Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 3, 2803 doi: 10.2166/ws.2021.435

Integrating water quality and streamflow into prediction of chemical dosage in a drinking water treatment plant using machine learning algorithms

Hui Wang*, Tirusew Asefa and Jack Thornburgh

Tampa Bay Water, 2575 Enterprise Road, Clearwater, FL 33763, USA
*Corresponding author. E-mail: hwang@tampabaywater.org

ABSTRACT

Understanding the relationship between raw water quality and chemical dosage is especially important for drinking water treatment plants (DWTP) that have multiple water sources where the ratio of different supply sources could change with seasons or in a matter of weeks in response to changing hydrologic conditions. In this study, the potential for deploying machine learning algorithms, including principal component regression (PCR), support vector regression (SVR) and long short-term memory (LSTM) neural network, are tested to build predictive models. These tools were used to estimate chemical dosage at a daily time-scale. Influent water quality such as pH, color, turbidity, and alkalinity, as well as chemical dosage including sulfuric acid, ferric sulfate and liquid oxygen were used to build and test these models. An 80/20 percent data split was used for training and testing model performance using correlation coefficients, relative mean square error, relative root mean square error and Nash–Sutcliffe efficiency. Results indicate, compared with PCR, both SVR and LSTM were able to capture the nonlinear relationship between chemical dose and source water quality changes and displayed higher predictive skills. These types of models have application in real-time operational support without requiring computationally expensive physics-based models.

Key words: chemical dosage prediction, domestic water supply, drinking water treatment plants, long short-term memory neural network, machine learning algorithms, support vector machine

HIGHLIGHTS

- A study that examines water quality and chemical use from a water treatment plant.
- Appropriate chemical dosage is essential to ensure safe potable water.
- Understanding the relationship between water quality and chemical use is a key step.
- Different algorithms are tested to build predictive models.
- Support vector machine and neural networks better capture nonlinear relationships.

INTRODUCTION

Drinking water treatment plants (DWTP), within design limits, can treat raw water from supply sources to meet regulatory water quality standards through a combination of physical, chemical and biological treatment processes. Chemical cost is a significant item of total operational cost for DWTP (Heberling *et al.* 2015; Price & Heberling 2018), which is primarily driven by source water quality for the same amount of raw water being treated.

Adjustment of chemical dosage based on influent water quality is often needed. A dramatic decline in water quality, often caused by extreme rainfall events, severe algal blooms or the presence of cyanobacteria, may require increased monitoring of the treatment process (Besmer & Hammes 2016), timely adjustment of chemical dosage or even temporary shutdown of the DWTP (Kansas Department of Health and Environment 2011). Although the DWTP has standard operating procedures and some flexibility to alter treatment processes within its design limit to cope with fluctuations in raw water quality, a wide range of water quality fluctuations still pose challenges for DWTP at times. A decision-supporting tool that can be used to help operators determine the right amount of chemical dosage depending on intake water quality can be instrumental in reducing chemical cost while achieving water quality standards. Such a tool is especially important for treatment plants that have multiple water supply sources because water quality at an intake is a mixture of water from different supply sources with distinct water quality.

Chemical dosage adjustment requires real-time data, understanding of the fluctuations of the physical, chemical and biological characteristics of the raw water, and in-house expertise on the specific treatment processes. Approaches that facilitate such adjustment can be categorized in two methods. The first is an on-site jar test, which requires treatment plant operators or laboratory technicians to conduct a pre-identified list of laboratory tests to determine the chemical dosage based on experience and professional judgment. The disadvantages of this approach include that it requires dedicated resources and time to conduct laboratory tests, and jar test findings may not correlate directly to real-time treatment. This would also require timing adjustment of the frequency of sample analysis whenever drastic water quality changes occur. The second is the application of statistical methods to build predictive models relating chemical dosage to source water quality changes.

Many case studies using artificial intelligence to predict treatment process performance (Çamdevýren *et al.* 2005; Veerapaneni *et al.* 2010) and facilitate chemical dosage in DWTP have been reported in the literature (Golfinopoulos *et al.* 1998; Maier *et al.* 2004; Lamrini *et al.* 2005; Bae *et al.* 2006; Veerapaneni *et al.* 2010; Heddam *et al.* 2011; Kim & Parnichkun 2016; O'Reilly *et al.* 2018). Nearly all of these studies focused on developing predictive tools for single chemical dosage. This study considers simultaneous dosing requirements of multiple chemicals used in the treatment process and aims to evaluate three machine learning algorithms in building a predictive tool.

This study is motivated by water quality management at a regional water supply agency, Tampa Bay Water, located on the west coast of Florida in the United States. Its drinking water treatment plant accepts a mixture of water from two surface water supply sources, the Hillsborough River and the Alafia River, and an offline reservoir that contains a blend of water from those two sources. Water quality in Hillsborough River and Alafia River is not the same due to land use and hydrogeologic differences in the watersheds. For instance, during the drought season from January to May, harvested water from the Hillsborough River resembles groundwater quality characteristics due to the interaction between surface and groundwater systems where groundwater dominates the total river flow, whereas water quality in an offsite reservoir is the time-averaged water quality of both sources. At any given time, a single source or a mixture of up to three sources may be used to supply water to the DWTP depending on season, water availability, and demand on the system. Examining the relationship between raw water quality and chemical dosage helps in the understanding of key parameters that drive chemical use. It can also play a crucial role in potentially rotating supply sources based on both water quantity and quality.

The goal of this study is to develop a predictive tool for chemical dosage based on intake water quality at the daily timescale. Specifically, it aims to: (1) examine the relationship between historical chemical dosage and influent water quality data; (2) evaluate dominant influent water quality parameters; and (3) compare performance of different models in predicting chemical dosage.

METHODOLOGY

In this section, three different approaches, namely principal component regression (PCR), support vector regression (SVR) and long short-term memory (LSTM) neural network, tested in this study are described briefly and references are provided for interested readers. Model evaluation metrics used for model comparison are also provided.

Principal component regression (PCR)

PCR is an approach used to build a regression model of a reduced dimension of original dependent variables, which uses principal component analysis. Principal component analysis (Shaw 2003; Abdi & Williams 2010) is a dimension reduction technique, aiming to replace a large number of correlated predictors with a reduced number of representative uncorrelated predictors, known as the principal components. Mathematically, each principal component is a linear combination of original predictors, coefficients of which can be obtained through the decomposition of the covariance matrix. A larger eigenvalue denotes that a higher variance can be explained in the direction represented by its corresponding eigenvector. The main advantage of such an approach is dimension reduction and its detailed mathematical derivations are widely reported in the literature. Rather than using the original time series, principal components are used as potential predictors in the PCR. PCR has been applied to build predictive models in water quantity and water quality (Çamdevýren *et al.* 2005; Wang *et al.* 2013).

