

Early warning system for the prediction of algal-related impacts on drinking water purification

A Swanepoel
10086978

Thesis submitted for the degree *Philosophiae Doctor* in Botany
at the Potchefstroom Campus of the North-West University

Promoter: Prof S Barnard
Co-promoter: Prof F Recknagel

May 2015

It all starts here™



TABLE OF CONTENT

i	Summary.....	i
ii	Opsomming	iv
iii	Lists of Figures	vii
iv	List of Tables	xv
v	List of Acronyms and Abbreviations.....	xvi
vi	Acknowledgements	xviii
1.	Chapter 1: Introduction.....	1
2.	Chapter 2: Literature Survey : The use of evolutionary algorithms for the development of forecasting models in water ecology	5
2.1	Introduction.....	5
2.2	Important input variables for algae (including cyanobacteria) related to forecasting models	6
2.3	Evaluation of prediction or forecasting models.....	12
2.4	Optimisation strategies for poorly performing models	14
2.4.1	Synergistic approach to certain input variables	14
2.4.2	Rule-based models.....	14
2.4.3	Hill climbing for parameter optimisation	15
2.4.4	Differential evolution for parameter optimisation.....	15
2.4.5	Support vector regression	16
2.4.6	Multi adaptative regression splines (MARS).....	16
2.4.7	Incorporating time lags into the input variables	16
2.5	Conclusions.....	17
3.	Chapter 3: Materials and Methods.....	19
3.1	Study area, sampling and analyses.....	19
3.2	Hybrid Evolutionary Algorithms (HEAs).....	23
3.3	Statistical analyses.....	26
3.4	Sensitivity analysis	27
3.5	Application of models.....	28
3.6	Development and validation of a model to determine chlorophyll-665 (total photosynthetic pigments concentration) from chlorophyll- <i>a</i> and <i>vice versa</i>	30
3.6.1	Model development	31

3.6.2	Model testing using “previously seen” data	32
3.6.3	Model testing using “unseen” data	34
3.6.4	Conclusions	36
4.	Chapter 4: An overview of the algae and associated nutrients in the Vaal Dam from 2000 – 2012	37
4.1	Introduction	37
4.2	Results and Discussion	38
4.3	Conclusions	51
5.	Chapter 5: Using Hybrid Evolutionary Algorithms for the development of compound rule-based forecasting models for cyanobacteria from all available historical laboratory generated data	53
5.1	Introduction	53
5.2	Models for the prediction of <i>Anabaena</i> sp. concentration.	53
5.2.1	<i>Anabaena</i> ^[0] prediction (in real-time)	54
5.2.2	<i>Anabaena</i> ^[7] prediction (7 days in advance)	55
5.2.3	<i>Anabaena</i> ^[14] prediction (14 days in advance)	58
5.2.4	<i>Anabaena</i> ^[21] prediction (21 days in advance)	59
5.3	Models for the prediction of <i>Microcystis</i> sp. concentration	61
5.3.1	<i>Microcystis</i> ^[0] prediction (in real-time)	62
5.3.2	<i>Microcystis</i> ^[7] prediction (7 days in advance)	63
5.3.3	<i>Microcystis</i> ^[14] prediction (14 days in advance)	65
5.3.4	<i>Microcystis</i> ^[21] prediction (21 days in advance)	67
5.4	Models for the prediction of microcystin concentration	69
5.4.1	Microcystin ^[0] prediction (in real-time)	70
5.4.2	Microcystin ^[7] prediction (7 days in advance)	72
5.4.3	Microcystin ^[14] prediction (14 days in advance)	74
5.4.4	Microcystin ^[21] prediction (21 days in advance)	77
5.5	Models for the prediction of geosmin concentration	79
5.5.1	Geosmin ^[0] prediction (in real-time)	79
5.5.2	Geosmin ^[7] prediction (7 days in advance)	81
5.6	Discussion and conclusions	84
6.	Chapter 6: Using Hybrid Evolutionary Algorithms for the development of limited input rule-based predictive models for cyanobacteria	90
6.1	Introduction	90

6.2	Models for the prediction of <i>Anabaena</i> sp. concentration	91
6.2.1	<i>Anabaena</i> ^[0] prediction (in real-time)	91
6.2.2	<i>Anabaena</i> ^[7] prediction (7 days in advance)	94
6.3	Models for the prediction of <i>Microcystis</i> sp. concentration.....	96
6.3.1	<i>Microcystis</i> ^[0] prediction (in real-time).....	96
6.3.2	<i>Microcystis</i> ^[7] prediction (7 days in advance).....	99
6.4	Models for the prediction of microcystin concentration.....	101
6.4.1	Microcystin ^[0] prediction (in real-time).....	101
6.4.2	Microcystin ^[7] prediction (7 days in advance).....	104
6.5	Discussion and conclusions	106
7.	Chapter 7: Application of forecasting models in the multi-barrier approach of water safety plans as part of the Blue Drop Certification Program in South Africa .	110
7.1	Introduction.....	110
7.2	SANS 241-1 (2011) – the latest South African Drinking Water Standard relating to cyanobacteria and cyanotoxins.....	112
7.3	The Blue Drop Certification Program for ensuring safe drinking water to South African consumers	114
7.4	Using a water safety plan (WSP) for risk assessment and risk reduction in DWTW	115
7.5	Possibility to incorporate prediction models into the WSP	122
7.6	Case study : Rand Water	126
7.7	Conclusions.....	131
8.	Chapter 8: General conclusions and recommendations.....	132
8.1	Introduction.....	132
8.2	General conclusions	132
8.3	Recommendations.....	134
9.	Chapter 9: References	136
	ADDENDUM A	146
	ADDENDUM B	Refer to attached CD

SUMMARY

Algae and cyanobacteria occur naturally in source waters and are known to cause extensive problems in the drinking water treatment industry. Cyanobacteria (especially *Anabaena* sp. and *Microcystis* sp.) are responsible for many water treatment problems in drinking water treatment works (DWTW) all over the world because of their ability to produce organic compounds like cyanotoxins (e.g. microcystin) and taste and odour compounds (e.g. geosmin) that can have an adverse effect on consumer health and consumer confidence in tap water. 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 of physical, chemical and biological water quality analyses, for their management decisions. But results of water quality analyses can be delayed from 3 hours to a few 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 use of on-line (*in situ*) instruments that can supply real-time results by the click of a button has become very popular in the past few years. On-line instruments were developed for analyses like pH, conductivity, nitrate, chlorophyll-*a* and cyanobacteria concentrations. Although, this real-time (on-line) data has given drinking water treatment managers a better opportunity to make sound management decisions around drinking water treatment options based on the latest possible results, it may still be “too little, too late” once a sudden cyanobacterial bloom of especially *Anabaena* sp. or *Microcystis* sp. enters the plant. Therefore the benefit for drinking water treatment management, of changing the focus from real-time results to future predictions of water quality has become apparent.

The aims of this study were 1) to review the environmental variables associated with cyanobacterial blooms in the Vaal Dam, as to get background on the input variables that can be used in cyanobacterial-related forecasting models; 2) to apply rule-based Hybrid Evolutionary Algorithms (HEAs) to develop models using a) all applicable laboratory-generated data and b) on-line measureable data only, as input variables in prediction models for harmful algal blooms in the Vaal Dam; 3) to test these models with data that was not used to develop the models (so-called “unseen data”), including on-line (*in situ*) generated data; and 4) to incorporate selected models into two cyanobacterial incident management protocols which link to the Water Safety Plan (WSP) of a large DWTW (case study : Rand Water).

During the current study physical, chemical and biological water quality data from 2000 to 2009, measured in the Vaal Dam and the 20km long canal supplying the Zuikerbosch DWTW of Rand Water, has been used to develop models for the prediction of *Anabaena* sp., *Microcystis* sp., the cyanotoxin microcystin and the taste and odour compound geosmin for different prediction or forecasting times in the source water. For the development and first stage of testing the models, 75% of the dataset was used to train the models and the remaining 25% of the dataset was used to test the models. Boot-strapping was used to determine which 75% of the dataset was to be used as the training dataset and which 25% as the testing dataset. Models were also tested with 2 to 3 years of so called “unseen data” (Vaal Dam 2010 – 2012) i.e. data not used at any stage during the model development. Fifty different models were developed for each set of “x input variables = 1 output variable” chosen beforehand. From the 50 models, the best model between the measured data and the predicted data was chosen. Sensitivity analyses were also performed on all input variables to determine the variables that have the largest impact on the result of the output.

This study have shown that hybrid evolutionary algorithms can successfully be used to develop relatively accurate forecasting models, which can predict cyanobacterial cell concentrations (particularly *Anabaena* sp. and *Microcystis* sp.), as well as the cyanotoxin microcystin concentration in the Vaal Dam, for up to 21 days in advance (depending on the output variable and the model applied). The forecasting models that performed the best were those forecasting 7 days in advance ($R^2 = 0.86, 0.91$ and 0.75 for *Anabaena*^[7], *Microcystis*^[7] and microcystin^[7] respectively). Although no optimisation strategies were performed, the models developed during this study were generally more accurate than most models developed by other authors utilising the same concepts and even models optimised by hill climbing and/or differential evolution. It is speculated that including “initial cyanobacteria inoculum” as input variable (which is unique to this study), is most probably the reason for the better performing models. The results show that models developed from on-line (*in situ*) measureable data only, are almost as good as the models developed by using all possible input variables. The reason is most probably because “initial cyanobacteria inoculum” – the variable towards which the output result showed the greatest sensitivity – is included in these models. Generally models predicting *Microcystis* sp. in the Vaal Dam were more accurate than models predicting *Anabaena* sp. concentrations and models with a shorter prediction time (e.g. 7 days in advance) were statistically more accurate than models with longer prediction times (e.g. 14 or 21 days in advance).

The multi-barrier approach in risk reduction, as promoted by the concept of water safety plans under the banner of the Blue Drop Certification Program, lends itself to the application of future predictions of water quality variables. In this study, prediction models of *Anabaena* sp., *Microcystis* sp. and microcystin concentrations 7 days in advance from the Vaal Dam, as well as geosmin concentration 7 days in advance from the canal were incorporated into the proposed incident management protocols. This was managed by adding an additional “Prediction Monitoring Level” to Rand Waters’ microcystin and taste and odour incident management protocols, to also include future predictions of cyanobacteria (*Anabaena* sp. and *Microcystis* sp.), microcystin and geosmin. The novelty of this study was the incorporation of future predictions into the water safety plan of a DWTW which has never been done before. This adds another barrier in the potential exposure of drinking water consumers to harmful and aesthetically unacceptable organic compounds produced by cyanobacteria.

Keywords:

Hybrid evolutionary algorithms, forecasting models, cyanobacteria, *Anabaena*, *Microcystis*, microcystin, geosmin, water safety plan, cyanobacterial incident management protocol, drinking water purification.

OPSOMMING

Alge en sianobakterië kom natuurlik in die bronwater van watersuiweringsaanlegte voor en is alombekend vir die groot probleme wat dit in drinkwatervoorsiening kan veroorsaak. Sianobakterië, veral *Anabaena* sp. en *Microcystis* sp., is verantwoordelik vir baie probleme in drinkwatersuiweringsaanlegte regoor die wêreld, a.g.v. hulle vermoë om organiese verbindings soos sianotoksiene (bv. mikrosistien) asook reuk- en smaakverbindinge (soos geosmien) te produseer. Hierdie organiese verbindings kan 'n negatiewe impak op die verbruiker se gesondheid hê, asook op sy vertroue in die kwaliteit en geskiktheid van kraanwater. As gevolg hiervan het die monitoring van sianobakterië in die bronwater van watersuiweringsaanlegte 'n uiters belangrike aspek van drinkwatervoorsiening geword.

Bestuurders van drinkwatersuiweringsaanlegte is afhanklik van die waterkwaliteitsresultate van fisiese, chemiese en biologiese analises om hulle in staat te stel om gesonde en tydige bestuursbesluite t.o.v. die suiweringsproses te neem. 'n Hele aantal faktore kan egter die tydige beskikbaarheid van hierdie waterkwaliteitsresultate beïnvloed. Hierdie faktore sluit in: water-eksemplaar versameling, afstand vanaf en toegang tot laboratoriumfasiliteite, die omset waarteen die laboratorium water eksemplare analiseer en data verwerk, die spesifieke metodes wat tydens analises gebruik word, ens. Resultate wat te laat aan bestuurders deurgegee is om sinvolle bestuursbesluite te kan neem, het gedurende die afgelope paar jaar aanleiding gegee dat aanlyn (*in situ*) instrumente, wat met die druk van 'n knoppie onmiddellike resultate kan voorsien, al hoe gewilder geword het. Aanlyn-instrumente is ontwikkel vir bepaling van veranderlikes soos pH, geleiding, nitrate, chlorofil-*a* en sianobakteriese selkonsentrasies. Alhoewel dit bestuurders in staat stel om tydige bestuursbesluite te neem gegrond op die mees onlangse resultate, mag dit steeds te laat wees vir 'n aanleg om te reageer wanneer die resultate van die bronwater 'n sianobakteriese opbloeitoon. Om dus die fokus te verskuif van huidige resultate (soos verskaf deur aanlyn-instrumente) na toekomstige voorspellings van waterkwaliteit, hou duidelik groot voordele vir die drinkwaterindustrie in.

Die doel van hierdie studie was 1) om 'n oorsig te doen oor die omgewingsveranderlikes in die Vaaldam wat betrekking het op sianobakteriese opbloei om sodoende agtergrond te verskaf oor die inset-veranderlikes wat gebruik kan word as basis vir die ontwikkeling van voorspellingsmodelle; 2) om hibried evolusionêre algoritmes te gebruik om voorspellingsmodelle op te stel deur gebruik te maak van a) alle toepaslike laboratorium-gegenereerde data en b) aanlyn-meetbare data alleen as inset-veranderlikes in die

voorspellingsmodelle vir skadelike sianobakteriese opbloeie in die Vaaldam; 3) om hierdie modelle te toets met sg. “onbekende” data wat nie voorheen gebruik is om die modelle te ontwikkel nie (wat aanlyn-gegenereerde data insluit); en 4) om sekere van hierdie modelle in twee sianobakteriese-insident-bestuurs-protokolle te inkorporeer, wat deel uitmaak van die waterveiligheidsplan van ‘n groot drinkwatersuiweringsaanleg (gevallestudie: Rand Water).

Tydens hierdie studie is fisiese, chemiese en biologiese waterkwaliteitsresultate wat vanaf 2000 tot 2009 in die Vaaldam en kanaal (wat water vanaf die Vaaldam na die Zuikerbosch drinkwatersuiweringsaanleg vervoer) geneem is, gebruik om voorspellingsmodelle mee te ontwikkel. Dit sluit modelle vir die voorspelling van die sianobakterië *Anabaena* sp., *Microcystis* sp., die sianotoksien mikrosistien, asook die reuk- en smaak-veroorsakende verbinding, geosmien in. Vir die ontwikkeling van die modelle en die eerste toetsfase, is 75% van die datastel gebruik om die modelle mee te ontwikkel en die oorblywende 25% is gebruik om die modelle mee te toets. Die tegniek, skoenlus-samevoeging (in Engels bekend as “bootstrapping”) is gebruik om te bepaal watter gedeelte van die datastel gebruik sou word vir ontwikkeling van die model en watter gedeelte vir die toets daarvan. Modelle is ook getoets met 2 tot 3 jaar onbekende data. Dit is data van die Vaaldam wat strek van 2010 – 2012, m.a.w. “nuwe” data wat in geen stadium van model-ontwikkeling gebruik is nie. Vyftig verskillende modelle is uiteindelik ontwikkel vir elke gekose datastel en die beste model van die 50 modelle is gekies om verder in die studie te gebruik. Sensitiwiteitsanalises is ook uitgevoer om te bepaal watter van die inset-veranderlikes die grootste invloed op die resultaat of uitset-veranderlikes het.

Resultate uit hierdie studie het aangetoon dat hibried evolusionêre algoritmes suksesvol gebruik kan word om relatief akkurate voorspellingsmodelle op te stel wat sianobakteriese selkonsentrasie (spesifiek *Anabaena* sp. en *Microcystis* sp.) sowel as die sianotoksien, mikrosistien, in die Vaaldam vir tot 21 dae vooruit kan voorspel. Die voorspellingsmodelle wat die beste presteer het, is modelle wat 7 dae vooruit voorspel ($R^2 = 0.86, 0.91$ en 0.75 vir onderskeidelik *Anabaena*^[7], *Microcystis*^[7] en mikrosistien^[7]). Alhoewel geen optimiseringstegnieke in hierdie studie toegepas is nie, was modelle wat opgestel is selfs meer akkuraat as modelle wat deur differensiële evolusie of d.m.v. trapgewyse voortbouingsmetodes (in Engels bekend as “hill climbing”) geoptimeer is. Daar word gespekuleer dat die insluiting van “aanvanklike sianobakteriese inokulum” as inset-veranderlike (wat uniek aan hierdie studie is), heel moontlik die rede vir die beter prestasie as ander modelle in die literatuur is. Die resultate toon dat modelle wat ontwikkel is van die beperkte aanlyn-meetbare inset-veranderlikes alleenlik, feitlik net so akkuraat is as modelle wat alle moontlike inset-

veranderlikes insluit. Die rede hiervoor is waarskynlik omdat “aanvanklike sianobakteriese konsentrasie” – die veranderlike ten op sigte waarvan die uitset-veranderlike die grootste sensitiwiteit getoon het – ook in hierdie modelle ingesluit word. Oor die algemeen is die modelle wat *Microcystis* sp. selkonsentrasies in die Vaaldam voorspel meer akkuraat as die modelle wat *Anabaena* sp. selkonsentrasies voorspel en modelle met ’n korter voorspellingstydperk (bv. 7 dae) was meer akkuraat as modelle met ’n langer voorspellingstydperk (bv. 14 of 21 dae).

Waterveiligheidsplanne, as deel van die Suid-Afrikaanse Bloudruppelsertifiseringsprogram, leen dit uitstekend tot die toepassing van hierdie waterkwaliteitsvoorspellingsmodelle. In hierdie studie is modelle wat *Anabaena* sp., *Microcystis* sp. en mikrosistien in die Vaaldam 7 dae vooruit voorspel, sowel as modelle wat geosmien in die kanaal 7 dae vooruit voorspel, in twee sianobakteriese-insident-bestuurs-protokolle van Rand Water geïnkorporeer. Dit is bereik deur ’n addisionele “voorspellingsvlak” in Rand Water se mikrosistien-, sowel as reuk-en-smaak-insident-bestuurs-protokolle te voeg om sodoende toekomstige voorspellings van sianobakterië (*Anabaena* sp. en *Microcystis* sp.), mikrosistien en geosmien in te sluit. Hierdie is ’n unieke toepassing omdat geen drinkwatersuiweringsaanleg nog vantevore toekomstige voorspellings in enige waterveiligheidsplan ingesluit het nie. Hierdie studie kan daartoe bydra om drinkwater steeds veiliger en meer esteties aanvaarbaar vir verbruikers te maak.

Sleutelwoorde:

Hibried evolusionêre algoritmes, voorspellingsmodelle, sianobakterië, *Anabaena*, *Microcystis*, mikrosistien, geosmien, waterveiligheidsplan, sianobakteriese-insident-bestuurs-protokolle, drinkwatersuiwering.

LIST OF FIGURES

Figure 3.1	Sampling points in the Vaal Dam (■) and canal (■) supplying untreated water to the Zuikerbosch Drinking Water Treatment Works (DWTW).....	20
Figure 3.2	Twenty kilometre long canal (with a hydraulic retention time of 4 – 6 hours), supplying source water to the Zuikerbosch DWTW.....	20
Figure 3.3	Google Earth picture of stations 3 and 4 at the Zuikerbosch DWTW (Google Earth, 2013).....	21
Figure 3.4	On-line (<i>in situ</i>) YSI 6600 V2 Multi-parameter water quality sonde in the canal measuring water temperature (°C), turbidity (NTU), conductivity (mS/m), dissolved oxygen (mg/L and % saturation), pH, chlorophyll- <i>a</i> (µg/L) and cyanobacterial cell concentration (cells/mL).....	22
Figure 3.5	Flow chart of the hybrid evolutionary algorithms (HEAs) performing rule structure optimisation by Genetic Programming and parameter optimisation by Genetic Algorithms – adapted from Cao et al., (2006).....	24
Figure 3.6	Conceptual diagram for using the hybrid evolutionary algorithm (HEA) for the discovery of predictive rule sets in water quality time-series data – adapted from Cao et al., 2006.....	25
Figure 3.7	Cyanobacterial incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Du Preez and Van Baalen, 2006).....	29
Figure 3.8	Regression analysis of chlorophyll- <i>a</i> vs. chlorophyll-665 at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012.....	31
Figure 3.9	Measured chlorophyll-665 vs. predicted chlorophyll-665 at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012 (“previously seen” dataset)..	32
Figure 3.10	Validation of model (Chlorophyll-665 = 1.24 * Chlorophyll- <i>a</i>) at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012 (“previously seen” dataset)..	33
Figure 3.11	Measured chlorophyll-665 vs. predicted chlorophyll-665 at M-B11_VG and M-Canal from 12 February 2012 to 11 November 2013 (“unseen” dataset).....	34
Figure 3.12	Validation of model (Chlorophyll-665 = 1.24 * Chlorophyll- <i>a</i>) at M-B11_VG and M-Canal from 12 February 2012 to 11 November 2013 (“unseen” dataset).....	35
Figure 4.1	Cyanobacteria scums observed in the Vaal Dam at Oranjeville on 20 September 2005.....	37
Figure 4.2	Percentage composition of the algae groups occurring in the Vaal Dam from 2000 - 2013.....	41

Figure 4.3	Box plot of chlorophyll- <i>a</i> (µg/L) in the Vaal Dam for the period 2000 – 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring)	42
Figure 4.4	Box plot of total cyanobacterial cell concentration (cells/mL) in the Vaal Dam for the period 2000 - 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring)	43
Figure 4.5	Box plot of microcystin (µg/L) in the Vaal Dam for the period 2000 - 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring)	44
Figure 4.6	Scatter plot with associated trend lines for <i>Microcystis</i> (Mic_s) and <i>Anabaena</i> (Ana) in cells/mL, in the Vaal Dam from 2000 – 2013	45
Figure 4.7	Scatter plot with associated trend lines for <i>Microcystis</i> sp. (Mic_s) in cells/mL and water temperature (Temp) in °C, in the Vaal Dam from 2000 - 2013	46
Figure 4.8	Canonical correspondence analysis (CCA) of the environmental variables and natural log of the phytoplankton data in the Vaal Dam for the period 2000 – 2013..	50
Figure 5.1	Forecasting model and associated sensitivity analyses of real-time <i>Anabaena</i> ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	54
Figure 5.2	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted real-time <i>Anabaena</i> ^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset. b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted real-time <i>Anabaena</i> ^[0] concentration using 3 years' unseen data from the Vaal Dam	55
Figure 5.3	Forecasting model and associated sensitivity analyses of <i>Anabaena</i> ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	56
Figure 5.4	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[7] concentration (7 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset. b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam	57
Figure 5.5	Forecasting model and associated sensitivity analyses of <i>Anabaena</i> ^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	58

Figure 5.6	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset	
	b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam	59
Figure 5.7	Forecasting model and associated sensitivity analyses of <i>Anabaena</i> ^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	60
Figure 5.8	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[21] concentration (21 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam	61
Figure 5.9	Forecasting model and associated sensitivity analyses of real-time <i>Microcystis</i> ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	62
Figure 5.10	a Comparison between the measured <i>Microcystis</i> sp. concentration and predicted real-time <i>Microcystis</i> ^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Microcystis</i> sp. concentration and predicted real-time <i>Microcystis</i> ^[0] concentration using 3 years' unseen data from the Vaal Dam	63
Figure 5.11	Forecasting model and associated sensitivity analyses for <i>Microcystis</i> ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	64
Figure 5.12	a Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[7] concentration (7 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam	65
Figure 5.13	Forecasting model and associated sensitivity analyses for <i>Microcystis</i> ^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity	

	analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	66
Figure 5.14 a	Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam	67
Figure 5.15	Forecasting model and associated sensitivity analyses for <i>Microcystis</i> ^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	68
Figure 5.16 a	Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[21] (21 days in advance) concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam	69
Figure 5.17	Forecasting model and associated sensitivity analyses for real-time microcystin ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	70
Figure 5.18 a	Comparison between the measured microcystin concentration and predicted real-time microcystin concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured microcystin concentration and predicted real-time microcystin concentration using 3 years' unseen data from the Vaal Dam	71
Figure 5.19	Forecasting model and associated sensitivity analyses for microcystin ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	73
Figure 5.20 a	Comparison between the measured microcystin concentration and predicted microcystin ^[7] concentration (7 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured microcystin concentration and predicted microcystin ^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam	74

Figure 5.21	Forecasting model and associated sensitivity analyses for microcystin ^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	75
Figure 5.22 a	Comparison between the measured microcystin concentration and predicted microcystin ^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured microcystin concentration and predicted microcystin ^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam	76
Figure 5.23	Forecasting model and associated sensitivity analyses for microcystin ^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	77
Figure 5.24 a	Comparison between the measured microcystin concentration and predicted microcystin ^[21] concentration (21 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured microcystin concentration and predicted microcystin ^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam	78
Figure 5.25	Forecasting model and associated sensitivity analyses of real-time geosmin ^[0] concentration in the canal ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	80
Figure 5.26 a	Comparison between the measured geosmin concentration and predicted real-time geosmin ^[0] concentration in the canal using 25% (boot-strapped) of the 10 years development dataset.	
b	Comparison between the measured geosmin concentration and predicted real-time geosmin ^[0] concentration using 3 years' unseen data from the canal.....	81
Figure 5.27	Forecasting model and associated sensitivity analyses of geosmin ^[7] concentration (7 days in advance) in the canal ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	82
Figure 5.28 a	Comparison between the measured geosmin concentration and predicted geosmin ^[7] concentration (7 days in advance) in the canal using 25% (boot-strapped) of the 10 years development dataset.	

	b Comparison between the measured geosmin concentration and predicted geosmin ^[7] concentration (7 days in advance) using 3 years' unseen data from the canal	83
Figure 6.1	Forecasting model and associated sensitivity analyses of real-time <i>Anabaena</i> ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	92
Figure 6.2	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted real-time <i>Anabaena</i> ^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted real-time <i>Anabaena</i> ^[0] concentration using 3 years' unseen data from the Vaal Dam.	
	c Comparison between the measured <i>Anabaena</i> sp. concentration and predicted real-time <i>Anabaena</i> ^[0] concentration using 2 years' <i>in situ</i> generated data from the canal	93
Figure 6.3	Forecasting model and associated sensitivity analyses for <i>Anabaena</i> ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).	94
Figure 6.4	a Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[7] concentration using 3 years' unseen data from the Vaal Dam.	
	c Comparison between the measured <i>Anabaena</i> sp. concentration and predicted <i>Anabaena</i> ^[7] concentration using 2 years' <i>in situ</i> generated data from the canal.....	95
Figure 6.5	Forecasting model and associated sensitivity analyses of real-time <i>Microcystis</i> ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	97
Figure 6.6	a Comparison between the measured <i>Microcystis</i> sp. concentration and predicted real-time <i>Microcystis</i> ^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Microcystis</i> sp. concentration and predicted real-time <i>Microcystis</i> ^[0] concentration using 3 years' unseen data from the Vaal Dam.	

	c Comparison between the measured <i>Microcystis</i> sp. concentration and predicted real-time <i>Microcystis</i> ^[0] concentration using 2 years' <i>in situ</i> generated data from the canal	98
Figure 6.7	Forecasting model and associated sensitivity analyses for <i>Microcystis</i> ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	99
Figure 6.8	a Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[7] concentration using 3 years' unseen data from the Vaal Dam.	
	c Comparison between the measured <i>Microcystis</i> sp. concentration and predicted <i>Microcystis</i> ^[7] concentration using 2 years' <i>in situ</i> generated data from the canal	100
Figure 6.9	Forecasting model and associated sensitivity analyses for real-time microcystin ^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	102
Figure 6.10	a Comparison between the measured microcystin concentration and predicted real-time microcystin ^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured microcystin concentration and predicted real-time microcystin ^[0] concentration using 3 years' unseen data from the Vaal Dam.	
	c Comparison between the measured microcystin concentration and predicted real-time microcystin ^[0] concentration using 2 years' <i>in situ</i> generated data from the canal	103
Figure 6.11	Forecasting model and associated sensitivity analyses of microcystin ^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch)	104
Figure 6.12	a Comparison between the measured microcystin concentration and predicted microcystin ^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.	
	b Comparison between the measured microcystin concentration and predicted microcystin ^[7] concentration using 3 years' unseen data from the Vaal Dam.	
	c Comparison between the measured microcystin concentration and predicted microcystin ^[7] concentration using 2 years' <i>in situ</i> generated data from the canal	105

Figure 7.1	Schematic representation of the relationship between the South African Blue Drop Certification Program, SANS 241, Water Safety Plans as well as Incident Management Protocols addressed in this study.....	111
Figure 7.2	A typical cyanobacterial incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Du Preez and Van Baalen, 2006).....	117
Figure 7.3	A typical taste and odour incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Swanepoel and Du Preez, 2008).....	121
Figure 7.4	Excel spreadsheet supplied to the DWTW for predicting <i>Anabaena</i> sp. concentrations in real-time, 7 days, 14 days and 21 days in advance. (The columns in grey will not be visible to the end-user	125
Figure 7.5	Proposed microcystin protocol to include future predictions of cyanobacteria (<i>Anabaena</i> sp. + <i>Microcystis</i> sp.) as well as microcystin concentration	128
Figure 7.6	Proposed taste and odour protocol to include future predictions of cyanobacteria (<i>Anabaena</i> sp. + <i>Microcystis</i> sp.) as well as microcystin concentration	130

LIST OF TABLES

Table 2.1	Input variables incorporated into models predicting chlorophyll- <i>a</i>	7
Table 2.2	Input variables incorporated into the best models predicting algae (including cyanobacteria) genera and groups	8
Table 2.3	Input variables incorporated into models predicting organic compounds associated with cyanobacteria	11
Table 3.1	Results from the paired students' t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model: (Chlorophyll-665 = 1.24 * Chlorophyll- <i>a</i>), with the "previously seen" dataset	32
Table 3.2	Results from the paired students' t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model: (Chlorophyll-665 = 1.24 * Chlorophyll- <i>a</i>) with the "unseen" dataset.....	34
Table 4.1	The descriptive statistical values of the measured variables in the Vaal Dam (2000 - 2012).....	38
Table 4.2	Factor-variable correlations (factor loadings) of Vaal Dam variables (2000 - 2012)	47
Table 4.3	Eigen values for the CCA done on all environmental variables in the Vaal Dam from 2000 – 2013 in Canoco	49
Table 5.1	Summary of all statistical results of models developed from historical laboratory generated data	85
Table 5.2	Frequency at which input variables were used in models to predict <i>Anabaena</i> sp., <i>Microcystis</i> sp., microcystin and geosmin concentrations.....	86
Table 6.1	Summary of all statistical results of models developed from historical laboratory generated data and tested with "unseen" laboratory generated data and "unseen" on-line (<i>in situ</i>) generated data	107
Table 6.2	Frequency at which input variables were used in models to predict <i>Anabaena</i> sp., <i>Microcystis</i> sp. and microcystin concentrations	108
Table A	Spearman Rank Order Correlations of environmental variables in the Vaal Dam for the period 2000 – 2013	146

LIST OF ACRONYMS AND ABBREVIATIONS

Ana	<i>Anabaena</i> sp. measured in cells/mL
APHA	American public health association
AWWA	American water works association
BMAA	β -N-methylamino-L-alanine
CCA	Canonical correspondence analysis
cells/mL	Cells per milli-litre, unit for measuring phytoplankton and cyanobacterial cell concentration
Chl _a	Chlorophyll- <i>a</i> measured in $\mu\text{g/L}$
Chl-665	Chlorophyll-665 (total photosynthetic pigments) measured in $\mu\text{g/L}$ (also see Tot pig)
CIMP	Cyanobacteria incident management protocol
COD	Chemical oxygen demand measured in mg/L
Cond	Electrical conductivity, measured in mS/cm
C-VD1	Catchment source water sampling point in the Vaal Dam at the dam wall
DO	Dissolved oxygen measured in mg/L
DOC	Dissolved organic carbon measured in mg/L
DWTW	Drinking water treatment works
DWA	Department of Water Affairs (of South Africa), currently known as DWS (Department of Water and Sanitation)
DWS	Department of Water and Sanitation (of South Africa), previously known as DWA (Department of Water Affairs)
EA(s)	Evolutionary algorithm(s)
ELISA	Enzyme linked immuno sorbent assay
GA	Genetic algorithms
GCMS	Gas chromatography mass spectrometry
GP	Genetic programming
HEA(s)	Hybrid evolutionary algorithm(s)
ICPMS	Inductively coupled plasma mass spectrometry
ILAC	International laboratory accreditation cooperation
IMP	Incident management protocol
MARS	Multi adaptative regression splines
M-B11_VG	Main source water sampling point in pipeline B11, supplying Rand Waters' Vereeniging drinking water treatment works

M-Canal	Main source water sampling point in the canal, supplying stations 3 and 4 at Rand Waters' Zuikerbosch drinking water treatment works
µg/L	Micro-gram per litre, unit for measuring chlorophyll- <i>a</i> , chlorophyll-665 and microcystin concentration
mg/L	Milli-gram per litre (unit for most inorganic chemical variables measured in water)
MIB/2-MIB	2-methylisoborneol measured in ng/L
Mic _n	Microcystin measured in µg/L
Mic _s	<i>Microcystis</i> sp. measured in cells/mL
mS/cm	Milli-Siemens per centimetre, unit for measuring electrical conductivity
ng/L	Nanno-gram per litre, unit for measuring geosmin and MIB concentrations
NHMRC	National health and medical research council (Australia)
NICD	National institute of communicable diseases
NTU	Nephelometric turbidity units, unit for measuring turbidity
NZMOH	New Zealand ministry of health
PCA	Principal component analysis
R ²	Square correlation coefficient
RMSE	Root mean squared error
SANAS	South African national accreditation system
SANS	South African national standard
Temp	Water temperature measured in °C
Tot pig	Total photosynthetic pigments measured in µg/L (also see Chl-665)
Tot alg	Total algae (including cyanobacteria) measured in cells/mL
Turb	Turbidity measured in NTU (also see NTU)
UKWIR	United Kingdom water industry research
WHO	World health organisation
WSP	Water safety plan

ACKNOWLEDGEMENTS

I would like to take the opportunity to sincerely thank the following people and institutions for their input, support and advice during this study:

- Rand Water for supplying the data for the project as well as allowing access to research tools and supplying a two year bursary for this study.
- The North West University for the platform to enrol for this study as well as excellent off-campus support regarding access to lecturers and scientific literature.
- Prof. Sandra Barnard from the North West University for guidance, support, excellent advice, scientific input and being a great mentor and friend.
- Prof. Hein du Preez from Rand Water for moral support, guidance, encouragement, scientific input and influencing the focus of the project. Without him, this project would have been impossible. It is a privilege to work for a manager/mentor of his calibre.
- Prof. Friedrich Recknagel and his students for assistance with the modelling techniques applied during this study.
- Dr. Hongqing Cao for assistance with the use of the super computer at Adelaide University as well as interpretation of the results.
- Dr. Carin van Ginkel for input regarding the sensitivity analysis technique applied.
- Ms. Ashvena Ramcharan, Ms. Ashvita Ramcharan, Ms. Carna Joubert, Ms. Elmarí Krüger, Ms. Ishana Dusrath, Ms. Lebohang Hanyane, Ms. Lelethu Bungu, Mr. Michael Sekonyela, Mr. Petrus Mofokeng, Dr. Rahzia Hendricks, and Ms. Rita Guglielmi from the Hydrobiology laboratory at Rand Water, for interest in this project, encouragement and accommodating me during the time I spent on it.
- My friends Hendrik Everts, Sanet Janse van Vuuren, Debbie de Wit and David de Wit for moral support and being a positive influence in my life, not only during this study.
- The Pienaar family and my uncle, Ron Coetzee, for moral support and putting up with me when I worked on the project during holidays and weekends with them.
- I wish my father was still alive to witness the completion of this study – he had such a big hand in motivating me to study science and believing in my abilities since under graduate years. He was my inspiration and my best friend. I truly miss him.
- The Kotzé family for making the three week stay in Adelaide so special, during the modelling phase of this project. Thank you for the memorable time me and my dad could spend with you. Our time there is one of the great memories of this project I will always cherish.

CHAPTER 1

INTRODUCTION

South Africa has an average rainfall of 497 mm per annum, which is clearly below the average rainfall of 860 mm for the rest of the world (South Africa.Info, 2013). This makes fresh water one of South Africa's most valuable and vulnerable natural resources. Although algae and cyanobacteria occur naturally in surface freshwaters (John et al., 2005), excessive growth of these organisms, called blooms, add to the pressure on water as a limited resource (Chorus, 2012). The severity, frequency, distribution, and impacts of harmful algal blooms mainly those of cyanobacteria and dinoflagellates, have increased worldwide (including in South Africa) in the recent decades (Harding and Paxton 2001; Codd et al., 2005; Guven and Howard, 2006; Van Ginkel and Silberbauer, 2007; Conradie and Barnard, 2012). These increasing algal blooms in source waters used for potable water purification result in extensive problems within the drinking water treatment works (DWTW). The occurrence of algae and cyanobacteria in the source water can cause a variety of different problems during the drinking water purification process (Knappe et al., 2004). These problems include the production of harmful cyanotoxins (Newcombe and Nicholson, 2004; Pantelić et al., 2013), the production of aesthetically unacceptable taste and odorous substances (Zoschke et al., 2011), the disruption of the coagulation, flocculation and sedimentation unit processes (Ewerts et al., 2013), the clogging of sand filters (Steynberg et al., 1998) and the formation of mud balls in the filter sand (Pieterse et al., 2000; Swanepoel et al., 2008b). Some of these problems in the DWTW contribute to the breakthrough of algae and cyanobacteria into the drinking water, which can be a vector for transporting cyanotoxins and other organic compounds into the drinking water (Pantelić et al., 2013). Their occurrence in the drinking water can also provide a carbon source for invertebrate and bacterial growth in the plant and the distribution system (Ferreira and Du Preez, 2012) as well as increase the potential for the formation of trihalomethanes – a toxic by-product of chlorination in the drinking water (Van der Walt et al., 2009; Pantelić et al., 2013).

The occurrence of especially high concentrations of cyanobacterial cells in the source water due to eutrophication is the reason for significant increases in drinking water purification costs (Graham et al., 2012). These increases are mostly due to the need for advanced treatment options (e.g. ozonation, powdered activated carbon, granular activated carbon, etc.) to deal with the organic compounds produced by cyanobacteria (Ho et al., 2011; Zoschke et al., 2011; Pantelić et al., 2013; Summers et al., 2013). These organic compounds include the notorious taste and odour compounds geosmin and 2-methylisoborneol (MIB) (Srinivasan and

Sorial, 2011) as well as cyanotoxins such as microcystin, cylindrospermopsin, anatoxin, saxitoxin, endotoxins (Du Preez and Van Baalen, 2006) and BMAA (β -N-methylamino-L-alanine) (Esterhuizen and Downing, 2008; Downing et al., 2011 and Pantelić et al., 2013).

Cyanobacterial blooms, as a symptom of eutrophication, have a notorious reputation to develop rapidly which usually catch water resource managers and even drinking water treatment managers off-guard (Van Ginkel, 2008). 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 can be delayed from 3 hours to fourteen 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. (Swanepoel et al., 2008a). Therefore the use of on-line (*in situ*) instruments that can supply real-time results by the click of a button and linked to real-time warning systems, have become very popular in the past few years (Randolph et al., 2008; Storey et al., 2011; Chang et al., 2012). On-line (*in situ*) instruments were developed, not only for chemical measurements such as pH, conductivity and nitrate concentration, but also for biological measurements such as chlorophyll-*a* and cyanobacterial cell concentration (Randolph et al., 2008) and even toxicity monitoring by utilising caged single species e.g. *Daphnia* or fish (Damásio et al., 2008). Although, this real-time (on-line) data has given drinking water treatment managers a better opportunity to make sound management decisions around drinking water treatment options, it may still be “too little, too late” once a sudden cyanobacterial bloom of especially *Anabaena* sp. or *Microcystis* sp. enters the DWTW. Therefore the benefit to managers and production chemists to be able to forecast future events of high cyanobacterial cell concentrations and high concentrations of cyanobacterial-related organic compounds in the source water has become evident.

The Vaal Dam (Figure 3.1) is the main source of raw water for the purification of drinking water for approximately 12 million consumers in South Africa (Viljoen, 2010). From the Vaal Dam, a 20 km long canal supplies source water to stations 3 and 4 at Rand Waters’ Zuikerbosch DWTW (Figures 3.1, 3.2 and 3.3). Occurrences of cyanobacterial blooms in the Vaal Dam can and have in the past, significantly increased the costs of drinking water treatment at this facility and may even impact millions of consumers when operators and managers are not vigilant regarding cyanobacterial blooms in the source water. The development of forecasting models for cyanobacterial cell concentrations and cyanobacterial-related organic compounds in the Vaal Dam to be applied at this large bulk DWTW, formed the basis of this study.

The research methodology applied in this study is evolutionary algorithms (EAs), which have been described as adaptive methods in search for suitable representations of models, to 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” (Cao et al., 2006). Hybrid evolutionary algorithms have been designed to discover predictive rule sets in complex ecological time series data by applying genetic programming for the optimisation of the rule structures. These rule structures were successfully used as prediction tools in many ecological studies (Talib et al, 2007; Chan et al., 2007; Recknagel et al, 2008; Van Ginkel, 2008; Welk et al, 2008; Recknagel et al., 2013 and Cao et al., 2013) and have also been applied on complex ecological time series data from the Vaal Dam to create forecasting models predicting cyanobacterial cell concentrations (*Anabaena* sp. and *Microcystis* sp.) as well as cyanobacterial-related organic compounds (microcystin and geosmin). These forecasting models were developed with the SANS 241:2011 guidelines (SANS 241-1, 2011; SANS 241-2, 2011) and the South African Blue Drop Certification Program (DWA, 2011; DWA, 2012) in mind. Models predicting *Anabaena* sp., *Microcystis* sp., microcystin and geosmin 7 days in advance were linked to the incident management protocols of Rand Water as case study (refer to Section 7.6) which, in turn, will impact positively on the application of the SANS 241 : 2011 drinking water standard (SANS 241-1, 2011; SANS 241-2, 2011) as well as the “incident response management”, “treatment process management” and “water monitoring program” aspects of the South African Blue Drop Certification Program (DWA, 2011; DWA, 2012).

To summarise, the aims of this study were the following:

- To review the physical, chemical and biological variables associated with cyanobacterial blooms in the Vaal Dam, as to get background on the input variables that can be used in cyanobacterial-related forecasting models.
- To apply rule-based Hybrid Evolutionary Algorithms (HEAs) on ecological time series data from the Vaal Dam and develop models using 1) all applicable laboratory-generated data and 2) on-line measureable data only, as input variables in forecasting models for cyanobacterial blooms in the Vaal Dam.
- To test these models with data that was not used to develop the models (so-called “unseen data”), including on-line (*in situ*) generated data.

-
- To incorporate selected models into two cyanobacterial incident management protocols which link to the Water Safety Plan (WSP) of a large DWTW (case study : Rand Water).

If these aims are met, it may ensure that managers and production chemists at large DWTW can prepare for incidents of high cyanobacterial cell concentrations in the source water, even before it occurs. This, in turn, will make a positive contribution to the adherence to the SANS 241 : 2011 drinking water standard (SANS 241-1, 2011; SANS 241-2, 2011), which will also filter through to supplied municipalities achieving good ratings in the South African Blue Drop Certification Program (DWA, 2011; DWA, 2012) and most of all, aid in ensuring the continuous supply of safe and aesthetically acceptable drinking water to consumers.

CHAPTER 2

LITERATURE SURVEY:

THE USE OF EVOLUTIONARY ALGORITHMS FOR THE DEVELOPMENT OF FORECASTING MODELS IN WATER ECOLOGY

2.1 *Introduction*

The utilisation of evolutionary algorithms (or hybrids thereof) with or without added mathematical optimisation strategies, for unravelling complex ecological relationships and forecasting or predicting certain problem-causing organisms have been widely published in especially limnological, marine and environmental modelling journals. Evolutionary algorithms (Chan et al., 2007) or hybrids thereof (Cao et al., 2013) or also called genetic algorithms (García Nieto et al., 2013), is an adaptive modelling technique based on the principles of biological evolution, natural selection and genetic variation. The principle of “survival of the fittest” applies to the models where rule sets and parameters are continually optimised to find the best fit relationship to the real data.

Models have been developed to utilise physico-chemical environmental variables together with certain biological variables (e.g. algal species cell concentration or algal species biovolume and chlorophyll-*a* concentration) (García Nieto, et al., 2013) to predict certain algae (including cyanobacteria) groups, genera or species as well as cyanotoxins (García Nieto et al., 2013) and taste and odour compounds (Dzialowski et al., 2009) in real time (Alonso Fernández et al., 2013). However, a few researchers endeavoured on longer-term forecasting e.g. 3 days in advance (Chan et al., 2007), 1 – 5 days in advance (Huang et al., 2012), 7 days in advance (Welk et al., 2008; Cao et al., 2013) and even 28 days in advance (Van Ginkel, 2008).

Forecasting models were mainly developed for problematic algal groups and species which include especially Cyanophyta and Dinophyta (Van Ginkel et al., 2007; Van Ginkel, 2008; Welk, et al., 2008; Cao et al., 2013; Recknagel et al., 2014), although other groups such as Chlorophyta and Bacillariophyta were also included in some studies (Recknagel et al., 2013). Forecasting models were also developed for organic compounds associated with cyanobacteria (García Nieto et al., 2013; Alonso Fernández et al., 2013) such as microcystin (Chan et al., 2007) and geosmin (Dzialowski et al., 2009).

Models that were developed seem to be relatively site-specific. Some models however, were applied (with varying success) to other water reservoirs of similar trophic status (Van Ginkel et al., 2007). Quantitative environmental modelling (which include forecasting or predictive modelling) in water ecology can play an important role in decision-making regarding reservoir management (Bennet et al., 2013) as well as drinking water purification management (Van Ginkel, 2008).

This literature survey is based on papers published in different journals on genetic modelling utilising evolutionary computation and dealing with predicting problematic algae (including cyanobacteria) and their related organic compounds, in the attempt to answer the following questions:

- Which ecological input variables were the most important and most frequently used in predicting algae (including cyanobacteria) and their related organic compounds?
- What strategies were used to evaluate model capabilities and accuracy?
- What optimisation strategies (if any) were utilised to enhance the forecasting ability of poor performing models?

2.2 Important input variables for algae (including cyanobacteria) related to forecasting models

Although a range of different physico-chemical and biological variables have been used as input components into the software for model development, not all of it have been incorporated into the “best fit” models during the evolutionary process of model “survival”. Even less has been identified as important when sensitivity analyses were performed. Generally the input variables ranged from the physical variables such as water temperature (Van Ginkel et al., 2007; Chan et al., 2007; García Nieto et al., 2013; Alonso Fernández et al., 2013; Cao et al., 2013; Recknagel et al., 2013) water flow or influx (Raine et al., 2010), turbidity, light availability (mostly determined by Secchi disk depth); chemical variables such as conductivity, pH, alkalinity, dissolved oxygen, a range of nutrients (e.g. total nitrogen, NO_3^- , NO_2^- , NH_4^+ , total phosphorous, PO_4^{3-} , total silica, SiO_2); and biological variables such as chlorophyll-*a* (Van Ginkel, 2007 and Van Ginkel et al., 2008) and different algal (including cyanobacteria) groups or species (García Nieto et al., 2013; Alonso Fernández et al., 2013) and even some invertebrate groups (Recknagel et al., 2013).

Models to predict chlorophyll-*a* concentration (Table 2.1) were developed by Welk et al., (2008) and Cao et al., (2013). The input variables for these models include only physico-chemical

variables such as water temperature, turbidity, DO, conductivity, pH, phosphorus, nitrogen and silica. Some of these variables are known to have a direct or indirect effect on chlorophyll-*a* concentration e.g. temperature, turbidity, conductivity and nutrients (Wetzel, 2001). The presence of excess nutrients (eutrophication) is particularly known to stimulate the growth of algae and cause the subsequent production of chlorophyll-*a*. Other variables for example DO and pH are known to be affected by especially photosynthesis of algae (which will increase with increasing chlorophyll-*a* concentration) (Wetzel, 2001). The presence of silica (SiO₂) as input variable in models predicting chlorophyll-*a* concentration suggests that diatoms or Bacillariophyta plays an important role in the prediction of chlorophyll-*a*, at least during certain seasons (Wetzel, 2001). The inclusion of water temperature suggests that cyanobacteria contribute significantly to chlorophyll-*a* concentration, as temperature usually correlates positively with cyanobacterial cell concentrations (Harding and Paxton, 2001; Conradie and Barnard, 2012).

Table 2.1 Input variables incorporated into models predicting chlorophyll-*a*.

Input variable	Authors
Water temperature	Welk et al., 2008; Cao et al., 2013;
Turbidity	Cao et al., 2013;
DO (dissolved oxygen)	Welk et al., 2008; Cao et al., 2013;
Total phosphorus (TP)	Cao et al., 2013;
EC (electrical conductivity)	Welk et al., 2008; Cao et al., 2013;
Total nitrogen (TN)	Cao et al., 2013;
SiO ₂	Cao et al., 2013;
pH	Cao et al., 2013;

Models predicting certain algal groups were developed by Van Ginkel et al., (2007), Van Ginkel, (2008) and Recknagel et al., (2013). Models predicting specific algal species were developed by Van Ginkel, (2008), Welk et al., (2008), Cao et al., (2013) and Recknagel et al., (2014). Models for the prediction of specific algae species or groups (Table 2.2) include physico-chemical variables very similar to those for the prediction of chlorophyll-*a* (Table 2.1). However, the use of chlorophyll-*a* and/or other algal or invertebrate groups seem to be included frequently in the models predicting specific groups or species. To be able to predict specific species (e.g. *Cylindrospermopsis* and *Microcystis*) turbidity seems to be more important than when only predicting the group to which these species belong (Cyanophyta) or other groups (Bacillariophyta, Chlorophyta or Dinophyta). It was expected that nitrogen fixing Cyanophyta such as *Cylindrospermopsis*, would not (unlike non-nitrogen fixing Cyanophyta e.g. *Microcystis*) have nitrogen included as input variable, but that does not seem to be the case. The inclusion of nitrogen in models predicting *Cylindrospermopsis*

(Cao et al., 2013; Recknagel et al., 2014), may be due to a secondary effect on other species affecting *Cylindrospermopsis* concentration and not necessarily on *Cylindrospermopsis* itself. Unfortunately the species of *Oscillatoria* studied by Welk et al., (2008) was not specified, since certain *Oscillatoria* species are known to fix nitrogen under certain conditions and others not, and therefore the inclusion of nitrogen in the *Oscillatoria* models cannot be interpreted one way or the other.

Comparing the prediction of the different groups (Cyanophyta, Bacillariophyta, Dinophyta and Chlorophyta – Table 2.2) it was observed that electrical conductivity played an important role in the prediction of Dinophyta. This was expected since different members of the Dinophyta are known to occur in a wide range of waters with different salinities and subsequently different electrical conductivity ranges (Wetzel, 2001). Dinophyta seems to be affected most by Chlorophyta and Cyanophyta (not by Bacillariophyta). Cyanophyta seems to be mostly affected by Chlorophyta and Dinophyta (not by Bacillariophyta). Bacillariophyta is mostly affected by Chlorophyta (not Cyanophyta or Dinophyta) (Recknagel et al., 2013). Chlorophyta is mostly affected by Bacillariophyta and Dinophyta (Recknagel et al., 2013). These observations may be due to the overlapping of certain groups of algae during certain seasons, which causes competition for resources. Chlorophyta, which is known to occur in a wide range of temperatures, seems to have the biggest effect on other algae (e.g. Dinophyta, Bacillariophyta and Cyanophyta), since their occurrence overlaps with almost all other groups of algae that are more temperature dependent. Bacillariophyta (usually occurring during lower temperature seasons) is only affected by (and affects) Chlorophyta, since Cyanophyta and Dinophyta are not present in abundance during the colder months. All groups of algae are affected by some or other invertebrates, which give an indication of the grazing capabilities of different invertebrates on the different algal groups (Recknagel et al., 2013).

Table 2.2 Input variables incorporated into the best models predicting algae (including cyanobacteria) genera and groups.

Output variable	Input variable	Authors
Cyanophyta	NO ₃ ⁻ inflow	Recknagel et al., 2013;
	TN:TP ratio	Recknagel et al., 2013;
	Chlorophyta	Recknagel et al., 2013;
	Dinophyta	Recknagel et al., 2013;
	Copepoda	Recknagel et al., 2013;
Chlorophyta	Epilimnion depth	Recknagel et al., 2013;
	Alkalinity	Recknagel et al., 2013;
	pH	Recknagel et al., 2013;
	PO ₄ ³⁻ inflow	Recknagel et al., 2013;

Table 2.2 (Cont.) Input variables incorporated into the best models predicting algae (including cyanobacteria) genera and groups.

Output variable	Input variable	Authors
Chlorophyta (cont.)	Dinophyta	Recknagel et al., 2013;
	Bacillariophyta	Recknagel et al., 2013;
	Rotifera	Recknagel et al., 2013;
Dinoflagellates / Dinophyta	Surface water temperature	Van Ginkel et al., 2007; Van Ginkel, 2008;
	EC (electrical conductivity)	Recknagel et al., 2013;
	DO	Recknagel et al., 2013;
	pH	Recknagel et al., 2013;
	DIP	Van Ginkel et al., 2007; Van Ginkel, 2008;
	Total N	Van Ginkel et al., 2007; Recknagel et al., 2013; Van Ginkel, 2008;
	Total P	Van Ginkel et al., 2007; Van Ginkel, 2008;
	TN:TP ratio	Recknagel et al., 2013;
	Chlorophyll- <i>a</i>	Van Ginkel et al., 2007; Van Ginkel, 2008;
	Chlorophyta	Recknagel et al., 2013;
	Cyanophyta	Recknagel et al., 2013;
	Copepoda	Recknagel et al., 2013;
Bacillariophyta	Water Temperature	Recknagel et al., 2013;
	DO	Recknagel et al., 2013;
	Total P	Recknagel et al., 2013;
	PO ₄ ³⁻	Recknagel et al., 2013;
	Chlorophyta	Recknagel et al., 2013;
	Cladocera	Recknagel et al., 2013;
	Copepoda	Recknagel et al., 2013;
	Total phosphorus (TP)	Cao et al, 2013;
	DO (dissolved oxygen)	Cao et al, 2013;
<i>Cylindrospermopsis</i> sp.	Water temperature	Cao et al, 2013; Recknagel et al., 2014;
	Turbidity	Cao et al, 2013; Recknagel et al., 2014;
	EC (electrical conductivity)	Cao et al, 2013; Recknagel et al., 2014;
	pH	Cao et al, 2013; Recknagel et al., 2014;
	SiO ₂	Cao et al, 2013; Recknagel et al., 2014;
	Total nitrogen (TN)	Cao et al, 2013; Recknagel et al., 2014;
	Total phosphorus (TP)	Recknagel et al., 2014;
	Dissolved oxygen (DO)	Recknagel et al., 2014;
	Chlorophyll- <i>a</i>	Recknagel et al., 2014;

Table 2.2 (Cont.) Input variables incorporated into the best models predicting algae (including cyanobacteria) genera and groups.

<i>Microcystis</i> sp.	Total phosphorus (TP)	Van Ginkel, 2008;
	Chlorophyll- <i>a</i>	Van Ginkel, 2008; Welk et al., 2008;
	Water (surface) temperature	Van Ginkel, 2008; Welk et al., 2008;
	Total nitrogen (TN)	Van Ginkel, 2008;
	pH	Van Ginkel, 2008;
	Secchi disk depth	Van Ginkel, 2008; Welk et al., 2008;
	Dissolved inorganic phosphorus (DIP)	Van Ginkel, 2008;
<i>Oscillatoria</i> sp.	Secchi disk depth	Welk et al., 2008;
	Water temperature	Welk et al., 2008;
	NO ₃ ⁻	Welk et al., 2008;
	SiO ₂	Welk et al., 2008;

Models for the prediction of organic compounds associated with cyanobacteria (e.g. cyanotoxins and geosmin – Table 2.3) have been developed by Chan et al., (2007), García Nieto et al., (2013) and Dzialowski et al., (2009). Models developed for microcystin (or cyanotoxin not specified by García Nieto et al., 2013), include physico-chemical as well as biological input variables. The physico-chemical variables predicting cyanotoxins are very similar to those predicting chlorophyll-*a* (Table 2.1) as well as the different cyanobacterial groups and species (Table 2.2). In addition to these common input variables (e.g. water temperature, turbidity, pH, DO and nutrients) the specific species known to produce these organic compounds, more often than not, also form part of the input variables. Chan et al., (2007) included another variable not previously discussed in this chapter, namely rainfall. This variable may be important to represent the factor showing inflow of nutrient-rich water into the aquatic ecosystem and the development and onset of a cyanobacterial bloom which may be associated with cyanotoxin production. García Nieto et al., (2013) and Alonso Fernández et al., (2013) included alkalinity as another variable that has not been incorporated into models for the prediction of chlorophyll-*a* and algal groups or species. Alkalinity therefore seems to play a role during the production of cyanotoxins rather than during cyanobacterial growth. García Nieto et al., (2013) and Alonso Fernández et al., (2013) also suggested a synergistic approach to using certain species data, namely the product of *Microcystis aeruginosa* and *Woronichinia naegeliana* concentrations, which produces a result greater than the sum of their individual effects (García Nieto et al., 2013). This seems to be a good approach if similar factors are both known to contribute to the same output variable and the sum or product of both contribute more significantly to the output variable than the factors individually.

No models developed by evolutionary computation could be found in literature for the prediction of geosmin concentrations. The only models found for predicting geosmin concentrations were done by Dzialowski et al., (2009), who used regression analysis in their model development. Therefore, due to a lack of geosmin models developed by evolutionary computation, the input variables of the models developed by regression analysis are discussed in this chapter, although it is not based on the same mathematical principles as the other models. As in the case of the microcystin models, the geosmin model also incorporates turbidity (in the form of Secchi disk depth), nutrients, pH, DO and the cyanobacterium known to produce geosmin (i.e. *Anabaena* sp.), together with total algae, total cyanobacteria and chlorophyll-*a*, which all represents the algae factor contributing to geosmin production.

Table 2.3 Input variables incorporated into models predicting organic compounds associated with cyanobacteria.

