19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Bayesian network model of Anabaena blooms in Grahamstown Lake

B.J. Williams[a] **and B. Cole**[b]

a. University of Newcastle, Callaghan, NSW 2308, Australia.

b. Hunter Water Corporation, Newcastle, Australia

Email: brian.williams@newcastle.edu.au

Abstract: Grahamstown Lake is an off-river storage supplying water to the city of Newcastle, Australia, with average depth, 7*m* and surface area 28km^2. Its catchment area is 100km^2, generating half its water and the remainder is pumped from the Williams River. Conventional water treatment processes as used by Hunter Water Corporation, including powdered activated carbon dosing, will not completely remove saxitoxins, which may be released from blooms of the cyanobacteria genus *Anabaena* in the lake. Management actions and considerations include water quality rules for pumping from the Williams River, catchment management and sediment treatment.

Since the lake has experienced an increase in *Anabaena* blooms over the last 20 years, a number of investigations have been undertaken. Previous modelling of water quality in Grahamstown Lake has used traditional process-based methods. Because there is very little data relative to the complexity of the system, these models could not be rigorously calibrated to generate accurate predictions and have been ineffective for decision-making purposes.

This paper describes the development of a data-driven, decision-focused Bayesian network model of Grahamstown Lake. This model meets the criteria of being decision-focussed, data driven, transparent, and capable of being used by non-expert modellers.

In the first stage of the development, all available data were arranged in a consistently formatted database from which the model could ‘learn’ probabilistic relationships between model elements such as pumped nutrient load, lake water column nutrient concentrations, *Anabaena* concentrations etc. This stage produced useful insights into ecosystem relationships and provided a basic model for later stages. The first stage model was static and took no account of the system dynamics. The stage 2 model uses the data sequentially and predicts *Anabaena* concentrations for some weeks ahead, following management interventions. The probabilistic nature of the models informs rational consideration of the uncertainty of predictions in this complex system.

The paper describes the Stage 1 model structure and modelling outcomes, Stage 2 dynamic modelling and elicitation of conditional probabilities to strengthen components of the model for which there is little data available at this time.

Keywords: *Bayesian networks, elicitation, data mining, water quality, reservoir management, cyanobacteria*

1. BACKGROUND

Hunter Water Corporation (Hunter Water) has managed Grahamstown Lake since the completion of the dam in 1964. The lake is an off-river storage supplying water to the city of Newcastle, Australia. It has an average depth of 7m and a surface area of 28km^2. Its catchment area is 100km^2, generating half the water it stores and the remainder is pumped from the Seaham Weir in the nearby Williams River.

Since the lake has experienced an increase in cyanobacteria blooms over the last 20 years, a number of investigations have been undertaken. Conventional water treatment processes as used by Hunter Water Corporation, including powdered activated carbon dosing, while effective for all other blue-green algal toxins will not completely remove saxitoxins, which may be released from *Anabaena* blooms in the lake.

Quantities of water pumped to the dam and extracted from it have been the subject of an on-going modelling effort which determines best pumping strategies for minimizing risk of water shortage. The Bayesian network (BN) modelling described in this report is aimed at determining best strategies for minimizing risk in relation to the *quality* of the water stored in Grahamstown Lake.

Previous modelling of water quality in Grahamstown Lake used traditional process-based methods (Williams, 2006). Because there is very little data relative to the complexity of the system, process-based models could not be calibrated to generate accurate predictions. Confidence limits on predictions were never determined so they could not effectively be used in decision-making processes.

Several BN models of Grahamstown Lake water quality are described in this paper. The simplest is an evidence-based 'minimal' model. The causal network specified in this model is its only subjective element. Its conditional probabilities are entirely 'learned' from Hunter Water's routine monitoring data. Elicitation of conditional probability estimates from industry experts was undertaken to bolster weak areas of data. Simple dynamic models have also been investigated.

