20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Data Driven Statistical Model for Manganese Concentration Prediction in Drinking Water Reservoirs

E. Bertone[a], R.A. Stewart[a], H. Zhang[a] and K. O'Halloran[b]

[a] *Griffith School of Engineering, Griffith University Gold Coast campus, Queensland*
[b] *Scientific Services and Data Systems, Seqwater, Brisbane, Queensland*
Email: edoardo.bertone@griffithuni.edu.au

Abstract: Continuously monitoring and managing manganese (Mn) concentrations in drinking water supply reservoirs are paramount for water suppliers, as high concentrations create discoloration of potable water supplied to the customers. Traditional Mn management approaches typically involve manual sampling and laboratory testing of raw water from supply reservoirs on a regular basis (typically weekly) and then treatment decisions are made based on the soluble Mn level exceeding an allowable threshold level; for the reservoir in this study the threshold level for treatment is 0.02 mg/L. Often Mn testing is conducted all year, but in the sub-tropical regions, such as the Gold Coast, Australia, where the reservoir of interest for this study (Hinze Dam) is located, high Mn concentrations only occur for a brief period during the dam destratification process which occurs at the beginning of winter. High concentrations of Mn, resulting from the destratification event, in water entering the water treatment plant are usually treated through pre filter chlorination for concentrations < 0.18 mg/L, or with addition of potassium permanganate for higher concentrations.

Recently, a vertical profiling system (VPS) has enabled the data collection of many water parameters, such as water temperature, dissolved oxygen, pH, conductivity and redox potential every 3 hours. Despite the abundance of physical and water quality data collected by the VPS, it cannot directly measure a range of water quality parameters such as Mn, thus manual sampling and testing are still required.

Since previous studies have shown significant links between the physicochemical parameters collected by VPS and Mn concentrations, a data driven model can be developed to predict Mn values accurately. A Multiple Linear Regression (MLR) with empirical equations for Hinze dam was trained using data from 2008 to 2011, and tested with an independent dataset from 2012. The model was able to predict one week ahead the average Mn concentration in the epilimnion, where the water is drawn, with a correlation coefficient higher than 0.83. The output is also displayed in form of probabilities of exceeding certain thresholds, for instance 0.02 mg/L (namely Mn treatment needed).

Successfully achieving the development of an autonomous and accurate tool for the data mining of VPS parameter datasets to predict levels of Mn provides several benefits for treatment operators: such a decision support system (DSS) would significantly reduce laboratory costs while concurrently enhancing treatment adaption response times.

Keywords: *Manganese, Decision Support System, Vertical Profiling System, Reservoir Destratification, Water Treatment*

1. INTRODUCTION

High manganese (Mn) concentrations are a widespread water quality issue impacting water utilities. In subtropical reservoirs such as Hinze dam, location of this study, soluble Mn is usually present in small amounts in the surface waters, where the water is typically drawn for human uses; however in winter, because of the lake turnover, its level sharply increases and the water treatment plant operators face the challenge of determining the appropriate quantity of oxidizer to be used. Utilities unable to complete the timely and appropriate level of treatment requirements for excess Mn concentrations could be passing this water through to consumers, potentially leading to water colour and odour issues.

Therefore, they are requested to set up efficient, reliable, practical and safe methods to predict and, when necessary, reduce high levels of Mn for potable purposes. Traditionally, operators rely on manual sampling of the water; nevertheless, in recent years the use of vertical profiling systems (VPS) to detect real-time values of a range of water quality and environmental parameters such as pH, redox potential or dissolved oxygen (DO) has enhanced the opportunity for a constant monitoring of variables of interest for the water supply operator. Unfortunately, Mn detection still currently relies on manual weekly sampling, because the VPS are unable to measure this chemical parameter.

Since many prior studies have proven the existence of correlations between some of the aforementioned parameters and other laboratory tested ones (such as Mn), there is an opportunity for an autonomous and intelligent tool to be developed that is able to predict future values of Mn with a high degree of accuracy, thereby reducing laboratory testing requirements and associated costs as well as improving operator decision making.

This paper firstly provides a background for the key issues related to the research study, followed by a description of the research methods applied for designing and building the Mn prediction model. Next, the model prediction accuracy (7 days ahead) is tested for the 2012 critical winter lake turnover period and its performance discussed. Finally, a discussion and study conclusions are provided.

