Smart Water

RESEARCH Open Access

Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model

Mo'tamad Bata[1], Rupp Carriveau[1*] and David S.-K. Ting[2]

* Correspondence: rupp@uwindsor.ca
[1]Civil and Environmental Engineering Department, Turbulence and Energy Laboratory, Ed Lumley Centre for Engineering Innovation, University of Windsor, 401 Sunset Ave, Windsor, ON, Canada
Full list of author information is available at the end of the article

Abstract

Regression Tree (RT) forecasting models are widely used in short-term demand forecasting. Likewise, Self-Organizing Maps (SOM) models are known for their ability to cluster and organize unlabeled big data. Herein, a combination of these two Machine Learning (ML) techniques is proposed and compared to a standalone RT and a Seasonal Autoregressive Integrated Moving Average (SARIMA) models, in forecasting the short-term water demand of a municipality. The inclusion of the Unsupervised Machine Learning clustering model has resulted in a significant improvement in the performance of the Supervised Machine Learning forecasting model. The results show that using the output of the SOM clustering model as an input for the RT forecasting model can, on average, double the accuracy of water demand forecasting. The Mean Absolute Percentage Error (MAPE) and the Normalized Root Mean Squared Error (NRMSE) were calculated for the proposed models forecasting 1 h, 8 h, 24 h, and 7 days ahead. The results show that the hybrid models outperformed the standalone RT model, and the broadly used SARIMA model. On average, hybrid models achieved double accuracy in all 4 forecast periodicities. The increase in forecasting accuracy afforded by this hybridized modeling approach is encouraging. In our application, it shows promises for more efficient energy and water management at the water utilities.

Introduction

Marching forward toward true sustainability, reliable and resilient water systems are essential in a world facing the challenge of water scarcity. Smart decision making in water systems is one of the two keys a Smart Water consists of (Joong 2018). The application of water demand forecasting is crucial for optimal operation and control of Smart Water Grids (SWG) (Public Utilities Board Singapore 2016). Supplying water at lower cost, with less energy, and lighter load on the network infrastructure is a primary goal of water utilities. This goal is achieved through multiple practices and applications in the water supply system; one of which is an accurate forecast of the systems' future demand. Short-term water demand forecasting has been employed by a plethora of utilities, researchers, and developers to tackle the imbalance between supply and demand. Short-term water demand forecasting can be used to manage water pressure, control leakage, schedule pumping operations, system maintenance, and infrastructure development (Zhou et al. 2002). However, developing a water demand forecasting model is challenging. That is because the accuracy of the model output controls the efficiency of the system response (Jamieson et al. 2007).

Forecasting models have been a topic of significant interest to researchers and developers alike. A wide spectrum of forecasting models has been featured in the literature. Kozlowski et al. 2018; Donkor et al. 2014; House-Peters and Chang 2011 presented an extensive review of methods and models used in water demand forecasting. Zhang 2001 largely classified these models into two groups, linear and nonlinear. Linear models are used extensively owing to their simplicity and the practicality of the required data acquisition. The ease of implementation and ability to update make these models very attractive. Autoregressive Integrated Moving Average (ARIMA) and univariate time series analysis models have been proposed by many researchers (Zhou et al. 2000; Maidment and Miaou 1986; Maidment et al. 1985; Perry 1981; Hughes 1980) to forecast water demand. ARIMA models can be employed to forecast water demand; however, the accuracy of the forecast can be unsatisfactory (Kozlowski et al. 2018). To overcome the unsatisfactory results of the linear models, researchers developed nonlinear models. Nonlinear models deemed to better capture the nonlinear patterns in water demand. Nonlinear models are costly and difficult to develop, implement, and update. However, their capacity to analyze big data with multiple parameters and concurrently find the nonlinearity relations between variables, have make them powerful prediction tools. Artificial Neural Networks (ANN), nonlinear regression models, fuzzy logic, and other nonlinear models are among the most popular for forecasting water demand (Bennett et al. 2013; Boguadis et al. 2005; Adamowski and Karapataki 2010; Adamowski et al. 2012; Tiwari and Adamowski 2015; Mitrea et al. 2009; Ghiassi et al. 2008; Ghiassi et al. 2005; Hippert et al. 2001; Jain et al. 2001; Cutore et al. 2008; Nasseri et al. 2011).

Two sub-groups can be further distinguished: Standalone and hybrid. Standalone are models that forecast the demand using one technique, where hybrid models are a combination of more than one technique. Standalone linear and nonlinear models are the dominant in the application of demand forecasting; however, researchers have also combined two or more models. Shamseldin and O'Connor 2001 combined a data-driven ANN model and a deterministic model to forecast demand. Abebe and Price 2003 added an ANN prediction model to improve rainfall-runoff model forecasts. The inclusion of the ANN model in both hybrid models was to calculate innovations by being calibrated to residual error time-series. Hiroyuki et al. 2001 combined multi-layer perceptron (MLP) of ANN model and a fuzzy regression tree model to forecast demand in power grid. While the MLP model was employed to forecast one step ahead, the fuzzy regression tree model assisted by determining a split value. This value helped organizing the input data in classes according to specific data rules. These two fused models were proven effective in forecasting power systems load. Mori and Takahashi 2011 proposed another example of hybrid model. A Regression Tree (RT) model was fused with a Relevance Vector Machine (RVM) model. The RT model used some distinct characteristic similarities to classify data into clusters. The classification technique abated the RVM prediction task in smaller clusters with common characteristics. The hybrid model was employed to successfully forecast the electric load in Japan. Tiwari and Adamowski 2013 assembled multiple ANN models developed using bootstrap sampling and wavelet analysis. Their hybrid model was shown to outperform standalone ANN, ARIMA, and ARIMA with exogenous variables (i.e. ARIMAX) models when deployed to forecast water demand.

Despite that hybrid models were mostly developed to forecast electric load; researchers have highlighted strong similarities between water and electricity demand patterns and forecasting approaches (Perry 1981). These similarities are represented by: demand driving factors, demand trend and seasonality, economic and socio-economic parameters, etc. Table 1 highlights short-term load forecasting studies in both water and power grids. Models listed in this table are distinguished by their mathematical representation (i.e. linear/nonlinear) and performance (i.e. standalone/hybrid).

