP models. Similar to the previous application, the GEP model gave the best accuracy with F5 operators and addition linking function. The equation of the GEP M1 model for PC concentration using TE, pH, SC and DO as inputs is given by:

$$\mathrm{PC} = \arctan\left(\frac{\cos\left(\frac{\mathrm{SC}-6.2}{\mathrm{TE}}\right)}{\mathrm{DO}+\sin(\mathrm{SC})}\right) + \frac{\mathrm{pH}^4}{\mathrm{SC}^2-\mathrm{TE}^2} + \frac{\mathrm{pH}}{2\mathrm{DO}+\mathrm{pH}^2+\frac{7.3+\mathrm{DO}}{5.5}} \quad (28)$$

Table 11 Testing results of different function sets and linkage function for developing GEP for USGS 14211720 station

Definition		RMSE
F1	$\{+,-,\times,\div\}$	0.065
F2	$\{+,-,\times,\div,\ln,e^x\}$	0.063
F3	$\left\{+,-,\times,\div,\sqrt{},\sqrt[3]{},x^2,x^3\right\}$	0.061
F4	$\left\{+,-,\times,\div,\ln,e^x,\sqrt{},\sqrt[3]{},x^2,x^3\right\}$	0.060
F5	$\left\{+,-,\times,\div,\ln,e^x,\sqrt{},\sqrt[3]{},x^2,x^3,\sin x,\cos x,\mathrm{arctg}x\right\}$	0.056
Linking functions		
Addition		0.056
Multiplication		0.057
Subtraction		0.060
Division		0.058

Figures 7, 8, 9 and 10 illustrate scatterplots of the computed versus measured PC for FFNN, ANFIS_GP, ANFIS_SC and GEP model M1, in the training, validation, test and all data. Comparison of the fit line equations and R^2 values shows that the ANFIS-SC model has less scattered

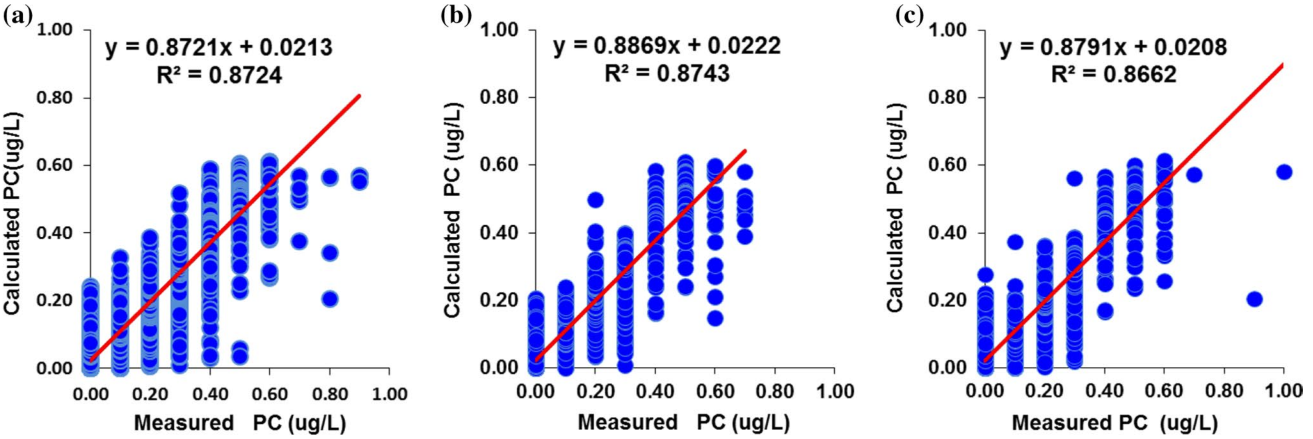

Fig. 7 Scatterplots of estimated versus measured values of PC for FFNN (M1) model: **a** training, **b** validation and **c** test data—Station ID: 14211720