Output variable	Input variable	Authors
Microcystin / unspecified cyanotoxin	Water temperature	Chan et al., 2007; García Nieto et al., 2013; Alonso Fernández et al., 2013;
	Rainfall	Chan et al., 2007;
	Turbidity	García Nieto et al., 2013; Alonso Fernández et al., 2013;
	pH	Chan et al., 2007;
	Alkalinity	García Nieto et al., 2013; Alonso Fernández et al., 2013;
	DO	Chan et al., 2007;
	Total phosphorus	García Nieto et al., 2013; Alonso Fernández et al., 2013;
	PO ₄ ³⁻	Chan et al., 2007;
	NO ₃ ⁻	Chan et al., 2007;
	Chlorophyll- <i>a</i>	Chan et al., 2007;
	<i>Microcystis wesenbergii</i> (cell concentration)	Chan et al., 2007;
	<i>Microcystis viridis</i> (cell concentration)	Chan et al., 2007;
	<i>Woronichinia naegeliana</i> (biovolume)	García Nieto et al., 2013; Alonso Fernández et al., 2013;
	<i>Microcystis aeruginosa</i> x <i>Woronichinia naegeliana</i>	García Nieto et al., 2013; Alonso Fernández et al., 2013;

Table 2.3 (cont.) Input variables incorporated into models predicting organic compounds associated with cyanobacteria.

Output variable	Input variable	Authors
Geosmin*	Secchi disk depth	Dzialowski et al., 2009;
	PO ₄ -P	Dzialowski et al., 2009;
	NO ₃ -N	Dzialowski et al., 2009;
	<i>Anabaena</i> concentration	Dzialowski et al., 2009;
	Chlorophyll- <i>a</i>	Dzialowski et al., 2009;
	Total algae	Dzialowski et al., 2009;
	Total cyanobacteria	Dzialowski et al., 2009;
	Total phosphorus	Dzialowski et al., 2009;
	pH	Dzialowski et al., 2009;
	DO (dissolved oxygen)	Dzialowski et al., 2009;

* In the absence of models developed by evolutionary algorithms, these models predicting geosmin concentrations were included that were developed by regression analysis and not by evolutionary algorithms.

2.3 Evaluation of prediction or forecasting models

The success of a model to predict or forecast a certain variable, with an accuracy as high as possible, may be measured in different ways. The most common indicator of “goodness” or “fitness” of an empirical model is determining the correlation coefficient (R^2) (Welk et al., 2008; Van Ginkel et al., 2007; Van Ginkel, 2008; Alonso Fernández et al., 2013; García Nieto et al., 2013; Recknagel, et al., 2013; Recknagel et al., 2014). A correlation is drawn between the results from the model and the “real” data – the larger the R^2 -value, the better the model. This is mostly done on a validation dataset that was part of the dataset used for model development (the so-called “previously seen” data) e.g. Chan et al., (2007); Recknagel et al., (2013) and Recknagel et al., (2014). However, the real test comes when models are tested with a validation data set not used for training the model (so-called “unseen data”) which does not refer to the “testing dataset” or “validation dataset” used by some (Cao et al., 2013; Recknagel et al., 2014), since this “testing dataset” may still be part of the interpolated data that has been used for model development and is not really “unseen” (Van Ginkel, 2008). However, the correlation coefficient is not the only measure of a good model, root mean square error (RMSE) and visual comparison between the predicted and measured data are both also used to evaluate the ability of a model to predict variables accurately (Chan et al., 2007; Cao et al., 2013; Bennet et al., 2013).

Not all models discussed in this chapter displayed the same accuracy by means of which it can predict or forecast its individual variables. Certain models performed exceptionally well, e.g. real-time cyanotoxin prediction ($R^2 = 0.98$ – García Nieto et al., 2013) and others less so, e.g. real-time Chlorophyta prediction ($R^2 = 0.46$ – Recknagel et al., 2013). The success of a model can depend on many factors:

- Accuracy and credibility of the results from the dataset used for training the models.
- Size of the dataset used for training the models (larger is not necessarily better, since models can be over-trained by using too much data – Van Ginkel, 2008). Note that over-training a model can be avoided by implementing certain stopping criteria for the algorithmic training.
- Specific variables chosen as input variables (Recknagel et al., 2013).
- Forecasting time (longer forecasting time may be less or more accurate – Van Ginkel, 2008).
- Combining certain variables (a single combined variable may prove to be better than incorporating all individual variables – García Nieto et al., 2013 and Alonso Fernández et al., 2013).
- Number of independent runs or generations by utilising evolutionary algorithms or hybrids thereof (Recknagel et al., 2013).
- Type of optimisation strategy used for improving the models' forecasting capability (Alonso Fernández et al., 2013; Cao et al., 2013).
- The application of a model can determine whether it is fit for use, even though the statistical accuracy may prove to be insignificant.

Variables incorporated into a model are also evaluated by performing sensitivity analyses on these input variables. Sensitivity analysis identifies the most important input variables within a model and assesses how variation in input variables affects the model output (Bennet et al., 2013). It also aids in checking if the model resembles the system under study (Salancinska et al., 2010) and can be used to determine unimportant input variables and may result in prioritisation of future research, experimental designs and input parameters. In short, sensitivity analysis is a tool to aid in the understanding of models (Salancinska et al., 2010). Different methods for performing sensitivity analyses exist (Norton, 2008; Ravalico et al., 2009; Campolongo et al., 2011; Makler-Pick et al., 2011; Nossent et al., 2011; Shahsavani and Grimvall, 2011; Yang, 2011; Annoni et al., 2012; Sun et al., 2012;). Most methods of sensitivity analyses are based on changing a single or multiple input variable(s) and observing its effect on the output variable. If the effect is large, it indicates that a small change in the input variable may have a large influence on the output variable. The results of a sensitivity analysis can also

supply additional ecological information, since sensitivity analyses will show if a certain variable has a positive or negative influence on a variable and can indicate certain thresholds between which its influence may vary (Recknagel et al., 2013).

2.4 Optimisation strategies for poorly performing models

If models do not perform as well as was expected, different optimisation strategies have successfully been applied to improve the models' prediction or forecasting capabilities.

2.4.1 Synergistic approach to certain input variables

García Nieto et al., (2013) and Alonso Fernández et al., (2013) applied a synergistic approach to certain variables of which the combined effect is regarded to be larger than the sum of the two individual effects. An example of such a synergistic variable is the product of *Microcystis aeruginosa* and *Woronichinia naegeliana* concentrations to be used in models predicting cyanotoxin concentration. Combining certain variables may enhance the prediction capabilities of models forecasting especially organic compounds produced by algae (or cyanobacteria), which involve influences that cannot be captured by mere external environmental (physico-chemical) variables.

2.4.2 Rule-based models

Many of the models discussed in this chapter, utilise the rule-base strategy, which separate mathematical optimisation in different ranges of certain input variables (Van Ginkel et al., 2007; Chan et al., 2007; Van Ginkel, 2008; Cao et al., 2013; Recknagel et al., 2013 and Recknagel et al., 2014). These rule sets indicate certain thresholds of input variables under which conditions, different models will represent the ecological relationships occurring under those conditions. For instance, a model may contain a single rule (Van Ginkel et al., 2007):

IF Chlorophyll-*a* > 190.453 This rule will determine which of the models will be used to best predict or forecast the variable:

THEN Dinoflagellate biomass = $TN * \{Chlorophyll-a / [62.266 / \ln(|TP|)]\}$

ELSE Dinoflagellate biomass = $\{Chlorophyll-a / [(Surface\ temp * 0.662) - 5.829]\} / \ln(|DIP|)$

The rule therefore identifies chlorophyll-*a* concentration of 190.453 µg/L as threshold above and below which different conditions apply and different models will be used to best represent it.

Cao et al. (2013) has taken this a step further incorporating multiple rules, which enhanced the forecasting ability of the models.

E.g.:

IF ((SiO₂ > 9.541) or (DO ≥ 9.541))

THEN *Cylindrospermopsis* = ((EC – 56.29) * Water temp * TN

ELSE IF (SiO₂ > 6.711)

THEN *Cylindrospermopsis* = (EC – Turb) * Water temp * TN

ELSE IF (SiO₂ > 6.396)

THEN *Cylindrospermopsis* = Water temperature * (-89.615) * ln (|TP|) etc., which can carry on for a set number of runs.

This rule set identifies certain ranges of SiO₂ concentrations where different conditions apply and multiple models will be used to represent those conditions.

2.4.3 Hill climbing for parameter optimisation

Cao et al., (2013) compared two parameter optimisation strategies, one being the hill climbing. This is a local search approach which starts with a random solution, and iteratively makes small changes to the solution each time improving it. Hill climbing enabled the authors to improve the forecasting ability of their models considerably.

2.4.4 Differential evolution for parameter optimisation

The other optimisation strategy investigated by Cao et al., (2013) is differential evolution for solving real-parameter optimisation problems proposed by Storn and Price, (1997). Differential evolution is an optimisation technique that extracts the differential information from the current population of solutions as to guide its search further. The scheme is completely self-organising, since no separate probability distribution has to be used. Cao et al., (2013) concluded that both hill climbing and differential evolution improved their models, with hill climbing resulting in a slightly better result.

2.4.5 *Support vector regression*

García Nieto et al., (2013) made use of support vector regression in combination with evolutionary or genetic algorithms, to develop a cyanotoxin diagnostic model in a certain reservoir. Support vector regression is a learning technique based on statistical learning theory that uses the structural risk minimization principle. This technique was applied very successfully by the authors, since a correlation coefficient of 0.98 was obtained by applying it to experimental data.

2.4.6 *Multi adaptative regression splines (MARS)*

Alonso Fernández et al. (2013), combined evolutionary algorithms with multi adaptative regression splines (MARS technique), to develop a model forecasting cyanotoxin presence in fresh waters. The MARS technique builds a model in two phases, first the forward and then the backward pass. In the forward pass, the MARS technique repeatedly adds basis functions in pairs to the mean of the response values. This process of adding terms continues until the change in the residual error is too small to continue or until the maximum number of terms is reached. The forward pass usually builds an overfit model, but the backward pass then prunes this model, by removing terms one by one. It deletes the least effective term at each step until it finds the best sub-model. In order to determine which of the basic functions should be included in the model, the MARS technique uses generalised cross-over validation (Alonso Fernández et al., 2013). The MARS technique was successfully combined with evolutionary algorithms by the authors, to develop a model forecasting cyanotoxin presence in fresh waters with a correlation coefficient of 0.98.

2.4.7 *Incorporating time lags into the input variables*

Recknagel et al., (2014) have optimised models for the forecasting of *Cylindrospermopsis raciborskii*, by performing a systematic screening experiment to test all combinations of time lags between 0 and 10 days for the 5 input variables associated with the models. They found that models could be improved by including a time lag of 5 days for EC (electrical conductivity) and 1 day for DO (dissolved oxygen). For the other 3 input variables (pH, turbidity and water temperature) no improvement could be achieved when introducing time lags into the input variable data. Time lag in the canal from abstraction in the Vaal Dam to when water reaches

the Zuikerbosch plant is approximately 4 - 6 hours, therefore time lag was not investigated during the current study, but may be worth investigating if the models do not perform as well as expected.

2.5 *Conclusions*

Applying evolutionary algorithms (or hybrids thereof) with or without other mathematical optimisation techniques has successfully been utilised to develop models for the prediction and forecasting of algae (including cyanobacteria) and cyanotoxins. The size of the datasets used for developing or training such models differed, but the average is about ten to twelve years of data. Certain authors removed parts of their datasets from the training set as to have “unseen” data by means of which they could properly test and validate the models, although not all have done this. From the papers it was obvious that testing the models with “unseen” data proved to be a significant step in validating the forecasting ability of the models.

Evolutionary algorithms was not only successful in constructing models to predict real-time (same day) results, but also to construct models for future forecasting of algae (including cyanobacteria) and related organic compounds. Mostly (although not in all cases), the accuracy of the models decreased with increasing forecasting time. Models also seemed to be relatively site-specific, although it can be applied to water bodies of similar trophic status with varying success.

The input variables incorporated into the models were surprisingly similar. Although models in certain papers include unique variables (like specific invertebrate groups, or the product of two cyanobacterial species combined into one synergistic variable), the majority of input variables were the ones known for decades to affect or get affected by algae (including cyanobacteria). These common input variables include water temperature, turbidity, electrical conductivity, pH, dissolved oxygen, nutrients and other algae (in the form of chlorophyll-*a* or other algal groups or species). Models for the prediction of organic compounds associated with algae (including cyanobacteria) were more specific and included unique variables such as alkalinity as well as the specific species responsible for the production thereof. However, these models also include the variables responsible for the growth of the species responsible for the production like temperature, turbidity, nutrients etc.

A variety of mathematical optimisation strategies have been discussed in this chapter and it seems that these strategies are able to improve the forecasting ability of models developed by evolutionary algorithms. Two methods proven to enhance the models in the papers reviewed are hill climbing and differential evolution, of which hill climbing proved to be slightly better than differential evolution.

From this literature survey, it is obvious that the application value of forecasting models developed by evolutionary computations are vast and should be exploited to aid in the management of more fresh water bodies and catchments as well as in drinking water purification management.

CHAPTER 3

MATERIALS AND METHODS

3.1 *Study area, sampling and analyses*

The Vaal Dam (Figure 3.1) is the main source of raw water for the purification of drinking water at Rand Water – South Africa’s largest bulk drinking water treatment facility (Viljoen, 2010). It is located on the border between the Gauteng and Free State Provinces and is approximately 150 km south of Johannesburg. The catchment area of the dam is approximately 38 500 km² with a wall height of 63.4 metres 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 (Matete and Hassan, 2006). This water is transported from Lesotho via the Liebenbergsvlei and Wilge Rivers (LHDP, 2013). Forty percent (40%) of the water in the Vaal Dam originates from the Lesotho Highlands Water Project¹, which is generally of pristine quality. However, point- and non-point sources of pollution along the transferring rivers, negatively impact the water quality until it reaches the Vaal Dam a few hundred kilometres downstream. This pollution results in annual cyanobacterial blooms in the Vaal Dam causing the trophic status of the dam to be classified as “mesotrophic” (DWA, 2013a). The trophic status classification was done according to the system used by the South African Department of Water Affairs (DWA, currently known as DWS – Department of Water and Sanitation) where mean total phosphate, mean chlorophyll-*a* concentration and percentage of time when chlorophyll-*a* is >30 µg/L is taken into consideration (DWA, 2013a).

From the Vaal Dam, a 20 km long canal (Figures 3.1 and 3.2) supplies water to South Africa’s largest bulk drinking water treatment facility (Figure 3.3) on the banks of the Suikerbosrand River. This facility can produce approximately 3800 ML of drinking water per day (depending on demand).

For the past fifteen years (and longer regarding certain variables), extensive water quality monitoring were performed by Rand Water on both the Vaal Dam (C-VD1 – Figure 3.1 ■) and the canal (M-Canal – Figure 3.1 ■). This was done in order to aid managers in making informed water treatment decisions regarding the quality of source water entering the DWTW.

¹ Personal communication: Mr. Francois van Wyk, Catchment Manager, Rand Water, Scientific Services.

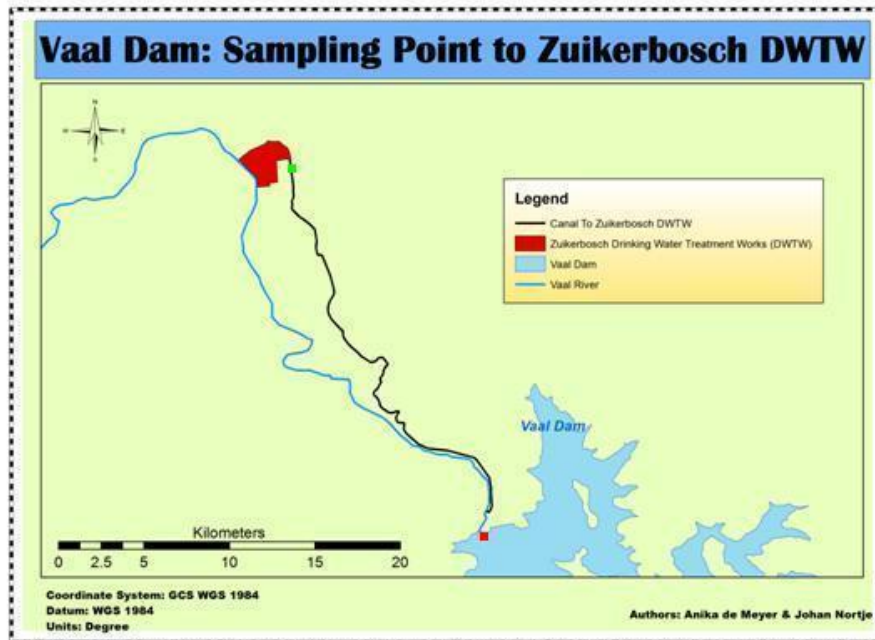


Figure 3.1 Sampling points in the Vaal Dam (■) and canal (■) supplying untreated water to the Zuikerbosch Drinking Water Treatment Works (DWTW).

Some of the variables were measured *in situ* in the Vaal Dam – namely Secchi disk depth (taken with a 30 cm diameter Secchi disk) together with water temperature, conductivity, pH, dissolved oxygen (concentration including percentage saturation) and turbidity (all taken with a YSI 556 handheld field instrument).



Figure 3.2 Twenty kilometre long canal (with a hydraulic retention time of 4 – 6 hours), supplying source water to the Zuikerbosch DWTW.



Figure 3.3 Google Earth picture of stations 3 and 4 at the Zuikerbosch DWTW (Google Earth, 2013).

The analyses (chemical and biological) were performed in the laboratories of Rand Water Analytical Services (Laboratory number: T0046), which is a SANAS (South African National Accreditation System) accredited laboratory. SANAS is affiliated internationally with ILAC (International Laboratory Accreditation Cooperation). Laboratory data used in this study have all been generated by SANAS accredited methods. Metals (Si, Fe, Mn) were determined simultaneously with the multi-element ICPMS (Inductively Coupled Plasma Mass Spectrometry) method, according to Standard Methods (APHA, 2013). The anions NO_2^- and NO_3^- (both measured as mg/L N) were determined by ion chromatography by means of the Metrohm 761 Compact IC-System according to Standard Methods (APHA, 2013). Ammonium (NH_4^+ - measured as mg/L N) and ortho phosphate (PO_4^{3-} - measured as mg/L P) were determined by an Aquakem automated colorimetric analysis system, according to Standard Methods (APHA, 2013). Alkalinity was determined by means of titration according to Standard Methods (APHA, 2013). Chlorophyll-*a* as well as total pigments (chlorophyll-665) were determined spectrophotometrically according to Swanepoel et al. (2008a). These two methods are based on methods originally described by Sartory (1982) and Steynberg (1986) respectively. Phytoplankton identification and enumeration was performed with the sedimentation technique, originally described by Lund et al., (1958) and adapted for the laboratory according to Swanepoel et al., (2008a). The percentage composition (i.e. individual contribution to the total number of cells enumerated) was determined for all phytoplankton groups (such as Cyanophyceae, Chlorophyceae, Bacillariophyceae, Dinophyceae, Euglenophyceae, Cryptophyceae and Chrysophyceae) and also for all genera within the groups. Geosmin and 2-methylisoborneol (MIB) were extracted using the purge and trap method and quantified by

means of the gas chromatography-mass spectrometry (GCMS) method based on Standard Methods (APHA, 2013) and adapted for the laboratory according to Swanepoel et al., (2008a). The concentration of the cyanotoxins, microcystin (total microcystin), was determined by means of the Enzyme Linked Immuno Sorbent Assay (ELISA) technique, also described by Swanepoel et al., (2008a), using Enviroligix ELISA kits (Enviroligix, 2013).

At the sampling point in the canal (M-Canal – Figure 3.1 ■) via which the source water is supplied to Stations 3 and 4 of the Zuikerbosch DWTW (Figure 3.3), an on-line (*in situ*) monitor (YSI 6600 V2, Multi-parameter Water Quality Sonde – Figure 3.4) was installed during 2008, which measures water temperature (°C), turbidity (NTU), pH, conductivity (mS/cm), dissolved oxygen (mg/L and % saturation), chlorophyll-*a* (µg/L) and cyanobacterial cell concentration (cells/mL). Data from this instrument was also used as test dataset for the models.



Figure 3.4 On-line (*in situ*) YSI 6600 V2 Multi-parameter water quality sonde in the canal measuring water temperature (°C), turbidity (NTU), conductivity (mS/m), dissolved oxygen (mg/L and % saturation), pH, chlorophyll-*a* (µg/L) and cyanobacterial cell concentration (cells/mL).

It is well known that estimating phytoplankton biomass, is a very useful measurement in limnology (Kasprzak, et al., 2008) and it proved to be vital during this study since it was incorporated into more than half of the models developed during this study. However, Rand Water (that supplied the data) uses two different methods to determine phytoplankton biomass namely chlorophyll-*a* and chlorophyll-665 (total pigments concentration) and therefore certain models incorporate chlorophyll-*a* as input variable while others include chlorophyll-665 as input variable. The two different analyses are performed for specific

reasons. Rand Water classifies their sampling sites according to 1) location, 2) what type of water it contains 3) the proximity to and influence on the DWTW and 4) the reason for sampling. Sampling sites are therefore classified as “catchment”, “drinking water mains”, “source water mains”, “projects”, “effluent”, “sanitation” etc., and on the basis of this classification certain analyses are carried out on certain sampling sites and not on others. For example, chlorophyll-*a* analysis is carried out on all catchment sampling sites (to be able to compare with chlorophyll-*a* results from catchments all over the world), while total pigments (chlorophyll-665) and geosmin analyses are carried out on all source and drinking water mains (to be able to evaluate algal and organic compound removal efficacy during drinking water treatment). During this investigation the classification of the sampling sites caused some challenges in the sense that C-VD1 (Figure 3.1 ■) is classified as a “catchment” site where chlorophyll-*a* is analysed; while M-Canal (Figure 3.1 ■) is classified as a “source water main” site where total pigments concentration (chlorophyll-665) is analysed. During this study, it was necessary to determine the relationship between these two variables as to be able to use chlorophyll-*a* and chlorophyll-665 interchangeably on models from different sampling sites. The determination and validation of this relationship is discussed in Section 3.6.

3.2 Hybrid Evolutionary Algorithms (HEAs)

“Ecological informatics” is defined as an interdisciplinary framework, promoting the use of advanced computational technology for the elucidation of principles of information processing at and between all levels of complexity of ecosystems – from genes to ecological networks – and aiding transparent decision-making in relation to important issues in ecology such as sustainability, biodiversity and global warming (Van Ginkel, 2008). Computation techniques such as fuzzy logic, cellular automata, artificial neural networks, evolutionary algorithms (EAs) and adaptive agents are considered as core concepts of ecological informatics (Van Ginkel, 2008).

Figure 3.5 represents the flowchart of hybrid evolutionary algorithms (HEAs) and Figure 3.6 shows the principle approach of using HEAs for the discovery of predictive rule-sets according to Cao et al., (2006). In the context of this research, a HEA designed for rule discovery in water quality time-series data was applied. The HEA uses 1) genetic programming to create and continually optimise the structure of the rule sets, and 2) a genetic algorithm to optimise the parameters of the rule set (Figure 3.5).

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 in Australia. This work was done with the assistance of Dr. H-C Cao and Prof. F. Recknagel at the University of Adelaide. Data from physical, chemical and biological water quality analyses performed on samples from the Vaal Dam (C-VD1 – Figure 3.1 ■) and from the canal (M-Canal – Figure 3.1 ■) for the ten year period 2000 to 2009, were used to develop the models. Data from analyses varied in frequency of sampling (from daily for certain variables, to monthly for other variables).

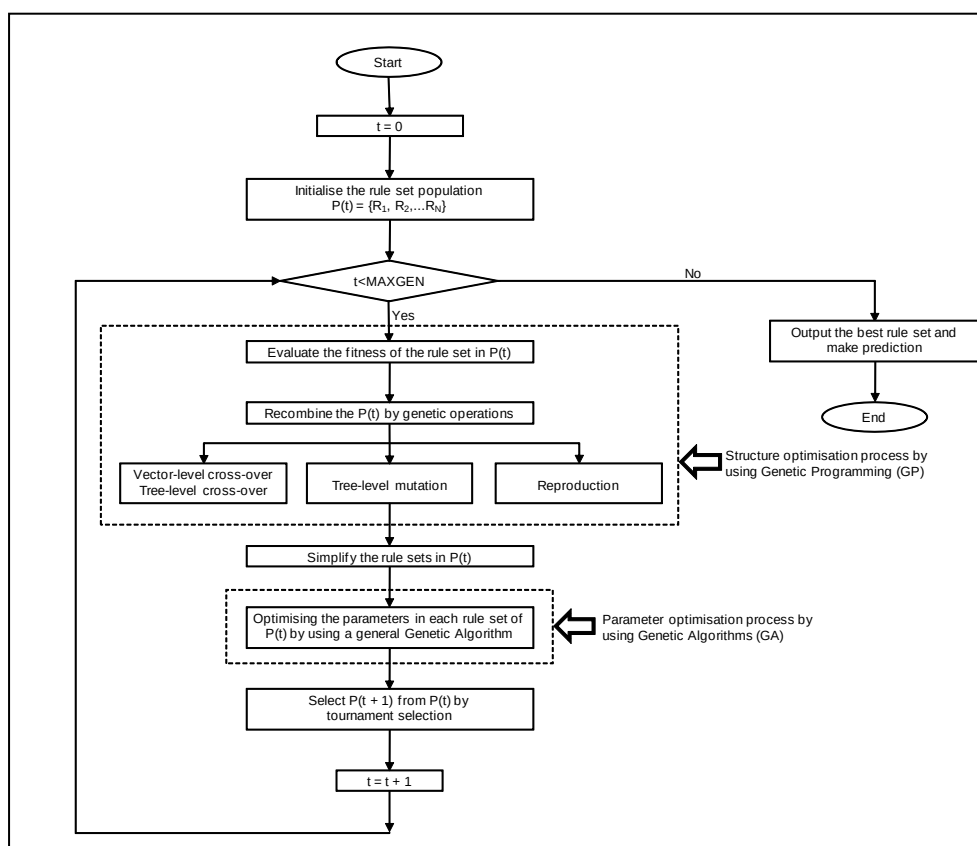


Figure 3.5 Flow chart of the hybrid evolutionary algorithms (HEAs) performing rule structure optimisation by genetic programming and parameter optimisation by genetic algorithms – adapted from Cao et al., (2006).

Because of the fact that samples were taken at different frequency intervals (in the Vaal Dam once a month and in the canal once a week or once a month for different variables), and prediction time necessitated intervals of 7, 14 and 21 days respectively, all the data was linearly interpolated to have corresponding results for all variables. A general linear interpolation algorithm was chosen since little variation in the data with few peaks occurring far apart were observed during the overview of algae and associated nutrients in the Vaal Dam (Chapter 4).

Interpolation may 1) cause the exclusion of peaks and troughs and 2) introduce bias in the data. However, since of ten years of data was used for training the models and three years of unseen data for testing the models, it was concluded that the exclusion of peaks and troughs was possible but acceptable in this study. The introduction of bias is also acknowledged, but since the data does not show a normal distribution, bias could not have been prevented with or without the introduction of linear interpolation. For the development and first stage of testing the models, 75% of the dataset was used to train the models and the remaining 25% of the dataset was used to test the models. Boot-strapping was 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 implies that a different 75% portion of the data will be used during training for all the different models. The reasons why boot-strapping was used in this study were because 1) different training data sets are used to obtain various models with a wider diversity and more robustness and 2) the best model is evaluated based on the validation results of the whole data set (Cao et al., 2013). Models were also tested with 2 to 3 years of so called “unseen data” (Vaal Dam 2010 – 2012) i.e. data not used at any stage during the model development (Bennet et al., 2013). Fifty different models were developed for each set of “x input variables = 1 output variable” chosen beforehand. From the 50 models, the best model between the measured data and the predicted data was chosen based on the following criteria: 1) root mean square error [RMSE], 2) the square correlation coefficient [R^2 -value] and 3) visual comparison between the predicted and measured data as according to Chan et al. (2007) and Bennet et al. (2013).

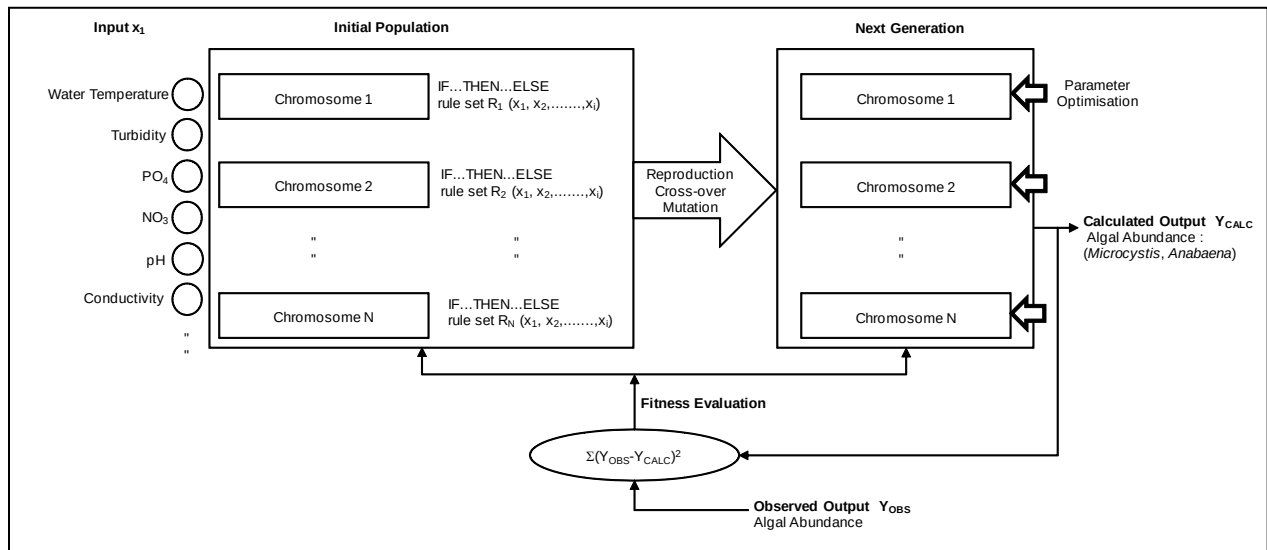


Figure 3.6 Conceptual diagram for using the hybrid evolutionary algorithm (HEA) for the discovery of predictive rule sets in water quality time-series data – adapted from Cao et al., 2006.

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 24h to 72h to complete the computation.

Input variables were consciously chosen to 1) either include all chosen variables (refer to Chapter 5) or 2) only variables that can be measured by on-line (*in situ*) instruments (refer to Chapter 6). This was done to specifically develop models that can be applied by using only data from the on-line (*in situ*) instruments, and to determine if these models were significantly less accurate than ones developed using all principle component variables as potential input variables.

The output variables for which predictive models were developed during this study (i.e. real time, 7 days in advance, 14 days in advance and 21 days in advance), were *Anabaena* sp., *Microcystis* sp., total cyanobacterial cell concentration, microcystin and geosmin, since these are the variables that can have an adverse effect on the drinking water treatment process as well as on consumer health (SANS 241-1, 2011; SANS 241-2, 2011) and consumer confidence in tap water (Knappe et al., 2004; Newcombe and Nicholson, 2004; Srinivasan and Sorial, 2011).

It is important to note that methods exist for optimising models developed by hybrid evolutionary algorithms, such as modelling single or multiple rules, optimising model parameters by hill climbing or differential evolution (Cao et al., 2013; Recknagel et al., 2013). However, these optimisation strategies were not employed during this investigation because most of the models developed in this study were already proven to be significantly more accurate than many models that were optimised by making use of these methods (Cao et al., 2013; Recknagel et al., 2013;).

3.3 Statistical analyses

Principal component analysis (PCA) was performed on the Vaal Dam dataset from 2000 - 2012 by using Statistica 10 for Windows (Chapter 4). This was done in order to get an in-depth view of the relationship between specific environmental variables that was thought to play a significant role in the development and onset of cyanobacterial blooms and their organic products in the Vaal Dam. A canonical correspondence analysis (CCA) was also performed to relate the effect of the environmental variables to the occurrence of certain phytoplankton

(including cyanobacteria) species. The phytoplankton concentrations were the only variables transformed to the natural log of the concentrations to reduce the large variability in the phytoplankton cell counts. The software package CANOCO, Version 4.5 was used (TerBraak, 1988) to perform the CCA. Ordinations were interpreted using the following rationale: Parameters are 1) positively correlated with each other if their arrows subtend a small angle, 2) not correlated if their arrows are 90° 3) negatively correlated if their arrows are directed oppositely (180°); 4) 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 (Chapter 5 and Chapter 6) tested with 1) 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”), 2) 3 years of “unseen data” (data not used in training the models – 2010 - 2012) and 3) 2 years of “unseen *in situ* generated data” (Chapter 6), were determined by XLSTAT, Version 2009.4.06.

During the development of the model to determine the relationship between chlorophyll-*a* and chlorophyll-665 (Section 3.6), regression analysis between chlorophyll-*a* and chlorophyll-665 was performed with Excel 2007. The paired students’ t-test (two-tailed) between the measured chlorophyll-665 and the predicted chlorophyll-665 (using the model) were done in GraphPad Prism Version 4.00 to determine if the correlation was statistically significant. The correlation coefficient (R^2 -value) was determined for the regression analysis (by using Excel 2007) and also for the model verification (by using XLSTAT). The Root Mean Squared Error (RMSE) between the measured chlorophyll-665 and the predicted chlorophyll-665 results was determined by means of XLSTAT, Version 2009.4.06.

3.4 Sensitivity analysis

Sensitivity analysis identifies the most important input variables within a model and assesses how variation in input variables affects the model output (Bennet et al., 2013). It also aids in checking if the model resembles the system under study (Salancinska et al., 2010) and can be used to determine unimportant input variables and may result in prioritisation of future research, experimental designs and parameters. In short, sensitivity analysis is a tool to aid in the understanding of models (Salancinska et al., 2010). During this study, sensitivity analyses were carried out for the best performing predictive rule sets as follow: 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) were 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 variable being tested was substituted with the range of values from 0% (which was effectively the minimum value of the input range), 5%, 10%, 15% etc. up to 100% (which was effectively the maximum value of the input range). This was done for each individual variable in the model and a graph was constructed with the results from this analysis (with % input range from 0 to 100% on the x-axis and the result from the model on the y-axis). The graph 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.

3.5 *Application of models*

To be able to apply these predictive models in the optimal functioning of a large DWTW, the application avenue of Water Safety Plans (WSP) were chosen. A Water Safety Plan (WSP), is a step-by-step risk management strategy to ensure consistent safe drinking water supply from catchment to consumer (Bartram et al., 2009). This study focused on two protocols that involve cyanobacteria, namely a microcystin protocol and a taste and odour protocol. Du Preez and Van Baalen, (2006) and Du Preez et al., (2007) call such protocols “step-by-step alert levels frameworks” setting in motion monitoring and management action sequences that drinking water utilities can proactively apply to deal with the problems cyanobacteria can cause in DWTW. The Du Preez and Van Baalen (2006) approach to such action sequences was applied in the WSP of Rand Water as study case (Section 7.6) where different variables, e.g. microcystin concentration (rounded to the first decimal), cyanobacterial cell concentration (either *Microcystis* sp., or *Anabaena* sp., or both), chlorophyll-*a* concentration, geosmin concentration and mouse bioassays in combination with each other, are used to trigger different action sequences in order to deal with increasing or decreasing concentrations of cyanobacteria and their related organic compounds in the source and drinking water. An example of such a protocol is displayed in Figure 3.7 (adapted from Du Preez and Van Baalen, 2006).

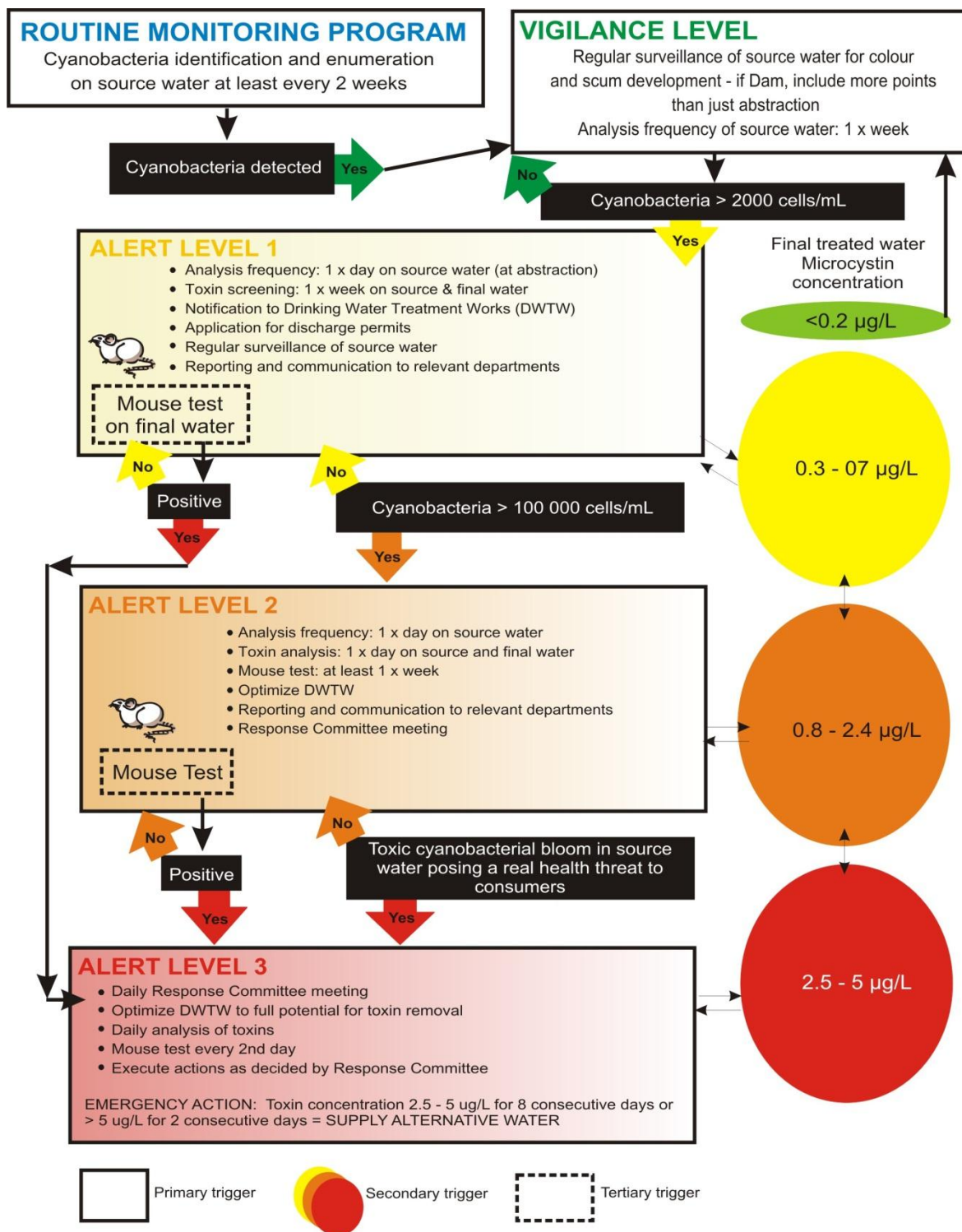


Figure 3.7 Cyanobacterial incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Du Preez and Van Baalen, 2006).

3.6 *Development and validation of a model to determine chlorophyll-665 (total photosynthetic pigments concentration) from chlorophyll-a and vice versa*

Some of the models in this study use chlorophyll-*a* as input variable (refer to Figure 5.7, Figure 5.9, Figure 5.15, Figure 5.17, Figure 5.19, Figure 5.21 and Figure 5.23) and others use chlorophyll-665 (total photosynthetic pigments concentration) as input variable (refer to Figure 5.25 – input variable “Tot pig”). The data supplied for this study however, only has one (not both) of these variables per sampling site. This may limit the application of the models if it is applied at a sampling site where the input variable is not measured. Therefore, it was necessary to 1) determine if there exists a statistical significant relationship between chlorophyll-*a* and chlorophyll-665, 2) develop a model by means of which this relationship can be calculated and 3) determine how comparable (or accurate) this calculated values would be when compared to the real values. A null hypothesis of Chlorophyll-665 > Chlorophyll-*a* was used which was true for every data point and therefore a probability failure was not determined.

Chlorophyll-*a* and chlorophyll-665 results from M-B11_VG and M-Canal (both are source water sampling points for Rand Water) for the period 29 September 2008 to 8 January 2012 (225 comparable data points) were used to develop the model. Regression analysis between chlorophyll-*a* and chlorophyll-665 was performed with Excel 2007. The linear regression equation (the simplest model that showed one of the highest R^2 -values when compared to other regression analyses) was used as the model. By using the model, the predicted chlorophyll-665 concentration was determined.

An additional test was performed to determine how well the model would perform with data that was not used to develop the model (so-called “unseen” data). Therefore it was tested with data from 12 February 2012 – 11 November 2013, (which included 254 comparable data points from M-B11_VG and M-Canal that have not been used to develop the model). The correlation coefficient (R^2 -value) and RMSE were both determined by using XLSTAT.

Since the relationship between chlorophyll-665 and chlorophyll-*a* was determined to be linear and constant, with no other variables influencing the model, the same statistical validations that apply to the determination of chlorophyll-665 also apply to the determination of chlorophyll-*a* and were therefore not presented in this section.

3.6.1 Model development

Using the data set of 29 September 2008 to 8 January 2012, a regression analysis done in Excel 2007, was performed between the chlorophyll-*a* and chlorophyll-665 results (Figure 3.8), which revealed a linear relationship, determinable by the following model:

$$\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$$

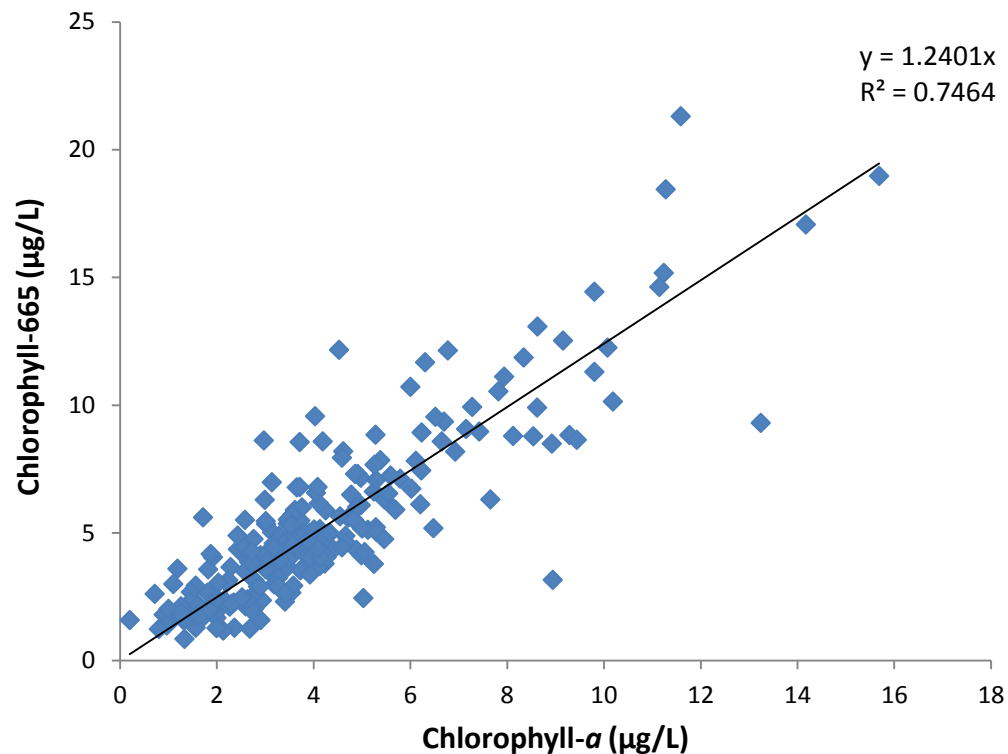


Figure 3.8 Regression analysis of chlorophyll-*a* vs. chlorophyll-665 at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012.

The correlation coefficient (R^2 -value) of this relationship (Figure 3.8) was determined to be 0.75 with $n = 225$.

3.6.2 Model testing using “previously seen” data

Using the model ($\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$), the measured chlorophyll-665 results were compared to the predicted chlorophyll-665 results (Figure 3.9 and Table 3.1) and a paired students’ t-test was performed on the results. The paired students’ t-test (Table 3.1) showed that there was a statistical significant correlation between the two and that the means were not significantly different ($P = 0.3089$).

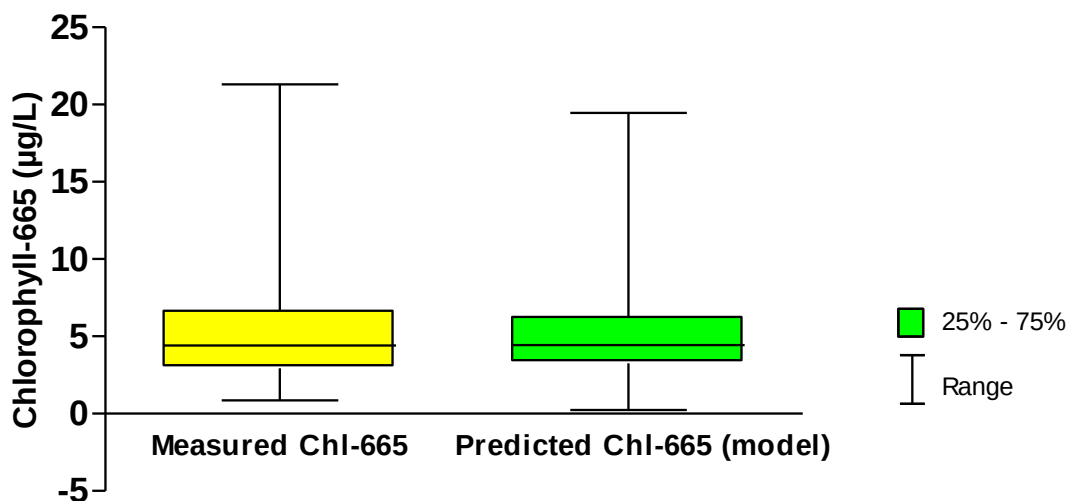


Figure 3.9 Measured chlorophyll-665 vs. predicted chlorophyll-665 at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012 (“previously seen” dataset).

Table 3.1 Results from the paired students’ t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model: ($\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$), with the “previously seen” dataset.

PARAMETER	VALUE
Paired t test	
P value	0.3089
P value summary	ns
Are means signif. different? ($P < 0.05$)	No
One- or two-tailed P value?	Two-tailed
t, df	$t=1.020$ $df=224$
Number of pairs	225
How big is the difference?	
Mean of differences	0.1167
95% confidence interval	-0.1076 to 0.3411
R squared	0.004622

Table 3.1 (cont.) Results from the paired students' t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model:
(Chlorophyll-665 = 1.24 * Chlorophyll-*a*), with the “previously seen” dataset.

PARAMETER	VALUE
How effective was the pairing?	
Correlation coefficient (r)	0.8665
P Value (one tailed)	P<0.0001
P value summary	***
Was the pairing significantly effective?	Yes

When the measured chlorophyll-665 results were compared to the predicted chlorophyll-665 results as determined by the model (Figure 3.10), the correlation coefficient was $R^2 = 0.75$ – the same as determined by the linear regression as was expected (Figure 3.8). The Root Mean Squared Error (RMSE) was 1.58 µg/L.

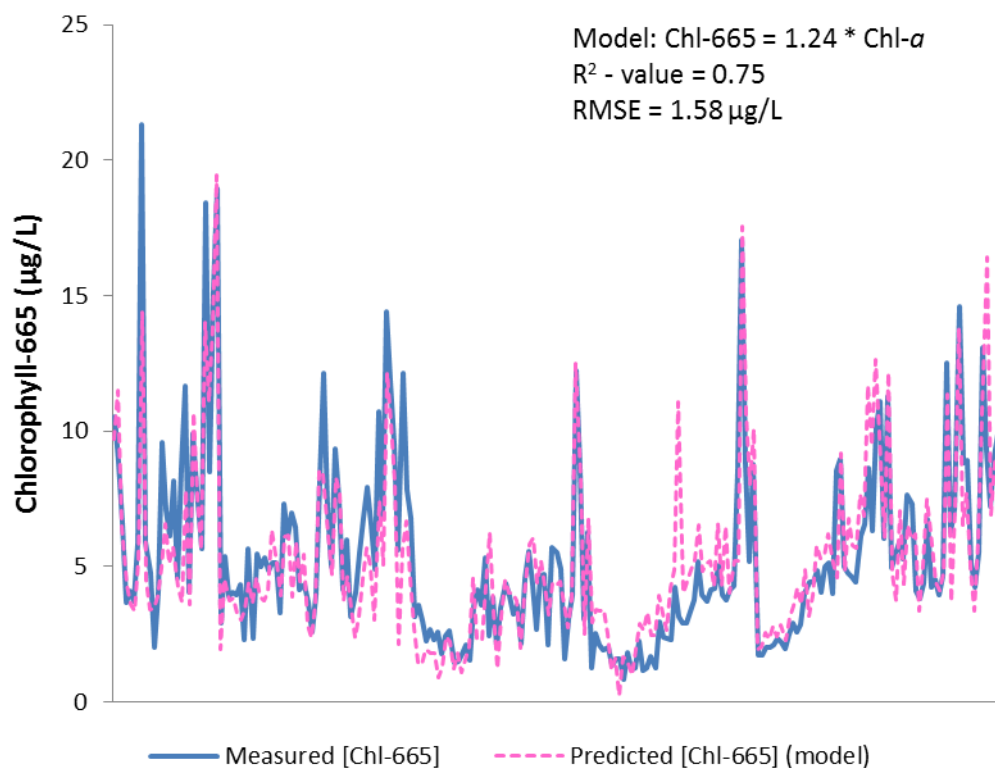


Figure 3.10: Validation of model (Chlorophyll-665 = 1.24 * Chlorophyll-*a*) at M-B11_VG and M-Canal from 29 September 2008 to 8 February 2012 (“previously seen” dataset).

3.6.3 Model testing using “unseen” data

Using the data set of 12 February 2012 – 11 November 2013 (i.e. data not used to develop the model – so called “unseen” data), a paired student’s t-test was performed between the measured chlorophyll-665 results and the predicted chlorophyll-665 results as determined by the model ($\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$). The results (Figure 3.11 and Table 3.2) revealed there was a statistical significant correlation between the two and that the means were not significantly different ($P = 0.1268$).

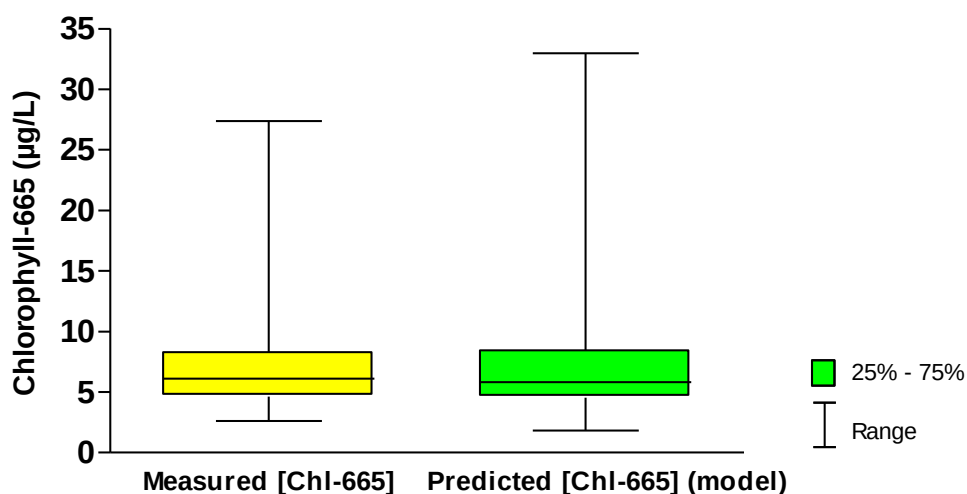


Figure 3.11 Measured chlorophyll-665 vs. predicted chlorophyll-665 at M-B11_VG and M-Canal from 12 February 2012 to 11 November 2013 (“unseen” dataset).

Table 3.2 Results from the paired students’ t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model:
($\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$) with the “unseen” dataset.

PARAMETER	VALUE
Table Analyzed	Data 1
Column A	Measured [Chl-665]
vs	vs
Column B	Predicted [Chl-665] (model)
Paired t test	
P value	0.1268
P value summary	ns
Are means signif. different? ($P < 0.05$)	No
One- or two-tailed P value?	Two-tailed
t, df	$t=1.532$ $df=253$
Number of pairs	254

Table 3.2 (cont.): Results from the paired students' t-test of the measured chlorophyll-665 vs. the predicted chlorophyll-665 result, using the model:
(Chlorophyll-665 = 1.24 * Chlorophyll-*a*) with the “unseen” dataset.

PARAMETER	VALUE
How big is the difference?	
Mean of differences	0.1545
95% confidence interval	-0.04319 to 0.3522
R squared	0.009189
How effective was the pairing?	
Correlation coefficient (r)	0.8639
P Value (one tailed)	P<0.0001
P value summary	***
Was the pairing significantly effective?	Yes

When the model was tested with “unseen” data that was not used to develop the model (Figure 3.12), the correlation coefficient (R^2 -value) was still 0.75 and the RMSE = 1.57 $\mu\text{g/L}$ where $n = 254$.

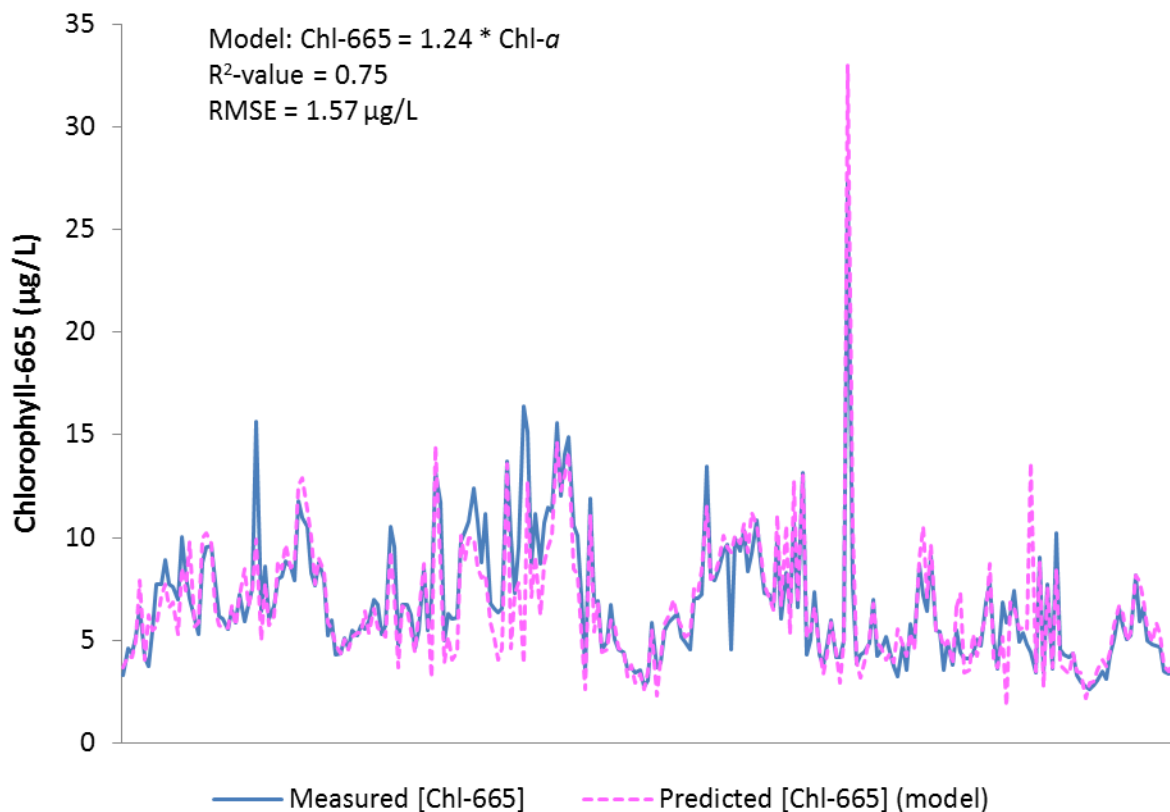


Figure 3.12: Validation of model (Chlorophyll-665 = 1.24 * Chlorophyll-*a*) at M-B11_VG and M-Canal from 12 February 2012 to 11 November 2013 (“unseen” dataset)

3.6.4 Conclusions

From the results it was evident that a statistical significant relationship exists between the chlorophyll-*a* and chlorophyll-665 concentrations. The model to represent this relationship was determined to be:

$$\text{Chlorophyll-665} = 1.24 * \text{Chlorophyll-}a$$

Although this model is not 100% accurate (R^2 -value between measured and predicted results are 0.75 when tested with “previously seen” data used to develop the model and also when tested with “unseen” data), the model is statistically significant and showed no significant differences between the means. Visual inspection also reveals a very good correlation for peaks and troughs in the data. This model may therefore be very useful to estimate the chlorophyll-665 from the chlorophyll-*a* results or *vice versa*. The root mean squared errors were determined to be 1.58 and 1.57 µg/L when tested with “previously seen” and “unseen” data respectively.

CHAPTER 4

AN OVERVIEW OF THE ALGAE AND ASSOCIATED NUTRIENTS IN THE VAAL DAM FROM 2000-2012

4.1 *Introduction*

The Vaal Dam serves as source to the largest bulk drinking water purification utility in South Africa which provides approximately 3800 ML of drinking water per day to approximately 12 million consumers within an 18 000 km² area (Viljoen, 2010). Cyanobacteria and other algae occur naturally in the Vaal Dam and during the warmer summer months scums of cyanobacteria are a common sight in the Vaal Dam (Figure 4.1). Therefore, the monitoring and management of this water body with regards to cyanobacteria and cyanotoxins are very important to the health of millions of people.



Figure 4.1 Cyanobacteria scums observed in the Vaal Dam at Oranjeville on 20 September 2005.

In this chapter physical, chemical and biological water quality data of samples collected from 2000 to 2012 in the Vaal Dam, are used to describe the cyanobacterial assemblages and the associated nutrients in the Vaal Dam. This is an important step for the later evaluation of the

forecasting models (Chapter 5 and Chapter 6), since these are the concepts that the models are based upon. Monitoring of the dam and the water quality was performed by the Analytical Services Department of Rand Water as described in Chapter 3. Samples were taken at the dam wall once a month during the study period. Multivariate statistical analyses for the overview were performed using Canoco version 4.5 and Statistica version 10 was used for the descriptive statistics as described in Chapter 3.

4.2 Results and Discussion

A summary of the basic descriptive data (minimum, maximum, average and standard deviation) of the environmental variables measured monthly in the Vaal Dam at the dam wall from 2000 - 2012 is given in Table 4.1. This summary includes the data that was used for 1) developing the models (2000 - 2009) and the laboratory-generated data used for testing the models (2010 - 2012) as described in Chapters 5 and 6. Spearman Rank order correlations were also determined and included in Addendum A. Most of the variables have been measured over the thirteen year period ($n = 165$), with the exception of one or two variables (e.g. the laboratory from where the results were obtained only started monitoring microcystin concentration in March 2002).