In this paper, short-term water demand forecasting application is achieved through a hybrid ANN model. The hybrid model consists of a clustering unsupervised and a predictive supervised machine learning models. The performance and complexity of the resulted fused model are investigated. The objective is to inspect whether adding an auxiliary clustering model to the forecasting model improves its performance or not. This is presented in the following flow: section 2 introduces the location and data of the studied water utility and its unique characteristics. Section 3 details the implemented forecasting methodology and forecasting models. The performance, computational load, and application of the hybrid model is presented and discussed in section 4. And finally, concluding remarks for this work is highlighted in section 5.

Study area and data

The data analyzed in this paper were collected from a water utility that services primarily rural areas located in Southwestern Ontario, Canada. The water utility supplies water to more than 65,000 residents, agricultural, and other commercial and industrial customers. On average, commercial greenhouses are the utility's dominant water consumers with an annual share of 80%. Since the legalization of cannabis in Canada in 2018, this annual percent share has been increasing according to the water utility's reports. The utility's reports show frequent requests by the commercial greenhouses on increasing their water demand drastically. This dynamic variance accompanied by other demand-driving factors have pressured the supply side, water supply systems, in the region.

Input data

The performance of a short-term water demand forecasting model is dependant on the input type and the temporal resolution of the fed data. Although a number of researchers (Boguadis et al. 2005; Ghiassi et al. 2008; Jain and Ormsbee 2002; Rice et al. 2017) showed that exogenous inputs can improve model predictability, a preliminary study (Bata et al. 2020) in the same area revealed that temporal inputs of the historical water demand are the main driver. The study has also highlighted that 4 recent months of 1-h resolution data is sufficient in the application of short-term water demand forecasting at the studied water utility. Therefore, the input data collected for this study is a 1-h resolution data that spans the last 4 months of 2017. Figure 1 shows the historical water outflow in (m^3/hr) for the studied water utility during the months of August (top

Table 1 Highlights of linear and nonlinear short-term forecasting models

Model Category	Performance	Purpose	Reference Number
Linear	Standalone	Forecasting daily urban water demand	(Zhou et al. 2000; Maidment and Miaou 1986; Maidment et al. 1985; Perry 1981; Hughes 1980)
Nonlinear	Standalone	Short-term water demand forecast	(Bennett et al. 2013; Boguadis et al. 2005; Adamowski and Karapataki 2010; Adamowski et al. 2012; Tiwari and Adamowski 2015; Mitrea et al. 2009; Ghiassi et al. 2008; Ghiassi et al. 2005; Hippert et al. 2001; Jain et al. 2001; Cutore et al. 2008; Nasseri et al. 2011)
	Hybrid	Short-term river flow forecast	(Shamseldin and O'Connor 2001)
	Hybrid	Short-term rainfall-runoff forecast	(Abebe and Price 2003)
	Hybrid	Forecast step-ahead for a power system	(Hiroyuki et al. 2001)
	Hybrid	Forecast the electric load	(Mori and Takahashi 2011)
	Hybrid	Short-term water demand forecast	(Tiwari and Adamowski 2013)

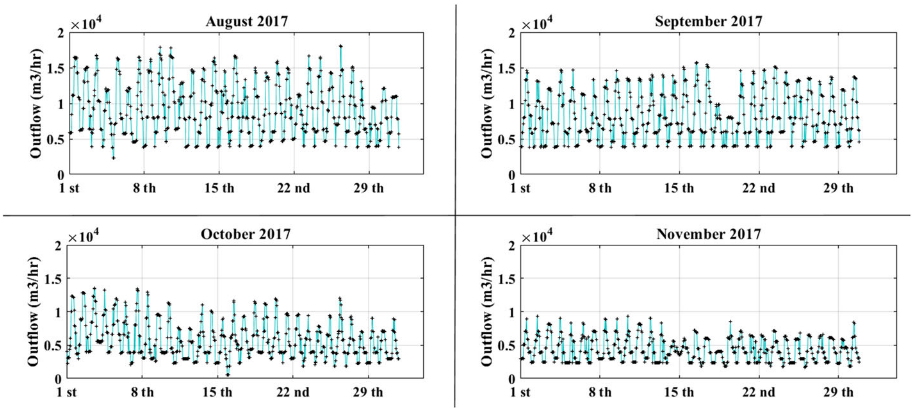


Fig. 1 The utility hourly water outflow (August – November), 2017. A 4-month (August–November) hourly water outflow dataset is shown in this figure, measured in m^3/hr. The hourly water outflow is the target dataset used for clustering, prediction, and performance assessment in the model

left), September (top right), October (bottom left), and November (bottom right) of 2017. The water demand can be seen from this figure, where the water outflow is at minimum around midnight and increases to reach its peak around noon. Although water demand, reflecting water consumption, is usually at peak during early morning and early evening, this is not the case here. That is because 80% of the total utility's demand is consumed by commercial greenhouses. These commercial greenhouses have their own water storage facilities that help them avoid consuming expensive-water during peak hours (note that water price fluctuates based on time-of-use in this region). Also observed is the gradual decrease in peak demand between the months of August and November. This is due to the seasonality of grown crops at the commercial greenhouses.

A unique indicator (i.e. predictor) has been assigned to each data point to address the abovementioned demand patterns and observations. The first indicator is hour of the day. This indicator assigned a value between 1 (represents 12:00 am – 01:00 am) and 24 (represents 11:00 pm – 12:00 am) to each data point. The second indicator is day of the month (e.g. values ranged between 1 and 31 for the month of August). The third indicator is month of the year where each data point was assigned a value between 8 (for August) and 11 (for November).

Data correlation

The correlation between the target, the current water outflow (Q_t), and other input data (i.e. predictors) was calculated. Pearson Correlation Coefficient (PCC) was calculated using eq. 1. PCC evaluates the linear relationship between two variables, and it ranges between − 1 to 1, where 0 indicates no correlation. Table 2 shows the investigated predictors ranked according to their correlation strength with the water outflow. Table 2 reveals that the strongest predictor with the highest PCC value is the K value for the same day previous hour water outflow. Although this is the most correlated predictor, it was not used in most of the models in this paper. That is because this predictor would be practical to use only in the 1 h ahead forecast. If the forecast is to be performed for other periodicities (8 h, 24 h, and 1 week ahead in this study), the values

Table 2 Pearson Correlation Coefficient (PCC) between the utility water Outflow and the predictors used in forecasting

Predictor	Rank	PCC
K[a] same day previous hour	1	0.843
Outflow previous day same hour	2	0.837
Outflow previous week same hour	3	0.821
K previous week same hour	4	0.793
Outflow average previous 24 h	5	0.562
Month of the year	6	−0.529[b]
Hour of the day	7	0.194
Day of the month	8	−0.058

[a] K is the cluster number obtained from the SOM model output

[b] Negative PCC means inverse correlation

for this predictor would be unknown. The same concept applies for other predictors chosen to train models that predicted the water demand 8 h, 24 h, and 1 week ahead.