Support vector regression

SVR represents the application of support vector machine (SVM), which was firstly developed by Vapnik (1995). SVM has been one of the emerging supervised learning techniques and gained wide application in the field of water resources engineering (Asefa *et al.* 2006; Lauer & Bloch 2008; Wang & Harrison 2014). A couple of characteristics owned by the SVM algorithm distinguish it from other machine learning approaches. For instance, it applies the structural risk minimization principle, while most statistical learning approaches only employ empirical risk minimization. This can potentially avoid over-fitting in the model-building process. In addition, the introduction of kernel functions and feature space allows SVM more capability in dealing with high-dimensional data. The kernel functions implicitly map the input data to a higher-dimensional feature space, where a linear solution in the higher-dimensional feature space corresponds to a nonlinear solution in the original space. It is for this reason that SVR is chosen to build the regression model between chemical dosage and water quality parameters, because a nonlinear relationship exists between the two based on preliminary analyses. Radial Basis Function is one class of kernel functions that has been demonstrated effective for many regression problems and it is applied in this study (Vapnik 1995). The following describes the general form of an RBF kernel and γ is a parameter to be determined in specific applications:

$$K(x, x_i) = \exp\left(-\gamma|x - x_i|^2\right) \tag{1}$$

SVR attempts to estimate the linear or nonlinear dependency $f(x)$ between a set of training samples $X = \{x_1, x_2, \ldots, x_l\}$ taken from R^n and target values $Y = \{y_1, y_2, \ldots, y_l\}$ with y_l taken from R^m. Each x_l contains n contributions and each y_l refers to m responses, herein, chemical dosage. Specifically, x_l include water quality parameters and streamflow information and y_l denotes the dosage of different chemicals used in the treatment process, corresponding to a water quality profile. Mathematically, SVR is a quadratic optimization problem with convex constraints. Therefore, a global optimal exists for the quadratic optimization problem. Details of the mathematical formulation can be found elsewhere (e.g., Vapnik 1995). In this study, LIBSVM (Chang & Lin 2011) is implemented in Matlab.

Long short-term memory neural network

A long short-term memory (LSTM) network is a specific type of recurrent neural network, in the configuration of which connections between neurons form a directed cycle (Sutskever *et al.* 2014). Using recurrent neural networks, information from previous time steps can be used as input for predicting the output of the current time step. The concept of information control is introduced, and distinct types of gates are used in its configurations (Hochreiter & Schmidhuber 1997). The salient feature of LSTM is that it can bring information in the past, e.g., crossing several time steps, and avoid earlier signals fading away in learning and simulating the underlying system dynamics. This is primarily because it does not have a problem with exploding and/or vanishing gradients. Although different variants of the LSTM network have been developed, the basic structure of an LSTM network (Gers 2001), typically consisting of three gates, is illustrated in Figure 1. The three different gates are the input gate, forget gate, and output gate, each of which is represented by a sigmoid neural network layer (σ) and a multiplicative unit ($\times$). Mathematical description of the LSTM structure depicted in Figure 1 can be found in Gers *et al.* (2000).

LSTM is implemented in Matlab using the stochastic gradient descent with momentum (SGDM) optimizer. Hyperparameters of the model, e.g., initial learning rate and momentum of the SGDM, are determined through Bayesian optimization.

Model evaluation metrics

A few metrics, including the coefficient of determination, relative root mean square error (RRMSE), and Nash–Sutcliffe coefficient of efficiency (NSE), are used to evaluate the performance of different models. Note that all criteria are applied to each chemical dosage for all predictive models. The coefficient of determination, R^2, is defined as the square of the correlation coefficient shown in Equation (2). The value of R^2 is within the range of 0 to 1. It represents the percentage of variance exhibited in the observed data that can be captured in the model output. It is often used to measure model accuracy and evaluate how well the model can simulate the observed data and predict future values.

$$R^2 = \frac{\left[\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})\right]^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2 \sum_{i=1}^{n}(P_i - \bar{P})^2} \tag{2}$$

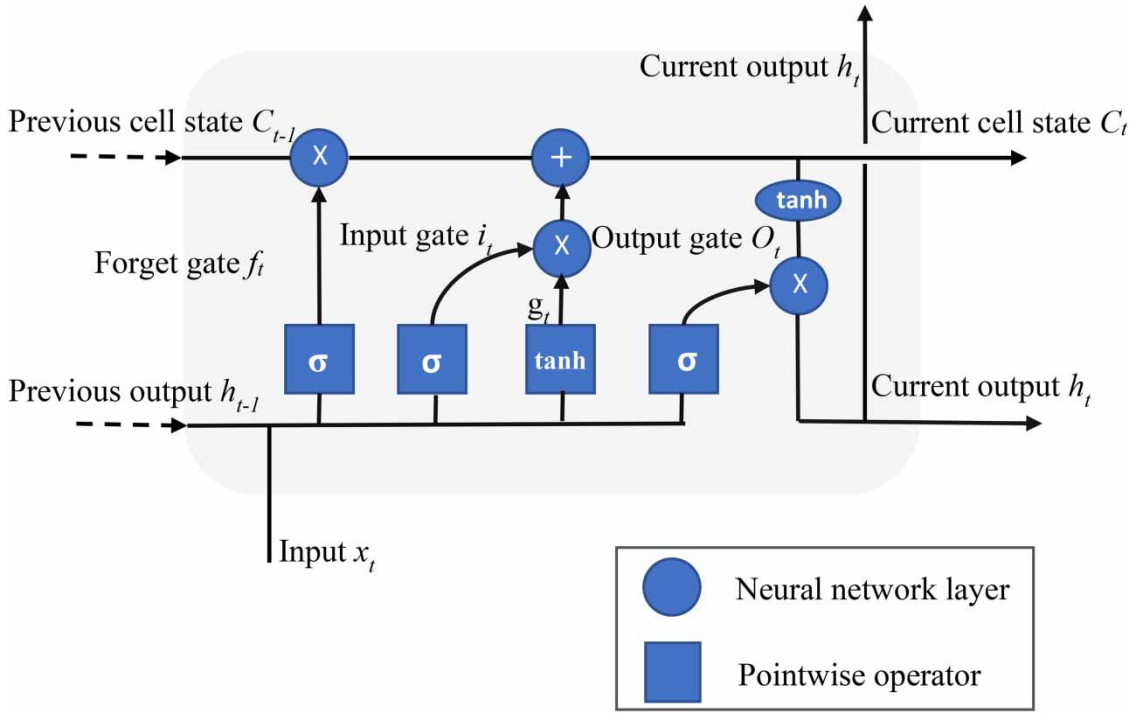

Figure 1 | Graphic representation of LSTM network. The primary configuration includes forget gate, input gate and output gate, which determines information from previous time steps and the current input used to determine output from the current time step.

where n is the length of time series of observed data, O_i is daily observed chemical dosage, P_i is predicted daily chemical dosage, $\bar{O}$ is the mean of observed chemical dosage and $\bar{P}$ is the mean of predicted chemical dosage.

The relative mean square error (in %) is defined based on Equations (3) and (4). Root mean square error is first calculated (Equation (3)) and then normalized by observed mean values. While root mean square error might differ largely among different chemical dosages, RRMSE offers a criterion to measure the percentage of RMSE compared with the average of the chemical dosage.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \tag{3}$$

$$RRMSE = \frac{RMSE}{\bar{O}} \times 100\% \tag{4}$$

The NSE is a measure that evaluates the performance of a model against an alternative model, which consistently uses the long-term mean of the variables as its prediction. Its calculation is shown in Equation (5).

$$NSE = \left[1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2}\right] \tag{5}$$

The value of NSE lies between 1.0 and $-\infty$. A value of 1.0 denotes perfect prediction with the numerator being zero in Equation (5). An efficiency of a negative value indicates that the model is not as good as using the mean value of chemical dosage for prediction.

