DEVELOPMENT OF DATA-DRIVEN MODELS FOR INFLUENT PREDICTION AT WASTEWATER TREATMENT PLANTS

By PENGXIAO ZHOU, B.E., B.Admin.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Applied Science

Master of Applied Science (2019)	McMaster University
(Civil Engineering)	Hamilton, Canada

TITLE:	Development of Data-Driven Models for Influent Prediction at Wastewater Treatment Plants
AUTHOR:	Pengxiao Zhou B.E., B.Admin. (Jilin University)
SUPERVISOR:	Zoe Li
NUMBER OF PAGES:	x, 70

ABSTRACT

Influent flow rate is essential to the operation and management of wastewater treatment plants (WWTPs). To support safe operation and effective management of WWTPs, a number of process-driven models were previously built for predicting the influent flow rate. However, in order to capture the complex nonlinear relationships in wastewater systems, these process-driven models require large-scale monitoring and complicated parameter tuning. In this research, to address those drawbacks, data-driven models are investigated for influent flow rate prediction. Three data-driven models, including multilayer perceptron (MLP), long short-term memory (LSTM) network, and random forest (RF), are introduced and developed. The developed models are applied to three WWTPs in Canada for influent flow rate prediction to demonstrate their applicability. Influent flow rate prediction with two temporal resolutions (i.e., daily and hourly) are provided. The results show that the proposed models have an overall good performance, especially the RF model. For both temporal resolutions, the performance of RF models is stable and satisfactory. In addition, an uncertainty analysis approach for the RF model is developed to provide more robust predictions. To the author's knowledge, this is the first Canadian study of wastewater influent flow rate prediction based on advanced data-driven techniques. The high temporal resolution prediction and the probabilistic prediction approach proposed in this research represent a unique contribution to methodologies related to wastewater modeling. This research can provide valuable support for WWTPs to improve operational efficiency and management effectiveness.

ACKNOWLEDGMENTS

The completion of the thesis is attributed to many people's support and encouragement. First of all, I would like to extend my sincere gratitude to my supervisor, Dr. Zoe Li, whose patient guidance, valuable suggestions, and constant encouragement make me successfully complete this thesis. She gives me much help throughout the process of selecting the research topic, writing the thesis, improving the outline and argumentation, and correcting grammatical errors, which has made my accomplishments possible.

I would like to thank Drs Yiping Guo, Mohamed Hussein, and Moataz Mohamed for serving on my Thesis Committee. I am also grateful for other faculty and staff at McMaster University. Their enlightening teaching provides me with a solid foundation to accomplish this thesis and their instructions help broaden my horizon and enrich my knowledge.

Special thanks should go to my friends and colleagues who provide me with valuable advice and support.

Finally, I would like to thank my parents for their absolute support and encouragement throughout these years.

PUBLICATIONS

This M.A.Sc thesis is organized in a sandwich style based on the following submitted/draft priori to submission papers:

1. **Pengxiao Zhou**, Zhong Li, Spencer Snowling, Brian Baetz, Dain Na, Gavin Boyd, “Development of a Random Forest Model for Inflow Prediction at Wastewater Treatment Plants” submitted on March 2019 to *Environmental Research and Risk Assessment*. Pengxiao Zhou processed the data, performed the computations and analysis, designed the figures, and drafted the manuscript. He made a significant original contribution and is the first author.

2. **Pengxiao Zhou**, Zhong Li, Spencer Snowling, Rajeev Goel, Qianqian Zhang, “Hourly Wastewater Influent Flow Prediction at Wastewater Treatment Plants” draft priori to submission. Pengxiao Zhou contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. He made a significant original contribution to this paper and is the first author.

CONTENTS

LIST OF FIGURES AND TABLES

ABBREVIATIONS

CDF	Cumulative Distribution Function
KNN	K-Nearest Neighbor
LSTM	Long Short-term Memory
MLP	Multilayer Perceptron
PDF	Probability Density Function
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machine
WWTPs	Wastewater Treatment Plants

CHAPTER 1 - INTRODUCTION

1.1 Background

Although often thought of as ordinary, water is the most valuable and remarkable substance. We are about two-thirds of water and life could not evolve without water (Chaplin 2001). Fortunately, the earth is originally allotted a finite amount of water. To sustain life, we must preserve and protect the water supply, and also purify and reuse the water we pollute (Chan 2006). As a result, wastewater treatment plants (WWTPs) emerge. Historically, the purpose of wastewater treatment is to protect the health and well being of communities (Chan 2006). With the continuous improvement of living standards and the awakening of the consciousness of protecting water resources, the purpose of wastewater treatment processes is not changed. However, the tasks of the wastewater treatment plants are becoming increasingly heavy and the effective management of wastewater treatment plants becomes more important.

Influent characteristics are important parameters for the effective management and the stable operation of WWTPs. With the development of monitoring techniques, some influent characteristics, such as influent flow rate and chemical oxygen demand (COD), could be monitored on a real-time basis. However, real-time monitoring is limited by a lack of adequate equipment, high cost, and it is relatively immature and unstable for some characteristics, such as biochemical oxygen demand (BOD) (Kim et al. 2015). Meanwhile, although online monitoring could provide essential information for assessing real-time influent characteristics, it is still hard to set aside operation time for WWTP engineers to cope with the quality and quantity fluctuations. Therefore, it is desired to develop reliable wastewater influent prediction models for process control and plan, so that WWTPs can meet discharge permit limits in an efficient way.

1.2 Wastewater Influent Prediction Models

Since 1960s, a number of mathematical models have been developed for predicting influent flow characteristics at WWTPs. Most of these models are process-driven models (also known as knowledge-driven or physically-based models), which predict the quality and quantity of wastewater influent based on the simulation of wastewater collection systems. The process-driven models are generally used as a tool to increase the knowledge on the process and system behavior for optimization and process control (Langergraber et al. 2004). However, the performance of process-driven models might vary depending on the target, expertise available and resource spent; furthermore, these process-driven models requires complicated parameter tuning and large-scale monitoring (Kim et al. 2015; Langergraber et al. 2004). Additionally, these models face challenges rising from numerous complexities and uncertainties, such as the complex connections of the combined sewer system and uncertainties due to aging infrastructure.

In the past decade, with the development of artificial intelligence technology, data-driven models are becoming popular and common for simulating various environmental systems. Meanwhile, alongside the improvement of sensor technology, a large amount of influent data from WWTPs becomes available. These data provide a critical basis for the development and application of data-driven models in the wastewater treatment management field.

Data-driven models are developed by analyzing the data pattern of one specific study area. With a limited number of assumptions about the physical behavior of a system, a model can be built based on the relationships between system state variables. In comparison with traditional process-driven models, data-driven models involve mathematical equations assessed not from the physical processes of a system but from the analysis of input and output data (Solomatine and Ostfeld 2007).

Although data-driven models may have some deviations from the ‘real model’ (or the physically-based model), their results are equivalent to the ‘real model’ to some extent and within the error tolerance. Application of the data-driven methods can be regarded as finding a substitution for the “real model” and solving problems in a real-world application before the mathematical relationships are completely clear. Data-driven models can be particularly advantageous for modeling complex systems such as municipal sewer systems. Municipal sewer systems collect and deliver the sanitary sewage directly from households and/or stormwater runoff to WWTPs. They are complex networks of pipes that are intrinsically difficult to model. With data-driven techniques, the wastewater collected by municipal sewer systems and delivered to WWTPs can be estimated using historical data without describing the underlying processes. However, the potential application of data-driven models on wastewater influent flow rate prediction is not well studied. To the author’s knowledge, there are no well-established data-driven wastewater influent flow rate prediction models available to WWTPs operators and managers in Canada.

1.3 Objective

The main objective of this thesis is to support stable operation and effective management of WWTPs by developing reliable wastewater influent flow rate prediction models based on data-driven techniques. This entails the following four tasks: 1) Developing data-driven models for wastewater influent flow rate prediction. Three models, including multilayer perceptron (MLP), long short-term memory (LSTM) networks, and random forest (RF), are investigated and developed. 2) Applying the proposed data-driven models for daily and hourly influent flow rate predictions at WWTPs in Canada. 3) Evaluating the performance of the proposed models using different statistical criteria. 4) Analyzing the uncertainties associated with the predicted results and providing more robust decision support.

1.4 Thesis Outline

Chapter 2 presents a literature review of the previous applications of data-driven models, particularly neural networks models, in the field of wastewater modeling and management. Advantages and limitations of existing data-driven models are discussed and summarized.

Chapter 3 explores the potential of RF models for influent flow rate prediction. An uncertainty analysis approach for RF models is developed for providing more robust decision support for the operation and management of WWTPs. This chapter includes a journal article that has been submitted and is under review.

Chapter 4 investigates the performance of different models on influent flow rate prediction with a high temporal resolution. LSTM and RF models are built and the traditional MLP model is used for comparison purposes. In addition, hourly interval results obtained from the RF models are provided for these three tested WWPTs. This chapter includes an article that will be submitted for publication.

Chapter 5 summarizes the conclusions and directions for future works.

CHAPTER 2 - LITERATURE REVIEW

2.1 Wastewater Simulation Models

Traditionally, the wastewater coming into a WWTPs can be simulated using process-driven models. For example, a benchmark simulation model (BSM2) was developed by the International Water Association for influent characteristics prediction (Vanrolleghem et al. 2007). Their study extended the prediction targets to energy usage in all units, sludge disposal costs, gas production, and use of chemicals. Kuo et al. (2010) developed a real-time storm sewer simulation system (RTS4) with a Storm Water Management Model (SWMM) for predicting the migration of sewer flows. Furthermore, Butler (2014) developed a model to predict the hydraulic conditions (variation of the flow in sewer networks) within sanitary and combined sewers during dry weather. Although these sophisticated models are available for simulating the quantity and quality of municipal wastewater, they often require a large amount of monitoring data as inputs, and the calibration processes are complex. As a result, these process-driven models are not widely used for influent predictions at WWTPs in Canada.

2.2 Alternative Data-Driven Models

Like many engineering modeling problems, wastewater prediction could be regarded as a regression problem, where the key is to find the mathematical relationships between input and target variable (Solomatine and Ostfeld 2007). As an alternative to the process-driven simulation models described in Section 1.2 and 2.1, data-driven models were demonstrated effective in finding such relationships in many previous studies.

Among various data-driven techniques, there are three most well-known and popular ones: multilayer perceptron (MLP), radial basis function (RBF), and support vector machine (SVM).

MLP is one of the most widely used artificial neural networks. The MLP model consists of multiple layers of elements (nodes) or neurons that interact through weighted connections (Pal 1992). Generally, the MLP model includes three layers: the input layer, the hidden layer(s), and the output layer. For the model building process, initial weights are firstly assigned to the input variables. Subsequently, through a set of activation functions, the value of one specific cost function related to the input variables is calculated. Then, one gradient descent method can be used to find the optimal value of the cost function, and the weights of the input variables are updated through backpropagation. Finally, the optimum weights of the input variables can be determined, and the relationship between the input variables and the output target can be defined.