Predictors ranked 2nd, 3rd, and 4th had a relatively strong correlation of approximately 0.8. These three predictors were used in training, testing, and validating the proposed models to forecast water demand 1 h, 8 h, 24 h, and 1 week ahead. Predictors ranked 5th and 6th had a relatively moderate correlation, however, they did not add any significance in models performance. This was also the case for the weakest studied correlated predictors ranked 7th and 8th.

$$PCC = \frac{n\sum XiYi - \sum Xi \sum Yi}{\sqrt{n\sum Xi^2 - (\sum Xi)^2}\sqrt{n\sum Yi^2 - (\sum Yi)^2} \tag{1}$$

Where,

n: is the number of data points

i: is the data point, ranges from i = 1 to i = n

Xi: is the i^{th} value of variable X

Yi: is the i^{th} value of variable Y

Forecasting methodology and models

This section presents a description of the forecasting methodology, the forecasting approach, the forecasting models, and the model performance assessment approach. Figure 2 illustrates the forecasting flowchart. The process begins with gathering available historical water demand data for the studied system. The data at this stage is usually raw. Here, raw data is referred to as pre-processed data. Raw data contains erroneous data (e.g. zero water outflow), noisy data (e.g. untrue water outflow), and/or missing data. Raw data is cleaned, smoothed, and imputed, where it becomes processed data. Then, 75% of the processed data, training set, is fed to two separate models, SARIMA model and Hybrid model. A description of the SARIMA model and the Hybrid model is presented in sections 3.1 and 3.2, respectively.

After the models are trained and calibrated, 15% of the processed data is used for testing while the reminder of 10% is deployed to validate the models. This data division configuration was used based on the guidelines of (Hunter et al. 2012). At this point, models can be fed with the time ahead input data to predict the target, water outflow.

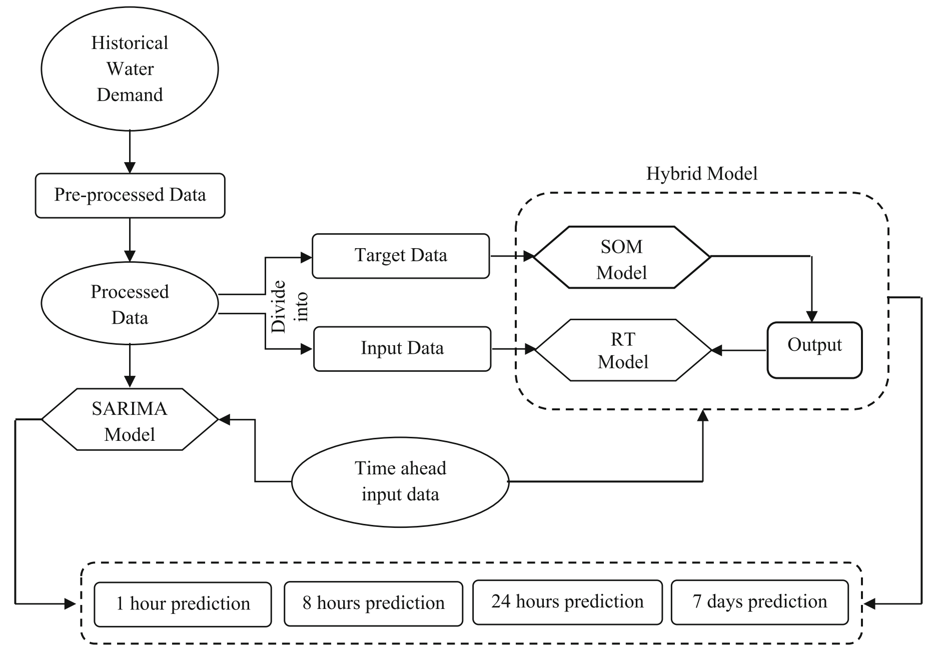


Fig. 2 Forecasting methodology flowchart. A flowchart of the steps developing, calibrating, and validating the models is shown. The historical water demand raw data is pre-processed, where missing and erroneous data are cleaned and imputed. Then, processed data is fed into the SARIMA model, whereas it is divided into two groups, target data and input data for the hybrid model. For the hybrid model, the target data is fed into the SOM model and the input data is sent directly to the RT model. In the SOM model, the target data is clustered and the output cluster number accompanied with the target data is added to the input data in the RT model. At this point, the RT model performs the prediction t time steps ahead. After predicting the target, the performance of the model is assessed (i.e. compared to the held back water demand data). If the performance is satisfactory, the model is implemented to forecast t time steps ahead. However, usually the desired performance cannot be attained from the first trial. In this case, more neurons can be added to the SOM model and/or more leaves and folds can be added to the RT model

SARIMA model

SARIMA model, denoted by ARIMA (p, d, q) x (P, D, Q) s, is a simple statistical model that is used to analyze and forecast time series data (Shumway and Stoffer 2000). The (p, d, q) non-seasonal order of the model is the number of Autoregressive (AR) parameters, differences, and Moving Average (MA) parameters. The (P, D, Q) s order of the seasonal order of the model is the AR parameters, differences, MA parameters, and periodicity. SARIMA model is formulated as (Shumway and Stoffer 2000):

$$\Phi_{\mathrm{P}}\left(\mathrm{B}^{\mathrm{S}}\right)\phi(\mathrm{B})\nabla_S^D\nabla^{\mathrm{d}}X_{\mathrm{t}} = \delta + \Theta_{\mathrm{Q}}\left(\mathrm{B}^{\mathrm{S}}\right)\theta(\mathrm{B})\,W_{\mathrm{t}} \tag{2}$$