1.1. Water quality issues

The *Anabaena* genus is of particular concern because of its potential for release of geosmin and the toxins, saxitoxin and possibly microcystin. While microcystin (also produced by the genus *Microcystis*) can be treated comparatively easy with particulated activated carbon (PAC) and chlorine, saxitoxin requires lengthy exposure to PAC(Ho et al., 2009), to an extent which is currently not practicable in Hunter Water's treatment facility. For this reason it has been thought advisable to specifically model this genus with a node in the Bayesian network where *Anabaena* is causally linked to the general algal concentration, stratification, water temperature and turbidity. *Anabaena* has an advantage in stratified conditions (Patterson, Hamilton and Ferris, 1994; Reynolds, 1976), because it has the capacity to adjust its buoyancy. Incoming data from Hunter Water's recently deployed thermistor chains will enable improved estimation of the strength of the probabilistic linkage between stratification and *Anabaena* blooms.

Nuisance tastes and odours can arise from the presence of two organic compounds, geosmin and 2-methyl-isoborneol (MIB) (Izaguirre and Taylor, 2007), which are released inter alia by a number of phytoplankton species, principally members of the cyanobacteria family. In recent years, Geosmin and MIB have been included in routine monitoring by Hunter Water.

1.2. Strategic management actions

- Pump rules: The decision to pump water of a particular quality from Seaham Weir into Grahamstown Lake, must impact at some level on the quality of water in the lake. Rules have been implemented to reduce pumping of water with high nutrient concentrations or algal cell counts.
- Catchment management: Reduction of nutrient inputs from the catchment impacts on potential algal growth in the lake. The model predicts probable impact of such reductions on water quality.
- Sediment treatment: There is no data describing the release or sequestration of nutrients at the sediment interface available at present, so the impact of sediment treatment is not included in the Stage 1 model. Laboratory work being undertaken will provide data which can be incorporated into the Stage 2 model to determine the impacts of various kinds of sediment treatment.
- Data monitoring: The model highlights the most critical data in the network and hence guides decisions for additional data gathering and it suggests trade-offs with less important data.

1.3. Operational management interventions

- Algal event alerts: Within the restrictions of limited data support, the model simulates the system's dynamics and allows probabilistic predictions of the concentration and duration of reportable events. This development will improve Hunter Water's capacity to inform regulators and stakeholders about critical events.
- PAC dosing: The model predicts probabilities of saxitoxin, MIB and geosmin concentrations of bloom events, providing better guidance for resource needs for the associated specialized treatment requirements.

1.4. Rationale for choice of model

The following criteria were specified for choice of the model.

1. The model should inform decision-making as directly as possible.
2. The model should use monitoring data as fully as possible.
3. Model predictions should acknowledge and describe uncertainty.
4. The model should be credible, accessible to, and 'owned' by Hunter Water's water quality scientists and engineers (Aber, 1997).
5. The model should be transparent to all stakeholders.

Bayesian network models meet these criteria. Process-based models have a loose connection to decision-making, only use monitored data selectively, do not acknowledge uncertainty, are generally too complex for non-expert modelers and are not inherently transparent.

1.5. Bayesian network applications in water environments in Australia

To the authors' knowledge, this is the first application of structured data-mining using a water authority's routine sampled data to assess probabilities for decision-making purposes. Bayesian networks are increasingly being used as a decision-making tool in a range of environmental/water management applications in Australia.

For example, a BN based decision support tool called CLAM (Ticehurst, Letcher and Rissik, 2008; Ticehurst et al., 2005) has been developed and used in nearly 30 coastal lakes and estuaries on the NSW coast. The Netica Software (Norsys Software Corp., 2009) used in the Grahamstown project has been widely used in studies of wetlands, environmental flows, marine algal blooms etc. (Chee, Burgman and Carey, 2005; Johnson, Fielding, Hamilton and Mengersen, 2010; Stewart-Koster et al., 2010). Pollino and Henderson (2010) have recently written an Australian government guide on the use of BN modelling in natural resource management.

1.6. Data quality

A complex ecosystem, such as that supported in Grahamstown Lake, requires far more data than are available to make predictions as accurately as we might desire. This is universally true of ecosystem models of natural systems. Process based models are rarely rigorously calibrated (Arhonditsis and Brett, 2004; Williams, 2006), because data are inadequate.