Similar to MLP, the RBF model is a multilayer network that can be used for fitting function and classification. However, the MLP and RBF networks are built in a totally different way. To build an RBF network, firstly, the input layer and the hidden layer are not connected by weights and thresholds, but by the distance between input samples and the hidden layer point (the distance from the center point). After obtaining the distance, the distance is substituted into the radial base function to obtain a numeric value. By multiplying the numeric value with the weight between the hidden layer and the output layer and then seeking the sum, the output of the corresponding input is obtained. It is worth mentioning that the number of center points, the location of center points, the “width” of radial base function, and the weights between the hidden layer and the output layer should be determined before training the model.

The Support Vector (SV) algorithm is a nonlinear algorithm developed in Russia. In recent decades, the SV machine was largely developed at the AT&T Bell Laboratories by Vapnik and his co-workers (SMOLA 2004). There are three important components that characterize the support vector machine (SVM) networks: 1) the solution technique which allows an expansion of the solution vector; 2) the solution that has been extended from linear to non-linear; 3) the soft margins which allow errors on the training samples (Farhat 2002). SVM can be used to solve both classification and regression problems.

2.3 Artificial Neural Networks for the Modeling of Wastewater Processes

Previously, various data-driven techniques were applied for modeling different geophysical processes including sewer processes. Among them, the artificial neural network (ANN) is the most common. For example, El-Din and Smith (2002) developed a neural network model for predicting wastewater inflow during rainfall events. A classical MLP model was built and rainfall at eight rain gauges which cover the major drainage basins of the city, an index to represent the day of the week, and another index to represent the hour of the day were chosen to be input variables. Additionally, rainfall data were also collected as inputs. The results demonstrated that the MLP model had great potential for providing excellent influent flow rate prediction. It was also found that the continuing training beyond 400 epochs in their case did not improve the model performance in term of Coefficient of Determination (R^2). Wei et al. (2013) used the MLP model for short-term prediction of wastewater influent flow rate and the prediction horizon was extended to 180 min. There were small time lags between the predicted and observed influent flow rate when the prediction horizon exceeded 30 min, and the lag increased as the prediction horizon increased. Meanwhile, feed forward neural networks (FFNNs) were widely used in many hydrological contexts. For instance, Aqil et al. (2007) did a comparative study of artificial neural networks and neuro-fuzzy in modeling rainfall-runoff dynamics. Although the proposed FFNN models showed satisfactory results, it was found that the neuro-fuzzy model showed comparative performance in comparison with two FFNN models including the Levenberg-Marquardt-FFNN and the Bayesian regularization-FFNN. The performance of their proposed models were examined on both hourly and daily bases. The effects of data transformation on model performance were also investigated, and it was found that there were no significant differences between using raw data and transformed data. Taormina et al. (2012) used a feed forward neural networks model for long modeling of hourly groundwater levels. The FFNN model was first trained to perform one-step ahead prediction using previous observed data as input. Then the ability of FFNN for long-period prediction was assessed by replacing previous observed data with previous outputs of the FFNN model. The results showed that the FFNN model outperformed the linear autoregressive with exogenous terms (ARX) model in terms of long-term prediction. Although FFNN is widely recognized as a powerful tool, it works only with static patterns (Charaniya and

Dudul 2012). When it comes to long-term prediction or dynamic prediction, FFNN usually suffers from its limited ability to address time lag issues.

To address the dynamic prediction problem, Charaniya and Dudul (2012) used a focused time delay neural networks (FTDNN) model for long-term rainfall prediction. In comparison with a classical static multilayer perceptron network (FFNN or MLP), the main difference is that FTDNN contains two special components tapped delay lines and recurrent connections. The main function of the delay line is storing the past results of the inputs. While for the recurrent connections, it reuses these past results as input variables for prediction. It was found that, in temporal problem, the observes are no longer an independent set of input but a function of time. Thus, the proper length of this time becomes important input variables for representing or predicting the target. Additionally, Verma et al. (2013) used five data-mining algorithms for predicting total suspended solids (TSS). And an iterative learning method (updating the input variables of the prediction model iteratively by using previous prediction results) based on MLP model was developed to reduce prediction error. Their results show that the week-ahead values of TSS can be predicted with about 68% accuracy. Furthermore, Wei and Kusiak (2015) used a dynamic neural network (DNN) with the online corrector for improving the prediction accuracy for a longer time horizon. The DNN model contains a memory structure and predictor. The memory structure captures the previous time series information, then the information is used by the predictor which make it learn the temporal pattern of the time series. Their results indicated that the prediction accuracy could not be improved but cost significant computation when more than five previous influent flow were captured as memories. Basically, this DNN model is similar to the FTDNN model and the iterative learning method mentioned above, and all of these models share similar principles with the recurrent neural network (RNN).

In recent years, to further improve long-term prediction accuracy and address time lag issues in time series forecasting problems, a number of advanced models based on the classical RNN structure were developed. LSTM is a special RNN model, and it was proved to have stable and powerful performance when modeling problems with relatively long-term dependencies (Hochreiter and Schmidhuber 1997; Ismail et al. 2018; Shi et al. 2015). The LSTM model has the same neural connection structure as RNN but a different neuron cell. The LSTM neuron incorporates three gates (forget gate, input gate, and output gate), which allow the network to have memory. The LSTM model was widely employed to solve the long-term prediction problems. For instance, Pisa et al. (2018) developed a LSTM model for predicting wastewater effluent concentrations and dealing with time series information and temporal data. LSTM was also widely used in other fields. For instance, Zhao et al. (2017) used LSTM for short-term traffic forecast. The comparison with other statistical models including Autoregressive integrated moving average (ARIMA) indicated that the proposed LSTM model showed a better performance. It was pointed out that although RNN was widely recognized as a suitable method to capture temporal and spatial evolution, traditional RNNs were not able to capture the long-term evolution, and training an RNN for long-term prediction became difficult because of the vanishing and exploding gradient. Compared with conventional

RNNs, LSTM network could better capture the features of time series within longer time span.

In summary, the artificial neural network technology is constantly improving. Advanced models based on the traditional neural network structure are constantly emerging, which brings valuable technical support for solving long-term prediction problems in a wide array of fields. However, the application of these advanced models in influent prediction area is limited and there are no well-established tools for the long-term prediction of wastewater influent characteristics.

2.4 Other Data-Driven Models for the Modeling of Wastewater Processes

Apart from artificial neural networks, there were also other data-driven models that have been used for the modeling of wastewater influent and other similar geophysical processes. Generally, data-driven models can be classified into the linear model and the nonlinear model, while some methods can convert between linear and nonlinear, such as neural networks. When the neural network uses a linear activation function, it is a linear model; while using a nonlinear activation function such as the sigmoid function, it turns to be a nonlinear model.

Autoregressive integrated moving average (ARIMA) is one of the most popular linear models for time series forecasting. Because it is an autoregression model, it requires fewer input variables, which makes this method more advantageous when the feature data is scarce. Berthouex and Box (1996) used an ARIMA model and an exponentially weighted moving average (EWMA) model for 1-5 days ahead effluent quality prediction. Ömer Faruk (2010) developed a hybrid neural network and ARIMA model for water quality prediction. Boyd et al. (2019) developed ARIMA models for daily inflow rate forecasting at a number of WWTPs in North America. Their results indicated that ARIMA was a simplistic model which could be interpreted and calculated easily, and although ARIMA relied only on historical data, this method was able to do time series analysis and provide support to plant operators. However, the main drawback of ARIMA was that it could only process a continuous time series, which means missing values in the dataset must be filled in and more time would be spent on preparing the data (Boyd et al. 2019).

On the other hand, a number of nonlinear models were also developed for modeling wastewater influent. Kim et al. (2015) developed a k-nearest neighbor (KNN) method for forecasting the influent characteristics at WWTPs. This KNN model was used for predicting influent flow rate, COD, suspended solid, total nitrogen, and total phosphorus. The KNN model was developed under the assumption that part of a past time series will reappear in the future. It was found that the calibration results depend on various factors, such as water quality, influent flow rate, and weather condition, which makes it difficult to tune the KNN model. Meanwhile, the model performance was affected by the uncertainty associated with the dataset. Recently, Nadiri et al. (2018) used an ensemble of fuzzy logic (FL) models for effluent quality parameters prediction. Instead of looking for the best FL prediction model, this study introduced a supervised committee FL (SCFL) ensemble model. Generally, three FL models were built firstly (including Takagi-Sugeno model, Mamdani model, and Larsen model). Then an ANN model was used to combine the prediction results from individual FL models. And they suggested that the fuzzy set can handle uncertainty in water quality parameters because the fuzzy logic theory is applicable to estimate inherently imprecise parameters. However, the discussion on how these methods can be used for long-term prediction is limited. Also, it is hard to determine each input variable’s contribution using any of the abovementioned methods.

In summary, there are only a few applications of data-driven modeling techniques for predicting wastewater influent. The existing models could be improved significantly in order to address issues such as determining input variable’s contribution, long-term prediction and model uncertainties. Thus, the development of more advanced and robust data-driven models for influent prediction is desired.

CHAPTER 3 - DAILY INFLUENT FLOW RATE PREDICTION

This chapter is organized based on the paper submitted on March 2019 to *Environmental Research and Risk Assessment*.

Development of a Random Forest Model for Inflow Prediction at Wastewater Treatment Plants

Pengxiao Zhou[a], Zhong Li[a,*], Spencer Snowling[b], Brian Baetz[a], Dain Na[a], Gavin Boyd[a]

[a]Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7

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

[*]Corresponding author:

Zhong Li (Tel: +19055259140 ext 21225; E-mail: zoeli@mcmaster.ca)

Abstract

Influent flow of wastewater treatment plants (WWTPs) is a crucial parameter for plant operation and management. In this study, a random forest (RF) model was developed for daily wastewater inflow prediction, and a new approach was proposed for quantifying the uncertainties associated with wastewater inflow forecasts. 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 (R^2) values for training and testing were 0.948 and 0.830, respectively. The R^2 values at the Humber WWTP were 0.958 and 0.582 for training and testing, respectively. In comparison with other approaches such as the multilayer perceptron neural networks (MLP) models and autoregressive integrated moving average (ARIMA) models, the results showed that the developed RF model performs well on forecasting inflow. In addition, probabilistic forecasts of daily inflow were generated to provide robust decision support for the operation, optimization, and management of WWTPs.

Keywords:

Random forest; WWTP; Wastewater prediction; Daily flow; Uncertainty analysis.

3.1 Introduction

It is well acknowledged that the wastewater inflow to a wastewater treatment plant (WWTP) is an essential parameter 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 control 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 forecast wastewater inflow is desired.

The accuracy of an influent flow forecasting 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 forecasting. In the past decades, alongside the development of artificial intelligence, numerous data-driven models have been applied to predict the inflow of WWTPs. For instance, 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 forecast 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. 2012; Zhe Zhou et al. 2015). Additionally, these methods cannot provide information on each input variable's contribution to the inflow (Wang et al. 2015), and they cannot tackle the uncertainties associated with the prediction process. 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 2001a). This method 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 2001a; 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 an 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 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. The RF method has been proven to be an effective method for building predictive models in many previous studies; however, the use of this method in wastewater inflow prediction is limited.

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) developing an uncertainty analysis approach to provide probabilistic inflow forecasts 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.

