A stormwater management framework for predicting first flush intensity and quantifying its influential factors

Cosimo Russo[1], Alberto Castro[2], Andrea Gioia[3], Vito Iacobellis[3] and Angela Gorgoglione[4*]

[1]Department of Electronics and Information, Politecnico di Milano, 32 Piazza Leonardo da Vinci, Milano, 20133, Italy.
[2]Department of Computer Science, Universidad de la República, 565 Ave Julio Herrera y Reissig, Montevideo, 11300, Uruguay.
[3]Department Department of Civil, Environmental, Land, Building Engineering and Chemistry, Politecnico di Bari, 126/b Via Amendola, Bari, 70126, Italy.
[4*]Department of Fluid Mechanics and Environmental Engineering, Universidad de la República, 565 Ave Julio Herrera y Reissig, Montevideo, 11300, Uruguay.

*Corresponding author(s). E-mail(s): agorgoglione@fing.edu.uy;
Contributing authors: cosimo.russo@mail.polimi.it; acastro@fing.edu.uy; andrea.gioia@poliba.it; vito.iacobellis@poliba.it;

Abstract

Despite numerous applications of Random Forest (RF) techniques in the water-quality field, its use to detect first-flush (FF) events is limited. In this study, we developed a stormwater management framework based on RF algorithms and two different FF definitions (30/80 and M(V) curve). This framework can predict the FF intensity of a single rainfall event for three of the most detected pollutants in urban areas (TSS, TN, and TP), yielding satisfactory results (30/80:

$\boldsymbol{accuracy}_{average}$ = 0.87; M(V) curve: $\boldsymbol{accuracy}_{average}$ = 0.75). Furthermore, the framework can quantify and rank the most critical variables based on their level of importance in predicting FF, using a non-model-biased method based on game theory. Compared to the classical physically-based models that require catchment and drainage information apart from meteorological data, our framework inputs only include rainfall-runoff variables. Furthermore, it is generic and independent from the data adopted in this study, and it can be applied to any other geographical region with a complete rainfall-runoff dataset. Therefore, the framework developed in this study is expected to contribute to accurate FF prediction, which can be exploited for the design of treatment systems aimed to store and treat the FF-runoff volume.

Keywords: Pollutant First Flush, 30/80, M(V) curve, Random Forest, SWMM, Feature Importance

1 Introduction

Worldwide, urbanization has led to an intensification of anthropogenic activities accompanied by an increase in impervious surfaces (Egodawatta et al, 2009; Dams et al, 2013; Guan et al, 2015). During a precipitation event with a particular duration and intensity, the first portion of the runoff contribution washes away such impervious surfaces, generating wastewater that is more concentrated in pollutants (Di Modugno et al, 2015; Liu et al, 2016). The so-called "first flush" (FF) has been recognized as a typical phenomenon of urban areas since it represents one of the most critical non-point source pollutions. Therefore, it can negatively impact the quality of receiving water bodies (Gorgoglione et al, 2021). Hence, the analysis and control of urban stormwater runoff have become key factors to protect surface water quality.

Consequently, model simulation and assessment represent a critical procedure to estimate the strength of the FF effect in urban areas and understand its characteristics. With this aim, complex physically-based models with high resolution have been developed (i.e., SWMM, InfoWorks, STORM). Overall, they show good performance in predicting pollutant concentration (Gorgoglione et al, 2016; Hur et al, 2018), but their application is limited by data availability (Rodríguez et al, 2021). In fact, they require meteorological data, information about catchment and drainage system characteristics as input. Recently, data-driven models, such as machine-learning models, have been attractive alternatives, particularly for those regions characterized by data scarcity, since they are more flexible, "quick learners," and they perform better in multi-source data prediction (Sun and Scanlon, 2019). Among such models, random forest (RF) has been widely adopted to tackle environmental matters (Creaco et al, 2016; Jeung et al, 2019; Perera et al, 2019; Wang et al, 2021; Vilaseca et al, 2021). It is an ensemble learning method based on decision trees ("weak learners") used for both classification and regression. Therefore, it has the capability not only to predict whether a rainfall event would generate FF but also, in case it does, to predict the exact pollution charge. However, machine-learning-model outcomes should always be interpreted with extra care since they do not have any knowledge about the mechanistic processes they are simulating.

The dynamic and random nature of urban runoff quality is demonstrated to be driven by multiple variables (features) (Gorgoglione et al, 2020a). Numerous studies have been undertaken to quantify such relationships. Li and Barrett (2008), Lee et al (2011), and Gorgoglione et al (2019b) have demonstrated that the FF phenomenon is particularly influenced by the antecedent dry period (ADP), total rainfall (TR), and runoff volume (RV), among the rainfall-runoff variables, and by the percentage of impervious area and the watershed slope, among the catchment characteristics. Gnecco et al (2005) and Kang et al (2006) have also highlighted the importance of maximum rainfall intensity (I_{max}) and rainfall duration (D) for FF occurrence. On the other hand, recently, Perera et al (2019) demonstrated that total rainfall depth has the highest importance in FF prediction, while *ADP* and impervious area fraction have relatively low influence. As far as we know, the feature-importance analysis carried out in previous studies is limited to feature-importance quantification, which can lead to biased findings depending on the model adopted ("model biased") since it is not able to gain deep insights into the processes of complex models. Compared with other methods, Shapley Additive exPlanations (SHAP), including a new class of additive feature-importance measures, presents enhanced computational performance and a better alignment with human interpretation (Lundberg and Lee, 2017). For these reasons, lately, SHAP has been considered a powerful model-interpretation technique adopted in several studies (Zhong et al, 2021; Padarian et al, 2020; Cross et al, 2020; Uusitalo et al, 2015; Wang et al, 2021).

Based on these considerations, the objective of this study is to develop a stormwater management framework able to: *i)* predict the FF occurrence, taking into account rainfall-runoff variables; *ii)* in case a precipitation event generates FF for a particular pollutant, predict the FF intensity; and *iii)* rank and interpret the rainfall-runoff variables in terms of their level of importance in predicting FF with a non-model-biased method, as SHAP (feature importance analysis).

The findings of this study are expected to contribute to the development of accurate and reliable stormwater-quality models in areas characterized by data scarcity, and, therefore, generate effective stormwater treatment design.

2 First flush definitions

In the last decades, several researchers have studied the FF phenomenon (Sartor et al, 1974; Alley and Smith, 1981; Egodawatta et al, 2007), and its definition is still a topic subject to debate. In the literature, numerous formulations have been proposed to assess the FF occurrence. Saget et al (1996) stated that an event generates FF when at least 80% of the pollutant load is washed off by the first 30% of the runoff volume. Helsel et al (1979), followed by Geiger (1984), introduced a dimensionless representation of the phenomenon. This representation, known as $M(V)$ *curve*, consists of drawing the curve that gives the variation of the cumulative pollutant load ($\sum M_i$) divided by the total pollutant load (M) in relation to the cumulative volume ($\sum V_i$) divided by the total volume (V). Given n measurements of flow rate Q_i and concentration C_i with a time interval (Δt_i), and by assuming that Q_i and C_i vary linearly between two measurements, the M(V) curve can be defined as follows (1, 2) Bertrand-Krajewski et al (1998):

$$M(t) = \frac{\sum_{i=1}^{n} C_i Q_i \Delta t_i}{M} = \frac{\sum M_i}{M} \tag{1}$$

$$V(t) = \frac{\sum_{i=1}^{n} Q_i \Delta t_i}{V} = \frac{\sum V_i}{V} \tag{2}$$

Even though this method was useful for comparing pollutant mass with the corresponding flow rate of different rainfall events, Helsel et al (1979) and Geiger (1984) simply defined FF occurrence when the M(V) curve lies above the 1:1 line, representing an excessive pollutant discharge in the first portion of the runoff event. Bertrand-Krajewski et al (1998) improved such M(V) curve definition by introducing four classes of FF intensity. It is known that every M(V) curve can be fitted approximately by a power function (Di Modugno et al, 2015):

$$M(t) = V(t)^b \tag{3}$$

where the coefficient b can be obtained as follows:

Table 1: M(V) curve first flush definition: four classes of first flush intensity.