Hunter Water's data set, while good by most standards, remains limited in terms of its predictive capability because of the system's complexity. Samples for analysis are collected at three locations in the lake: the north end, near the inlet from the Balickera Canal, centre and south end, near the off-take outlet. The stations are roughly 3 kilometres apart. Sampling has mostly been undertaken weekly since 1985, though not all parameters were sampled from that time nor are all parameters always sampled at weekly frequency. As we might expect in the 26 year period only a limited number of 'extreme' events have occurred, so for them there is weak data support. The minimal BN model 'learns' conditional probabilities of blooms entirely from this set of monitored data, with the exception of the catchment loads which result from a simple catchment model of the nutrient (nitrogen and phosphorus) loads. A program was written to generate a set of consistent 'cases' which could be used for this purpose from the Hunter Water data set.

An implicit assumption in using the weekly samples in this way is that samples are independent. In fact there will be correlation of water quality within events. The probability of events of some prescribed concentration will be less than the probability of the samples taken at that concentration because events will in general contain more than one sample exceeding the prescribed concentration.

1.7. Bayesian network models

The Bayesian paradigm is founded on the notion that probabilities of events may be modified by knowledge of some additional information. Such modified probabilities are called conditional probabilities because they are 'conditioned' on this additional information.

A Bayesian network is a directed causal network (Pearl, 2000) in which probabilities are assigned for all internal (*'child'*) nodes, conditional on the states of their *'parent'* nodes (linked by arrows into the child node). An important assumption in the construction of Bayesian networks is that there are no cycles within the network. Formally, such networks are called *directed acyclic graphs* (DAG's). For computational reasons, the probability distributions for the nodes must be described by a finite set of mutually exclusive 'states', such as 'very low', 'low', 'moderate', 'high', 'very high'. In this model concentration ranges have been defined for each state. The data and computational requirements increase combinatorially with the number of states for each node and its parents. For *Anabaena*, the Alert level concentrations required by the EPA (NSW) have been used. All child nodes have associated with them a *conditional probability table* (CPT) each entry of which provides the probability of the child node being in one of its discrete states, given that the parent nodes are in a particular combination of their possible states. The dimensions of the CPT are thus the product of the possible states of the parent nodes and the child node.

The CPT's in the Bayesian network in practice are determined ("learned") from the data, by various methods, the simplest of which is to count the number of instances of each of the child states occurring for combinations of parent states. Netica offers two other methods of CPT determination, namely conjugate gradient descent and the expectation maximization (EM) algorithm. Both these methods are more time consuming. They work with an iterative process in which a candidate net is determined, its log likelihood estimated and then incremental changes are made using the case data to find a better net. These methods are described at length in Neapolitan (2004).

2. GRAHAMSTOWN LAKE MODELS

Figure 1 shows the 'minimal' Bayesian network for the Grahamstown Lake system. This network was selected by trial from amongst a group of proposed models, in consultation with Hunter Water's water quality scientists and engineers. In particular it allows testing of pumping strategies and the impact of catchment management.

