

Unveiling Sediment Deposition Thresholds in Discontinuous Emergent Vegetation Patches Using a SHAP-Interpretable Ensemble Machine Learning

Reza Javidi Sabbaghian

reza.js.civil.eng@gmail.com

Hakim Sabzevari University <https://orcid.org/0000-0001-6093-3273>

Reza Arabshahi

University of Tehran <https://orcid.org/0009-0008-4806-1685>

Mohammad Ali Bahrabadi

Tarbiat Modares University <https://orcid.org/0009-0006-5950-9567>

Seyed Ali Akbar Salehi Neyshabouri

Tarbiat Modares University <https://orcid.org/0000-0002-8485-0525>

Mohammad Ali Banihashemi

banihash@ut.ac.ir

University of Tehran <https://orcid.org/0000-0001-5399-5329>

Research Article

Keywords: Sediment deposition, Vegetated channels, Ensemble machine learning, SHAP interpretability, Riverine hydrodynamics

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11485176/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Unveiling Sediment Deposition Thresholds in Discontinuous Emergent
Vegetation Patches Using a SHAP-Interpretable Ensemble Machine Learning

Authors' name: Reza Arabshahi ^a, Mohammad Ali Bahrabadi ^b, Seyed Ali Akbar Salehi
Neyshabouri ^b, Reza Javidi Sabbaghian ^{c,d,*}, Mohammad Ali Banihashemi ^{a,*}

* Corresponding authors: Reza Javidi Sabbaghian, Mohammad Ali Banihashemi

^a School of Civil Engineering, University College of Engineering, University of Tehran,
Tehran, Iran

^b School of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran

^c Department of Civil Engineering, Hakim Sabzevari University, Sabzevar, Iran

^d Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

R. Arabshahi (reza.arabshahi@ut.ac.ir,  0009-0008-4806-1685)

M.A. Bahrabadi (m.bahrabadi@modares.ac.ir,  0009-0006-5950-9567)

S.A.A. Salehi Neyshabouri (salehi@modares.ac.ir,  0000-0002-8485-0525)

R. Javidi Sabbaghian (r.javidi.s@hsu.ac.ir, rezajs.civil.eng@gmail.com  0000-0001-6093-
3273)

M.A. Banihashemi (banihash@ut.ac.ir,  0000-0001-5399-5329)

Abstract:

Predicting sediment deposition in vegetated channels is challenging due to nonlinear flow-vegetation-sediment interactions that are poorly captured by empirical models. While physics-based equations (e.g., Saint-Venant, advection-diffusion) govern these processes, their application to discontinuous vegetation patches faces parameterization difficulties and computational demands. This study introduces a novel interpretable ensemble machine learning (ML) framework as a complementary approach, integrating novel feature engineering that defines spatial positioning relative to vegetation patches and vegetation-induced hydrodynamic effects. Five ensemble algorithms, including random forest (RF), gradient boosting (GB), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and categorical boosting (CatBoost), were evaluated using 498 observations from 11 laboratory flume experiments under varying hydraulic and vegetation conditions. Bayesian hyperparameter optimization with 5-fold cross-validation ensured robust hyperparameter tuning. A multi-criteria ranking for selecting top performers, with sensitivity analysis showed CatBoost and LightGBM consistently ranking highest. Instead of focusing solely on prediction, the discovery of a quantitative physical threshold is targeted. Through Shapley additive explanations (SHAP) analysis, a previously undocumented threshold between deposition-promoting and deposition-inhibiting regimes is identified and physically interpreted. CatBoost achieves the highest generalization (unseen holdout: $R^2 = 0.953$). SHAP reveals a crisp threshold at ≈ 0.3 m "*cumulative vegetation exposure*", which aligns with the theoretical adjustment length scale $1/(C_D \cdot a_v)$ for emergent canopies. "*Local flow velocity*" shows consistently weak influence (mean $|\text{SHAP}| = 0.049$ to 0.063), indicating that it is not a primary control within dense patches. This study demonstrates that interpretable ML can generate testable physical hypotheses from existing experimental data.

44 **Keywords:** Sediment deposition; Vegetated channels; Ensemble machine learning; SHAP
45 interpretability; Riverine hydrodynamics

Nomenclature

Symbol	Description	Unit
A	Cross-sectional area	m^2
a_v	Vegetation density	m^{-1}
c	Constriction coefficient	—
C	Sediment concentration	kg/m^3
C_D	Drag coefficient	—
D	Dispersion coefficient	m^2/s
Dep/Dep_{avg}	Deposition sediment rate in the experimental flume	—
E	Erosion rate from bed	$\text{kg}/\text{m}^2/\text{s}$
g	Gravitational acceleration	m/s^2
H	Water depth	cm
h	Water depth	m
L_t	Length of gaps between patches	m
L_v	Length of vegetation patches	m
l_x	The distance of the adjacent stems	cm
Q_0	Discharge	m^3/h
R_{iv}	The ratio of gap length to patch length	—
S	Deposition rate onto bed	$\text{kg}/\text{m}^2/\text{s}$
t	Time	s
U_0 and V	Flow velocity	m/s
VP_1 and VP_3	The starting points of vegetation patches	m
VP_2 and VP_4	The ending points of vegetation patches	m
X	Longitudinal distance along the channel	m
z_b	Bed elevation	m
ρ	Water density	kg/m^3
τ_b	Bed shear stress	N/m^2
ΔU	Local flow velocity change	m/s

1. Introduction

Aquatic vegetation reshapes velocity profiles, turbulence, and boundary shear stress in rivers, making it a central control on sediment transport and deposition (Huai et al., 2019; L. Yang et al., 2023). Once regarded mainly as a source of flow resistance (López & García, 2001), aquatic plants are now recognized as key elements of flood control and river restoration, enhancing biodiversity (Hori et al., 2009) while modulating hydrodynamic and morphodynamic processes (Huai et al., 2021; J. Wang et al., 2024; W.-J. Wang et al., 2019). Elevated local velocity and turbulence around isolated plants or patches promote erosion and constrain patch expansion (Kim et al., 2015; Kitsikoudis et al., 2016; Tranmer et al., 2024), while reduced flow within vegetated areas enhances deposition and reinforces vegetation establishment (Shi et al., 2023; L. Yang et al., 2021; J. Zhang et al., 2020). Recent experiments have improved our understanding of these processes. (Tranmer et al., 2024) showed that patch-adjacent erosion occurs regardless of sediment supply, whereas interior deposition depends strongly on supply conditions, and (Yagci et al., 2024) found that elongation of emergent patches amplifies global scour while restricting far-field deposition, implying hydrodynamic limits on further patch growth. Sediment accumulation around plants thus drives riverbed evolution, sensitive to vegetation density and spatial arrangement (Gao et al., 2023; Huai et al., 2021; X. Liu et al., 2024; Vargas-Luna et al., 2019). Yet the nonlinear interactions among vegetation configuration, flow hydrodynamics, and sediment dynamics, particularly across channels with multiple discontinuous patches, remain incompletely understood, motivating the development of complementary approaches that can efficiently capture emergent behavior.

Process-based modeling of these interactions rests on the Saint-Venant equations coupled with advection-diffusion formulations for sediment transport. Applying these to discontinuous vegetation patches is difficult: vegetation drag requires parameterizing complex wake dynamics, patch heterogeneity creates three-dimensional flow structures beyond one-

dimensional formulations, and deposition-growth feedbacks operate on timescales that challenge numerical stability. High-fidelity simulations that resolve these complexities remain computationally prohibitive for design and scenario exploration. This gap between mathematical rigor and practical feasibility motivates the complementary, data-driven approach pursued here.