Where,

Φ_{P} (B^{S}): is the seasonal AR parameter of order P

$\phi(\mathrm{B})$: is the ordinary non-seasonal AR parameter

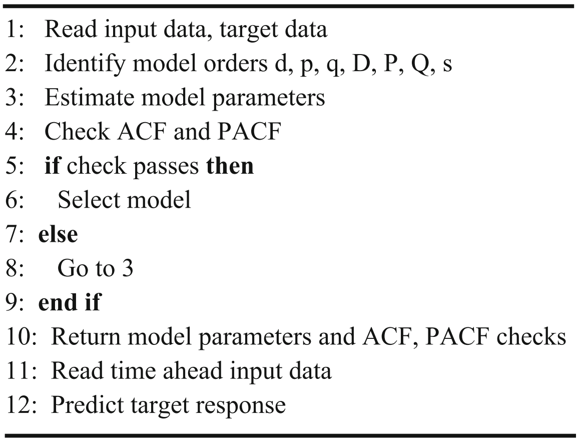
∇_S^D: is the seasonal difference component (∇_S^D = {1 - B^{S}} $^{\mathrm{D}}$)

∇^{d}: is the ordinary non-seasonal difference component (∇^{d} = {1 – B} $^{\mathrm{d}}$)

X_{t}: is the measured time series denoted by time t

δ: is the intercept

Θ_{Q} (B^{S}): is the seasonal MA parameter of order Q



```
1:  Read input data, target data
2:  Identify model orders d, p, q, D, P, Q, s
3:  Estimate model parameters
4:  Check ACF and PACF
5:  if check passes then
6:      Select model
7:  else
8:      Go to 3
9:  end if
10: Return model parameters and ACF, PACF checks
11: Read time ahead input data
12: Predict target response
```

Fig. 3 SARIMA model algorithm. This figure shows the algorithm used to develop SARIMA forecasting models. Processed data is fed to the model where the model order parameters are identified by plotting the Autocorrelation Factor (ACF) and the Partial Autocorrelation Factor (PACF). After reaching a satisfying performance, the time ahead input data is read to predict the response (i.e. the water demand)

$\theta(B)$: is the ordinary non-seasonal MA parameter

W_t: is the usual Gaussian noise process

Figure 3 shows the algorithm used to determine these parameters and develop SARIMA models in this paper. SARIMA seasonal and non-seasonal parameters are estimated iteratively through plotting the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). SARIMA is a simple traditional model that can be trained and fitted on a small dataset. Arandia et al. 2016 proposed six SARIMA models that were trained by three small datasets (24 h and 7 days windows, and 15 min, 1 h, and 24 h resolutions). A 7-day window was showed to be sufficient to train the SARIMA model. In this paper, two SARIMA models with two different training windows were fitted to forecast water demand. These models are displayed in Table 3, where they are distinguished by their seasonal period.

Hybrid model

The hybrid model in this study comprises of two models, SOM model and RT model. The practice for the proposed hybrid model is to simply feed the output of the SOM clustering model, accompanied by other desired correlated inputs, to the RT forecasting model.

SOM, also known as Kohonen Neural Networks (Kohonen 1982), is an unsupervised learning technique that reduces data dimensionality. SOM uses competitive learning to cluster input data into groups while preserving the topology and the distribution of the input data. Simply stated, an n-dimensional grid of neurons compete to win data points according to how close these points are in the input pattern. The patterns that are close in the input space will be mapped to units that are close in the output space (i.e. grid) (Bação et al. 2005). Figure 4 illustrates the mechanism of a 2-dimensional (2D) grid of 3 neurons competing to map input data into an output of 9 clusters. As shown in this figure, input data points (I_1, I_2, …, I_n) are fed to the network where a same initial weight is assigned. Each data point multiplied by its assigned weight is called a node. Then, the Euclidean distance is computed between each node and all competing

Table 3 SARIMA model structure identified

Model identifier	Resolution (hr)	f (hr^{-1})	Seasonal period (hr)	Structure
S-24	1	1	24	ARIMA (0, 1, 2) X (0, 1, 1)$_{24}$
S-168	1	1	168	ARIMA (0, 1, 2) X (0, 1, 1)$_{168}$

neuron (3 neurons in this example). The data point in this specific node is won by only the neuron with the shortest computed distance. The neurons that did not win this data point are topologically mapped using a neighbourhood function. This function determines how close each neuron should be to other competing neurons. At the end of this mapping technique and before processing the next input data point, weights are adjusted according to the previous neighbourhood topology. By the end of the training, the input data is grouped in 9 clusters that have the same topology as the input space.

In this study, 2D-SOM are developed and implemented to cluster the water outflow input data using the algorithm shown in Fig. 5a. In the hybrid model, four 2D-SOM clustering models were developed. These four models vary with the number (N) of neurons used in the 2D grid layer. For example, HYB-2 N is a hybrid model with 2 neurons in each dimension of the 2D grid layer. For our purpose, N ranged between 2 and 5. If N were to be less than 2 (i.e. $N = 1$), the resulted number of clusters is N^2 which equals to 1. This 1 cluster would basically have the same input dataset topology. Therefore, the minimum number of neurons was set to 2. The upper limit in this study is 5 neurons was the limit in this study. That is because using 6 or more neurons did not yield in any significant efficiency increase in the performance of the SOM model and the overall hybrid model.

RT is a supervised learning technique that is used for prediction. RT is the numeric outcome model of the general classification and regression tree (CART) introduced by (Breiman et al. 1984). The model is constructed with an assembly of rules based on variables extracted from the dataset (i.e. predictors). These rules are represented by values that are selected to form the best possible splits to differentiate instances (i.e.

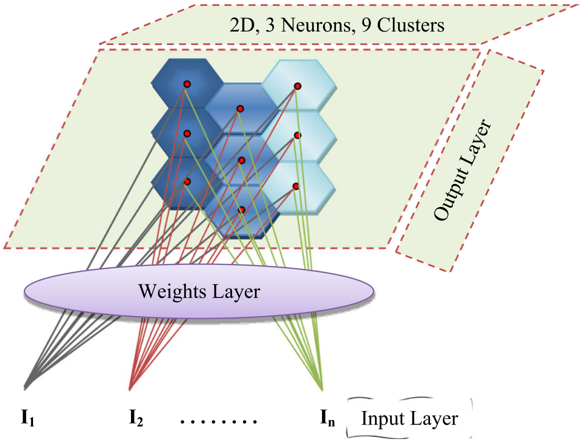


Fig. 4 2-dimentional (2D) SOM structure with number of neurons (N) equal 3. This 2D network structure presents a brief introduction to the mechanism of the SOM. The inputs are fed to the map where initial weights are assigned. Then, by calculating the Euclidean Distance, the indices are arranged in clusters. The number of resulted clusters depends on the number of neurons selected when the map is structured. Each index is won by one neuron