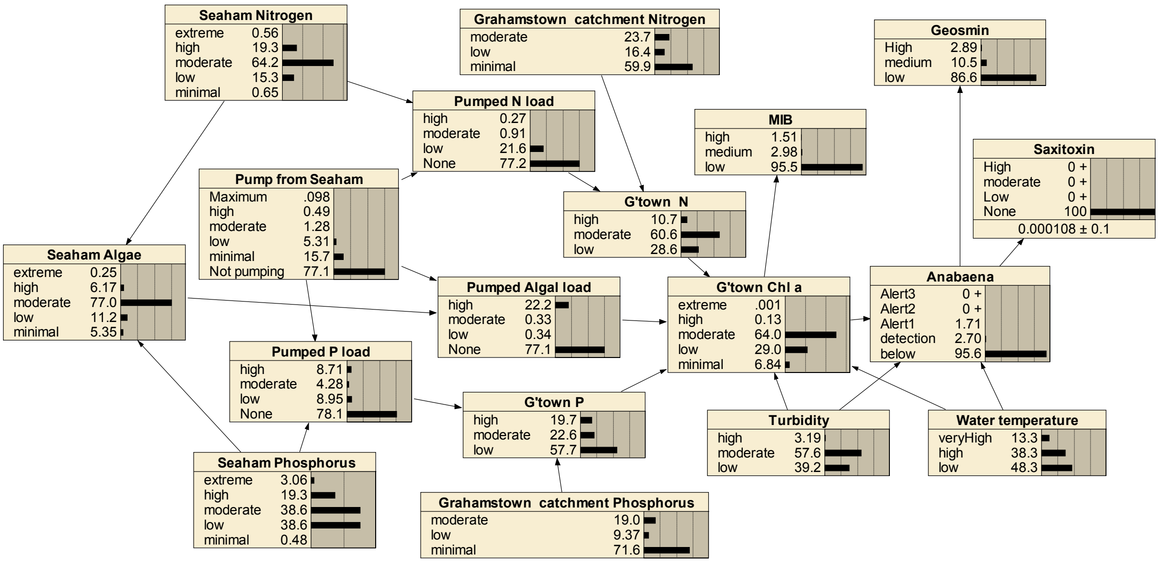

Figure 1. Belief bars of minimal Bayesian network model of Grahamstown Lake system

The nodes include water quality in the Seaham Weir (algae, nitrogen and phosphorus), the level of pumping, the associated pumped load of algae and nutrients, catchment loads of nitrogen and phosphorus and the resulting environmental conditions in the reservoir itself – nitrogen (G'town N), phosphorus (G'town P) and algae (G'town Chl a). MIB is assumed to be dependent on the chlorophyll 'a' concentration and *Anabaena* is conditioned on the chlorophyll ‘a’, turbidity, water temperature and stratification. Saxitoxin is causally linked to the *Anabaena* concentration. The linkages express causality, but information can also be inferred from ‘downstream’ nodes. The boxes indicate the probabilities associated with each node. Notice that the Alert 3 and Alert 2 conditions for *Anabaena* have associated probabilities of 0+ - the software’s indication of a very small number. Hunter Water’s data set, in fact, has no examples of concentrations exceeding Alert 1.

Fig 2 shows a slightly expanded version of the minimal model in which there is an additional node (Stratification) which is an additional ‘parent’ node, conditioning *Anabaena*. There are several other less obvious differences resulting from elicited information from three external experts. Elicitation of expert values (O'Hagan et al., 2006) is of primary importance for Bayesian networks with limited data. It offers a means of transparently incorporating additional information which can be critically peer-reviewed prior to the model’s use in decision-making.

Notice in Fig 2 that there are now small probabilities associated with the Alert3 and Alert2 categories of *Anabaena*. These are derived from the *Anabaena* CPT which now contains stratification as a condition and for which, in extreme situations, values were elicited from the experts. Hunter Water has only about 18 months of stratification data, so elicitation was necessary for these CPT’s.

