ISEIS

Journal of Environmental Informatics Letters 1(2) 87-93 (2019)

Journal of Environmental Informatics Letters

www.iseis.org/jeil

Short-Term Wastewater Influent Prediction Based on Random Forests and Multi-Layer Perceptron

P. Zhou[1], Z. Li[1 *], S. Snowling[2], R. Goel[2], and Q. Zhang[1, 3]

[1] *Department of Civil Engineering, McMaster University, Hamilton, Ontario, L8S 4L7, Canada*
[2] *Hydromantis Environmental Software Solutions, Inc., 407 King Street West, Hamilton, Ontario, L8P 1B5, Canada*
[3] *School of Management, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China*

Received 09 March 2019; revised 23 May 2019; accepted 27 May 2019; published online 30 June 2019

ABSTRACT. Influent flow rate is a crucial parameter closely related to the plant-wide control of wastewater treatment plants (WWTPs). In this study, a random forest (RF) model and a multi-layer perceptron (MLP) model are developed for hourly influent flow rate prediction at a confidential WWTP in Canada. Both models perform well on predicting influent flow rate one-step ahead. The coefficient of determination (R^2) values of MLP and RF for the testing data set are 0.927 and 0.925, respectively. Furthermore, the multi-step ahead prediction accuracy of the proposed models is discussed. To improve the multi-step ahead prediction accuracy of the RF model, time-tag information is transformed to numerical values and then fed into the RF model as input. The R^2 values of the RF model for the testing data set with and without time-tag information are 0.334 and 0.811, respectively. The results show that the RF model's performance for multi-step ahead prediction is heavily affected by the time-tag information. Including time-tag information as input could dramatically improve the multi-step ahead prediction accuracy. In this study, the RF model shows more robust performance than the MLP model on solving short-term wastewater influent prediction problems.

Keywords: short-term prediction, multi-step ahead, hourly, influent flow, WWTPs

1. Introduction

Influent flow rate is a crucial parameter which significantly affects the management and operation of wastewater treatment plants (WWTPs). Wastewater characteristics, such as total suspended solids (TSS) and biochemical oxygen demand (BOD), are strongly correlated to influent flow rate (Bechmann et al., 1999). To maintain stable effluent quality, it is critical to adjust the treatment process according to the influent flow rate, as well as the pollutant concentrations of the influent (Wei and Kusiak, 2015).

In recent decades, studies on plant-wide control of WWTPs have gained a lot of attention. Alongside the development of sensor technology, plant-wide monitoring networks have been widely implemented, and feedback control of wastewater treatment unit processes has been widely used at WWTPs around the world (Olsson et al., 1998; Jeppsson et al., 2002; Dürrenmatt and Gujer, 2012; Campisano et al., 2013). These monitoring networks can provide abundant data for the feedback control of unit processes with fast dynamics (Ma et al., 2014). However, control delay is a common challenge for unit processes with slow dynamics, such as biochemical treatment processes (Schütze et al., 2004). For relatively slow treatment processes, feed-forward signals are needed for supporting realtime feedback control, especially when there are disturbances of various time scales (Shen et al., 2009; Ma et al., 2014). Therefore, it is desirable to know the influent flow rate in advance, and this is when the prediction of the influent flow rate becomes valuable.

Previously, models of different complexity levels have been studied for multi-step ahead prediction (Ismail et al., 2018). However, as the prediction horizon increases, the prediction accuracy of those models usually decreases drastically. Wei and Kusiak (2015) compared the performance of a Multi-layer Perceptron (MLP) model and a Dynamic Neural Network (DNN) model on short-term prediction of influent flow rate. The results showed that the DNN model has a better performance than the MLP model when it comes to a longer prediction horizon. On the other hand, a modified long short-term memory (LSTM) model focusing on long-term prediction was studied by Ismail et al. (2018). The results showed that although the LSTM model and the recurrent neural network (RNN) model perform better than the traditional MLP model on short-term prediction, the prediction accuracy decreases when the prediction horizon expands beyond a few time steps in the future. Overall, many models were proposed for short-term influent flow rate prediction and neural network models such as MLP, DNN and LSTM have been widely studied for short-term or even long-term prediction. Although these neural network models are able to partly address the multi-step ahead prediction

*Corresponding author. Tel.: +1 (905) 5259140; fax: +1 (905) 5727944.
E-mail address: zoeli@mcmaster.ca (Z. Li).