The three modeling techniques described in this section have been tested in this study to evaluate which has better predictive skills in determining daily chemical use based on influent water quality. Although the modeling techniques are general and can be applied to any other DWTP, results presented in later sections may only apply to the study area in this study, given that water quality parameters and water treatment process may not be the same.

CASE STUDY

Study area and data

Tampa Bay Water (TBW), formed in 1998, is a regional water utility in southwest Florida that provides wholesale water to six of its member governments including the cities of Tampa, St Petersburg and New Port Richey and the counties of Hillsborough, Pinellas and Pasco, serving about 2.5 million residents. TBW has a diverse water supply system, with 13 groundwater wellfields of up to 454,628 cubic metres per day (120.1 million gallons per day (mgd)) permitted capacity on an annual basis, 454,249 cmd (120 mgd) surface water treatment plant and 5.678×10^7 cubic metre (15×10^9 gallon) off-stream reservoir. In addition, a seawater desalination plant of 94,635.30 cmd (25 mgd) has been in operation since 2005.

The surface water treatment plant accepts a mixture of water from two surface water supply sources and an offline reservoir, depending on source rotation from those sources. The two surface water supply sources are the Tampa Bypass Canal connecting to the Hillsborough River, and the Alafia River. A mixture of different portions from the supply sources is based on regional demand, water availability and expert judgment. To ensure high-quality effluent, the plant incorporates a three-stage water treatment process. The first stage is a Veolia patented process that aims to remove the color and particles from the raw water. The second stage is ozone disinfection that kills microorganisms including bacteria, viruses, and protozoa. The third stage is biologically active filtration where remaining organic molecules are removed. After the three-stage process, treated water is disinfected again using chlorine and chloramines. Each of those processes demands chemical use and the chemical dosage is dependent on source water quality.

In this study, water quality data, normalized chemical dosage in the units of milligrams per litre (mg/L), and streamflow at the two surface water rivers during the years 2015–2019 were collected. All data are available at the daily time-scale. Water quality parameters include color, pH, turbidity, alkalinity, temperature, conductivity, chloride and sulfate. Chemicals used in the treatment processes include sulfuric acid, ferric sulfate, polymers, hydrated lime, liquid oxygen, microsand, sodium hydroxide and sodium hypochlorite. Only a subset of chemicals is identified for predictive model building after preliminary analysis. Water quality data at the intake were obtained from supervisory control and data acquisition (SCADA) readings; chemical dosage was obtained from operators and streamflow was retrieved from Tampa Bay Water's database, which gathers hydrologic data from its service area.

Results

Influent water quality exhibits strong seasonality as shown in Figure 2. Color, measured in platinum cobalt units (PCU), is typically introduced by dissolved organic materials from decaying plants and leaves, dissolved minerals and/or industrial discharges. Presence of color can reduce both the quantity and quality of light penetration into the water column. Color measured in the source water varies between 0 and 260 PCU on a daily basis with high values during the months of August, September, and October (Figure 2(a)). Figure 2(b) shows the daily fluctuation of alkalinity, which is a measure of the water's capacity to neutralize acids. It serves as a buffer against rapid pH changes. The buffering capacity comes from the dissolution of compounds in rocks and minerals, e.g., calcium carbonate, magnesium carbonate, phosphates and hydroxides. Because of this, the alkalinity of groundwater is typically greater than that of surface waters. Figure 2(b) indicates that it varies from 40 to 130 mg/L and has higher values in the months of January to May, corresponding to the dry season when influent is a mixture of reservoir water and surface water from the Tampa Bypass Canal, which is connected to the Hillsborough River.

There is a statistically significant correlation among some of the water quality parameters, as shown in Figure 3. The scatterplots of different combinations of water quality parameters are shown in the grids and corresponding correlation values are shown in ovals. The correlation values between different pairs of water quality parameters are shown in the ovals. The red ovals indicate negative and blue indicate positive correlation that is statistically significant at the 0.05 significance level. Otherwise, the correlation value is presented in a grey oval, as shown for the correlation between temperature and turbidity. One of the water quality parameters, color, is negatively correlated with pH, alkalinity, conductivity, calcium hardness, chloride and sulfate. Alkalinity is positively correlated with conductivity, calcium hardness, chloride and sulfate. The correlation

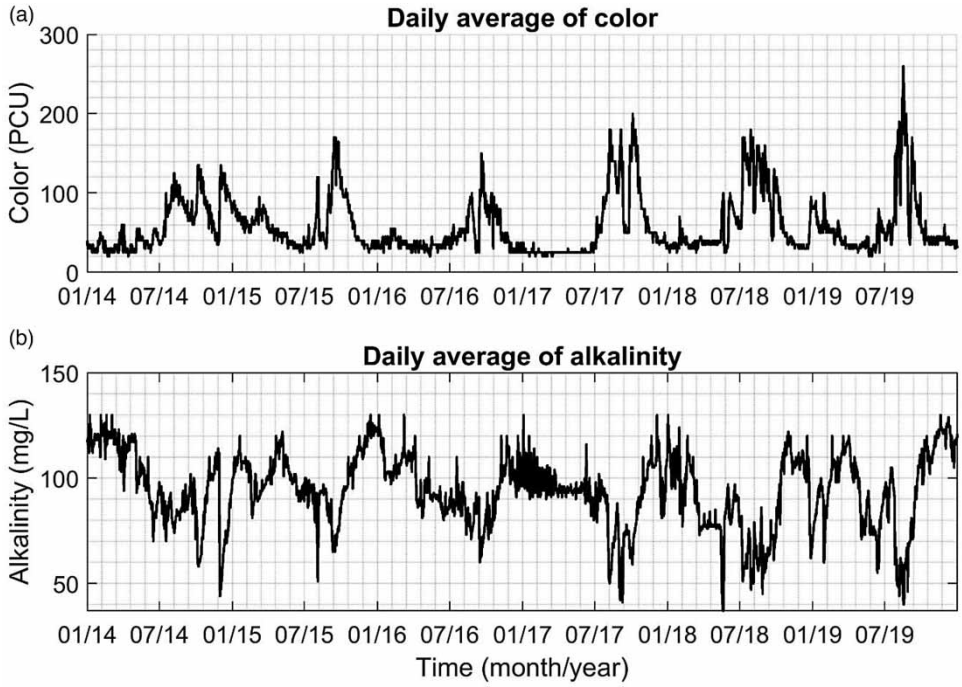

Figure 2 | Panel (a) displays the daily variation of one water quality parameter, color, at the intake of the drinking water treatment plant between January 2014 and December 2019; and panel (b) shows daily alkalinity variation for the same period.