More than a century of laboratory research has produced over a hundred empirical sediment-transport formulas (C. T. Yang, 1996; C. T. Yang et al., 2009), yet classical predictors such as Meyer-Peter-Muller and van Rijn struggle under real-world conditions (Haddadchi & Amir A, 2012; Meyer-Peter & Müller, 1948; Van Rijn, 1984), prompting growing use of artificial intelligence, including neural networks and hybrid ML methods, for pattern recognition in complex hydrological datasets (Achite et al., 2022; Habib et al., 2024; Khosravi et al., 2024; Kisi & Yaseen, 2019; Noori et al., 2022; Sahoo et al., 2022). Among supervised approaches, ensemble methods (stacking, voting, bagging, boosting) have gained particular traction for their efficiency, stability, and predictive power over single-model techniques (Abba et al., 2019; Yu et al., 2020; Zounemat-Kermani et al., 2020). Recent work has also combined physical modeling with ML. (Li, 2024) showed that Rouse-like sediment-flux formulations require revision for coarse suspended particles in vegetated flows, while ensemble models such as LightGBM and XGBoost have proven effective for long-term suspended-sediment prediction, and random forest for short-term extremes (Yue et al., 2025)

Despite their predictive strength, ensemble models remain largely "black-box," limiting their contribution to process understanding (El Bilali et al., 2023; Sun & Scanlon, 2019; S. Wang et al., 2022). In vegetated channels, where interactions among configuration, hydrodynamics, and sediment dynamics are still poorly resolved, extracting physically interpretable insight is as important as accuracy, since interpretability enables plausibility checks and translation into design guidance (Ding et al., 2020; Krishnan, 2020; Schmidt et al., 2020). SHAP, grounded in

cooperative game theory (Shapley, 1953), offers a theoretically consistent framework for attributing predictions to input variables (Rodríguez-Pérez & Bajorath, 2020; Viana et al., 2021) and has seen growing use in hydrology and sediment transport (Lund et al., 2022; S. Band et al., 2026; Z. Zhang et al., 2022). (Meddage et al., 2022) used SHAP to show that energy slope has the most impact on bulk-average velocity, while (Yue et al., 2025) found that discharge and precipitation jointly explain roughly 57 % and 19 % of suspended-sediment variability, respectively. Still, studies applying SHAP specifically to sediment deposition in channels with discontinuous vegetation patches remain scarce, and SHAP-derived insights must be interpreted cautiously given the method's inherent disregard for variable interactions.

The existing literature on vegetated channel sedimentation can be organized into three strands (Table 1), each with a distinct limitation this study addresses. First, mechanistic flume experiments provide indispensable physical grounding but remain resource-intensive and configuration-specific, with derived thresholds tied to a single flume geometry and rarely validated across independent setups. Second, predictive ensemble-ML studies achieve strong generalization but offer no pathway from prediction to mechanism, precisely the limitation that motivated the shift toward explainable AI in hydrological modeling more broadly. And finally, SHAP-based interpretability studies demonstrate that feature attribution can be extracted from ensemble models, yet typically reaffirm already-known controls rather than reveal new regime shifts, and none addresses discontinuous, multi-patch vegetation specifically.

Table 1. Overview of existing research on sedimentation in vegetated channels: three main research strands

Experimental / Mechanistic	Study	Focus Area	Methods	Key Strengths	Limitations	Performance	Interpretability
	(Z. Chen et al., 2026)	Suspended sediment carrying capacity in vegetated flow	Diffusion theory + energy theory, 2-D model)	Mechanistically grounded, vegetation-aware formulas; embedded in a 2-D sediment transport model	Theoretical/laboratory closure only; field-scale generalization not established; no ML or interpretability layer	N/A (theory-to-model validation, not ML accuracy)	High (fully mechanistic)
	(Gui et al., 2024)	Turbulence and deposition under floating vegetation	TKE-based deposition model	Connects near-bed TKE to deposition; validated on lab data and real Eichhornia crassipes canopies	Floating-canopy condition only; narrow scope; not transferable to emergent rigid vegetation	N/A (mechanistic model validation)	Moderate-High (mechanistic)
	(Tranmer et al., 2024)	Erosion/deposition around riparian vegetation patches	Full-scale flume experiment	Live emergent willow patch; isolates sediment supply as controlling variable for patch-interior deposition	Single patch geometry; no predictive/ML framework; mechanism not generalized beyond this flume	N/A	High (mechanistic)
	(Yagci et al., 2024)	Patch-pattern effects on wake scour/deposition	Flume experiment	Shows patch geometry/elongation controls global scour and limits far-field deposition	Pattern-specific; no quantitative threshold derived; no cumulative/multi-patch sequence tested	N/A	Moderate (mechanistic)
	(L. Yang et al., 2021)	Stochastic deposition modeling in vegetated channels	Lagrangian random-displacement model	Probabilistic deposition/resuspension rules driven by vegetation-induced TKE; threshold-like behavior	Model-derived, parameterized threshold; not validated as a generalizable, data-driven criterion	N/A	Moderate
	(C. Liu et al., 2021)	Stem size, turbulence and sediment resuspension	Flume experiment	Identifies a critical-turbulence resuspension threshold; shows stem size has little effect vs. density	Sparse emergent canopies and unidirectional flow only; no predictive model	N/A	High
	(P. Kumar & Sharma, 2025b)	Turbulence and bed morphology in partially vegetated channels	3-D flume measurements	Direct link between turbulence reduction inside vegetated zone and enhanced deposition	Single continuous canopy, not discontinuous patch-gap sequence; purely descriptive	N/A	High (mechanistic)
	(P. Kumar & Sharma, 2025a)	Hydrodynamics/bed morphology, double-layered vegetated mobile-bed channel	Mobile-bed flume experiment	Captures genuine two-way flow-sediment feedback in a mobile (erodible) bed, not a fixed-bed approximation	Submerged, continuous-canopy vegetation, not discontinuous emergent patches; no ML/interpretability	N/A	Moderate

Table 1. (continued)

Study		Focus Area	ML Methods	Key Strengths	Limitations	Performance	Interpretability
Prediction-focused ML / Black-box	(S. Kumar et al., 2026)	Vegetation-influenced stream-velocity estimation	REPT, RF, RT and hybrid MS/RC variants	Broad model comparison; hybrid models outperform standalone models and empirical equations	Strongly metric-dependent; interpretability limited; velocity-focused, not deposition	Hybrid RF-based models best on reported metrics (correlation, NSE, AI, RMSE, SI, PBias, KGE)	Low-Moderate
	(Jia et al., 2025)	Vegetation drag-coefficient prediction	9 models incl. SVM, XGBoost, BP, PSO/WSO/RIME-optimized variants	Large 490-sample database; robust under missing-parameter scenarios	Accuracy-focused; sensitivity analysis only, no SHAP-based mechanistic discovery	XGBoost best among 9 models (exact metric not isolated)	Moderate
	(Bezak et al., 2025)	Catchment-scale suspended sediment load under climate change	XGBoost (with rainfall-runoff model)	Integrates RCP2.6/4.5/8.5 climate scenarios; strong catchment-scale application; outperforms sediment rating curve	Catchment-scale, not vegetated channels; no SHAP or interpretability analysis	XGBoost ~15 % lower bias than sediment rating curve	Low (no interpretability framework)
SHAP-based Interpretable ML	(Li et al., 2024)	Coarse suspended sediment closure in vegetated turbulence	Physics-constrained ML (diffusive/settling-flux constraints)	Embeds physical constraints directly in the ML closure; improves SSC prediction near canopy edges	Scope limited to tested canopy-density/particle-size regime; not benchmarked against multiple ensemble algorithms	Not reported as single metric (closure-based)	Moderate-High
	(Basumatary et al., 2025)	Bed load transport prediction	RF, Extra Trees, XGBoost, CatBoost	SHAP waterfall/summary plots purpose-built for interpretation, not just accuracy reporting	General hydraulic/sediment context, not vegetated channels; no engineered spatial-memory features	XGBoost highest accuracy (RMSE, NSE, PCC, R ² reported)	High (dedicated SHAP framework)
	(Mir et al., 2024)	Bedload transport prediction	XGBoost, RF, ETR, KNN, LR, BR (metaheuristic-optimized)	Multi-algorithm ensemble comparison; metaheuristic tuning; basic SHAP attribution; Shields-parameter sensitivity	Limited to bedload (no vegetation context); no dimensionless/scale-independent predictors; SHAP used only for ranking	XGBoost: R ² = 0.99, RMSE = 0.11	Limited (basic SHAP)
	Current Study	Sediment deposition in discontinuous emergent vegetated channels	RF, GB, XGB, LightGBM, CatBoost (Bayesian-optimized)	Cross-experiment dataset; novel spatial-memory features; SHAP used for threshold discovery, not just ranking; threshold matches theoretical $1/(C_D \cdot a_v)$ scale	Rigid emergent vegetation and steady flow only; field-scale validation still required	CatBoost: R ² = 0.953 ± 0.009 , RMSE = 0.149, IA = 0.987 (unseen holdout)	SHAP-derived physical threshold validated against canopy hydrodynamic theory

