Intelligent Data Mining of Vertical Profiler Readings to Predict Manganese Concentrations in Water Reservoirs

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

*Griffith School of Engineering, Griffith University, Gold Coast Campus, Queensland, Australia.
**Scientific Services and Data Systems, Seqwater, Queensland, Australia
[a]Corresponding author: edoardo.bertone@griffith.edu.au

Abstract

Continuously monitoring and managing manganese (Mn) concentrations in drinking water supply reservoirs are paramount for water suppliers since high soluble Mn loads lead to discoloration of potable water. Traditional Mn management approaches involve water samplings and laboratory analyses typically on a weekly basis; if critical thresholds are exceeded, then appropriate treatment procedures are exploited. Despite the Mn level currently being manually sampled throughout the year, in subtropical monomictic lakes such as Hinze dam, critical Mn concentrations in the epilimnion, where the water is drawn, are typically recorded only during winter during lake circulation. Vertical profiling system (VPS) installed can continuously collect physical parameters such as water temperature, pH or dissolved oxygen, which determine the transport process of Mn in the lake. Therefore, a long-term historical database gives opportunities for the development of a data driven prediction model to autonomously forecast future Mn concentration values.

In the present study, VPS and samplings data were collected and analysed, and prediction models applying nonlinear regression techniques and data-driven equations were developed and assessed; they were able to forecast future Mn concentrations from 1 to 7 days ahead with correlation coefficients higher than 0.83 on an independent test dataset. Importantly, the peak concentrations in the epilimnion during the lake destratification were accurately predicted. The model also displays the probabilities of the Mn to exceed certain key-thresholds, thus assisting operators in Mn treatment decision-making. Such a tool is very beneficial for the water supplier, since costly and time consuming water samplings for monitoring Mn concentrations can be avoided, thus relying only on the real time VPS-based model outputs.

Keywords: Water reservoirs; vertical profiling systems; manganese; reservoir destratification; water treatment

Introduction

An elevated Manganese (Mn) level in drinking water supply reservoirs is a widespread water quality problem faced by many water utilities. Mn concentration in a lake or reservoir is controlled by chemical, physical and biogeochemical processes with, in the case of eutrophic, warm monomictic lakes, which are thermally stratified for most of the year, usually high concentrations in the anoxic hypolimnion, which is the deepest layer of the lake, and low concentrations in the well-oxygenated epilimnion or top layer (Kohl and Medlar, 2007). As a consequence, in many reservoirs, the water is usually drawn from the epilimnion rather than from the nutrient-rich bottom waters, since the dissolved Mn level is in a safe range. The Mn monitoring is usually performed by weekly manual water samplings and subsequent laboratory analyses. However, with winter approaching, the thermal stratification becomes weaker and weaker until a destratification occurs, driven by wind and/or thermal convection. The lake circulation leads to an almost uniform distribution of chemical and biological constituents throughout the water column, with the top layers enriched in nutrients from the hypolimnion (Nürnberg, 1988) through mechanisms such as turbulent diffusion. Mn concentrations not acceptable for drinking purposes are typically recorded during this event (Calmano et al., 1993). As a result, the water supplier must treat the raw water accordingly: pre-filter chlorination and for higher concentrations the addition of potassium permanganate are the most widely applied strategies to oxidise the soluble Mn, which precipitates and can be easily removed. If the raw water were not treated to a required standard, the Mn-rich water would be distributed to the customers, leading to discoloration and possibly odour issues that can diminish the confidence of the customers in the water supplier. Hence, understanding the Mn cycle and predicting critical events is one of the major concerns for drinking water suppliers.

Interestingly, few studies have been conducted to try to fully model the Mn cycle. Importantly, to the author's knowledge, no studies attempted to predict future Mn concentrations. Several prediction models have been applied over the years for different environmental problems, but as pointed out by Maier et al. (2010), with regards to recently widely applied models such as Artificial Neural Networks (ANN), the vast majority of them deals with water quantity more than water quality issues. Besides, the typically modelled water quality parameters are pH or salinity (e.g. Zhang and Stanley, 1997; Bastarache et al., 1997) with few studies related to nutrients. An interesting model was created by Bowden (2003), which adopted an ANN to predict the peak concentrations of cyanobacteria in the River Murray, Australia. Process-based models have been widely applied in the environmental sector whenever enough data were made available. Nevertheless, attempts to model the Mn cycle, with particular focus on the rapid transport processes towards the epilimnion during the lake destratification, were not present. One of the few studies found in the literature was completed by Johnson et al. (1991), who created a mathematical model for simulating the Mn cycle in a Swiss lake. The model made use of differential equations including 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, because of the several inputs required, it is not ideal for short-term forecasts.

In summary, the creation of a model focusing on the prediction, up to one week ahead, of the soluble epilimnetic Mn concentration represents a novel research goal. Moreover, the creation of such an intelligent tool that is able to remotely collect and numerically process readily available VPS provisioned data to predict Mn concentrations would result in a cost benefit to the water supplier through a reduction in the number of water samplings and laboratory analyses required.

Methods

Research domain and data collection

The study domain is Hinze dam, which is also called Advancetown Lake (153.28°E, 28.06°S), which is a subtropical, eutrophic monomictic reservoir located in South-East Queensland, Australia (Fig.1). It is the largest water supply reservoir servicing the Gold Coast region, which has a population greater than 500,000.