2. BACKGROUND

2.1. The manganese cycle in a lake

The total Mn load into a lake or reservoir is not constant over time, but it depends on many internal and external factors; besides, it is typically very different between the surface waters (i.e. the epilimnion) and the deep bottom waters (i.e. the hypolimnion). The epilimnion is usually warmer and richer in oxygen than the hypolimnion, because of the solar radiation that, particularly during summer, determines water heating and photosynthesis. The presence of algae and photosynthesis usually means a higher pH too (because of the removal of CO_2 forms such as HCO_3^-, which is acidic). Since with high pH and redox potential values the soluble divalent Mn is unstable (Hem, 1963), this is oxidised into more complex insoluble compounds that precipitate into the hypolimnion. Hence, the Mn concentration in the epilimnion is usually low: this is one of the reasons why the water to be directed to the treatment plant is typically drawn from this layer, since the concentration of nutrients is usually low thereby reducing the degree of water treatment processing required. In the hypolimnion, the radiation cannot penetrate and the photosynthesis does not occur: the much lower levels of DO (Tundisi and Matsumura, 2011) in these waters is used by respiration of bacteria, which also contribute to lowering the pH with acidic reactions such as denitrification; this reducing environment (Calmano et al., 1993) along with the anoxic conditions (Chiswell and Huang, 2003) makes the soluble divalent Mn the most stable Mn form and a reduction of the insoluble Mn, coming mainly from the bottom sediments (Dojlido and Best, 1993) occurs. The Mn can diffuse around the hypolimnion, but it usually cannot reach the epilimnion, since its warmer, less dense waters are hardly mixed with the cooler hypolimnetic ones. Nevertheless, sometimes during the year the stratification is broken and lake circulation occurs. For warm monomictic lakes such as Hinze dam, this happens once per year during winter, when the solar radiation is weaker and the epilimnion becomes cooler: gradually the thermal gradient gets smaller and smaller until winds are able to apply the shear needed for mixing the whole water column (Tundisi and Matsumura, 2011). Lake circulation leads to an even distribution of chemical and biological constituents throughout the water column, with the top layers enriched in nutrients (e.g. Mn) from the hypolimnion (Nürnberg, 1988). This Mn is oxidized and the insoluble Mn precipitates downwards or is washed away with the outflow, but these processes are slow, therefore for some days/weeks high epilimnetic soluble Mn concentrations persist.

Interestingly, few studies have been conducted to try to fully model or even predict the Mn cycle. As pointed by Maier et al. (2010), with regards to recently widely applied statistical models such as Artificial Neural Networks (ANN), the vast majority of the environmental models deals with water quantity more than water

quality issues. Besides, the typical water quality parameters that have been modelled are pH or salinity (e.g. Zhang and Stanley, 1997; Bastarache et al., 1997) with few or no studies related to nutrients. An interesting model was created by Bowden (2003), which adopted an ANN to predict the peak concentrations of cyanobacteria in River Murray, Australia. Process-based models have been widely applied in the environmental sector whenever enough data were made available. Nevertheless, to the author's knowledge, attempts of modelling the Mn cycle, with particular regards to the quick transport processes towards the epilimnion during the lake destratification, were not present. One of the few studies found in the literature was done by Johnson et al. (1991), who created a mathematical, time-dependent model for simulating the Mn cycle in a Swiss lake. The model made use of differential equations including all the main processes affecting the formation and transport of soluble and particulate Mn, such as eddy diffusion, outflow, flux from the sediment, oxidation in the water column and coagulation with subsequent sedimentation. However, the equations in this model did not include the mixing in the epilimnion, whose depth changes over the year, thus implicitly excluding the winter lake circulation mixing processes.

2.2. Hinze dam

Hinze Dam, also known as Advancetown Lake (153.28° E, 28.05° S), supplies most of the water provided to Gold Coast City (Queensland, Australia). The dam was originally constructed in 1976 (42,400 ML water storage capacity) and was raised in 1989 (161,070 ML). In 2011, the AUD$395 million Stage Three project raised the dam wall a further 15 m (from 93.5 to 108 m), doubling the dam's capacity and providing increased water security and flood mitigation.

Hinze Dam is located 15 km southwest of Nerang, immediately downstream of the confluence of the Nerang River and Little Nerang Creek (Figure 1). The 600 m long dam can currently hold 310,730 ML of water across a surface area of 9.72 km^2, while the catchment area covers 207 km^2. The raw water is drawn and directed to the closest water treatment plant, located in Molendinar (about 10 km north-east).Since Hinze Dam is the major water source for the Gold Coast area, which is home to over 500,000 people, providing adequately treated water from this water source is extremely important.

