UNCERTAINTIES OF MACHINE LEARNING IN PREDICTING THE HYDROLOGICAL RESPONSES OF LID PRACTICES

YANG YANG
The University of Hong Kong, Hong Kong SAR, China, yyang90@connect.hku.hk

TING FONG MAY CHUI
The University of Hong Kong, Hong Kong SAR, China, maychui@hku.hk

ABSTRACT

Low impact development (LID) practices, such as green roofs and bioretention cells, are regarded as environmentally-friendly alternatives to the conventional drainage infrastructures. It is essential to accurately predict the hydrological responses of LID practices for assessing and optimizing LID designs. However, the accuracy of the commonly used process-based hydrological models is sometimes affected by model assumptions and the availability of field measurements for calibration. Machine learning methods can potentially avoid these issues by directly modeling the correlation between the input (e.g., rainfall time series) and the response (e.g., outflow hydrograph) of a system. However, considerable uncertainties are involved when training machine learning models. As a case study, the correlation between rainfall time series and outflow rates in an LID site in the U.S. is modeled using 11 commonly used machine learning models, including random forest, k-nearest neighbors, and gradient boosting machine. These models are trained on high temporal resolution data using formal machine learning procedures, which include feature engineering, pre-processing, model tuning, and resampling. Different methods are used in the training procedures for assessing the involved uncertainties. For example, in feature engineering, the original high-resolution time series is transformed into different sets of features (e.g., mean and peak rainfall intensity in the past two hours) which are used as input to machine learning models, and different types of transformations are used to pre-process these features. The results show that some machine learning models can achieve comparable or better prediction accuracy when compared to process-based models, and performance of different machine learning models can vary significantly. The feature engineering and the resampling procedures are found to have significant impacts on the quality of the trained models. Evaluating multiple machine learning models and using various methods in model training are crucial for assessing the uncertainties involved in machine learning.

Keywords: Low impact development, sustainable drainage system, machine learning, stormwater management, uncertainty analysis

1. INTRODUCTION

Low impact development (LID) practices, also known as sustainable urban drainage systems or green infrastructures, are nature-based solutions to urban stormwater drainage problems (Fletcher et al., 2015). The commonly used LID practices include green roofs, bioretention cells, porous pavements, and grass swales. LID practices are commonly constructed throughout urban catchments and their primary function is to promote on-site interception, storage, infiltration, evapotranspiration, and reuse of stormwater (Ahiablame et al., 2012). As they mimic natural drainage processes, they are often regarded as environmentally friendly alternatives to conventional drainage networks.

Various numerical models have been employed to predict the hydrological responses of LID practices under different rainfall conditions (Elliott and Trowsdale, 2007). Such procedure is critical to the evaluation and optimization of LID designs (Yang and Chui, 2018b). Commonly used numerical models are mostly process-based, where the hydrological processes involved are characterized using physically-based or empirical equations. However, due to the high complexity of the hydrological processes involved, the prediction accuracy of process-based models can be affected by model assumptions and the knowledge of the studied system (Niazi et al., 2017).

Machine learning models are able to learn the statistical correlations between the input and output of a system from observation data without knowing the underlying physical processes (Solomatine and Ostfeld, 2008). Thus, machine learning models may be applied to learn the correlation between input forcing and hydrological

responses of LID practices (e.g., the rainfall-runoff correlation), especially when the process-based models are insufficient. Multiple types of machine learning models, such as linear regressions, k-nearest neighbors, and support vector machines have been used in studies in many subfields of hydrology (Papacharalampous et al., 2019). However, the application of machine learning in LID related studies has been limited. Multiple linear regression models were built in Li (2015) to study the correlation between rainfall depth and runoff volume in a catchment that implemented LID practices. Yang and Chui (2018a) adopted several machine learning models to predict the occurrences of overflow and flow rates of two LID sites in the U.S, where considerable differences in prediction accuracy were observed among different machine learning models.