# Class	Intensity	b values
0	no FF	$b > 1$
1	weak FF	$0.862 < b \leq 1$
2	medium FF	$0.185 < b \leq 0.862$
3	strong FF	$0 < b \leq 0.185$

$$b = \frac{\ln M(t)}{\ln V(t)} \tag{4}$$

FF intensity varies depending on the b value. Four classes of FF intensity can be identified (Table 1).

In this study, both the *30/80* and the improved M(V) curve methods were considered to define FF. In this way, we will also test the performance of the stormwater management framework in a two-class scenario (30/80) and in a multiclass one (M(V) curve).

3 Materials and Methods

3.1 Study area

An urban residential watershed located in Southern Italy, Sannicandro di Bari (SB), was selected as the study site (Di Modugno et al, 2015). It has a surface equal to 31.24 ha, an average slope equal to 1.56%, and its average elevation is 169 m above sea level. The land-use information of this area was obtained from the land use map of 2011 downloaded from SIT.Puglia (2021). 70% of the entire catchment is covered by impervious surfaces (e.g., streets, roofs), while green areas cover only 3.80% of the watershed (e.g., gardens, parks). From the climatic point of view, the study site is characterized by a mean annual temperature equal to 15.0 °C and a mean annual rainfall equal to 586 mm.

The entire stormwater drainage network is 1.96 km long. It collects runoff into a concrete rectangular channel (dimensions: $1.20m \times 1.70m$).

3.2 Data collection

The dataset adopted in this study includes three data sources: observations, simulations, and generations. In the following sections, an in-depth description of each of them is provided.

3.2.1 Observation subset

Hereafter, we will call "observations" those five events monitored at SB with complete rainfall, runoff, and water quality records. The monitoring station consisted of a rain gauge (ISCO 674 model) to record the precipitation, a bubble flowmeter (ISCO 730 model) to monitor flow rate, and an automatic sampler with 24 bottles of 0.5 L each to measure water quality. The latter was evaluated with the standardized methods reported in Baird et al (2017). A detailed description of the data-collection process and the equipment used can be found in Di Modugno et al (2015). A total of five rainfall-runoff events were monitored: 10 Nov 2006, 22 Nov 2006, 1 Dec 2006, 24 Jan 2007, and 10 Feb 2007. The water-quality variables considered for this study were total suspended solids (TSS), total nitrogen (TN), and total phosphorus (TP). This is justified by the fact that nitrogen and phosphorus are the main nutrients presented in urban wash-off (Gorgoglione et al, 2021; Yang and Lusk, 2018), and their adsorption to sediment particles or loose soil is the primary form by which their offsite movement takes place; therefore, TSS were also considered for this study. A summary of the observed rainfall-runoff events (antecedent dry period (ADP), total rainfall (TR), event duration (D), maximum rainfall intensity (I_{max}), average rainfall intensity (I_{ave}), runoff volume (RV), runoff peak (RP)) with the corresponding water-quality data (minimum, maximum,

Table 2: Summary of the rainfall-runoff information of the observation subset.

Event	**ADP (days)**	**TR (mm)**	**ED (min)**	I_{max} **(mm/h)**	I_{ave} (mm/h)	**RV** (m^3)	**RP** (m^3/s)
10 Nov 2006	6	2.4	50	24	0.94	113.49	0.04
22 Nov 2006	11	4.3	112	6	1.10	148.86	0.04
1 Dec 2006	18	5.9	251	12	0.93	286.88	0.05
24 Jan 2007	19	1.6	37	12	0.76	111.62	0.05
10 Feb 2007	6	12.9	398	36	1.57	460.11	0.05

Table 3: Summary of the water-quality information of the observation subset.

Event	**TSS**				**TN**				**TP**			
	min (mg/L)	**max (mg/L)**	**EMC (mg/L)**	**EML (kg)**	**min (mg/L)**	**max (mg/L)**	**EMC (mg/L)**	**EML (kg)**	**min (mg/L)**	**max (mg/L)**	**EMC (mg/L)**	**EML (kg)**
10 Nov 2006	224.0	420.0	19.54	21.32	7.0	8.3	0.47	0.59	0.70	1.00	0.05	0.06
22 Nov 2006	124.0	2160.0	86.40	231.44	3.6	14.0	0.45	1.73	0.24	2.96	0.11	0.25
1 Dec 2006	6.0	217.0	6.04	51.00	-	-	-	-	-	-	-	-
24 Jan 2007	177.0	807.0	47.96	73.66	5.4	10.0	0.48	0.91	0.65	0.99	0.03	0.07
10 Feb 2007	541.0	2090.0	40.00	410.06	6.3	13.0	0.25	2.42	2.08	3.63	0.08	1.02

event mean concentration (EMC), event mean load (EML)) are reported in Tables 2 and 3, respectively. The following equations were adopted to calculate EMC and EML (Gorgoglione et al, 2018, 2021):

$$EMC = \frac{\sum_{i=1}^{n} C_i V_i}{V} \tag{5}$$

$$EML = \sum_{i=1}^{n} C_i V_i = EMC \cdot V \tag{6}$$

where C_i is the average pollutant concentration at time step i [mg/L], V_i is the runoff volume during time increment i [L], V is the total runoff volume per event [L], and n is the total number of samples collected during a precipitation event.

3.2.2 Simulation subset

Thereafter, "simulations" will be those events characterized by recorded rainfall (precipitation data described in Section 3.2.1) and simulated discharge and water quality. The Storm Water Management Model (SWMM) was adopted for water quantity and quality simulations (Rossman, 2015). SWMM was successfully implemented, calibrated, and validated at SB in our previous work

(Di Modugno et al, 2015). For this study, the five observed precipitations, along with the drainage and watershed characteristics, were the input of the SWMM model. Therefore, a total of five simulations were obtained, simulated flow rate and pollutant concentration were collected and constituted the simulation subset. Further information about the adopted model and how it simulates pollutant build-up/wash-off and transport processes are reported in the appendix A.1.

3.2.3 Generation subset

From now on, we will call "generations" those events characterized by synthetic rainfall, obtained by the Iterated Random Pulse (IRP) model (developed by Veneziano and Iacobellis (2002)), and simulated flow rate and pollutant concentration, produced by SWMM model. Specifically, the IRP model was implemented at SB catchment and it generated a 15-year-long rainfall time series with 15 minutes of aggregation. Based on the regional regulation (RegionePuglia, 2013), a criterion of 48 h of ADP was adopted to identify single rainfall events. As a result, 567 synthetic precipitation events were detected and used as input of the calibrated SWMM model to simulate the correspondent discharge and pollutant concentration (TSS, TN, and TP). A thorough description of the IRP model and its implementation in the study area can be found respectively in Veneziano et al (2002) and Gorgoglione et al (2016, 2019b).

It is important to highlight that SWMM and IRP models are exploited in this study as data generators. The objective is to build a proper dataset to assess the management framework capability in predicting the FF occurrence along with its intensity. As mentioned before, both models were properly validated for the study area in our previous works (Di Modugno et al, 2015; Veneziano and Iacobellis, 2002).

