First Flush Occurrence Prediction and Ranking of Its Influential Variables in Urban Watersheds: Evaluation of XGBoost and SHAP Techniques

Angela Gorgoglione[1](✉), Cosimo Russo[2], Andrea Gioia[3], Vito Iacobellis[3], and Alberto Castro[1]

[1] Universidad de la República, 11300 Montevideo, Uruguay
{agorgoglione,acastro}@fing.edu.uy
[2] Politecnico di Milano, 20133 Milan, Italy
cosimo.russo123@gmail.com
[3] Politecnico di Bari, 70126 Bari, Italy
{andrea.gioia,vito.iacobellis}@poliba.it

Abstract. In the last decades, the quality of the surface-water bodies has been increasingly threatened mainly due to anthropogenic activities (i.e., urbanization and the consequent increase of impervious surfaces). This leads to the generation of runoff even more rich in pollutants and difficult to adequately treat in the specific first-flush treatment plant. Consequently, accurate water-quality predictions in urban areas and related detections of events that may generate first flush are the key to enhancing urban water management and pollution control. In this study, the ability of a supervised machine-learning technique (XGBoost) to predict the occurrence of first flush was demonstrated. The model showed outstanding performance in predicting first flush for three of the most detected pollutants in urban areas (total suspended solids, total nitrogen, and total phosphorus) by using rainfall-runoff variables as input. Furthermore, by exploiting a non-model-biased method based on game theory (SHAP), such variables were quantified and ranked based on their level of importance in pollutant first-flush predictions. The findings of this work proved that the XGBoost model is a functional tool for enhancing the accuracy of first-flush predictions in urban watersheds and, thus, contributes to the development of an effective design of first-flush treatment plants.

Keywords: First flush · Urban runoff · XGBoost · Hydroinformatics

1 Introduction

Accurate prediction of water quality in urban areas plays a crucial role in minimizing stream pollution and protecting the health of aquatic ecosystems. Due to anthropogenic activities, such as agriculture, urbanization, sewage discharge, and water extraction, the quality of the surface-water bodies has been increasingly threatened by affecting the ecosystems and populations that benefit from such resources [1, 2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13379, pp. 423–434, 2022.
https://doi.org/10.1007/978-3-031-10545-6_29

The water-quality field counts of a long history of developing discharge and water-quality prediction models: for several watershed scales, based on different datasets, and evaluated with different performance criteria. However, it is still challenging to identify "the best model" under certain conditions [3].

To simulate and predict water quality in urban regions, several physically-based models have been successfully adopted, such as SWMM, STORM, FLUPOL, and InfoWorks [4, 5]. However, such models are limited by several factors: *i)* data availability since they require rainfall-runoff information along with catchment and drainage-network characteristics as input; *ii)* they are time-consuming and labor-intensive since they require a meticulous implementation in a specific geographic area and an adequate data pre-processing [6, 7].

Supervised machine-learning algorithms have been proving to be a very useful tool for predicting surface-water quality [8, 9]. Even though researchers were initially diffident about adopting machine-learning models since they are often black-box predictors, during the last years, more and more scholars have been using data-driven models to predict water quality [10, 11]. One of the most widespread algorithms in machine learning is ensemble learning. It employs the basic idea of gathering weak learners (*decision trees*) to create a powerful model with a higher prediction rate, to enhance the robustness and the generalization ability by combining multiple learners [12]. Currently, ensemble learning such as random forest (RF) and Gradient Boosting Decision Tree (GBDT) are widely adopted [13]. Extreme gradient boosting (XGBoost) is also an excellent algorithm for ensemble learning [14]. Compared with other ensemble models, XGBoost can enhance the model's robustness by introducing terms for regularization and column sampling. When each decision tree chooses the split point, a parallelization strategy is adopted to improve the model's running speed [15]. Furthermore, it can be exploited for both classification and regression, requiring less training and prediction time. However, to our knowledge, XGBoost's ability to predict the occurrence of first flush (FF) in urban watersheds was never investigated.

An important aspect to take into account in water-quality modeling is the dynamic and random nature of urban runoff quality [16]. Numerous studies demonstrated that it is influenced by a multitude of variables, such as pollutant type, physical water-quality characteristics, and spatial and temporal factors [2, 17, 18]. However, only a few works ranked such variables based on their level of importance in predicting FF.