3.2 Methodology

3.2.1 Random Forest

3.2.1.1 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 2001b). 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.

3.2.1.2 Regression Trees

A regression tree is a forecasting model that can be described as a decision tree, and it deals with the forecasting of an output variable y, given a vector of input variables x (Loh 2008). The output variable y can be continuous or discrete (e.g., the value of the inflow rate in this study). 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:

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

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.

The root node and internal nodes are split using the best variable among a subset of m variables which are chosen randomly from all M input predictor variables. The best variable is the one that results in the lowest impurity 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. A completed regression tree is presented in Fig. 3.1 as an example. The development of a wastewater influent flow forecasting regression tree is summarized as follows: (1) Start with an original training set including N samples. Select N samples, with replacement, from the original training set to form a new training set. Theoretically, for a training data set which includes N samples, a maximum of N^N new training sets can be generated. In this study, k new training sets were created using the bootstrap method from the original training set. (2) Grow a regression tree for each of the k new training set. In this study, the CART approach with MSE as the impurity criterion is used to grow internal nodes and eventually, leaf nodes. (3) After k regression trees are formed, use all trees to generate predictions. Each DT produces one value, and the mean value of these k values is taken as the predicted value.

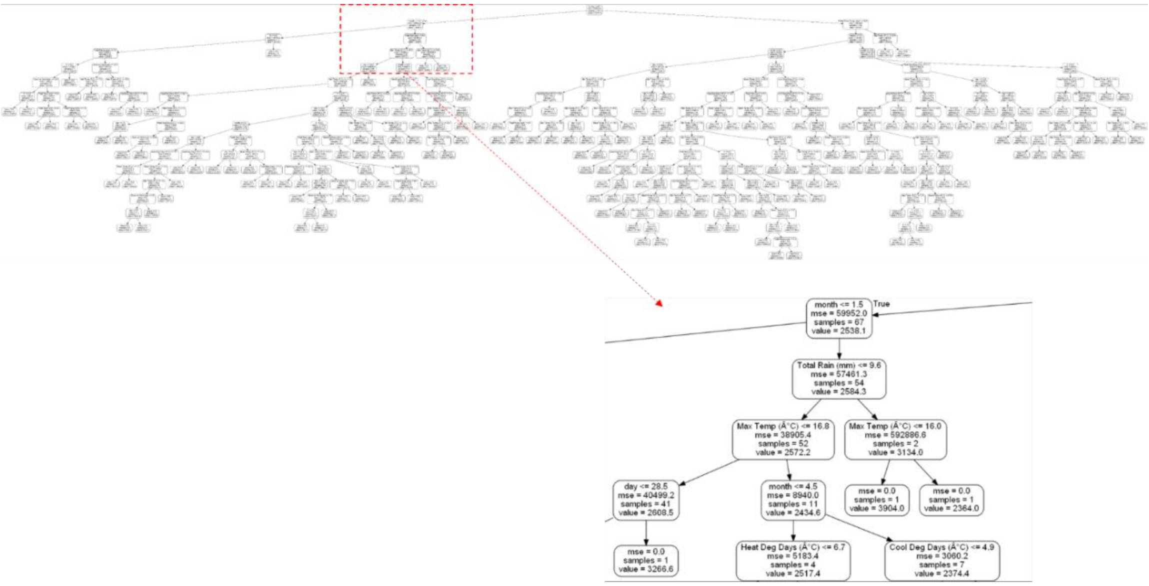

Fig. 3.1 A sample of the regression trees

3.2.1.3 Variable Importance

RF became popular because of its ability to address large numbers of variables with relatively small-scale observations and the advantages over other existing data-driven methods in terms of assessing variable importance (Grömping 2009). 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.

3.2.2 Model Development

The representativeness of training datasets is important to the effectiveness and overall performance of an RF model (Wang et al. 2015). To reflect the impacts of weather conditions and domestic water usage patterns on wastewater inflow, numerous weather and date/time variables are selected as predictor variables. 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. More details regarding the weather features are given in Section 3.3.2. 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 correlation analysis and a literature review. A list of the selected weather features is given in Table. 3.1. 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. This random selection process is considered effective in improving on the generalization error.

Table. 3.1 The selected input features for the two WWTPs

Feature category	Feature	WWTP
Weather Features	Maximum Temperature (℃)	Humber, Confidential plant
	Minimum Temperature (℃)	Humber, Confidential plant
	Mean Temperature (℃)	Humber, Confidential plant
	Heating Degree Days (℃)	Humber, Confidential plant
	Cooling Degree Days (℃)	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

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. In this study, three different numbers (300, 1000, 3000) are first assigned to k, and three different numbers (M, $Sqrt(M)$ and $Log(M)$) are considered for m. Subsequently, the best combination of k and m for each WWTP can be identified using a grid search and a 3-folds cross validation. The best combination of parameters is then used to build a random forest for predicting wastewater inflow. A flowchart of the training and testing processes is shown in Fig. 3.2.

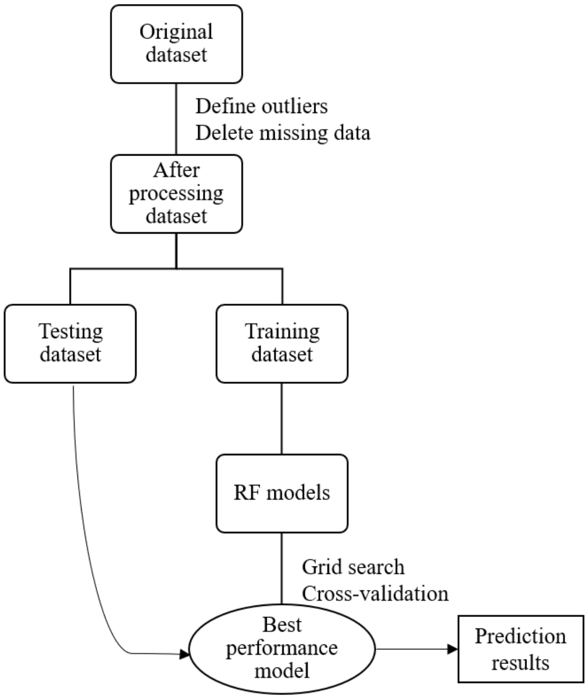

Fig. 3.2 Flow chart of the training and testing process

3.2.3 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 Equation 4.

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

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 Equation 5 is the squared root of the MSE, which prevents positive and negative errors to cancel each other out in order to express the error metric in the same units as the original data (Bennett et al. 2013).

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

R^2, given by Equation 6, isthe Coefficient of Determinination, and measures the correlation of the observed and modeled values. R^2 ranges from 0 to 1, with 1 corresponding to the strongest correlation.

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

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 Equation 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.

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

3.3 Case Study

3.3.1 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. 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.

3.3.2 Data

The influent flow data were 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 was used. For the confidential WWTP, flow daily data from November 1, 2015 to October 30, 2016 was collected. Time-series flow plot for Humber WWTP and the confidential WWTP are presented in Fig. 3.3.

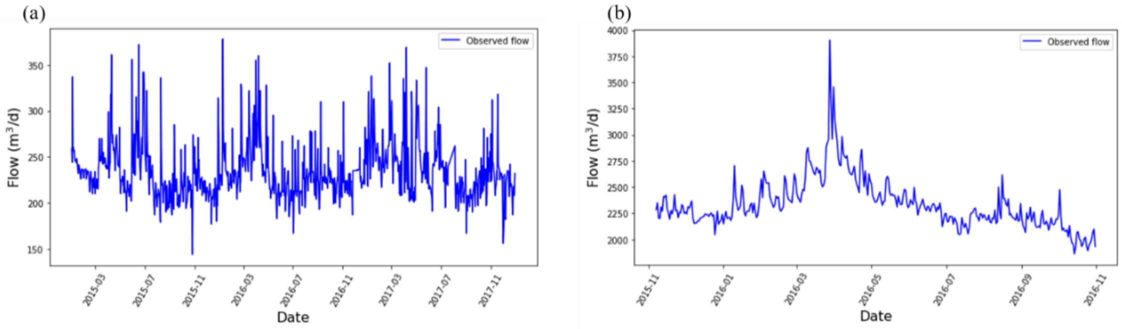

Fig. 3.3 Time-series flow graph for Humber (a) and confidential plant (b)

The weather data were obtained from Weather Canada (https://weather.gc.ca/canada_e.html). The weather data was 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 Equation 8), Cooling Degree Days (°C, defined by Equation 9), Total Rain (mm), Total Snow (mm), and Total Precipitation (mm).

$$HDD = (1\ day) \sum_{days} (T_b - T_m)^{+} \qquad (8)$$

$$CDD = (1\ day) \sum_{days} (T_m - T_b)^+ \qquad (9)$$

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).

3.4 Result Analysis and Discussion

3.4.1 Modeling Performance

A RF model was built for each of the two WWTPs using the approach described above. For each WWTP, the parameters used for each RF model were different. For Humber, the original dataset had a total of 1,080 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 1,053 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 3,000 trees, and the number of the features tried at each split m was equal to $\log M$. Fig. 3.4 shows the scatter plot of predicted and observed flow.

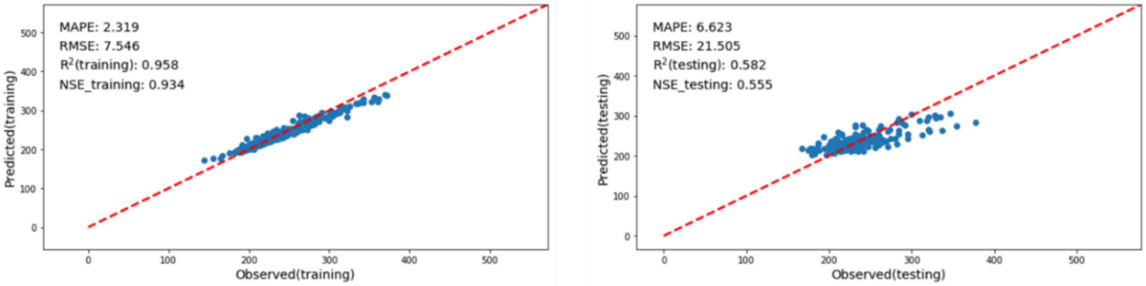

Fig. 3.4 Scatter plot results of training and testing for Humber

At the confidential WWTP, flow data from November 1, 2015 to October 30, 2016 was 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 number of the trees k equal to 1,000, 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 are illustrated in Fig. 3.5.

