water

MDPI

Article

Machine Learning Surrogate Modeling in R for Rapid Screening of Green Infrastructure Hydrological Performance in Urban Stormwater Management: A Proof-of-Concept Study Using Synthetic Data

Raghad Awad *, Štefan Stanko, Danka Barloková, Ján Ilavský and Ivona Škultétyová

Department of Sanitary and Environmental Engineering, Slovak University of Technology in Bratislava, 812 43 Bratislava, Slovakia; stefan.stanko@stuba.sk (Š.S.); danka.barlokova@stuba.sk (D.B.); jan.ilavsky@stuba.sk (J.I.); ivona.skultetyova@stuba.sk (I.Š.)

* Correspondence: raghad.awad@stuba.sk

Abstract

Background: Physically-based, coupled hydrological–low-impact-development (LID) models, such as the U.S. EPA Storm Water Management Model (SWMM), estimate green infrastructure (GI) performance in detail but are computationally expensive to run across many catchment, storm, and typology combinations. Methods: This methodological proof-of-concept develops an open-source R workflow (randomForest, xgboost, caret) on a synthetic dataset of 2000 catchment–storm–typology scenarios generated from prescribed non-linear equations of imperviousness, storm return period, and the coverage of four GI typologies. None of the scenarios were generated by SWMM-LID simulation or field monitoring. Random forest, XGBoost, and a linear baseline were trained to predict synthetic peak-flow attenuation and suspended-solid (TSS) removal. Results: On the held-out synthetic test set, XGBoost and random forest reached R^2 values of 0.96 and 0.92 for peak-flow attenuation (linear baseline: 0.90) and 0.94 and 0.89 for TSS reduction (linear baseline: 0.79). These values show that the models learned the synthetic response surface; they do not measure predictive skill for real GI systems. Feature importance reproduced the typology weighting embedded in the data-generating equations. Surrogate inference took about 2–4 ms (measured), compared with tens of minutes typically reported in the literature for SWMM-LID runs (not measured here); this is an indicative comparison, not a controlled benchmark. Conclusions: This study is a methodological demonstration only and is not a validated hydrological surrogate. Real application requires retraining and validation using SWMM-LID simulation ensembles or field-monitored data.

Keywords: green infrastructure; machine learning; random forest; XGBoost; R programming; surrogate modeling; urban stormwater management; synthetic data; proof of concept

Check for updates

Academic Editor: Gwo-Fong Lin

Received: 31 August 2026
Revised: 21 September 2026
Accepted: 26 September 2026
Published: 28 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Physically-based, coupled hydrological–hydraulic models extended with a low-impact-development (LID) module—such as the U.S. EPA Storm Water Management Model (SWMM)—are the dominant tool for quantifying the peak-flow attenuation, runoff-volume reduction, and pollutant-removal performance of green infrastructure (GI) [1] at the catchment scale [2]. These models are mechanistically grounded and, given adequate calibration data, can represent the storage, infiltration, and treatment processes of individual GI

typologies with considerable fidelity [3,4]. They remain, for good reason, the reference standard against which any faster approximation must ultimately be judged: their governing equations are derived from established principles of open-channel and pipe hydraulics, infiltration theory, and pollutant transport, so their outputs carry a physical interpretability that purely statistical approaches cannot claim on their own terms.

Their principal practical limitation is computational cost. Evaluating a single catchment under a single design storm and a single GI-retrofit configuration already requires meaningful runtime—typically minutes to tens of minutes per run, depending on catchment size, time-step resolution, and the number of LID sub-catchments represented—and city-scale planning frequently requires comparing many candidate configurations before a shortlist of sites is carried forward to detailed engineering design. A realistic screening exercise might need to compare four or more GI typologies, several plausible coverage densities for each, and multiple design storms across dozens or hundreds of sub-catchments; the combinatorial product of these dimensions can easily reach into the thousands of distinct scenarios, a scale at which repeated physically-based simulation becomes impractical within a normal municipal planning timeline, however desirable it would be to evaluate every combination exhaustively. This screening burden grows further where a composite, multi-typology program is being planned rather than a single intervention [5], since the typologies do not act independently and their combined performance cannot simply be summed from individually simulated results.

Machine learning (ML) surrogate models offer a complementary approach to this screening problem. Once trained on a sufficient body of input–output examples from a physically-based model (or, where available, from monitored field data), a surrogate model can approximate the same input–output relationship at a small fraction of the computational cost, enabling a near-instantaneous evaluation of new scenarios [6]. This general strategy has already demonstrated value across water-resource applications, ranging from rainfall–runoff forecasting to water-quality prediction [6–9], but its application specifically to GI hydrological performance screening, implemented in a fully open, reproducible statistical computing environment such as R, remains comparatively underexplored. R is a natural choice of implementation environment for this purpose: it is free and open source, has a mature and actively maintained ecosystem of machine learning packages, and is already widely used in the environmental and water-engineering research community for statistical analysis, making a surrogate modeling workflow built in R more readily adoptable by other researchers and practitioners than one built in a closed, licensed platform.

It is important to distinguish three conceptually different kinds of "surrogate" model in this literature: (i) a conventional surrogate trained on the outputs of a high-fidelity physically-based simulation ensemble (e.g., a calibrated SWMM–LID model run many times); (ii) a data-driven model trained directly on monitored field observations; and (iii) a synthetic analytical emulator, trained on data generated from prescribed closed-form equations, as used in the present study. This distinction is central to judging novelty and evidentiary strength, since only the first two provide evidence about real hydrological behavior. Recent work has already pursued the first route for GI/LID screening: Bao et al. generated 10,000 samples from a calibrated and validated SWMM and trained machine learning surrogates to predict total and peak catchment outflows for low-impact-development sizing and optimization [10]; Kim et al. combined SWMM with a long short-term memory (LSTM) network to evaluate stormwater-runoff reduction by green infrastructure in a monitored urban catchment in Seocho-gu, Seoul, explicitly targeting computational acceleration of SWMM-based screening [11]; and a design-based case study of green roofs in a small (about 2 ha) urban area reported a reduction of about 16% in the maximum design flow rate relative to conventional roofs [12]. Positioned against this

emerging body of SWMM-trained and field-calibrated ML surrogates, the contribution of the present article is more modest and more precisely stated: it is a transparent, fully open source R-based demonstration that multiple linear regression, random forest, and XGBoost can recover a deliberately non-linear, interacting synthetic GI performance relationship, together with a reproducible workflow that can later be re-pointed at a genuine SWMM-LID ensemble or monitored data, rather than a new surrogate-modeling methodology or a validated predictive tool for real catchments. It should also be emphasized that GI hydrological performance in practice depends on considerably more than the six catchment, storm, and typology-coverage variables considered here: soil properties and infiltration parameters, ground slope, drainage-network configuration, LID internal geometry and media composition, antecedent soil moisture, the temporal shape of the rainfall hyetograph (not only its return period), and the spatial placement of GI assets within the catchment can all materially affect performance [3,4,13]. Reducing the representation of the system to six variables is therefore a deliberate simplification adopted for this proof-of-concept demonstration, not a claim that these six variables are sufficient to characterize GI performance in general; extending the framework to a richer, physically-based feature set is identified as necessary future work in Section 4.

