Coupled Data-Driven Evolutionary Algorithm for Toxic Cyanobacteria (Blue-Green Algae) Forecasting in Lake Kinneret

Article *in* Journal of Water Resources Planning and Management · April 2015

DOI: 10.1061/(ASCE)WR.1943-5452.0000451

CITATIONS
5

READS
237

6 authors, including:

Avi Ostfeld
Technion - Israel Institute of Technology
304 PUBLICATIONS **12,466** CITATIONS
SEE PROFILE

Meir Rom
Mekorot
12 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Lea Kronaveter
Peak-Dynamics Ltd
8 PUBLICATIONS **65** CITATIONS
SEE PROFILE

Tamar Zohary
Israel Oceanographic and Limnological Research Institute (IOLR)
185 PUBLICATIONS **9,974** CITATIONS
SEE PROFILE

Coupled Data-Driven Evolutionary Algorithm for Toxic Cyanobacteria (Blue-Green Algae) Forecasting in Lake Kinneret

Avi Ostfeld, F.ASCE[1]; Ariel Tubaltzev[2]; Meir Rom[3]; Lea Kronaveter[4]; Tamar Zohary[5]; and Gideon Gal[6]

Abstract: Cyanobacteria blooming in surface waters have become a major concern worldwide, as they are unsightly, and cause a variety of toxins, undesirable tastes, and odors. Approaches of mathematical process-based (deterministic), statistically based, rule-based (heuristic), and artificial neural networks have been the subject of extensive research for cyanobacteria forecasting. This study suggests a new framework of linking an evolutionary computational method (a genetic algorithm) with a data driven modeling engine (model trees) for external loading, physical, chemical, and biological parameters selection, all coupled with their associated time lags as decision variables for cyanobacteria prediction in surface waters. The methodology is demonstrated through trial runs and sensitivity analyses on Lake Kinneret (the Sea of Galilee), Israel. Model trials produced good matching as depicted through the results correlation coefficient on verification data sets. Temperature was reconfirmed as a predominant parameter for cyanobacteria prediction. Model optimal input variables and forecast horizons differed in various solutions. Those in turn raised the problem of best variables selection, pointing towards the need of a multiobjective optimization model in future extensions of the proposed methodology. **DOI: 10.1061/(ASCE)WR.1943-5452.0000451.** *© 2014 American Society of Civil Engineers.*

Author keywords: Cyanobacteria; Model trees; Genetic algorithm; Model; Forecasting.

Introduction

Cyanobacteria (blue-green algae) refer to a class of photosynthetic bacteria commonly found in surface waters such as rivers, lakes, and reservoirs. Under certain conditions, these bacteria can multiply to large amounts that can dominate phytoplankton communities (Reynolds 1984).

The presence of cyanobacteria has a range of potential harmful effects such as the capability to produce toxins and undesirable tastes and odors. These byproducts can potentially cause severe gastrointestinal illness and, in extreme cases, death, in humans and animals. The toxic substances produced by cyanobacteria also create unique problems, which substantially increase water treatment costs. As a result, it has become crucial to control cyanobacteria blooms in surface waters.

To control the growth of cyanobacteria, forecasting tools must be developed. These tools will enable the understanding of those mechanisms underlying cyanobacteria growth, in particular its relationships with the surrounding physical, chemical, and biological environments. Such understanding is still lacking, thus efforts ranging from deterministic process-based to data-driven modeling attempts have been the subject of extensive research during the last two decades.

This study suggests a coupled data driven (model trees) evolutionary (genetic) algorithm scheme for forecasting toxic cyanobacteria blooms in Lake Kinneret (the Sea of Galilee), Israel. This work is based on the conference paper of Ostfeld et al. (2006), but has been substantially extended to include multiple runs and sensitivity analyses.

Using model trees coupled with a genetic algorithm scheme for optimizing the model trees' input variables and lag times, provides a simple set of linear rules for cyanobacteria forecasting. Using sensitivity analyses, multiple model trees are derived, which hold similar matching results, but differ in the optimal selected input variables and the cyanobacteria prediction horizon. Users can thus select a model tree (MT) based not only on matching results, but also on more reliable variable measurements and a forecast horizon.

Literature Review

Most of the studies cited below incorporate an artificial neural network (ANN) in their methodology. The literature review herein thus summarizes a few general applications on utilizing ANN's for water resources and civil engineering applications, followed by modeling attempts for cyanobacteria predictions in surface waters.

The latest state of the art on ANN theory and applications for civil engineering was provided at the editorial of Flood (2006).

[1]Faculty of Civil and Environmental Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (corresponding author). E-mail: ostfeld@technion.ac.il

[2]Faculty of Civil and Environmental Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

[3]Lake Kinneret Watershed Unit, Mekorot, Israel National Water Co., Jordan District, Eshkol (North site), P.O. Box 610, Upper Nazareth 17105, Israel.

[4]Lake Kinneret Watershed Unit, Mekorot, Israel National Water Co., Jordan District, Eshkol (North site), P.O. Box 610, Upper Nazareth 17105, Israel.

[5]Kinneret Limnological Laboratory, Israel Oceanographic and Limnological Research, P.O. Box 447, Migdal 14950, Israel.

[6]Kinneret Limnological Laboratory, Israel Oceanographic and Limnological Research, P.O. Box 447, Migdal 14950, Israel.

Note. This manuscript was submitted on January 16, 2014; approved on March 4, 2014; published online on July 24, 2014. Discussion period open until December 24, 2014; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496/04014069(13)/$25.00.

Flood (2006) argued that although the ANN technology was intensively employed in civil engineering since the early 1990s, its application sophistication was very limited. Flood (2006) called upon the establishments of neural networks which are vastly more complex in structure. Those were suggested to be built using genetic algorithms (GAs) and related methods.

Ranjithan et al. (1993) were amongst the early researchers to suggest the pattern classification capability of a neural network for designing reliable groundwater remediation strategies to cope with the spatial variability uncertainty of hydraulic conductivity. Flood and Kartam (1994a, b) were the first to generalize the potential of utilizing ANNs for civil engineering applications. Flood and Kartam (1994a) overviewed the theory of ANNs and demonstrated their basic capabilities on a simple structural analysis problem. Flood and Kartam (1994b) extended the basics described in Flood and Kartam (1994a) through exploring the versatility of ANN's as a problem-solving tool, and through showing how they can be applied on different civil engineering problems. Wen and Lee (1998) developed a neural network-based multiobjective optimization framework of water quality management for water pollution control and river basin planning. The neural network scheme was used to predict the decision maker preference structure. Zou et al. (2007) developed an ANN scheme coupled with a multiobjective optimization model for solving an inverse surface water quality eutrophication problem. Lately, Asadollahfardi et al. (2012) developed an ANN model to predict total dissolved solids as a water quality indicator for rivers management.

Guven and Howard (2006) provided a thorough literature review on cyanobacteria prediction models for lakes and rivers. Their review divided the literature into deterministic process-based and ANN models (Burke 1991). The later ANN category fits into the broader data-driven modeling class of methods in which this manuscript resides. This part of the literature review concentrates on data-driven modeling techniques for cyanobacteria forecasting, and in particular on ANNs utilization. Fuzzy logic and remote sensing studies for cyanobacteria predictions are also reviewed, and the utilization of evolutionary algorithms. Examples of deterministic process-based models for cyanobacteria modeling can be found in Guven and Howard (2007, 2011).

Maier (1995) suggested ANNs for modeling water resources systems; he reviewed ANNs and proposed model frameworks for salinity and cyanobacteria forecasting for the River Murray, Australia. Results from this study was summarized in Maier and Dandy (1996, 1997). Maier and Dandy (1996) utilized ANNs to forecast salinity in the River Murray fourteen days in advance to better schedule pump operations to reduce high salinity levels and incurred costs; Maier and Dandy (1997) used ANNs for predicting the cyanobacteria *Anabaena sp.* blooms in the River Murray. Recknagel (1997) applied ANNs for cyanobacteria prediction in Lake Kasumigaura, Japan, demonstrating the usefulness of employing phytoplankton models in forecasting cyanobacteria blooms. Yabunaka et al. (1997) used the same ANN approach as Recknagel (1997) for forecasting the growth of phytoplankton in the Kasumigaura Lake. Recknagel et al. (1997) showed that ANNs were superior to other physically-based methods in predicting cyanobacteria blooms for four different freshwater systems. Whitehaed et al. (1997) compared time series analysis, dynamic mass balance, and neural network techniques for six sites along the River Thames, U.K. ANNs were found to suggest a new approach, but their predictive capability has not suppressed other approaches. Maier et al. (1998) improved the study of Maier and Dandy (1997) by using back-propagation type of ANNs. Maier et al. (2000) developed a methodology based on B-spline associative memory networks (AMNs) combined with fuzzy membership functions to explore the relationship between model inputs and outputs for cyanobacteria concentration predictions. Wei et al. (2001) used ANNs to quantify the interactions between abiotic factors and algal genera for the Kasumigaura Lake. Ibelings et al. (2003) combined a deterministic process-based simulation model for cyanobacteria growth with fuzzy logic. Chang et al. (2004) developed an empirical model to detect algal blooms phenomena in the Techi Reservoir, Taiwan, based on multiple linear regression models. Bowden et al. (2005) built upon Maier et al. (1998) by utilizing a clustering self-organizing map (SOM) (Kohonen 1982) procedure for ANN input preprocessing. Once the input dimensionality was reduced using the SOM, a GA (Holland 1975; Goldberg 1989) was utilized to determine the inputs that have the most significant relationship with the modeled output. Muttil and Chau (2006) developed a coupled ANN - genetic programming (GP) model for long term predictions of algal biomass in Tolo Harbour, Hong Kong. The GP was utilized to establish the empirical physical related equations, and ANN - for forecasting. Kutser et al. (2006) used fuzzy logic coupled with ocean color satellite images for estimating cyanobacteria blooms in the Baltic Sea. Lilover and Laanemets (2006) developed a fuzzy logic model for predicting the maximum biomass of the toxic cyanobacteria *Nodularia spumigena* bloom in the Gulf of Finland. Teles et al. (2006) used ANNs for predicting cyanobacteria blooms in the Crestuma Reservoir, Portugal, using a three year fortnightly periodicity database. Kingston et al. (2006) used Bayesian and deterministic ANNs to examine the influence of the inputs uncertainty to an ANN on its resulted output. Yao et al. (2007) developed a synergetic multioperator, multipopulation GA mechanism for neural networks construction. The proposed algorithm was successfully applied for pattern recognition of blue-green algae in lakes. Devillers et al. (2007) developed a methodology-entitled PASS (prediction of activity spectra for substances) for predicting the biological and toxicological activities of cyanobacteria species using the structural formulas of their chemicals. Qingyu et al. (2008) employed remote sensing for predicting cyanobacteria blooms in the Taihu Lake, China, based on the optical characteristics of the four periods of algae blooms. Torres et al. (2011) used general regression neural networks to forecast the density of cyanobacteria blooms in Torrão reservoir, Portugal. Ahn et al. (2011) linked an ANN, an SOM, and a multilayer perceptron for cyanobacteria prediction in the Daechung Reservoir, Korea. The water temperature and total dissolved nitrogen were found to be the most dominant factors on predicting cyanobacteria growth.

