Data-driven recursive input–output multivariate statistical forecasting model: case of DO concentration prediction in Advancetown Lake, Australia

Edoardo Bertone, Rodney A. Stewart, Hong Zhang and Cameron Veal

ABSTRACT

A regression model integrating data pre-processing and transformation, input selection techniques and a data-driven statistical model, facilitated accurate 7 day ahead time series forecasting of selected water quality parameters. A core feature of the modelling approach is a novel recursive input–output algorithm. The herein described model development procedure was applied to the case of a 7 day ahead dissolved oxygen (DO) concentration forecast for the upper hypolimnion of Advancetown Lake, Queensland, Australia. The DO was predicted with an $R^2 > 0.8$ and a normalised root mean squared error of 14.9% on a validation data set by using 10 inputs related to water temperature or pH. A key feature of the model is that it can handle nonlinear correlations, which was essential for this environmental forecasting problem. The pre-processing of the data revealed some relevant inputs that had only 6 days' lag, and as a consequence, those predictors were in-turn forecasted 1 day ahead using the same procedure. In this way, the targeted prediction horizon (i.e. 7 days) was preserved. The implemented approach can be applied to a wide range of time-series forecasting problems in the complex hydro-environment research area. The reliable DO forecasting tool can be used by reservoir operators to achieve more proactive and reliable water treatment management.

Key words | data processing, dissolved oxygen, input algorithm, time-series forecasting, water resources management

Edoardo Bertone
Rodney A. Stewart (corresponding author)
Hong Zhang
Griffith School of Engineering,
Griffith University,
Gold Coast Campus,
Brisbane,
Queensland 4222,
Australia
E-mail: *r.stewart@griffith.edu.au*

Cameron Veal
Seqwater,
Brisbane,
Queensland 4002,
Australia

INTRODUCTION

Dissolved oxygen cycle in drinking water reservoirs

Oxygen is an essential element for most forms of life; specifically, the oxygen dissolved in a body of water (dissolved oxygen (DO)) influences the majority of the biogeochemical processes occurring in it. The solubility of oxygen depends partly on the water temperature, with higher solubility in colder waters. Typically, subtropical lakes and reservoirs such as Advancetown Lake, which was the case study location for this study, are thermally stratified for most of the year, forming a top layer of warmer, less dense waters called the epilimnion, and a colder, denser bottom called the hypolimnion. The transition layer, where the highest temperature gradient is recorded, is called the metalimnion.

Oxygen production occurs in the epilimnion, where solar radiation enables photosynthesis to occur. Oxygen consumption, on the other hand, is much higher in the hypolimnion, where the DO level is markedly lower than in the epilimnion, despite its higher solubility due to colder waters (Tundisi & Matsumura 2011). In the hypolimnion light penetration is limited and photosynthesis, in turn, is limited too, due to the fact that the algae can hardly develop, even though they can move between layers using internal buoyancy. This implies a lower pH (because of high CO_2 and composite forms, not dissociated because of the absence of photosynthesis) and no production of

doi: 10.2166/hydro.2015.131

oxygen. Moreover, DO is completely consumed (in the case of productive lakes) by the microbial oxidation of organic matter (Bertoni 2011), with a rate of depletion depending on the volume of the reservoir and the amount of organic matter present for metabolism (Macdonald 1995).

As a consequence, during the stratification season, hypolimnetic waters usually record a gradual depletion of available DO; in cases of eutrophic lakes, this leads to anoxia, with important consequences including additional challenges to water treatment. Previous research reports that DO plays an important role for the release of bottom sediments, which occurs during stratification periods when the DO drops below critical levels; for example, Chiswell & Huang (2003) reported a remarkable release of manganese from the bottom sediments for DO levels below 1.5 mg/L. Other studies also found similar relationships between increases in manganese and DO depletion at the sediment–water interface (Delfino & Lee 1971). During winter, when epilimnetic waters cool down and sink and a full lake circulation occurs, anoxic, nutrient-rich waters can reach the top layer, leading to high nutrient loads throughout the whole water column. Since nutrients must be removed before they reach consumers' taps, it is essential for water treatment organisations to understand the storages dynamic DO cycle and its implications for other element cycles. The purpose of this research was to predict 7 days ahead DO concentrations in the upper hypolimnion of Advancetown Lake, which is located in south-east Queensland, Australia.

Review of DO prediction models

There are a number of previous studies that seek to predict DO concentrations in lakes and reservoirs. Akkoyunlu *et al.* (2011) applied artificial neural networks (ANN) to estimate DO concentrations in a Turkish lake with respect to depth rather than time. Jayaweera & Asaeda (1993) built a mathematical model for DO in lakes, but the model cannot predict future concentrations unless all future input values are known. Moreover, Rocha *et al.* (2009) applied multiple regression models to predict DO and chlorophyll-*a* in 25 lakes of the Upper Paranà River floodplain. Their study found that pH, nitrate and lake area all had relevant positive correlations with DO values, while water temperature and electrical conductivity had negative correlations. Ranković *et al.* (2012) attempted to predict DO concentrations in reservoirs using adaptive network-based fuzzy inference systems with the optimal combination of model inputs selected based on final model performance. El-Shaarawi (1984) applied regression models to predict the DO depletion rate and probability of anoxia in the hypolimnion of the Central Basin of Lake Eire, noting how the main variables affecting these phenomena are water level and total phosphorus. Since the model was applied only during stratification seasons, the importance of water temperature (particularly the temperature differential between the top and bottom of the lake, which is a good detector of lake circulation) was only partially recognised. However, later studies on the same lake (Patterson *et al.* 1985) coupled a vertical mixing model with a DO budget model, thus recognising the importance of turbulent mixing in the distribution of DO throughout the water column.

All these models aimed to understand the relative importance of the possible inputs, but they did not aim to forecast future concentrations. Only a few research studies have tried to achieve such a goal: Xu & Liu (2013) predicted DO concentrations in a freshwater pond 1 hour ahead ($t = 1$) with the use of a wavelet ANN fed with the input at time $t = 0$. Another example is given by Coopersmith *et al.* (2011), who forecasted DO and probability of hypoxia in multiple points of a reservoir 1 day ahead. However, both of these studies had short-term prediction horizons, while the aim of this project is to enable a 7 days ahead DO forecast.

Review of hydrological forecasting models

To the authors' knowledge, there has been no attempt to build a DO forecasting model with prediction horizons as long as 7 days. However, in the hydro-environmental field, medium- to long-term forecasting models have been created to predict a number of parameters. To achieve that, in recent decades multi-step-ahead (MS) techniques have been applied; these MS techniques can predict time series values several time steps into the future, and can be divided into direct (where a model predicts directly n steps ahead with n the prediction horizon) or recursive (where the output of a one-step-ahead model is given as an input to the same model to predict two steps ahead and so on until n is reached) variations. Direct models usually perform better (e.g., Ji *et al.* 2005), since in recursive models the performance decreases with an increase in the prediction horizon due to the accumulation of error.

An application in the hydrologic field is given by Cheng *et al.* (2008), who predicted 12 steps ahead the monthly discharge from a Chinese hydropower plant using an improved, ANN-based MS model. However, for both the Ji *et al.* (2005) and Cheng *et al.* (2008) models only a single dependent variable was used. A second example is provided by Zaldívar *et al.* (2000), who forecasted high waters at the Venice lagoon using both direct and recursive nonlinear neural networks; it was underlined how, for recursive univariate models, the error propagates and performance dramatically decreases, every time the forecasted output is given as a new input. Lekkas *et al.* (2001) tried to overcome the problem by using real-time error updating techniques. On the other hand, however, Bertone *et al.* (2014) recently developed a data-driven model to predict manganese concentrations in Advancetown Lake 7 days ahead, also using predicted inputs (i.e., water and air temperatures). Limiting the number of dependent variables in the model, especially those based on less reliable future forecasts (e.g., rain or wind), provided better performance than the more complex modelling approaches, such as physical, process-based models (Helfer *et al.* 2011; Bertone *et al.* 2014).

Persuasion for data-driven models for forecasting applications

Despite the more complex models (such as, often, process-based models) being accurate for real-time modelling since they include several inputs and processes, they usually perform poorly for medium-term forecasts (such as 7 days), since all the input variables must also be forecasted, adding a degree of uncertainty even before running the model itself. Moreover, several studies have been conducted where data-driven models have been successfully applied to estimate or forecast hydrological variables. Such data-driven models often incorporate pre-processing algorithms or optimisation techniques in order to enhance their reliability and performance. Among the models, neural networks are widely used, for instance for algal blooms forecasting (Recknagel *et al.* 2002) even though for the same problem, other models such as regression trees can be successfully applied (Jung *et al.* 2010). Regarding model optimisation, Abrahart *et al.* (1999) used an iterative learning approach to remove unimportant components, as well as optimisation strategies to make the neural network more efficient to run. Similarly, quicker simulations can be obtained by removing superfluous or redundant points in the input time series, with use of subtractive clustering methods or genetic algorithms (Sannasiraj *et al.* 2004; Doan *et al.* 2005). Other data pre-processing techniques include data transformation (Bowden *et al.* 2003; Joorabchi & Zhang 2007), which is particularly important for neural network models, and the use of singular spectrum analysis (Sivapragasam *et al.* 2001) in the case of input time series containing several repetitive values (such as rain, which is 0 for several time steps of a time series). Giustolisi & Savic (2006, 2009) recently proposed a new data-driven technique that is based on evolutionary polynomial regression (EPR), where genetic algorithms are used for estimating the best combination of parameter inputs (i.e., which inputs to include, and how to combine them together); it is also possible to use nonlinear transformations of the inputs, but not a combination of different transformations at the same time.