This study is positioned at the intersection of these three strands. Relative to the flume literature, it draws on a varied dataset, spanning patch length, inter-patch spacing, density, and water depth across multiple experimental runs, benchmarked against the theoretical adjustment length scale $1/(C_D a_v)$, allowing broader exploration of configurations than additional flume construction would permit. Critically, encoding these configuration parameters as engineered model inputs, rather than treating each combination as a separate physical experiment to be constructed and instrumented, is intended to make it computationally feasible to evaluate a far broader space of vegetation-patch configurations than would be practical through additional flume construction alone, since each new configuration in a machine-learning framework requires only a new combination of input values rather than a new physical setup, instrumentation, and data-collection cycle. Relative to predictive ensemble-ML work, Bayesian-optimized, cross-validated benchmarking of five algorithms tests whether interpretability can be achieved without sacrificing accuracy. Relative to the SHAP literature, SHAP is applied here not as a post hoc ranking exercise but as a hypothesis-generation tool, probing for unspecified regime shifts, an approach not yet undertaken for vegetated-sediment systems, supported by two engineered features encoding the spatial memory of sequential patch-gap arrangements and translating recent mechanistic findings into a predictive, threshold-generating framework. Building on these contributions, the objectives of this study are to:

- Benchmark five ensemble machine learning algorithms (Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost) for predicting sectional sediment deposition under varying vegetation and flow conditions;
- Employ SHAP analysis to identify the dominant physical controls on deposition and the mechanistic thresholds governing deposition-promoting versus deposition-inhibiting regimes;

- Demonstrate how interpretable ML can generate hypotheses and advance understanding of sedimentation in channels with discontinuous vegetation patches, as a computationally efficient complement to laboratory and high-fidelity numerical investigations.

The paper is structured as follows: The data and methodology are explained in Section 2, the results and comparison are presented in Section 3, the results are discussed and interpreted in Section 4, and the main conclusions are presented in Section 5.

2. Materials and methods

2.1. Experimental dataset

The laboratory dataset reported by (Y. Liu et al., 2025), which investigates sediment deposition in an open-channel flume with discontinuous, emergent vegetation patches, was used in this study. The flume experiments were conducted under controlled hydraulic conditions with rigid vertical cylinders (acrylic rods) of uniform diameter (5 mm), arranged in staggered configurations to mimic emergent vegetation. The dataset contains 498 longitudinal observations of sediment deposition collected from 11 experimental runs (Y. Liu et al., 2025). (please refer to the original reference for detailed flume dimensions and measurement protocols). For reproducibility, all variables are expressed in consistent SI units, and the same experimental case labels are retained for data partitioning.

2.2. Feature engineering and new variables

To model the influence of discontinuous vegetation patches on sediment deposition, several physically based features were derived from the original experimental variables. The first group of features describes the spatial location of each measurement point relative to the vegetation patches. A second group of features captures vegetation effects on flow and sediment transport.

A summary of the new features is provided in Table 2 (more detailed derivations are given in the Supplementary Materials).

Table 2. Summary of input features used in machine learning models

Feature Name (Units)	Description	Derivation
Longitudinal distance, X (m)	Distance along the flume from the upstream reference	Direct measurement from experiments.
Relative Position in Vegetation (-)	Normalized position within a vegetation patch	e.g. $(X - 0.5)/0.5$ for the first patch, or -1 for points outside vegetation
Distance to Vegetation (m)	Shortest longitudinal distance to nearest patch edge	upstream of the 1 st patch: $VP_1 - X$; inside a patch: 0; between patches: $\min(X - VP_1, VP_2 - X)$; downstream of the 2 nd patch: $X - VP_3$
Cumulative Vegetation Exposure (m)	Total length of vegetation upstream of the measurement point	Sum of patch lengths encountered upstream.
Upstream Turbulence Index (-)	Ordinal index reflecting cumulative turbulence from upstream patches	0 (upstream of patches), 1 (after first patch), 2 (after both patches).
Vegetation Presence	Whether the point is inside the vegetation patch or not	point inside a vegetation patch (1) or outside the patch (0)
Vegetation Drag Index (m^2/s)	Aggregated vegetation-induced resistance	$Presence \times a_v \times U_0 \times H/100$
Local Flow Velocity (m/s)	Effective flow velocity at the measurement point	$c \cdot U_0$ inside vegetation; U_0 outside.

Intermediate features such as "*Vegetation Presence*" and "*Vegetation Density*", while essential for constructing the "*Vegetation Drag Index*" and related variables, were not included simultaneously as direct model inputs to avoid redundancy and unnecessary dimensionality. The final set of input features was selected as follows: "*longitudinal distance X (m), Relative Position in Vegetation, Distance to Vegetation (m), Cumulative Vegetation Exposure (m), Upstream Turbulence Index, Vegetation Drag Index, and Local Flow Velocity (m/s)*". The target variable is the normalized sediment deposition rate (Dep/Dep_{avg}).

All variables are expressed in SI units. The dimensionless target (Dep/Dep_{avg}) eliminates scale dependence across experiments. No data transformation (e.g., log) was applied because tree-based models are robust to non-normality. Descriptive statistics for the seven input features

and the target variable, along with histograms (including kernel density estimates), scatter plots, and the correlation matrix, are provided in the Supplementary Materials.

2.2.1. Feature independence assessment

To evaluate the independence of the seven input features (a requirement for stable SHAP feature importance), the variance inflation factor (VIF) for each feature were computed. VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. As a general rule, $VIF > 5$ indicates moderate multicollinearity, and $VIF > 10$ indicates severe multicollinearity that may distort feature importance rankings (James et al., 2021; Kutner, 2005; O'brien, 2007).

Table 3 presents the VIF values calculated from the 498 observations. Most features exhibit VIF values below 5, indicating acceptable collinearity. However, *"Relative Position in Vegetation"* shows a VIF of 9.10, which exceeds the commonly used threshold of 5. This elevated VIF arises because this feature is mathematically related to several other vegetation-geometry variables (e.g., *"Cumulative Vegetation Exposure, Distance to Vegetation"*). Consequently, SHAP importance for this feature should be interpreted with caution; its contribution may be partially shared with correlated features. Nevertheless, the remaining six features have $VIF < 5$, and the overall model performance remains strong (see Section 3). Therefore, all seven features were retained to preserve physical interpretability, while the observed collinearity was acknowledged as a limitation.

205

Table 3. Variance Inflation Factors (VIFs) for the seven input features

Feature (Units)	VIF
Longitudinal distance, X (m)	4.57
Relative Position in Vegetation (-)	9.10
Distance to Vegetation (m)	2.30
Cumulative Vegetation Exposure (m)	3.84
Upstream Turbulence Index (-)	2.97
Vegetation Drag Index (m ² /s)	4.78
Local Flow Velocity (m/s)	4.48

206

2.2.2. Justification of feature simplifications

207

It is acknowledged that the engineered features (e.g., "*Upstream Turbulence Index*" as an

208

ordinal variable and "*Local Flow Velocity*" based on a constriction coefficient, c) are

209

simplifications of complex three-dimensional turbulent processes. However, two points justify

210

their adequacy for the present laboratory conditions. First, in emergent rigid vegetation with

211

negligible stem Reynolds number effects, the dominant turbulent structures are shear-layer

212

vortices at the canopy top (Ghisalberti & Nepf, 2002) and wake turbulence behind individual

213

stems (Nepf, 1999); the ordinal index (0, 1, and 2) captures the cumulative effect of successive

214

patch leading edges; a first-order approximation validated by the strong predictive performance

215

(Section 3). Second, the "*Local Flow Velocity*" feature uses a calibrated constriction coefficient

216

($c \approx 0.2$) that reproduces measured velocities within $\pm 15\%$ error based on validation against

217

point velocity measurements reported in the original experimental study. For the present

218

laboratory conditions, these features are sufficient to capture the dominant controls on sediment

219

deposition.