Bobbin and Recknagel (2001) were the first to suggest an evolutionary algorithm for cyanobacteria predictions in lakes through a knowledge-discovery model of pattern dynamics. Whigham and Recknagel (2001a) presented a GP model for rules and equations discovery using time series data of phytoplankton. Whigham and Recknagel (2001b) extended the study of Whigham and Recknagel (2001a) by applying several machine learning techniques, including GAs for model calibration. Recknagel et al. (2002) used a hybrid ANN-GA framework for seven-days-ahead forecasts of algal blooms in Lake Kasumigaura (Japan). Cao et al. (2006) described a hybrid evolutionary algorithm which links a GP model for rules discovery with a GA for parameters calibration. Chan et al. (2007) applied a hybrid ANN and evolutionary algorithms framework to model limnological time-series data of Lake Suwa (Japan). Welk et al. (2008) applied rules discovery for cyanobacteria prediction through a link of a hybrid evolutionary algorithm and *k*-fold cross-validation, and through rule-based agents of merged time-series data of lakes having the same lake category. Recknagel et al. (2008a) used adaptive agents techniques combined with GP for early warning and operational control of algae blooms.

Recknagel et al. (2008b) showed the advantages of using an object-oriented implementation of mass balance equations, and a rule-based agent methodology for the design and validation of ecosystem models. Recknagel and Cao (2009) provided a summary on evolutionary and machine learning approaches for data analysis, modeling, synthesis, and forecasting of limnological processes in lakes and rivers. Lately, Vilán et al. (2013), Fernández et al. (2013), and Nieto et al. (2013) developed a cyanotoxin diagnostic model for the Trasona reservoir in northern Spain. Their model utilized a support vector machine (SVM) coupled with a multilayer perceptron network. The model was able to signify the importance of biological and physical-chemical variables on the cyanotoxins presence in the reservoir, and to predict its formation.

Rolland et al. (2013) investigated the temporal and spatial variations in the cyanobacterial community structure of Lake St Charles, the drinking water supply for Quebec City, Canada, during 2007–2011. Results showed that temperature played a key role in controlling bloom-forming cyanobacteria and total nitrogen and total phosphorus. Predicting the occurrence of blooms given the temporal variability in cyanobacteria and other phytoplankton species, was reaffirmed by the authors as a major challenge for water quality management of drinking water supplies. Quiblier et al. (2013) reviewed the current knowledge on benthic cyanobacteria and its potential risks to contribute toxins to the planktonic (water column) of water bodies. Their conclusion is of an immediate need of exploring new research for investigating drivers of benthic bloom formations and toxin productions. Those efforts should ultimately lead to the development of models that can be used to assess benthic cyanobacteria formations and their role as toxin sources. The study of Quiblier et al. (2013) emphasizes the need for development of modeling tools for toxic cyanobacteria predictions, regardless of their sources, thus strengthening the need for the development of tools for toxins cyanobacteria forecasting, such as this work effort.

Methodology

The methodology couples model trees and a genetic algorithm for predicting cyanobacteria in Lake Kinneret. A description of model trees, genetic algorithms, and the developed coupled framework are presented herein.

Model Trees

A MT is a special case of a decision tree (DT), where a DT represents a classifier which is depicted in a tree structure flow chart. DT induction algorithms introduce several advantages over other learning algorithms, such as robustness to noise, low computational cost for model generation, and ability to deal with redundant attributes (Barros et al. 2012). The main difference between a DT and a MT is attributed to the tree leaf node outcome: in a DT each output is associated with a single discrete/continuous value, whereas in a MT each leaf node holds a piecewise linear function.

Model trees assume a collection of a set of training cases. Each case is specified by its set of attribute values, either discrete or numeric, and has an associated target/objective value. The MT process constructs a model which relates the objective values of the training cases to their attribute values. The merit of the model outcome is quantified by its prediction accuracy on new unseen cases.

DTs/MTs (Solomatine and Ostfeld 2008) techniques build on partitioning the parameter space into areas (subspaces) and constructing in each of them a separate regression model of zero or first order. The tree is fitted to a training data set by splitting the data into homogeneous subsets based on data attributes. The tree is constructed so that the target variable of all training cases is correctly predicted in the tree leaves. Each leaf is a linear regression model, which incorporates the numerical decision attributes and predicts continuous values for the target variable. The tree is then pruned bottom-up and transformed into a set of if-then rules, a process which simplifies its structure and improves its ability to classify new instances.

In particular we utilize herein the M5 (model trees version 5) algorithm (Quinlan 1992, 1993) implemented in Cubist (http://rulequest.com/cubist-info.html). In the M5 model the data set is either associated with a leaf, where a regression model is built, or with a node, where some test is selected which recursively splits the data set into subsets corresponding to the test outcomes. An M5 outcome yields a committee of linear models spanning certain subsets of the training set which belong to particular regions of the input space. Those simple linear models/rules are easy to follow and implement as a prediction tool. Having a set of linear rules as a prediction instrument for a decision maker is much more intuitive and thus appealing than running, for instance, an ANN model.

The splitting criterion in M5 is based on the standard deviation of the target values that reach a node. The expected error reduction at a node is entitled the standard deviation reduction (SDR) measure, and is calculated through

$$\mathrm{SDR} = \mathrm{sd}(T) - \sum_i \frac{|T_i|}{|T|}\mathrm{sd}(T_i) \tag{1}$$

where T = set of examples that reach the splitting node; T_1, T_2, ... = particular subsets of T which results from splitting the node according to the selected input variable; sd (T) = standard deviation of T; and $|T_i|/|T|$ = represents a weight corresponding to the fraction of the examples which belong to subset T_i. After examining all possible splits of the data through an exhaustive search, the M5 model selects the partition which maximizes SDR. Fig. 1 describes a simple illustrative example of the splitting outcome of M5 for two inputs X_1 and X_2.

For further technical information on DTs/MTs the reader is referred to a state of the art survey on evolutionary algorithms for decision-tree induction by Barros et al. (2012) and to Witten and Frank (2005). A detailed Pseudo-Code for M5 is provided at Wang and Witten (1996) at Section 3.4.

Genetic Algorithms

GAs are heuristic search procedures based on the mechanisms of genetics and Darwin's natural selection principles, combining an artificial survival of the fittest with genetic operators abstracted from nature (Holland 1975; Goldberg 1989). GAs differ from other search techniques in that they search among a population of elements and use probabilistic rather than deterministic transition rules. As a result, GAs search more globally (Wang 1997; Haupt and Haupt 1998).

A typical genetic algorithm incorporates three main stages: (1) initial population generation: the GA generates a bundle of strings (entitled a population), with each string being a coded representation of the decision variables; (2) computation of the strings fitness: the GA evaluates each string's fitness (i.e., the value of the objective function corresponding to each string); and (3) generation of a new population: the GA generates the next population by performing: selection, crossover, and mutation, where selection involves the process of choosing chromosomes from the current

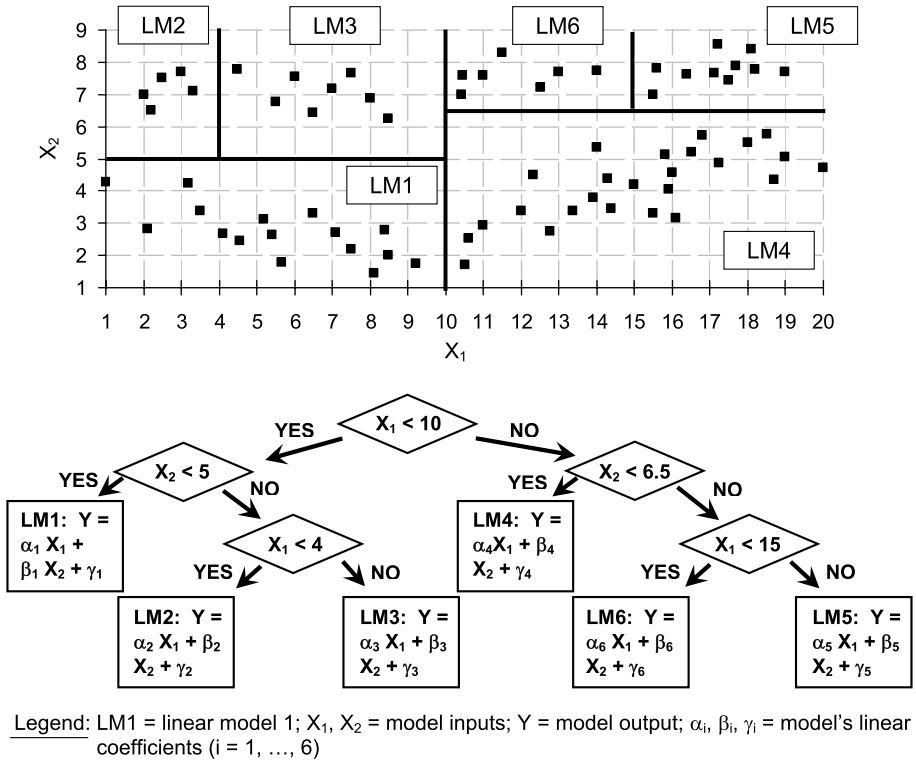


Fig. 1. Simple illustrative example for a model tree outcome

population for reproduction according to their fitness values, crossover involves partial exchange of information between pairs of strings, and mutation–a random change in one of the strings locations. Strings may have binary, integer, or real values.