However, in cases where a sufficiently robust model needs to utilise a number of forecasted inputs (calculated by associated sub-models) as presented by Bertone *et al.* (2014), there are presently limited reported studies focussing on the hydrological field. Han *et al.* (2007) worked on a flood-forecasting model and noted how the inclusion of unlagged rain input data would have substantially increased the prediction accuracy, but no attempt to predict rainfall was completed by the authors. In other studies, Tsanis *et al.* (2008) and Joorabchi *et al.* (2009) attempted to forecast the groundwater level by including rainfall forecasts linked to historical seasonal averages. While this approach has some merit, since seasonal averages may be adequate for slower processes such as groundwater replenishment, such an approach would be inadequate where daily or weekly forecasts are needed since there is high rainfall variability within each season.

Applying a novel data-driven modelling approach for forecasting DO in reservoirs

The use of modelled, forecasted inputs proved to be a promising inclusion to the current model as, unlike the traditional MS recursive techniques, the error propagation is kept limited due to the following two main reasons:

- The model is not univariate, but multivariate; therefore it does not rely on the previous output forecast only, but

mainly on other inputs which can be both forecasted or not. Also, it is not the output to be forecasted for the following time step, but only some inputs are predicted in case a lower lag would drastically increase the accuracy.

- There is no summation of n models' errors, where n is the number of time steps ahead required for prediction. There is only one direct model forecasting the inputs and one direct final model forecasting the required output; hence only two errors need to be summed thereby limiting the error propagation.

For the present study, since DO concentration in the hypolimnion has been shown to be affected by only a few main parameters and decreases approximately linearly throughout the stratification (Dobson & Gilbertson 1971; Charlton 1979; Anderson *et al.* 1984), a data-driven model was deemed appropriate for this forecasting problem. A 7 days ahead statistical model was applied, extending the methodology of Bertone *et al.* (2014). This data-driven approach includes algorithms for data pre-processing, a variable input selection process, and variable input forecasts in those cases where the near real-time inputs do not yield a satisfactory performance. The inclusion of secondary input-forecasting sub-models, besides being novel in this particular formulation, follows the recommendation of Solomatine & Ostfeld (2008) who emphasised the importance of hybrid models and in particular the use of sub-models to better identify sub-processes. An advantage of the proposed forecasting approach is that it is completely automated, thus not requiring human intervention (e.g., visual inspection of correlations as in previous studies such as Lees (2000)). Moreover, unlike EPR, different linear and nonlinear input transformations are considered during the same model run, thus allowing a more comprehensive investigation of potential outputs to be conducted in a relatively short amount of time.

METHODS

Research domain

The location of this study is Advancetown Lake (153.28 °E, 28.06 °S), also known as Hinze Dam (Figure 1). It is located

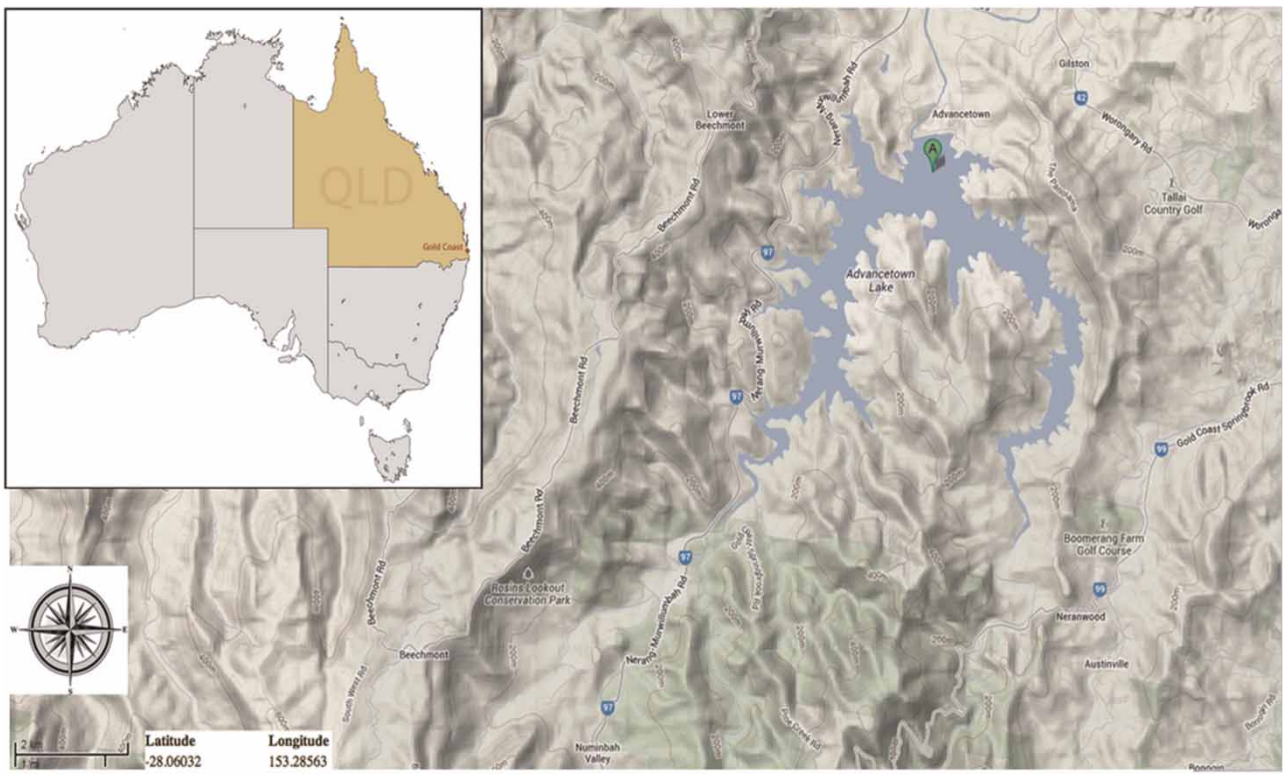

Figure 1 | Advancetown Lake map; A is the location of the vertical profiling system (VPS).

in South East Queensland (SEQ), Australia. Its current capacity is 310,730 ML, after a third upgrade of the dam was completed in 2011, raising the dam wall a further 15 m. This reservoir services the City of Gold Coast, which has a population greater than 500,000. Moreover, it is connected to the SEQ Water Grid, a water supply scheme covering all of SEQ, designed to respond to drought seasons by redirecting water from regions with an oversupply to regions in need (Spiller 2008).

The lake's average depth at full capacity is 32 m, while its surface area is approximately 15 km^2. The catchment area, mostly within national parks, covers 207 km^2 of terrain; the two main inflows are the Nerang River and the Little Nerang Creek, the latter also being the outflow from another dam, Little Nerang Dam, the oldest reservoir servicing the Gold Coast region. An intake tower, located next to the Hinze Dam wall, draws the water from the most suitable depth based on reservoir water quality at the time of extraction and redirects it to the closest treatment plant in Molendinar, which is approximately 10 km north-east of the reservoir. The reservoir treatment operations are managed by Seqwater, which is a Queensland Government-owned authority, which is responsible for bulk water supply for the SEQ region.

Data collection

Currently, water quality is primarily monitored though laboratory analysis of water samples collected manually on a weekly basis. However, in 2008 a vertical profiling system (VPS) was installed near the intake tower (point A in Figure 1). This VPS consists of a YSI Sonde suspended by a cable to a floating buoy which is automatically winched up and down through the water column and collects water quality parameters every metre including: water temperature, DO, pH, specific conductivity, redox potential and turbidity. Collected data for the whole profile are transmitted via telemetry every 2 hours back to a central data repository for collation, analysis and display. Through an effective collaboration with Seqwater, the data from the VPS and from the manual water samplings were made available for this research project. Data from the manual water samplings, which are more comprehensive, exist for a longer period of time (i.e., 2000–2013), but are limited to a depth of 24 m, collected at 3 m intervals. VPS data typically were collected for the whole water column, but cannot measure the complete range of parameters relevant to this study. Hence, for the sake of consistency, the data used in this study come from the period 2008–2013 from depths of 0–24 m. In addition, weather data for the same period were collected from the Australian Bureau of Meteorology and from a small weather station installed on the VPS buoy. Finally, the river inflow data were obtained from the Department of Energy and Resources Management of the Queensland Government.

Model development

To predict DO concentrations in the upper hypolimnion of Advancetown Lake, a number of techniques for data pre-processing, variable input selection, data post-processing, and a multiple regression model, were combined to create the final forecasting model named ALMO (pre-processing ALgorithms + regression Models). Although ALMO was specifically applied for DO prediction in Advancetown Lake in this present paper, it can be applied for the prediction of other independent variables at any predetermined number of time steps ahead. The logical flow chart of ALMO is shown in Figure 2.