Table 4.1 The descriptive statistical values of the measured variables in the Vaal Dam (2000 - 2012).

	Minimum	Maximum	Average	Standard Deviation	n (number of samples included)
Chlorophyll- <i>a</i> ($\mu\text{g/L}$)	<2	194	12.0	19.30	164
Dissolved Oxygen (mg/L)	3.52	22.5	7.83	2.24	156
pH	5.83	13.52	7.88	0.91	156
Secchi disk depth (cm)	12	128	31.6	12.41	164
Water Temperature ($^{\circ}\text{C}$)	8.82	26	17.50	4.43	163
Turbidity (NTU)	8.57	141	57.22	26	163
NH_4^+ (mg/L)	<0.002	7.31	0.08	0.57	165
PO_4^{3-} (mg/L)	<0.001	3.06	0.05	0.24	165
Total Phosphate (mg/L)	<0.001	3.09	0.12	0.28	165
NO_2^- (mg/L)	<0.001	1.17	0.05	0.13	163
NO_3^- (mg/L)	0.01	7.12	0.34	0.67	165
Conductivity (mS/m)	13.90	318.50	21.69	23.73	163
Alkalinity (mg/L CaCO_3)	27.00	253.70	72.78	18.15	165
SO_4^{2-} (mg/L)	9.70	68.59	15.78	7.90	149

Table 4.1 (cont.) The descriptive statistical values of the measured variables in the Vaal Dam (2000 - 2012).

	Minimum	Maximum	Average	Standard Deviation	n (number of samples included)
Si (mg/L)	0.20	17.84	5.25	2.76	162
Hardness (mg/L CaCO ₃)	39.78	290.21	69.18	23.43	161
COD (mg/L)	0.01	34	14.02	5.51	151
Br (mg/L)	<0.001	8.6	0.23	0.76	138
Cl (mg/L)	<0.001	88	6.64	7.63	150
F (mg/L)	0.029	1	0.24	0.12	150
Al (mg/L)	<0.001	10.44	0.94	1.39	162
Ca (mg/L)	7.42095	73.1	15.22	6.17	162
B (mg/L)	<0.001	4.60	0.04	0.36	162
Cu (mg/L)	<0.001	0.06	0.01	0.01	162
Fe (mg/L)	<0.001	3.36	0.53	0.61	162
K (mg/L)	<0.001	10.02	2.27	1.04	162
Mg (mg/L)	4.18224	26.18	7.67	2.32	162
Mn (mg/L)	<0.001	0.14	0.02	0.02	162
Mo (mg/L)	<0.001	0.04	0.00	0.01	162
Na (mg/L)	3.077	55.9	9.76	4.47	162
Ni (mg/L)	<0.001	0.12	0.01	0.02	162
Pb (mg/L)	<0.001	6.94	0.21	0.88	162
S (mg/L)	<0.001	53.7	5.47	4.44	161
V (mg/L)	<0.001	0.09	0.01	0.01	162
Zn (mg/L)	<0.001	0.10	0.01	0.01	163
Cyanobacteria (cells/mL)	<36	194 921	7 950	21 884	165
<i>Anabaena</i> sp. (cells/mL)	<36	66 263	2 456	7 098	165
<i>Microcystis</i> sp. (cells/mL)	<36	194 921	5 486	19 202	165
Microcystin (µg/L)	<0.18	5.0	0.34	0.70	140

During the study period, the *in situ* measured pH in the Vaal Dam ranged from 5.8 to 13.5 with an average of 7.88. The average chlorophyll-*a* during the study period is 12 µg/L with an average Secchi disk depth of 31.6 cm and an average water temperature of 17.5 °C. The ions Mg, Ca and Na occur in significant concentrations and therefore hardness also appear to be an important physico-chemical characteristic. Nutrients such as NH₄⁺ (0.08 mg/L) and NO₃⁻ (0.34 mg/L) are lower when compared to that in the hypertrophic dam, Hartbeespoort, (total N 2.7 mg/L and Total P 0.112 mg/L – Van Ginkel, 2008) but Total P occur in a similar range (0.12 mg/L) than that of the Hartbeespoort Dam (0.112 mg/L – Van Ginkel, 2008).

The percentage composition of the algae groups in the Vaal Dam (Figure 4.2), shows the consecutive cyanobacteria (Cyanophyceae) and diatom (Bacillariophyceae) succession in the Vaal Dam, where cyanobacteria (indicated in light blue) are mostly dominant in the warmer months (summer and autumn), while diatoms (indicated in orange) are mostly dominant in winter. The cyanobacteria were mostly comprised of *Microcystis* (69%) and to a lesser extent *Anabaena* spp. (31%) and the diatoms were mostly comprised of the filamentous *Aulacoseira* (52%) and the centric *Cyclotella* (38%), and to a lesser extent the pennates, *Navicula* and *Nitzschia* spp. (8%), as well as a few other pennates that occurred infrequently. During the time when the green algae (Chlorophyceae – indicated in green) were dominant, it was mostly due to *Chlamydomonas* sp. which proved to be the dominant Chlorophyte for 78% of the time. For the remainder of the green algae dominance, the dominant genera varied between *Actinastrum*, *Oocystis* and *Micractinium*. The Euglenophyceae (indicated in pink) occurring in the Vaal Dam was mostly *Trachelomonas* sp., while *Cryptomonas major* was the only member of the Cryptophyceae (indicated in navy blue). The Dinophyceae (indicated in yellow) occurring in the Vaal Dam mostly comprised of *Ceratium hirundinella* and to a lesser extent *Peridinium* and *Sphaerodinium* spp. During 6 sampling occasions from January 2009 to December 2011 the grab samples taken from the Vaal Dam at the dam wall (as indicated in Figure 4.2) contained algae below the detection limit of the method (no colour indication). It is important to note that no natural ecosystem is uniform and that the results (especially of the algae) are specific to the time and locality where the samples were taken, and may differ significantly within a short time and space. Therefore, reporting algae “below the detection limit” does not imply that there were no algae in the Vaal Dam during those times.

By grouping the data into the four different seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring – Figure 4.3) it was evident that the highest chlorophyll-*a* concentration occurred during summer and autumn, although the averages between all four seasons did not vary by more than 10 µg/L.

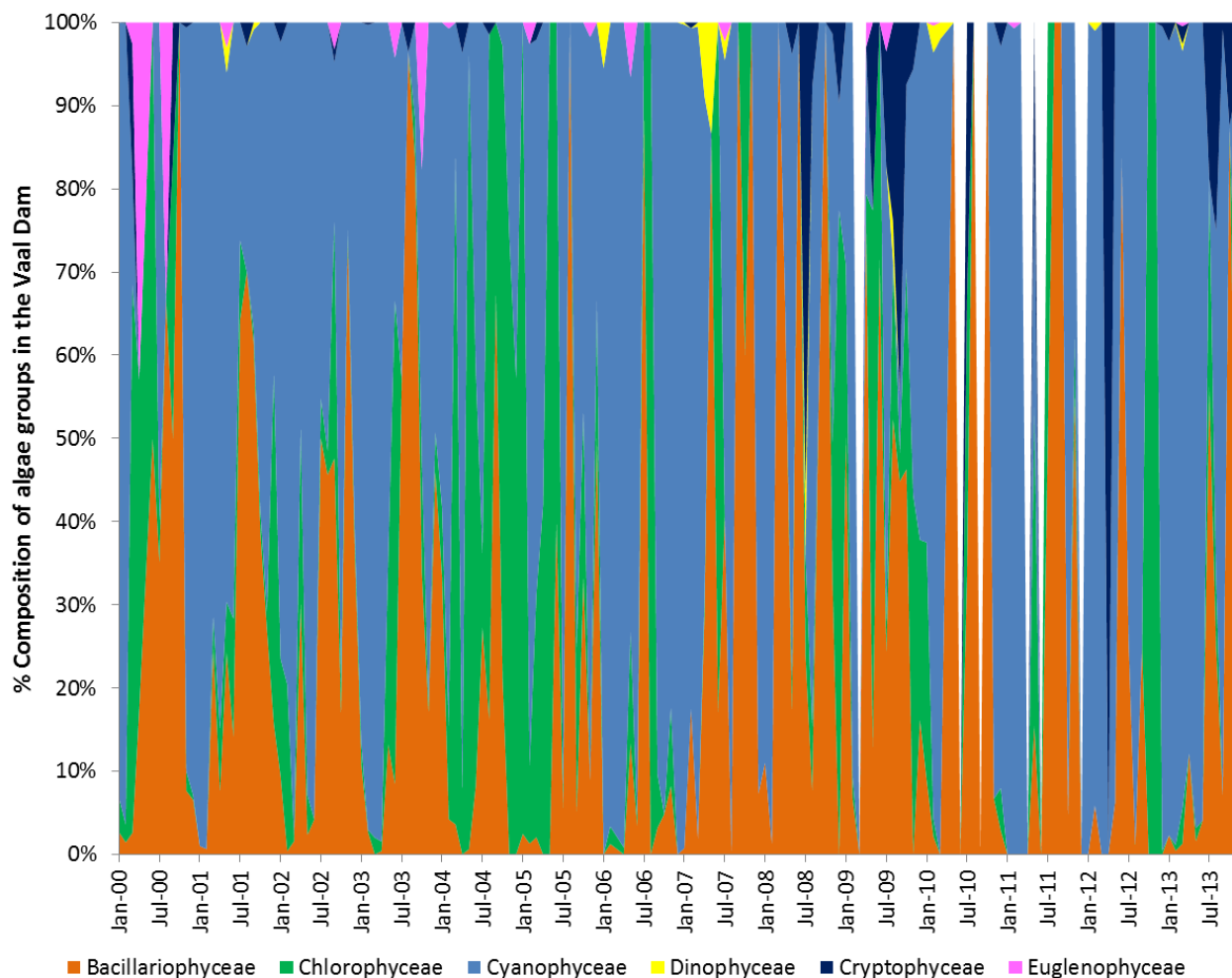


Figure 4.2 Percentage composition of the algae groups occurring in the Vaal Dam from 2000 – 2013.

Even though the temperature ranges during autumn and spring are almost identical, the higher algal inoculum carried over from summer to autumn (when compared to the lower carry over from winter to spring) probably resulted in higher chlorophyll-*a* concentrations during autumn than during spring (Figure 4.3). The maximum chlorophyll-*a* concentration (194 µg/L) was detected during the autumn of 2004.

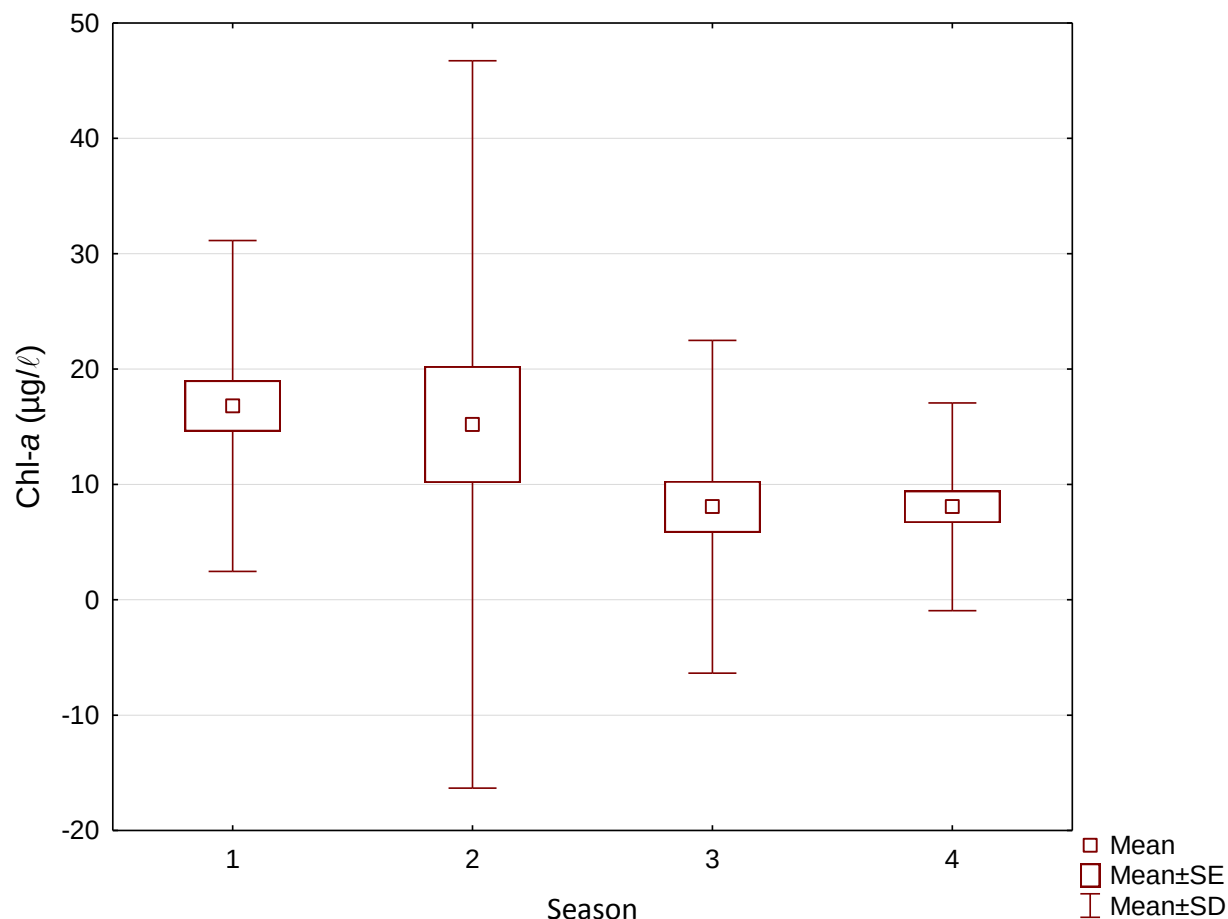


Figure 4.3 Box plot of chlorophyll-*a* (µg/L) in the Vaal Dam for the period 2000 – 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring).

The total cyanobacterial cell concentration (Figure 4.4) in the Vaal Dam during the study period, showed the highest concentration during summer (average $\pm 20\,000$ cells/mL), followed by autumn (average ± 8000 cells/L), spring (average ± 3000 cells/mL) and the lowest concentration during winter. Cyanobacteria are well known to occur during periods of high water temperature (Harding and Paxton, 2001; García Nieto, 2013) and the Vaal Dam is no exception. The fact that the cyanobacterial cell concentrations are higher during autumn than during spring is most probably due to a higher carry-over from summer to autumn than from winter to spring. The higher cyanobacterial cell concentrations during autumn than during spring is also observed in other impoundments known for high cyanobacterial cell concentrations such as Hartbeespoort Dam (Van Ginkel and Silberbauer, 2007). *Microcystis* sp. appeared in much higher cell numbers than *Anabaena* sp. and reached a maximum of 194 921 cells/mL during the past 13 years.

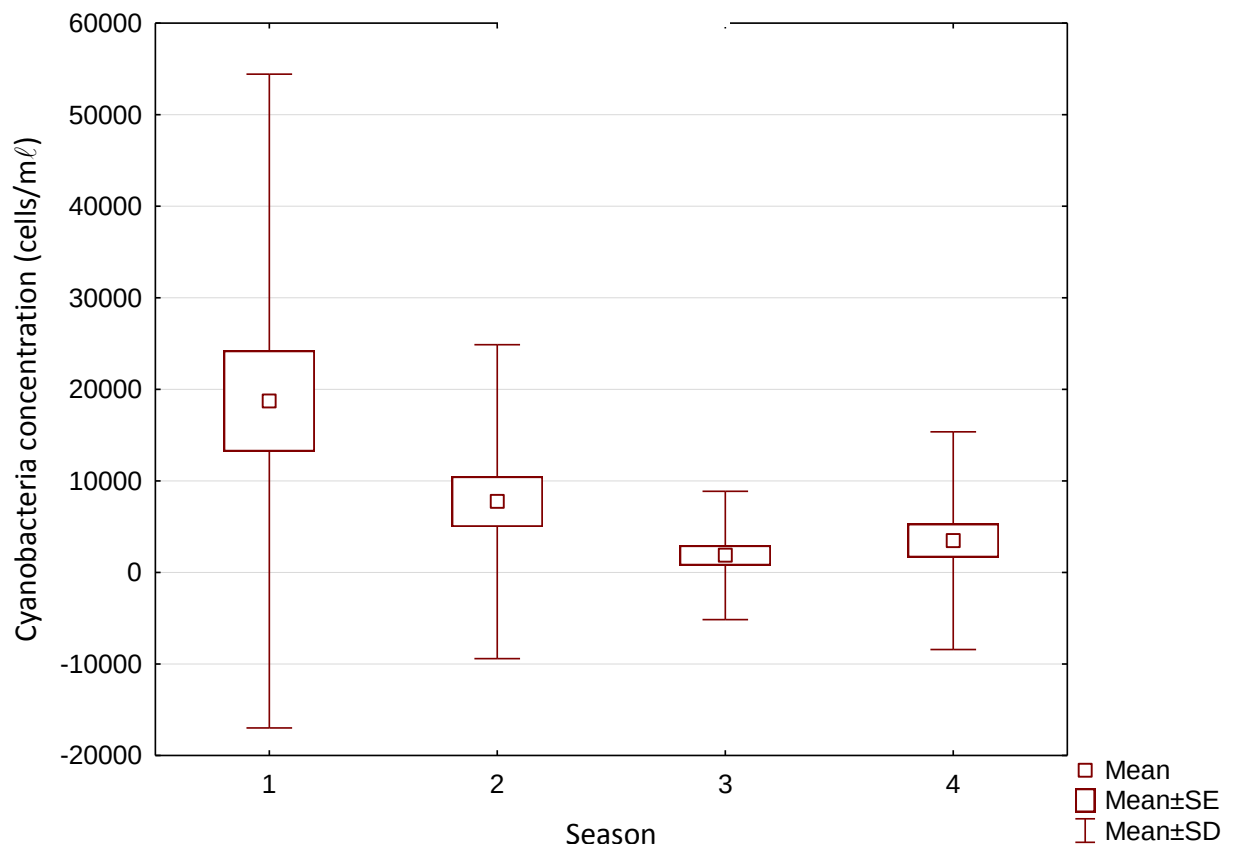


Figure 4.4 Box plot of total cyanobacterial cell concentration (cells/mL) in the Vaal Dam for the period 2000 - 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring).

The microcystin concentration (Figure 4.5) in the Vaal Dam during the study period showed the highest concentrations during summer, followed by autumn, spring and the lowest concentration during winter. This confirms that cyanobacteria, known for the production of the cyanotoxin microcystin (such as *Anabaena* sp. and *Microcystis* sp. – WHO, 2011), mostly occur during the summer and autumn months. The succession of cyanobacteria in the Vaal Dam always begin with *Anabaena* sp. during spring and summer, followed by *Microcystis* sp. during summer and autumn². The higher microcystin concentration in autumn when compared to spring seems to indicate that microcystin production in the Vaal Dam is closer related to the presence of *Microcystis* sp. rather than *Anabaena* sp. From literature it is also evident that the production of microcystin is very closely related to water temperature, where microcystin is mostly produced where water temperatures are >23 °C (Conradie and Barnard, 2012), which is

² Personal observation during a nine year period in the Vaal Dam.

most probably the reason why the microcystin concentration during summer is approximately double the concentration detected during autumn (Figure 4.5).

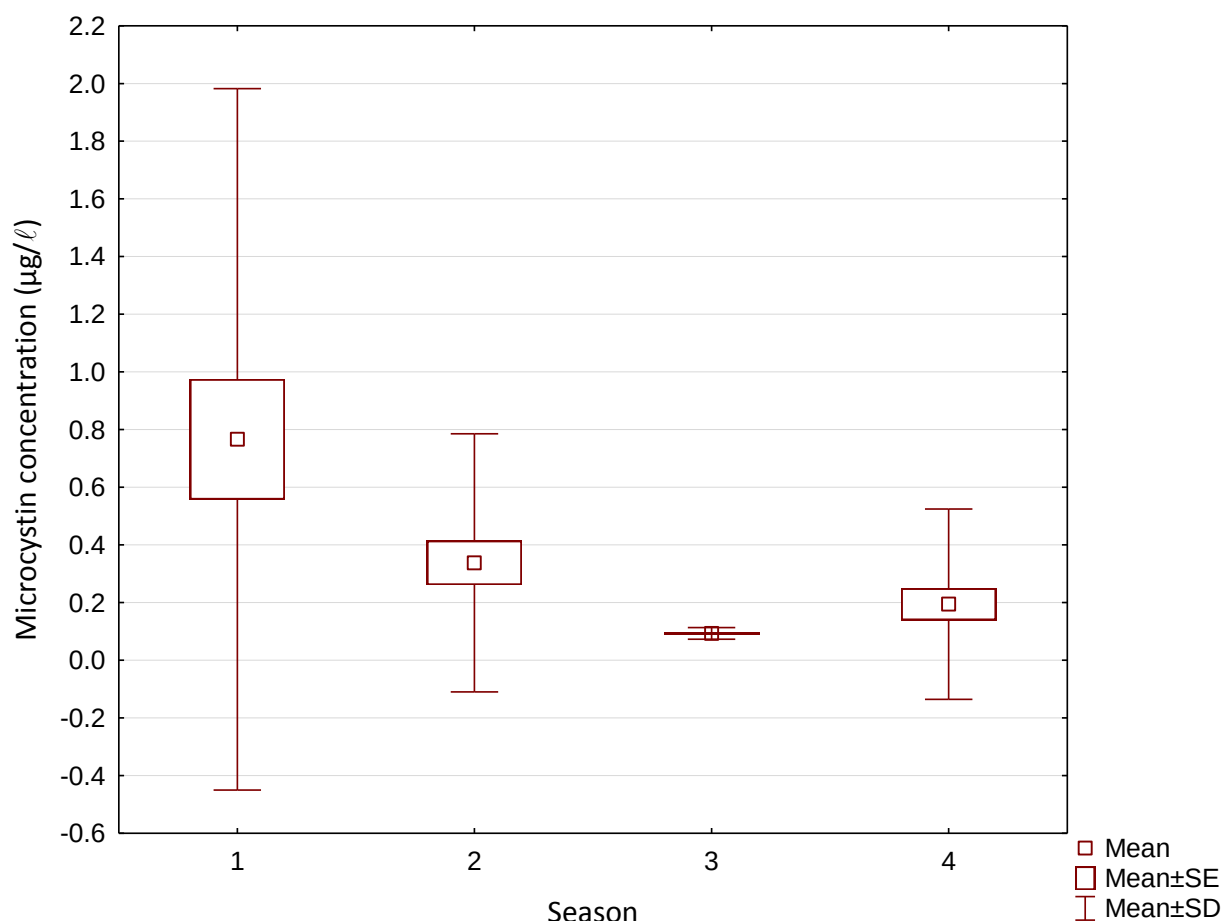


Figure 4.5 Box plot of microcystin (µg/L) in the Vaal Dam for the period 2000 - 2012, separated into seasons (1 = summer, 2 = autumn, 3 = winter and 4 = spring).

A scatter plot for *Anabaena* sp. and *Microcystis* sp. with associated trend lines in the Vaal Dam (Figure 4.6) as well as the scatter plot for microcystin and temperature with associated trend lines (Figure 4.7), indicate that there were incidences of sporadic high concentrations of cyanobacterial cells together with sporadic incidences of high microcystin concentrations during the study period. Therefore the development of forecasting models able to predict these incidences in the source water to the Zuikerbosch DWTW (Figure 3.3), will significantly contribute to the production of safe and aesthetically acceptable drinking water to consumers. The ability to predict high concentrations of cyanobacterial cells and high concentrations of microcystin may also aid in saving treatment costs when only sporadic incidences call for advanced treatment options (like the dosing of PAC) to be applied.

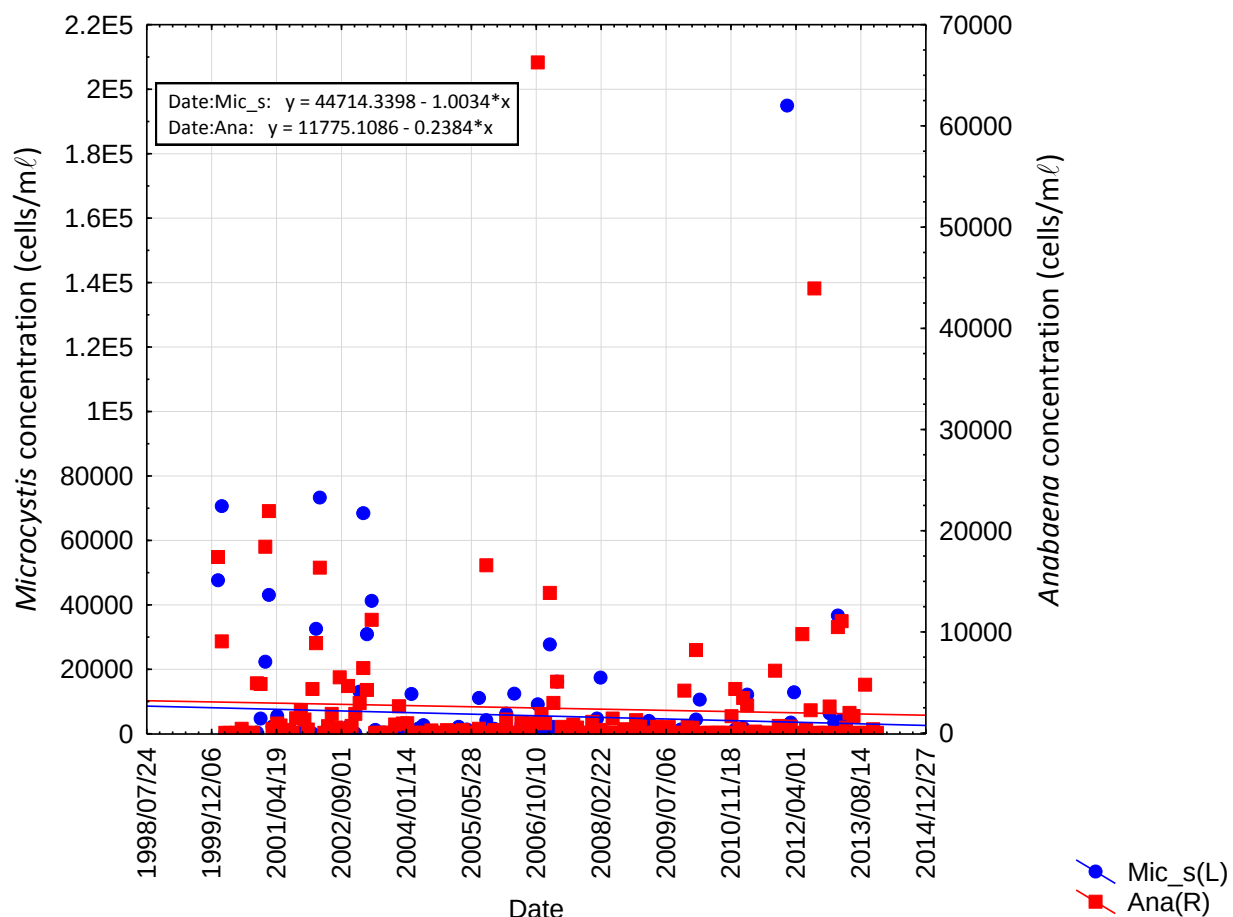


Figure 4.6 Scatter plot with associated trend lines for *Microcystis* (Mic_s) and *Anabaena* (Ana) in cells/mL, in the Vaal Dam from 2000 - 2013.

The trend lines associated with *Microcystis* sp. and *Anabaena* sp. (Figure 4.6), as well as the trend line associated with microcystin (Figure 4.7) don't show a significant temporal change during the study period. This suggests that anthropogenic activity in and around the Vaal Dam has not changed significantly during the past thirteen years. Most of the Vaal Dam catchment area is agricultural land and some coal mining activities occur. These anthropogenic activities are not anticipated to change much in the foreseeable future. A study done on the soils of the Vaal Dam catchment in 1993, indicated that it has a twofold greater magnitude of soluble salts (and sulphate) when compared to soils from Natal (Fey and Guy, 1993), which may be due to mining activity in the catchment. The Vaal Dam is also extensively used for recreational activities such as boating, fishing and water skiing which may increase rather than decrease in future.

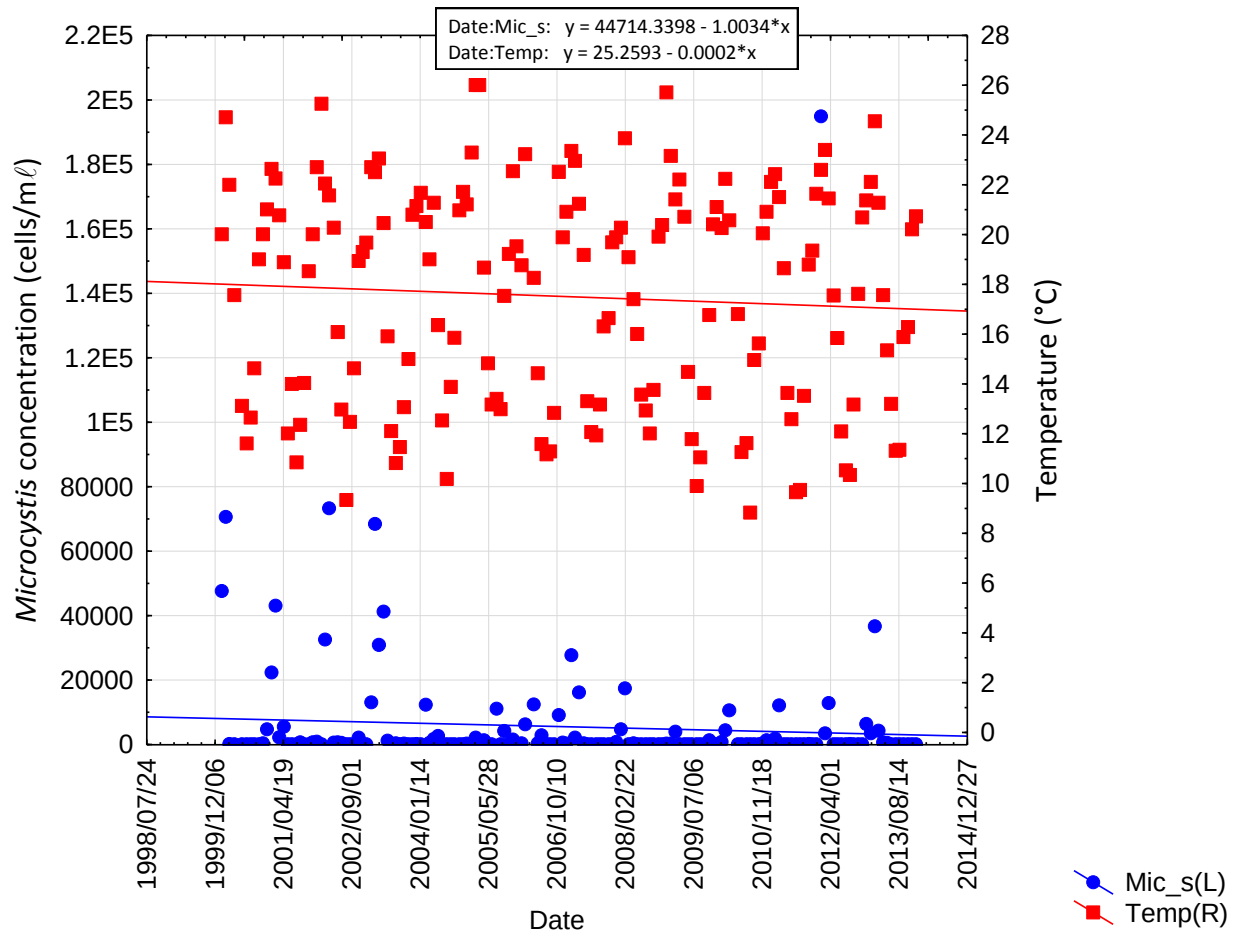


Figure 4.7 Scatter plot with associated trend lines for *Microcystis* sp. (Mic_s) in cells/mL and water temperature (Temp) in °C, in the Vaal Dam from 2000 - 2013.

According to the PCA performed in Statistica (Table 4.2), 39 different principal components or factors could be extracted from the data set containing 41 variables. According to the scatter plot the first 6 factors cumulatively explain 55% of the variation in the data.

According to Hair et al., (2009), factor loadings (Table 4.2) represent the correlation between the variable and its factor and therefore variables with the highest factor loadings are more important to that factor. Factor loadings of $\pm 0.30 - \pm 0.40$ are considered to meet minimal level for interpretation of structure, loadings of ± 0.50 or greater are considered significant and loadings exceeding ± 0.70 are considered well-defined and the goal of any factor analysis. However, in a database with a sample size of approximately 150 (like in this case) the factor loading needs to be ± 0.45 to be significant (Hair et al., 2009).

Table 4.2 Factor-variable correlations (factor loadings) of Vaal Dam variables (2000 - 2012).

Variable	Factor-variable correlations (factor loadings), based on correlations					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Chlorophyll- <i>a</i>	-0.040283	-0.145854	-0.309129	0.249602	-0.108374	-0.114305
DO	-0.054146	-0.036626	0.128968	-0.002415	0.029819	0.304259*
pH	-0.039388	-0.006727	-0.330769	0.027258	-0.277459	-0.010176
Secchi disk depth	-0.226920	-0.306745	0.401001	0.230316	0.199705	-0.083116
Temperature	0.113011	-0.324156	-0.188759	0.278240	-0.174753	-0.121602
Turbidity	0.451804	0.337502	-0.408113*	-0.240680	-0.223574	-0.105227
NH ₄ ⁺	-0.252162	0.652625*	0.070466	0.618230*	0.203218	0.098297
PO ₄ ³⁻	-0.233654	0.670757*	0.042622	0.596713*	0.199115	0.088866
TP	-0.169183	0.688734*	0.004126	0.502087*	0.219635	0.147106
NO ₂ ⁻	-0.321895	0.604292*	-0.014773	0.474700	0.151892	-0.047320
NO ₃ ⁻	-0.349045	0.244756	0.126721	0.197840	-0.678711*	0.010425
Conductivity	-0.086918	-0.048234	0.162600	0.137709	0.030323	-0.187742
Alkalinity	-0.320811	-0.204832	0.259497	0.252818	-0.680178	0.061592
SO ₄ ²⁻	-0.839307*	0.030818	-0.123068	-0.262151	-0.126531	-0.122744
Si	0.491144	0.469738	-0.562045*	-0.291039	-0.171798	0.041691
Total Si	0.491144	0.469738	-0.562045*	-0.291039	-0.171798	0.041691
Hardness	-0.905000*	-0.022498	-0.166750	-0.285680	-0.005781	0.015998
COD	0.030351	0.242559	-0.200683	0.247362	-0.013944	-0.171336
Br	-0.181138	0.021460	0.262480	0.277410	-0.783963	0.042255
Cl	-0.873708*	0.203707	-0.144334	-0.145386	-0.022545	-0.185215
F	-0.060576	0.186085	0.107864	0.316645	-0.641939*	0.185440
Al	0.340487	0.357574	-0.339395	-0.265256	-0.127196	-0.014881
Ca	-0.906723*	-0.010393	-0.176676	-0.298102	-0.012268	-0.006746
B	-0.036518	-0.035362	0.038909	-0.056069	0.013091	0.033938
Cu	-0.211851	-0.155746	-0.017388	-0.164531	0.000867	0.697079*
Fe	0.435631	0.436085	-0.429108	-0.276172	-0.119370	-0.079212
K	-0.569184	0.392059	-0.218986	-0.001980	0.150088	-0.135163
Mg	-0.862708*	-0.040003	-0.144281	-0.254187	0.004381	0.050038
Mn	-0.171489	0.137707	-0.165526	0.133313	0.148336	0.373425*
Mo	-0.232022	-0.098754	-0.099355	-0.070299	0.058324	0.298263
Na	-0.924800*	0.138737	-0.128522	-0.133057	0.052274	-0.075254
Ni	-0.117518	-0.109852	-0.185803	0.007905	-0.006269	0.008442
Pb	0.008981	-0.066002	0.088839	-0.109431	0.181629	0.349662*
S	-0.874853*	-0.006377	-0.203867	-0.296772	-0.021778	-0.093547
V	0.049940	0.118119	-0.054887	-0.143804	-0.138481	0.357007*
Zn	-0.042100	0.026051	-0.130343	-0.179638	-0.010693	0.419060*
Cyanobacteria	-0.085430	-0.461130	-0.686392*	0.444291	0.050998	0.082165
Total Algae	-0.096537	-0.465216	-0.700105*	0.459801	0.031590	0.051527
<i>Anabaena</i> sp.	-0.079637	-0.201044	-0.389997*	0.101141	0.027247	0.052779
<i>Microcystis</i> sp.	-0.065575	-0.438155	-0.617102*	0.456720	0.046533	0.071490
Microcystin	0.015698	-0.343083	-0.344341*	0.428459	-0.003567	-0.132700

* Variables indicated in red display the highest correlation to the factor determined by the PCA. Variables displayed in green may not be regarded as significant by Hair et al., (2009), but may aid in the identification of the factor.

The first factor accounted for 18.4% of the variance and can be termed “water softness” since it is negatively correlated with hardness as well as with all the components associated with it, namely Ca, Mg, SO_4^{2-} , S, Na and Cl. The second factor can be interpreted as “nutrients” and is represented by the positive relationship with algal nutrients (NH_4^+ , Total phosphate, PO_4^{3-} and NO_2^-). It accounted for 10% of the variance and is regarded to play a significant role in the development and proliferation of algae (including cyanobacteria). The third factor can be termed “light penetration” and is represented by the positive relationship with Secchi disk depth and the negative relationship with Si, total silica and turbidity as well as the algae, cyanobacteria and microcystin. This factor (which accounts for 9% of the variance) will have a positive influence on the photosynthetic rate of algae and might reduce the competitive advantage that cyanobacteria have over other algae when regulating their buoyancy for optimum light availability (Wetzel, 2001). The fourth factor is also represented by certain nutrients (NH_4^+ , total phosphate and PO_4^{3-}), but since nutrients were already captured in a factor, this factor was named “algae” since it is represented by total algae, cyanobacteria, *Microcystis* sp. and microcystin. The fifth factor is termed “acidity” since it is represented by a negative relationship to alkalinity, NO_3^- , Br and F. Factor 6 is termed “metals” since it is represented by a variety of the metals (Cu, Mn, Pb, V and Zn) together with dissolved oxygen. Factors 5 and 6 may not display a general positive or negative effect on algae, although different variables within the factors may affect certain algae and cyanobacteria, such as NO_3^- in factor 5, which would increase algal growth.

A canonical correspondence analysis (CCA – Figure 4.8) was performed in Canoco version 4.5 where the phytoplankton data was plotted in comparison to the environmental data. Note that inflating environmental variables were removed before the CCA was performed in the following way: NO_3^- was chosen to represent the nitrogen nutrient factor (therefore NO_2^- and NH_4^+ were removed); Conductivity was chosen to represent the contributing ions (therefore SO_4^{2-} and Na were removed); Hardness was chosen to represent Mg and Ca; and Si was chosen to represent the metals (therefore Fe and Al were removed). The phytoplankton species that occurred <5 times during the study period were not included in the analysis and all centric and pennate diatoms were combined as to reduce the number of species that needed to be removed. The natural log (Ln) of the phytoplankton data was used in the analysis. The Eigenvalues for the first 4 axes of the CCA is displayed in Table 4.3. The Monte Carlo Permutation test on all canonical axes (number of permutations = 499) showed a P-value of 0.022 and an F-ratio of 1.541.

Table 4.3 Eigen values for the CCA done on all environmental variables in the Vaal Dam from 2000 – 2013 in Canoco.

Axes	1	2	3	4
Species-environment correlations	0.687	0.613	0.498	0.465
Cumulative percentage variance of species data	8.4	13.0	16.2	18.3
Cumulative percentage variance of species-environment relation	38.2	59.5	74.0	83.7
Sum of Eigen values = 1.454;				
Sum of all canonical Eigen values = 0.318				

The factors identified in the PCA done in Statistica (Table 4.2) correlated with the variables closely associated with one another in Figure 4.8. The first factor (“water softness”) negatively relates to the variables ordinated in the first quadrant of Figure 4.8 (Hardness). Factor 2 (“nutrients”) positively relate to the variables ordinated in the fourth quadrant of Figure 4.8 (PO_4^{3-} and NO_3^-). Factor 3 (“light penetration”) relates positively to Secchi disk depth (quadrant 2) and negatively with Si and turbidity (quadrant 3). Factor 4 (“algae”) also relates positively with the nutrients (quadrant 4) and chlorophyll-*a* (quadrant 3). Factor 5 (“acidity”) negatively relates to alkalinity (quadrant 1) and NO_3^- (quadrant 4). Factor 6 (“metals”) was not included in the input data used for the CCA, since the metals identified to represent factor 6 (Cu, Mn, Pb, V and Zn) do not have a significant influence on the algae and cyanobacteria at such low concentrations as detected in the Vaal Dam (Table 4.1).

The CCA (Figure 4.8), which accounts for 61% in the variation of the species vs. environmental variables relationship, shows that the cyanobacteria namely *Microcystis* sp. and *Anabaena* sp. closely associate with water temperature, pH and COD (chemical oxygen demand). It is important to note that when the species data is included in the multivariable analysis, temperature seem to become a very important variable (refer to the length of the temperature arrow in Figure 4.8). This is most probably due to the important role that temperature plays in the occurrence of cyanobacteria (during periods of high temperatures) and diatoms (during the periods of low temperatures – Wetzel, 2001).

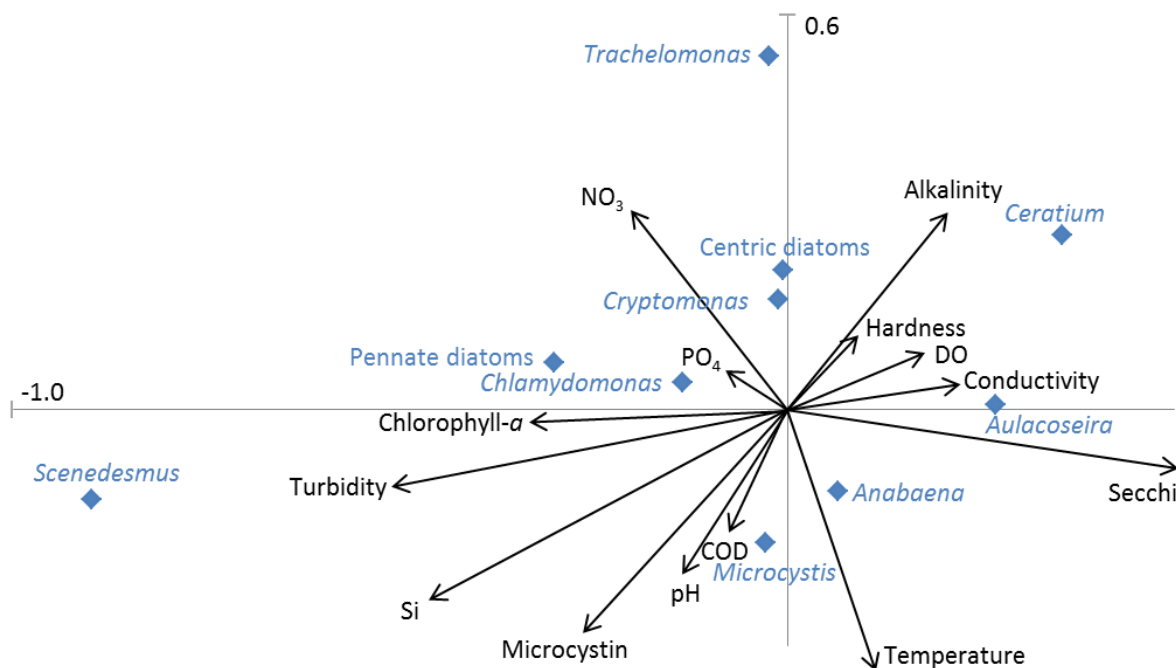


Figure 4.8 Canonical correspondence analysis (CCA) of the environmental variables and natural log of the phytoplankton data in the Vaal Dam for the period 2000 - 2013.

During periods of high *Anabaena* sp. and *Microcystis* sp. concentrations in the Vaal Dam, the utilisation of CO₂ during photosynthesis results in an increase in pH. The association of these cyanobacteria with high chemical oxygen demand was most probably due to decomposition of algal material during and after the bloom. *Anabaena* sp. and *Microcystis* sp. also correlate negatively with the nutrients, which indicate that during summer times when these cyanobacteria are abundant in the Vaal Dam, their utilisation of nutrients has a significant impact on the concentration of the nutrients. García Nieto and co-workers (2013) found that high turbidity favours the occurrence of cyanobacteria over other algal groups since cyanobacteria are able to manipulate their position in the water column due to gas vacuole formation (Wetzel, 2001). This was not as apparent in the Vaal Dam (since turbidity forms a $\pm 90^\circ$ angle with *Anabaena* sp. and *Microcystis* sp.), most probably because of the inherently turbid nature of the Vaal Dam where cyanobacteria were favoured continuously whenever the other environmental variables (e.g. water temperature) were conducive for cyanobacterial development. Reynolds (2006) found that high alkalinity favours cyanobacterial growth, this too was not found in the Vaal Dam, most probably because the average alkalinity of the Vaal Dam was ± 72 mg/L CaCO₃.

Microcystin also correlated more closely with *Microcystis* sp. rather than *Anabaena* sp., indicating that in the Vaal Dam, microcystin is mostly produced by *Microcystis* sp. rather than *Anabaena* sp. This was also observed in Figure 4.5, when succession of *Anabaena* sp. during spring and summer followed by *Microcystis* sp. in summer and autumn, with higher microcystin concentrations detected during autumn, is taken into account.

The diatoms (centric diatoms, pennate diatoms and *Aulacoseira* sp.) correlate negatively with temperature which is well-known to occur all over the world (Wetzel, 2001; García Nieto, 2013) and the Vaal Dam is no exception. *Aulacoseira* sp. and centric diatoms correlate negatively with Si, which indicate that during periods of diatom abundance, silica is utilised in the Vaal Dam to be incorporated into their siliceous frustules. Pennate diatoms do not follow the same negative correlation with silica. This indicates that pennate diatoms (on average 48 cells/mL in the 13 year period) do not have a significant negative effect on the silica concentration (as opposed to *Aulacoseira* sp. and centric diatoms) that occurred in higher abundance (average 247 cells/mL and 192 cells/mL respectively in the 13 year period) and with higher frequency occurring in the Vaal Dam. The positive correlation of *Chlamydomonas* sp. and *Scenedesmus* sp. with chlorophyll-*a* (Chla) indicates that these chlorophyceae contribute significantly to the chlorophyll content in the Vaal Dam. *Ceratium* sp. although known to contribute significantly to chlorophyll-*a* concentration in other water bodies (Ewerts et al., 2013), seem to have a negative relationship with chlorophyll-*a* concentration in the Vaal Dam (Figure 4.8). This dinoflagellate, do not occur frequently or in high abundance in the Vaal Dam and therefore its contribution to chlorophyll-*a* concentration in the Vaal Dam is not observed.

4.3 Conclusions

During the thirteen year study period, the Vaal Dam has been relatively stable regarding water quality. This is most probably due to very little change in the anthropogenic activity in the Vaal Dam catchment during this period. During the whole study period of thirteen years, it was clear that although the dam was not plagued by cyanobacterial blooms, incidences of high cyanobacterial cell concentrations (*Microcystis* sp. and *Anabaena* sp.) and microcystin concentrations did occur in the Vaal Dam. These occurrences were mostly observed during summer when temperatures were favourable for the development of cyanobacterial blooms and associated organic compounds. The cyanotoxin, microcystin, is produced during summer months and is mostly associated with the occurrence of *Microcystis* sp. rather than *Anabaena* sp. It is therefore evident that forecasting models able to predict incidences of high

cyanobacterial cell concentrations and high cyanotoxin concentrations in the Vaal Dam will significantly aid managers at the Zuikerbosch DWTW, in anticipating water purification problems associated with cyanobacteria. This will also contribute to the production of safe and aesthetically acceptable drinking water to consumers, with possible cost saving implications.

From the overview discussed in this chapter, it seems that variables such as water temperature, light penetration (as indicated by Secchi disk depth), nutrients (e.g. NO_3^- and PO_4^{3-}) and pH may play a significant role in the prediction of cyanobacteria such as *Anabaena* sp. and *Microcystis* sp. as well as the prediction of microcystin. Models for the prediction of these variables will be discussed in Chapters 5 and 6, while the application thereof in the cyanobacterial incident management protocols as part of the Water Safety Plan (WSP) will be discussed in Chapter 7.

CHAPTER 5

USING HYBRID EVOLUTIONARY ALGORITHMS FOR THE DEVELOPMENT OF COMPOUND RULE-BASED FORECASTING MODELS FOR CYANOBACTERIA FROM ALL AVAILABLE HISTORICAL LABORATORY GENERATED DATA

5.1 Introduction

In this chapter, the models developed from all available laboratory generated data for predicting different cyanobacteria genera, as well as two of their associated organic compounds namely the cyanotoxin “microcystin” as well as the taste and odour compound “geosmin” will be discussed. Ten years of laboratory generated data (2000 – 2009) in the Vaal Dam supplying source water to the Zuikerbosch drinking water treatment works (DWTW) was used for the development of these models. Four different prediction-time scenarios were investigated namely *real-time* (same day – indicated by ^[0]), *7 days in advance* (indicated by ^[7]), *14 days in advance* (indicated by ^[14]) and *21 days in advance* (indicated by ^[21]). Fifty rule-based models were developed for each prediction time, at 1) the Vaal Dam only 2) canal only and 3) Vaal Dam and canal combined, of which only the best model from all three sites (those with the highest R-squared in comparison to the measured data) were described in this chapter (refer to page 25).

5.2 Models for the prediction of *Anabaena* sp. concentration

Anabaena sp. is a cyanobacterium well-known for the production of cyanotoxins namely microcystin, cylindrospermopsin, anatoxin-a, anatoxin-a(S) and saxitoxin (Du Preez and Van Baalen, 2006; Newcombe et al., 2010). The production and presence of especially microcystin in the source water is very important from a drinking water treatment perspective, since it is one of the variables specified in the SANS 241 : 2011 drinking water standard of South Africa (SANS 241-1, 2011; SANS 241-2, 2011). In the Vaal Dam and canal investigated in this study, *Anabaena* sp. is commonly found in relatively large concentrations during summer (Ewerts et al., 2013) and known for the production of the taste and odour compound geosmin (Swanepoel and Du Preez, 2008) as is the case in many other water bodies in the world (Su et al., 2013). Therefore, models predicting the occurrence of *Anabaena* sp. can be of great

benefit to the management of the Zuikerbosch DWTW to prepare for incidents of high *Anabaena* sp. and potentially high geosmin or cyanotoxin concentrations in the source water.

5.2.1 *Anabaena*^[0] prediction (in real-time)

The IF criterium of the model (Figure 5.1) is determined by the initial cyanobacteria inoculum.

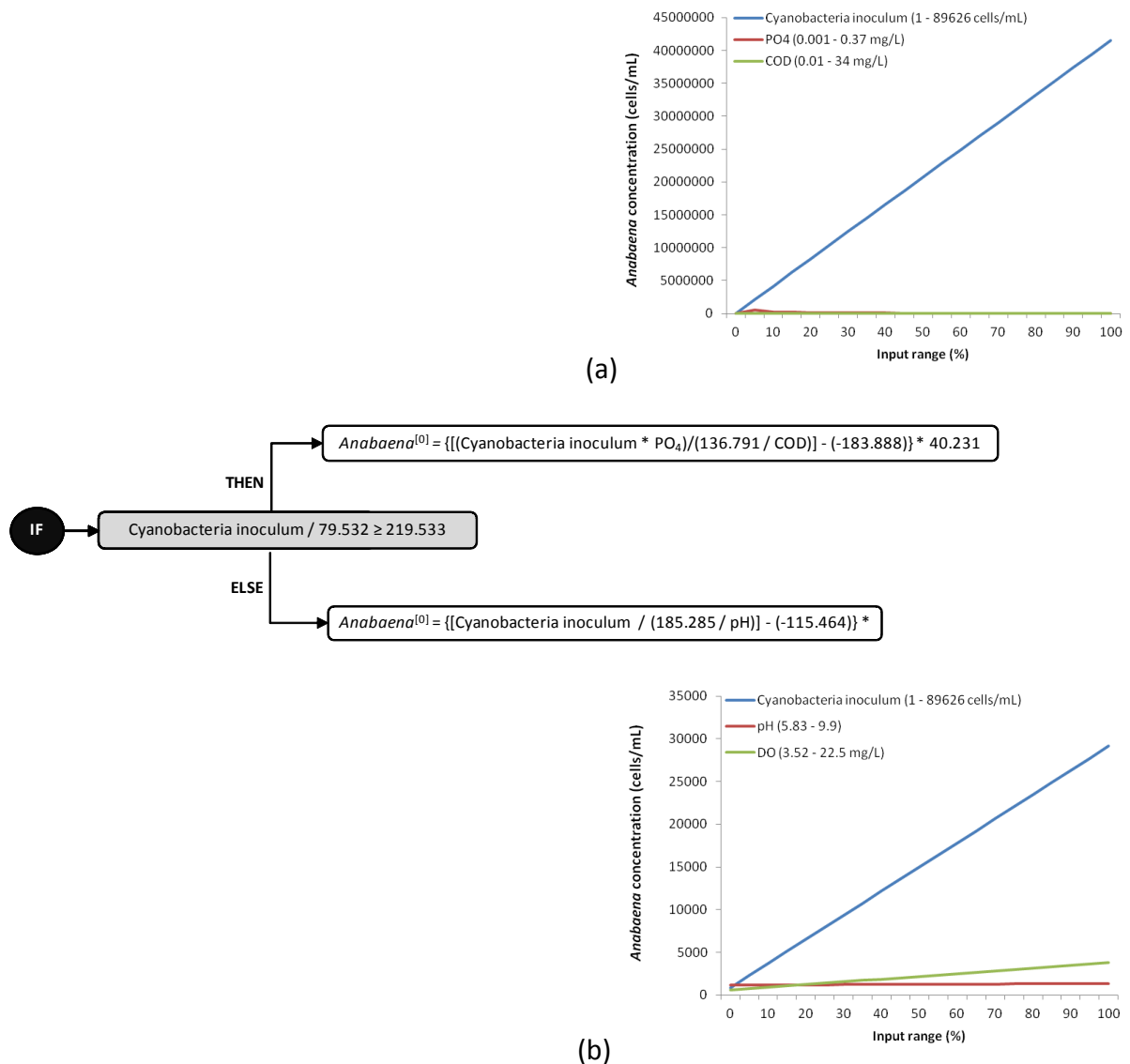


Figure 5.1 Forecasting model and associated sensitivity analyses of real-time *Anabaena*^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The THEN branch of the model (Figure 5.1a) represents the high range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 5.1b) represents the low range rule set and shows the greatest sensitivity towards the

initial cyanobacteria inoculum and less so to the DO concentration. The other variables used in the model (PO_4^{3-} and COD in the THEN branch and pH in the ELSE branch) display very little influence on the predicted *Anabaena*^[0] concentration.

The comparison between the measured *Anabaena* sp. concentration and that of the results from the models predicting real-time *Anabaena*^[0] concentration when using the 25% bootstrapped testing dataset (Figure 5.2a), shows a R^2 -value of 0.89 and a RMSE (root mean square error) of 1944 cells/mL. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.79 and a RMSE of 1026 cells/mL (Figure 5.2b), indicating that both the time prediction and the magnitude of increased *Anabaena* sp. concentrations were significantly correlated.

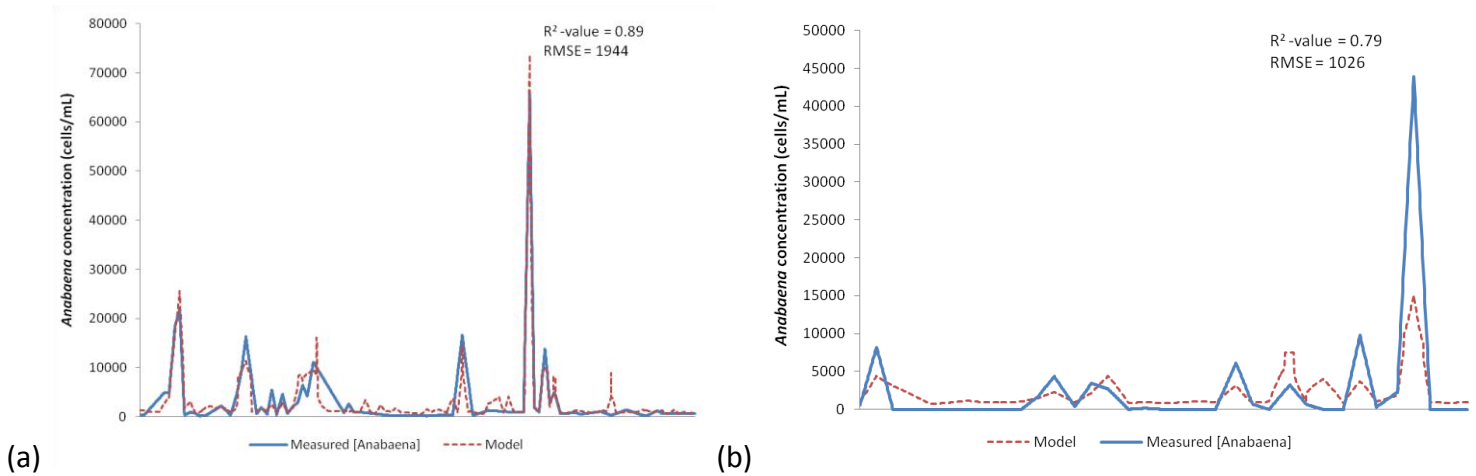


Figure 5.2 a Comparison between the measured *Anabaena* sp. concentration and predicted real-time *Anabaena*^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Anabaena* sp. concentration and predicted real-time *Anabaena*^[0] concentration using 3 years' unseen data from the Vaal Dam.

5.2.2 *Anabaena*^[7] prediction (7 days in advance)

The IF criterium of the model predicting *Anabaena*^[7] concentration (Figure 5.3) is determined by the Fe divided by the NO_2^- concentration, or the turbidity multiplied by the PO_4^{3-} concentration. The THEN branch of the model (Figure 5.3a) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 5.3b) represents the high range rule set and shows the greatest sensitivity

towards the initial cyanobacteria inoculum and less so to the PO_4^{3-} concentration. The other variables used in the model (pH, COD and Mn in the THEN branch and pH, COD, NO_3^- and PO_4^{3-} in the ELSE branch) display very little influence on the predicted *Anabaena*^[7] concentration.