3.3 Data analysis

Prior to model implementation, the exploratory data analysis (EDA) was carried out (whose results are reported in section 4.2). It had the objective of detecting possible outliers, understanding the pollutant behavior in our study area, and selecting appropriate rainfall-runoff characteristics for model development with the aim of preventing correlated variables from overshadowing critical relationships between rainfall-runoff characteristics and the FF phenomenon (Gorgoglione et al, 2018). To compute EDA, the Python package *pandas-profiling* was used. Since the processes under study are non-linear, we adopted the Spearman index to evaluate the existence of correlations among the variables taken into account. For confirming and/or adding extra information, the Kendall and Phik indices were computed as well. A description of these three indices is provided in the appendix A.2. Correlation coefficients can be plotted in a *correlation heatmap*, a graph in which variables are associated pairwise and the strength of each correlation is represented by the darkness of the color: the darker the color, the higher the correlation (closer to 1 or -1).

Furthermore, prior to model development, we normalized the input variables to deal with their different measurement-units scale and give equal weight to each of them (Gorgoglione et al, 2020b). Since no outliers were detected in our dataset and all input variables were positive, the MinMaxScaler class from the *sklearn.preprocessing* package was used for min-max normalization, which brings all the variables to the interval $[0, 1]$.

3.4 Random forest classifier

RF is a supervised learning algorithm able to represent non-linear relationships (Breiman, 2001). It is an ensemble method, i.e., it is composed of many weak learners (*decision trees*) that are used to predict a class (classification) or a value (regression). Its response is the most predicted class in case of classification, or the average of the predicted values in case of regression. This study will exploit the RF capability as a classifier, using a random forest classifier (RFC).

To reduce model variance, ensemble algorithms like RF, use bootstrap aggregating (bagging) methods (Breiman, 1996). Such methods build several instances of random subsets of the original training set and then aggregate their individual predictions to form a final prediction. Furthermore, when building such instances, RF also randomizes the set of model features. By randomizing the construction procedure of the weak learner and then making an ensemble out of it, model variance is reduced. The injected randomness decouples the prediction error of individual weak learners. By taking an average of those predictions, some errors can cancel out. For this reason, bagging methods work best with strong and complex models (RF *vs.* decision tree) (Breiman, 2001).

The occurrence of FF and its intensity were predicted using RFCs. Particularly, the occurrence (two classes: no and yes FF) was predicted with both FF definitions (30/80 and M(V) curve), while the intensity was predicted with the M(V) curve classification (four classes: no, low, medium, and strong FF) (Table 1). Both FF definitions were considered and analyzed independently. For each of them, the three pollutants (TSS, TN, and TP) were considered individually, for a total of six RFCs. The Python class *sklearn.ensemble.RandomForestClassifier*, from the *sklearn* package Pedregosa et al (2011), was used for the implementation of the RFC algorithm.

3.5 Model cross-validation and testing

The RFCs were cross-validated with 75% of the rainfall-runoff events (training set) and then tested using the remaining 25% of the dataset (testing set). The two subsets were randomly selected for the two processes (cross-validation and testing) with the aim of minimizing the potential bias that may be introduced in the model assessment.

We randomly divided data into k groups (folds) of approximately equal size for running the cross-validation process. The first fold is used as the validation set and the rest as the training set. Then repeat k times and find the average of the k loss-function values. Considering the input matrix, we adopted 5-fold cross-validation.

During cross-validation, a hyperparameter tuning process that aims at obtaining a reliable model is executed. In general, exploring the entire hyperparameter-domain space is not feasible; therefore, several methods that use sampling and/or heuristics are employed. We used the open-source Python library *Optuna* for hyperparameter optimization, which aims to balance the sampling and pruning algorithms. In Akiba et al (2019), authors introduced it under a new design constituted by three criteria: *i*) define-by-run programming, which allows the user to construct the search space in a dynamic way; *ii*) efficient implementation, which focuses on the optimal functionality of sampling strategies as well as pruning algorithms; *iii*) easy-to-setup, versatile architecture that can be deployed for several types of tasks. Optuna is also framework agnostic, i.e., it can be easily integrated with any of the machine learning frameworks (e.g., Scikit-Learn).

In this study, we considered the following hyperparameters for model cross-validation: *i*) the number of decision trees in the forest (*n_estimators*), *ii*) the maximum depth of each tree in the forest (*max_depth*), *iii*) the maximum number of features that a tree can consider (*max_features*). *n_estimators* is related to the bagging process by controlling the number of instances in the ensemble. *max_depth* controls the decision-tree growth: the deeper the tree, the more splits it has and it captures more information about the data. *max_features* is related to the randomization of model features and represents the number of features to consider when looking for the best split.

To prevent over-fitting, a regularization mechanism was applied to RF. Most regularization parameters prune the trees with different rules. In this case, the parameter used for pruning was *max_ depth*.

After identifying the best hyperparameters for the forest, the final scoring was calculated by exploiting a new RF on the training dataset using the best hyperparameters found and evaluating it on the testing dataset.

To evaluate the learner's performance, two loss functions were used. In a classification problem, the *accuracy* is defined as the number of correctly classified samples (i.e., true positives (T_P) and false negatives (F_N)) divided by the total number of samples (TOT_s) (Eq.7). This value represents the percentage of correctly classified samples. It gives an immediate and intuitive representation of how the model behaves.

$$accuracy = \frac{T_P + F_N}{TOT_s} \tag{7}$$

A more informative value that behaves well even in the case of unbalanced classes is the *F1 score*, defined as follows (Eq. 8):

$$F1score = \frac{1}{\sum_{l \in L} |\hat{y_l}|} \sum_{l \in L} |\hat{y_l}| F1(y_l, \hat{y_l}) \tag{8}$$

where L is the maximum number of classes (l), y and $\hat{y}$ represent the observed and predicted values respectively, y_l and $\hat{y_l}$ are the subset of y and $\hat{y}$ that belong to the class l, and F1 is defined as follows (Eq. 9):

$$F1 = 2 * \frac{P * R}{P + R} \tag{9}$$

where P is the precision (Eq. (10)) and R is the recall (Eq. (11)). In both Eqs. (10) and (11), T_P is the number of true positives, F_P the false positives, and F_N the false negatives.

$$P = \frac{T_P}{T_P + F_P} \tag{10}$$

$$R = \frac{T_P}{T_P + F_N} \tag{11}$$

The F1 score was used as the objective function and the accuracy was computed for validation. F1 score ranges between $[0, 1]$, the closer to 1, the better the model.

We used *stratified* and *uniform* models as baselines. The stratified model generates predictions by respecting the training set's class distribution; the uniform model returns predictions uniformly at random.

3.6 Feature-importance analysis

SHAP was adopted to carry out the feature-importance analysis (Lundberg and Lee, 2017). It is based on the cooperative game theory solution, Shapley Values (Shapley, 1997). The SHAP objective is to explain the prediction of an instance by computing the contribution of each feature to the prediction. SHAP is, therefore, a technique for estimating the expected marginal contribution of a factor among all possible contributions. In this study, we selected SHAP over the given feature importance calculated by RF because *i)* it not only provides the contribution of each predictor but is also able to compute the positive or negative relationship of each variable with the output; *ii)* it is not "model biased," i.e., it can be calculated for any machine-learning classification and regression model, allowing a fair comparison among different models; *iii)* several studies on environmental matters have recently proved its efficiency (Padarian et al, 2020; Uusitalo et al, 2015; Zhong et al, 2021; Cross et al, 2020).

To run this analysis, the SHAP python package was used.

The stormwater management framework implemented in this study is summarized in the flowchart reported in Figure 1. The first-flush analysis, modeling, and feature-importance study were conducted using an Intel core i7 10th generation, 32GB RAM DDR4, and 1TB SSD.

