Evaluation of models generated via hybrid evolutionary algorithms for the prediction of *Microcystis* concentrations in the Vaal Dam, South Africa

A Swanepoel[1], S Barnard*[2], F Recknagel[3] and H Cao[3]
[1]*Scientific Services, Rand Water, PO Box 3526, Vereeniging 1930, South Africa*
[2]*School for Biological Sciences, North-West University, South Africa*
[3]*School of Biological Sciences, University of Adelaide, Australia*

ABSTRACT

Cyanobacteria are responsible for many problems in drinking water treatment works (DWTW) because of their ability to produce cyanotoxins that potentially can have an adverse effect on consumer health. Therefore, the monitoring of cyanobacteria in source waters entering DWTW has become an essential part of drinking water treatment management. Managers of DWTW rely heavily on results from physical, chemical and biological water quality analyses, from grab samples, for their management decisions. However, results of water quality analyses may be delayed from 3 h to 14 days depending on a magnitude of factors such as sampling, distance and accessibility to laboratory, laboratory sample turn-around times, specific methods used in analyses, etc. Therefore, the benefit to managers and production chemists to be able to forecast future events of high cyanobacterial cell concentrations in the source water is evident. During this study, physical, chemical and biological water quality data from samples taken from 2000 to 2009 in the Vaal Dam, supplying South Africa's largest bulk drinking water treatment facility, were used to develop models for the prediction of the cyanobacterium *Microcystis* sp. in the source water (real-time prediction together with 7, 14 and 21 days in advance). Water quality data from the Vaal Dam from 2010–2012 were used to test these models. The model showing the most promising results for incorporation into a 'Cyanobacterial Incident Management Protocol' is the one predicting *Microcystis* sp. 7 days in advance. This model showed a square correlation coefficient (R^2) of 0.90 when tested with the testing dataset (chosen by bootstrapping from the 2000–2009 input dataset) and a R^2 of 0.53 when tested with the 3-year 'unseen' dataset from 2010–2012.

Keywords: cyanobacteria, drinking water treatment works, prediction models, cyanobacterial incident management protocol, water safety plan

INTRODUCTION

Algae and cyanobacteria occur naturally in source waters worldwide. However, certain species are known to form harmful blooms (Harding and Paxton, 2001), which can cause extensive problems in the drinking water treatment industry (Knappe et al., 2004; Meriluoto and Codd, 2005; Zoschke et al., 2011). Cyanobacteria (especially *Microcystis* sp.) are widely responsible for many water treatment problems due to their ability to produce organic compounds. These organic compounds include the cyanotoxin microcystin (Conradie and Barnard, 2012), which can have an adverse effect on consumer health, as well as taste and odour compounds (like geosmin and 2methylisoborneol) that decrease consumer confidence in drinking water (Zoschke et al., 2011). Therefore, the monitoring of cyanobacteria in source waters entering drinking water treatment works (DWTW) has become an essential part of drinking water treatment management (Swanepoel et al., 2008).

Recently Cyanobacterial Incident Management Protocols (Du Preez and Van Baalen, 2006; Du Preez et al., 2007) and Water Safety Plans (Bartram et al., 2009) have been used to manage incidents of, for example, high cyanobacteria concentrations in source water destined for drinking water purification. In order to fully utilise these management tools (protocols and safety plans), managers and production chemists of DWTW, rely heavily on results of physical, chemical and biological water quality analyses for their water treatment and management decisions. However, results of water quality analyses can be delayed from 3 h to 7 days or longer, depending on factors such as sampling, distance and accessibility to laboratories, laboratory sample turnaround times, and specific methods used in the analysis, etc. (Swanepoel et al., 2008). Therefore, the application value of models that are able to predict the cyanobacteria concentration in source waters, a few days or weeks in advance, is evident. Such models will enable managers and production chemists of DWTW to prepare for a cyanobacteria-related incident before it occurs.

Previous studies have demonstrated that highly complex ecological time-series data can be successfully probed to develop rule sets as prediction tools, by using hybrid evolutionary algorithms (HEAs) (Talib et al, 2007; Chan et al., 2007; Recknagel et al, 2008; Van Ginkel, 2008; Welk et al, 2008; Recknagel et al., 2013 and Recknagel et al., 2014). Ecological data is considerably more prone to observational and/or measurement noise and the ecological interactions are inherently more complex and nonlinear. In a previous study by Van Ginkel et al. (2010), different ecological informatics modelling techniques were compared. The rule set discovered by hybrid evolutionary algorithms (HEA) proved to be highly applicable to the hypertrophic reservoirs of South Africa. During the current study, physical, chemical and biological water quality data from samples collected from 2000 to 2009 in the Vaal Dam were used to develop models for the prediction of

* To whom all correspondence should be addressed.
☎ +018-299 2508; e-mail: sandra.barnard@nwu.ac.za
Received: 26 July 2014; accepted in revised form 8 March 2016

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

Microcystis sp. in the source water. The aim of this study was to evaluate the suitability of *Microcystis* sp. prediction models in the Vaal Dam (real-time, 7, 14 and 21 days in advance), for application to a large bulk drinking water treatment facility and possible incorporation into its 'Cyanobacterial Incident Management Protocol' (Du Preez and Van Baalen, 2006). This will enable the DWTW to initiate preventative measures for dealing with source water containing high concentrations of *Microcystis* sp. cells, before it even reaches the plant.

MATERIALS AND METHODS

Study site

The Vaal Dam (Fig. 1) is approximately 150 km south of Johannesburg, South Africa. The catchment area of the dam is approximately 38 500 km^2 with a wall height of 63.4 m above the lowest foundation (DWA, 2013b). The Lesotho Highlands Water Project pumps water into the system in order to supply water to the industrial hub of Gauteng. This water is being transported from Lesotho via the Liebenbergsvlei and Wilge Rivers (LHDP, 2013). The Vaal Dam is classified as mesotrophic, according to the classification system used by the South African Department of Water Affairs (DWA), where mean total phosphate (0.077 mg/L), mean chlorophyll-*a* concentration (14.8 µg/L) and percentage of time where chlorophyll-*a* is >30 µg/L (17%) is taken into account (DWA, 2013a).

From the Vaal Dam, a 20 km long canal supplies water to Stations 3 and 4 at the Zuikerbosch DWTW – South Africa's largest bulk drinking water treatment facility (Fig. 1). This facility can produce approx. 3 000 ML of drinking water per day (depending on demand). Samples for analyses are collected at the dam wall (coordinates: X: 28.12059553; Y: –26.88444867); the lake behind the dam wall has a surface area of about 320 km^2 and is 47 m deep at full capacity (DWA, 2013b). Results from physical, chemical, and biological analyses done by Rand Water's Analytical Services Laboratory on water samples from the Vaal Dam supplying the Zuikerbosch DWTW, for the period 2000 to 2012, were used in this study (Fig. 1).

Physical, chemical and biological analyses of water

Sampling and laboratory analyses of samples from the Vaal Dam took place once a month. All chemical and biological analyses were carried out according to SANAS (South African National Accreditation System – affiliated at ILAC), accredited standard methods (APHA, 2013).