220

221

2.3. Ensemble machine learning models

A suite of ensemble machine learning algorithms was employed to capture nonlinearities, threshold effects, and generalizability in sediment deposition prediction. The ensemble models can be broadly categorized into two groups on the basis of their underlying learning strategy: bootstrap aggregating (bagging) and boosting.

- Bagging reduces model variance by training multiple base learners independently on bootstrapped subsets of the data and aggregating their predictions. Random Forest was selected as the primary bagging-based model in this study.
- Boosting sequentially builds learners, with each new model focusing on correcting errors made by previous ones, thereby improving predictive accuracy and handling complex patterns. In this work, several state-of-the-art boosting algorithms, including Gradient Boosting, XGBoost, LightGBM, and CatBoost, were evaluated.

Decision Tree was used as baseline model, despite its known tendency to overfit, to contrast its performance with that of ensemble approaches. The specific models used in this study are as follows (more details of each model are provided in the Supplementary Materials):

- Random Forest: variance reduction by bagging and random feature sampling (Heddam, 2021b, 2021a).
- Gradient Boosting: residual-based boosting framework (Friedman, 2001; Géron, 2022).
- XGBoost: efficient boosting with regularization and parallelization (T. Chen & Guestrin, 2016; Georganos et al., 2018).
- LightGBM: computationally efficient boosting for high-dimensional data (Al Mamlook et al., 2020; T. Chen et al., 2020).

- CatBoost: ordered boosting to reduce gradient bias and improve generalizability (Friedman, 2002; Prokhorenkova et al., 2018).

This diverse selection enables systematic comparisons of predictive accuracy, computational efficiency, and interpretability for modeling.

2.4. Data partitioning strategy

To rigorously evaluate the predictive and generalization capabilities, the dataset was partitioned at the experimental level. All observations from experiment A2 were designated as a holdout (unseen) dataset and were excluded from both model training and hyperparameter tuning. This holdout scenario provides a stringent test of model performance under new hydraulic and vegetation conditions and reduces the risk of overfitting. From the remaining 10 experiments, the data were randomly split into 70 % for training and 30 % for testing. Random partitioning was performed via a fixed random seed to ensure reproducibility. All the models were trained under identical conditions to ensure fair comparison.

2.5. Hyperparameter optimization and model evaluation

The hyperparameters were optimized via Bayesian optimization, with 5-fold cross validation applied during tuning, which efficiently explores parameter interactions compared with grid or random search (Bergstra et al., 2013; Snoek et al., 2012). The parameter ranges are summarized in Table 4. The performance was evaluated via evaluation metrics as described in Table 5.

Table 4. Hyperparameters of the machine learning models used in this study

Models	Hyperparameters [Ranges]
Random Forest	Maximum number of features considered for splitting [1, number of features]
CatBoost	Learning rate [0.01, 0.5] - Subsample ratio [0.5, 1]
Gradient Boosting	
Light GBM	Learning rate [0.01, 0.5] - Subsample ratio [0.5, 1] - Column sample by tree [0.5, 1.0] - L1 regularization [0, 10] - L2 regularization [0, 10]
XGBoost	Learning rate [0.01, 0.5] - Subsample ratio [0.5, 1] - Column sample by tree [0.5, 1.0] - Gamma [0, 5] - L1 regularization [0, 10] - L2 regularization [0, 10]

*Note: For all models: Minimum number of samples in a leaf [1, 20] - Minimum number of samples required to split a node [2, 20] - Maximum depth of each tree [3, 30] - Number of estimators [100, 1000]

Table 5. Evaluation metrics and their mathematical formulations

Evaluation Metrics	Equation
Coefficient of determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$
Mean absolute percentage error (MAPE)	$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{ y_i - \hat{y}_i }{ y_i } \times 100 \% \quad (3)$
Scatter Index (SI)	$SI = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n ((y_i - \hat{y}_i) - (y_i - \bar{y}))^2}}{\frac{1}{n} \sum_{i=1}^n y_i} \quad (4)$
Bias	$Bias = \frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i \quad (5)$
Index of Agreement (IA)	$IA = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y} + y_i - \hat{y}_i)^2} \quad (6)$

*Note: n represents the number of samples, y_i represents the observed values, $\hat{y}_i$ represents the values of the i -th sample predicted by the models, $\bar{y}$ represents the mean of the observed values, and $\hat{\bar{y}}$ represents the mean of the predicted values.

2.6. Shapley additive explanations (SHAP)

In this study, SHAP was used to: (i) quantify the global importance of spatial, hydrodynamic, and vegetation-related features; (ii) examine feature effects and interactions at the local (sample) scale; and (iii) assess whether the learned relationships are physically consistent with the established understanding of vegetation–flow–sediment interactions. These diagnostics complement traditional performance metrics by revealing nonlinear and threshold-type

behavior, building on recent applications of interpretable ML in hydrology (Ding et al., 2020; El Bilali et al., 2023; Schmidt et al., 2020; S. Wang et al., 2022).

SHAP decomposes predictions into additive feature contributions based on Shapley values from cooperative game theory (Shapley, 1953), providing both magnitude and direction of effects. This capability is particularly relevant in hydro-sedimentary contexts, where nonlinear thresholds (e.g., cumulative vegetation exposure vs. deposition) govern deposition. SHAP was chosen over alternatives such as LIME or partial dependence because of its consistency and quantitative rigor. Importantly, SHAP insights were cross-validated against established sediment transport theory, ensuring that model explanations align with physical plausibility (Lund et al., 2022; Molnar et al., 2020; Rahaeimehr et al., 2025; Štrumbelj & Kononenko, 2014; S. Wang et al., 2022). The following section presents the results of model training, evaluation, and interpretability analysis, highlighting the comparative performance of ensemble learners and the physical insights derived from SHAP explanations. The overall methodological framework adopted in this study is illustrated in the Figure 1.

