Autonomous intake selection optimisation model for a dual source drinking water treatment plant
E. Bertone, K. O’Halloran, M. Bartkow and K. Mann

ABSTRACT
The Mudgeeraba drinking water treatment plant, in Southeast Queensland, Australia, can withdraw raw water from two different reservoirs: the smaller Little Nerang dam (LND) by gravity, and the larger Advancetown Lake, through the use of pumps. Selecting the optimal intake is based on water quality and operators’ experience; however, there is potential to optimise this process. In this study, a comprehensive hybrid (data-driven, chemical, and mathematical) intake optimisation model was developed, which firstly predicts the chemicals dosages, and then the total (chemicals and pumping) costs based on the water quality at different depths of the two reservoirs, thus identifying the cheapest option. A second data-driven, probabilistic model then forecasts the volume of the smaller LND 6 weeks ahead in order to minimise the depletion and spill risks. This is important in case the first model identifies this reservoir as the optimal intake solution, but this could lead in the long term to depletion and full reliance on the electricity-dependent Advancetown Lake. Both models were validated and proved to be accurate, and with the potential for substantial monetary savings for the water utility.

Key words | data-driven modelling, optimisation model, water level forecast, water treatment

INTRODUCTION
The delivery of safe drinking water is an essential task for any bulk water supplier charged with providing water that is clear, cool, good tasting, reasonably soft, stable, plentiful, and cheap (Sarai 2006). The treatment of raw water from lakes, rivers, or wells is, therefore, required in order to meet the drinkability guidelines defined by different national regulators.

In case of the Mudgeeraba water treatment plant (WTP), in Southeast Queensland, Australia, raw water is drawn from two different reservoirs: (1) Little Nerang dam (LND; maximum capacity: 6,705 mL) by gravity, and (2) the upper intake of Hinze dam (HUI; 310,730 mL), which bounds Advancetown Lake, through three electric pumps. Figure 1 illustrates the size and location of the two reservoirs and the treatment plant.

Both lakes have a vertical profiling system (VPS) which can measure, among others, water temperature, pH, dissolved oxygen, turbidity, etc., for the whole water column every hour. The historical VPS data, as well as other WTP and lake data, provide a unique opportunity to correlate raw water quality with the dosage of treatment chemicals (e.g. lime, alum, chlorine, carbon dioxide, sodium hydroxide, polymers, etc.).

Previous attempts of treatment chemicals predictions (e.g. van Leeuwen et al. 2001; Maier et al. 2004; Abdullahi 2013) were limited to one or a few chemicals, and often based on jar tests, thus limiting their usefulness in terms of real-time deployment. The large amount of data available, as well as the successful recent development of several data-driven models for similar applications (e.g. Bertone
et al. 2015), made it feasible to create a comprehensive treatment costs prediction model. This will allow the WTP operators to avoid jar tests in case of sudden weather/water quality changes by relying on the model prediction and, importantly, to proactively select the optimal intake depth/lake in order to minimise the treatment costs.

If such prediction model can be developed, another issue to be addressed for an effective water treatment management is to make sure that the intake prediction is compatible with current storage levels constraints. In particular, whenever LND is selected as being the optimal reservoir, water treatment decision-makers will have to ensure that high withdrawal rates from LND would not lead to quick depletion of this small reservoir. This is especially important as LND, by not having associated pumping costs, is often kept as a backup reservoir in case of failure of the HUI pumping system. As a consequence, another goal of this project was to develop a medium term (6 weeks) dam level forecast for LND in order to provide the decision-makers with different depletion and spill risks given different withdrawal scenarios. Such long-term type of prediction models cannot eliminate the uncertainty, but they have the capacity to reduce it (Krzysztofowicz 2001) or quantify it; in fact, the accurate measurement of the hydro meteorological variables linked with water level predictions often comes with large ambiguity (Buyukyildiz et al. 2014). Such uncertainty must therefore be accounted for in the model; the deployment of techniques such as Monte Carlo simulations is deemed ideal for this kind of model, as their application for the assessment of parameter uncertainty in hydrologic models substantially increased in recent times with the advances in computing technology (Hassan et al. 2009).