In this study, the following GA operations are used: *Selection*—using weighted random pairing; where the better the fitness of a chromosome, the higher is its likelihood to be selected as a parent. *Crossover*—the one point crossover method where the offspring is a linear combination of its two parents. *Mutation*—through randomly altering one of the chromosome's parameter values. *Elitism*—the best chromosome in each generation is moved unchanged to the next.

A genetic algorithm pseudo code may take the following form:

1. Initialization
   - Set the generation counter $k = 0$;
   - Generate an initial population $G(0)$;
   - Evaluate $G(0)$.
2. Main Scheme Repeat
   - Set $k = k + 1$;
   - Generate $G(k)$ using $G(k-1)$;
   - Evaluate $G(k)$.

   Until stopping conditions are met.

During the last decade, GAs became one of the more successful robust optimization techniques employed for water resources and environmental engineering management (Nicklow et al. 2010). The genetic algorithm implemented in this study is optiGA (Salomons 2002).

Coupled Model Tree–Genetic Algorithm Scheme

The coupled approach takes advantage of the model trees' strength in solving classification problems, and applies the proven capabilities of a genetic algorithm in optimization to polish its performance. In each iteration, the model trees' performance is used as the genetic algorithm objective (fitness) function. Thus, optimization is guided by the accuracy of the MT model's predictions, and improves it continuously.

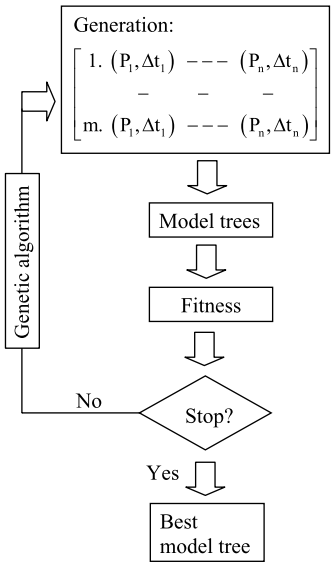
The database used for this study [1.1.1986–31.12.2004 courtesy of the Kinneret limnological laboratory (KLL) and Mekorot Co. Kinneret watershed unit] is summarized in Appendix SI. It consists of 87 variables partitioned into physical (e.g., water level) chemical (e.g., nitrate), biological (e.g., Peridinium wet weight biomass), and external loading (e.g., total phosphorus) sets. The algorithm described herein refers to this database. The algorithm is summarized in Fig. 2 and in the following steps:

Initialization

1. Randomly select from the database a predefined number of variables [e.g., four variables: one from each database group (i.e., one from the physical, one from the chemical, one from the biological, and one from the external loading) or any four from the entire database].
2. Randomly select a time lag (in weeks) for each of the picked variables.
3. Repeat 1 and 2 until a set of variables and time lags are chosen (i.e., the initial GA population).

Model Tree Construction

1. Construct a MT for each of the GA strings using Cubist (Quinlan 1993) for cyanobacteria blooms prediction.
2. Assign each string an objective (fitness) value equal to its resulted correlation coefficient computed on an external independent cross validation dataset (i.e., evaluate the MT prediction quality on an independent data set). It should be noted that in Cubist the Pearson's r correlation coefficient is utilized. Other correlation coefficients such as the root mean square error (RMSE), the Theil u-statistic, or the Nash sutcliffe efficiency (NSE), could be likely employed.



Legend: $(P_1,\Delta t_1)$ = variable and lag time one, respectively; n = number of variables in a single string (solution); m = number of population strings

Fig. 2. Proposed methodology (schematic)

Genetic Algorithm

1. Perform selection, crossover, and mutation using optiGA (Salomons 2002).
2. Construct a new population of strings (i.e., variables and time lags).
3. Check if stopping conditions are met (i.e., if no improvement is gained through a predefined number of generations or if the maximum number of generations is attained). If stopping conditions are met, define the corresponding highest correlation coefficient string as the optimal solution (i.e., the optimal model tree), otherwise go back to Step B.

The following GA parameters (Salomons 2002) were used in this application: string–integer; selector–roulette; crossover–one point; elitism–the best string in each generation is included unchanged in the next generation; crossover probability – 0.95; mutation pprobability – 0.02; generation number – 50; and population size – 50. Running time of a single trial was 3.2 min on a PC Lenovo T7300@2.00 GHz, 778 MHz, 1.98 GB of RAM.

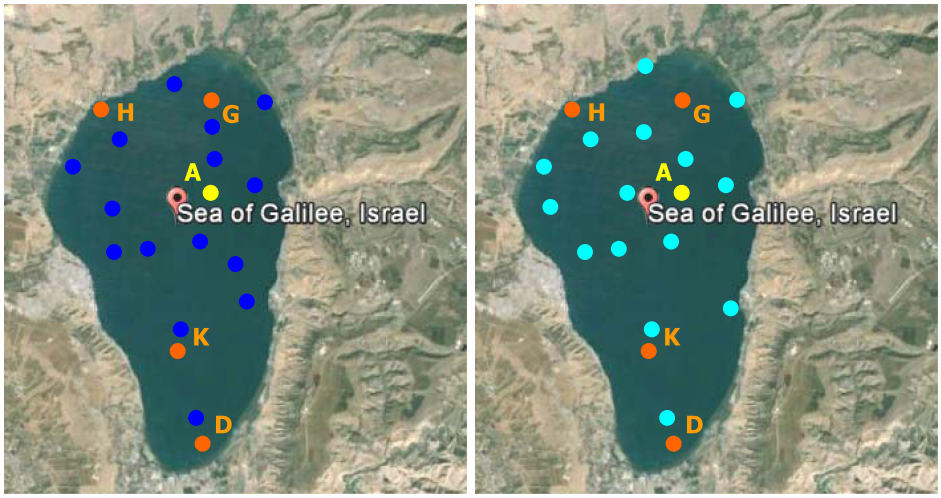
Results and Discussion

Results and discussion of the proposed methodology are described through exploring the forecast of cyanobacteria blooms in Lake Kinneret, Israel. Cyanobacteria blooms unexplainably appeared in 1994, after 30 years during which there had been no major problems involving the lake's water quality (Hadas et al. 2012). Sources of pollution are from anthropogenic sources and water abstraction for domestic and agricultural uses (e.g., apple orchards, cotton and winter wheat) and commercial fish ponds, dairy operation and cattle grazing, all contributed through surface and sub-surface flows from the lake watershed (Preis et al. 2006).

Lake Kinneret, also known as the Sea of Galilee, is a warm monomictic lake located at the northern end of the Afro Syrian Rift Valley in northern Israel: 32:50 N, 35:35 E, 209 m below sea level. It occupies a surface area of approximately 170 km^2; a volume of 4 km^3; maximum and minimum depth of 44 and 26 m, respectively; a shoreline length of 53 km; and a residence time of approximately five years. The lake drains the Kinneret watershed, which resides in Israel and Lebanon at an area of 2,730 km^2 (Markel and Shamir 2002).

The lake is monitored (Fig. 3) by Mekorot Kinneret watershed unit on a fortnight basis and by the Kinneret limnological laboratory (KLL) at five additional weekly monitored stations (A, G, H, K, and D). A meteorological station, installed and operated by Mekorot and located at the vicinity of station H, records radiation, wind velocities, and air and surface water temperatures. Monitoring station A, placed at the deepest point in the lake, is the most intensively monitored station.