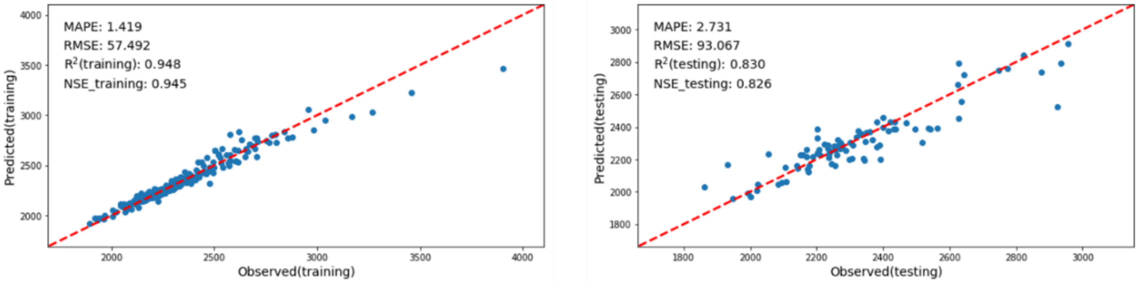

Fig. 3.5 Scatter plot results of training and testing for confidential plant

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 applied for wastewater inflow prediction, other algorithms used in previous studies, such as ARIMA used by Abunama and Othman (2017) and MLP proposed by Qianqian Zhang et al. (2018), are compared to RF. The results illustrate that RF can predict the wastewater inflows competently. Compared with ARIMA and MLP, the RF model for the confidential station shows better performance in terms of the statistical criteria. Although the RF model for the Humber station is not very good with regards to the values of NSE and R^2, inflow prediction results based on RF (MAPE=6.623) show better performance in terms of MAPE when compare to the ARIMA model (Abunama and Othman 2017) which has a MAPE value of 8.012. To summarize, RF has significant potential to predict wastewater inflows, and it usually performs better than ARIMA and MLP.

3.4.2 Variable Importance Analysis

The variable importance was calculated by the sum of the MSE decrease as described in Section 2.1.3. This provides valuable support for decision makers to understand each variable's contribution to the flow volume. Fig. 3.6 shows the variable importance of each station. For Humber, 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 & Smith (2002), where the authors suggested that the influent flow to a WWTPs 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. It is recommended to carefully select input variables through literature review, system characaterizaion, and correlation analysis when building a RF model. To some extent, the variable importance analysis can also help identify the proper input variables (Wang et al. 2015).

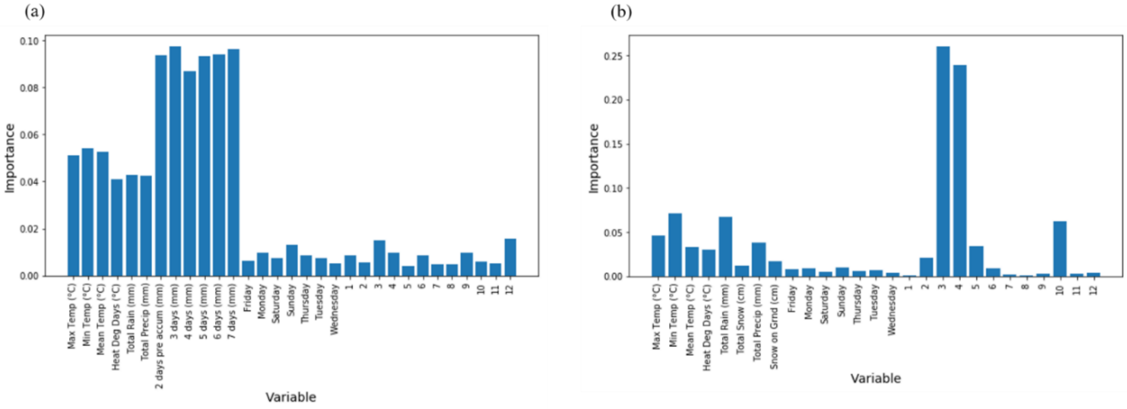

Fig. 3.6 Variable importance for Humber (a) and confidential plant (b)

3.4.3 Uncertainty Analysis

In the developed RF models, each tree can generate one predicted value under no prone conditions. The final predicted inflow rate was estimated using the mean value of predicted values generated by all the trees in the forest. In general, uncertainties associated with the predicted results can be quantified by analyzing all the generated predicted values. For example, the probability density function (PDF) and cumulative distribution function (CDF) of the predicted inflow could be generated using results from k trees as samples. Similarly, the upper and lower bounds of predicted inflow could also be found to create the inflow ranges.

As an example, the PDF and CDF graphs at a randomly selected point from the confidential plant's testing dataset are presented in Fig. 3.7. Following the traditional RF modeling approach as described in Section 2.1.2, the predicted value is 2,383.7 m^3/hr. Using the proposed uncertainty analysis approach, as the final prediction value, the PDF graph illustrates that the relative likelihood of around 2,350 m^3/hr is the highest. Furthermore, the CDF graph shows that the cumulative probability that inflow is less than or equal to 2,300 m^3/hr is zero, while that for an inflow of greater than or equal to 2,450 m^3/hr is one. This implies that the range of the predictive values is [2,300, 2,450] m^3/hr. Furthermore, with the CDF graph, the probability that the predicted inflow exceeds a certain threshold can be assessed. For example, from the CDF graph shown in Fig. 3.7, the corresponding cumulative probability of flow at 2,400 m^3/hr is approximately 0.7. Thus, the probability that the predicted inflow exceeds 2,400 m^3/hr 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.

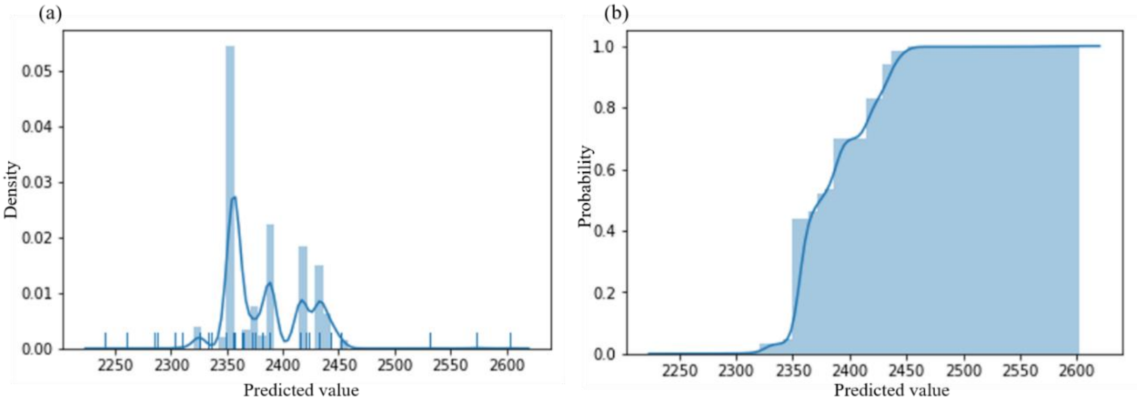

Fig. 3.7 Probability Density Function (a) and Cumulative Distribution Function (b)

Additionally, the predicted daily inflow interval solutions for Humber and the confidential station during the testing period are presented in Fig. 3.8. These were slightly over-estimated for some points, especially the upper bounds of some points at the confidential plant (for instance, one point near November 2015). Overall, though the predicted interval solutions are slightly large for some data points, the interval solutions did capture almost all the observed flow values within the interval ranges.

To evaluate the accuracy of the predicted intervals, the relative error of the interval solution is introduced as follows (Li et al. 2015):

$$REIS(\%)\begin{cases} \dfrac{c_i^{max}-q_i}{q_i}\times 100, & if\ q_i > c_i^{max};\\ 0, & if\ c_i^{min} < q_i < c_i^{max}\ ;\\ \dfrac{c_i^{min}-q_i}{q_i}\times 100, & if\ q_i > c_i^{max} \end{cases}$$

where c_i^{max} and c_i^{min} are maximum and minimum predicted results on sample i generated by k trees, respectively; and q_i is the observed value.

For the Humber station, among the 264 samples used for testing, 248 samples of their observed values fall into its corresponding interval solution generated by the RF model, accounting for 92.4% of the total testing samples. The percentage of samples with the absolute REIS less than 5%, 10%, and 20% are 94.7%, 96.6% and 99.6%, respectively. For the confidential plant, 61 observed values of the total 90 samples fall into its corresponding intervals, accounting for 67.8% of the total testing samples. The percentage of samples with the absolute REIS less than 5%, 10%, and 20% are 82.2%, 95.6% and 100%, respectively.

Moreover, it can be seen from Fig. 3.8 that all the upper bounds of interval solutions for Humber are relatively stable in comparison to the confidential plant. For the confidential plant, the large upper bound points may be explained due to the large rate of flow change at the plant. For Humber, most of the observed influent flow values fall into the range from 150 MLD to 350 MLD, whereas for the confidential plant, the observed influent flow values change from 1750 m^3/hr to 4000 m^3/hr. The interval solution was generated by using the maximum and minimum values generated by k trees, and not all the trees were built using proper samples. Thus, some trees may become disturbances, and large changes of scale may lead to a large variance.

The predicted flow interval solution provides an upper bound and a lower bound using the maximum value and the minimum value generated by k trees. This work is the first attempt to analyze the uncertainty of predicted flow by using this method. In this case study, most of the testing points fell into the predicted interval. The interval solutions 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 during extreme weather events and domestic water consumption rush hours.

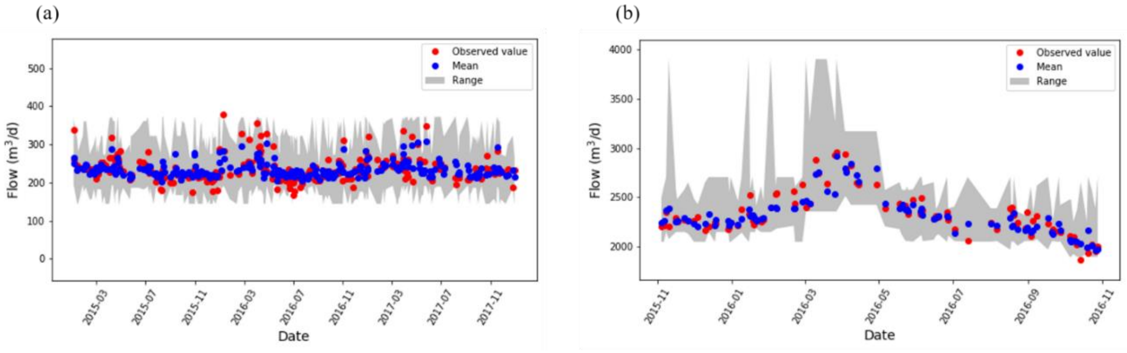

Fig. 3.8 Range prediction for Humber (a) and confidential plant (b)

3.5 Conclusions

In this study, a RF model was developed 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 avoids overfitting and analyzes each input variable's contribution. 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 uncertainty analysis method was proposed to quantify the uncertainties with RF forecasts 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 Humber and the confidential plant for training were 0.916 and 0.949, respectively; while those for testing were 0.606 and 0.827, respectively. The NSE values of Humber for the training and testing were 0.896 and 0.597, respectively; while those for the confidential station were 0.946 and 0.822, respectively. The results demonstrate that the developed 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.

Acknowledgments

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.

CHAPTER 4 - HOURLY INFLUENT FLOW RATE PREDICTION

This chapter is organized based on the draft priori to submission paper.