One of the key challenges facing machine learning in hydrological studies is the creation and selection of suitable input variables (Taormina and Chau, 2015). Since the hydrological responses of LID practices are governed by the antecedent meteorological conditions over an extended period of time, the high-resolution long-term time series of the past weather conditions should be used as input variables of these models, provided that LID practices' responses at finer time scales are interested. However, many machine models have limited abilities in modeling high-dimensional input. Therefore, lower dimensional variables derived from the original high-dimensional time series are often used instead as input variables. For instance, Yang and Chui (2018a) computed aggregated rainfall depths over different time periods in the past based on high-resolution rainfall time series and used them as input variables of multiple machine learning models. Lower dimensional variables may be derived in various ways. There are no restrictions on the input variable creation methods, and the usefulness of a specific method in different studies could vary significantly. However, in many existing studies, only one or a small number of these methods are used without explicit reasoning. Therefore, it is important to estimate the uncertainties associated with the input variable creation methods.

This study aims to investigate the following questions.

1. Are machine learning models with lower dimensional input variables derived from the high dimensional input variables better than those with the original input variables in terms of attaining higher prediction accuracy?
2. Do different types of machine learning models respond similarly when using the same methods for creating lower dimensional input variables?
3. To what extent can the input variable creation methods influence the prediction accuracy of machine learning models?

2. METHODS AND MATERIALS

2.1 Methods

Let Y_t be the random variable that represents the hydrological response of an LID site at time step t, and X_t be the time series of the antecedent meteorological and hydrological conditions recorded on and before time step t. Then, Y_t may be represented as a function of X_t (Eq. 1).

$$Y_t = f(X_t) \tag{1}$$

X_t has elements corresponding to multiple time steps, that at time step $t - i$ its element is a random vector I_{t-i}, which is the meteorological and hydrological measurements taken at time step $t - i$ (Eq. 2).

$$X_t = [I_{t-0} \quad I_{t-1} \quad I_{t-2} \quad ...] \tag{2}$$

The goal of machine learning is to find a function to approximate f based on samples of the input-output pairs, $\{y_t, x_t\}$.

The high dimensional variable, X_t, is sometimes transformed to lower dimensional inputs variables (or variables that are expected to improve the quality of trained models) through some function ϕ. Machine learning models are trained to approximate function g, which maps $\phi(X_t)$ to Y_t, as shown in Eq. 3.

$$Y_t = g(\phi(X_t)) \tag{3}$$

$\phi(X_t)$ are often called features, and the process of deriving these features are called feature engineering. ϕ are often defined using domain knowledge or through experiments (Kuhn and Johnson, 2019). The feature engineering process is highly flexible and can be subjective, which may result in large uncertainties.

Summary statistics computed for X_t or its elements are often used as lower dimensional input variables in current studies, such as Li (2015) and Yang and Chui (2018a). In this study, D_k, the sum of the elements in different sets of I_{t-i}, are used as input variables. D_k can be computed using Eq. 4.

$$D_k = \sum_{i \in S_k} I_{t-i} \tag{4}$$

where, S_k is a set of non-negative integers.

S_k are defined through a stochastic process described as follow. Suppose that the hydrological response of an LID site is affected by the elements in set $\{I_{t-k} | 0 \leq k \leq m\}$. Then, L integers, n_1 through n_L, are randomly sampled from the integers between 1 and $m-1$. The L integers then cut the range between 0 and m into $L+1$ intervals, such as $[0, n_1)$, $[n_1, n_2)$, and $[n_L, m]$. The integers fall within the range of an interval form a set, S_k. D_k can then be computed for each subset of I_{t-i} indexed by S_k.