It should be noted that the data used in the development of this model were comprised of information routinely collected by the Kinneret limnological laboratory (KLL) and Mekorot. Consequently, these data were not collected specifically for the purpose of this study. For model development, all data were converted to weekly averages. It is important to note that that there is a large



Legend: A, D, K, G, H = KLL (Kinneret Limnological Laboratory) monitoring stations;
●,● = Mekorot beginning and midmonth, respectively, monitoring campaign stations

Fig. 3. Lake Kinneret monitoring stations (© Google, Map data © 2014 Google, Mapa GISrael, ORION_ME)

Table 1. Results Summary

Run	NOV	Type	Coding	Physical (6)	Chemical (48)	Biology (15) (CTXF in m)	External loading (15)	Forecast horizon (weeks)	BCVC
1	1	C	1 0 0 0	WTemp (2)	NA	NA	NA	2	0.54
2			0 1 0 0	NA	PTOT 10–15 (6)	NA	NA	6	0.59
3			0 0 1 0	NA	NA	CTXF 1 (4)	NA	4	0.90
4			0 0 0 1	NA	NA	NA	PO_4 (3)	3	0.48
5	4		1 2 0 1	ATemp (8)	NO_3 5–10 (1) ALK 10–15 (1)	NA	SS10 (7)	1	0.81
6	5		1 3 0 1	WTemp (3)	ALK 5–10 (1) NH_4 1 m (1) NH_4 0–5 m (1)	NA	NTOT (8)	1	0.88
7	6		1 4 0 1	WTemp (3)	NH_4 0–5 m (8) NTOT 0–5 m (8) NORG 5–10 (8) PO_4 5–10 (8)	NA	NTOT (8)	3	0.83
8	6		2 3 0 1	WTemp (4) RHumid (4)	NH_4 10–15 (5) CHLP 10–15 (5) ALK 0–5 (5)	NA	K (8)	4	0.81
9	1	F	NA	NA	NA	CTXF 1 (6)	NA	6	0.90
10	2			NA	NA	CHLP 10–15 (1) CTXF 1 (8)	NA	1	0.93
11	4			ATemp (6)	pH 5–10 (4)	CTXF 1 (8)	PTOT (2)	2	0.93
12	6			NA	CO_2 5–10 (2) ALK 1 m (2)	Rotifera (8) CTXF 5 (8) CTXF 1 (8)	ALK (4)	2	0.95
13	8			Windv (1) RHumid (1)	NO_3 1 m (2) NORG 10–15 (2)	CTXF 1 (6) CTXF 1 (6) CTXF 5 (6)	Cl (3)	1	0.95

Note: NOV = number of variables; C = constrained; F = free; Coding = 1 0 0 0–one constrained physical variable, zero chemical, zero biology, zero external loading; Physical (6) = six physical variables in database; BCVC = best cross validation correlation coefficient, utilizing Pearson's r correlation coefficient; NA = not applicable; WTemp (2) = water temperature (lag of two weeks); rest variables notation–see supplementary dataset metadata.

degree of uncertainty in the biological data of the database. This uncertainty stems from the inherent sampling and counting errors of these measurements. A summary of the data and metadata of the database used in this work is given in Appendix SI: database metadata and notation.

Model Runs

The methodology was tested through trial runs (Table 1) and sensitivity analyses (Tables 2–4 and Figs. 4–10) on selecting decision variables, modifications to the training and verification durations, to the model predicted variables, and on the GA generation/population sizes. The design of the run experiments was aimed at testing the model optimal solution behavior to modifications made in the data or the constraints. Conclusions on the model capabilities as a result of the trails are summarized below and in the conclusions section. The model main graphical user interface (GUI) was built using Microsoft Visual Basic 6.0.

Although this part is targeted at elaborating on the model results, it should be stressed that at some instances a full physical interpretation of the final model outcomes is hard to make. This is attributed to the approach undertaken of a data-driven-evolutionary algorithm framework. This observation was also noted by Bowden et al. (2005).

Trial Runs

The trial run results are described in Table 1. In runs 1–4, the model is constrained to select one variable at a time from each of the database groups. In run 1, the water temperature (WTemp) is selected as the most significant physical variable with a forecast horizon of two weeks and best cross validation correlation

Table 2. Sensitivity Analyses Results

	Data					Results		
	Genetic algorithm							
Run	Population	Generations	Training period	Verification period	Predicted variable (CTXF in m)	Selected variables (CTXF in m)	Forecast horizon (weeks)	BCVC
Run 11 (Table 1)	50	50	1994–2001	2002–2004	CTXF 1	ATemp (P, 6) pH 5–10 (C, 4) CTXF 1 (B, 8) PTOT (EL, 2)	2	0.93
SA1	100	100	1994–2001	2002–2004	CTXF 1	CHLP 10–15 (C, 1) PTOD 0–5 (C, 1) CTXF 1 (B, 7) CTXNF 5 (B, 7)	1	0.95
SA2	50	50	1994–2001	2002–2004	CTXF 10	ATemp (P, 1) NH_4 1 m (C, 3) CTXF 5 (B, 7) PTOT (EL, 1)	1	0.93
SA3	50	50	1994–1998	1999–2004	CTXF 1	Glrad (P, 1) CHLP 5–10 (C, 3) CTXF 5 (B, 5) SO_4 (EL, 4)	1	0.91
SA4	50	50	2002–2004	1994–2001	CTXF 1	CORG 5–10 (C, 5) NO_3 0–5 (C, 5) NO_3 10–15 (C, 5) CTXF 1 (B, 8)	5	0.93
SA5	50	50	1990–2002	2003–2004	Coppoda	pH 5–10 (C, 2) PTOD 1 m (C, 2) Copepoda (B, 6) CTXF 10 (B, 6)	2	0.93

Note: SA1 = sensitivity analysis 1; ATemp (P, 6) = air temperature (physical, lag 6 weeks); C = chemical; B = biology; EL = external loading; BCVC = best cross validation correlation coefficient, utilizing Pearson's r correlation coefficient; rest variables notation–see Appendix SI.

Table 3. Results of Training and Validation for Predicting CTXF 1 m {Toxic, Nitrogen-Fixing Cyanobacteria [Wet Weight Biomass ($\mu g\,L^{-1}$)] at a Depth of 1 m}

	Years 1994–2004											Training		Validation		Results [selected four variables (free, no constraints)]				
Run	94	95	96	97	98	99	00	01	02	03	04	RE	CC	RE	CC	Physical	Chemical	Biological (CTXF in m)	External loading	FH
1	T	V	V	V	V	V	V	V	V	V	V	0.27	0.87	0.24	0.89	WindV (2)	CORG 0–5 (3)	CTXF 1 (3)	NH_4 (2)	2
2	T	T	V	V	V	V	V	V	V	V	V	0.29	0.87	0.23	0.90	None	CORG 10–15 (2)	CTXF 1 (7)	Alk (1); Mg (1)	1
3	T	T	T	V	V	V	V	V	V	V	V	0.25	0.88	0.27	0.90	ATemp (4)	CHLP 10–15 (3)	CTXF 1 (8)	Mg (7)	3
4	T	T	T	T	V	V	V	V	V	V	V	0.19	0.91	0.28	0.93	None	CORG 5–10 (1) CHLP 1 m (1)	CTXF 1 (6) Microzoopl (6)	None	1
5[a]	T	T	T	T	T	V	V	V	V	V	V	0.25	0.92	0.34	0.91	Glrad (1)	CHLP 5–10 (3)	CTXF 5 (5)	SO_4 (4)	1
6	T	T	T	T	T	T	V	V	V	V	V	0.21	0.93	0.31	0.93	None	CHLP 0–5 (3) NTOT 0-5 (3) pH 5–10 (3)	CTXF 1 (8)	None	3
7	T	T	T	T	T	T	T	V	V	V	V	0.23	0.90	0.31	0.93	None	None	CTXF 1 (7) Cladocera (7) CNTX 1 (7) CTXNF 5 (7)	None	7
8[b]	T	T	T	T	T	T	T	T	V	V	V	0.21	0.89	0.28	0.93	ATemp (6)	pH 5–10 (4)	CTXF 1 (8)	PTOT (2)	2
9	T	T	T	T	T	T	T	T	T	V	V	0.20	0.94	0.47	0.93	None	NH_4 0–5 (4)	CTXF 1 (6) Per 5 (6) CTXNF 5 (6)	None	4
10	T	T	T	T	T	T	T	T	T	T	V	0.34	0.93	0.35	0.95	None	CORG 1 m (7)	CTXF 5 (7) Coppoda (7) Per 1 (7)	None	7

Note: 94 = 1994; T = training year; V = validation year; RE = relative error; CC = cross validation correlation coefficient utilizing Pearson's r correlation coefficient; ATemp (6) = air temperature (lag of 6 weeks); FH = forecast horizon (weeks).
[a]See also Table 2, SA3.
[b]See also Table 2, Run 11; rest variables notation–see manuscript in Appendix SI.

Table 4. Validation Ranking for Predicting CTXF 1 m {Toxic, Nitrogen-Fixing Cyanobacteria [Wet Weight Biomass ($\mu g\,L^{-1}$)] at a Depth of 1 m}

	Years 1994–2004											Validation ranking (see also Table 3)				Results [selected 4 variables (free, no constraints)]				
Run	94	95	96	97	98	99	00	01	02	03	04	RE	CC	Sum	Rank	Physical	Chemical	Biological (CTXF in m)	External loading	FH
1	T	V	V	V	V	V	V	V	V	V	V	2	5	7	3	WindV (2)	CORG 0–5 (3)	CTXF 1 (3)	NH_4 (2)	2
2	T	T	V	V	V	V	V	V	V	V	V	1	4	5	1	None	CORG 10–15 (2)	CTXF 1 (7)	Alk (1); Mg (1)	1
3	T	T	T	V	V	V	V	V	V	V	V	3	4	7	3	ATemp (4)	CHLP 10–15 (3)	CTXF 1 (8)	Mg (7)	3
4	T	T	T	T	V	V	V	V	V	V	V	4	2	6	2	None	CORG 5–10 (1) CHLP 1 m (1)	CTXF 1 (6) Microzoopl (6)	None	1
5[a]	T	T	T	T	T	V	V	V	V	V	V	6	3	9		Glrad (1)	CHLP 5–10 (3)	CTXF 5 (5)	SO_4 (4)	1
6	T	T	T	T	T	T	V	V	V	V	V	5	2	7	3	None	CHLP 0–5 (3) NTOT 0–5 (3) pH 5–10 (3)	CTXF 1 (8)	None	3
7	T	T	T	T	T	T	T	V	V	V	V	5	2	7	3	None	None	CTXF 1 (7) Cladocera (7) CNTX 1 (7) CTXNF 5 (7)	None	7
8[b]	T	T	T	T	T	T	T	T	V	V	V	4	2	6	2	ATemp (6)	pH 5–10 (4)	CTXF 1 (8)	PTOT (2)	2
9	T	T	T	T	T	T	T	T	T	V	V	8	2	10	5	None	NH_4 0–5 (4)	CTXF 1 (6) per 5 (6) CTXNF 5 (6)	None	4
10	T	T	T	T	T	T	T	T	T	T	V	7	1	8	4	None	CORG 1 m (7)	CTXF 5 (7) Coppoda (7) per 1 (7)	None	7

Note: 94 = 1994; T = training year; V = validation year; RE = relative error; CC = cross validation correlation coefficient utilizing Pearson's r correlation coefficient; ATemp (6) = air temperature (lag of 6 weeks); FH = forecast horizon (weeks).
[a]See also Table 2, SA3.
[b]See also Table 2, Run 11; rest variables notation–see Appendix SI.

