Water Research 166 (2019) 115075

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Taxonomy of influential factors for predicting pollutant first flush in urban stormwater runoff

Thamali Perera [a, b], James McGree [a], Prasanna Egodawatta [a], K.B.S.N. Jinadasa [c], Ashantha Goonetilleke [a, *]

[a] Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane, 4001, Queensland, Australia
[b] Department of Mathematics, University of SriJayewardenepura, Nugegoda, 10250, Sri Lanka
[c] Department of Civil Engineering, University of Peradeniya, Peradeniya, 20400, Sri Lanka

ARTICLE INFO

Article history:
Received 3 June 2019
Received in revised form
3 September 2019
Accepted 9 September 2019
Available online 9 September 2019

Keywords:
Pollutant first flush
Stormwater treatment design
Classification and regression tree
Random forest
Stormwater quality
Stormwater pollutant processes

ABSTRACT

Pollutant first flush in urban stormwater runoff is an important phenomenon influenced by a range of rainfall and catchment related variables. Even though numerous studies have been undertaken to mathematically define the first flush and the influential variables of first flush, limited research have been carried out to rank such variables in terms of their level of importance in generating first flush. Identifying the degree of importance of the variables is critical for accurate predictions of first flush occurrence and understanding the main drivers of first flush. This research study undertook a comprehensive analysis of the variables influencing the predictions of first flush occurrence and their relative importance. The study results are expected to contribute to more accurate predictions of first flush by affording greater importance to the highly ranked factors and their impacts. The study outcomes confirmed that total rainfall depth was the most important variable influencing the prediction of first flush events while the maximum intensity was the second. Rain duration, runoff depth, runoff peak and average intensity were the next four most important variables. Antecedent dry period and effective impervious area fraction had relatively low ranking while the time of concentration and the event mean concentration were found to be the least important variables. Furthermore, the study outcomes highlight that the use of a combination of variables and due consideration of their interactions can yield better results than considering their individual roles.

1. Introduction

Urbanisation results in an increase in anthropogenic activities, leading to increased pollutant generation. Impervious surfaces common to urban areas provide the primary platform for the deposition and accumulation of pollutants which are subsequently washed-off by surface runoff. Wash-off of pollutants deposited on urban surfaces can create detrimental impacts on urban receiving water quality (Goonetilleke and Thomas, 2003; Jacobson, 2011; Liu et al., 2016; Walsh, 2000).

Pollutant "first flush", which is commonly defined as the disproportionate discharge of either higher pollutant concentrations or load in the initial part of a runoff event relative to its latter part, is of major interest in stormwater treatment design. If the first flush can be defined accurately, then the treatment systems can be designed to store and treat the most polluted part of runoff volume (first flush volume), while bypassing the remainder directly with minimal or no treatment (Bach et al., 2010; Goonetilleke et al., 2005; Gupta and Saul, 1996; Kang et al., 2008; Lee et al., 2002; Sansalone and Cristina, 2004). This approach would help to reduce the space required and the construction and maintenance costs of treatment systems. Extensive studies have been undertaken over the years to define the occurrence of first flush and the influential variables in generating first flush (Bach et al., 2010; Geiger, 1987; Gupta and Saul, 1996; Sansalone and Cristina, 2004). Research has shown that pollutant first flush can be influenced by a range of variables which can be categorised as pollutant, site and rainfall related factors. However, there is no universal set of variables clearly defined which can be reliably used in predicting first flush occurrence. Lee et al. (2002) concluded that the type of pollutant, catchment area, contributing impervious area and rainfall intensity are the key variables that influence first flush occurrence. Most researchers have highlighted the importance of maximum rainfall

* Corresponding author. Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane, 4001, Australia.
E-mail address: a.goonetilleke@qut.edua.au (A. Goonetilleke).

https://doi.org/10.1016/j.watres.2019.115075
0043-1354/

intensity, rainfall duration and antecedent dry period for first flush predictions (Gnecco et al., 2005; Gupta and Saul, 1996; Kang et al., 2006; Li et al., 2007).

Furthermore, though previous studies have evaluated the variables influencing first flush occurrence, the relative degree of importance of these variables or ranking have not been discussed. Most researchers have limited their analyses to one or two variables giving similar importance to each and discussed only the individual roles of such variables in defining first flush occurrence (Kang et al., 2008; Schiff et al., 2016). Generally, it is exceptionally difficult to consider all the relevant variables because of the highly varying nature of rainfall and site-specific characteristics (Kayhanian and Stenstrom, 2005). The limitations in data analysis tools commonly used in water quality research is also a disadvantage in understanding the degree of influence exerted by the diversity of variables in relation to first flush.

Furthermore, for accurate stormwater quality modelling and design of treatment systems, it is important to identify the critical variables and take into account the significance of these variables in first flush occurrence. Accordingly, the objectives of this study were, to identify the critical variables influencing the pollutant first flush behaviour and to rank these variables in terms of their degree of influence (or the level of importance) in predicting first flush occurrence. The study outcomes are expected to contribute to the development of robust models for stormwater quality prediction and thereby effective stormwater treatment system design.

2. Materials and methods

2.1. Study sites

Two urban residential catchments located in Coomera and Nerang in the Gold Coast region, together with the domestic and international aprons of the Brisbane airport and the Direct Factory Outlet (DFO) car park located within the Brisbane airport land were selected as the study sites. All sites are located in Queensland State, Australia (see Fig. 1).

The catchments in Coomera and Nerang were subdivided into three subcatchments each and encompass a mix of land use and land cover. Therefore, the overall mixture of study sites selected permitted the in-depth investigation of the influence of catchment characteristics on stormwater quality (see Fig. 1 for site characteristics). Further, the extensive monitoring data available for the study sites, as discussed in Section 2.2 enabled the in-depth investigation of the influence of rainfall characteristics on stormwater quality.

2.2. Data collection

The research study used rainfall data recorded from established pluviograph stations located in the vicinity. For the selected rainfall events, stormwater samples were captured using automatic monitoring stations. The monitoring stations consisted of an automatic sampler to collect samples during the rising and falling limb of flow events so that pollutant loads could be calculated, a probe for the continuous measurement of basic water quality parameters such as pH and electrical conductivity and v-notch weir and pressure transducer for flow measurement to generate hydrographs and a data logger. A detailed description of data collection process can be found elsewhere (Alias, 2013; Liu, 2011; Mangangka, 2013).

Monitored storm events with complete records of rainfall, runoff and water quality were selected for the study. Water quality records were considered complete when at least one water quality measurement at the initial part, middle part and the later part of the runoff event were available. Accordingly, a total of 63 rainfall-runoff events were selected. For the analysis, hydrologic variables widely cited in research literature in relation to first flush analysis, namely, total rainfall depth (RD), average rainfall intensity (AvgI), maximum 5 min intensity (MaxI), rain duration (D), antecedent dry period (ADP), runoff depth (RoD), runoff volume (RoV), runoff peak (RoP) and event mean concentration (EMC) were chosen (Geiger, 1987; Gupta and Saul, 1996; Lee et al., 2002; Sansalone and Cristina, 2004).