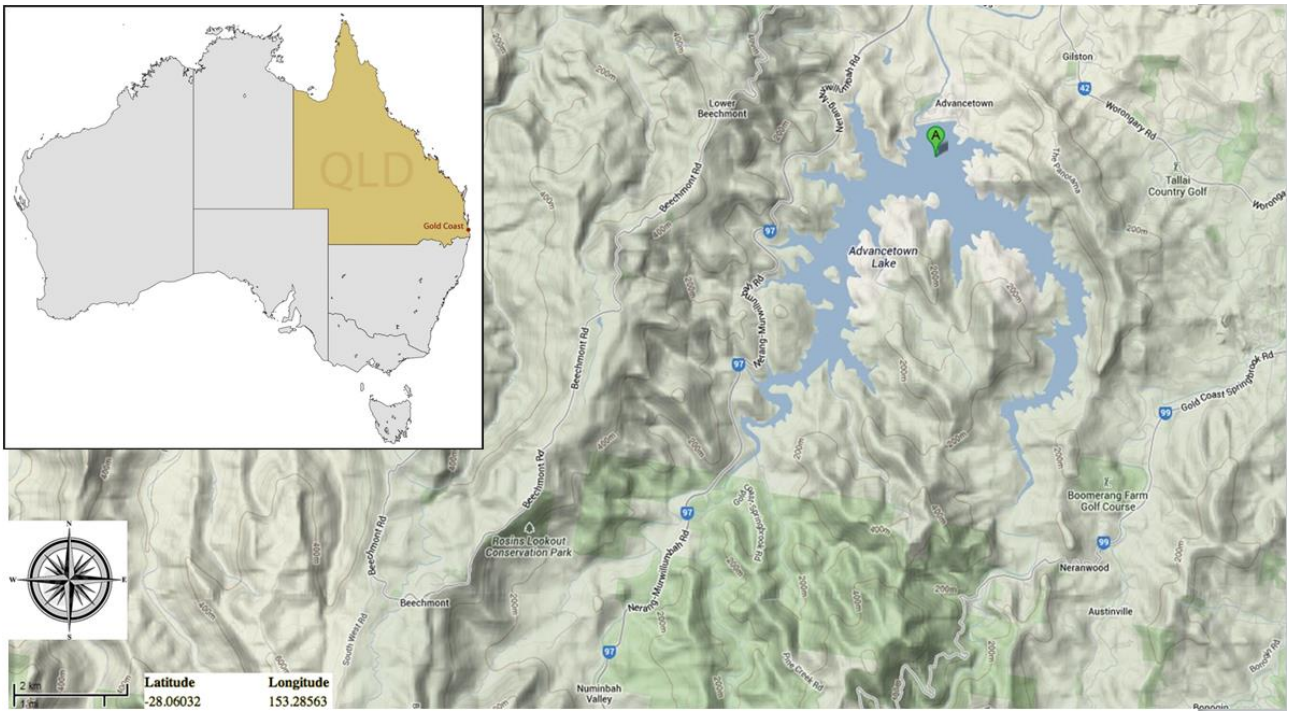

Fig.1 – Hinze dam map; green "A" represents VPS location

Its current capacity is 310,730 ML after a recent upgrade ("Stage 3", 2011) that doubled the previous volume and the average residence time (which was previously 920 days) The average depth is 32 metres and the surface area is 9.72 km^2, while the catchment area covers 207 km^2 of terrain which is mostly national parks. The two main inflows are the Nerang River and Little Nerang creek, coming from another smaller dam named Little Nerang Dam. The water is drawn from the most convenient depth (typically around 3-6 metres below the surface) through an intake tower located close to the dam wall, and it is distributed to the nearest potable water treatment plant located 10 km northeast.

Through an effective collaboration with Seqwater, the main bulk water supplier for South East Queensland, data collected from Hinze dam were made accessible. Data was collected from different sources, mainly from manual water samplings and from a YSI Vertical Profiling System (VPS). Manual water samplings were performed next to the dam wall from 0 to 24 metres on a 3 metres interval on a weekly basis and the results from 2000 to 2013 were collected. This dataset included water temperature, pH, conductivity, dissolved oxygen (DO), redox potential (ORP), turbidity, colour, soluble and particulate Mn. On the other hand, data from the VPS begins in 2008, when the VPS was placed close to the intake tower. The VPS consists of a buoy and a probe that can be located at any depth in the water column for effective data acquisition. The collected information can then be transmitted back to Seqwater's computer systems. At Hinze dam, VPS collects data at a time interval of 3 hours and a vertical spatial interval of 1 metre. The parameters include water temperature, conductivity, pH, DO, turbidity, ORP, chlorophyll-a and blue green algae. Because of the smaller time interval for the collected VPS data, manually sampled Mn data was aligned to this dataset using linear interpolation.

Weather data was collected from the Australian Bureau of Meteorology (BoM). Some weather data was available from a local weather station located next to Hinze dam, but this dataset was incomplete so recalibration of data from the nearest BOM weather station located at the Gold Coast Seaway (30km northeast) was performed. Weather data included air temperature, wind speed and direction, solar radiation and rainfall. The dam inflow data was received from the Queensland Government Department of Energy and Resources Management (DERM).

Data pre-processing and analysis

Time-series graphs for all of the variables were visually inspected, in order to check whether the relationships described in the literature are confirmed by the real data for Hinze dam. Once each relationship was examined, statistical analyses were performed, in order to reveal any significant correlation between variables. The performed analyses included: linearity tests; stationarity tests; general statistics; autocorrelograms and crosscorrelograms; scatter plots; bimodality tests; seasonality tests; heteroskedasticity tests; and normality tests.

Specifically, the presence of nonlinearities was assessed through the BDS test (Brock et al., 1987), which can also test the presence of nonlinear dependencies between variables, provided that linear ones have been removed. Moreover, two different tests to assess normality were performed, namely the Jarque-Bera test (Jarque and Bera, 1987), based on skewness and kurtosis, and the Lilliefors test (Lilliefors, 1967), which is an adaptation of the Kolmogorov-Smirnov test. Finally, the bimodality can be assessed by use of different coefficients, such as the bimodality coefficient (BC), Hartigan's dip statistics (HDS), or the difference in Akaike's information criterion (AIC) (Freeman and Dale, 2013). Since the BC is based on skewness and kurtosis, this test was chosen to be the most appropriate index, consistent with other tests such as aforementioned Jarque-Bera test.

As expected for an environmental system, time series analysis presented several complexities. The usually high skewness and kurtosis lead to positive results for the test for bimodality, with BC higher than 0.555 (i.e. the threshold level to detect bimodality) for most of the time series of Mn, dissolved oxygen, turbidity and hypolimnetic water temperature. Besides, bimodality is a sign of non-normality. As a consequence, all of the bimodal time series are also not normal according to the Jarque-Bera and Lilliefors tests. Moreover, other variables, which did not show evidence of bimodality, proved themselves to be not normal with the aforementioned tests. Ultimately, the application of the BDS test to detrended and deseasonalised data, along with visual analysis of autocorrelograms and crosscorrelograms, shows the presence of nonlinearities and nonlinear relationships between most of the time series. On the other hand, high linear correlations were found between air temperature and water temperature time series, despite that a hysteresis cycle was noticed, mainly because of the thermal inertia of the water mass. However, the hysteresis is rather small, because of the relatively small volume of the reservoir.

Despite the presence of nonlinearities, through plotting the soluble Mn values against all the possible predictors it was possible to understand which type of nonlinearity links the variables together. The most interesting results are shown in Figure 2, representing a clear hyperbolic correlation between the soluble epilimnetic Mn data against the difference of water temperature between the surface and the bottom (ΔT_w) after normalization. This result could have been expected because of the theoretical cycles of those variables in warm monomictic lakes found in the literature and confirmed in Hinze data (Fig. 3). No significant changes in the variables' cycles can be noticed after the volume upgrade. Soluble Mn shows high concentrations in the epilimnion only during winter turnover, when the water temperature of the whole water column reaches the same value and hence ΔT_w goes to zero. The analysis of other variables such as dissolved oxygen (DO) and pH (always high in the epilimnion) showed that the production of soluble Mn is not supported by the oxic, alkaline epilimnetic environment, thus justifying the low Mn level throughout the

stratification season; however they could not provide significant correlation, since the winter peaks are unaffected by the biogeochemistry of the epilimnion. Iron appeared to have a similar but slower cycle to that of Mn, but it could not be sufficiently correlated for a statistically significant Mn forecast.

In addition, a relationship was found between the percentage of hypolimnetic Mn going into the epilimnion during the winter circulation and the water column temperature (T_w) just before the onset of the circulation, which is constant throughout the water column (Fig. 4). Higher water temperatures lead to less dense waters and stronger mixing processes such as turbulent diffusion, therefore more soluble Mn can go to the surface layers before oxidation and precipitation occur throughout the water column because of the temporary presence of dissolved oxygen at any depths. Since, because of the recent dam capacity upgrade, the reservoir has evolved towards a new equilibrium, which among other changes showed smaller epilimnetic soluble Mn peaks. The equation related to the last five years was considered to be more appropriate and reliable to describe the future peak events. By analysing DO (decreasing right after the turnover until fully depleted in early summer), pH (typically acidic) and ORP (decreasing during the stratification season) it was evident that the production of soluble hypolimnetic Mn is predominately resulting from the biogeochemistry relationships of these three factors. However, our exploratory studies revealed that epilimnetic Mn prediction was predominately a function of the temperature gradient of the water column (i.e. mixing momentum) and these hypolimnetic Mn biogeochemistry relationships were of much less significance in this upper layer where water was drawn for treatment.