4 Results

4.1 First flush analysis

An initial FF analysis was conducted to identify whether the 577 events (observations + simulations + generations) generate or not FF on the basis of the two definitions adopted in this study. For each event, this was done for TSS, TN, and TP separately. As for the 30/80 classification, the result of this analysis is binary: "FF yes" or "FF no," if the event generates or does not generate FF, respectively. While, for the M(V) curve definition, the outcome is one of the four classes represented in Table 1, which, not only informs about the occurrence of FF but also, specifies its intensity (weak, normal, strong) in case of FF existence. The output of this FF analysis will be used as ground truth for the training/testing process of RFC.

In Figure 2, we reported the class distribution for 30/80 and M(V) curve FF definitions. A class imbalance characterizes the dataset for both FF classifications. In the 30/80 FF, for TSS and TP, almost 30% of the events generates FF, while for TN, almost 70% of the events produces FF. For the M(V) curve definition, Class 2 (medium FF) is the most populated one for TSS (59%) and TP (72%), while the second most populated for TN (42%) after Class 3 (strong

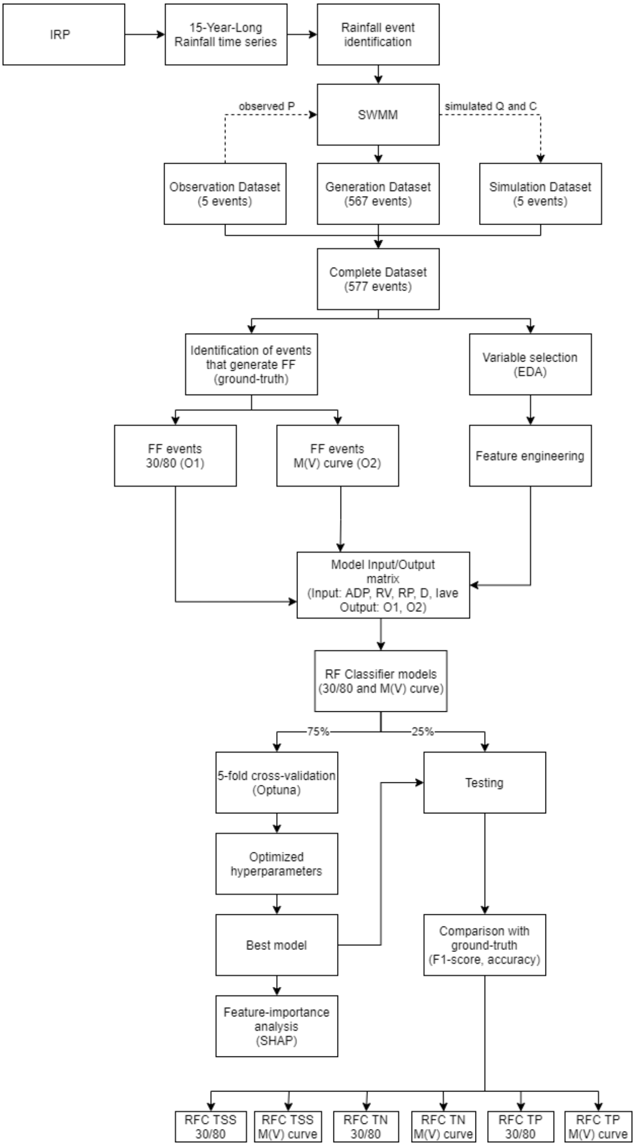
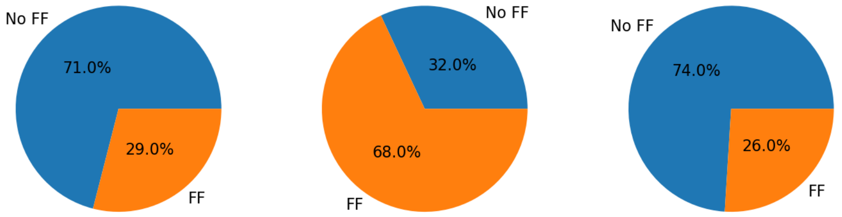
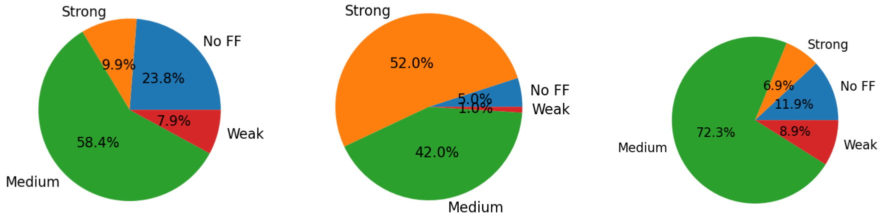

IRP
15-Year-Long Rainfall time series
Rainfall event identification
observed P
SWMM
simulated Q and C
Observation Dataset (5 events)
Generation Dataset (567 events)
Simulation Dataset (5 events)
Complete Dataset (577 events)
Identification of events that generate FF (ground-truth)
Variable selection (EDA)
FF events 30/80 (O1)
FF events M(V) curve (O2)
Feature engineering
Model Input/Output matrix (Input: ADP, RV, RP, D, Iave Output: O1, O2)
RF Classifier models (30/80 and M(V) curve)
75%
25%
5-fold cross-validation (Optuna)
Testing
Optimized hyperparameters
Best model
Comparison with ground-truth (F1-score, accuracy)
Feature-importance analysis (SHAP)
RFC TSS 30/80
RFC TSS M(V) curve
RFC TN 30/80
RFC TN M(V) curve
RFC TP 30/80
RFC TP M(V) curve

Fig. 1: Flowchart of the stormwater management framework implemented in the study.

FF) (52%); Class 1 (weak FF) is the least populated for the three pollutants (8% for TSS, 1% for TN, 9% for TP).



(a) TSS - 30/80 (b) TN- 30/80 (c) TP - 30/80



(d) TSS - M(V) curve (e) TN - M(V) curve (f) TP - M(V) curve

Fig. 2: Sample distribution per FF class for 30/80 (a, b, c) and M(V) curve (d, e, f) FF definitions.

4.2 Exploratory data analysis

The data matrix (577x13) was the input for the EDA, where 577 are the rainfall events (observations + simulations + generations), and 13 are the rainfall-runoff and water quality variables (ADP, TR, RV, RP, D, I_{ave}, I_{max}, EMC_{TSS}, EMC_{TN}, EMC_{TP}, EML_{TSS}, EML_{TN}, and EML_{TP}).

The EDA revealed significant correlations among the input/output variables. This information was not only used for gaining useful insights about the pollutant behavior in our study area (output variables), but also for excluding some of the input variables for the modeling part to reduce model complexity and to prevent them from overshadowing critical relationships between rainfall characteristics and the wash-off process.

In Figure 3, the correlation heatmap calculated with Spearman coefficient is represented. In this regard, two variables were considered to be significantly correlated if the Spearman coefficient was grater than 0.95 ($p - value = 0.05$). As for the input variables, TR and RV showed the highest direct correlation (0.99), followed by I_{max} and RP (0.97). Other strong relationships that confirm the previous ones are between I_{max} and I_{ave} (0.90), I_{ave} and TR (0.89), I_{ave} and RP (0.88), I_{ave} and RV (0.88). However, the latter are lower than 0.95. Based on these results, the variables TR and I_{max} were excluded from the modeling process since they are respectively represented by RV and RP.

As for the output variables, interesting insights were revealed. A high inverse correlation was detected between EML_{TN} respectively with TR (-0.84) and RV (-0.84). While, EML_{TSS} and TP show direct strong relationships with TR (0.83 and 0.82 respectively) and RV (0.83 and 0.82 respectively). It can be observed that, in our study area, TP has a higher particle-bound component than TN, which, instead, tends to reduce its concentration when precipitation and, therefore, runoff increase (dilution process). Moreover, the amount of TSS in stormwater is determined by weathering or displacement processes (high-energy processes) that mobilize suspended solids and make them available to be washed off. Therefore, the higher TR and RV, the greater EML_{TSS}.