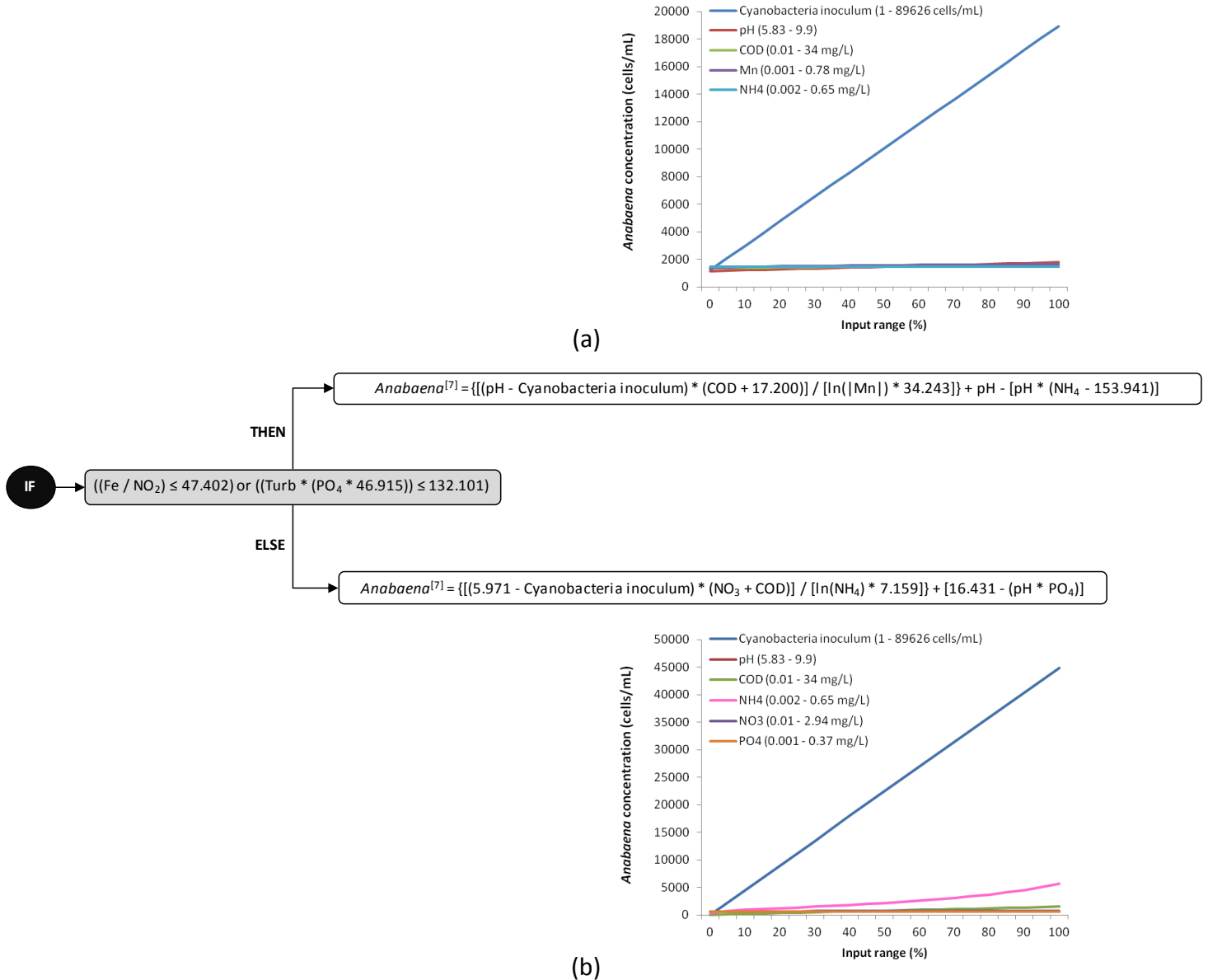


Figure 5.3 Forecasting model and associated sensitivity analyses of *Anabaena*^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Anabaena* sp. concentration and that of the results from the models predicting *Anabaena*^[7] (7 days in advance), when using the 25% boot-strapped

testing dataset (Figure 5.4a), shows a R^2 -value of 0.86 and decreased to 0.35 (Figure 5.4b) when tested with the unseen dataset. The RMSE determined when using the testing dataset was 2127 cells/mL and 1745 cells/mL when using the unseen dataset (Figure 5.4b). Visual inspection of the correlation with the unseen dataset (Figure 5.4b) reveals that the time prediction of increased *Anabaena*^[7] concentration is relatively accurate; however, the magnitude of the concentration is not predicted with such good accuracy. From a drinking water treatment perspective, the magnitude of the increasing cyanobacterial concentration may not be as important as the time-prediction of when increased concentrations (>2000 cells/mL) can be expected. Predicting the correct time of when >2000 cells/mL can be expected will enable the DWTW to start preparations to deal with higher concentrations in the source water, irrespective of the specific expected cyanobacterial concentration and therefore this model is still regarded as very useful from a drinking water treatment perspective.

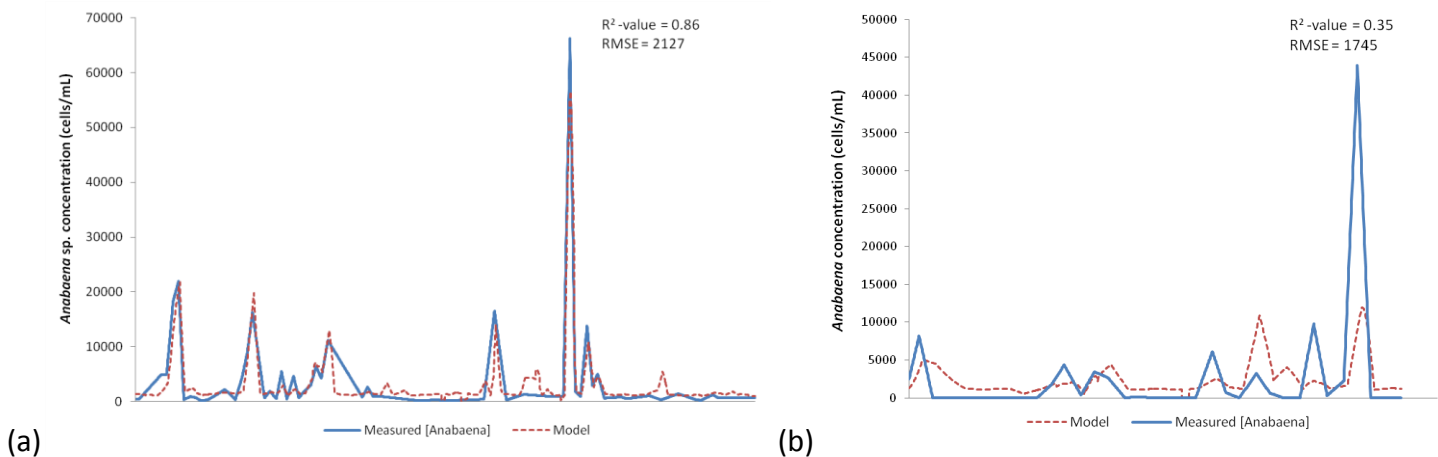


Figure 5.4 a Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[7] concentration (7 days in advance) in the Vaal Dam using 25% (bootstrapped) of the 10 years development dataset.
 b Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam.

5.2.3 *Anabaena*^[14] prediction (14 days in advance)

The IF criterium of the model predicting *Anabaena*^[14] concentration (Figure 5.5) is determined by the Si concentration. The THEN branch of the model (Figure 5.5a) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum, and also sensitivity towards NH_4^+ concentration during the first 5% of the NH_4^+ input range.

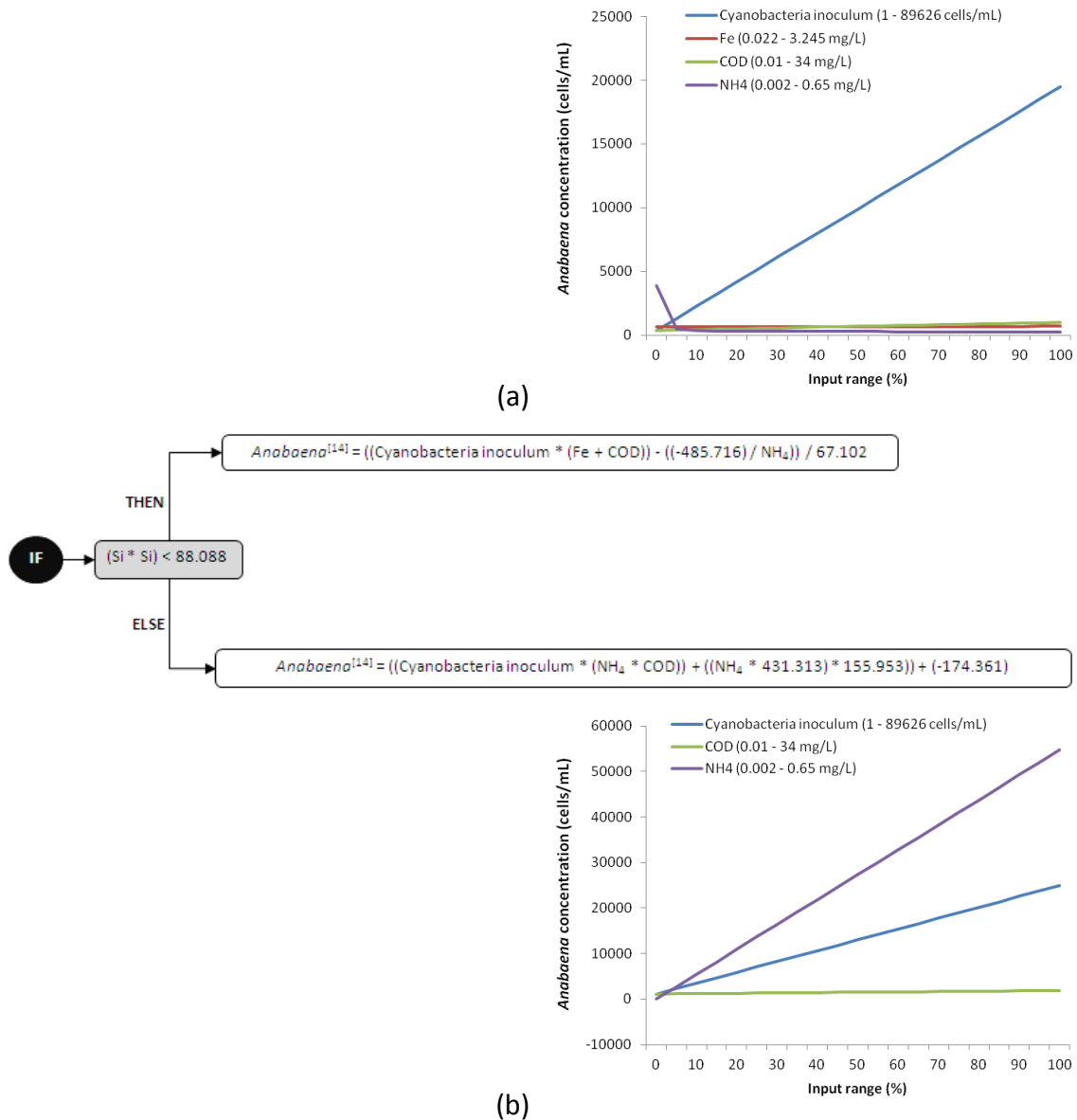


Figure 5.5 Forecasting model and associated sensitivity analyses of *Anabaena*^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The ELSE branch of the model (Figure 5.5b) represents the high range rule set and shows the greatest sensitivity towards NH_4^+ concentration and less towards the initial cyanobacteria inoculum. The other variables used in the model (Fe and COD in the THEN branch and COD in the ELSE branch) display very little influence on the predicted *Anabaena*^[14] concentration.

The comparison between the measured *Anabaena* sp. concentration and that of the results from the models predicting *Anabaena*^[14] (14 days in advance), when using the 25% bootstrapped testing dataset (Figure 5.6a), shows a R^2 -value of 0.80 and a RMSE of 2575 cells/mL. When the model was tested with 3 years of unseen data (Figure 5.6b), the model showed a R^2 -value of 0.37 and a RMSE of 3726 cells/mL. Visual inspection of the correlation (Figure 5.6b) reveals that the model does not seem to predict increased *Anabaena*^[14] concentration well, although there may be some application value in the >15 000 cells/mL range.

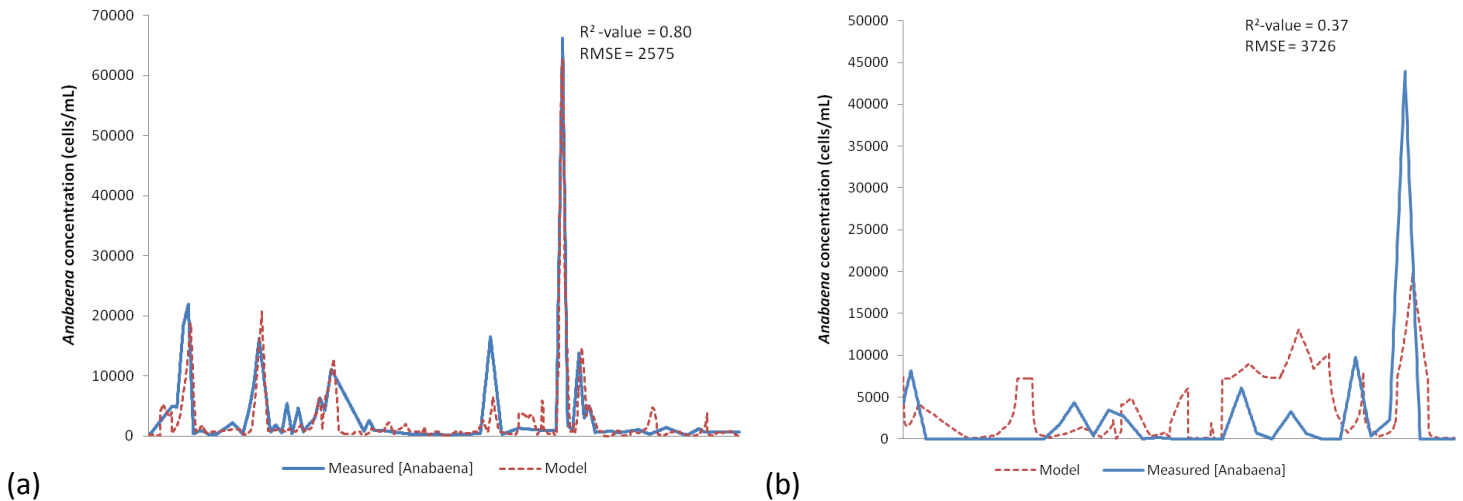


Figure 5.6 a Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam.

5.2.4 *Anabaena*^[21] prediction (21 days in advance)

The IF criterium of the model predicting *Anabaena*^[21] concentration (Figure 5.7) is determined by the Si concentration. The THEN branch of the model (Figure 5.7a) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 5.7b) represents the high range rule set and shows the greatest

sensitivity towards initial cyanobacteria inoculum and to a lesser extent to NO_2^- and NH_4^+ concentrations especially during the first 10 % of the input range thereof. The other variables used in the model (Secchi disk depth, in the THEN branch and chlorophyll-*a* and conductivity in the ELSE branch) display very little influence on the predicted *Anabaena*^[21] concentration.

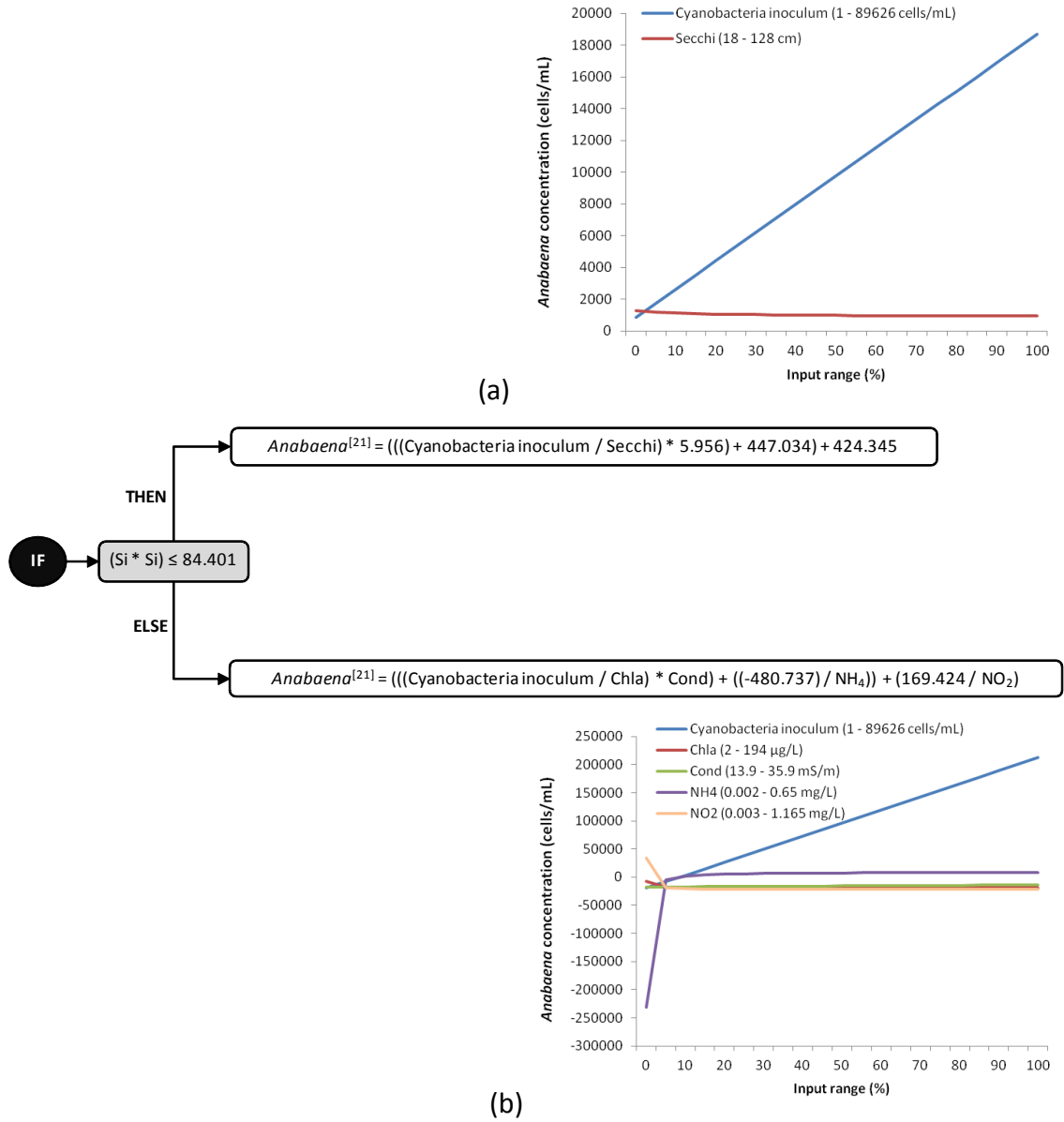


Figure 5.7 Forecasting model and associated sensitivity analyses of *Anabaena*^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Anabaena* sp. concentration and that of the results from the models predicting *Anabaena*^[21] (21 days in advance), when using the 25% bootstrapped testing dataset (Figure 5.8a), shows a R^2 -value of 0.76 which decreased to 0.04 when tested with the unseen dataset (Figure 5.8b). The RMSE when using the 25% bootstrapped testing dataset (Figure 5.8a) was 2786 cells/mL and decreased to 1213 cells/mL when tested with the unseen dataset (Figure 5.8b). Visual inspection of the correlation when using the unseen data (Figure 5.8b) reveals that the model does not predict increased *Anabaena*^[21] concentration well.

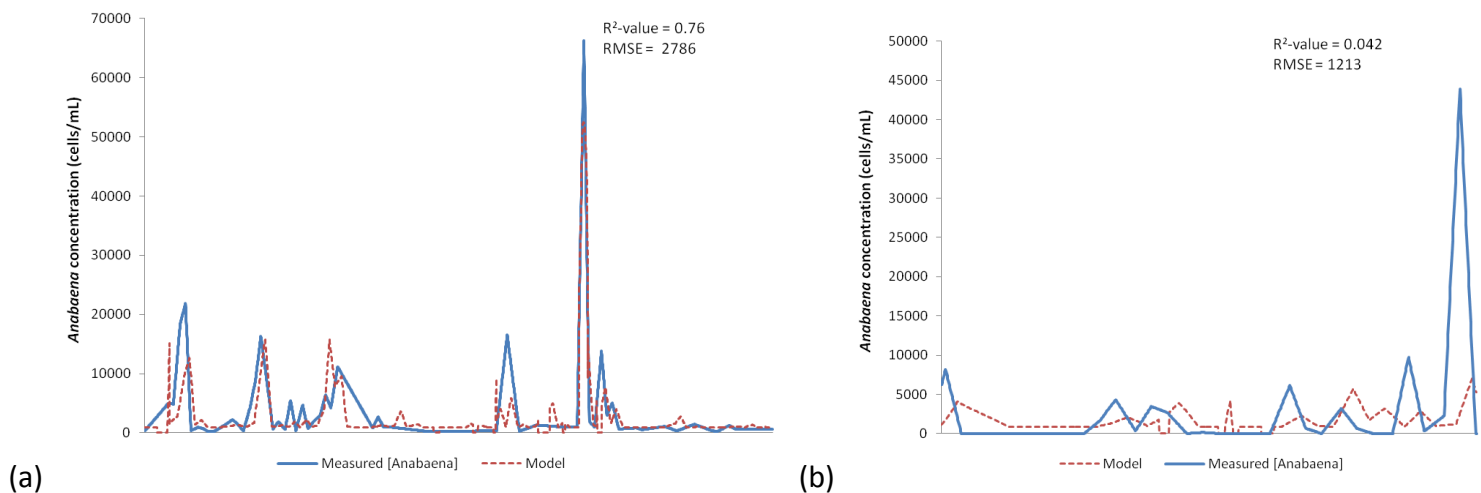


Figure 5.8 a Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[21] concentration (21 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam.

5.3 Models for the prediction of *Microcystis* sp. concentration

Microcystis sp. is a cyanobacterium well-known all over the world for the production of the cyanotoxin, microcystin (WHO, 2011; Du Preez and Van Baalen, 2006; Conradie and Barnard, 2012). The presence of microcystin in the source water is very important from a drinking water treatment perspective, since it is one of the variables specified in the SANS 241 : 2011 drinking water standard of South Africa (SANS 241-1, 2011; SANS 241-2, 2011). In the Vaal Dam and canal investigated in this study, *Microcystis* sp. is most commonly found (Ewerts et al., 2013) and associated with the production of the taste and odour compound geosmin, as well as the cyanotoxins, microcystin (Swanepoel et al., 2008a). Therefore, models

predicting the occurrence of *Microcystis* sp. will be of great benefit to the DWTW using the Vaal Dam as source, in order to prepare for incidents of high *Microcystis* sp. concentrations in the source water.

5.3.1 *Microcystis*^[0] prediction (in real-time)

The IF criterium of the model (Figure 5.9) is determined by the Fe concentration.

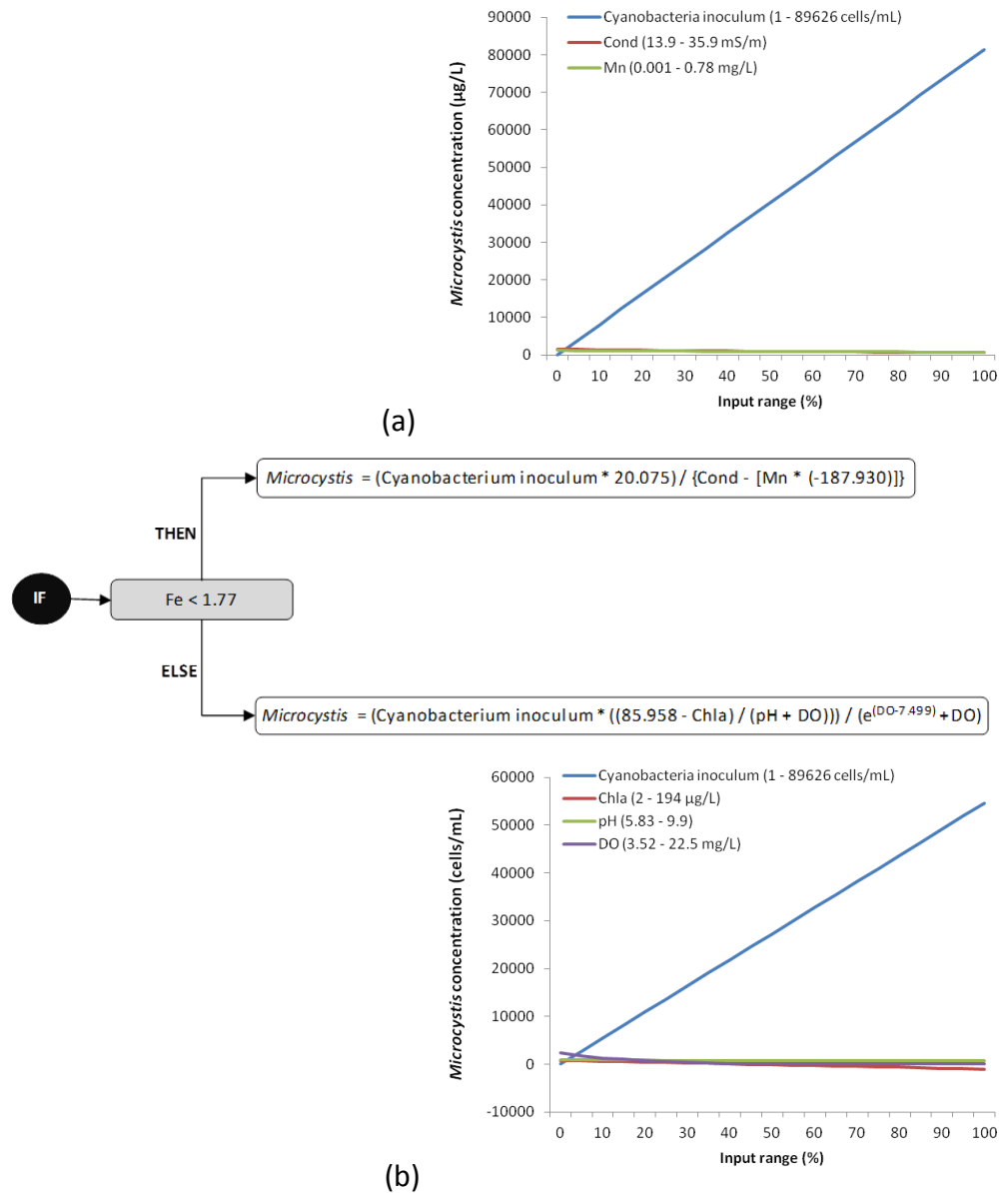


Figure 5.9 Forecasting model and associated sensitivity analyses of real-time *Microcystis*^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The THEN branch of the model (Figure 5.9a) represents the low range rule set and the ELSE branch of the model (Figure 5.9b) represents the high range rule set. Both the THEN and ELSE branches of the model show the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables used in the model (conductivity and Mn in the THEN branch and chlorophyll-*a*, pH and DO in the ELSE branch) display very little influence on the predicted *Microcystis*^[0] concentration when compared to the influence of the initial cyanobacteria inoculum.

The comparison between the measured *Microcystis* sp. concentration and the results from the models predicting real-time *Microcystis*^[0] when using the 25% boot-strapped testing dataset (Figure 5.10a), shows a R²-value of 0.95 and a root mean square error (RMSE) of 4262 cells/mL. When the model was tested with 3 years of unseen data, the correlation showed an even higher R²-value (0.97) and a RMSE of 2462 cells/mL (Figure 5.10b), indicating that the event prediction of increased *Microcystis*^[0] concentration together with the magnitude of the event displayed an excellent correlation.

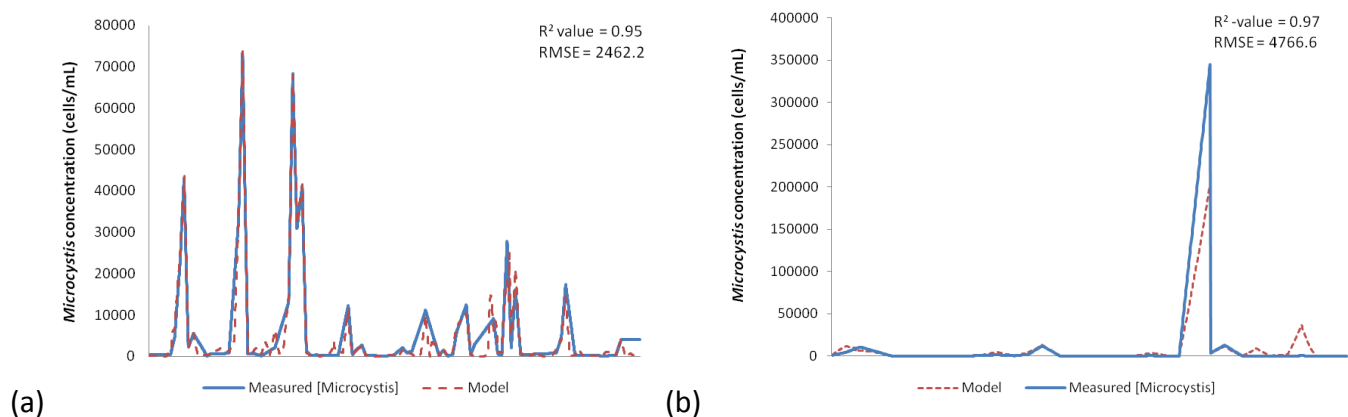


Figure 5.10 a Comparison between the measured *Microcystis* sp. concentration and predicted real-time *Microcystis*^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.

b Comparison between the measured *Microcystis* sp. concentration and predicted real-time *Microcystis*^[0] concentration using 3 years' unseen data from the Vaal Dam.

5.3.2 *Microcystis*^[7] prediction (7 days in advance)

For the best model developed for the prediction of *Microcystis*^[7] (7 days in advance), the IF criterium of the model is determined by a combination of dissolved oxygen (DO), NO₃⁻ and

turbidity (Figure 5.11). The THEN branch of the model (Figure 5.11a) represents the high range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 5.11b) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables used in this model (namely Secchi disk depth, PO_4^{3-} , turbidity, Fe and DO) in comparison to the initial cyanobacteria inoculum, display very little influence on the predicted *Microcystis*^[7] concentration.

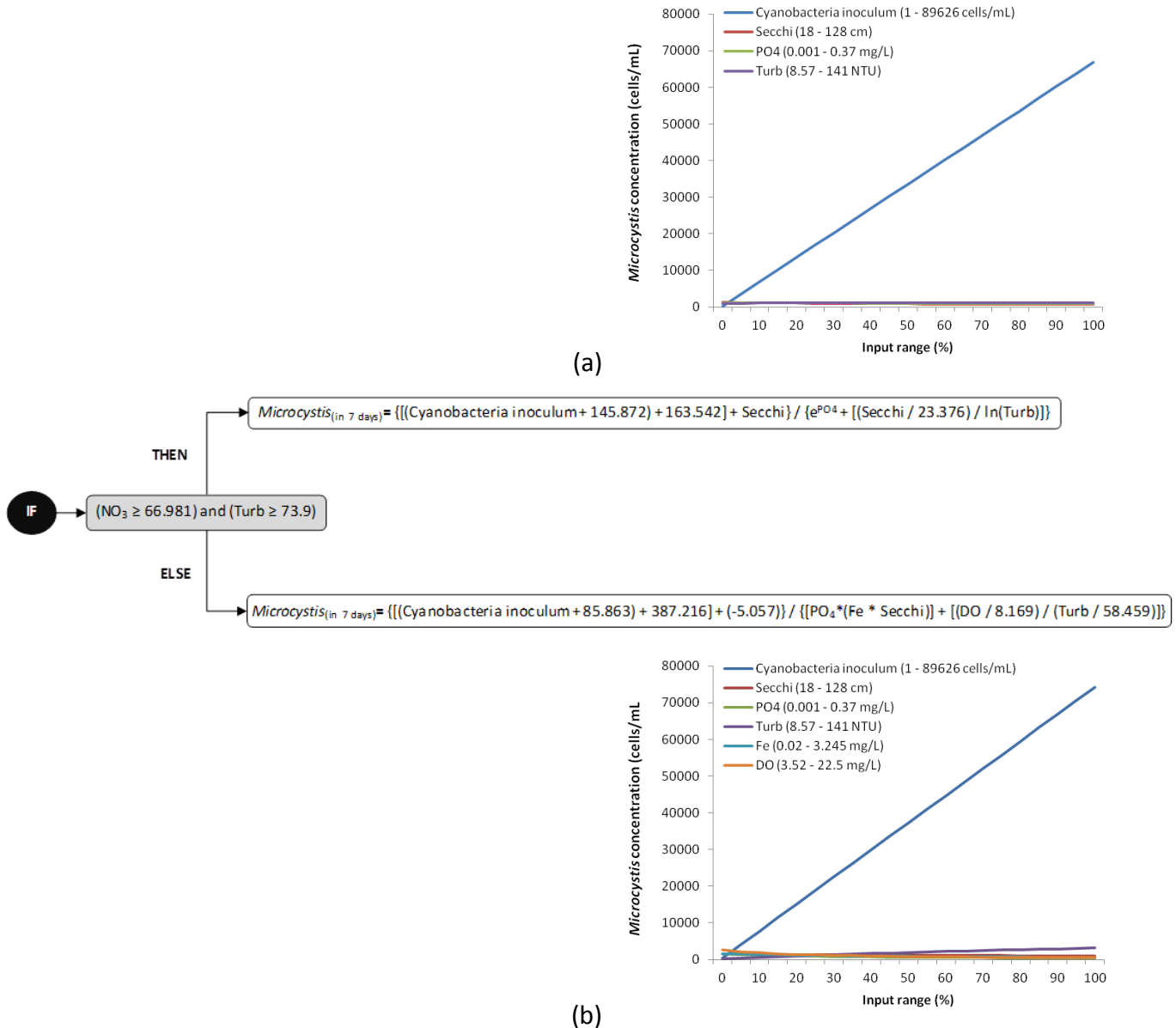


Figure 5.11 Forecasting model and associated sensitivity analyses for *Microcystis*^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis*^[7] (7 days in advance), when using the 25% bootstrapped testing dataset (Figure 5.12a), shows a R^2 -value of 0.91 and a RMSE of 3136 cells/mL. When the model was tested with 3 years of unseen data (Figure 5.12b), the model showed a R^2 -value of 0.59 and a RMSE of 43 479 cells/mL. The event prediction of increased *Microcystis* sp. concentration showed a significant correlation (Figure 5.12b); however, it seems that the *Microcystis*^[7] concentration is over-estimated somewhat by the model.

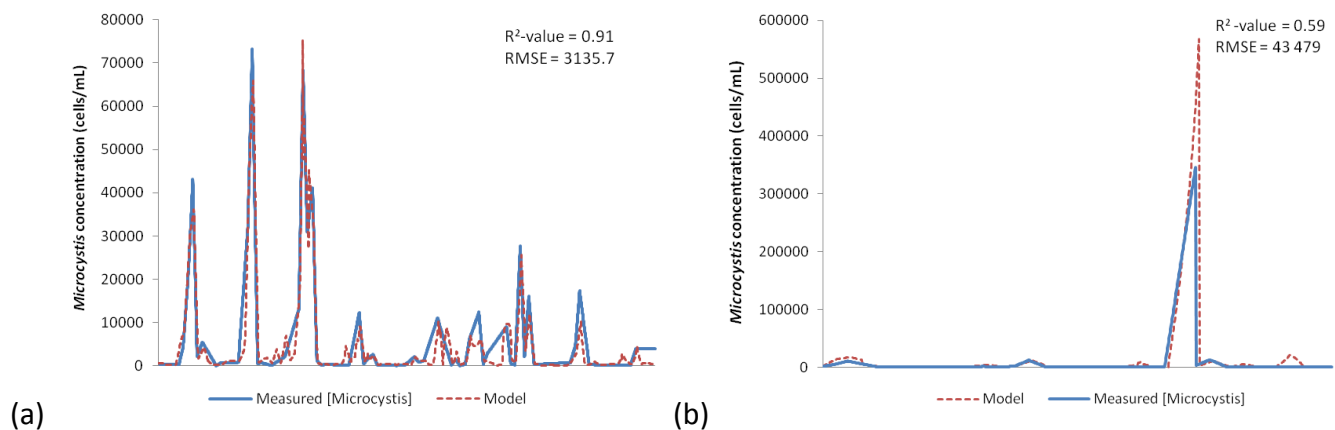


Figure 5.12 a Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[7] concentration (7 days in advance) in the Vaal Dam using 25% (bootstrapped) of the 10 years development dataset.
 b Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam.

5.3.3 *Microcystis*^[14] prediction (14 days in advance)

The IF criterium of the best model developed for the prediction of *Microcystis*^[14] (14 days in advance), is determined by the chemical oxygen demand (COD) (Figure 5.13). The THEN branch of the model (Figure 5.13a) represents the high range rule set and the ELSE branch of the model (Figure 5.13b) represents the low range rule set. Both of the branches (Figure 5.13a and 5.13b) show the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables used 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*^[14] concentration in comparison to the initial cyanobacteria inoculum.

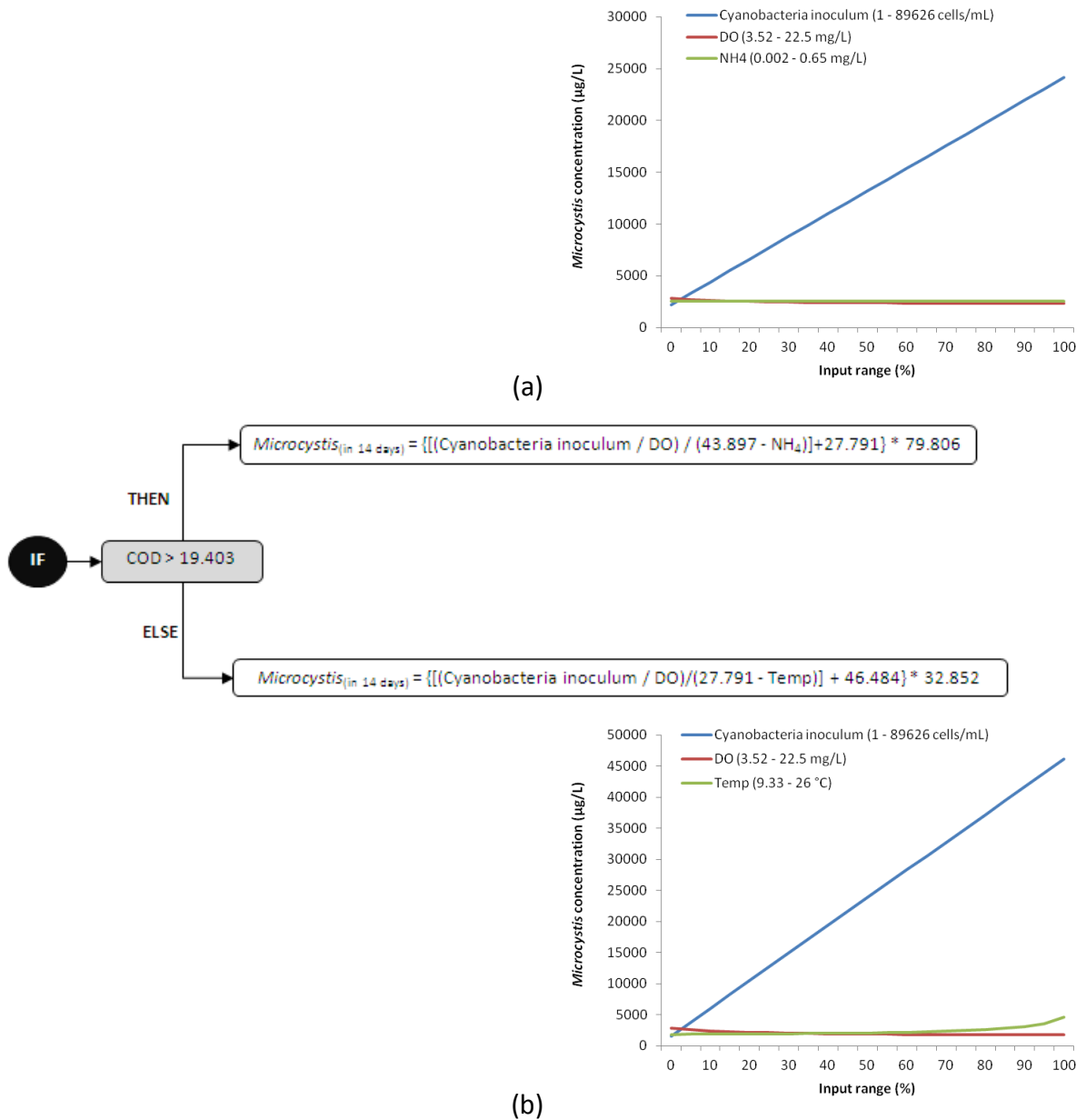


Figure 5.13 Forecasting model and associated sensitivity analyses for *Microcystis*^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis*^[14] 14 days in advance, when using the 25% bootstrapped testing dataset (Figure 5.14a), shows a R^2 -value of 0.79 and a RMSE of 4494 cells/mL. When the model was tested with 3 years of unseen data (Figure 5.14b), the model showed a R^2 -value of 0.38 and a RMSE of 48 130 cells/mL. The event prediction of increased

Microcystis sp. concentration together with the magnitude of the event showed a significant correlation, although the model seem to over-estimate high concentrations of *Microcystis* sp.

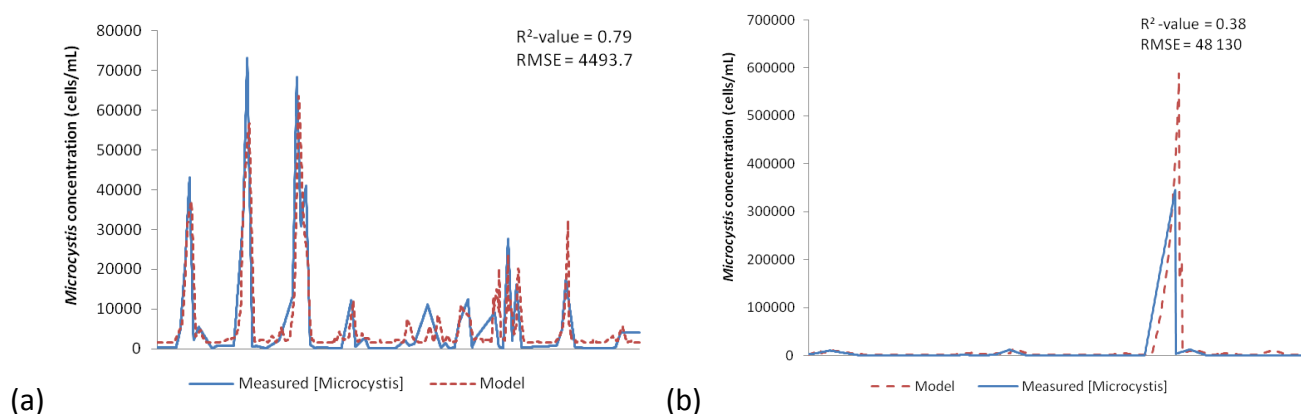


Figure 5.14 a Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam.

5.3.4 *Microcystis*^[21] prediction (21 days in advance)

The IF criterium of the best model developed for the prediction 21 days in advance (Figure 5.15), is determined by a combination of nutrients (PO_4^{3-} and NH_4^+). The THEN branch of the model (Figure 5.15a) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 5.15b) 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.

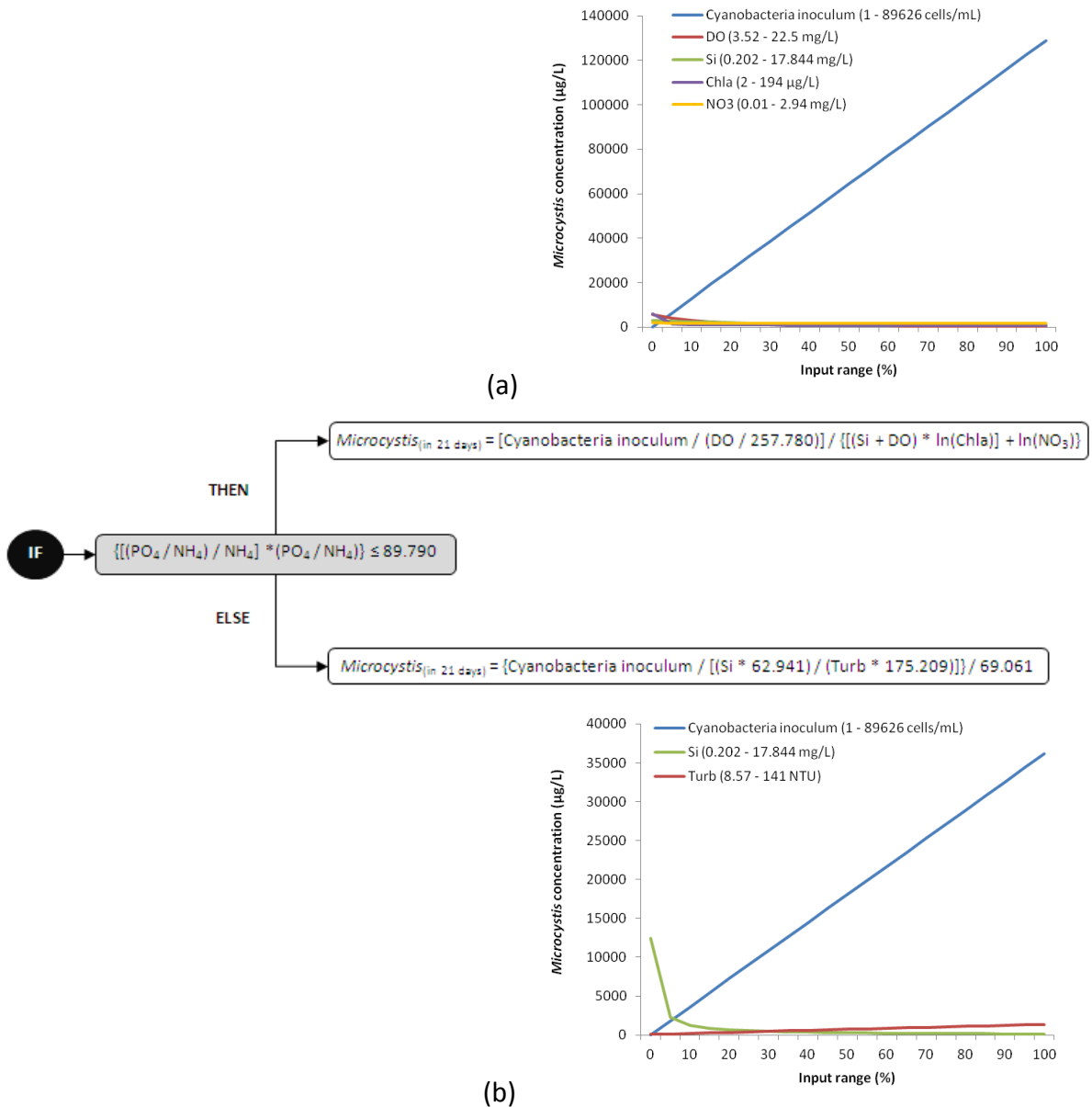


Figure 5.15 Forecasting model and associated sensitivity analyses for *Microcystis*^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis*^[21] 21 days in advance, when using the 25% bootstrapped testing dataset (Figure 5.16a), shows a R^2 -value of 0.74 and a RMSE of 4994 cells/mL. When the model was tested with 3 years of unseen data (Figure 5.16b), the model showed a RMSE of 18 494 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 unseen data was tested on the model.

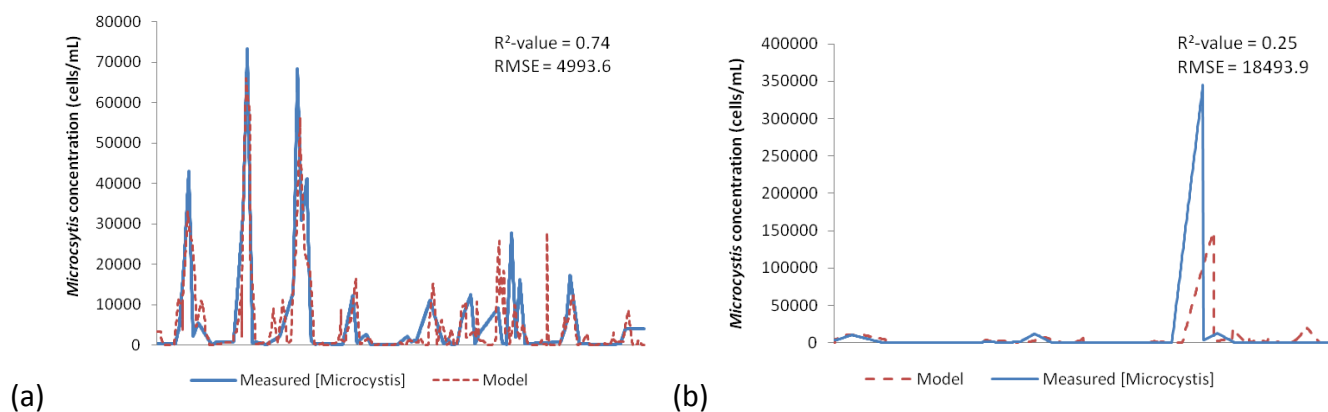


Figure 5.16 a Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[21] (21 days in advance) concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam.

5.4 Models for the prediction of microcystin concentration

Microcystin, the most commonly known cyanotoxin in the world, is produced by cyanobacteria such as *Microcystis* sp., *Anabaena* sp., *Oscillatoria* sp., *Planktothrix* sp. and *Nostoc* sp. (WHO, 2011). Microcystin (as other cyanotoxins) usually remain contained inside the cyanobacteria cells, and is only released upon cell lysis (breakdown of cyanobacteria cells) and cell death (Du Preez and Van Baalen, 2006). However, during the drinking water purification process, the addition of coagulant chemicals (such as hydrated lime) or disinfection chemicals (such as ozone or chlorine) may cause cell lysis, resulting in the release of cyanotoxins into the water (Knappe et al., 2004; Newcombe, 2009). Additionally, the presence of microcystin in the source water is very important from a drinking water treatment perspective, as it is one of the variables specified in the SANS 241:2011 drinking water standard of South Africa (SANS 241-1, 2011; SANS 241-2, 2011). Therefore, the ability to be able to predict the occurrence of high microcystin concentrations will enable DWTW to adjust their processes in advance to reduce the risk to consumers.

5.4.1 Microcystin^[0] prediction (in real-time)

The IF criterium of the model is mainly determined by the turbidity of the water with an added contribution by the initial cyanobacteria inoculum (Figure 5.17).

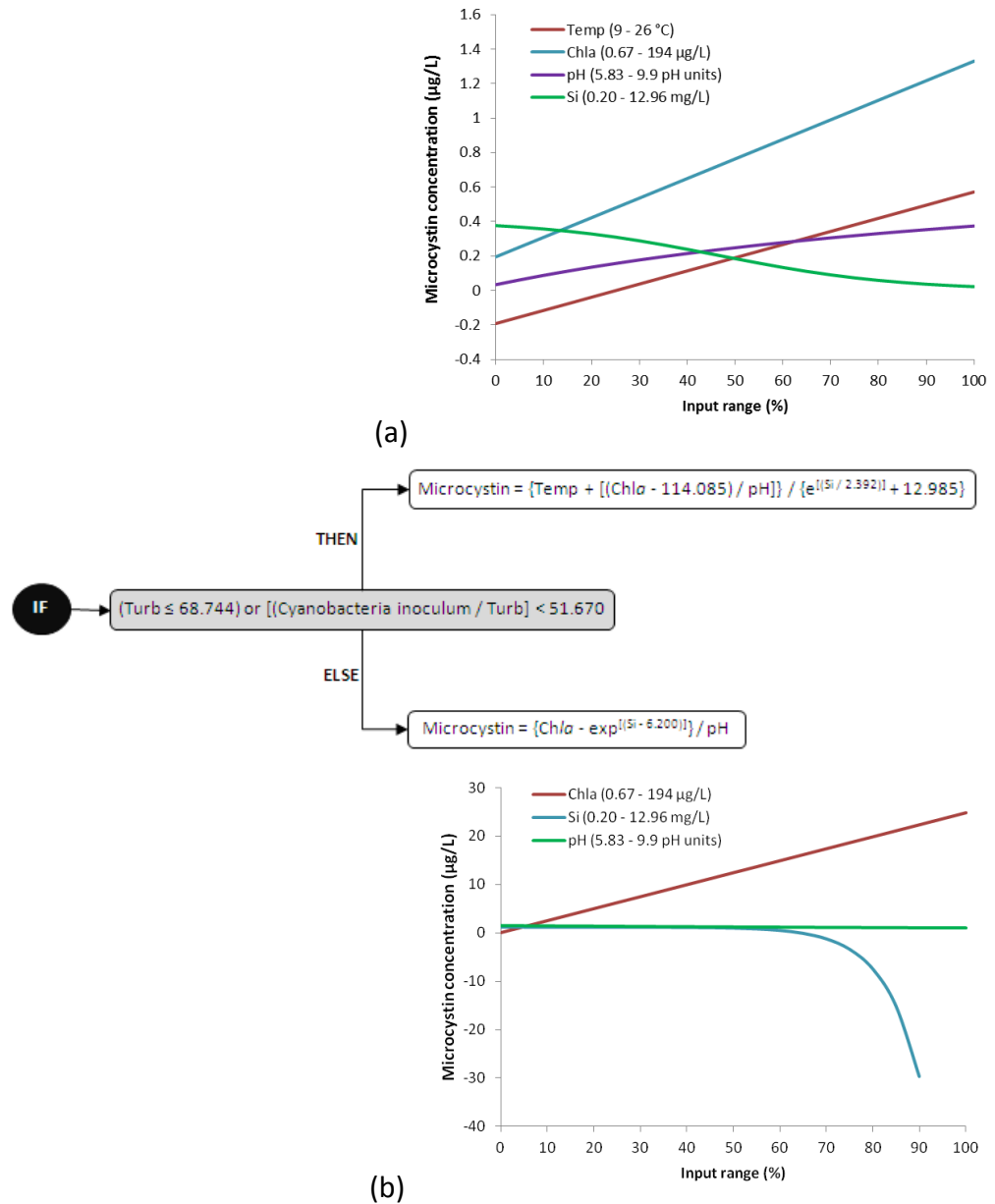


Figure 5.17 Forecasting model and associated sensitivity analyses for real-time microcystin^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The THEN branch of the model (Figure 5.17a) represents the low range rule set and shows the greatest sensitivity towards the chlorophyll-*a* concentration, temperature and pH and to a lesser extent the Si concentration. The ELSE branch of the model (Figure 5.17b) represents the high range rule set and shows the greatest sensitivity towards chlorophyll-*a* concentration as well as the Si concentration particularly during the higher 30% of the Si input range. The water pH (in the ELSE branch of the model) display very little influence on the predicted microcystin concentration.

The comparison between the measured microcystin concentration and that of the results from the models predicting the real-time microcystin^[0] concentration when using the 25% bootstrapped testing dataset (Figure 5.18a), shows a R^2 -value of 0.88 and RMSE of 0.216 $\mu\text{g/L}$. When the model was tested with 3 years of unseen data, the correlation the R^2 -value decreased to 0.23 and the RMSE increased to 0.604 $\mu\text{g/L}$ (Figure 5.18b).

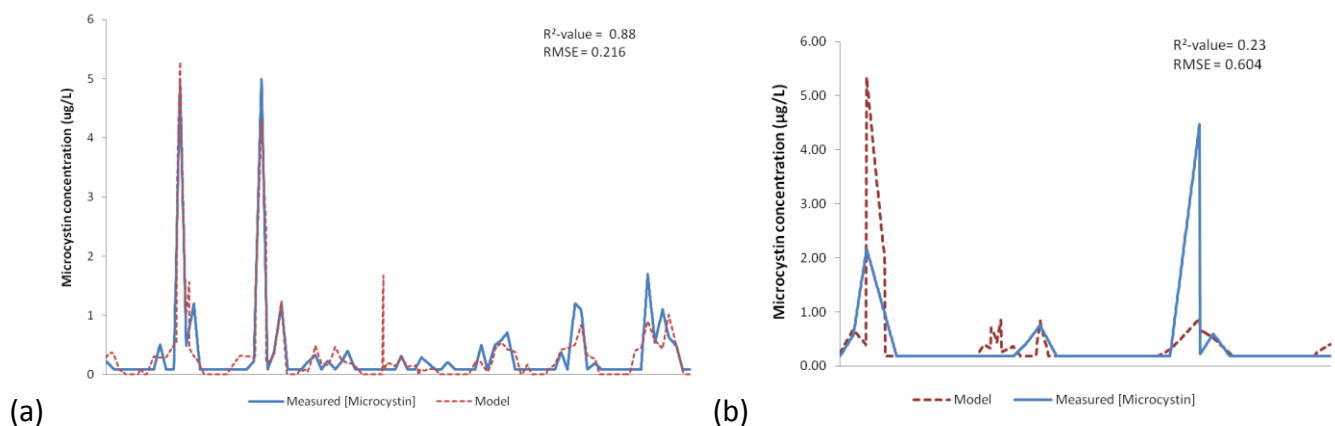


Figure 5.18 a Comparison between the measured microcystin concentration and predicted real-time microcystin concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.

- b Comparison between the measured microcystin concentration and predicted real-time microcystin concentration using 3 years' unseen data from the Vaal Dam.

Visual inspection of the correlation (Figure 5.18b) reveals that the time prediction of increased microcystin^[0] concentration is relatively accurate; however, the magnitude of the concentration is not predicted with good accuracy. From a drinking water treatment perspective, the magnitude of the increasing microcystin concentration may not be as important as the time-prediction of when increasing concentrations can be expected. Predicting the correct time of when high microcystin concentrations can be expected will enable the DWTW to start with preparations to deal with it, irrespective of the specific expected microcystin concentration and

therefore this model is still regarded as very useful from a drinking water treatment perspective.

5.4.2 *Microcystin^[7] prediction (7 days in advance)*

The IF criterium of the model is determined by the turbidity of the water (Figure 5.19). The THEN branch of the model (Figure 5.19a) represents the high range rule set and shows the greatest sensitivity towards initial cyanobacteria inoculum.

The ELSE branch of the model (Figure 5.19b) represents the low range rule set and shows the greatest sensitivity towards water temperature, and to a lesser extent to conductivity. The other variables used in the model (chlorophyll-*a*, pH, water temperature and DO in the THEN branch) display very little influence on the predicted microcystin concentration.

The comparison between the measured microcystin concentration and that of the results from the models predicting the microcystin concentration 7 days in advance when using the 25% boot-strapped testing dataset (Figure 5.20a), shows a R^2 -value of 0.75 and RMSE of 0.27 $\mu\text{g/L}$. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.27 and RMSE of 0.525 $\mu\text{g/L}$ (Figure 5.20b). Although the magnitude of the concentration is not predicted with such good accuracy, the time prediction of increased microcystin concentration is predicted with relatively good accuracy (Figure 5.20b). As stated previously, the magnitude of the cyanobacteria concentration or the specific concentration of increased microcystin may not be as important as the time-prediction of when it can be expected from a drinking water treatment perspective and therefore this model is still regarded useful.