ISSN: 2663-6859 print/2663-6867 online
 doi:10.3808/jeil.201900010

problems, they are also facing many challenges, such as time-consuming calibration and overfitting problems.

Meanwhile, the Random Forest (RF) model has gained a lot of attention recently, and it has been applied in a wide range of areas (Pal, 2005; D áz-Uriarte and Alvarez de Andr és, 2006; Abdel-Rahman et al., 2013). In comparison with neural network models, RF has a distinct advantage in avoiding overfitting because of the bootstrap method it uses. Additionally, the RF model is able to illustrate the contribution of each input variable to the predicted target. However, as an effective and promising machine learning model, the use of this promising method in wastewater influent prediction is limited. To the best of the authors' knowledge, RF has not yet been used for the multi-step ahead prediction of wastewater inflow rate.

Therefore, the objective of this study is to explore the potential of RF for short-term wastewater influent flow rate prediction. This entails the following four tasks: (1) developing RF and MLP models with different prediction horizons for influent flow rate prediction; (2) applying the developed models and predicting the hourly influent flow rate for different prediction horizons at a real-world WWTP; (3) evaluating the performance of the proposed models using different statistical criteria; (4) exploring the approaches to improve RF's performance for multi-step ahead prediction. This study will provide valuable decision support for wastewater treatment process control. It will also provide an insight into the application of RF models in short-term wastewater prediction.

2. Data and Study Area

A confidential WWTP is used in this study to illustrate the performance of the proposed short-term influent flow rate prediction models. The inflow of this WWTP consists of sanitary sewage, infiltration, a small portion of stormwater. The influent flow rate data were provided by Hydromantis, which is a software development company in the water and wastewater sector. Influent flow rate data with 15-minute intervals from 1st November 2015 to 31st October 2016 are collected. Then the data with 15-minute intervals are resampled to hourly data, resulting in a total of 8,783 samples. The time series plot of hourly influent flow rate is presented in Figure 1. Meanwhile, hourly weather data are provided by Dark Sky, a company that specializes in weather forecasting and visualization. The weather data are matched with the hourly influent flow rate data with the same data length and frequency.

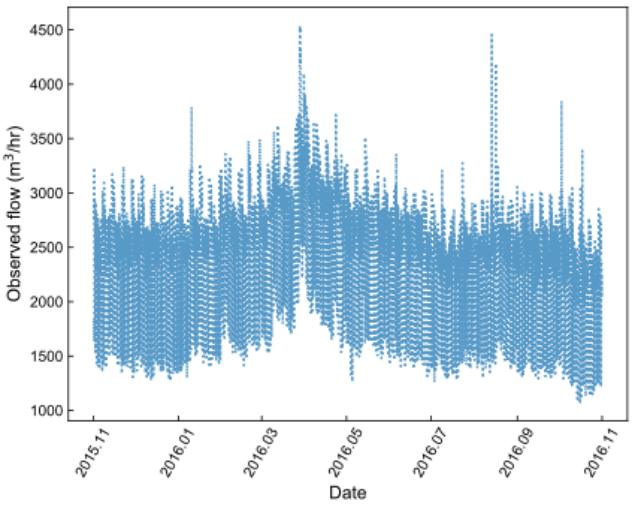

Figure 1. Time series of hourly influent flow rate at the confidential WWTP.

3. Methodology

3.1. Random Forests

The random forest (RF) method was systematically proposed by Breiman in 2001. A RF is an ensemble of decision tree classifiers (Breiman, 2001). Each decision tree is constructed using parts of samples taken from the original data set through a bootstrap method. After forest is developed, each decision tree can make a prediction, and then a majority vote or an average value can be taken as the predicted value of the RF (Figure 2).