Based on these considerations, this study aims to *i)* assess the XGBoost algorithm's ability to predict FF occurrence in urban areas; *ii)* identify and rank the input variables in terms of their importance in predicting FF.

The water-quality variables considered for this study were total suspended solids (TSS), total nitrogen (TN), and total phosphorus (TP) since they are the most detected pollutants in urban runoff [19, 20].

2 Materials and Methods

2.1 Study Site

The study area selected for this study is an urban watershed located in Sannicandro di Bari, Southern Italy. It is a catchment with the following characteristics: surface =

31.24 ha, slope = 1.56%, average elevation = 169 m a.s.l., mean annual rainfall = 586 mm, mean annual temperature = 15.0 °C, impervious area = 70% of the total area.

The sewage network is 1.96 km long. It collects water into a rectangular concrete channel (1.20 m × 1.70 m).

The watershed under study, along with its sewage network and the discharge point, is represented in Fig. 1.

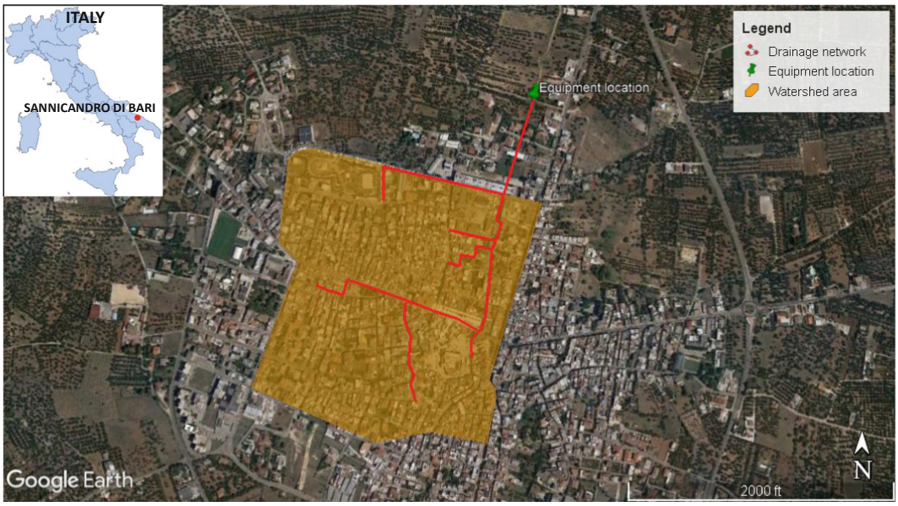

Fig. 1. Study area located in Southern Italy [2].

2.2 Dataset Creation

A thorough description of the dataset creation process can be found in Gorgoglione et al. [2].

It is essential to highlight that the dataset is constituted by 141 rainfall-runoff events (data points) from three different data sources: *observations*, *simulations*, and *generations*. The *observations* are time series of discharge (Q_{obs}), precipitation (P_{obs}), and pollutant concentration (C_{obs}) observed during a monitoring campaign carried out at the study site. The *simulations* are rainfall-runoff events characterized by recorded rainfall (P_{obs}) and simulated discharge (Q_{sim}), and pollutant concentration (C_{sim}). The simulations were obtained with the Stormwater Management Model (SWMM) [21] implemented and calibrated for the study site [22]. P_{obs} along with watershed and drainage-network characteristics were used as model input. The *generations* are rainfall-runoff events constituted by synthetic precipitation (P_{gen}) obtained from the Iterated Random Pulse (IRP) model [23], used as input of SWMM, and simulated discharge (Q_{sim}) and pollutant concentration (C_{sim}). It is important to highlight that the dataset used for this study is a subset of the one described in Gorgoglione et al. [2]. A summary of the dataset creation is reported in Table 1.

Table 1. Dataset creation.