Figure 1. Hinze dam location map

Currently, one vertical profiler is collecting data in Hinze dam next to the intake tower. Vertical profilers are automatic recording systems that provide a fast, direct and reliable means for analysing the response of a reservoir to slow or sudden weather variations (Rouen et al., 2005). A VPS usually consists of one or more monitoring stations, used to constantly monitor the weather and the vertical variations of the chemical–physical parameters of the reservoir; one remote station that receives, analyses and stores the information collected by the monitoring stations; and communications tools, used for data transfer (e.g. Wireless). However, important parameters for water treatment cannot be provided 'live' by the VPS such as Mn, which still requires costly and time-consuming manual weekly water samplings and laboratory analysis.

3. RESEARCH METHODS

3.1. Data collection

This collaborative project between Griffith University and Seqwater enabled over 12 years of water quality and environmental data to be made available. Data utilised to populate the Mn prediction model includes the following:

- Three hourly VPS data for Hinze dam (parameters described in section 2,3) at a location near the water intake for the period of 2008-13.
- Manually sampled and laboratory tested data from the surface to a depth of 24 m for the period of 2000-2013 including those parameters also provided by the VPS (e.g. water temperature, dissolved oxygen, etc.) but also many others such as Mn or Fe.

- Daily river inflow data collected from the Queensland Government Department of Energy and Resources Management.
- Daily weather data for the period 2000-2013 provided by the Australian Bureau of Meteorology (BoM).

The main problem with dealing with such a large and heterogeneous dataset is the variance in the collection frequencies of the different parameters. Given this issue, a representative value has been determined for each day through a range of techniques (e.g. averaging, interpolation, etc.). It should be noted that daily variations in the independent Mn variable have still been maintained in the model. A study limitation relates to not having Mn data below 24 m depth (approx. half of Hinze dam full depth level). Historical full depth Mn vertical profile records would have helped to fully explain the complex Mn transport mechanisms occurring during the period surrounding lake circulation. In order to partially overcome the aforementioned limitations, extra Mn samples were collected during the 2013 lake turnover period. As a consequence, whole water column data was collected every 2nd day in the weeks surrounding the critical lake turnover period where Mn transportation across the water column is highly dynamic.

3.2. Data analysis and model choice

The first step of analysis for the aforementioned data was through visual interpretation through numerous time series plots between different variables. This inductive research process was beneficial as it not only enabled existing relationships reported in the literature to be confirmed or otherwise, but also revealed other unexpected potential relationships. The second step of analysis involved rigorous statistical analysis in order to derive the appropriate model input variables.

Before selecting the most appropriate analysis technique(s) for the model, an extensive review of the pertinent literature was performed, focusing on three model categories: statistical models (such as multiple linear regression, neural networks, regression tree), probabilistic models (such as Bayesian networks) and physical models (using software such as DYRESM or MIKE). The advantages and disadvantages of each modelling technique were then assessed and step-by-step testing procedures were followed in order to select those technique(s) which were deemed to be the most suitable for this problem.

3.3. Model validation

In order to test the performance of a model, the historical data set was divided into a training set, used for setting up the model, and an independent testing set. This latter data set was kept separate and only utilised after the training process was completed. The testing set will determine the accuracy of the model developed. A key requirement of the model was that it had some adaptability in order to take into account the influence of future potential changes in environmental conditions that can dramatically affect the equilibrium of the lake. Other authors have stated that ”stationarity is dead” (Milly et al., 2008), meaning that future environmental models need to be able to learn and adapt to changed circumstances.

4. RESULTS

4.1. Identifying model input variables

As confirmed in the literature, it was noticed how in Hinze dam, critical levels of Mn in the epilimnion are reached only during winter turnovers. High inflow events, related to heavy rains, increased the concentrations sometimes, but without reaching critical threshold levels. Table 1 provides a summary of the relevance of the main input variables analysed: after plotting the Mn against each possible predictor, the most appropriate data transformation (e.g. hyperbolic, exponential) was applied and the correlation coefficient (R) at the most relevant lag computed. Also moving averages were considered since they can increase the correlation. Apart for the water temperature differential, no variable was found to have a relevant impact in Mn prediction. The other nutrient cycles had some common features to the Mn one, but data were available for only two years on a monthly interval. DO and pH followed the cycles described in the literature (i.e. alkaline, oxygenated epilimnion and acidic, anoxic hypolimnion during stratification) but they were not useful in Mn prediction. Turbidity showed peaks after heavy rains, while chlorophyll-a values were higher in the epilimnion; however they did not seem to affect the Mn cycle.