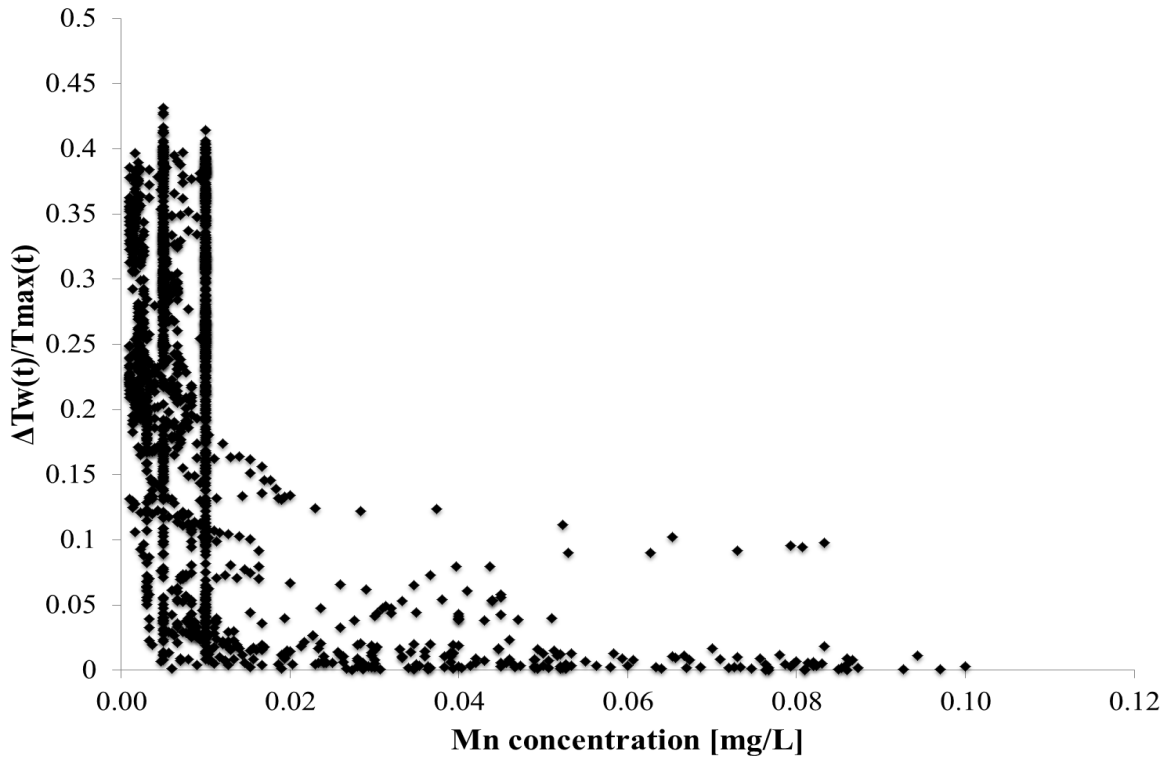

Figure 2: Scatter plot of soluble epilimnetic Mn with $\Delta T_w(t)/T_{max}(t)$ for Hinze Dam from 2008-2013. ΔT_w is the difference of water temperature between surface and bottom; T_{max} is the maximum water temperature of the water column at time

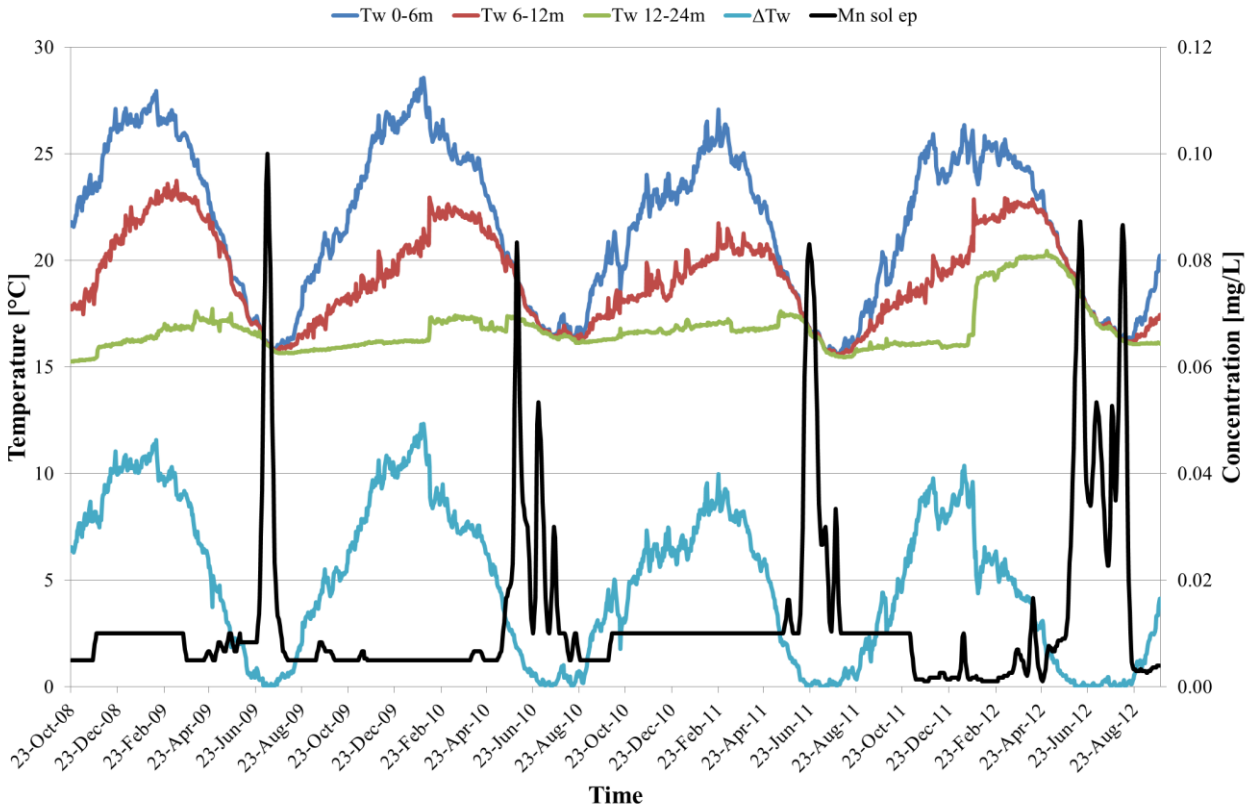

Figure 3: Time series soluble epilimnetic Mn and water temperature at different depths for Hinze Dam between 2008-2012

t.