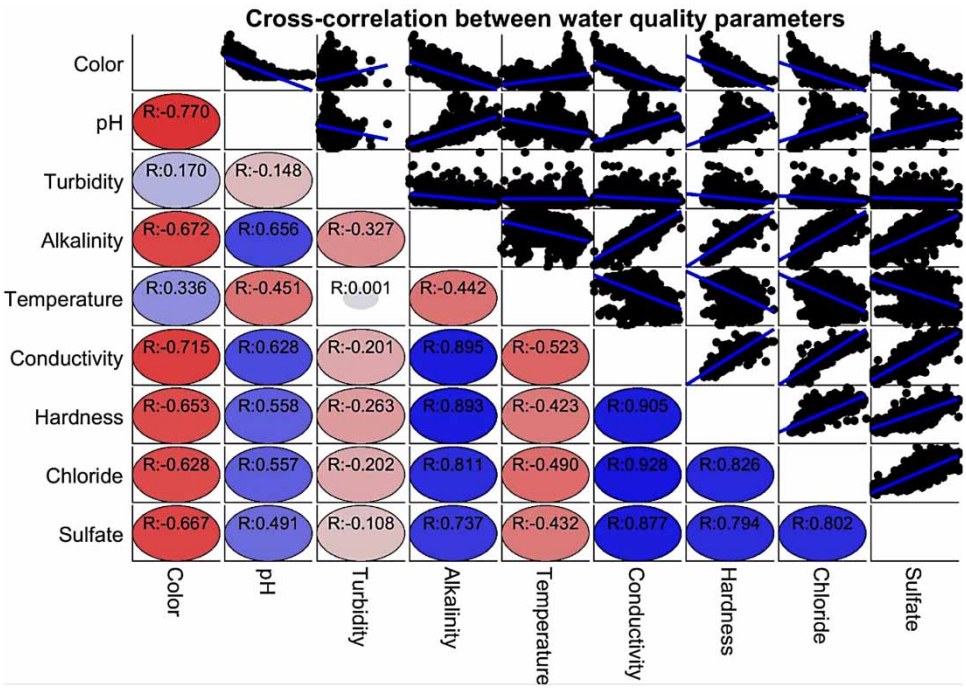

Figure 3 | Cross-correlation between water quality parameters and boxplots for each pair of water quality parameters. The red ovals denote negative and the blue ovals indicate positive correlation. If the correlation is not statistically significant at the 5% significance level, it is shown as a gray color. Values in the ovals represent the correlation between the corresponding water quality parameters.

ranges from 0.74 to 0.90. Conductivity, reflecting the capacity of water to conduct an electronic current, is a measure of ionized substances in water. Strong correlation among water quality parameters provides a basis to employ principal component analysis, using a smaller number of variables to represent the total variabilities exhibited in those variables to reduce the dimensions of a predictive model.

Streamflow data can also be used as predictors for chemical use in the treatment processes since strong correlation has been identified between streamflow and water quality parameters. For the study area, the months of July, August, and September are the rainy season during which time extra surface water can be harvested and stored in the offline reservoir. In this study, both water quality parameters and streamflow data at the two surface water sources are incorporated as predictors. Figure 4 shows the cross-correlation between predictors and chemical use at the daily time-scale. All the correlation values displayed in Figure 4 are statistically significant at the significance level of 0.05.

The pattern of cross-correlation is the same for ferric sulfate, Actiflo polymer and liquid oxygen, as indicated by the colors of the ovals in the last three columns in Figure 4(a). Sulfuric acid is positively correlated with pH, alkalinity, chloride and

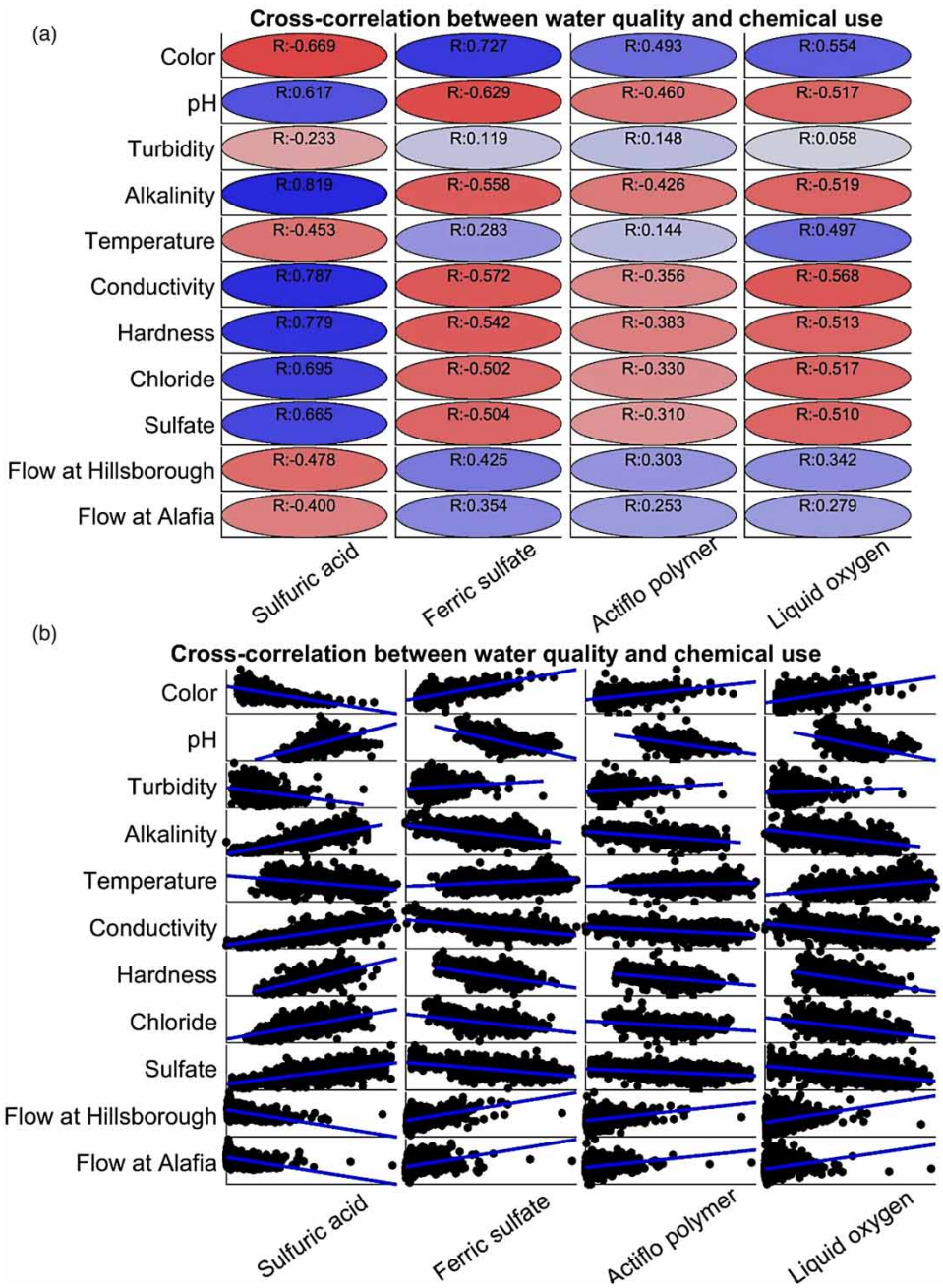

Figure 4 | Panel (a) displays the correlation between predictors, including water quality parameters and streamflow at the two surface rivers, and chemical use in the water treatment processes; panel (b) shows corresponding boxplots.

sulfate. Most of the correlation values are within the range of 0.6 to 0.8. It is, however, negatively correlated with streamflow at Hillsborough River and Alafia River, with the correlation value of −0.48 and −0.40, respectively. Ferric sulfate is negatively correlated with streamflow data and several water quality parameters, including pH, alkalinity, conductivity, calcium hardness, chloride and sulfate. Linear correlation between Actiflo polymer and water quality/quantity is not as strong as that between sulfuric acid and water quality parameters.