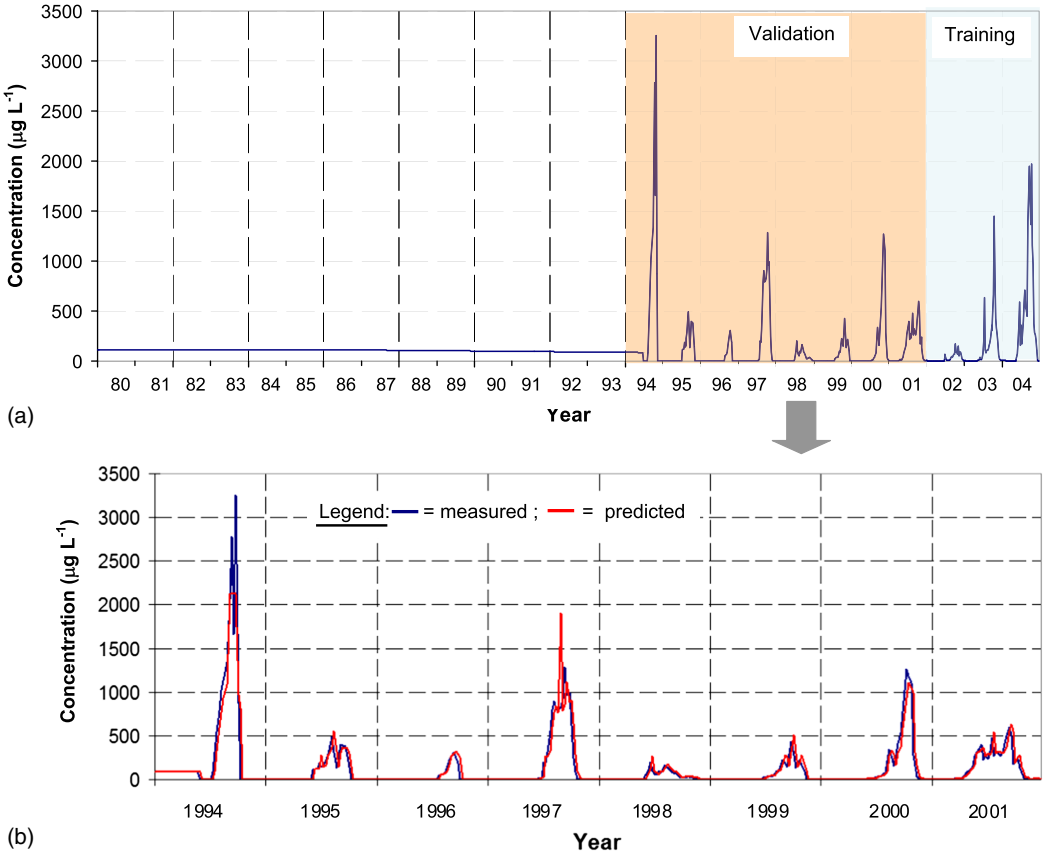


Fig. 4. Toxic, nitrogen-fixing cyanobacteria [wet weight biomass ($\mu g\,L^{-1}$)] at a depth of 1 m at monitoring Station A (Fig. 3): (a) data time series 1980–2004; (b) results for sensitivity analysis 4 (SA4, Table 2)

coefficients (i.e., best = out of several trails) of only 0.54. In run 2, the average total phosphate at a depth 10–15 m (PTOT 10–15) is selected from the chemical group with a slightly higher correlation coefficient of 0.59 and a forecast horizon of six weeks. In run 3, toxic, nitrogen-fixing cyanobacteria at a depth of 1 m (CTXF 1 m) are picked from the biology set with a correlation coefficient of 0.90 and a prediction horizon of four weeks. In run 4, orthophosphate (PO_4) from the external loading variables is selected with a forecast horizon of three weeks and a correlation coefficient of 0.48.

Runs 1–4 reveal the foremost variables of each of the sets. Some results are obvious (e.g., the most important variable for predicting CTXF 1 m is CTXF 1 m itself, resulting in the highest correlation coefficient among all four runs of 0.9). Temperature as a predominant variable for cyanobacteria prediction was also reported by Maier et al. (1998).

In run 5, the model is constrained to four variables: one from the physical set, two from the chemical, and one from the external loading. The model selected air temperature (ATemp) from the physical set with a lag of eight weeks; average nitrate of depth 5–10 m (NO_3 5–10) and average alkalinity at a depth of 10–15 m (ALK 10–15) with lags of one week from the chemical; and suspended solids (SS10) from the external loading with a lag of seven weeks. The best correlation coefficient is 0.81 and the prediction horizon one week (i.e., the minimum among all individual variable lag times of 8, 1, and 7). In run 6, the model is constrained to one physical, three chemical, and one external loading, resulting in a prediction horizon of one week and an increase of the correlation coefficient to 0.88 compared with run 5. In run 7, an additional chemical variable is constrained to be selected compared with run 6. This yields a decrease in the correlation coefficient to 0.83, but an increase to three weeks of the prediction horizon. In runs 8, 2, 3, and 1, variables are constrained from the physical, chemical, and external loading sets, respectively, resulting in a correlation coefficient of 0.81 and a forecast horizon of four weeks.

Runs 9–13 are unconstrained (i.e., the model is free to select any predefined number of variables from the entire dataset). In run 9, the model is run for selecting one variable. The variable selected is CTXF 1 m with a lag of six weeks and a correlation coefficient of 0.9. It is thus very similar to the result of constrained run number 3. In run 10, two biological variables are picked; the forecast horizon is reduced to one, but the correlation coefficient increases to 0.93, as compared with run 9. In runs 11, 12, and 13, 4, 6, and 8, variables are selected, yielding correlation coefficients of 0.93 and 0.95, and forecast horizons of one and two weeks.

Observing the results presented in Table 1, one would inevitably question which run to select. The answer here is not clear, because one must make a tradeoff between the forecast horizon and the reliability of measurements of the selected variables. For example, there is a large degree of uncertainty in the biological data derived from these types of measurements, considering their inherent sampling and counting errors. Thus, although CTXF 1 m may be attractive, its measurements involve a high degree of uncertainty. In this study, run 11, which holds a correlation coefficient of 0.93, a variable from each of the groups, and a prediction horizon of two weeks, is selected. Table 2 describes sensitivity analyses on running the model unconstrained for four variables (i.e., sensitivity analyses on run 11).

IF
CTXF 1m (lag 8 weeks) $\leq$ 181
THEN
Predicted CTXF 1m (5 weeks forecast) = 3.1 + 1.05 CTXF 1m (lag 8 weeks)

IF
CTXF 1m (lag 8 weeks) > 181 and CTXF 1m (lag 8 weeks) $\leq$ 427
THEN
Predicted CTXF 1m (5 weeks forecast) = 157.8 + 0.52 CTXF 1m (lag 8 weeks) + 374 NO3 0-5 (lag 5 weeks)

IF
NO3 10-15 (lag 5 weeks) > 0.0425 and CTXF 1m (lag 8 weeks) > 427
THEN
Predicted CTXF 1m (5 weeks forecast) = -111.3 + 0.74 CTXF 1m (lag 8 weeks)

IF
NO3 10-15 (lag 5 weeks) $\leq$ 0.0425 and CORG 5-10 (lag 5 weeks) > 3.25 and CTXF 1m (lag 8 weeks) > 427
THEN
Predicted CTXF 1m (5 weeks forecast) = 196.3 + 0.72 CTXF 1m (lag 8 weeks)

IF NO3 10-15 (lag 5 weeks) $\leq$ 0.0425 and CORG 5-10 (lag 5 weeks) $\leq$ 3.25 and CTXF 1m (lag 8 weeks) > 427
THEN
Predicted CTXF 1m (5 weeks forecast) = 9012.2 - 58598 NO3 10-15 (lag 5 weeks) - 2250 CORG 5-10 (lag 5 weeks)

Legend: CTXF 1m (lag 8 weeks) = toxic, nitrogen-fixing cyanobacteria [wet weight biomass ($\mu g\ L^{-1}$)] at a depth of 1m at monitoring Station A (Fig. 1) (corresponding to measurements of a lag of 8 weeks); NO3 0-5, NO3 10-15 = average nitrate ($mg\ L^{-1}$) of depth 0-5 m and 10-15 m at monitoring Station A, respectively; and CORG 5-10 = average organic carbon ($mg\ L^{-1}$) of depth 5-10 m at monitoring Station A