Further correlation results obtained from other non-linear coefficients, Kendall and Phik, confirmed the relationships found in Figure 3. These correlation heatmaps are reported in the appendix A.2.

4.3 Classification of first flush events

Six classifiers were independently developed (2 FF definitions × 3 pollutants). For each of them, the matrix (577 x 6) was the input, where 577 are the rainfall events and 6 are the input/output variables (five inputs: ADP, RV, RP, D, I_{ave}; one of the output: FF TSS 30/80, FF TN 30/80, FF TP 30/80, FF TSS M(V), FF TN M(V), FF TP M(V)).

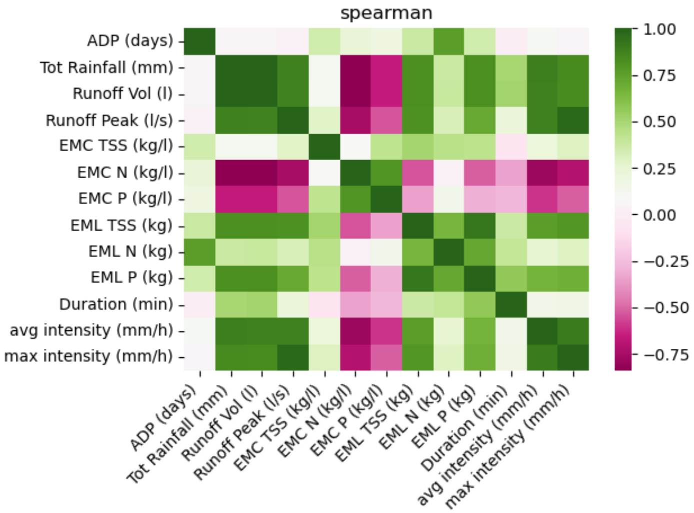


Fig. 3: Correlation heatmap computed with Spearman coefficient.

As mentioned earlier, the input/output variables were min-max normalized to deal with the different order of magnitudes and avoid biased modeling results. Afterward, the RFCs were cross-validated by using 75% of the dataset (training dataset). Three hyperparameters were tuned during the model cross-validation: *n_ estimators*, *max_features*, *max_ depth*. The range of variation and the values chosen for these parameters are shown in Table 4. Since *n_ estimators* control the number of trees in the forest, a high value of this hyperparameter was chosen for all the classifiers to reduce the variance of the ensemble models ($n\_estimators > 1000$). To improve the generalization capability of the algorithms, we wanted to prevent fully grown trees by keeping good model performance. Therefore, we chose low *max_ depth* values. By only considering a random subset of features in each tree, the entropy of the forest increases, further reducing the variance. This is the reason why the *max_features* hyperparameter is generally lower than the total number of variables.

Table 4: Hyperparameter optimization and best values chosen.

Hyperparameter	**Description**	**Range of variation**	**FF definition**	**Pollutant**	**Value chosen**
n_ estimators	The number of trees in the forest.	$[1, +\infty)$	30/80	TSS	1900
				TN	1200
				TP	1800
			M(V) curve	TSS	1400
				TN	1700
				TP	1500
max_ features	The number of features to consider when looking for the best split.	$[1, n]$ with $n = \#$ of input variables (5 in our case)	30/80	TSS	3
				TN	5
				TP	4
			M(V) curve	TSS	4
				TN	2
				TP	4
max_ depth	The maximum depth that each tree is allowed to reach.	$[1, m]$ with $m = \#$ of dataset samples (577 in our case)	30/80	TSS	6
				TN	4
				TP	8
			M(V) curve	TSS	16
				TN	6
				TP	28

Table 5: RFCs results for 30/80 and M(V) curve FF definitions, for TSS, TN, and TP.

FF definition	**Pollutant**	**Baselines**				**Results**	
		F1 score stratified	**Accuracy stratified**	**F1 score uniform**	**Accuracy uniform**	**F1 score**	**Accuracy**
30/80	TSS	0.24	0.60	0.33	0.48	0.79	0.88
	TN	0.63	0.50	0.65	0.57	0.92	0.89
	TP	0.32	0.67	0.32	0.48	0.71	0.83
M(V) curve	TSS	0.42	0.42	0.27	0.23	0.65	0.67
	TN	0.42	0.41	0.31	0.23	0.83	0.83
	TP	0.59	0.58	0.24	0.19	0.73	0.75

5-fold cross-validation, run in the Optuna framework, was adopted for the hyperparameter optimization, using the F1 score as the objective function. The average F1 score of the five experiments was returned as the final result. The Optuna was set up to maximize the objective function, performing 500 experiments with early stopping of 100 runs in case the F1 score was not improved. The six best classifiers found were then tested with the remaining 25% of the data (testing dataset). The correspondent results are reported in Table 5.

For the three pollutants and both FF definitions, the F1 score and accuracy are higher than those obtained if the stratified and uniform predictors were used (baselines). Overall, all the predictions were very satisfactory

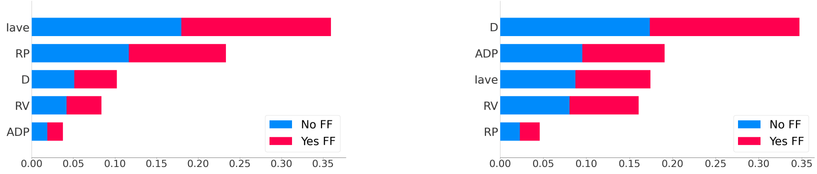
($accuracy_{average}$ = 0.87 for the 30/80 FF definition, $accuracy_{average}$ = 0.75 for the M(V) curve FF definition). Slightly lower performance was found for the M(V) curve classification compared to the 30/80 one due to the class imbalance detected (Figure 2).

4.4 Feature importance for classification model

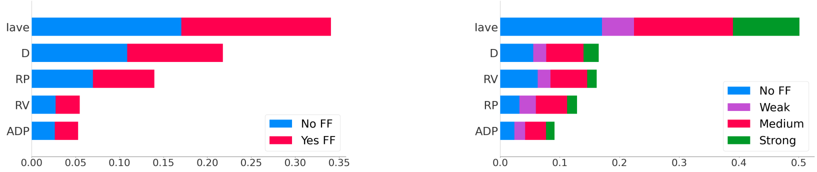
Based on the best RFC models, the SHAP values were computed for the input variables of each model, and the key features for predicting FF occurrence were identified. Figure 4 shows the feature ranking for the three pollutants and for the 30/80 and M(V) curve FF definitions. The different colors in Figure 4 depict the classes described in section 2 and represented in Figure 2.

Based on the mean absolute SHAP values, it is important to remark that, independently from the FF definition adopted and the pollutant considered, I_{ave} is always the most important or among the most significant predictors of FF occurrence. In particular, it is the most critical predicting variable for TSS and TP FF, followed by RP for TSS (30/80) and by D for TP (for both definitions) and for TSS (M(V) curve). In comparison, D is the most significant predictor in TN FF occurrence, followed by ADP (30/80) and I_{ave} (M(V) curve). For TSS and TP, ADP is the least important variable for model prediction for both FF definitions, along with RP only for M(V) curve definition. While, for TN, RP always has the lowest importance.

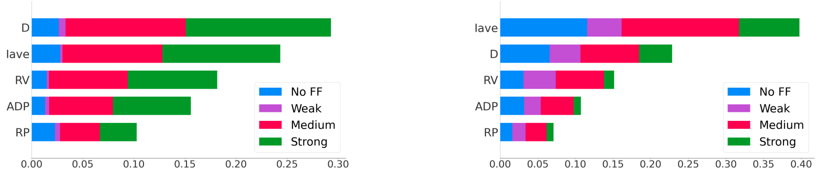
It is essential to highlight that each variable has a different weight on each class, particularly for M(V) curve FF definition. This may be affected by the data distribution in each class (Figure 2). In any case, Class 2 (medium FF) is always well represented for the three pollutants. Class 0 (no FF) is also well characterized for TSS and TP; while Class 3 (strong FF) is also well pictured for TN.