a

```
1:  Read input data, target data
2:  Identify model orders d, p, q, D, P, Q, s
3:  Estimate model parameters
4:  Check ACF and PACF
5:  if check passes then
6:      Select model
7:  else
8:      Go to 3
9:  end if
10: Return model parameters and ACF, PACF checks
11: Read time ahead input data
12: Predict target response
```

b

```
1:  Read input data, target data
2:  Choose training function
3:  Create the regression tree
4:  Define cross-validation function
5:  Set the number of folds
6:  Estimate predictor's importance
7:  Enable Principal Component Analysis
8:  Run performance checking
9:  if passes then
10:     Select tree
11: else
12:     Go to 5
13: end if
14: Return tree components and checks' results
15: Read time ahead input data
16: Predict target response
```

Fig. 5 a SOM model algorithm. This figure shows the algorithm used to develop SARIMA forecasting models. Processed target data is fed to the model where the model is trained, tested, and validated. After reaching a satisfying performance, the time ahead input data is read to predict the response (i.e. the cluster number). **b** RT model algorithm. This figure shows the algorithm used to develop RT forecasting models. Processed target data is fed to the model where the model is trained, tested, and validated. After reaching a satisfying performance, the time ahead input data is read to predict the response (i.e. the water demand)

observations). Once a rule, also called decision, is selected, a split is applied at a specific node. This process continues to be applied to each node in the tree through a recursive procedure. RT models are obtained by repeatedly dividing the data space and fitting a simple prediction model within each split. As a result, the data division can be represented graphically as a decision tree (Loh 2011). This splitting process continues until a predefined limit is reached. This limit could be where no further information gain can be achieved. Alternatively, splitting can be left to continue where the tree is pruned at the end of the process. Pruning is a technique that establish stopping rules to prevent the growth of tree sections that do not seem to improve the accuracy of the predicting model.

RT model development begins with feeding the input data to the tree root, then the data is filtered and sent to a branch and then to another branch until it reaches the leaf. The leaf is where the final decision is made, is called the Response. In this study, five RT models were developed to forecast 1 h, 8 h, 24 h, and 7 days ahead. The first RT model is a standalone model. This model is not fed any of the SOM output (i.e. no K inputs) as a model input. The rest of the models (HYB-N2, HYB-N3, HYB-N4, and HYB-N5) are hybridized models; all predictors are fed to the model every time the model predicts the future outflow demand. Although the standalone RT model is fed with fewer input parameters, however, all models in this group are structurally identical in terms of the input data time span, the number of tree leaves, and the cross-validation folds. The RT model algorithm shown in Fig. 5b was used to develop the RT predicting model.

Model performance

The performance of the proposed models was measured based on the deviation of the predicted outflow from the actual outflow. Both over-estimated and under-estimated predictions were considered as inaccurate model performance, therefore, included in error measurements. The performance was measured with: (1) Mean Absolute Percentage Error (MAPE) and (2) Root Mean Squared Error (RMSE). The Normalized Root Mean Squared Error (NRMSE) was calculated and shown along with MAPE in Table 4. The inclusion of NRMSE was to account for the variation of the means of datasets used in forecasting the water outflow. Equations 3, 4, and 5 represent the MAPE, RMSE, and NRMSE, respectively. Here n denotes the number of data points, $\overline{Y_i}$ is the data set mean, and $\hat{Y}_i$ and Y_i represent the forecasted and the actual water outflow, respectively.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Yi - \hat{Y}i}{Yi}\right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i^{\,2} - Y_i^2\right)} \tag{4}$$

$$NRMSE = \frac{RMSE}{\overline{Y_i}} \tag{5}$$

Table 4 Models overall performance

Model	Model identifier	MAPE (%)				NRMSE (%)			
		1 h	8 h	24 h	7 days	1 h	8 h	24 h	7 days
SARIMA	S-24	15.37	17.84	17.81	21.38	18.93	18.76	19.08	24.17
	S-168	14.85	15.72	17.61	18.73	17.31	17.95	19.02	19.83
RT	RT	11.48	12.93	13.72	16.75	12.43	13.19	17.81	21.04
Hybrid	HYB-N2	08.62	09.58	11.41	13.89	10.83	12.83	14.47	18.62
	HYB-N3	06.73	07.34	08.92	09.74	07.97	09.51	10.73	11.86
	HYB-N4	04.97	06.03	06.57	06.93	05.54	06.86	08.73	09.83
	HYB-N5	04.84	05.82	06.12	06.80	05.18	06.24	06.78	08.62

Results and discussion

Models overall performance

The results for the three proposed models forecasting 1 h, 8 h, 24 h, and 7 days ahead are shown in Table 4. It can be observed that the forecast in all time horizons have a relatively better overall performance when RT, and HYB models are deployed compared to the simple linear SARIMA models. On average, the MAPE for RT model was 15% to 25% less compared to the SARIMA models, forecasting 1 h, 8 h, 24 h, and 7 days ahead. Likewise, HYB models had a MAPE of 35% to 70% less than SARIMA models forecasting the same four time horizons ahead. As can be seen, the nonlinear models have outperformed the linear SARIMA models. That is due to the following two reasons. First, the utility's outflow has a nonlinear relationship over service time, where the change in water demand per unit time is variable. RT and HYB models are naturally nonlinear models that have a higher capability in capturing nonlinear patterns in a time series. These models perform extensive computations to extract relationships between parameters in the current, previous, and subsequent time steps. On the other hand, SARIMA models are linear models that linearly approximate a time series with a trend and seasonality regardless of the time series' nature. This process of estimating trend and seasonality parameters grows into a harder task when the time series has a random nonlinear change in patterns. Although, it is apparent that RT and HYB models have the advantage over SARIMA in this study, the time series degree of nonlinearity might change that. With a lower degree of nonlinearity, SARIMA model can better capture the time series' trend and seasonality orders, which is translated to a better performance in the forecasting application. Secondly, RT and HYB models were fed with the correlated inputs mentioned in Table 2 in addition to the water outflow. Meanwhile, SARIMA models were trained and fit with the water outflow time series only. Even though the extra correlated parameters fed to the RT and HYB models are a rearrangement time series of the water outflow or the temporal data, these reproduced time series aided models in extracting interesting nonlinear patterns.