Data type	Composition	Data source
Observations	P_{obs}, Q_{obs}, C_{obs}	Monitoring campaign (*obs*)
Simulations	P_{obs}, Q_{sim}, C_{sim}	Monitoring campaign (*obs*), SWMM (*sim*)
Generations	P_{gen}, Q_{sim}, C_{sim}	IRP (*gen*), SWMM (*sim*)

The water-quality variables considered for this study are total suspended solids (TSS), total phosphorus (TP), and total nitrogen (TN). This is justified by the fact that TP and TN are among the primary nutrients detected in urban runoff, and their adsorption to particles is the main form by which their offsite movement takes place; therefore, TSS were also taken into account for this work.

The rainfall-runoff and water-quality variables (hereafter also called *features* or *attributes*) calculated and used as input/output of the model are the following: antecedent dry period (*ADP*), average rainfall intensity (I_{ave}), rainfall duration (*D*), total rainfall (*TR*), maximum rainfall intensity (I_{max}), runoff volume (*RV*), and TSS, TN, and TP event mean load (EML_{TSS}, EML_{TN}, EML_{TP}). It is noteworthy mentioning that prior data analysis was adequately carried out with the aim of choosing appropriate rainfall-runoff features to avoid correlated variables from overshadowing meaningful relationships between rainfall-runoff characteristics and the FF phenomenon.

2.3 First Flush Analysis: 30/80 First Flush Definition

A FF analysis was conducted to identify how many rainfall-runoff events among the 141 generate FF for the three pollutants under study. This was done on the basis of the 30/80 FF definition [24]: an event generates FF for a particular pollutant when at least 80% of such pollutant load is washed off by the first 30% of the runoff volume. In Fig. 2, we reported, as an example, the FF occurrence calculated for TN for the rainfall event of June 19, 2008 (a data point of the 141 that constitute our dataset). In this plot, the blue curve V(t) represents the cumulative runoff volume divided by the total volume; the red curve M(t) represents the cumulative pollutant mass (TN in this case) divided by the total TN mass. Since the 80% of TN load (red dotted line) "occurs" before the 30% of runoff volume (blue dotted line), this event generates TN FF.

This FF definition was applied to the entire dataset. Such results are presented in Table 2. Almost 70% of the events does not generate FF for TSS and TP (71% and 74%, respectively). In contrast, almost 70% of the events generates FF for TN (68% precisely).

This information was used as ground truth to train the XGBoost model.

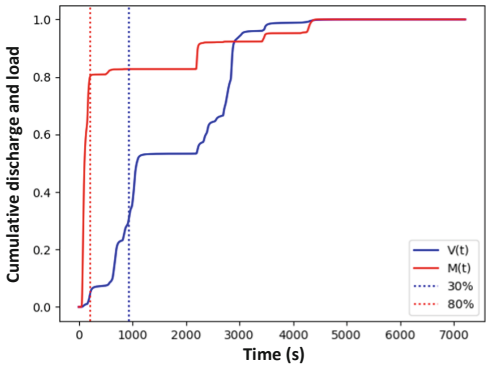

Fig. 2. TN FF occurrence for the event of June 19, 2008. (Color figure online)

Table 2. Datapoint distribution in the two classes "No FF" and "Yes FF" based on the 30/80 FF definition.

Pollutant	No FF occurrence (No FF)	FF occurrence (Yes FF)
TSS	71.0%	29.0%
TN	32.0%	68.0%
TP	74.0%	26.0%

2.4 eXtreme Gradient Boosting (XGBoost) Model

XGBoost model, proposed by Chen and Guestrin [25], belongs to decision tree-based machine-learning models. Such models have the advantages of handling missing data well and reaching fast high performance by just building the decision tree.

XGBoost is one of the most adopted supervised machine-learning models in the environmental field since it has been showing good performance [13, 26, 27]. It is able to couple several weak learners (*decision trees*) into one robust learning machine by generating and iterating multiple trees. This algorithm is characterized by the following features: *i) parallelization*: it can automatically use the CPU multithreading for parallelism while enhancing the algorithm to improve accuracy (this represents the most notable feature of XGBoost). *ii) Regularization:* it includes different regularization penalties to avoid overfitting and adequately generalize the model. *iii) Non-linearity*: it can detect and learn from non-linear data patterns. *iv)* Big amounts of data can be handled at high speed according to block technology. *v)* It is a learning algorithm based on the decision-tree model and can automatically process scarce data [27].