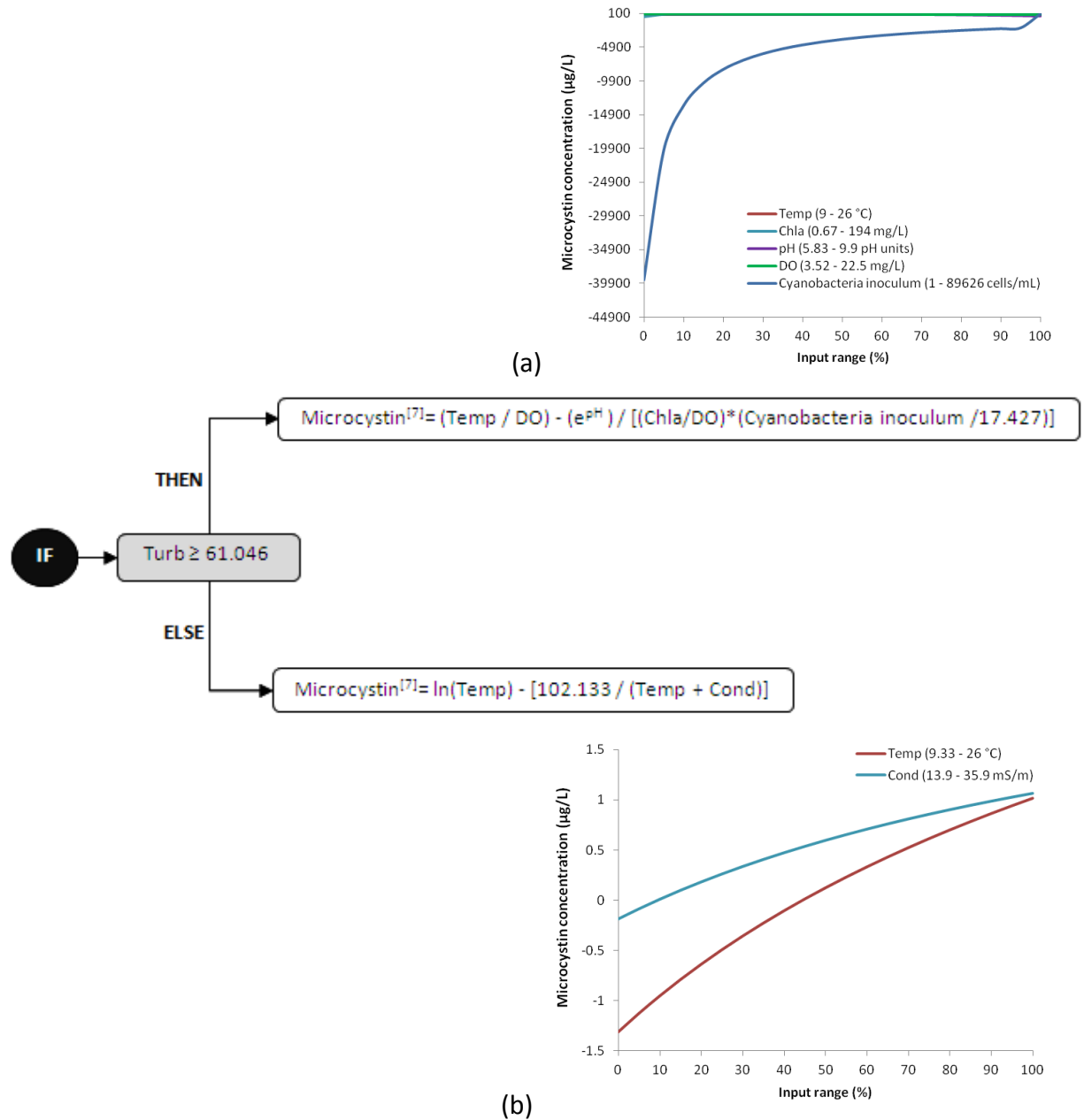


Figure 5.19 Forecasting model and associated sensitivity analyses for microcystin^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

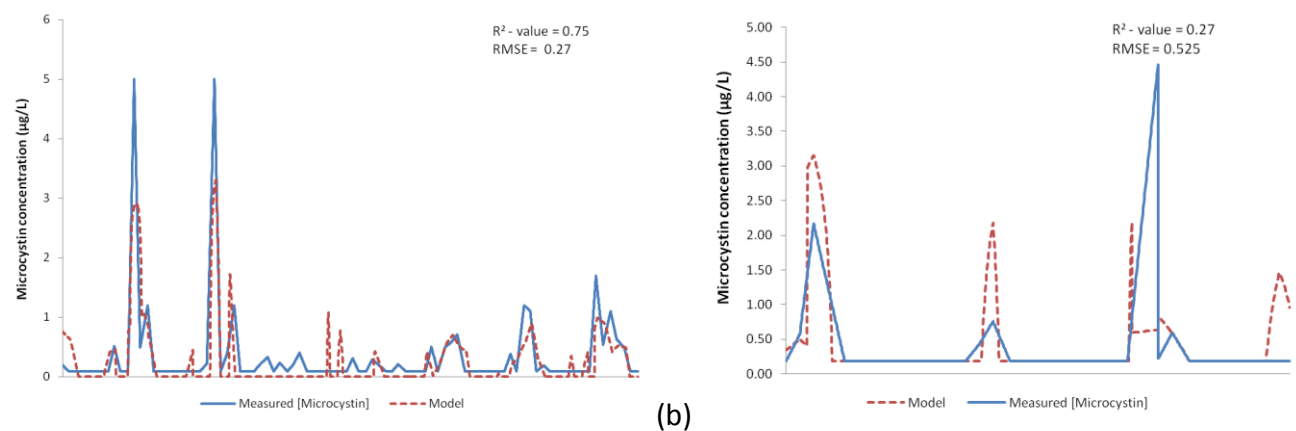


Figure 5.20 a Comparison between the measured microcystin concentration and predicted microcystin^[7] concentration (7 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured microcystin concentration and predicted microcystin^[7] concentration (7 days in advance) using 3 years' unseen data from the Vaal Dam.

5.4.3 Microcystin^[14] prediction (14 days in advance)

The IF criterium of the model is determined by the concentration of the nutrient NH_4^+ , together with the initial cyanobacteria inoculum (Figure 5.21).

The THEN branch of the model (Figure 5.21a) represents the low range rule set and shows sensitivity towards temperature (particularly during the 70 - 100 % temperature input range) and chlorophyll-*a* concentration (particularly during the lower 20% of the chlorophyll-*a* input range). The ELSE branch of the model (Figure 5.21b) represents the high range rule set and shows the greatest sensitivity towards Si concentration and to a lesser extent to pH.

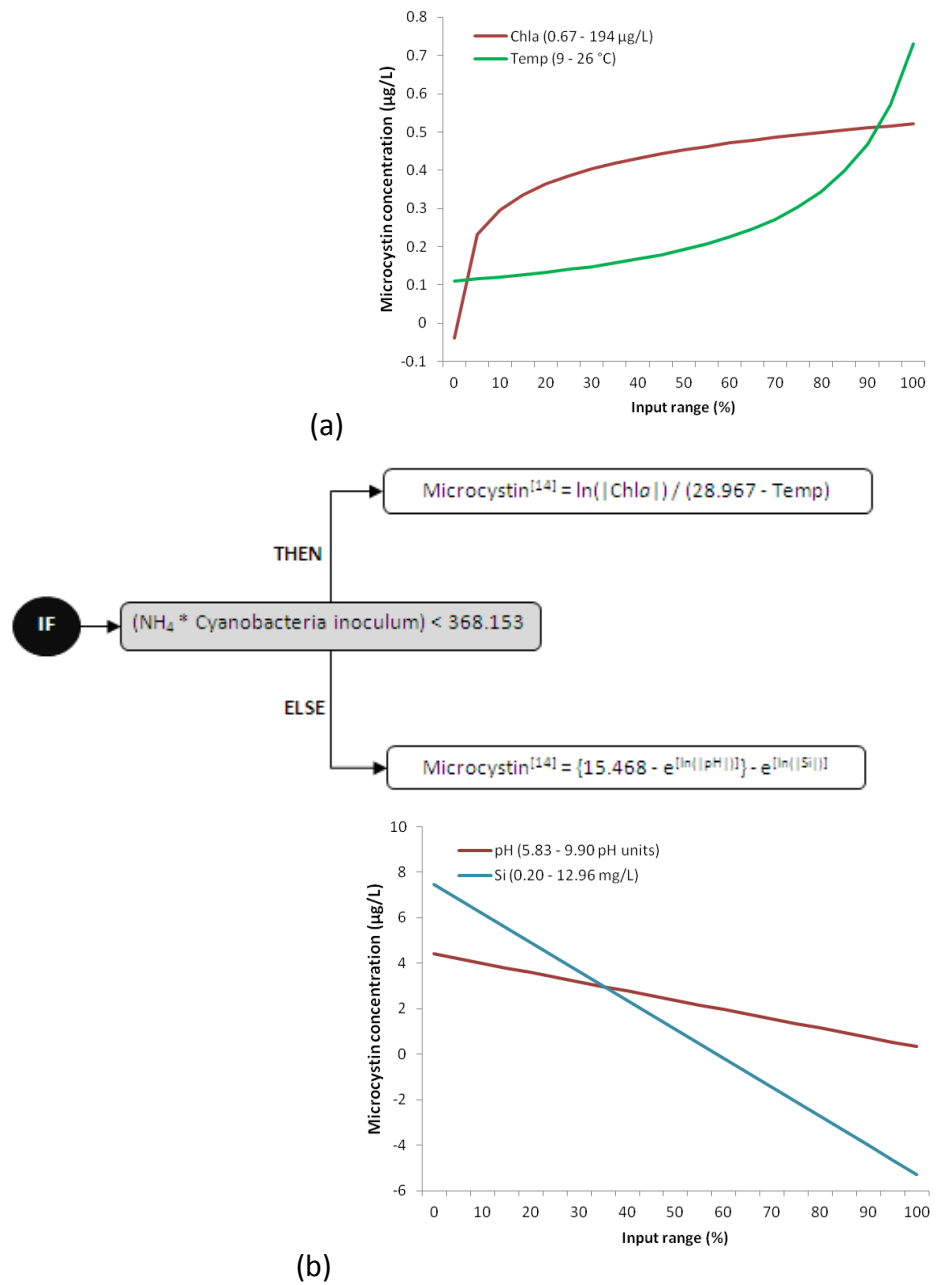


Figure 5.21 Forecasting model and associated sensitivity analyses for microcystin^[14] concentration (14 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured microcystin concentration and that of the results from the models predicting the microcystin^[14] concentration 14 days in advance when using the 25% boot-strapped testing dataset (Figure 5.22a), shows a R^2 -value of 0.63 and RMSE of 0.31 $\mu\text{g/L}$. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.28 and RMSE of 0.57 $\mu\text{g/L}$ (Figure 5.22b). From visual inspection it is evident that although the magnitude of the concentration is not predicted with such good accuracy, the time prediction of increased microcystin concentration is predicted with relatively good accuracy (Figure 5.22b) and therefore, from a drinking water treatment perspective, definitely useful for application in early warning.

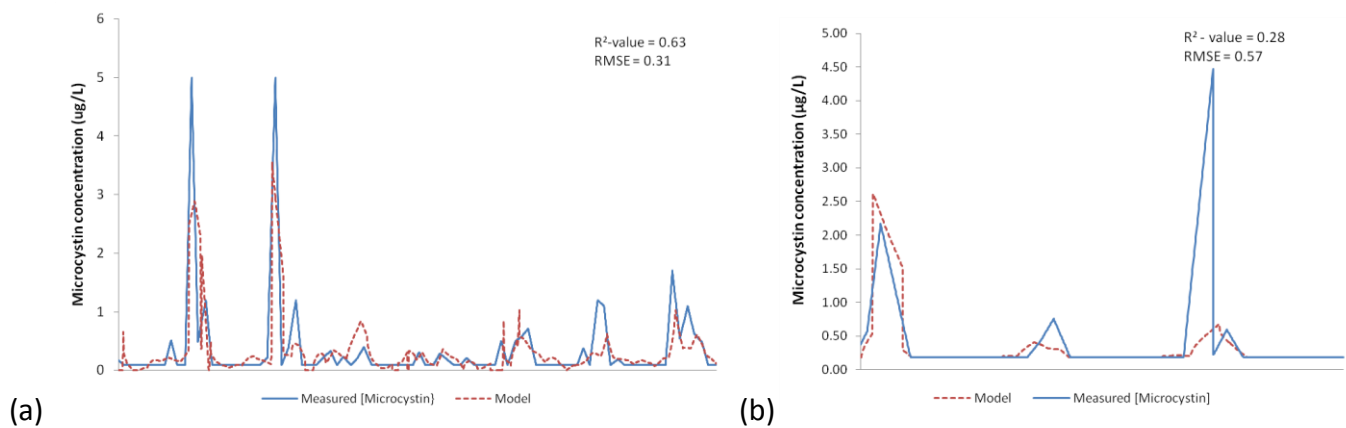


Figure 5.22 a Comparison between the measured microcystin concentration and predicted microcystin^[14] concentration (14 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured microcystin concentration and predicted microcystin^[14] concentration (14 days in advance) using 3 years' unseen data from the Vaal Dam.

5.4.4 Microcystin^[21] prediction (21 days in advance)

The IF criterium of the model is determined by the Si concentration (Figure 5.23).

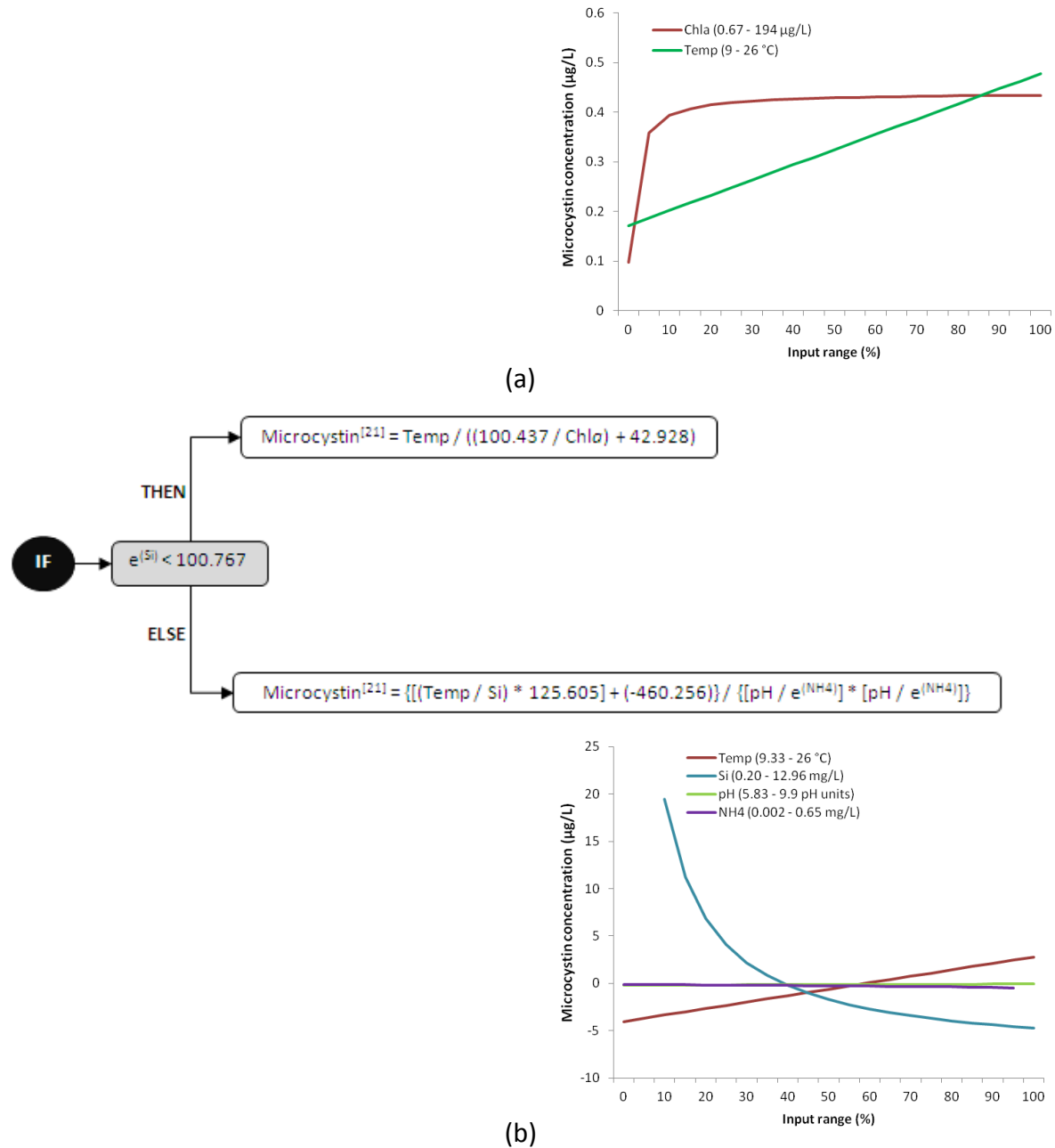


Figure 5.23 Forecasting model and associated sensitivity analyses for microcystin^[21] concentration (21 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The THEN branch of the model (Figure 5.23a) represents the low range rule set and shows the greatest sensitivity towards temperature and to a lesser extent chlorophyll-*a* concentration (particularly during the lower 10% of the chlorophyll-*a* input range). The ELSE branch of the model (Figure 5.23b) represents the high range rule set and shows the greatest sensitivity towards Si concentration and to a lesser extent to temperature. Changes in the other variables used in the model (pH and NH_4^+ in the ELSE branch) do not seem to have a significant influence on the predicted microcystin concentration.

The comparison between the measured microcystin concentration and that of the results from the models predicting the microcystin^[21] concentration 21 days in advance when using the 25% boot-strapped testing dataset (Figure 5.24a), shows a R^2 -value of 0.58 and RMSE of 0.33 $\mu\text{g/L}$. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.41 and RMSE of 0.52 $\mu\text{g/L}$ (Figure 5.24b). Visual inspection (Figure 5.24b) indicates that time prediction of concentrations $>1.5 \mu\text{g/L}$ is relatively accurate although magnitude of high concentrations is not predicted with such accuracy.

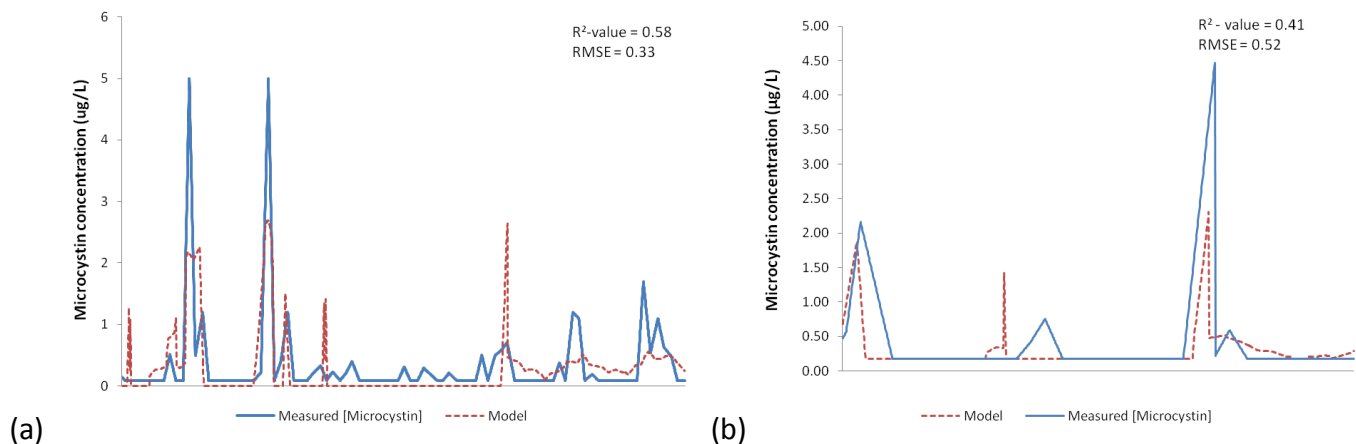


Figure 5.24 a Comparison between the measured microcystin concentration and predicted microcystin^[21] concentration (21 days in advance) in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured microcystin concentration and predicted microcystin^[21] concentration (21 days in advance) using 3 years' unseen data from the Vaal Dam.

5.5 Models for the prediction of geosmin concentration

Geosmin, one of the very common taste and odour-related organic compounds produced by cyanobacteria (e.g. *Anabaena* sp., *Microcystis* sp., *Aphanizomenon* sp. and *Oscillatoria* sp. – Swanepoel et al., 2008a) and to a lesser extent actinomycetes (Zaitlin and Watson, 2006), is the compound most frequently produced in the Vaal Dam. During the drinking water treatment process, the addition of coagulant chemicals (such as hydrated lime) or disinfection chemicals (such as ozone or chlorine) may cause cyanobacteria and actinomycetes cell lyses, resulting in the release of geosmin into the water. Geosmin (unlike certain cyanotoxins e.g. microcystin), cannot be oxidised by chlorine (Knappe et al., 2004) and when a drinking water treatment facility only has conventional drinking water treatment processes in place (like most of the DWTW in South Africa) the release of these organic compounds into the water is a major concern. Although taste and odour-related organic compounds (such as geosmin and 2-methylisoborneol) are not at all harmful to consumer health, it is notorious for decreasing consumer confidence in tap water (Zoschke et al., 2011). Therefore, the ability of a water treatment facility to be able to predict the occurrence of high geosmin concentrations will enable DWTW to adjust their processes in advance. This will aid to reduce the presence of taste and odour compounds in the treated water, which may be even more detrimental to the water business than the presence of microcystin, since microcystin can be oxidised by chlorine and is undetectable to human senses.

The data used to develop the geosmin models was not from the Vaal Dam as in all the other models, but from the canal since geosmin is only measured at sampling sites classified as “source water mains” (such as M-Canal – refer to Figure 3.1 ■ page 20), and not at sampling sites classified as “catchment” (such as C-VD1 – refer to Figure 3.1 ■ page 20).

5.5.1 Geosmin^[0] prediction (in real-time)

The IF criterium of the model is determined by the silica concentration (Figure 5.25). The THEN branch of the model (Figure 5.25a) represents the low range rule set and shows the greatest sensitivity towards initial cyanobacteria inoculum and to a lesser extent to DOC (dissolved organic carbon) and water temperature (Temp). The ELSE branch of the model (Figure 5.25b) represents the high range rule set and shows the greatest sensitivity towards total photosynthetic pigments concentration (Tot pig, also called chlorophyll-665), especially during the 25 – 40% input range. This may be due to different cyanobacteria or other phytoplankton

species present (contributing to total photosynthetic pigments concentration), but only certain of these species contribute to the geosmin concentration. Competition between phytoplankton species (including cyanobacteria) some of which contribute to geosmin concentration and some not, may also have impacted on this relationship between total photosynthetic pigments (Tot pig) and geosmin concentration. To a lesser extent, initial cyanobacteria inoculum also affects geosmin concentration positively.

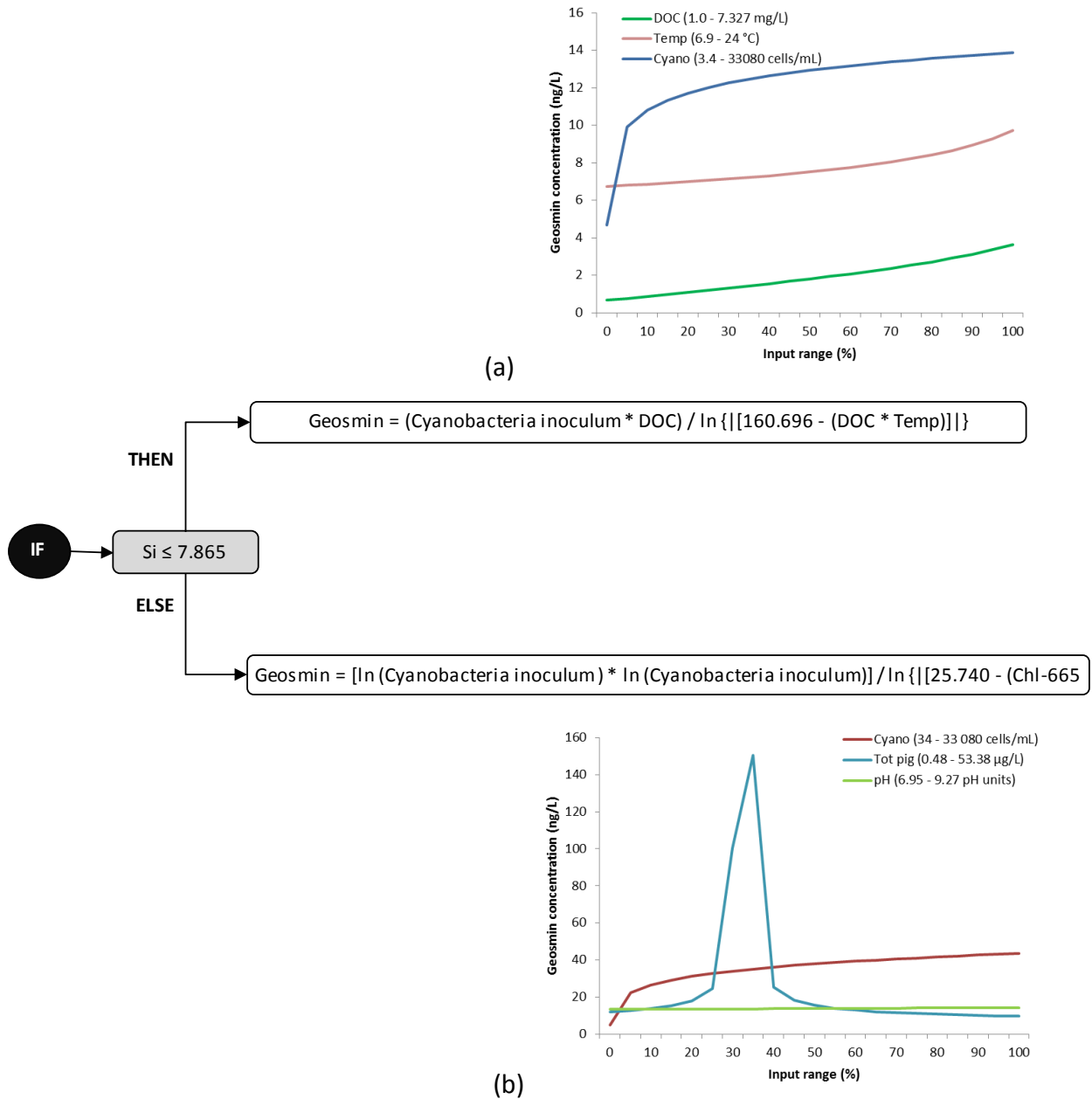


Figure 5.25 Forecasting model and associated sensitivity analyses of real-time geosmin^[0] concentration in the canal ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured geosmin concentration and that of the results from the models predicting real-time geosmin^[0] when using the 25% boot-strapped testing dataset (Figure 5.26a), shows a R^2 -value of 0.49 and a RMSE of 5.61 ng/L. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.02 and a RMSE of 7.54 ng/L (Figure 5.26b), indicating that neither the event prediction of increased geosmin concentration nor the magnitude of the event displayed a significant correlation. This may be due to actinomycetes (Zaitlin and Watson, 2006) and not cyanobacteria producing geosmin in the canal.

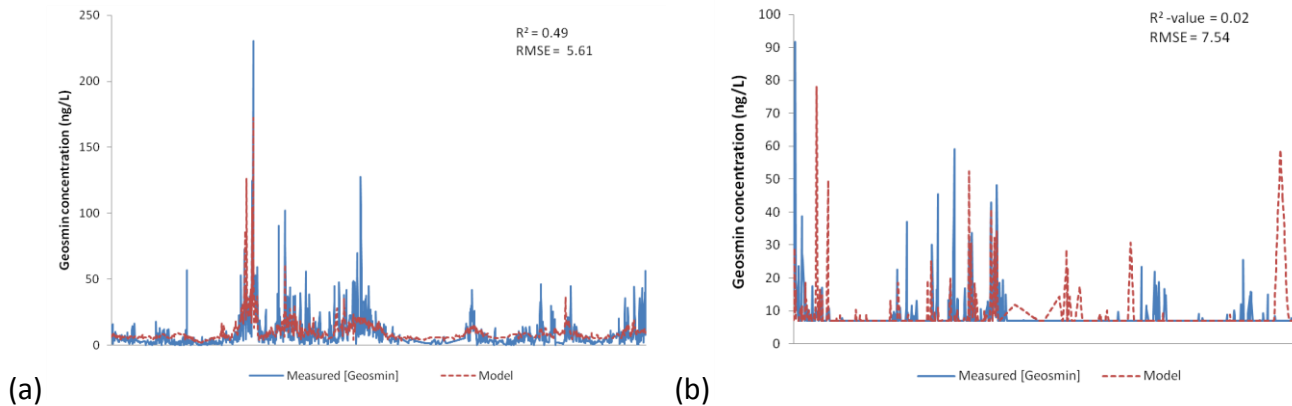


Figure 5.26 a Comparison between the measured geosmin concentration and predicted real-time geosmin^[0] concentration in the canal using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured geosmin concentration and predicted real-time geosmin^[0] concentration using 3 years' unseen data from the canal.

From the correlation results in Figure 5.26 it is evident that geosmin^[0] prediction is not as good as would be desired. Therefore, only prediction for real-time and 7 days in advance were investigated further.

5.5.2 Geosmin^[7] prediction (7 days in advance)

The IF criterium of the model predicting geosmin^[7] concentration (Figure 5.27) is determined by the cyanobacteria inoculum. The THEN branch of the model (Figure 5.27a) represents the low range rule set and shows the greatest sensitivity towards water temperature (Temp – especially during the first 20% of the input range) and to a lesser extent to turbidity (Turb). The ELSE branch of the model (Figure 5.27b) represents the high range rule set and shows the greatest sensitivity towards cyanobacteria inoculum, especially during the first 10% of the input range.

The relationship between cyanobacteria and geosmin is negative, which may be due to actinomycetes and not cyanobacteria producing geosmin in the canal (Zaitlin and Watson, 2006). However, if the significance of the model is taken into account (R^2 -values of 0.25 and 0.007) this negative relationship between cyanobacteria and geosmin may explain why the model did not prove to be effective in predicting geosmin^[7] concentrations.

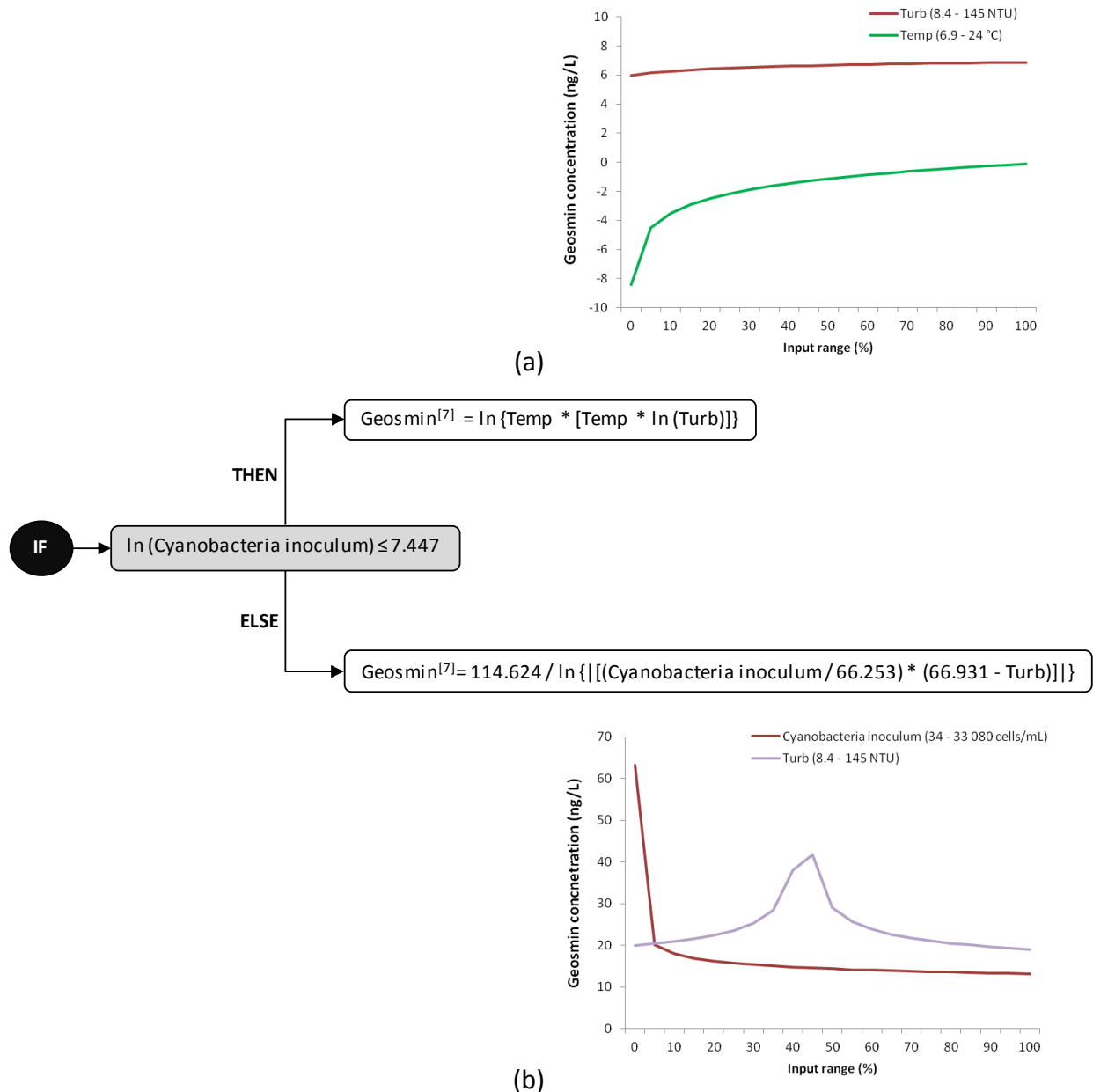


Figure 5.27 Forecasting model and associated sensitivity analyses of geosmin^[7] concentration (7 days in advance) in the canal ((a) represents the sensitivity analysis of the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

It also shows sensitivity towards turbidity (Turb) and especially during the 35 – 55% input range. This may be due to geosmin producing cyanobacteria preferring this turbidity range or that the geosmin producing cyanobacteria and actinomycetes contributed to the turbidity at this input range.

The comparison between the measured geosmin concentration and that of the results from the models predicting geosmin^[7] (7 days in advance) when using the 25% boot-strapped testing dataset (Figure 5.28a), shows a R^2 -value of 0.25 and a root mean square error (RMSE) of 5.78 ng/L. When the model was tested with 3 years of unseen data, the correlation showed a R^2 -value of 0.007 and a RMSE of 3.67 ng/L (Figure 5.28b), indicating that neither the event prediction of increased geosmin concentration nor the magnitude of the event displayed a significant correlation. This may be due to actinomycetes (Zaitlin and Watson, 2006) and not cyanobacteria producing geosmin in the canal.

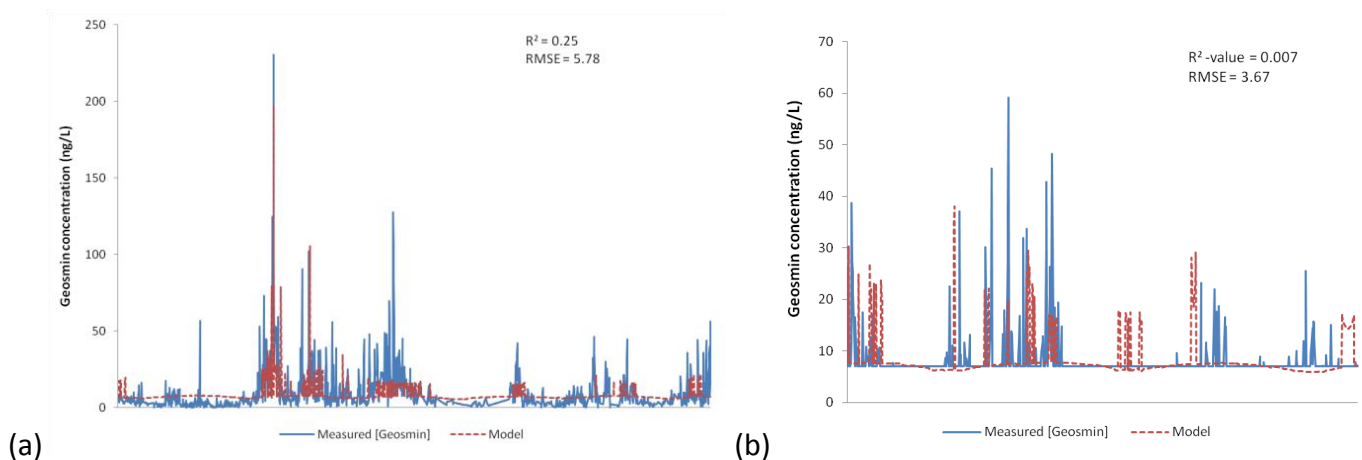


Figure 5.28 a Comparison between the measured geosmin concentration and predicted geosmin^[7] concentration (7 days in advance) in the canal using 25% (boot-strapped) of the 10 years development dataset.

b Comparison between the measured geosmin concentration and predicted geosmin^[7] concentration (7 days in advance) using 3 years' unseen data from the canal.

From the correlation results in Figures 5.26 and 5.28 it is evident that geosmin^[0] and geosmin^[7] predictions were unsuccessful and more research is necessary into 1) the contribution of actinomycetes to geosmin production in the canal as well as 2) the specific environmental variables involved in triggering cyanobacteria to produce geosmin in the canal. If cyanobacteria are responsible for the production of most of the geosmin, it seems that the variables involved in the production of geosmin have not been captured in developing these models.

5.6 Discussion and conclusions

From the models discussed in this chapter, it is evident that prediction models for cyanobacteria and certain cyanobacterial-related organic compounds can be effectively developed for application in the drinking water treatment industry.

The summary of the results from all the models discussed in this chapter is presented in Table 5.1. From the results it is evident that models with a shorter prediction time are statistically more accurate, as expected. Testing the models with previously “seen” data (i.e. 25% of the boot-strapped results from the testing dataset that was part of the 10 years development dataset) gave a better correlation than testing it with “unseen” data. Although both validations (using the “seen” and “unseen” data) were taken into account when deciding if the models were fit for use, more emphasis was placed on the validation with the “unseen” data as this will give a better indication of the robustness of the models. The models predicting *Anabaena* sp., *Microcystis* sp. and microcystin gave relatively good results (except for the *Anabaena*^[21] model, which was not significant). The geosmin models were also not regarded as significantly correlated – indicating that the variables used to develop the models, did not include the most important drivers causing cyanobacteria to produce geosmin and that actinomycetes (Zaitlin and Watson, 2006) may probably also be involved in producing geosmin in the canal. The root mean square errors (RMSE) in general (Table 5.1) also seem to increase with increasing prediction time (except for the *Anabaena*^[21], *Microcystis*^[21], microcystin^[0], microcystin^[21] and geosmin^[7] models tested with the “unseen” data).

The *Anabaena*, *Microcystis* and microcystin models for real-time prediction and forecasting proved to be very successful. Welk and co-workers (2008) published similar single rule models for Chlorophyll-*a* and *Microcystis* sp. 7 days in advance as well as a model for the prediction of *Oscillatoria* sp. 5 days in advance, in different reservoirs in Australia. According to Welk et al. (2008) these models showed R²-values of 0.42 and 0.38 (Chlorophyll-*a* 7 days in advance), 0.39 and 0.48 (*Microcystis* sp. 7 days in advance) and 0.63 and 0.62 (*Oscillatoria* sp. 5 days in advance). Cao and co-workers (2013) developed similar single rule models 7 days in advance for Chlorophyll-*a* in the Wivenhoe reservoir in Queensland Australia, and found R²-values of 0.32 and 0.24 which could be optimised with multiple rules to 0.37 and 0.27 and with hill climbing and differential evolution to reach R²-values of 0.50, 0.38, 0.59 and 0.44 respectively. They also developed single rule models for the prediction of *Cylindrospermopsis*, which gave R²-values of 0.53 and 0.46. These models could also be enhanced by employing multiple rules which increased the R²-values to 0.57 and 0.47. Optimisation with hill climbing and differential

evolution resulted in R^2 -values of 0.69, 0.55, 0.72 and 0.58 respectively (Cao et al., 2013). It is noteworthy that the models by Welk and co-workers (2008) as well as by Cao and co-workers (2013) were not tested with “unseen” data but all correlations reported were that of the 25% boot-strapped testing data set that was not incorporated into the training dataset. The single rule models with no optimisation for *Anabaena*^[0], *Anabaena*^[7], *Microcystis*^[0], *Microcystis*^[7], Microcystin^[0] and Microcystin^[7] (as summarised in Table 5.1) proved to perform even better than the optimised models published by Cao et al., (2013). The reason for the good performance of the models in this study is most probably because of the reliability of input data as well as the inclusion of initial cyanobacteria inoculum (Table 5.2) as input variable. The inclusion of initial cyanobacteria inoculum is unique when compared to similar models developed by other authors (Chan et al., 2007; Van Ginkel, et al., 2007; Recknagel, et al. 2008, Van Ginkel, 2008; Welk et al., 2008; Cao et al., 2013; Recknagel et al., 2013) and its inclusion as input towards which the output is most sensitive to, is most probably the reason why the models in this study performed better than most of the models published by other authors, even those optimised by Cao et al., (2013).

Table 5.1 Summary of all statistical results of models developed from historical laboratory generated data.

MODEL	Tested with 25% boot-strapped data		Tested with 3 years “unseen” data		Significantly correlated?
	R^2 -value	RMSE	R^2 -value	RMSE	
<i>Anabaena</i> ^[0]	0.89	1 944 cells/mL	0.79	1 026 cells/mL	✓
<i>Anabaena</i> ^[7]	0.86	2 127 cells/mL	0.35	1 745 cells/mL	✓
<i>Anabaena</i> ^[14]	0.80	2 575 cells/mL	0.37	3 726 cells/mL	✓
<i>Anabaena</i> ^[21]	0.76	2 786 cells/mL	0.042	1 213 cells/mL	✗
<i>Microcystis</i> ^[0]	0.95	2 462 cells/mL	0.97	4 767 cells/mL	✓
<i>Microcystis</i> ^[7]	0.91	3 136 cells/mL	0.59	43 479 cells/mL	✓
<i>Microcystis</i> ^[14]	0.79	4 494 cells/mL	0.37	48 130 cells/mL	✓
<i>Microcystis</i> ^[21]	0.74	4 994 cells/mL	0.25	18 494 cells/mL	✓
Microcystin ^[0]	0.88	0.216 µg/L	0.23	0.604 µg/L	✓
Microcystin ^[7]	0.75	0.27 µg/L	0.27	0.525 µg/L	✓
Microcystin ^[14]	0.63	0.31 µg/L	0.28	0.57 µg/L	✓
Microcystin ^[21]	0.58	0.33 µg/L	0.41	0.42 µg/L	✓
Geosmin ^[0]	0.49	5.61 ng/L	0.02	7.54 ng/L	✗
Geosmin ^[7]	0.25	5.78 ng/L	0.007	3.67 ng/L	✗

The frequencies of the different input variables used in the models to predict *Anabaena* sp., *Microcystis* sp., microcystin and geosmin are displayed in Table 5.2 ranging from the most

frequently included variable to the least frequently included in the models. The variables used to predict the *Anabaena* sp. concentrations in the Vaal Dam were the initial cyanobacteria inoculum, NH_4^+ , COD, pH, Fe, PO_4^{3-} , NO_2^- , chlorophyll-*a*, Si, DO, turbidity, conductivity, NO_3^- , Secchi disk depth and Mn (in order of most to least frequently incorporated into the models – Table 5.2).

Table 5.2 Frequency at which input variables were used in models to predict *Anabaena* sp., *Microcystis* sp., microcystin and geosmin concentrations.

Variable used in models	Number of occurrences in <i>Anabaena</i> models	Number of occurrences in <i>Microcystis</i> models	Number of occurrences in microcystin models	Number of occurrences in geosmin models	Frequency
Cyanobacteria inoculum (Cyano)	4	4	3	2	13/14 = 93%
Phytoplankton biomass (Chl <i>a</i> / Chl665)	1	2	4	1	8/14 = 57%
pH	2	1	4	1	8/14 = 57%
Water Temperature (Temp)	-	1	4	2	7/14 = 50%
NH_4^+	3	1	2	-	6/14 = 43%
Si	1	1	3	1	6/14 = 43%
Dissolved Oxygen (DO)	1	4	1	-	6/14 = 43%
Turbidity (Turb)	1	2	2	1	6/14 = 43%
Fe	2	2	-	-	4/14 = 29%
Chemical Oxygen Demand (COD)	3	1	-	-	4/14 = 29%
Conductivity (Cond)	1	1	1	-	3/14 = 21%
PO_4^{3-}	2	1	-	-	3/14 = 21%
NO_3^-	1	2	-	-	3/14 = 21%
Secchi disk depth (Secchi)	1	1	-	-	2/14 = 14%
Mn	1	1	-	-	2/14 = 14%
NO_2^-	2	-	-	-	2/14 = 14%
Dissolved Organic Carbon (DOC)	-	-	-	1	1/14 = 7%

Initial cyanobacteria inoculum was incorporated into all *Anabaena* sp. models and according to the sensitivity analyses showed to have the largest influence on these models. It should be noted however, that this may only be true if the largest percentage of cyanobacteria are comprised of *Anabaena* sp. cells (as is usually the case in the Vaal Dam during the first part of the summer). The nutrients (PO_4^{3-} , NO_3^- , NO_2^- , NH_4^+) separately or in combination with each other, were also included in all *Anabaena* sp. models. At least one of the nutrients (either NH_4^+ , PO_4^{3-} , NO_3^- or NO_2^-) are incorporated into all *Anabaena* sp. models, with NH_4^+ occurring as the most frequently incorporated nutrient. Nutrient concentrations are usually associated with future occurrences of high cyanobacterial cell concentrations – as high concentrations may

cause growth of cells. Therefore it was unexpected to find PO_4^{3-} occurring in the *Anabaena*^[0] model. However, the occurrence of PO_4^{3-} in the *Anabaena*^[0] model, may be due to dimensionality and in the data (Cabestany et al., 2005). In addition, according to the sensitivity analysis of the *Anabaena*^[0] model, the effect of the PO_4^{3-} concentration seem to be very small on the final result of the model.

The variables used to predict *Microcystis* sp. concentration in the Vaal Dam, were initial cyanobacteria inoculum, DO, Chl_a, turbidity, Fe, NO_3^- , pH, water temperature, NH_4^+ , Si, COD, conductivity, PO_4^{3-} , Secchi and Mn (in order of most to least frequently incorporated into the models) (Table 5.2). Initial cyanobacteria inoculum and dissolved oxygen (DO) were incorporated in all of the models. At least one of the nutrients (either NH_4^+ , PO_4^{3-} , or NO_3^-) are incorporated in all *Microcystis* sp. models except for the real-time *Microcystis*^[0] model. The reasons for this may be that the real-time *Microcystis* sp. concentration (*Microcystis*^[0]) cannot be influenced by the change in nutrient concentration of 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 *Microcystis* sp. concentrations show the greatest sensitivity towards the initial cyanobacteria inoculum, provided that the total cyanobacteria inoculum comprise mostly of *Microcystis* sp. cells (as is usually the case during the latter part of the summer season in the Vaal Dam). The *Microcystis*^[21] model also shows sensitivity towards the silica concentration and particularly during the first 10% of the silica input range. Silica might represent a secondary effect on *Microcystis* sp. concentration, since Si is mostly utilised by diatoms for the production of their frustules (Davidson et al., 2012) and this usually occur in the Vaal Dam in winter when cyanobacteria such as *Microcystis* sp. are not abundant.

The variables used to predict the microcystin concentration in the Vaal Dam were water temperature, chlorophyll-*a* concentration, pH, initial cyanobacteria inoculum, Si, NH_4^+ , turbidity, DO and conductivity (in order of most to least frequently incorporated into the models – Table 5.2). Water temperature, chlorophyll-*a* and pH were incorporated into all models predicting microcystin concentration and initial cyanobacteria inoculum into 3 of the 4 models (all except the microcystin^[21] model). Proxies for toxic cyanobacterial cell biomass (such as chlorophyll-*a* and to a lesser extent initial cyanobacteria inoculum) are regarded to sometimes give misleading information regarding microcystin concentration (Jähnichen et al., 2011). This fact is acknowledged, since microcystin is strongly correlated with the cell number of toxic cyanobacterial cells (Joung et al., 2011), which only form part of the chlorophyll-*a* concentration and the initial cyanobacteria inoculum in the Vaal Dam. The nutrient NH_4^+ was incorporated into the microcystin^[14] and microcystin^[21] models, with no nutrients included in

the microcystin^[7] model. This may indicate that nutrient levels (and particularly that of NH_4^+) only affect microcystin concentration after 14 days. The sensitivity analyses of the models predicting microcystin concentration reveals the greatest sensitivity towards water temperature, cyanobacteria inoculum and chlorophyll-*a* concentration, conductivity and Si concentration. According to literature, the microcystin concentration increases with increasing nutrient concentration, light intensity, and temperature (Jacoby et al., 2000; Sekadende et al., 2005; Wu et al., 2006; Heisler et al., 2008), as well as cell number (Joung et al., 2011). From the studies of Joung et al. (2011) it seems that nutrients such as TP (total phosphorus) induce the growth of toxic as well as non-toxic *Microcystis* strains, which may be the reason why nutrients do not have such a large impact on the microcystin models in the Vaal Dam. Joung et al., (2011) also found that water temperature was the single most important variable favouring the occurrence of toxic strains (and therefore also the production of microcystin), where higher water temperatures fostered a faster growth of toxic *Microcystis* strains rather than non-toxic *Microcystis*. Therefore the conclusion made by Paerl and Huisman (2008), that global warming could lead to more toxic cyanobacterial blooms, may also be valid in the Vaal Dam context, since all of the microcystin models from the Vaal Dam show great sensitivity towards increasing water temperatures. The sensitivity towards Si concentration (as in the case of *Microcystis* sp. predictions models), may be secondary in nature due to the utilization of Si during winter when diatoms are abundant in the Vaal Dam.

The models for the prediction of *Microcystis* sp., *Anabaena* sp. and microcystin have proved to be very successful for as far as 7 days in advance and show the most promising application value for incorporation into the Water Safety Plan for the early warning of increasing cyanobacterial cell concentrations and increasing microcystin concentrations of the Zuikerbosch DWTW. This will benefit the DWTW as to enable them to start preparing for a high cyanobacterial or microcystin incident before it happens. The models for the prediction of geosmin concentrations, however, were unsuccessful and more research is necessary into the contribution of actinomycetes to geosmin as well as the specific environmental variables involved in triggering cyanobacteria to produce geosmin in the canal.

The models discussed in this chapter utilised all possible input variables e.g. nutrients such as PO_4^{3-} , NH_4^+ , NO_3^- , NO_2^- as well as variables such as chemical oxygen demand (COD), dissolved organic carbon (DOC), Mn, Fe etc., which is mostly only measureable in a large water chemistry laboratory. These models that can incorporate so many input variables obviously have the best opportunity to forecast the cyanobacterial cell concentrations and cyanobacterial-related organic compounds more accurately than models with limited input variables. However, many

DWTW (especially those from rural areas and small municipalities) do not have large laboratory facilities able to perform all of these analyses on a regular basis, or even if they do have the capacity for such analyses it may take too long for the results to be generated and incorporated into such forecasting models. Therefore, another aim of this study was to (in addition to the compound models described in this chapter) also develop simpler models to forecast *Anabaena* sp., *Microcystis* sp. and microcystin with consciously chosen variables that can be measured by on-line (*in situ*) instruments which do not require expensive laboratory facilities or can be more readily available for incorporation into the models. These variables include water temperature, turbidity, dissolved oxygen, pH, conductivity, chlorophyll-*a* and initial cyanobacteria inoculum, all of which can be measured by relatively inexpensive (when compared to laboratory analyses) YSI instrumentation such as the YSI 6600 V2 Multi-parameter water quality sonde, or the YSI EXO range of water quality sondes. Models developed to include only these limited input variables are discussed in Chapter 6.

CHAPTER 6

USING HYBRID EVOLUTIONARY ALGORITHMS FOR THE DEVELOPMENT OF LIMITED INPUT RULE-BASED FORECASTING MODELS FOR CYANOBACTERIA

6.1 Introduction

In this chapter, the forecasting models developed from a limited range of specifically chosen variables for predicting different cyanobacteria genera, as well as their associated cyanotoxin “microcystin” are discussed. The input variables were specifically chosen because they are measureable by on-line (*in situ*) instruments (i.e. water temperature, turbidity, conductivity, pH, dissolved oxygen, chlorophyll-*a* and initial cyanobacteria inoculum) and the application value for using on-line (*in situ*) generated data was investigated. Ten years of laboratory generated data (2000 – 2009) in the Vaal Dam supplying the Zuikerbosch DWTW (drinking water treatment works) was used for the development of these models. As in the case of the compound models discussed in Chapter 5 (which include all possible input variables) four different predicting-time scenarios were investigated namely *real-time* (same day), *7 days in advance*, *14 days in advance* and *21 days in advance*. Due to the fact that especially with the *in situ* data, increasing prediction time culminated into decreased statistical significant correlations between measured and forecasted data, only the *real-time* (indicated by ^[0]) and *7 days in advance* (indicated by ^[7]) forecasting models are discussed in this chapter. Fifty rule-based models were developed for each predicting-time, at 1) the Vaal Dam only 2) canal only and 3) Vaal Dam and canal combined, of which only the best (highest R-squared in comparison to measured data) were chosen from all three to be discussed in this chapter. Models were tested with the 25% boot-strapped testing dataset, 3 years unseen laboratory generated data and 2 years unseen on-line (*in situ*) generated data from the YSI 6600 V2, Multi-parameter water quality sonde situated in the canal (refer to Figure 3.1 page 22). The reason for using only 2 years of unseen on-line (*in situ*) generated data, was that initial problems were experienced with the on-line instrument regarding installation, calibration and verification and only 2 years of uninterrupted data was available for testing. The other models (e.g. the *14 days in advance* and *21 days in advance* prediction models) are included on the CD at the back page of this thesis.

6.2 Models for the prediction of *Anabaena* sp. concentration

Anabaena sp. is a cyanobacterium well-known for the production of cyanotoxins (Du Preez and Van Baalen, 2006; Newcombe et al, 2010). In the Vaal Dam and canal investigated in this study, *Anabaena* sp. is often the dominant phytoplankton species during spring and summer in the source water to the Zuikerbosch DWTW (Ewerts et al., 2013) and has been linked to the taste and odour compound geosmin (Swanepoel et al., 2008a). Therefore, models predicting the occurrence of *Anabaena* sp. will be of great benefit to the Zuikerbosch DWTW in order for it to prepare for incidents of high *Anabaena* sp. and potentially high geosmin or cyanotoxin concentrations in the water.

6.2.1 *Anabaena*^[0] prediction (in real-time)

The IF criterium of the model (Figure 6.1) is determined by the conductivity of the water. The THEN branch of the model (Figure 6.1a) represents the low range rule set and the ELSE branch (Figure 6.1b) represents the high range rule set. Both the THEN and ELSE branches of the model show the greatest sensitivity towards the initial cyanobacteria inoculum. Sensitivity towards dissolved oxygen (DO) in the THEN and ELSE branches as well as turbidity in the ELSE branch of the model have also been established.

The comparison between the measured *Anabaena* sp. concentration and that of the results from the models predicting real-time *Anabaena*^[0] when using the 25% boot-strapped testing dataset (Figure 6.2a), shows a R^2 -value of 0.89 and a RMSE (root mean square error) of 1912 cells/mL. When the model was tested with 3 years of unseen laboratory generated data, the correlation showed a R^2 -value of 0.01 and a RMSE of 9458 cells/mL (Figure 6.2b). When the model was tested with 2 years of unseen on-line (*in situ*) generated data, the correlations showed a R^2 -value of 0.01 and RMSE of 9587 cells/mL (Figure 6.2c). From these results it is evident that the data generated by laboratory analyses (Figure 6.2a and 6.2b) shows a much better correlation than data generated by the *in situ* YSI instrument (Figure 6.2c).

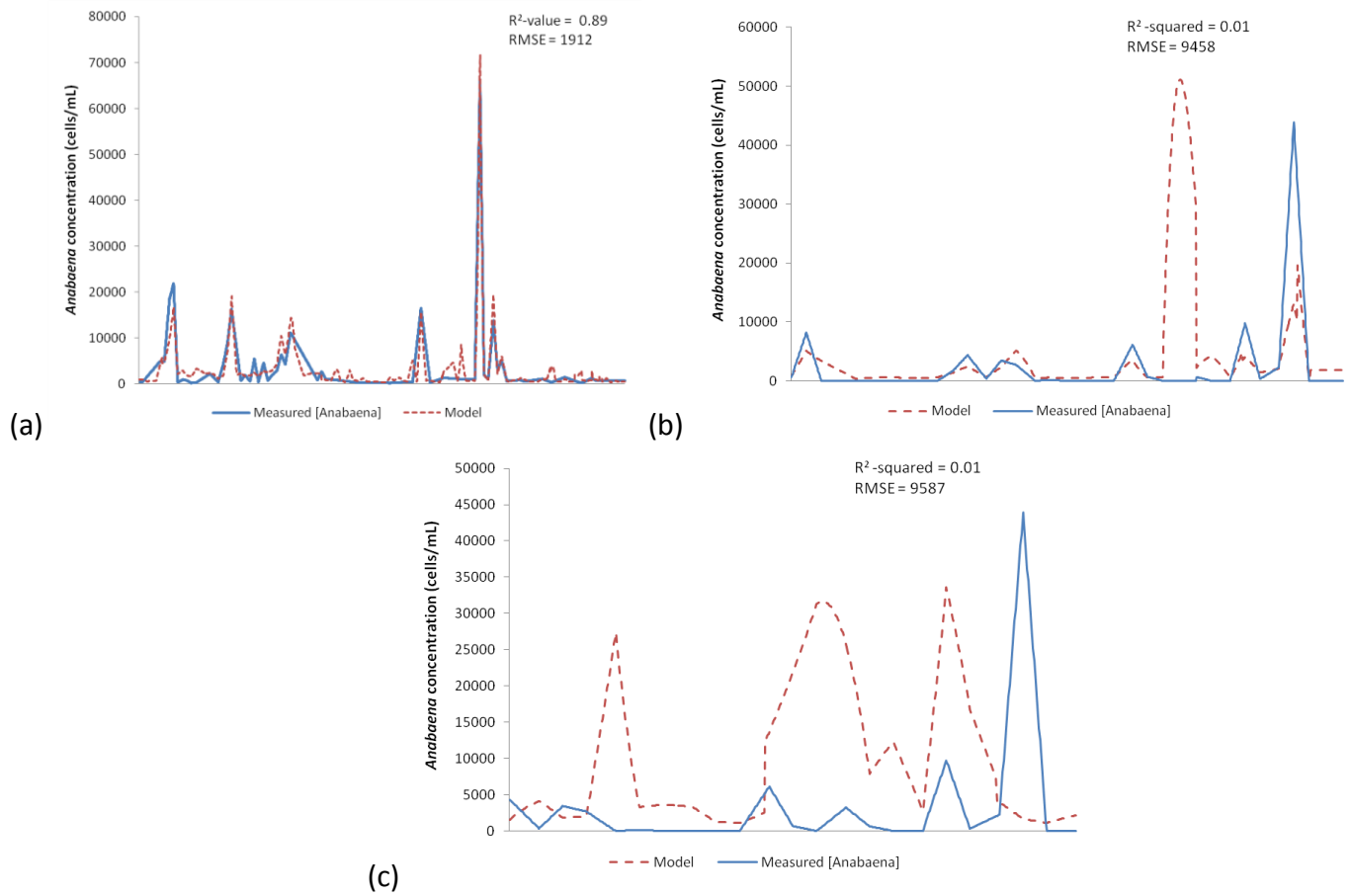


Figure 6.2 a Comparison between the measured *Anabaena* sp. concentration and predicted real-time *Anabaena*^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured *Anabaena* sp. concentration and predicted real-time *Anabaena*^[0] concentration using 3 years' unseen data from the Vaal Dam.
 c Comparison between the measured *Anabaena* sp. concentration and predicted real-time *Anabaena*^[0] concentration using 2 years' *in situ* generated data from the canal.