Figure 4(b) shows scatterplots of potential predictors and chemical dosage. The blue line is a linear regression of the two variables in each boxplot. It provides insights that otherwise might be ignored by only examining the numbers. For instance, the correlation between alkalinity and sulfuric acid is 0.82; and the correlation between color and ferric sulfate is 0.73. Although the two correlation values are similar, the scatterplot between alkalinity and sulfuric acid shows a different feature compared with the scatterplot between color and ferric sulfate. The variability of alkalinity is smaller at the lower end of sulfuric acid; the variability of color is higher at the lower end of ferric sulfate. Understanding such features facilitates an examination of the performance of predictive models. Preliminary investigation found that lag-1 correlation, i.e., correlation between chemical dosage at the current day and chemical dosage the previous day, is significant. Therefore, lagged-1 chemical dosages, one for each predictand, were also included as predictors. In total, 15 predictors, including nine water quality parameters, streamflow data at two surface water rivers, and four lagged-1 chemical dosages, are used for building predictive models.

As shown earlier, there is a strong cross-correlation between water quality parameters. Figure 5 below shows the percentage of variance explained by each principal component in the 14 predictors. The first five principal components, in total, explain 82.7% of the variance exhibited in all the predictors. Each principal component is a combination of all the 15 predictors.

Since predictors are of different units and the magnitude of values varies, each predictor was normalized to ensure data has a zero mean. In building predictive models, 80% of the data was used for model calibration and the remaining 20% was set aside for testing model performance. For the PCR model, the first five principal components were used. Figure 6 shows the scatterplots between observed data and model prediction for both training and testing periods. The nonlinear relationship between predictors and chemical dosage for sulfuric acid and ferric sulfate can be inferred from Figure 6(a) and 6(b). At the higher end of actual sulfuric acid dosage, when it is over 120 mg/L, the PCR model underpredicts observations. At the lower end of actual ferric sulfate, when it is less than 100 mg/L, the PCR model overpredicts these values.

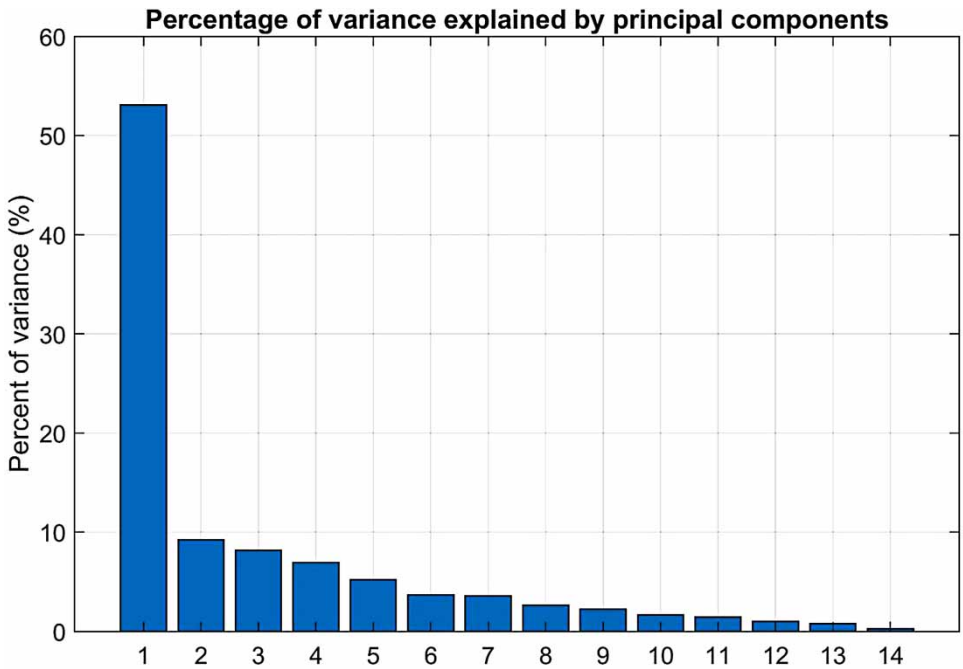

Figure 5 | Percentage of variance exhibited in all predictors that is explained by different principal components. The first five principal components explain 82.5% of the total variance, indicating high correlation between the predictors.

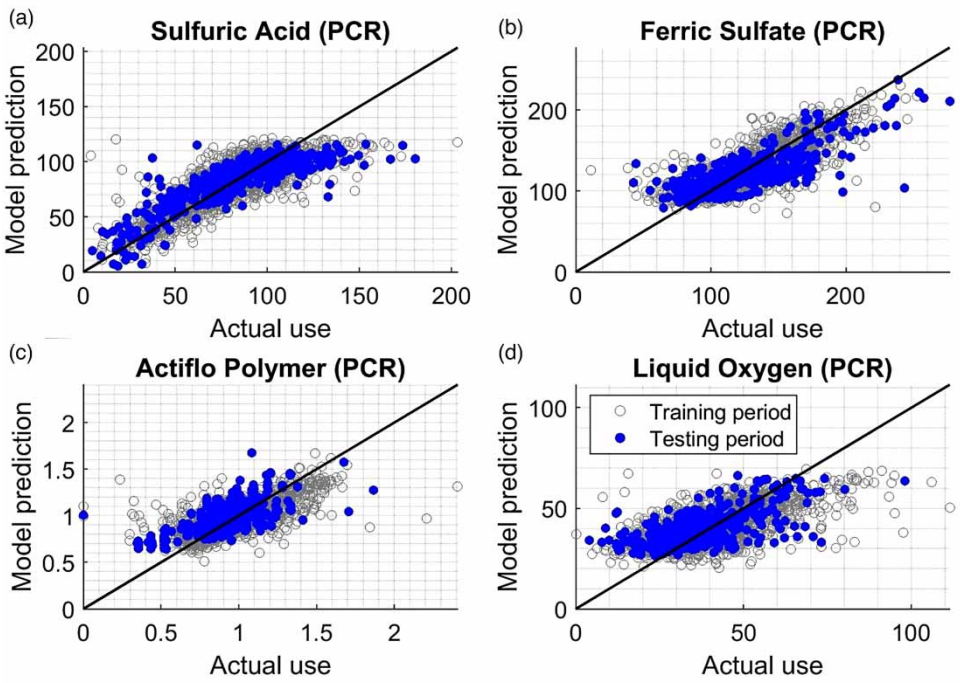

Figure 6 | Boxplots between actual chemical dosage and predicted values using principal component regression for sulfuric acid (panel a); ferric sulfate (panel b); Actiflo polymer (panel c); and liquid oxygen (panel d). The training data are shown as gray circles and testing data are shown as blue filled circles.