This article develops and demonstrates an R-based machine learning surrogate modeling workflow for emulating a synthetic GI hydrological performance response surface defined by catchment, storm, and typology-coverage features, using a synthetic, proof-of-concept dataset described in Section 2.1. In line with this scope, this manuscript is presented as a methodological proof-of-concept and technical note rather than as a full hydrological validation study: its contribution is a reproducible workflow and an algorithm comparison on a known synthetic response surface, and it makes no claim to have validated the models against SWMM-LID simulations or field observations. Given that random forest and XGBoost surrogates for hydrological applications are already well established methodologically, the principal contribution claimed here is not a new surrogate-modeling technique, but a transparent, fully open source R-based demonstration comparing linear regression, random forest, and XGBoost for emulating a deliberately non-linear, interacting GI performance relationship, together with a reproducible workflow, clearly distinguished from an already-validated hydrological surrogate. The framework is deliberately structured around three widely used, well-documented modeling approaches—multiple linear regression as an interpretable baseline, random forest as a robust bagged-tree ensemble, and XGBoost as a state-of-the-art gradient-boosted tree ensemble—so that the incremental value of tree-based, non-linear modeling over a simple linear approximation can be assessed directly and transparently, rather than assumed. Four research questions structure the analysis: (i) can random forest and XGBoost models, implemented in R, reproduce a synthetic GI peak-flow-attenuation response surface, defined from catchment and typology-coverage features, with acceptable accuracy? (ii) How does prediction accuracy compare between random forest, XGBoost, and a naive linear baseline? (iii) Which catchment and typology features are most influential in determining GI hydrological performance, according to feature-importance analysis? (iv) What indicative computational efficiency gain does surrogate inference offer relative to typical literature runtimes for physically-based hydrological–LID simulations (an illustrative comparison rather than a controlled benchmark)? Sections 2 and 3 answer these questions in turn, and Section 4 discusses what the answers do, and do not, license in terms of practical municipal application.

2. Materials and Methods

The overall workflow, ranging from synthetic dataset generation through model training to evaluation and feature-importance analysis, is summarized schematically in Figure 1 and described in detail in Sections 2.1–2.5 below.

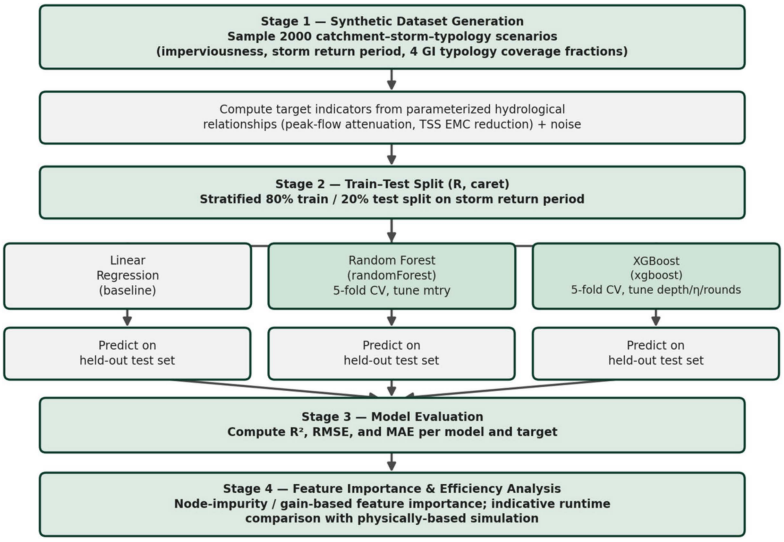


Figure 1. Schematic workflow of the R-based surrogate modeling framework, from synthetic dataset generation (Stage 1) through train–test splitting (Stage 2), model training (three parallel branches), evaluation (Stage 3), and feature-importance and computational-efficiency analysis (Stage 4). Green-shaded boxes denote the four main workflow stages and the two tree-based models; grey boxes denote intermediate steps and the linear baseline; arrows indicate the direction of the workflow.

2.1. Synthetic Training Dataset Generation

A training dataset of 2000 catchment–storm–typology scenarios was generated to support model development. Each scenario was defined by six input features: baseline catchment imperviousness I (sampled uniformly between 30% and 80%), design storm return period T (sampled from the discrete set {2, 10, 100} years), and the coverage fraction (0–100% of feasible area) assigned to each of four GI typologies—bioretention cells (C_{bio}), green roofs (C_{gr}), permeable pavement (C_{pp}), and vegetated swales (C_{sw}). Sampling was performed independently and uniformly at random for each feature, so that the training data spanned the full joint parameter space rather than being concentrated around a small number of representative catchment archetypes, which is important for a surrogate model that is later expected to interpolate across arbitrary combinations of these features. Each typology's coverage fraction was defined relative to its own typology-specific feasible implementation area within the catchment (for example, the area of suitable rooftops for green roofs, or the area of parking and low-traffic pavement for permeable pavement), not relative to a single shared footprint; the four feasible areas were treated as distinct and non-competing in this proof-of-concept dataset, so that scenarios in which several typologies simultaneously approached 100% of their own respective feasible area remained internally consistent. This is nonetheless a simplification: in a real catchment, the feasible areas for different typologies can overlap (e.g., a parcel suitable for bioretention may also be suitable for permeable pavement), so the true joint feasibility constraint is more restrictive than assumed here, and this is noted as a limitation of the synthetic design in Section 4.

Rather than assigning each typology an independent, additive contribution to hydrological performance, the dataset-generation process combined the four coverage fractions into a single weighted aggregate coverage effect, with weights set in proportion to each typology's unit-coverage attenuation performance (permeable pavement: 0.45, bioretention

cells: 0.38, vegetated swales: 0.25, green roofs: 0.22). These weights are illustrative rather than calibrated: no specific real catchment or calibration dataset underlies them. They were set to reflect, in relative rank order only, the comparatively higher unit-area hydrological performance reported for permeable pavement and bioretention relative to green roofs and vegetated swales in previous physically-based and field-monitored studies [14–17], and should not be read as parameters fitted to any particular site. This is a modeling simplification adopted for the proof-of-concept synthetic dataset and is revisited as a limitation in Section 4:

$$E_{cov} = (0.38 \cdot C_{bio} + 0.22 \cdot C_{gr} + 0.45 \cdot C_{pp} + 0.25 \cdot C_{sw})/100 \quad (1)$$

This aggregate coverage effect was then modified by two further terms before being passed through a saturating (exponential) response function. First, an imperviousness penalty term reduces the effective contribution of a given unit of GI coverage as baseline catchment imperviousness increases, reflecting the physically reasonable expectation that GI retrofits achieve less runoff control, per unit of coverage, on catchments that are more heavily sealed to begin with:

$$P_{imp} = 1.35 - 0.012 \cdot (I - 30) \quad (2)$$

Second, a storm-severity factor reduces effective performance as the design storm return period increases, consistent with the diminishing relative effectiveness of fixed-capacity storage and infiltration measures against progressively larger design events:

$$F_{storm} = 1.22 - 0.095 \cdot \ln(T) \quad (3)$$

The two target hydrological indicators—peak-flow attenuation (A, in percent) and total-suspended-solid (TSS) event-mean-concentration reduction (R_{TSS}, in percent)—were then computed from a saturating exponential function of the product of the aggregate coverage effect, the imperviousness penalty, and the storm-severity factor, so that performance rose steeply for small increases in coverage from a low base and flattened out (diminishing returns) as coverage approached full feasible extent:

$$A = 74 \cdot [1 - \exp(-2.6 \cdot E_{cov} \cdot P_{imp} \cdot F_{storm})] \quad (4)$$

An analogous expression, using pollutant-specific typology weights (bioretention: 0.78, permeable pavement: 0.71, vegetated swales: 0.58, and green roofs: 0.55, reflecting each typology's relative filtration performance for particulate pollutants) and a separately calibrated storm-severity term, was used to generate the TSS-reduction target; the full expression is given in Appendix A (Equations (A1)–(A3)). Independent Gaussian noise, with a standard deviation equal to a range of 6–7% of the signal's own standard deviation, was added to each computed target to emulate the parameter and model structural uncertainty present in real hydrological simulations, and both targets were subsequently clipped to a physically plausible 0–98% range. Critically, the imperviousness penalty term (Equation (2)) enters the target-generating function multiplicatively with the coverage effect rather than additively, which means the true relationship between imperviousness and hydrological performance depends on the level of GI coverage already present—a genuine feature interaction that a linear model using only main effects (i.e., a model without an explicit imperviousness × coverage interaction term) cannot represent, and which therefore provides a principled, non-arbitrary reason to expect the tree-based models to outperform the linear baseline, rather than a difference manufactured simply by adding unstructured noise.