The decision support system combining the intake optimisation model and the LND volume forecasting model will lead to considerable monetary benefits and optimised water treatment for the Mudgeeraba WTP.

MATERIAL AND METHODS

The Mudgeeraba WTP is the second largest drinking water treatment facility in the Gold Coast region, Queensland, Australia, as it can treat a maximum of 110 mL/day of raw
water (Rogers et al. 2008). This is withdrawn from HUI and LND, which are located about 3 and 8 kilometres west and southwest of Mudgeeraba, respectively.

Firstly, historical data were collected. Seqwater, the main bulk water supplier of the Southeast Queensland region and currently custodian of the Mudgeeraba WTP, provided historical data, from 2008 to 2014, for lake water quality (weekly manual samplings, \( n = 311 \)) and chemicals dosages and costs (daily, \( n = 2,187 \)). Also, raw water quality data (daily) were provided, as well as the amount of water withdrawn from LND and from HUI. Additionally, energy costs were also made available. Finally, data from VPSs were also provided. These remote sensing tools, through a number of probes, can provide hourly water quality data for the whole water column. These were installed in LND and HUI only recently (2014); however, the ultimate goal of this project would be to have an intake optimisation mode relying solely on VPS data, so that if there is a sudden water quality change (e.g. turbidity current), the operators can promptly change the intake according to the model’s updated predictions. At this stage however, the developed model relies on manual sampling data. For more details on available data, the reader can refer to Bertone et al. (2016).

The aforementioned data were then analysed using, among others, self-organising maps (Kohonen 1990; Mounce et al. 2014). Self-organising maps, a form of artificial neural networks, allow the user to quickly and visually detect correlations and interdependencies between multiple variables, thus they are very effective for data analysis of complex, multi-variable systems. As a confirmation, a number of correlations between water quality and chemicals were found, some of them unexpected from a simple treatment process point of view (e.g. alkalinity with coagulation aids such as polydadmac); this facilitated the second step, i.e. deeper regression analysis between the selected variables and model development (Bertone et al. 2016). Subsequently, based on historical correlations, each chemical dosage was predicted using a different model (Table 1). Data was divided between a training set, and a test set to check model performance: in some cases (e.g. alum or chlorine), through a data-driven approach by directly associating the dosage to some raw water quality variables; in other cases (e.g. lime, carbon dioxide) in order to correct alkalinity or pH through well-known relationships (e.g. Caldwell-Lawrence diagrams; Caldwell & Lawrence 1953). Dosages were then converted to costs according to the unitary cost of each chemical. Pumping costs were also accurately estimated based on daily flows from HUI (i.e. where the pumps are used).

For the second modelling activity, i.e. storage volume forecasting for LND, more historical data needed to be collected. These were inflow, spill, storage volume, WTP intake, and environmental flow. Data frequency ranged from daily to weekly, and some input data (e.g. rainfall) was available since 1926. However, the period where all the input variables were available and with only few missing data points was 1999–2015. For this project’s purpose, it was decided that weekly data were suitable (\( n = 814 \)) to achieve a

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Cost prediction model step-by-step analysis processing procedure (Bertone et al. 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>Parameter predicted</td>
</tr>
<tr>
<td>1</td>
<td>Lime</td>
</tr>
<tr>
<td>2</td>
<td>KMnO4</td>
</tr>
<tr>
<td>3</td>
<td>CO2</td>
</tr>
<tr>
<td>4</td>
<td>Alum</td>
</tr>
<tr>
<td>5</td>
<td>Polydadmac</td>
</tr>
<tr>
<td>6</td>
<td>NaClO</td>
</tr>
<tr>
<td>7</td>
<td>NaOH</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Total WT cost</td>
</tr>
<tr>
<td>9</td>
<td>Electricity cost</td>
</tr>
<tr>
<td>10</td>
<td>Total cost</td>
</tr>
</tbody>
</table>
credible compromise between accuracy and computational time. Based on weekly inflow, a threshold nonlinear regression model was able to predict the expected variation in LND volume with good accuracy (Bertone et al. 2017). The real variation is obtained by subtracting the expected outflows (WTP intake, environmental flow, and predicted spills). Equation (1) represents the results of the two different models developed for low and high inflows.