Numerous studies have discussed the importance of catchment characteristics such as the total area of the catchment and impervious fraction in influencing first flush occurrence (Alias et al., 2014; Bertrand-Krajewski et al., 1998; Lee et al., 2002). Effective impervious area (EIA) fraction, which is the fraction of hydraulically connected impervious area where water travels over an entirely impervious pathway to a stormwater drainage system inlet has proven its importance in influencing stormwater quality (Boyd et al., 1993; Ebrahimian, 2015). Therefore, data on total impervious fraction (Imp), together with EIA and the fractions of individual impervious surface types such as streets, roofs and driveways were determined for the analysis undertaken. Time of concentration (TC) is another important catchment characteristic influencing first flush behaviour (Kang et al., 2008). Accordingly, TC was calculated using the available catchment data and Kirpich formula and the Friend's equation were used for the calculations where appropriate (Queensland Urban Drainage Manual, 2017; Thompson, 2006).

Suspended solids is the most common pollutant found in urban stormwater runoff and it has been recognised as a primary indicator of stormwater quality (Williamson and Crawford, 2011). Accordingly, suspended solids concentration for the selected storm events were obtained as the key stormwater quality parameter. The range of hydrologic parameters corresponding to each catchment is provided in the Supplementary Information, Table S1.

2.3. Study approach

The analytical approach in this study consisted of three main stages. The data set was subjected to initial analysis to reduce the dimensionality. The analysis focused on removing statistically not significant and correlating variables which were found to influence first flush predictions. The second stage was to identify the storm events which have led to the occurrence of first flush. Classification of events as first flush and non first flush events in the Stage II was used in Classification and Regression Tree (CART) algorithm to build decision trees in Stage III.

The third stage was to comprehensively analyse the key variables which were found to be important in generating first flush and to rank them in terms of their level of importance. The data was analysed using several sophisticated data analysis tools and techniques as appropriate, including principal component analysis (PCA) and the machine learning algorithms such as classification and regression trees (CART) and random forest (Breiman et al., 1984a; Breiman et al., 1984b; Jolliffe and Cadima, 2016).

2.3.1. Stage I: Identifying the possible correlations between variables

Including the rainfall and catchment variables, the initial data set comprised of twelve variables. Therefore, a technique to reduce the dimensionality of the dataset was necessary to increase interpretability while avoiding data redundancy. In this context PCA was undertaken, which is commonly used to identify the correlations between a large set of variables in a data set and consequently used as a variable reduction method with potentially minimum loss of information (Jolliffe and Cadima, 2016).

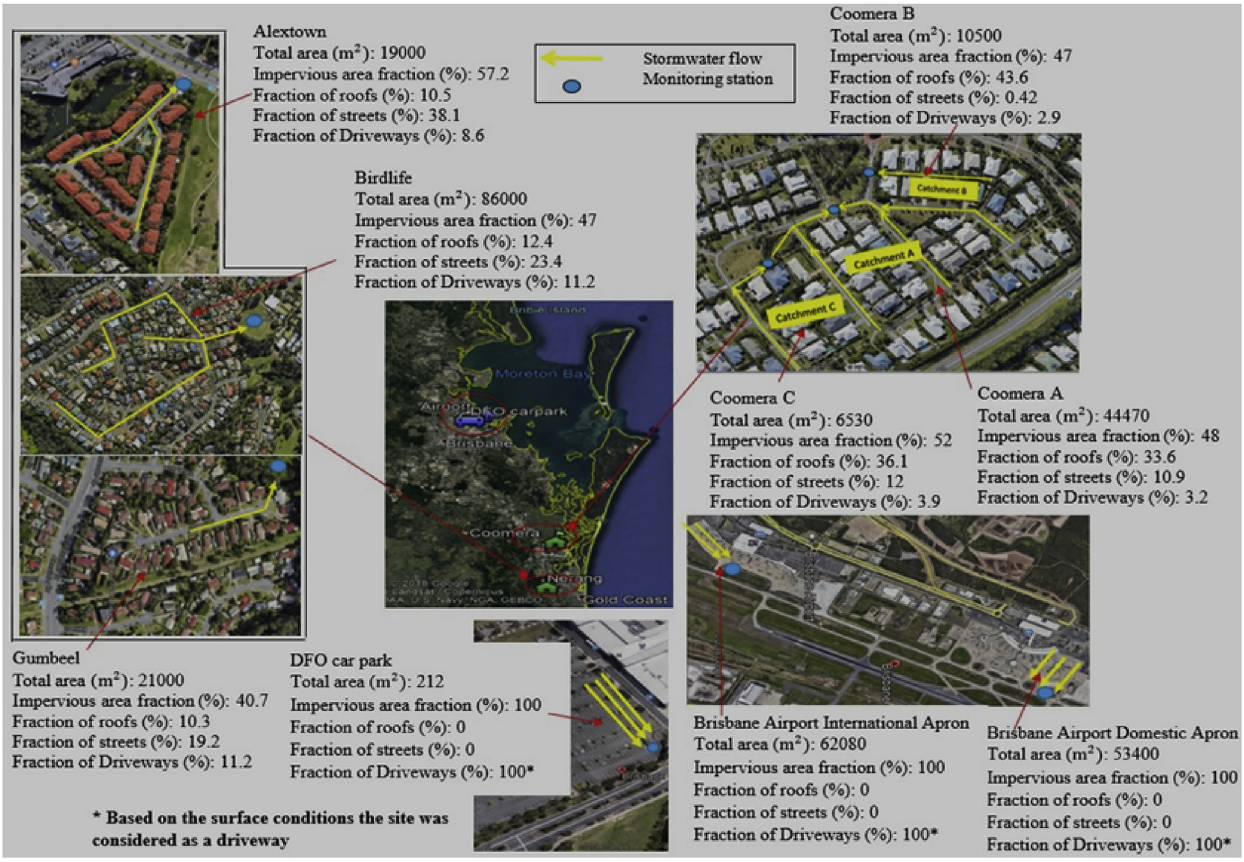

Fig. 1. Selected study sites and the stormwater flow directions with the monitoring stations (Google earth-pro, 2018).

2.3.2. Stage II: Identification of first flush events

The first flush behaviour was initially analysed using the dimensionless cumulative pollutant load ($M(t)$) vs dimensionless cumulative runoff volume ($V(t)$) curve, which is generally known as the $M(V)$ curve. By assuming that the flow rate and the pollutant concentration vary linearly between two successive measurements, $M(t)$ and $V(t)$ can be defined as given in Equations (1) and (2), respectively (Bertrand-Krajewski et al., 1998; Lee et al., 2002; Massoudieh et al., 2008; Sansalone and Cristina, 2004).

$$M(t) = \sum_{i=1}^{n} C_i Q_i \Delta t_i \Big/ M \quad (1)$$

$$V(t) = \sum_{i=1}^{n} Q_i \Delta t_i \Big/ V \quad (2)$$

where, Q_i and C_i are the respective flow rate and concentration at time t_i corresponding to the i^{th} measurement of an event having n number of measurements and M and V are the total pollutant load and the total runoff volume discharged.