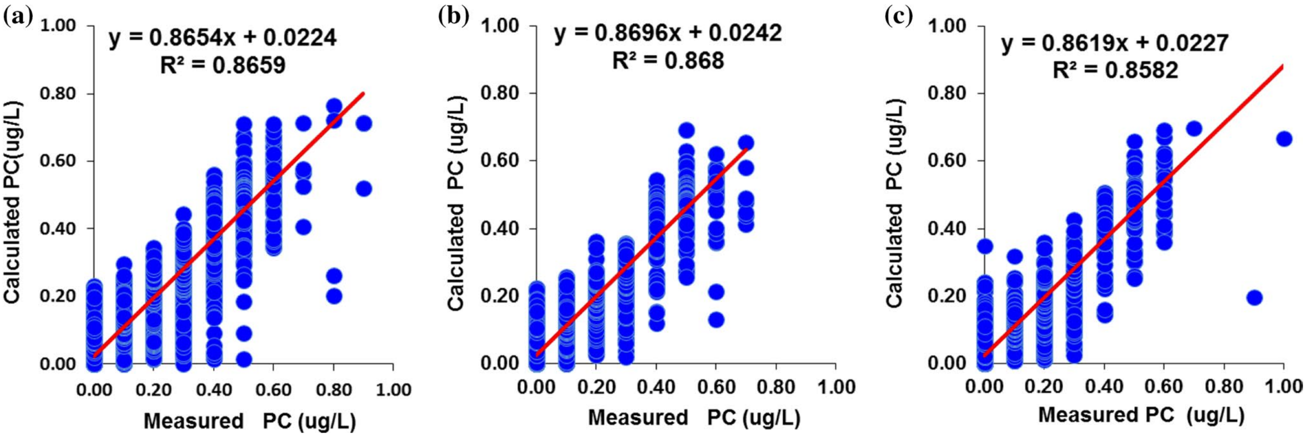

Fig. 8 Scatterplots of estimated versus measured values of PC for ANFIS_GP (M1) model: **a** training, **b** validation and **c** test data—station ID: 14211720

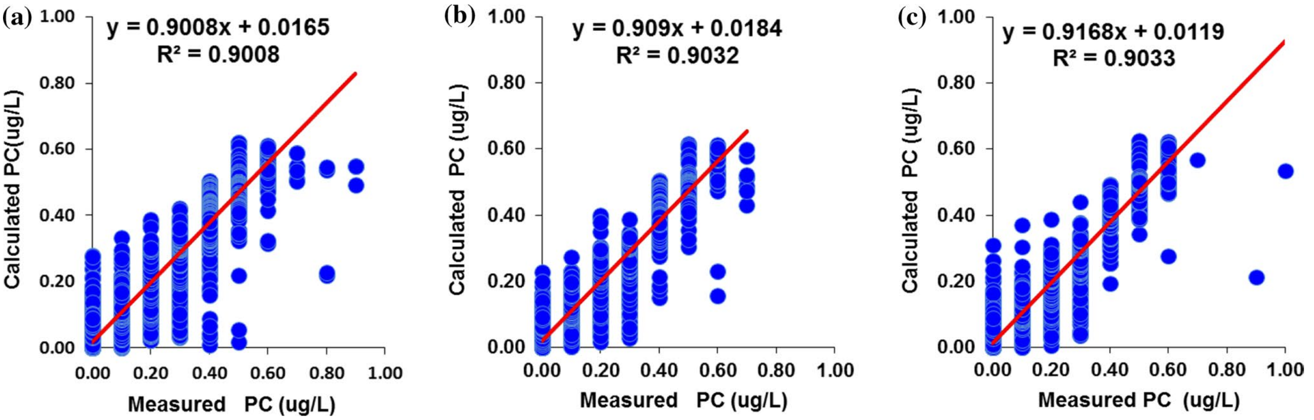

Fig. 9 Scatterplots of estimated versus measured values of PC for ANFIS_SC (M1) model: **a** training, **b** validation and **c** test data—station ID: 14211720