(a) TSS - 30/80 (b) TN - 30/80



(c) TP - 30/80 (d) TSS - M(V) curve



(e) TN - M(V) curve (f) TP - M(V) curve

Fig. 4: SHAP values for 38/80 FF definition: (a) TSS, (b) TN, and (c) TP, and for M(V) curve FF definition: (d) TSS, (e) TN, and (f) TP.

It is worth remarking that sediments transport plays a critical role in the TP mobilization from impervious surfaces at our study area. This is justified by the fact that the same rainfall-runoff variables influence TSS and TP FF prediction. TN shows a different behavior from TSS, confirming its low particle-bound component. These results confirmed the ones obtained with the EDA.

5 Discussion

5.1 Model performance

The first contribution introduced by this study is represented by the prediction of FF intensity (weak, medium, strong) based on the M(V) curve definition. The average accuracy for the three pollutants under study is equal to 75% for the six RFCs. Such level of accuracy is a good step forward, considering that the current approach for predicting pollutant FF is based on a graphical representation. Therefore, FF classes have respectively 25% accuracy in prediction considering the four classes defined by the M(V) curve FF definition, or 50% accuracy regarding the 30/80 FF definition (two classes). Furthermore, Perera et al (2019) found an accuracy of 71% for predicting TSS FF occurrence in a two-class scenario (FF yes or FF no). We were able to improve such prediction reaching an accuracy of 88%. Another aspect to take into account is the comparison of data requirements for RF with a classical physically-based model, such as SWMM. The SWMM inputs include rainfall and other meteorological data, catchment characteristics, and drainage system information along with storage/treatment system characteristics. The inputs of the RF models implemented in the management framework simply include rainfall-runoff variables. Based on SHAP outcomes, considering that rainfall variables are the most significant predictors of pollutant FF occurrence, it may be possible that RF model inputs can be limited to rainfall characteristics without significantly decreasing model performance.

5.2 Influencing variables for FF prediction

This study also identified and quantified the most important variables in predicting sediment and nutrient FF in urban areas by adopting a non-model biased method. Such capability of SHAP was demonstrated by the fact that, independently from the FF definition (30/80 or M(V) curve) used, the results yielded were very similar. In fact, it was found that I_{ave} was always the most important predictor of TSS and TP FF occurrence. In contrast, D was the most critical predictor for TN FF. The latter was also proved by Jeung et al (2019), who stated that TN showed a strong correlation with rainfall duration, but the range in importance rate was wider than the ones found for the other variables. Therefore, the correlation TN-rainfall characteristics always requires further research compared to other pollutants. Numerous studies also demonstrated that there is a higher correlation between rainfall characteristics (more than runoff characteristics) and TSS and TP FF existence. Perera et al (2019) found that TSS FF was mainly influenced by rainfall depth and maximum rainfall intensity. Jeung et al (2019) proved the strong relationship existing between rainfall intensity and TSS and TP concentration. Despite the low SHAP values of ADP, confirmed by several authors (Perera et al, 2019; Jeung et al, 2019), the impact of such variable cannot be neglected since its influence on pollutant FF has been proved in past researches (Lee et al, 2011; Sartor et al, 1974; Gorgoglione et al, 2019a) and it was then found for the EML_{TN} prediction.

6 Conclusion

This study developed a stormwater management framework based on the RF algorithm and two different FF definitions (30/80 and M(V) curve). Such framework was able to predict rainfall events that generate FF and, in case they produce FF, it can also predict its intensity. The framework was developed for TSS, TN, and TP. Normalized rainfall-runoff variables were considered as model input, including ADP, RV, RP, D, and I_{ave}. For the three pollutants and for both FF definitions, the predictions were very satisfactory

($accuracy_{average} = 0.87$ for the 30/80 FF definition, $accuracy_{average} = 0.75$ for the M(V) curve FF definition).

Furthermore, exploiting a non-model biased method (SHAP), the framework ranked the model input variables to investigate their importance in predicting pollutant FF. It was found that I_{ave} was always the most critical predictor of TSS and TP FF occurrence. While D was the most critical variable for TN FF. For TSS and TP, ADP is the least important variable for model prediction for both FF definitions. For TN, RP always has the lowest importance. These outcomes show that, in our study area, sediment transport plays a key role in the TP mobilization from the impervious portion of the watershed. Furthermore, the results also highlight that, instead of studying the role of an individual variable, analyzing the interactions among variables can return more robust predictions.

This study demonstrated the potential of the stormwater management framework developed as a tool to estimate FF and help to better understand stormwater quality processes in urban areas. Since computing resources and algorithms have quickly advanced in the last decade, machine learning methods are expected to be more frequently adopted in hydro-environmental studies particularly when a rapid analysis and solution are required with a limited number of observations.

References

Akiba T, Sano S, Yanase T, et al (2019) Optuna: A next-generation hyperparameter optimization framework. Association for Computing Machinery, New York, NY, USA, p 2623–2631, https://doi.org/10.1145/3292500.3330701

Alley WM, Smith PE (1981) Estimation of accumulation parameters for urban runoff quality modeling. Water Resources Research 17(6):1657–1664. https://doi.org/https://doi.org/10.1029/WR017i006p01657

Baak M, Koopman R, Snoek H, et al (2020) A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. Computational Statistics & Data Analysis 152:107,043. https://doi.org/https://doi.org/10.1016/j.csda.2020.107043

Baird R, Eaton A, Rice E (2017) Standard Methods for the Examination of Water and Wastewater, 23rd edn. American Public Health Association, American Water Works Association, and Water Environment Federation

Bertrand-Krajewski JL, Chebbo G, Saget A (1998) Distribution of pollutant mass vs volume in stormwater discharges and the first flush phenomenon. Water Research 32(8):2341–2356. https://doi.org/https://doi.org/10.1016/S0043-1354(97)00420-X

Breiman L (1996) Bagging predictors. Machine Learning 24:123–140. https://doi.org/10.1007/BF00058655

Breiman L (2001) Random forests. Machine Learning 45:32–45. https://doi.org/10.1023/A:1010933404324

Creaco E, Berardi L, Sun S, et al (2016) Selection of relevant input variables in storm water quality modeling by multiobjective evolutionary polynomial regression paradigm. Water Resources Research 52(4):2403–2419. https://doi.org/https://doi.org/10.1002/2015WR017971

Cross T, Sathaye K, Darnell K, et al (2020) Predicting Water Production in the Williston Basin Using a Machine Learning Model, pp 3492–3503.

https://doi.org/10.15530/urtec-2020-2756

Dams J, Dujardin J, Reggers R, et al (2013) Mapping impervious surface change from remote sensing for hydrological modeling. Journal of Hydrology 485:84–95. https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.09.045, hydrology of peri-urban catchments: processes and modelling

Di Modugno M, Gioia A, Gorgoglione A, et al (2015) Build-up/wash-off monitoring and assessment for sustainable management of first flush in an urban area. Sustainability 7(5):5050–5070. https://doi.org/https://doi.org/10.3390/su7055050

Egodawatta P, Thomas E, Goonetilleke A (2007) Mathematical interpretation of pollutant wash-off from urban road surfaces using simulated rainfall. Water Research 41(13):3025–3031. https://doi.org/https://doi.org/10.1016/j.watres.2007.03.037

Egodawatta P, Thomas E, Goonetilleke A (2009) Understanding the physical processes of pollutant build-up and wash-off on roof surfaces. Science of The Total Environment 407(6):1834–1841. https://doi.org/https://doi.org/10.1016/j.scitotenv.2008.12.027