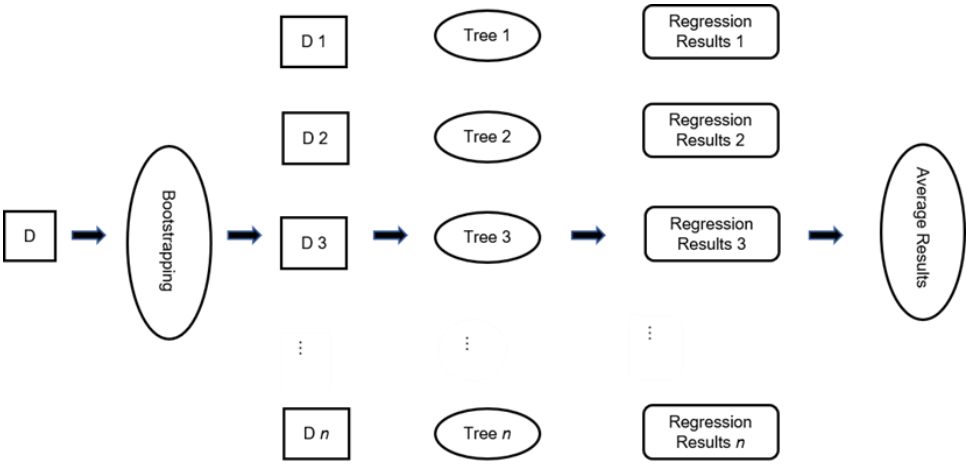

Figure 2. Development process of a RF model.

There are two types of decision trees (classification tree and regression tree). They can be used to solve classification and regression problems, respectively. Decision trees deal with the prediction of an output variable *y* given a vector of predictor variables *x*. If *y* is taking a real value which is continuous or discrete (e.g., the weight of a car or the number of accidents), the problem is regression. Otherwise, if the domain of *y* is a finite set of unordered values (e.g., the type of car or its country of origin), the problem is classification (Loh, 2008). In this study, our target is influent flow rate, which means regression trees are performed. The development process of a RF is summarized as follows:

1) For an original training data set including *A* samples, *A* samples are randomly drawn, with replacement, from the original data set. These *A* samples form a new training set for growing a regression tree.
2) A classification and regression tree (CART) algorithm with mean square error (MSE) as the split criterion is used to build the regression tree. CART is a procedure for binary recursive partitioning (Speybroeck, 2012). Readers are referred to Breimin (1984) and Steinberg (2015) for more details about CART. The number of trees in the forests (*n*) is an important parameter. The values of *n* is determined by a grid search cross validation method in this study. Combined with cross validation and given a prior range of *n*, this grid search method is able to obtain the best value of *n* by evaluating the performance of each value among the given ranges. Other parameters that can affect model performance, such as the number of variables tried at each split node (*m*), can be also determined by the grid search cross validation method. It is also worth mentioning that there is no pruning process during constructing each regression tree, and each regression tree is grown to the largest in this study.
3) After the regression trees are built, each tree can produce one predicted value. Thus, *n* predicted values are generated by *n* trees.
4) In this study, the average value of the *n* predicted values is taken as the final predicted value of the RF.

3.2. Multi-Layer Perceptron

Artificial neural networks are proved to be a useful tool for predictive modeling in many disciplines (Paegelow, 2018). In this study, a typical artificial neural network model called multi-layer perceptron (MLP) is tested and compared with the RF model. The MLP model can approximate virtually any measurable function and there is no assumptions regarding the distribution of input data (Gardner and Dorling, 1998). It is widely used in hydrologic modeling and has been proved to be effective (Coulibaly and Evora, 2007; Solaimany-Aminabad et al., 2014; Wei and Kusiak, 2015). In comparison with other traditional data-driven models, such as dynamic neural networks (DNN) and long short-term memory (LSTM) neural networks, MLP is easier to construct and tune, which makes it ideal to be used as a baseline model option for comparison purposes.

MLP consists of multiple layers of neurons that interact with weighted connections (Pal, 1992). Generally, an MLP model includes an input layer, a number of hidden layer(s), and an output layer. The number of hidden layers (n_l) and the number of neurons in each layer (n_i) are two essential parameters of MLP. How to optimize n_l and n_i has been extensively studied. Many previous studies suggested that one hidden layer is sufficient for most problems (Rajasekaran and Amalraj, 2002; Xu, 2008; Panchal et al., 2011; Student, 2012). As for determining the value of n_i, there are many rule-of-thumb methods. For example, n_i should be 2/3 the size of the input layer, plus the size of the output layer. Meanwhile, other approaches such as Akaike's Information Criterion can also be adopted (Panchal et al., 2011). In this study, the values of n_l and n_i were determined according to previous studies and combined with the grid search method mentioned in section 3.1. Since weights are the key to the MLP models, the process of determining weights for an MLP model which contains one hidden layer and one neuron is presented in Figure 3 and is articulated in three steps:

1) Initial weights close to zero are randomly assigned to the input variables.
2) Through an activation function, the cost function related to the input variables is calculated using:

$$y_i = f\left(\sum_{i=1}^{i} w_i x_i\right) \tag{1}$$

$$J = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

where $f(x)$ is the activation fucntion; J is the cost function; m is the number of training samples; x_i is the i^{th} input variable; y_i is the i^{th} predicted value; and $\hat{y}_i$ is the i^{th} observed value.

3) A gradient descent algorithm is employed to find the optimal value of the cost function. Then the corresponding optimal weights can be updated. The descent range (D) can be determined as follows:

$$D = \alpha \frac{\partial J}{\partial x_i} \tag{3}$$

where α is the learning rate.

3.3. Model Training and Testing

Both the RF and the MLP models are performed for influent flow rate prediction at a confidential WWTP. Firstly, 75% of the hourly data samples are marked as training data and the other 25% of samples are testing data. It is worth mentioning that, in this study, three-folds cross validation is used to tune the parameters during the training process. Both weather information and historic influent flow rate data are used as model input. A single-step ahead model (predicting the influent flow rate at time $t + 1$) is built, using the weather information at time t, as well as the historic influent flow rate at time t, $t - 1$, and $t - 2$, as input data. The previous study of Wei and Kusiak

(2015) indicated that including influent flow data five or more time steps back could hardly improve the model's prediction accuracy. On the contrary, it may lead to significant computation and thus decrease the modeling efficiency. Furthermore, according to the correlation analysis, influent flow rate at time t, t - 1, and t - 2 show the highest correlation with the influent flow rate at time t. Thus, only 1-, 2-, and 3-hour prior influent flow rate data are used as training input in this study.

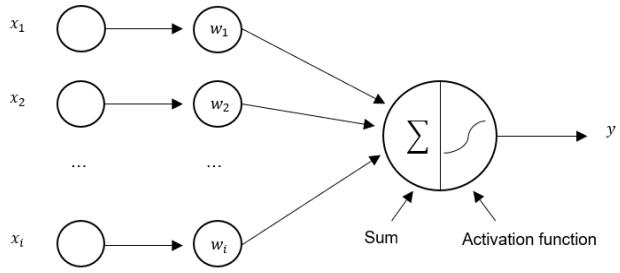

Figure 3. Schematic diagram of a hidden neuron in an MLP model. x_i is the i^{th} input variable, w_i is the i^{th} weight, and y is the output target.

RF and MLP models are also built for predicting the influent flow rate at t + 2, t + 3, t + 4, t + 5, t + 6, t + 24, t + 48, and t + 72 using the same model development procedures described above. These models are mainly used for testing the performance of wastewater influent flow rate prediction with different prediction horizons. For different prediction horizon models from t + 1 ~ t + 72, the input variables are the same, including weather variables (temperature, humidity, precipitation, wind speed, and wind bearing) at time t and historic influent flow rate at t, t - 1, and t - 2.

4. Results and Discussion

4.1. Single-Step Ahead Prediction Results

For the single-step ahead RF model, a total of 8,779 samples is used for training and testing. Figure 4 is the scatter plot of predicted and observed influent flow rate. The root mean square errors (RMSE) and coefficient of determination (R^2) values are also showed in the figure. It is indicated that, overall, the proposed single-step ahead RF model is satisfactory. However, it can be observed that for there is an underestimation when the observed value goes above 3,500 m^3/hr. This may be because the range of predicted values heavily depends on the range of observed values in the training set, which limits the model's capability of capturing extreme values. More specifically, in this study, the values of training samples range from 1,500 m^3/hr to 3,500 m^3/hr, so the predicted values are more likely to fall in this range. This implies that it is very important to increase the representativeness and diversity of training samples when building RF models.

