Stochastic Environmental Research and Risk Assessment (2019) 33:1781–1792
https://doi.org/10.1007/s00477-019-01732-9

ORIGINAL PAPER

A random forest model for inflow prediction at wastewater treatment plants

Pengxiao Zhou[1] · Zhong Li[1] · Spencer Snowling[2] · Brian W. Baetz[1] · Dain Na[1] · Gavin Boyd[1]

Published online: 20 September 2019

Abstract

Influent flow of wastewater treatment plants (WWTPs) is a crucial variable for plant operation and management. In this study, a random forest (RF) model was applied for daily wastewater inflow prediction, and a new probabilistic prediction approach was, for the first time, applied for quantifying the uncertainties associated with wastewater inflow prediction. The RF model uses regression trees to capture the nonlinear relationship between wastewater inflow and various influencing factors, such as weather features and domestic water usage patterns. The proposed model was applied to the daily wastewater inflow prediction for two WWTPs (i.e., Humber and one confidential plant) in Ontario, Canada. For the confidential WWTP, the coefficient of determination ($\boldsymbol{R}^2$) values for training and testing were 0.971 and 0.722, respectively. The $\boldsymbol{R}^2$ values at the Humber WWTP were 0.957 and 0.584 for training and testing, respectively. In comparison with other approaches such as the multilayer perceptron neural networks (MLP) models and autoregressive integrated moving average models, the results show that the RF model performs well on predicting inflow. In addition, probabilistic prediction of daily inflow was generated. For the Humber station, 93.56% of the total testing samples fall into its corresponding predicted interval. For the confidential plant, 78 observed values of the total 89 samples fall into its corresponding interval, accounting for 87.64% of the total testing samples. The results show that the probabilistic approach can provide robust decision support for the operation, management, and optimization of WWTPs.

Keywords Random forest · WWTP · Wastewater prediction · Daily flow · Uncertainty analysis

1 Introduction

It is well acknowledged that the wastewater inflow to a wastewater treatment plant (WWTP) is an essential variable for plant operation and management. The rate of wastewater inflow depends on local drainage characteristics, domestic water usage patterns, and meteorological conditions (Abunama and Othman 2017; El-Din and Smith 2002; Szelag et al. 2017). In recent decades, in order to implement advanced control strategies, plant-wide monitoring networks and controlling systems have been widely used in WWTPs (Campisano et al. 2013; Dürrenmatt and Gujer 2012). A large amount of data are collected by these monitoring networks. The data collected could provide important information for wastewater inflow prediction and treatment process control. Therefore, utilizing these data to predict wastewater inflow is desired.

The accuracy of an influent flow prediction model depends on how the relationships are described in the model between inflow and various influencing factors, such as meteorological conditions, sewer system characteristics, and human factors (Amatya et al. 1997; Li et al. 2015; Pagano et al. 2009). However, these relationships are often nonlinear and complex, which leads to challenges in wastewater inflow prediction. In the past decades, alongside the development of artificial intelligence, numerous data-driven models (Table 1) have been applied to predict the inflow of WWTPs (Boyd et al. 2019; Kim et al. 2006; Szelag et al. 2017; Wei et al. 2013; Wei and Kusiak 2015; Djebbar and Kadora 1998; Zhang et al. 2018). For instance,

✉ Zhong Li
zoeli@mcmaster.ca

[1] Department of Civil Engineering, McMaster University, Hamilton, ON L8S 4L7, Canada

[2] Hydromantis Environmental Software Solutions, Inc., 407 King Street West, Hamilton, ON L8P 1B5, Canada

Table 1 Typical models for wastewater inflow prediction

S/n	Study	Method
1.	Djebbar and Kadora (1998)	Artificial neural networks (ANN)
2.	El-Din and Smith (2002)	ANN
3.	Kim et al. (2006)	Autoregression integrated moving average (ARIMA)
4.	Wei et al. (2013) and Wei and Kusiak (2015)	Multi-layer perceptron (MLP), dynamic neural networks (DNN)
5.	Kim et al. (2016)	K-nearest neighbor (KNN)
6.	Szelag et al. (2017)	Support vector machine (SVM), random forest (RF), KNN, kerner regression (K)
7.	Zhang et al. (2018)	Recurrent neural networks (RNN), nonlinear autoregressive with exogenous inputs (NARX), long short-term memory (LSTM)

El-Din and Smith (2002) used artificial neural networks (ANNs) to predict wastewater inflow during storm events. Moreover, Kim et al. (2016) proposed a k-nearest neighbor (k-NN) method to predict the influent characteristics of WWTPs. Although these methods can better solve the nonlinear problems in inflow prediction, there are still some drawbacks. For example, the ANN method often has over-learning and low speed of convergence problems (Wang et al. 2015; Yeh and Li 2002). The k-NN method is affected by the search range and could be computationally expensive as the size of the problem increases (Ponomarenko et al. 2014; Zhe Zhou et al. 2015). Additionally, these methods cannot provide information on each input variable's contribution to the inflow (Wang et al. 2015). In order to solve these problems, alternative and effective methods are still required.

More recently, random forest (RF) has gained a lot of attention as an effective predictive modeling technique. RF is an ensemble classifier, proposed by Breiman in 2001, and comprises a collection of tree-structured classifiers (Breiman 2001). RF can be regarded as a modified version of bagging, which uses a similar but improved way of bootstrapping (Gislason et al. 2006). It has certain advantages compared to the traditional bagging method in terms of accuracy and computational intensity (Breiman 2001; Gislason et al. 2006). In addition, there are variable importance measurements in the RF method, which help to determine each input variable's contribution. As a promising method, RF has been applied in a wide range of areas. For instance, Pal (2005) used a RF classifier for land cover classification. His study concluded that the RF classifier, compared with Support Vector Machines (SVMs), requires less user-defined parameters and is easier to define the parameters. Díaz-Uriarte and Alvarez de Andrés (2006) investigated the use of RF for gene selection and classification based on microarray data. The RF model showed a comparable performance to other methods such as diagonal linear discriminant analysis (DLDA), K-nearest neighbor (KNN), and SVMs. Abdel-Rahman et al. (2013) proposed a spectral band selection method for predicting sugarcane leaf nitrogen concentration using RF regression algorithm. The results showed that sugar leaf nitrogen concentration can be predicted by RF regression algorithm with a coefficient of determination (R^2) value of 0.67. Szelag et al. (2017) used several nonlinear models including RF for wastewater inflow prediction and their results indicated that RF model is competitive in comparison with SVM and KNN. Dai et al. (2018) successfully applied optimized random forest regression model for deformation monitoring of concrete dam and Zahedi et al. (2018) used the random forest regression model for predicting solid particle erosion in elbows. The RF method has been proven to be an effective method for building predictive models in many previous studies. However, the performance of using RF models in wastewater inflow prediction still needs to be demonstrated and improved through more case studies.