2.5 Model Training and Testing

A 5-fold cross-validation was adopted to calibrate and test the XGBoost algorithm. In particular, 75% of the rainfall-runoff events were used for training the model and the remaining 25% for testing it.

During cross-validation, a hyperparameter optimization process that aims at obtaining an accurate and reliable model was carried out. For this purpose, we adopted the open-source Python library *Optuna* [28].

Generally, the XGBoost hyperparameters are divided into four groups: general parameters, boosting parameters, learning task parameters, and command-line parameters. Before running an XGBoost model, the first three types of parameters must be set up. In contrast, the command-line parameters are only used in the console version of XGBoost and, therefore, they are not considered for optimization. The hyperparameters considered in this work were *n_estimators* and *reg_lambda*. In Table 3, we described such hyperparameters, highlighting their role in the model and their range of variation. *reg_lambda* was used to handle the regularization part of XGBoost since increasing this value will make the model more conservative.

Table 3. Selected hyperparameters for optimization.

Group	Hyperparameter	Description	Range of variation
Boosting parameters	*reg_lambda*	It is responsible for L2 regularization on leaf weights	0.00–1.00
Learning task parameters	*n_estimators*	It is the number of decision trees	1000–2000

After identifying the best set of hyperparameters, the final scoring was computed by using such a set on the training dataset (25% of the dataset).

2.6 Goodness-of-Fit Indicators

Model performance was evaluated by computing the *F1 score* and the *Accuracy*. The latter represents the number of correctly classified samples (i.e., true positives (*TP*) and true negatives (*TN*)) (*CORRs*) divided by the total number of samples (*TOTs*) (Eq. 1):

$$Accuracy = \frac{CORR_s}{TOT_s} \tag{1}$$

It represents the percentage of correctly classified samples related to the total amount of samples. Therefore, it ranges between 0 and 1: the closer to 1, the higher the model performance. Since the *Accuracy* is not particularly descriptive on its own in case of class imbalance and, therefore, it does not take into account false positives (*FP*) and false negatives (*FN*), it is often complemented with the *F1 score* (Eq. 2):

$$F1score = 2\frac{P * R}{P + R} \tag{2}$$

where the precision (*P*) and the recall (*R*) are respectively represented in Eqs. 3 and 4:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

F1 score varies between 0 and 1. Also in this case, the closer to 1, the better the model performance.

2.7 Shapley Additive exPlanations (SHAP) Analysis

SHapley Addictive exPlanation (SHAP) was adopted to carry out the feature-importance analysis [29].

It is based on the Shapley Values, the cooperative game-theory solution [30]. The main objective of SHAP is to estimate the expected marginal contribution of a feature among all possible contributions. In other words, it explains the prediction of an instance by calculating the contribution of each variable to the prediction.

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

3 Results and Discussion

3.1 Hyperparameter Optimization

Optuna was set up to run T = 1000 trials, with a sliding window size t = 100. F1 score was used as the objective function during the training process. The search space definition and the best hyperparameters found for the XGBoost models for the three pollutants under study are provided in Table 4.

Table 4. Hyperparameters optimization.

Hyperparameter	Range of variation	Pollutant	Value chosen
reg_lambda	0.00–1.00	TSS	0.65
		TN	0.30
		TP	1.00
n_estimators	1000–2000	TSS	2000
		TN	2000
		TP	1000

3.2 Model Performance

The XGBoost model was tested using the best hyperparameter configuration found.

To summarize its performance, the three confusion matrices (one per pollutant) resulting from the model testing are reported in Fig. 3. Following Eq. 3, P is equal to 0.73, 0.90, and 0.72, respectively for TSS, TN, and TP. Following Eq. 4, R is equal to 0.86, 0.95, and 0.59, respectively for TSS, TN, and TP. In other words, this means that the probability of predicting a FF- and non-FF event as they actually are for the three pollutants is relatively high (low false positives and false negatives, respectively).