The DO forecast model's input parameters are the relevant data, such as pH or water temperature, which can be defined as $x_i = \{x_i(t_{j-l}) \ldots, x_i(t_{j-2}), x_i(t_{j-1}), x_i(t_j)\}$, where $i = 1, 2, 3 \ldots$ is the index for various parameters, t_j represents the current time, and l is the maximum lag to be considered in the inputs, according to the knowledge the researcher has of the system and the computational time that can be accepted. The higher the l chosen, the bigger the input matrix will be and the longer the simulation; however, for some particular prediction problems, where it is known that there might be a high correlation between the independent variable and a certain dependent input variable at a high lag, it can be useful to select a high l. To further reduce computational time, a preliminary analysis of the available data is recommended in order to exclude those variables that are not correlated with the target vector.

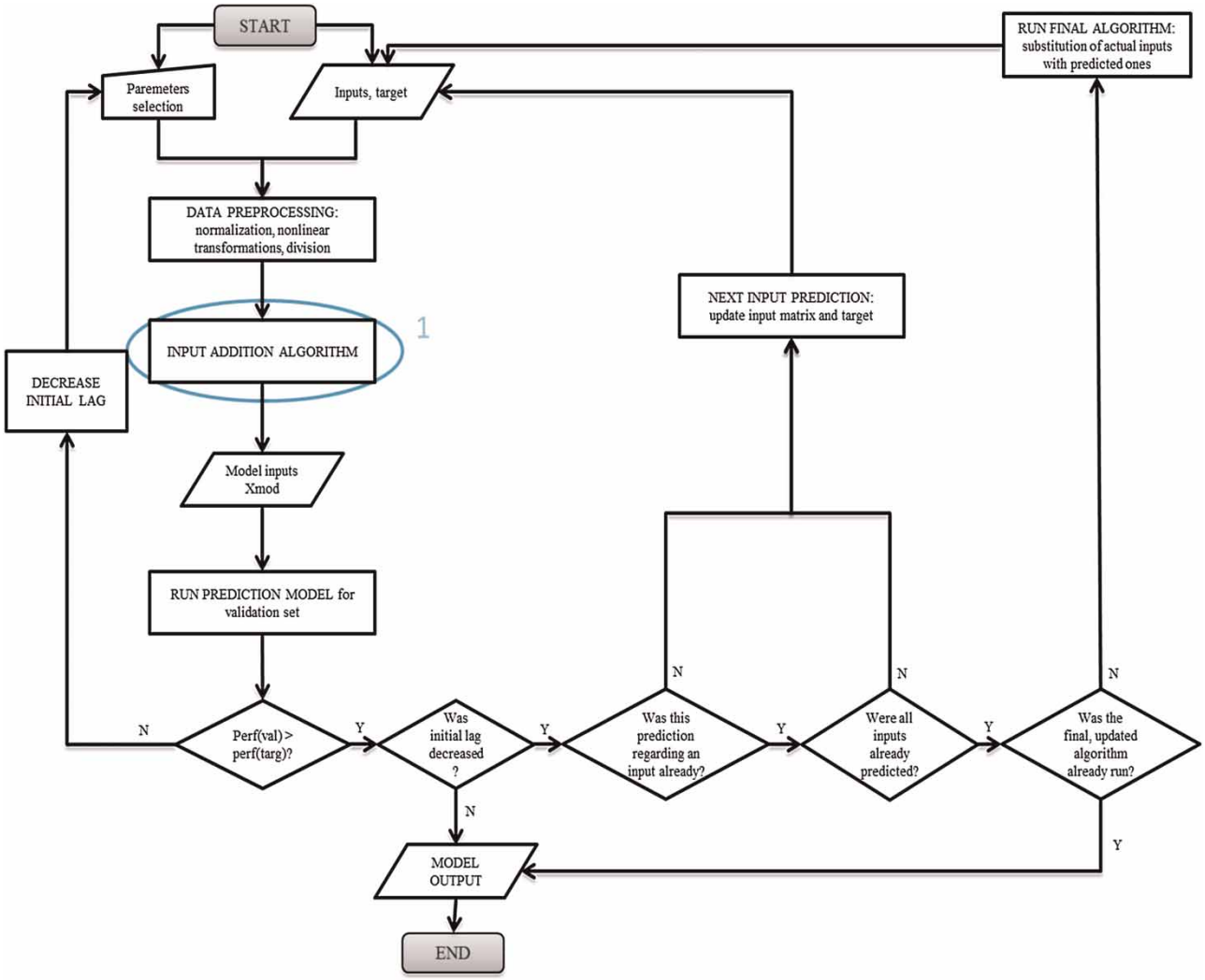

Figure 2 | ALMO flow chart.

The target or independent variable to be predicted (i.e., DO in this model) can be defined as:

$$y_O = \{y(t_{j+1}), y(t_{j+2}), \ldots, y(t_{j+n})\} \quad (1)$$

where n is the number of time steps ahead of the forecast (i.e., the prediction horizon). The desired statistical performance of the forecasting model is fixed by the user. After this, the inputs are pre-processed, including data normalisation and division between a calibration and a validation set. Normalising the data within a certain range (e.g., between 0 and 1) provides the opportunity to assign the same importance to each input, rather than overestimating the importance of inputs with a higher magnitude and underestimating inputs with a lower magnitude. The elements of the input variable vectors $\boldsymbol{x_i}$ can be normalised as:

$$x_i{*}(t_j) = \frac{x_i(t_j) - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (2)$$

where t_j represents any time step of the input time series and $\max(x_i)$ and $\min(x_i)$ are, respectively, the maximum and minimum values of the time series x_i considered.

To account for nonlinear correlations, a number of nonlinear transformations are applied to the inputs. These

created time series will update the input matrix. Although a number of default nonlinear transformations (e.g., hyperbolic, logarithmic, exponential, etc.) are predetermined, the user can modify or add more nonlinear transformations according to the results of the data analysis.

The normalised input $x_i*(t_j)$ is transformed to $u_i(t_j)$, $v_i(t_j)$, $w_i(t_j)$ using, among others, the following nonlinear transformation equations:

$$u_i(t_j) = e^{x_i*(t_j)} \tag{3}$$

$$v_i(t_j) = \log\left(x_i*(t_j)\right) \tag{4}$$

$$w_i(t_j) = \frac{1}{\left(1 + x_i*(t_j)\right)^3} \tag{5}$$

The processed input matrix after the nonlinear transformations will include the original time series as well as all the nonlinear elaborations:

$$X = [X_1, X_2,\ X_i\ ,\ \ldots] \tag{6}$$

where:

$$X_i = [x_i*\ u_i\ v_i\ w_i \ldots\ ] \tag{7}$$

represents the original and nonlinear transformation time series for each normalised input.

The final aspect of the pre-processing of the data is the division of the inputs and target time series into a calibration and a validation set. The percentages of the data in the calibration and validation sets are decided by the researcher; this procedure allows for testing of the prediction performance on an independent data set in order to avoid problems such as over-fitting (i.e., the model fitting the calibration set very well, but performing much more poorly when new data are presented).

Subsequently, an iterative input selection algorithm is applied (circle 1 in Figure 2), following previous studies by Castelletti *et al.* (2011). This is based on estimating the relative contribution of each input candidate in predicting the target y through the application of regression models. To do so, the normalised root mean squared error (NRMSE) is calculated using the following formula:

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{i=1}^{m} \frac{(y_{Oi} - y_{Mi})^2}{y_{Oi}}}}{m} \tag{8}$$

where m is the number of time steps included in the calibration set, y_O is the target time series for the calibration set and y_M is the time series of the output of the performed regression analysis.

The inputs are ranked according to their performance; the best performing input (i.e., the one yielding the least NRMSE, called $\mathrm{NRSME_{min}}$) is selected and stored, and the residuals $\varepsilon_{\min} = y_{Oi} - y_{Mi}$ are calculated. If the target NRMSE is not achieved, a new iteration of the algorithm is then performed, by estimating which input among the remaining candidates best fits the residuals through the use of the same predefined regression model. The iterative process is repeated until the target performance, defined by a certain predetermined value of NRSME, is reached, and thus the set of selected input variables is defined and stored in a matrix called X_{mod}. One of the advantages of this procedure and in particular of the application of the residuals as the next target is avoiding multicollinearity; in fact, once one input has been selected, all the variables highly correlated with that input will become useless when the next residuals will be targeted, and thus the rankings will have to be reconsidered. The input addition algorithm logical steps are illustrated in Figure 3. According to the inputs that have been selected, the equivalent of X_{mod} for the validation set is prepared by another algorithm.

When the iterative input selection procedure is completed, the model's forecasting performance is then assessed using the validation data set. Model users can decide whether to use NRMSE as their performance index, or to also calculate R^2. For this specific case study, an $R^2 \geq 0.8$ was selected as a target performance; this standard of performance was chosen because in environmental systems, the level of uncertainty is often high. In the case where the target statistical performance level cannot be achieved in the desired forecasting time frame (7 days ahead), the model will gradually reduce n in order to assess whether an acceptable performance can be achieved by reducing the prediction horizon (as can be sensibly expected, since a forecast closer in time will involve less uncertainty).