Therefore, the objective of this study is to explore the potential of RF for wastewater inflow prediction. This entails the following four tasks: (1) developing a data-driven model based on random forest for wastewater inflow prediction; (2) applying the developed RF model and predict the daily inflow at two WWTPs in Ontario, Canada; (3) evaluating the performance of the proposed model using different statistical criteria; (4) applying a uncertainty analysis approach to provide probabilistic inflow predictions for more robust decision support. This study will provide valuable support for WWTP management, as well as an insight into the uncertainties involved in wastewater treatment systems.

2 Methodology

2.1 Random forest

2.1.1 The principle of random forests

The RF method was proposed by Breiman, who was inspired by the papers on written character recognition, the random subspace method, and random split selection (Amit and Geman 1997; Dietterich 2000; Ho 1998). A random forest is an ensemble classifier comprising a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$, where the $\{\Theta_k\}$ are independent and identically distributed random vectors, and x is an input vector (Breiman 2001). Each tree-structured classifier is a decision tree (DT). Each DT is independently constructed during the training process using a bootstrap sample of the original data set, and each node of the DT is split using the best variable among a subset of predictors (Liaw and Wiener 2002). After the ensemble classifier is constructed and finalized, a simple majority vote or an average value is taken for prediction.

2.1.2 Regression trees

A regression tree is a prediction model that can be described as a decision tree, and it deals with the prediction of an output target y, given a vector of input variables x (Loh 2014). The output variable y of a regression tree can be continuous or discrete (e.g., the value of inflow rate in this study or the number of stations). A regression tree consists of a root node, internal nodes, and leaf nodes. A classification and regression tree (CART) approach with mean squared errors (MSE) as the node impurity criterion was used when growing a regression tree in this study. The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the number of samples; y_i is the observed value on sample i; and $\hat{y}_i$ is the predicted value on sample i. In this study, $\hat{y}_i$ equals the mean value of the samples in the node. Thus, MSE can be regarded as the variance of the samples in the node.

In this study, there was no pruning for each tree. Therefore, each leaf node was labeled with one predicted value. While there are various software applications that can be used to build RFs, in this study, the implementation of RF is conducted by *Scikit-learn* in *Python* (Fabian et al. 2011). The building process of random forests is summarized as follows and these trees are adopted for wastewater inflow prediction in this study:

1. k new training date sets are created by conducting the bootstrapping method on the original training set.
2. A regression tree is grown for each new training data set.
3. After k regression trees are formed, each regression tree produces one predicted value, and the mean value of these k values is taken as the final prediction.

2.1.3 Variable importance

RF became popular because of its numerous appealing properties, such as the significant advantages over other existing data-driven methods in terms of assessing variable importance (Grömping 2009; Tyralis et al. 2019a). Variable importance illustrates each input variable's contribution to the target during the node split (Wang et al. 2015). There are four different methods to determine the variable importance in a random forest. Readers are referred to Breiman (2002) for more details. In this study, the sum of impurity criterion decreases is used to measure the variable importance. At every node split, one variable is used to form the split and as a result, there is a decrease in the splitting criterion. The sum of all decreases in all trees due to a given variable, normalized by the total number of trees, is the sum of impurity criterion decreases (Breiman 2002). The importance of a node j on feature f in a DT k (I_{kjf}) is computed as:

$$I_{kjf} = C_j - \frac{M_{left(j)}}{M_j} * C_{left(j)} - \frac{M_{right(j)}}{M_j} * C_{right(j)} \tag{2}$$

where C_j is the measure of the impurity of the node j; $M_{left(j)}$ and $M_{right(j)}$ are the number of instances in the left and right subset of node j, respectively; M_j is the number of the instances in the node j; and $C_{left(j)}$ and $C_{right(j)}$ are the impurity of the left and right subset of node j, respectively.

The variable importance of feature f (F_f) can then be calculated as:

$$F_f = \frac{\sum_1^k \sum_1^j I_{kjf}}{k} \tag{3}$$

where k is the number of regression trees; and j is the total number of nodes in a DT.

2.2 Model development

The representativeness of training datasets is important to the effectiveness and overall performance of a RF model (Wang et al. 2015). To reflect the impacts of weather conditions and domestic water usage patterns on wastewater inflow, numerous weather observations and date/time variables are selected as predictor variables. Unlike one-step ahead or multiple-step ahead predictions (Papacharalampous et al. 2018, 2019), in this study, the historical weather data are used for model training and testing. For instance, when inflow at time t were predicted, historical weather features at time t were used as model inputs. While for real-world applications, forecasted weather data would be used. The weather features include maximum temperature (°C), minimum temperature (°C), mean temperature (°C), heating degree days (°C), cooling degree days (°C), total rain (mm), total snow (mm), total precipitation (mm), and accumulated precipitation (mm); the date/time variables are months of the year, and days of the week. One-hot encoding is applied on date/time variables firstly and then they are fed to RF model as categorical variables. Including date/time variables as input features is able to reflect data pattern, especially for time sequences date (González and Zamarreño 2005; Singh et al. 2012). For instance, water usage during weekends and holidays of domestic residents is different, which has a close relationship with the influent flow rate at wastewater treatments. More details regarding the weather features are given in Sect. 3.2. It is worth mentioning that the selection of weather features changes from one study area to another due to the different characteristics of each plant (Tehrany et al. 2013). In this study, the weather features were selected separately for each WWTP based on a correlation analysis and a literature review. A list of the selected weather features is given in Table 2. In this study, 75% of the data in the original dataset are selected randomly to generate a training dataset, while the other 25% are used to form the corresponding testing dataset. In comparison with selecting training set sequentially for a time series dataset, random selection can avoid neglecting some major data patterns. For instance, predicting the wastewater inflow in winter based on samples only from summer would lead a poor performance.

The number of trees (k), and the number of features tested at each split (m) are the two most important parameters when building a RF model. For regression problems, Probst and Boulesteix (2018) suggested that the expected out-of-bag MSE and mean absolute error decrease when increasing k; meanwhile, the first 100 trees usually achieve the biggest performance gain. In this study, the performance of the RF model becomes stable after the first 300 trees and it was found that $k = 300$ sometimes results better or equivalent performance in comparison with $k = 3000$ in terms of MSE curve. Therefore, to maintain satisfactory model performance while saving the computation time when we later use proposed model in real-world, three different numbers (300, 1000, 3000) are first assigned to k, and three different numbers (M, $Sqrt(M)$ and $\log_2 M$), where M is the total number of input features, are considered for m. Subsequently, the best combination of k and m for each WWTP can be identified using a grid search and a threefolds cross validation. Then the best combination of parameters is used to build a random forest for predicting wastewater inflow. A flowchart of the training and testing processes is shown in Fig. 1.

2.3 Evaluation of modeling performance

Four statistical criteria, including mean absolute percentage error (MAPE), root mean square error (RMSE), coefficient of determination (R^2), and Nash–Sutcliffe efficiency (NSE) are used to evaluate the performance of the RF model. MAPE is defined by Eq. 4.