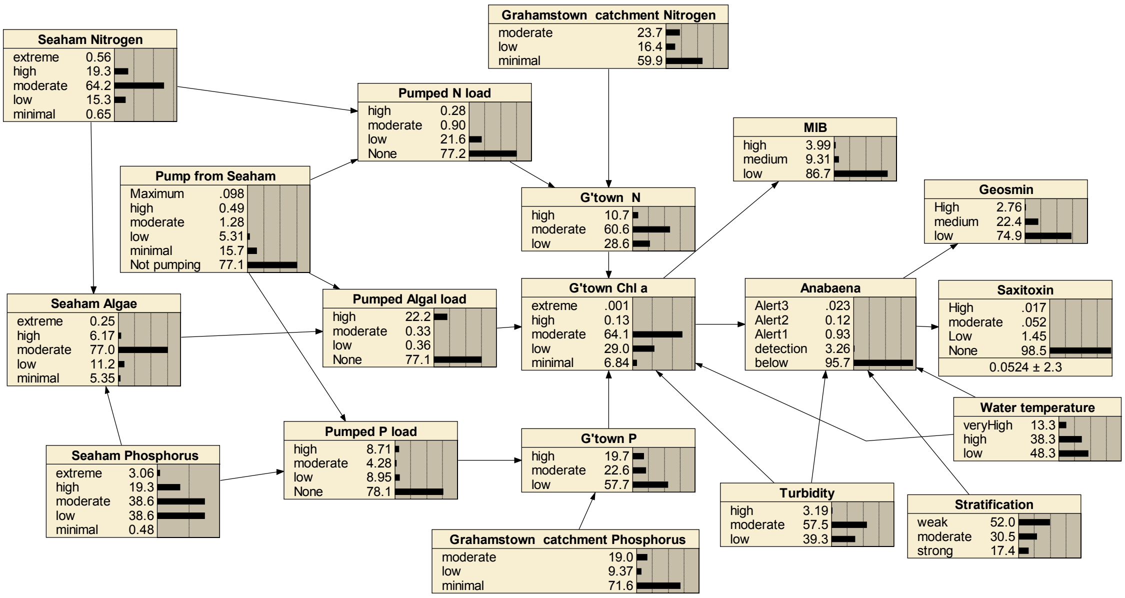

Fig 2. Expanded model with elicited CPT values for *Anabaena* Alert 2 and 3 and MIB, geosmin and saxitoxin

Secondly, notice that there are now values for high and moderate categories of saxitoxin. This follows from the increased *Anabaena* probability, but has also been affected by elicited conditional probabilities. Similarly, there are modified values for geosmin and MIB. Elicitation of extreme components of the CPT for Anabaena and the CPT’s for Geosmin, MIB and Saxitoxin was undertaken using a questionnaire sent to three well known Australian blue-green algae authorities. Conditional probabilities estimated by these experts were arithmetically averaged and then added directly to the CPT’s in Netica.

With these modifications to the CPT’s, estimates of impacts of critical conditions for *Anabaena* can now be made as shown in Figure 3. Notice that the nodes conditioning *Anabaena* have all been set to critical levels. The software then estimates the probabilities of an *Anabaena* bloom as shown. Recall that Hunter Water has never had an Alert 2 bloom. The explanation for the high probabilities (50% chance of Alert 2 or higher) is that the conditions specified are very rare. The extreme chlorophyll ‘a’ event estimated using the EM algorithm has a likelihood of around 1 in 2000 years.

The model in this form has been used to consider the impact of high pumping events when there is poor quality water in Seaham Weir, catchment management to reduce catchment nutrient loads and seasonal risks.

Attempts to construct conventional time series models have been unsuccessful because of the ephemeral nature of blooms. Modelling was also attempted with a variety of Dynamic Bayesian Networks (DBN). A lag-one model is shown in Figure 4, cropped to show the detail of the *Anabaena* and geosmin trajectories. An Alert 3 category is set and the following sequence of nodes (Anabaena1, Anabaena2 etc) shows its decay week by week. The uniform probabilities in Anabaena3 (occurring 3 weeks after the bloom initiates) are thought to arise from the sparsity of real data in the CPT and the approximate nature of the EM algorithm. No attempt has been made at this time to interpolate or smooth the expert estimates inserted in the CPT's.