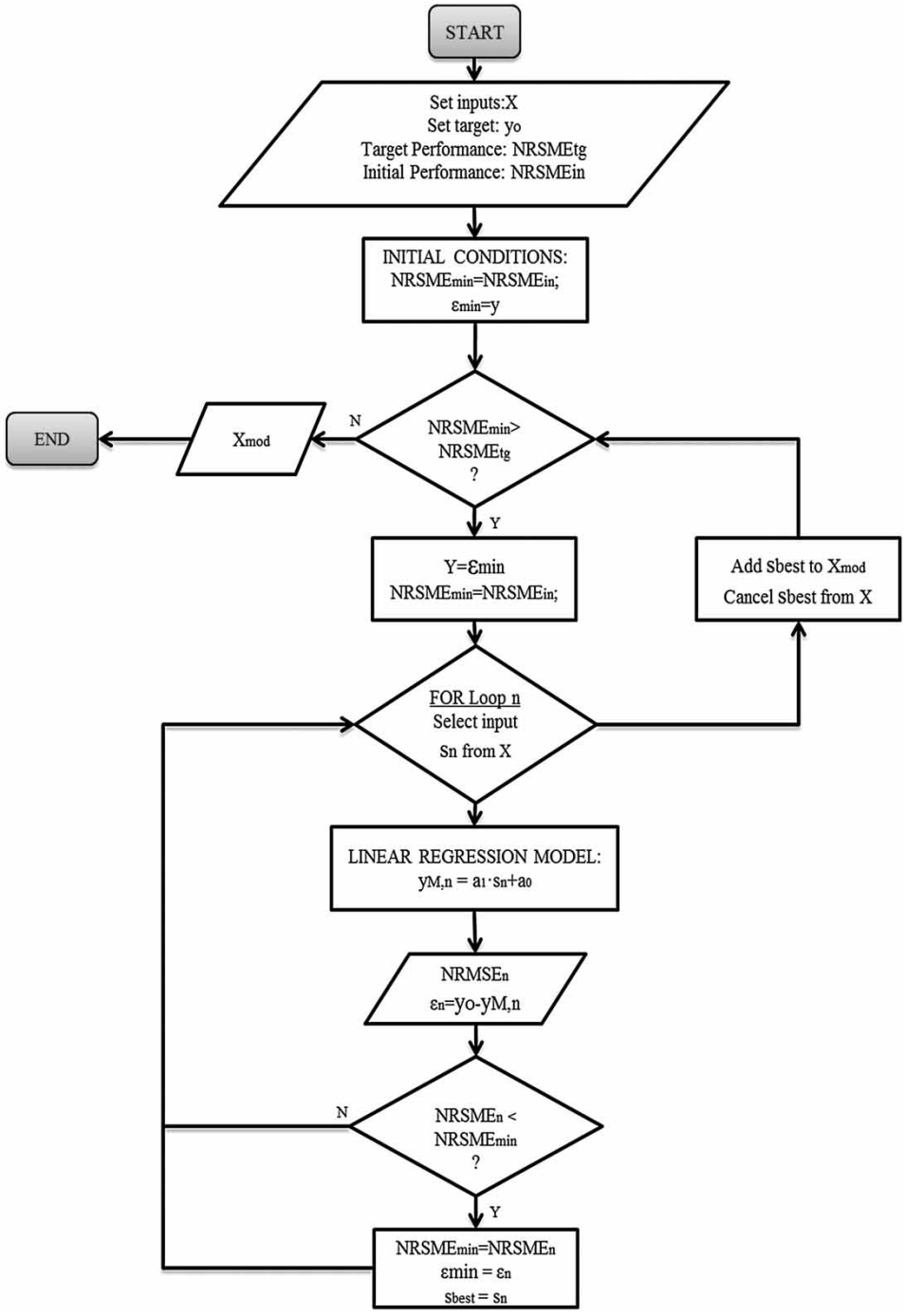

Figure 3 | Input selection algorithm flow chart (circle 1 in Figure 2).

Hence, n is iteratively reduced and the previously described procedure repeated, until the input selection algorithm yields an acceptable validation performance.

When the desired R^2 is reached, in order to guarantee that the prediction will still be valid as many time steps ahead as was originally decided despite the n reduction, the same forecast procedure will be applied to each input selected after the reduction of n, to show a delay from y_O which is lower than the originally established prediction horizon. The number of time steps ahead of each input

forecast will be equal to the number of n reductions before that input was found (e.g., if n was reduced twice before finding a certain relevant input, then this means that the selected input must be forecasted two steps ahead). The same input selection algorithm will be adopted. The outputs of this input forecast procedure will substitute the original associated inputs in X_{mod}. So if, for instance, we have to predict $y_O(t_{j+7})$, and ALMO found that the relevant inputs are $u_i(t_{j-1})$, $v_j(t_{j-3})$ and $w_k(t_{j+2})$, then obviously $w_k(t_{j+2})$ must also be calculated using inputs at time t or older; hence ALMO will then also be able to select the best inputs and run a prediction model for $w_k(t_{j+2})$, and the results will be used as one of the inputs for the prediction of $y_O(t_{j+7})$. Finally, the regression model is run again for the validation set by using the updated X_{mod}. The final model output and performance are displayed and saved.

In the present study, the model used for prediction was a multivariable linear regression model, since nonlinearities were already considered by the initial nonlinear input transformations, as in Equations (3)–(5). Castelletti *et al.* (2011) take nonlinearities into account through the application of a nonlinear (tree-based) model; however, since Bertone *et al.* (2014) showed that nonlinear models such as ANN may yield high volatility in the residuals, this could be detrimental in cases where forecasted variables must be used as inputs for the target prediction. In this case, the volatility would increase exponentially, since the inputs themselves will already show high volatility in their residuals, before even running the nonlinear model for the target forecast. Thus, following the outcomes of Bertone *et al.* (2014), a linear regression model, which typically yields smoother outputs, was chosen to be more appropriate for this research study. This model, if applied to time series forecasting, can be described by the following equation:

$$y_M(t_{j+n}) = a_0 + \sum_{i=1}^{l} a_{i,1}x_i* + \sum_{i=1}^{l} a_{i,2}u_i + \sum_{i=1}^{l} a_{i,3}v_i + \sum_{i=1}^{l} a_{i,4}w_i + \ldots + \varepsilon \quad (9)$$

where n is the prediction horizon and l is the maximum lag, $a_{i,k}$ are the regression coefficients for the k regressors, a_0 is the intercept and ε represents the error term. The model is linear because Equation (9) is a linear function of the unknown parameters a_i (Montgomery *et al.* 2012). The equation can be summarised as:

$$y = A \times X_{\mathrm{mod}} + \varepsilon \quad (10)$$

where A is a vector containing the unknown regression coefficients (with a_0 in the first row), X_{mod} is a matrix containing the input time series after the input selection procedure and a time series of 1 in the first column, and ε is the residuals' time series.

The method used to determine the model parameters is the method of least squares, which aims to determine A in order to minimise the sum of squares of the difference between the target values and the model outputs. To do so, the following equation will be solved:

$$A = (X_{\mathrm{mod}}{}'X_{\mathrm{mod}})^{-1}X_{\mathrm{mod}}{}'y \quad (11)$$

The target performances chosen had a NRMSE of 0.096 during the calibration process and a R^2 of 0.8 for the validation set. The calibration set counted for 70% of the data and the validation set for the remaining 30%. These proportions for data division have been widely applied in several previous studies (see Ismail *et al.* 2011), proving to be an efficient division for calibration and validation purposes. The chosen n was seven time steps, while l was decided to be three to contain the computational time. Since daily data were adopted, one time step corresponds to 1 day.