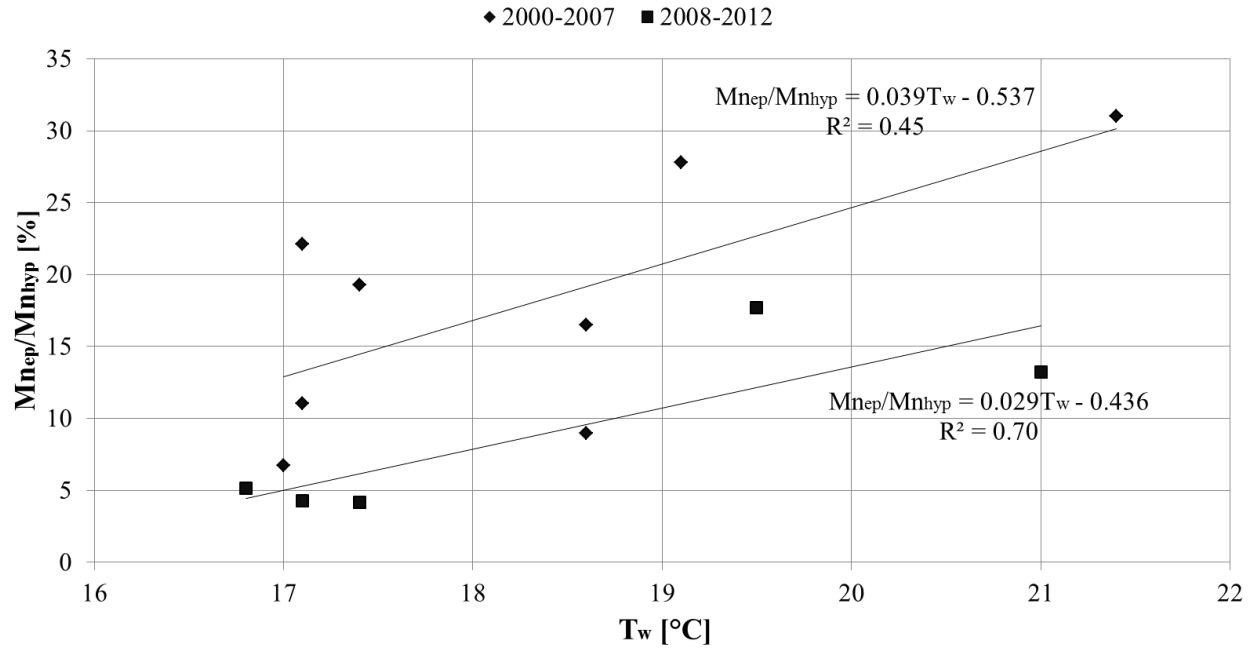

Figure 4: Relationship between Mn_{ep} /Mn_{hyp} and T_w for Hinze dam over the 2000 – 2012 period. Mn_{ep} is the peak concentration of soluble Mn in the epilimnion during the turnover event; and Mn_{hyp} is the concentration of soluble Mn stored in the hypolimnion (data limited to 24m) just before the

beginning of the circulation event. T_w is the temperature of the water column at the beginning of the circulation event (when $\Delta T_w \cong 0$ hence water column temperature is assumed uniform)

Model development

Once the statistical analysis was completed and the features and correlations of the data assessed, it was possible to derive the most appropriate model and its key input parameters (see Figure 5). According to the results of the statistical analysis, no good correlation between Mn and any predictors at lag > 7 days was found; hence no individual variable is able to predict the soluble Mn in the epilimnion one week ahead. However, a strong correlation between Mn and ΔT_w was evident at lag 0 days. Also, very high dependencies were found between air and water temperature, with correlation decreasing with increasing lag. After careful assessment, the model required three data processing modules or parts to be created to ensure an accurate and reliable Mn prediction. The three model parts are described below:

- Model part 1: completes analysis of the current water column temperature difference and the forecasted air temperature up to one week ahead (collected from the BoM) and outputs the water column temperature difference one week ahead.
- Model part 2: takes the output of Part 1 (i.e. $\Delta T(t+7)$) and by using the hyperbolic correlation relationship shown in Figure 3, it will yield a prediction of the soluble Mn in the epilimnion 7 days ahead.
- Model part 3: where part 2 predicts the beginning of the lake turnover event, through the correlations shown in Figure 4, the future peak Mn value will be corrected using the amount of Mn stored in the hypolimnion at the beginning of the turnover event, measured through manual water sampling.

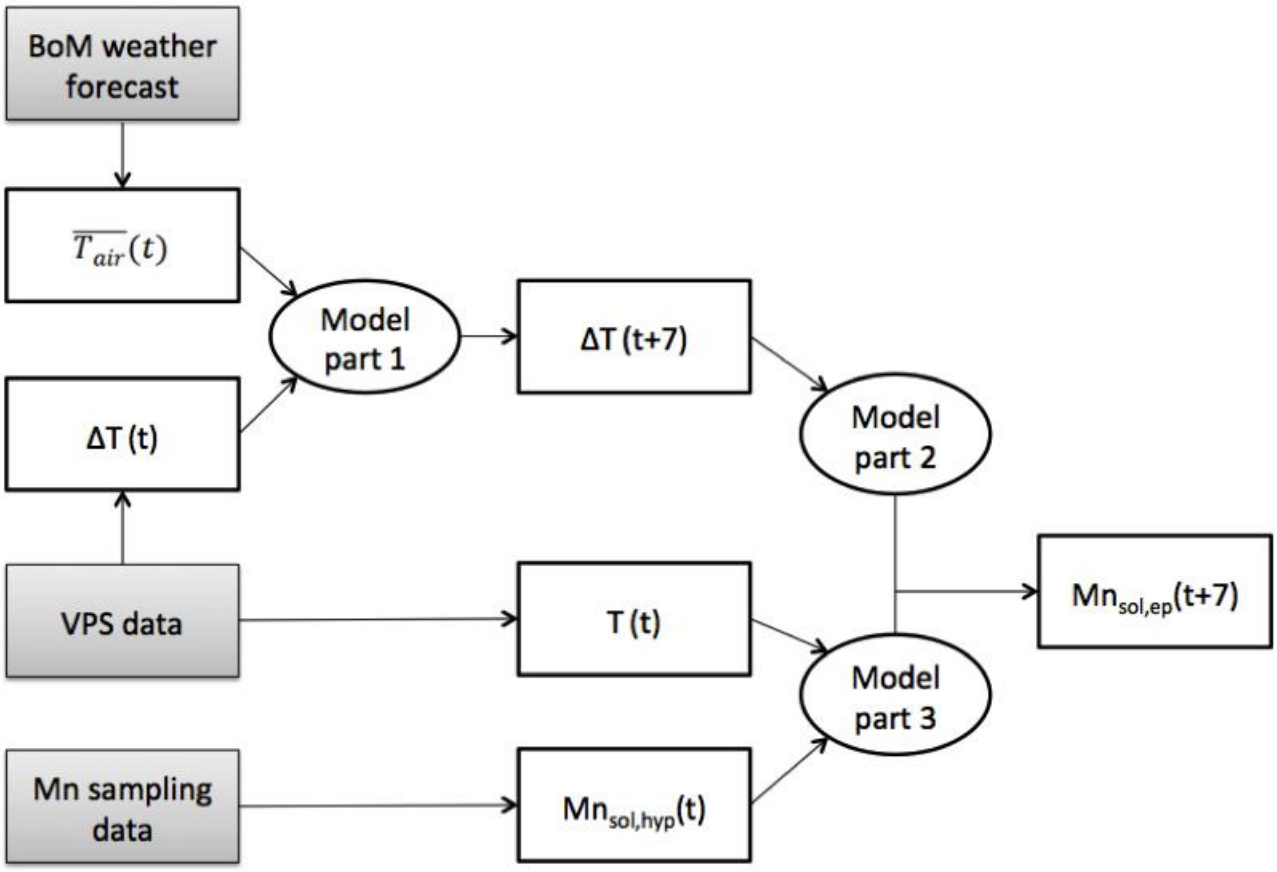

Figure 5: Model structure and core analysis parts

The variables shown in Figure 5 are defined as:

- $\overline{T_{air}}(t) = \frac{1}{7} \cdot \sum_{n=1}^{7} T_{air}(t+n)$ = air temperature forecast from 1 to 7 days ahead, collected from the BoM;
- $\Delta T(t) = \frac{\sum_{z=1}^{6} T(z,t)}{6} - \frac{\sum_{z=12}^{24} T(z,t)}{13}$ = current water temperature difference between epilimnion (simplified as the top 6 m) and hypolimnion (simplified as the layer between 12 and 24 m, since there is no relevant temperature difference below 24m);
- $\Delta T(t+7)$ = water temperature difference prediction 7 days ahead;
- $T(t) = \frac{\sum_{z=1}^{24} T(z,t)}{24}$ = average water column temperature (first 24 m yields better correlation);
- $Mn_{sol,hyp}(t)$ = current average soluble Mn stored between 12 and 24 m; and
- $Mn_{sol,ep}(t+7)$ = average soluble Mn stored between 0 and 6 m, 7 days ahead.