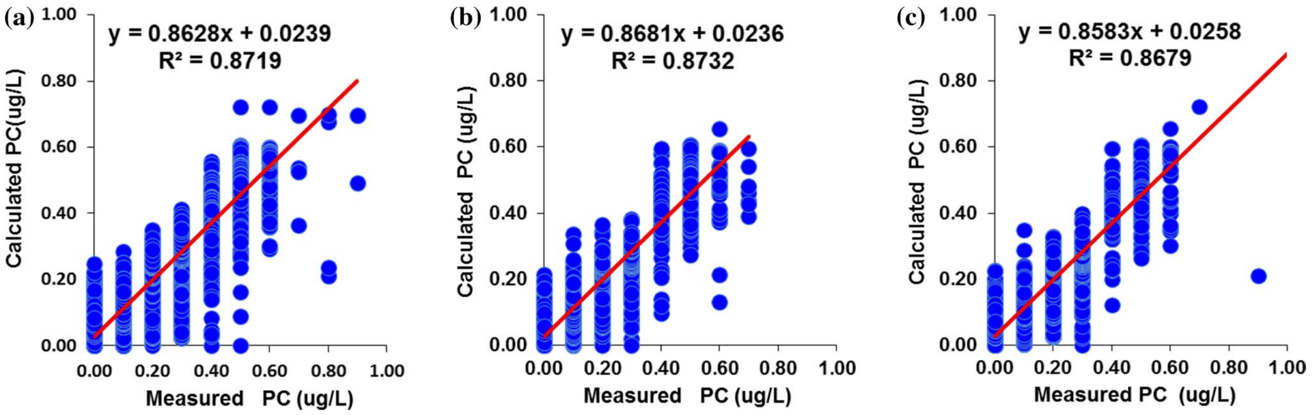

Fig. 10 Scatterplots of estimated versus measured values of PC for GEP (M1) model: **a** training, **b** validation and **c** test data—station ID: 14211720

PC estimates than the other models. The expression tree of the best GEP M1 model is shown in Fig. 10.

Conclusions

In this study, four of the most powerful artificial intelligence (AI) techniques, namely feedforward neural networks (FFNN), gene expression programming (GEP), adaptive neuro-fuzzy inference system with grid partition (ANFIS-GP) and adaptive neuro-fuzzy inference system with subtractive clustering (ANFIS-SC), have been proposed to predict the phycocyanin pigment concentration as a function of several water quality variables. Data used for developing the models were selected from two USGS water quality stations. Water temperature, pH, specific conductance and dissolved oxygen were used as predictors. From the results obtained, it can be concluded that all the AI models proposed herein are very promising and provided good results and ANFIS_SC has shown high accuracy in comparison with the all others models. Among six different combinations of the input variables, we have also demonstrated that the proposed ANFIS_SC model can predict PC concentration with high accuracy using only few inputs. Hence, the proposed models can be successfully used for estimating PC concentration in the absence of direct measurement.

Acknowledgements The authors would like to thank the staff of the United States Geological Survey (USGS) for providing the data that made this research possible.

References

Aqil M, Kita I, Yano A, Nishiyama S (2007) Analysis and prediction of flow from local source in a river basin using a Neuro-fuzzy modeling tool. J Environ Manag 85:215–223. https://doi.org/10.1016/j.jenvman.2006.09.009

Backer LC (2002) Cyanobacterial harmful algal blooms: developing a public health response. Lake Reserv Manag 18:20–31. https://doi.org/10.1080/07438140209353926

Banzhaf W, Nordin P, Keller RE, Francone FD (1998) Genetic programming. Kaufmann, San Francisco

Chiu S (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:267–278. https://doi.org/10.3233/IFS-1994-2306

Dekker A (1993) Detection of the optical water quality parameters for eutrophic waters by high resolution remote sensing. Ph.D. thesis, Amsterdam Free University, Amsterdam, The Netherlands

Ferreira C (2001) Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst 13(2):87–129

Ferreira C (2006) Gene expression programming: mathematical modeling by an artificial intelligence. Springer, Berlin, p 478

Gregor J, Maršálek B, Šípková H (2007) Detection and estimation of potentially toxic cyanobacteria in raw water at the drinking water treatment plant by in vivo fluorescence method. Water Res 41:228–234. https://doi.org/10.1016/j.watres.2006.08.011

Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, Upper Saddle River

Heddam S (2014) Modelling hourly dissolved oxygen concentration (DO) using two different adaptive neuro-fuzzy inference systems (ANFIS): a comparative study. Environ Monit Assess 186:597–619. https://doi.org/10.1007/s10661-013-3402-1

Heddam S (2016a) Multilayer perceptron neural network based approach for modelling phycocyanin pigment concentrations: case study from Lower Charles River Buoy, USA. Environ Sci Pollut Res 23:17210–17225. https://doi.org/10.1007/s11356-016-6905-9

Heddam S (2016b) Simultaneous modelling and forecasting of hourly dissolved oxygen concentration (DO) using radial basis function neural network (RBFNN) Based approach: a case study from the Klamath River, Oregon, USA. Model Earth Syst Environ 2:135. https://doi.org/10.1007/s40808-016-0197-4