Comparing RT and HYB models reveals that hybridizing the RT model has significantly improved its predictivity. Fusing RT model with SOM model has increased the accuracy of forecasting water outflow 1 h, 8 h, 24 h, and 7 days ahead. Table 4 shows that for forecasting water outflow 1 h ahead, MAPE for the RT model dropped by 25%, 41%, 56%, and 58% for HYB-N2, HYB-N3, HYB-N4, and HYB-N5, respectively. Similarly, with a small variance, MAPE for the 8 h, 24 h, and 7 days ahead forecast dropped by 25% to 55%, 16% to 55%, and 17% to 59%, respectively. The SOM clustering model grouped the water outflow dataset into smaller datasets with a smaller space. This has helped the hybridized RT model in extracting patterns from spaces with instances close in intensity.

Inspecting the HYB models divulges that the number of neurons in the SOM models affects its predictivity proportionally. That means HYB models with higher number of neurons performed with less forecasting error. For example, HYB-N5 with 5 neurons in the infused SOM model had a less MAPE than HYB-N4, HYB-N3, and HYB-N2 in all forecasting horizons. Here, increasing the number of neurons reaches a limit where it no longer affects the model performance. In this study, this limit was five neurons, where HYB-6 N model had approximately the same performance as HYB-N5 with a slight decrease of less than 0.5% in MAPE. This asymptote reached by the model is due to having most of datasets patterns explained by other used neurons. In other words, the model at that limit have had reduced data dimensionality to a level where the increase in information gain ratio is negligible.

The last observation that can be drawn from the overall performance results in Table 4 is regarding the time horizon. Here, a forecast time horizon refers to the number of hours ahead a water outflow is to be predicted. Water outflow is a variable function that depends on a plethora of factors. These factors are, but not limited to, the number and growth of the population, consumer type, consumption seasonality, water price, socio-economic factors, etc. All these factors vary with time and therefore have a level of uncertainty. This can be noticed in the better performance of all proposed models on shorter time horizons. For instance, MAPE increased for all models forecasting 8 h ahead compared to 1 h ahead.

Computational load

Although model overall forecasting performance is a good measure for model accuracy, it does not incorporate model deployment complexity. Therefore, two more measures were calculated and considered for model selection. The first measure is the Akaike Information Criteria (AIC), which penalizes models that use more parameters. AIC consists of two terms, likelihood and number of parameters. The second measure is the time spent during data training steps. As a control of these two measures, all models were built using the same tool. The computational tool used in this research was HP Pavilion TS 14 Notebook PC with a 1.6 GHz Intel Core i5 processor and 8 GB memory. Table 5 displays the results of AIC and training time for the seven proposed models.

AIC and training time results reveal that the SARIMA model had the lowest and most preferred performance. SARIMA-24 has had a relatively low AIC due to the shortest data span, 7 days, used to feed the model. SARIMA-168 with a slightly higher AIC and training time ranked second. RT model had a 20 -fold higher AIC, and triple training time compared to SARIMA model. This drastic increase in AIC was due to the extra inputs fed to the model which increased the number of parameters. HYB models had a 10% to 110% higher AIC, and 40% to 135% longer training time compared to the RT model. Again, the addition of the SOM model to the RT model increased the number of parameters which lead to more complex model.

Table 5 Computation load of forecasting models on training data

Model	Historical data	Identifier	Indicators	
			AIC	Time (s)
SARIMA	7 days	S-24	1281	13
	7 days	S-168	1362	16
RT	4 months	RT	26,738	41
Hybrid	4 months	HYB-N2	29,643	57
	4 months	HYB-N3	35,692	67
	4 months	HYB-N4	46,421	82
	4 months	HYB-N5	55,729	96

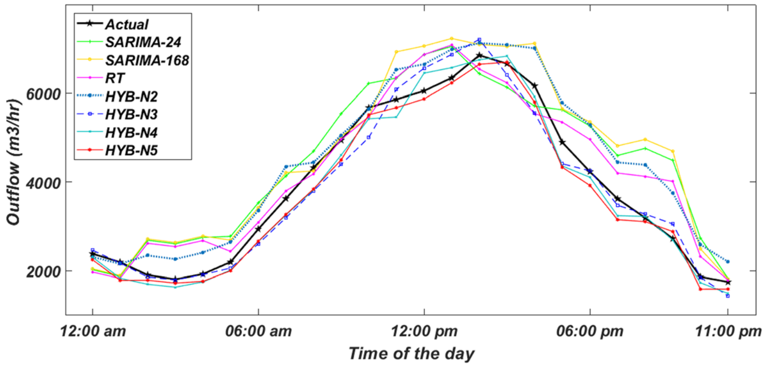


Fig. 6 Models forecasting water outflow 1 h ahead on a weekday in November 2017. This figure displays 1 h forecast results of the models when implemented and tested on a held back dataset for a random weekday on November 2017. The figure shows that Hybrid models forecasting 1 h ahead have performed better compared to SARIMA and standalone RT models. In addition, Hybrid models with more neurons outperformed those with less neurons on the tested dataset

Model application

Models predictivity was investigated using datasets that were not used in the training, testing, and validating of the proposed models. Figure 6 shows the performance of SARIMA, RT, and HYB models forecasting 1 h ahead for a held back dataset of a weekday in November 2017. Similar to the overall performance, nonlinear RT and HYB models have shown better forecasting performance compared to the linear SARIMA model. Also, when RT model performance is compared to HYB models performance, it can be seen that HYB-N5 and HYB-N4 have shown better fits over the tested time span. Here, it can be concluded again that: (1) the RT model performed better when fused with the SOM clustering model, and (2) the HYB model performance improved as the number of clusters in SOM was increased.