As a consequence, it was noticed how the prediction of critical Mn concentrations is strictly connected to the turnover prediction, and a good match between the beginning of the Mn peaks and the attainment of the same water temperature throughout the water column was found (Figure 2). Hence, a transformation of the water column data was calculated though (1), yielding a correlation coefficient with the real values of 0.82.

Table 1. Predictors relevance for soluble epilimnetic Mn, literature/reality/statistical analysis

Main predictors	Relevance in prediction				
	Literature	Hinze Dam dataset	Statistical Analysis		
			Best R	Lag	MA
Water column T differential	H	H	0.82	0	1
$Mn_{sol,\ hyp}$	H	L	-0.21	58	14
DO	M	L	-0.29	17	7
Fe	M	M	x	x	x
NOx	M	M	x	x	x
pH	L	L	-0.42	17	7
Rain	M	L	0.11	0	1
Inflow	M	L	0.09	0	1

H = high; M= medium; L = low. Lag expressed in days. MA= moving average.

$$Mn_{sol,ep}(t) = \left[\frac{\left(\left(\frac{1}{1+\Delta T_w(t)}\right)-min\left(\frac{1}{1+\Delta T_w}\right)\right)}{\left(max\left(\frac{1}{1+\Delta T_w}\right)-min\left(\frac{1}{1+\Delta T_w}\right)\right)}\right]^3 \cdot \left(max\left(Mn_{sol,ep}\right) - min\left(Mn_{sol,ep}\right)\right) + min\left(Mn_{sol,ep}\right) \quad (1)$$

Where:

$Mn_{sol,ep}(t) =$ value of soluble Mn in the epilimnion at time t [mg/L];

$max\left(Mn_{sol,ep}\right)$ = maximum value of soluble Mn in the epilimnion within the historical set [mg/L];;

$min\left(Mn_{sol,ep}\right)$ = minimum value of soluble Mn in the epilimnion within the historical set [mg/L];;

$$\Delta T_w = \left(\frac{\sum_{z=0}^{2} T_w(3\cdot z)}{3}\right) - \left(\frac{\sum_{z=4}^{8} T_w(3\cdot z)}{5}\right) \quad (2)$$

Where:

$T_w(3 \cdot z) =$ water temperature at depth 3z [°C].

4.2. Model development

Accordingly, the choice of a model was related to the water temperature prediction. Different options have been considered and explored. Many physical models (e.g. DYRESM) are able to model the lake water temperature with good accuracy, but they require an exhaustive number of lake property variables as inputs. More importantly, since the model is required to make predictions one week ahead, many of those inputs should be forecasted in-turn, thus summing up the error of each forecast leading to an unacceptable overall error. As an alternative, since high correlation coefficients were found between the average air temperature of the previous week and the water column temperature (Figure 2), simple statistical models such as Multiple Linear Regression (MLR) or Artificial Neural Networks (ANN) for predicting water temperature using only air temperature as an input were determined to have potential for a better performance.

Hence, it was decided to develop three different models, in order to predict the water column temperature one week ahead: two statistical models (MLR and ANN), with the average air temperature of the past 7 days ($\overline{T_{air}}(t)$) and the current water column temperature as inputs, and a physical model (namely DYRESM) requiring multiple physical inputs. Once the water column temperature is predicted, it is given as an input to the data-driven equation, thus yielding the epilimnetic soluble Mn prediction.

4.3. Model testing

Model testing utilised an independent dataset for years 2011-12. The most accurate water column prediction model was proven to be MLR, since it shows superior performances of ANN, which yields higher volatility, detrimental for the application of the data-driven Mn equation. The physical model (i.e. DYRESM) was also tested. Forecast seven days ahead for all the input variables were collected for one month from the Australian Bureau of Meteorology (BoM). Within the historical data (i.e. assuming input forecast = input real values), the model performance was similar to the MLR. Nevertheless, as expected, when presenting the forecasted input the model performance was poorer than MLR, because of the errors in the input predictions by the BoM.