Heddam S (2016c) New modelling strategy based on radial basis function neural network (RBFNN) for predicting dissolved oxygen concentration using the components of the Gregorian calendar as inputs: case study of Clackamas River, Oregon, USA. Model Earth Syst Environ 2:167. https://doi.org/10.1007/s40808-016-0232-5

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4(2):251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jang JR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685. https://doi.org/10.1109/21.256541

Jang JR (2016) Frequently asked questions-ANFIS in the fuzzy logic toolbox. http://www.cs.nthu.edu.tw/jang/anfisfaq.htm. Accessed 26 June 2017

Jang JR, Sun C, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice Hall Inc., Englewood Cliffs

Kisi O, Zounemat-Kermani M (2014) Comparison of two different adaptive neuro-fuzzy inference systems in modelling daily reference evapotranspiration. Water Resour Manag 28:2655–2675. https://doi.org/10.1007/s11269-014-0632-0

Kitsikoudis V, Spiliotis M, Hrissanthou V (2016) Fuzzy regression analysis for sediment incipient motion under turbulent flow conditions. Environ Process 3:663–679. https://doi.org/10.1007/s40710-016-0154-2

Kong Y, Lou I, Zhang Y, Lou CU, Mok KM (2014) Using an online phycocyanin fluorescence probe for rapid monitoring of cyanobacteria in Macau freshwater reservoir. Hydrobiologia 741:33–49. https://doi.org/10.1007/s10750-013-1759-3

Kotti IP, Sylaios GK, Tsihrintzis VA (2016) Fuzzy modeling for nitrogen and phosphorus removal estimation in free-water surface constructed wetlands. Environ Process. https://doi.org/10.1007/s40710-016-0177-8

Kuo YM, Yang J, Liu WW, Zhao E, Li R, Yao L (2018) Using generalized additive models to investigate factors influencing cyanobacterial abundance through phycocyanin fluorescence in East Lake, China. Environ Monit Assess 190(10):599. https://doi.org/10.1007/s10661-018-6981-z

Le CF, Li YM, Zha Y, Sun DY (2009) Specific absorption coefficient and the phytoplankton package effect in Lake Taihu, China. Hydrobiologia 619:27–37. https://doi.org/10.1007/s10750-008-9579-6

Le CF, Li YM, Zha Y, Wang Q, Zhang H, Yin B (2011) Remote sensing of phycocyanin pigment in highly turbid inland waters in Lake Taihu, China. Int J Remote Sens 32(23):8253–8269. https://doi.org/10.1080/01431161.2010.533210

Li L, Sengpiel RE, Pascual DL, Tedesco LP, Wilson JS, Soyeux E (2010) Using hyperspectral remote sensing to estimate chlorophyll-a and Phycocyanin in a mesotrophic reservoir. Int J Remote Sens 31(15):4147–4162. https://doi.org/10.1080/01431161003789549

Li L, Li L, Shi K, Li Z, Song K (2012) A semi-analytical algorithm for remote estimation of phycocyanin in inland waters. Sci Total Environ 435–436:141–150. https://doi.org/10.1016/j.scitotenv.2012.07.023

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

Mishra S, Mishra DR, Schluchter WM (2009) A novel algorithm for predicting phycocyanin concentrations in cyanobacteria: a proximal hyperspectral remote sensing approach. Remote Sens 1:758–775. https://doi.org/10.3390/rs1040758

Noori R, Abdoli MA, Farokhnia A, Abbasi M (2009) Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. Expert Syst Appl 36:9991–9999. https://doi.org/10.1016/j.eswa.2008.12.035

Olden JD, Joy MK, Death RG (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model 178:389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

Patel HM, Rastogi RP, Trivedi U, Madamwar D (2018) Structural characterization and antioxidant potential of phycocyanin from the cyanobacterium Geitlerinema sp. H8DM. Algal Res 32:372–383. https://doi.org/10.1016/j.algal.2018.04.024

Rad HN, Jalali Z, Jalalifar H (2015) Prediction of rock mass rating system based on continuous functions using Chaos-ANFIS model. Int J Rock Mech Min Sci 73:1–9. https://doi.org/10.1016/j.ijrmms.2014.10.004