Figures 7 presents the results of forecasting a weekend day for SARIMA, RT, HYB models 8 h ahead. Again, HYB models performed better than SARIMA and RT models,

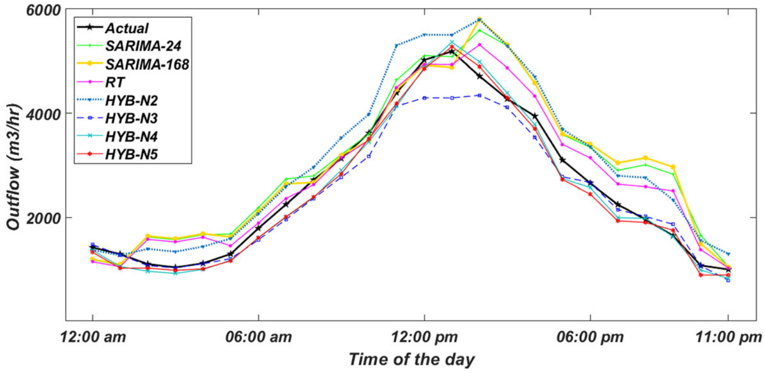


Fig. 7 Models forecasting water outflow 8 h ahead on a weekend-day in November 2017. This figure displays 8 h forecast results of the models when implemented and tested on a held back dataset for a random weekend-day on November 2017. Analogously to the 1 h forecast in Fig. 6, Hybrid models forecasting 8 h ahead have performed better compared to SARIMA and standalone RT models. Also, Hybrid models with more neurons outperformed those with less neurons on the tested dataset

and HYB models with higher number of clusters showed the best performance among proposed models.

Concluding remarks

Three models were presented to forecast a water utility outflow 1 h, 8 h, 24 h, and 1 week ahead. The main objective of this study was to investigate the influence of fusing a Supervised and Unsupervised Machine Learning techniques in the application of short-term water demand forecasting. This hybrid model (i.e. HYB) comprised of RT forecasting model and SOM clustering model. The performance of the HYB models was assessed and compared to a standalone RT model, and a SARIMA model. In virtue of its adequate performance, simple structure, and wide use in the application of short-term forecasting, SARIMA model was appended as a baseline model.

The results of this study have highlighted the following concluding remarks:

- Fusing the RT Supervised Machine Learning model with the SOM Unsupervised Machine Learning model improved models predictivity. HYB models have shown better prediction performance, less forecasting error, when compared to the standalone RT model. The Mean Absolute Percentage Error (MAPE) for HYB models was shown to be 15% to 60% less than the MAPE for the standalone RT model.
- Increasing the number of clusters in HYB models has led to a significant decrease in forecasting error. For example, the MAPE dropped, on average, by 25% when the number of clusters increased from 4 (for $N = 2$) to 9 (for $N = 3$), and by 25% when the number of clusters increased from 9 (for N = 3) to 16 (for $N = 4$). Including more clusters will eventually not affect the performance after a steady state is reached. This can be seen when HYB models with16 clusters are compared to those of 25 clusters.
- Nonlinear models (i.e. RT, and HYB) have shown better forecasting performance than the simple linear SARIMA model. However, they are more complicated to build and interpret, and perform the forecast on a slower pace.

To conclude, the HYB model would be the best selection of the investigated models if forecasting accuracy is prioritized over model simplicity. A water utility should consider the implementation of HYB model in order to obtain an accurate forecast. This significant increase in forecasting accuracy could help the water utility meet its demand requirements efficiently, increase its service reliability and customer satisfaction. In particular, water treatment and distribution processes along with maintenance and system development could be improved through optimized pumping schedules and storage. Maintenance and system development could also be improved by having a better understanding of system loads. Decreasing the demand side uncertainties through accurate prediction could also help the water utility avoid over−/under-loading the water supply system. Although this paper only investigated the model performance on short-term water demand load, the proposed model can also be applied to energy, commercial, and industrial loads.

Nomenclature

The following symbols are used in this chapter:

RT = Regression tree
SOM = Self-organizing map
ANN = Artificial neural network
SARIMA = Seasonal autoregressive integrated moving average
MAPE = Mean absolute percentage error
NRMSE = Normalized root mean squared error
SWG = Smart water grids
ARIMA = Autoregressive integrated moving average
MLP = Multi-layer perceptron
RVM = Relevance vector machine
ARIMAX = Autoregressive integrated moving average with exogenous variables
m3/hr. = Cubic meter per hour
am = Ante Meridiem (i.e. Before midday)
Q_t = The current water outflow
PCC = Pearson correlation coefficient
K = The cluster number
ACF = Autocorrelation function
PACF = Partial autocorrelation function
hr. = Hour
f = Frequency
S-24 = SARIMA model with a seasonal period equal to 24 h
S-168 = SARIMA model with a seasonal period equal to 168 h
n = A number (i.e. 1, 2, 3 … etc.)
2D = Two dimensional
I_n = Input number n
N = Number of neurons
HYB-2 N = A hybrid model with 2 neurons in each dimension of the grid
HYB-3 N = A hybrid model with 3 neurons in each dimension of the grid
HYB-4 N = A hybrid model with 4 neurons in each dimension of the grid
HYB-5 N = A hybrid model with 5 neurons in each dimension of the grid
RMSE = Root mean squared error
$\overline{Y_i}$ = Data set mean
$\widehat{Y_i}$ = Forecasted water outflow
Y_i = Actual water outflow
AIC = Akaike information criterion

Acknowledgements
Authors would like to express their sincere appreciation to Rodney Bouchard of Union Water Supply System, and Klaus Dohring of Green Sun Rising Inc. for their perpetual succor.

Conflict of interest
The authors declare no conflict of interest.

Authors' contributions
M.H.B devised the abstraction for the manuscript and performed the majority of initial and final draft development under the guidance and supervision of R.C. and D.S.-K.T. R. C delivered proficiency in water utility operations. D.S.-K.T. provided expertise in data illustration, interpretation, and model validation practices. The interactions of the

mentioned authors represented a true collaborative effort in this publication. The author(s) read and approved the final manuscript.

Funding

This research was funded mutually by the Union Water Supply Systems and the Turbulence and Energy Lab, Ontario, Canada.

Availability of data and materials

The data used in this study was obtained from a water utility that does not want this proprietary data shared outside of the approved manuscript.

Competing interests

The authors declare no competing interest.

Author details

[1]Civil and Environmental Engineering Department, Turbulence and Energy Laboratory, Ed Lumley Centre for Engineering Innovation, University of Windsor, 401 Sunset Ave, Windsor, ON, Canada. [2]Mechanical, Automotive and Materials Engineering Department, Turbulence and Energy Laboratory, Ed Lumley Centre for Engineering Innovation, University of Windsor, 401 Sunset Ave, Windsor, Ontario, Canada.

