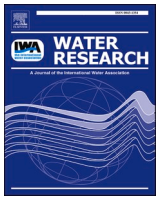
Water Research 304 (2026) 126379

Contents lists available at ScienceDirect

Water Research

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



Non-point source pollution prediction and dynamics simulation in urban runoff: a physics-informed neural network approach

Sijie Tang [a,b], Jiping Jiang [a,c,*], Shuo Wang [b,d], Yi Zheng [a,e,f], Dragan Savic [g,h], Aijie Wang [c]

[a] School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China
[b] State Key Laboratory of Climate Resilience for Coastal Cities, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, 999077, China
[c] State Key Laboratory of Urban-Rural Water Resource and Environment, Shenzhen Key Laboratory of Organic Pollution Prevention and Control, School of Eco-Environment, Harbin Institute of Technology, Shenzhen, 518055, China
[d] Research Institute for Land and Space, The Hong Kong Polytechnic University, Hong Kong, 999077, China
[e] Shenzhen Municipal Engineering Lab of Environmental IoT Technologies, Southern University of Science and Technology, Shenzhen, 518055, China
[f] State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, Southern University of Science and Technology, Shenzhen, 518055, China
[g] KWR Water Research Institute, Nieuwegein, 3430 BB, the Netherlands
[h] Centre for Water Systems, Department of Engineering, University of Exeter, Exeter, EX4 4QF, United Kingdom

HIGHLIGHTS

- Proposes PIWON integrating wash-off ODEs into RNN architecture.
- Simulates continuous pollutographs from event-based tabular data.
- Achieves NSE 0.65 and GA 0.94, outperforming five ML baselines.
- Identifies nonlinear drivers; rainfall dilutes EMC but intensifies first flush.
- Enables shift from volume-based to dynamic quality-based stormwater control.

ARTICLE INFO

Dataset link: National Stormwater Quality Database (Original data)

Keywords:
Physics-informed neural networks
Non-point source
Total suspended solids
First flush intensity
Stormwater management

ABSTRACT

Urban non-point source (NPS) pollution poses a significant threat to water environments, yet modeling its complex dynamics remains constrained by the trade-off between the extensive data requirements of process-based models and the limited interpretability of machine learning approaches. This study introduces the physics-informed wash-off network, a hybrid architecture that embeds the differential equations governing pollutant accumulation and wash-off into a recurrent neural network. Leveraging a tabular event dataset, the model generates continuous pollutographs, achieving improved predictive performance with a Nash-Sutcliffe Efficiency of 0.65 and generalization score of 0.94 compared to five state-of-the-art data-driven baselines, where the highest values were 0.33 and 0.89, respectively. Beyond prediction, the model employs interpretability analysis to identify the non-linear drivers of total suspended solids dynamics. Results reveal a distinct divergence: while land use and imperviousness consistently drive both event mean concentration and first flush intensity, precipitation oppositely affects them. Specifically, heavier rainfall dilutes average concentrations but intensifies the first flush. This opposing relationship explains the negative correlation observed between the two metrics and highlights the limitations of uniform stormwater regulations. Consequently, we propose a differentiated management framework: catchments with high first flush potential are strong candidates for rapid diversion and separation technologies, whereas those with low first flush potential are better suited for volume-based retention strategies. These findings advocate for a paradigm shift from static, volume-based controls to dynamic, quality-based management.

* Corresponding author.
E-mail address: jjp_lab@sina.com (J. Jiang).

https://doi.org/10.1016/j.watres.2026.126379
Received 7 April 2026; Received in revised form 4 June 2026; Accepted 27 June 2026
Available online 28 June 2026

1. Introduction

As traditional point-source emissions become increasingly controlled, non-point source (NPS) pollution has emerged as the primary threat to urban water environments (Gunawardena et al., 2014; Masoner et al., 2019; Risch et al., 2018). The rapid expansion of impervious surfaces and intensive anthropogenic activities have fundamentally altered urban hydrology, increasing direct runoff volumes and degrading water quality with accumulated particulates like total suspended solids (TSS) (Charters et al., 2016; Novotny, 1999; Shao et al., 2020). Consequently, understanding and dynamically mitigating urban NPS pollution is now a critical priority in modern stormwater management and urban renewal (Nguyen et al., 2019; Wang et al., 2025).

Current modeling approaches to simulate NPS accumulation and wash-off face a fundamental trade-off between physical fidelity and data dependency (Launay et al., 2016; Yang et al., 2016). Traditional process-based models—such as SWMM, SWAT, and the MIKE series (Li et al., 2017)—are grounded in conservation laws and theoretically capable of simulating full pollutographs. However, their practical application is heavily constrained by the need for extensive, site-specific calibration data, frequently leading to over-parameterization and equifinality in unmonitored catchments (Xiong et al., 2022). Conversely, the recent paradigm shift toward machine learning (ML) and deep learning (Shen, 2018) offers high computational efficiency and robust predictive performance (Behrouz et al., 2022; Zuo et al., 2025). Yet, these purely data-driven approaches act as "black boxes" (Sun and Scanlon, 2019). Lacking physical consistency, they are prone to learning spurious correlations and typically fail to simulate interpretable water quality dynamics when extrapolating beyond their training domain.

Although post hoc methods explain black-box model outputs (Erion et al., 2021; Samek et al., 2019; Sundararajan et al., 2017), their paradigm remains end-to-end, prioritizing label prediction over interpretable water quality dynamics. Furthermore, given that the vast majority of available datasets report only aggregated metrics like Event Mean Concentration (EMC), regression-based empirical models predominate; consequently, these models often discard the temporal evolution of storm events. Physics-Informed Neural Networks (PINNs) bridge the gap between these paradigms by leveraging the mathematical equivalence between the discrete update steps of a recurrent network and the numerical integration of a differential equation (Niu et al., 2019). PINNs have proven effective for solving inverse problems across diverse fields, including nano-optics (Chen et al., 2020), topology optimization (Zhang et al., 2022), multiphase porous media flow (Abbasi and Andersen, 2024; Serebrennikova et al., 2022), and high-speed fluid flow (Jagtap et al., 2022; Raissi et al., 2020). Building on this potential, recent applications in watershed-scale runoff modeling have achieved notable success (Jiang et al., 2020).

To address this dichotomy, this study proposes the Physics-Informed Wash-Off Network (PIWON). By integrating recurrent neural networks (RNNs) with the differential equations that govern conventional build-up and wash-off processes, the model captures the accumulation and transport dynamics of urban NPS pollution in rainfall runoff. The broader scientific significance of this hybrid architecture resides in its capacity to ensure physical consistency within a deep learning framework, thereby mitigating the risk of the network learning spurious noise. PIWON introduces a novel modeling paradigm that not only predicts TSS loads but also extracts interpretable, continuous water quality dynamics directly from highly aggregated and sparse tabular data. Specifically, this study addresses three pivotal questions: 1) How does this hybrid modeling approach compare with state-of-the-art methods? 2) Which non-linear factors exert the most substantial influence on pollution loads and the first flush effect? and 3) In what ways can this physics-informed framework revolutionize the design and management of stormwater control measures?

2. Model development, method and data

2.1. Physics-informed wash-off networks

The feasibility of applying PINNs to time-dependent processes such as pollutant wash-off is supported by the conceptual similarity between recurrent neural networks (RNNs) and ordinary differential equations (ODEs). RNNs, as well as some residual networks and normalizing flows, can be viewed as a discrete approximation of a continuous-time dynamical system, where the hidden state evolves according to a difference equation as shown below:

$$S_{t+1} = h_t + f_\theta(h_t, x_t)\Delta t \tag{1}$$

where S_t represents the hidden state at time t, x_t represents the input at time t, θ represents the model parameters. Previous research has proven that these iterative updates can be seen as a Euler discretization of a continuous transformation (Haber and Ruthotto, 2017; Lu et al., 2017; Ruthotto and Haber, 2020). In the limit as the time step $\Delta t \to 0$, this update becomes the continuous ODE as below:

$$\frac{dS_t}{dt} = f_\theta(S_t, x_t) \tag{2}$$

This equivalence forms the foundation of the emerging field of neural ODEs, which treats neural networks as parameterized differential operators. Existing research has demonstrated a theoretical connection between the network architecture of the RNN family and numerical methods for ODEs, providing theoretical support to address problems involving system dynamics (Niu et al., 2019). In this context, a PINN can be interpreted as an explicit neural ODE that embeds known physical laws directly into the learning process. Instead of relying solely on sequential data propagation as in RNNs, PINNs leverage automatic differentiation to compute continuous derivatives with respect to time and enforce that these derivatives satisfy the governing physical ODEs.

ODEs are extensively employed by scientists for studying system dynamics such as seismic signal analysis (Peng et al., 2014) and global climate modeling (Kaper and Engler, 2013). An urban catchment is a typical dynamic system where input (e.g., rainfall, pollutant buildup), output (e.g., runoff, pollutant washoff), and storage (e.g., soil moisture, road deposit) evolve over time whereby the entire physical process can be reflected in a series ODEs and transformed into a recurrent neural network. The wash-off process describes the detachment and transport of pollutants from impervious surfaces during rainfall–runoff events. Empirical studies have shown that the remaining surface pollutant mass M_t decreases exponentially as runoff progresses. A widely used discrete formulation of this process is expressed as:

$$M_{t+1} = M_t\left(1 - aQ_t^b\right) \tag{3}$$

where M_t represents the pollutant mass remaining on the surface at time step t, Q_t is the runoff volume during the same interval, and a and b are empirical coefficients representing the wash-off rate and hydraulic sensitivity, respectively. This recursive form reflects the exponential decay of surface pollutant mass due to successive rainfall–runoff interactions and serves as a discrete-time analog to the continuous exponential model:

$$\frac{dM_t}{dt} = -kq_tM_t \tag{4}$$

where q_t represents the flow rate at time step t, and k is an empirical coefficient.