Geiger W (1984) Characteristics of combined sewer runoff. In: Proceeding de la 3ème conférence internationale «Urban Storm Drainage», Göteborg, pp 4–8

Gnecco I, Berretta C, Lanza L, et al (2005) Storm water pollution in the urban environment of genoa, italy. Atmospheric Research 77(1):60–73. https://doi.org/https://doi.org/10.1016/j.atmosres.2004.10.017, precipitation in Urban Areas

Gorgoglione A, Gioia A, Iacobellis V, et al (2016) A rationale for pollutograph evaluation in ungauged areas, using daily rainfall patterns: Case studies of the apulian region in southern italy. Applied and Environmental Soil Science 2016. https://doi.org/10.1155/2016/9327614

Gorgoglione A, Bombardelli FA, Pitton BJL, et al (2018) Role of sediments in insecticide runoff from urban surfaces: Analysis and modeling. International Journal of Environmental Research and Public Health 15(7). https://doi.org/https://doi.org/10.3390/ijerph15071464

Gorgoglione A, Bombardelli FA, Pitton BJ, et al (2019a) Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of sediment transport in urban areas. Environmental Modelling & Software 111:170–181. https://doi.org/https://doi.org/10.1016/j.envsoft.2018.09.022

Gorgoglione A, Gioia A, Iacobellis V (2019b) A framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. Sustainability 11(18). https://doi.org/https://doi.org/10.3390/su11184933

Gorgoglione A, Castro A, Gioia A, et al (2020a) Application of the self-organizing map (som) to characterize nutrient urban runoff. In: Gervasi O, Murgante B, Misra S, et al (eds) Computational Science and Its Applications – ICCSA 2020. Springer International Publishing, Cham, pp 680–692, https://doi.org/https://doi.org/10.1007/978-3-030-58811-3_49

Gorgoglione A, Gregorio J, Ríos A, et al (2020b) Influence of land use/land cover on surface-water quality of santa lucía river, uruguay. Sustainability 12(11). https://doi.org/https://doi.org/10.3390/su12114692

Gorgoglione A, Castro A, Iacobellis V, et al (2021) A comparison of linear and non-linear machine learning techniques (pca and som) for characterizing urban nutrient runoff. Sustainability 13(4). https://doi.org/https://doi.org/10.3390/su13042054

Guan M, Sillanpää N, Koivusalo H (2015) Modelling and assessment of hydrological changes in a developing urban catchment. Hydrological Processes 29(13):2880–2894. https://doi.org/https://doi.org/10.1002/hyp.10410

Helsel DR, Kim JI, Grizzard TJ, et al (1979) Land use influences on metals in storm drainage. Journal (Water Pollution Control Federation) 51(4):709–717

Hur S, Nam K, Kim J, et al (2018) Development of urban runoff model ffc-qual for first-flush water-quality analysis in urban drainage basins. Journal of Environmental Management 205:73–84. https://doi.org/https://doi.org/10.1016/j.jenvman.2017.09.060

Jeung M, Baek SS, Beom J, et al (2019) Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. Journal of Hydrology 575:1099–1110. https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.05.079

Kang JH, Kayhanian M, Stenstrom MK (2006) Implications of a kinematic wave model for first flush treatment design. Water Research 40(20):3820–3830. https://doi.org/https://doi.org/10.1016/j.watres.2006.09.007

Lee JY, Kim H, Kim Y, et al (2011) Characteristics of the event mean concentration (emc) from rainfall runoff on an urban highway. Environmental Pollution 159(4):884–888. https://doi.org/https://doi.org/10.1016/j.envpol.2010.12.022

Li MH, Barrett ME (2008) Relationship between antecedent dry period and highway pollutant: Conceptual models of buildup and removal processes. Water Environment Research 80(8):740–747. https://doi.org/10.2175/106143008x296451

Liu A, Gunawardana C, Gunawardena J, et al (2016) Taxonomy of factors which influence heavy metal build-up on urban road surfaces. Journal of Hazardous Materials 310:20–29. https://doi.org/https://doi.org/10.1016/j.jhazmat.2016.02.026

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., p 4765–4774

Padarian J, McBratney AB, Minasny B (2020) Game theory interpretation of digital soil mapping convolutional neural networks. SOIL 6(2):389–397. https://doi.org/10.5194/soil-6-389-2020

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Perera T, McGree J, Egodawatta P, et al (2019) Taxonomy of influential factors for predicting pollutant first flush in urban stormwater runoff. Water Research 166:115,075. https://doi.org/https://doi.org/10.1016/j.watres.2019.115075

RegionePuglia (2013) Regional Regulation, 9 December 2013, nº26, “Stormwater runoff and first flush regulations” (implementation of article 13 of Legislative Decree nº152/06 and subsequent amendments).

Rodríguez R, Pastorini M, Etcheverry L, et al (2021) Water-quality data imputation with a high percentage of missing values: A machine learning approach. Sustainability 13(11). https://doi.org/https://doi.org/10.3390/su13116318

Rossman LA (2015) Storm Water Management Model User's Manual Version 5.1. U.S. Environmental Protection Agency (EPA), National Risk Management Research Laboratory Office of Research and Development U.S. Environmental Protection Agency, Cincinnati, OH, USA

Saget A, Chebbo G, Bertrand-Krajewski JL (1996) The first flush in sewer systems. Water Science and Technology 33(9):101–108. https://doi.org/https://doi.org/10.1016/0273-1223(96)00375-7, solids in Sewers

Sartor JD, Boyd GB, Agardy FJ (1974) Water pollution aspects of street surface contaminants. Journal (Water Pollution Control Federation) 46(3):458–467

Shapley LS (1997) A value for n-person games. Classics in game theory 69

SIT.Puglia (2021) SIT Puglia. http://www.sit.puglia.it/, accessed: 2021-12-15

Sun A, Scanlon B (2019) How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters 14(7):073,001. https://doi.org/10.1088/1748-9326/ab1b7d

Uusitalo L, Lehikoinen A, Helle I, et al (2015) An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software 63:24–31. https://doi.org/https://doi.org/10.1016/j.envsoft.2014.09.017

Veneziano D, Iacobellis V (2002) Multiscaling pulse representation of temporal rainfall. Water Resources Research 38(8):13–1–13–13. https://doi.org/https://doi.org/10.1029/2001WR000522

Veneziano D, Furcolo P, Iacobellis V (2002) Multifractality of iterated pulse processes with pulse amplitudes generated by a random cascade. Fractals 10(02):209–222. https://doi.org/https://doi.org/10.1142/S0218348X02001026

Vilaseca F, Castro A, Chreties C, et al (2021) Daily rainfall-runoff modeling at watershed scale: A comparison between physically-based and data-driven models. In: Gervasi O, Murgante B, Misra S, et al (eds) Computational Science and Its Applications – ICCSA 2021. Springer International Publishing, Cham, pp 18–33

Wang F, Wang Y, Zhang K, et al (2021) Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environmental Research 202:111,660. https://doi.org/https://doi.org/10.1016/j.envres.2021.111660

Yang YY, Lusk MG (2018) Nutrients in urban stormwater runoff: Current state of the science and potential mitigation options. Current Pollution Reports 4:112–127. https://doi.org/10.1007/s40726-018-0087-7

Zhong S, Zhang K, Wang D, et al (2021) Shedding light on “black box” machine learning models for predicting the reactivity of ho radicals toward organic compounds. Chemical Engineering Journal 405:126,627. https://doi.org/https://doi.org/10.1016/j.cej.2020.126627

Statements and Declarations

- Funding

  Cosimo Russo's visiting at Universidad de la República was partially funded by the exchange program "Tesi all'estero" from Politecnico di Milano.