Received: 13 March 2020 Accepted: 18 May 2020
Published online: 29 May 2020

References

Abebe AJ, Price RK (2003) Managing uncertainty in hydrological models using complementary models. Hydrological Sci J 48(5):679–692

Adamowski J, Chan HF, Prasher SO, Ozga-Zielinski B, Sliusarieva A (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban demand forecasting in Montreal, Canada. Water Resour Res 48(1):W01528

Adamowski J, Karapataki C (2010) Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. J Hydrol Eng 15(10):729–743

Arandia E, Eck B, McKenna S (2016) Tailoring seasonal time series models to forecast short-term water demand. J Water Res Plann Manage 142(3):04015067

Bação F, Lobo V, Painho M (2005) Self-organizing maps as substitutes for k-means clustering, Lecture Notes. In: Sunderam VS, Albada G v, Sloot P, Dongarra JJ (eds) Computer Science, vol 3516. Springer-Verlag, Berlin Heidelberg, pp 476–483, ISSN 0302–9743

Bata M, Carriveau R, Ting D (2020) Short-Term Water Demand Forecasting Using Nonlinear Autoregressive Artificial Neural Networks (ANN). Journal of Water Resources Planning and Management 146(3):04020008. https://doi.org/10.1061/(asce)wr.1943-5452.0001165.

Bennett C, Stewart R, Beal C (2013) ANN-based residential water end-use demand forecasting model. Expert Syst Appl 40(4): 1014–1023

Boguadis J, Adamowski K, Diduch R (2005) Short-term municipal water demand forecasting. Hydrol Process 19(1):137–148

Breiman L, Friedman JH, Olstren RA, Stone CJ (1984) Classification and regression trees. Woodsworth International, Los Angeles

Cutore P, Campisano A, Kapelan Z, Modica C, Savic D (2008) Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. Urban Water J 5(2):125–132

Donkor E, Mazzuchi T, Soyer R, Roberson J (2014) Urban water demand forecasting: a review of methods and models. J Water Resour Plan Manag 140:146–159

Ghiassi M, Saidane H, Zimbra DK (2005) A dynamic artificial neural network model for forecasting time series events. Int J Forecast 21(2):341–362

Ghiassi M, Zimbra DK, Saidane H (2008) Urban water demand forecasting with a dynamic artificial neural network model. J Water Resour Plan Manag 134(2):138–146

Hippert HS, Pedriera CE, Souza RC (2001) Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans Power Syst 16(1):44–55

Hiroyuki M, Noriyuki K, Kenta I (1948-1953) Short-term load forecasting with fuzzy regression tree in power systems. In: IEEE International Conference on Systems, Man and Cybernetics, October 7, 2001 - October 10, 2001. Tucson, pp. 1948-1953. Institute of Electrical and Electronics Engineers Inc

House-Peters LA, Chang H (2011) Urban water demand modeling: review of concepts, methods, and organizing principles. Water Resour Res 47(5):W05401. https://doi.org/10.1029/2010WR009624

Hughes T (1980) Peak period design standards for small western U. S. Water Supply 16(4):661–667

Hunter D, Yu H, Pukish S, Kolbusz J, Wilamowski M (2012) Selection of proper neural network sizes and architectures—a comparative study. IEEE Trans Ind Inform 8:228–240

Jain A, Ormsbee L (2002) Short-term water demand forecast modeling techniques — conventional methods versus AI. J Am Water Works Assoc 94(7):64–72

Jain A, Varshney A, Joshi U (2001) Short-term water demand forecast modeling at IIT Kanpur using artificial neural networks. Water Resour Manag 15(5):299–321

Jamieson DG, Shamir U, Martinez F, Franchini M (2007) Conceptual design of a generic, real time, near-optimal control system for water-distribution networks. J Hydroinf 9(1):3–14

Joong K (2018) Accelerating our current momentum toward smart water. Smart Water 3:2

Kohonen T (1982) Clustering, Taxonomy, and Topological Maps of Patterns. Proceedings of the 6th International Conference on Pattern Recognition

Kozlowski E, Kowalska B, Kowalski D (2018) Water demand forecasting by trend and harmonic analysis. Arch Civil Mech Eng 18:140–148

WY. Loh, Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery (2011), 1(1): 14–23

Maidment D, Miaou S (1986) Daily water use in nine cities. J Water Resour Plan Manag 110(1):90–106

Maidment D, Miaou S, Crawford M (1985) Transfer function models of daily urban water use. Water Resour Res 21(4):425–432

Mitrea C, Lee C, Wu Z (2009) A comparison between neural networks and traditional forecasting methods: a case study. Int J Eng Bus Manag 1(2):19–24

Mori A, Takahashi A (2011) Hybrid intelligent method of relevant vector machine and regression tree for probabilistic load forecasting. In: 2nd IEEE PES international conference and exhibition on innovative smart grid technologies, ISGT Europe.

Nasseri M, Moeini A, Tabesh M (2011) Forecasting monthly urban water demand using extended Kalman filter and genetic programming. Exp Syst Appl 38(6):7387–7395

Perry P (1981) Demand forecasting in water supply networks. J Hydraulic Division 107(9):1077–1987

Public Utilities Board Singapore (2016) Managing the water distribution network with a smart water grid. Smart Water 1:4

Rice D, Carriveau R, Ting D, Bata M (2017) Evaluation of Crop to Crop Water Demand Forecasting: Tomatoes and Bell Peppers Grown in a Commercial Greenhouse. Agriculture 7(12):104

Shamseldin AY, O'Connor KM (2001) A non-linear neural network technique for updating of river flow forecast. Hydrol Earth Syst Sci 5(4):577–597

Shumway R, Stoffer D (2000) Time series analysis and its applications. Springer, Berlin

Tiwari K, Adamowski J (2013) Water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Res Research 49(10):6486–6507

Tiwari M, Adamowski J (2015) Medium term urban water demand forecasting with limited data using an ensemble waveletbootstrap machine-learning approach. J Water Resour Plan Manag 141(2):04014053

Zhang G (2001) An investigation of neural networks for linear time-series forecasting. Comput Oper Res 28(12):1183–1202

Zhou S, McMahon TA, Walton A, Lewis J (2000) Forecasting daily urban water demand: a case study of Melbourne. J Hydrol 236(3–4):153–164

Zhou S, McMahon TA, Walton A, Lewis J (2002) Forecasting operational demand for an urban water supply zone. J Hydrol 259(1–4):189–202

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.