It is reasonable to assume that the hydrological responses of LID practices are affected more by recent meteorological and hydrological conditions than by that from the distant past. To allow more D_k to be created for recent events, the distribution of the L integers should be right-skewed, i.e., there are more smaller integers than larger integers. In this study, p_i, the probability weight that integer i is selected is calculated as Eq. 5.

$$p_i = \frac{1}{i^q} \tag{5}$$

where, $q > 1$.

After the D_k features have been created, machine learning models are trained to approximate the function that maps D_k to Y_k. 11 commonly used machine learning models are considered in this study. They are: (1) linear regression (LM), (2) bagged multivariate adaptive regression splines (bagMARS), (3) partial least squares regression (PLS), (4) ridge regression (Ridge), (5) lasso regression (Lasso), (6) classification and regression trees (CART), (7) k-nearest neighbors (KNN), (8) cubist regression model (Cubist), (9) support vector machine with polynomial kernel (SVM), (10) random forests (RF), and (11) extreme gradient boosting (XGBoost).

The model training and hyperparameter optimization are performed though k-fold cross-validation. The caret package in the R programming language is used for these tasks (Kuhn, 2008). The considered hyperparameters are the recommended or default values of caret or the packages for the aforementioned models in R. More details on the model training and evaluation procedures can be found in Kuhn and Johnson (2013) and Hastie et al. (2009).

2.2 Materials and experiments

The study site locates in a commercial lot in Washington Street, Geauga County, Ohio, the U.S. (Darner et al., 2015). Multiple types of LID practices, including green roofs, bioretention basins, and porous pavements were constructed to treated stormwater runoff from the roof of a commercial building and a parking lot. Darner et al. (2015) evaluated the hydrological performance of the LID practices at this site using a statistical approach. Process-based models were not built in their study as the detailed design and construction configurations were not available. The rainfall and runoff conditions of this site were monitored by the U.S. Geological Survey since 2009. Rainfall depth was monitored on a 10-minute interval. Stormwater runoff from this site was collected by three flumes, where water levels were recorded at least every 10 minutes. The water levels were then converted to flow rates. The rainfall and runoff data collected between April 2011 and October 2011 were used in this study.

Y_t in this study is the sum of the instantaneous flow rates measured at the three flumes on the regular 10-minute intervals, and X_t is the rainfall time series of the past. Group-wise cross-validation was used, that rainfall-runoff data collected for the same event are grouped together, such that data belonging to the same group were used only for training or validation at each cross-validation iteration. Storms events of different intensities were distributed relatively evenly within each training, validation, or test set by adopting the stratified sampling method. 44 independent rainfall events with runoffs were identified using a 24-hour dry period threshold (Joo et al., 2014). 11 rainfall events were randomly selected and used as test dataset for evaluating the performance of fitted machine learning models. The other 33 events were used to train machine learning models, where five-fold cross-validation was used to optimize the hyperparameters of the models. D_k were pre-processed before feeding to the models. The default pre-processing method applied in this study is centering, scaling and also the Yeo-Johnson transformation (Bishara and Hittner, 2012).

A few experiments were conducted in this study. In the first experiment, X_t was the rainfall depth time series measured between time steps $t-144$ and $t-0$, i.e., the past one day. L, which controlled the number of input variables to be created, varied from 5 to 144. Since input variables were created in a stochastic manner, that for each L five sets of input variables were created randomly and used in training machine learning models to evaluate the associated uncertainties. q was set to 1.05 (Eq. 5). q controls the distribution of input

variables along the time axis, which may also affect the quality of trained models. The uncertainties related to q were not investigated in this study. Cubist and linear regression models were trained on these randomly created input variables. In the second experiment, X_t was the rainfall time series recorded in the past two days. L was set to 20, and q was set to 1.1. The 11 machine learning models listed above were trained and their performances were compared. The feature creation process was also repeated five times, which was same as the previous experiment. In the third experiment, cubist models were trained again but this time with the principal component analysis (PCA) added to the processing procedure. The importance of pre-processing procedures was then evaluated by comparing the results to that of the corresponding cubist models obtained in experiment 2.