6.2.2 *Anabaena*^[7] prediction (7 days in advance)

The IF criterium of the model predicting *Anabaena*^[7] concentration (Figure 6.3) is determined by the turbidity of the water. The THEN branch represents the high range rule set and the ELSE branch represents the low range rule set. Both the THEN and ELSE branches of the model (Figure 6.3a and 6.3b) show the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables in the model (pH, conductivity, and turbidity) display very little influence on the predicted *Anabaena*^[7] concentration.

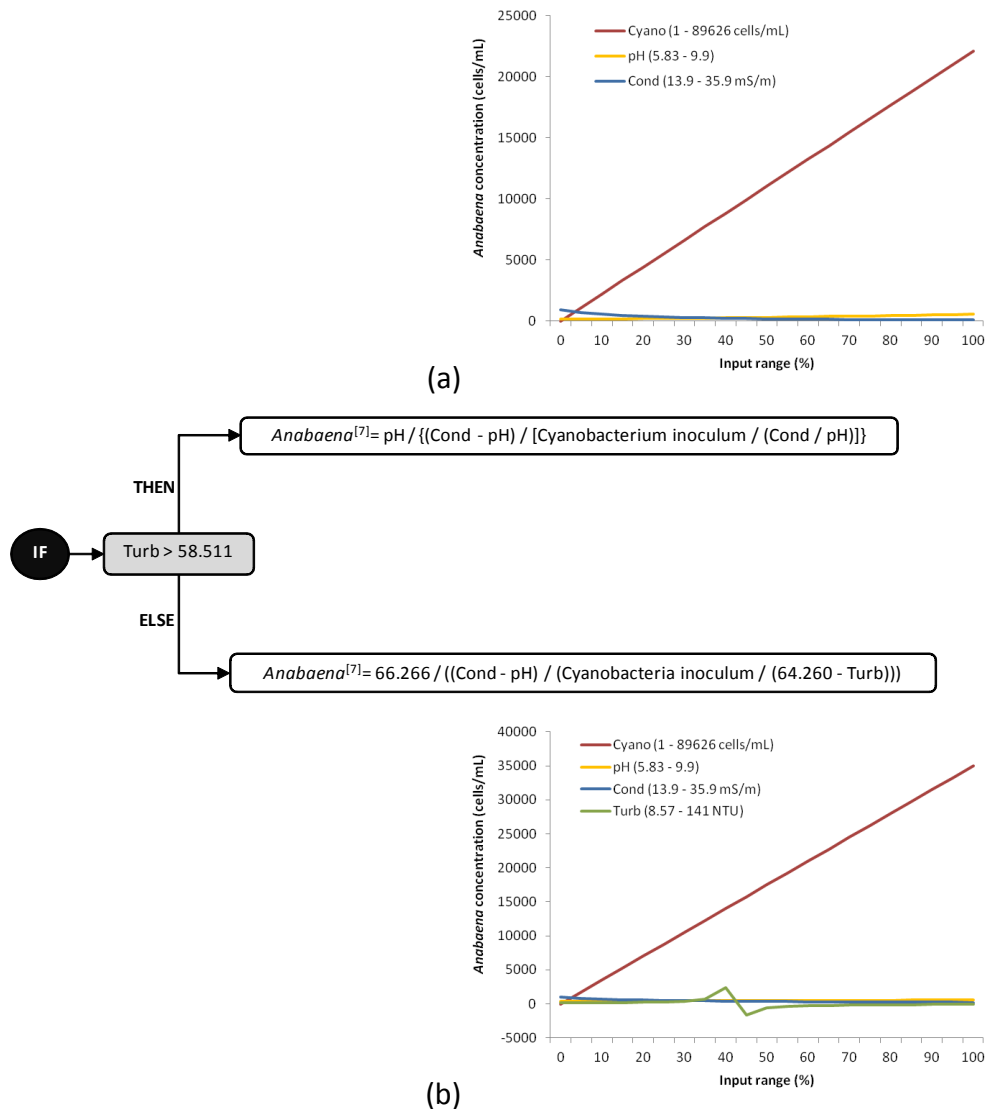


Figure 6.3 Forecasting model and associated sensitivity analyses for *Anabaena*^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the real *Anabaena* sp. concentration and that of the results from the models predicting *Anabaena*^[7] concentrations (7 days in advance) when using the 25% bootstrapped testing dataset (Figures 6.4a), showed R^2 -value of 0.76 and a RMSE of 2771 cells/mL. When the models were tested with 3 years of unseen laboratory generated data, the models showed R^2 -values of 0.002 and a RMSE of 19 018 cells/mL (Figure 6.4b). When tested with 2 years of unseen on-line (*in situ*) generated data (Figure 6.4c), it showed a R^2 -value of 0.02 and a RMSE of 3597 cells/mL.

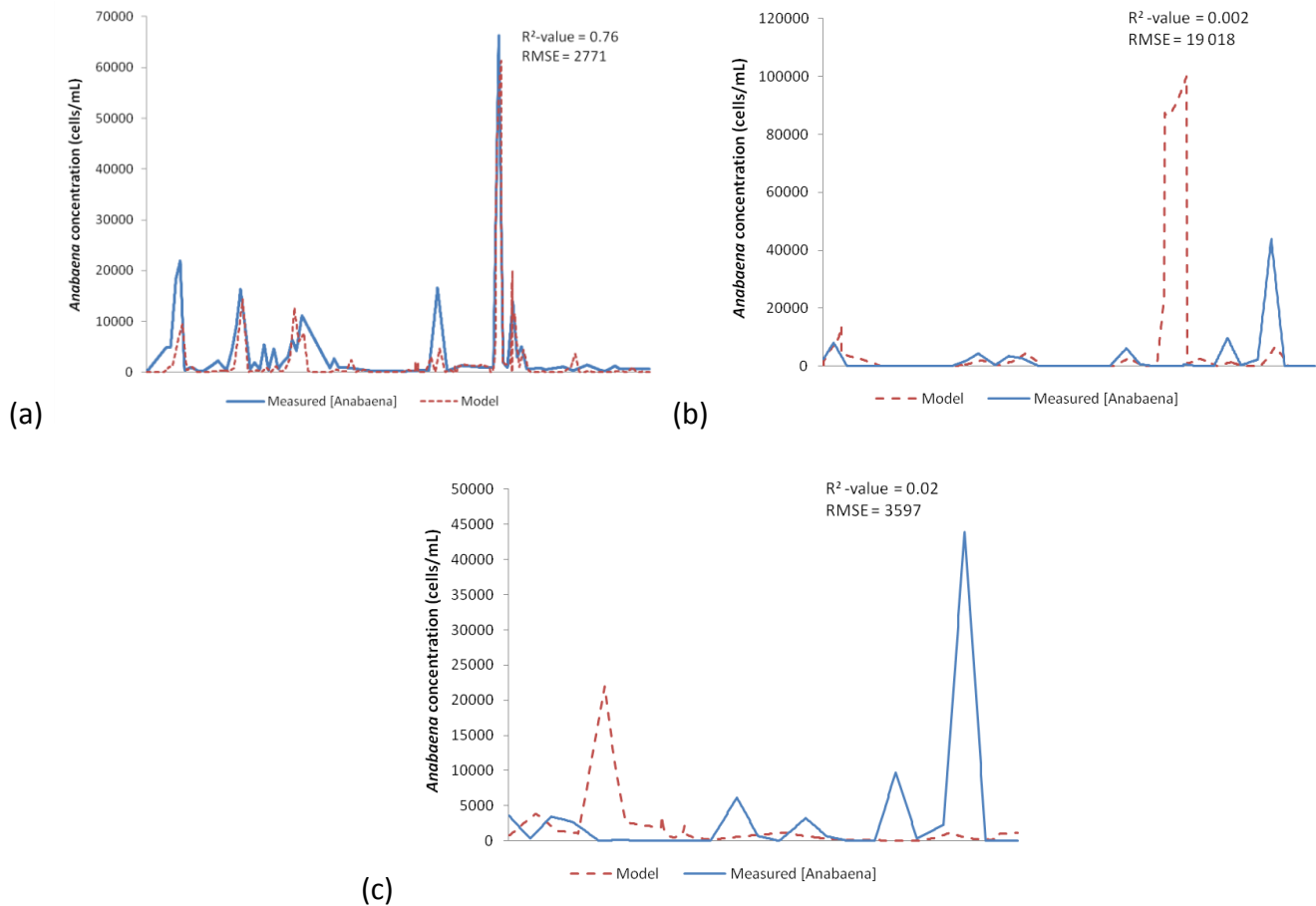


Figure 6.4 a Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[7] concentration using 3 years' unseen data from the Vaal Dam.
c Comparison between the measured *Anabaena* sp. concentration and predicted *Anabaena*^[7] concentration using 2 years' *in situ* generated data from the canal.

The time prediction of increased *Anabaena* sp. concentration seems to be predicted with relatively good accuracy when the input data is laboratory generated (Figure 6.4b); although the magnitude of the concentration is not predicted with such good accuracy. From a drinking water treatment perspective, the magnitude of the increasing cyanobacterial cell concentrations may not be as important as the time-prediction of when increased concentrations (>2000 cells/mL) can be expected. Predicting the correct time of when >2000 cells/mL can be expected will enable the DWTW to start preparations to deal with higher concentrations in the source water, irrespective of the specific expected cyanobacterial cell concentration and therefore this model is still regarded as very useful. Testing the model with on-line (*in situ*) generated data however, does not give a good correlation with the measured *Anabaena* sp. concentration.

6.3 Models for the prediction of *Microcystis* sp. concentration

In the Vaal Dam and canal investigated in this study, *Microcystis* sp. is most commonly known for the production of the taste and odour compound geosmin as well as the cyanotoxin microcystin (Swanepoel et al., 2008a). This organism is also known to occur in relatively large numbers in the Vaal Dam and canal during summer and autumn each year (Ewerts et al., 2013). Therefore, models predicting the occurrence of *Microcystis* sp. can be of great benefit to the Zuikerbosch DWTW, to prepare for incidents of high *Microcystis* sp. concentrations in the source water.

6.3.1 *Microcystis*^[0] prediction (in real-time)

The IF criterium of the model (Figure 6.5) is determined by the conductivity and dissolved oxygen (DO) concentration. The THEN branch of the model (Figure 6.5a) represents the low range rule set and shows the greatest sensitivity towards the initial cyanobacteria inoculum. The ELSE branch of the model (Figure 6.5b) represents the high range rule set and, as in the case of the THEN branch, shows the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables in this model namely conductivity, turbidity, DO and pH display very little influence on the predicted *Microcystis*^[0] concentration.

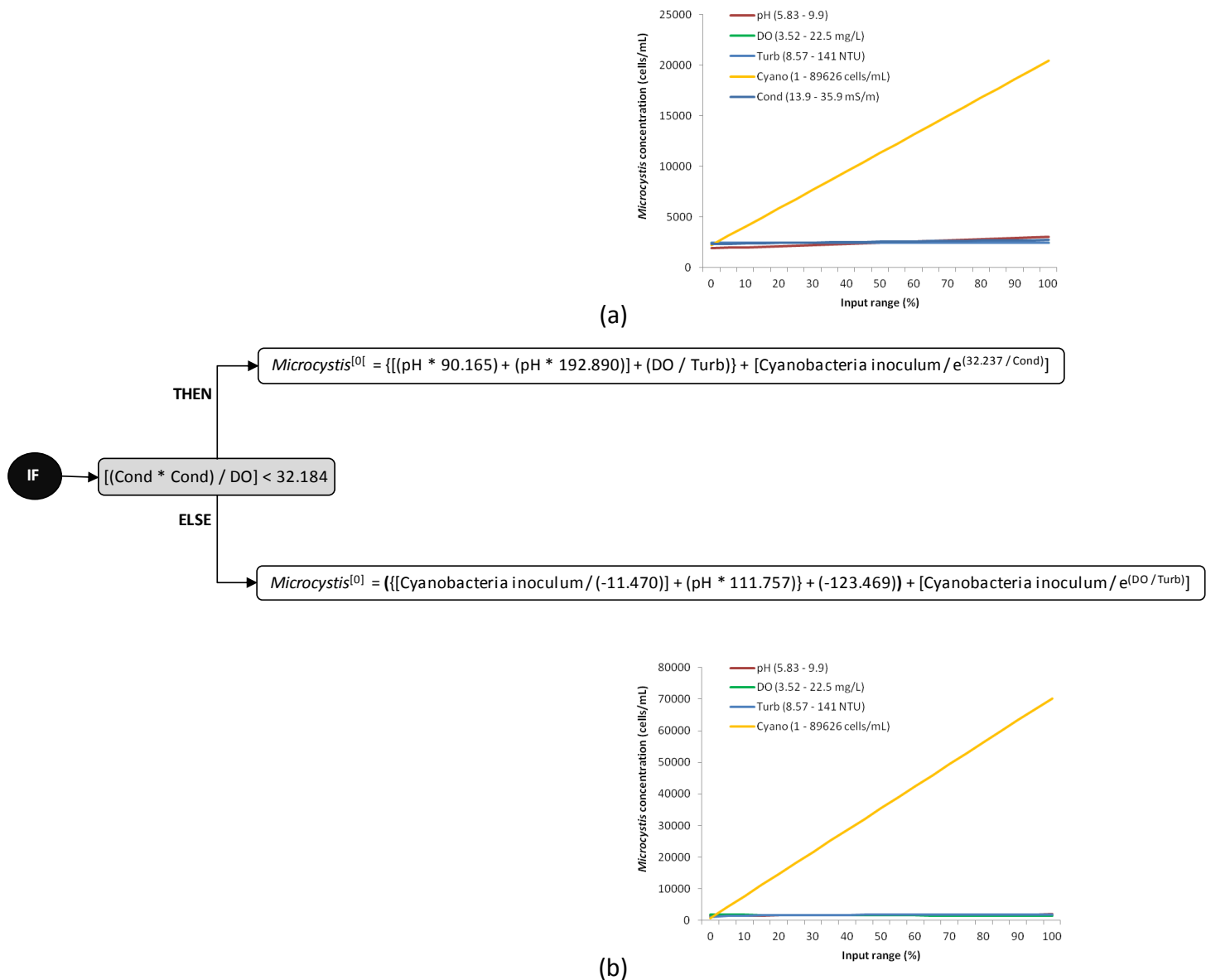


Figure 6.5 Forecasting model and associated sensitivity analyses of real-time *Microcystis*^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting real-time *Microcystis*^[0] when using the 25% boot-strapped testing dataset (Figure 6.6a), shows a R^2 -value of 0.94 and a RMSE of 2522 cells/mL.

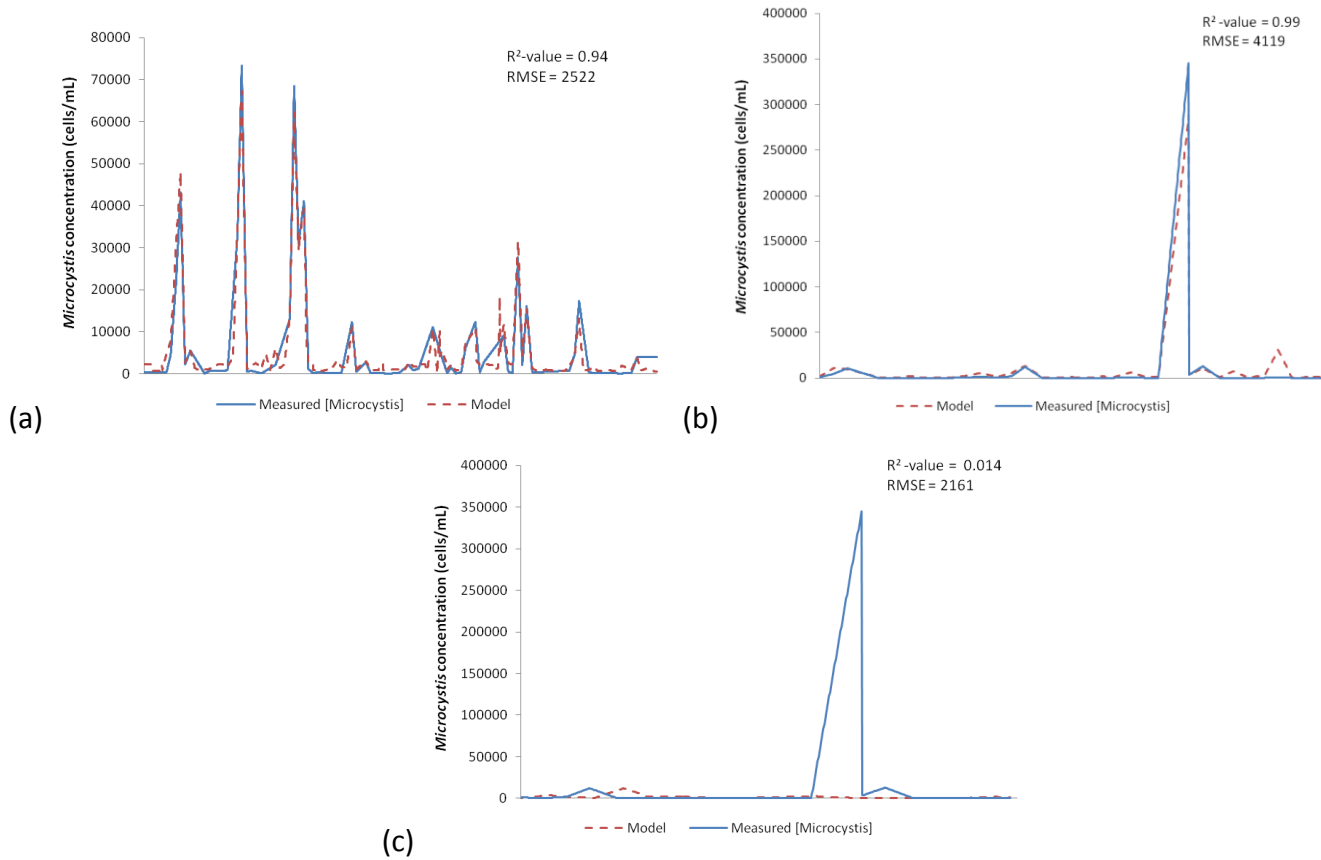


Figure 6.6 a Comparison between the measured *Microcystis* sp. concentration and predicted real-time *Microcystis*^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Microcystis* sp. concentration and predicted real-time *Microcystis*^[0] concentration using 3 years' unseen data from the Vaal Dam.
c Comparison between the measured *Microcystis* sp. concentration and predicted real-time *Microcystis*^[0] concentration using 2 years' *in situ* generated data from the canal.

When tested with three years of unseen laboratory generated data (Figure 6.6b), the correlation coefficient (R^2) increases to 0.99, indicating an excellent correlation. The RMSE is 4119 cells/mL when the results from the model predicting *Microcystis*^[0] is compared to three years unseen laboratory generated data. When the model was tested with 2 years of unseen (*in situ*) generated data, the correlation showed a R^2 -value of 0.014 and a RMSE of 2161 cells/mL (Figure 6.6c). As in the case of the *Anabaena* models, it seems that when testing the prediction models with laboratory generated data, the predicted results compare very well to the measured results, but when tested with the on-line (*in situ*) generated data the models do not perform well.

6.3.2 *Microcystis*^[7] prediction (7 days in advance)

The best model developed for the prediction of *Microcystis*^[7] (7 days in advance) displays an IF criterium that is determined by conductivity. The THEN branch of the model (Figure 6.7a) represents the low range rule set and the ELSE branch of the model (Figure 6.7b) the high range rule set.

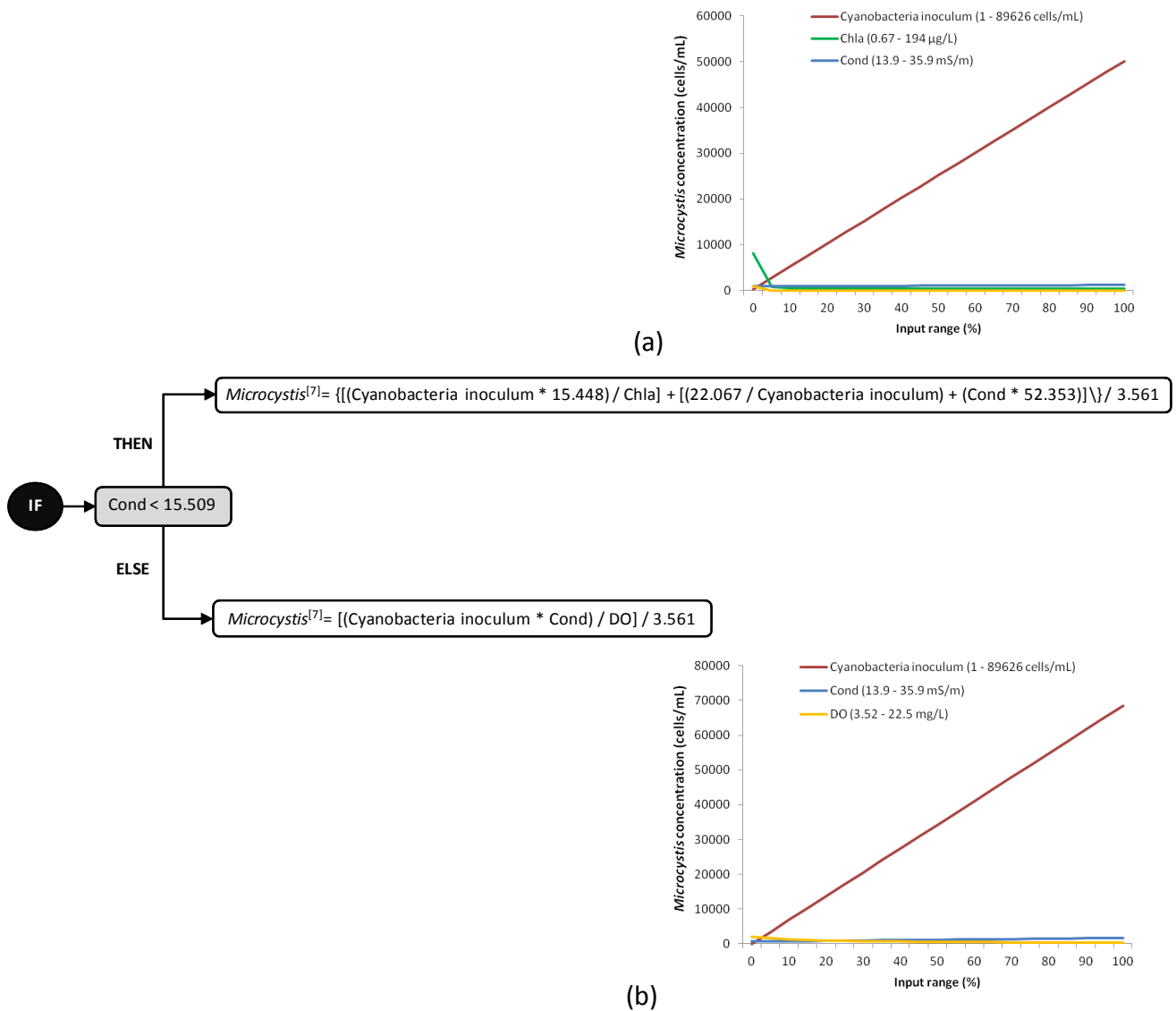


Figure 6.7 Forecasting model and associated sensitivity analyses for *Microcystis*^[7] concentration (7 days in advance) in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

Both the THEN and ELSE branches of the model (Figures 6.7a and 6.7b) show the greatest sensitivity towards the initial cyanobacteria inoculum. The other variables in this model (namely chlorophyll-*a*, conductivity and DO) in comparison to the initial cyanobacteria inoculum, display very little influence on the predicted *Microcystis*^[7] concentration.

The comparison between the measured *Microcystis* sp. concentration and that of the results from the models predicting *Microcystis*^[7] (7 days in advance), when using the 25% bootstrapped testing dataset (Figure 6.8a), shows a R^2 -value of 0.91 and a RMSE of 3354 cells/mL. When the model was tested with 3 years of unseen laboratory generated data, the model showed a R^2 -value of 0.52 and a RMSE of 39 098 cells/mL (Figure 6.8b).

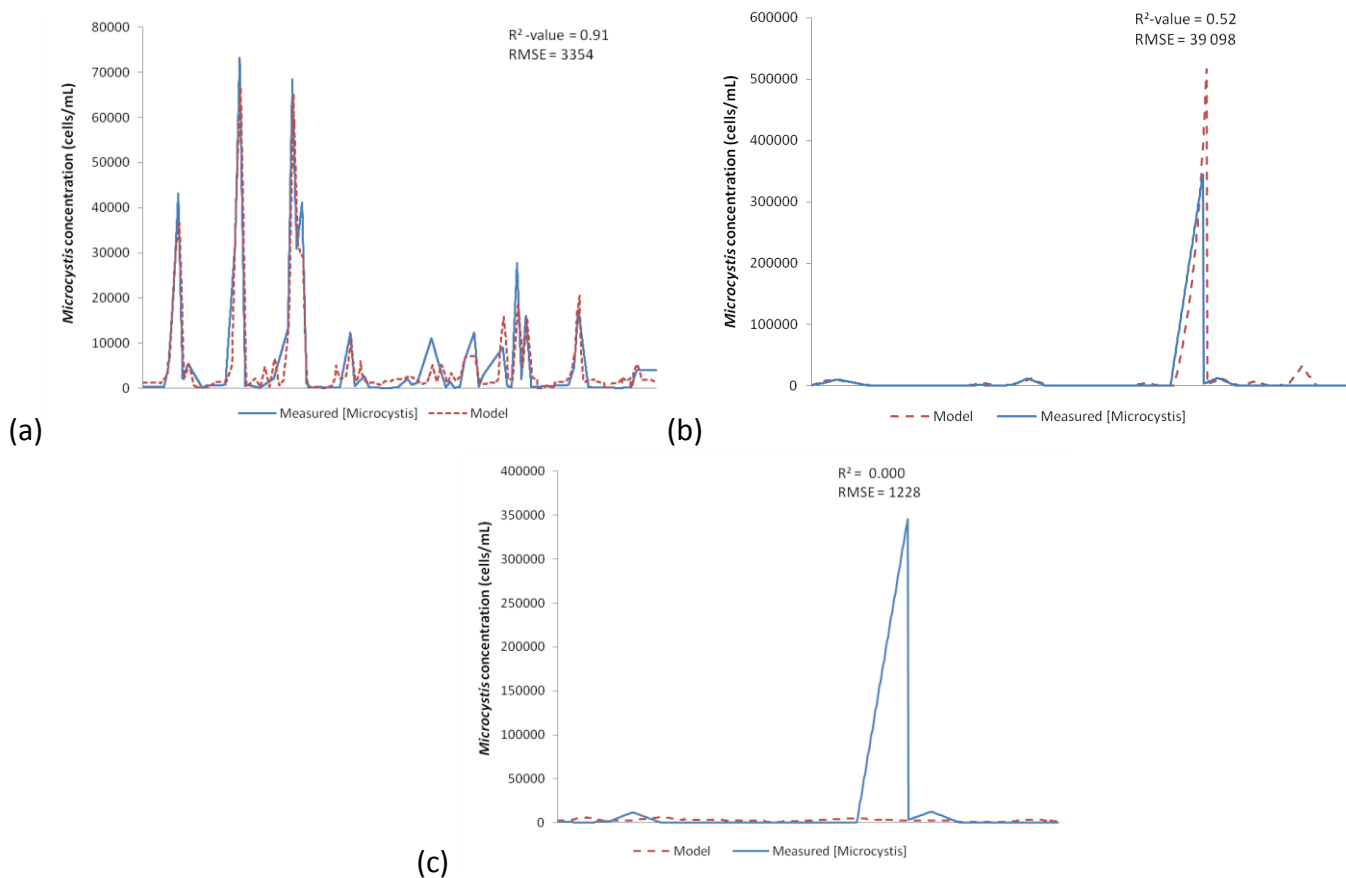


Figure 6.8 a Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
b Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[7] concentration using 3 years' unseen data from the Vaal Dam.
c Comparison between the measured *Microcystis* sp. concentration and predicted *Microcystis*^[7] concentration using 2 years' *in situ* generated data from the canal.

When the model was tested with 2 years of unseen on-line (*in situ*) generated data, the model showed a R^2 -value of 0.00 and a RMSE of 1228 cells/mL (Figure 6.8c), indicating that the model did not perform well when tested with on-line (*in situ*) generated data when compared to laboratory generated data.

6.4 *Models for the prediction of microcystin concentration*

Microcystin, the most commonly known cyanotoxin in the world, is produced by a variety of cyanobacteria (WHO, 2011), some of which are very common to the Vaal Dam and Vaal River especially during the warmer months (Ewerts et al., 2013). The monitoring of this variable has become very important, especially since the incorporation of a guideline limit of 1 µg/L into SANS 241 – the South African National Standard for drinking water (SANS 241-1, 2011). Therefore, the ability to predict the occurrence of high microcystin concentrations in the Vaal Dam will enable the Zuikerbosch DWTW to adjust their processes in advance and reduce the risk to their consumers.

6.4.1 *Microcystin^[0] prediction (in real-time)*

The IF criterium of the model is determined by a combination of the initial cyanobacteria inoculum and the pH (Figure 6.9). The THEN branch of the model (Figure 6.9a) represents the high range rule set and shows the greatest sensitivity towards chlorophyll-*a* and to a lesser extent DO concentration. The ELSE branch of the model (Figure 6.9b) represents the low range rule set and shows the greatest sensitivity towards chlorophyll-*a* concentration. It also shows a significant sensitivity towards turbidity during the 50-70% input range. This phenomenon of a positive-negative-positive sensitivity towards turbidity during the 50-70% input range was unexpected and is hard to explain.

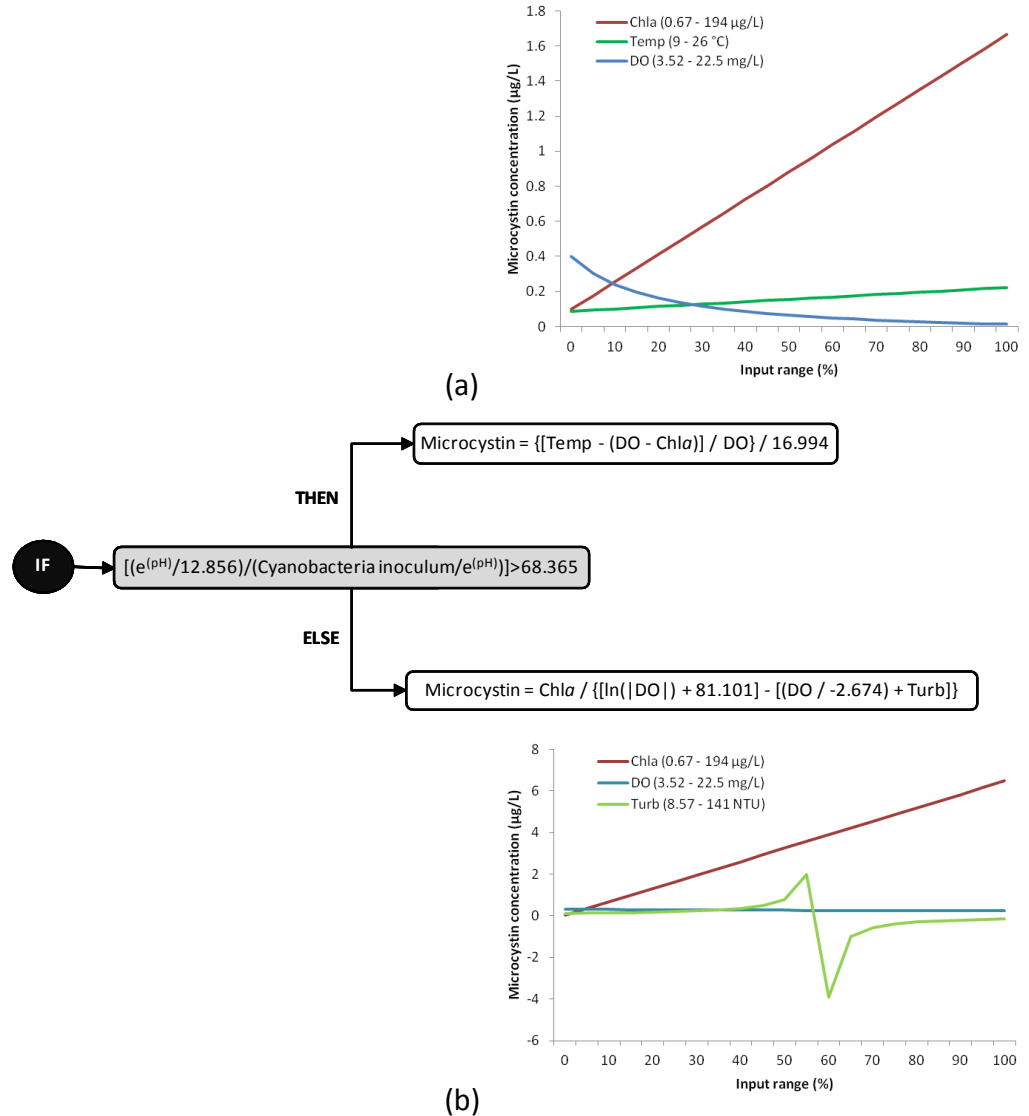


Figure 6.9 Forecasting model and associated sensitivity analyses for real-time microcystin^[0] concentration in the Vaal Dam ((a) represents the sensitivity analysis for the THEN branch and (b) represents the sensitivity analysis of the ELSE branch).

The comparison between the measured microcystin concentration and that of the results from the models predicting real-time microcystin^[0] concentration when using the 25% boot-strapped testing dataset (Figure 6.10a), shows a R^2 -value of 0.83 and a RMSE of 0.23 $\mu\text{g/L}$. When the model was tested with 3 years of unseen laboratory generated data, the model showed a R^2 -value of 0.37 and a RMSE of 0.47 $\mu\text{g/L}$ (Figure 6.10b). Visual inspection of Figure 6.10b reveals that the timing of incidents is predicted fairly well, while the magnitude may not be predicted with such great accuracy. From a DWTW perspective, the magnitude of an incident of high microcystin concentration may not be as important as the time prediction indicating

when such an incident will occur. Therefore this model may still be very useful to apply in the cyanobacterial incident management protocol of a DWTW. When the model was tested with 2 years of unseen on-line (*in situ*) generated data (Figure 6.10c), the R^2 -value decreased to 0.009 and the RMSE was 0.04 $\mu\text{g/L}$. It is evident that while the model performed relatively well when tested with unseen laboratory generated data it did not perform well at all when tested with data from the on-line (*in situ*) instrument.

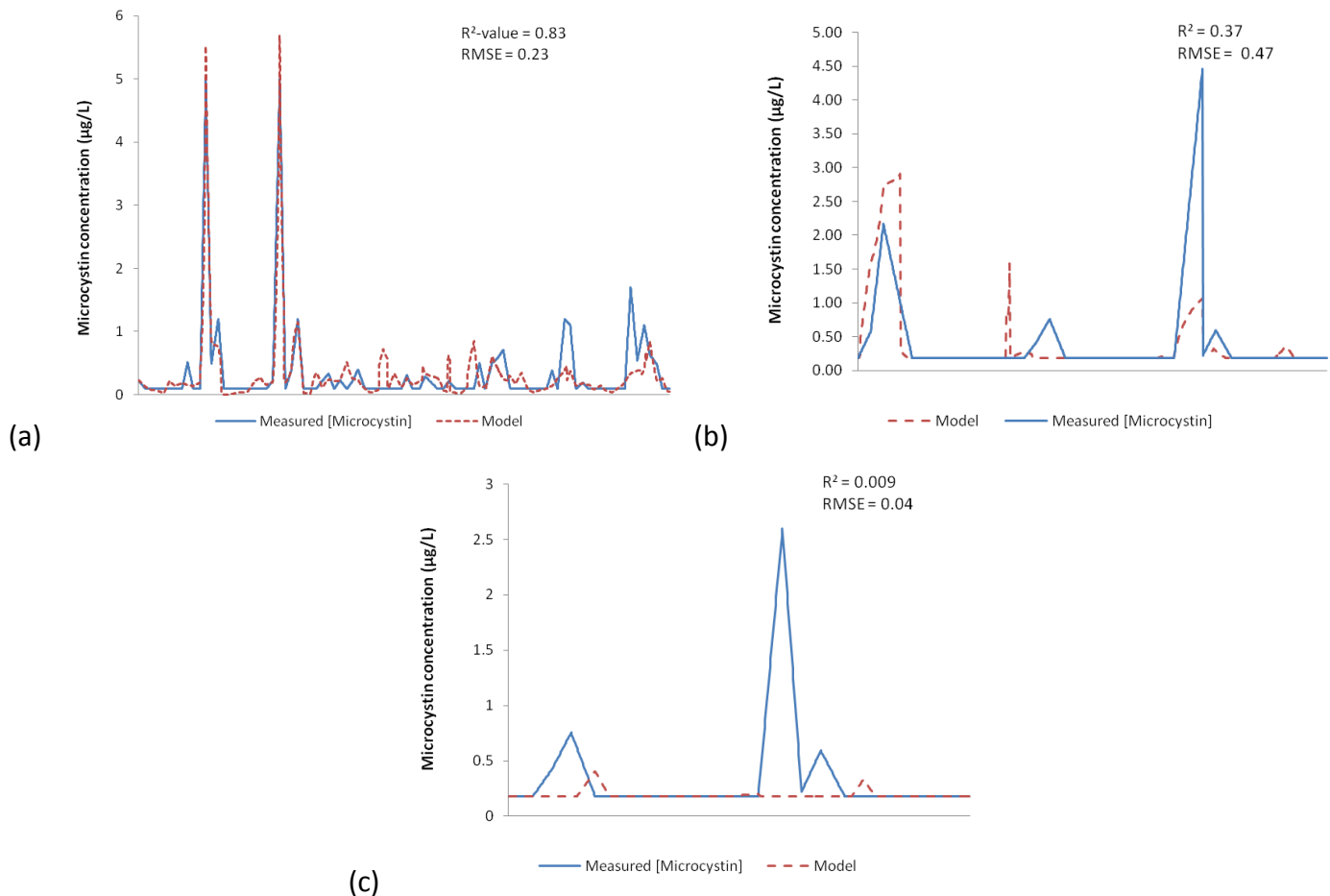
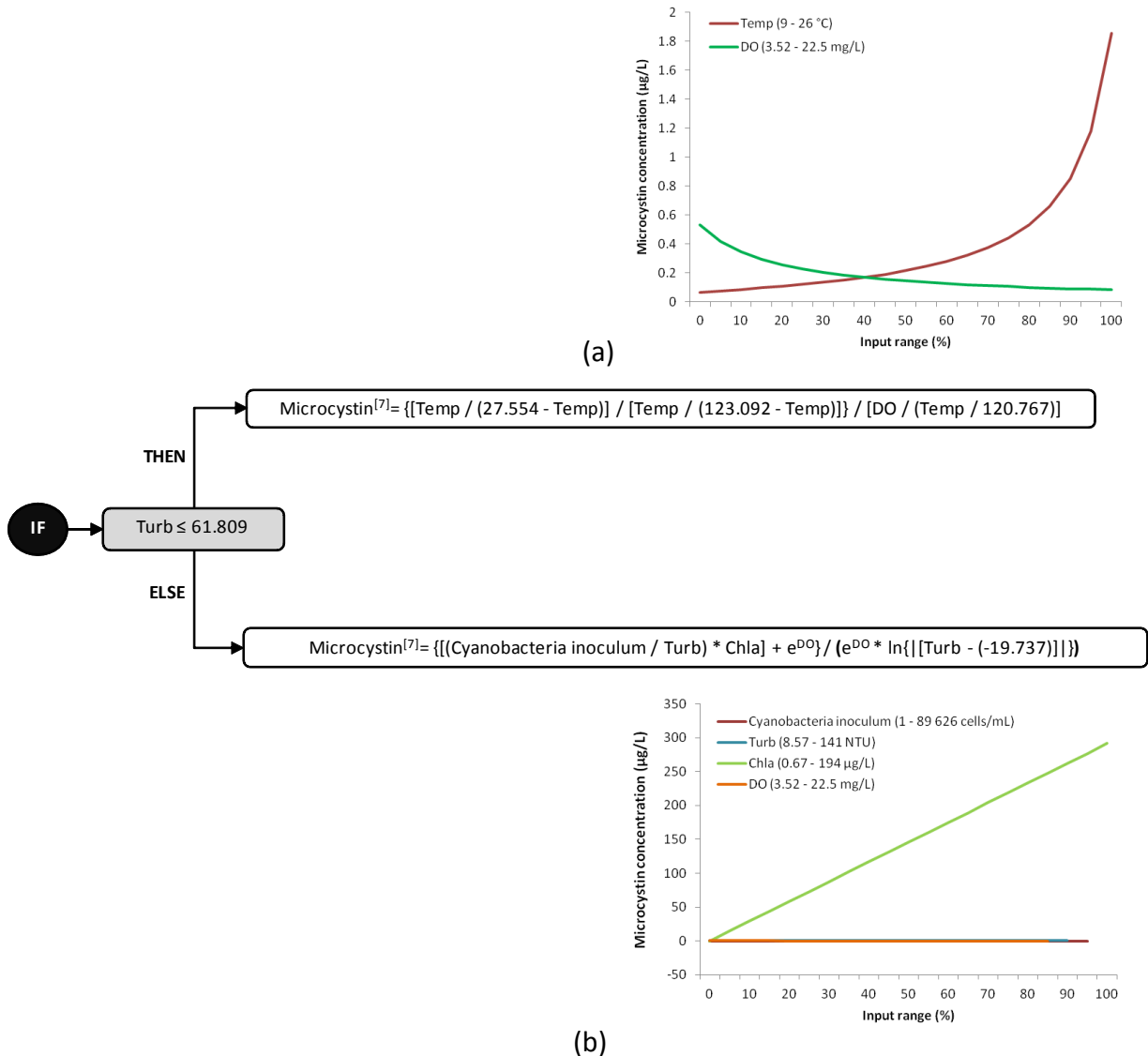


Figure 6.10 a Comparison between the measured microcystin concentration and predicted real-time microcystin^[0] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured microcystin concentration and predicted real-time microcystin^[0] concentration using 3 years' unseen data from the Vaal Dam.
 c Comparison between the measured microcystin concentration and predicted real-time microcystin^[0] concentration using 2 years' *in situ* generated data from the canal.

6.4.2 Microcystin^[7] prediction (7 days in advance)

The IF criterium of the model is determined by the turbidity of the water (Figure 6.11). The THEN branch of the model (Figure 6.11a) represents the low range rule set and shows the greatest sensitivity towards temperature and dissolved oxygen concentration. The ELSE branch of the model (Figure 6.11b) represents the high range rule set and shows the greatest sensitivity towards the chlorophyll-*a* concentration.



In comparison to the chlorophyll-*a* concentration the other variables included in the THEN branch (turbidity, DO and initial cyanobacteria inoculum) display very little influence on the predicted microcystin^[7] concentration.

The comparison between the measured microcystin concentration and that of the results from the models predicting the microcystin^[7] concentration 7 days in advance when using the 25% boot-strapped testing dataset (Figure 6.12a), shows a R^2 -value of 0.75 and a RMSE of 0.29 $\mu\text{g/L}$. When the model was tested with 3 years of unseen laboratory generated data, the model showed a R^2 -value of 0.31 and a RMSE of 0.096 $\mu\text{g/L}$ (Figure 6.12b).

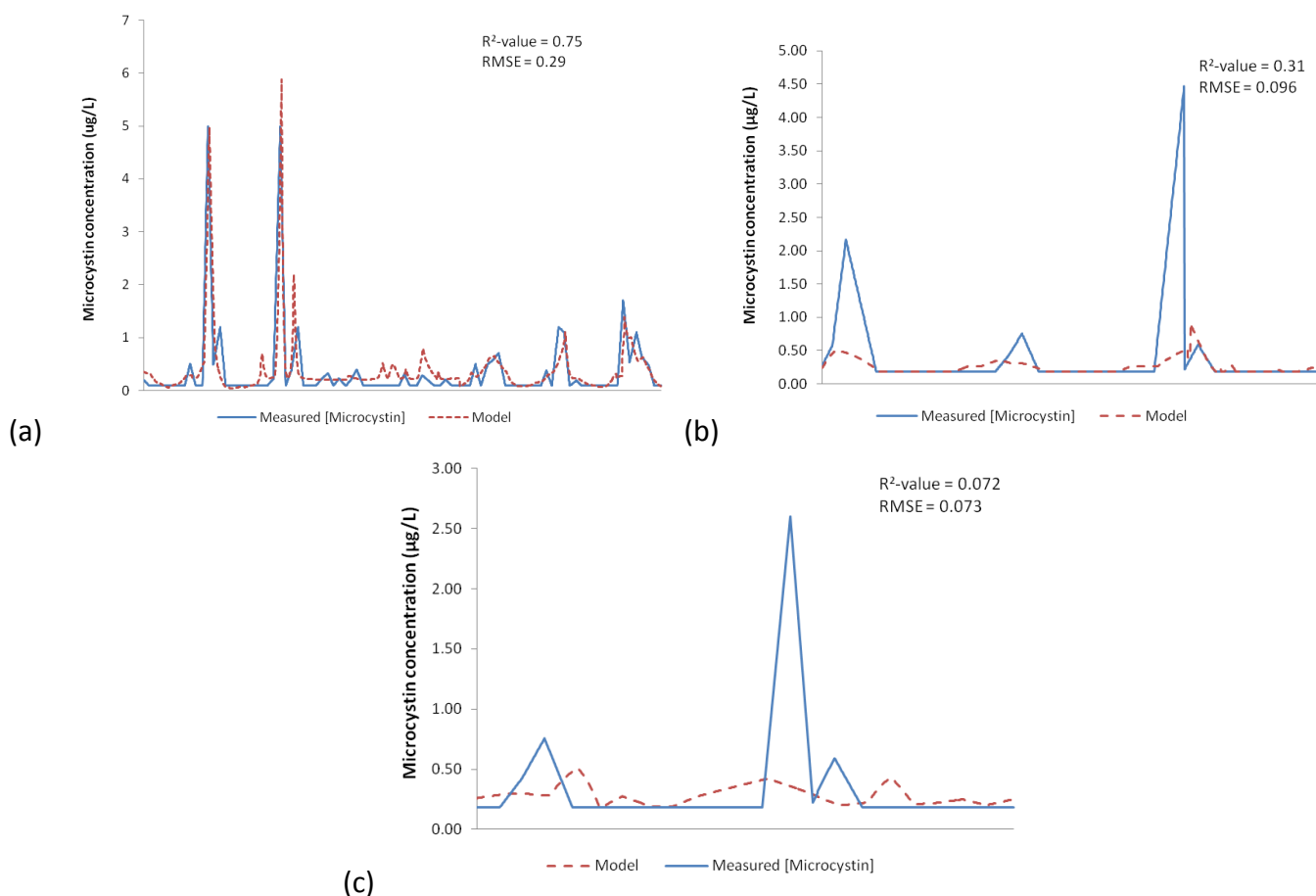


Figure 6.12 a Comparison between the measured microcystin concentration and predicted microcystin^[7] concentration in the Vaal Dam using 25% (boot-strapped) of the 10 years development dataset.
 b Comparison between the measured microcystin concentration and predicted microcystin^[7] concentration using 3 years' unseen data from the Vaal Dam.
 c Comparison between the measured microcystin concentration and predicted microcystin^[7] concentration using 2 years' *in situ* generated data from the canal.

Visual inspection of Figure 6.10b reveals that the timing of incidents is predicted fairly well, while the magnitude may not be predicted with such great accuracy. As discussed in Chapter 5, from a drinking water treatment perspective the magnitude of an incident of high microcystin concentration may not be as important as the time prediction indicating when such an incident will occur. Therefore this model may still be very useful to apply in the cyanobacterial incident management protocol of the Zuikerbosch DWTW. When the model was tested with 2 years of unseen on-line (*in situ*) generated data (Figure 6.12c), the R^2 -value decreased to 0.072 and the RMSE was 0.073 $\mu\text{g/L}$. It is evident that as in the case of the real-time microcystin^[0] prediction, the model performed relatively well when tested with unseen laboratory generated data, but not well at all when tested with data from the on-line (*in situ*) instrument.

6.5 Discussion and conclusions

From the models discussed in this chapter, it was evident that forecasting models for *Anabaena* sp., *Microcystis* sp. and microcystin can be effectively developed for application in the drinking water treatment industry, even if input variables were limited to on-line (*in situ*) measureable variables (i.e. temperature, turbidity, conductivity, pH, dissolved oxygen (DO) concentration, chlorophyll-*a* and initial cyanobacteria inoculum). However, if the models were initially developed and trained with laboratory generated data the application value in on-line (*in situ*) generated data seems to be limited. From these results it was evident that if on-line (*in situ*) generated data is to be used as input into the models, such models will have to be developed and trained by on-line (*in situ*) generated data to be successful. This is mostly due to the greater variability found in the on-line (*in situ*) data, when compared to that of the laboratory generated data (Swanepoel et al., 2014). The on-line (*in situ*) instruments have been found to be relatively stable if maintenance and cleaning are done at least once a week and re-calibration of probes is done with vigilance when data drifts are detected. Within one week (from one cleaning or maintenance occasion to the next) the variability within the data is higher than what is found in laboratory generated data (Swanepoel et al., 2014) and it may be this variability that contributed to the poor correlation between the measured data and the predicted data when the models were tested with the on-line (*in situ*) generated data. Although the data from the on-line instrument may not be useful for incorporation into the models as they currently are, it can still be used as part of the early warning system detecting trends of increasing cyanobacterial cell concentrations in the source water and should therefore not be dismissed as a failed experiment.

The summary of the results from all the models discussed in this chapter is presented in Table 6.1. The results show that models developed from on-line (*in situ*) measureable data only (as discussed in this chapter), are almost just as good as the models developed by using all possible input variables (as discussed in Chapter 5) when tested with laboratory generated “previously seen” and “unseen” data. The reason is most probably because the most important variable (the one towards which the output result showed the greatest sensitivity towards) is included in these models, namely initial cyanobacteria inoculum.

Table 6.1 Summary of all statistical results of models developed from historical laboratory generated data and tested with “unseen” laboratory generated data and “unseen” on-line (*in situ*) generated data.

MODEL	Tested with 25% boot-strapped data		Tested with 3 years “unseen” laboratory generated data		Tested with 2 years “unseen” on-line (<i>in situ</i>) generated data		Significant when tested with laboratory generated data?	Significant when tested with on-line (<i>in situ</i>) generated data?
	R ² -value	RMSE	R ² -value	RMSE	R ² -value	RMSE		
<i>Anabaena</i> ^[0]	0.89	1 912 cells/mL	0.01	9 458 cells/mL	0.01	9 587 cells/mL	✓	✗
<i>Anabaena</i> ^[7]	0.76	2 771 cells/mL	0.002	19 018 cells/mL	0.02	3 597 cells/mL	✓	✗
<i>Microcystis</i> ^[0]	0.94	2 522 cells/mL	0.99	4 119 cells/mL	0.014	2 161 cells/mL	✓	✗
<i>Microcystis</i> ^[7]	0.91	3 354 cells/mL	0.52	39 098 cells/mL	0.000	1 228 cells/mL	✓	✗
Microcystin ^[0]	0.83	0.23 µg/L	0.37	0.47 µg/L	0.009	0.04 µg/L	✓	✗
Microcystin ^[7]	0.75	0.29 µg/L	0.31	0.096 µg/L	0.072	0.073 µg/L	✓	✗

Testing the models with “previously seen” data (i.e. 25% of the boot-strapped results from the testing dataset that was part of the 10 years development dataset) gave a better correlation than testing it with “unseen” data. Testing the models with the on-line (*in situ*) generated data did not give a good correlation and therefore data from this instrument cannot be used in the “Prediction Monitoring” step in the cyanobacterial incident management protocols (refer to Sections 7.5 and 7.6, pages 122 – 131). However, data from these on-line (*in situ*) instruments can still be useful to indicate increasing trends in cyanobacterial cell concentrations. The models predicting *Microcystis* sp. concentration showed the best correlations, followed by the microcystin models and lastly the *Anabaena* sp. models. Although the *Anabaena* sp. prediction models, when tested with unseen laboratory generated data, did not show a good correlation coefficient (R²-value), visual inspection of the models indicated that both the *Anabaena*^[0] and *Anabaena*^[7] prediction models can be utilized since the time-prediction of increased concentrations are predicted with relatively good accuracy, although the magnitude of the concentration is sometimes over- or under-estimated. The RMSE (root mean square error) generally shows an increase with increasing prediction time (except in the microcystin^[7] model tested with the “unseen” laboratory generated data).

The frequencies of the different input variables used in the models to predict *Anabaena* sp., *Microcystis* sp. and microcystin are displayed in Table 6.2, ranging from the most frequently included to the least frequently included variable in the models. Initial cyanobacteria inoculum was incorporated into all models discussed in this chapter (as was the case in most of the models discussed in Chapter 5), and revealed to be the variable towards which the model shows the greatest sensitivity. Dissolved oxygen was important in *Microcystis* sp. and microcystin models, while temperature only played a role in microcystin models but not in those forecasting *Microcystis* sp. and *Anabaena* sp., most probably because temperature is regarded as an important variable for the production of microcystin (Conradie and Barnard, 2013). As in the research done by Chan et al. (2007) chlorophyll-*a* was determined to be an important driving variable for the prediction of microcystin (more so than for the prediction of *Microcystis* sp. and *Anabaena* sp.). Conductivity on the other hand, played an important role in the models forecasting *Microcystis* sp. and *Anabaena* sp. but not in the microcystin models.

Table 6.2 Frequency at which input variables were used in models to predict *Anabaena* sp., *Microcystis* sp. and microcystin concentrations.

Variable used in models	Number of occurrences in <i>Anabaena</i> models	Number of occurrences in <i>Microcystis</i> models	Number of occurrences in microcystin models	Frequency
Cyanobacteria inoculum (Cyano)	2	2	2	6/6 = 100%
Turbidity (Turb)	2	1	2	5/6 = 83%
Dissolved Oxygen (DO)	-	2	2	4/6 = 67%
Conductivity (Cond)	2	2	-	4/6 = 67%
Phytoplankton biomass (Chl <i>a</i>)	-	1	2	3/6 = 50%
pH	1	1	1	3/6 = 50%
Water temperature (Temp)	-	-	2	2/6 = 33%

The models for the prediction of *Anabaena* sp., *Microcystis* sp. and microcystin have proved to be very successful for as far as 7 days in advance and can be incorporated into the cyanobacteria incident management protocols as discussed in Chapter 7, provided that the input variables are those from laboratory generated data. This will benefit the DWTW as to enable them to start preparing for high cyanobacterial cell concentrations or a high microcystin incident in the source water before it happens.

It is stressed that the results from the on-line (*in situ*) instrument display great variability in the results and if on-line (*in situ*) generated data is to be used as input, models need to be developed and trained with on-line (*in situ*) generated data. It is anticipated that the results and validation of such models will give much better results than what is currently the case. The fact that the correlation coefficients of the models developed with limited-input variables (when tested with the laboratory generated data) were almost just as good as the models developed with all possible input variables shows that models developed for on-line (*in situ*) instruments will most probably work well, provided that such models are developed with data from these on-line (*in situ*) instruments. This may especially be applicable in the case of small DWTW that do not have laboratory facilities or limited laboratory facilities.

Some of the models that were discussed in Chapters 5 and 6 of this manuscript have been chosen to be incorporated into the cyanobacterial incident management protocols of Rand Water as a case study (Chapter 7, Section 7.6).

CHAPTER 7

APPLICATION OF FORECASTING MODELS IN THE MULTI-BARRIER APPROACH OF WATER SAFETY PLANS AS PART OF THE BLUE DROP CERTIFICATION PROGRAM IN SOUTH AFRICA

7.1 Introduction

All over the world, the management of surface water to supply drinking water of acceptable quality has become a priority (Du Preez and Van Baalen, 2006). Over the past few years, management actions regarding surface water have become more and more sophisticated as new developments in computer, GIS, satellite and analyses technologies have become common place.

The introduction of the concept of water safety plans (WSP) used by drinking water production utilities has opened up the opportunity to apply some of these highly sophisticated technological advancements as part of the multi-barrier approach in order to supply safe and aesthetically acceptable drinking water to consumers. This includes reducing the risk of supplying “bad water quality” (negatively affecting consumer health) and/or “perceived bad water quality” (not necessarily affecting consumer health) to consumers.

The introduction of the incentive-based Blue Drop Certification Program by the Department of Water Affairs (DWA, currently known as DWS – Department of Water and Sanitation) provided the perfect platform for the South African drinking water treatment industry to apply principles and concepts of water safety planning which include the development of incident management protocols (DWA, 2011) (Figure 7.1). In this chapter the application of prediction models discussed in Chapters 5 and 6 is proposed to be incorporated into two cyanobacterial incident management protocols (one relating to the cyanotoxin microcystin and one relating to the taste and odour compound geosmin) at Rand Water as a case study (refer to Section 7.6). As displayed in Figure 7.1 these incident management protocols form part of the water safety planning of a drinking water treatment works (DWTW) and is covered in the South African Blue Drop Certification Program according to the guidelines and standards set in SANS 241 (SANS 241-1, 2011) as well as other guidelines and standards.