This dataset-generation approach was adopted because a large ensemble of full physically-based simulation runs was not available for this study; the resulting dataset should therefore be understood as a proof-of-concept training corpus, constructed to be consistent with established hydrological relationships—including the attenuation and pollutant-removal formulations used in physically-based hydrological–LID modeling studies [3,4,14], and field-monitored bioretention hydrology and nutrient-removal performance [15,16] and green-roof runoff-quantity performance [17]—rather than as an ensemble of independently run, high-fidelity SWMM–LID simulations. This is a material limitation discussed further in Section 4. Section 3.1 reports the descriptive statistics for the resulting dataset.

2.2. Machine Learning Models and R Implementation

All modeling was carried out in R (version 4.4) [18], run on a standard desktop workstation, using a fully scripted, non-interactive workflow so that the entire pipeline—ranging from dataset generation through final evaluation—can be re-executed deterministically given the same random seed. Random forest models [19] were built using the randomForest package [20], and gradient-boosted tree models using the xgboost package [21]; the caret package [22] was used to manage the train–test split, cross-validation, and hyperparameter tuning workflow in a unified interface across both tree-based model types, and the tidyverse collection of packages [23] was used for data wrangling, feature engineering, and the visualizations reported in Section 3. The dataset was partitioned into training (80%, n = 1600) and held-out test (20%, n = 400) subsets using stratified random sampling on the storm return-period variable via caret's createDataPartition function, so that all three return periods were proportionally represented in both subsets and model evaluation was not biased by an accidental imbalance in storm severity between the training and test data. A multiple linear regression model, fit on the same training data using the same six input features with no interaction or polynomial terms, was retained as an interpretable baseline against which the two tree-based models were compared; this baseline was deliberately kept simple (main effects only) so that the comparison in Section 3.2 isolates the value added specifically by the tree-based models' capacity to learn non-linear feature interactions, rather than confounding that comparison with a more elaborately engineered linear model.

Hyperparameters for both tree-based models were tuned via 5-fold cross-validation on the training subset only (the held-out test subset was not used at any point during tuning, to avoid optimistic bias in the reported test-set performance), implemented through caret's train function with a pre-defined tuning grid. For the random forest model, the number of predictors randomly sampled as candidates at each split (mtry) was tuned over the 2–6 integer range, with the number of trees fixed at 500 following inspection of the out-of-bag error stabilization curve, beyond which additional trees produced a negligible further reduction in out-of-bag error. For the XGBoost model, a grid search was conducted over maximum tree depth (3, 4, 5, 6, 8), learning rate/eta (0.01, 0.05, 0.08, 0.1, 0.3), and number of boosting rounds (100, 200, 300, 400, 500), selecting the combination that minimized cross-validated root-mean-square error (RMSE); the selected configuration used a maximum depth of 5, a learning rate of 0.08, and 300 boosting rounds. Both tree-based models used their package defaults for all other hyperparameters (e.g., minimum node size for random forest; subsample ratio and column-sampling ratio for XGBoost), since a full joint search over every available hyperparameter was judged to offer limited additional benefit relative to the added computational cost and risk of overfitting the cross-validation folds themselves.

For the avoidance of ambiguity, storm return period (T) was retained as a single numeric predictor on its natural scale (years, $T \in \{2, 10, 100\}$) throughout data generation and model training, consistent with its use inside the continuous logarithmic term $\ln(T)$ in Equations (2)–(3); it was not one-hot encoded or treated as an unordered categorical factor by any of the three models. Accordingly, T is treated as a numerical variable (drawn from a discrete set of three values) rather than as a categorical one. As a final pre-processing step, all six input features were left on their natural physical scales (percent for imperviousness and the four coverage fractions; years for storm return period) rather than being standardized or normalized, since tree-based models are invariant to monotonic rescaling of individual features and the linear baseline's coefficients remain directly interpretable in their original physical units when features are left unscaled. No missing data occurred in the synthetic dataset by construction, so no imputation step was required.

2.3. Model Evaluation Metrics

Model performance was evaluated exclusively on the held-out test subset (never seen during training or hyperparameter tuning) using three complementary metrics: the coefficient of determination (R^2), which summarizes the proportion of test-set variance explained by the model; the root-mean-square error (RMSE), which penalizes large individual errors more heavily and is reported in the same physical units (percentage points) as the target indicator; and the mean absolute error (MAE), which is more robust to occasional large residuals and gives a more directly interpretable typical error magnitude. All three metrics were computed separately for each of the two target indicators (peak-flow attenuation and TSS event-mean-concentration reduction) and for each of the three candidate models (linear regression, random forest, XGBoost), yielding the six model–target combinations reported in Section 3.2.

2.4. Feature-Importance Analysis

For the random forest model, feature importance was quantified using the mean decrease in node impurity, here the reduction in residual sum of squares attributable to splits on each feature, as appropriate for a regression forest (this measure should not be confused with "Gini importance": Gini impurity is the node-splitting criterion used for classification forests in the R randomForest implementation, whereas regression forests, as used throughout this study, split on residual-sum-of-squares reduction, and it is this residual-sum-of-squares-based node impurity that is reported in Section 3.3), summed across all splits on a given feature across all 500 trees and normalized to sum to 100% across the six input features. For the XGBoost model, feature importance was quantified using the gain-based importance metric, representing the average improvement in the loss function attributable to splits on each feature, likewise normalized to sum to 100%. Both importance measures were computed on the final trained models (using their respective tuned hyperparameters) and are reported in Section 3.3 for the peak-flow attenuation target; the two measures are conceptually related but not numerically identical, since they are computed from different underlying tree ensembles with different structures, so a degree of disagreement between the two rankings is expected and is itself informative about how sensitive the importance ranking is to modeling choices.

2.5. Indicative Computational Efficiency Comparison

To quantify the practical screening-speed advantage of the surrogate modeling approach, indicative per-scenario runtime was compared between the physically-based hydrological–LID simulation approach (based on typical runtimes reported for single-catchment, single-storm SWMM–LID simulations of comparable scale) and the trained R-based surrogate models, timed on the same desktop workstation used for model training.