The *Microcystis* sp. counts were performed according to the phytoplankton identification and enumeration method described by Swanepoel et al. (2008). During sample preparation, the gas vacuoles of cyanobacteria were pressure-deflated using a specially-designed mechanical hammer that exerts a pressure of 49.5 kPa on the sample (Walsby, 1971, 1994), which is approximately the pressure needed to collapse the gas vacuoles of cyanobacteria. The sample was then homogenised at 13 000 r/min for ±15 s after which 3 mL of sample was pipetted into a sedimentation chamber. The sedimentation chambers were then centrifuged for 10 min at 3 500 r/min to allow phytoplankton cells to settle to the bottom thereof. After settling, all phytoplankton cells were identified and enumerated with an inverted light microscope, using the technique described by Lund et al. (1958) and adapted for Rand Water by Swanepoel et al. (2008). One of the eyepieces of the microscope contains a Whipple grid to delineate the counting area (called a 'field'). The glass bottoms of the sedimentation chambers were examined in 'fields' covering most parts of the sedimentation chamber, while counting all algal cells inside the grid or 'field'. The original sub-sample volume that was transferred to the sedimentation chamber, the area of the sedimentation chamber, the area of a 'field' as well as the number of 'fields' counted, were used to calculate the concentration of individual phytoplankton genera as cells per millilitre (cells/mL).

Statistical analyses

Principal component analysis (PCA) was carried out on the input dataset used for the model development in order to characterise the water in the dam according to the relationships between variables. All physical, chemical and biological variables were used as concentrations but centred and standardised to compensate for unit differences in the PCA. The cyanobacteria concentration was the only variable transformed to the natural log of the concentrations to reduce the large variability in the cyanobacteria counts. The computer package CANOCO, Version 4.5 was used (Ter Braak, 1988) to perform the PCA. Ordinations were interpreted using the following rationale: Parameters are (i) positively correlated with each other if their arrows subtend a small angle, (ii) not correlated if their arrows are 90°, (iii) negatively correlated if their arrows are directed oppositely (180°); (iv) parameters with the longest arrow relative to an axis have the greatest influence on that axis.

Square correlation coefficients (R^2) and root mean square error (RMSE) of the models were tested with (i) 25% of the data from the original database (2000 – 2009) that was used for training the models (chosen by bootstrapping and called the 'testing database') and (ii) 3 years of 'unseen data' (data not used in training the models – 2010–2012), were determined by XLSTAT, Version 2009.4.06.

Hybrid evolutionary algorithms (HEAs)

Evolutionary algorithms (EAs) are adaptive methods used in search of suitable representations of models, which recognise patterns in data sets. EAs mimic the processes of biological evolution, natural selection and genetic variation based on the principle of 'survival of the fittest' (Welk et al., 2008, from Cao et al., 2006). EAs have been designed to discover predictive rule

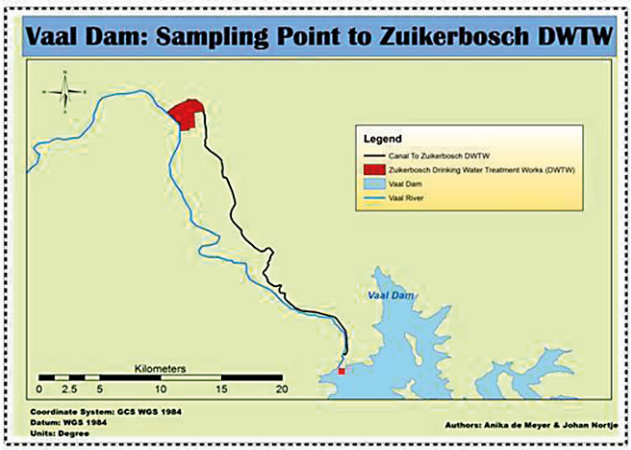

Figure 1
Sampling point (n) in the Vaal Dam supplying untreated water to the Zuikerbosch Drinking Water Treatment Works (DWTWs)

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

sets in complex ecological time-series data by applying genetic programming for the optimisation of the rule structures (IF *x*, THEN *y*, ELSE *z*) and genetic algorithms for the optimisation of parameters of the rule sets (Cao et al., 2006; Recknagel et al., 2008). For this study a Hybrid EA (HEA) designed for rule discovery in water-quality time-series was applied (Cao et al., 2006). Hybridisation was used in order to improve the performance of the evolutionary algorithm and to improve the quality of the solutions obtained by the algorithm (Grosan and Abraham, 2007). Improvement of the models was achieved by structure optimisation using genetic programming as well as parameter optimisation by using genetic algorithms (Welk et al., 2008).

The HEA was applied for short-time forecasting of *Microcystis* sp. concentrations in the Vaal Dam using physical data (turbidity, water temperature and Secchi disk depth) and chemical data (conductivity, pH, dissolved oxygen, PO_4^{3-}, NO_3^-, NH_4^+, Si, Fe^{2+}, Mn^{2+}, chemical oxygen demand), as well as biological data (chlorophyll–*a* concentration and initial cyanobacterium inoculum). Table 1 displays the descriptive statistical values of the input data used for the model development.

Because of the fact that samples were only taken once a month, and prediction time necessitated date ranges of 7, 14 and 21 days, the data were linearly interpolated to have corresponding results for all variables at a frequency of 360 days per year. For the development and first stage of testing of the models, 75% of the dataset was used to train the models and 25% of the dataset was used to test the models. Boot-strapping (i.e. random selection) were used to determine which 75% of the dataset was to be used as the training dataset and which 25% as the testing dataset. Boot-strapping also implies that a different 75% portion of the data will be used during training for all the different models. Fifty different models were developed for each set of '*x* input variables = 1 output variable' chosen beforehand. From the 50 models, the best model relating the measured data and the predicted data were chosen based on the following criteria: (i) root mean square error (RMSE), (ii) the square correlation coefficient (R^2-value) and (iii) visual comparison between the predicted and measured data as according to Chan et al. (2007) and Bennet et al. (2013).

For applications of the HEAs an initial population of 100 and a maximum number of generations and repetitive runs of 80 were chosen, because the database was relatively large and 80 repetitive runs could take anything from 24 h to 72 h to complete. The rule-sets were discovered and optimised using a large-scale parallel computational device and relevant software developed in the Ecoinformatics and Watershed Ecology Laboratory at the University of Adelaide, Australia.

Sensitivity analyses were carried out for the best performing predictive rule sets as follows: The minimum, maximum and median of all input variables used to develop the model were determined. A linear range of all the variables used in each model (either in the THEN or the ELSE branch) was constructed ranging from the minimum (at 0%) to the maximum (at 100%) in increments of 5%. To determine the sensitivity of the model towards a specific variable, the model was tested by substituting all variables with the median thereof, except for the variable being tested. The tested variable was substituted with the range of values from 0% to 100% on the *x*-axis and the result from the model on the *y*-axis. The curve with the steepest slope (either positive or negative) was identified as the variable towards which the model showed the greatest sensitivity. This implies that small changes in a variable towards which the model shows a high sensitivity will have a bigger influence on the result of the model when compared to a variable towards which the model shows a low sensitivity.