First flush can be observed when the $M(V)$ curve lies completely above the 45^0 line (see Fig. 2) indicating a disproportionately high discharge of pollutants in the initial stage of the runoff event (Bertrand-Krajewski et al., 1998; Geiger, 1987). Furthermore, $M(V)$ curves can be fitted using the power relationship as given in Equation (3), where $b(>0)$ is the first flush exponent that directly describes the shape of the curve and the first flush is said to occur if $b<1$ (Bertrand-Krajewski et al., 1998).

$$M(t) = V(t)^b \quad (3)$$

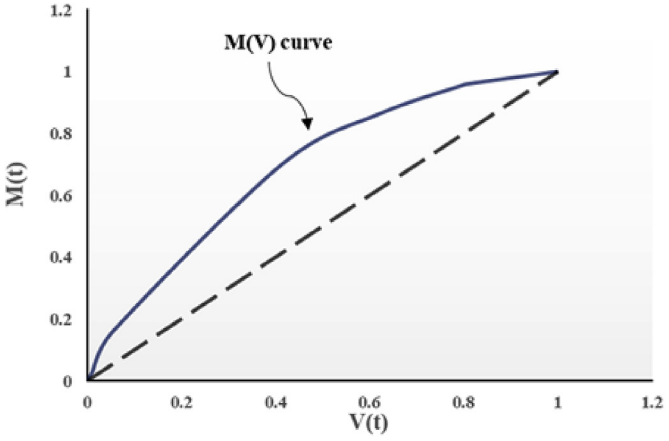

Fig. 2. Hypothetical representation of a $M(V)$ curve of a first flush event.

Exponent b can be calculated using the linear model given in Equation (4) that gives the logarithm of Equation (3). Depending on the value of b, the strength of the first flush can be explained where, the low values of b indicate relatively stronger first flush compared to values closer to 1 (Bertrand-Krajewski et al., 1998; Minervini, 2012; Sharifi et al., 2011).

$$\ln M = b \ln V(t) \quad (4)$$

2.3.3. Stage III: Regression tree analysis

For this stage, Classification and Regression Tree (CART) algorithm was employed with the objective of identifying the important variables which influence the occurrence of first flush. Regression trees (or classification trees) are an alternate method to the traditional statistical techniques. Regression trees perform well in predicting nonlinear relationships with high-order interactions between covariates. Furthermore, it is not necessary to ensure statistical assumptions of the data are met, and the results are easily interpretable (De'ath and Fabricius, 2000). Regression trees recursively partition the data space in such a way that the resulting groups are homogeneous as much as possible and then, prediction models are fitted at each partition. A binary splitting criterion is used when partitioning, which minimises the residual sum of the squared error of the predictions at each node (Breiman et al., 1984; De'ath and Fabricius, 2000).

When a tree is 'growing', a subset of variables is extracted randomly from the original set of variables and the variable that result in the highest information gain from the data is chosen to split the data at the node. The information gain is based on increasing the homogeneity of the branches after a dataset is split on an attribute (Rokach and Maimon, 2005). Because of this random selection of variables, a set of variables is kept out when the tree is growing and therefore, predictions of the model can be subject to error which is called the "out of bag error". However, construction of many trees and taking the average result has been proven to minimise the out of bag error, thereby maximising the accuracy. This leads to the use of random forests which reclusively builds a large number of classification or regression trees, i.e., a forest. Each tree in the forest is constructed by selecting a subset of variables and a sample of observations from the data. The predictions from the individual trees within the forest are combined by taking the average, and this is the prediction from the random forest. In undertaking such a procedure, it is possible to identify the relative importance of each variable by determining how often it appears in each tree and where it appears in the tree (variables that appear 'higher' in the tree are given more importance). This approach was adopted to determine the relative importance of variables which are regarded as influential factors in predicting the occurrence of first flush.

3. Results and discussion

3.1. Correlation between variables

PCA was performed using the R statistical package (version 3.5.1). Fig. 3 (a) gives the individual plot and Fig. 3(b) gives the variable plot resulting from PCA. As shown in Fig. 3(b), the first two PCs describe 70% of the total variance of the data. In the individual plot, a vector represents a variable and the length of a vector is proportional to the variance of the corresponding variable. The angle between two vectors indicates the degree of correlation between the two variables, where an acute angle represents a close relationship. However, for better interpretation, the results derived from the variable plot should be supported by a correlation matrix.

Accordingly, the correlation matrix developed which is provided in the Supplementary information, Table S2, was used for the interpretations. In this regard, any two variables were considered to be significantly correlated if the correlation coefficient was greater than 0.8 and the angle between the corresponding vectors was less than 30°.

It is evident that, Imp and EIA are strongly correlated with a correlation coefficient of 0.98 and the angle between the corresponding vectors being less than 30°. Moreover, the strong correlation of EIA with water quality variables compared to Imp has also been shown in past research (Boyd et al., 1993). Therefore, EIA was chosen to represent the variables, ImP and EIA. The relationship between RoD, RoV, RoP and MaxI can be interpreted similarly by assessing the angles between the corresponding vectors and their correlation coefficients. RoD and RoV show a significant correlation between each other with a correlation coefficient of 0.99. Even though, RoD and RoP show a correlation coefficient of 0.91, the corresponding angle is more than 30°. Therefore, RoP cannot be considered as a significantly correlating variable with RoD. Similarly, MaxI and RoP have a correlation coefficient of 0.86, but the corresponding angle does not satisfy the criteria defined above. Although, MaxI and RoD show similar behavioural patterns in the biplot, the correlation coefficient is 0.67 indicating that their relationship is not significant. Accordingly, RoV was the only variable which proves its significant correlation with RoD. Therefore, RoD was selected for subsequent analysis instead of RoV, since it is interpretable and compatible with RD.

As shown in Fig. 3(a), it can be observed that storm events are clustered (as circled) based on catchment land cover. All the events from residential sites fall into one cluster, scattered along the negative PC1 and positive PC2 axes. This is the same quadrant of the plot where relatively long ADP and TC vectors are pointed. The events from totally impervious sites (Airport aprons and the car park) formed another cluster located along the positive PC1 and negative PC2 axes, which is the same quadrant of the plot where long EIA and AvgI vectors are pointed. These outcomes highlight the catchment specific behaviour of runoff event characteristics and reinforces the need for an in-depth analysis of the relationships between the catchment characteristics and first flush for accurate urban stormwater quality modelling.

3.2. First flush event identification

The $M(V)$ curve was drawn for each runoff event using the hydrograph and pollutograph data. Subsequently, the relationship between $M(t)$ and $V(t)$ was modelled using the relationship given in Equation (3) and the exponent b was calculated using Equation (4) using MATLAB (R2017a).

An event was defined as a first flush event if the two criteria were satisfied (i.e. the $M(V)$ curve lies totally above the 45^0 line and $0 < b < 1$). Table 1 summarises the results of this stage of the analysis which identified the first flush events from the selected 63 rainfall events and the range of the first flush exponent for each catchment. It can be noted that, 65% of the total events experienced first flush behaviour (irrespective of land cover), and more than half of the events from each catchment exhibited first flush, except the international apron.