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4}$$

where n is the number of samples; y_i is the observed value on sample i; $\hat{y}_i$ is the predicted value on sample i.

RMSE defined by Eq. 5 is the squared root of the MSE, which prevents positive and negative errors to cancel each

Table 2 The selected input features for the two WWTPs

Feature category	Feature	WWTP
Weather features	Maximum temperature (°C)	Humber, confidential plant
	Minimum temperature (°C)	Humber, confidential plant
	Mean temperature (°C)	Humber, confidential plant
	Heating degree days (°C)	Humber, confidential plant
	Cooling degree days (°C)	Humber
	Total rain (mm)	Humber, confidential plant
	Total snow (mm)	Humber
	Total precipitation (mm)	Humber, confidential plant
	2-day accumulate precipitation (mm)	Confidential plant
	3-day AP	Confidential plant
	4-day AP	Confidential plant
	5-day AP	Confidential plant
	6-day AP	Confidential plant
	7-day AP	Confidential plant
Date Features	Month (Jan–Dec)	Humber, confidential plant
	Workdays (Mon–Sun)	Humber, confidential plant

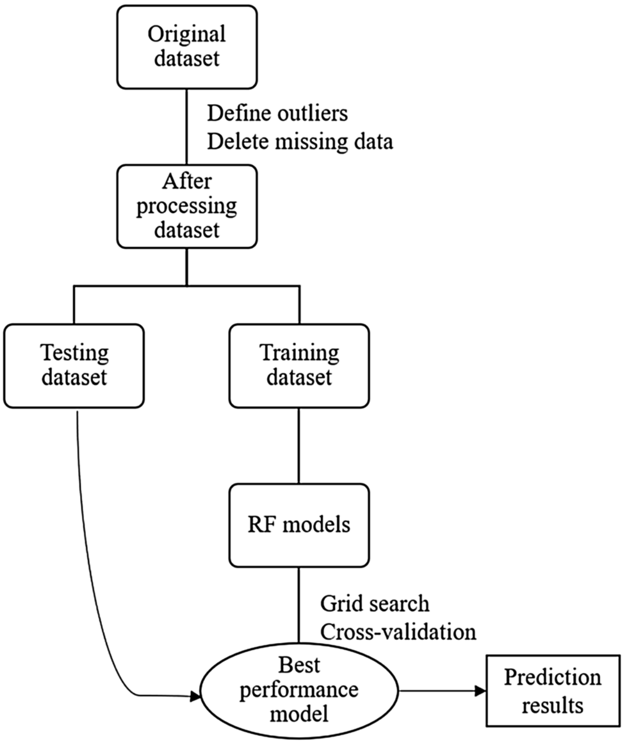

Fig. 1 Flow chart of the training and testing process

other out in order to express the error metric in the same units as the original data (Bennett et al. 2013).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

R^2, given by Eq. 6, is the squared of Pearson product-moment correlations, and measures the correlation of the observed and modeled values. R^2 ranges from 0 to 1, with 1 corresponding to the strongest correlation.

$$R^2 = \left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \tilde{y})}{\sqrt[2]{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt[2]{\sum_{i=1}^{n}(\hat{y}_i - \tilde{y})^2}}\right]^2 \tag{6}$$

where $\tilde{y}$ is the mean of predicted values; and $\bar{y}$ is the mean of observed values.

NSE (Nash and Sutcliffe 1970) defined by Eq. 7 is a widely used criterion for calibration and evaluation of hydrological models (Gupta et al. 2009). The range of NSE can vary from negative infinity to 1, which indicates a perfect fit.

$$\mathrm{NSE} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

2.4 Probabilistic prediction

Let y be the output target and x be a vector of input variables. In a RF regression model that includes k trees, each end node of a tree includes one predicted value under the no prune scenario. The predicted value of tree i is expressed as y_i. Let the probability distribution of the variable y be

$$P\{y = y_i\} = p_i = \frac{1}{k}, \quad i = 1, 2, \ldots k \tag{8}$$

For the traditional conditional mean approach, the final predicted result y is estimated using the mean of y_i which is generated by k trees and it can be expressed as follows:

$$E(y|x) = \sum_{i=1}^{k} y_i p_i, \quad i = 1, 2, \ldots k \tag{9}$$

However, the conditional mean shows only one aspect of the distribution of a target y and ignores other features; thus, this deficiency promotes the development of probabilistic prediction (Nicolai Meinshausen 2006). In fact, a probability distribution function (PDF) and a cumulative distribution function (CDF) of the target y can be generated using the predicted results from k trees. Meanwhile, the probability of target y that does not exceed one certain threshold (Y) can be calculated.

$$P(y \leq Y|x) = \sum_{i=1}^{k}(p_i|y_i \leq Y), \quad i = 1, 2, \ldots k \tag{10}$$

Moreover, an interval prediction based on the quantiles of target y can be built. To evaluate the accuracy of the predicted intervals, a criterion which is specifically for scoring predicted interval is introduced as follows (Gneiting and Raftery 2007; Dunsmore 1968; Winkler 1972). For a central $(1 - \alpha) \times 100\%$ prediction interval, where α is the quantile level and $\alpha \in (0, 1)$, the upper and the lower bounds of this prediction interval are the predictive quantiles at levels $1 - \frac{\alpha}{2}$ and $\frac{\alpha}{2}$, respectively. And the scoring criterion is expressed as:

$$S_{\alpha}^{int}(l, u; x) = (u - l) + \frac{2}{\alpha}(l - x)1\{x < l\} + \frac{2}{\alpha}(x - u)1\ \{x > u\} \tag{11}$$

where u and l are values representing the $1 - \frac{\alpha}{2}$ and $\frac{\alpha}{2}$ quantiles, respectively, and x is the observed value.

3 Case study

3.1 Study area

Two wastewater treatment plants in Ontario, Canada (i.e., the Humber WWTP and one confidential WWTP), were used to demonstrate the applicability and performance of the proposed RF model. The Humber WWTP is situated on the mouth of the Humber River, and is Toronto's second largest WWTP. It serves a population of approximately 680,000 with a capacity of 473,000 m^3/d (www.toronto.ca/services-payments/water-environment/). The confidential WWTP serves a population of approximately 141,500, and it consists of preliminary treatment, primary treatment, secondary treatment and tertiary treatment. This confidential WWTP is designed to collect only sanitary sewage. However, a significant amount of flow in the sanitary sewer system originates from sources like downspouts and illegal sump pump connections during storm events, and infiltration during rainfall events.

3.2 Data

The influent flow data was obtained from Hydromantis Environmental Software Solutions, Inc., a software development company in the water and wastewater treatment sector. For the Humber WWTP, daily flow data from January 2, 2015 to December 31, 2017 were used. For the confidential WWTP, flow daily data from November 1, 2015 to October 30, 2016 were collected. Time-series flow plot for the Humber WWTP and the confidential WWTP are presented in Fig. 2.