\[
\begin{align*}
\text{LNDVol}_{\text{fin}} &= \begin{cases} 
\text{LNDVol}_{\text{in}} + 11.15 \cdot \text{Infl}^{0.4494} - \text{Sp} - \text{enfl} - \text{WTPint} & \text{Infl} \leq 2000 \\
\text{LNDVol}_{\text{in}} + 6 \cdot 10^{-5} \cdot \text{Infl}^2 - 0.11 \cdot \text{Infl} - \text{enfl} - \text{WTPint} + 185.7 & \text{Infl} > 2000
\end{cases}
\end{align*}
\] (1)

where:

- \( \text{LNDVol}_{\text{fin}} \) = Final volume of LND (at the end of the week) [mL]
- \( \text{LNDVol}_{\text{in}} \) = Initial volume of LND (at the start of the week) [mL]
- \( \text{Sp} \) = Spill amount from the top of the dam wall [mL/week]
- \( \text{enfl} \) = Environmental flow [mL/week]
- \( \text{Infl} \) = Inflow from Little Nerang Creek [mL/week]
- \( \text{WTPint} \) = Raw water from LND drawn to the Mudgeeraba WTP [mL/week. Maximum withdrawal rate: 55 mL/day].

The threshold (2,000 mL) is the result of an optimisation process in which the value was changed until the best overall model accuracy (quantified with \( R^2 \)) was reached. Although evaporation is not directly quantified, the first equation shows that the weight of the inflow parameters is reduced, by applying an exponent lower than 1. This can be seen as an indirect contribution of evaporation. In case of high inflows, evaporation plays a minor role and in fact is not clearly accounted for in the equation. More details on model development and performance can be found in Bertone et al. (2017).

The threshold regression model is then used to forecast the LND volume 6 weeks ahead by using the Bureau of Meteorology (BoM) Seasonal Streamflow Forecast (SSF). These provide the likelihood of the river flow to be below, near or above the median value for the next three months. A Monte Carlo-based approach was used where \( n \) quasi-random values (\( n = 1,000,000 \)) of expected weekly inflows were extracted from an exponential distribution fitted to the historical inflow data probability density function. The exponential distribution with \( \lambda = 0.001274 \) was selected as it yielded the best fit among a number of different options (including generalised Pareto and generalised extreme value). The values are quasi-random, as the number of inflow values falling into different intervals (i.e. below/near/above median) is dictated by the SSF. For example, if the probability of low flow is 20%, then only 20% of the random numbers will be extracted from a region of the exponential distribution corresponding to low flow values. Subsequently, the quasi-random generated inflow is used as input for the threshold regression model. The volume is predicted one week ahead, with a quasi-random error (as it also follows within a probability density function fitted to the model’s errors distribution) applied to the result to account for uncertainty in the prediction. After the value is predicted, environmental flow and pre-selected withdrawal amount are subtracted; if the value is above the maximum dam volume, the final volume is set to be equal to the maximum dam volume and the difference accounts as spill. The model is re-applied six times to obtain the 6-week ahead prediction. No serial correlation between historical inflow values was found, so the random inflow values are independent from each other. It was estimated that a longer forecasting horizon would lead to a too high cumulative error; 6 weeks ahead at the same time provides enough planning time for the treatment decision-makers (Bertone et al. 2017).

RESULTS AND DISCUSSION

Figure 2 shows how the model could reliably (\( R^2 = 0.71 \)) predict the total (treatment and pumping) variable costs of treatment (Bertone et al. 2016). Since historical raw water quality for alternative depths/locations were available, by running the optimisation model over historical data it was found that monetary savings of about AUD$100,000 per year could have been achieved with an optimal raw water
source choice. This was mainly associated with an increased usage of LND which does not have pumping costs. It was in fact estimated that even in the case of much worse water quality, the extra pumping costs from HUI would imply higher total costs. Increased sludge disposal costs due to increased use of alum were also accounted for; an increased usage of LND was still recommended in most cases. This is made clear by Figure 3, which shows how total treatment costs almost doubled when HUI was selected and associated pumping costs were added.