For Part 1, different categories were taken into account, focusing particularly on statistical and physical models. Modelling water temperature is an issue that has involved many research studies in the past. A good amount of them made use of physical models: for instance, Caissie et al. (2007) modelled river water temperature using a deterministic model, obtaining good results, with errors concentrated during the snowmelt season. Interestingly, Helfer et al. (2011) made use of an existing, worldwide-applied processed-based model called DYRESM to calculate the water temperature and evaporation rates of a reservoir in order to assess different strategies for evaporation reduction. The use of deterministic models provides a comprehensive study of the system, but the application of the physical equations involved (e.g. net heat flux) requires a remarkable amount of inputs (e.g. saturated vapour pressure, air-water vapour pressure, air temperature, wind velocity, atmospheric pressure, cloud cover, boundary conditions etc.). Moreover, the vast majority of the previous researches did not try to forecast the water temperature into the future, which would be a harder task; future water temperature prediction requires the calculation of the forecasted value of each input value, thus compounding the level of uncertainty associated to each forecast to the final model error.

Hence, a simpler statistical model, with a reduced number of inputs, was considered to be the best option. Reducing the number of inputs was achieved by calculating the cross-correlation coefficients at different lags for smoothed data with a moving average method and variable span. Despite the cross-correlation coefficient gradually decreased by increasing the lag, it was discovered that smoothing data increases the level of correlation. In particular, a 7-days ahead air temperature moving average was found to provide the best correlation with the epilimnetic water temperature, and also with the whole water column temperature. Smoothed solar radiation provided good correlations too, but it was assessed to be redundant latent variable of air temperature so was not required. Wind, rainfall and river inflow proved to have only short-term effects; besides, high rainfall and inflow events are typically recorded during the wet season in summer, thus not affecting the winter turnover event. In conclusion, a model based only on the air temperature forecast proved to have the potential for good performance, since air temperature is one of the easiest meteorological variables to forecast, and a one-day error can be greatly reduced in importance by smoothing the whole week of forecasts.

For Part 2 of the model, a relatively simplistic statistical nonlinear regression model was selected. This model applied the hyperbolic correlation (Figure 3) to yield a good estimate of the soluble Mn in the epilimnion by making use of the known water column temperature.

Where in cases that Part 2 of the model predicts the onset of a critical event (i.e. prediction above 0.02 mg/L), then Part 3 of the model will make use of the data-driven equation shown in Figure 4 to produce a correction coefficient that is applied to the output of Part 2. The developed equation related to the last 5 years of data was decided to be more appropriate, because of the recent Hinze

Dam upgrade works which has slightly changed the dams' hydrodynamic and biogeochemical properties.

All prediction models need to be validated. In most time series forecasting studies, this is usually achieved by dividing the dataset into a training set, where the model is built, and a test set, were the performance is assessed. In this way, problems such as "overfitting" (i.e. where a model has very high accuracy for the training set, but making predictions with new data is very poor) are avoided. This type of validation has been completed in this study. Moreover, a second form of validation has also been performed in this study, where the tested model is also used to predict the lake turnover event in 2013, and the accuracy examined.

Model validation was considered fundamental because, as Milly et al. (2008) asserted, "stationarity is dead" since a changing climate and human activities means that it is more likely that future environmental cycles will be different from those of the past. Interestingly, this is particularly true for Hinze dam, since the reservoir was recently upgraded and its volume gradually increased over the last wet seasons, thereby changing some of the characteristics of the dam mixing processes. The creation of a model that is flexible enough to handle change is crucial for the long-term validity of the intended intelligent Mn prediction tool.

The last part of this research is related to the model deployment. It is important that any developed Mn prediction tool is user-friendly and its outputs can be easily interpreted by treatment plant operators for decision-making. The creation of a prototype user-friendly Graphical User Interface (GUI) that could be utilised by dam operators was completed in the final stage of this project.

Results and Discussion

For the Part 1 of the model, because of the hysteresis cycles between air and water temperature (Fig. 6 and 7), a threshold seasonal autoregressive model (TSAM) was created.

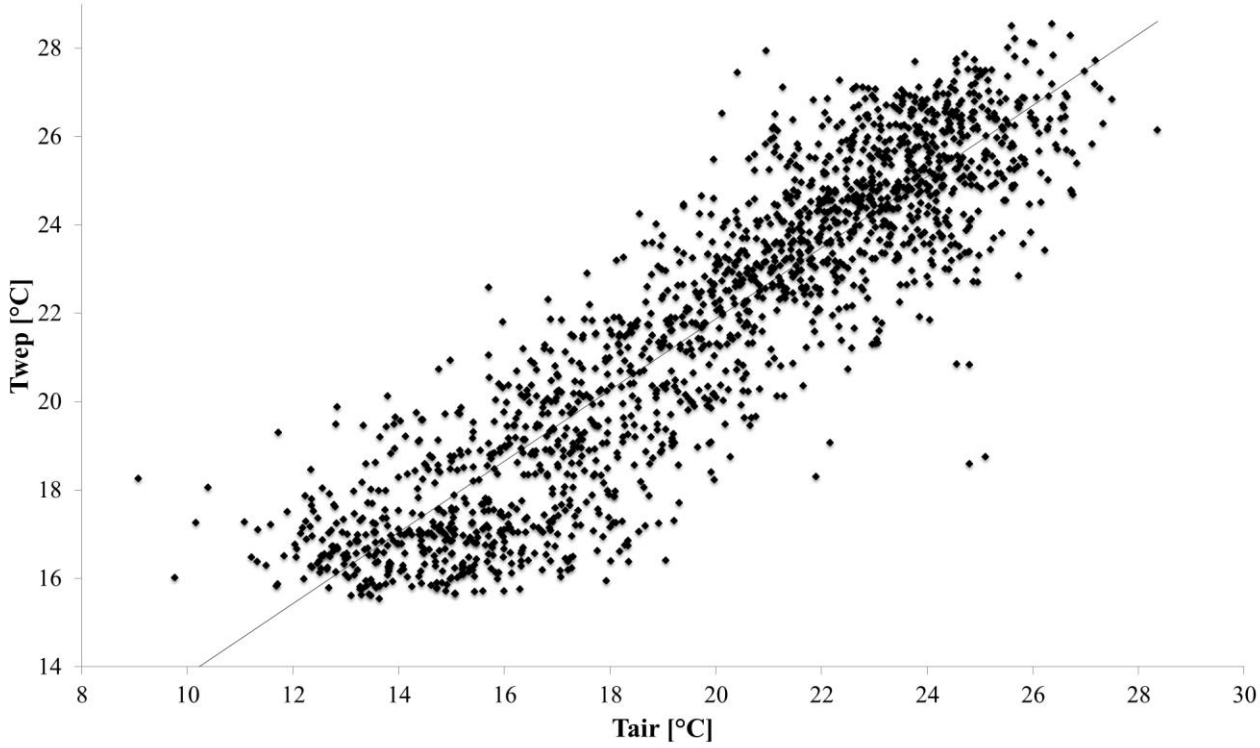

Figure 6: Scatter plot air temperature – epilimnetic water temperature, Hinze dam, 2008-2013