The weather data were obtained from Weather Canada (https://weather.gc.ca/canada_e.html). The weather data were collected and matched with the corresponding flow data with the same data length and frequency. The weather variables include maximum temperature (°C), minimum temperature (°C), mean temperature (°C), heating degree days (°C, defined by Eq. 12), cooling degree days (°C, defined by Eq. 13), total rain (mm), total snow (mm), and total precipitation (mm).

$$HDD = (1day)\sum_{days}(T_b - T_m)^+ \qquad (12)$$

$$CDD = (1day)\sum_{days}(T_m - T_b)^+ \qquad (13)$$

where T_b is the base temperature; T_m is the daily mean temperature; and the plus signs indicate that only positive values count (Büyükalaca et al. 2001).

4 Result analysis and discussion

4.1 Modeling performance

A RF model was built for each of the two WWTPs using the approach described above. For the Humber station, the original dataset had a total of 1080 samples. Outliers in the original dataset were detected using the three-standard deviation (3σ) method and samples that included missing values were deleted, which resulted in a total of 1053 samples. After pre-processing, 789 data points were selected randomly to form the training set, and the remaining 264 data points were used for testing. The model with the best training results had 3000 trees, and the number of the features tried at each split m was equal to $\log_2 M$. At the confidential WWTP, flow data from

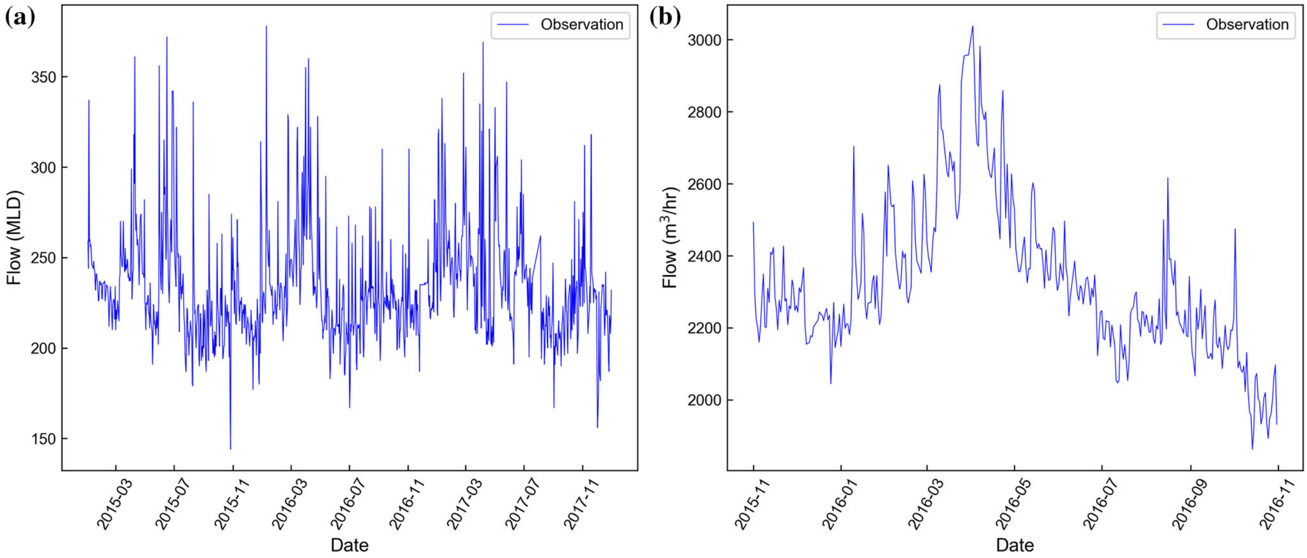

Fig. 2 Time-series flow graph for the Humber (**a**) and the confidential plant (**b**)

November 1, 2015 to October 30, 2016 were collected. Outliers in the data were identified manually after consulting the engineers at the WWTP, and samples with missing values were deleted. The pre-processing resulted in a total of 359 data points. 269 data points were randomly selected as training data, and the remaining 90 points were used as testing data. After using the grid search method, it was found that the best performance model had the number of the trees k equal to 1000, and the number of features tried at each split m was equal to M. The results of MAPE, RMSE, R^2 and NSE, as well as the scatter plots of the predicted and observed flows generated by the RF model for each plant are illustrated in Fig. 3.

Generally, the effectiveness of hydrologic models can be estimated by statistical parameters, such as NSE and R^2. The required minimum value of NSE is 0.5, and R^2 with values greater than 0.5 are considered acceptable (Mello et al. 2008; Moriasi et al. 2007). In addition, scatter plots were employed as NSE alone is not an adequate indicator (Jain and Sudheer 2008). In this study, according to the values of NSE and R^2, the proposed RF models for the Humber station and the confidential station are considered satisfactory.

Furthermore, to evaluate the performance of the proposed RF model, other algorithms used in previous studies, including multi-layer perceptron (MLP) and autoregression integrated moving average (ARIMA) are compared with RF. MLP is a classical artificial neural networks model. It has been used in many disciplines and has been proved to be a useful predictive model (Olmedo et al. 2018). ARIMA is a time series analysis model that has been used for several decades and has been used for wastewater inflow prediction (Boyd et al. 2019). Although RF, MLP, and ARIMA are all able to predict wastewater inflow, there are some significant differences in terms of categories of input variables and formats of inputs. For instance, ARIMA is an autoregression model which heavily depends on prior data and thus has a higher requirement about the data