- Competing Interests

  The authors have no relevant financial or non-financial interests to disclose.

- Ethics approval

  Not applicable.

- Consent to participate

  Not applicable.

- Consent for publication

  Not applicable.

- Availability of data and materials

  The water quality dataset described in this article can be accessed from https://gitlab.com/fing-hydroinformatics/first-flush-rfc/-/tree/main/data.

- Code availability

  The stormwater management framework developed for this work is freely available at https://gitlab.com/fing-hydroinformatics/first-flush-rfc. It was implemented in Phyton3 using Conda (two scripts, one for Linux and one for MS Windows, can be found to generate the software environment with all its requirements). This framework can be run in any general-purpose computer.

- Authors' contributions

  Cosimo Russo: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Funding acquisition.

  Alberto Castro: Conceptualization, Methodology, Investigation, Writing - Review & Editing, Supervision, Funding acquisition.

Andrea Gioia: Resources, Writing - Review & Editing.

Vito Iacobellis: Resources, Writing - Review & Editing.

Angela Gorgoglione: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

All authors read and approved the final manuscript.

Appendix A Supplementary information

A.1 SWMM model description and implementation

SWMM simulates the hydrograph and pollutograph for a real storm event (for a single and long-term event) based on the rainfall and other meteorological inputs, and system characteristics (catchment, conveyance, and storage/treatment) for urban and peri-urban watersheds. SWMM has been designed in blocks or operating units. Each block can be used individually or in a cascade, and an executive block coordinates its outputs. The runoff block, as well as the transport block, were utilized for this study. By using inlet hydrographs generated from the runoff unit, the transport block executes the flow and pollutant routing through the drainage network.

To simulate the runoff from urban surfaces, the kinematic-wave equation was chosen. Furthermore, the water losses taken into account were represented by the depression storage on the impervious portion of the watershed and the infiltration process. The latter was modeled by evaluating, for each subcatchment, the percentage of the impervious and pervious area obtained from the land-use map. The infiltration model adopted in this work was based on Horton's equation, whose parameter values have been chosen according to the representative values reported in the literature in relation to soil type. Eight parameters of the runoff block of SWMM were used to calibrate the hydraulic-hydrologic model: the depth of depression storage on impervious ($Dstore - Imperv$) and pervious ($Dstore - Perv$) portions of the subcatchment, Manning's coefficient for overland flow over the impervious ($N - Imperv$) and pervious ($N - Perv$) portions of the subcatchment, the percent of the impervious area without depression storage ($\%ZeroImperv$), and the infiltration parameters of Horton's equation.

Pollutant build-up within a land-use category is described by a mass per unit of subcatchment area. The amount of build-up is a function of the number of dry weather days antecedent to the rainfall event. The build-up function follows a growth law that asymptotically approaches a maximum limit:

$$M_a(d_{adp}) = \frac{Accu}{Disp} \cdot A \cdot P_{imp}(1 - e^{Disp \cdot d_{adp}}) \tag{A1}$$

where $M_a(d_{adp})$ represents the pollutant build-up during the antecedent dry period [kg/ha]; $Disp$ is the parameter that measures the disappearance of accumulated solids due to the action of wind or vehicular traffic [1/d]; P_{imp} is the impervious area fraction; $Accu$ the parameter that characterizes the solids build-up rate [kg/(ha d)]; $\frac{Accu}{Disp} \cdot A \cdot P_{imp}$ presents the maximum asymptotic limit of the build-up curve. The pollutant wash-off over different land uses takes place during wet periods, and it is described by the differential equation:

$$\frac{dM_d(t)}{dt} = -Arra \cdot i(t)^{wash} \cdot M_a(t) \tag{A2}$$

where $\frac{dM_d(t)}{dt}$ is the wash-off load rate [kg/h]; $Arra$ is the wash-off coefficient [mm^{-1}]; $i(t)$ is the runoff rate [mm/h]; $wash$ is the wash-off exponent, a parameter that controls the influence of rainfall intensity on the amount of leached pollutants. Four parameters of the runoff block were identified for the calibration of the water-quality model. For the build-up function: the parameter that characterizes the solids build-up rate ($Accu$) and the parameter that identifies the disappearance of accumulated sediments due to the wind or vehicular traffic ($Disp$). For the wash-off function: the wash-off coefficient ($Arra$) and the wash-off exponent ($wash$).

A.2 Exploratory data analysis: indices and results

Spearman, Kendall and Phik coefficients are able to capture non-linear correlations. Spearman exploits monotonicity, while Kendall measures ordinal associations. Both coefficients have values in domain $[-1, 1]$, where -1 indicates a perfect negative correlation, $+1$ a perfectly positive correlation, and 0 no correlation. The formulas are defined in equations A3 and A4, respectively.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{A3}$$

$$\tau = \frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{\binom{n}{2}} \tag{A4}$$

For Spearman's ρ, d_i is the difference between the two ranks of each observation, and n is the number of observations. For Kendall's τ, the definition of concordant and discordant pairs is needed: a pair of values $(x_i, y_i), (x_j, y_j), i < j$ is concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$.

The Phik coefficient Baak et al (2020) is also a non-linear correlation coefficient that was refined to work consistently with continuous and categorical variables.

The corresponding correlation heatmaps are reported in Figures A1 and A2

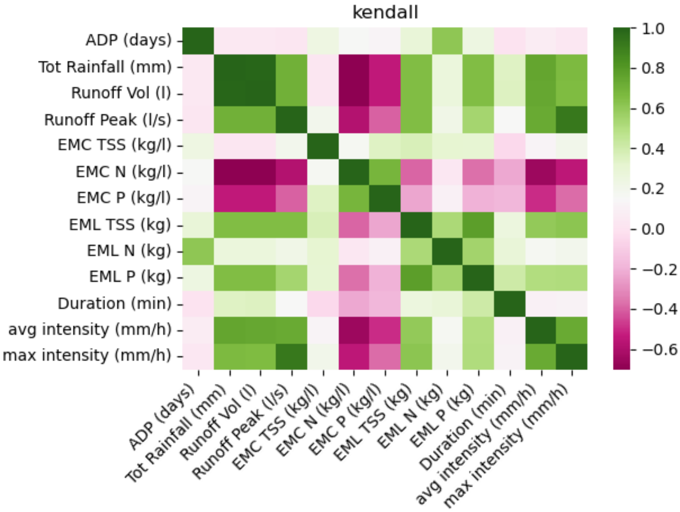


Fig. A1: Correlation heatmap computed with Kendall coefficient.

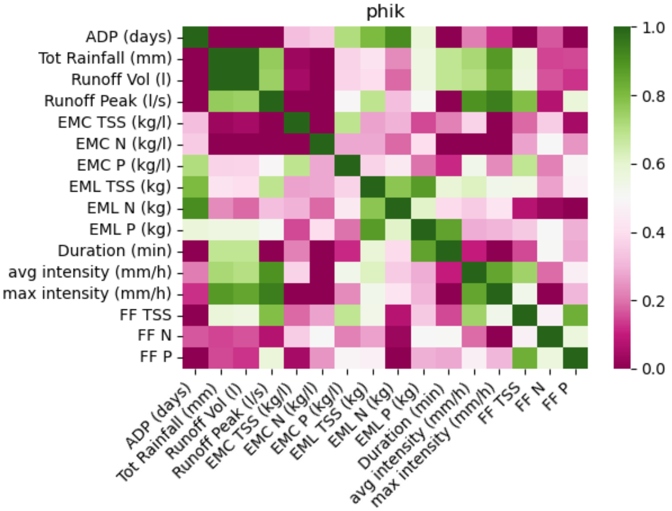


Fig. A2: Correlation heatmap computed with Phik coefficient.