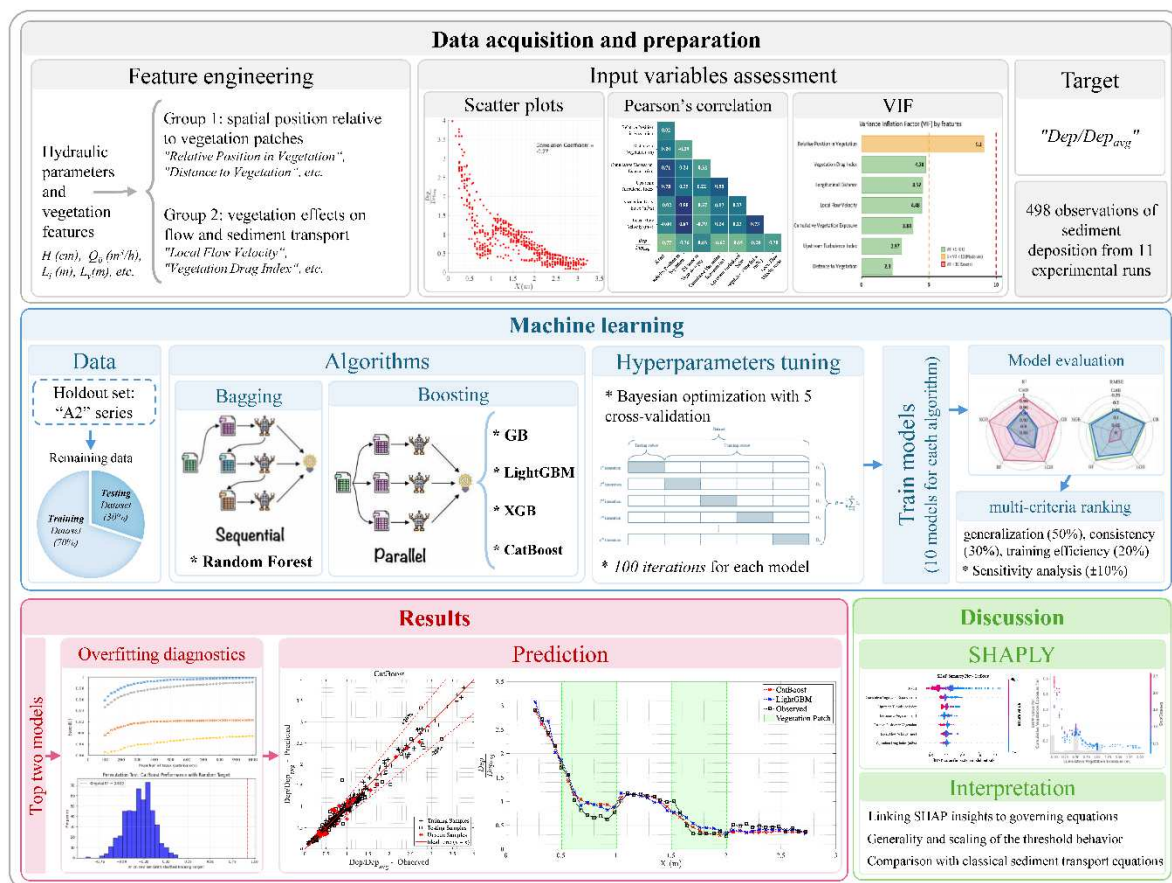


Figure 1. Workflow of the proposed ensemble machine learning and SHAP framework

3. Results

3.1. Overall model performance

Figure 2 summarizes the performance of the five ensemble machine learning models; CatBoost (CatB), Gradient Boosting (GB), LightGBM (LGB), Random Forest (RF), and XGBoost (XGB), for predicting longitudinal sediment deposition rates across the datasets. During training, all the models achieved high predictive accuracy, with R^2 values ranging from approximately 0.988 for Random Forest to 0.997 for XGBoost, reflecting their ability to capture the complex relationships among vegetation, flow, and sediment variables. XGBoost exhibited the best apparent fit while Gradient Boosting, CatBoost, and LightGBM also performed strongly, indicating the effectiveness of boosting-based ensemble methods in this application.

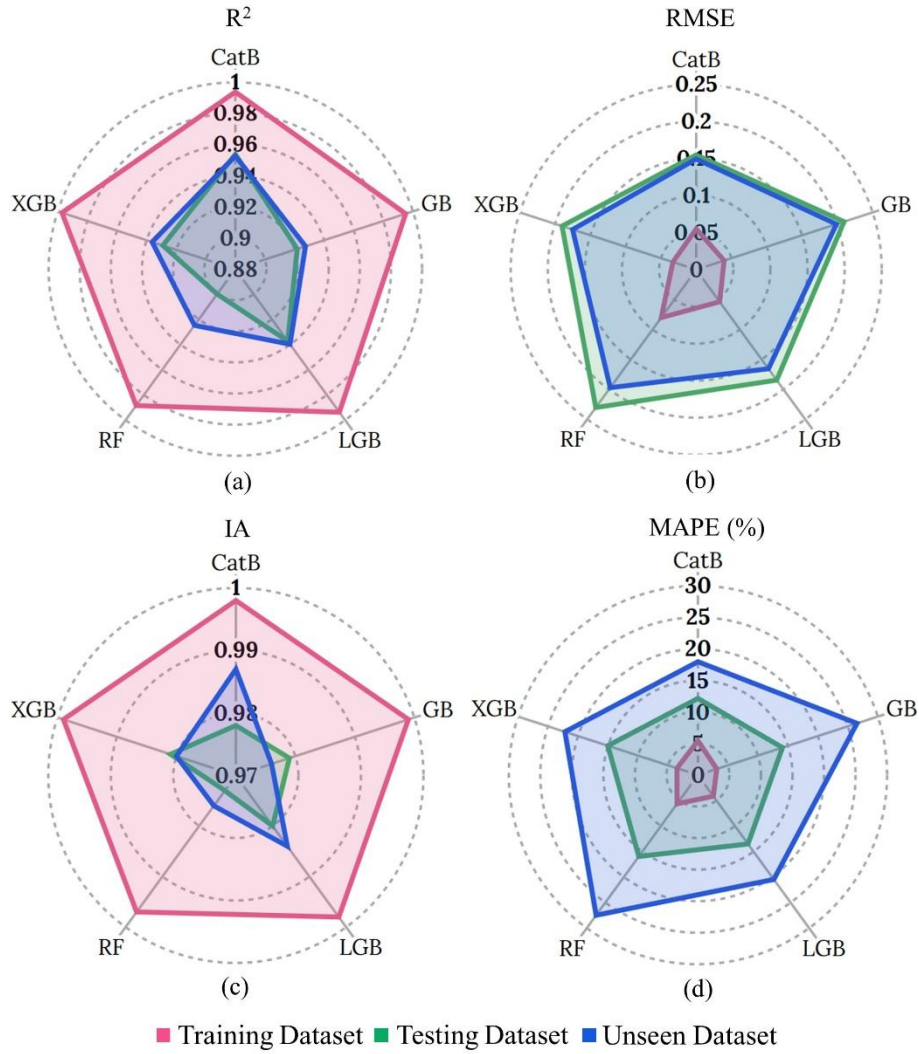


Figure 2. Performance evaluation of five machine learning models; (a) R^2 , (b) RMSE, (c) IA, and (d) MAPE (%)

However, testing phase revealed critical differences in generalization capability beyond training performance. CatBoost delivered the most robust predictions, achieving the highest $R^2 \approx 0.954$ and lowest RMSE, IA, and MAPE on the test set, followed by LightGBM, XGBoost, and Gradient Boosting.

The unseen holdout (Experiment A2) provided the ultimate generalization test. CatBoost again led with $R^2 \approx 0.953$ and lowest errors, demonstrating exceptional extrapolation to novel vegetation-flow conditions. LightGBM and XGBoost maintained strong performance ($R^2 > 0.93$), while Gradient Boosting and Random Forest exhibited declining accuracy.

3.2. Multicriteria model ranking

To objectively rank the models, a multicriteria evaluation framework was employed, with 50 % weight assigned for generalization, 30 % weight assigned for consistency, and 20 % weight assigned for training efficiency, reflecting practical priorities in sediment progradation modeling. Key metrics, such as the normalized R^2 , RMSE, IA, and MAPE of the unseen dataset, form the basis of generalized scoring (Table 6). Model consistency was assessed through the standard deviation of these metrics across datasets (Table 7), whereas training efficiency was evaluated by performance on training data (Table 8), with the recognition that efficiency has less influence than does predictive reliability. Sensitivity analyses (Table 9) confirmed the stability of model rankings under varying weights, underscoring the robustness of the selection criteria.

Table 6. Generalization performance of the machine learning models

Models	Normalized evaluation metrics				
	R^2	RMSE	IA	MAPE	Avg
CatB	1.000	1.000	1.000	1.000	1.000
GB	0.973	0.752	0.989	0.677	0.848
LGB	0.986	0.902	0.997	0.878	0.941
RF	0.971	0.758	0.989	0.654	0.843
XGB	0.983	0.853	0.993	0.810	0.910

Table 7. Consistency scores of the evaluated models

Models	Normalized standard deviation of evaluation metrics				
	R^2	RMSE	IA	MAPE	Avg
CatB	1.000	1.000	0.944	1.000	0.986
GB	0.574	0.591	0.756	0.536	0.614
LGB	0.734	0.795	1.000	0.766	0.824
RF	0.513	0.712	0.723	0.574	0.631
XGB	0.626	0.648	0.884	0.665	0.706

333 Table 8. Training efficiency scores of the evaluated models

Models	Normalized evaluation metrics				
	R ²	RMSE	IA	MAPE	Avg
CatB	0.997	0.602	0.999	0.578	0.794
GB	0.998	0.836	1.000	1.000	0.958
LGB	0.997	0.616	0.999	0.757	0.842
RF	0.991	0.413	0.998	0.557	0.740
XGB	1.000	1.000	1.000	0.898	0.974

334

335 Table 9. Result of the sensitivity analysis

Models	Multiplied Scores						
	Supposed weights	Sensitivity analysis ($\pm 10\%$ in weight assignments)					
	50-30-20	40-30-30	40-40-20	60-30-10	60-20-20	50-40-10	50-20-30
CatB	0.955	0.934	0.953	0.975	0.956	0.974	0.935
GB	0.800	0.811	0.777	0.789	0.823	0.766	0.834
LGB	0.886	0.876	0.874	0.896	0.898	0.884	0.888
RF	0.759	0.748	0.737	0.769	0.780	0.748	0.769
XGB	0.861	0.868	0.841	0.855	0.882	0.835	0.888

336 In summary, this rigorous assessment validates CatBoost as the optimal model for accurately
337 and stably predicting sediment deposition rate, balancing high precision, reliability, and
338 operational efficiency. LightGBM complements this with superior adaptability and reduced
339 uncertainty, establishing these ensemble methods as practical tools for advanced sediment
340 transport modeling.