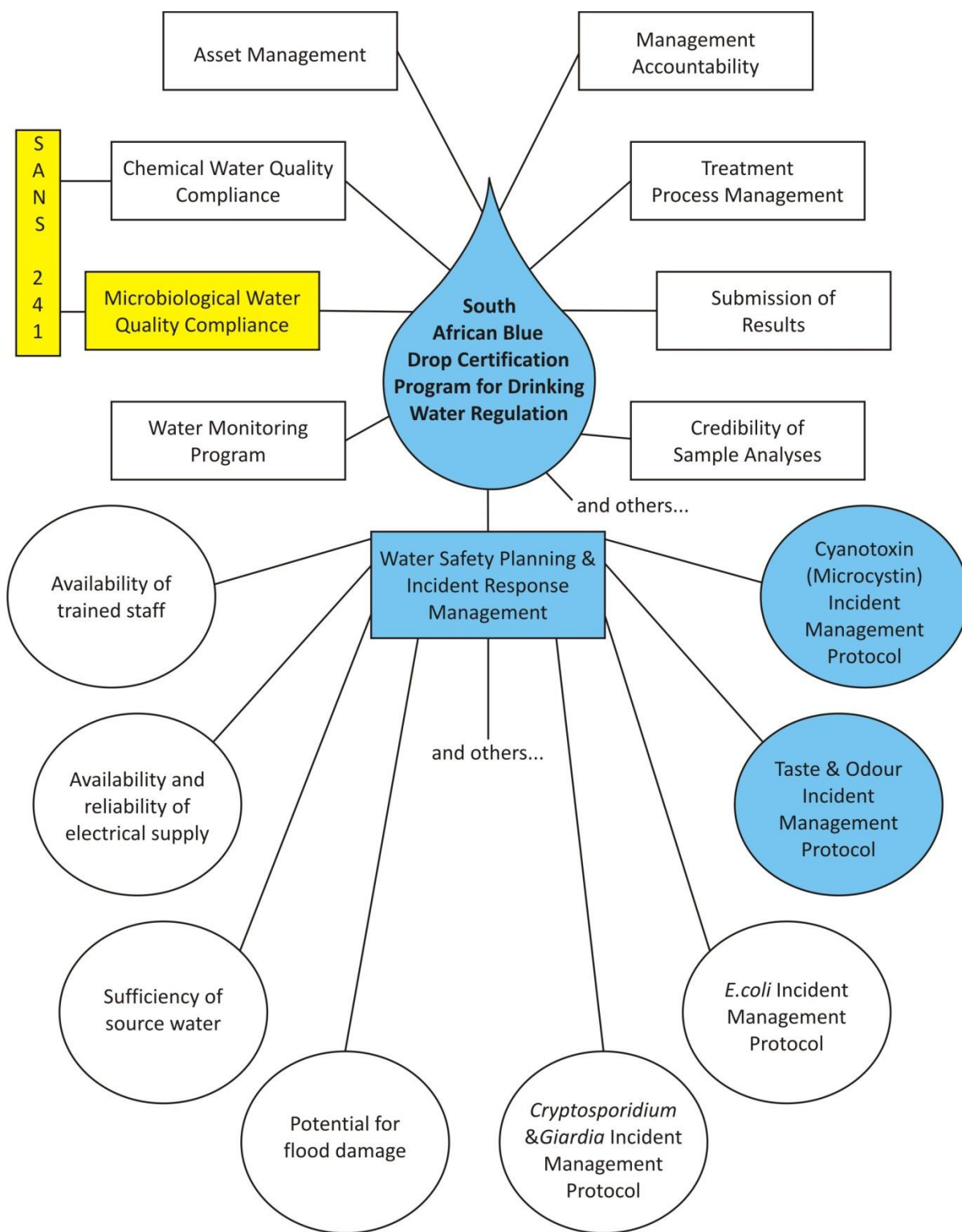


Figure 7.1 Schematic representation of the relationship between the South African Blue Drop Certification Program, SANS 241, Water Safety Plans as well as incident management protocols addressed in this study.

7.2 SANS 241-1 (2011) – the latest South African Drinking Water Standard relating to cyanobacteria and cyanotoxins

Before 2011, the South African drinking water treatment industry had to rely on guidelines and standards from different sources for the determination of applicable guidelines and standards regarding cyanobacterial-related variables. These guidelines and standards (e.g. $\leq 1 \mu\text{g/L}$ microcystin as LR) were adopted from (amongst others) the World Health Organisation (WHO, 2011 and earlier editions), Australian water quality guidelines (NHMRC, 2011 and earlier editions) and New Zealand water quality guidelines (NZMOH, 2001).

During the past few years however, the Department of Water Affairs (currently known as the Department of Water and Sanitation), the South African Bureau of Standards together with scientists from academia and from the drinking water treatment industry have accomplished a great deal to try and ensure the distribution of safe drinking water to the South African public. Since the development of the South African National Standard for drinking water, the 2011 version was the first to include a guideline limit for the cyanotoxin microcystin (SANS 241-1, 2011). This limit ($\leq 1 \mu\text{g/L}$ microcystin as LR), is based on the guideline set by the World Health Organisation which was developed for an exposure of a person of 60 kg consuming 2 L of drinking water per day (WHO, 2011). SANS 241-1 (2011) states that microcystin only needs to be measured where an algal bloom of $> 20\,000$ cyanobacteria cells/mL are present in the raw (source) water. Since most drinking water treatment utilities in South Africa (with the exception of a hand full of large ones) do not have the capacity of performing algal identification in their source water (let alone enumerating the cyanobacterial cells), an addition was made to the National Standard that states: “In the absence of algal monitoring an algal bloom is deemed to occur where the surface water is visibly green in the vicinity of abstraction, or samples taken have a strong musty odour.” It is obvious that these guidelines may be left open to interpretation and as can be expected, may cause not only confusion, but also deniability of a microcystin incident taking place at a DWTW. SANS 241-2, (2011) entitled “*The application of SANS 241-1*”, states that one critical control point in the drinking water needs to be analysed for microcystin at least 6 times a year during the summer months. This guideline may also not be sufficient and cause confusion, since it does not include any source water analyses or specify how many of the drinking water lines need to be analysed. Measuring the concentration of microcystin in the source water is of utmost importance since chlorine disinfection, that is usually able to oxidise microcystin (Chorus, 2012), can fail (especially in small DWTW where unskilled operators are often performing these vital tasks), causing the microcystin in the final treated water to be equal to

that of the source water. Furthermore, when taken into account that cyanobacterial blooms might stretch over a period longer than the six warmest months of the year and also may change very rapidly over a short period of time (especially in warm areas like the Magaliesberg area) 6 monitoring occasions a year is obviously not sufficient. It must however, be commended that the microcystin guideline has finally been incorporated into the SANS 241 : 2011 drinking water standard and that future revisions of this standard now have a basis to build from.

Currently no guideline or suggestion is made regarding the taste and odour compounds geosmin or MIB in the SANS 241 : 2011 drinking water standard. This is most probably because the taste and odour compounds are not harmful to consumers. Although not harmful to consumers, DWTW value the monitoring of taste and odour compounds just as high as the monitoring of toxic substances since aesthetically unacceptable drinking water will inevitable reduce the consumers' confidence in tap water, impacting on their core business (Zoschke et al., 2011).

Rand Water that provided water quality data for this study supplies drinking water to more than 12 million consumers, with the safety of their final product very high on their priority list (Viljoen, 2010). They have therefore started with microcystin analyses and guidelines more than ten years ago and still monitors microcystin not only at one critical point in the drinking water (as stipulated by SANS 241-2, 2011), but in all drinking water mains. Currently Rand Water has 7 different stations in operation and their water quality monitoring program also includes microcystin analyses at critical points in the source water as well as at critical points in the distribution network. These analyses are performed twice or four times per month (depending on the point) during the eight warmest months of the year (September to April). The total number of routine microcystin analyses done by Rand Water adds up to 581 per year at a cost of R 888 per sample³. Since this bulk DWTW regards the health of their consumers as priority (as is indicated by the more than half a million Rand per year spent on microcystin analyses alone), they have not only introduced microcystin analyses into their monitoring program but also incorporated a microcystin protocol as well as a taste and odour protocol into their water safety plan (WSP) (Viljoen, 2010).

³ As per financial year 2013 – 2014.

7.3 *The Blue Drop Certification Program for ensuring safe drinking water to South African consumers*

During 2007/2008 the South African Department of Water Affairs (DWA, currently known as DWS) came up with an incentive-based regulation strategy where conscious rewards as well as penalties are used to encourage performance in excellence and continuous improvement in drinking water quality and supply in South Africa, called the Blue Drop Certification Program. The South African Minister of Water Affairs introduced this concept to the South African water sector at the National Municipality Indaba in Johannesburg on the 11th of September 2008 (DWA, 2011). The Blue Drop Certification Program which addresses **drinking water** quality management regulation is supported by the Green Drop Certification Program, which addresses **wastewater** quality management regulation.

The Blue Drop Certification Program measures and compares the results of the performance of Water Services Authorities and their providers in South Africa. These performances include many aspects such as: Water Safety Planning Process and Incident Response Management; Treatment Process Management; Management Accountability; Water Monitoring Program; Credibility of Sample Analyses; Submission of Results; Microbiological Water Quality Compliance; Chemical Water Quality Compliance as well as Asset Management (DWA, 2011; DWA, 2012) (Figure 7.1).

Achieving Blue Drop status is a competitive and sought after achievement that most municipalities strive for and the outcome thereof also reflects positively or negatively on the bulk drinking water providers. Therefore most bulk drinking water providers are not only involved in ensuring water of good quality to their customers (many different municipalities) but they are also involved in assisting their customer municipalities to draft their individual water safety plans (WSP), report their results of water quality analyses to DWA/DWS, as well as aiding these municipalities in completing their reports to DWA/DWS so that they can achieve their Blue Drop status.

It is evident from the current study that incorporating predictive models for certain important water quality variables into the WSP of small or bulk DWTW, will have a positive impact far beyond the ability of DWTW to prepare for the treatment of high cyanobacterial cell concentrations in the source water. It will also positively influence the ability of customer municipalities (as in the case of bulk DWTW) or individual municipalities to attain and keep their Blue Drop status.

7.4 Using a water safety plan (WSP) for risk assessment and risk reduction in DWTW

Over the past few years, suppliers of drinking water have recognised the limitations of relying solely on end-product monitoring to ensure safe drinking water to their consumers. Therefore, they have sought to reinforce their approach by adopting preventative strategies where hazards and risks are proactively identified, assessed and managed. This led to the development of water safety plans (Hamilton et al., 2006). A water safety plan (WSP), is a step-by-step risk management strategy to ensure consistent safe drinking water supply from catchment to consumer (Davison et al., 2005; Bartram et al., 2009). A WSP encompass many diverse aspects, most of which will not be addressed in this study e.g. the availability of trained staff, the availability and reliability of electrical supply, the sufficiency of source water, the potential for flood damage, appropriate treatment processes and chemicals, the reliability of communication procedures, etc. (Figure 7.1).

All over the world, WSPs have been developed and implemented by drinking water suppliers to ensure safe drinking water e.g. in Australia (NHMRC, 2011), New Zealand (NZMOH, 2001), the United Kingdom (UKWIR, 2003) and other countries (WHO, 2011). The “indirect potable reuse industry” that, in part, utilises discharge of wastewater effluent has also adopted the WSP approach for ensuring safe drinking water to their consumers (Dominguez-Chicas and Scrimshaw, 2010). Even the cruise ship industry has developed their own WSP to ensure ship water safety to all passengers and crew members (Mouchtouri et al., 2012). However, no WSP developed thus far, include future predictions for hazard and risk management in drinking water supply and distribution.

Mercer and Bartram (2011) mentioned that as technology develop and improve, the way in which water quality is ensured and managed also need to evolve. Therefore this study was undertaken to use the outcomes from the prediction models to add another barrier in the multi-barrier approach of WSPs in order to ensure drinking water of “good quality” and “perceived good quality”. This novel approach to include future predictions as part of the cyanobacterial incident management protocols is described in Sections 7.5 and 7.6.

Protocols dealing with water quality incidents for different variables are usually included in the WSP of a DWTW (Figure 7.1) e.g. *E. coli* protocol, *Cryptosporidium* and *Giardia* protocol, taste and odour protocol, microcystin protocol, etc. (Viljoen, 2010). This study will focus on the two protocols that involve cyanobacteria namely the microcystin and taste and odour protocols. Du Preez and Van Baalen, (2006) call such protocols “step-by-step alert levels frameworks” setting

in motion monitoring and management action sequences that drinking water utilities can use proactively to deal with the problems cyanobacteria can cause in a DWTW. Different variables (e.g. microcystin concentration, cyanobacterial cell concentration, chlorophyll-*a* concentration, geosmin concentration, MIB concentration and mouse bioassays in combination with each other) are used to trigger different action sequences in order to deal with increasing or decreasing concentrations of cyanobacteria and their related organic compounds in the source and drinking water. An example of such a protocol is displayed in Figure 7.2 (adapted from Du Preez and Van Baalen, 2006).

During the “Routine Monitoring Level” (Figure 7.2) variables such as cyanobacteria identification and enumeration as well as microcystin concentrations are monitored at least once in two weeks. When cyanobacteria are detected during the routine analysis, then the alert is stepped up to the “Vigilance Level”. The “Vigilance Level” mostly impacts on actions in the laboratory, where scientists and analysts are alerted to pay closer attention to the results that are generated by the laboratory on a daily basis. Once the cyanobacterial cells increase to concentrations of > 2000 cells/mL in the source water, or the microcystin concentration in the final treated water increase to levels > 0.2 µg/L (but still < 0.8 µg/L), the alert is raised to “Alert Level 1”. At “Alert Level 1” a written memorandum will be compiled and sent to all stakeholders involved and the DWTW will start preparations for a potential incident.

The preparations taken by the DWTW may include:

- Ensuring that infrastructure and staff are available for additional sampling and capability of taking depth profile samples at the intakes.
- Determining the potential holding capacity at the sludge dumping site, if conditions necessitate supernatant retention.
- Ensuring that the application process for discharge permits from the sludge dumping site is in order, if discharge of excess water from site is necessary.
- Testing the PAC (powdered activated carbon) dosing facility and ensuring that it is in a state of immediate operational functionality if needed.
- Ensuring that the sludge removal process is effective and that sludge can be removed more frequently if needed.
- Ensuring that CO₂ is available for pH correction so that lime can be used as coagulant chemical. (This may not be a problem in other countries but South Africa only has one supplier that can produce large quantities of CO₂ for pH correction.)
- Ensuring that load shifting is possible from one unit/plant to another when necessary.
- Ensuring that abstraction points can be lowered at the dam intake when necessary.

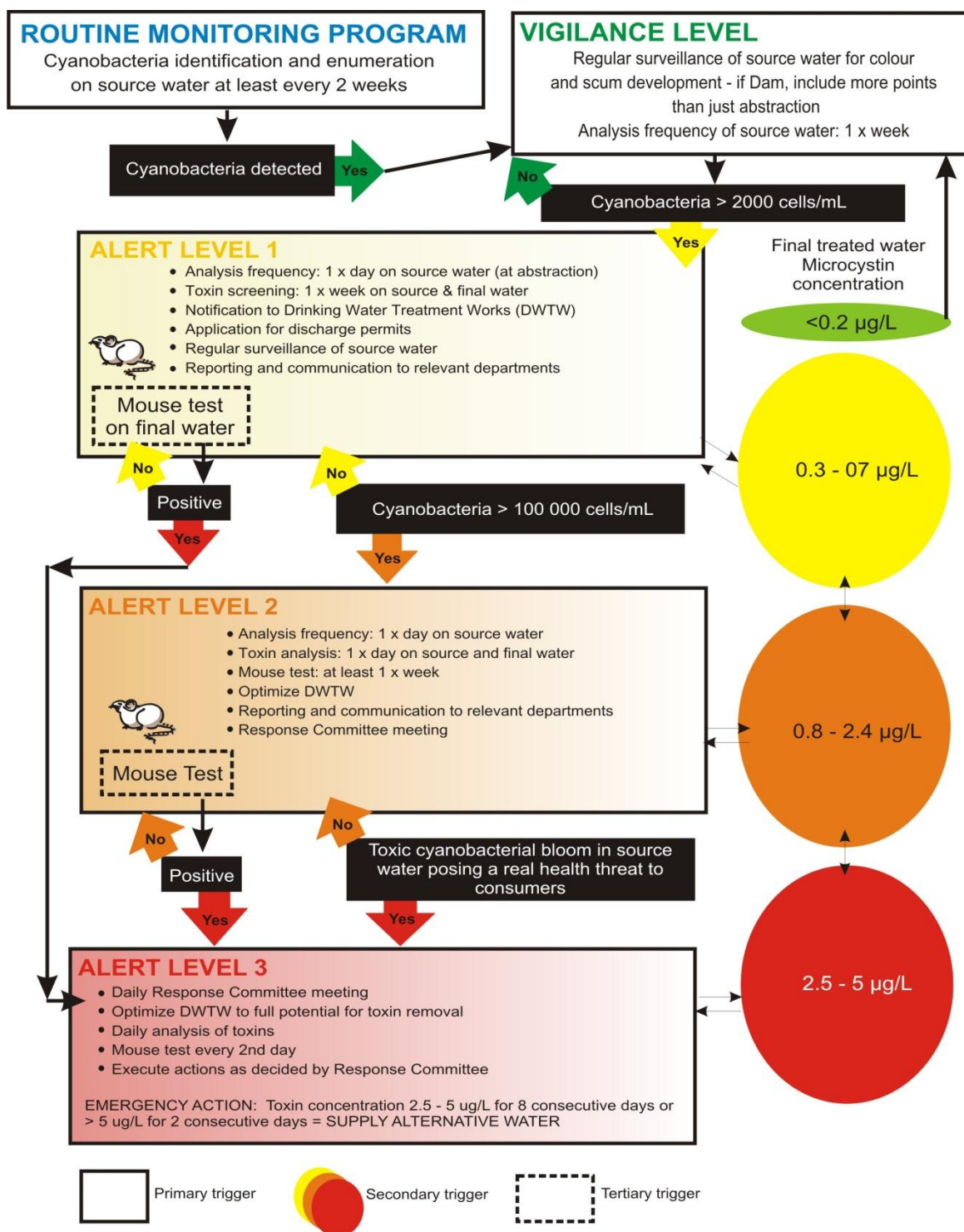


Figure 7.2 A typical cyanobacterial incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Du Preez and Van Baalen, 2006).

A mouse bioassay by oral ingestion of sampled water (performed at the Toxicology Department at Onderstepoort Veterinary Institute in Pretoria) may also be performed on the final treated water and if the result is negative it will not change the “Alert Level”. A mouse bioassay which will be an indication of mammal response to the water is classified as a “tertiary trigger” indicating that it overrules all other triggers and once is positive will immediately raise the state of the protocol to “Alert Level 3”. A positive mouse bioassay is based on observations regarding “signs of poisoning”, which refer to 1) behavioural abnormalities 2) *post mortem* investigation of the liver and 3) possible toxicosis resulting in death (Mohamed et al., 2006). If the cyanobacterial cells increase to > 100 000 cells/mL in the source water, or the microcystin concentration in the final treated water increase to > 0.8 µg/L (but still < 2.5 µg/L), the state of the protocol will raise to “Alert Level 2”. At “Alert Level 2”, the analysis frequency of cyanobacteria identification and enumeration in the source water as well as microcystin concentration in the final treated water will increase to once a day, mouse bioassays will increase to at least once a week, reporting and communication will continue and a response committee meeting will be convened where different strategies will be discussed and implemented.

The strategies discussed and implemented by the response committee may include:

- Optimising the intensity and scope of the monitoring program.
- Optimising the use of best quality source water (e.g. change abstraction to points where cyanobacterial cell concentrations are lowest and move load to stations where less/no problems are experienced).
- Conducting depth profile analyses at the dam intake.
- Lowering abstraction levels in the dam intake to reduce cyanobacteria load into the DWTW.
- Commissioning the PAC (powdered activated carbon) dosing plant.
- Optimising operations at the sludge dumping site (e.g. terminating or reducing recycling of supernatant into the source water if and where it is applicable).
- If concentrations at the sludge dumping site are high enough to impact on the final water quality after treatment, then discharge of the supernatant to the river should be considered, instead of redirecting it to the head of works (blending with the source water).
- Where operationally possible, terminating the re-routing of filter backwash water to head of works (blending with the source water).
- Optimising the coagulation, flocculation and sedimentation processes by:

-
- Increasing frequency of performing jar stirring experiments for the determination of optimum coagulant dosage concentrations.
 - Minimising any overdosing of organic coagulant, since high dosages of organic coagulants, especially over prolonged periods, may result in filter clogging and mud ball formation.
 - Ensuring that the flocculators are not overloaded to prevent short circuiting of water which will inevitably lead to poor coagulation and flocculation.
 - Ensuring that the sludge is frequently removed from the sedimentation tanks and that the process is effective.
 - Optimising the filtration and filter cleaning processes by:
 - Keeping doors to filter houses closed at all times to reduce light availability in the filter house for algae and cyanobacterial growth as well as preventing insects from entering.
 - Turning lights to minimum illumination to ensure that algae and cyanobacteria cannot grow in the filters.
 - Ensuring that the filter beds are maintained according to specifications (relating to e.g. filter media and dressing of the filter beds).
 - Reducing filter run times.
 - Preventing mud ball formation to ensure that filters are effective as the last barrier to remove the cyanobacteria that were able to avoid the sedimentation process.
 - Considering a double instead of single filter backwash cycle to remove more of the mucilaginous cyanobacterial cells that tend to adhere to the sand granules.
 - Monitoring the turbidity of individual filters to ensure that filters are functioning properly.
 - Increasing backwash frequency where necessary.
 - Optimising the disinfection process by:
 - Monitoring chlorine and residual chlorine levels.
 - Establishing optimum oxidation conditions for certain cyanotoxins, as far as possible.
 - Under certain circumstances it may be advisable to consider draining a large percentage of the reservoir where source water may temporarily be stored before it enters the plant, to allow fresh (non-stagnant) water to fill up the reservoir again.

A mouse bioassay (by oral ingestion of sampled water) will be performed on the final treated water at least once a week and if the result is negative it will not change the state of the protocol or “Alert Level”. If, however, the mouse bioassay is positive it will trigger the DWTW

to move to “Alert Level 3” immediately. If the cyanobacterial bloom in the source water poses a real threat to consumers, or the microcystin concentration in the final treated water increase to $> 2.5 \mu\text{g/L}$, the state of the protocol will increase to “Alert Level 3”. At “Alert Level 3”, daily response committee meetings will be held. The optimisation actions in “Alert Level 2” will carry on, sampling and analysis frequency will increase to twice a day, microcystin toxin analyses will be performed once a day, mouse bioassays will be carried out every second day and decisions by the response committee will take priority for implementation on the plant and elsewhere. When the microcystin concentration in the final treated water reaches a level of between $2.5 - 5 \mu\text{g/L}$ for 8 consecutive days or $> 5 \mu\text{g/L}$ for 2 consecutive days an alternate supply of drinking water should be provided for consumers.

A typical taste and odour protocol addresses the aesthetic properties (mostly taste and odours caused by geosmin and 2-methylisoborneol (MIB)) and is displayed in Figure 7.3. Although geosmin and MIB are not harmful to consumer health, occurrences of these organic compounds in the final treated water may be just as damaging to the image of a DWTW (Zoschke et al., 2011) and therefore just as much emphasis is placed on the prevention of these compounds, than on the prevention of cyanotoxins to occur in the final treated water.

During the “Routine Monitoring Program” level (Figure 7.3), variables such as cyanobacteria identification and enumeration as well as geosmin and 2-methylisoborneol (MIB) concentrations are monitored at least once in two weeks. When cyanobacterial cells increase to concentrations $> 2000 \text{ cells/mL}$, and/or geosmin or MIB $> 15 \text{ ng/L}$ are detected during the routine analysis, then the state of the protocol is raised to “Alert Level 1”.

At “Alert Level 1” a written memorandum will be compiled and sent to all stakeholders involved, on-site inspections of the DWTW will be carried out, sampling will be increased and the DWTW will start preparations for a potential incident. Refer to page 116 for an example of possible preparations the DWTW can take.

If the cyanobacterial cell concentration increase to $> 50\,000 \text{ cells/mL}$ in the source water, or the geosmin and or MIB concentrations in the final treated water increase to $> 30 \text{ ng/L}$, the state of the protocol will be raised to “Alert Level 2”. At “Alert Level 2”, a response committee meeting will be convened where different strategies will be discussed and implemented. Refer to pages 116 as well as 118-119 for possible strategies to implement.

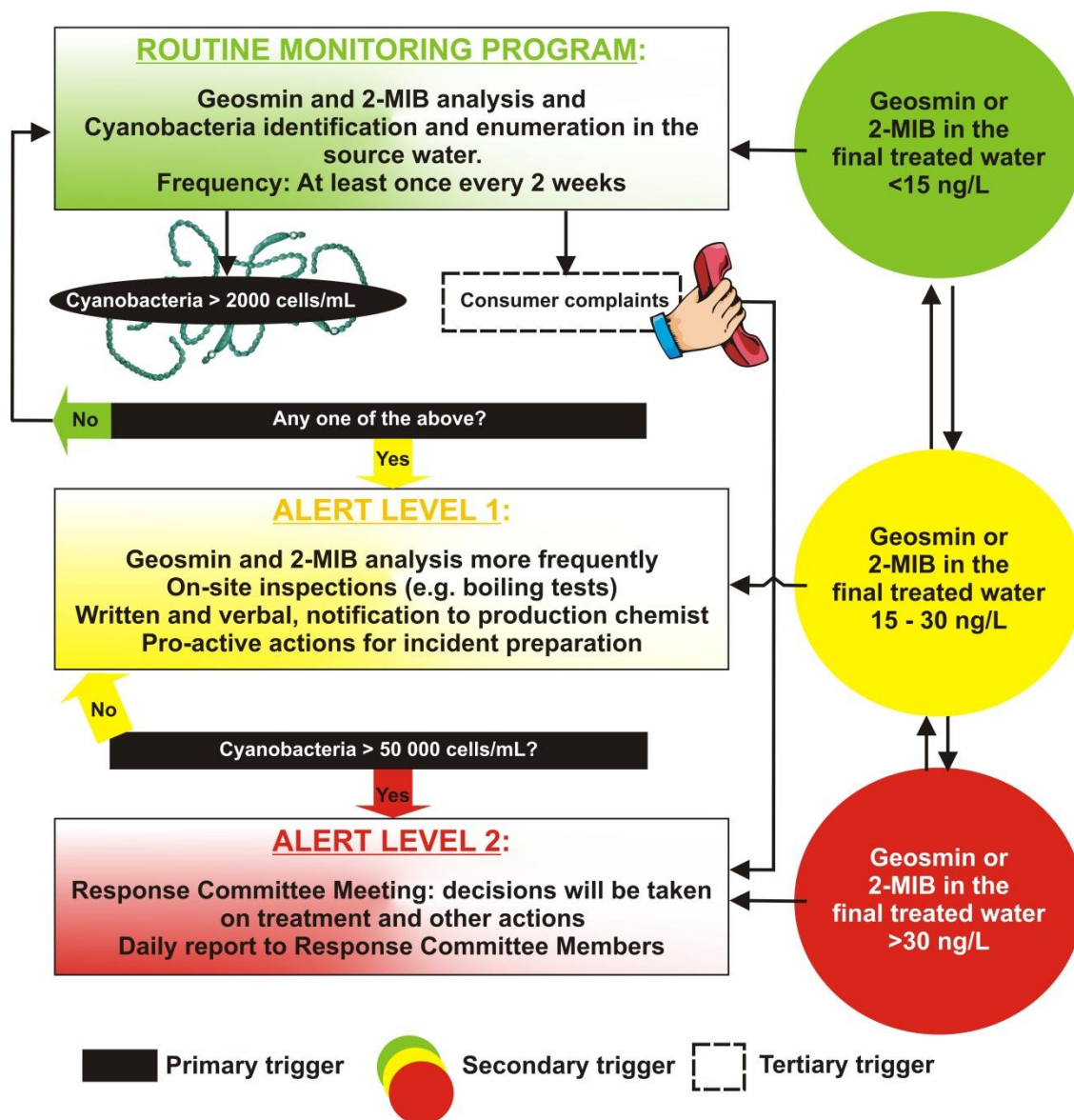


Figure 7.3 A typical taste and odour incident management protocol using cyanobacterial cell concentration as the primary trigger (adapted from Swanepoel and Du Preez, 2008).

Because this protocol deals with the aesthetic aspect of the final treated water, the tertiary trigger overruling all other triggers is consumer complaints. Since this “perceived perception” of drinking water quality can adversely affect consumer confidence in tap water as well as the public image of the DWTW (Zoschke et al., 2011), it is treated with the same importance as the presence of microcystin in the final treated water. Therefore “Alert Level 2” can be reached

immediately if a substantial number of consumer complaints are received. It should be noted that consumer perception can be affected by many external factors especially news broadcasts of bad water quality incidents happening elsewhere. During the Delmas-incident in 2005 when the outbreak of typhoid fever (NICD, 2005) caused a large panic in the South African media, DWTW got many more consumer complaints than usual, even though the Delmas-incident occurred due to the contamination of a bore hole and had no impact on the water quality at other DWTW⁴. Incidents such as these can have a detrimental effect on the application value of the incident management protocols, but it is usually during such incidents that consumers need reassurance. Speedy responses to queries (related to consumers' own drinking water or not) can cause either a positive or negative perception of water quality as well as the DWTW willingness to address issues. Consumer complaints should therefore never be ignored or dismissed, since it can impact severely on consumer confidence in tap water, even though water quality may not have been impacted at all.

7.5 *Possibility to incorporate prediction models into the WSP*

According to the WHO WSP Manual (Bartram et al., 2009), hazard identification requires historic information and events (e.g. events where the plant had to deal with increased cyanobacterial cell concentrations and increased cyanotoxin concentrations), as well as predictive assessment of historic information based on utility data and knowledge of particular aspects of the treatment and supply systems. Identification of these hazards and risks as well as applying knowledge on historical incidents and data to aid in the predictive capabilities of a DWTW, forces lateral thinking and innovative solutions to the challenge. In that sense prediction models are a new and innovative way of addressing possible and even probable future risks and start initiating solutions before the problem arises. As part of the risk assessment, the "likelihood of the occurrence" should also be determined (Bartram et al., 2009) and therefore all models used in this study have been tested against real unseen data which provides the "likelihood of occurrence" as referred to by Bartram et al., (2009). Taking a multi-barrier approach increases the capability of the DWTW to reduce these risks, since the adding of every single barrier increases the prevention capability for each potential breach. Therefore the inclusion of a future prediction aspect of certain variables in the water quality protocols of

⁴ Personal communication Prof. Hein du Preez, Head Biology, Rand Water Analytical Services.

WSPs adds another barrier to prevent breaches of previously defined thresholds (e.g. SANS 241 guidelines) that might have an impact on water quality or perceived water quality.

The prediction models developed during this study can effectively be incorporated into the incident management protocol (which form part of the Water Safety Plan – Figure 7.1) by adding a “Prediction Monitoring Level” which will be performed in association with the “Routine Monitoring Level” (Figures 7.5 and 7.6). The “Prediction Monitoring Level” will be carried out where results of future predictions of cyanobacterial cell concentrations (*Anabaena* sp. and *Microcystis* sp.), microcystin and geosmin concentrations are utilised which 1) preferably incorporate the models utilising the full suite of input variables (as described in Chapter 5) are utilised, as those models are more accurate or 2) if not possible incorporating the models utilising the limited suite of input variables (as described in Chapter 6). The application thereof is depended on whether the capability to execute this future prediction exists in the DWTW and therefore it will be performed in conjunction with the “Routine Monitoring Level” and not replace it. If the capacity does not exist, the normal “Routine Monitoring Level” will still function and “Alert Level 1”, can still be reached without the “Prediction Monitoring Level”. It is important to note that Rand Water, as case study, and that provided the data for this study, measures most of the input variables on a daily basis in their source water with a data turn-around time of approximately 24h (due to prioritising result output). This quick availability of important results significantly simplifies the utilisation of the models and spreadsheets (Figure 7.4). However, it should be noted that water utilities that do not measure the input variables on a daily basis, or who may have to wait weeks for the results of analyses may encounter significant problems when trying to utilise the spreadsheets (Figure 7.4) in their context. If the input variables are measured on the Monday of week 1, the results are received from the laboratory on Thursday of week 1, data is entered into the spreadsheets on Friday of week 1, the output will still only be forecasting what will be happening in the source water on the Monday of week 2 and will therefore not allow a full week to prepare for a possible incident. Therefore, if drinking water treatment utilities want to incorporate these forecasting models into their incident management protocols, the availability of data needs to be taken into account. DWTW may opt to rather use the models forecasting results 14 days in advance due to data constraints, or opt to use the models developed for limited input variables (as described in Chapter 6) which may be easier to perform in a shorter period of time.

According to the Du Preez and Van Baalen incident management protocols (Du Preez and Van Baalen, 2006 and Du Preez et al., 2007), the “Routine Monitoring Level” refers to monitoring of certain variables (specifically the variables chosen to be the triggers in these protocols) on an on-going basis (all year round). Monitoring of these variables is proposed to be done at least once every two weeks (or more frequently where possible). However, results from these analyses may be delayed due to a range of different factors e.g. sampling, distance of sample point to laboratory, laboratory sample turn-around time, number of samples to be analysed, analytical staff availability, instrument capability and availability etc. Therefore, the addition of the “Prediction Monitoring” may speed up the process considerably and allow the DWTW to step up to “Alert Level 1” sooner, as to be able to start preparing for a potential incident. During “Alert Level 1” in both the microcystin and taste and odour protocols, most of the DWTW incident preparations are being completed and therefore from “Alert Level 1” onwards, results of the daily analyses (and not those of future predictions) will be used to determine Alert Level reaction and response. Therefore, although the models cannot give a 100% accurate prediction, it will not have a major financial impact on the DWTW, because at “Alert Level 1” most of the preparations for a potential bloom do not have major financial implications. However, major financial losses can be prevented if the DWTW is alerted well in advance that it needs to adjust its treatment procedures since high cyanobacterial cell concentrations are expected in the source water.

Spreadsheets set up to give the real-time, 7 days, 14 days and 21 days prediction by only completing the current days analytical results (Figure 7.4) can be supplied to the station chemist of the DWTW. This can be utilised to determine the results for future predictions of cyanobacterial cell concentrations and cyanobacterial-related organic compounds. The part of the spreadsheet in grey (Figure 7.4) will not be visible to the end-user. All spreadsheets to be used for this purpose are available on the CD attached to the back of this thesis (ADDENDUM B). In certain large bulk DWTW the responsibility for the initial reporting of results is situated within the laboratory and therefore the responsibility of the daily updating of the spreadsheets may not be situated at the plant, but at the laboratory. This will differ from institution to institution and should be described in the WSP of each DWTW.

Real Time																	7 Days				14 Days				21 Days				Anaabaena concentration					
The best ruleset is: IF ((Cyno/79.532)*>219.533) THEN Ana=(((((Cyno*PO4)/{136.791/COI ELSE Ana=(((((Cyno/{185.285/pH})-{-115.464})																	The best ruleset is: IF (((Fe/NO2)<=47.402)OR((Turb*{PO4*46.915})<=132.101)) THEN Ana=(((((pH-Cyno)*{COD+17.200})/{ln{Mn}}*34.243)+{ln ELSE Ana=(((((5.971-Cyno*{NO3+COD})/{ln{ln{NH4}}*7.159})+{16.431-{-				The best ruleset is: IF ((Si*Si)<88.088) THEN Ana=(((((Cyno*{Fe+COD})+{(-485.7 ELSE Ana=(((((Cyno*{NH4*COD})+{((NH4*431.3 THEN Ana=(((((Cyno/Chla)*Cond)+{(-160.737)/NH4})+{169.424/NO2}))				The best ruleset is: IF ((Si*Si)<84.401) THEN Ana=(((((Cyno/Secchi)*5.956)+447.034)+424.345) ELSE Ana=(((((Cyno/Chla)*Cond)+{(-480.737)/NH4})+{169.424/NO2}))									
total error=2038.4597 total R ² value=0.85 total error=2319.0669 total R ² value=0.86																	total error=2840.7964 total R ² value=0.80 total error=3135.7087 total R ² value=0.76				R ² =0.89 R ² =0.86 R ² =0.80 R ² =0.76													
Date	Turb	Cond	Temp	Chla	Cyno	pH	PO4	NO3	Si	Fe	Mn	COD	Secchi	NO2	DO	NH4	IF	THEN	ELSE	IF 1	IF 2	IF both	THEN	ELSE	IF	THEN	ELSE	IF	THEN	ELSE	0 days	7 days	14 days	21 days
26-01-2010	58.60	18.00	20.26	2.81	1256	7.34	0.001	0.104	5.12	0.591	0.002	9.00	30.00	0.025	6.48	0.001	FALSE	7401.3227	1070.62558	TRUE	TRUE	TRUE	1289.53435	246.548252	TRUE	7417.99484	-95.792444	TRUE	1120.73687	-465772.4	1071	1290	7418	1121
27-01-2010	58.41	17.96	20.33	4.42	1658.4	7.36	0.001	0.106	5.11	0.580	0.003	9.43	29.75	0.025	6.49	0.002	FALSE	7402.59694	1176.85195	TRUE	TRUE	TRUE	1358.32861	376.436168	TRUE	3516.34203	9.20587318	TRUE	1203.39919	-203559.63	1177	1358	3516	1203
28-01-2010	58.21	17.93	20.40	6.04	2060.9	7.39	0.001	0.107	5.10	0.568	0.004	9.86	29.50	0.025	6.50	0.003	FALSE	7403.97264	1284.02048	TRUE	TRUE	TRUE	1432.42411	520.352095	TRUE	2431.41262	125.908896	TRUE	1287.46256	-127383.76	1284	1432	2431	1287
29-01-2010	58.02	17.89	20.47	7.65	2463.3	7.41	0.001	0.109	5.10	0.557	0.004	10.29	29.25	0.026	6.51	0.005	FALSE	7405.44978	1392.13381	TRUE	TRUE	TRUE	1511.58288	680.531493	TRUE	1957.08892	255.573186	TRUE	1372.96292	-91169.279	1392	1512	1957	1373
30-01-2010	57.83	17.86	20.54	9.27	2865.7	7.43	0.001	0.110	5.09	0.546	0.005	10.71	29.00	0.026	6.51	0.006	FALSE	7407.02837	1501.19456	TRUE	TRUE	TRUE	1595.65037	857.652864	TRUE	1716.71671	399.455308	TRUE	1459.93742	-70035.044	1501	1596	1717	1460
31-01-2010	57.64	17.82	20.61	10.88	3268.1	7.46	0.001	0.112	5.08	0.534	0.006	11.14	28.75	0.026	6.52	0.007	FALSE	7408.70841	1611.20538	TRUE	TRUE	TRUE	1684.52557	1052.1266	TRUE	1592.35376	558.811823	TRUE	1548.42453	-56202.566	1611	1685	1592	1548
01-02-2010	57.44	17.79	20.68	12.50	3670.6	7.48	0.001	0.113	5.07	0.523	0.007	11.57	28.50	0.027	6.53	0.008	FALSE	7410.4899	1722.16889	TRUE	TRUE	TRUE	1778.14284	1264.29567	TRUE	1535.19638	734.899294	TRUE	1638.46403	-46456.998	1722	1778	1535	1638
02-02-2010	57.25	17.75	20.76	14.11	4073	7.51	0.001	0.115	5.06	0.512	0.008	12.00	28.25	0.027	6.54	0.010	FALSE	7412.37283	1834.08772	TRUE	TRUE	TRUE	1876.46082	1494.0387	TRUE	1521.39314	928.974285	TRUE	1730.09716	-39228.635	1834	1876	1521	1730
03-02-2010	57.06	17.71	20.83	15.73	4475.4	7.53	0.001	0.117	5.05	0.500	0.009	12.43	28.00	0.027	6.55	0.011	FALSE	7414.35722	1946.96452	TRUE	TRUE	TRUE	1979.45545	1742.0088	TRUE	1537.90527	1142.29336	TRUE	1823.36659	-33660.022	1947	1979	1538	1823
04-02-2010	56.86	17.68	20.90	17.34	4877.9	7.55	0.001	0.118	5.04	0.489	0.009	12.86	27.75	0.028	6.56	0.012	FALSE	7416.44305	2060.80191	TRUE	TRUE	TRUE	2087.11537	2010.18761	TRUE	1577.0032	1376.11307	TRUE	1918.31655	-29543.174	2061	2087	1577	1918
05-02-2010	56.67	17.64	20.97	18.95	5280.3	7.58	0.001	0.120	5.04	0.478	0.010	13.29	27.50	0.028	6.57	0.013	FALSE	7418.63033	2175.60253	TRUE	TRUE	TRUE	2199.43876	2296.33645	TRUE	1633.81392	1631.69	TRUE	2014.99288	-26268.021	2176	2199	1634	2015
06-02-2010	56.48	17.61	21.04	20.57	5682.7	7.60	0.001	0.121	5.03	0.467	0.011	13.71	27.25	0.029	6.57	0.014	FALSE	7420.91906	2291.369	TRUE	TRUE	TRUE	2316.43116	2601.77134	TRUE	1705.11302	1910.28069	TRUE	2113.44308	-22692.886	2291	2316	1705	2113
07-02-2010	56.29	17.57	21.11	22.18	6085.1	7.62	0.001	0.123	5.02	0.455	0.012	14.14	27.00	0.029	6.58	0.016	FALSE	7423.30924	2408.10397	TRUE	TRUE	TRUE	2438.10384	2926.80789	TRUE	1788.68186	2213.14171	TRUE	2213.71644	-20202.191	2408	2438	1789	2214
08-02-2010	56.09	17.54	21.18	23.80	6487.6	7.65	0.001	0.124	5.01	0.444	0.013	14.57	26.75	0.029	6.59	0.017	FALSE	7425.80087	2525.81006	TRUE	TRUE	TRUE	2654.47265	3271.76103	TRUE	1882.9438	2541.52964	TRUE	2315.86406	-18082.533	2526	2544	1883	2316
09-02-2010	55.90	17.50	21.25	25.41	6890	7.67	0.001	0.126	5.00	0.433	0.014	15.00	26.50	0.030	6.60	0.018	FALSE	7428.39394	2644.4899	TRUE	TRUE	TRUE	2965.55713	3636.94483	TRUE	1986.74763	2896.70101	TRUE	2419.939	-16258.447	2644	2696	1987	2420
10-02-2010	55.71	17.46	21.32	27.03	7292.4	7.69	0.001	0.128	4.99	0.421	0.014	15.43	26.25	0.030	6.61	0.019	FALSE	7431.08847	2764.14614	TRUE	TRUE	TRUE	2831.37979	4022.67233	TRUE	2099.23317	3279.91241	TRUE	2525.99632	-14673.572	2764	2831	2099	2526
11-02-2010	55.51	17.43	21.39	28.64	7694.9	7.72	0.001	0.129	4.98	0.410	0.015	15.86	26.00	0.030	6.62	0.020	FALSE	7433.88444	2844.7814	TRUE	TRUE	TRUE	2971.96565	4429.25544	TRUE	2219.74473	3692.42039	TRUE	2634.0932	-13284.968	2885	2972	2220	2634
12-02-2010	55.32	17.39	21.46	30.26	8097.3	7.74	0.001	0.131	4.98	0.399	0.016	16.29	25.75	0.031	6.63	0.022	FALSE	7436.78186	3006.39932	TRUE	TRUE	TRUE	3117.34175	4857.00492	TRUE	2347.77379	4135.48152	TRUE	2744.28905	-12059.352	3006	3117	2348	2744
13-02-2010	55.13	17.36	21.53	31.87	8499.7	7.76	0.001	0.132	4.97	0.387	0.017	16.71	25.50	0.031	6.63	0.023	FALSE	7439.78073	3128.99583	TRUE	TRUE	TRUE	3267.53687	5306.23042	TRUE	2482.9199	4610.35236	TRUE	2856.6456	-10920.525	3129	3268	2483	2857
14-02-2010	54.94	17.32	21.60	33.49	8902.1	7.79	0.001	0.134	4.96	0.376	0.018	17.14	25.25	0.032	6.64	0.024	FALSE	7442.88104	3252.58768	TRUE	TRUE	TRUE	3442.58124	5777.23996	TRUE	2624.86338	5118.28946	TRUE	2971.22703	-9997.5878	3253	3423	2625	2971
15-02-2010	54.74	17.29	21.67	35.10	9304.6	7.81	0.001	0.135	4.95	0.365	0.019	17.57	25.00	0.032	6.65	0.025	FALSE	7446.08281	3377.16534	TRUE	TRUE	TRUE	3582.5063	6270.34104	TRUE	2773.346	5660.5494	TRUE	3088.1001	-9123.667	3377	3583	2773	3088
16-02-2010	54.55	17.25	21.75	36.71	9707	7.84	0.001	0.137	4.94	0.353	0.020	18.00	24.75	0.032	6.66	0.027	FALSE	7449.36082	3502.73592	TRUE	TRUE	TRUE	3747.34458	6785.83956	TRUE	2928.15683	6238.38874	TRUE	3207.33423	-8334.99	3503	3747	2928	3207
17-02-2010	54.36	17.21	21.82	38.33	10109	7.86	0.001	0.139	4.93	0.342	0.020	18.43	24.50	0.032	6.67	0.028	FALSE	7452.79069	3629.29981	TRUE	TRUE	TRUE	3917.12395	7324.04044	TRUE	3048.91292	6853.06404	TRUE	3329.00172	-7620.2062	3629	3917	3089	3329
18-02-2010	54.16	17.18	21.89	39.94	10512	7.88	0.001	0.140	4.92	0.331	0.021	18.86	24.25	0.033	6.68	0.029	FALSE	7456.2968	3756.86206	TRUE	TRUE	TRUE	4091.89525	7885.24757	TRUE	3256.09652	7505.83186	TRUE	3453.17781	-6969.8775	3757	4092	3256	3453
19-02-2010	53.97	17.14	21.96	41.56	10914	7.91	0.001	0.142	4.91	0.319	0.022	19.29	24.00	0.033	6.69	0.030	FALSE	7459.90436	3885.48249	TRUE	TRUE	TRUE	4271.67673	8469.76383	TRUE	3428.9592	8197.94876	TRUE	3579.9409	-6376.0931	3885	4272	3429	3580
20-02-2010	53.78	17.11	22.03	43.17	11317	7.93	0.001	0.143	4.91	0.308	0.023	19.71	23.75	0.034	6.69	0.031	FALSE	7463.61336	4014.98924	TRUE	TRUE	TRUE	4456.50941	9077.89116	TRUE	3607.60736	8930.67131	TRUE	3709.3727	-5832.1729	4015	4457	3608	3709
21-02-2010	53.59	17.07	22.10	44.79	11719	7.95	0.001	0.145	4.90	0.297	0.024	20.14	23.50	0.034	6.70	0.033	FALSE	7467.42382	4145.55955	TRUE	TRUE	TRUE	4646.42929	9709.93067	TRUE	3791.95368	9705.25607	TRUE	3841.55836	-5332.4381	4146	4646	3792	3842
22-02-2010	53.39	17.04	22.17	46.40	12122	7.98	0.001	0.146	4.																									

7.6 Case study : Rand Water

Rand Water, initially used the World Health Organisation (WHO) framework as guide to develop its own WSP and implemented it in 2003 (Viljoen, 2010), years before it was required under the Blue Drop Certification Program. Water quality and the safety of their consumers have always been a high priority and to date Rand Water has never recorded a health-related water quality incident (Viljoen, 2010). However, with increasing eutrophication and deteriorating quality of source waters all over the world, Rand Water has always been aware of potential risks to their consumers (Steynberg et al., 1998) and has been open for new ideas and new technologies to aid in their constant supply of safe and aesthetically acceptable drinking water. Therefore, in the South African context, the most logical place to start pursuing the application of forecasting models in cyanobacteria and cyanobacterial-related organic compounds would be Rand Water.

The incident management protocols utilised by Rand Water are those for 1) *E. coli*, 2) pathogenic protozoa (*Cryptosporidium* and *Giardia*), 3) algal toxins (microcystin), 4) taste and odour (geosmin and MIB) and 5) dirty water (turbidity) (Viljoen, 2010). For the purpose of this study, the prediction models will be applied to the one for algal toxins (microcystin) and taste and odour (geosmin and MIB). The models forecasting different *Microcystis*^[7] (Figure 5.11), *Anabaena*^[7] (Figure 5.3), microcystin^[7] (Figure 5.19) and geosmin^[7] (Figure 5.27) seven days in advance have been chosen to be incorporated into the protocols. These models were chosen because of the following reasons: 1) they have the highest R^2 -value when compared to the 25% boot-strapped testing dataset as well as the three years unseen dataset, 2) the data availability in the case of the Rand Water source water is available within 24h and 3) because seven days allow most DWTW enough time to prepare for a cyanobacterial-related incident in the source water⁵. Since geosmin is not analysed in the Vaal Dam, the only model available for incorporation was the one in the canal predicting geosmin concentrations 7 days in advance. Although this model did not prove to be statistically significant, it was decided to include it in the protocol for completeness. This model can be optimised in future and then it will be easy to slot into the correct position and utilise it alongside the other models in this protocol. The models predicting the total cyanobacterial cell concentrations in each case, have a lower R^2 -value than those predicting the *Anabaena* sp. and *Microcystis* sp. concentrations individually, although 99.47% of the “total cyanobacteria” in the Vaal Dam and canal is made up

⁵ Personal communication Mr. Danie du Plessis, Operations Manager Potable Water, Zuikerbosch, Rand Water.

of *Anabaena* sp. + *Microcystis* sp. Therefore, one main optimising strategy applied during this study was to use the models predicting *Anabaena* sp. and *Microcystis* sp. separately, determine the sum and use that value as “total cyanobacteria” input in the WSP instead of the models predicting total cyanobacterial concentrations (Figures 7.5 and 7.6). It is evident that there are very specific (and different) environmental drivers influencing the occurrence of *Anabaena* sp. and *Microcystis* sp. respectively and therefore the sum of the models predicting *Anabaena* sp. and *Microcystis* sp. is more accurate than the model predicting “total cyanobacteria”.

In the Rand Water context, the initial reporting responsibility is situated within the Biology Section of the Analytical Services Department and therefore the responsibility of the daily updating of the spreadsheets will not be situated at the plant, but at the laboratory. The protocols in the WSP require input of “total cyanobacteria”, “geosmin”, “MIB” and “microcystin” concentrations. In the Vaal Dam and canal, the occurrence of MIB is very rare and do not pose a significant risk to the Rand Water DWTW (in fact, 86% of the MIB results in the canal show results below the detection limit) and therefore no models predicting MIB were developed with data from the Vaal Dam or canal. However, at certain DWTW, MIB might be of importance and therefore models predicting MIB results can be developed and also incorporated into the “Prediction Monitoring” of the taste and odour incident management protocol if deemed necessary.

The changes proposed to be implemented into the microcystin protocol are displayed in Figure 7.5, where an additional “Prediction Monitoring Level” (either situated with the laboratory personnel or with the station chemist at the DWTW) will be established and will run simultaneously with the “Routine Monitoring Program”. Since the “Vigilance Level” (Figure 7.2), mostly only apply to reaction within the laboratory and in reality do not differ much from the “Routine Monitoring Program”, it was removed from the proposed protocol (Figure 7.5) for simplification purposes. Therefore the “Routine Monitoring Program” does not reach “Alert Level 1” through the “Vigilance Level” (Figure 7.2), but directly, as in the case of the taste and odour protocol (Figure 7.3).

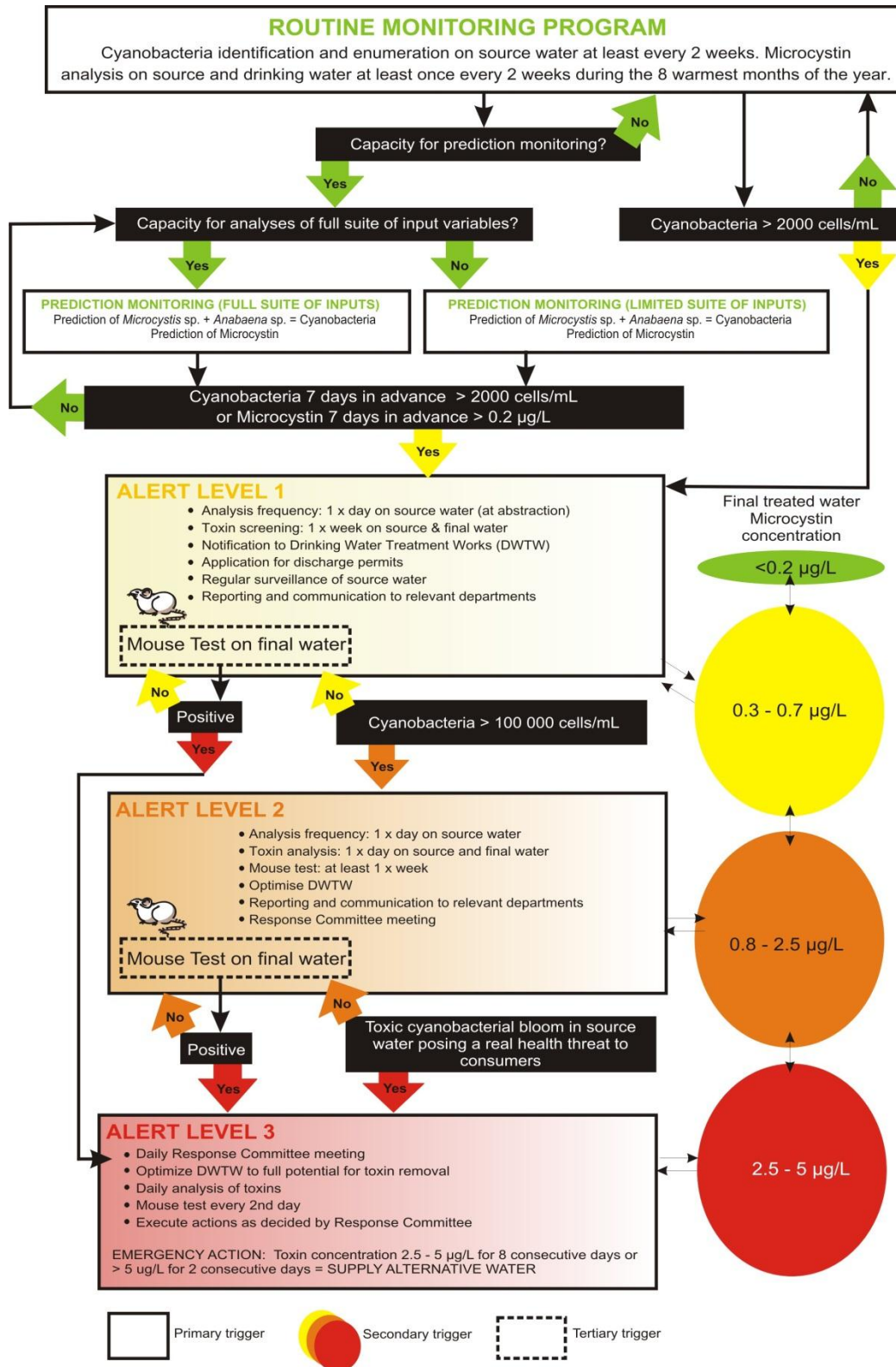


Figure 7.5 Proposed microcystin protocol to include future predictions of cyanobacterial cell concentrations (*Anabaena* sp. + *Microcystis* sp.) as well as microcystin concentration.

In the proposed microcystin protocol (Figure 7.5), “Alert Level 1”, can therefore either be reached through the “Routine Monitoring Program” or through “Prediction Monitoring” (and here preferably through the “full suite of inputs” rather than the “limited suite of inputs”). The “Prediction Monitoring” will not influence the actions associated with “Alert Level 2” or “Alert Level 3”, because by then it has already fulfilled its purpose of an early warning system.

The changes proposed to be implemented into the taste and odour protocol is displayed in Figure 7.6, where an additional “Prediction Monitoring Program” will be established (in the case of Rand Water situated with the laboratory personnel and not with the station chemist at the DWTW). In this protocol (Figure 7.6), “Alert Level 1” can be reached through 1) the routine monitoring program or 2) prediction monitoring (either utilising the models incorporating the full suite of input variables or the models utilising the limited suite of input variables) or 3) consumer complaints (as tertiary trigger). The “Prediction Monitoring” will not influence the actions associated with “Alert Level 2” because when “Alert Level 2” is activated the “Prediction Monitoring” would already have fulfilled its early warning purpose.

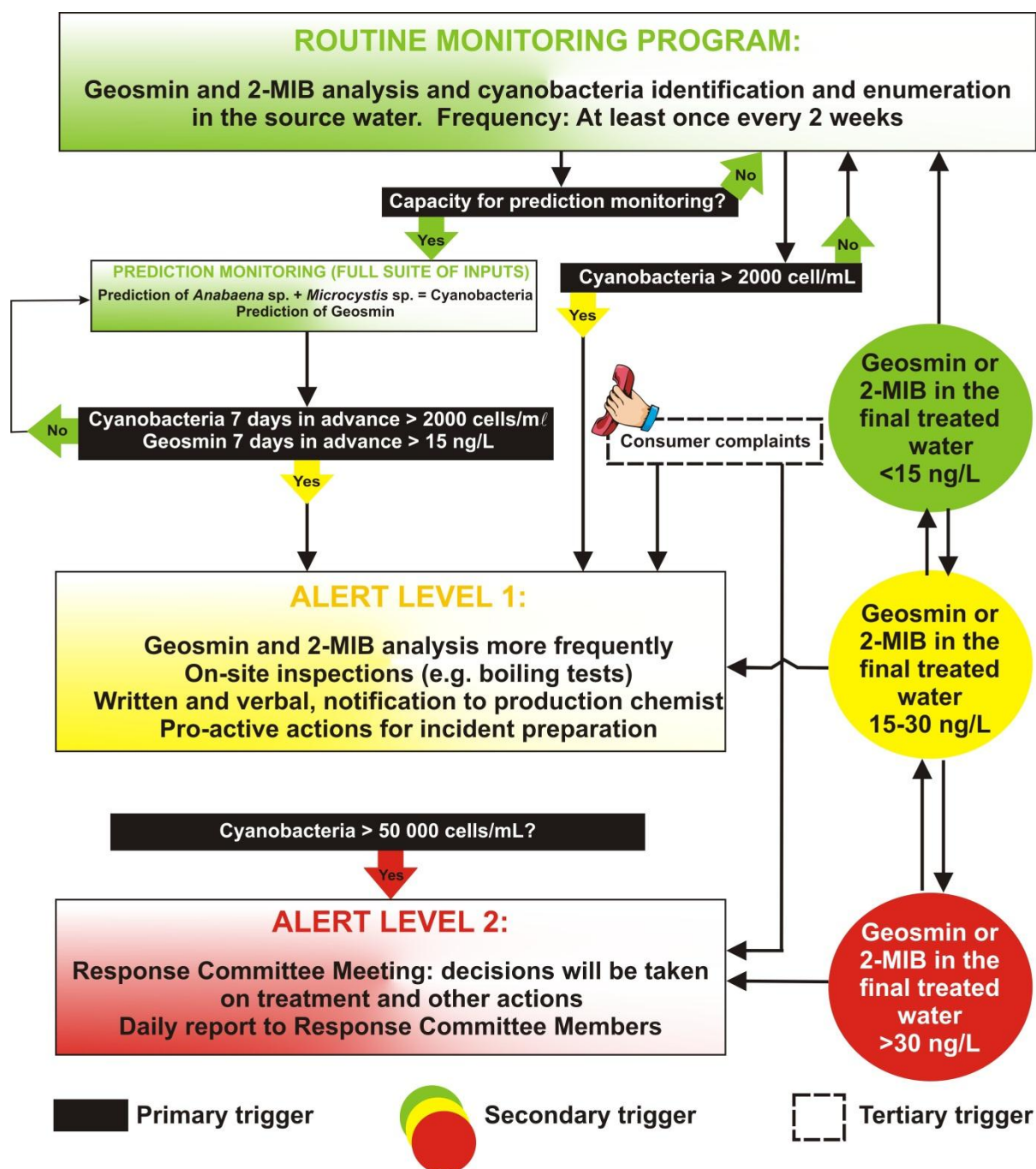


Figure 7.6 Proposed taste and odour protocol to include future predictions of cyanobacterial cell concentrations (*Anabaena* sp. + *Microcystis* sp.) as well as microcystin concentration.