Surrogate runtime was measured as the wall-clock time for a single call to the trained model's predict function given one new scenario's six input features, averaged over 1000 repeated calls to obtain a stable estimate; this runtime excludes model training time itself, which is a one-time cost incurred once per surrogate model rather than once per screened scenario, and is therefore not part of the per-scenario efficiency comparison relevant to repeated screening use. It should be stated clearly that this is not a controlled, equivalent-hardware computational benchmark: the surrogate runtimes were measured directly by the authors, whereas the physically-based SWMM-LID runtime is not a value measured on the same workstation for an equivalent scenario, but a typical range taken from the literature for single-catchment, single-storm SWMM-LID runs of comparable scale [2,24]. A calibrated SWMM-LID model of an equivalent real catchment was not available within the scope of this proof-of-concept study, so the physically-based simulation could not be executed on the same hardware for direct, scenario-matched timing. The reported five-order-of-magnitude figure should therefore be read as an indicative, order-of-magnitude illustration of the efficiency advantage inherent to surrogate inference versus iterative physically-based simulation, not as a rigorously controlled benchmark; a same-hardware, scenario-matched comparison against an actual SWMM-LID model is identified as necessary future work in Section 4. The comparison also covers inference only: it excludes the one-time cost of running a SWMM-LID simulation ensemble to train a real surrogate and of retraining it, which would have to be counted in any assessment of net computational savings.

3. Results

3.1. Synthetic Dataset Characteristics

Table 1 summarizes descriptive statistics for the six input features and two target indicators across the full 2000-scenario synthetic dataset. Because all six input features were sampled independently and uniformly (or, for storm return period, from a fixed discrete set) across their respective ranges, the dataset is approximately balanced across the sampled parameter space, with mean coverage values for each of the four GI typologies close to the theoretical midpoint of 50%, as expected under independent uniform sampling; this balance matters for model training because it ensures that the tree-based models are not implicitly biased toward a narrow, over-represented region of the parameter space at the expense of others. The two target indicators, by contrast, are not uniformly distributed—both are concentrated toward the middle-to-upper portion of their respective 0–98% ranges, a direct consequence of the saturating exponential response function used in their construction (Equations (1)–(4)), under which many combinations of moderate-to-high coverage fractions produce broadly similar, already-substantial attenuation, while only combinations with very low aggregate coverage or very severe storms produce the small values in the lower tail.

Table 1. Descriptive statistics for the six input features and two target hydrological indicators across the full 2000-scenario synthetic training and test datasets.

Feature	Min	Mean	Max
Imperviousness (%)	30.0	55.1	80.0
Storm return period (years)	2	— (numerical, discrete set: 2, 10, 100)	100
Bioretention coverage (%)	0.0	50.1	100.0
Green roof coverage (%)	0.0	49.7	100.0
Permeable pavement coverage (%)	0.0	50.2	100.0
Vegetated swale coverage (%)	0.0	50.0	100.0
Peak-flow attenuation (%, target)	17.5	58.3	73.7
TSS EMC reduction (%, target)	34.1	82.6	92.6

3.2. Model Performance Comparison

Tables 2 and 3 summarize test-set performance for the three candidate models against the two target indicators, computed as described in Section 2.3. The dataset and modeling pipeline were executed under a fixed, documented random seed, and the corresponding code and data are publicly available (see Supplementary Materials). For peak-flow attenuation (Table 2), the linear baseline already achieves a respectable R^2 of 0.895, reflecting the fact that the underlying saturating response function, while genuinely non-linear, remains reasonably well approximated by a linear model across the specific, moderate range of the input parameter space sampled here; nonetheless, both tree-based models improve further on this baseline, with random forest reaching R^2 = 0.924 and XGBoost reaching R^2 = 0.964, the latter corresponding to a 42% reduction in RMSE relative to the linear model (from 3.10 to 1.80 percentage points). For TSS event-mean-concentration reduction (Table 3), the gap between the linear baseline and the tree-based models is substantially larger—R^2 = 0.789 for the linear model versus 0.893 for random forest and 0.939 for XGBoost—consistent with the TSS-reduction target's construction using a separate, more sharply saturating exponential function and a different, less evenly distributed set of typology weights, which together produce a relationship the linear model's main-effects-only structure captures less completely than it does for peak-flow attenuation. Across both targets, XGBoost consistently achieved marginally higher R^2 and lower error than random forest, consistent with the generally reported advantage of gradient-boosted trees over bagged trees on structured, moderate-sized tabular datasets of this kind [21]. These R^2 values must be interpreted cautiously. They were obtained on a held-out subset of the same synthetic data-generating process that produced the training data, and they therefore quantify how closely each model learned the prescribed synthetic response surface (including its added noise), not how well it would predict the performance of real green infrastructure. They are not comparable with R^2 values reported for surrogates trained on SWMM-LID outputs or monitored data.

Table 2. Test-set performance for peak-flow attenuation prediction across the three candidate models. Values refer to the held-out synthetic test set (n = 400) and describe recovery of the synthetic response surface, not performance on real GI systems.

Model	R^2 (Attenuation)	RMSE (Attenuation, %)	MAE (Attenuation, %)
Linear regression	0.895	3.10	2.17
Random forest	0.924	2.63	1.84
XGBoost	0.964	1.80	1.25

Table 3. Test-set performance for total-suspended-solid (TSS) event-mean-concentration reduction prediction across the three candidate models. Values refer to the held-out synthetic test set (n = 400) and describe recovery of the synthetic response surface, not performance on real GI systems.

Model	R^2 (TSS Reduction)	RMSE (TSS, %)	MAE (TSS, %)
Linear regression	0.789	3.84	2.59
Random forest	0.893	2.73	1.73
XGBoost	0.939	2.06	1.25

Figure 2 visualizes this performance directly as scatter plots of predicted versus observed peak-flow attenuation on the held-out test set, for random forest (left panel) and XGBoost (right panel), with the diagonal dashed line marking perfect prediction. Both panels show test-set points clustered tightly around the 1:1 line across the full observed range of roughly 25–75% attenuation, with visibly less scatter in the XGBoost panel, consistent with its higher reported R^2 and lower RMSE in Table 2; neither model shows a

systematic tendency to over- or under-predict at either the low or high end of the observed range, which would have suggested a specific structural mis-specification (for example, insufficient flexibility to capture the saturating upper tail of the response function) rather than simple residual noise.

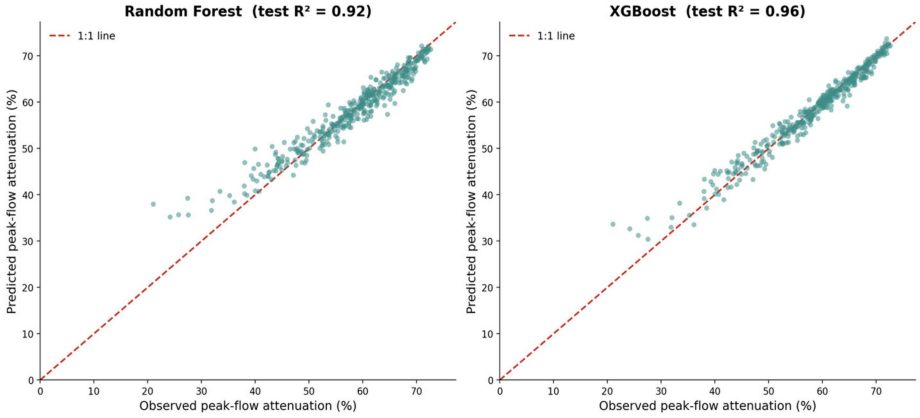


Figure 2. Predicted versus observed peak-flow attenuation on the held-out test set (n = 400), for random forest (**left**, test R^2 = 0.92) and XGBoost (**right**, test R^2 = 0.96). The dashed red line marks perfect 1:1 agreement between predicted and observed values. Results are for the synthetic dataset only.