For the single-step ahead MLP model, the same 8,779 samples are used for training. The MLP model with the best testing results has one hidden layer and 25 hidden neurons. Figure 5 is the scatter plot of predicted and observed influent flow rate during the training and testing periods. It can be found that, though some extreme points are not predicted precisely, this single-step ahead model shows an overall satisfactory prediction performance, with an R^2 value of 0.927 during the testing period.

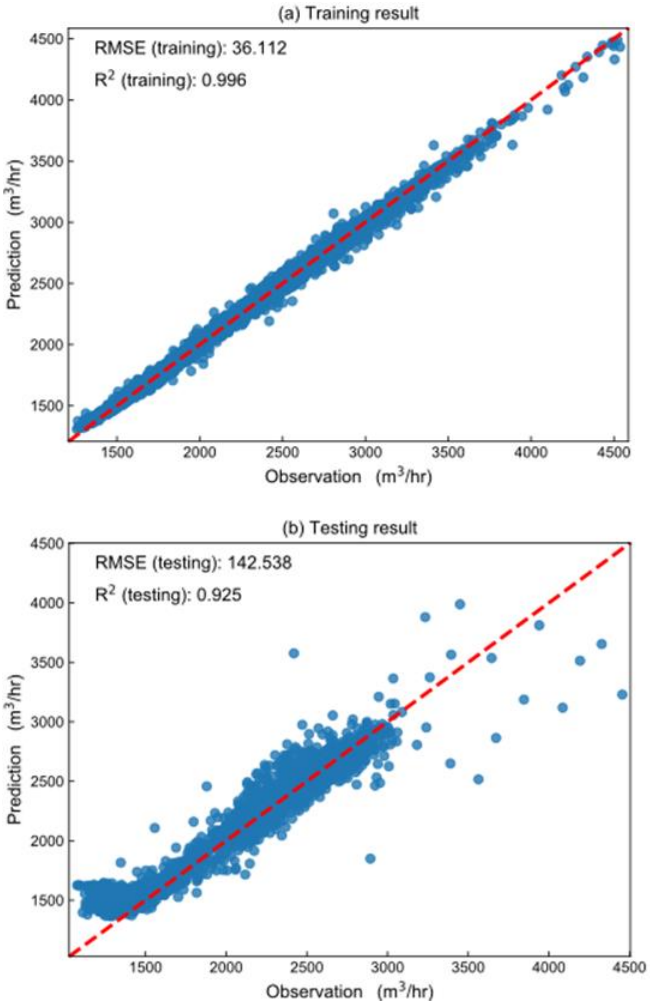

Figure 4. Training (a) and testing (b) results of the RF model with a prediction horizon of t + 1.

4.2. Multi-Step Ahead Prediction Results

Multi-step ahead models based on MLP and RF are built to describe the relationships between the input variables at time t and the output target at time $t + n$. Overall, the prediction accuracy worsens significantly as the prediction horizon increases. Figure 6 shows the changes of R^2 and RMSE results during the testing period as the prediction horizon increases. The RMSE of MLP model increases from 141.47 to 508.60 when the prediction horizon changes from t + 1 to t + 6. Meanwhile, the value of R^2 decreases dramtically from more than 0.92 to less than 0.18. For the RF model, RMSE increses from 142.54 to 446.20 and R^2 decreases from 0.93 to 0.33. Moreover, the increase of prediction horizon has a stronger negative impact on the MLP method than the RF method.

Interestingly, when the time horizon changes to t + 24, t +

48, and $t + 72$, the performance of both methods improve slightly in terms of RMSE and R^2. This indicates that the influent flow rate follows a recurring daily pattern.