Schalles JF, Yacobi YZ (2000) Remote detection and seasonal patterns of phycocyanin, carotenoid, and chlorophyll pigments in eutrophic waters. Arch Hydrobiol Spec Issues Adv Limnol 55:153–168

Sharaf N, Bresciani M, Giardino C, Faour G, Slim K, Fadel A (2019) Using Landsat and in situ data to map turbidity as a proxy of cyanobacteria in a hypereutrophic Mediterranean reservoir. Ecol Inform 50:197–206. https://doi.org/10.1016/j.ecoinf.2019.02.001

Simis SGH, Peters SWM, Gons HJ (2005) Remote sensing of the cyanobacterial pigment Phycocyanin in turbid inland water. Limnol Oceanogr 50(1):237–245. https://doi.org/10.4319/lo.2005.50.1.0237

Simis SG, Huot Y, Babin M, Seppala J, Metsamaa L (2012) Optimization of variable fluorescence measurements of phytoplankton communities with cyanobacteria. Photosynth Res 112:13–30. https://doi.org/10.1007/s11120-012-9729-6

Sivapragasam C, Muttil N, Muthukumar S, Arun VM (2010) Prediction of algal blooms using genetic programming. Mar Pollut Bull 60:1849–1855. https://doi.org/10.1016/j.marpolbul.2010.05.020

Song K, Li L, Li S, Tedesco L, Hall B, Li Z (2012) Hyperspectral retrieval of phycocyanin in potable water sources using genetic algorithm-partial least squares (GA-PLS) modeling. Int J Appl Earth Obs Geoinf 18:368–385. https://doi.org/10.1016/j.jag.2012.03.013

Song K, Li L, Tedesco L, Clercin N, Hall B, Li S, Shi K, Liu D, Sun Y (2013a) Remote estimation of phycocyanin (PC) for inland waters coupled with YSI PC fluorescence probe. Environ Sci Pollut Res 20:5330–5340. https://doi.org/10.1007/s11356-013-1527-y

Song K, Li L, Li Z, Tedesco L, Hall B, Shi K (2013b) Remote detection of cyanobacteria through phycocyanin for water supply source using three-band model. Ecol Inform 15:22–33. https://doi.org/10.1016/j.ecoinf.2013.02.006

Song K, Li L, Tedesco L, Li S, Hall B, Du J (2014) Remote quantification of phycocyanin in potable water sources through an adaptive model. ISPRS J Photogramm Remote Sens 95:68–80. https://doi.org/10.1016/j.isprsjprs.2014.06.008

Sun D, Li Y, Wang Q, Le C, Lv H, Huang C, Gong S (2012) A novel support vector regression model to estimate the phycocyanin concentration in turbid inland waters from hyperspectral reflectance. Hydrobiologia 680:199–217. https://doi.org/10.1007/s10750-011-0918-7

Sylaios GK, Gitsakis N, Koutroumanidis T, Tsihrintzis VA (2008) CHLfuzzy: a spreadsheet tool for the fuzzy modeling of chlorophyll concentrations in coastal lagoons. Hydrobiologia 610:99. https://doi.org/10.1007/s10750-008-9358-4

Tebbs EJ, Remedios JJ, Harper DM (2013) Remote sensing of chlorophyll-a as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline-alkaline, flamingo lake, using Landsat ETM+. Remote Sens Environ 135(2013):92–106. https://doi.org/10.1016/j.rse.2013.03.024

Vasileva-Stojanovska T, Vasileva M, Malinovski T, Trajkovik V (2015) An ANFIS model of quality of experience prediction in education. Appl Soft Comput 34:129–138. https://doi.org/10.1016/j.asoc.2015.04.047

Wei M, Bai B, Sung AH, Liu Q, Wang J, Cather ME (2007) Predicting injection profiles using ANFIS. Inf Sci 177:4445–4461. https://doi.org/10.1016/j.ins.2007.03.021

Xiaoling Z, Gaofang Y, Nanjing Z, Ruifang Y, Jianguo L, Wenqing L (2019) Chromophoric dissolved organic matter influence correction of algal concentration measurements using three-dimensional fluorescence spectra. Spectrochim Acta Part A Mol Biomol Spectrosc 210:405–411. https://doi.org/10.1016/j.saa.2018.10.050

Yager R, Filev D (1994) Generation of fuzzy rules by mountain clustering. J Intell Fuzzy Syst 2(3):209–219

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.