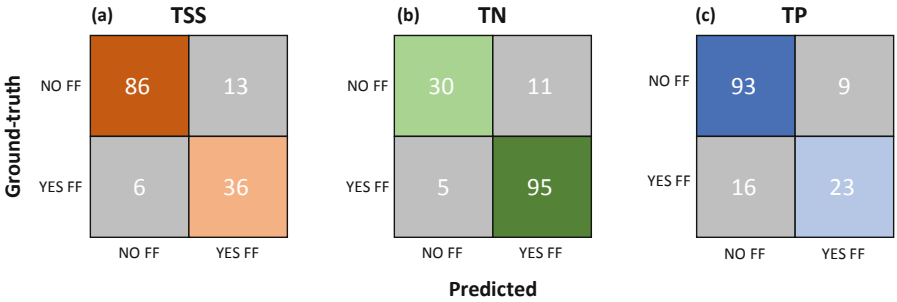

Fig. 3. Confusion matrix for (a) TSS, (b) TN, and (c) TP.

Goodness-of-fit indicator values were calculated for the three pollutants under study and are summarized in Table 5. Overall, the XGBoost model developed shows very satisfactory results. In particular, slightly lower performance was found for TP. The level of *Accuracy* reached in this study (always higher than 80%) is a good step forward, considering that the current approach for predicting pollutant FF is based on a graphical representation with 50% of *Accuracy* (FF occurrence or not). Furthermore, Perera et al. [31] in their work found an *Accuracy* of 71% for predicting TSS FF occurrence in a two-class scenario like the one under study. We were able to enhance such state-of-the-art, reaching an *Accuracy* of 87%.

Table 5. XGBoost model performance.

Goodness-of-fit indicator	Pollutant	Results
Accuracy	TSS	0.87
	TN	0.89
	TP	0.82
F1 score	TSS	0.79
	TN	0.92
	TP	0.65

3.3 Feature-Importance Analysis

Based on the best XGBoost model, the SHAP values were calculated for the input features, and the critical variables for predicting FF occurrence were ranked. The variables are ranked in terms of their degree of influence in predicting FF occurrence. The feature ranking for the three pollutants under study is reported in Fig. 4. Overall, high values of all the input variables have a positive impact on the model outcomes (red dots); while low values may not have an influence or have a negative impact on model results (blue dots). In particular, it is possible to see that I_{ave} is always among the most important