3. RESULTS AND DISCUSSIONS

3.1 Influences of feature engineering methods on model performance metrics

The results of experiment 1 are shown in Figure 1. The cubist regression models, in general, had higher prediction accuracies than the linear regression models, as indicated by the higher R^2 values and lower root mean square error. In this experiment, the data used for each cross-validation iteration and the test set were fixed. Five sets of input variables were created for each L using the stochastic method described in the method section. Thus, the performance variations of each set of models with the same number of input variables were only caused by the stochastic feature creation process, which are shown by the vertical distances within each bar of dots in Figure 1. The results suggest that cubist regression models are more sensitive to modifications in input variables than the linear regression models. Due to the considerable sensitivities of the models to the stochastic feature creation processes, it is hard to determine the optimal number of input variables. However, the models performed differently when the number of input variables changed. For instance, linear regression models with fewer numbers of input variables are more sensitive to the stochastic feature creation process than the linear models with more input variables. Nevertheless, all these models could be useful for this site, where process-based models are unlikely to be set up due to limited field measurements.

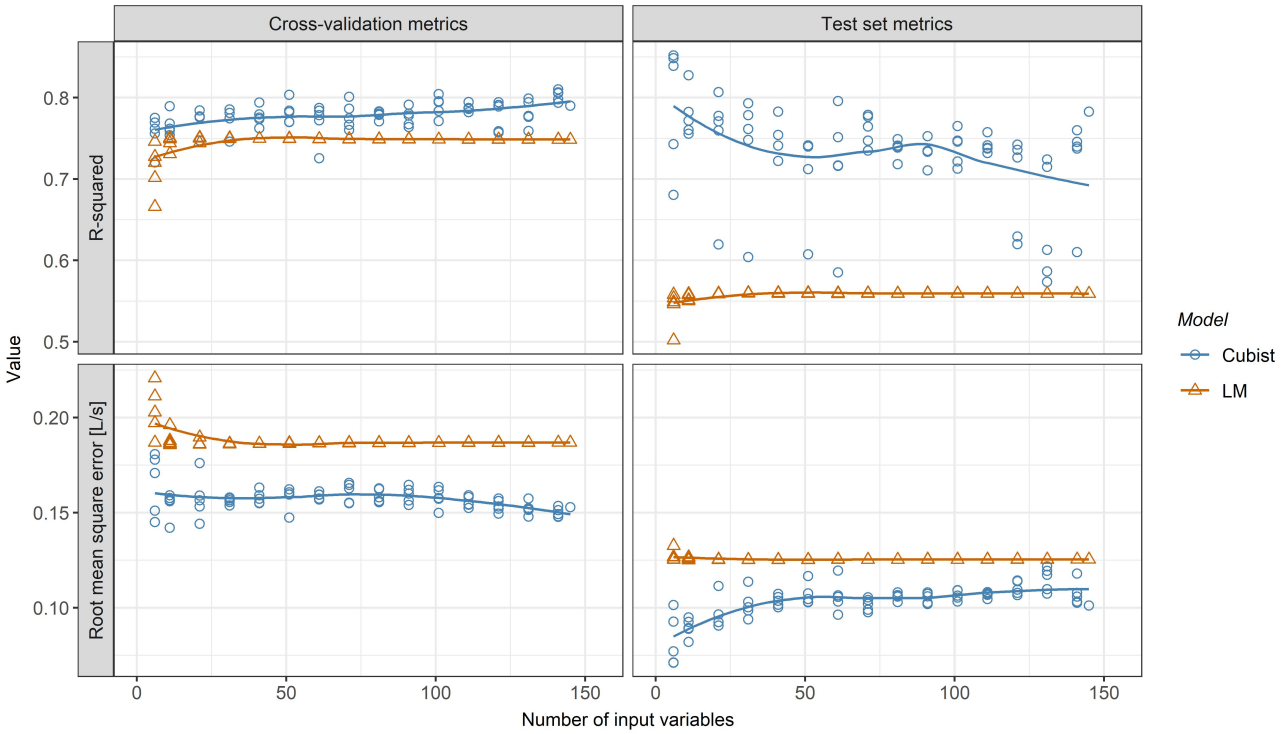

Figure 1. The performance of cubist regression models (Cubist) and linear regression models (LM) with different numbers of input variables evaluated at the cross-validation iterations and on the test set. Each dot corresponds to a model trained on a randomly generated set of input variables.