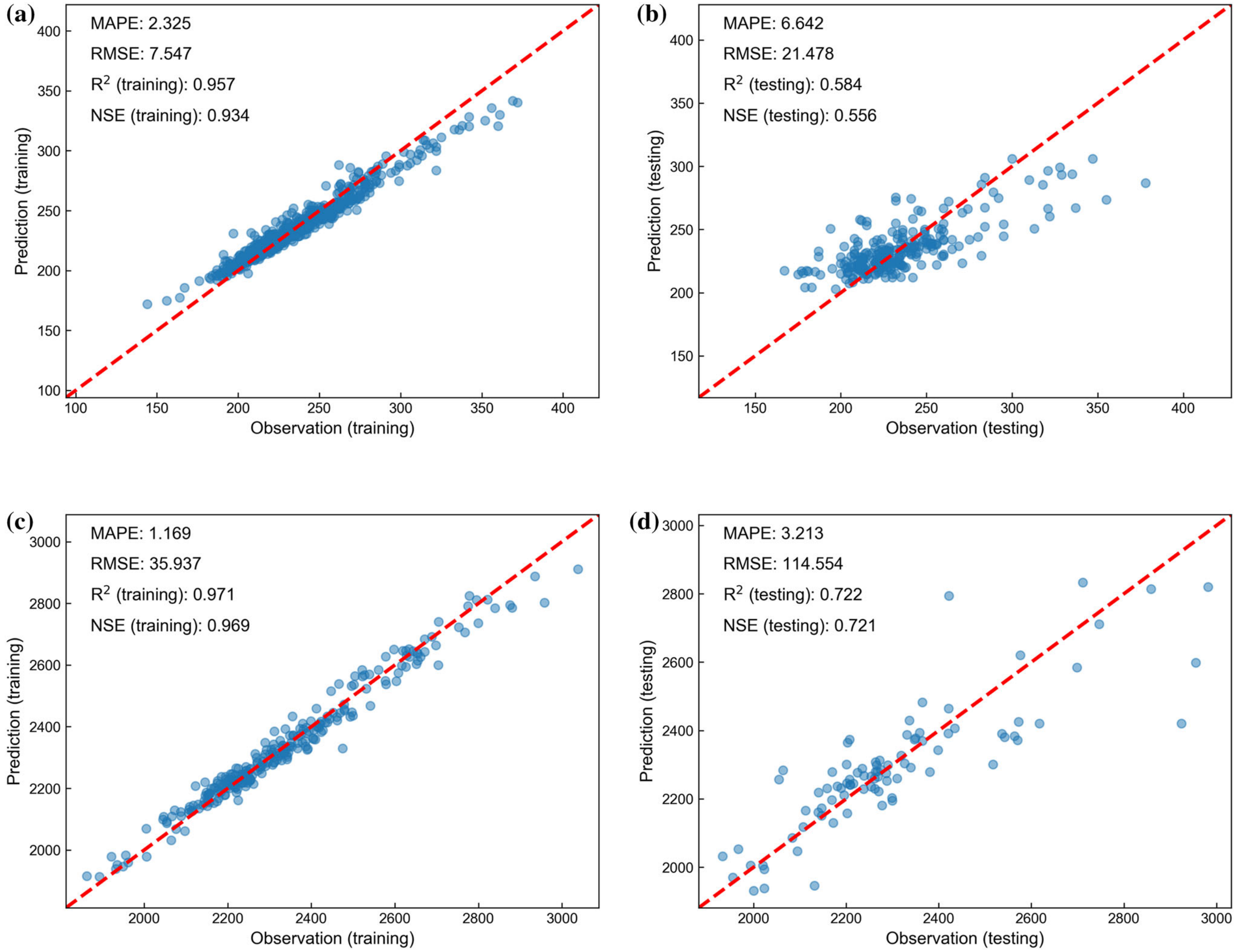

Fig. 3 Scatter plots of RF: **a** training of the Humber station, **b** testing of the Humber station, **c** training of the confidential plant, **d** testing of the confidential plant

continuity. Additionally, among these three methods, only the RF model can address continuous and categorical input variables simultaneously (Tyralis et al. 2019a). In this study, to compare performance of different models, MAPE and R^2 are used. The performance of the MLP model for each WWTP is presented in Fig. 4

Overall, the results illustrate that RF can predict the wastewater inflows competently. Compared with ARIMA, the RF model for the confidential station shows outstanding performance. Although the RF model for the Humber station is not as good as the confidential station with regards to NSE and R^2, the MAPE value (6.623) is lower than that of the ARIMA model (8.012) (Abunama and Othman 2017). Additionally, the performance of RF models in this study are more stable among different stations when compared with the ARIMA model by Boyd et al. (2019). It is worth mentioning that RF's capability of capturing extreme values (high and low values) is not as good as MLP in this study, although the overall results of RF are slightly better than MLP. This may be because the range of prediction results of RF model is determined by the range of training dataset. The RF model in this study can not produce a prediction result which exceeds the range of training data. To better predict the extreme values, there are some methods that could be further tested in future studies. For example, improving data quality and including more peak flow events while training the model could be helpful. In this study, randomly selecting training data points from time series dataset instead of selecting sequentially is a way to increase the possibility of covering more peak points at different timestamps. Additionally, developing separate models for dry and wet seasons, as well as wavelet transformation, could also help enhance the model's performance in capturing the peak values (Jothiprakash and Kote 2011; Tiwari and Chatterjee 2011).

4.2 Variable importance analysis

The variable importance was calculated using the sum of the MSE decrease as described in Sect. 2.1.3. This provides