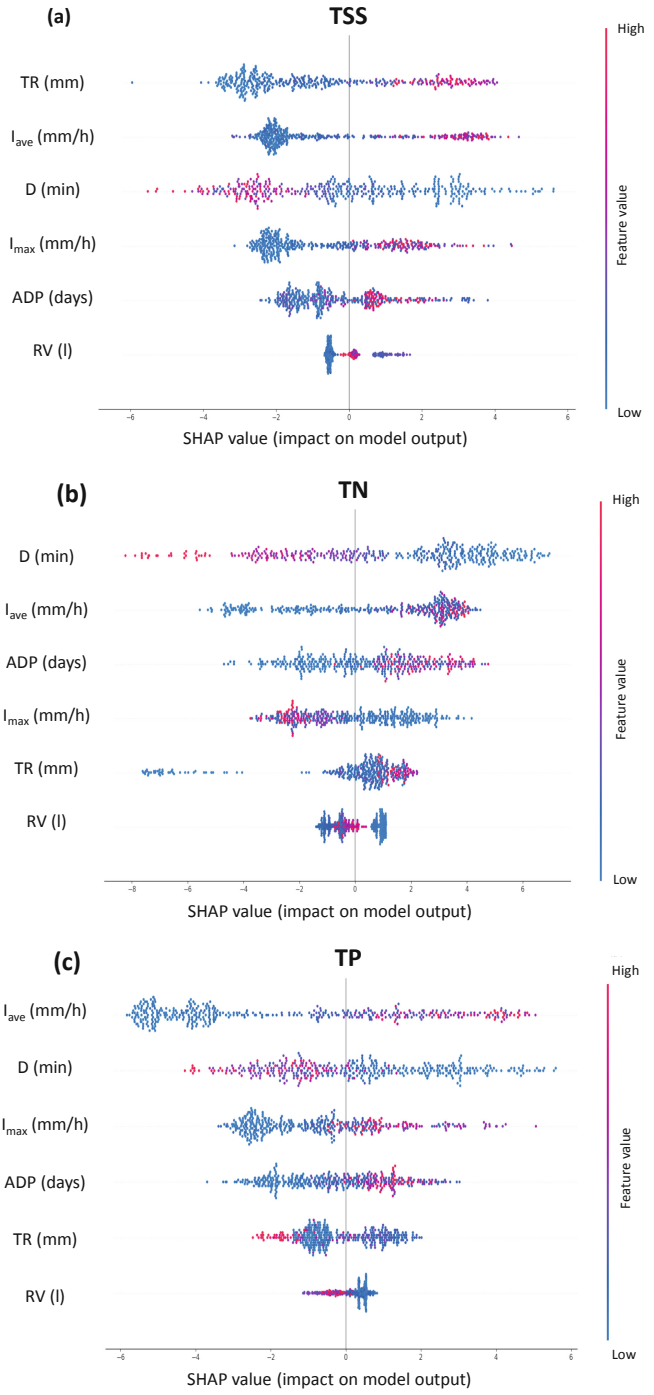

Fig. 4. SHAP values to explain the prediction of FF occurrence for (a) TSS, (b) TN, and (c) TP. (Color figure online)

variables for the three pollutants, as well as *D*. While, *RV* is always the least important variable in the prediction of FF occurrence for sediment and nutrients. Furthermore, it is interesting to highlight that *TR* is the most critical variable for sediment FF prediction; while it is not that important for nutrient FF.

Based on such results, we can state that rainfall characteristics are more critical than the runoff ones in FF occurrence prediction in urban areas. This represents an advantage of machine-learning models compared to classical physically-based models. In fact, the latter require rainfall-runoff data, watershed and drainage network characteristics. While, machine learning models, XGBoost in this case, performs well only considering rainfall-runoff input and, based on SHAP results, it may be possible that by reducing the input variables to only rainfall characteristics, its performance would continue being good even though a slight decrease may be detected.

4 Conclusions

This research study assessed the ability of the XGBoost algorithm in predicting the occurrence of FF in urban watersheds of the three most detected pollutants in urban runoff (TSS, TN, and TP), based on the 30/80 FF definition. *ADP*, I_{ave}, *D*, *TR*, I_{max}, and *RV* were rainfall-runoff variables considered as input. For the three pollutants, the model performance were very satisfactory: $Accuracy_{TSS} = 0.87$, $Accuracy_{TN} = 0.89$ and $Accuracy_{TP} = 0.82$. The robustness of the model is also confirmed by the resulting confusion matrices that show a high probability of correctly predicting true positives and false negatives.

Moreover, with the aid of a non-model biased method (SHAP) based on game theory, the rainfall-runoff input variables were ranked on the basis of their level of importance in predicting FF occurrence. It was found that I_{ave} and *D* were always among the most important variables for the three pollutants. While, *RV* was always the least significant feature in the prediction of FF occurrence for sediment and nutrients. Furthermore, it was found that *TR* was the most critical variable for sediment FF prediction; while it was not a critical one for nutrient FF.

Since algorithms and computing resources have been progressing rapidly, machine-learning techniques are expected to be more often exploited in water-quality studies, mostly when quick analysis and solutions are needed with the limited observation availability. However, the use of such techniques should be handle with extra care taking into account their limitations which could mislead results. For instance, XGBoost is sensitive to outliers, constraining its application to some environmental fields, and it tends to overfit the model if the latter is not early stopped (as it was done for this study).

References

1. Hounslow, A.: Water Quality Data: Analysis and Interpretation. CRC Press, Boca Raton (2018)
2. Gorgoglione, A., Castro, A., Iacobellis, V., Gioia, A.: A comparison of linear and non-linear machine learning techniques (PCA and SOM) for characterizing urban nutrient runoff. Sustainability **13**(4) (2021)