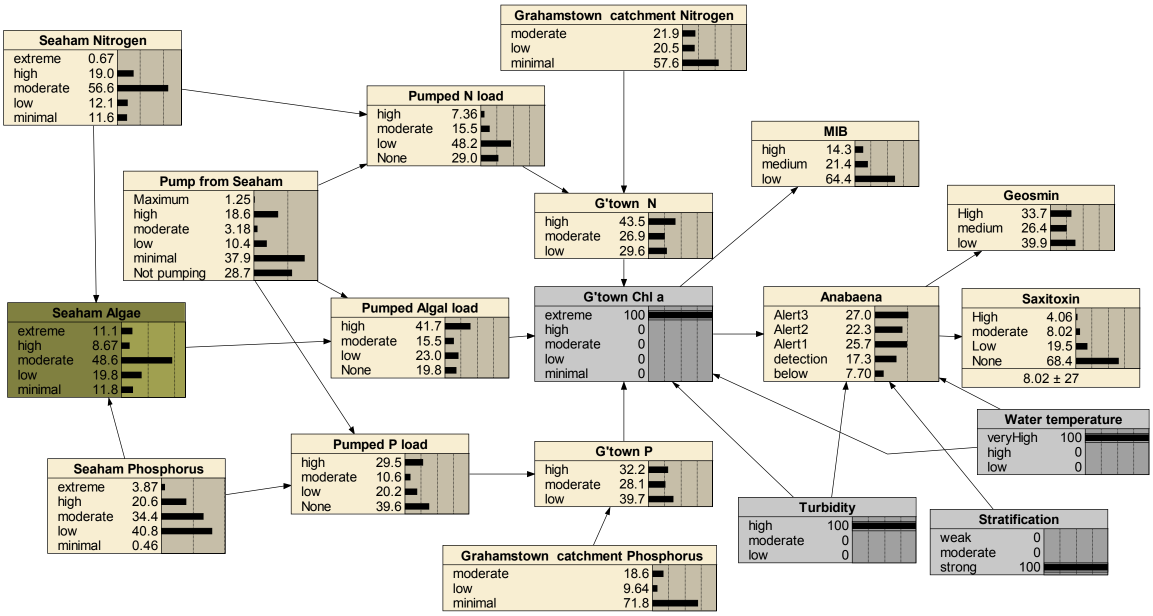

Figure 3. Prediction of *Anabaena* concentration probability distribution under critical conditions

Interpolation (Das, 2004) in the CPT's to reduce such 'noisy' behavior will be undertaken in future development of the model.

Comprehensive scenario assessment has not yet been undertaken, but already improved predictive insights and confidence arising from the use of the Bayesian network models have allowed Hunter Water to cancel development of mooted treatment facilities with capital costs of around $12 million and annual operational costs of around $1 million. Preliminary runs of pumping strategies, suggests that there are only small risks associated with relaxation of pumping strategies.

Perhaps the most outstanding shortcoming of Bayesian network models is that while they estimate probabilities of states, learnt from data, current technology does not allow estimates of the uncertainty of the estimates.

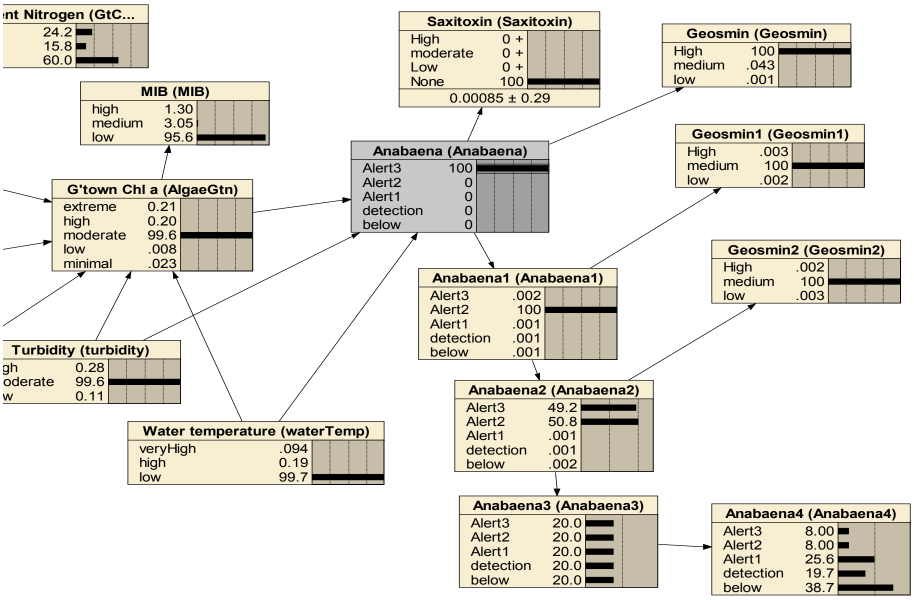

Figure 4. Example of weekly interval, dynamic network (cropped to show dynamic nodes)

3. CONCLUSIONS

Bayesian network models are an effective way of using routine monitoring data to improve decision-making in complex water quality management systems in the water supply industry. Elicited data provides a transparent, defensible means of extending data. The nature of these models is probabilistic and they account for the uncertainty in these data-poor complex systems. The relative simplicity of Bayesian network models allows both ecologists and managers to assess and test the causal structure of the model, investigate scenarios and interpret outcomes in relation to decision-making.