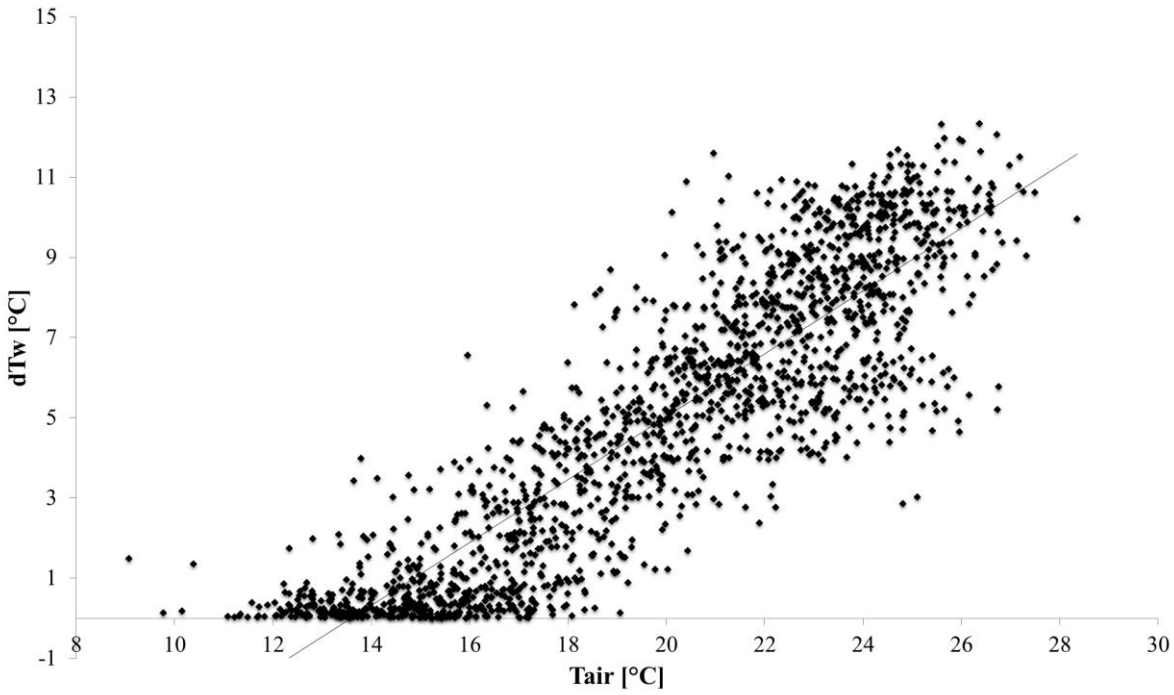

Figure 7 Scatter plot air temperature – water column temperature difference, Hinze dam, 2008-2013

The model is able to detect seasonal change (i.e. warming, cooling, winter), and to relate air and water temperature accordingly, using relevant equations. In this way, nonlinearities related to the hysteresis cycle are accounted, thus improving the final model performance. The model was calibrated using a training set of data (i.e. 2008-11) and the performance was tested on an independent set of data (i.e. 2011-12). Equation (1) presented below describes the TSAM:

$$\Delta T(\mathrm{t}+7) = \Delta T(\mathrm{t}) + A_i \cdot \left(\overline{T_{air}}(t) - T_{ep}(\mathrm{t})\right) + B_i \quad (1)$$

where:

$\Delta T(t+7)$ is the predicted water column temperature difference 7 days ahead;
$\Delta T(t)$ is the current water column temperature difference;
$\overline{T_{air}}(t) = \frac{1}{7}\sum_{n=1}^{7} T_{air}(t+n)$ is the mean predicted air temperature (collected from the BoM) from one to 7 days ahead;
$T_{ep}(t)$ is the current water temperature in the epilimnion;
A_i and B_i are coefficients calculated by linear regression analysis, which change according to the season, expressed by the index i described below:

- i = 1: warm season (for Hinze Dam, September to February);
- i = 2: cooling season (for Hinze Dam, March to May); and
- i = 3: winter season (for Hinze Dam, June to August).

The derived regression coefficients for TSAM are A_1 = 0.4272, A_2 = 0.2701, A_3 = 0.2248, B_1 = 0.7505, B_2 = 0.2328 and B_3 = 0.4367.

Hence, by calculating the difference in temperature between the surface waters and the surrounding average air for the next 7 days, the model can interpret historical data and the seasonal characteristics in order to simulate the degree of heat exchange and predict the water column temperature with high accuracy. The correlation coefficient for the independent test set was 0.97. Figure 8 illustrates that there is a very good match between the historical and modelled data.

Although the only input is air temperature, previous studies (Livingstone and Padisàk, 2007) already stated that it can be considered the main variable affecting the heat balance of the surface layers of a lake, implicitly including other factors, thus the high correlation found by the model is reasonable. There are very few prediction discrepancies evident in this figure, except on the few occasions when there were sudden drops in water column temperature differential that occurred during summer. These unusual events were not related to the air temperature, but due to mixing processes instigated by very high precipitation events. Evidently, these unusual events are not critical to the models goal to predict unacceptable Mn concentrations in the epilimnion, as very high precipitation events do not typically occur in the study region during the critical winter turnover event, when Mn peaks are recorded in the epilimnion.

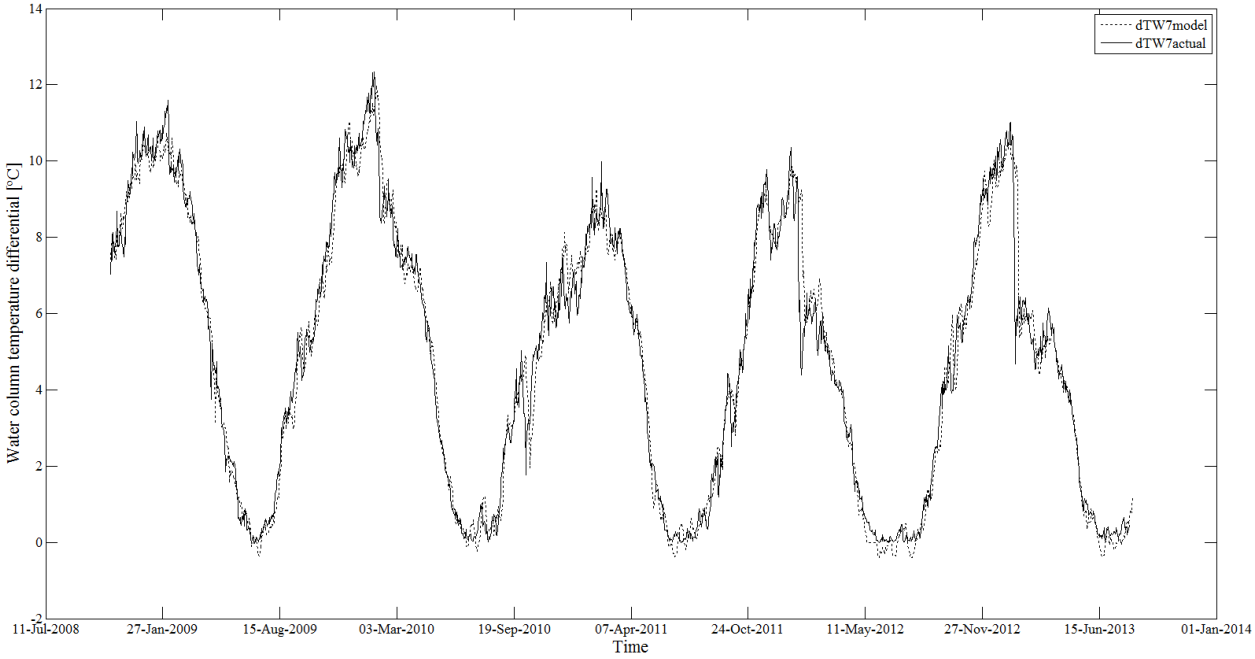

Figure 8: The actual and predicted by TSAM water column temperature difference for Hinze dam, 2008-13

The application of Artificial Neural Networks (ANN) was also explored and tested for prediction accuracy comparison. ANN were reasonable accurate overall, but had much higher variability than the simpler empirical model: with the application of the hyperbolic transformation included in Part 2 of the model, the variability would increase exponentially. Moreover, a physical model was developed using DYRESM software, yielding good results for the training set but poor results for the test set, where there was much more uncertainty related to the multiple input forecast data that needed to be included thereby compounding errors. This limitation applies to any process-based models, where complexity and the required multitude of inputs is a performance-limiting factor for forecasting problems. In summary, exploratory assessments of more complex predictions tools revealed that the developed empirically derived time-series forecast models may more fit-for-purpose for this particular forecasting problem.