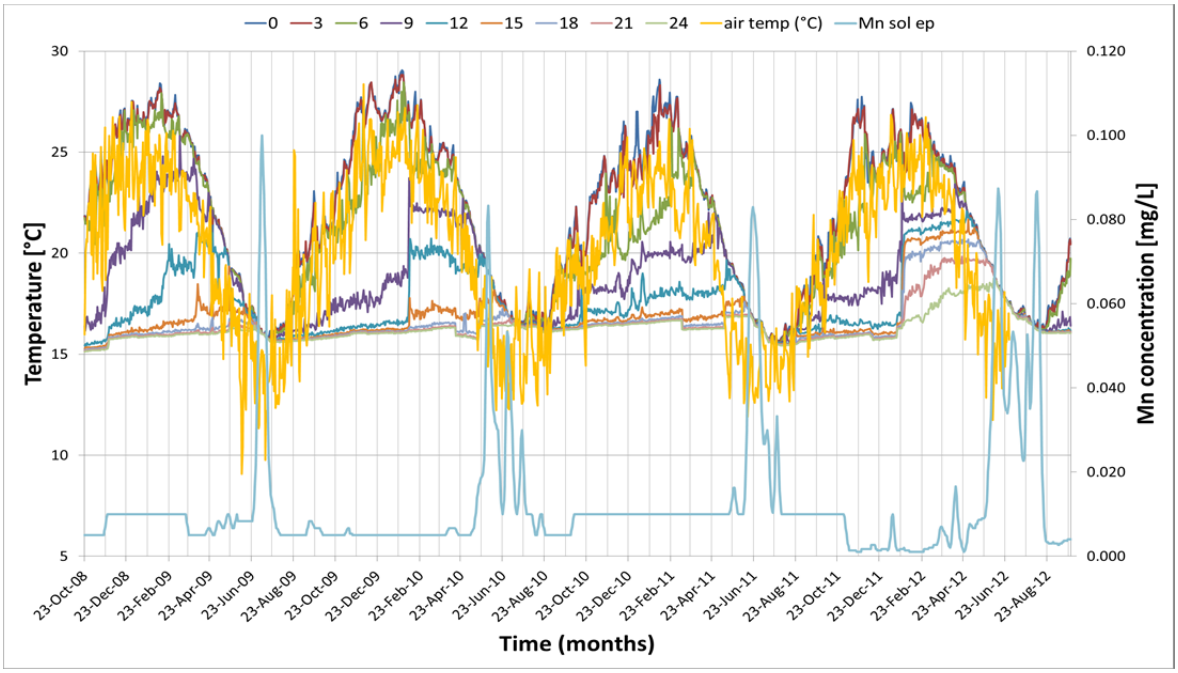

Figure 2. Soluble epilimnetic Mn, air and water column temperature time series, 2008-12

Once the MLR was chosen to be the most appropriate water column temperature prediction model, the data-driven equation was applied, yielding the final model: MLR with Data Driven Equation (MLRDDE). The predicted peak concentration is corrected through its correlation with the Mn load in the hypolimnion prior turnover and to the water column temperature. Table 2 provides the final structure and performance of the model; on the independent test set, MLRDDE was able to predict one week ahead the Mn concentration with a correlation coefficient of 0.83 (Figure 3). Similar models for shorter-term predictions were computed too, yielding similar performances. In particular, it can be noticed how the beginning of the peak (0.087 mg/L) is accurately predicted. Moreover, despite the end of the peak is lagged from the real values, also the second peak (similar maximum value) which is an atypical event never encountered into the training set, was predicted.

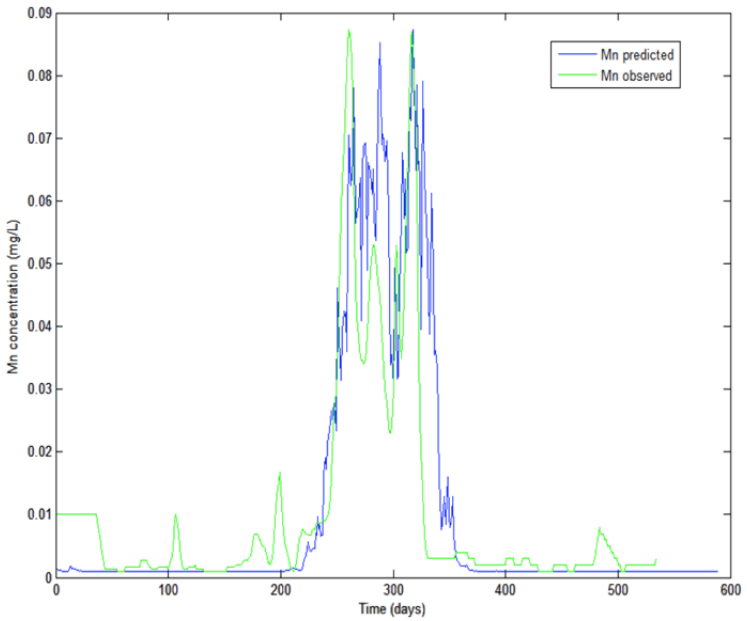

Figure 3. One week ahead Mn prediction using MLRDDE

Thus the model proved to be accurate but flexible enough to take into account unforeseen phenomena.