REFERENCES

Aber, J. D., (1997): Why don't we believe the models. *Bulletin of the Ecological Society of America*, **78,** 232-233.

Arhonditsis, G. B., and M. T. Brett, (2004): Evaluation of the current state of mechanistic aquatic biogeochemical modeling. *Marine Ecology Progress Series*, **271,** 13–26.

Chee, Y. E., M. Burgman, and J. Carey, (2005): Use of a Bayesian network decision tool to manage environmental flows in the Wimmera River, Victoria. LWA/MDBC Project UMO43 Delivering sustainability through risk management. Report No 4, 76 pp.

Das, B., (2004): Generating conditional probabilities for Bayesian networks: easing the knowledge acquisition problem. [Available online at http://arxiv.org/abs/cs.AI/0411034.]

Ho, L., P. Tanis-Plant, N. Kayal, N. Slyman, and G. Newcombe, (2009): Optimising water treatment practices for the removal of *Anabaena circinalis* and its associated metabolites, geosmin and saxitoxins. *Journal of Water and Health*, **7,** 544-556.

Izaguirre, G., and W. D. Taylor, (2007): A guide to geosmin- and MIB-producing cyanobacteria in the United States. *Water Science & Technology*, **55,** 9-14.

Johnson, S., F. Fielding, G. Hamilton, and K. Mengersen, (2010): An integrated Bayesian network approach to *Lyngbya majuscula* bloom initiation. *Marine Environmental Research*, **69,** 27-37.

Neapolitan, R. E., (2004): *Learning Bayesian networks.* Pearson Prentice Hall.

Norsys Software Corp., (2009). http://www.norsys.com/netica.html.

O'Hagan, A., C. E. Buck, A. Daaneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, (2006): *Uncertain judgements : eliciting experts' probabilities.* Wiley.

Patterson, J. C., D. P. Hamilton, and J. M. Ferris, (1994): Modelling of cyanobacterial blooms in the mixed layer of lakes and reservoirs. *Australian Journal of Marine and Freshwater Research*, **45,** 829-845.

Pearl, J., (2000): *Causality : models, reasoning, and inference.* Cambridge University Press.

Pollino, C. A., and C. Henderson, (2010): Bayesian networks: a guide for their application in natural resource management and policy. Technical Report 14, 48 pp.

Reynolds, C. S., (1976): Succession and vertical distribution of phytoplankton in response to thermal stratification in a lowland mere, with special reference to nutrient availability. *The Journal of Ecology*, **64,** 529-551.

Stewart-Koster, B., S. E. Bunn, S. J. Mackay, N. L. Poff, R. J. Naiman, and P. S. Lake, (2010): The use of Bayesian networks to guide investments in flow and catchment restoration for impaired river ecosystems. *Freshwater Biology*, **55,** 243-260.

Ticehurst, J. L., R. A. Letcher, and D. Rissik, (2008): Integration modelling and decision support: A case study of the Coastal Lake Assessment and Management (CLAM) Tool. *Mathematics and Computers in Simulation*, **78,** 435-449.

Ticehurst, J. L., D. Rissik, R. A. Letcher, L. T. H. Newham, and A. J. Jakeman, (2005): Development of decision support tools to assess the sustainability of coastal lakes. *Proceedings for MODSIM Conference*, Melbourne, http://www.mssanz.org.au/modsim05/.

Williams, B. J., (2006): *Hydrobiological modelling.* University of Newcastle. www.lulu.com, 700 pp.