The results of TSAM were provided as input to Part 2 of the Mn prediction model. A hyperbolic transformation of the normalized values of the difference in the water column temperature was computed, yielding Equation 2.

$$Mn_{sol,ep}(t) = \left[\frac{\left(\left(\frac{1}{1+\Delta T(t)}\right) - min\left(\frac{1}{1+\Delta T(t)}\right)\right)}{\left(max\left(\frac{1}{1+\Delta T(t)}\right) - min\left(\frac{1}{1+\Delta T(t)}\right)\right)}\right]^3 \cdot \left(max\left(Mn_{sol,ep}\right) - min\left(Mn_{sol,ep}\right)\right) + min\left(Mn_{sol,ep}\right) \quad (2)$$

Where:

$Mn_{sol,ep}(t)$ is the value of soluble Mn in the epilimnion at time *t* [mg/L];
$max\left(Mn_{sol,ep}\right)$ is the maximum value of soluble Mn in the epilimnion within the historical set [mg/L]; and
$min\left(Mn_{sol,ep}\right)$ is the minimum value of soluble Mn in the epilimnion within the historical set [mg/L].

$$\Delta T(t) = \frac{\sum_{z=1}^{6} T(z,t)}{6} - \frac{\sum_{z=12}^{24} T(z,t)}{13} \quad (3)$$

Where:

$T(z,t) =$ water temperature at depth *z* at time *t* [°C]; and
z = water depth [m].

Equation 2 produced a correlation coefficient (*R*) with the soluble epilimnetic Mn of 0.74 for the test data set. With the application of the peak soluble Mn correction coefficient (Part 3 of model), the correlation coefficient results further improved.

Figure 9 shows the output of the model for the test set, which includes the peaks related to the untrained 2012 winter Hinze Dam circulation event. The correlation coefficient achieved for the test data set was higher than 0.86. Moreover, it can be seen in the figure how all of the 3 main peaks occurring during the lake turnover event were predicted well. This prediction outcome was very pleasing since the historical data set used for training only had one main peak per turnover event recorded. The prediction of this unique event is a robust validation of the model, since it was able to adapt to the evolving physical environment of the reservoir.

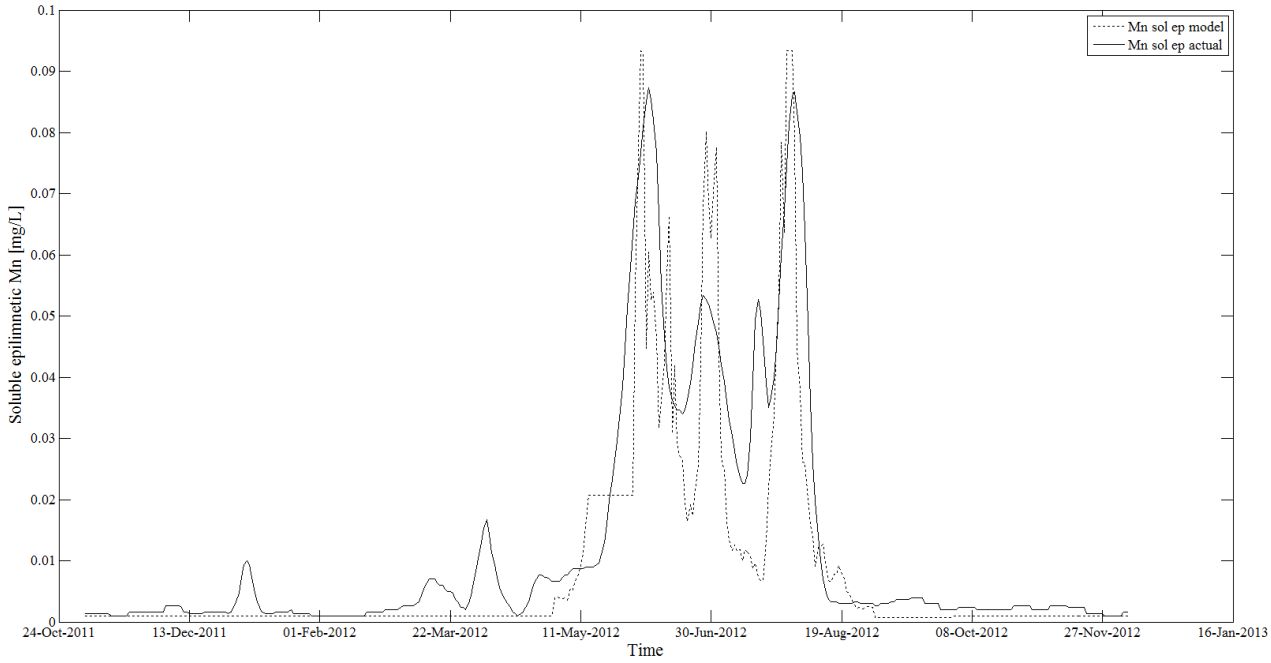

Figure 9: Prediction of soluble epilimnetic Mn of the final model. Test set (Nov 2011- Jan 2013) Hinze dam

Given that the overarching goal of the prediction model was to provide water treatment operators with a user-friendly tool to know well in advance when elevated soluble Mn concentrations would need to be treated, another performance index was derived. Noting that pre-filter chlorination is applied for soluble Mn concentrations in the raw water when it is higher than 0.02 mg/L, the following indexes were determined to better quantify the performance of the model:

A. Number of correct no warnings: when both the actual and predicted concentrations were lower than 0.02 mg/L;
B. Number of correct warnings: when both the actual and predicted concentrations were higher than 0.02 mg/L;
C. Number of false no warnings: when the predicted concentration was lower than 0.02 mg/L, but the actual value was higher; and
D. Number of false warnings: when the predicted concentration was higher than 0.02 mg/L, but the actual value was lower.

Table 1 presents the findings of an assessment of the prediction model for the above soluble Mn thresholds indexes for the Hinze Dam test data set. Count categories C and D are incorrect threshold predictions.

Table 1: Soluble Mn threshold range warnings count

Warning type	**No. of events**	**Percent of total events (%)**
A. Correct no warning	204	69.9
B. Correct warning	63	21.6
C. False no warning	16	5.4
D. False warning	9	3.1
TOTAL	**292**	**100.0**

Hence, with only 8.5% (5.4 + 3.1 = 8.5%) of incorrectly predicted warnings for operators, the prediction model shows strong promise for being a very robust Mn decision-making tool. Following the above same principles, other models were built in order to predict the soluble epilimnetic Mn at shorter timeframes, namely 6, 5, 4, 3, 2 and 1 day ahead. Slightly different coefficients in the equations were derived, and similar prediction accuracy was obtained to the 7-days ahead prediction.