RESULTS

Characterisation of Vaal Dam water

A PCA was performed on the same dataset used to develop the models (Vaal Dam monthly collected physical, chemical

TABLE 1
The descriptive statistical values of the measured data in the Vaal Dam (2000–2009) used as input data for model development (*n* = 165)

	Minimum	Maximum	Average	Standard deviation
Chlorophyll-*a* (µg/L)	0.67	194	12.82	17.72
Chemical Oxygen Demand (COD – mg/L)	0.01	34	14.41	4.76
Conductivity (Cond – mS/m)	13.9	35.9	19.87	3.18
Dissolved Oxygen (DO – mg/L)	3.52	22.5	7.95	2.42
Fe^{2+} (mg/L)	0.022	3.245	0.536	0.56
Mn^{2+} (mg/L)	0.001	0.078	0.014	0.013
NH_2^- (mg/L)	0.002	0.65	0.044	0.074
NO_2^- (mg/L)	0.003	1.165	0.058	0.108
NO_3^- (mg/L)	0.01	2.94	0.286	0.349
pH	5.83	9.9	7.862	0.578
PO_4^{3-} (mg/L)	0.001	0.37	0.037	0.044
Secchi disk depth (cm)	18	128	32.718	13.748
Si (mg/L)	0.202	17.844	5.416	2.771
Turbidity (Turb – NTU)	8.57	141	55.079	27.400
Water temperature (Temp – °C)	9.33	26	17.762	4.140
Initial Cyanobacteria inoculum (cells/mL)	0	89 626	6 996.003	14 440.55

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

and biological data from 2000–2009) in order to characterise the water in the dam according to the relationships between variables. The results from the PCA included the following (i) physical variables: turbidity (Turb), water temperature (Temp) and Secchi disk depth (Sec); (ii) chemical variables: dissolved oxygen (DO), pH, electrical conductivity (Cond), Fe^{2+}, Si, NO_3^-, NO_2^-, NH_4^+, PO_4^{3-}, chemical oxygen demand (COD), and (iii) biological variables: chlorophyll-*a* (Chl*a*) and cyanobacteria (LnCyano). The ranges for these variables are summarised in Table 1. The Eigen values from the PCA are displayed in Table 2 and the results are represented in Fig. 2.

From the results in Table 2 and Fig. 2, it is evident that the first axis, which accounts for 22% of the variation, mostly explains the variation in the nutrients (NO_2^-, NO_3^-, NH_4^+ and PO_4^{3-}) as well as chemical oxygen demand (COD), Mn^{2+} and water temperature (Temp). The second axis, which accounts for an additional 18% of the variation, mostly explains the variation in turbidity (Turb), Secchi disk depth (Sec), Fe^{2+}, Si, pH, electrical conductivity (Cond), chlorophyll-*a* (Chl*a*) and cyanobacteria (LnCyano).

Nutrients (NH_4^+, NO_2^-, NO_3^- and PO_4^{3-}) together with Mn^{2+} are higher during the colder winter months, since the arrows representing them lie in the opposite direction to the arrow representing temperature. High turbidity associates closely with high pH, high Si, high Fe^{2+} and high chlorophyll-*a* (Chl*a*), while high cyanobacteria (LnCyano) concentrations associate with low conductivity (Cond), low dissolved oxygen (DO) and low Secchi disk depth (Sec). The arrow representing chlorophyll-*a* subtends a ±90° angle with water temperature (Temp) indicating that high chlorophyll-*a* concentrations do not only occur during summer or winter, but vary throughout the year in the Vaal Dam. One can therefore deduce that the chlorophyll-*a* level is not solely caused by the presence of cyanobacteria, but other phytoplankton as well. Chlorophyll-*a* shows a positive correlation with pH indicating that, during periods where algal blooms occur, pH increases, most probably due to the consumption of CO_2 during photosynthesis. Chlorophyll-*a* (Chl*a*) and cyanobacteria (LnCyano) show a negative correlation with dissolved oxygen (DO).

Models and related sensitivity analyses

Real-time Microcystis *sp. prediction*

For the best model developed for real-time *Microcystis* sp. prediction the IF criterion of the model (Fig. 3) is determined by the Fe^{2+} concentration. The THEN branch of the model (Fig. 3a) represents the low-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Fig. 3b) represents the high-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables used in the model (conductivity and Mn^{2+} in the THEN branch and pH, DO and chlorophyll-*a*, in the ELSE branch) display very little influence on the predicted *Microcystis* sp. concentration.

The comparison, between the measured *Microcystis* sp. concentration and that resulting from the models predicting real-time *Microcystis* sp. when using the 25% boot-strapped testing dataset (Fig. 4a), shows a R^2-value of 0.95 and a root mean square error (RMSE) of 4 262.2 cells/mL. When the model was tested with 3 years of 'unseen data', the correlation showed a R^2-value of 0.97 and a RMSE of 4 766.6 cells/mL (Fig. 4b), indicating that the event prediction of increased *Microcystis* sp. concentration together with the magnitude of the event displayed a significant correlation.

Microcystis *sp. prediction 7 days in advance*

For the best model developed for the prediction 7 days in advance the IF criterion of the model is determined by a combination of conductivity, PO_4^{3-} and pH (Fig. 5). The THEN branch of the model (Fig. 5a) represents the high-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Fig. 5b) represents the low-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables in this model (namely conductivity and DO), in comparison to the initial cyanobacteria concentration, display very little influence on the predicted *Microcystis* sp. concentration.

The comparison, between the measured *Microcystis* sp. concentration and that resulting from the models predicting *Microcystis* sp. 7 days in advance, when using the 25% boot-strapped testing dataset (Fig. 6a), shows a R^2-value of 0.90 and a RMSE of 3 135.7 cells/mL. When the model was tested with 3 years of 'unseen data' (Fig. 6b), the model showed a R^2-value of 0.53 and a RMSE of 44 559 cells/mL. The event prediction

TABLE 2
Eigen values for the PCA

Axes	1	2	3	4	Total variance
Eigen values	0.220	0.177	0.107	0.087	1.000
Cumulative percentage variance of data	22.0	39.7	50.5	59.2	

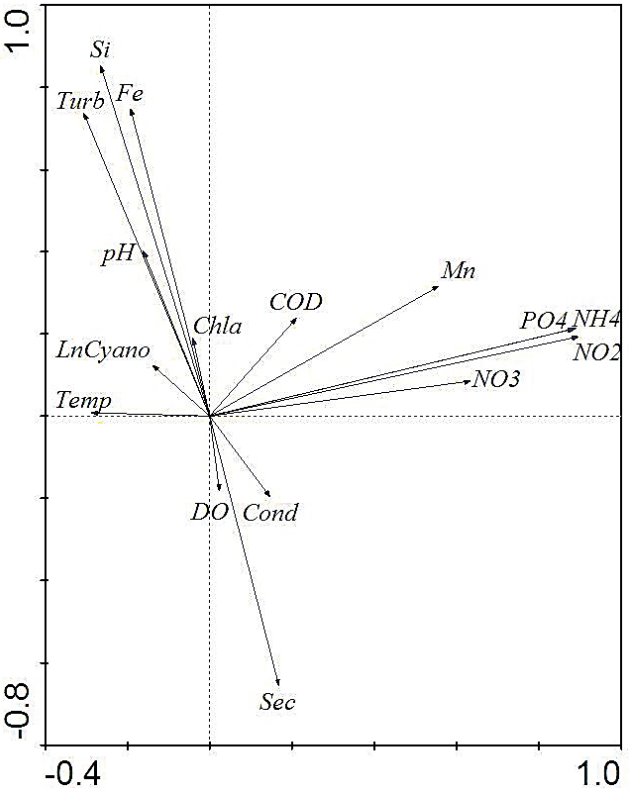

Figure 2
Principle component analysis (PCA) performed on physical, chemical and biological variables measured in the Vaal Dam from 2000–2009.

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

of increased *Microcystis* sp. concentration showed a significant correlation; however, it seems that the *Microcystis* sp. concentration is over-estimated somewhat by the model.

Microcystis *sp. prediction 14 days in advance*

The IF criterion of the best model developed for the prediction 14 days in advance is determined by the chemical oxygen demand (COD) (Fig. 7). The THEN branch of the model (Fig. 7a) represents the high-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Fig. 7b) represents the low-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables in this model, namely dissolved oxygen (DO) and NH_4^+ in the THEN branch and DO and water temperature (Temp) in the ELSE branch, display very little influence on the predicted *Microcystis* sp. concentration.