Hourly Wastewater Influent Flow Prediction at Wastewater Treatment Plants

Pengxiao Zhou[a], Zhong Li[a,*], Spencer Snowling[b], Rajeev Goel[b], Qianqian Zhang[a,c]

[a]Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7

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

[c]School of Management, Chengdu University of Information Technology, Chengdu, 610225, China

[*]Corresponding author:

Zhong Li (Tel: +19055259140 ext 21225; E-mail: zoeli@mcmaster.ca)

Abstract

Influent flow rate prediction is of great importance to the operation and management of wastewater treatment plants (WWTPs). However, due to the complexities of wastewater collection systems, predicting WWTP influent flow rate with a high temporal resolution is a daunting challenge. In this study, three machine learning models, including multilayer perceptron (MLP), long short-term memory network (LSTM), and random forest (RF), are developed for hourly influent flow rate forecasting. The proposed models are applied for influent forecasting at three WWTPs (i.e., Humber, Woodward, and one confidential plant) in Ontario, Canada. For the confidential WWTP, all of the three proposed models show good performance. The Coefficient of Determination for the testing period are 0.829, 0.787, and 0.906, respectively. For the other two WWTPs, the performance of MLP and LSTM is barely satisfactory; however, the RF model results in a good fit between the observed and predicted hourly flow rate. After the model performance evaluation, interval predictions of hourly influent flow at each WWTPs are generated. To the authors' knowledge, this work is the first attempt to predict wastewater influent with an hourly time step for WWTPs in North America.

Keywords:

Random forest; Long short-term memory; WWTP; Wastewater prediction; Hourly flow; Uncertainty analysis.

4.1 Introduction

Considering the fluctuations in the influent flow rate is important for the stable operation and management of a wastewater treatment plant (WWTP). The influent flow rate entering a WWTP not only has an impact on effluent quality, such as total suspended solids and biochemical oxygen demand, it also affects the total energy consumption of the plant (Bechmann et al. 1999; Wei and Kusiak 2015). Particularly, for the combined sewer system, the influent flow rate is closely related to facility security during storm events. Therefore, the accurate prediction of wastewater influent flow rate is essential for improving treatment efficiency, optimizing energy consumption, and maintaining facility security.

The forecasting of influent characteristics was a major challenge for wastewater treatment simulation and optimization (Wei and Kusiak 2015). In the recent decade, alongside the development of sensor technologies and advanced control strategies, plant-wide controlling systems and monitoring networks were widely used in WWTPs (Campisano et al. 2013; Dürrenmatt and Gujer 2012). As a result, high-frequency and high-quality data were made available to plant operators and managers, and these data are extremely valuable for improving influent forecasting accuracy (Boyd et al. 2019; Chiang et al. 2018). Therefore, data mining approaches that could utilize wastewater monitoring data and provide reliable influent forecasts are desired.

Previously, a number of data-driven models have been developed to predict the influent flow rate. For instance, El-Din and Smith (2002) applied artificial neural networks (ANNs) to forecast the wastewater influent flow rate during extreme weather events. Fernandez et al. (2009) developed a neurofuzzy wastewater flow-rate forecasting model (NFWFFM) using only two input variables, day of the week and average daily flow-rate of day before. Wei and Kusiak (2015) considered the spatial feature of influent flow and compared the performance of static multi-layer perceptron (MLP) neural networks and dynamic neural network (DNN) on the short-term prediction of influent flow rate. Kim et al. (2016) evaluated the k-nearest neighbor (k-NN) method for forecasting influent flow rate, chemical oxygen demand, suspended solid, total nitrogen and total phosphorus. Although these models could provide fairly good influent forecasts, they do have some major drawbacks. For example, these models have difficulties in making long-term predictions and addressing the associated time lag problems (Charaniya and Dudul 2012). Also, the temporal resolution of prediction is mostly daily, which is not ideal for real-time facility operation and dynamic process control. Additionally, very few of the previous studies tackled the various uncertainties associated with the wastewater forecasting process. Therefore, more advanced data-driven modeling techniques that are reliable for high temporal resolution and long-term prediction and can provide robust decision support based on uncertainty analysis are still desired.

More recently, recurrent neural networks (RNNs) based on Long Short-Term Memory (LSTM) cells emerged as a popular data-driven technique (Ismail et al. 2018). LSTM was systematically proposed by Hochreiter and Schmidhuber in 1997. The main advantage of LSTM is that can bridge long time lags (Hochreiter and Schmidhuber 1997). LSTM was applied to a wide range of areas. For instance, Shi et al. (2015) developed a convolutional LSTM network to predict the rainfall intensity in the near future and the proposed model performs well. Zhao et al. (2017) used LSTM network for short-term traffic forecast and the proposed LSTM achieved better performance in comparison with some other models. In the meanwhile, another model called random forest (RF) has also gained a lot of attention. RF is an ensemble classifier, which comprises a collection of tree-structured classifiers (Breiman 2001a). It has certain advantages in terms of modeling accuracy and computational intensity (Breiman 2001a; Gislason et al. 2006). Furthermore, each input variable’s contribution could be analyzed in a quick and straightforward manner with the tree structure. As a result of these advantages, RF was wildly used in various areas in the past decade (Abdel-Rahman et al. 2013; Díaz-Uriarte and Alvarez de Andrés 2006; Pal 2005). Both LSTM and RF have great potential to be used as effective techniques for wastewater influent prediction at a high temporal resolution. However, there are very limited researchs on the application of LSTM and RF in this area.

Therefore, the objective of this study is to develop advanced data-driven models based on LSTM and RF for hourly wastewater influent prediction. This entails the following four tasks: 1) developing influent prediction models using LSTM and RF, as well as a traditional neural network technique, i.e., MLP; 2) applying the developed models to predict hourly influent flow rate at three WWTPs in Ontario, Canada; (3) evaluating the performance of the proposed models using different statistical criteria; (4) providing interval prediction of hourly flow rate results for each WWTPs using the RF models. This study will provide innovative and robust tools for wastewater influent flow rate prediction at a high temporal resolution, and thus provide valuable support for effective WWTP operation and management.

4.2 Study Area and Data Collection

Three WWTPs in Ontario, Canada (i.e., Woodward, Humber, and a confidential WWTP), were selected for this study. The Woodward WWTP is located in the City of Hamilton, Ontario. Both sanitary and combined sewage is processed by this WWTP with an average capacity of 409 MLD. The Humber WWTP is Toronto's second largest WWTP and it serves a population of about 680,000 with a capacity of 473,000 m^3/d. The confidential WWTP serves a population of approximately 140,000. 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.

The influent flow rate data are obtained from Hydromantis Environmental Software Solutions, Inc., a software development company that focuses on the water and wastewater treatment sector. For the Woodward WWTP, time series data in 5-minute intervals from January 1, 2015 to December 31, 2017 are collected. For the Humber WWTP, hourly flow rate data from January 2, 2015 to December 31, 2017 are used. For the confidential WWTP, hourly data from November 1, 2015 to October 30, 2016 are collected.

The weather data are obtained from Dark Sky, a company that specializes in weather forecasting and visualization. The weather data are collected and matched with the corresponding flow data with the same data length and frequency. The weather variables include temperature, humidity, precipitation, wind speed, and wind bearing.

4.3 Methodology

4.3.1 Multi-Layer Perceptron

Artificial neural networks were proved to be a useful tool to build generalizable models in many disciplines (H.Taud and J.F.Mas 2018). One of the most widely used artificial neural networks, multi-layer perceptron (MLP), is used as a baseline model in this study. The MLP model comprises multiple layers of nodes (or neurons) that interact through weighted connections (Pal 1992). Generally, the MLP model consists of three layers: the input layer, the hidden layer(s), and the output layer. In this study, initial weights are first randomly assigned to the input variables. Subsequently, through a set of activation functions, the value of a cost function related to the input variables is calculated. Then, a gradient descent algorithm is used to find the optimal value of the cost function, and the weights of the input variables are updated through backpropagation. Finally, the optimum weights of the input variables are determined, and the relationship between the input variables and the output target can be defined.

4.3.2 Long Short-Term Memory Networks

Although traditional neural networks such as the MLP model perform well in many disciplines, they can not well address problems with time series features. To address this issue, recurrent neural networks (RNNs) are developed. While full connections only exist between adjacent layers in MLP models, connections among nodes within the same layer are built in RNN models (Zhao et al. 2017). These connections allow information to persist, which makes it capable of retaining time series information. RNN's unique architecture allows it to use all previous input information up to current time and the depth of these connections can be adjusted case by case (Zhao et al. 2017). However, there are still some problems with RNNs. For example, as the inter-layer connections grow, it may be computationally demanding for an RNN to use previous information. In practice, RNN models usually can not learn long-term dependencies (Bengio et al. 1994; Universit and Betreuer 1991). In order to solve this problem, Long Short-Term Memory (LSTM) network is developed. LSTM is a special RNN, and it has been proven to have stable and powerful performance when modeling problems with relatively long-term dependencies (Hochreiter and Schmidhuber 1997; Ismail et al. 2018; Shi et al. 2015). LSTM models have the same neural connection structure as a typical RNN but a different neuron cell. The LSTM neuron incorporates three gates (forget gate, input gate, and output gate), which allows this network to have memory. A typical LSTM cell is presented in Fig. 4.1.

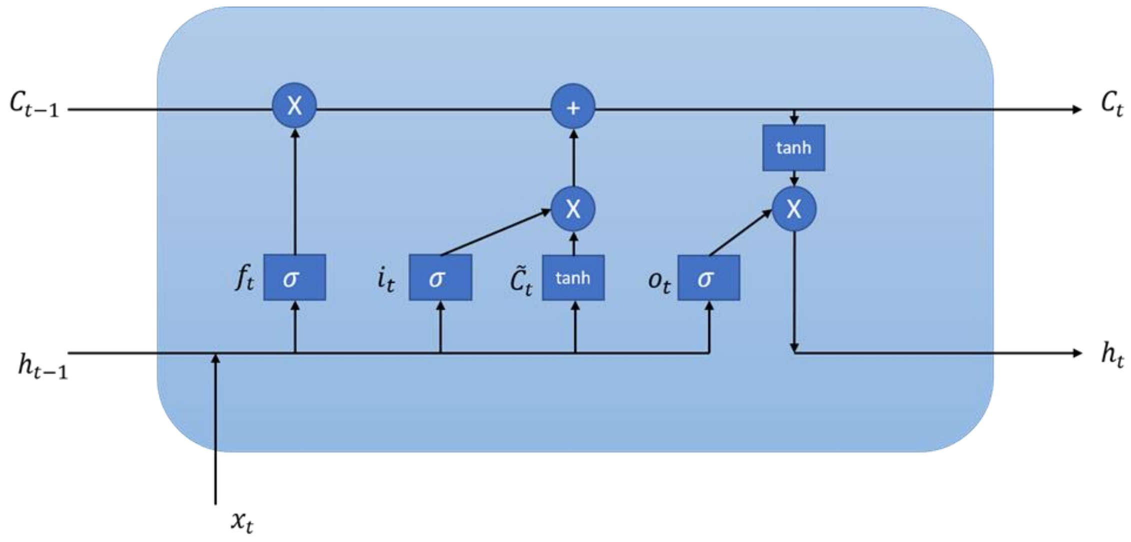

Fig. 4.1 Structure of a typical Long Short-Term Memory Neuron

The gates in the LSTM neuron are mainly used for information selection. For instance, in this study, the forget gate layer selects information by using the sigmoid function. When information x_t and h_{t-1} enter the neuron, the result of this forget gate layer can be expressed as follows:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{1}$$

where x_t is the new input information at time t; h_{t-1} is the output information at time $t-1$; w_f is the assigned weight for the forget gate; b_f is the bias for the forget gate; and σ is the sigmoid function.