3. Vilaseca, F., Narbondo, S., Chreties, C., Castro, A., Gorgoglione, A.: A comparison between lumped and distributed hydrological models for daily rainfall-runoff simulation. IOP Conf. Ser. Earth Environ. Sci. **958**(1), 012016 (2022)
4. Hur, S., Nam, K., Kim, J., Kwak, C.: Development of urban runoff model c-qual for first-flush water-quality analysis in urban drainage basins. J. Environ. Manag. **205**, 73–84 (2018)
5. Gorgoglione, A., Gioia, A., Iacobellis, V., Piccinni, A.F., Ranieri, E.: A rationale for pollutograph evaluation in ungauged areas, using daily rainfall patterns: case studies of the apulian region in southern Italy. Appl. Environ. Soil Sci. **2016** (2016)
6. Rodríguez, R., et al.: Water-quality data imputation with a high percentage of missing values: a machine learning approach. Sustainability **13**(11) (2021)
7. Vilaseca, F., Castro, A., Chreties, C., Gorgoglione, A.: Daily rainfall-runoff modeling at watershed scale: a comparison between physically-based and data-driven models. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12955, pp. 18–33. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87007-2_2
8. Sun, A., Scanlon, B.: How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. **14**(7) (2019)
9. Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H.: Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. **202** (2021)
10. Russo, C., Castro, A., Gioia, A., Iacobellis, V., Gorgoglione, A.: A stormwater management framework for predicting first flush intensity and quantifying its influential factors. Earth Space Sci. Open Arch. **44** (2022). https://doi.org/10.1002/essoar.10510381.1
11. Jeung, M., Baek, S.-S., Beom, J., Cho, K., Her, Y., Yoon, K.: Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. J. Hydrol. **575**, 1099–1110 (2019)
12. Gudiyangada, T., Piralilou, S.T., Gholamnia, K., Ghorbanzadeh, O., Blaschke, T.: Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using dempster shafer theory. J. Hydrol. **590**(125275) (2020)
13. Ma, M., et al.: XGBoost-based method for flash flood risk assessment. J. Hydrol. **598**(126382) (2021)
14. Hosseiny, H., Nazari, F., Smith, V., Nataraj, C.: A framework for modeling flood depth using a hybrid of hydraulics and machine learning. Sci. Rep. **10**(1) (2020)
15. Budholiya, K., Shrivastava, S.K., Sharma, V.: An optimized XGBoost based diagnostic system for effective prediction of heart disease. J. King Saud Univ. Comput. Inf. Sci. (2020)
16. Gorgoglione, A., Castro, A., Gioia, A., Iacobellis, V.: Application of the self-organizing map (SOM) to characterize nutrient urban runoff. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12252, pp. 680–692. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58811-3_49
17. Gorgoglione, A., Gioia, A., Iacobellis, V.: A framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. Sustainability **11**, 4933 (2019)
18. Gorgoglione, A., Gregorio, J., Ríos, A., Alonso, J., Chreties, C., Fossati, M.: Influence of land use/land cover on surface-water quality of Santa Lucía river, Uruguay. Sustainability **12**(11) (2020)
19. Pitt, R.: Characterizing and controlling urban runoff through street and sewerage cleaning. Water Eng. Res. Lab. (1985)
20. Sartor, J.D., Boyd, G.B., Agardy, F.J.: Water pollution aspects of street surface contaminants. J. (Water Pollut. Control Fed.) **46**, 458–467 (1974)

21. Rossman, L.A.: 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 (2015)
22. Di Modugno, M., et al.: Build-up/wash-off monitoring and assessment for sustainable management of first flush in an urban area. Sustainability **7**, 5050–5070 (2015)
23. Veneziano, D., Iacobellis, V.: Multiscaling pulse representation of temporal rainfall. Water Resour. Res. **38**, 131–1313 (2002)
24. Saget, A., Chebbo, G., Bertrand-Krajewski, J.L.: The first flush in sewer systems. Water Sci. Technol. **33**(9), 101–108 (1996)
25. Chen, T., Guestrin, C.: XGboost: a scalable tree boosting system. In: Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2016)
26. Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F., El-Shafie, A.: Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng. J. **12**, 1545–1556 (2021)
27. Lu, H., Ma, X.: Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere **249**(126169) (2020)
28. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework, pp. 2623–2631. Association for Computing Machinery, New York (2019)
29. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: I. Guyon, U.V., Luxburg, S., Bengio, H., Wallach, R., Fergus, S., Vishwanathan, R.G. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774. Curran Associates, Inc. (2017)
30. Shapley, L.S.: A value for n-person games. In: Classics in Game Theory (1997)
31. Perera, T., McGree, J., Egodawatta, P., Jinadasa, K., Goonetilleke, A.: Taxonomy of influential factors for predicting pollutant first flush in urban stormwater runoff. Water Res. **166**(115075) (2019)