In the cell of Physics-Informed Wash-off Networks (PIWON), we embed this discrete physical law into a neural network framework. Similar to the conventional RNN architecture shown in Fig. 1a (Rumelhart et al., 1986), the framework of the PIWON layer consists of recurrent units capable of providing memory of past sequences (Fig. 1b). Within the recurrent units of the PIWON layer, connections between

neurons (input x, state s, and output y) are explicitly defined in a discrete form through state space representation, replacing the architecture parameters of a regular RNN (i.e., weights and biases) with physically meaningful parameters (denoted as *a, b* in Equation 3). The PIWON cell takes hydrological inputs (i.e., surface runoff Q_t) and predicts the surface pollutant mass M_{t+1} through a neural mapping $\widehat{M}_{t+1} = f_\theta(M_t, Q_t)$, where θ denotes the trainable network parameters. This integration of physical equations and neural networks bridges the gap between discrete-time sequence models and continuous-time physical systems, allowing PIWON to represent the pollutant wash-off dynamics as a smooth, physically consistent evolution over time while retaining the expressive power of deep neural networks.

Based on the PIWON layer, we constructed the framework embedding water quality dynamics into deep learning architecture as shown in Fig. 1c. For each input sample, we have the records of total precipitation and total surface runoff during the rainfall event in an urban catchment. Also, the catchment-specific hydrological characteristics, such as land use and antecedent dry days, have also been recorded, and are used as model input in this study. Due to the lack of continuous rainfall intensity observation in the dataset, we generated the time series of rainfall for a given total rain volume. The details can be found in Text S1 in Supporting Information.

With the exception of the input module, the model framework comprises four primary blocks: a seq2seq block, two fully connected blocks, and one physics-informed block. The seq2seq block employs a GRU (Gated Recurrent Unit, a variant of RNN)-based encoder-decoder architecture for runoff generation, utilizing the designed precipitation time series and fundamental ground information. The two fully connected blocks consist of three common neural network layers each. The first fully connected block generates the initial hidden state, representing the initial load build-up before and during the antecedent days of the rainfall event. The second fully connected block generates physically meaningful parameters, namely *a* and *b* in the exponential model of NPS pollution wash-off process. Both fully connected blocks are tailored to map case-dependent attributes (e.g., land use and antecedent dry days) onto the model parameters of the PIWON layer. The final block utilizes the outputs from the other three blocks to simulate the wash-off process of NPS pollution in urban runoff, through the proposed RNN-like PIWON layers. A time series recording the pollution mass on the ground is anticipated to be provided throughout the entire framework.

2.2. Data description and preprocessing

The National Stormwater Quality Database (NSQD) remains the most comprehensive and authoritative repository of stormwater quality data (Pitt et al., 2004). We utilized NSQD v4.02, which documents assessments for over 9,051 storm events across approximately 300 urban catchments, categorized by land use. Given that the prohibitive costs of monitoring campaigns often preclude the acquisition of continuous, high-quality time series, the NSQD employs Event Mean Concentrations (EMCs) to facilitate the comparison of pollution loads across varying runoff events and sampling strategies. EMC is defined as the ratio of the total pollutant mass to the total runoff volume for a specific event:

$$EMC = \frac{M}{V} = \frac{\sum C_i Q_i \Delta t}{\sum Q_i \Delta t} = \frac{\sum C_i V_i}{\sum V_i} \quad (5)$$

where M is the mass of pollutants (mg); V is the total volume of runoff (L); Q_i is the runoff flowrate (L/s); C_i is the pollutant concentration (mg/L); V_i is the volume of runoff within each time interval (L), and Δt is the time interval of the incremental sample (sec). This study focuses on Total Suspended Solids (TSS) in surface runoff, which comprises water-insoluble inorganic materials (e.g., sediment, clay) and organic matter (e.g., fecal matter, plankton, vegetation, microorganisms). Anthropogenic NPS pollution elevates TSS loads, significantly altering the physical, chemical, and biological properties of receiving waters (Bilotta and Brazier, 2008). Moreover, TSS serves as a primary vector for other pollutants—including nutrients, heavy metals, and organic compounds—that adsorb onto solid particles (Mahbub et al., 2011;

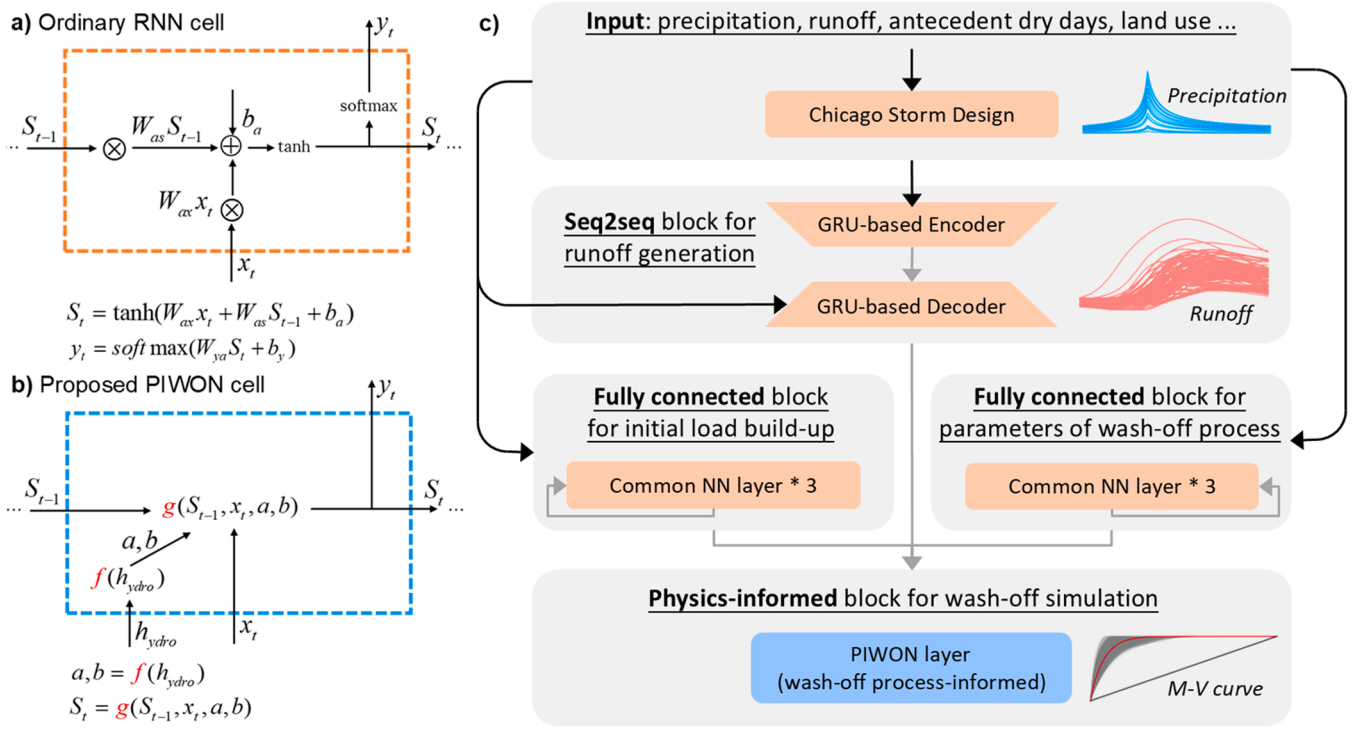


Fig. 1. Architectures of the a) ordinary RNN cell, and the b) proposed PIWON cell. The empirical coefficients *a* and *b* are generated based on local hydrological variables h_{ydro}. The hidden state S_t, representing the available pollutant mass on the land surface, is derived from S_{t-1}, *a*, *b*, and the surface runoff x_t. c) The complete model framework illustrating the simulation of continuous wash-off time series from aggregated tabular inputs via the physics-informed layer.

Mahbub et al., 2010; Vaze and Chiew, 2002). Consequently, TSS concentration is a standard indicator of NPS pollution. The NSQD contains 7,482 records reporting valid TSS EMCs, making it the most extensively documented parameter in the database.

Each storm event in the NSQD is characterized by site-specific parameters, such as catchment size, percent imperviousness, and dominant land use, as well as temporal factors like precipitation timing and the antecedent dry period. In total, 17 variables (including land use, rainfall depth, runoff depth, antecedent dry days, catchment area, EPA rain zone, conveyance type, and season) were selected as input features for the PIWON model, with two serving as targets (Table 1). Preprocessing involved encoding categorical features as monotonically increasing integers (treating missing values as distinct categories). For numerical features, the missing values were imputed using the mean.

In addition, two independent datasets from China were collected to evaluate the model's transferability and its capability for pollutograph simulation. The first dataset, collected in the Longgang District of Shenzhen, contains 70 rainfall events with corresponding EMC measurements. The second dataset, obtained from the Guangming District of Shenzhen, includes 32 storm events with TSS time-series data. Detailed descriptions of these datasets are provided in Tang et al. (2021) and Shang et al. (2025). The processed NSQD dataset and two independent datasets can be found in the GitHub repository at https://github.com/nantekoto/piwon.

2.3. First flush intensity

Storm runoff volumes significantly exceed domestic wastewater generation. Consequently, the concept of the "first flush", where the initial portion of runoff carries the majority of the pollution load, was developed to facilitate the targeted treatment of concentrated pollutants rather than the entire runoff volume. This phenomenon is generally characterized using the *M(V)* curve (Griffin Jr et al., 1980) (Fig. S2); a first flush exists if the curve lies above the 45° bisector. To strictly define the phenomenon, a widely accepted definition requires that at least 80% of the pollutant mass be transported within the first 30% of the runoff volume (Bertrand-Krajewski et al., 1998). However, these thresholds are largely empirical, and alternative criteria (e.g., 20/80 or 25/50) have also been proposed (Sansalone and Buchberger, 1997). A quantitative version of such definition is mass first flush ratio (MFF), which is generally calculated as the ratio of the normalized cumulative pollutant mass to the normalized cumulative runoff volume at a specific point in the runoff event:

Table 1
Variables and corresponding descriptions.

Type	Variable	Description
Land use	%Res	% of residential area
	%Inst	% of institutional area
	%Comm	% of commercial area
	%Ind	% of industrial area
	%Open	% of open area
	%Fwy	% of freeway area
	%Water	% of water area
	%UNK	% of unknown area
	Prim LU	Primary land use type
	Sec LU	Secondary land use type
Catchment	%Imp	% of impervious area
	Area	Catchment area (ha)
	Zone	EPA rain zone (refer to Fig. S1)
	Conv	Conveyance type
Hydrology	Precip	Precipitation depth (mm)
	ADD	Antecedent dry days
	Season	Season
Label	TSS	EMC of TSS (mg/L)
	Runoff	Runoff volume (mm)

$$\mathrm{MFF}_n = \frac{M_n}{V_n} \tag{6}$$

where n is the percentage of the total runoff volume considered.