341 Model ranking consistency across phases highlights modern gradient boosting's suitability for
342 vegetated channel deposition, where nonlinear interactions dominate. CatBoost and LightGBM
343 excelled through ordered boosting and robust regularization, maintaining stable accuracy from
344 training to unseen data. XGBoost showed training overfitting (near-perfect fit $\rightarrow$ larger unseen
345 degradation), while RF underperformed.

346 **3.2.1. Overfitting diagnostics**

To address concerns about overfitting given the modest dataset size (498 observations), three complementary diagnostics were performed, each targeting a different aspect of model generalization.

First, learning curves were generated for CatBoost and LightGBM by increasing the number of trees from 100 to 1000 while monitoring the test R^2 on the 70/30 split of the 10 training experiments (unseen holdout A2 excluded). As shown in Figure 3, for both models, the test R^2 increased rapidly during the first few hundred trees and then gradually levelled off. Critically, the test R^2 never decreased with additional trees; it remained stable or increased only marginally beyond 500 trees (CatBoost: from 0.922 at 500 trees to 0.923 at 1000 trees; LightGBM: from 0.886 at 500 trees to 0.895 at 1000 trees). Meanwhile, the training R^2 continued to improve slowly (e.g., CatBoost training R^2 from 0.995 to 0.998). This behavior is characteristic of a well generalized boosting model that means adding more trees does not degrade performance on unseen data, which is a strong indicator that overfitting is not occurring.

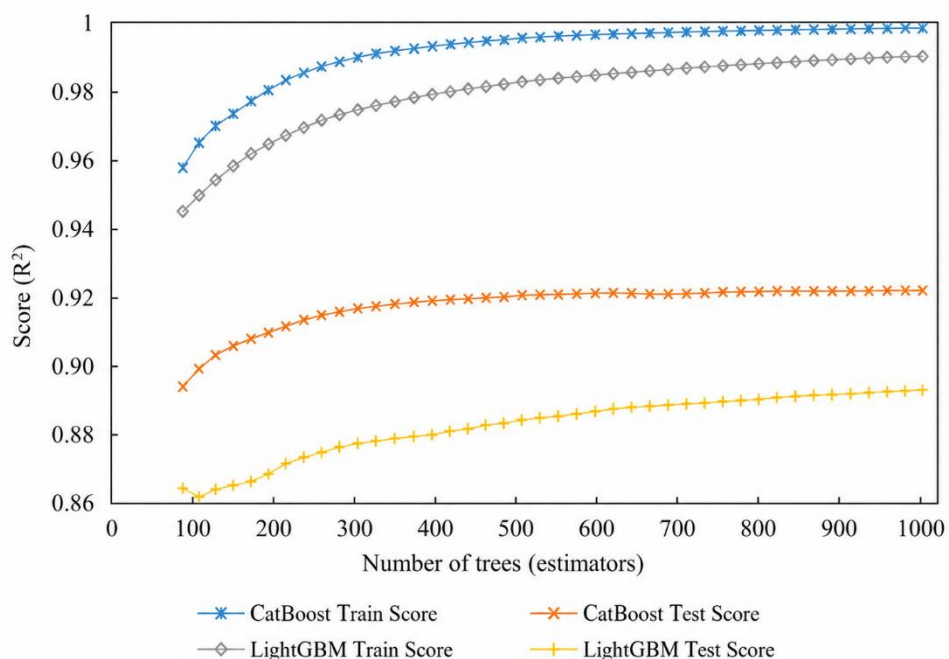
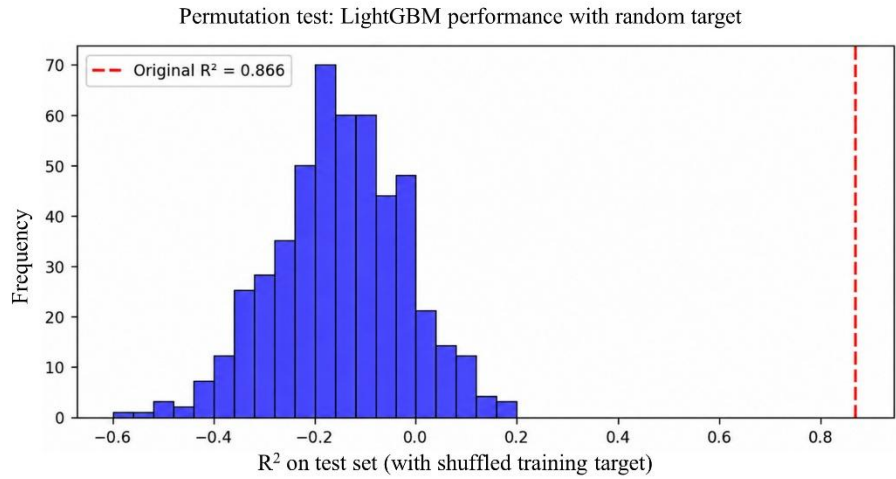
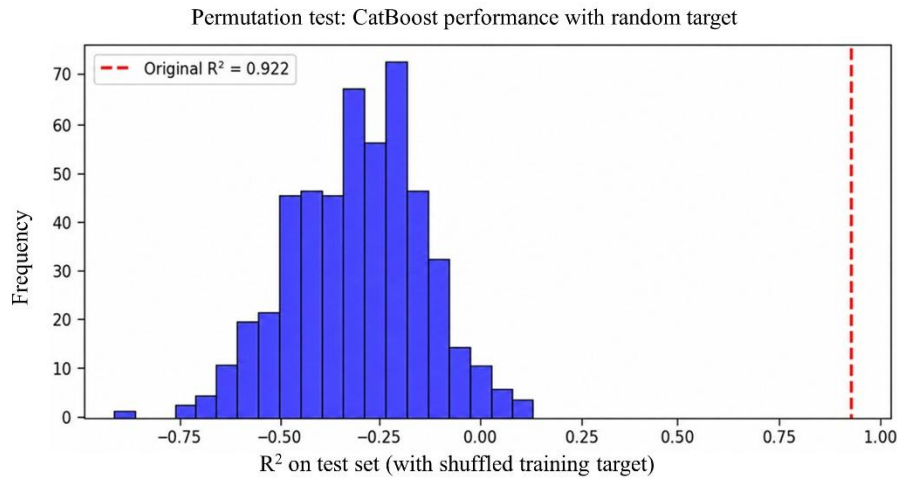


Figure 3. Learning curves of the CatBoost and LightGBM models

363 Second, permutation tests were performed (Figure 4) to verify that the models capture genuine
 364 physical relationships rather than spurious correlations. The target variable was randomly
 365 shuffled 500 times while keeping the features unchanged. For each shuffled training set, the
 366 model (with the same hyperparameters) was retrained and evaluated on the original
 367 (unshuffled) test set. The original CatBoost model achieved a test R^2 of 0.922, whereas the
 368 shuffled models yielded a mean R^2 of -0.301 ± 0.160 , with a maximum of 0.134. For
 369 LightGBM, the original test R^2 was 0.866, while the shuffled models gave a mean R^2
 370 of -0.149 ± 0.128 , with a maximum of 0.200. In both cases, the near-zero or negative mean R^2
 371 of the shuffled models confirms that the high predictive performance of the original models is
 372 not due to chance or random correlations.