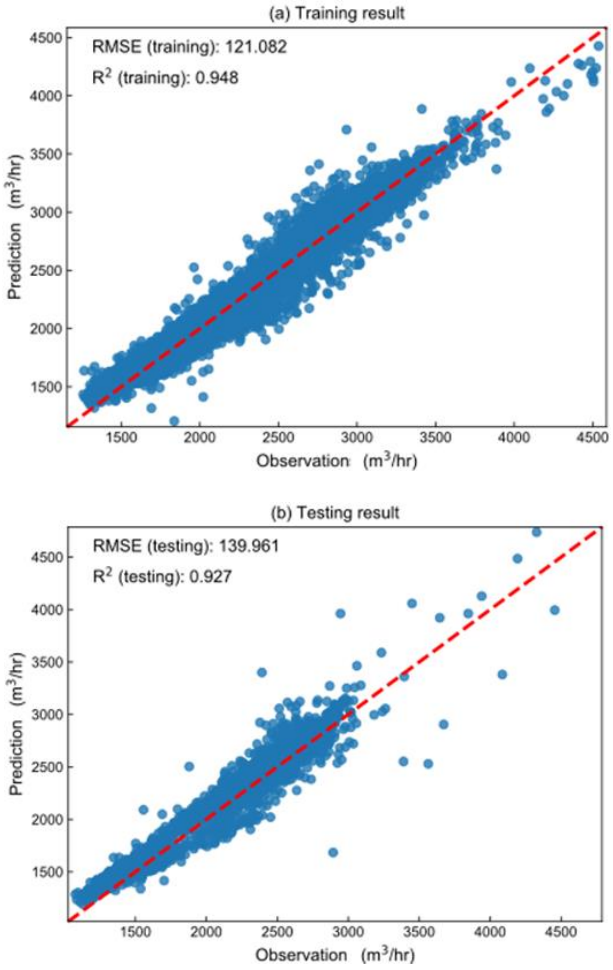

Figure 5. Training (a) and testing (b) results of the MLP model with a prediction horizon of $t + 1$.

4.3. Discussion on Improving Multi-Step Prediction Accuracy

To improve the accuracy of multi-step prediction, many strategies have been employed. For instance, in the field of neural networks, dynamic neural network (DNN) models and long short-term memory network (LSTM) models show better performance than the traditional MLP method for multi-step ahead prediction (Wei and Kusiak, 2015; Ismail et al., 2018). However, for the RF method, discussions on how to improve the accuracy of multi-step ahead prediction are limited. Therefore, possibilities to improve multi-step ahead prediction are explored.

The reason why DNN and LSTM generally performs better while solving multi-step ahead prediction problems is that these models include time series information as input variables and they are able to address time series information in an efficient manner. Inspired by this, time-tag information is further introduced to the proposed RF model in this study. The time series information is first transformed from time tags to numeric values, such as hour of the day (1, 2, … , 24), day of the week (1, 2, … , 7), and month of the year (1, 2, … , 12). Then, these values are introduced to the RF model as input features. To demonstrate the feasibility of this method, the RF model with a prediction horizon of $t + 6$ is rebuilt. Weather and flow data at time t, together with the time-tag information, are used as input to predict the output target at time $t + 6$. Figure 7 shows the performance of the RF models before and after including time-tag information as input. It can be observed that the accuracy of the RF model improves dramatically after including the time-tag information. The RMSE value decreases from 436.21 to 228.13, while the R^2 value increases from 0.33 to 0.81. The main reason for the improvement is that the time-tag information enables the RF model to find the data pattern more effectively and address time series data more properly. Take the prediction of the influent flow rates at 9:00 am on Monday mornings for example. After introducing the time-tag information, such as hour of the day (9), day of the week (1), RF models tend to classify all flow rate at 9:00 am on Monday mornings into the same subset. Thus, the data pattern can be better described using the resulted regression trees in the RF, and the prediction accuracy can be improved.