A more generalized approach fits experimental *M(V)* curves to a power function $F(X) = X^b$, where b=1 indicates uniform wash-off and lower values signify a stronger first flush (Bertrand-Krajewski et al., 1998). Nevertheless, studies indicate that *M(V)* curves vary significantly across catchments, and the fit between observed data and the power function is frequently inadequate (e.g., the mixed *M(V)* curve in Fig. S2).

In this study, we proposed a First Flush Intensity (FFI) index to represent the wash-off inequality within a storm runoff event, inspired by the Gini index in economics (Gini, 1936). The FFI is defined as the ratio of the area that lies under the *M(V)* curve $A_{M(V)}$ over the area under the line at 45 degrees A_{diag}:

$$\mathrm{FFI} = \frac{A_{M(V)}}{A_{diag}} = 2A_{M(V)} \tag{7}$$

FFI ranges from 0 to 2. A uniform wash-off process implies $A_{M(V)}$=0.5, resulting in an FFI of 1. Higher FFI values signify a more pronounced first flush. Conversely, an FFI<1 indicates a "later flush" phenomenon, where the majority of the pollution load is carried by the latter part of the runoff, corresponding to the lagging *M(V)* curve shown in Fig. S2.

2.4. Physics-constrained training strategy

To enforce physical consistency, we incorporated multiple constraints in the training objective. The total loss is composed of three terms:

$$\mathscr{L} = \mathscr{L}_{\mathrm{load}} + \lambda_{\mathrm{runoff}}\mathscr{L}_{\mathrm{runoff}} + \lambda_{\mathrm{exp}}\mathscr{L}_{\mathrm{exp}} \tag{8}$$

where $\mathscr{L}_{\mathrm{load}}$ is a pollution load loss term, which minimizes the difference between observed and predicted EMC via mean squared error (MSE) function; $\mathscr{L}_{\mathrm{runoff}}$ is a runoff loss term, which minimizes the difference between observed and predicted total runoff volume via MSE function; $\mathscr{L}_{\mathrm{exp}}$ is a soft constraint to penalize deviations from the exponential curves. Coefficients $\lambda_{\mathrm{runoff}}$ and λ_{exp}balance data fidelity and physics enforcement. This implementation combines the interpretability and parameter parsimony of the exponential wash-off model with the representational flexibility of neural networks, enabling robust reproduction of water quality dynamics from the tabular dataset.

The preprocessed TSS dataset was partitioned into training, validation, and test sets using an 80:10:10 ratio. Specifically, 10% of the samples were reserved as a test set to evaluate final generalization. To ensure robustness and stability (Mienye and Sun, 2022), five distinct random seeds were used to extract a 10% validation set; for each seed, a model was trained on the remaining 80% using an early stopping strategy, with the best-performing model selected to mitigate overfitting (Prechelt, 2002).

Addressing the slow convergence characteristic of PINN architectures (Jahani-Nasab and Bijarchi, 2024), a two-stage training strategy was implemented: an initial phase testing various initializations and optimizers to ensure convergence, followed by a fine-tuning phase using a lower learning rate to enhance accuracy. The model was trained for 400 epochs (processing 7,637 samples per epoch) to capture the full convergence profile, and the epoch achieving the best validation performance was selected for evaluation. The complete training process required approximately 70 minutes on a single NVIDIA GeForce RTX 3060 (12 GB).

2.5. Benchmarking and performance metrics

To rigorously benchmark the predictive capabilities of the PIWON framework, we implemented and fine-tuned five state-of-the-art baseline models, selecting representative algorithms from both gradient boosting decision trees (GBDTs) and emerging tabular deep learning architectures. The baseline ensemble included XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018), which currently represent the industry standard for tabular regression tasks (Grinsztajn et al., 2022; McElfresh et al., 2023). Additionally, we integrated two novel deep learning foundation models: TabM (Tabular Multi-Layer Perceptrons) (Hollmann et al., 2025), which utilizes an efficient ensemble of MLPs with shared backbones to capture complex feature interactions, and TabPFN (Tabular Prior-Data Fitted Network) (Hollmann et al., 2025), a transformer-based architecture pre-trained on synthetic datasets to enable in-context learning for small-sample regimes (Chen et al., 2023; Gorishniy et al., 2022).

To ensure a rigorous comparison and maximize the performance of the black-box baselines, we adopted a multi-layer stacking strategy combined with extensive hyperparameter optimization. Instead of relying on single model instances, we utilized k-fold cross-validation bagging, where models were trained on distinct data folds and their predictions averaged to reduce variance and mitigate overfitting. The out-of-fold predictions generated by these base learners were subsequently concatenated and used as input features for a secondary meta-learning layer, allowing the system to learn optimal non-linear combinations of the diverse inductive biases offered by the tree-based and deep learning models. This training protocol ensures that the baseline results represent the upper bound of performance achievable by purely data-driven approaches on the available dataset. Consistent with the PIWON protocol, these baselines were evaluated over five repetitions using different random seeds on the identical training dataset. AutoGluon (Erickson et al., 2020) was employed for hyperparameter optimization; detailed configurations are provided in Table S2.

We use Nash–Sutcliffe Efficiency (NSE) coefficient, root mean square error (RMSE) and mean absolute percentage error (MAPE) as evaluation metrics. NSE represents the proportion of variance in the observations that the model can predict (Nash and Sutcliffe, 1970). The definition of NSE is as follow:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \tag{9}$$

where $\overline{O}$ represents the average of observations. NSE ranges from $-\infty$ to 1. The closer the NSE is to 1, the better the simulation result is. RMSE is the root of the average squared magnitude of errors between the predicted and actual values (Hyndman and Koehler, 2006). It can be calculated as follow:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \tag{10}$$

where N is the total sample number. MAPE is another popular evaluation metric for regression problems, which is sensitive to relative errors (De Myttenaere et al., 2016):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{S_i - O_i}{O_i}\right| \tag{11}$$

Generalizability is a crucial issue for data-driven models, indicating that the models can maintain strong performance when transferred to new datasets. A relevant criterion for assessing generalization ability (GA) evaluates this aspect by comparing the model's performance on the test set to its performance on the training set (Gorgij et al., 2023):

$$GA = \frac{RMSE_{train}}{RMSE_{test}} = \frac{\sqrt{MSE_{train}}}{\sqrt{MSE_{test}}} \tag{12}$$

If the GA value exceeds one, the model is considered undertrained or underfitting; conversely, if the GA value is less than one, the model is regarded as overtrained or overfitting. Therefore, the closer the GA value is to one when it is less than one, the better the model's generalization.

2.6. Interpretability framework

To transcend the "black-box" limitations inherent in data-driven modeling and verify the physical plausibility of the learned relationships, we employed SHAP (SHapley Additive exPlanations), a unified framework grounded in cooperative game theory (Lundberg, 2017). Specifically, we utilized the Kernel SHAP estimator, a model-agnostic method that approximates feature importance by training a weighted local linear regression on perturbed data coalitions. This approach was selected for its mathematical consistency and its ability to interpret both the custom PIWON architecture and the heterogeneous baseline ensemble.

Given the computational complexity of calculating exact Shapley values in high-dimensional feature spaces, we implemented a rigorous approximation strategy. We constructed a representative background dataset by summarizing the training distribution via k-means (k=50) clustering, which served as the reference for integrating out "missing" features during the coalition sampling process. This allowed us to compute the marginal contribution of each hydrological driver (e.g., rainfall intensity, ADD) to the model output, decomposing the prediction into additive feature attribution values. This process ensures that the resulting SHAP values accurately reflect the physical drivers of water quality dynamics rather than artifacts of the sampling variability.

3. Results and discussion

3.1. Comparative analysis of predictive performance

The predictive performance of the proposed PIWON framework was benchmarked against five state-of-the-art data-driven models (XGBoost, LightGBM, CatBoost, TabM, and TabPFN). The quantitative evaluation, summarized in Table 2, demonstrates that PIWON achieves an improved balance between accuracy and generalizability. In the test phase, PIWON attained an NSE of 0.65±0.007, significantly outperforming the baselines, where the highest Test NSE was 0.33 (TabPFN). Notably, while baseline models exhibited overfitting, evidenced by sharp performance declines from training to testing (e.g., XGBoost dropped from 0.51 to 0.32), PIWON maintained high stability. This is quantified by the generalization metric. PIWON achieved a GA of 0.94, approaching the ideal value of 1.0, whereas baselines ranged from 0.81 to 0.89. This stability suggests that incorporating physical laws acts as a robust regularizer, preventing the neural network from fitting spurious noise in the training data.

The superiority of the physics-informed approach is visually demonstrated by the scatter plots and error distributions in Fig. 2. As shown in Fig. 2a, baseline models fail to capture the full dynamic range of TSS concentrations, exhibiting high variance and a tendency to cluster predictions around the mean. This limitation is further highlighted in Fig. 2b, which analyzes relative errors across concentration deciles. Purely data-driven models display a systematic regression-to-the-mean bias: consistently overestimating loads in low-concentration events while significantly underestimating them in high-concentration events. In contrast, PIWON effectively eliminates this bias, maintaining a stable error distribution centered near zero across all ranges. Consequently, the error frequency distribution for PIWON (Fig. 2c) follows a symmetric, Gaussian-like profile, whereas baselines exhibit skewed residuals. These

Table 2
Quantitative performance evaluation of PIWON and baseline models. The table reports the mean ± standard deviation of performance metrics across 5 random seeds.