The performance metrics obtained at the cross-validation iterations and for the test set are expected to be similar. However, considerable divergences can be observed between the two sets of metrics, suggesting that more data were desired for more reliably assessing models' generalization error. This result also indicates that the uncertainties involved the model performance evaluation were large. Nested cross-validation may be performed to investigate the uncertainties involved in the splitting of the training and testing sets, which are associated with the small amount of data. Such procedures were not adopted in this study because experiment 1 was designed mainly to reveal the uncertainties related to the feature engineering process.

3.2 Performance of different types of machine learning models

In experiment 2, the prediction accuracy of 11 types of machine learning models was evaluated, and the results are shown in Figure 2. Each type of model was evaluated five times on different sets of input variables generated using the stochastic method described above. Considerable variations in prediction accuracies can be observed when the input variables varied, e.g., the cubist models (Cubist) and extreme gradient boosting models (XGBoost). Figure 2 shows the overall rank of the models in terms of prediction accuracy with the better models being placed on the top of the figure. The models were ranked based on their mean prediction accuracies of the two metrics on the test set. Despite the noticeable variations in different realizations of the feature creation process, some models are found to be superior as compared to the others. For instance, the prediction accuracies of the classification and regression trees (CART) were significantly lower than the random forest models (RF) and support vector machines with polynomial kernel (SVM). However, the model performance may vary when different L values are used, as shown in Figure 1. The hyperparameters of each model were selected from a fixed set, enriching the coverage of these sets may also affect the prediction accuracy of the machine learning models. Nevertheless, the performance metrics show that some of these models have good prediction accuracy, as the $R^2 > 0.70$ and $RMSE < 0.20$ in both cross-validation and the test case. As previously discussed, the evaluation results were also affected by the uncertainties involved in sampling cross-validation folds and the test set. In conclusion, some machine learning models were found to be more useful than the other models, suggesting that more types of machine learning models should be evaluated during application. The multiple sources of uncertainties should be considered when trying to identify the optimal model(s).