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis* sp. 14 days in advance, when using the 25% bootstrapped testing dataset (Fig. 8a), shows a R^2-value of 0.79 and a RMSE of 4 493.7 cells/mL. When the model was tested with 3 years of 'unseen data' (Fig. 8b), the model showed a R^2-value of 0.39 and a RMSE of 48 129.6 cells/mL. The event prediction of increased *Microcystis* sp. concentration together with the magnitude of the event showed a significant correlation.

Microcystis *sp. prediction 21 days in advance*

The IF criterion of the best model developed for the prediction 21 days in advance, is determined by a combination of nutrients (PO_4^{3-} and NH_4^+) concentrations (Fig. 9). The THEN branch of the model (Fig. 9a) represents the low-range rule set and shows

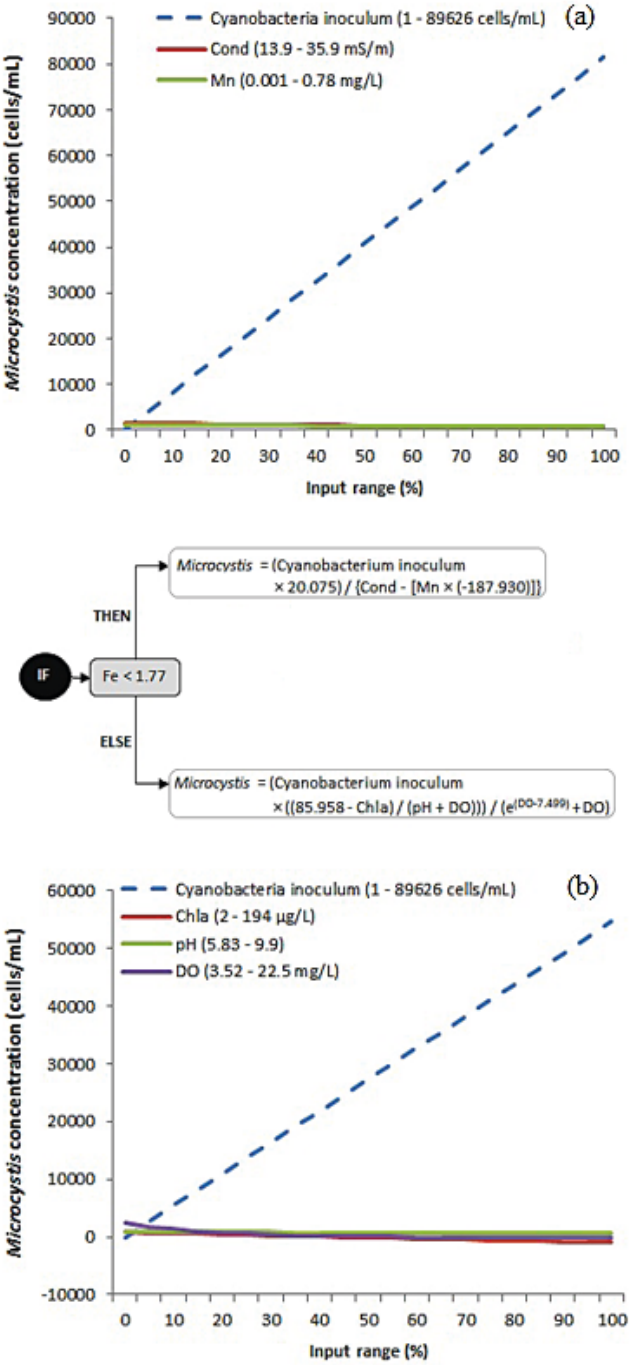

Figure 3

Prediction model and associated sensitivity analyses of real-time Microcystis *sp. concentration in the Vaal Dam: (a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch*

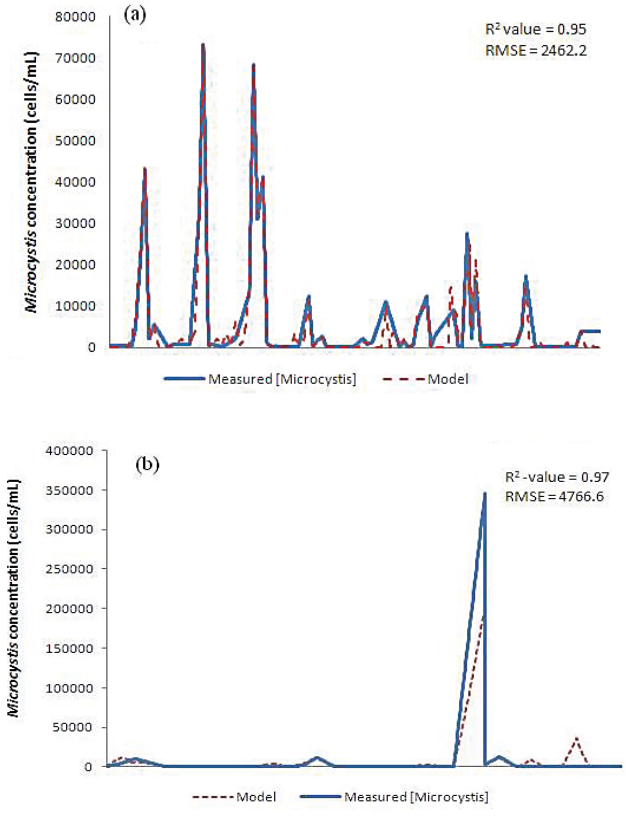

Figure 4

(a) Comparison between the measured Microcystis *sp. concentration and predicted real-time* Microcystis *sp. concentration in the Vaal Dam using 25% (boot-strapped) of the 10-year development dataset . (b) Comparison between the measured* Microcystis *sp. concentration and predicted real-time* Microcystis *sp. concentration using 3 years' 'unseen data' from the Vaal Dam*

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

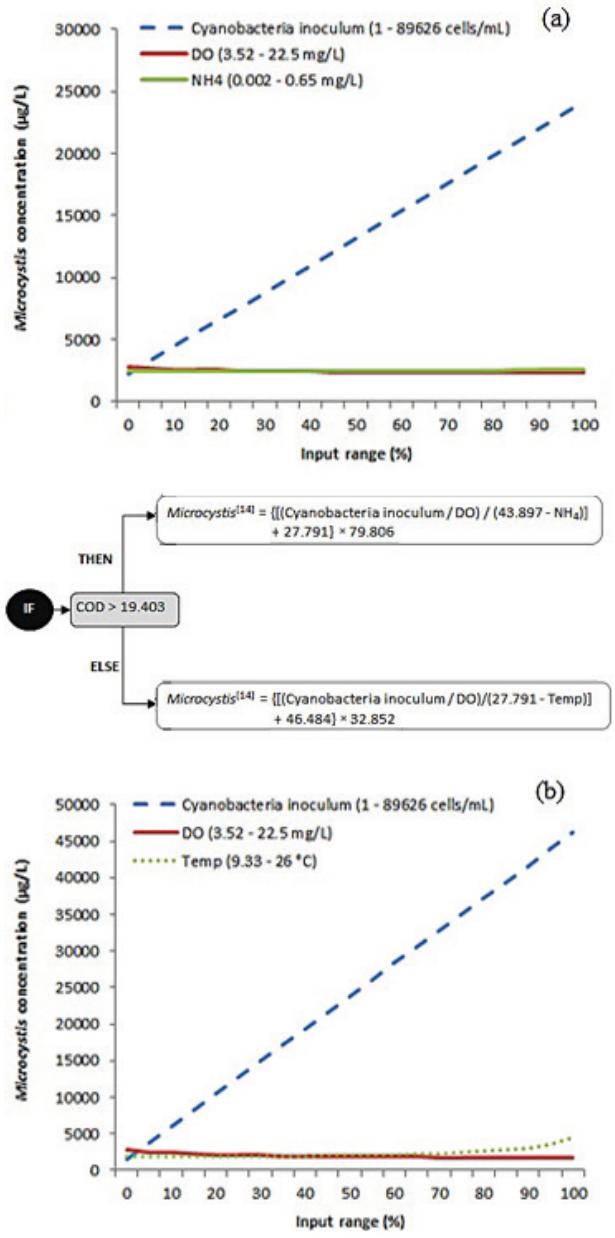

Figure 5

The 7-day prediction model and associated sensitivity analyses for Microcystis *sp. concentration in the Vaal Dam: (a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch*

the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Fig. 9b) represents the high-range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum as well as the Si concentration, particularly during the lower 10% of the input range. The other variables in this model (dissolved oxygen (DO), Si, chlorophyll-*a*, and NO_3^- in the THEN branch, and turbidity in the ELSE branch), display very little influence on the predicted *Microcystis* sp. concentration.