3.3. Feature Importance

Table 4 and Figure 3 report feature importance for the peak-flow attenuation target from both tree-based models. Permeable pavement coverage and bioretention coverage rank as the two most important individual features for random forest (28.1% and 23.2%, respectively). For XGBoost, storm return period is the single most important predictor (31.0% gain-based importance), with permeable pavement and bioretention next (22.8% and 18.5%). Permeable pavement and bioretention together account for just over half of total importance in the random forest model (51.3%) and just over 41% in XGBoost, consistent with their higher unit-coverage attenuation weights in the underlying data-generating function (Equation (1)) relative to green roofs and vegetated swales. Baseline imperviousness ranks third for random forest (18.3%) and fourth for XGBoost (11.6%), while storm return period shows the sharpest divergence between the two models: storm return period ranks fourth for random forest (11.3%) but first for XGBoost (31.0%)—a genuine disagreement between the two models' learned importance rankings for storm return period and imperviousness, rather than a reporting artifact, which most plausibly reflects the different ways bagged trees (random forest) and sequentially fit boosted trees (XGBoost) partition credit between correlated or interacting predictors during training, since the two features enter the true data-generating process multiplicatively rather than independently (Equations (2)–(4)) and can therefore substitute for one another to some extent in explaining a given prediction. Green roof coverage ranks last in both models (9.2% and 7.7%), consistent with its lowest unit-coverage weight (0.22) in the underlying data-generating function among the four typologies.

Table 4. Feature importance for the peak-flow attenuation target, from random forest (node-impurity importance, based on residual-sum-of-squares reduction) and XGBoost (gain-based importance), expressed as a percentage of total importance across the six input features.

Feature	RF Importance (%)	XGBoost Importance (%)
Permeable pavement coverage	28.1	22.8
Bioretention coverage	23.2	18.5
Baseline imperviousness	18.3	11.6

Table 4. *Cont.*

Feature	RF Importance (%)	XGBoost Importance (%)
Storm return period	11.3	31.0
Vegetated swale coverage	10.0	8.4
Green roof coverage	9.2	7.7

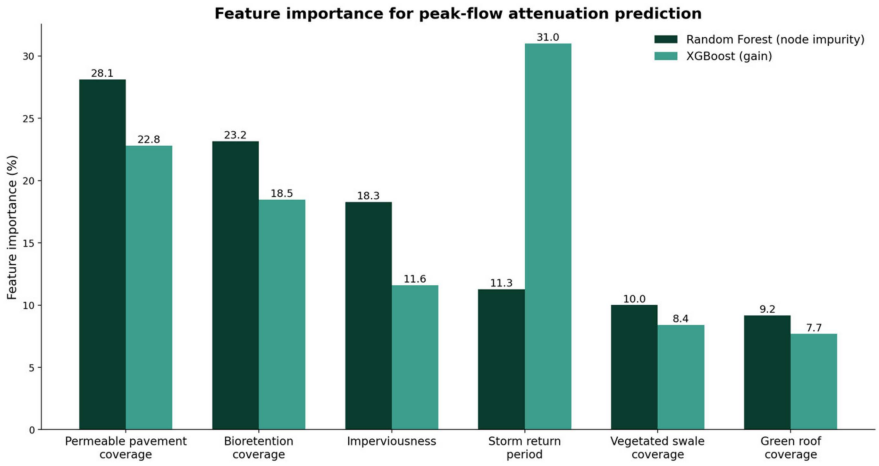


Figure 3. Feature importance for peak-flow attenuation prediction, comparing random forest (node-impurity importance) and XGBoost (gain-based importance) across all six input features, ranked by random forest importance.

Taken together, the pattern in Table 4 and Figure 3 is broadly consistent with the storm-magnitude sensitivity reported in physically-based GI modeling studies, where performance is known to vary systematically with return period [13]. This agreement should not, however, be presented as independent external validation of the surrogate models. Permeable pavement and bioretention were explicitly assigned the highest unit-coverage weights in the target-generating Equation (1) (Section 2.1), so their consistently strong ranking among the top predictors in both models largely reflects information that was directly embedded in the synthetic dataset by construction, rather than a hydrological relationship the models discovered independently. The correct reading of Table 4 and Figure 3 is therefore methodological: it shows that both tree-based models correctly recover the relative typology weighting that was built into the data-generating function, which is a necessary sanity check on model fitting and feature-importance computation, not evidence that the models have learned something about GI performance beyond what was specified in Equations (1)–(4).

3.4. Computational Efficiency

Table 5 compares indicative per-scenario runtime between the physically-based hydrological–LID simulation approach and the trained R-based surrogate models, measured as described in Section 2.5. Once trained, both surrogate models return a prediction in a few milliseconds (directly measured on the workstation used for model training; approximately 4 ms for random forest and 2 ms for XGBoost, the latter benefiting from its typically shallower, sequentially optimized tree structure), an indicative ratio of roughly five orders of magnitude relative to the typical literature runtime for a single catchment–storm–typology SWMM-LID configuration (not measured in this study). As explained in Section 2.5, this figure is an indicative, order-of-magnitude comparison rather than a controlled, same-hardware benchmark, since the physically-based simulation time is taken from typical literature values rather than measured for an equivalent scenario on

the same workstation. The comparison also excludes the one-time cost of generating a SWMM-LID training ensemble for a real application, so that net savings arise only when a trained surrogate is reused across many screening queries. This efficiency gain is the central practical motivation for the surrogate modeling approach: it converts a screening task that would be computationally prohibitive to perform exhaustively across a full city-scale parameter space—for example, evaluating several thousand combinations of catchment, storm, and typology-coverage configurations across a city's sub-catchments, which would require, on the basis of the literature runtimes assumed here, on the order of days to weeks of continuous physically-based simulation—into one that can be completed for the same number of candidate configurations in well under a second of surrogate-model inference time, leaving the physically-based simulation time budget free to be concentrated entirely on detailed evaluation of the small shortlist of configurations the surrogate screening identifies as most promising.

Table 5. Indicative per-scenario computational runtime for the physically-based simulation approach versus the trained R-based surrogate models. The physically-based runtime is a typical literature range and was not measured in this study, whereas the surrogate runtimes were measured directly; the comparison is therefore indicative only and is not a controlled, same-hardware benchmark.

Approach	Runtime per Scenario	Indicative Ratio
Physically-based hydrological–LID simulation	~12–18 min (literature-typical; not measured)	1 (baseline)
Random forest surrogate (R)	~4 ms (measured)	~2×10^5
XGBoost surrogate (R)	~2 ms (measured)	~4×10^5