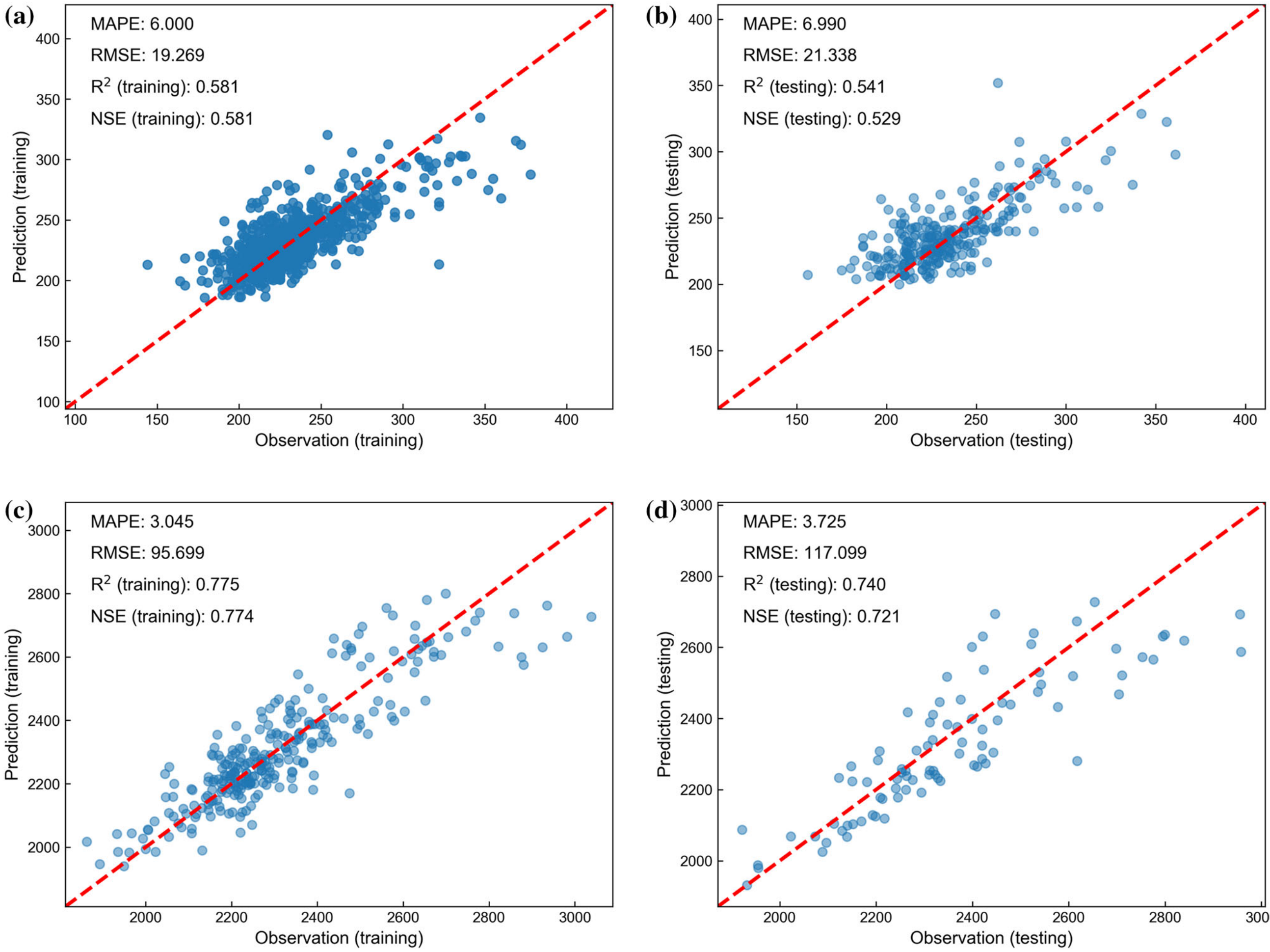

Fig. 4 Scatter plots of MLP: **a** training of the Humber station, **b** testing of the Humber station, **c** training of the confidential plant, **d** testing of the confidential plant

valuable support for decision makers to understand each variable's contribution to the flow volume. Figure 5 shows the variable importance of each station. For the Humber station, it is shown that 2-day accumulative precipitation (2DAP) and the 3-day accumulative precipitation (3DAP) are the main contributing factors. This is consistent with the work of El-Din and Smith (2002), where the authors suggested that the influent flow to a WWTP may increase substantially during storm events. However, the results of variable importance for the confidential plant imply a very different pattern. The month of the year has the highest variable importance. It is worth mentioning that the input variables used for these two stations are different. When using the Humber input variables for the prediction of the confidential plant, although month of the year is still the most important variable, the testing results are worse, with a R^2 value of 0.589. If month of the year is not included as an input, the goodness of fit would be even worse. This illustrates that the selection of input variables has a significant impact on the model performance. Variable importance metrics (VIMs) can help select the most relevant input information (Wang et al. 2015). For example, variables with VIMs around zero could be excluded (Tyralis et al. 2019a). In addition, the integrated tool proposed by Tyralis et al. (2019b) which uses random forests and linear models can also help find important predictors. The RF model could also perform better if recently lagged predictors are used (Tyralis and Papacharalampous 2017). It is recommended to carefully select input variables through literature review, system characterization, and correlation analysis when building a RF model.

4.3 Probabilistic prediction and uncertainty analysis

As an example, the PDF and CDF graphs at a randomly selected point from the confidential plant's testing dataset are presented in Fig. 6. Following the traditional RF modeling approach as described in Sect. 2.1.2, the predicted value is 2383.7 m^3/h. Using the proposed probabilistic prediction approach for the final prediction value, the PDF graph illustrates that the probability at around 2350 m^3/h is the highest. Furthermore, the CDF graph shows that the cumulative probability that inflow is less than or equal to 2300 m^3/h is zero, while that for an inflow of greater than or equal to 2450 m^3/h is one. This implies that the range of the predictive values is [2300, 2450] m^3/h. Additionally, with the CDF graph, the probability that the predicted inflow exceeds a certain threshold can be assessed. For instance, from the CDF graph shown in Fig. 7, the corresponding accumulative probability of flow at 2400 m^3/h is approximately 0.7. Thus, the probability that the predicted inflow exceeds 2400 m^3/h is approximately 0.3. To summarize, the CDF graph can provide probability information about the risk of extreme inflow for each time step. Hence, knowing the probability of extreme events occurring will better support with the management and operation of WWTPs.

To provide more information about the inflow for WWTPs, the predicted daily inflow interval prediction results for the Humber station and the confidential station during the testing period are presented in Fig. 7. Results of the scoring criterion for interval prediction mentioned in

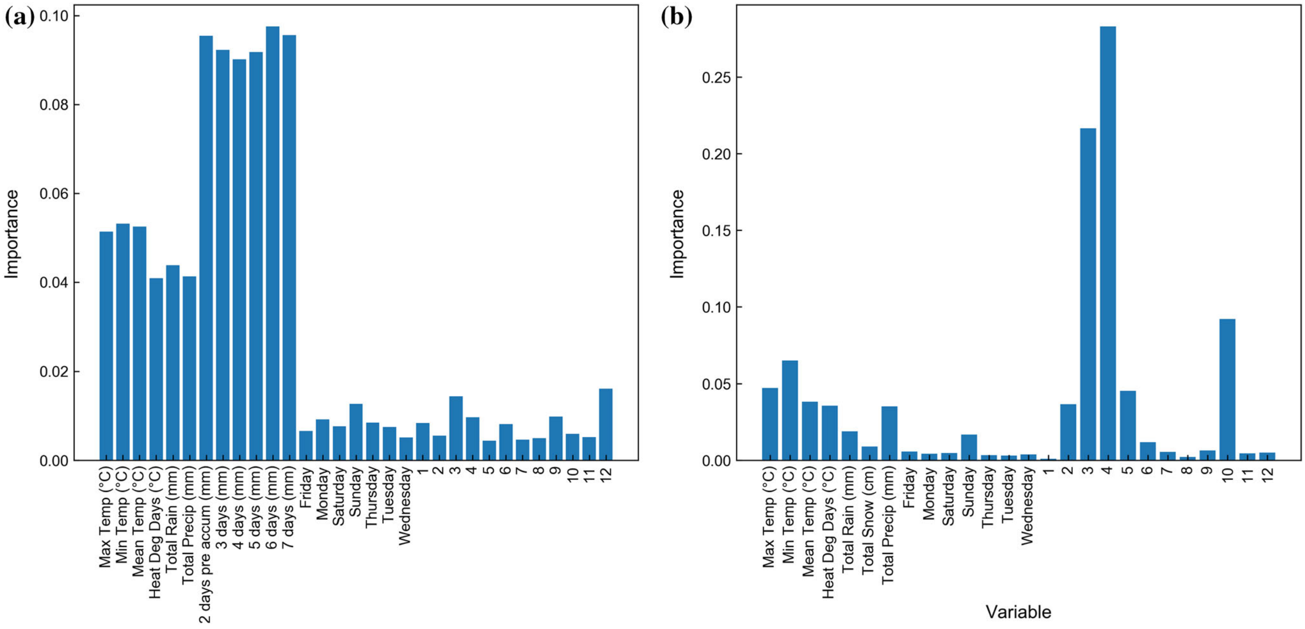

Fig. 5 Variable importance for Humber (**a**) and confidential plant (**b**)