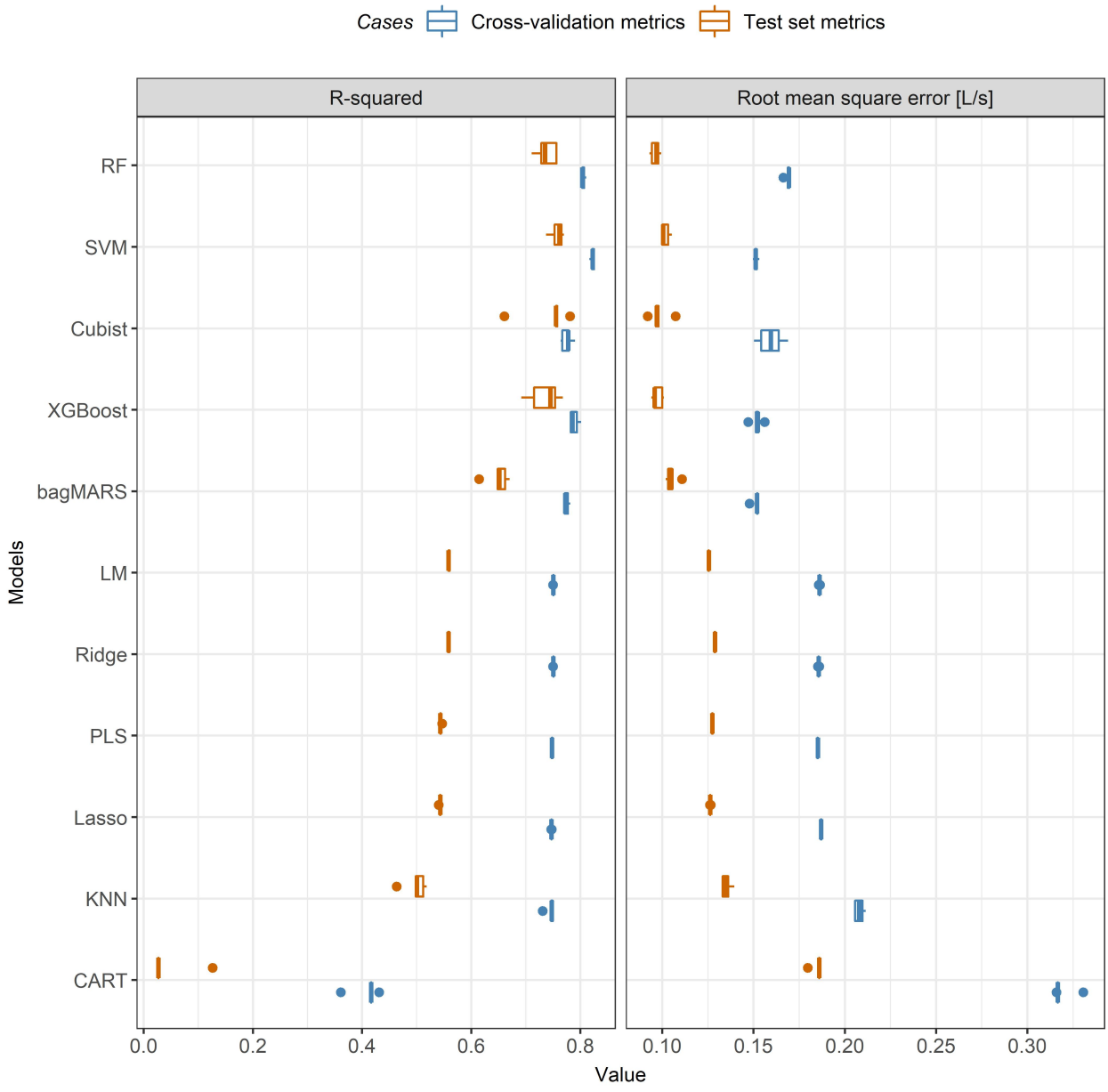

Figure 2. The performance of multiple types of machine learning models evaluated at the cross-validation iterations and on the test set. Each bar of the boxplot corresponds to models trained on multiple randomly generated sets of input variables.

3.3 Influences of preprocessing procedures

The result of experiment 3 showed another factor that potentially affects prediction accuracies, that is the pre-processing method. The cubist models were trained again using the same model set up procedures as experiment 2, except adding the PCA to the pre-processing procedure to further reduce the number of input variables. The resulted models (Cubist-w/-PCA) performed noticeably better than the original cubist models (Cubist) in experiment 2, as shown in Figure 3. This result shows that pre-processing techniques can be an important source of uncertainties that affects the prediction accuracy of machine learning models. Thus, this study recommends multiple pre-processing procedures to be evaluated for deriving higher quality models.

Figure 3. The performance of cubist models evaluated at the cross-validation iterations and on the test set. Each bar of the boxplot corresponds to models trained with different pre-processing procedures. Cubist-w/-PCA corresponds to cubist models trained with the PCA added to the pre-processing procedure, and Cubist corresponds to the cubist models obtained in experiment 2.