In using SVR, different kernel functions including the linear and RBF kernel are tested to obtain better performance. A grid search method is used to identify the optimal hyperparameters in training the SVR model. Mapping the predictors to a higher-dimensional space allows SVR to better capture the relationship between predictors and chemical dosage. Figure 7 shows scatterplots between actual chemical dosage and predictions from the SVR model for both the training and testing periods. Compared with principal component regression, the SVR model better captures the variabilities in sulfuric acid and ferric sulfate.

In training the LSTM model, an input layer of 14 nodes is first added, followed by an LSTM layer with ten nodes and a fully connected layer of four nodes, corresponding to the dosage of the four chemicals. The mean square error of the model during the training period is used to optimize the selection of hyperparameters of the LSTM model. Figure 8 shows the scatterplots between actual chemical dosage and predictions from the SVR model for both the training and testing periods. Model performance for both the training and testing periods is summarized in Table 1.

All the three models have better predictive skill for sulfuric acid and ferric sulfate compared with Actiflo polymer and liquid oxygen. Although performance in the training period is not the ultimate objective of model building, it can provide insights into how well a model could perform for the testing period. For the training data, the R^2 between model predictions and observation ranges between 0.62 and 0.75 for sulfuric acid and ferric sulfate, indicating 62%–75% variance can be explained by the predictive models. The R^2 reduces to within the range of 0.42 and 0.66 for Actiflo polymer and liquid oxygen. This indicates that the dosages of Actiflo polymer and liquid oxygen are not fully determined by influent water quality; it may also relate to other treatment processes that are not explicitly captured in this study. This can also be reflected in the metric of RRMSE, the value of which range between 0.10 and 0.16 for sulfuric acid and ferric sulfate during the testing period. RRMSE is within the range of 0.11 and 0.23 for Actiflo polymer and liquid oxygen. For the testing data, NSE is greater than 0.5 for sulfuric acid and ferric sulfate for all the three models.

Among the three predictive models, both SVR and LSTM are generally better than PCR, whereas the performance of SVR and LSTM is comparable, although there is some difference between the two depending on the type of chemicals. LSTM better captures variabilities for sulfuric acid and ferric sulfate when the chemical dosage is relatively higher, i.e., sulfuric acid over 120 mg/L and ferric sulfate over 200 mg/L. SVR, on the other hand, has better performance for Actiflo polymer when the observation is less than 0.75 mg/L.

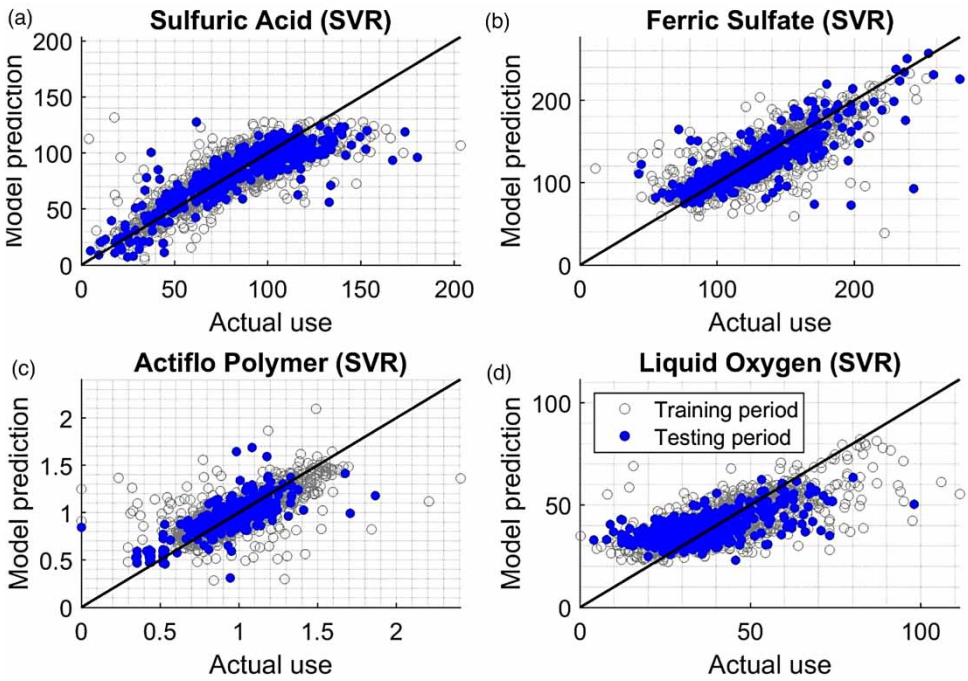

Figure 7 | Boxplots between actual chemical dosage and predicted values using support vector regression for sulfuric acid (panel a); ferric sulfate (panel b); Actiflo polymer (panel c); and liquid oxygen (panel d). The training data are shown as gray circles and testing data are shown as blue dots.

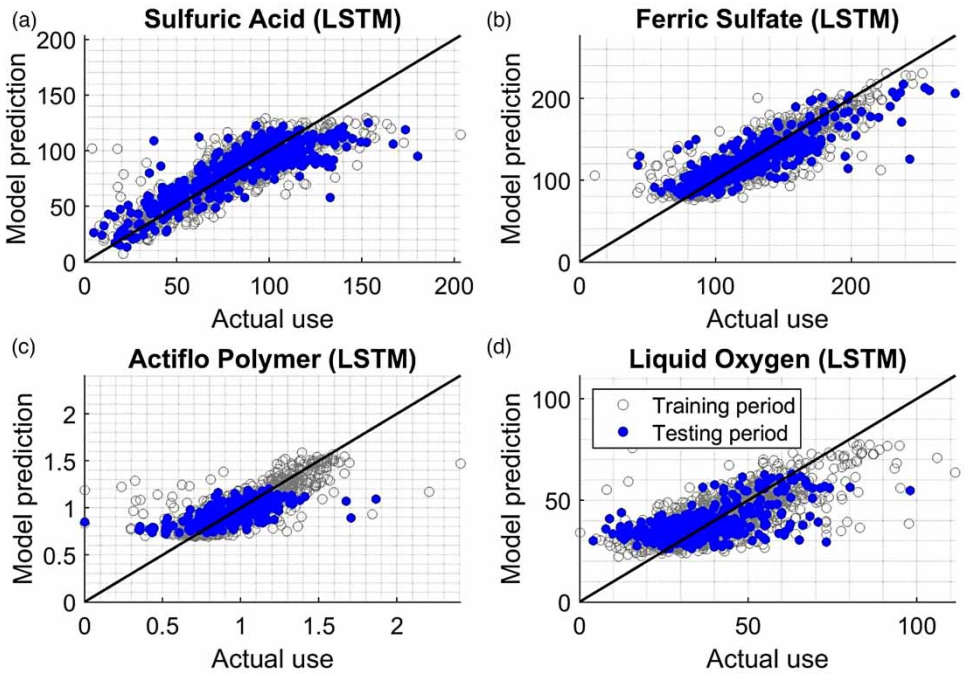

Figure 8 | Boxplots between actual chemical dosage and predicted values using the long short-term memory model for sulfuric acid (panel a); ferric sulfate (panel b); Actiflo polymer (panel c); and liquid oxygen (panel d). The training data are shown as gray circles and testing data are shown as blue dots.