	PIWON		XGBoost		LightGBM		CatBoost		TabM		TabPFN	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
NSE	0.69 ±0.01	0.65 ±0.007	0.51 ±0.005	0.32 ±0.001	0.56 ±0.009	0.32 ±0.001	0.53 ±0.004	0.31 ±0.001	0.51 ±0.009	0.31 ±0.002	0.47 ±0.000	0.33 ±0.000
RMSE	0.65 ±0.01	0.69 ±0.007	0.81 ±0.004	0.95 ±0.000	0.77 ±0.007	0.95 ±0.000	0.8 ±0.003	0.96 ±0.001	0.81 ±0.008	0.96 ±0.002	0.84 ±0.000	0.95 ±0.000
MAPE	0.13 ±0.00	0.13 ±0.002	0.18 ±0.001	0.21 ±0.000	0.17 ±0.002	0.21 ±0.000	0.18 ±0.001	0.21 ±0.000	0.18 ±0.003	0.21 ±0.000	0.19 ±0.000	0.21 ±0.000
GA	0.94 ±0.016		0.85 ±0.004		0.81 ±0.007		0.83 ±0.003		0.84 ±0.007		0.89 ±0.000	

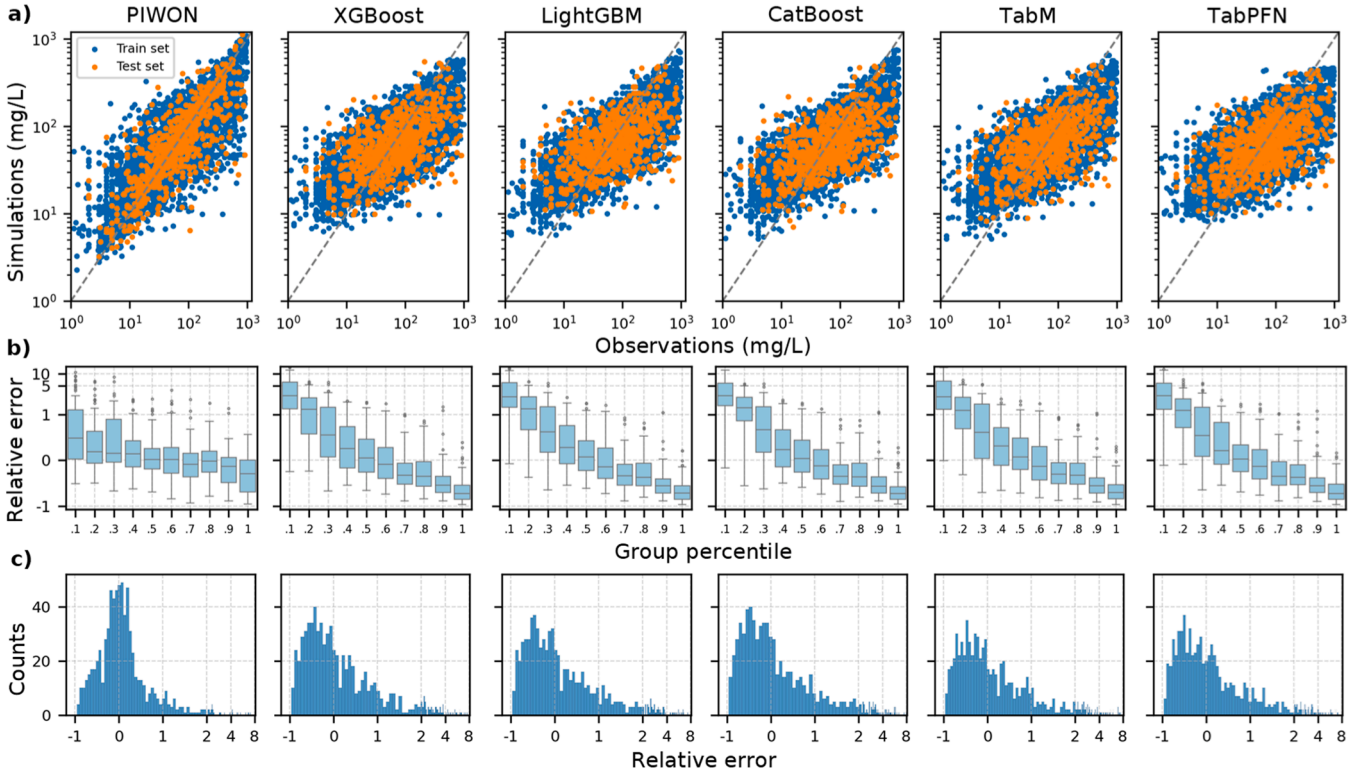


Fig. 2. Comparison of predictive performance between the proposed PIWON model and five baseline models. (a) Scatter plots of simulated versus observed TSS EMC on a logarithmic scale. The dashed line indicates the 1:1 line of perfect agreement. (b) Boxplots of relative error distribution across ten deciles (0.1–1.0) of the observed concentration range. (c) Histograms of relative error frequencies. All results are aggregated from 5 independent random seeds to ensure statistical robustness.

findings indicate that by embedding the exponential wash-off mechanism, PIWON transcends simple data fitting to capture the underlying non-linear dynamics of pollutant mobilization, enabling accurate predictions even for extreme events that challenge conventional machine learning algorithms.

To evaluate the robustness and transferability of PIWON, two fully independent datasets from Shenzhen, China, were used for external validation (Fig.s S3–S4). These datasets differ substantially from the NSQD training database in terms of spatial scale, hydrometeorological conditions, urban morphology, monitoring protocols, and data completeness, with several important predictors (e.g., %Imp) unavailable for the Shenzhen catchments. Therefore, the model was tested under a genuine out-of-distribution scenario without regional retraining.

Despite these differences, PIWON maintained reasonable predictive skill for EMC estimation (Fig. S3). For the Longgang dataset, the model achieved a correlation coefficient of 0.53 with relatively low prediction errors (MSE = 0.47; MAPE = 0.19). For the Guangming dataset, which exhibited greater concentration variability and a wider EMC range, the model achieved a higher correlation coefficient of 0.69, indicating that PIWON successfully captured the overall scaling behavior of TSS EMC across storm events. Although NSE values remained relatively low, this is expected for highly heterogeneous stormwater quality datasets with limited sample sizes and strong event-to-event variability.

The Guangming dataset was further used to validate pollutograph simulations for 31 independent storm events (Fig. S4). Many monitored catchments were small urban drainage areas (< 0.1 ha), where runoff quality responses are highly stochastic due to rapid hydraulic response times and localized surface heterogeneity. Consequently, substantial variability in simulation performance was observed among events. While some events achieved strong temporal agreement with observations ($R > 0.8$), others showed weak or negative correlations. Nevertheless, the overall distribution yielded a median correlation coefficient of 0.48, suggesting that PIWON generally captured the dominant

temporal evolution patterns of TSS concentrations despite the severe domain shift and absence of site-specific calibration. These results indicate that the embedded physical constraints within PIWON improve model generalizability and help preserve physically realistic wash-off behavior under highly heterogeneous urban conditions.

3.2. Identification of drivers for TSS EMC

Beyond simple correlation analysis, SHAP enables the exploration of the mechanisms underlying the TSS wash-off process. The global feature importance analysis (Fig. 3a) establishes the hierarchy of drivers for TSS EMC. Land use indicators, specifically Primary Land Use (Prim LU) and Percentage Industrial (%Ind), are the dominant predictors. This importance is physically aligns with the concept of mass supply. In urban catchments, the magnitude of a pollution event is fundamentally constrained by the availability of surface pollutants. Industrial zones, characterized by heavy vehicular traffic, material storage, and atmospheric deposition, function as continuous sources where the particulate supply is rarely exhausted (Li et al., 2015). Conversely, open spaces and residential areas typically operate as mass-limited systems with finite pollutant loads (Chow et al., 2013).

The SHAP dependence plot for Primary Land Use (Fig. 3b) illustrates this stratification. Industrial (Ind), Commercial (Comm), Freeway (Fwy), and Institutional (Inst) land uses exhibit high positive SHAP values, driving predictions toward higher TSS concentrations. In contrast, Open Space (Open) and Residential (Res) areas show negative SHAP values, reducing predicted loads. This divergence results from both land use activity and physical characteristics; commercial and industrial zones typically possess higher connected imperviousness, which facilitates efficient pollutant transport (Paton and Haacke, 2021). This pattern is consistent in the Secondary Land Use analysis (Fig. 3e), confirming that the model captures the overall character of the catchment. Although the percentage contributions of individual land uses (e.g., %Res, %Comm in Fig. S5) rank lower globally, this is likely due to multicollinearity (Salih, 2024), where the dominant Primary Land Use feature displaces correlated fractional inputs.

Catchment morphology also influences predictions. The Area feature (Fig. 3d) displays a negative correlation with SHAP values. As catchment size increases, predicted TSS concentration generally decreases. This is likely due to longer concentration times in larger catchments, which allow for the desynchronization of peak flows and the settling of suspended solids (Deletic, 2005). Additionally, larger catchments in this dataset are more likely to contain pervious cover that reduces overall imperviousness. Imperviousness (%Imp, Fig. 3f) shows a positive linear relationship with SHAP values, reflecting increased runoff efficiency and limited infiltration (Wang et al., 2021). However, its global ranking (Fig. 3a) is lower than anticipated, potentially due to a high proportion of missing values (Vo et al., 2024) (41.85%). Despite this, the model