After the identification of first flush events, two attributes of the first flush could be defined. From this point onwards, '1' refers to the existence of the first flush and '0' refers to the non-existence of first flush unless specifically mentioned. Subsequently, probability density distribution of each variable was plotted breaking down by the two attributes. Probability density plots were drawn with the purpose of identifying the distinguishable range of values of variables which can detect first flush. Fig. 4 shows the distribution of each variable by the two attributes. The horizontal axis of each plot represents the values of the corresponding variable and the vertical axis is the probability of being '0' or '1'.

It is evident that attribute values overlap with all variables. Therefore, it is difficult to observe potentially useful predictors of first flush. For example, the density of AvgI indicates that, for an event having a value between 0 and 25 mm/h (shown circled in Fig. 4), there is a high probability of first flush occurrence. However, at the same time there is a certain possibility of a non-first flush event. Therefore, it is difficult to make exact predictions of the

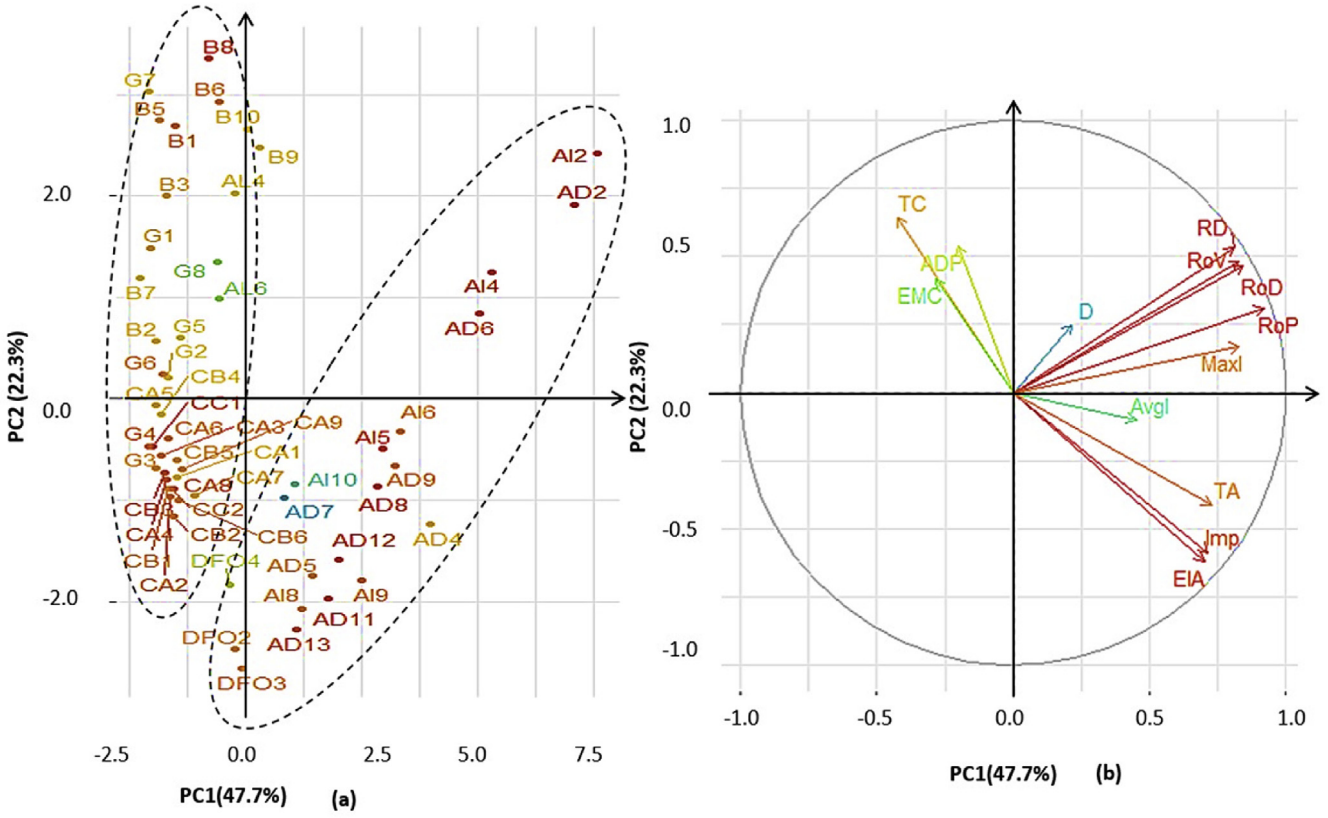

Fig. 3. (a) Individual plot given by PCA; (b) Variable plot given by PCA.
Note: Storm events were labelled based the study site: CA- Coomera A, CB- Coomera B, CC- Coomera C, AL- Alextown, B- Birdlife, G- Gumbeel, AD- Apron domestic, AI- Apron international. For example, CA1 means the first event of Coomera A catchment.

Table 1
Summary of the $M(V)$ curve analysis in identifying first flush events.

Catchment	Total events	First flush events	Percentage of first flush Occurrence (%)	Range of ***b***
Coomera A	9	8	88	0.66–1.18
Coomera B	6	6	100	0.77–0.97
Coomera C	2	2	100	0.77–0.91
Alextown	2	2	100	0.92–0.93
Birdlife	10	6	60	0.56–1.21
Gumbeel	8	5	62.5	0.45–1.28
Apron (Domestic)	13	7	53	0.26–1.35
Apron (International)	10	3	30	0.71–2.12
DFO car park	3	2	66	0.27–1.87
Total	63	41	65	N/A

existence of first flush or non-existence within a specific range of values of the variable. This behaviour highlights the fact, that an exact prediction of first flush occurrence is difficult by considering only a single variable. Accordingly, the analysis was extended to the regression tree approach for identifying a set of variables which are critical in predicting the occurrence of first flush.

3.3. Classification tree analysis

CART algorithm was adopted to identify the variables which are the most important in predicting a first flush event. Since the response variable can have one of two possible outcomes, 0 or 1, the model was a classification tree. The classification tree was built using the tools in R (version 3.5.1). When implementing CART with R, initially a fully grown tree is fitted with the entire data and eventually the tree is pruned to the smallest tree with lowest misclassification error. The data is then split into randomly selected subsets called folds (say k). Then k-fold cross-validation is performed and for each training fold, a sub tree is grown. Therefore, it is not necessary to use an additional validation set. Accordingly, the complexity parameter (cp) is selected from the set of sub trees grown which gives the smallest risk of misclassification.