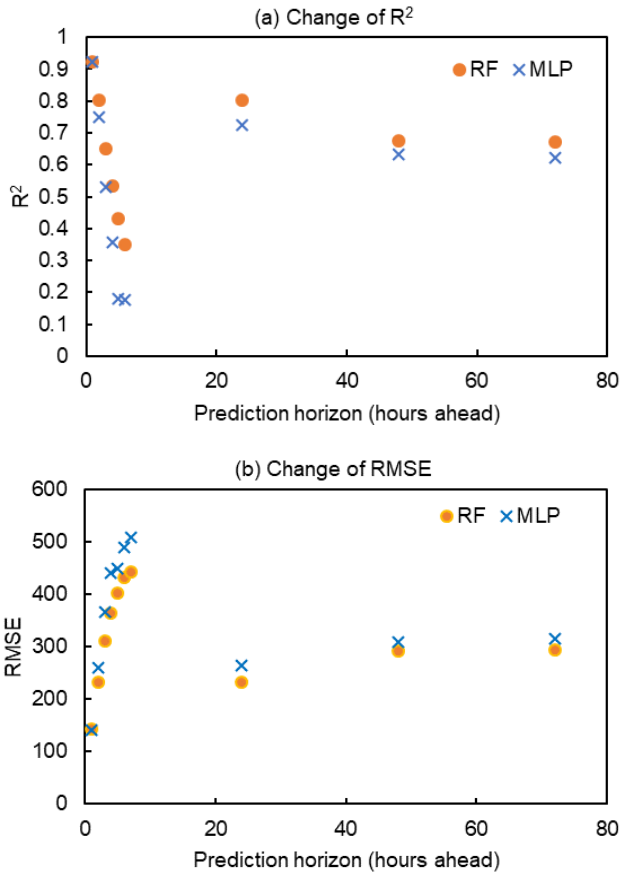

Figure 6. Changes of R^2 (a) and RMSE (b) as prediction horizon increases.

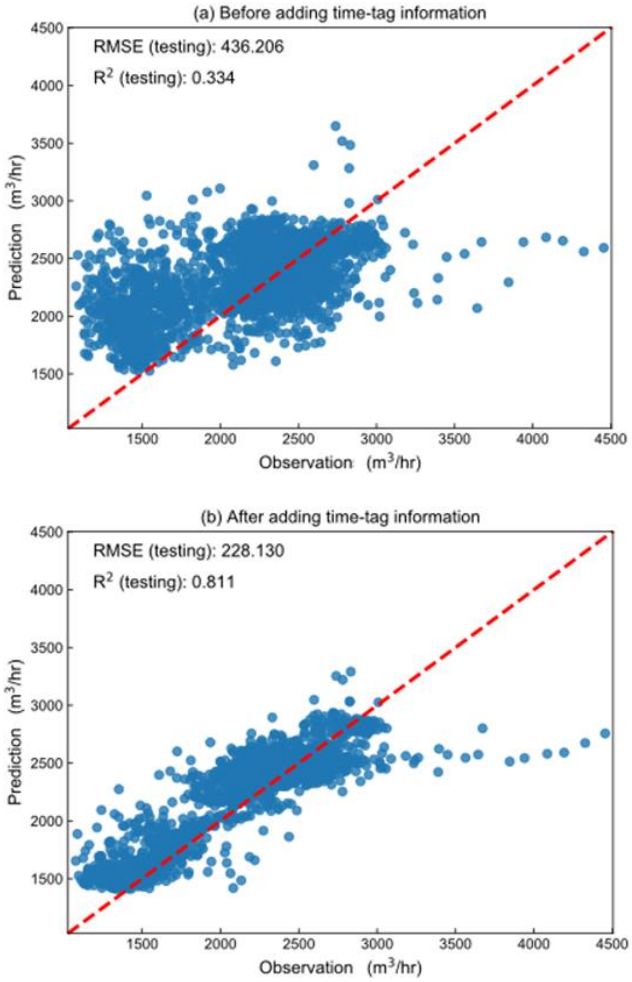

Figure 7. Performance of RF models with a prediction horizon of $t + 6$: (a) before adding time-tag information and (b) after adding time-tag information.

5. Conclusions

In this study, a number of RF and MLP models were developed for short-term hourly wastewater influent flow rate prediction at a confidential WWTP in Canada. The proposed models were both able to capture the nonlinear weather-flow relationships for single-step ahead prediction. However, the accuracy of multiple-step ahead prediction were unsatisfactory. To address this issue, time-tag information, such as hour of the day, day of the week, and month of the year, were transformed into numeric values to be included as training inputs for the RF model. The RF model that includes time-tag information was tested and compared with the RF model built in the traditional way. For six-step ahead prediction, the root mean square error (RMSE) value decreased from 436.21 to 228.13; while the coefficient of determination (R^2) increased from 0.33 to 0.81. The results indicated that including time-tag information could help enhance RF's multi-step ahead prediction accuracy. For future studies, more methods to decode time series information and improve RF's performance should be explored.

Acknowledgments. This research was supported by the Southern Ontario Water Consortium and Hydromantis Environmental Software Solutions, Inc. The authors would like to thank the engineers at the WWTP for their comments.

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*, 34(2), 712-728. https://doi.org/10.1080/01431161.2012.713142.

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. https://doi.org/10.1016/S1462-0758(99)00007-2.