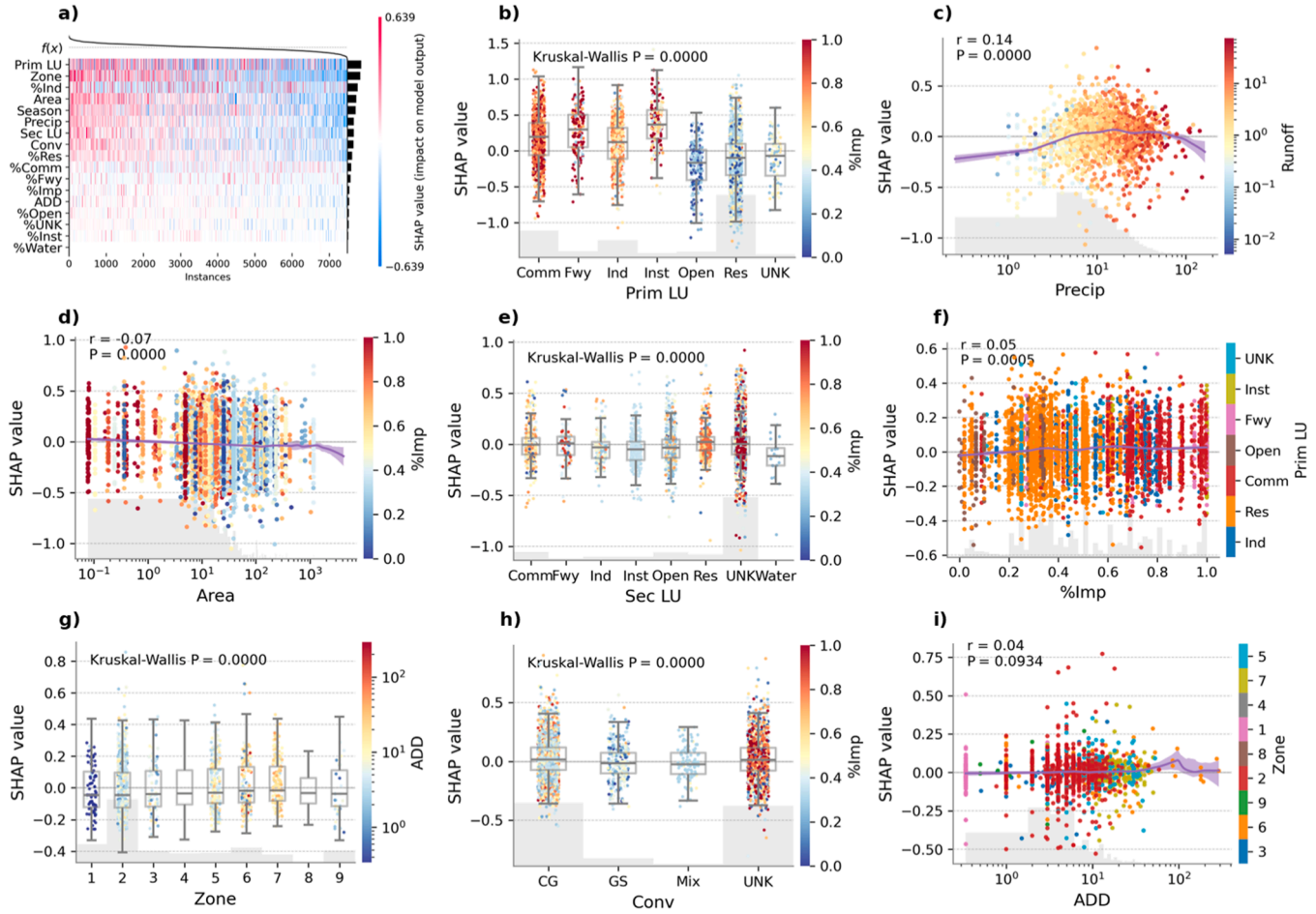


Fig. 3. Model interpretation of non-linear drivers for TSS EMC. (a) Global feature importance heatmap ranking predictors by their marginal contribution to the TSS EMC for all test instances; (b-i) feature importance distribution and dependence plots for specific drivers. For categorical features, boxplots illustrate the distribution of SHAP values; statistical significance was evaluated using the Kruskal-Wallis test to detect differences between groups (followed by a pairwise Dunn's test to determine which specific groups are different from each other, shown in Fig. S6). For numeric features, a locally weighted scatterplot smoothing (LOWESS) regression with a 95% confidence interval (shaded band) is overlaid to visualize non-linear trends, and Pearson correlation coefficients with p values are provided to quantify linear relationships.

correctly identifies that high imperviousness accelerates transport.

Infrastructure type is also significant. Analysis of Conveyance (Conv, Fig. 3h) highlights the impact of green infrastructure: runoff conveyed through Grass Swales (GS) is associated with significantly lower SHAP values compared to Curb and Gutter (CG) systems. This supports the physical function of swales, which reduce flow velocity, promote sedimentation, and filter particles through vegetation (Stagge et al., 2012).

Regional climate patterns, represented by EPA Rain Zone (Zone, Fig. 3g), show significant variation. Arid and semi-arid regions (Zones 5 and 6, see Fig. S1) are associated with higher TSS predictions. This is linked to the Antecedent Dry Days (ADD) mechanism, where longer inter-event dry periods allow pollutants to accumulate. In contrast, wetter climates (Zones 1 and 2) experience frequent rainfall that regularly washes surfaces. Generally, pollutant buildup increases initially and then plateaus as removal rates balance deposition (Vaze and Chiew, 2002). In this study, the direct ADD feature (Fig. 3i) does not show a significant trend (Vo et al., 2024), likely due to substantial missing data (78.29%).

Finally, the Precipitation feature (Fig. 3c) exhibits a non-linear trend. At low to medium rainfall depths, SHAP values increase, indicating pollutant mobilization as rainfall energy detaches particles. As precipitation increases further (typically $>$ 20 mm), the curve trends downward. This captures the dilution effect: once the available pollutant mass is washed off, additional rainfall volume reduces the overall EMC (Lee et al., 2011).

3.3. Mechanisms governing wash-off dynamics

Leveraging PIWON's capability to simulate continuous pollutographs from tabular data allows for an exploration of wash-off physics. The *M(V)* curves generated for the dataset (Fig. 4a) predominantly arch above the 45-degree bisector, indicating a significant first flush effect where the rate of mass accumulation exceeds volume accumulation (Griffin Jr et al., 1980). This geometry reflects a mass-limited regime: the finite supply of surface pollutants is mobilized by initial rainfall kinetic energy, leaving subsequent runoff to dilute the overall concentration. Conversely, a subset of events (grey lines, Fig. 4a) exhibits a moderate profile, implying a transport-limited regime where pollutant supply exceeds the early transport capacity. The variability across the ensemble confirms that the first flush is not a static catchment property but a dynamic outcome of storm characteristics, antecedent conditions, and morphology.

To quantify first flush strength, this study employs several metrics. The dataset confirms a dominant first flush phenomenon for TSS, with an average First Flush Intensity (FFI) of 1.77, and average MFF_{20} and MFF_{30} values of 4.06 and 3.14, respectively (indicating that ~81% and

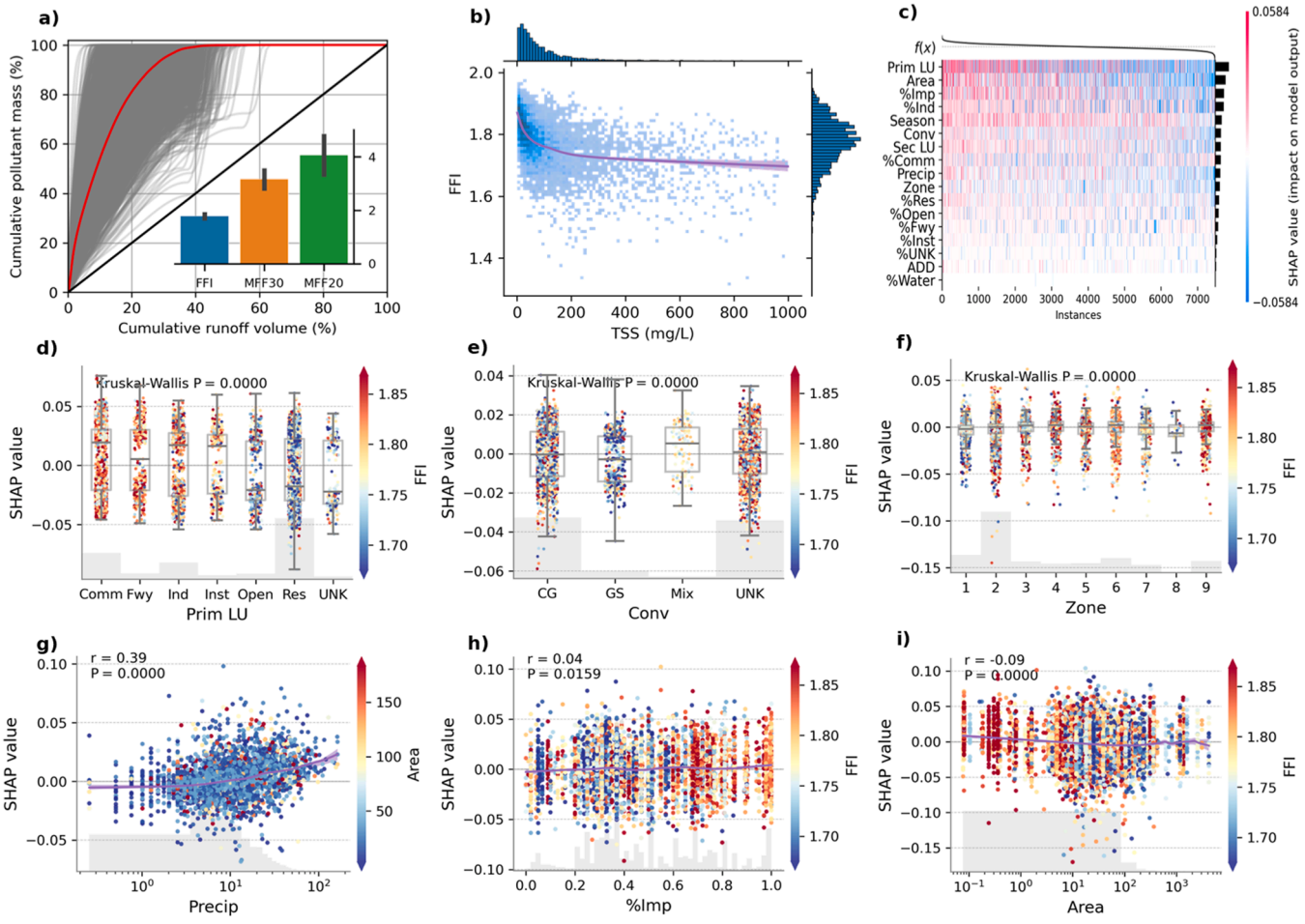


Fig. 4. Characterization and driver analysis of the FFI (First Flush Intensity). (a) Dimensionless cumulative mass-volume *M(V)* curves for TSS. The grey lines represent individual storm events, while the red curve illustrates the averaged pollutograph representing the aggregate behaviour. (b) Scatter plot of FFI versus TSS EMC, overlaid with marginal histograms. (c) Global feature importance heatmap ranking predictors by their marginal contribution to the FFI for all test instances; (d-i) Feature importance distribution and dependence plots for specific drivers. For categorical features, boxplots illustrate the distribution of SHAP values; statistical significance was evaluated using the Kruskal-Wallis test to detect differences between groups (followed by a pairwise Dunn's test to determine which specific groups are different from each other, shown in Fig. S9). For numeric features, a locally weighted scatterplot smoothing (LOWESS) regression with a 95% confidence interval (shaded band) is overlaid to visualize non-linear trends, and Pearson correlation coefficients with p values are provided to quantify linear relationships.