4. CONCLUSIONS

This research shows that machine learning methods can be valuable tools to predict the hydrological responses of LID practices. The models can be applied in cases when the information for setting up process-based models is not available. In LID-related studies, the dimension of input variables can be high. Feature engineering methods can effectively reduce the dimension of the input variables and can potentially improve the prediction accuracy of machine learning models.

The feature engineering processes are found to have significant influences on the prediction accuracy of the trained models. This research shows that different machine models can respond differently when the input variable change. However, it would be difficult to predict the consequence of using a specific feature engineering method without training and evaluating the models. This research also shows that pre-processing methods, such as applying the PCA to the input variables, can be useful for improving the quality of the models. As it is difficult to predict the effectiveness of different types of machine learning models, feature engineering process, and pre-processing methods in attaining higher prediction accuracy, future studies should investigate as many combinations of these modeling configurations as possible in order to identify higher quality models. Other types of machine learning models that relied less on feature engineering, such as deep learning models, may be applied in future LID-related studies.

This research shows that while it is relatively easy to set up machine learning models that deliver reasonably good prediction accuracy, it can be challenging to establish the optimal model (or models) due to the large uncertainties related to feature engineering, pre-processing, and hyperparameter optimization. Therefore, this research calls for further investigations on methods for automatically optimizing the structure of machine learning models, especially under the condition of having a limited amount of data.

ACKNOWLEDGMENTS

This study was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU17255516). We would like to thank Robert A. Darner from the U.S. Geological Survey for providing information of the study site.

REFERENCES

Ahiablame, L. M., Engel, B. A., and Chaubey, I. (2012). Effectiveness of low impact development practices: literature review and suggestions for future research. *Water, Air, & Soil Pollution,* 223**,** 4253-4273.

Bishara, A. J. and Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological methods,* 17**,** 399.

Darner, R. A., Shuster, W. D., and Dumouchelle, D. H. (2015). *Hydrologic Characteristics of Low-impact Stormwater Control Measures at Two Sites in Northeastern Ohio, 2008-13*, US Department of the Interior, US Geologic Survey Reston.

Elliott, A. and Trowsdale, S. A. (2007). A review of models for low impact urban stormwater drainage. *Environmental modelling & software,* 22**,** 394-405.

Fletcher, T. D., Shuster, W., Hunt, W. F., Ashley, R., Butler, D., Arthur, S., Trowsdale, S., Barraud, S., Semadeni-Davies, A., and Bertrand-Krajewski, J.-L. (2015). SUDS, LID, BMPs, WSUD and more–The evolution and application of terminology surrounding urban drainage. *Urban Water Journal,* 12**,** 525-542.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Joo, J., Lee, J., Kim, J., Jun, H., and Jo, D. (2014). Inter-event time definition setting procedure for urban drainage systems. *Water,* 6**,** 45-58.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software,* 28**,** 1-26.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, Springer.

Kuhn, M. and Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*, Chapman and Hall/CRC.

Li, S. (2015). Data driven comprehensive assessment of the performance of stormwater best management practices. *PhD Dissertation*, University of Louisville.

Niazi, M., Nietch, C., Maghrebi, M., Jackson, N., Bennett, B. R., Tryby, M., and Massoudieh, A. (2017). Storm water management model: Performance review and gap analysis. *Journal of Sustainable Water in the Built Environment,* 3**,** 04017002.

Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment,* 33**,** 481-514.

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics,* 10**,** 3-22.

Taormina, R. and Chau, K.-W. (2015). Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and extreme learning machines. *Journal of Hydrology,* 529**,** 1617-1632.

Yang, Y. and Chui, T. F. M. (2018a). Hydrologic Performance Simulation of Green Infrastructures: Why Data-Driven Modelling Can Be Useful? International Conference on Urban Drainage Modelling. Springer, 480-484.

Yang, Y. and Chui, T. F. M. (2018b). Integrated hydro-environmental impact assessment and alternative selection of low impact development practices in small urban catchments. *Journal of environmental management,* 223**,** 324-337.