The value of the sigmoid function is between 0 and 1. Thus, if $f_t = 0$, it means no information can pass the gate. If $f_t = 1$, it means that all of the information is allowed to enter the gate. The input gate decides the information that will be updated and entered:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(w_c[h_{t-1}, x_t] + b_c) \tag{3}$$

where i_t decides the information that will be updated at time t; $\tilde{C}_t$ is a created candidate vector at time t; w_i and w_c are the assigned weight for the processes i and c, respectively; b_i and b_c are the biases for the processes i and c, respectively. Then the neuron state can be updated as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

Finally, the output gate decides the information that will be outputted:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_0) \tag{5}$$

where h_t is the output information; o_t is the output gate for determining the information that will be outputted; w_o is the assigned weight for the process o; b_o is the bias for the process o.

4.3.3 Random Forest

RF is an ensemble classifier that consists of a cluster of decision trees (DT) $\{h(x, \Theta_k),\ k = 1, \ldots\}$, where x is the input vector, and $\{\Theta_k\}$ are independent distributed random vectors and subsets of *x* (Breiman 2001a). The RF method can be used to address both classification and regression problems. When applying the RF model to regression problems, such as the wastewater influent flow rate prediction, regression trees can be constructed. The construction process of the RF model in this study is presented in the Fig. 4.2. Each regression tree is developed using a bootstrap sample of the original data set. The regression tree deals with the forecasting of an continuous or discrete output variable y, given a vector of input variables x (Loh 2008). In this study, the classification and regression tree (CART) algorithm is used when growing the regression tree (Breiman 2017). Mean squared errors (MSE) as the variance of the samples in the node is used for node impurity criterion. After the trees and the forest are constructed and finalized, the average of all values predicted by the trees can be calculated as the predicted value of the random forest. Moreover, the sum of MSE decreases can be used to measure the variable importance, which quantifies each input variable's contribution to the target (Breiman, 2002; Zhou et.al).

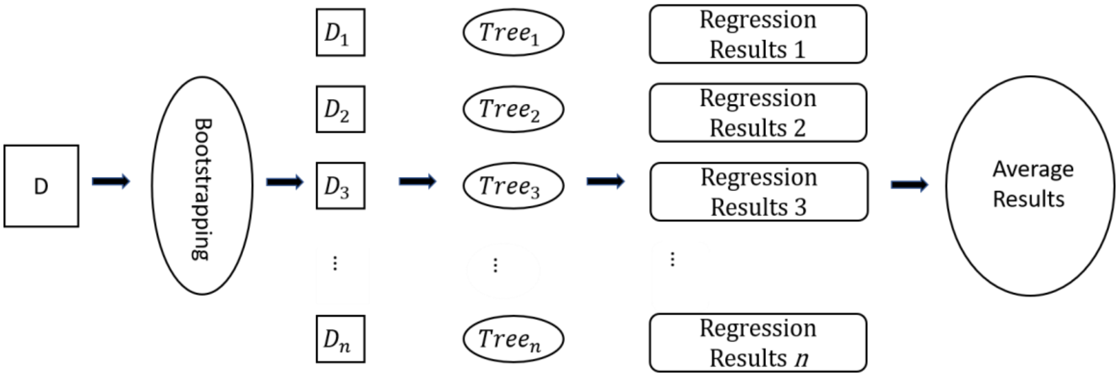

Fig. 4.2 The process of developing an RF model

4.3.4 Model Development and Evaluation Criteria

The accuracy of modeling results depends on how the non-linear relationship between the influent flow rate and other predictors is captured (Li et al. 2015). In addition, the representativeness of training datasets is critical to a data-driven model

(Wang et al. 2015). Therefore, the selection of predictors and training datasets has significant impacts on the overall performance of a model. Wastewater influent flow rate is affected by numerous factors such as the local sewage system, weather conditions, and domestic water usage pattern. In this study, temperature, humidity, precipitation, pressure, wind speed, and wind direction are used to describe the change of weather conditions. Months of the year, days of the week, and hours of the day are used to decode local water usage pattern. Seventy-five percent of the original data are randomly selected to develop and train the MLP and RF models; while the first seventy-five percent of the original data are directly selected for the LSTM model, in order to keep the sequence information intact. Furthermore, a grid search and cross-validation approach are used to calibrate the model. When there are several options for one parameter, the combined grid search and cross-validation approach can help find the best parameter. In this study, a total of 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 proposed models (Bennett et al. 2013; Nash and Sutcliffe 1970).

4.3.5 Interval Forecasts Based on the RF Model

To further quantify the uncertainties associated with RF modeling outputs, interval flow rate forecasts are generated using a bootstrapping approach. In the developed RF model, each tree produces one predicted value without pruning the tree. While an exact influent flow rate value can be estimated using the mean value of predicted values generated by all of the trees in the forest, all these predicted values can also be used to generate the distribution and provide probabilistic forecasting results. Therefore, the associated uncertainties can be quantified by analyzing the distribution of predicted values. For instance, the probability distribution, cumulative distribution, and confidence intervals of influent flow rates can be generated.

4.4 Results and Discussion

4.4.1 Modeling Results Analysis

Three different models have been used in this study, including two neural networks model (MLP and LSTM) and one random forest model (RF). All three models were applied for hourly influent flow rate forecasting for three different WWTPs. For the confidential station, from 00:00:00 on November 1st, 2015 to 23:45:00 on October 31st, 2016, a total of 35,133 data samples were collected at a 15-minutes time step. Outliers were manually detected by engineers at the WWTP. After removing the outliers, the data were upscaled to hourly, resulting in a total of 8,783 data samples. For the MLP model and the RF model, 75% (6,587 samples) of the hourly data were randomly selected and marked as training data, and the rest 2,196 samples were used for testing. Meanwhile, to build the LSTM model, the first three-quarter of the original time series were used as training data, and the rest were testing data. Hourly data from 21:00:00 on January 2nd, 2015 to 00:00:00 on December 31st, 2017 at the Humber station, as well as 5-minute data from 00:00:00 at January 1st, 2015 to 23:55:00 on December 31th, 2017 at Woodward, were collected. After removing outliers using the three times standard deviation method, in a total of 24,509 samples and 25,986 samples were fed to models, respectively. The modeling results of MLP, LSTM, and RF are presented in Figs. 4.3-4.5, respectively.

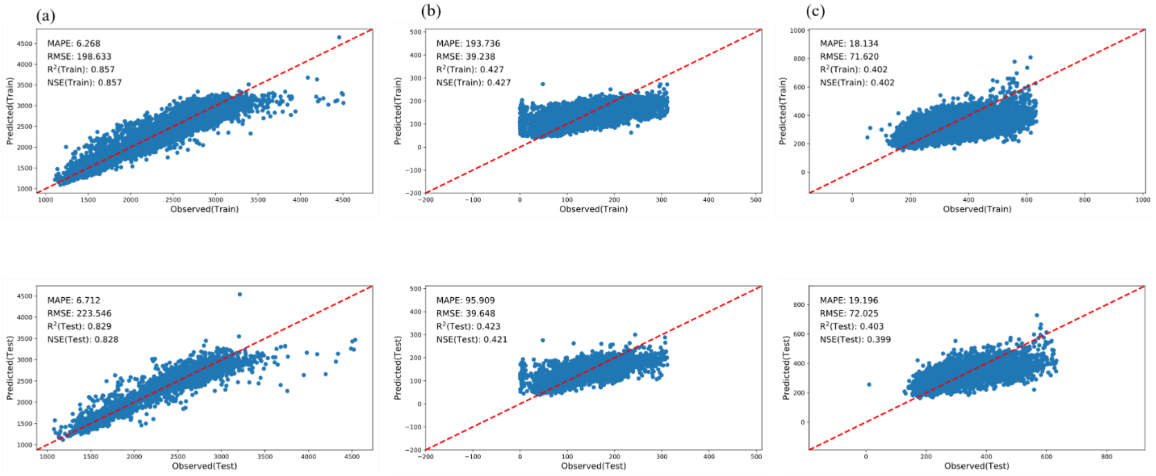

Fig. 4.3 Results of the MLP model: (a) the confidential station, (b) the Humber station, and (c) the Woodward station

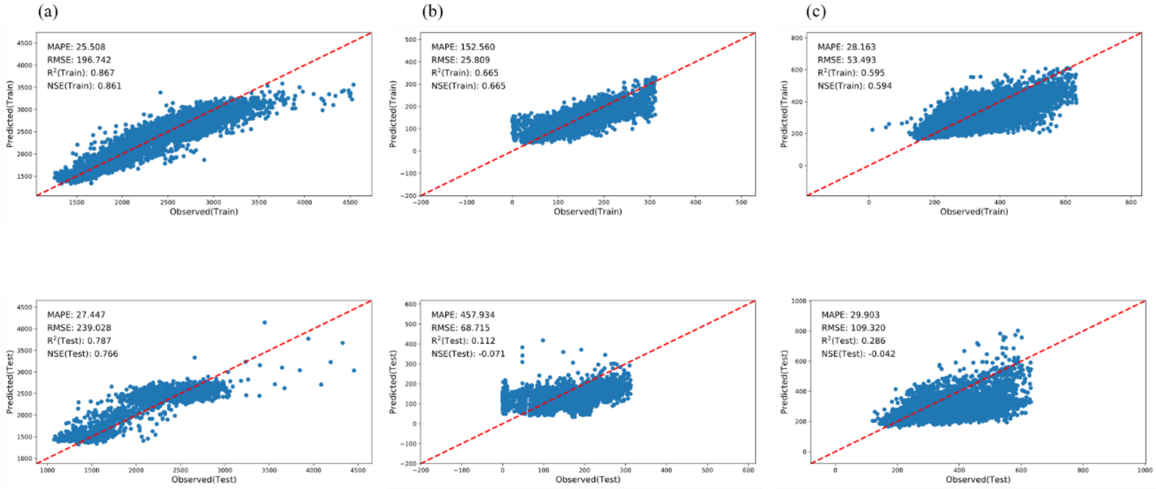

Fig. 4.4 Results of the LSTM model: (a) the confidential station, (b) the Humber station, and (c) the Woodward station