MODEL RESULTS

Raw data analysis and inputs pre-selection

In the present study, the above-presented ALMO is used to forecast DO, one of the most important water quality parameters for water storage dams. In lake dynamics analysis, the lake is always divided into three typical layers, the epilimnion (0–6 m in the case of this study), metalimnion (6–12 m) and upper hypolimnion (12–24 m). In Figures 4–6, the time series of the depth-averaged DO in the upper hypolimnion of Advancetown Lake is plotted along with other relevant time series. The described variables can be defined as follows:

$$X_{\mathrm{hyp}} = \int_{12}^{24} X(z)dz \quad (12)$$

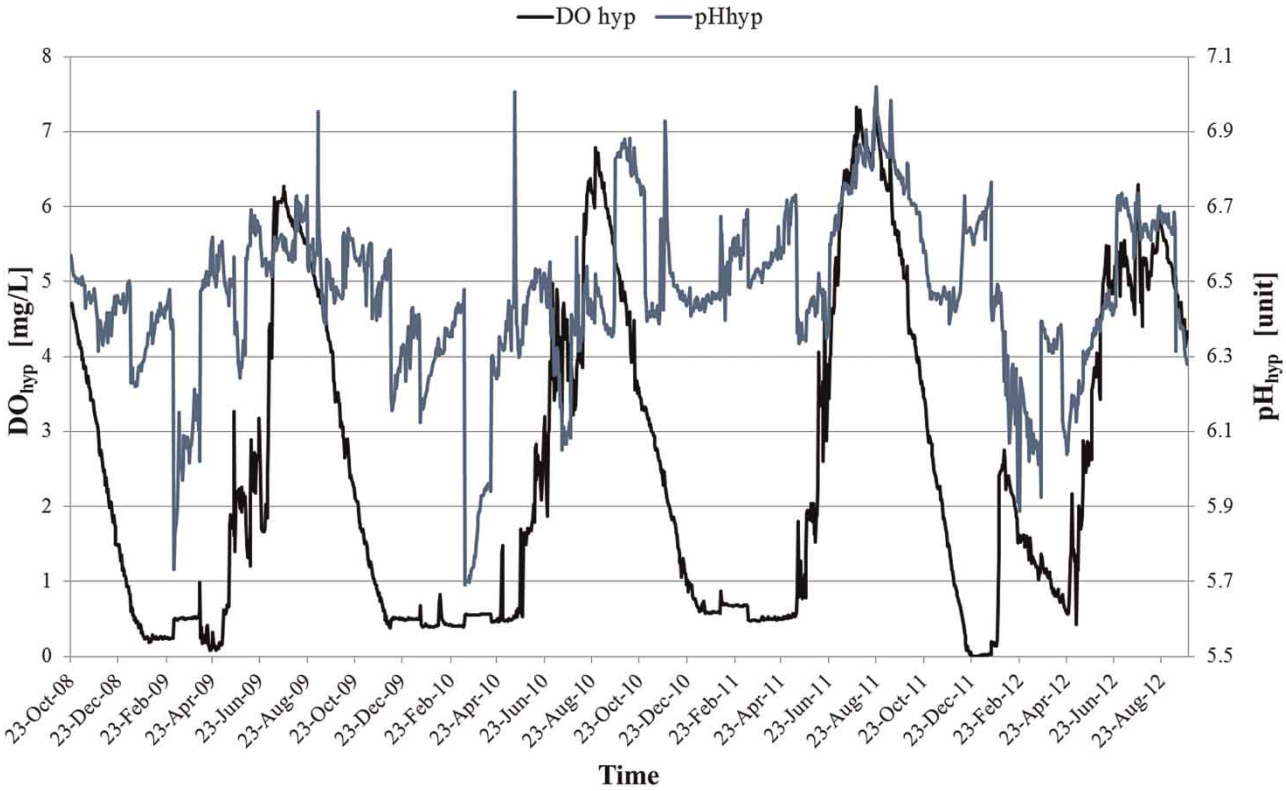

Figure 4 | DO_{hyp} and pH_{hyp} time series, Advancetown Lake, 2008–2012.

$$X_{\text{met}} = \int_{6}^{12} X(z)dz \tag{13}$$

$$X_{\text{ep}} = \int_{0}^{6} X(z)dz \tag{14}$$

where X represents the generic variable; subscripts hyp, met and ep stand for upper hypolimnion, metalimnion and epilimnion; and z represents the depth in metres. In this study, data were collected at every 3 m, so that the depth vector Z was used for discretisation:

$$Z = [z_1\ z_2\ z_3\ z_4\ z_5\ z_6\ z_7\ z_8] = [0\ 3\ 6\ 9\ 12\ 15\ 18\ 21\ 24] \tag{15}$$

The water column temperature differential between the upper hypolimnion and epilimnion layers was defined as follows:

$$\Delta T_w = T_{w_{\text{ep}}} - T_{w_{\text{hyp}}} \tag{16}$$

From Figure 4, the behaviour of the average DO concentration in the upper hypolimnion (12–24 m) for the target time series can be analysed. It can be noted that the depletion rate follows a linear pattern, and progresses from its maximum value occurring during stratification, when upper, well-oxygenated waters enter the hypolimnion, to around December, when the hypolimnion becomes anoxic and the increase in nutrients (such as manganese) can be noticed as soon as the DO level drops to about 2 mg/L (Figure 5). As expected, the pH behaviour, even if less pronounced in its extremes, follows the DO trends: during winter circulation, as the upper waters with higher pH enter the layer in question, the pH rises. Nevertheless, bacteria activity causes the pH to decrease when the circulation is over.

During the stratification season, the DO shows a linear, gradual decrease. However, at the onset of the circulation period, its level sharply increases. It was decided that a good variable describing the strength of the stratification and, in turn, the commencement of the lake circulation was the water column temperature differential ΔT_w between

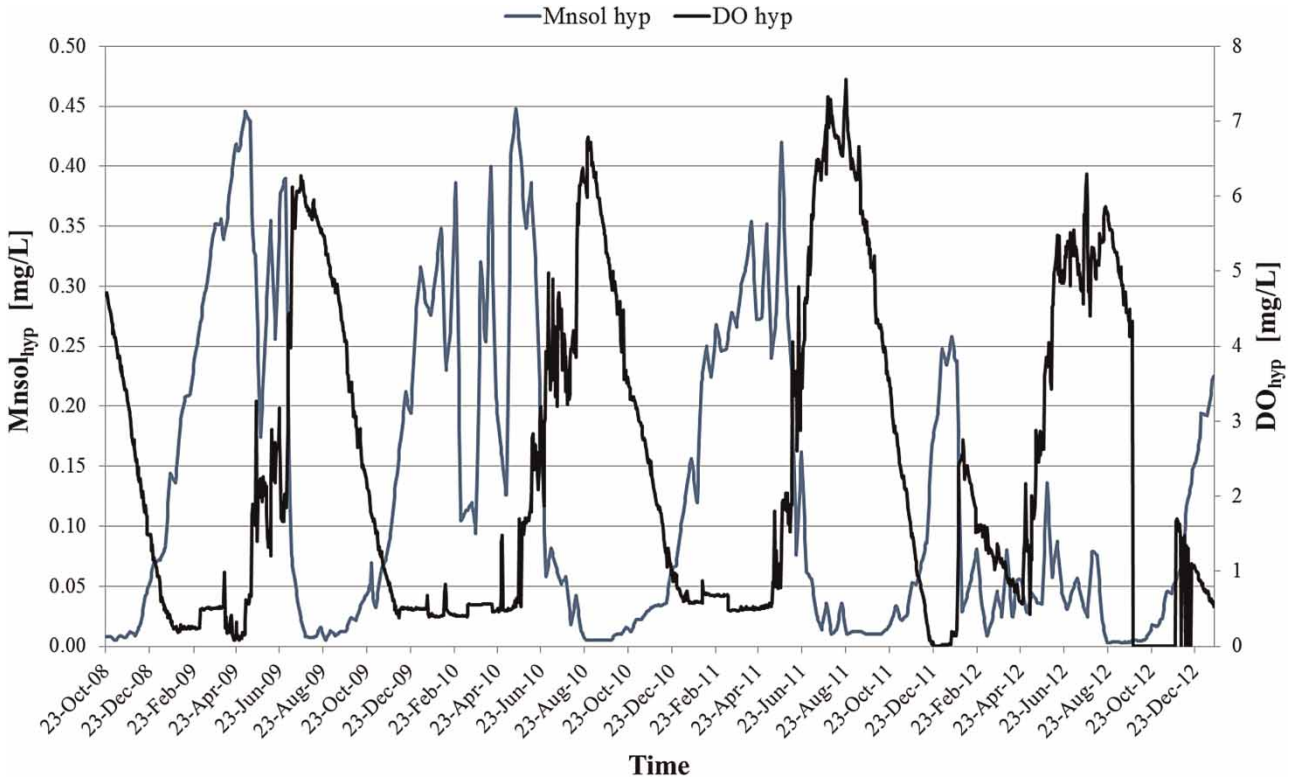

Figure 5 | DO_{hyp} and $Mnsol_{hyp}$ time series, Advancetown Lake, 2008–2012.

epilimnion and upper hypolimnion, as described in Equation (16). When ΔT_w is high, the stratification is strong and the DO level decreases until anoxia is reached. If ΔT_w is at approximately zero, meaning that the epilimnetic waters have similar temperature (and density) to the hypolimnion, then lake circulation is occurring and the DO concentration increases. Figure 6 clearly displays this inverse relationship, which, through proper linear and nonlinear transformations (e.g., hyperbolic), will be detected by the model.

All the other available time series were visually inspected. In order to reduce computational time, the most relevant ones were selected to be part of the model's input matrix (Table 1).

Prediction model outcomes

With all the model parameters determined and the input matrix prepared, ALMO was run in order to predict depth-averaged DO concentrations in the upper hypolimnion of Advancetown Lake. The model used 16 different transformations of the eight pre-selected DO predictor variables, creating a total of 128 input candidates time series, each counting 1,435 data point with the time step of a day (from October 2008 to October 2012). For model training and testing purposes, the 1,435 daily time steps were divided in a 70–30% ratio (i.e., 1,005/430) as described in the Model development section). Using Matlab R2012a software on an Intel®Core™ i5-2400 CPU @ 3.10 GHz processor, ALMO took approximately 16 minutes to perform the simulation, with most of the time (more than 15 minutes) spent over the calibration phase. Nine variables were selected as relevant inputs. The general equation describing the ALMO regression model outputs is:

$$\mathrm{DO}_{\mathrm{hyp}}(t_{j+7}) = \sum_{i=1}^{9} a_i s_i(t_i) + a_0 \tag{17}$$

where s_i are the time series vectors contained in X_{mod} created after ALMO input selection, a_i are the respective regression coefficients, and t_i the lags of the i relevant inputs s_i selected by ALMO.