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis* sp. 21 days in advance, when using the 25% boot-strapped testing dataset (Fig. 10a), shows a R^2-value of 0.74 and

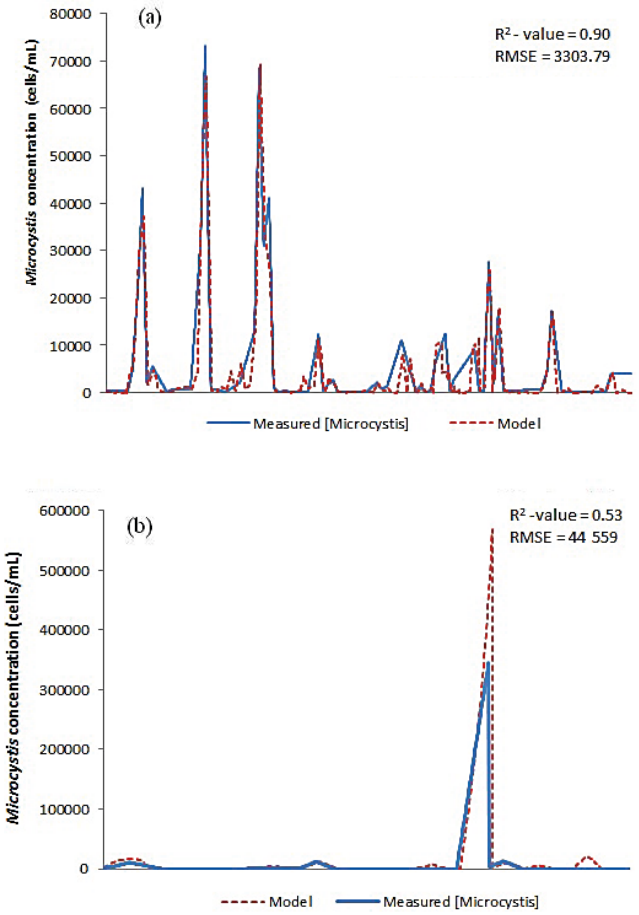

Figure 6

(a) Comparison between the measured Microcystis *sp. concentration and predicted 7 days in advance* Microcystis *sp. concentration in the Vaal Dam using 25% (boot-strapped) of the 10-year development dataset . (b) Comparison between the measured* Microcystis *sp. concentration and predicted 7 days in advance* Microcystis *sp. concentration using 3 years' 'unseen data' from the Vaal Dam*

a RMSE of 4 993.6 cells/mL. When the model was tested with 3 years of 'unseen data' (Fig. 10b), the model showed a RMSE of 18 493.9 cells/mL and a R^2-value of 0.25, which is not a good correlation. Neither event prediction nor the magnitude of the increased *Microcystis* sp. concentration demonstrated a significant correlation when the 3 years of 'unseen data' were tested on the model.

Comparison between models

The frequencies of the different input variables used in the models to predict *Microcystis* sp. concentrations are displayed in Table 3, ranging from the most frequently included variable to that least frequently included in the models.

The frequency distribution table (Table 3) indicates that the initial cyanobacteria concentration and the dissolved oxygen concentration were the variables most frequently used in the models to predict *Microcystis* sp. concentrations. Turbidity, and nutrients (NH_4^+, NO_3^- and PO_4^{3-}), as well as Fe^{2+} and chlorophyll-*a* concentration, were used in 50% of the models. The rest of the variables (water temperature, conductivity, pH, Si, Mn^{2+} and COD) were only used in 25% of the models.

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

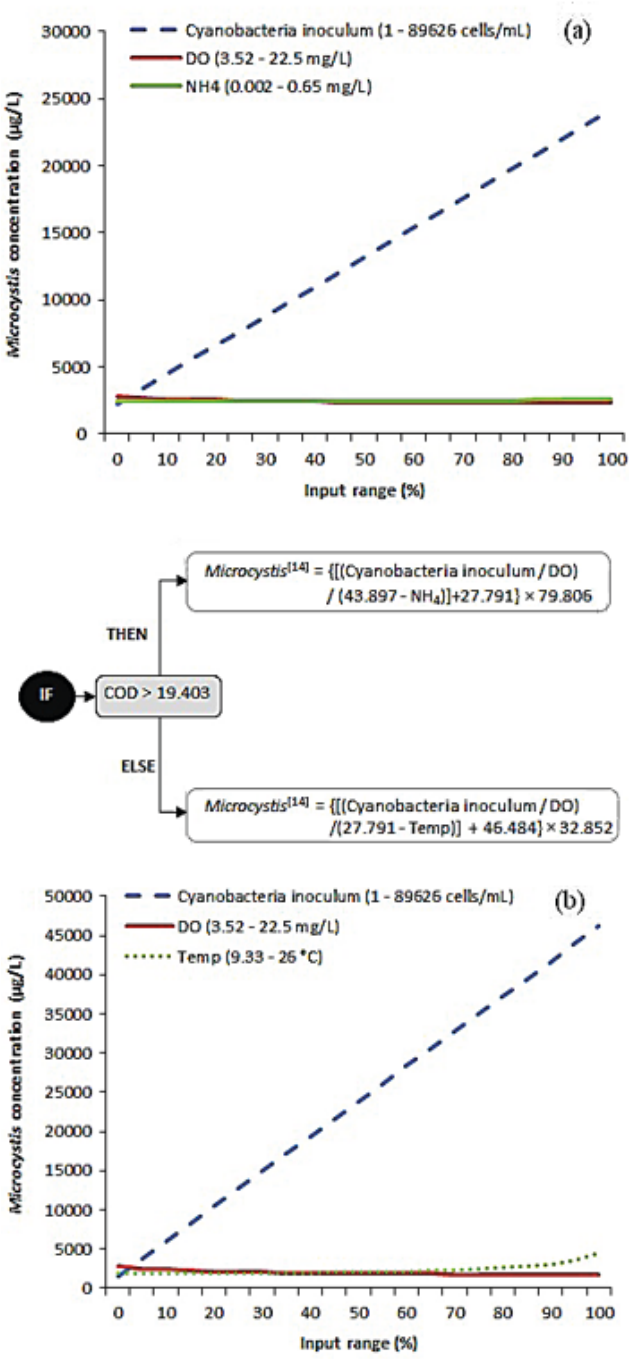

Figure 7

The 14-day prediction model and associated sensitivity analyses for Microcystis *sp. concentration in the Vaal Dam: (a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch*