(a)



(b)

Figure 4. Results of permutation test for (a) LightGBM and (b) CatBoost

Third, model stability against random initializations was assessed by repeating the entire training-evaluation cycle (including the 70/30 split) 10 times on the same 10 experiments, each time with a different random seed, and then evaluating each trained model on the completely unseen holdout (Experiment A2). The low standard deviations (CatBoost $R^2 = 0.953 \pm 0.009$, RMSE = 0.149 ± 0.019 , IA = 0.987 ± 0.003) confirm that the model's performance is stable and not an artefact of a particular random seed.

Additionally, Bayesian hyperparameter optimization with 5-fold cross-validation (Section 2.5) was used during model development, which inherently reduces overfitting by validating on multiple data subsets. Collectively, these diagnostics provide strong evidence that the reported generalization is not an artefact of overfitting and that the models reliably learn physically meaningful patterns from the data.

3.3. Predicting "*Dep/Dep_{avg}*" via ML models

Figure 5 compares the experimental measurements of "*Dep/Dep_{avg}*" with predictions from two leading machine learning models: CatBoost and LightGBM. The CatBoost predictions closely follow the ideal 1:1 line across training, testing, and unseen datasets, remaining within $\pm 20\%$ error margins. This consistency underscores CatBoost's superior precision and stability in capturing the nonlinear relationships inherent in sediment deposition dynamics. In contrast, LightGBM predictions exhibit greater scatter, particularly in the testing data, indicating higher variance and sensitivity to noise or suboptimal parameter tuning. This visual analysis underscores CatBoost's superior precision, consistent with quantitative metrics.

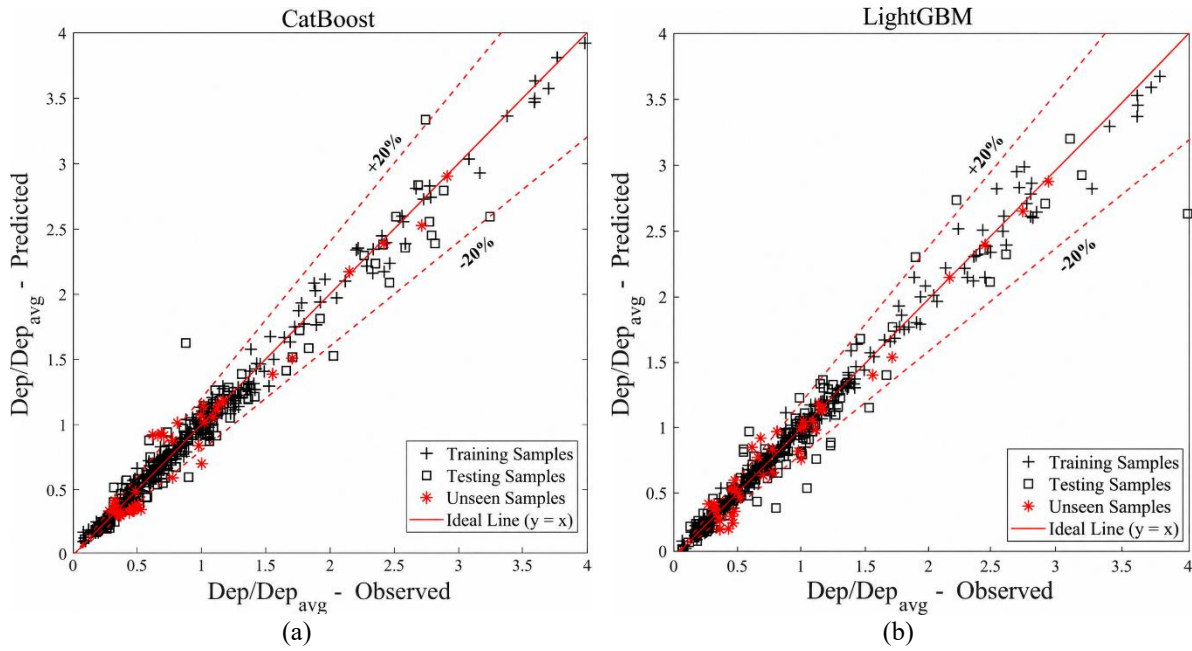


Figure 5. Relationship between experimental and predicted sediment deposition rates for (a) CatBoost and (b) LightGBM models

The observed variability in LightGBM may result from differences in tree construction methods or regularization strategies, suggesting the need for further hyperparameter optimization to enhance its consistency. The performance metrics presented in Table 9 further substantiate these observations. This trade-off in performance stresses the importance of choosing a model on the basis of specific application needs, whether prioritizing accuracy or generalizability in hydrological engineering issues.

Table 9. Evaluation metrics (mean $\pm$ standard deviation) for CatBoost and LightGBM models

Model	Dataset	R^2	RMSE	MAPE (%)	IA	BIAS	SI
CatBoost	Training	0.994 $\pm$ 0.004	0.055 $\pm$ 0.019	5.429 $\pm$ 1.865	0.998 $\pm$ 0.001	-5.35E-06 $\pm$ 1.024E-5	0.062 $\pm$ 0.023
	Testing	0.954 $\pm$ 0.011	0.155 $\pm$ 0.023	12.150 $\pm$ 1.329	0.978 $\pm$ 0.013	-0.009 $\pm$ 0.017	0.218 $\pm$ 0.062
	Unseen	0.953 $\pm$ 0.009	0.149 $\pm$ 0.019	17.967 $\pm$ 2.180	0.987 $\pm$ 0.003	0.011 $\pm$ 0.015	0.164 $\pm$ 0.018
LightGBM	Training	0.994 $\pm$ 0.005	0.054 $\pm$ 0.023	4.146 $\pm$ 1.634	0.998 $\pm$ 0.002	-2.54E-04 $\pm$ 8E-4	0.072 $\pm$ 0.028
	Testing	0.937 $\pm$ 0.013	0.185 $\pm$ 0.029	13.507 $\pm$ 0.922	0.980 $\pm$ 0.005	-0.006 $\pm$ 0.025	0.216 $\pm$ 0.027
	Unseen	0.940 $\pm$ 0.016	0.166 $\pm$ 0.022	20.466 $\pm$ 2.829	0.984 $\pm$ 0.004	0.041 $\pm$ 0.034	0.174 $\pm$ 0.022

Figure 6 shows model predictions on the unseen dataset, confirming that both CatBoost and LightGBM effectively capture sediment deposition rate trends over the spatial coordinate " Dep/Dep_{avg} " in meters. The analysis reveals that both models broadly capture the overall trend of the observed data, displaying a steep initial decline, intermittent fluctuations, and stabilization at lower values toward the latter spatial range. Both the CatBoost and LightGBM predictions generally track this trend, but their accuracies vary across distinct spatial regions, particularly within the vegetation patches identified in the plot at $X = 0$ to 0.5 m and $X = 1.5$ to 2 m, including a rapid drop in the first patch and a pronounced dip in the second.

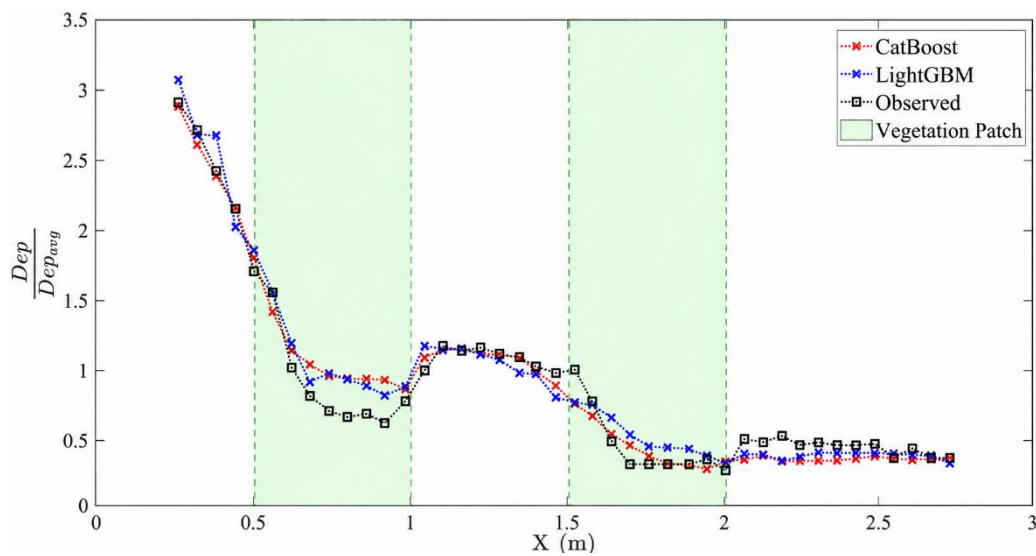


Figure 6. Predicted vs. observed " Dep/Dep_{avg} " for the unseen dataset (Experiment A2) via the CatBoost and LightGBM models