Fig. 5. Example of model tree rules for sensitivity analysis 4 (SA4) (Table 2 and Fig. 4)

Sensitivity Analyses

In sensitivity analysis 1 (SA1), the genetic algorithm population size and number of generations is increased to 100 (50 in run 11). As a result, the correlation coefficient is increased to 0.95, the forecast horizon reduced to one week, and the MT variables modified, excluding CTXF 1 m. In SA2, the target variable is altered to CTXF 10 m (i.e., toxic, nitrogen-fixing cyanobacteria at a depth of 10 m) yielding a forecast horizon and correlation coefficient of one week, and 0.93, respectively. In SA3, the training and validation periods are changed: the training period is reduced by three years to 1998–1999, as compared with run 11, and the verification duration increased to 1999–2004. As a result, the model correlation coefficient is slightly reduced and the prediction horizon set to one week. In SA4, the training and verification periods are switched as compared with run 11. The model yields a correlation coefficient of 0.93 with a forecast horizon of five weeks. The results for SA4 are graphically presented in Fig. 4. Fig. 4 describes a constant concentration of the toxic, nitrogen-fixing cyanobacteria until 1994, after which blooming started to occur (part A of Fig. 4). The resulting outcome for the training dataset period 2002–2004, and the validation dataset for 1994–2001, are shown in part B of Fig. 4, demonstrating a close correspondence between measured and predicted outcomes. Fig. 5 presents the MT rules for SA4. In SA5, the methodology is tested to predict the concentration of the zooplankton group Coppoda. Fig. 6 shows the matching results for SA5 for the verification duration of 2003–2004. The obtained correlation coefficient and forecast horizon are 0.93 and two weeks, respectively.

Tables 3 and 4 show results of training and validation for predicting CTXF 1 m with four free (i.e., no constraints) variables. Model run results are evaluated through altering the training and validation durations between 1994 and 2004. Table 3 presents the relative error (RE) (i.e., the ratio of the average error magnitude to the error magnitude that would result from always predicting the mean value), the cross validation correlation coefficient (CC) utilizing Pearson's r correlation coefficient, and the optimal variables, time lags, and forecast horizons. Table 4 describes the validation RE and CC ranking of the runs shown in Table 3, assigning the minimum RE and maximum CC the highest rank of 1. Summing-up the ranking of RE and CC for each of the runs (Table 4), yields runs scoring for which run number 2 receives the highest ranking. The MT outcome of run number 2, according to this measure, can thus be considered as the preferable forecasting model.

In Figs. 7–10 sensitivity analysis on the model genetic algorithm generation/population sizes are explored through statistics on multiple trial runs. Fig. 7 shows statistics on the best cross validation correlation coefficient (BCVC) for 30 model trial runs on run 12 (Table 1) with different genetic algorithm generation/population sizes: 100/100, 50/50, 30/30, and 20/20. Fig. 7 demonstrates that as the number of generation/population numbers decrease, so is the maximum (i.e., best), average, and minimum (worse), model results. Fig. 8 describes the standard deviations of the results shown in Fig. 7, presenting an increase in the standard deviation of the obtained results as the generation/population numbers decrease. Fig. 9 gives the results for CTXF 1 m for the validation period

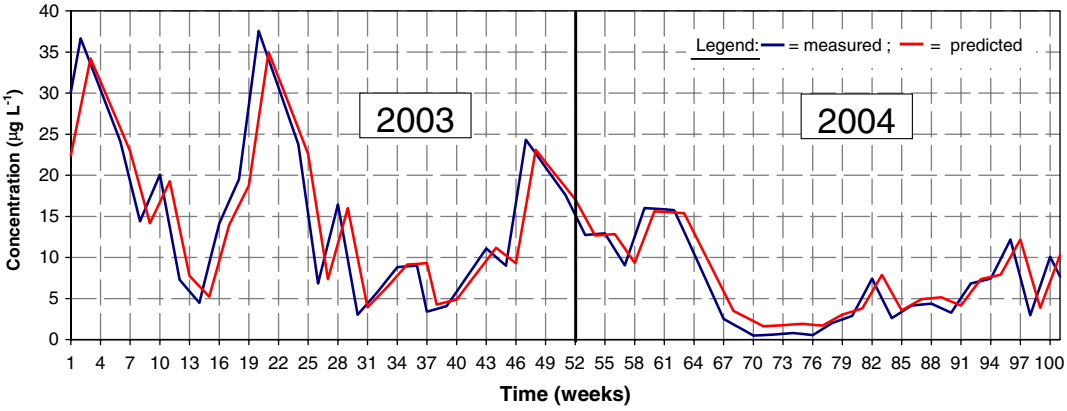


Fig. 6. Coppoda prediction results for 2003 and 2004 [sensitivity analysis 5 (SA5) (Table 2)]

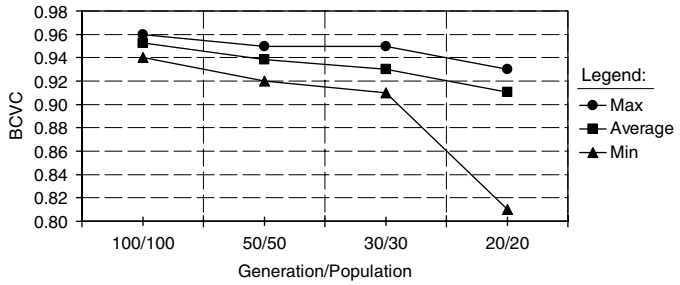


Fig. 7. Statistics of the best cross validation correlation coefficient (BCVC) for 30 model runs of run 12 (Table 1) with different genetic algorithm generation/population sizes

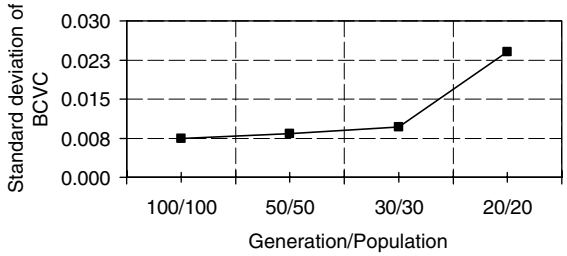


Fig. 8. Standard deviation of the best cross validation correlation coefficient (BCVC) for 30 model runs of run 12 (Table 1 and Fig. 7) with different genetic algorithm generation/population sizes

of 2002–2004 for one of the best cross-validation correlation coefficients outcomes of 0.96, corresponding to a generation/population of 100/100 (Fig. 7). In Fig. 10 the algorithm convergence performance to 0.96 for the run described in Fig. 9 is presented. Fig. 10 shows that the model converges in a *stair-like* pattern to 0.96. The convergence pattern shown in Fig. 10 is typical to all runs. The stopping criterion selected in all runs was for attaining the maximum predefined number of generations. A no-improvement criterion for several subsequent generations could also be likely used.

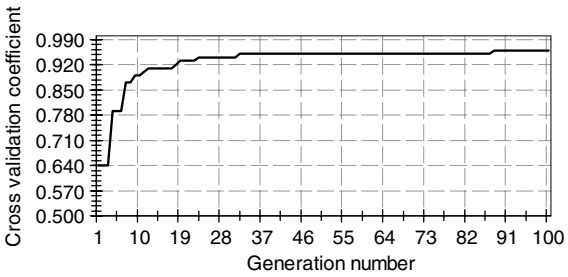


Fig. 10. Algorithm convergence to 0.96 cross validation coefficient for the run described in Fig. 9

Conclusions

To control cyanobacteria blooming in surface waters both the ability to model its growth and forecast its appearance are required. Reviewing the literature, most attempts, thus far, to make predictions regarding the physical growth of cyanobacteria have been limited.

This study, based on Ostfeld et al. (2006), suggests a new approach linking model trees with a genetic algorithm for both modeling and forecasting the biomass of toxic cyanobacteria in surface waters. The model results in a simple set of empirical linear rules and a prediction horizon. The correlation coefficient on a cross validation dataset estimated the accuracy of the prediction. The methodology was tested on Lake Kinneret (Sea of Galilee), Israel, through multiple trails with different parameters/constraints settings.

The developed model is generic and can be applied for any water quality lake prediction problem. To follow the proposed methodology, a vast database is required incorporating Physical, Chemical, Biology, and other (e.g., external loading) measurements, a data driven model (e.g., an M5 or an ANN engine), and a GA program. The reader is referred to Figs. 1 and 3 for more schematic details on the methodology and a possible interface, respectively. Limitations are expected if the database is small thus spanning only part of the decision space, or long running times of the GA if the duration of a single data driven model construction is substantial.

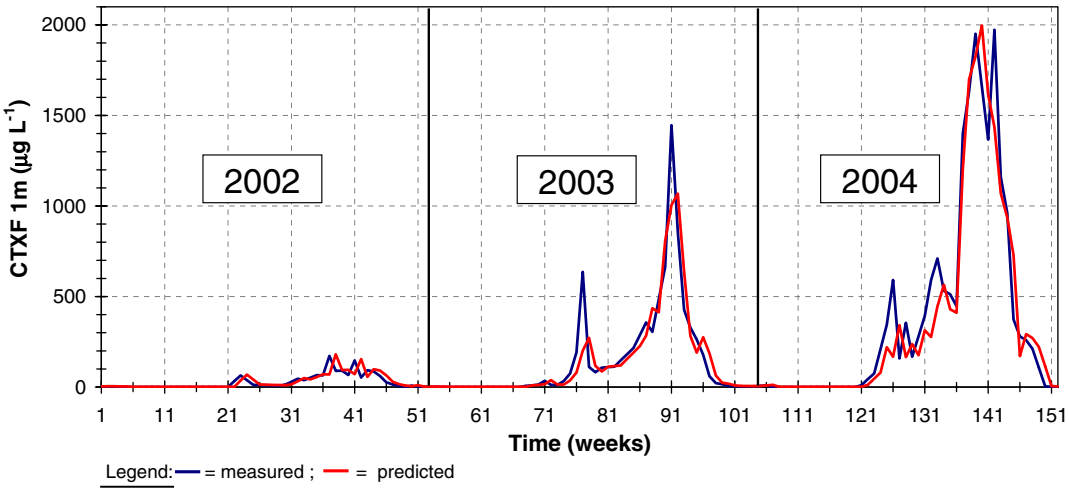


Fig. 9. CTXF 1 m [toxic, nitrogen-fixing cyanobacteria (in wet weight biomass, $\mu g\,L^{-1}$) at a depth of 1 m] prediction results for 2002–2004 for BCVC (best obtained cross validation correlation coefficient, also Fig. 7) of 0.96

Many of the model trials produced good matching as depicted by the correlation coefficient (Fig. 7). Those, however, incorporated different optimal input variables and forecast horizons.