In order to evaluate the performance, CART algorithm was repeatedly implemented by varying the input parameters "minsplit", "minbucket" and "cp" (minsplit is the minimum number of observations that must exist in a node for a split to be attempted and minbucket is the minimum number of observations in any terminal node). Accordingly, minsplit was set to 10, 15 and 20, minbucket was set to 3, 5 and 6 and cp was set to 0.01, 0.001, and 0.0001. Therefore, CART was implemented 27 times using the different permutations of the selected parameter values. The summary of the results are given in Table S4 in the supplementary

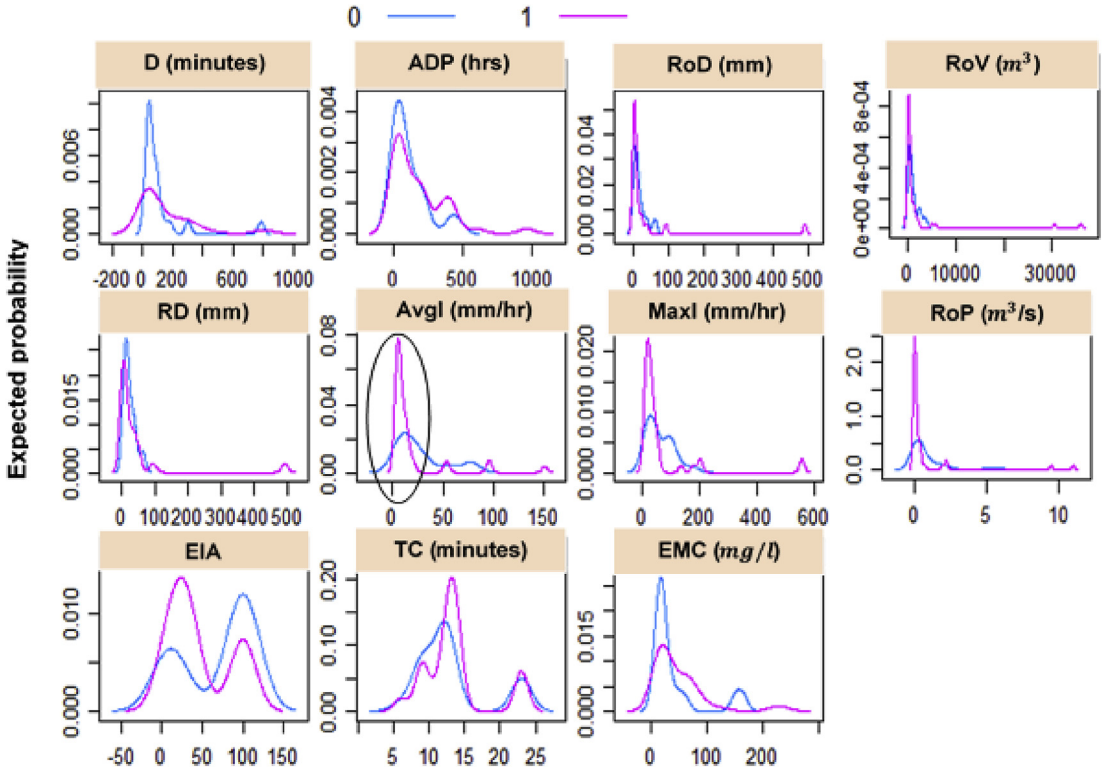

Fig. 4. Density distribution plot of variables of the two attributes.

information.

The best tree is grown with the selected cp which is given in Figure (5) which corresponds to the data set used in this study. The distribution of cross-validated error based on the number of splits is given in Fig. S1 and the summary of cross-validation is given in Table S5 in supplementary information.

The classification tree split the entire data space into five groups as labelled, which are at the bottom of the tree (leaf nodes of the tree). In each leaf node, prediction of first flush occurrence for the events in the group, either 0 or 1 is given with a probability of an event being a first flush event. The splitting was based on the variables AvgI, RD and RoD. Therefore, these variables gain relatively more information from the data. Accordingly, the role of AvgI, RD and RoD can be considered as relatively more important compared to the other variables in predicting first flush occurrence. As shown in Fig. 5, the Group 1 events (if the AvgI of an event is $>=16.56\ mm/h$ and RoD $<21.83\ mm$, the event falls into Group 1) have relatively low RoD and high AvgI with 18% probability of first flush occurrence, which is very low. Accordingly, the prediction of first flush for this group is 0. These events can have sufficient energy to detach the pollutants from the surfaces because of the high AvgI, but may not have sufficient runoff to wash-off the pollutants. Therefore, there is limited possibility of first flush occurrence. However, events in Group 2 with RoD $>21.83\ mm$ and AvgI $>=16.56\ mm/h$ have high probability (55%) of first flush occurrence relative to Group 1 and the model has predicted 1 as the state of first flush for this group. The behaviour of events in Group 1 and 2 suggests that an increase of only a single variable, either RoD or AvgI would not directly influence the accurate clustering of first flush events. However, increase in RoD and AvgI together would result in an increase in the probability of identifying the event as a first flush event. Therefore, the interaction of both variables needs to be considered when predicting the occurrence of first flush.

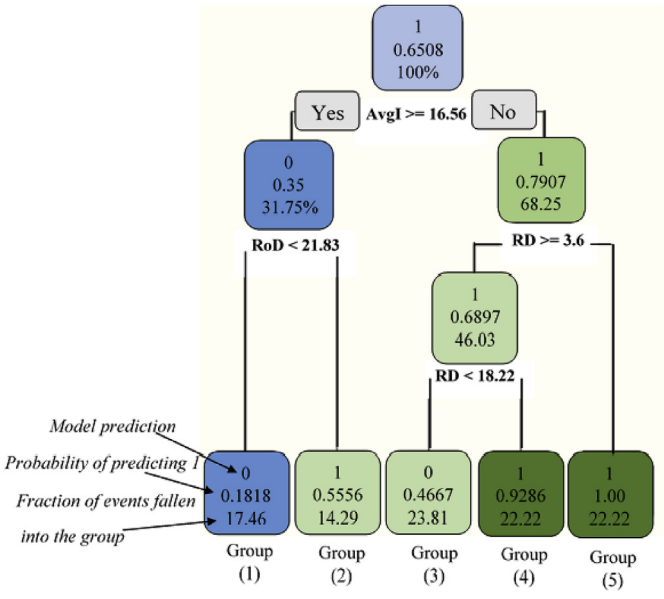

Fig. 5. Classification tree for predicting first flush occurrence.

In Group 3, events have RD $<18.2\ mm$ and the AvgI is $<16.56\ mm/h$ and have only a 46% possibility of first flush occurrence and the model prediction for this group is 0. However, the events in Group 4 have relatively high RD ($>18.2\ mm$) with low AvgI ($<16.56\ mm/h$) and the corresponding probability of an event being a first flush event is 92%. This behaviour suggests that having a high RD can significantly impact on first flush occurrence. Therefore, RD plays an important role in identifying the events as first flush events. However, the impact of other variables in relation to first flush should also be considered. For example, variables such

as ADP and MaxI are relatively high for the events in Group 4. Therefore, the initial pollutant loads available for wash-off would be high because of the longer ADP and the high RD with high MaxI resulting in rapid pollutant wash-off leading to significant first flush occurrence. Therefore, in predicting first flush occurrence, evaluating the impact of RD together with its interactions with other variables can lead to more accurate predictions.