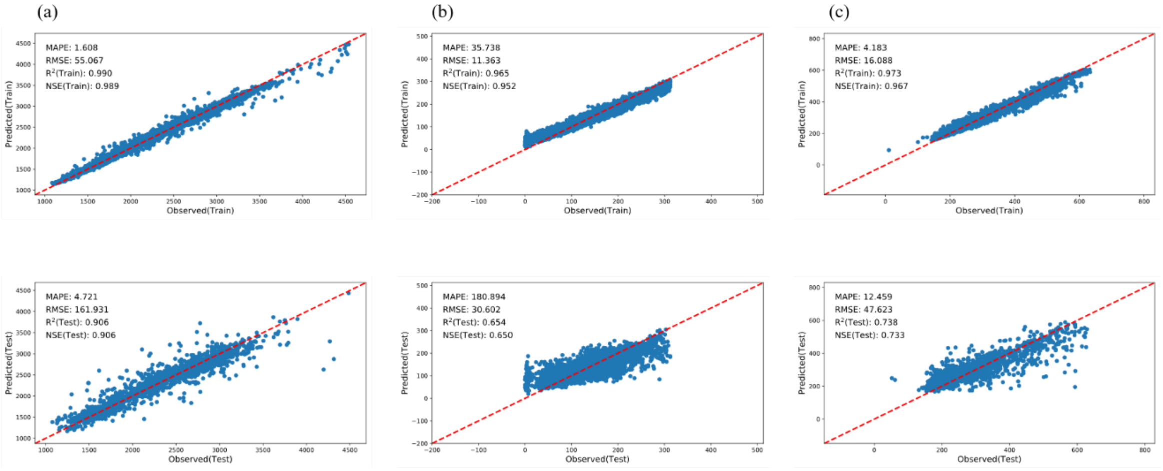

Fig. 4.5 Results of the RF model: (a) the confidential station, (b) the Humber station, and (c) the Woodward station

These scatter plots indicate that all of the three methods can obtain a satisfactory forecast of influent flow rate with an hourly time step for the confidential station. The NSE values of the MLP, LSTM, and RF models during the testing period are 0.828, 0.766, and 0.906, respectively; those of R^2 are 0.829, 0.787, and 0.906, respectively; while MAPE scores 6.712%, 27.447%, and 4.721%, respectively. Typically, for modeling purposes, the required minimum value of NSE is 0.5, and only R^2 values greater than 0.5 are acceptable (Mello et al. 2008; Moriasi et al. 2007). Thus, it is illustrated that these models could provide satisfactory hourly influent flow rate forecasts for the confidential WWTP.

When applied to the Humber and Woodward WWTPs, the accuracy of the three models decreases drastically. The decrease is particularly significant for MLP and LSTM. The R^2 values of MLP and LSTM during testing are both under 0.5. In addition, results of the LSTM model indicate an obvious poor performance of testing. That may be the result of the non-linear relationship between hourly influent flow rate and the fact that the selected predictors are not representative enough. To better describe the complex nonlinear relationship accurately, more relevant input data should be collected. Despite the poor performance of MLP and LSTM, the RF model is more robust for resisting disturbance. Although the results are not as good as those at the confidential station, they still indicate that the RF model can obtain satisfactory forecasts at Humber and Woodward with R^2 equals 0.654 and 0.738 for testing, respectively.

It is worth mentioning that none these models included previous wastewater influent flow rates as predictors. When including previous flow as predictors, it would be the dominant impact factor of the flow to be predicted, and the modeling performance would improve dramatically, especially for hourly influent flow prediction (El-Din and Smith 2002; Herrera et al. 2010; Shen et al. 2009). That is because the autocorrelation of hourly influent rate between adjacent time steps is strong. Although introducing previous flow as predictors could improve the modeling results, it is not ideal for operation and management practices as it limits the prediction horizon. Long-horizon forecasting is very important for the design, operation, and management of WWTPs. One-day forecast of influent flow is generally not sufficient for operators to improve process control planning accordingly. This is why the LSTM model, which could provide long-term forecasts, presents a unique advantage. On top of that, in this study, it is implied that the RF model also has great potential for addressing long-horizon forecasting problems.

4.4.2 Interval Prediction Results

The interval forecast of hourly influent flow rate during the testing period are presented in Fig. 4.6. For each time step, the interval forecasts are obtained from the maximum and minimum values generated by all the trees in the established forest. Overall, almost all of the observed data fall within the predicted ranges. For the confidential station and Woodward, the predicted intervals not only provide the range of influent flow rate but also present the pattern. While for the Humber station, although the intervals capture the observed influent flow rate very well, the lower bound of many predicted values is of 0 m^3/s. That is because there are some nearly zero training samples in the original dataset. These data points could lead one or more predicted values of zero or near zero.

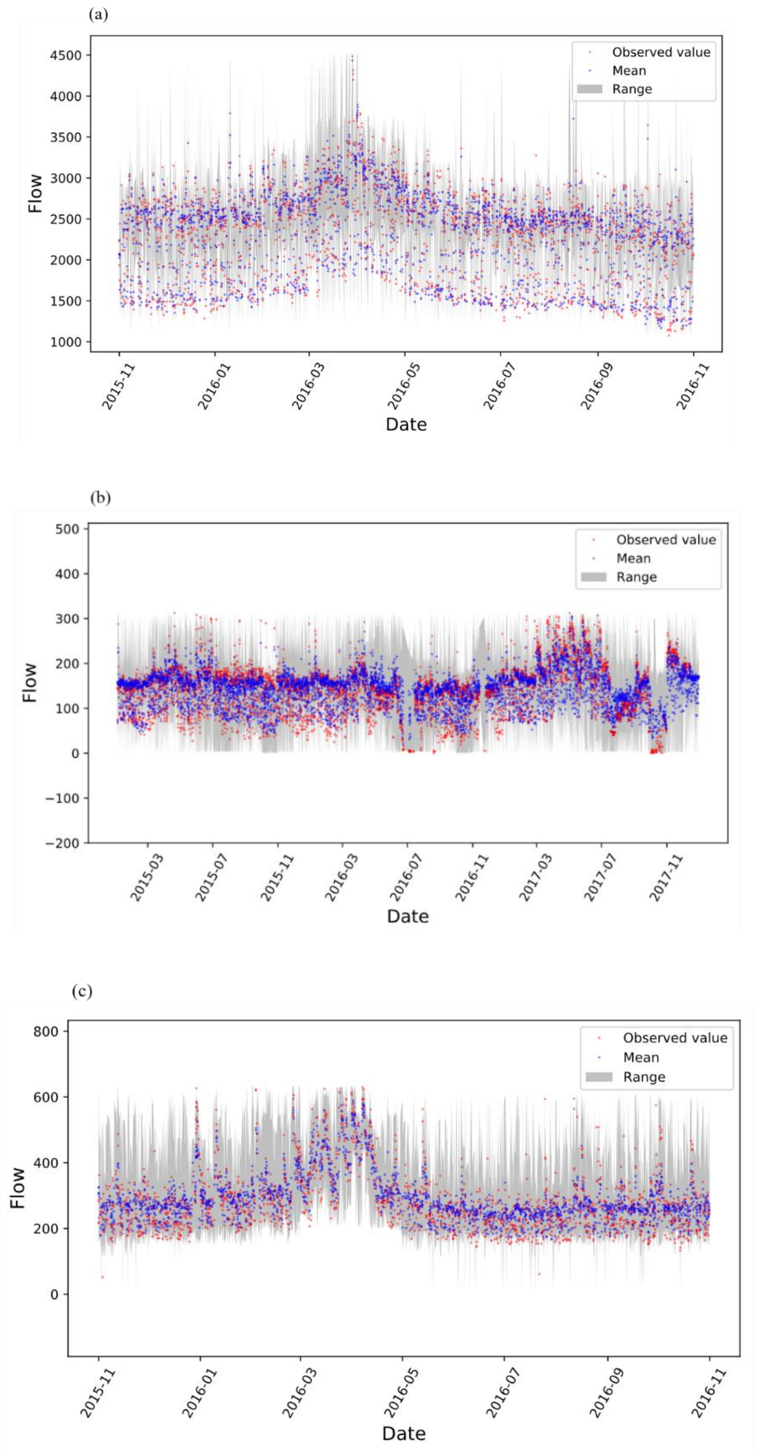

Fig. 4.6 Probabilistic results of (a) the confidential station, (b) the Humber station, and (c) the Woodward station

4.5 Conclusions

In this study, three data-driven models, including MLP, LSTM, and RF, were developed for hourly wastewater influent flow prediction. The developed models were tested for three WWTPs in Canada, including Humber, Woodward, and a confidential plant. The proposed MLP model in this study showed a good performance at the confidential plant with the R^2 value equals 0.829 for the testing period. The performance of MLP model deteriorates when applied to the other two stations. The results of the LSTM model showed a satisfactory fit between observations and predictions at the confidential plant. However, it is hard for the proposed LSTM model to capture the relationship between the influent flow rate and the predictors at the other two stations. The performance of the RF model was more stable compared to MLP and LSTM:, the R^2 values are 0.906, 0.650, and 0.733 for the confidential plant, Humber, and Woodward, respectively. Overall, the advanced machine learning models developed in this study can generate high temporal resolution forecasts of hourly influent flow rate, which is rare in previous studies. This study provides a set of effective tools for predicting the influent flow rate, which can provide valuable decision support for the management of WWTPs.

Acknowledgment

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 engineers at Hydromantis and the WWTPs for their suggestions and comments.

CHAPTER 5 - CONCLUSIONS

The need for effective management of wastewater treatment plants is very clear due to the increasing water scarcity across the globe, which highlights the importance of wastewater influent prediction for wastewater treatment plants (WWTPs). Previous studies indicate that data-driven models have great potential to provide reliable influent flow forecasts and thus to support the operation and management of WWTPs. This thesis explored the practical applications of advanced data-driven techniques in the field of wastewater influent prediction. A number of advanced data-driven techniques were used to predict wastewater influent flow one- and multiple-step ahead using historical flow data, meteorological conditions and time series information such as day of the week. Several WWTPs in Canada were used as case studies, and uncertainty analysis was also conducted to provide more robust support for wastewater management.

Firstly, a random forest (RF) model was developed for daily influent flow prediction. The results showed that RF models perform well in terms of accuracy. Compared to other influent flow rate prediction models, such as autoregression integrated moving average (ARIMA) and multi-layer perceptron (MLP), RF models have the advantage of quantifying each input variable’s contribution. Furthermore, an uncertainty analysis approach was proposed based on the structure of RF model, which allows it to generate probabilistic predictions and provide decision-makers with more information about the risks of extreme influent flow events.

Secondly, three data-driven models including multi-layer perceptron (MLP), long short-term memory neural network (LSTM), and random forest (RF) were developed to predict hourly influent flow rate. These models provided satisfactory influent flow forecasts with a high temporal resolution (i.e., hourly), which addresses a major research gap in the area of wastewater influent prediction. The results suggested that the RF model might be more robust than MLP and LSTM in terms of prediction accuracy when it comes to forecasting with high temporal resolution.

In this research, a set of effective wastewater influent prediction tools were developed for WWTPs. These tools could help WWTPs make reliable and accurate predictions of performance and operating cost. The developed models could also be added to dynamic wastewater modeling software to help improve the design, operation, and management of WWTPs in Canada. This work could provide valuable technical support for making the maximum use of Canada’s existing and future wastewater treatment facilities.

In future research, the prediction accuracy of the proposed models could be further improved by integrating with other models. Improving data quality could also be helpful. In addition, testing for more WWTPs is still needed to further demonstrate the reliability of the proposed models. More work could also be done to investigate other topics such as predictions with a longer horizon, predictions with smaller data samples, and model uncertainties.

REFERENCES

Abdel-Rahman, E. M., Ahmed, F. B., and Ismail, R. (2013). "Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data." *International Journal of Remote Sensing*.