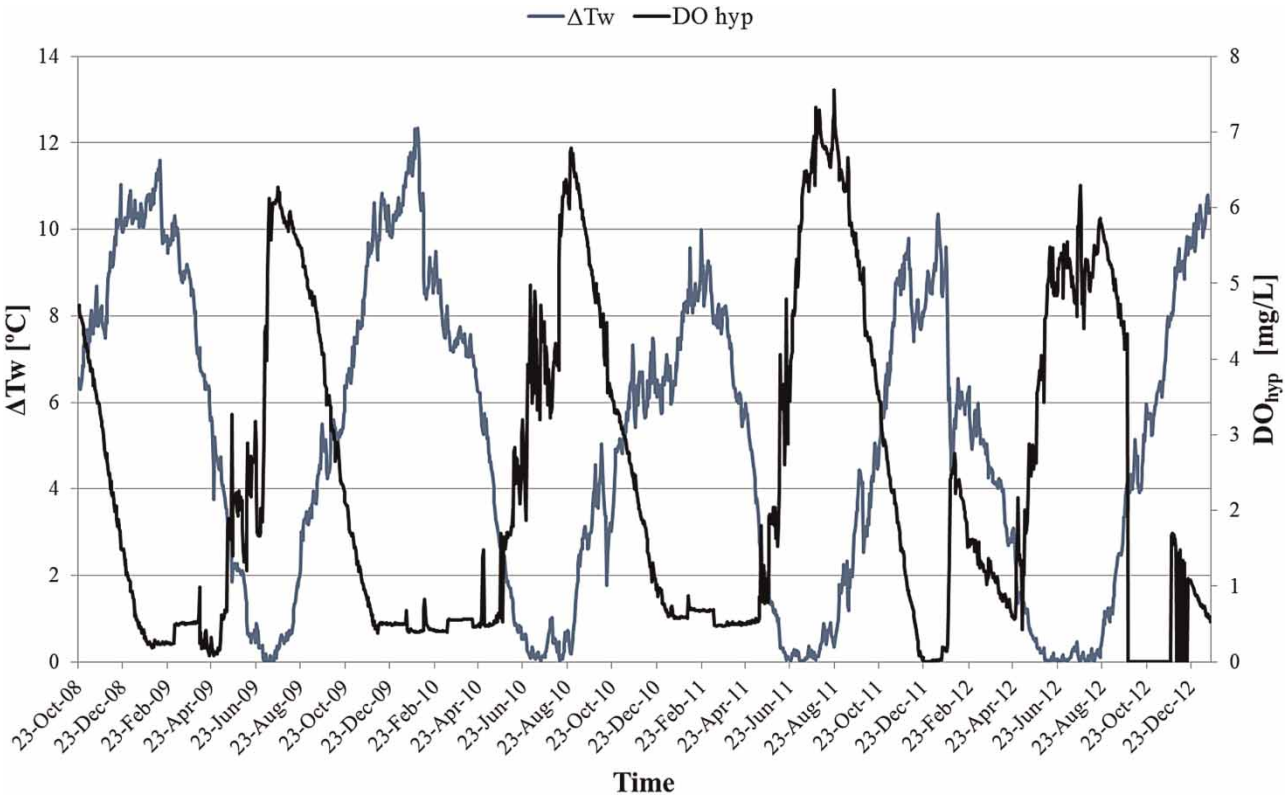

Figure 6 | DO_{hyp} and ΔT_w time series, Advancetown Lake, 2008–2012.

Table 1 | Input matrix

Input	Variable
X_1	ΔT_w
X_2	Air temperature
X_3	Water temperature hyp
X_4	pH hyp
X_5	pH met
X_6	Conductivity hyp
X_7	Redox potential hyp
X_8	Turbidity hyp

Table 2 gives a list of the final inputs s_i selected by the model with the respective lag t_i and regression coefficients a_i, described in Equation (17). The respective NRMSE reached when that input was added to the regression is also shown in Figure 7. The order reflects the ranks of the input, as can be noticed by observing the relative NRMSE reduction (Figure 7); the first input addition implies most of the performance achieved with the addition of s_2 producing an NMRSE reduction of 0.032, while the addition of s_9 only implies a reduction of 0.001, and from s_{10} the performance improvement was negligible. As a consequence, for this case study, it was decided to retain all the first nine inputs, as s_{10} did not imply any noticeable increase in performance; however, if the user was willing to accept only a slight reduction in forecasting accuracy, a simpler model could have been developed with just a few key variables.

It can be noted how only nonlinear transformations of the original inputs were considered relevant by the model, and this reflects well the uncertainty and lack of linearity of environmental systems such as lakes. The most important inputs (five of the first six), as expected from the preliminary visual inspection of the data, are related to the lake circulation, and altogether help in modelling the nonlinear spikes in DO concentrations at the beginning of the turnover event. The pH, in both the upper hypolimnion and metalimnion, was also a relevant input. Other variables such as specific conductivity, turbidity or redox potential did not help in improving the model performance.

Table 2 | Prediction model's final inputs

Input name	Variable	a_i	t_i	NRMSE
s_1	$(1+\Delta T_w)^{-5}$	6.131	t_{j-3}	0.152
s_2	$(1+Tw_{hyp})^{-5}$	4.244	t_{j-3}	0.110
s_3	pH_{hyp}^3	1.608	t_{j+1}	0.106
s_4	$(1+\Delta T_w)^{-5}$	−1.654	t_{j+1}	0.103
s_5	ΔTw^3	−0.540	t_{j-1}	0.101
s_6	$(1+T_{air})^{-5}$	−0.621	t_{j+1}	0.099
s_7	$(1+pH_{hyp})^{-5}$	0.994	t_{j-2}	0.098
s_8	$(1+pH_{met})^{-5}$	0.984	t_{j+1}	0.097
s_9	ΔTw^3	−1.160	t_{j+1}	0.096
s_{10}^a	$(1+T_{air})^{-5}$	−0.498	t_{j-3}	0.096

a_0 = intercept = −1.278.
[a] = first excluded variable, as it did not decrease NRMSE.

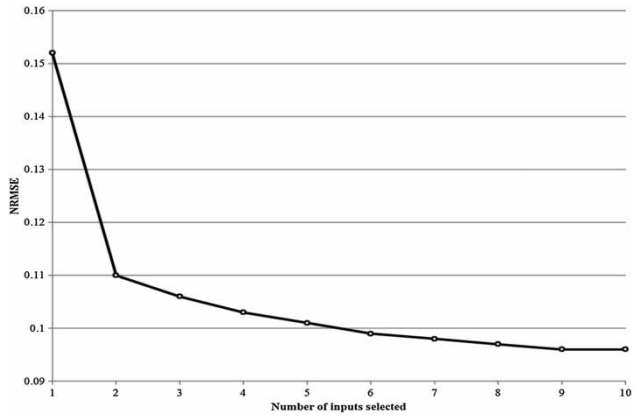

Figure 7 | NRMSE reduction after each input variable is added to the forecasting model.

Although the forecast must be performed for 7 days ($t+7$), the model had to use as inputs several variables with a reduced lag ($t+1$), thus implying the necessity for their forecast too. Hence, each of those variables was forecasted in turn (1 day ahead) and the results were substituted into the recorded time series in order to estimate the real final performance.

Figure 8 displays the model's forecasted DO concentrations in comparison with the measured DO values for a validation set (from August 2011 to October 2012, corresponding to the final 30% of the period considered). Validation yielded a respectable $R^2 = 0.804$ and $NRMSE = 0.149$ level of accuracy. Interestingly, the least performing segment is caused by an unpredicted rapid spikezin concentration at the end of January 2012. The discrepancy in model accuracy is made even clearer when assessing seasonal performances, as detailed in Table 3.

Figure 9 plots the model results against the observed values for the three different periods of time detailed in Table 3. DO was forecasted with high accuracy during the period where DO was decreasing. However, during the period of very low DO concentrations, there was a lack of any appreciable correlation due to the aforementioned spike. A consistent underestimation can also be noticed for low DO levels during the increasing DO concentration period, probably still related to the same unusual, sharp increase.