In this instance the models incorporated into the protocols of Rand Waters' water safety plan (WSP) is from the models derived from laboratory analyses. If however, certain DWTW do not have the same capacity for analyses as Rand Water, the models where *in situ* determinable variables were used can be incorporated into such a WSP. Although models that were initially developed by data from laboratory analyses (such as the models from this study) can be used by inputting data from *in situ* instruments, the prediction results are not as accurate as would be desired. And therefore it became evident during this study that if *in situ* measured data (and not that of laboratory analyses) were to be used in the prediction models, the models need to be developed and trained by data from the *in situ* instruments.

7.7 Conclusions

The multi-barrier approach in risk reduction as promoted by the concept of water safety plans under the banner of the Blue Drop Certification Program, lends itself to the application of future predictions of water quality variables. In this study, prediction models of *Anabaena* sp., *Microcystis* sp. and microcystin concentrations 7 days in advance from the Vaal Dam, as well as geosmin concentration 7 days in advance from the canal were incorporated into the proposed incident management protocols. This was managed by adding an additional "Prediction Monitoring Level" of the microcystin and taste and odour incident management protocols, to also include future predictions of cyanobacterial cell concentrations (*Anabaena* sp. + *Microcystis* sp.), microcystin and geosmin. The incorporation of future predictions, adds another barrier in the potential exposure of drinking water consumers to harmful and aesthetically unacceptable organic compounds produced by cyanobacteria.

CHAPTER 8

GENERAL CONCLUSIONS AND RECOMMENDATIONS

8.1 *Introduction*

With increasing eutrophication and pollution in source waters all over the world, the impact that algae (including cyanobacteria) and cyanobacterial-related organic compounds have on the effectiveness of the water treatment process is becoming more profound each year. Algal blooms in the source water (especially blooms of certain notorious cyanobacteria) are known to cause major problems in the drinking water treatment works (DWTW), such as filter clogging, disruption of the coagulation and flocculation unit processes, production of cyanotoxins, production of taste and odorous substances and penetration into the final drinking water with obvious aesthetical and health implications to the consumer. Cyanobacterial blooms in the source water are also the cause of major increases in drinking water treatment costs due to the fact that conventional treatment methods are very limited in its ability to remove or oxidise organic compounds produced during a cyanobacterial bloom. A problem that DWTW operators and managers have to face currently is that there is no indication of when such a cyanobacterial bloom will occur and how large the impact thereof will be on the purification process. Therefore, water purification operators are mostly unprepared when such blooms suddenly appear and it then takes a few days or more to adjust the water purification process to such an extent that it can deal effectively with the changes in water chemistry as caused by the cyanobacterial bloom, which by then, may have already impacted negatively on the consumer. From this study it was apparent from the onset that the application value of models developed by Hybrid Evolutionary Algorithms to predict cyanobacterial cell concentrations and cyanobacterial-related organic compounds can aid substantially in the ability of a large or small DWTW to prepare for such an incident.

8.2 *General conclusions*

This study have shown that hybrid evolutionary algorithms can successfully be used to develop relatively accurate forecasting models that can predict cyanobacterial cell concentrations (particularly *Anabaena* sp. and *Microcystis* sp.) as well as the cyanotoxin microcystin in the Vaal Dam, for up to 21 days in advance (depending on the output variable and the model applied).

The models developed during this study (although no optimisation strategies like hill climbing or differential evolution were performed) were generally more accurate than most models developed by other authors utilising the same concepts and even models optimised by hill climbing and/or differential evolution. The forecasting models that performed the best were those forecasting 7 days in advance ($R^2 = 0.86, 0.91$ and 0.75 for *Anabaena*^[7], *Microcystis*^[7] and microcystin^[7] respectively). It is speculated that including “initial cyanobacteria inoculum” as input variable (which is unique to this study), is most probably the reason for the better performing models. The results show that models developed from on-line (*in situ*) measureable data only, are almost as good as the models developed by using all possible input variables. The reason is most probably because “initial cyanobacteria inoculum” – the variable towards which the output result showed the greatest sensitivity towards – is included in these models. Generally models predicting *Microcystis* sp. in the Vaal Dam were more accurate than models predicting *Anabaena* sp. concentrations and models with a shorter prediction time (e.g. 7 days in advance) were statistically more accurate than models with a longer prediction time (e.g. 14 or 21 days in advance). Testing the models with previously “seen” data (i.e. 25% of the boot-strapped results from the testing dataset that was part of the 10 years development dataset) gave a better correlation than testing it with “unseen” data.

The only two optimisation strategies applied during this study were 1) using rule-based strategy together with hybrid evolutionary algorithms, which separate mathematical optimisation in different ranges of certain input variables (i.e. high range rule set and low range rule set) and 2) rather than using the models predicting “cyanobacterial cell concentration”, the models predicting *Anabaena* sp. and *Microcystis* sp. separately were used and the sum thereof was used as the “total cyanobacteria” input in the incident management protocol since it was more accurate.

The geosmin models developed with data from the canal, were not regarded as significantly correlated – indicating that 1) the variables used to develop these models did not include the most important drivers causing cyanobacteria to produce geosmin and/or 2) that actinomycetes may probably also be involved in producing geosmin in the canal.

The models that were specifically developed with the real time, on-line (*in situ*) instrument in mind did not perform well at all when data from the on-line instrument was used to test the models. It was obvious from this exercise that if data from on-line (*in situ*) measuring instruments is to be used in these models, the training of the models needs to be done with data from the instruments and not from laboratory generated data as was done in this study.

However, data from the on-line (*in situ*) measuring instrument can still be useful in indicating trends in increasing cyanobacterial cell concentrations in source waters and can still be utilised as part of the early warning system for cyanobacteria and cyanobacterial-related organic compounds.

This study illustrated that incorporating the models into the incident management protocols under “Prediction Monitoring” which will run simultaneously with the “Routine Monitoring” of a DWTW, is most probably the best way to utilise these models in practice. The Blue Drop Certification Program developed by DWA (currently known as DWS), provides the perfect platform under which these models can be applied to aid in early warning regarding cyanobacteria and cyanobacterial-related organic compounds. The novelty of this study lies in the application of prediction models in the water safety plan of a DWTW, something that has never been done before.

8.3 Recommendations

- The models presented in this study were performed on water from the Vaal Dam and canal supplying the Zuikerbosch DWTW. It is therefore important to note that these models are most probably site specific and can not necessarily be applied to source waters elsewhere. Therefore, before models from this study are used or applied in other catchments it is important to validate it with data from that catchment.
- If these models are tested and shown not to perform as well with data from other catchments, models can be developed with data from those catchments which would most probably result in more accurate forecasting.
- The models forecasting geosmin concentrations in the canal did not perform as well as the others and therefore a few remedies may be considered:
 - Investigate the influence of actinomycetes in the production of geosmin in the canal.
 - Re-train the models for geosmin prediction and also include “initial geosmin concentration” as input variable for future predictions. This might capture the variable (not currently measured in the canal), which might be crucial in explaining the production of geosmin by cyanobacteria.
 - Consider performing hill climbing and differential evolution as optimisation strategies on the geosmin models.

-
- If data from remote sensing on-line (*in situ*) instruments are planned to be used in the forecasting models, the training dataset needs to consist of data from such instruments.
 - DWTW that does not have data readily available for the application into the WSP should consider the following:
 - Prioritise results needed from the laboratory as input variables to be available in time for prediction monitoring to still be useful to the DWTW.
 - Rather than utilising the models with longer prediction times (e.g. 14 or 21 days in advance), DWTW should first opt for models predicting 7 days in advance but utilising limited input variables such as described in Chapter 6.
 - If utilising models with limited input variables does not provide enough time for prediction monitoring then models with longer prediction times can be considered.
 - Application of forecasting models in the avenue of recreational use of water bodies should be investigated. Forecasting models can aid organisers of recreational events to anticipate water quality when events are planned in impoundments not utilised for such activities on a daily basis (e.g. Midmar Mile, Duzi Canoe Marathon or Berg River Canoe Marathon).

CHAPTER 9

REFERENCES

- Alonso Fernández, J.R., Díaz Muñoz, C., Garcia Nieto, P.J. and De Cos Juez, F.J. Forecasting the cyanotoxins presence in fresh waters: A new model based on genetic algorithms combined with the MARS technique. *Ecological Engineering*, **53**, 68 – 78, (2013).
- Annoni, P., Bruggemann, R. and Saltelli, A. Random and quasi-random designs in variance-based sensitivity analysis for partially ordered sets. *Reliability Engineering and System Safety*, **107**, 184 – 189, (2012).
- APHA *Standard Methods for the Examination of Water and Wastewater*. 22nd Edition. American Public Health Association, Washington D.C., (2013).
- Bartram, J., Corrales, A., Davison, A., Deere, D., Drury, D., Gordon, B., Howard, G., Rinehold, A. and Stevens, M. *Water safety plan manual step-by-step risk management for drinking water suppliers*. World Health Organisation, Geneva, (2009).
- Bennet, N.D., Croke, F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D. and Andreassian, V. Characterising performance of environmental models. *Environmental Modelling and Software*, **40**, 1 - 20, (2013).
- Cabestany, J., Prieto, A., Sandoval, D.F. (Eds.). *The Curse of Dimensionality in Data Mining and Time Series Prediction*. IWANN 2005, LNCS 3512, Springer-Verlag, Berlin, Heidelberg, 758 – 770, (2005).
- Campolongo, F., Saltelli, A. and Cariboni, J. From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, **182**, 978 – 988, (2011).
- Cao, H., Recknagel, F. and Orr, P.T. Enhanced functionality of the redesigned hybrid evolutionary algorithm HEA demonstrated by predictive modelling of algal growth in the Wivenhoe Reservoir, Queensland (Australia), *Ecological Modelling*, **252**, 32 – 43, (2013).
- Cao, H., Recknagel, F., Welk A., Kim, B. and Takamura, N. 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 Edition*. Springer-Verlag Berlin, Heidelberg, New York, 330 - 342, (2006).

-
- Chang, D-W, Hobson, P., Burch, M. and Lin, T-F. Measurement of cyanobacteria using in-vivo fluoroscopy - Effect of cyanobacterial species, pigments, and colonies. *Water Research*, **46**, 5037 – 5048, (2012).
- Chan, W.S., Recknagel, F., Cao, H. and Ho_dong, P. Elucidation and short-term forecasting of Microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms. *Water Research*, **41**, 2247 - 2255, (2007).
- Chorus, I. *Current approaches to cyanotoxin risk assessment, risk management and regulations in different countries*. Federal Environment Agency, Wörlitzer Platz 1, 06833 Dessau-Roßlau, Germany, (2012).
- Codd, G.A., Azevedo, S.M.F.O., Bagchi, S.N., Burch, M.D., Carmichael, W.W., Harding, W.R., Kaya, K. and Utkilen, H.C. CYANONET A global network for cyanobacterial bloom and toxin risk management. IHP-IV. Technical Documents in Hydrology, **76**, UNESCO, Paris, 138pp., (2005).
- Conradie, R.C. and Barnard, S. The dynamics of toxic *Microcystis* strains and microcystin production in two hypertrophic South African reservoirs. *Harmful Algae*, **20**, 1 – 10, (2012).
- Damásio, J. Tauler, R., Teixidó, E., Rieradevall, M, Prat, N., Riva, M.C, Soares, A.M. and Barata, C. Combined use of *Daphnia magna* in situ bioassays, biomarkers and biological indices to diagnose and identify environmental pressures on invertebrate communities in two Mediterranean urbanized and industrialized rivers (NE Spain). *Aquatic Toxicology*, **87** (4), 310 - 320, (2008).
- Davidson, K., Gowen, R.J., Tett, P., Bresnan, E., Harrison, P.J., McKinney, A., Milligan, S., Mills, D.K., Silke, J. and Crooks A-M. Harmful algal blooms: How strong is the evidence that nutrient ratios and forms influence their occurrence? *Estuarine, Coastal and Shelf Science*, **115**, 399 – 413, (2012).
- Davison, A., Howard, G., Stevens, M., Callan, P., Fewtrell, L., Deere, D. and Bartram, J. Water Safety Plans: Managing Drinking-water Quality from Catchment to Consumer. Report WHO/SDE/WSH/05.06. *Water Sanitation and Health*, Geneva, Switzerland, (2005).
- Dominguez-Chicas, A. and Scrimshaw, M.D. Hazard and risk assessment for indirect potable reuse schemes : An approach for use in developing Water Safety Plans. *Water Research*, **44**, 6115 – 6123, (2010).

-
- Downing, S., Banack, S.A., Metcalf, J.S., Cox, P.A. and Downing, T.G. Nitrogen starvation of cyanobacteria results in the production of β -N-methylamino-L-alanine. *Toxicon*, **58**, 187 – 194, (2011).
- Du Preez, H. and Van Baalen, L. *Generic Incident Management Framework for toxic blue-green algal blooms, for application by potable water suppliers*. Water Research Commission, South Africa. WRC Report No. : T263/06, (2006).
- Du Preez, H., Swanepoel, A., Van Baalen, L. and Oldewage A. Cyanobacterial Incident Management Frameworks (CIMFs) for application by drinking water suppliers. *Water SA* **33** (5), 643 – 652, (2007).
- DWA (Department of Water Affairs) *2011 Blue Drop Report*. Department of Water Affairs, Pretoria, South Africa, (2011).
- DWA (Department of Water Affairs) *2012 Blue Drop Report*. Department of Water Affairs, Pretoria, South Africa, (2012).
- DWA (Department of Water Affairs) *Trophic status of South African Impoundments*. From the National Eutrophication Monitoring Programme, Department of Water Affairs (DWA), Pretoria, South Africa, <http://www.dwa.gov.za/iwqs/eutrophication/NEMP/nempdams.htm> (date accessed: 05-07-2013a).
- DWA (Department of Water Affairs). *The Orange River basin in South Africa*. <http://www.dwaf.gov.za/orange/Vaal/vaaldam.htm> (date accessed: 10-07-2013b).
- Dzialowski, A.R., Smith, V.H., Huggins, D.G., deNoyelles, F., Lim, N-C., Baker, D.S. and Beury, J.H. Development of predictive models for geosmin-related taste and odor in Kansas, USA, drinking water reservoirs. *Water Research*, **43**, 2829 – 2840, (2009).
- EnvirologixTM QuantiPlateTM Kit for Microcystins Leaflet. Catalog No.: EP 022, 500 Riverside Industrial Parkway, Portland, ME 04103-1486, USA, 7 pp., (2013).
- Esterhuizen, M. and Downing, T.G. β -N-methylamino-L-alanine (BMAA) in novel South African cyanobacterial isolates. *Ecotoxicology and Environmental Safety*, **71**, 309 - 313, (2008).
- Ewerts, H., Swanepoel, A. and Du Preez, H.H. Efficacy of conventional drinking water treatment processes in removing problem-causing phytoplankton and associated organic compounds. *Water SA*, **39** (5), 739 – 750, (2013).
- Ferreira, L. and Du Preez, H.H. Investigation into the occurrence of aquatic invertebrates throughout drinking water purification plants. *Water Science and Technology: Water Supply*, **12** (2), 250 – 257, (2012).

-
- Fey, M.V. and Guy, S.A. The capacity of soils in the Vaal Dam catchment to retain sulphate from atmospheric pollution. Water Research Commission, South Africa. WRC Report No. : 414/1/93 87 pp., (1993).
- García Nieto, P.J., Alonso Fernández, J.R., De Cos Juez, F.J., Sánchez Lashera, F. and Díaz Muñoz, C. Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). *Environmental Research*, **122**, 1 – 10, (2013).
- Google Earth, Version 7.1.2.2041, Google Inc. (2013). <https://earth.google.com> (date accessed: 16-07-2013).
- Graham, M., Blignaut, J., De Villiers, L., Mostert, D., Sibande, X., Gebremedhin, S., Harding, W., Rossouw, N., Freese, N.S., Ferrer, S. and Browne, M. *Development of a generic model to assess the costs associated with eutrophication*. Water Research Commission, South Africa. WRC Report No. : 1568/1/12, (2012).
- Guven, B. and Howard, A. Modelling the growth and movement of cyanobacteria in river systems. *Science of the Total Environment*, **368**, 898 – 908, (2006).
- Hair, J.F. (Jr.), Black, W.C., Babin, B.J. and Anderson, R.E. *Multivariate data analysis*. ISBN-10: 0138132631 | ISBN-13: 9780138132637, Prentice Hall, Upper Saddle River, NJ 07458, 116 - 120, (2009).
- Hamilton, P.D., Gale, P. and Pollard, S.J.T. A commentary on recent water safety initiatives in the context of water utility risk management. *Environment International*, **32**, 958 - 966, (2006).
- Harding, W.R. and Paxton, B.R. *Cyanobacteria in South Africa: A Review*. Water Research Commission, South Africa. WRC Report No. : TT 153/01, (2001).
- Heisler, J., Glibert, P.M., Burkholder, J.M., Anderson, D.M., Cochlan, W., Dennison, W.C., Dortch, Q., Gobler, C.J., Heil, C.A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H.G., Sellner, K., Stockwell, D.A., Stoecker, D.K. and Suddleson, M. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae*, **8**, 3 – 13, (2008).
- Ho, L., Lambling, P., Bustamante, H., Duker, P. and Newcombe, G. Application of powdered activated carbon for the adsorption of cylindrospermopsin and microcystin toxins from drinking water supplies. *Water Research*, **45**, 2954 – 2964, (2011).
- Huang, J., Gao, J. and Hörmann, G. Hydrodynamic-phytoplankton model for short-term forecasts of phytoplankton in Lake Taihu, China. *Limnologia*, **42**, 7 – 18, (2012).

-
- Jacoby, J.M., Collier, D.C., Welch, E.B., Hardy, F.J. and Crayton, M. Environmental factors associated with a toxic bloom *Microcystis aeruginosa*. *Can. J. Fish. Aquat. Sci.*, **57**, 231 - 240, (2000).
- Jähnichen, S., Long, B.M. and Petzoldt, T. Microcystin production by *Microcystis aeruginosa*: Direct regulation by multiple environmental factors. *Harmful Algae*, **12**, 95 – 104, (2011).
- John, D.M., Whitton, B. A. and Brook, A.J. (Eds). *The Freshwater Algal Flora of the British Isles - An Identification guide to freshwater and terrestrial algae*. ISBN: 0 521 77051 3. Cambridge University Press, The Edinburgh Building Cambridge, CB2 2RU UK, 707 pp., (2005).
- Joung, S-H., Oh, H-M., Ko, S-R., and Ahn, C-Y. Correlations between environmental factors and toxic and non-toxic *Microcystis* dynamics during bloom in Daechung Reservoir, Korea. *Harmful Algae*, **10**, 188 – 193, (2011).
- Kasprzak, P., Padisák, J., Koschel, R., Krienitz, L. and Gervais, F. Chlorophyll *a* concentration across a trophic gradient of lakes: An estimator of phytoplankton biomass? *Limnologia*, **38**, 327 – 338, (2008).
- Knappe, D.R.U., Belk, R.C, Briley, D.S., Gandy, S.R., Rastogi, N., Rike, A.H., Galsgow, H., Hannon, E., Frazier, W.D., Kohl, P. and Pugsley, S. *Algae Detection and Removal Strategies for Drinking Water Treatment Plants*. Report number 90971, AWWA Research Foundation, U.S.A., 466 pp., (2004).
- LHDP (Lesotho Highlands Development Project) *Overview of the Lesotho Highlands Water Project*. <http://www.lhwp.org.ls/overview/default.htm> (date accessed: 10-07-2013).
- Lund, J.W.G., Kipling, C. and Le Cren, E.D. The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting. *Hydrobiologia*, **11**, 143 - 170, (1958).
- Makler-Pick, V., Gal, D., Gorfine, M., Hipsey, M. R. and Carmel, Y. Sensitivity analysis for complex ecological models – a new approach. *Environmental Modelling & Software*, **26**, 124 – 134, (2011).
- Matete, M. and Hassan R. Integrated ecological economics accounting approach to evaluation of inter-basin water transfers: An application to the Lesotho Highlands Water Project. *Ecological Economics*, **60**, 246 – 259, (2006).
- Mercer, J. and Bartram, J. *Worldwide Regulatory Strategies and Policies for Drinking Water*. Elsevier B.V., (2011).

-
- Mouchtouri, V.A., Bartlett, C.L.R., Diskin, A. and Hadjichristodoulou, C. Water Safety Plan on cruise ships: A promising tool to prevent waterborne diseases. *Science of the Total Environment*, **429**, 199 – 205, (2012).
- Mohamed, Z.A., El-Sharouny, H.A. and Ali, W.S.M. Microcystin production in benthic mats of cyanobacteria in the Nile River and irrigation canals, Egypt. *Toxicon*, **47**, 584 - 590, (2006).
- Newcombe, G. and Nicholson, B., Water treatment options for dissolved cyanotoxins. *Water Supply: Research and Technology-aqua*, **53** (4), 227 – 239, (2004).
- Newcombe, G., (ed.) *International Guidance Manual for the Management of Toxic Cyanobacteria*. GWRC (Global Water Research Coalition), Alliance House, London, SW1H0QS, UK, (2009).
- Newcombe, G., House, J., Ho, L., Baker, P. and Burch, M. *Management Strategies for Cyanobacteria (Blue-Green Algae): A Guide for Water Utilities*, Water Quality Research Australia Research, Report No. 74, (2010).
- NHMRC (National Health and Medical Research Council – Australia) *Australian drinking water guidelines*. Canberra: NHMRC, (2011).
- NICD (National Institute of Communicable Diseases) Typhoid fever outbreak in Delmas. *Communicable Disease Communiqué, September 2005*, (2005).
- Norton, J.P. Algebraic sensitivity analysis of environmental models. *Environmental Modelling & Software*, **23**, 963 – 972, (2008).
- Nossent, J., Elsen, P. and Bauwens, W. Sobol' sensitivity analysis of a complex environmental model. *Environmental Modelling & Software*, **26**, 1515 – 1525, (2011).
- NZMOH (New Zealand Ministry of Health) *How to prepare and develop public health risk management plans for drinking-water supplies*. Wellington: Ministry of Health, (2001).
- Paerl, H.W. and Huisman, J. Blooms like it hot. *Science*, **320**, 57 – 58, (2008).
- Pantelić, D., Svirčev, Z., Simeunović, J., Vidović, M. and Trajković, I. Cyanotoxins: Characteristics, production and degradation routes in drinking water treatment with reference to the situation in Serbia. *Chemosphere*, **91**, 421 – 441, (2013).
- Pieterse, A.J.H., Basson, N.D., Cloot, A., Mienie, L.J., Traut, D.F. and Visser, R. *The occurrence, distribution and removal of algal species and related substances in a full-scale water purification plant*. Water Research Commission, South Africa. WRC Report No. : 567/1/00, (2000).

-
- Raine, R., McDermott, G., Silke, J., Lyons, K., Nolan, G. and Cusack, C. A simple short range model for the prediction of harmful algal events in the bays of southwestern Ireland. *Journal of Marine Systems*, **83** (3 – 4), 150 – 157, (2010).
- Randolph, K., Wilson, J., Tedesco, L., Li, L.D., Pascual, L. and Soyeux, E. Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll *a* and phycocyanin. *Remote Sensing of Environment*, **112**, 4009 – 4019, (2008).
- Ravalico, J.K., Maier, H.R. and Dandy, G.C. Sensitivity analysis for decision-making using the MORE method – A Pareto approach. *Reliability Engineering and System Safety*, **94**, 1229 - 1237, (2009).
- Recknagel, F., Van Ginkel, C., Cao, H., Cetin, L. and Zhang, B. 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. *Ecological Modelling*, **215**, 144 – 158, (2008).
- Recknagel, F., Orr, P. and Cao, H. Inductive reasoning and forecasting of population dynamics of *Cylindrospermopsis raciborskii* in three sub-tropical reservoirs by evolutionary computation. *Harmful Algae*, **31**, 26 – 34, (2014).
- Recknagel, F., Ostrovsky, I., Cao, H., Zohary, T. and Zhang, X. Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computations and wavelets. *Ecological Modelling*, **255**, 70 – 86, (2013).
- Reynolds, C.S. *Ecology of Phytoplankton*. Cambridge University Press, New York, (2006).
- SANS 241-1. *South African National Standard for Drinking Water*, (2011).
- SANS 241-2. *Application of SANS 241-1 South African National Standard for Drinking Water*, (2011).
- Salacinska, K., El Serafy, G.Y., Los, F.J. and Blauw, A. Sensitivity analysis of the two dimensional application of the Generic Ecological Model (GEM) to algal bloom prediction in the North Sea. *Ecological Modelling*, **221**, 178 – 190, (2010).
- Sartory, D.P. *Spectrophotometric analysis of chlorophyll-a in freshwater phytoplankton*. Department of Environmental Affairs Technical Report, South Africa, Report number : TR 115, (1982).
- Sekadende, B.C., Lyimo, T.J. and Kurmayer, R. Microcystin production by cyanobacteria in the Mwanza Gulf (Lake Victoria, Tanzania). *Hydrobiologia*, **543**, 299 – 304, (2005).

-
- Shahsavani, D. and Grimvall, A. Variance-based sensitivity analysis of model outputs using surrogate models. *Environmental Modelling & Software*, **26**, 723 – 730, (2011).
- South Africa.Info, South Africa's weather and climate. Brand South Africa Country Portal <http://www.southafrica.info/travel/advice/climate.htm> (accessed 22/07/2013).
- Srinivasan, R., and Sorial, G.A. Treatment of taste and odor causing compounds 2-methyl isoborneol and geosmin in drinking water: A critical review. *Journal of Environmental Sciences*, **23** (1), 1 – 13, (2011).
- Steynberg, M.C., Adam, K. and Pieterse, A.J.H. An algal monitoring protocol: The strategic link between reservoir and treatment process. *Water Science and Technology – Reservoir Management and Water Supply*, **37** (2), 153 – 159, (1998).
- Steynberg, M.C. Aspekte van die invloed van eutrofikasie op die Vaalrivierbarrage. Dissertation submitted as fulfilment for the degree, Magister Scientiae in the Faculty of Natural Science, Department of Botany, University of the Free State, South Africa, 52 – 53, (1986).
- Storey, M.V., Van der Gaag, B and Burns, B.P. Advances in on-line drinking water quality monitoring and early warning systems. *Water Research*, **45**, 741 – 747, (2011).
- Storn, R. and Price, K. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, **11** (4), 341 – 359, (1997).
- Su, M., Gaget, V., Giglio, S., Burch, M., An, W. and Yang, M. Establishment of quantitative PCR methods for the quantification of geosmin-producing potential and *Anabaena* sp. in freshwater systems. *Water Research*, **47**, 3444 – 3454, (2013).
- Summers, R.S., Kim, S.M., Shimabuku, K., Chae, S-H. and Corwin, C.J. Granular activated carbon adsorption of MIB in the presence of dissolved organic matter. *Water Research*, **47**, 3507 - 3513, (2013).
- Sun, W.Y., Newham, L.T.H., Croke, B.F.W. and Norton, J.P. Three complementary methods for sensitivity analysis of a water quality model. *Environmental Modelling & Software*, **37**, 19 - 29, (2012).
- Swanepoel, A. and Du Preez, H.H. Application of a Geosmin Incident Management Framework (GIMF) during an *Anabaena circinalis* bloom in source water. Paper published as part of the conference proceedings of the *International Water Association Conference (IWA 2008)*, Vienna Austria, (2008).

-
- Swanepoel, A., Du Preez, H.H. and Barnard, S. Verification of remote sensing (*in situ*) measuring instruments for application in algae monitoring and research. Poster presented at the *National Laboratory Association Conference* (NLA, 2014), Muldersdrift, South Africa, (2014).
- Swanepoel, A., Du Preez, H., Schoeman, C., Janse van Vuuren, S. and Sundram, A. *Condensed Laboratory Methods for the Monitoring of phytoplankton, including Cyanobacteria, in South African Freshwaters*. Water Research Commission, South Africa. WRC Report No. : TT 323/08, 108 pp., (2008a).
- Swanepoel, A., Du Preez, H.H., Dusrath, I. and Rajele, M. *Ceratium hirundinella* reveals algal penetration into the potable water at Rand Water. Paper published as part of the conference proceedings of the *Water Institute of Southern Africa Conference* (WISA 2008), Sun City, South Africa, (2008b).
- Talib, A., Recknagel, F. and Van der Molen, D. Patternising phytoplankton dynamics of two shallow lakes in response to restoration measures by applying non-supervised artificial neural networks. *Environmentalist*, **27**, 195 – 205, (2007).
- TerBraak, C.J.F. 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, (1988).
- UKWIR (United Kingdom Water Industry Research) *Managing microbial and chemical risks from source to tap*. Report Ref. No. 03/DW/02/31. London UKWIR Limited, (2003).
- Van der Walt, M., Krüger, M. and Van der Walt, C. *The South African oxidation and disinfection manual*. WISA Oxidation and Disinfection Division. Water Research Commission, South Africa. WRC Report No. : TT 406/09, (2009).
- Van Ginkel, C.E. and Silberbauer, M.J. Temporal trends in total phosphorous, temperature, oxygen, chlorophyll *a* and phytoplankton populations in Hartbeespoort Dam and Roodeplaat Dam, South Africa. *African Journal of Aquatic Science*, **32** (1), 63 – 70, (2007).
- Van Ginkel, C.E., Cao., H., Recknagel, F. and Du Plessis, S. Forecasting of dinoflagellate blooms in warm-monomictic hypertrophic reservoirs in South Africa by means of rule-based agents. *Water SA*, **33** (4), 531 – 538, (2007).

-
- Van Ginkel, C.E. *Investigating the applicability of Ecological informatics modelling techniques for predicting harmful algal blooms in hypertrophic reservoirs of South Africa*. Thesis submitted for the degree, Doctor of Philosophy at the North West University, Potchefstroom campus, South Africa, (2008).
- Viljoen, F.C. The World Health Organization's water safety plan is much more than just an integrated drinking water quality management plan. *Water Science & Technology*, IWA Publishing, **61** (1), 173 – 179, (2010).
- Welk, A., Recknagel, F., Cao, H., Chan, W-S. and Talib, A. Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. *Ecological Informatics*, **3**, 46 – 54, (2008).
- Wetzel, R.G. *Limnology*, 3rd Edition, San Diego. Academic Press (Elsevier), 1006 pp., (2001).
- WHO (World Health Organisation) *Guidelines for Drinking Water Quality*, 4th Edition. World Health Organisation, Geneva, 541 pp., (2011).
- Wu, S.K., Xie, P., Liang, G.D., Wang, S.B. and Liang, X.M. Relationships between microcystins and environmental parameters in 30 subtropical shallow lakes along the Yangtze River, China. *Freshwater Biol*, **51**, 2309 – 2319, (2006).
- Yang, J. Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. *Environmental Modelling & Software*, **26**, 444 – 457, (2011).
- Zaitlin, B. and Watson, S.B. Actinomycetes in relation to taste and odour in drinking water: Myths, tenets and truths. *Water Research*, **40**, 1741 – 1753, (2006).
- Zoschke, K., Engel, C., Börnick, H. and Worch, E. Adsorption of geosmin and 2-methylisoborneol onto powdered activated carbon at non-equilibrium conditions: Influence of NOM and process modelling. *Water Research*, **45**, 4544 – 4550, (2011).

ADDENDUM A

Table A: Spearman Rank Order Correlations of environmental variables in the Vaal Dam for the period 2000 – 2013.

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta)					
	MD pairwise deleted Marked correlations are significant at p < 0.05000					
	Chl- <i>a</i>	DO	pH	Secchi	Temp	Mic_n
Chl- <i>a</i>	1.000000	-0.163882	0.234818	0.140446	0.483287	0.475114
DO	-0.163882	1.000000	0.052010	0.109565	-0.474567	-0.347556
pH	0.234818	0.052010	1.000000	0.050387	0.290662	0.247057
Secchi	0.140446	0.109565	0.050387	1.000000	0.038974	0.089207
Temp	0.483287	-0.474567	0.290662	0.038974	1.000000	0.630651
Mic_n	0.475114	-0.347556	0.247057	0.089207	0.630651	1.000000
Turb	-0.197899	-0.078699	-0.016392	-0.637389	0.068824	-0.001247
NH ₄ ⁺	-0.022615	-0.138259	-0.215347	-0.080398	-0.116208	-0.048902
PO ₄ ³⁻	-0.074195	0.091511	-0.000605	-0.233991	-0.158725	-0.047476
TP	-0.033224	0.147047	-0.003108	-0.132553	-0.130490	-0.036802
NO ₂ ⁻	0.023536	-0.152620	-0.128976	-0.147057	-0.077985	0.103725
NO ₃ ⁻	-0.311401	0.005524	-0.236924	-0.274613	-0.303323	-0.098828
Cond	0.156234	0.108269	0.289528	0.429292	0.017123	0.011234
Alk	0.137446	0.109627	0.323553	0.470438	0.008356	0.049179
SO ₄ ²⁻	0.157792	0.074030	-0.038051	0.278838	-0.081824	0.128380
Si	-0.237608	0.054836	0.010963	-0.501162	-0.064376	-0.128084
Total Si	-0.220456	0.055347	0.011409	-0.498754	-0.066565	-0.128084
Hardness	0.124585	0.132632	0.150319	0.347988	-0.113185	-0.042985
COD	0.221975	0.158268	0.122189	-0.047983	0.057205	0.106326
Br	0.122716	0.069589	0.064273	0.205896	0.004382	0.079408
Cl	0.268749	-0.024848	0.091609	0.469712	0.046600	0.144173
F	0.138231	-0.076132	0.206977	-0.047533	0.170443	0.147300
Al	-0.215052	-0.009019	-0.043840	-0.466247	-0.006129	-0.078201
Ca	0.084239	0.144350	0.117425	0.363413	-0.137061	-0.061030
B	0.093899	0.042676	0.002409	-0.147974	-0.058026	-0.016114
Cu	0.108204	0.073013	0.073784	0.017537	0.001520	-0.032963
Fe	-0.235161	0.017038	-0.031405	-0.489620	-0.058895	-0.127181
K	-0.017412	0.052104	0.003733	-0.104807	-0.038692	-0.042751
Mg	0.124276	0.134626	0.158459	0.329072	-0.106296	-0.053976
Mn	-0.005730	0.050119	-0.197846	-0.175741	-0.138102	-0.015026
Mo	0.119119	0.090155	0.122016	0.089747	-0.018957	0.088751
Na	0.067246	0.097268	0.148991	0.339433	-0.093298	0.032693
Ni	0.028539	0.154433	0.140193	0.047675	-0.044126	-0.041580
Pb	-0.030406	0.093304	-0.015318	0.005924	-0.079055	0.001206
S	0.190881	0.068360	0.089437	0.324712	0.045720	0.041950
V	0.023381	-0.016026	-0.017390	-0.221942	-0.033899	0.015973
Zn	0.017620	0.118373	0.008400	-0.167831	-0.043348	-0.151276
Cyano	0.523061	-0.191807	0.283004	0.125462	0.517395	0.493249
Tot Alg	0.669108	-0.141931	0.307954	0.154470	0.488707	0.504454
Ana	0.382490	-0.056966	0.225251	0.122877	0.339409	0.200794
Mic_s	0.477987	-0.195645	0.248858	0.077641	0.564921	0.579356

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta) MD pairwise deleted Marked correlations are significant at $p < 0.05000$					
	Turb	NH ₄ ⁺	PO ₄ ³⁻	TP	NO ₂ ⁻	NO ₃ ⁻
Chl- <i>a</i>	-0.197899	-0.022615	-0.074195	-0.033224	0.023536	-0.311401
DO	-0.078699	-0.138259	0.091511	0.147047	-0.152620	0.005524
pH	-0.016392	-0.215347	-0.000605	-0.003108	-0.128976	-0.236924
Secchi	-0.637389	-0.080398	-0.233991	-0.132553	-0.147057	-0.274613
Temp	0.068824	-0.116208	-0.158725	-0.130490	-0.077985	-0.303323
Mic_n	-0.001247	-0.048902	-0.047476	-0.036802	0.103725	-0.098828
Turb	1.000000	0.126613	0.188587	0.000532	0.199804	0.333376
NH ₄ ⁺	0.126613	1.000000	0.188842	-0.040892	0.242043	0.163772
PO ₄ ³⁻	0.188587	0.188842	1.000000	0.148030	0.085333	0.296224
TP	0.000532	-0.040892	0.148030	1.000000	0.064061	0.122150
NO ₂ ⁻	0.199804	0.242043	0.085333	0.064061	1.000000	0.242249
NO ₃ ⁻	0.333376	0.163772	0.296224	0.122150	0.242249	1.000000
Cond	-0.489527	-0.209576	-0.166261	-0.031216	-0.113964	-0.190011
Alk	-0.570913	-0.223988	-0.260110	-0.082500	-0.120176	-0.391352
SO ₄ ²⁻	-0.310555	-0.087927	-0.050435	-0.132875	0.056381	0.090278
Si	0.576709	0.070683	0.193552	0.185951	0.086518	0.145894
Total Si	0.574421	0.086421	0.207648	0.192427	0.070322	0.137980
Hardness	-0.552414	-0.142186	-0.053465	0.092019	-0.052653	-0.136862
COD	0.058481	-0.028677	0.036776	0.124923	-0.101619	0.011922
Br	-0.189377	0.020912	-0.106716	0.083598	0.111996	0.030384
Cl	-0.514854	-0.117720	-0.157745	-0.012088	0.027937	-0.052433
F	0.072894	-0.049608	-0.017010	0.092103	0.169456	-0.036538
Al	0.580412	0.110449	0.308361	0.136109	0.175881	0.292809
Ca	-0.499523	-0.090673	-0.056751	0.019109	-0.060851	-0.138419
B	0.136492	0.170626	0.102718	0.097488	0.152077	-0.032507
Cu	-0.134940	-0.055896	-0.136946	0.028330	0.057408	-0.131903
Fe	0.610235	0.063422	0.248522	0.136567	0.171339	0.248602
K	0.056260	0.047424	-0.005744	0.297463	0.064176	0.109321
Mg	-0.528655	-0.147283	-0.056833	0.105221	0.004050	-0.079685
Mn	0.177368	0.231777	-0.001903	-0.075386	0.218149	0.155504
Mo	-0.144260	0.095225	-0.028213	0.062597	0.098075	0.014625
Na	-0.400962	-0.129355	-0.145185	0.139278	0.015861	0.036055
Ni	-0.101852	-0.111387	-0.122067	0.161797	-0.075392	-0.040304
Pb	-0.050479	-0.011743	-0.078834	-0.050595	-0.112714	0.043283
S	-0.472505	-0.123345	-0.106249	-0.103313	-0.126383	-0.058142
V	0.210814	0.175556	0.000624	0.201255	0.149249	0.006206
Zn	0.223901	0.117041	-0.044102	0.046104	0.111545	0.002117
Cyano	-0.183156	-0.159091	-0.130408	-0.065901	-0.253669	-0.263921
Tot Alg	-0.169531	-0.159818	-0.162189	-0.031975	-0.158640	-0.269748
Ana	-0.204852	-0.193432	-0.155251	-0.075869	-0.293556	-0.244181
Mic_s	-0.067237	-0.034828	-0.081055	-0.017041	-0.168878	-0.276083

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta) MD pairwise deleted Marked correlations are significant at p <0.05000					
	Cond	Alk	SO ₄ ²⁻	Si	Total Si	Hardness
Chl- <i>a</i>	0.156234	0.137446	0.157792	-0.237608	-0.220456	0.124585
DO	0.108269	0.109627	0.074030	0.054836	0.055347	0.132632
pH	0.289528	0.323553	-0.038051	0.010963	0.011409	0.150319
Secchi	0.429292	0.470438	0.278838	-0.501162	-0.498754	0.347988
Temp	0.017123	0.008356	-0.081824	-0.064376	-0.066565	-0.113185
Mic_n	0.011234	0.049179	0.128380	-0.128084	-0.128084	-0.042985
Turb	-0.489527	-0.570913	-0.310555	0.576709	0.574421	-0.552414
NH ₄ ⁺	-0.209576	-0.223988	-0.087927	0.070683	0.086421	-0.142186
PO ₄ ³⁻	-0.166261	-0.260110	-0.050435	0.193552	0.207648	-0.053465
TP	-0.031216	-0.082500	-0.132875	0.185951	0.192427	0.092019
NO ₂ ⁻	-0.113964	-0.120176	0.056381	0.086518	0.070322	-0.052653
NO ₃ ⁻	-0.190011	-0.391352	0.090278	0.145894	0.137980	-0.136862
Cond	1.000000	0.774843	0.429918	-0.381996	-0.375893	0.623690
Alk	0.774843	1.000000	0.278110	-0.363283	-0.364943	0.618471
SO ₄ ²⁻	0.429918	0.278110	1.000000	-0.471913	-0.471822	0.515730
Si	-0.381996	-0.363283	-0.471913	1.000000	0.999901	-0.347998
Total Si	-0.375893	-0.364943	-0.471822	0.999901	1.000000	-0.345955
Hardness	0.623690	0.618471	0.515730	-0.347998	-0.345955	1.000000
COD	-0.035401	-0.139532	-0.142638	0.119231	0.141896	-0.164517
Br	0.019007	0.191988	-0.061303	-0.194659	-0.194940	0.110022
Cl	0.496678	0.511625	0.464596	-0.490704	-0.491149	0.574856
F	0.036816	0.141721	0.078488	-0.002240	-0.002014	0.104638
Al	-0.424959	-0.444816	-0.369903	0.730924	0.725825	-0.426575
Ca	0.584795	0.571996	0.543159	-0.381717	-0.371334	0.927431
B	-0.151646	-0.059636	-0.057417	0.239279	0.265752	0.058272
Cu	0.153557	0.199863	0.175886	-0.031653	-0.044518	0.261047
Fe	-0.542042	-0.494359	-0.456417	0.739089	0.731923	-0.519518
K	-0.041121	-0.110589	-0.195764	0.037285	0.017288	-0.049267
Mg	0.578101	0.591823	0.410231	-0.313380	-0.323985	0.934942
Mn	-0.198064	-0.288401	0.137773	0.196260	0.201953	-0.039198
Mo	0.182601	0.145083	0.244296	-0.049500	-0.068238	0.191455
Na	0.557713	0.478464	0.343726	-0.415254	-0.430498	0.641172
Ni	0.253032	0.228679	-0.098045	0.082900	0.094842	0.177468
Pb	0.109066	0.055515	0.135930	-0.143623	-0.170859	0.012610
S	0.549080	0.448317	0.646257	-0.453062	-0.437489	0.643983
V	-0.209066	-0.154307	-0.212249	0.453739	0.454611	-0.078944
Zn	-0.010036	-0.047312	-0.177963	0.266121	0.272486	0.049287
Cyano	0.242890	0.257848	-0.017938	-0.038952	-0.026289	0.155638
Tot Alg	0.260957	0.275693	0.033924	-0.105783	-0.098576	0.165151
Ana	0.287764	0.285210	-0.018670	-0.042399	-0.031521	0.153126
Mic_s	0.074514	0.131879	-0.132798	0.062828	0.080912	0.075153

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta) MD pairwise deleted Marked correlations are significant at p <0.05000					
	COD	Br	Cl	F	Al	Ca
Chl- <i>a</i>	0.221975	0.122716	0.268749	0.138231	-0.215052	0.084239
DO	0.158268	0.069589	-0.024848	-0.076132	-0.009019	0.144350
pH	0.122189	0.064273	0.091609	0.206977	-0.043840	0.117425
Secchi	-0.047983	0.205896	0.469712	-0.047533	-0.466247	0.363413
Temp	0.057205	0.004382	0.046600	0.170443	-0.006129	-0.137061
Mic_n	0.106326	0.079408	0.144173	0.147300	-0.078201	-0.061030
Turb	0.058481	-0.189377	-0.514854	0.072894	0.580412	-0.499523
NH ₄ ⁺	-0.028677	0.020912	-0.117720	-0.049608	0.110449	-0.090673
PO ₄ ³⁻	0.036776	-0.106716	-0.157745	-0.017010	0.308361	-0.056751
TP	0.124923	0.083598	-0.012088	0.092103	0.136109	0.019109
NO ₂ ⁻	-0.101619	0.111996	0.027937	0.169456	0.175881	-0.060851
NO ₃ ⁻	0.011922	0.030384	-0.052433	-0.036538	0.292809	-0.138419
Cond	-0.035401	0.019007	0.496678	0.036816	-0.424959	0.584795
Alk	-0.139532	0.191988	0.511625	0.141721	-0.444816	0.571996
SO ₄ ²⁻	-0.142638	-0.061303	0.464596	0.078488	-0.369903	0.543159
Si	0.119231	-0.194659	-0.490704	-0.002240	0.730924	-0.381717
Total Si	0.141896	-0.194940	-0.491149	-0.002014	0.725825	-0.371334
Hardness	-0.164517	0.110022	0.574856	0.104638	-0.426575	0.927431
COD	1.000000	0.143890	-0.041730	-0.054827	0.094316	-0.197273
Br	0.143890	1.000000	0.256986	0.323945	-0.175215	0.005501
Cl	-0.041730	0.256986	1.000000	0.130460	-0.409398	0.574371
F	-0.054827	0.323945	0.130460	1.000000	-0.013501	0.065035
Al	0.094316	-0.175215	-0.409398	-0.013501	1.000000	-0.472282
Ca	-0.197273	0.005501	0.574371	0.065035	-0.472282	1.000000
B	0.023725	-0.002928	-0.029012	0.142534	0.120451	0.086816
Cu	-0.076213	-0.130263	0.157349	0.109244	-0.135202	0.282154
Fe	0.126696	-0.170088	-0.478323	-0.049356	0.886210	-0.548617
K	0.120065	0.163371	-0.089379	-0.040203	0.111844	-0.091971
Mg	-0.083368	0.215952	0.535977	0.104297	-0.344927	0.776367
Mn	0.087766	-0.199061	-0.159405	0.037627	0.131803	0.043258
Mo	-0.128044	-0.072987	0.117920	0.025922	-0.085174	0.181046
Na	-0.032408	0.160848	0.632750	0.010889	-0.374670	0.568381
Ni	0.086862	-0.007485	0.171395	-0.020518	0.017504	0.159311
Pb	-0.017792	-0.146013	0.020806	-0.142769	-0.036193	0.050811
S	-0.106420	-0.010818	0.500140	0.117277	-0.502946	0.655512
V	-0.001455	0.004827	-0.156935	0.053194	0.381081	-0.125049
Zn	-0.001713	-0.006852	-0.119465	0.210413	0.111947	0.086334
Cyano	0.173413	0.005299	0.131835	0.076613	-0.160437	0.079770
Tot Alg	0.172662	0.117866	0.193032	0.107404	-0.158358	0.076703
Ana	0.144726	-0.032277	0.125250	0.005636	-0.171395	0.096586
Mic_s	0.157767	0.034618	0.004501	0.085688	-0.054730	0.021618

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta) MD pairwise deleted Marked correlations are significant at p <0.05000					
	B	Cu	Fe	K	Mg	Mn
Chl- <i>a</i>	0.093899	0.108204	-0.235161	-0.017412	0.124276	-0.005730
DO	0.042676	0.073013	0.017038	0.052104	0.134626	0.050119
pH	0.002409	0.073784	-0.031405	0.003733	0.158459	-0.197846
Secchi	-0.147974	0.017537	-0.489620	-0.104807	0.329072	-0.175741
Temp	-0.058026	0.001520	-0.058895	-0.038692	-0.106296	-0.138102
Mic_n	-0.016114	-0.032963	-0.127181	-0.042751	-0.053976	-0.015026
Turb	0.136492	-0.134940	0.610235	0.056260	-0.528655	0.177368
NH ₄ ⁺	0.170626	-0.055896	0.063422	0.047424	-0.147283	0.231777
PO ₄ ³⁻	0.102718	-0.136946	0.248522	-0.005744	-0.056833	-0.001903
TP	0.097488	0.028330	0.136567	0.297463	0.105221	-0.075386
NO ₂ ⁻	0.152077	0.057408	0.171339	0.064176	0.004050	0.218149
NO ₃ ⁻	-0.032507	-0.131903	0.248602	0.109321	-0.079685	0.155504
Cond	-0.151646	0.153557	-0.542042	-0.041121	0.578101	-0.198064
Alk	-0.059636	0.199863	-0.494359	-0.110589	0.591823	-0.288401
SO ₄ ²⁻	-0.057417	0.175886	-0.456417	-0.195764	0.410231	0.137773
Si	0.239279	-0.031653	0.739089	0.037285	-0.313380	0.196260
Total Si	0.265752	-0.044518	0.731923	0.017288	-0.323985	0.201953
Hardness	0.058272	0.261047	-0.519518	-0.049267	0.934942	-0.039198
COD	0.023725	-0.076213	0.126696	0.120065	-0.083368	0.087766
Br	-0.002928	-0.130263	-0.170088	0.163371	0.215952	-0.199061
Cl	-0.029012	0.157349	-0.478323	-0.089379	0.535977	-0.159405
F	0.142534	0.109244	-0.049356	-0.040203	0.104297	0.037627
Al	0.120451	-0.135202	0.886210	0.111844	-0.344927	0.131803
Ca	0.086816	0.282154	-0.548617	-0.091971	0.776367	0.043258
B	1.000000	0.170764	0.138602	-0.099005	-0.032911	0.431409
Cu	0.170764	1.000000	-0.121400	0.057258	0.235844	0.218789
Fe	0.138602	-0.121400	1.000000	0.179973	-0.436297	0.107368
K	-0.099005	0.057258	0.179973	1.000000	0.042463	-0.031170
Mg	-0.032911	0.235844	-0.436297	0.042463	1.000000	-0.107764
Mn	0.431409	0.218789	0.107368	-0.031170	-0.107764	1.000000
Mo	0.097055	0.257860	-0.149491	0.001715	0.193799	0.204700
Na	-0.232602	0.192225	-0.396589	0.224794	0.671180	-0.233782
Ni	0.041257	0.214923	0.006773	0.173921	0.172148	0.008501
Pb	-0.090696	0.122157	-0.094601	-0.035741	0.013016	0.149535
S	0.088335	0.146681	-0.594469	-0.184730	0.537782	0.070722
V	0.299476	0.279456	0.359348	0.103612	-0.032009	0.254792
Zn	0.261362	0.425648	0.115472	0.026690	0.049407	0.322610
Cyano	-0.005757	0.121784	-0.144742	0.007205	0.181611	-0.109277
Tot Alg	-0.016621	0.154636	-0.156799	0.056722	0.214517	-0.158038
Ana	-0.113292	0.113049	-0.161316	0.021412	0.165997	-0.141203
Mic_s	0.105359	0.018340	-0.052781	-0.008661	0.076250	-0.063748

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta)					
	MD pairwise deleted Marked correlations are significant at p <0.05000					
	Mo	Na	Ni	Pb	S	V
Chl- <i>a</i>	0.119119	0.067246	0.028539	-0.030406	0.190881	0.023381
DO	0.090155	0.097268	0.154433	0.093304	0.068360	-0.016026
pH	0.122016	0.148991	0.140193	-0.015318	0.089437	-0.017390
Secchi	0.089747	0.339433	0.047675	0.005924	0.324712	-0.221942
Temp	-0.018957	-0.093298	-0.044126	-0.079055	0.045720	-0.033899
Mic_n	0.088751	0.032693	-0.041580	0.001206	0.041950	0.015973
Turb	-0.144260	-0.400962	-0.101852	-0.050479	-0.472505	0.210814
NH ₄ ⁺	0.095225	-0.129355	-0.111387	-0.011743	-0.123345	0.175556
PO ₄ ³⁻	-0.028213	-0.145185	-0.122067	-0.078834	-0.106249	0.000624
TP	0.062597	0.139278	0.161797	-0.050595	-0.103313	0.201255
NO ₂ ⁻	0.098075	0.015861	-0.075392	-0.112714	-0.126383	0.149249
NO ₃ ⁻	0.014625	0.036055	-0.040304	0.043283	-0.058142	0.006206
Cond	0.182601	0.557713	0.253032	0.109066	0.549080	-0.209066
Alk	0.145083	0.478464	0.228679	0.055515	0.448317	-0.154307
SO ₄ ²⁻	0.244296	0.343726	-0.098045	0.135930	0.646257	-0.212249
Si	-0.049500	-0.415254	0.082900	-0.143623	-0.453062	0.453739
Total Si	-0.068238	-0.430498	0.094842	-0.170859	-0.437489	0.454611
Hardness	0.191455	0.641172	0.177468	0.012610	0.643983	-0.078944
COD	-0.128044	-0.032408	0.086862	-0.017792	-0.106420	-0.001455
Br	-0.072987	0.160848	-0.007485	-0.146013	-0.010818	0.004827
Cl	0.117920	0.632750	0.171395	0.020806	0.500140	-0.156935
F	0.025922	0.010889	-0.020518	-0.142769	0.117277	0.053194
Al	-0.085174	-0.374670	0.017504	-0.036193	-0.502946	0.381081
Ca	0.181046	0.568381	0.159311	0.050811	0.655512	-0.125049
B	0.097055	-0.232602	0.041257	-0.090696	0.088335	0.299476
Cu	0.257860	0.192225	0.214923	0.122157	0.146681	0.279456
Fe	-0.149491	-0.396589	0.006773	-0.094601	-0.594469	0.359348
K	0.001715	0.224794	0.173921	-0.035741	-0.184730	0.103612
Mg	0.193799	0.671180	0.172148	0.013016	0.537782	-0.032009
Mn	0.204700	-0.233782	0.008501	0.149535	0.070722	0.254792
Mo	1.000000	0.119807	0.231937	0.270434	0.212079	0.190738
Na	0.119807	1.000000	0.263962	-0.045994	0.407138	-0.169287
Ni	0.231937	0.263962	1.000000	0.031710	0.092198	0.065449
Pb	0.270434	-0.045994	0.031710	1.000000	0.027193	0.044328
S	0.212079	0.407138	0.092198	0.027193	1.000000	-0.188936
V	0.190738	-0.169287	0.065449	0.044328	-0.188936	1.000000
Zn	0.201820	-0.004275	0.207246	-0.043521	-0.080711	0.295167
Cyano	0.029052	0.089749	0.228433	-0.065500	0.194971	-0.026478
Tot Alg	0.084689	0.153999	0.231694	-0.077824	0.174606	-0.017911
Ana	0.002890	0.102181	0.211215	0.048625	0.123211	-0.076330
Mic_s	0.035333	-0.049713	0.194614	-0.102829	0.153801	0.055819

Variable	Spearman Rank Order Correlations (Spreadsheet all data 2000-13.sta) MD pairwise deleted Marked correlations are significant at p <0.05000				
	Zn	Cyano	Tot Alg	Ana	Mic_s
Chl- <i>a</i>	0.017620	0.523061	0.669108	0.382490	0.477987
DO	0.118373	-0.191807	-0.141931	-0.056966	-0.195645
pH	0.008400	0.283004	0.307954	0.225251	0.248858
Secchi	-0.167831	0.125462	0.154470	0.122877	0.077641
Temp	-0.043348	0.517395	0.488707	0.339409	0.564921
Mic_n	-0.151276	0.493249	0.504454	0.200794	0.579356
Turb	0.223901	-0.183156	-0.169531	-0.204852	-0.067237
NH ₄ ⁺	0.117041	-0.159091	-0.159818	-0.193432	-0.034828
PO ₄ ³⁻	-0.044102	-0.130408	-0.162189	-0.155251	-0.081055
TP	0.046104	-0.065901	-0.031975	-0.075869	-0.017041
NO ₂ ⁻	0.111545	-0.253669	-0.158640	-0.293556	-0.168878
NO ₃ ⁻	0.002117	-0.263921	-0.269748	-0.244181	-0.276083
Cond	-0.010036	0.242890	0.260957	0.287764	0.074514
Alk	-0.047312	0.257848	0.275693	0.285210	0.131879
SO ₄ ²⁻	-0.177963	-0.017938	0.033924	-0.018670	-0.132798
Si	0.266121	-0.038952	-0.105783	-0.042399	0.062828
Total Si	0.272486	-0.026289	-0.098576	-0.031521	0.080912
Hardness	0.049287	0.155638	0.165151	0.153126	0.075153
COD	-0.001713	0.173413	0.172662	0.144726	0.157767
Br	-0.006852	0.005299	0.117866	-0.032277	0.034618
Cl	-0.119465	0.131835	0.193032	0.125250	0.004501
F	0.210413	0.076613	0.107404	0.005636	0.085688
Al	0.111947	-0.160437	-0.158358	-0.171395	-0.054730
Ca	0.086334	0.079770	0.076703	0.096586	0.021618
B	0.261362	-0.005757	-0.016621	-0.113292	0.105359
Cu	0.425648	0.121784	0.154636	0.113049	0.018340
Fe	0.115472	-0.144742	-0.156799	-0.161316	-0.052781
K	0.026690	0.007205	0.056722	0.021412	-0.008661
Mg	0.049407	0.181611	0.214517	0.165997	0.076250
Mn	0.322610	-0.109277	-0.158038	-0.141203	-0.063748
Mo	0.201820	0.029052	0.084689	0.002890	0.035333
Na	-0.004275	0.089749	0.153999	0.102181	-0.049713
Ni	0.207246	0.228433	0.231694	0.211215	0.194614
Pb	-0.043521	-0.065500	-0.077824	0.048625	-0.102829
S	-0.080711	0.194971	0.174606	0.123211	0.153801
V	0.295167	-0.026478	-0.017911	-0.076330	0.055819
Zn	1.000000	-0.016647	0.027848	-0.068452	0.063822
Cyano	-0.016647	1.000000	0.890968	0.789995	0.780863
Tot Alg	0.027848	0.890968	1.000000	0.677940	0.697975
Ana	-0.068452	0.789995	0.677940	1.000000	0.409354
Mic_s	0.063822	0.780863	0.697975	0.409354	1.000000