~95% of the total load is transported by the first 20% and 30% of the volume). While MFF ratios are widely used, they are limited as "point" measurements. They capture a snapshot at an arbitrary threshold, creating a saturation effect for intense events. For example, MFF_{20} saturates at 5 regardless of how early the mass is delivered (3.33 for MFF_{30}, as shown in Fig. S7). In contrast, the FFI index integrates the entire *M(V)* profile, capturing wash-off behavior across the full event duration.

We applied SHAP to the FFI index to dissect the non-linear drivers of wash-off intensity. The impact of land use on FFI (Fig. 4c-d) closely mirrors the drivers of EMC, suggesting that physical characteristics promoting high loads also facilitate rapid mobilization. Industrial (%Ind), Freeway (%Fwy), and Commercial (%Comm) land uses are primary drivers of high FFI values. This is attributed to high connected imperviousness (smooth surfaces like concrete and asphalt) and high pollutant buildup, which allow rainfall energy to efficiently detach particulates (Zhang et al., 2020). Conversely, Residential (%Res) and Open Space (%Open) land uses dampen first flush intensity due to disconnected impervious surfaces and higher roughness from green infrastructure (Zhang et al., 2020).

This structural influence extends to infrastructure and climate. Conveyance analysis (Fig. 4e) supports the efficacy of green infrastructure: Grass Swales (GS) are associated with lower FFI values compared to Curb and Gutter (CG) systems, confirming their role in flow regulation and early pollutant interception (Stagge et al., 2012). Similarly, Percentage Imperviousness (%Imp) shows a significant, though weaker, positive correlation with FFI (Fig. 4h). The EPA Rain Zone impact (Fig. 4f) aligns with the EMC results reported in Section 3.2, likely linked to mass-limited accumulation regimes. The Antecedent Dry Days (ADD) feature alone shows no significant relationship with FFI (Fig. S8k), and individual land use contributions rank low globally (Fig. S8a-h). Catchment Area shows a negative relationship with FFI (Fig. 4i). This is mechanically driven by the mixing and desynchronization of flows rather than simple dilution. As the catchment scale increases, the temporal lag between upstream and downstream runoff homogenizes the wash-off profile, reducing the intensity of the first flush (Lin et al., 2024).

Notably, while most drivers influence FFI and EMC in the same direction, Precipitation (Precip) reveals an inverse relationship (Fig. 4g). For TSS EMC, higher precipitation generally causes dilution (Fig. 3c); however, for FFI, higher precipitation correlates with increased intensity. This divergence explains the negative correlation observed between EMC and FFI (Fig. 4b). In high-rainfall (mass-limited) events, initial runoff exhausts the available pollutants; the subsequent "clean tail" increases the FFI but dilutes the final EMC (Lee et al., 2011). Conversely, in low-precipitation or transport-limited events, pollutant delivery is more uniform. Without a clean tail to dilute the mean, the EMC remains high, but the lack of a sharp initial peak results in a low FFI (Al Mamoon et al., 2019). This confirms that the first flush is a dynamic outcome of the interplay between available mass and rainfall volume.

3.4. Implications for urban water management

PIWON's capacity to not only predict EMC but also simulate wash-off temporal dynamics offers significant advantages for urban water management. Current regulations typically mandate the treatment of a fixed Water Quality Volume (e.g., the first 0.5 inches of runoff for 90% pollutant removal (Grisham, 1995; Schueler, 1994)), often ignoring catchment-specific wash-off characteristics. The observed variability in FFI suggests that such uniform sizing approaches may be inefficient for certain highly dynamic catchments. The distinction in SHAP drivers for EMC and FFI supports a differentiated selection of Best Management Practices (BMPs) tailored to specific wash-off regimes. In catchments with high FFI potential, such as industrial zones with high connected imperviousness, pollutant loads are highly concentrated in the early storm stages. These areas emerge as highly suitable candidates for first flush diversion systems, where technologies like hydrodynamic separators, baffle boxes, or off-line settling tanks are most effective (Hoss et al., 2016; Yu et al., 2013). Conversely, in catchments where the model predicts low FFI (e.g., residential areas with swales), pollutant delivery is more uniform due to vegetative attenuation. Here, volume-based treatment approaches, such as bioretention cells and constructed wetlands, are likely more appropriate and cost-effective (Hoss et al., 2016; Yu et al., 2013). Table S3 summarizes these divergent drivers and their management implications, providing a heuristic for engineers to select practices based on predicted FFI characteristics.

Traditional static infrastructure (e.g., passive weirs, fixed orifices) lacks the adaptability required for dynamic storm variability. Static systems prioritize volume capture uniformly, often treating minor, highly polluted events and large, dilute events identically, which can exacerbate downstream Combined Sewer Overflows (CSOs) (Peng et al., 2017). By integrating PIWON-generated pollutographs M_t into Model Predictive Control (MPC) frameworks, operators can instead implement adaptive quality-based interception strategies that explicitly target pollutant peaks.

As illustrated in Fig. 5, the effectiveness of treatment strategies strongly depends on the temporal synchronization between runoff and pollutant dynamics. Figs. 5a and 5b present two hypothetical storm events designed for conceptual comparison, where the total runoff depth (15 mm) and pollutant load (20 g/m^2) were controlled to be identical using exponential runoff and pollutant functions. During the representative First-Flush Event (Fig. 5a), pollutant concentrations peak early and generally coincide with runoff generation, resulting in comparable pollutant removal efficiencies between the volume-based and quality-based strategies (both 58.3%). In contrast, during the Lagging-Flush Event (Fig. 5b), pollutant peaks occur substantially later than the hydrograph peak. Under this condition, the conventional volume-based strategy exhausts the available storage capacity (5 mm) on relatively clean early runoff, achieving only 7.6% pollutant removal efficiency, while the quality-based strategy increases pollutant removal efficiency to 77.3% under the same treatment volume.

A similar phenomenon was observed in the real NSQD storm event shown in Fig. 5c, where multiple delayed concentration peaks were decoupled from runoff maxima. Although the conventional volume-based strategy achieved only 37.8% pollutant removal efficiency, the PIWON-enabled quality-based strategy increased efficiency to 86.3% by selectively intercepting high-concentration runoff periods. Across the NSQD dataset, quality-based treatment consistently outperformed conventional volume-based capture, with an average efficiency improvement of 9.42% (Fig. 5d).

These results demonstrate that physics-informed pollutograph prediction can provide actionable information for real-time stormwater management. Such strategies could be implemented through sensor-triggered diversion, smart detention systems, or dynamically controlled green-gray infrastructure. During intense first-flush events, systems could proactively divert highly polluted runoff into storage or treatment facilities, whereas in lagging-flush scenarios, cleaner runoff could be strategically bypassed to preserve limited treatment capacity. This adaptive operation may improve pollutant removal efficiency while reducing unnecessary hydraulic loading and infrastructure stress, thereby prolonging the service life of water infrastructure assets (Hilliges et al., 2013).

A critical bottleneck in urban hydrological modeling is the scarcity of high-resolution water quality data. Calibrating process-based models (e.g., SWMM) demands extensive, site-specific field campaigns to determine buildup/wash-off coefficients. Conversely, purely data-driven models are prone to overfitting and poor generalization, as evidenced by the low performance of baseline models in this study (Table 2). PIWON addresses this by embedding governing physical laws, specifically mass conservation and exponential decay, directly into the network architecture. This physical regularization prevents the learning of spurious correlations (e.g., runoff generation without pollutant supply) even when training data is sparse or noisy (Chen et al., 2025).