The unusual peak was caused by an extremely heavy precipitation event, which led to a partial mixing of the reservoir and the intrusion of oxygenated waters from the river and the epilimnion. This atypical event did not occur previously within the calibration data set; hence the model could not predict it. It is reasonable to predict or hypothesise that, over a validation set that does not include extreme weather events, the model performance would be even higher. Alternatively, if both river inflow and temperature data were available, it would be possible to use them, or a combination of them, as extra inputs for the model. However, a similar event would still have to be present in order to allow the model to learn to detect similar event occurrences in future validation data. In case such an event was included in the calibration data set, ALMO would have given more importance to meteorological variables such as rain at time $t+6$ or $t+7$, thus implying the difficult challenge of reliably forecasting rain 6–7 days ahead. This could compromise the accuracy of the final DO prediction, as rain, unlike, e.g., air temperature, is difficult to be accurately predicted. Nevertheless, as only very extreme wet weather events have the power to induce mixing and affect DO in Advancetown Lake, it might be necessary to predict only those extreme events (e.g., rainfall >100 mm) without having to correctly quantify its exact amount. Hence, the use of imperfect weather forecast potentially would not substantially jeopardise the final reliability of ALMO for DO prediction in Advancetown Lake.

Predictably, the output time series is seven time steps longer. Therefore, ALMO can be effectively used for real-time DO prediction 7 days ahead, given that all the required inputs are provided.

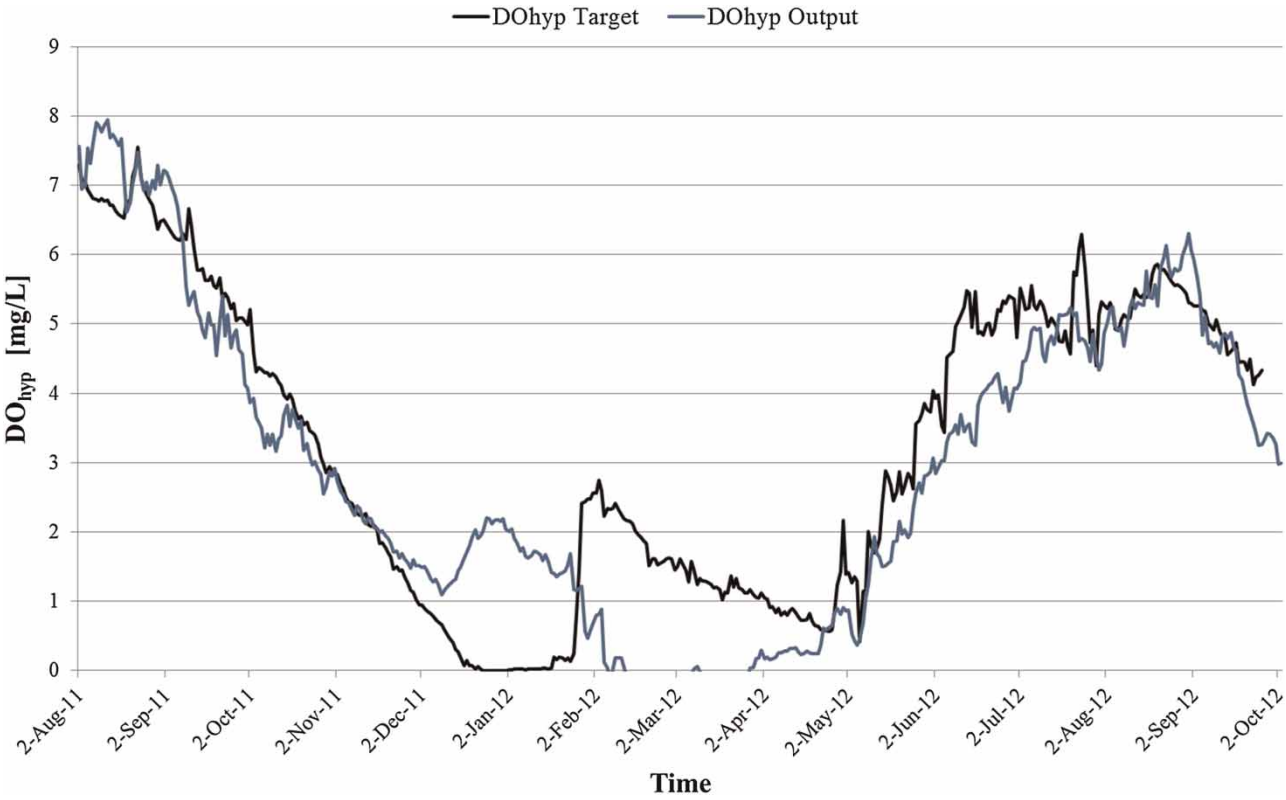

Figure 8 | DO_{hyp} real/forecasted time series for the Advancetown Lake validation data set.

Table 3 | Seasonal NRMSE

Period	Dates	NRMSE
Winter/Spring DO depletion	2 August 2011–2 December 2011; 20 August 2012–25 September 2012	0.071
Summer/Autumn hypoxia	2 December 2011–2 May 2012	0.197
Autumn/Winter oxygenation	2 May 2012–20 August 2012	0.115

DISCUSSION

The Model results section showed how ALMO was successfully applied for the DO forecasting case study. Forecasting DO in an intermediate layer of a reservoir is challenging due to the complexity of the mixing and biogeochemical reactions occurring in it. Yet, the ability to reliably estimate DO 7 days ahead is highly beneficial for the Advancetown Lake water operators, as it would enable proactive raw water offtake gate level selection to occur.

The installation of remote water quality monitoring instrumentation (e.g., VPS) in drinking water reservoirs is becoming more commonplace and will exponentially expand the quantum of lake and weather data available for analysis and decision-making. Owing to the high installation and running cost of these systems, there is increasing pressure from the water utilities for exploiting such databases for enhanced reservoir and treatment management purposes. The formulation and application of data-driven modelling approaches such as ALMO provide bulk water utility operators with a reliable forecasting tool that selects its own model predictor variables and sufficiently accounts for different types of nonlinearities that may be evident in the data. Moreover, its user-friendliness enables the modeller to adjust, for example, the target performance or the number of nonlinear transformations, in order to optimise computer processing time. Moreover, since ALMO is a

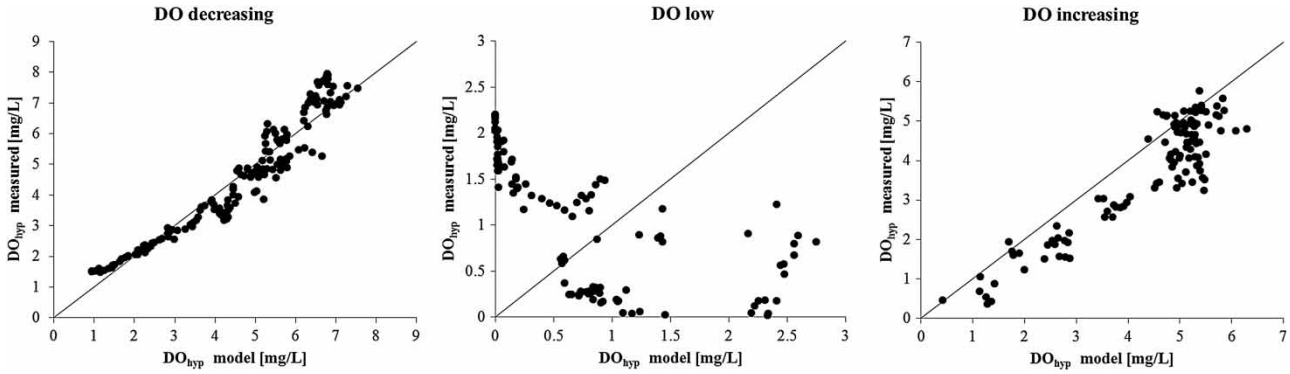

Figure 9 | Scatter plots of observed vs. forecasted DO values during the decreasing DO period, the low DO period and the increasing DO period.

purely data-driven approach, able to detect a number of linear or nonlinear correlations at different lags between any time series presented, it can be applied to forecast a wide range of independent variables for a number of feasible time step horizons. However, readers should be cautioned that the approach requires a basic knowledge of the environmental system being investigated as well as sufficient data availability in order for ALMO to derive accurate forecasts of the independent variable.

ALMO is a novel, practical and user-friendly approach for time series forecasting of independent water quality variables. After a careful review of contemporary modelling approaches in this field, a hybrid model was constructed comprising analytical techniques/approaches such as nonlinear data transformations, linear regression and recursive input–output forecasting modelling techniques. Other popular approaches were trialled to examine their fitness-for-purpose for this forecasting problem but exhibited deficiencies that prevented their inclusion; two examples include: (a) ANN models were not suitable for this forecasting application since sub-model outputs are used as inputs for another sub-model, with ANN outputs typically presenting higher variance and thus tending to increase the likelihood of error propagation, as argued in Bertone *et al.* (2014); (b) ALMO better handled the numerous data transformations needed for this problem when compared to the EPR. Overall, ALMO is a sensible addition to the current suite of environmental modelling techniques available, and has particular merit for similar applications to that presented herein (i.e., week ahead forecast of a particular independent variable using a remote lake monitoring system).

CONCLUSIONS

An innovative ensemble of data processing algorithms coupled with a prediction model (ALMO) has been built. Its main features are: takes into account nonlinearities; only relevant inputs are selected; the user can define the prediction horizon and the maximum lag, along with the target performance; if the target performance is not reached, the model automatically reduces the prediction horizon in order to reach a better correlation by using inputs closer in time with the output; and if and when the target performance is reached, then the model will forecast those added inputs too.