4. Discussion

The results indicate that tree-based machine learning models, implemented entirely within the open-source R environment, can recover the prescribed synthetic relationship between catchment, storm, and GI typology-coverage features and the two synthetic performance indicators with high fidelity (R^2 = 0.89–0.96 across both targets and both tree-based models), and that this learned relationship, once trained, can be evaluated in milliseconds. These statements refer to the synthetic response surface only: they do not establish that the models can predict the performance of real green infrastructure, nor that they approximate SWMM-LID outputs. The gap between the linear baseline and the tree-based models was larger for TSS reduction (R^2 improving from 0.789 to 0.939, a relative RMSE reduction of nearly half) than for peak-flow attenuation (R^2 improving from 0.895 to 0.964), which is consistent with the two targets having been constructed from data-generating functions with different degrees of curvature and different typology-weighting schemes (Section 2.1); this difference is itself a useful methodological observation, since it shows that the relative advantage of tree-based over linear modeling cannot be assumed to be constant across different hydrological performance indicators and should be checked separately for each target of interest rather than generalized from a single indicator. The dominance of permeable-pavement and bioretention coverage in the feature-importance results (Table 4, Figure 3) reproduces the relative typology weighting that was explicitly built into the target-generating Equation (1) (Section 2.1), rather than constituting independent confirmation of a physical relationship; as discussed in Section 3.3, this result should be read as a check that model fitting and feature-importance computation behaved as expected, not as external validation. What the computational exercise genuinely demonstrates is narrower and more concrete: that the R-based implementation of random forest and XGBoost can be fitted, tuned, and evaluated correctly on a moderately sized, non-linear, interacting tabular dataset; that the two tree-based algorithms behave comparably but not identically, both in predictive accuracy and in how they distribute importance across

correlated or interacting features; and that a substantial body of hydrological interpretation, model retraining, and external validation would be required before these comparative algorithm behaviors could be read as statements about real GI hydrological performance. At the same time, the notable disagreement between random forest and XGBoost on the relative importance of storm return period versus imperviousness (Section 3.3) is a useful reminder that feature-importance rankings from a single tree-based model should not be over-interpreted as a definitive statement about the true underlying physical relationship, particularly where two features enter that relationship through a multiplicative interaction and can partially substitute for one another during tree construction.

Interpretation of the reported R^2 values. The R^2, RMSE, and MAE values in Tables 2 and 3 should be interpreted cautiously and in a strictly methodological sense. Because the test subset is drawn from the same synthetic data-generating process as the training subset, high R^2 values show that the models learned the synthetic response surface (Equations (1)–(4) and (A1)–(A3), plus noise). They are not estimates of predictive accuracy for real green infrastructure performance, where the input–output relationship is unknown, site-specific, and influenced by many factors that are not represented here. A value such as $R^2 = 0.96$ must therefore not be read as evidence that the surrogates would reproduce SWMM-LID outputs or monitored performance to a comparable degree.

Article type and contribution. For these reasons, the present work is best classified as a methodological proof of concept (technical note) rather than a full hydrological validation study. Its contribution is limited to (i) an open, reproducible R workflow; (ii) a like-for-like comparison of linear regression, random forest, and XGBoost on a known non-linear, interacting response surface; and (iii) an explicit statement of the steps required to turn the workflow into a validated screening tool. It does not provide new hydrological knowledge about the performance of any GI typology.

Computational-efficiency comparison. The efficiency comparison should be read with the same caution. The surrogate runtimes were measured directly, whereas the SWMM-LID runtime is a typical literature range (Section 2.5); hardware, model configuration, catchment size, and time step were therefore not controlled, and the reported ratio of roughly 10^5 is an order-of-magnitude illustration of the difference between surrogate inference and iterative simulation, not a benchmark. The comparison also omits the one-time cost of generating the SWMM-LID ensemble needed to train a real surrogate: for example, an ensemble of 10,000 simulations, the size used in [10], at 12–18 min each would amount to a range of roughly 2000–3000 h of single-core simulation time before parallelization, so net savings materialize only when the trained surrogate is reused across many screening queries.

Synthetic training data are the framework's most significant limitation and are worth dwelling on in some detail rather than noting only in passing. The dataset used here was generated from parameterized analytical relationships with added noise, rather than from an ensemble of independently run, high-fidelity physically-based simulations or from field-monitored performance records. This choice was made because such an ensemble was not available for this proof-of-concept study, but it has a specific and important consequence for how the reported results should be read: the R^2, RMSE, and MAE values in Tables 2 and 3 describe how well the surrogate models recover the analytical relationships (Equations (1)–(4)) used to generate their own training data, not how well they would predict real, monitored GI performance in an actual city. A surrogate model can, in principle, achieve a near-perfect fit to a synthetic data-generating function while still performing poorly against real-world data, if the analytical function itself is a poor representation of true catchment hydrology—a risk that echoes the uncertainty that has been reported directly for SWMM-based LID modeling itself, where different plausible parameterizations of the same physical catchment can produce materially different runoff predictions [24].

Before any operational deployment for municipal decision-making, the surrogate modeling approach demonstrated here should therefore be retrained and revalidated against either a genuine ensemble of physically-based hydrological–LID simulation runs spanning a realistic range of catchments and climates—including the kind of urbanization- and climate-change-sensitive conditions surveyed in broader reviews of sustainable urban drainage systems [25]—or against field-monitored GI performance records where these become available, consistent with the broader observation that GI performance modeling evidence remains uneven across the literature [26]. Until that revalidation is carried out, the appropriate reading of this study's numerical results is methodological rather than predictive: they demonstrate that the proposed R-based workflow is capable of learning a complex, non-linear, interacting hydrological relationship accurately and efficiently when such a relationship exists in the training data, not that the specific attenuation and pollutant-removal values reported here apply to any particular real catchment.

This proof-of-concept demonstration should be benchmarked explicitly against the growing body of ML surrogate studies trained directly on physically-based simulation or monitored data, rather than judged in isolation. Bao et al. generated 10,000 SWMM-based samples from a calibrated and validated model and trained ML surrogates to predict total and peak catchment outflows for LID sizing and optimization, illustrating the preferred workflow of calibrated SWMM → scenario ensemble → surrogate training → independent validation → optimization/screening that the present study has not yet completed [10]. Kim et al. combined SWMM with an LSTM network to evaluate GI stormwater-runoff reduction for a monitored catchment in Seocho-gu, Seoul, explicitly addressing computational acceleration of SWMM-based screening under different rainfall conditions, and represent the kind of real study area, SWMM-based simulation ensemble, and rainfall-scenario evaluation that would be required before operational claims could be made from the present framework [11]. Against this benchmark, the R^2 = 0.89–0.96 values reported here demonstrate only that the chosen R implementation can fit a known synthetic function accurately; they are not evidence of comparable predictive skill against SWMM-LID outputs or field observations, and the present study's contribution should be judged as a transparent, reproducible workflow rather than as a competing surrogate to these SWMM-trained approaches.

A second limitation concerns extrapolation: the surrogate model's predictions are reliable only within the range of catchment imperviousness (30–80%), storm return period ({2, 10, 100} years), and typology-coverage values (0–100%) represented in its training data, and should not be trusted for scenarios that fall substantially outside that range—a general caveat for tree-based ML models, which partition their input space based on the training data they observe and do not extrapolate smoothly beyond the bounds of those data in the way a parametric physical model, whose functional form is specified independently of any particular training sample, is able to. A municipality with catchments more than 80% impervious, or interested in intermediate storm return periods (such as 5- or 25-year events) not represented in the discrete {2, 10, 100} set used here, would need to either retrain the surrogate on an expanded sampling range or treat its predictions for such scenarios with appropriate caution. A third consideration is interpretability: while the node-impurity- and gain-based feature-importance measures reported in Section 3.3 provide a useful first-order, model-wide account of which features drive predictions on average, they do not explain any individual prediction in the way a full physically-based simulation trace does—a municipal engineer reviewing a specific, unusual-looking surrogate prediction for one candidate site has no equivalent of the step-by-step hydraulic reasoning a SWMM output report would provide, and may reasonably want more granular, case-specific explanation before acting on that prediction. Future extensions of this framework could incorporate

SHAP (SHapley Additive exPlanations) values [27] for instance-level interpretability, which would decompose each individual prediction into the specific contribution of each input feature for that scenario rather than reporting only model-wide average importance; could integrate with sensor-actuated real-time control of installed GI assets to extend the surrogate's usefulness beyond initial siting into ongoing operational management [28]; and could couple the R-based surrogate with a geographic information system to support spatially explicit, city-scale screening directly from parcel- or sub-catchment-level input data, consistent with watershed-scale evidence that composite, multi-typology deployment outperforms isolated interventions [29].