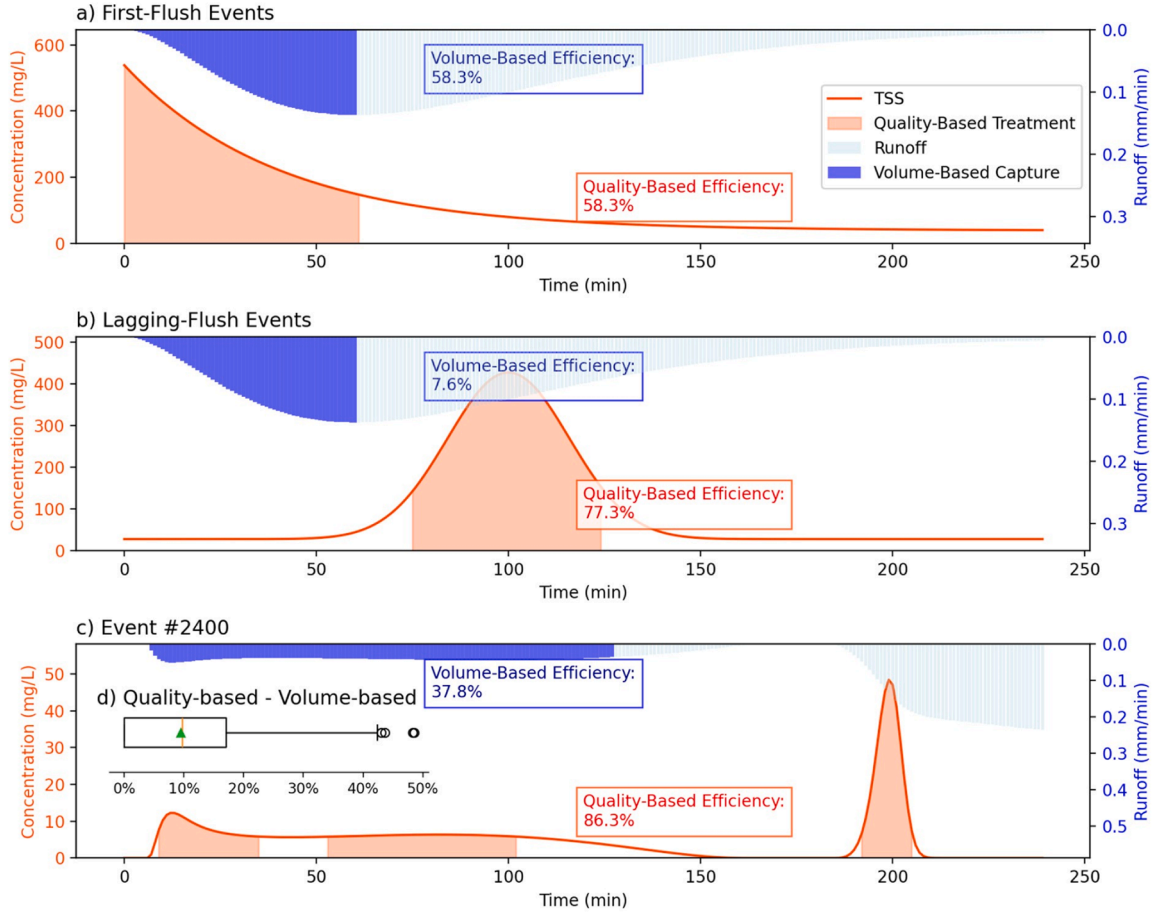


Fig. 5. Illustrative comparison of volume-based versus quality-based runoff control strategies under different runoff pollution dynamics restricted by a fixed treatment capacity of 5 mm. (a) Synthetic first-flush event, where pollutant concentrations peak at the beginning of the runoff event. (b) Synthetic lagging-flush event, where peak pollutant concentrations occur after the initial runoff peak. (c) Event #2400 from the NSQD dataset illustrating a realistic multi-stage pollutograph with delayed concentration peaks. Red curves indicate TSS concentrations, while blue shaded areas represent runoff hydrographs. Orange shaded regions denote periods selected for quality-based treatment, whereas blue shaded regions indicate runoff captured under conventional volume-based strategies. Reported efficiencies represent pollutant removal efficiency achieved under equivalent treatment volumes (5 mm). (d) Distribution of efficiency improvements achieved by quality-based treatment relative to volume-based capture across the NSQD dataset.

Results indicate PIWON achieves a high GA of 0.94, significantly outperforming baselines (GA = 0.89), suggesting it captures the underlying wash-off process rather than merely memorizing the dataset. For data-scarce regions, such as developing nations or municipalities with limited monitoring budgets, a PIWON model pretrained on comprehensive datasets offers greater reliability than standard deep learning models. Its physical structure constrains predictions to physically plausible bounds, facilitating transferability from well-instrumented regions to unmonitored catchments for improved planning and risk assessment (Chen et al., 2025).

3.5. Discussions

Despite their widespread popularity and integration into standard commercial software, such as SWMM (Gironás et al., 2010), due to their mathematical simplicity and computational efficiency, traditional empirical wash-off models are frequently cited as unreliable in complex urban catchments (Bonhomme and Petrucci, 2017). These conceptual models are heavily abstracted; they fundamentally assume a spatially uniform distribution of pollutants and rely on lumped, static coefficients to represent highly heterogeneous physical processes. Consequently, the standard exponential decay equation often fails to account for critical mechanical realities, such as the differential wash-off of varying particle sizes (Tong et al., 2025), the changing kinetic energy of dynamic rainfall (Kozak et al., 2019), and the influence of localized surface roughness (Abouelsaad et al., 2024).

While the proposed PIWON framework successfully mitigates much of this rigidity by allowing the neural network to dynamically map catchment-specific attributes to the wash-off parameters, the architecture remains bounded by the foundational assumptions of these lumped conceptual equations. To further transcend these limitations, future iterations of this physics-informed framework could substitute the empirical exponential model with more rigorous mechanistic alternatives. For instance, future frameworks could embed shallow-water hydrodynamic equations coupled with advection-dispersion sediment transport models to explicitly simulate particle detachment and transport. Additionally, shifting from a lumped approach to a spatially distributed framework—potentially by integrating Graph Neural Networks (GNNs) to map the physical topology of the urban drainage network—would allow for a more precise, physically explicit simulation of non-point source pollution dynamics.

It is important to emphasize that PIWON does not uniquely reconstruct historical pollutographs from EMC observations. Because multiple intra-event concentration trajectories may yield similar event mean concentrations, the problem is inherently equifinal. Therefore, the generated pollutographs should be interpreted as physics-constrained realizations that are consistent with observed event-scale statistics and wash-off dynamics rather than exact reproductions of historical

concentration time series. In this study, synthetic Chicago Design Storms were adopted as standardized forcing conditions to isolate pollutant wash-off responses under controlled hydrological scenarios. While this approach improves experimental consistency, it cannot fully represent the diversity and intermittency of natural rainfall structures.

Consequently, the derived *M(V)* curves, First Flush Index, and SHAP-based interpretations should be viewed as physics-informed diagnostic analyses and hypothesis-generating tools rather than direct observational confirmation of first-flush mechanisms. Nevertheless, validation against an independent monitoring dataset from Shenzhen demonstrated that PIWON was able to reproduce meaningful temporal concentration dynamics despite substantial domain shifts and the absence of site-specific retraining. Future work should incorporate high-resolution observed rainfall and continuous water quality monitoring datasets to further constrain temporal uncertainty and improve pollutograph identifiability.

CRediT authorship contribution statement

Sijie Tang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Jiping Jiang:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Shuo Wang:** Conceptualization, Project administration, Supervision, Writing – review & editing. **Yi Zheng:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Dragan Savic:** Conceptualization, Methodology, Validation, Writing – review & editing. **Aijie Wang:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

This research was supported by the National Key Research and Development Program of China (Grant No. 2023YFC3207504), the National Natural Science Foundation of China (Grant No. 52321005 and 51979136), and the Guangdong Overseas Master Program (Grant No. MS202600118).

The computational resources in this study were supported by the Center for Computational Science and Engineering at Southern University of Science and Technology.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2026.126379.

Data availability

Research Link Provided
National Stormwater Quality Database (Original data) (PIWON)

References

Abbasi, J., Andersen, P.Ø., 2024. Application of physics-informed neural networks for estimation of saturation functions from countercurrent spontaneous imbibition tests. SPE J. 29 (04), 1710–1729.
Abouelsaad, O., Hassan, A., Omar, M., Hinkelmann, R., 2024. Identifying manning roughness coefficient using automatic calibration method and simulation of pollution incidents in the Nile River, Egypt. J. Hydrol.: Reg. Stud. 55, 101908.
Al Mamoon, A., Jahan, S., He, X., Joergensen, N.E., Rahman, A., 2019. First flush analysis using a rainfall simulator on a micro catchment in an arid climate. Sci. Total Environ. 693, 133552.
Behrouz, M.S., Yazdi, M.N., Sample, D.J., 2022. Using Random Forest, a machine learning approach to predict nitrogen, phosphorus, and sediment event mean concentrations in urban runoff. J. Environ. Manag. 317, 115412.
Bertrand-Krajewski, J.-L., Chebbo, G., Saget, A., 1998. Distribution of pollutant mass vs volume in stormwater discharges and the first flush phenomenon. Water Res. 32 (8), 2341–2356.
Bilotta, G.S., Brazier, R.E., 2008. Understanding the influence of suspended solids on water quality and aquatic biota. Water Res. 42 (12), 2849–2861.
Bonhomme, C., Petrucci, G., 2017. Should we trust build-up/wash-off water quality models at the scale of urban catchments? Water Res. 108, 422–431.
Charters, F.J., Cochrane, T.A., O'Sullivan, A.D., 2016. Untreated runoff quality from roof and road surfaces in a low intensity rainfall climate. Sci. Total Environ. 550, 265–272.
Chen, K.-Y., Chiang, P.-H., Chou, H.-R., Chen, T.-W., Chang, T.-H., 2023. Trompt: Towards a Better Deep Neural Network for Tabular Data.
Chen, Y., Lu, L., Karniadakis, G.E., Dal Negro, L., 2020. Physics-informed neural networks for inverse problems in nano-optics and metamaterials. Opt. Express 28 (8), 11618–11633.
Chen, Y., Wang, H., Chen, Z., 2025. Sensitivity analysis of physical regularization in physics-informed neural networks (PINNs) of building thermal modeling. Build. Environ. 273, 112693.
Chen, T. and Guestrin, C. 2016 Xgboost: a scalable tree boosting system, pp. 785-794.
Chow, M., Yusop, Z., Shirazi, S., 2013. Storm runoff quality and pollutant loading from commercial, residential, and industrial catchments in the tropic. Environ. monit. assess. 185 (10), 8321–8331.
De Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean absolute percentage error for regression models. Neurocomputing 192, 38–48.
Deletic, A., 2005. Sediment transport in urban runoff over grassed areas. J. Hydrol. 301 (1-4), 108–122.
Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A., 2020. Autogluon-Tabular: Robust and Accurate Automl For Structured Data.
Erion, G., Janizek, J.D., Sturmfels, P., Lundberg, S.M., Lee, S.-I., 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat. Mach. Intell. 3 (7), 620–631.
Gini, C., 1936. On the measure of concentration with special reference to income and statistics. Colo. Coll. Publ. Gen. ser. 208 (1).
Gironás, J., Roesner, L.A., Rossman, L.A., Davis, J., 2010. A new applications manual for the Storm Water Management Model(SWMM). Environ. Model. Softw. 25 (6), 813–814.
Gorgij, A.D., Askari, G., Taghipour, A., Jami, M., Mirfardi, M., 2023. Spatiotemporal forecasting of the groundwater quality for irrigation purposes, using deep learning method: long short-term memory (LSTM). Agric. Water Manag. 277, 108088.
Gorishniy, Y., Rubachev, I., Babenko, A., 2022. On embeddings for numerical features in tabular deep learning. Adv. Neural Inf. Process. Syst. 35, 24991–25004.
Griffin Jr, D., Grizzard, T., Randall, C., Helsel, D., Hartigan, J., 1980. Analysis of non-point pollution export from small catchments. J. (Water Pollut. Control Fed.) 780–790.
Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Process. Syst. 35, 507–520.
Grisham, D.M., 1995. Designing for the" first flush". Civ. Eng. 65 (11), 67.
Gunawardena, J., Ziyath, A.M., Egodawatta, P., Ayoko, G.A., Goonetilleke, A., 2014. Mathematical relationships for metal build-up on urban road surfaces based on traffic and land use characteristics. Chemosphere 99, 267–271.
Haber, E., Ruthotto, L., 2017. Stable architectures for deep neural networks. Inverse Probl. 34 (1), 014004.
Hilliges, R., Schriewer, A., Helmreich, B., 2013. A three-stage treatment system for highly polluted urban road runoff. J. Environ. Manag. 128, 306–312.
Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F., 2025. Accurate predictions on small data with a tabular foundation model. Nature 637 (8045), 319–326.
Hoss, F., Fischbach, J., Molina-Perez, E., 2016. Effectiveness of best management practices for stormwater treatment as a function of runoff volume. J. Water Resour. Plan. Manag. 142 (11), 05016009.
Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22 (4), 679–688.
Jagtap, A.D., Mao, Z., Adams, N., Karniadakis, G.E., 2022. Physics-informed neural networks for inverse problems in supersonic flows. J. Comput. Phys. 466, 111402.
Jahani-Nasab, M., Bijarchi, M.A., 2024. Enhancing convergence speed with feature enforcing physics-informed neural networks using boundary conditions as prior knowledge. Sci. Rep. 14 (1), 23836.
Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. Geophys. Res. Lett. 47 (13), e2020GL088229.
Kaper, H., Engler, H., 2013. Mathematics and Climate. SIAM.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
Kozak, C., Fernandes, C.V.S., Braga, S.M., do Prado, L.L., Froehner, S., Hilgert, S., 2019. Water quality dynamic during rainfall episodes: integrated approach to assess diffuse pollution using automatic sampling. Environ. Monit. Assess. 191 (6), 402.