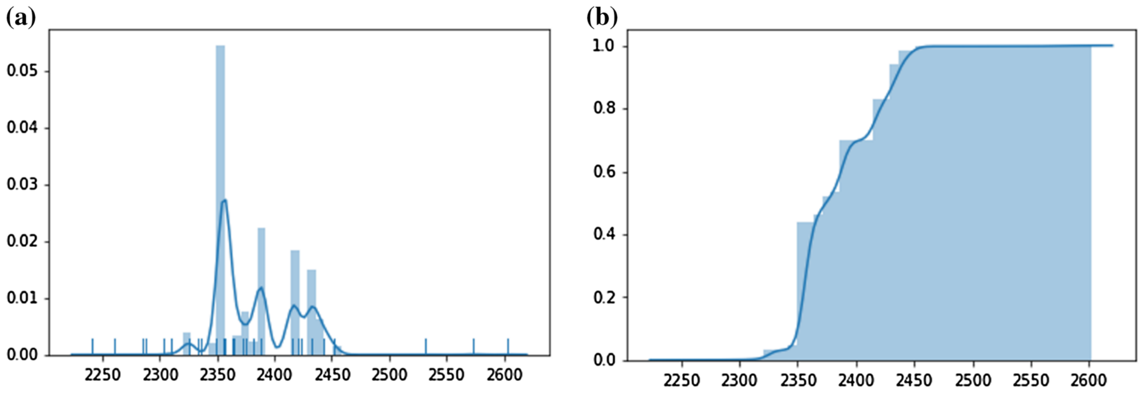

Fig. 6 Probability distribution function (**a**) and accumulative distribution function (**b**)

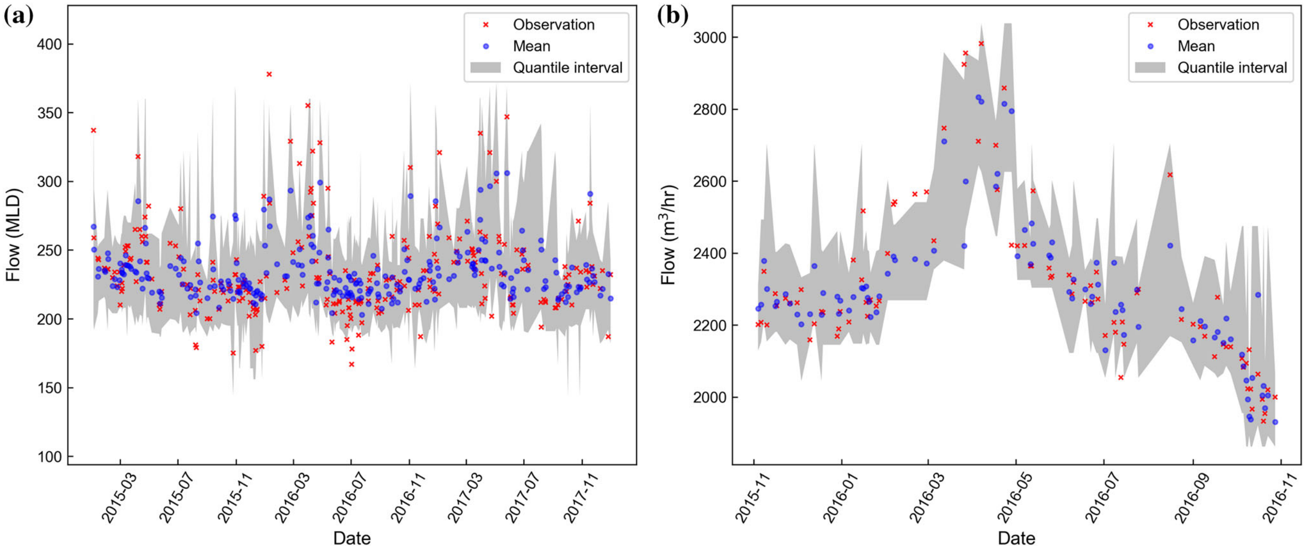

Fig. 7 Range prediction for the Humber station (**a**) and confidential plant (**b**)

Sect. 2.4 is presented in Table 3. Additionally, to cancel the effects of measurement unit, the average interval score divided by an average inflow rate is presented as well.

Overall, though the predicted intervals are slightly large for some data points, the interval prediction results capture almost all the observed flow values within the interval ranges. For the Humber station, among the 264 samples used for testing, 247 samples of their observed values fall into its corresponding interval generated by the RF model, accounting for 93.56% of the total testing samples. For the confidential plant, 78 observed values of the total 89 samples fall into its corresponding intervals, accounting for 87.64% of the total testing samples. Moreover, it can be observed from Fig. 7 that all the upper and lower bounds of interval prediction for Humber seem like relatively stable in comparison to the confidential plant and the interval coverage of Humber is higher than that of the confidential station. For the confidential plant, the large variance of bounds may be explained by the measurement unit. Most of the observed influent flow values of Humber fall into the range from 150 MLD to 350 MLD, whereas for the confidential plant, the observed influent flow values change from 1750 to 4000 m^3/h. The interval prediction was generated by results from *k* trees, and not all the trees were built using proper samples. Thus, some trees may become disturbances, which may lead to a large variance. After cancelling the unit effect, the interval prediction for the confidential station shows slightly better performance (10.60%) than that of the Humber (13.88%).

This work is the first attempt to analyze the uncertainty of predicted wastewater inflow using this probabilistic method. In this case study, most of the testing points fell into the predicted interval. The interval prediction results combined with CDF graph analysis can not only provide range solutions of the predicted wastewater inflows, but also identify the probability of inflows exceeding a certain threshold. Thus, this strategy offers an excellent support for decision-makers and operators of WWTPs, especially

Table 3 Score of central 95% interval inflow prediction for each WWTP

WWTP	Coverage (%)	Average interval score	$\frac{\text{Average interval score}}{\text{Average inflow rate}}$ (%)
Humber station	93.56	32.91 (MLD)	13.88
The confidential station	87.64	247.73 (m^3/h)	10.60

during extreme weather events and domestic water consumption rush hours.

5 Conclusions

In this study, a RF model was applied for wastewater inflow prediction at WWTPs. A RF model is an ensemble model which comprises a collection of DTs. This model shows its significant potential for wastewater inflow prediction, as it analyzes each input variable's contribution and provide valuable probabilistic prediction results. The proposed model could address the nonlinear relationships between the influent flow of WWTPs and various influencing factors such as weather features, and domestic water usage patterns. In addition, a new probabilistic prediction method was applied to quantify the uncertainties with RF predictions and thus, provide more robust support for the operation and management of WWTPs.

The proposed model was applied to predict the daily influent flow at the Humber and the confidential WWTPs in Ontario, Canada. The R^2 values for the Humber station and the confidential plant for training were 0.957 and 0.971, respectively; while those for testing were 0.584 and 0.722, respectively. The results demonstrate that the RF models could perform well for wastewater inflow prediction. Compared to other inflow prediction models such as the ARIMA and MLP, the RF model has the advantage of determining each variable's contribution, an important factor for decision-makers. Furthermore, using the proposed uncertainty analysis approach, the PDF and CDF of wastewater inflow at each time step were generated. This can provide decision-makers with more information about the risks of extreme inflows. Performance of the RF regression model could be enhanced by increasing the quality and quantity of input data. For future studies, the RF model's capability for predictions with a higher temporal resolution (e.g., hourly prediction) should be further investigated.