Despite minor events cannot be predicted by this model, MLRDDE with its few inputs is able to focus on the main event and predict it correctly. Other models (i.e. DYRESM) that rely in the forecast of multiple inputs were proved to bring to a poorer performance overall, and this could mean a possible prediction of some minor unimportant events but a poorer forecast of the critical peaks. Because of these considerations, MLRDDE was chosen to be a suitable model for the prediction problem faced in this research.

Table 2. Model structure and performance

Model part	Inputs	Output	R on the test set
Part 1 (MLR)	$\Delta T_w(t)$	$\Delta T_w(t+7)$	0.96
	$\overline{T_{air}}(t)$		
Part 2 (DDE)	$\Delta T_w(t+7)$	$Mn_{sol,ep}(t+7)$	0.83
	$Mn_{sol,hyp}(t)$		

5. DISCUSSION AND CONCLUSIONS

A seven-days ahead Mn prediction model has been created in order to support the planned development of a DSS that will assist Seqwater operators in their Mn treatment decisions. The final model (i.e. MLRDDE) was able to accurately predict the occurrence of the turnover event and the peak Mn concentration. The creation of a DSS based on MLRDDE will lead to cost savings for Seqwater as it will mean that there is a much lower requirement for manual sampling and laboratory testing of water quality parameters such as Mn. It will also provide operators with a near real time dashboard showing Mn values in the water column at the source water intake to the treatment plant.

Future research seeks to further improve the model validity through including more frequent Mn samples of the entire water column. More complete Mn datasets will help to explain some of the unique variations in peak Mn concentrations during the lake turnover period from one year to the next.

ACKNOWLEDGMENTS

The authors are grateful to Seqwater for their financial and technical support of this collaborative project.

REFERENCES

Bastarache, D., El-Jabi, N., Turkham, N. and Clair, T.A. 1997. Predicting conductivity and acidity for small streams using neural networks. *Canadian Journal of Civil Engineering*, **24**(6), 1030- 1039.

Bowden, G.J. (2003). *Forecasting water resources variables using Artificial Neural Network*. PhD Thesis, University of Adelaide, Australia.

Calmano, W., Hong, J. and Förstner, U. (1993). Binding and mobilization of heavy metals in contaminated sediments affected by pH and redox potential. *Water Science and Technology,* 28, 223-235.

Chiswell, B. and Huang, D. (2003). *Evaluation of North Pine dam sediment geochemistry*. Report prepared for South East Queensland Water Corporation, 11 May 2003.

Dojlido, J.R., Best, G.A. (1993). Chemistry of water and water pollution'. Ellis Horwood Series in Water and Wastewater Technology, pp. 360.

Hem, J.D. (1963). Chemical equilibria affecting the behaviour of manganese in natural water. *Hydrological Science Journal*, 8, 30-37.

Johnson, C.A., Ulrich, M., Sigg, L., Imboden, D. M. (1991). A mathematical model of the manganese cycle in a seasonally anoxic lake. *Limnology and Oceanography*, **36**(7), 1415-1426.

Maier, H.R., Jain, A., Dandy, G.C. and Sudheer, K.P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environmental Modelling and Software,* 25(8), 891-909.

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J. (2008). Stationarity is dead: whiter water management? *Science,* 319, 5863, 573-574.

Nürnberg, G.,K. (1988). A simple model for predicting the date of fall turnover in thermally stratified lakes. *Limnology and Oceanography,* 33, 5, 1190-1195.

Rouen, M., George, G., Kelly, J., Lee, M., Moreno-Ostoa, E. (2005). High-resolution automatic water quality monitoring systems applied to catchment and reservoir monitoring. *Freshwater Forum*, 23, 20–37.

Tundisi, J.S., Matsumura, T. (2011). Limnology. *Taylor and Francis Group, LLC*

Zhang, Q. and Stanley, S. J. (1997). Forecasting Raw-water Quality Parameters for the North Saskatchewan River by Neural Network Modelling. *Water Research*, 31(9), 2340- 2350.