Path to real application. Moving from this proof of concept to a screening tool requires, at a minimum: (i) a calibrated and validated SWMM-LID model for the target catchment or catchment class; (ii) a designed simulation ensemble spanning the relevant catchment, storm (including hyetograph shape and antecedent conditions), soil, and LID-design variables; (iii) retraining and re-tuning of the surrogate on that ensemble, or on field-monitored performance records; (iv) validation on held-out simulations and, where possible, independent monitored events, with performance and uncertainty reported separately from training results; and (v) a same-hardware, scenario-matched runtime comparison that also accounts for the cost of generating the ensemble.

Used appropriately—as a rapid first-pass screening tool that narrows a large candidate space down to a shortlist for detailed physically-based modeling, rather than as a substitute for that detailed modeling—the R-based workflow developed here could, once retrained and validated on SWMM-LID simulation ensembles or field-monitored data, serve as a practical, fully open source complement to existing GI impact-modeling workflows; in its present synthetic-data form, it demonstrates a method and is not a screening tool, consistent with the broader case for machine learning as a complement to, rather than a replacement for, physically-based hydrological science [6]. This complementary framing also extends naturally to the wider multi-benefit case for GI as a climate-adaptation strategy [30] and to the economic screening question that ultimately determines whether a shortlisted site proceeds to construction [31]: a fast hydrological pre-screen of the kind this workflow could support, once validated, could help ensure that the more time-consuming detailed modeling and cost–benefit analysis stages, which remain the appropriate final word before any real capital commitment, are applied only to the candidate configurations most likely to justify that investment, rather than being spread thinly and inefficiently across every combination a municipality might in principle wish to consider.

5. Conclusions

This article presented a methodological proof of concept, in the form of an R-based machine learning workflow, for emulating a synthetic green infrastructure hydrological performance response surface defined by catchment, storm, and typology-coverage features. Random forest and XGBoost models, trained on a synthetic dataset of 2000 catchment–storm–typology scenarios generated from parameterized, non-linear hydrological relationships with a deliberately embedded imperviousness × coverage interaction term, achieved R^2 values in the range of 0.89–0.96 on the held-out synthetic test set across peak-flow attenuation and total-suspended-solids removal targets, consistently outperforming a linear regression baseline (R^2 = 0.79–0.90), and returned predictions in milliseconds, compared with tens of minutes per scenario typically reported in the literature for the SWMM-LID simulation. This last comparison is indicative only, because the surrogate runtime was measured whereas the SWMM-LID runtime was not. The R^2 values show that the models learned the synthetic response surface; they do not demonstrate the ability to predict real green infrastructure performance. Feature-importance analysis identified permeable-pavement

and bioretention coverage as the two dominant predictors of peak-flow attenuation, with storm return period and baseline imperviousness contributing meaningfully but with some disagreement in relative ranking between the two tree-based models, a ranking that largely reflects the typology weights embedded in the data-generating equations and therefore serves as a consistency check on model fitting, not as independent validation.

These results support five conclusions, all of which apply within the scope of a methodological proof of concept. First, tree-based machine learning models implemented in R can reproduce the non-linear synthetic GI performance relationships considered in this proof-of-concept study, with high fidelity to that synthetic response surface and negligible marginal inference cost once trained. This is a necessary but not sufficient condition for later use in rapid screening; whether such models can serve that purpose for real catchments can only be established after validation against real hydrological data. Second, the size of the advantage tree-based models offer over a simple linear baseline is not fixed but depends on the specific target indicator and the degree of non-linearity and feature interaction present in its true underlying relationship, as shown by the materially larger linear-to-tree-based performance gap observed for TSS reduction than for peak-flow attenuation in this study; this indicates that the choice of modeling approach for any new target indicator should be justified empirically for that indicator specifically, rather than assumed by analogy from a different, previously studied indicator. Third, the synthetic nature of the training data is the principal limitation of this study and restricts the reported numerical results to a methodological, rather than a predictive, interpretation. The reported R^2 values describe how well the models learned the synthetic response surface, not how well they would predict real green infrastructure performance. Real application therefore requires retraining and validation of the surrogates using SWMM-LID simulation ensembles or field-monitored performance data; until this has been done, none of the attenuation or pollutant-removal figures reported here should be relied upon for any real site or planning decision. Fourth, surrogate modeling of this kind is best understood as a complement to, rather than a substitute for, detailed physically-based hydrological–LID modeling—a rapid first-pass filter that narrows a large candidate space before detailed engineering evaluation of a shortlist, freeing the more computationally expensive physically-based modeling and cost–benefit analysis stages to be concentrated where they matter most. Fifth, the computational-efficiency comparison is indicative only: the surrogate runtime was measured, whereas the SWMM-LID runtime was taken from typical literature values, so the ratio of roughly 10^5 is an order-of-magnitude illustration and not a controlled benchmark; it also excludes the cost of generating a real SWMM-LID training ensemble.

Future work should therefore give first priority to retraining and validating the surrogates on SWMM-LID simulation ensembles from calibrated models, or on field-monitored performance data, accompanied by a same-hardware, scenario-matched runtime comparison. Further extensions, namely instance-level explainability using SHAP values, an enriched physically-based feature set, and integration with geographic information systems for spatially explicit, parcel-level screening, would then strengthen the practical applicability of the workflow to municipal green infrastructure planning and would represent the natural next steps in moving this proof-of-concept methodology toward a decision-support tool that could be defended for use in a real capital-planning process.

Supplementary Materials: The full 2000-scenario synthetic dataset and all R scripts used in this article (dataset generation; model training and tuning for random forest and XGBoost; computation of R^2/RMSE/MAE; and generation of all reported figures and tables) are openly and permanently available in the public Zenodo repository at https://doi.org/10.5281/zenodo.22794460.

Author Contributions: Conceptualization, R.A. and Š.S.; methodology, R.A.; software, R.A.; validation, R.A., Š.S., D.B., J.I. and I.Š.; formal analysis, R.A.; investigation, R.A.; resources, Š.S.; data curation, R.A.; writing—original draft preparation, R.A.; writing—review and editing, R.A., Š.S., D.B., J.I. and I.Š.; visualization, R.A.; supervision, Š.S.; project administration, Š.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Young Researchers Support Program of the Slovak University of Technology in Bratislava under the project "Use of QGIS for the Calculation of the Blue-Green Factor and the Evaluation of Environmental Qualities of Urbanized Areas" (project No. 1648). This work was also supported by the Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences (VEGA) under project No. 1/0682/23 and by the Slovak Research and Development Agency (APVV) under project No. VV-MVP-24-0180.

Data Availability Statement: The synthetic dataset and R code supporting the reported results are openly and permanently available in the public Zenodo repository at https://doi.org/10.5281/zenodo.22794460. This study collected no new data from human participants or animals.