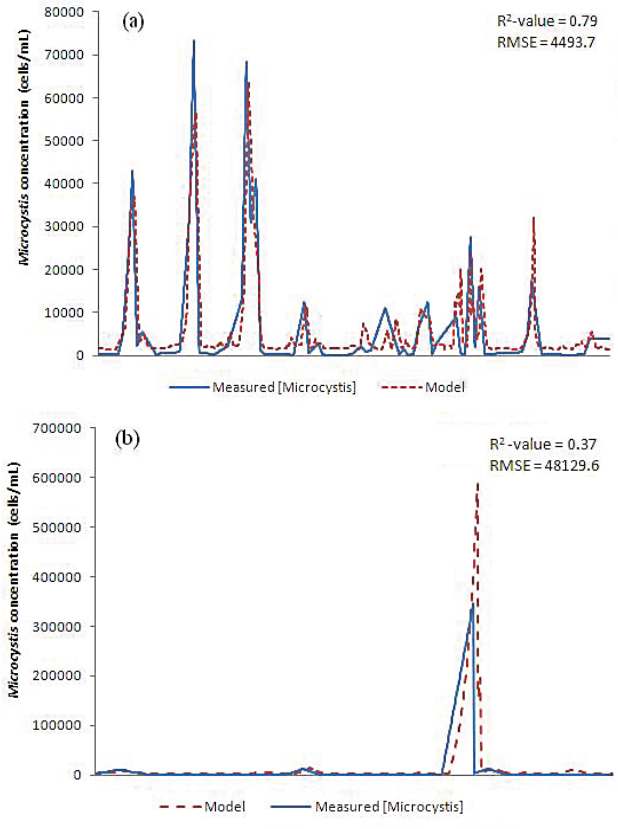

Figure 8

(a) Comparison between the measured Microcystis *sp. concentration and predicted 14 days in advance* Microcystis *sp. concentration in the Vaal Dam using 25% (boot-strapped) of the 10-year development dataset. (b) Comparison between the measured* Microcystis *sp. concentration and predicted 14 days in advance* Microcystis *sp. concentration using 3 years' 'unseen data' from the Vaal Dam*

Table 4 indicates the summary of the statistical and visual comparisons between the models tested with (i) the 25% of the original dataset chosen as testing dataset by bootstrapping and (ii) 3 years of 'unseen data' from follow-up years that were not used in the model development.

The square correlation coefficients (R^2) decrease with increasing time prediction and overall the square correlation coefficients when testing the models with 'unseen data' did not correlate as well when compared to tests with the 25%

TABLE 3
Frequency at which input variables were used in models to predict *Microcystis* sp. concentrations

Variable used in models	Number of occurrences in *Microcystis* sp. prediction modelss	Frequency (%)
Cyanobacteria inoculum (Cyano)	4	100
Dissolved oxygen (DO)	4	100
Turbidity (Turb)	2	50
NH_4^+	2	50
PO_4^{3-}	2	50
NO_3^-	2	50
Fe^{2+}	2	50
Chlorophyll-*a* (Chl*a*)	2	50
pH	2	50
Conductivity (Cond)	2	50
Water temperature (Temp)	1	25
Si	1	25
Mn^{2+}	1	25
COD	1	25

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

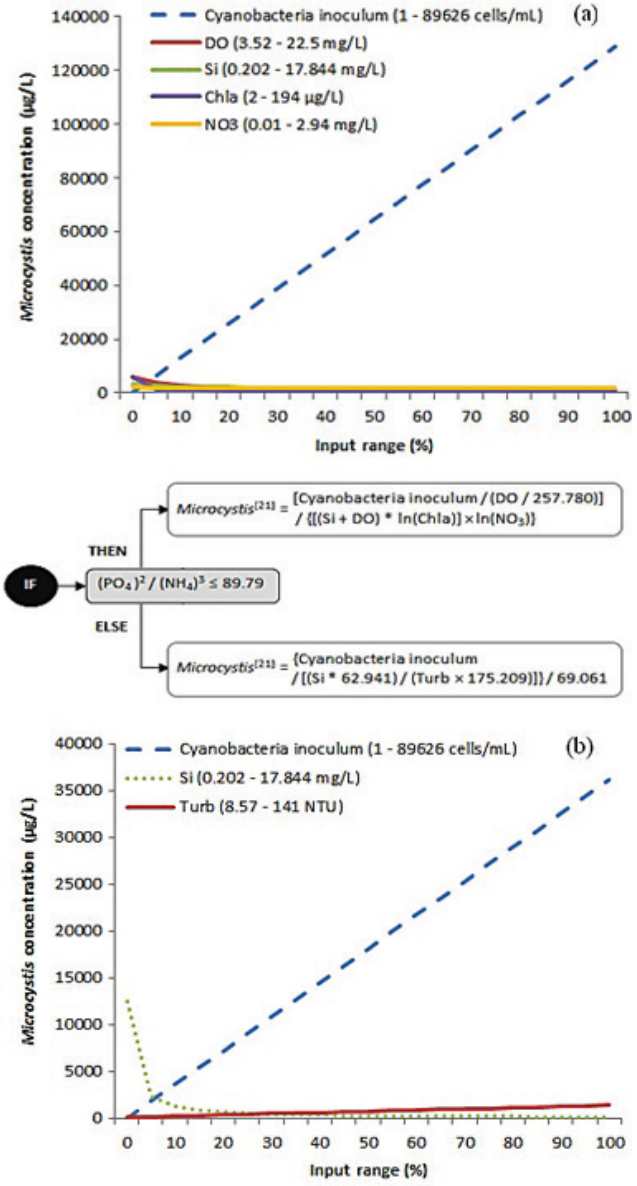

Figure 9

The 21-day prediction model and associated sensitivity analyses for Microcystis *sp. concentration in the Vaal Dam: (a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch*

randomly-chosen (bootstrapped) dataset. However, the visual inspection of the models showed good event prediction (with the exception of the model for 21 days in advance tested with the 3-year 'unseen data'). The RMSE of the models increased with increasing prediction time (with the exception of the model for 21 days in advance tested with the 3-year 'unseen data').

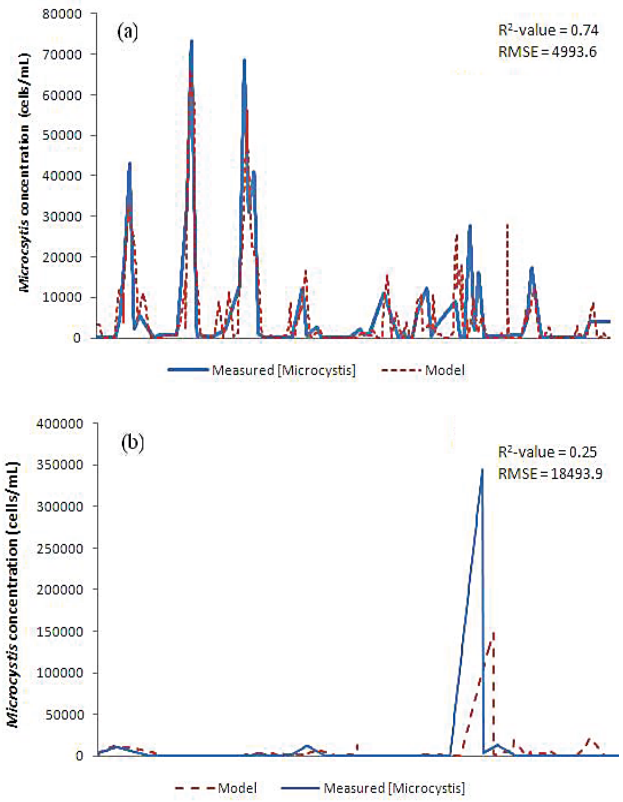

Figure 10

(a) Comparison between the measured Microcystis *sp. concentration and predicted 21 days in advance* Microcystis *sp. concentration in the Vaal Dam using 25% (boot-strapped) of the 10-year development dataset. (b) Comparison between the measured* Microcystis *sp. concentration and predicted 21 days in advance* Microcystis *sp. concentration using 3 years' 'unseen data' from the Vaal Dam*