Table 1 | Performance comparison of the three predictive models during training and testing periods.

		Training data			Testing data		
Evaluation metric	**Chemical dosage**	**PCR**	**SVR**	**LSTM**	**PCR**	**SVR**	**LSTM**
R^2	Sulfuric acid	0.67	0.71	0.75	0.71	0.74	0.71
	Ferric sulfate	0.62	0.69	0.77	0.61	0.66	0.70
	Actiflo polymer	0.49	0.55	0.67	0.49	0.64	0.54
	Liquid oxygen	0.42	0.56	0.53	0.42	0.38	0.36
RRMSE	Sulfuric acid	0.20	0.18	0.17	0.21	0.20	0.21
	Ferric sulfate	0.16	0.14	0.12	0.18	0.16	0.16
	Actiflo polymer	0.18	0.17	0.14	0.20	0.16	0.19
	Liquid oxygen	0.26	0.23	0.23	0.29	0.30	0.30
NSE	Sulfuric acid	0.67	0.71	0.75	0.69	0.73	0.71
	Ferric sulfate	0.62	0.69	0.77	0.59	0.65	0.69
	Actiflo polymer	0.49	0.54	0.67	0.42	0.63	0.46
	Liquid oxygen	0.42	0.55	0.53	0.36	0.34	0.34

DISCUSSION

Adjustment of chemical dosage in the treatment processes in drinking water treatment plants (DWTP) based on influent water quality is essential in producing quality water and maintaining operational cost at the targeted level. Data analyses provided in the Results section provide insights into what water quality conditions would require a higher dosage of different chemicals, indicating higher chemical cost and potential challenges for the treatment processes. It is well recognized that improved source water quality generally leads to a lower chemical cost for DWTP (Holmes 1988; Piper 2003; Forster & Murray 2007; Abildtrup *et al.* 2013; Mosheim & Ribaudo 2017; Warziniack *et al.* 2017). Therefore, source water protection, broadly defined as actions that protect source water quality conditions from adverse impact prior to the intake, and real-time water quality monitoring have received growing attention from practitioners in recent years (Price & Heberling 2018).

Historically, chemical dosage was determined by an implicit model, i.e., mainly experience and professional judgement. Therefore, such chemical dosage might not be optimal given water quality conditions at that time. Predictive models explored in this study aim to provide an explicit and alternative model in determining chemical dosage. Both predictive models built using SVR and LSTM have excellent performance for sulfuric acid and ferric sulfate. These models can be used by water managers to predict chemical dosage. For Actiflo polymer and liquid oxygen, predictive models developed in this study can be further fine-tuned to incorporate treatment processes, e.g., outlet water quality from prior treatment units. For instance, the dosage of liquid oxygen is mostly related to what is coming out of the Actiflo and overfeed of chemicals in Actiflo can lead to much higher ozone demand. Operators do this in real-time by making an adjustment, then watching to see whether it affects the ozone residual. Incorporating intermediate water quality in the upstream treatment processes to predict chemical dosage in the downstream processes is a potential extension of the current study.

Analyses presented in this study also have implications for water resources management to incorporate both water quantity and water quality. This is especially true for water utilities that have multiple supply sources with varying water quality to feed their DWTP. For the case study of Tampa Bay Water, water quality in Hillsborough River and Alafia River is not the same due to land use and hydrogeologic differences in the watersheds. For instance, during the drought season from January to May, harvested water from Hillsborough river resembles groundwater quality characteristics due to the interaction between surface and groundwater systems. Figure 9 shows the temporal variation of one water quality parameter, color, and portions of influent from different supply sources. A decision tool that incorporates both water quantity and water quality at each supply source can be developed to determine water withdrawal from each, which is beyond the scope of this study. Real-time monitoring of water quality at the source water, which is currently unavailable at the time of conducting this study, would facilitate the developing of a decision support tool based on the current study. This can be another extension of this study.

CONCLUSIONS

This study presented data-driven approaches to predict chemical dosage in DWTP based on influent water quality. Three machine learning algorithms, namely principal component regression, support vector machine and long short-term

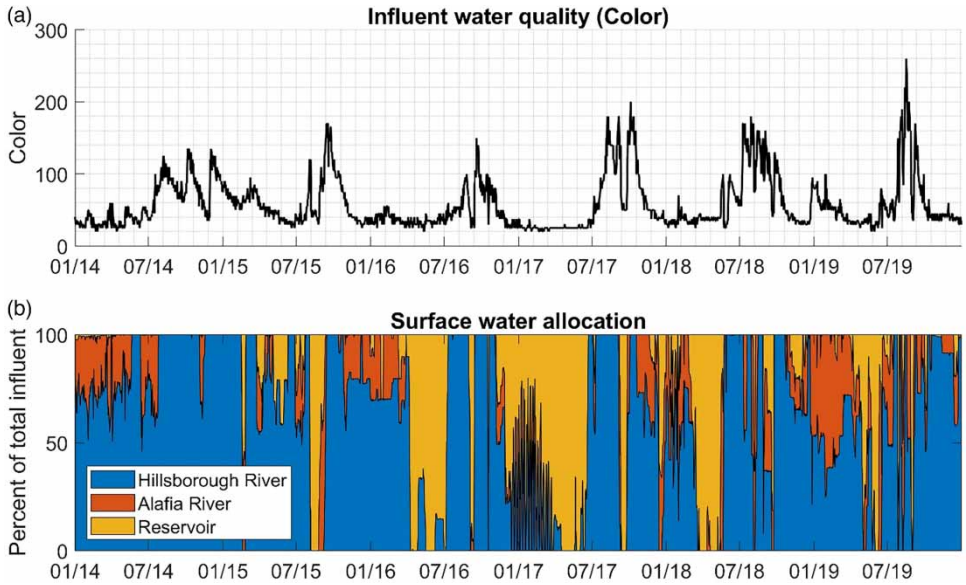

Figure 9 | Panel (a) displays daily variation of color in platinum cobalt units during the time period January 2014 to December 2019; and panel (b) displays estimated percent of source water (%) that comprised the influent at the drinking water treatment plant.

memory neural network, are tested to build predictive models for chemical dosage at the daily time-scale. Predictors include water quality parameters, e.g., pH, color and turbidity, and streamflow information. An 80%/20% split was used for the training and testing datasets. Four evaluation metrics, namely correlation, relative mean square error, relative root mean square error, and Nash–Sutcliffe efficiency, were used. Both SVR and LSTM were found to better capture the nonlinear relationship than PCR. This was demonstrated through better predictive skills. The performance of SVR and LSTM was comparable, although there was a slight difference between the two. LSTM better captured variabilities for sulfuric acid and ferric sulfate when the chemical dosage was relatively higher, i.e., sulfuric acid over 120 mg/L and ferric sulfate over 200 mg/L. SVR, on the other hand, had better performance for Actiflo polymer when the observation was less than 0.75 mg/L. Extension of the analyses presented in this study includes further incorporation of treatment processes to improve predictive skill for Actiflo polymer and liquid oxygen dosages. Results from this study can be further utilized in building a decision support tool that incorporates both water quantity and quality to determine withdrawal from different supply sources to feed the DWTP. Although the analyses were conducted for a case study presented in this paper, the methodology is general and can be useful for other regions.