However, it is important to note that even with low RD (<3.6 mm) and low intensity (AvgI < 16.56 mm/h), first flush can be observed in Group 5 events. The events in Group 5 belong to the three small residential catchments (Coomera A, B and C) having relatively very low values for all rainfall parameters (see Table S1 in the supplementary information). However, these storms have exhibited first flush with 100% probability. This is attributed to the impact of other factors such as site characteristics which dominate over the runoff characteristics. More specifically, these three catchments have a high percentage of roof surfaces relative to the other residential catchments and the catchments are wide and short (see Fig. 1). Therefore, the build-up of high loads of pollutants on the roof surfaces and their rapid wash-off with rainfall will result in rapid introduction of pollutant loads to the drainage network (Egodawatta et al., 2009). The wide nature of these catchments can further accelerate the pollutant flush out along the drainage network. Therefore, this can result in significant first flush even for a relative small rainfall event. Therefore, it can be concluded that the assessment of the implications of land cover together with other physical characteristics such as catchment shape and slope on pollutant wash-off process is vital for accurate stormwater quality predictions.

3.4. Random forest analysis

In the classification tree, variables actually used in tree construction are RD, AvgI and RoD. This means that these variables describe more variability in data and gain more information and therefore, critical in making predictions. However, it is generally the case that random forests provide more accurate predictions when compared to CARTs. Therefore, to improve the predictive ability of the classification tree and to minimise the out of bag error, the random forest algorithm was employed. Fig. 6 shows one example of the outcomes of the application of the random forest algorithm with the corresponding mean decrease accuracy (MDA) values. MDA is an index which indicate how much the accuracy of the predictions of a trained model would decrease by removing the corresponding variable from that model. Variables having higher MDA values are generally more important and gain more information from the data than the other variables. However, the observations made from building a single random forest are inadequate to draw robust conclusions about ranking the influential variables.

When training a random forest model, the algorithm tries to create uncorrelated trees extracting only a subset of variables and a sample of observations randomly. Therefore, the outcome can vary depending on the variable selection and also on the number of trees built in a forest. However, variable importance measures can become asymptotically stable with increasing numbers of trees (Grömping, 2009). Therefore, the random forest model was retrained by varying the number of trees constructed in a forest and the number of variables extracted in a single iteration. Accordingly, 501, 1001 and 10001 trees were constructed separately, while varying the number of variables to 2, 3 and 6 which were extracted at once. Fig. 7 gives the boxplots corresponding to the distribution of MDA values of each variable resulting from the random forest outcomes at each iteration.

It can be noted from Figs. 6 and 7, that RD is ranked as the most important variable according to the MDA index followed by MaxI. Thereafter, D, RoD, RoP and AvgI are the next four most important variables with close MDA values between each other (shown as the range marked by a dash lined box on the MDA axis in Fig. 6). ADP and EIA are ranked 7th and 8th in relation to their level of importance having relatively low MDAs. There is a significant decline in MDA values after EIA. Accordingly, EMC and TC are the two least important variables for model prediction (shown as the range marked by a dotted lined box on the MDA axis in Fig. 6). It can be argued that the set of variables, RD, MaxI, D, RoD, RoP and AvgI are critical for predicting the existence or non-existence of first flush.

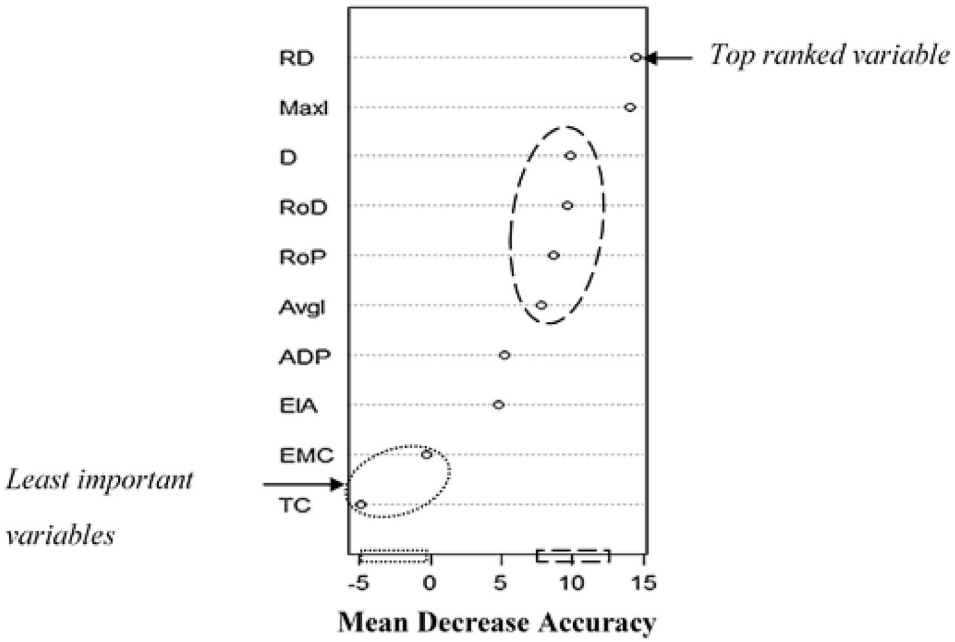

Fig. 6. Random forest used to identify the importance of parameters.

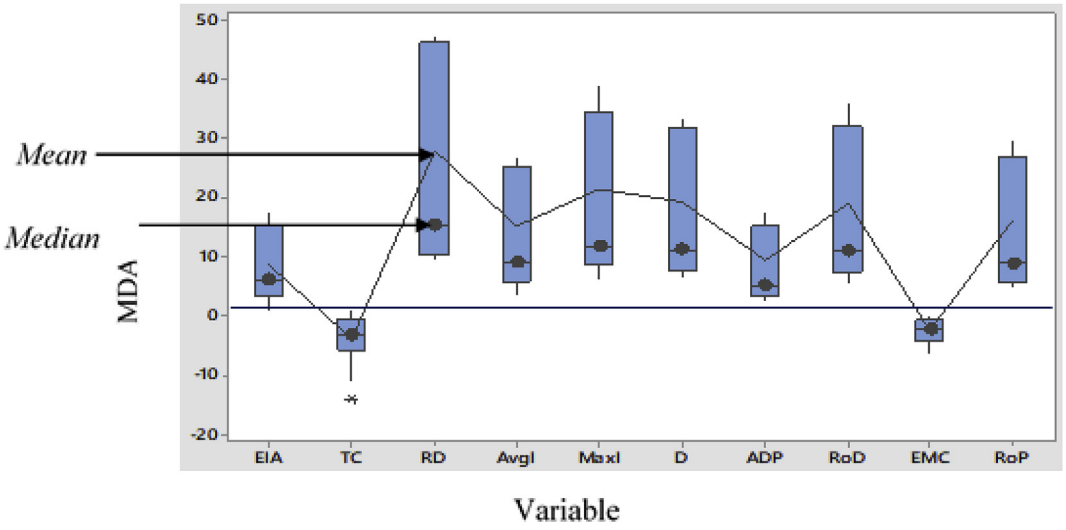

Fig. 7. Boxplots of the relative importance values for each variable.

This is also in agreement with the previous research outcomes (Alias et al., 2014; Gupta and Saul, 1996; Li et al., 2015; Schiff et al., 2016). Even though, ADP and EIA have low MDA values, the impact of these variables cannot be ignored since their influence on first flush have been confirmed in past research (Gupta and Saul, 1996; Lee et al., 2002; Sartor and Boyd, 1972). EMC and TC does not have a significant impact on the predictions of first flush due to relatively low MDAs.