Acknowledgements This study was supported by the Southern Ontario Water Consortium and the Natural Science and Engineering Research Council of Canada. The authors would like to thank the contributions of engineers at Hydromantis and the WWTPs for their suggestions and comments.

References

Abdel-Rahman EM, Ahmed FB, Ismail R (2013) Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 hyperion hyperspectral data. Int J Remote Sens 34(2):712–728

Abunama T, Othman F (2017) Time series analysis and forecasting of wastewater inflow into Bandar Tun Razak Sewage Treatment Plant in Selangor, Malaysia. In: IOP conference series: materials science and engineering, vol 210(1)

Amatya DM, Skaggs RW, Gregory JD (1997) Evaluation of a watershed scale forest hydrologic model. Agric Water Manag 32(3):239–258

Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Comput 9(7):1545–1588

Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V (2013) Characterising performance of environmental models. Environ Model Softw 40:1–20

Boyd G, Na D, Li Z, Snowling S, Zhang Q, Zhou P (2019) Influent forecasting for wastewater treatment plants in North America. Sustainability 11(6):1764

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L (2002) Manual on setting up, using, and understanding random forests v3.1. http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf

Büyükalaca O, Bulut H, Yılmaz T (2001) Analysis of variable-base heating and cooling degree-days for Turkey. Appl Energy 69(4):269–283

Campisano A, Cabot Ple J, Muschalla D, Pleau M, Vanrolleghem PA (2013) Potential and limitations of modern equipment for real time control of urban wastewater systems. Urban Water J 10(5):300–311

Dai B, Gu C, Zhao E, Qin X (2018) Statistical model optimized random forest regression model for concrete dam deformation monitoring. Struct Control Health Monit 25(6):1–15

Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinform 7(1):3

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and ranomization. Mach Learn 40(2):139–157

Djebbar Y, Kadora PT (1998) Estimating sanitary flows using neural networks. Water Sci Technol 38(10):215–222

Dunsmore IR (1968) A bayesian approach to calibration. J R Stat Soc 30(2):396–405

Dürrenmatt DJÔ, Gujer W (2012) Data-driven modeling approaches to support wastewater treatment plant operation. Environ Model Softw 30:47–56

El-Din AG, Smith DW (2002) A neural network model to predict the wastewater inflow incorporating rainfall events. Water Res 36(5):1115–1126

Fabian P, Gael V, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27(4):294–300

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

González PA, Zamarreño JM (2005) Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy Build 37(6):595–601

Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. Am Stat 63(4):308–319

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol 377(1–2):80–91

Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844

Jain SK, Sudheer KP (2008) Fitting of hydrologic models: a close look at the Nash–Sutcliffe index. J Hydrol Eng 13(10):981–986

Jothiprakash V, Kote AS (2011) Improving the performance of data-driven techniques through data pre-processing for modelling daily reservoir inflow. Hydrol Sci J 56(1):168–186

Kim JR, Ko JH, Im JH, Lee SH, Kim SH, Kim CW, Park TJ (2006) Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model. Water Sci Technol 53(4–5):185–192

Kim M, Kim Y, Kim H, Piao W, Kim C (2016) Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. Front Environ Sci Eng 10(2):299–310

Li Z, Huang G, Han J, Wang X, Fan Y, Cheng G, Zhang H, Huang W (2015) Development of a stepwise-clustered hydrological inference model. J Hydrol Eng 20(10):04015008

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22

Loh WY (2014) Classification and regression tree methods. Wiley StatsRef: Statistics Reference Online

Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7:983–999

Mello CR, Viola MR, Norton LD, Silva AM, Weimar FA (2008) Development and application of a simple hydrologic model simulation for a Brazilian headwater basin. CATENA 75(3):235–247

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50(3):885–900

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 10(3):282–290

Olmedo MTC, Paegelow M, Mas JF, and Escobar F (eds) (2018) Geomatic approaches for modeling land change scenarios. Springer, Switzerland

Pagano TC, Garen DC, Perkins TR, Pasteris PA (2009) Daily updating of operational statistical seasonal water supply forecasts for the Western U.S. J Am Water Resour Assoc 45(3):767–778

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222

Papacharalampous G, Tyralis H, Koutsoyiannis D (2018) One-step ahead forecasting of geophysical processes within a purely statistical framework. Geosci Lett 5(1):1–19

Papacharalampous G, Tyralis H, Koutsoyiannis D (2019) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stochast Environ Res Risk Assess 32(2):481–514

Ponomarenko A, Avrelin N, Naidan B, Boytsov L (2014) Comparative analysis of data structures for approximate nearest neighbor search. In: Data analytics, pp 125–130

Probst P, Boulesteix A-L (2018) To tune or not to tune the number of trees in random forest? J Mach Learn Res 18:1–18

Singh RP, Gao PX, Lizotte DJ (2012) On hourly home peak load prediction. In: 2012 IEEE 3rd international conference on smart grid communications, SmartGridComm 2012. IEEE, pp 163–166

Szelag B, Bartkiewicz L, Studziński J, Barbusiński K (2017) Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear. Arch Environ Protect 43(3):74–81

Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 504:69–79

Tiwari MK, Chatterjee C (2011) A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. J Hydroinform 13(3):500–519

Tyralis H, Papacharalampous G (2017) Variable selection in time series forecasting using random forests. Algorithms 10(4):114

Tyralis H, Papacharalampous G, Langousis A (2019a) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910

Tyralis H, Papacharalampous G, Tantanee S (2019b) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. J Hydrol 574:628–645

Wang Z, Lai C, Chen X, Yang B, Zhao S, Bai X (2015) Flood hazard risk assessment model based on random forest. J Hydrol 527:1130–1141

Wei X, Kusiak A (2015) Short-term prediction of influent flow in wastewater treatment plant. Stoch Env Res Risk Assess 29(1):241–249

Wei X, Kusiak A, Sadat HR (2013) Prediction of influent flow rate: data-mining approach. J Energy Eng 139:118–123

Winkler RL (1972) A decision-theoretic approach to interval estimation. J Am Stat Assoc 67(337):187–191

Yeh AG, Li X (2002) Urban simulation using neural networks and cellular automata for land use planning. In: Advances in spatial data handling. pp 451–464.

Zahedi P, Parvandeh S, Asgharpour A, McLaury BS, Shirazi SA, McKinney BA (2018) Random forest regression prediction of solid particle Erosion in elbows. Powder Technol 338:983–992

Zhang D, Martinez N, Lindholm G, Ratnaweera H (2018) Manage sewer in-line storage control using hydraulic model and recurrent neural network. Water Resour Manag 32(6):2079–2098

Zhou Z, Wen C, Yang C (2015) Fault detection using random projections and k-nearest neighbor rule for semiconductor manufacturing processes. IEEE Trans Semicond Manuf 28(1):70–79