It must be said that the model was created using actual air temperature values, thus assuming a perfect air temperature forecast. In order to overcome this limitation, a sensitivity analysis was carried out. Firstly, the weather forecast provided by the BoM were collected over a period of two months, and an average error of 1.1 ºC was recorded, both for the minimum and the maximum air temperature forecast calculation. Secondly, white noise (i.e., a series of uncorrelated random values with constant mean and constant variance) with values included between [-2.2; 2.2] was added to the air temperature time series used in the test set, in order to account for the uncertainty related to the weather forecast. Although the calculation of $\Delta T_w(t+7)$ was not affected, the correlation coefficient for the Mn prediction 7 days ahead dropped from 0.86 to 0.80. However, to counteract this slight reduction in reliability, more specific and reliable weather forecast data can be requested from BoM, thus reducing this error and increasing prediction accuracy closer to those for the perfect air temperature forecasts.

The model was prepared by May 2013, and it was used for predicting soluble Mn concentrations during the course of the 2013 winter turnover event (Figure 10). This was the second stage of model validation using newly collected data. Although a small, early spike in soluble epilimnetic Mn, not

associated with the turnover dynamics, was not detected, the model managed to predict 7 days ahead the beginning of the soluble epilimnetic Mn main peak event with only one-day discrepancy. Since the inputted actual sampled Mn concentration data in the model were collected on a weekly basis, the model was trained using weekly data interpolated to a daily series; thus an error of one day is understandable given this limitation in the sampled input data. Moreover, the model predicted a peak soluble Mn concentration of 0.12 mg/L whereas the actual peak soluble concentration was 0.13 mg/L. The difference between the predicted and actual result is only 7.7 % providing further evidence of the robustness of the model. From Figure 10 it can be noticed how the 2013 winter turnover was slower than the previous ones; this could be due to the increased volume of the dam. In fact, for the first time, this dam reached its new full capacity during the 2013 wet season, thus increasing its thermal stability with follow-on influences to transport mechanisms. Future data collection and turnover analyses will help in understanding if a recalibration, leading to a dam post-upgrade version of the model, is necessary.

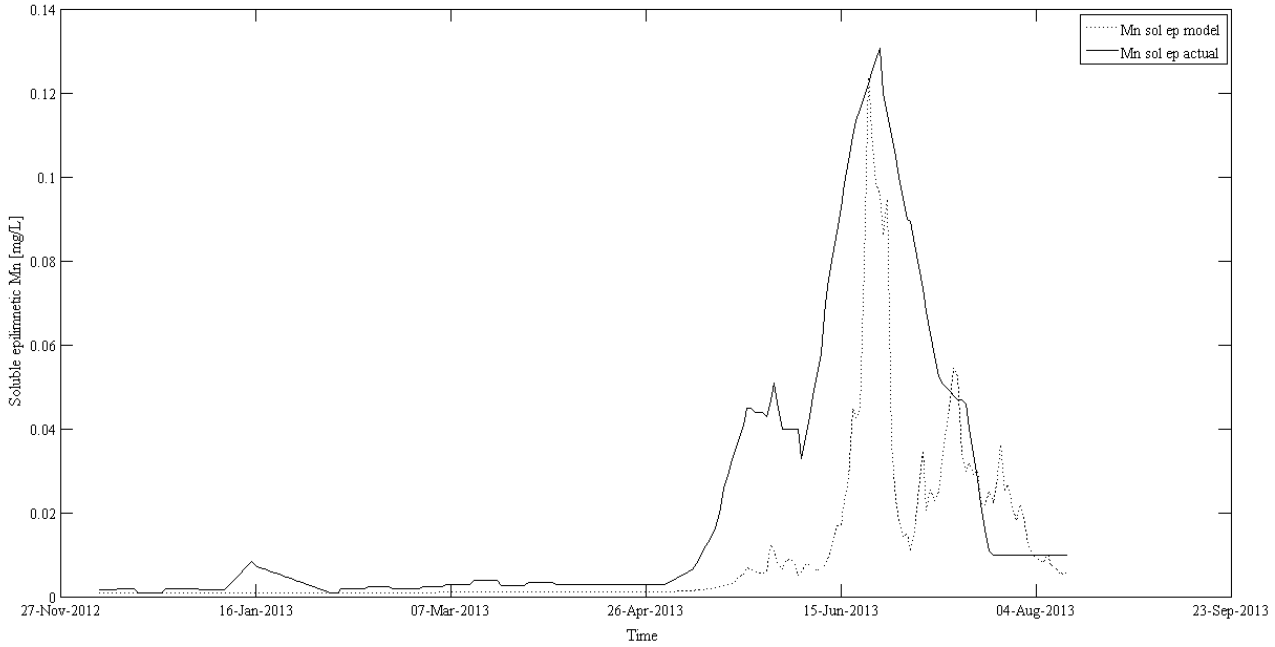

Figure 10: Prediction of soluble epilimnetic Mn of the final model. Test set 2 (Nov 2012- Aug 2013) Hinze dam

Conclusions

A Threshold Seasonal Autoregressive Model (TSAM) incorporated with high-hyperbolic data-driven equations and peak correction equations was built in order to forecast up to 7 days ahead the concentration of soluble Mn in the epilimnion of Hinze dam. The model, based on water temperature and forecasted air temperatures, achieved a correlation coefficient of 0.86 when predicting the winter 2012 lake circulation test set key features of the critical peak event. Specifically, the model was able to suitably predict the beginning of the critical event, the peak Mn concentrations, and the presence of multiple peaks. A second validation step applied the model to predict the 2013 winter turnover event, also achieving a high correlation coefficient of 0.76 despite the different dynamics of the upgraded dam. The model forecasted the 2013 winter turnover main peak event with high accuracy, both in terms of the timing of the event and in the peak concentration estimation. Importantly, the key feature of the model is that it will become even more accurate over time as more data becomes available. With its simple structure, the model yielded better Mn forecasting performance than the more complex and widely used process-based models. Thus, autonomous data-driven approaches for forecasting some of the key water quality parameters

such as epilimnetic Mn offers a novel alternative to traditional methods and also more pro-active water quality operational management and decision-making.

The model can also display outputs for dam operator decision-making purposes. Presently, a simple soluble Mn threshold warning system has been developed, but in the future a user-friendly GUI will be developed and utilised by water treatment plant operators. It will have a range of cost and time saving benefits for the water supplier, including a reduction in the amount of unnecessary pre-filter chlorination, much less costly and time-consuming weekly in-situ Mn samplings and associated laboratory analysis will be required in the future, and more reliable operator decision making.

Future work aims to further validate the model accuracy with 2014 data, build a user-friendly Mn decision support system, and also expand the model capabilities by applying it to other reservoirs where different Mn cycles present new potable water treatment challenges.

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.

Brock, W. A., J. Scheinkman, W. Dechert, and B. LeBaron. 1996. A test for independence based on the correlation dimension'. *Econometric Reviews*, **15**(3), 197-235.

Caissie, D., Satish, M.G. and El-Jabi, N. 2007. Predicting water temperature using a deterministic model: application on Miramichi river catchments (New Brunswick, Canada). *Journal of Hydrology,* **336**, 303-315.

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.

Freeman, J.B. and Dale, R. 2013. Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, **45**(1), 83-97.

Helfer, F., Zhang, H. and Lemckert, C. 2011. Modelling of lake mixing induced by air-bubble plumes and the effects on evaporation. *Journal of Hydrology,* **406**, 182-198.

Jarque, C.M. and Bera, A.K. 1987. A test for normality of observations and regression residuals. *International Statistical Review*, **55** (2), 163–172.

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.

Kohl, P. and Medlar, S. 2007. *Occurrence of Manganese in Drinking Water and Manganese Control.* IWA Publishing.

Lilliefors, H. 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**. 399–402.

Livingstone, D.M. and Padisàk, J. 2007. Large-scale coherence in the response of lake surface-waters temperatures to synoptic-scale climate forcing during summer. *Limnology and Oceanography*, **52**, 896-902.

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. and 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.

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.