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

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. https://doi.org/10.1080/1573062X.2013.763996.

Coulibaly, P. and Evora, N. D. (2007). Comparison of neural network methods for infilling missing daily weather records, *Journal of Hydrology*, 341(1-2), 27-41. https://doi.org/10.1016/j.jhydrol.2007.04.020.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest, *BMC bioinformatics*, 7, 3. https://doi.org/10.1186/1471-2105-7-3.

Dürrenmatt, D.J.Ô., and Gujer, W. (2012). Data-driven modeling approaches to support wastewater treatment plant operation, *Environmental Modelling and Software*, 30, 47-56. https://doi.org/10.1016/j.envsoft.2011.11.007

Gardner, M.W. and Dorling, S.R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences, *Atmospheric Environment*, 32(14-15), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

Ismail, A.A., Wood, T., and Bravo, H. C. (2018). *Improving Long Horizon Forecasts with Expectation-Biased LSTM Networks*, Xiv: 1804.06776.

Jeppsson, U., Alex, J., Pons, M.N., Spanjers, H., and Vanrolleghem, P.A. (2002). Status and future trends of ICA in wastewater treatment - A European perspective, *Water Science and Technology*, 45(4-5), 485-494.https://doi.org/10.2166/wst.2002.0653.

Loh, W. (2008). *Encyclopedia of Statistics in Quality and Reliability*.

Ma, S., Zeng, S., Dong, X., Chen, J., and Olsson, G. (2014). Short-term prediction of influent flow rate and ammonia concentration in municipal wastewater treatment plants, *Frontiers of Environmental Science and Engineering*, 8(1), 128-136. https://doi.org/10.1007/s11783-013-0598-9.

Olsson, G., Aspegren, H., and Nielsen, M.K. (1998). Operation and control of wastewater treatment - A Scandinavian perspective over 20 years, *Water Science and Technology*, 37(12), 1-3. https://doi.org/10.2166/ wst.1998.0484

Paegelow, M. (2018). *Geomatic Approaches for Modeling Land Change Scenarios*.

Pal, M. (2005). Random forest classifier for remote sensing classification, *International Journal of Remote Sensing*, 26(1), 217-222. https://doi.org/10.1080/01431160412331269698.

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

Panchal, G., Ganatra, A., Kosta, Y.P., and Panchal, D. (2011). Behaviour Analysis of Multilayer Perceptrons with Multiple Hidden Neurons and Hidden Layers, *International Journal of Computer Theory and Engineering*, 3(2), 332-337. https://doi.org/10.7763/IJCTE.

2011.V3.328

Rajasekaran, S. and Amalraj, R. (2002). Predictions of design parameters in civil engineering problems using SLNN with a single hidden RBF neuron, *Computers and Structures*, 80(31), 2495-2505. https://doi.org/10.1016/S0045-7949(02)00213-4.

Schütze, M., Campisano, A., Colas, H., Schilling, W., and Vanrolleghem, P.A. (2004). Real time control of urban wastewater systems Where do we stand today? *Journal of Hydrology*, 299(3-4), 355-348. https://doi.org/10. 1016/j.jhydrol.2004.08.010

Shen, W., Chen, X., Pons, M.N., and Corriou, J.P. (2009). Model predictive control for wastewater treatment process with feedforward compensation, *Chemical Engineering Journal*, 155(1-2), 161-174. https://doi.org/10. 1016/j.cej.2009.07.039

Shuxiang Xu and L. C. (2008). A Novel Approach for Determining the Optimal Number of Hidden Layer Neurons for FNN's and Its Application in Data Mining, *5th International Conference on Information Technology and Applications*, Icita, 683-686.

Solaimany-Aminabad, M., Maleki, A., and Hadi, M. (2014). Application of artificial neural network (ANN) for the prediction of water treatment plant influent characteristics, *Journal of Advances in Environmental Health Research*, 1(2), 1-12. https://doi.org/10. 22102/jaehr.2013.40130

Speybroeck, N. (2012). Classification and regression trees, *International Journal of Public Health*, 57(1), 243-6. https://doi.org/10. 1007/s00038-011-0315-z

Student, S.K. (2012). Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture, *International Journal of Engineering Trends and Technology*, 3, 714-717.

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. https://doi.org/10.1007/s004 77-014-0889-0