It can be noted that, RD and MaxI have the highest MDAs, thus indicating their significance. Similar interpretations can be made regarding the other variables as discussed previously using Fig. 6 by comparing the mean and median in the boxplot with the output given in Fig. 6. Therefore, the model outcome of a random forest remains invariant irrespective of how the random forest was constructed (i.e. the variations in the input parameters which are the number of trees and the subset of variables used for the construction, do not change the outcome). Therefore, the variables investigated in this study can be ranked according to their level of importance in predicting the first flush, as follows:

$$RD > MaxI > D > RoD > RoP > AvgI > ADP > EIA > EMC > TC.$$

In order to validate the random forest outcomes, cross validation was subsequently performed. Accordingly, Leave One Out (LOO) cross validation approach was adopted which leaves out one observation, builds a random forest, then predicts the omitted observation. For our data set, this process can be repeated 63 times, yielding 63 predictions of 'unseen' observations. Table 2 gives the corresponding confusion matrix resulted from LOO cross validation.

With reference to Table 2, the overall accuracy of the random forest predictions is 71.4%. This level of accuracy is not unreasonable since current approaches for predicting first flush is based on graphical illustrations such as pollutographs and hydrographs and these approaches have only 50% accuracy in predictions (a prediction using a graph may be either true or false). Sensitivity and specificity are 85.3% and 45.45%, respectively. Equations for calculating accuracy, sensitivity and specificity are given in Equations (5)–(7), respectively. The specificity of the model is relatively lower which means that probability of predicting a non first flush event as a first flush event is relatively high. However, from a practical perspective, treating low concentrated runoff (i.e. non first flush event) and discharging would not cause a risk more than the risk associated with discharging strongly concentrated volume (i.e. first flush event) without treatment. However, an index to weight sensitivity and specificity relative to their importance can be used if necessary to improve the performance (Li et al., 2013).

Table 2
Confusion matrix of LOO cross validation.

		Observed	
		0	1
Predicted	0	10	6
	1	12	35

$$\text{Accuracy} = \frac{\text{Correctly predicted observations}}{\text{Total observations}}\% = \frac{10+35}{63}\% = 71.4\% \quad (5)$$

$$Sensitivity = \frac{Correctly\ predicted\ as\ true\ (1)}{Total\ true\ observations}\% = \frac{35}{(35+6)}\% = 85.3\% \quad (6)$$

$$Specificity = \frac{Correctly\ predicted\ as\ false\ (0)}{Total\ false\ observations}\% = \frac{10}{(10+12)}\% = 45.5\% \quad (7)$$

Subsequently, a receiver operating characteristic curve (ROC) was drawn (Fig. 8), which is a graphical illustration of the diagnostic ability of the random forest. ROC curve indicates that at the optimal point, the random forest sensitivity is 82.5% and specificity is 45.5%, which are close to the LOO cross validation results. Therefore, it can be concluded that the random forest performs well with the data used in the analysis with relatively high prediction accuracy.

3.5. Practical value of the study

This study was conducted to identify the variables, which are important for predicting the occurrence of first flush, and to rank them according to their level of importance. Pollutant load

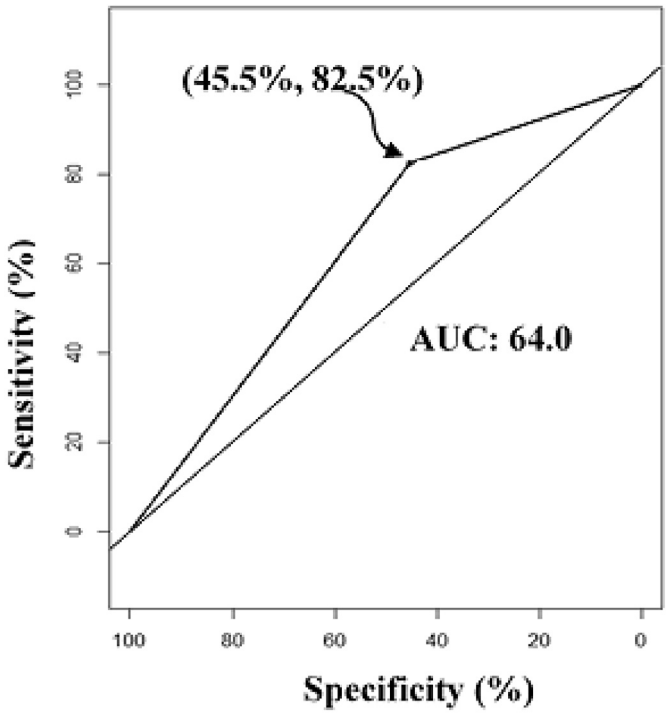

Fig. 8. ROC curve corresponding to the random forest predictions.

distribution was analysed for each event using the $M(V)$ curve for the identification of first flush events, but the quantitative definitions were not used as these were found to be arbitrary. For modelling and making predictions, machine learning algorithms such as CART and random forest were used which have proven their ability in predicting nonlinear relationships with high-order interactions between covariates. This type of analytical tools have rarely been used in the field of urban stormwater quality research.

Enhanced prediction performance of CART and random forest have also been discussed in past studies (Jeung et al., 2019). However, the analysis can also be performed using standard methods such as regression models, but subject to changes in performance. Accordingly, binomial logistic regression approach was used for the comparison of performance. Binomial regression was used because the first flush has only two attributes; 0 and 1. The outcomes are provided in Supplementary Information Table S3. Linear and nonlinear relationships between variables were considered for the logistic model. As given in Table S3, variables which proved their significance in predictions are RoP and EIA only. However, when considering the confusion matrix given in Table 3, the accuracy of the model was 63.4%, and the sensitivity and specificity were 78% and 36.3%, respectively. The results are not as robust as the performance of the CART and random forest approaches. CART and random forest provide substantial improvements in predicting first flush occurrence compared to the standard statistical methods and ranks the significance of variables, which cannot be easily handled using logistic models.

Table 3
Confusion matrix of binomial logistic regression.

		Actual	
		0	1
Predicted	0	8	9
	1	14	32

The study outcomes constitute the initial step in the pathway to providing a mathematically robust definition of first flush, which is still a disputed phenomenon in the stormwater quality arena. Current approaches for predicting first flush is either through the use of hydrographs and pollutographs or the use of non dimensional arbitrary measures in terms of runoff volume. This study provides a new insight to first flush prediction using machine learning algorithms based on rainfall and site characteristics. Though the study is based on data from selected geographical areas, this novel methodology can be applied to other geographical locations with a complete data set. This methodology can be used for providing an accurate quantitative definition of first flush based on site and rainfall characteristics. It is important to note that the methodology developed is generic and independent of the data used. The data collected was essentially to demonstrate the application of the methodology. The study outcomes can contribute first flush prediction which can be extracted by considering an appropriate modelling strategy which takes into account the significance of variables highlighted in this study.