Abunama, T., and Othman, F. (2017). "Time Series Analysis and Forecasting of Wastewater Inflow into Bandar Tun Razak Sewage Treatment Plant in Selangor, Malaysia." *IOP Conference Series: Materials Science and Engineering*, 210(1).

Amatya, D. M., Skaggs, R. W., and Gregory, J. D. (1997). "Evaluation of a watershed scale forest hydrologic model." *Agricultural Water Management*.

Amit, Y., and Geman, D. (1997). "Shape Quantization and Recognition with Randomized Trees." *Neural Computation*, 9(7), 1545–1588.

Aqil, M., Kita, I., Yano, A., and Nishiyama, S. (2007). "A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff." *Journal of Hydrology*, 337(1–2), 22–34.

Bechmann, H., Nielsen, M. K., Madsen, H., and Kjølstad Poulsen, N. (1999). "Grey-box modelling of pollutant loads from a sewer system." *Urban Water*, 1(1), 71–78.

Bengio, Y., Simard, P., and Frasconi, P. (1994). "Learning Long-Term Dependencies with Gradient Descent is Difficult." *IEEE Transactions on Neural Networks*.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V. (2013). “Characterising performance of environmental models.” *Environmental Modelling and Software*, Elsevier Ltd, 40, 1–20.

Berthouex, P. M., and Box, G. E. (1996). “Time series models for forecasting wastewater treatment plant performance.” *Water Research*, water research, 30(8), 1865–1875.

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q., and Zhou, P. (2019). “Influent Forecasting for Wastewater Treatment Plants in North America.” *Sustainability*, 11(6), 1764.

Breiman, L. (2001a). “Random Forests.” *Machine Learning*.

Breiman, L. (2001b). “Random forests.” *Machine Learning*, 45(1), 5–32.

Breiman, L. (2017). Classification and regression trees. Routledge.

Butler, D., and Graham, N. J. D. (1995). “Modeling Dry Weather Waste-Water Flow in Sewer Networks.” *Journal of Environmental Engineering-Asce*, 121(2), 161–173.

Büyükalaca, O., Bulut, H., and Yılmaz, T. (2001). “Analysis of variable-base heating and cooling degree-days for turkey.” *Applied Energy*, 69(4), 269–283.

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

Chan, et al. (2006). *Handbook of Water and Wastewater Treatmenrt Plant Operations*.

Chaplin, M. F. (2001). “Water: Its importance to life.” *Biochemistry and Molecular Biology Education*, 29(2), 54–59.

Charaniya, N. A., and Dudul, S. V. (2012). “Focused time delay neural network model for rainfall prediction using Indian ocean dipole index.” *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*, 851–855.

Chiang, Y.-M., Hao, R.-N., Zhang, J.-Q., Lin, Y.-T., and Tsai, W.-P. (2018). “Identifying the Sensitivity of Ensemble Streamflow Prediction by Artificial Intelligence.” *Water*, 10(10), 1341.

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). “Gene selection and classification of microarray data using random forest.” *BMC Bioinformatics*.

Dietterich, T. G. (2000). “An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and ranomization.” *Machine Learning*, 40(2), 139–157.

Dürrenmatt, D. J. Ô., and Gujer, W. (2012). “Data-driven modeling approaches to support wastewater treatment plant operation.” *Environmental Modelling and Software*, Elsevier Ltd, 30, 47–56.

El-Din, A. G., and Smith, D. W. (2002). “A neural network model to predict the wastewater inflow incorporating rainfall events." *Water Research*, 36(5), 1115–1126.

Farhat, N. H. (2002). "Photonic neural networks and learning machines." *IEEE Expert*, 7(5), 63–72.

Fernandez, F. J., Seco, A., Ferrer, J., and Rodrigo, M. A. (2009). "Use of neurofuzzy networks to improve wastewater flow-rate forecasting." *Environmental Modelling and Software*, Elsevier Ltd, 24(6), 686–693.

Fuller, M. F., Conover, W. J., and Gibbons, J. D. (1973). "Practical Nonparametric Statistics." *The Statistician*.

Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). "Random forests for land cover classification." *Pattern Recognition Letters*, 27(4), 294–300.

Grömping, U. (2009). "Variable importance assessment in regression: Linear regression versus random forest." *American Statistician*, 63(4), 308–319.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling." *Journal of Hydrology*, Elsevier B.V., 377(1–2), 80–91.

Ho, T. K. (1998). "The random subspace method for constructing decision forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hochreiter, S., and Schmidhuber, J. (1997). "Long Short-Term Memory." *Neural*

Computation.

Ismail, A. A., Wood, T., and Bravo, H. C. (2018). “Improving Long-Horizon Forecasts with Expectation-Biased LSTM Networks.”

Jain, S. K., and Sudheer, K. P. (2008). “Fitting of Hydrologic Models: A Close Look at the Nash–Sutcliffe Index.” *Journal of Hydrologic Engineering*, 13(10), 981–986.

Khan, M. S., Coulibaly, P., and Dibike, Y. (2006). “Uncertainty analysis of statistical downscaling methods.” *Journal of Hydrology*, 319(1–4), 357–382.

Kim, M., Kim, Y., Kim, H., Piao, W., and Kim, C. (2015). “Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant.” (December).

Kim, M., Kim, Y., Kim, H., Piao, W., and Kim, C. (2016). “Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant.” *Frontiers of Environmental Science and Engineering*, 10(2), 299–310.

Kuo, J., Chen, Y., Hsieh, S., Lin, S., and Liao, Y. (2010). “A pattern-oriented approach to development of a real-time storm sewer simulation system with an SWMM model.” *Journal of Hydroinformatics*, 12(4), 408–423.

Langergraber, G., Rieger, L., Winkler, S., Alex, J., Wiese, J., Owerdieck, C., Ahnert, M., Simon, J., and Maurer, M. (2004). “A guideline for simulation studies of wastewater treatment plants.” *Water Science and Technology*, 50(7), 131–138.

Levene, H. (1960). “Robust tests for equality of variances.” *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*.

Li, Z., Huang, G., Han, J., Wang, X., Fan, Y., Cheng, G., Zhang, H., and Huang, W. (2015). “Development of a Stepwise-Clustered Hydrological Inference Model.” *Journal of Hydrologic Engineering*, 20(10), 04015008.

Liaw, a, and Wiener, M. (2002). “Classification and Regression by randomForest.” *R news*, 2(December), 18–22.

Loh, W. (2008). “Encyclopedia of Statistics in Quality and Reliability.” 315–323.

Mello, C. R., Viola, M. R., Norton, L. D., Silva, A. M., and Weimar, F. A. (2008). “Development and application of a simple hydrologic model simulation for a Brazilian headwater basin.” *Catena*, Elsevier B.V., 75(3), 235–247.

Moriasi, D. N., Arnold, J. G., Liew, M. W. Van, Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). “MODEL EVALUATION GUIDELINES FOR SYSTEMATIC QUANTIFICATION OF ACCURACY IN WATERSHED SIMULATIONS.” *Transactions of the ASABE*, 50(3), 885–900.

Nadiri, A. A., Shokri, S., Tsai, F. T. C., and Asghari Moghaddam, A. (2018). “Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model.” *Journal of Cleaner Production*, Elsevier Ltd, 180, 539–549.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models part I - A discussion of principles." *Journal of Hydrology*.

Ömer Faruk, D. (2010). "A hybrid neural network and ARIMA model for water quality time series prediction." *Engineering Applications of Artificial Intelligence*.

Pagano, T. C., Garen, D. C., Perkins, T. R., and Pasteris, P. A. (2009). "Daily updating of operational statistical seasonal water supply forecasts for the Western U.S." *Journal of the American Water Resources Association*.

Pal, M. (2005). "Random forest classifier for remote sensing classification." *International Journal of Remote Sensing*, 26(1), 217–222.

Pal, S. K. (1992). "NSankar 1992.pdf." *IEEE Transactions on Neural Networks*.

Pisa, I., Sant, I., Vicario, J. L., Morell, A., Vilanova, R., Pisa, I., Santin, I., Vicario, J., Morell, A., and Vilanova, R. (2018). "A Recurrent Neural Network for Wastewater Treatment Plant effluents ' prediction." 1(1), 5–7.

Ponomarenko, A., Avrelin, N., Naidan, B., and Boytsov, L. (2012). "Comparative Analysis of Data Structures for Approximate Nearest Neighbor Search." *Journal of Mathematical Sciences,*.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., and Woo, W. (2015). "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting." 1–9.

SMOLA, A. J. and B. S. (2004). “A tutorial on support vector regression - art%3A10.1023%2FB%3ASTCO.0000035301.49549.88.pdf.” *Statistics and Computing,* 199–222.

Solomatine, D. P., and Ostfeld, A. (2007). “Data-driven modelling: some past experiences and new approaches.” *Journal of Hydroinformatics,* 10(1), 3–22.

Szelag, B., Bartkiewicz, L., Studziński, J., and 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.” *Archives of Environmental Protection*, 43(3), 74–81.

Taormina, R., Chau, K. W., and Sethi, R. (2012). “Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon.” *Engineering Applications of Artificial Intelligence,* Elsevier, 25(8), 1670–1676.

Taud, H., & Mas, J. F. (2018). Multilayer perceptron (MLP). In Geomatic Approaches for Modeling Land Change Scenarios (pp. 451-455). Springer, Cham.

Tehrany, M. S., Pradhan, B., and Jebur, M. N. (2013). “Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS.” *Journal of Hydrology*, Elsevier B.V., 504, 69–79.

Universit, I. T., and Betreuer, B. (1991). “IM FACH INFORMATIK Untersuchungen zu dynamischen neuronalen Netzen." *Iclr*, 14.

Vanrolleghem, P. A., Gernaey, K. V., Rosen, C., Pons, M.-N., Alex, J., Copp, J., and Jeppsson, U. (2007). "Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs." *Water Science and Technology*, 53(1), 287–295.

Verma, A., Wei, X., and Kusiak, A. (2013). "Predicting the total suspended solids in wastewater: A data-mining approach." *Engineering Applications of Artificial Intelligence*, Elsevier, 26(4), 1366–1372.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X. (2015). "Flood hazard risk assessment model based on random forest." *Journal of Hydrology*, Elsevier B.V., 527, 1130–1141.

Wei, X., and Kusiak, A. (2015). "Short-term prediction of influent flow in wastewater treatment plant." *Stochastic Environmental Research and Risk Assessment*, 29(1), 241–249.

Wei, X., Kusiak, A., and Sadat, H. R. (2013). "prediction of influent flow rate: data-mining approach." 139(February), 118–123.

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S. (1998). "Statistical downscaling of general circulation model output: A comparison of methods." *Water Resources Research*.

Yeh, A. G. O., and Li, X. (2002). “Urban simulation using neural networks and cellular automata for land use planning.” *Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling*.

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., and Liu, J. (2017). “LSTM network: a deep learning approach for short-term traffic forecast.” *IET Intelligent Transport Systems*, 11(2), 68–75.

Zhe Zhou, Chenglin Wen, and Chunjie Yang. (2015). “Fault Detection Using Random Projections and k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes.” *IEEE Transactions on Semiconductor Manufacturing*, 28(1), 70–79.