The formulated modelling procedure has far-reaching potential for application in a range of time series forecasting problems in different fields. In the case study presented here, the model effectively predicted DO concentrations a week ahead in Advancetown Lake with $R^2 > 0.8$. The high correlation was reached by decreasing the prediction horizon from 7 to 6 days and building a secondary 1 day ahead forecasting model for those new selected inputs. Hence, the indirect approach first introduced in Bertone *et al.* (2014) and formalised in a comprehensive manner herein represents a novel way to increase forecasting model performance. Furthermore, the introduction of nonlinear input transformations was essential for an accurate forecast,

as could have been expected for an environmental system. It should be stressed that user knowledge of the environmental system being investigated is essential to develop a prediction model using the herein described approach since the inclusion of the right inputs that are physically or chemically related to the output is essential for performing a reliable model simulation that avoids unjustified random high correlations.

Successfully predicting DO concentrations in a reservoir 1 week ahead is of paramount importance, since this parameter is highly correlated with critical nutrient cycles. The implemented data-driven forecasting tool can serve as an early warning system for Seqwater operators to more proactively address critical DO conditions. Future work will focus on the application of the model to different reservoirs and water parameters, on the reduction of the computational time when using large input matrices, and on the creation of a user-friendly interface that clearly displays outputs and warnings.

ACKNOWLEDGEMENTS

The authors are grateful to Griffith University for support for Mr Bertone and to Seqwater for their technical and financial support to this collaborative project.

REFERENCES

Abrahart, R. J., See, L. & Kneale, P. 1999 Using pruning algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. *J. Hydroinform.* **1** (2), 103–114.

Akkoyunlu, A., Altun, H. & Cigizoglu, H. 2011 Depth-integrated estimation of dissolved oxygen in a lake. *J. Environ. Eng.* **137** (10), 961–967.

Anderson, J. E., El-Shaarawi, A. H., Esterby, S. R. & Unny, T. E. 1984 Dissolved oxygen concentrations in Lake Erie (U.S.A.-Canada), 1. Study of spatial and temporal variability using cluster and regression analysis. *J. Hydrol.* **72**, 209–229.

Bertone, E., Stewart, R. A., Zhang, H. & O'Halloran, K. 2014 Intelligent data mining of vertical profiler readings to predict manganese concentrations in water reservoirs. *J. Water Supply Res.Technol–AQUA* **63** (7), 541–552.

Bertoni, R. 2011 Limnology of rivers and lakes. In: *Limnology. Encyclopedia of Life Support Systems (EOLSS)*. Developed under the auspices of the UNESCOEolss Publishers, Oxford, UK.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2003 Data transformation for neural networks models in water resources application. *J. Hydroinform.* **5** (4), 245–258.

Castelletti, A., Galelli, S., Restelli, M. & Soncini-Sessa, R. 2011 Tree-based variable selection for dimensionality reduction of large-scale control systems. *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Paris, France, 11–15 April*.

Charlton, M. N. 1979 *Hypolimnetic oxygen depletion in central Lake Erie: Has there been any change*? Scientific Series, no. 110, Inland Waters Directorate, National Water Research Institute, Canada Centre for Inland Waters, Burlington, Ontario.

Cheng, C., Xie, J., Chau, K. & Layeghifard, M. 2008 A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction. *J. Hydrol.* **361** (1–2), 118–130.

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

Coopersmith, E. J., Minsker, B. & Montagna, P. 2011 Understanding and forecasting hypoxia using machine learning algorithms. *J. Hydroinform.* **13** (1), 64–80.

Delfino, J. J. & Lee, G. F. 1971 Variation of manganese, dissolved oxygen and related chemical parameters in the bottom waters of Lake Mendota, Wisconsin. *Water Res.* **5** (12), 1207–1217.

Doan, C. D., Liong, S. Y. & Karunasinghe, D. S. K. 2005 Derivation of effective and efficient data set with subtractive clustering method and genetic algorithm. *J. Hydroinform.* **7** (4), 219–233.

Dobson, H. H. & Gilbertson, M. 1971 Oxygen depletion in the hypolimnion of the Central Basin of Lake Erie, 1929–1970. In: *Proc. 14th Conf. Great Lakes Res.*, Int. Assoc. Great Lakes Res., pp. 743–748.

EI-Shaarawi, A. H. 1984 Dissolved oxygen concentrations in Lake Erie (U.S.A.-Canada), 2. A statistical model for dissolved oxygen in the Central Basin of Lake Erie. *J. Hydrol.* **72**, 231–243.

Giustolisi, O. & Savic, D. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinform.* **8** (3), 207–222.

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPR-MOGA. *J. Hydroinform.* **11** (3–4), 225–236.

Han, D., Chan, L. & Zhu, N. 2007 Flood forecasting using support vector machines. *J. Hydroinform.* **9** (4), 267–276.

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

Ismail, M. J., Ibrahim, R. & Ismail, I. 2011 Development of neural network prediction model of energy consumption. *World Acad. Sci. Eng. Technol.* **5**, 10–27.

Jayaweera, M. & Asaeda, T. 1993 Modeling of dissolved oxygen in lakes. *Environ. Syst. Res.* **21**, 413–418.

Ji, Y., Hao, J., Reyhani, N. & Lendasse, A. 2005 Direct and recursive prediction of time series using mutual information selection. In: *8th International Work Conference on Artificial Neural Networks*, IWANN 2005, Barcelona, Spain, *8–10 June*, pp. 1010–1017.

Joorabchi, A. & Zhang, H. 2007 Application of artificial neural networks in flow discharge prediction for the Fitzroy River, Australia. *J. Coast. Res. SI* **50**, 287–291.

Joorabchi, A., Zhang, H. & Blumenstein, M. 2009 Application of artificial neural networks to groundwater dynamics in coastal aquifers. *J. Coast. Res. SI* **56**, 966–970.

Jung, N., Popescu, I., Kelderman, P., Solomatine, D. P. & Price, R. K. 2010 Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. *J. Hydroinform.* **12** (3), 272–274.

Lees, M. J. 2000 Data-based mechanistic modelling and forecasting of hydrological systems. *J. Hydroinform.* **2** (1), 15–34.

Lekkas, D. F., Imrie, C. E. & Lees, M. J. 2001 Improved non-linear transfer function and neural network methods of flow routing for real-time forecasting. *J. Hydroinform.* **3** (3), 153–164.

Macdonald, R. H. 1995 *Hypolimnetic withdrawal from a shallow eutrophic lake*. Doctoral Thesis, the University of British Columbia, Canada.

Montgomery, D. C., Peck, E. A. & Vining, G. G. 2012 *Introduction to Linear Regression Analysis*. 5th edn. Wiley series in Probability and Statistics. Wiley, Hoboken, NJ, p. 821.

Patterson, J. C., Allanson, B. R. & Ivey, G. N. 1985 A dissolved oxygen budget model for Lake Erie in summer. *Freshwater Biol.* **15** (6), 683–694.

Ranković, V., Radulović, J., Radojević, I., Ostojić, A. & Ĉomi, L. 2012 Prediction of dissolved oxygen in reservoirs using adaptive network-based fuzzy inference system. *J. Hydroinform.* **14** (1), 167–179.

Recknagel, F., Bobbin, J., Whigham, P. & Wilson, H. 2002 Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwaters lakes. *J. Hydroinform.* **4** (2), 125–133.

Rocha, R. R., Thomaz, S. M., Carvalho, P. & Gomes, L. C. 2009 Modeling chlorophyll-a and dissolved oxygen concentration in tropical floodplain lakes (Paraná River, Brazil). *Braz. J. Biol.* **69** (2), 491–500.

Sannasiraj, S. A., Zhang, H., Babovic, V. & Chan, E. S. 2004 Enhancing tidal prediction accuracy in a deterministic model using chaos theory. *Adv. Water Res.* **27**, 761–772.

Sivapragasam, C., Liong, S. & Pasha, M. F. K. 2001 Rainfall and runoff forecasting with SSA-SVM approach. *J. Hydroinform.* **3** (3), 141–152.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *J. Hydroinform.* **10** (1), 3–22.

Spiller, D. 2008 Water for today, water for tomorrow: establishment and operation of the SEQ water grid. *Australian Econ. Rev.* **41** (4), 420–427.

Tsanis, I. K., Coulibaly, P. & Daliakopoulos, I. N. 2008 Improving groundwater level forecasting with a feedforward neural network and linearly regressed projected precipitation. *J. Hydroinform.* **10** (4), 317–330.

Tundisi, J. S. & Matsumura, T. 2011 *Limnology*. Taylor and Francis (CRC Press), Boca Raton, Florida.

Xu, L. & Liu, S. 2013 Study of short-term water quality prediction model based on wavelet neural network. *Math. Computer Model.* **58** (3–4), 807–813.

Zaldívar, J. M., Gutiérrez, E., Galván, I. M., Strozzi, F. & Tomasin, A. 2000 Forecasting high waters at Venice Lagoon using chaotic time series analysis and nonlinear neural networks. *J. Hydroinform.* **2** (1), 61–84.

First received 8 December 2014; accepted in revised form 28 April 2015. Available online 6 June 2015