ACKNOWLEDGEMENTS

The authors thank the constructive comments from the two anonymous reviewers that have improved the manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Abdi, H. & Williams, L. J. 2010 Principal component analysis. *WIREs Computational Statistics* **2** (4), 433–459.

Abildtrup, J., Garcia, S. & Stenger, A. 2013 The effect of forest land use on the cost of drinking water supply: a spatial econometric analysis. *Ecological Economics* **92**, 126–136.

Asefa, T., Kemblowski, M., McKee, M. & Khalil, A. 2006 Multi-time scale stream flow predictions: the support vector machines approach. *Journal of Hydrology* **318** (1–4), 7–16. doi:10.1016/j.jhydrol.2005.06.001.

Bae, H., Kim, S. & Kim, Y. J. 2006 Decision algorithm based on data mining for coagulant type and dosage in water treatment systems. *Water Science and Technology* **53** (4–5), 321–329. doi:10.2166/wst.2006.137.

Besmer, M. D. & Hammes, F. 2016 Short-term microbial dynamics in a drinking water plant treating groundwater with occasional high microbial loads. *Water Research* **107**, 11–18. doi:10.1016/j.watres.2016.10.041.

Çamdevýren, H., Demýr, N., Kanik, A. & Keskýn, S. 2005 Use of principal component scores in multiple linear regression models for prediction of *Chlorophyll*-a in reservoirs. *Ecological Modelling* **181** (4), 581–589.

Chang, C.-C. & Lin, C.-J. 2011 LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2** (3), 27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Forster, D. L. & Murray, C. 2007 Effects of pesticide use and farming practices on water treatment costs in Maumee River basin communities. In: *Economic Valuation of River Systems* (Hitzhusen, F. J., ed.), Edward Elgar, Cheltenham, UK and Northampton, MA, USA, pp. 115–128.

Gers, F. A. 2001 *Long Short-Term Memory in Recurrent Neural Networks*. PhD thesis, EPFL, Lausanne, Switzerland.

Gers, F. A., Schmidhuber, J. & Cummins, F. 2000 Learning to forget: continual prediction with LSTM. *Neural Computation* **12** (10), 2451–2471.

Golfinopoulos, S. K., Xilourgidis, N. K., Kostopoulou, M. N. & Lekkas, T. D. 1998 Use of a multiple regression for predicting trihalomethane formation. *Water Research* **32** (9), 2821–2829.

Heberling, M. T., Nietch, C. T., Thurston, H. W., Elovitz, M., Birkenhauer, K. H., Panguluri, S., Ramakrishnan, B., Heiser, E. & Neyer, T. 2015 Comparing drinking water treatment costs to source water protection costs using time series analysis. *Water Resources Research* **51**, 8741–8756. doi:10.1002/2014WR016422.

Heddam, S., Bermad, A. & Dechemi, N. 2011 Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water-treatment plant: comparative study. *Journal of Environmental Engineering* **137** (12), 1209–1214.

Hochreiter, S. & Schmidhuber, J. 1997 LSTM can solve hard long time lag problems. In: *NIPS'96: Proceedings of the 9th International Conference on Neural Information Processing Systems* (Jordan, M. I. & Petsche, T., eds), MIT Press, Cambridge, MA, USA, pp. 473–479.

Holmes, T. P. 1988 The offsite impact of soil erosion on the water treatment industry. *Land Economy* **64** (4), 356–366.

Kansas Department of Health and Environment 2011 *Water Quality Standards White Paper: Chlorophyll-a Criteria for Public Water Supply Lakes or Reservoirs*. Kansas Department of Health and Environment Bureau of Water, Topeka, KS, USA.

Kim, C. M. & Parnichkun, M. 2016 MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant. *Journal of Water Supply: Research and Technology – Aqua* **66** (1), 49–61. doi:10.2166/aqua.2016.022.

Lamrini, B., Benhammou, A., Le Lann, M.-V. & Karama, A. 2005 A neural software sensor for online prediction of coagulant dosage in a drinking water treatment plant. *Transactions of the Institute of Measurement and Control* **27** (3), 195–213. doi:10.1191/0142331205tm141oa.

Lauer, F. & Bloch, G. 2008 Incorporating prior knowledge in support vector regression. *Machine Learning* **70** (1), 89–118.

Maier, H. R., Morgan, N. & Chow, C. W. K. 2004 Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling & Software* **19**, 485–494.

Mosheim, R. & Ribaudo, M. 2017 Costs of nitrogen runoff for rural water utilities: a shadow cost approach. *Land Economics* **93** (1), 12–39.

Piper, S. 2003 Impact of water quality on municipal water price and residential water demand and implications for water supply benefits. *Water Resources Research* **39** (5), 1127.

Price, J. I. & Heberling, M. T. 2018 The effects of source water quality on drinking water treatment costs: a review and synthesis of empirical literature. *Ecological Economics* **151**, 195–209. doi:10.1016/j.ecolecon.2018.04.014.

O'Reilly, G., Bezuidenhout, C. C. & Bezuidenhout, J. J. 2018 Artificial neural networks: applications in the drinking water sector. *Water Supply* **18** (6), 1869–1887. https://doi.org/10.2166/ws.2018.016.

Shaw, P. J. A. 2003 *Multivariate Statistics for the Environmental Sciences*. Hodder Arnold, London, UK.

Sutskever, I., Vinyals, O. & Le, V. Q. 2014 Sequence to sequence learning with neural networks. In: *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, eds), MIT Press, Cambridge, MA, USA, pp. 3104–3112.

Vapnik, V. N. 1995 *The Natural of Statistical Learning Theory*. Springer Verlag, New York, USA.

Veerapaneni, S., Budd, G., Bond, R. & Horsley, M. 2010 Using neural networks to predict treatment process performance. *Journal – American Water Works Association* **102** (4), 38–44. https://doi.org/10.1002/j.1551-8833.2010.tb10083.x.

Wang, H., Reich, B. & Lim, Y. H. 2013 A Bayesian approach to probabilistic streamflow forecasts. *Journal of Hydroinformatics* **15** (2), 381–391. doi:10.2166/hydro.2012.080.

Wang, H. & Harrison, K. W. 2014 Improving efficiency of the Bayesian approach to water distribution contaminant source characterization with support vector regression. *Journal of Water Resources Planning and Management* **140** (1), 3–11. doi:10.1061/(asce)wr.1943-5452.0000323.

Warziniack, T., Sham, C. H., Morgan, R. & Feferholtz, Y. 2017 Effect of forest cover on water treatment costs. *Water Economics and Policy* **3** (4), 1750006.

First received 30 August 2021; accepted in revised form 29 November 2021. Available online 14 December 2021