Launay, M.A., Dittmer, U., Steinmetz, H., 2016. Organic micropollutants discharged by combined sewer overflows–characterisation of pollutant sources and stormwater-related processes. Water Res. 104, 82–92.
Lee, J.Y., Kim, H., Kim, Y., Han, M.Y., 2011. Characteristics of the event mean concentration (EMC) from rainfall runoff on an urban highway. Environ. Pollut. 159 (4), 884–888.
Li, C., Zheng, X., Zhao, F., Wang, X., Cai, Y., Zhang, N., 2017. Effects of urban non-point source pollution from Baoding City on Baiyangdian Lake, China. In: Water, 9, p. 249.
Li, D., Wan, J., Ma, Y., Wang, Y., Huang, M., Chen, Y., 2015. Stormwater runoff pollutant loading distributions and their correlation with rainfall and catchment characteristics in a rapidly industrialized city. PLoS One 10 (3), e0118776.
Lin, F., Ren, H., Qin, J., Wang, M., Shi, M., Li, Y., Wang, R., Hu, Y., 2024. Analysis of pollutant dispersion patterns in rivers under different rainfall based on an integrated water-land model. J. Environ. Manag. 354, 120314.
Lu, Y., Zhong, A., Li, Q., Dong, B., 2017. Beyond Finite Layer Neural Networks: Bridging Deep Architectures And Numerical Differential Equations.
Lundberg, S., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv: 1705.07874.
Mahbub, P., Ayoko, G.A., Goonetilleke, A., Egodawatta, P., 2011. Analysis of the build-up of semi and non volatile organic compounds on urban roads. Water Res. 45 (9), 2835–2844.
Mahbub, P., Ayoko, G.A., Goonetilleke, A., Egodawatta, P., Kokot, S., 2010. Impacts of traffic and rainfall characteristics on heavy metals build-up and wash-off from urban roads. Environ. Sci. Technol. 44 (23), 8904–8910.
Masoner, J.R., Kolpin, D.W., Cozzarelli, I.M., Barber, L.B., Burden, D.S., Foreman, W.T., Forshay, K.J., Furlong, E.T., Groves, J.F., Hladik, M.L., 2019. Urban stormwater: An overlooked pathway of extensive mixed contaminants to surface and groundwaters in the United States. Environ. Sci. Technol. 53 (17), 10070–10081.
McElfresh, D., Khandagale, S., Valverde, J., Prasad, CV., Ramakrishnan, G., Goldblum, M., White, C., 2023. When do neural nets outperform boosted trees on tabular data? Adv. Neural Inf. Process. Syst. 36, 76336–76369.
Mienye, I.D., Sun, Y., 2022. A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10, 99129–99149.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290.
Nguyen, T.T., Ngo, H.H., Guo, W., Wang, X.C., Ren, N., Li, G., Ding, J., Liang, H., 2019. Implementation of a specific urban water management-Sponge City. Sci. Total Environ. 652, 147–162.
Niu, M.Y., Horesh, L., Chuang, I., 2019. Recurrent Neural Networks In The Eye Of Differential Equations.
Novotny, V., 1999. Integrating diffuse/nonpoint pollution control and water body restoration into watershed management. JAWRA J. Am. Water Resour. Assoc. 35 (4), 717–727.
Paton, E., Haacke, N., 2021. Merging patterns and processes of diffuse pollution in urban watersheds: a connectivity assessment. Wiley Interdiscip. Rev.: Water 8 (4), e1525.
Peng, H., Kitagawa, G., Takanami, T., Matsumoto, N., 2014. State-space modeling for seismic signal analysis. Appl. Math. Model. 38 (2), 738–746.
Peng, H.-Q., Liu, Y., Gao, X.-L., Wang, H.-W., Chen, Y., Cai, H.-Y., 2017. Calculation of intercepted runoff depth based on stormwater quality and environmental capacity of receiving waters for initial stormwater pollution management. Environ. Sci. Pollut. Res. 24 (31), 24681–24689.
Pitt, R., Maestre, A. and Morquecho, R. 2004 The national stormwater quality database (NSQD, version 1.1), pp. 13-51.
Prechelt, L., 2002. Neural Networks: Tricks of the trade. Springer, pp. 55–69.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.
Raissi, M., Yazdani, A., Karniadakis, G.E., 2020. Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. Science 367 (6481), 1026–1030.
Risch, E., Gasperi, J., Gromaire, M.-C., Chebbo, G., Azimi, S., Rocher, V., Roux, P., Rosenbaum, R.K., Sinfort, C., 2018. Impacts from urban water systems on receiving waters–how to account for severe wet-weather events in LCA? Water Res. 128, 412–423.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.
Ruthotto, L., Haber, E., 2020. Deep neural networks motivated by partial differential equations. J. Math. Imaging Vis. 62, 352–364.
Salih, A.M., 2024. Explainable Artificial Intelligence And Multicollinearity: A Mini Review Of Current Approaches.
Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature.
Sansalone, J.J., Buchberger, S.G., 1997. Partitioning and first flush of metals in urban roadway storm water. J. Environ. eng. 123 (2), 134–143.
Schueler, T.R., 1994. First flush of stormwater pollutants investigated in Texas. Watershed Prot. Tech. 1 (2), 88.
Serebrennikova, A., Teubler, R., Hoffellner, L., Leitner, E., Hirn, U., Zojer, K., 2022. Transport of organic volatiles through paper: physics-informed neural networks for solving inverse and forward problems. Transp. Porous Media 145 (3), 589–612.
Shang, F., Tang, S., Wang, H., Yang, R., Hou, Z., Ping, Y., Zhang, Z., Chen, H., Yu, Y., Goonetilleke, A., 2025. Assessing the effectiveness of non-point source pollution models in data-limited urban areas. J. Hydrol. 661, 133636.
Shao, M., Zhao, G., Kao, S.-C., Cuo, L., Rankin, C., Gao, H., 2020. Quantifying the effects of urbanization on floods in a changing environment to promote water security—a case study of two adjacent basins in Texas. J. Hydrol. 589, 125154.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54 (11), 8558–8593.
Stagge, J.H., Davis, A.P., Jamil, E., Kim, H., 2012. Performance of grass swales for improving water quality from highway runoff. Water Res. 46 (20), 6731–6742.
Sun, A.Y., Scanlon, B.R., 2019. 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), 073001.
Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks. PMLR, pp. 3319–3328.
Tang, S., Jiang, J., Zheng, Y., Hong, Y., Chung, E.-S., Shamseldin, A.Y., Wei, Y., Wang, X., 2021. Robustness analysis of storm water quality modelling with LID infrastructures from natural event-based field monitoring. Sci. Total Environ. 753, 142007.
Tong, X., Liang, Q., Sander, G., Wang, G., Lai, X., 2025. A Physically based model for non-point source pollutant wash-off process over impervious surfaces. Water Resour. Res. 61 (6), e2024WR038791.
Vaze, J., Chiew, F.H., 2002. Experimental study of pollutant accumulation on an urban road surface. Urban Water 4 (4), 379–389.
Vo, T.L., Nguyen, T., Lopez-Ramos, L.M., Hammer, H.L., Riegler, M.A., Halvorsen, P., 2024. Explainability of machine learning models under missing data.
Wang, Y., Li, C., Qiao, J., Hu, Y., Zhang, Q., Yin, J., Slater, L., 2025. Meta-analysis of urban non-point source pollution from road and roof runoff across China. Earth's Future 13 (3), e2024EF005296.
Wang, Y., Zhang, X., Xu, J., Liang, C., She, D., Xiao, Y., 2021. Evaluating effects of urban imperviousness connectivity on runoff with consideration of receiving pervious area properties. Urban Water J. 18 (8), 598–607.
Xiong, R., Zheng, Y., Chen, N., Tian, Q., Liu, W., Han, F., Jiang, S., Lu, M., Zheng, Y., 2022. Predicting dynamic riverine nitrogen export in unmonitored watersheds: leveraging insights of AI from data-rich regions. Environ. Sci. Technol. 56 (14), 10530–10542.
Yang, X., Liu, Q., Fu, G., He, Y., Luo, X., Zheng, Z., 2016. Spatiotemporal patterns and source attribution of nitrogen load in a river basin with complex pollution sources. Water res. 94, 187–199.
Yu, J., Yu, H., Xu, L., 2013. Performance evaluation of various stormwater best management practices. Environ. Sci. Pollut. Res. 20 (9), 6160–6171.
Zhang, E., Dao, M., Karniadakis, G.E., Suresh, S., 2022. Analyses of internal structures and defects in materials using physics-informed neural networks. Sci. adv. 8 (7), eabk0644.
Zhang, T., Xiao, Y., Liang, D., Tang, H., Xu, J., Yuan, S., Luan, B., 2020. A physically-based model for dissolved pollutant transport over impervious surfaces. J. Hydrol. 590, 125478.
Zuo, D., Cao, J., Lin, Y., Hu, M., Luo, P., He, B., Xu, J., 2025. A hybrid physics-machine learning modeling framework enhances nonpoint source pollution forecasting. ACS ES&T Water.