WTP operators, whenever LND is selected to be the most appropriate choice, can establish how much raw water they can draw in the next month and a half based on the volume prediction model. The model was validated by running it over five different historical scenarios, with different combinations of initial dam level, withdrawal rate, and rainfall pattern. Results are displayed in Figures 4 and 5. It can be seen how the model prediction was very close to the real volume variation, apart from Scenario 2, where it underestimated the
increase in volume, with the real final volume outside the confidence intervals set by 1 standard deviation (SD). In defence of the model, this was an atypical scenario, with a very dry 6-week period, and thus such an increase in dam level was largely unexpected. More details about model prediction and validation can be found in Bertone et al. (2011). In terms of the deployed probabilistic approach and confidence with the predictions, it can be seen that for certain, predictably wet scenarios (e.g. 1 and 4), the model confidently forecasted high volume and spill risk, with SD usually below 50 mL. The uncertainty grew in less predictable, drier scenarios (e.g. 2, 3 and 5) where SD increased to up to 150 mL, resulting in a flatter distribution.

In order to facilitate the deployment of the LND volume prediction model, a graphical user interface (GUI) was developed (Figure 6). WTP operators, in order to obtain a prediction, only have to enter the current month and dam level, while a link is provided in order to get the SSF. Subsequently, they can run a number of simulations based on different hypothesised withdrawal amounts, until they find the amount which makes them most comfortable given calculated depletion and spill risks.

An example is provided in Figure 7. Two pie charts help summarise and interpret the probability curve. One provides the risk of high and medium spill, while the other one tells the probability of low, medium and high volumes. The threshold for ‘low’ was selected based on hydraulic calculations (below which it is not possible to draw water; see Hamilton 2015); all the other thresholds were decided in accordance to the operators’ indications.

**CONCLUSIONS**

A comprehensive optimisation model in support of drinking water treatment operations was developed. Firstly, it enables to select the optimal reservoir and intake depth to draw raw water from, based on water quality and electricity costs. This assessment also includes the estimation of the dosage of
chemicals required. Then, it quantifies the risk of depletion and spill from the smaller LND reservoir, for a thorough risk assessment and management of different withdrawal scenarios. Given the historical conservative withdrawal approach from LND, kept as a backup source despite spilling regularly, this tool enables an increased, but safe, withdrawal from LND, which has the potential of leading to substantial monetary savings as well as a more proactive water treatment management.

To the authors’ knowledge, although prediction models were developed in the past that can estimate, for example, a specific chemical dosage (e.g. van Leeuwen et al. 2001; Maier et al. 2004), few of them have been able to estimate the total treatment costs, and typically these are not comprehensive as they do not include all the chemicals (e.g. Abdullahi 2013). Similarly, although water level forecasting models have been extensively studied and developed for both the very short term (i.e. days, e.g. Kisi et al. 2012; Afiq et al. 2013) and long term (i.e. years, e.g. Privalsky 1992; simulations over decades, e.g. Bertone et al. 2014) prediction, very few have focused on a forecasting horizon of 1–2 months, and such hybrid approach used for this research can be considered novel for this specific application (Bertone et al. 2017). Most importantly, there is no evidence of previous research combining such two different types of models in order to provide guidance to WTP operators for optimum raw water intake selection.

Although, at this stage, the accuracy of the treatment chemicals prediction model component might not be enough to eliminate the traditional, reliable jar tests routine, the model is still accurate enough to be used in case of sudden weather/water quality variations; importantly, it was also crucial to identify the potential for reducing costs through enhanced usage of LND (Bertone et al. 2016). Similarly, although the water level fluctuation model cannot be as accurate as a short-term model, for example flood risk prediction, it still reliably assists the end-users for medium-term operational decision.

Even though the model accuracy might change based on historical data available, such methodology can be applied to any drinking WTP using the same treatment chemicals, and to any dam with similar operational configuration.

Future work will focus on refining the models based on new data (especially VPS data) and updated BoM SSF.
models, as well as on developing a mobile App version of the GUI, in order to increase the engagement with the end-users and in turn the deployment of the outputs of this research project.

REFERENCES


Bertone, E., Stewart, R. A., Zhang, H. & O’Halloran, K. 2015 An autonomous decision support system for manganese...