Acknowledgments: This article is based on a part of the first author's doctoral dissertation, entitled "Green Infrastructure Impact Modelling on the Water Management in Urban Areas," conducted at the Department of Sanitary and Environmental Engineering, Slovak University of Technology in Bratislava. The authors gratefully acknowledge the support of the Department of Sanitary and Environmental Engineering and the Slovak University of Technology in Bratislava in the preparation of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GI	Green Infrastructure
LID	Low-Impact Development
TSS	Total Suspended Solids
EMC	Event-Mean Concentration
CV	Cross-Validation
RMSE	Root-Mean-Square Error
MAE	Mean Absolute Error
SHAP	SHapley Additive exPlanations
ML	Machine Learning
RF	Random Forest
SWMM	Storm Water Management Model
LSTM	Long Short-Term Memory

Appendix A

This appendix gives the full analytical expression used to generate the total-suspended-solid (TSS) event-mean-concentration reduction target described in Section 2.1, in the same form as the peak-flow attenuation expression given in the main text as Equations (1)–(4). The aggregate coverage effect for TSS uses pollutant-specific typology weights, reflecting the relative particulate-filtration performance of each typology (bioretention: 0.78, permeable pavement: 0.71, vegetated swales: 0.58, green roofs: 0.55):

$$E_{TSS} = (0.78 \cdot C_{bio} + 0.55 \cdot C_{gr} + 0.71 \cdot C_{pp} + 0.58 \cdot C_{sw})/100 \quad \text{(A1)}$$

A separately calibrated, milder storm-severity term is used for the TSS target, reflecting the comparatively weaker sensitivity of particulate filtration performance to storm return period relative to peak-flow attenuation:

$$F_{\text{storm,TSS}} = 1.10 - 0.045 \cdot \ln(T) \quad \text{(A2)}$$

The TSS reduction target is then computed from the same saturating exponential form used for peak-flow attenuation, using the TSS-specific aggregate coverage effect and storm-severity term together with the imperviousness penalty term already defined in the main text as Equation (2):

$$R_{\text{TSS}} = 92 \cdot [1 - \exp(-1.95 \cdot E_{\text{TSS}} \cdot P_{\text{imp}} \cdot F_{\text{storm,TSS}})] \quad \text{(A3)}$$

As with the peak-flow attenuation target, independent Gaussian noise (6–7% of the signal's own standard deviation) was added to R_{TSS} prior to clipping to the physically plausible 0–98% range described in Section 2.1.

References

1. Fletcher, T.D.; Shuster, W.; Hunt, W.F.; Ashley, R.; Butler, D.; Arthur, S.; Trowsdale, S.; Barraud, S.; Semadeni-Davies, A.; Bertrand-Krajewski, J.-L.; et al. SUDS, LID, BMPs, WSUD and more—The evolution and application of terminology surrounding urban drainage. *Urban Water J.* **2015**, *12*, 525–542. [CrossRef]
2. Rossman, L.A. *Storm Water Management Model User's Manual, Version 5.1*; U.S. Environmental Protection Agency: Cincinnati, OH, USA, 2015. Available online: https://www.epa.gov/water-research/storm-water-management-model-swmm (accessed on 27 August 2026).
3. Dietz, M.E. Low impact development practices: A review of current research and recommendations for future directions. *Water Air Soil Pollut.* **2007**, *186*, 351–363. [CrossRef]
4. Ahiablame, L.M.; Engel, B.A.; Chaubey, I. Effectiveness of low impact development practices: Literature review and suggestions for future research. *Water Air Soil Pollut.* **2012**, *223*, 4253–4273. [CrossRef]
5. Jia, H.; Yao, H.; Tang, Y.; Yu, S.L.; Field, R.; Tafuri, A.N. LID-BMPs planning for urban runoff control and the case study in China. *J. Environ. Manag.* **2015**, *149*, 65–76. [CrossRef]
6. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [CrossRef]
7. Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; De Marinis, G. Machine learning algorithms for the forecasting of wastewater quality indicators. *Water* **2017**, *9*, 105. [CrossRef]
8. Sun, A.Y.; Scanlon, B.R. How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett.* **2019**, *14*, 073001. [CrossRef]
9. Xu, T.; Liang, F. Machine learning for hydrologic sciences: An introductory overview. *WIREs Water* **2021**, *8*, e1533. [CrossRef]
10. Bao, Z.; Bain, D.J.; Musharrat, S. Surrogate-based multiobjective optimization to rapidly size low impact development practices for outflow capture. *J. Hydrol.* **2022**, *616*, 128848. [CrossRef]
11. Kim, D.; Kim, D.; Chung, G. Evaluating the stormwater runoff reduction effect of green infrastructure using a SWMM-based LSTM model: A case study of Seocho-gu, Seoul. *Urban Clim.* **2025**, *64*, 102692. [CrossRef]
12. Silva, F.; Calheiros, C.S.C.; Valle, G.; Pinto, P.; Albuquerque, A.; Antão-Geraldes, A.M. Influence of green roofs on the design of a public stormwater drainage system: A case study. *Sustainability* **2023**, *15*, 5762. [CrossRef]
13. Qin, H.; Li, Z.; Fu, G. The effects of low impact development on urban flooding under different rainfall characteristics. *J. Environ. Manag.* **2013**, *129*, 577–585. [CrossRef]
14. Zhang, S.; Guo, Y. SWMM simulation of the storm water volume control performance of permeable pavement systems. *J. Hydrol. Eng.* **2015**, *20*, 06014010. [CrossRef]
15. Davis, A.P. Field performance of bioretention: Hydrology impacts. *J. Hydrol. Eng.* **2008**, *13*, 90–95. [CrossRef]
16. Hunt, W.F.; Jarrett, A.R.; Smith, J.T.; Sharkey, L.J. Evaluating bioretention hydrology and nutrient removal at three field sites in North Carolina. *J. Irrig. Drain. Eng.* **2006**, *132*, 600–608. [CrossRef]
17. Berndtsson, J.C. Green roof performance towards management of runoff water quantity and quality: A review. *Ecol. Eng.* **2010**, *36*, 351–360. [CrossRef]
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: https://www.R-project.org/ (accessed on 27 August 2026).

19. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
20. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
21. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
22. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
23. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]
24. Zhang, Z.; Hu, W.; Wang, W.; Zhou, J.; Liu, D.; Qi, X.; Zhao, X. The hydrological effect and uncertainty assessment by runoff indicators based on SWMM for various LID facilities. *J. Hydrol.* **2022**, *613*, 128418. [CrossRef]
25. Zhou, Q. A review of sustainable urban drainage systems considering the climate change and urbanization impacts. *Water* **2014**, *6*, 976–992. [CrossRef]
26. Eckart, K.; McPhee, Z.; Bolisetti, T. Performance and implementation of low impact development—A review. *Sci. Total Environ.* **2017**, *607–608*, 413–432. [CrossRef]
27. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
28. Kerkez, B.; Gruden, C.; Lewis, M.; Montestruque, L.; Quigley, M.; Wong, B.; Bedig, A.; Kertesz, R.; Braun, T.; Cadwalader, O.; et al. Smarter stormwater systems. *Environ. Sci. Technol.* **2016**, *50*, 7267–7273. [CrossRef]
29. Ahiablame, L.; Shakya, R. Modeling flood reduction effects of low impact development at a watershed scale. *J. Environ. Manag.* **2016**, *171*, 81–91. [CrossRef]
30. Almaaitah, T.; Appleby, M.; Rosenblat, H.; Drake, J.; Joksimovic, D. The potential of blue-green infrastructure as a climate change adaptation strategy: A systematic literature review. *Blue-Green Syst.* **2021**, *3*, 223–248. [CrossRef]
31. Liu, W.; Chen, W.; Feng, Q.; Peng, C.; Kang, P. Cost-benefit analysis of green infrastructures on community stormwater reduction and utilization: A case of Beijing, China. *Environ. Manag.* **2016**, *58*, 1015–1026. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.