TABLE 4
Summary statistics and comparisons of the *Microcystis* sp. prediction models tested with (a) 25% bootstrapped results of the original dataset and (b) 3 years of 'unseen data'

Models tested with...	*Microcystis* sp. prediction model	R^2	RMSE	Visual inspection: Satisfactory event prediction?
(a) 25% of the training dataset chosen by boot-strapping	Real time	0.95	2 462.2	✓
	7 days in advance	0.90	3 303.8	✓
	14 days in advance	0.79	4 493.7	✓
	21 days in advance	0.74	4 993.6	✓
(b) 3 years 'unseen' dataset not used in the model development	Real time	0.97	4 766.6	✓
	7 days in advance	0.5	44 559.3	✓
	14 days in advance	0.37	48 129.6	✓
	21 days in advance	0.25	18 493.9	✗

R^2 = square correlation coefficient; RMSE = root mean square error

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

DISCUSSION

Approximately 40% of the variation in the physical, chemical and biological data in the Vaal Dam from 2000–2009 could be explained by the first two axes of the two principle component analyses (Fig. 2) performed on the dataset used to develop the models. The PCA was performed in order to determine which of the variables would influence cyanobacteria and most probably be important as input variables for model development. The negative correlation of nutrients (NO_2^-, NO_3^-, NH_4^+ and PO_4^{3-}) to temperature (Temp), might be due to large cyanobacteria blooms in summer utilising and depleting the nutrients causing nutrient concentrations to be higher during winter times, when cyanobacteria or other phytoplankton concentrations are lower in the Vaal Dam. Chlorophyll-*a* and pH correlate positively indicating that increasing photosynthesis will inevitably increase the pH as CO_2 is removed from the aquatic environment. The negative correlation between chlorophyll-*a* and dissolved oxygen is most probably due to temperature since the highest levels of DO were observed during winter, although it may also be due to aerobic bacterial activity when large numbers of phytoplankton cells are decomposed during and after blooms. The fact that the arrow representing chlorophyll-*a* subtends a ±90° angle towards the arrow representing temperature, indicates that high chlorophyll-*a* concentrations are not limited to a specific season (either high or low temperatures) but can vary throughout the year. The chemical and biological results from the Vaal Dam indicate that it is in a mesotrophic state, but chlorophyll-*a* concentrations as high as 194 µg/L have been detected in the Vaal Dam during this period (Table 1).

The variables used by the HEA to predict the *Microcystis* sp. concentration in the Vaal Dam were: the initial cyanobacteria inoculum, dissolved oxygen (DO), turbidity (Turb), NH_4^+, PO_4^{3-}, NO_3^-, Fe^{2+}, chlorophyll-*a* (Chl*a*), water temperature (Temp), conductivity (Cond), Secchi disk depth (Secchi), pH, Si, Mn^{2+}, and chemical oxygen demand (in order of most to least frequently incorporated into the models – Table 3). The importance of these variables was also evident in the PCA which indicated that the nutrients, especially PO_4^{3-}, NO_2^-, NH_4^+ and NO_3^-, could explain the variation in the *Microcystis* sp. concentration and indirectly that of DO, due to photosynthesis and temperature. Initial cyanobacteria inoculum and dissolved oxygen were incorporated in all of the models predicting *Microcystis* sp. concentration, with the nutrient concentrations (NH_4^+, PO_4^{3-} or NO_3^-) used separately or in combination in 50% of the models. It should be noted, however, that at least one of the nutrients (either NH_4^+, PO_4^{3-} or NO_3^-) is incorporated in all *Microcystis* sp. models, except the model predicting the real-time concentration (Fig. 3). The reason for this may be that the real-time *Microcystis* sp. concentration cannot be influenced by a change in the nutrient concentration on that day, while the future occurrence of *Microcystis* sp. will inevitably be influenced by the nutrient concentration in the water.

The sensitivity analyses of the models predicting the *Microcystis* sp. concentration indicate that the greatest sensitivity is towards the initial cyanobacteria inoculum, Figs 3a and 4b, Fig. 5a and 5b, Figs 7a and 7b as well as Figs 9a and 9b. The initial cyanobacteria inoculum will have a large influence on the *Microcystis* sp. concentration, provided that the total cyanobacteria inoculum mostly comprises of *Microcystis* sp. cells (as is usually the case in the Vaal Dam). The 21 days in advance *Microcystis* sp. model also shows sensitivity towards the Si concentration, particularly during the first 10% of the silica input range. Silica might represent a secondary effect on *Microcystis* sp. concentrations, since Si is mostly utilised by diatoms in winter (Wetzel, 2001), when cyanobacteria like *Microcystis* sp. are not abundant in the Vaal Dam.

The models predicting the occurrence of *Microcystis* sp. (Figs 4a, 6a, 8a and 10a) show relatively good square correlation coefficients (R^2-values range from 0.95 at real-time prediction to 0.74 at 21-days prediction) when tested with the 25% boot-strapped testing dataset from 2000–2009. Although the testing with the 3-year 'unseen data' (Figs 4b, 6b, 8b and 10b) did not show square correlation coefficients as high as when tested with the 25% boot-strapped results (R^2values range from 0.97 at real-time prediction to 0.25 at 21-days prediction), it was still regarded as a significant correlation (with the exception of the model for 21 days in advance). Overall, the square correlation coefficients decrease and the RMSE increase with increasing prediction times (Table 4), displaying the increase in uncertainty over longer prediction periods. The visual inspection of the models was essential in determining the suitability of the model for further application (Bennet et al., 2013).

Currently the Zuikerbosch DWTW managers and production chemists are solely reliant on laboratory analyses of cyanobacteria cell counts, which (depending on sampling, distance and accessibility to laboratory, laboratory sample turn-around time and various other facts) may delay results for up to a week or even longer (Swanepoel et al., 2008). By the time the results become available, consumers might already have been exposed to cyanotoxins in their drinking water. With the prediction models, the managers and production specialists at DWTW can anticipate the occurrence of *Microcystis* sp. in the source water and start preparations before it happens. The models that would most probably have the greatest value when incorporated into the 'Cyanobacteria Incident Management Protocol' of the Zuikerbosch DWTW (Du Preez and Van Baalen, 2006) are the models predicting the *Microcystis* sp. 7 days in advance. A 7-day advance warning gives the plant sufficient time to prepare for incidences of high cyanobacteria and their related metabolites (e.g. microcystin) in the source water.

CONCLUSIONS

The most important variables for predicting of *Microcystis* sp. in the Vaal Dam were shown to be initial cyanobacteria inoculum and dissolved oxygen as they occur in 100% of the models. Initial cyanobacteria inoculum will determine how many cells are available for further bloom development. Dissolved oxygen is probably included due to the significant negative correlation with cyanobacteria which usually blooms during higher temperatures. Nutrients (either PO_4^{3-}, NH_4^+ or NO_3^-) are also important in predicting *Microcystis* sp. concentrations in advance (7–21 days).

The models that most probably would have the greatest value when applied at the Zuikerbosch DWTW are the models predicting *Microcystis* sp. 7 days in advance, since those were the most accurate. Seven days is sufficient time to prepare for treatment of source water containing cyanobacteria and their related metabolites.