4. Conclusions

This research study ranked variables in terms of their relative importance in predicting first flush occurrence. The study outcomes revealed that total rainfall depth is the highest ranked factor, followed by the maximum 5 min rainfall intensity, rainfall duration, runoff depth, runoff peak and average rainfall intensity, respectively. Antecedent dry period and the effective impervious area fraction have relatively low rank and the time of concentration and the event mean concentration have the lowest ranking among all the variables investigated in this study. Furthermore, the outcomes of the study showed that, rather than considering the role of an individual variable, investigating the interactions between variables can yield more robust predictions. Additionally, the outcomes of this study also highlight the importance of considering the land cover and the physical characteristics of the catchment in predicting first flush occurrence.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the University Grant Commission of Sri Lanka for the financial support provided to the first author to carry out the postgraduate research studies. We also thankful to the Queensland University of Technology (QUT) for providing the opportunity to undertake this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watres.2019.115075.

References

Alias, N., 2013. *Flirst Flush Behaviour In Urban Residential Catchments.* (PhD). Queensland University of Technolohy.

Alias, N., Liu, A., Egodawatta, P., Goonetilleke, A., 2014. Sectional analysis of the pollutant wash-off process based on runoff hydrograph. J. Environ. Manag. 134, 63–69.

Bach, P.M., McCarthy, D.T., Deletic, A., 2010. Redefining the stormwater first flush phenomenon. Water Res. 44 (8), 2487–2498. https://doi.org/10.1016/

j.watres.2010.01.022.
Bertrand-Krajewski, J.-L., Chebbo, G., Saget, A., 1998. Distribution of pollutant mass vs volume in stormwater discharges and the first flush phenomenon. Water Res. 32 (8), 2341–2356.
Boyd, M., Bufill, M., Knee, R., 1993. Pervious and impervious runoff in urban catchments. Hydrol. Sci. J. 38 (6), 463–478.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees, 37. Wadsworth Int. Group, pp. 237–251, 15.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.
Ebrahimian, A., 2015. *Determination Of Effective Impervious Area in Urban Watersheds*. (PhD). University of Minnesota.
Egodawatta, P., Thomas, E., Goonetilleke, A., 2009. Understanding the physical processes of pollutant build-up and wash-off on roof surfaces. Sci. Total Environ. 407 (6), 1834–1841.
Geiger, W., 1987. Flushing effects in combined sewer systems. In: Paper Presented at the Proceedings of the 4th International Conference Urban Drainage, Lausanne, Switzerland, 1987.
Gnecco, I., Berretta, C., Lanza, L., La Barbera, P., 2005. Storm water pollution in the urban environment of Genoa, Italy. Atmos. Res. 77 (1–4), 60–73.
Goonetilleke, A., Thomas, E., Ginn, S., Gilbert, D., 2005. Understanding the role of land use in urban stormwater quality management. J. Environ. Manag. 74 (1), 31–42.
Goonetilleke, A., Thomas, E.C., 2003. Water Quality Impacts of Urbanisation: Evaluation of Current Research. Centre for Built Environment and Engineering Research, Faculty of Built Environment and Engineering, Queensland University of Technology. Technical Report.
Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63 (4), 308–319.
Gupta, K., Saul, A.J., 1996. Specific relationships for the first flush load in combined sewer flows. Water Res. 30 (5), 1244–1252.
Jacobson, C.R., 2011. Identification and quantification of the hydrological impacts of imperviousness in urban catchments: a review. J. Environ. Manag. 92 (6), 1438–1448.
Jeung, M., Baek, S., Beom, J., Cho, K., Her, Y., Yoon, K., 2019. Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. J. Hydrol. 575, 1099–1110.
Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2065), 20150202.
Kang, J.-H., Kayhanian, M., Stenstrom, M.K., 2006. Implications of a kinematic wave model for first flush treatment design. Water Res. 40 (20), 3820–3830.
Kang, J.-H., Kayhanian, M., Stenstrom, M.K., 2008. Predicting the existence of stormwater first flush from the time of concentration. Water Res. 42 (1–2), 220–228.
Kayhanian, M., Stenstrom, M.K., 2005. Mass loading of first flush pollutants with treatment strategy simulations. Transp. Res. Rec. 1904 (1), 133–143.
Lee, J., Bang, K., Ketchum Jr., L., Choe, J., Yu, M., 2002. First flush analysis of urban storm runoff. Sci. Total Environ. 293 (1–3), 163–175.
Li, D.-l., Shen, F., Yin, Y., Peng, J.-x., Chen, P.-y., 2013. Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. Chin. Med. J. 126 (6), 1150–1154.
Li, D., Wan, J., Ma, Y., Wang, Y., Huang, M., Chen, Y., 2015. Stormwater runoff pollutant loading distributions and their correlation with rainfall and catchment characteristics in a rapidly industrialized city. PLoS One 10 (3), e0118776.
Li, Q.L., Cheng, Q.Y., Qing-Ci, H., L.-L, K., 2007. First flush of storm runoff pollution from an urban catchment in China. J. Environ. Sci. 19, 295–299.
Liu, A., 2011. *Influence Of Rainfall And Catchment Characteristics On Urban Stormwater Quality*. (PhD). Queensland University of Technology.
Liu, A., Gunawardana, C., Gunawardena, J., Egodawatta, P., Ayoko, G.A., Goonetilleke, A., 2016. Taxonomy of factors which influence heavy metal build-up on urban road surfaces. J. Hazard Mater. 310, 20–29.
Mangangka, I.R., 2013. *Role Of Hydraulic Factors in Constructed Wetland And Bioretention Basin Treatment Performence*. (PhD). Quennsland University of Technology.
Massoudieh, A., Abrishamchi, A., Kayhanian, M., 2008. Mathematical modeling of first flush in highway storm runoff using genetic algorithm. Sci. Total Environ. 398 (1–3), 107–121.
Minervini, W.P., 2012. Of: a stochastic stormwater quality volume-sizing method with first flush emphasis. Water Environ. Res. 84 (3), 282.
Queensland Urban Drainage Manual. (2017). Institute of Public Works Engineering Australasia, Queensland..
Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers-a survey. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 35 (4), 476–487.
Sansalone, J.J., Cristina, C.M., 2004. First flush concepts for suspended and dissolved solids in small impervious watersheds. J. Environ. Eng. 130 (11), 1301–1314.
Sartor, J.D., Boyd, G.B., 1972. Water Pollution Aspects of Street Surface Contaminants, vol. 81. US Government Printing Office.
Schiff, K., Tiefenthaler, L., Bay, S., Greenstein, D., 2016. Effects of rainfall intensity and duration on the first flush from parking lots. Water 8 (8), 320.
Sharifi, S., Massoudieh, A., Kayhanian, M., 2011. A stochastic stormwater quality volume-sizing method with first flush emphasis. Water Environ. Res. 83 (11), 2025–2035.
Thompson, D.B., 2006. The Rational Method. David B. Thompson Civil Engineering Department Texas Tech University, pp. 1–7.
Walsh, C.J., 2000. Urban impacts on the ecology of receiving waters: a framework for assessment, conservation and restoration. Hydrobiologia 431 (2–3), 107–114.
Williamson, T.N., Crawford, C.G., 2011. Estimation of suspended-sediment concentration from total suspended solids and turbidity data for Kentucky, 1978-19951. J. Am. Water Resour. Assoc. 47 (4), 739–749.