Receiving a variety of good solutions raises a new problem of model selection: which of the model solutions need to be chosen? To quantitatively address this question a one step further in a multiobjective space needs to be undertaken, taking into consideration other objectives such as the reliability of the measured inputs and the forecast horizon. Development of a multiobjective model through utilizing the nondominated sorting genetic algorithm II (NSGA-II) (Deb et al. 2002), for example, to replace the current single objective GA, can be a nice extension of the proposed methodology.

The analysis in this work was limited to the Pearson's r correlation coefficient constrained through Cubist. Additional runs utilizing other correlation coefficients such as the RMSE, the Theil u-statistic, or the NSE were not explored, and should be the subject of future testing of the proposed methodology.

Other research challenges stemming from this study involve the development of methodologies for incorporating deterministic mathematical models (e.g., Guven and Howard 2011) into data-driven models, such as model trees or ANNs, and their comparison with physically based models such as that of Gal et al. (2009). Incorporation of ANNs, would, however, be more difficult to interpret. A MT provides simple linear rules as a prediction outcome and has certain advantageous over an ANN model, as observed by Solomatine and Xue (2004): "... model trees, being analogous to piecewise linear functions, have certain advantages compared with ANNs - they are more transparent and hence acceptable by decision makers, are very fast in training and always converge." In addition, modification/extension of the developed model to predict cyanobacteria over a given time horizon is warranted.

Dating the selection of the optimal variables for prediction, either for an ANN or in this study, relies on an algorithmic optimization search engine, with almost no consideration of the related physical variables. Inclusion of physical considerations is expected to enhance both the understanding of cyanobacteria growth and reduce the uncertainty of its prediction. Such considerations may include a weighted selection of input variables, augmentation of the utilized model data through physically-based model runs, and linking inputs and outputs for different time scales between data driven models (DDMs) and physically based models. In addition, a limitation which *distorts* the spatial variability of the variables, and thus the model ability to correctly capture the most influencing parameters, is the averaging of the measured variables (e.g., chemical parameters). Future extensions of this work should incorporate spatial variability considerations of the data such as standard deviation and median.

Another observation is related to the input variables selection which is a very important problem when developing DDMs because the choice of input variables has a tremendous effect on model complexity and performance. There are numerous methods employed to undertake the problem of input variable selection which can be broadly classified into three main classes: *wrappers, embedded, or filter algorithms* (Blum and Langley 1997; Guyon and Elisseeff 2003): (1) *wrapper algorithms* search through the set of combinations of input variables and select the subset that optimizes the performance of a trained DDM, (2) *embedded algorithms* directly incorporate input variable selection problem into the DDM training algorithm, and (3) *filter algorithms* distinctly separate the input variable selection problem from the DDM training and adopt an auxiliary statistical analysis technique to measure the relevance of individual or combinations of input variables.

In this study, the embedded approach (2) was selected in which selection of the input variables (including selecting input variables and corresponding time lags) is incorporated into the model training. Exploring wrapper and filter algorithms as described above should be the subject of further research.

Acknowledgments

This research was supported by the Fund for the Promotion of Research at the Technion, and by the Technion Grand Water Research Institute (GWRI). Data for this study was obtained through the courtesy of Mekorot-Israel National Water Company Co. and the Kinneret Limnological Laboratory (KLL), Israel Oceanographic & Limnological Research Ltd. (IOLR). The authors extend special gratitude to Dr. Alon Rimmer and Yury Lechinsky from KLL for providing meteorological data for this research.

Supplemental Data

Appendix S1 is available online in the ASCE Library (www.ascelibrary.org).

References

Ahn, C.-Y., Oh, H.-M., and Park, Y.-S. (2011). "Evaluation of environmental factors on cyanobacterial bloom in eutrophic reservoir using artificial neural networks." *J. Phycol.*, 47(3), 495–504.

Asadollahfardi, G., Taklify, A., and Ghanbari, A. (2012). "Application of artificial neural network to predict TDS in Talkheh Rud River." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000402, 363–370.

Barros, R. C., Basgalupp, M. P., Carvalho, A., and Freitas, A. A. (2012). "A survey of evolutionary algorithms for decision-tree induction." *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev.*, 42(3), 291–312.

Blum, A., and Langley, P. (1997). "Selection of relevant features and examples in machine learning." *Artif. Intell.*, 97(1–2), 245–271.

Bobbin, J., and Recknagel, F. (2001). "Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms." *Ecol. Model.*, 146(1–3), 253–262.

Bowden, G., Graeme, D., and Maier, H. (2005). "Forecasting cyanobacteria (blue-green algae) using artificial neural networks." Artificial neural networks in water supply engineering, S. Lingireddy and G. M. Brion, eds., ASCE, Reston, VA, 71–96.

Burke, I. J. (1991). "Introduction to artificial neural systems for pattern recognition." *Comput. Oper. Res.*, 18(2), 211–220.

Cao, H., Recknagel, F., Welk, A., Kim, B., and Takamura, N. (2006). "Hybrid evolutionary algorithm for rule set discovery in time-series data to forecast and explain algal population dynamics in two lakes different in morphometry and eutrophication." *Ecological informatics*, F. Recknagel, ed., 2nd Ed., Springer, Heidelberg, Germany, New York, 347–368.

Chan, W. S., Recknagel, F., Cao, H., and Park, H. D. (2007). "Elucidation and short-term forecasting of microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms." *Water Res.*, 41(10), 2247–2255.

Chang, K.-W., Shen, Y., and Chen, P.-C. (2004). "Predicting algal bloom in the Techi reservoir using Landsat TM data." *Int. J. Remote Sens.*, 25(17), 3411–3422.

Deb, K., Pratap, A., Agrawal, S., and Meyarivan, T. (2002). "A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: NSGA-II." *Proc., Parallel Problem Solving from Nature VI Conf.*, Paris, 849–858.

Devillers, J., et al. (2007). "Prediction of biological activity profiles of cyanobacterial secondary metabolites." *SAR QSAR Environ Res*, 18(7–8), 629–643.

Fernández, J. R. A., Nieto, P. J. G., Torres, J. M., and Muñiz, C. D. (2013). "Analysis of cyanotoxins presence from experimental cyanobacteria

concentrations in the Trasona reservoir (northern Spain) using support vector regression." *Int. J. Nonlinear Sci. Numer. Simul.*, 14(2), 103–112.

Flood, I. (2006). "Next generation artificial neural networks for civil engineering." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)0887-3801(2006)20:5(305), 305–307.

Flood, I., and Kartam, N. (1994a). "Neural networks in civil engineering. I: Principles and understanding." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)0887-3801(1994)8:2(131), 131–148.

Flood, I., and Kartam, N. (1994b). "Neural networks in civil engineering. II: Systems and application." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)0887-3801(1994)8:2(149), 149–162.

Gal, G., Hipsey, M. R., Parparov, A., Wagner, U., Makler, V., and Zohary, T. (2009). "Implementation of ecological modeling as an effective management and investigation tool: Lake Kinneret as a case study." *Ecol. Modell.*, 220(13–14), 1697–1718.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, New York.

Guven, B., and Howard, A. (2006). "A review and classification of the existing models for cyanobacteria." *Prog. Phys. Geogr.*, 30(1), 1–24.

Guven, B., and Howard, A. (2007). "Identifying the critical parameters of a cyanobacterial growth and movement model by using generalised sensitivity analysis." *Ecol. Modell.*, 207(1), 11–21.

Guven, B., and Howard, A. (2011). "Sensitivity analysis of a cyanobacterial growth and movement model under two different flow regimes." *Environ. Model. Assess.*, 16(6), 577–589.

Guyon, I., and Elisseeff, A. (2003). "An introduction to variable and feature selection." *J. Mach. Learn. Res.*, 3, 1157–1182.

Hadas, O., et al. (2012). "Appearance and establishment of diazotrophic cyanobacteria in Lake Kinneret, Israel." *Freshwater Biol.*, 57(6), 1214–1227.

Haupt, R. L., and Haupt, S. E. (1998). *Practical genetic algorithms*, Wiley, Canada.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Arbor, MI.

Ibelings, B. W., Vonk, M., Los, H. F. J., van der Molen, D. T., and Mooij, W. M. (2003). "Fuzzy modeling of Cyanobacterial surface waterblooms: Validation with NOAA-AVHRR satellite images." *Ecol. Appl.*, 13(5), 1456–1472.

Kingston, G. B., Maier, H. R., and Lambert, M. R. (2006). "Forecasting cyanobacteria with Bayesian and deterministic artificial neural networks." *Int. Joint Conf. on Neural Networks*, Vancouver, Canada, 4870–4877.

Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps." *Biol. Cybern.*, 43(1), 59–69.

Kutser, T., Metsamaa, L., Strombeck, N., and Vahtmae, E. (2006). "Monitoring cyanobacterial blooms by satellite remote sensing." *Estuarine Coastal Shelf Sci.*, 67(1–2), 303–312.

Lilover, M.-J., and Laanemets, J. (2006). "A simple tool for the early prediction of the cyanobacteria Nodularia spumigena bloom biomass in the Gulf of Finland." *Oceanologia*, 48(S), 213–229.

Maier, H. R. (1995). "Use of artificial neural networks for modeling multivariate water quality time series." Ph.D. thesis, Univ. of Adelaide, Adelaide, Australia.

Maier, H. R., and Dandy, G. C. (1996). "The use of artificial neural networks for the prediction of water quality parameters." *Water Resour. Res.*, 32(4), 1013–1022.

Maier, H. R., and Dandy, G. C. (1997). "Modeling cyanobacteria (blue green algae) in the River Murray using artificial neural networks." *Math. Comput. Simul.*, 43(3–6), 377–386.

Maier, H. R., Dandy, G. C., and Burch, M. D. (1998). "Use of artificial neural networks for modeling cyanobacteria Anabaena spp. in the River Murray, South Australia." *Ecol. Model.*, 105(2–3), 257–272.

Maier, H. R., Sayed, T., and Lence, B. J. (2000). "Forecasting cyanobacterial concentrations using B-spline networks." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)0887-3801(2000)14:3(183), 183–189.

Markel, D., and Shamir, U. (2002). "Monitoring lake Kinneret and its watershed: Forming the basis for management of a water supply lake." *Water resources quality preserving the quality of our water resources*, H. Rubin, P. Nachtnebel, J. Fuerst, and U. Shamir, eds., Springer, Berlin, Heidelberg, 177–190.

Muttil, N., and Chau, K.-W. (2006). "Neural network and genetic programming for modeling coastal algal blooms." *Int. J. Environ. Pollut.*, 28(3–4), 223–238.

Nicklow, J., et al. (2010). "State of the art for genetic algorithms and beyond in water resources planning and management." *J. Water Resour. Plann. Manage. Div.*, 10.1061/(ASCE)WR.1943-5452.0000053, 412–432.

Nieto, P. J. G., Fernández, J. R. A., de Cos Juez, F. J., Lasheras, F. S., and Muñiz, C. D. (2013). "Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (northern Spain)." *Environ. Res.*, 122, 1–10.

Ostfeld, A., Tubaltzev, A., Rom, M., and Kronaveter, L. (2006). "A hybrid model tree (MT)–Genetic algorithm (GA) scheme for toxic cyanobacteria predictions in Lake Kinneret." *Examining the Confluence of Environmental and Water Concerns, Proc., World Environmental and Water Resources Congress*, ASCE, Reston, VA.

Preis, A., Tubaltzev, A., and Ostfeld, A. (2006). "Kinneret watershed analysis tool: A cell based decision tree model for watershed flow and pollutants predictions." *Water Sci. Technol.*, 53(10), 29–35.

Qingyu, W., Nanb, J., Hengc, L., and Bin, H. (2008). "The modeling for dynamic algae blooms prediction based on remote sensing." *The international archives of the photogrammetry, remote sensing and spatial information sciences*, Vol. XXXVII, 1543–1548. ⟨http://www.isprs.org/proceedings/XXXVII/congress/4_pdf/270.pdf⟩ (June 9, 2014).

Quiblier, C., Wood, S., Echenique-Subiabre, I., Heath, M., Villeneuve, A., and Humbert, J.-F. (2013). "A review of current knowledge on toxic benthic freshwater cyanobacteria—Ecology, toxin production and risk management." *Water Res.*, 47(15), 5464–5479.

Quinlan, J. R. (1992). "Learning with continuous classes." *Proc. A-92*, Adams, and Sterling, eds., World Scientific, Singapore, 343–348.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA.

Ranjithan, S., Eheart, J. W., and Garrett, J. H. (1993). "Neural network-based screening for groundwater reclamation under uncertainty." *Water Resour. Res.*, 29(3), 563–574.

Recknagel, F. (1997). "ANNA–Artificial neural network model for predicting species abundance and succession of blue-green algae." *Hydrobiologica*, 349(1–3), 47–57.

Recknagel, F., Bobbin, J., Whigham, P., and Wilson, H. (2002). "Comparative application of artificial neural networks and genetic algorithms for multivariate time series modelling of algal blooms in freshwater lakes." *J. Hydroinf.*, 4(2), 125–134.

Recknagel, F., and Cao, H. (2009). "Ecological informatics by means of neural, evolutionary and object-oriented computation." Chapter 9, *Handbook of ecological modelling and informatics*, WIT Transactions on State of the Art in Science and Engineering, Vol. 34, ⟨http://library.witpress.com/pages/PaperInfo.asp?PaperID=22703⟩ (June 9, 2014).

Recknagel, F., Cao, H., van Ginkel, C., van der Molen, D., Park, H., and Takamura, N. (2008a). "Adaptive agents for forecasting seasonal outbreaks of blue-green algal populations in lakes categorised by circulation type and trophic state." *Proc., Int. Assoc. Theor. Appl. Limnol.*, 30(2), 191–197.

Recknagel, F., French, M., Harkonen, P., and Yabunaka, K. L. (1997). "Artificial neural network approach for modeling and prediction of algal blooms." *Ecol. Modell.*, 96(1–3), 11–28.

Recknagel, F., van Ginkel, C., Cao, H., Cetin, L., and Zhang, B. (2008b). "Generic limnological models on the touchstone: Testing the lake simulation library SALMO-OO and the rule-based microcystis agent for warm-monomictic hypertrophic lakes in South Africa." *Ecol. Modell.*, 215(1–3), 144–158.

Reynolds, C. S. (1984). *The ecology of freshwater phytoplankton*, Cambridge University Press, Cambridge, U.K.

Rolland, D. C., Bourget, S., Warren, A., Laurion, I., and Vincent, A. F. (2013). "Extreme variability of cyanobacterial blooms in an urban drinking water supply." *J. Planktonic Res.*, 35(4), 744–758.

Salomons, E. (2002). "optiGA: An ActiveX control (OCX) for genetic algorithms (GA)." ⟨http://www.optiwater.com/optiga.html⟩ (June 10, 2014).

J. Water Resour. Plann. Manage.

Solomatine, D., and Ostfeld, A. (2008). "Data driven modeling: Some past experiences and new approaches." *J. Hydroinf.*, 10(1), 3–22.
Solomatine, D. P., and Xue, Y. (2004). "M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China." *J. Hydrol. Eng.*, 10.1061/(ASCE)1084-0699(2004)9:6(491), 491–501.
Teles, L. O., Vasconcelos, V., Pereira, E., and Saker, M. (2006). "Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks." *Environ. Manage.*, 38(2), 227–237.
Torres, R., Pereira, E., Vasconcelos, V., and Teles, L. O. (2011). "Forecasting of cyanobacterial density in Torrão reservoir using artificial neural networks." *J. Environ. Monit.*, 13(6), 1761–1767.
Vilán, J. A. V., Fernández, J. R. A., Nieto, P. J. G., Lasheras, F. S., de Cos Juez, F. J., and Muñiz, C. D. (2013). "Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (northern Spain)." *Water Resour. Manage.*, 27(9), 3457–3476.
Wang, Q. J. (1997). "Using genetic algorithms to optimise model parameters." *Environ. Modell. Softw.*, 12(1), 27–34.
Wang, Y., and Witten, I. H. (1996). "Induction of model trees for predicting continuous classes." *Working Paper Series ISSN 1170-487X*, Dept. of Computer Science, Univ. of Waikato, Hamilton, New Zealand.
Wei, B., Sugiura, N., and Maekawa, T. (2001). "Use of artificial neural network in the prediction of algal blooms." *Water Res.*, 35(8), 2022–2028.
Welk, A., Recknagel, F., Cao, H., Chan, W. S., and Talib, A. (2008). "Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms." *Ecol. Inf.*, 3(1), 46–54.
Wen, C.-G., and Lee, C.-S. (1998). "A neural network approach to multiobjective optimization for water quality management in a river basin." *Water Resour. Res.*, 34(3), 427–436.
Whigham, P., and Recknagel, F. (2001a). "An inductive approach to ecological time series modelling by evolutionary computation." *Ecol. Modell.*, 146(1–3), 275–287.
Whigham, P., and Recknagel, F. (2001b). "Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms." *Ecol. Modell.*, 146(1–3), 243–251.
Whitehaed, P. G., Howard, A., and Arulmani, C. (1997). "Modeling algal growth and transport in rivers—A comparison of time seres analysis, dynamic mass balance and neural network techniques." *Hydrobiologica*, 349(1–3), 39–46.
Witten, H. I., and Frank, E. (2005). *Data mining practical machine learning tools and techniques*, 2nd Ed., Morgan Kaufmann Publishers, San Francisco, CA.
Yabunaka, K. I., Hosomi, M., and Murakami, A. (1997). "Novel application of a back-propagation artificial neural network model formulated to predict algal bloom." *Water Sci. Technol.*, 36(5), 89–97.
Yao, Z., Fei, M., Li, K., Kong, H., and Zhao, B. (2007). "Recognition of blue-green algae in lakes using distributive genetic algorithm-based neural networks." *Neurocomputing*, 70(4–6), 641–647.
Zou, R., Lung, W.-S., and Wu, J. (2007). "An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling." *Water Resour. Res.*, 43(8), in press.