It is evident that these predictive models will contribute significantly in anticipating and managing high *Microcystis* sp. concentrations in the source water supplying the Zuikerbosch DWTW. These models might also have application value to recreational water users, where event managers of large and small water-sport events can use such models to predict the *Microcystis* sp. concentration in the water whenever recreational events are planned.

http://dx.doi.org/10.4314/wsa.v42i2.08
Available on website http://www.wrc.org.za
ISSN 1816-7950 (On-line) = Water SA Vol. 42 No. 2 April 2016

ACKNOWLEDGEMENTS

This paper reflects the views of the authors who thank Rand Water, South Africa for supporting the investigation. A special word of thanks to Prof Hein du Preez, Head Biology, Analytical Services and Mr Danie du Plessis, Operational Manager (Potable Water).

REFERENCES

APHA (2013) *Standard Methods for the Examination of Water and Wastewater* (22nd edn). American Public Health Association, Washington D.C.

BARTRAM J, CORRALES A, DAVISON A, DEERE D, DRURY D, GORDON B, HOWARD G, RINEHOLD A and STEVENS M (2009) *Water Safety Plan Manual Step-By-Step Risk Management for Drinking Water Suppliers.* World Health Organisation, Geneva.

BENNET ND, CROKE FW, GUARISO G, GUILLAUME JHA, HAMILTON SH, JAKEMAN AJ, MARSILI-LIBELLI S, NEWHAM LTH, NORTON JP, PERRIN C, PIERCE SA, ROBSON B, SEPPELT R, VOINOV AA, FATH BD and ANDREASSIAN V

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. In: Recknagel F (ed.) *Ecological Informatics* (2nd edn.) Springer-Verlag, Berlin. 330–342. http://dx.doi.org/10.1007/3-540-28426-5_17

CHAN WS, RECKNAGEL F, CAO H and HO-DONG P (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** 2247–2255. http://dx.doi.org/10.1016/j.watres.2007.02.001

CONRADIE RC and BARNARD S (2012) The dynamics of toxic *Microcystis* strains and microcystin production in two hypertrofic South African reservoirs. *Harmful Algae* **20** 1–10. http://dx.doi.org/10.1016/j.hal.2012.03.006

DU PREEZ H and VAN BAALEN L (2006) *Generic Incident Management Framework for toxic blue-green algal blooms, for application by potable water suppliers.* WRC Report No.: TT263/06. Water Research Commission, Pretoria, South Africa.

DU PREEZ H, SWANEPOEL A, VAN BAALEN L and OLDEWAGE A (2007) Cyanobacterial Incident Management Frameworks (CIMFs) for application by drinking water suppliers. *Water SA* **33** (5) 643–652.

DWA (DEPARTMENT OF WATER AFFAIRS, SOUTH AFRICA) (2013a) Trophic status of South African Impoundments. National Eutrophication Monitoring Programme, Department of Water Affairs (DWA), Pretoria, South Africa. URL: http://www.dwa.gov.za/iwqs/eutrophication/NEMP/nempdams.htm (Accessed 5 July 2013).

DWA (DEPARTMENT OF WATER AFFAIRS, SOUTH AFRICA) (2013b) The Orange River basin in South Africa. URL: http://www.dwaf.gov.za/orange/Vaal/vaaldam.htm (Accessed 10 July 2013).

GROSAN C and ABRAHAM A (2007) Hybrid evolutionary algorithms: Methodologies, architectures, and reviews. *Stud. Comput. Intell.* **75** 1–17. http://dx.doi.org/10.1007/978-3-540-73297-6_1

HARDING WR and PAXTON BR (2001) Cyanobacteria in South Africa: A Review. WRC Report No.: TT 153/01. Water Research Commission, Pretoria. 165 pp.

KNAPPE DRU, BELK RC, BRILEY DS, GANDY SR, RASTOGI N, RIKE AH, GALSGOW H, HANNON E, FRAZIER WD, KOHL P and PUGSLEY S (2004) *Algae Detection and Removal Strategies for Drinking Water Treatment Plants.* Report No. 90971, AWWA Research Foundation, Denver, USA. 466 pp.

LHDP (LESOTHO HIGHLANDS DEVELOPMENT PROJECT) (2013) Overview of the Lesotho Highlands Water Project. URL: http://www.lhwp.org.ls/overview/default.htm (Accessed 10 July 2013).

LUND JWG, KIPLING C and LE CREN ED (1958) The inverted microscope method of estimating algal numbers and statistical basis of estimation by counting. *Hydrobiologia* **11** 143–170. http://dx.doi.org/10.1007/BF00007865

MERILUOTO J and CODD GA (Eds) (2005) *TOXIC: Cyanobacterial Monitoring and Cyanotoxin analysis.* Abo Akademi University, Turku, Finland. 149 pp.

RECKNAGEL F, ORR P and CAO H (2014) Inductive reasoning and forecasting of population dynamincs of *Cylindrospermopsis raciborskii* in three sub-tropical reservoirs by evolutionary computation. *Harmful Algae* **31** 26–34. http://dx.doi.org/10.1016/j.hal.2013.09.004

RECKNAGEL F, OSTROVSKY I, CAO H, ZOHARY T and ZHANG X (2013) Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computations and wavelets. *Ecol. Model.* **255** 70–86. http://dx.doi.org/10.1016/j.ecolmodel.2013.02.006

RECKNAGEL F, VAN GINKEL C, CAO H, CETIN L and ZHANG B (2008) Generic limnological models on the touchstone: Testing the lake simulation library SALMO-OO and the rule-based *Microcystis* agent for warmmonomictic hypertrophic lakes in South Africa. *Ecol. Model.* **215** 144–158. http://dx.doi.org/10.1016/j.ecolmodel.2008.02.035

SWANEPOEL A, DU PREEZ H, SCHOEMAN C, JANSE VAN VUUREN S and SUNDRAM A (2008) Condensed laboratory methods for the monitoring of phytoplankton, including cyanobacteria, in South African freshwaters. WRC Report No.: TT32308. Water Research Commission, Pretoria. 108 pp.

TALIB A, RECKNAGEL F and VAN DER MOLEN D (2007) Patternising phytoplankton dynamics of two shallow lakes in response to restoration measures by applying non-supervised artificial neural networks. *Environmentalist* **27** 195–205. http://dx.doi.org/10.1007/s10669-007-9023-x

TER BRAAK CJF (1988) CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (Version 4.5). Report LWA-88-02, Agricultural Mathematics Group, Wageningen, The Netherlands. 1–95.

VAN GINKEL CE (2008) Investigating the applicability of Ecological informatics modeling techniques for predicting harmful algal blooms in hypertrophic reservoirs of South Africa. PhD thesis, North West University, Potchefstroom, South Africa.

VAN GINKEL CE, DU PLESSIS S and BEZUIDENHOUT JJ (2010) Investigating the applicability of ecological informatics modelling techniques for predicting harmful algal blooms in hypertrophic reservoirs of South Africa. WRC Report No.: TT45110. Water Research Commission, Pretoria. 119 pp.

WALSBY AE (1971) The pressure relationships of gas vacuoles. *Proc. R. Soc. Lond. B* **178** 301–326. http://dx.doi.org/10.1098/rspb.1971.0067

WALSBY AE (1994) Gas vesicles. *Microbiol. Rev.* **58** 94–144.

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. Inform.* **3** 46–54. http://dx.doi.org/10.1016/j.ecoinf.2007.12.002

WETZEL RG (2001) *Limnology, Lake and River Ecosystems* (3rd edn). Academic Press. New York. 1006 pp.

ZOSCHKE K, ENGEL C, BÖRNICK H. and WORCH E (2011) Adsorption of geosmin and 2methylisoborneol onto powdered activated carbon at non-equilibrium conditions: Influence of NOM and process modelling. *Water Res.* **45** 4544–4550. http://dx.doi.org/10.1016/j.watres.2011.06.006