The vegetation patches, highlighted as critical regions, appear to challenge the models' predictive accuracy, as the abrupt changes in " Dep/Dep_{avg} " are not perfectly replicated, suggesting that the influence of vegetation may not be fully encapsulated in the models' feature sets or training data. Comparing the two models, CatBoost consistently outperforms LightGBM by delivering predictions that are closer to the observed data with reduced scatter, indicating superior generalizability to this unseen dataset. In contrast, LightGBM's greater variability and tendency to overestimate, particularly in vegetation patches and transition

zones, may reflect higher sensitivity to noise or potential overfitting during training. This differential performance underscores the importance of model-specific tuning and feature engineering to enhance predictive fidelity on complex, unseen spatial data.

Overall, both CatBoost and LightGBM demonstrate reasonable predictive performance across the transect, effectively capturing the general declining trend and vegetation patch effects. CatBoost exhibits a slightly closer fit to the observed values, particularly in the later sections, suggesting a potential advantage in modeling stabilized deposition. LightGBM, while showing greater variability and overestimation in the early sections, aligns well in the mid-to-late range.

Error patterns also align with physical expectations. Models tend to achieve higher accuracy in regions where flow and vegetation conditions are more homogeneous and feature signals are strong (e.g., within well-developed patches or far from patch edges), whereas discrepancies are typically greater near transitions between vegetated and non-vegetated zones, or where local turbulence and recirculation effects are more complex. These transition regions are known to present challenges for both empirical formulas and numerical models, and the residual errors here emphasize the importance of interpretable diagnostics, such as SHAP, for understanding how vegetation structure and hydrodynamics jointly influence sediment deposition.

4. Discussion

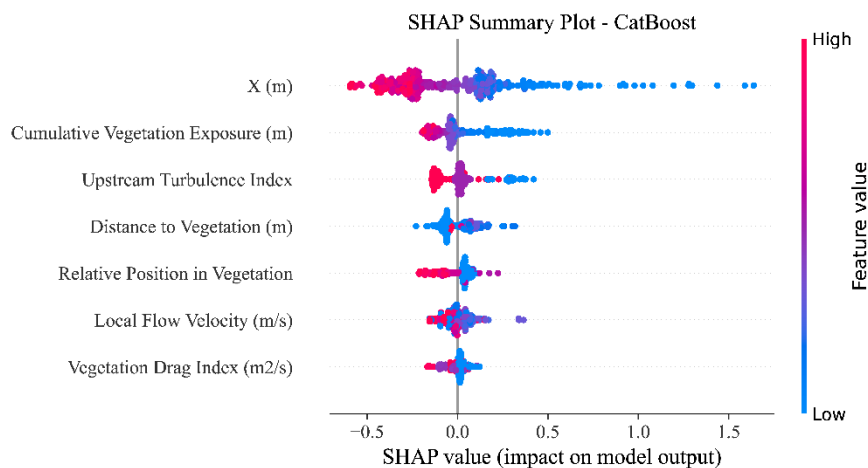
4.1. Investigation of the effects of input variables on " Dep/Dep_{avg} "

Global SHAP importance (as shown in Figure 7), measured as the mean absolute SHAP value is summarized in Table 10. For both CatBoost and LightGBM, longitudinal distance " X " is the most influential feature (mean $|SHAP| = 0.297$ and 0.352 , respectively). "*Cumulative Vegetation Exposure*" ranks second in both models (0.116 in CatBoost, 0.205 in LightGBM), confirming that the total length of vegetation encountered upstream strongly controls deposition. This ranking is physically reasonable, as it reflects the progressive adjustment of

the flow and sediment fields along the flume, combined with the cumulative effects of vegetation patches on velocity reduction, turbulence generation, and residence time. However, the ranking of other features differs between the two models. In CatBoost, the *"Upstream Turbulence Index"* ranks third (0.081), followed by *"Distance to Vegetation"* (0.074). In LightGBM, *"Distance to Vegetation"* ranks third (0.122), while the *"Upstream Turbulence Index"* ranks sixth (0.042). Despite these differences, *"Local Flow Velocity"* consistently shows weak influence in both models (0.049–0.063), supporting the conclusion that within densely vegetated patches, velocity is so effectively reduced that it is not a primary predictor. The *"Vegetation Drag Index"* has the smallest importance in both models (0.033 to 0.027). Table 11 summarizes SHAP dependencies and model divergences for input features.

Table 10. Mean absolute SHAP values (feature importance) for CatBoost and LightGBM models

Feature	CatBoost	LightGBM
X (m) – Longitudinal distance	0.297	0.352
Relative Position in Vegetation	0.068	0.075
Distance to Vegetation (m)	0.074	0.122
Cumulative Vegetation Exposure (m)	0.116	0.205
Upstream Turbulence Index	0.081	0.042
Vegetation Drag Index (m ² /s)	0.033	0.027
Local Flow Velocity (m/s)	0.049	0.063



(a)

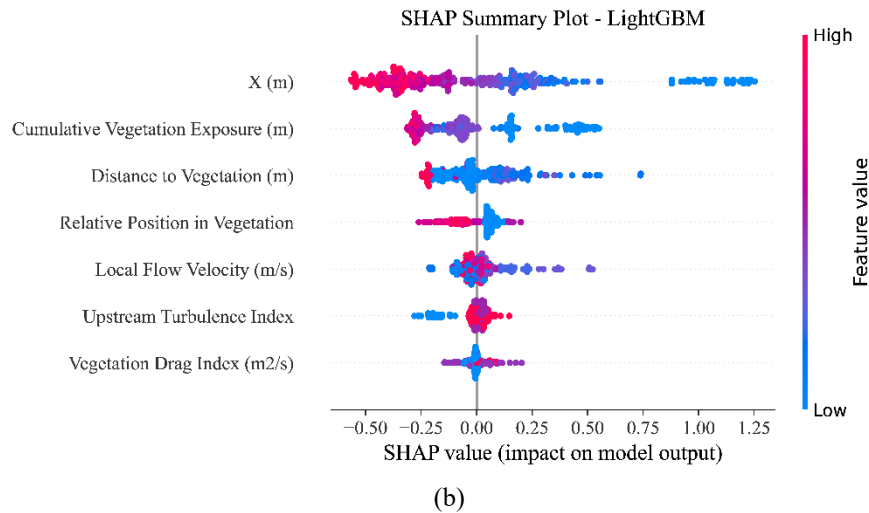


Figure 7. SHAP summary plots for (a) CatBoost (b) LightGBM

Table 11. Summary of SHAP dependence patterns for key hydrodynamic and vegetation-related features

Feature	High-Value Effect	Physical Mechanism	Model Divergence
X (Downstream Distance)	↓ Deposition	↑ Velocity /Turbulence/Resuspension	CatBoost captures nonlinear downstream effects
Cumulative Veg. Exposure	↓ Deposition	Vegetation-induced turbulence	Consistent high importance across models
Upstream Turbulence	↓ Deposition	Low-moderate. turbulence → ↑ settling	CatBoost emphasizes turbulence more
Distance to Vegetation	↑ Deposition	Flow deceleration at distance	LightGBM broader range/higher rank
Relative Position Veg.	↓ Deposition	Boundary deceleration vs. interior turbulence	CatBoost stronger boundary sensitivity
Local Flow Velocity	↓ Deposition	Low velocity → settling	LightGBM more velocity-sensitive
Vegetation Drag Index	Variable	Drag retardation effects	LightGBM slightly higher drag sensitivity

Collectively, these differences highlight that while both models broadly agree on the primacy of spatial and vegetation-related predictors, they diverge in how they quantify, rank, and interpret their influence. These modeling distinctions carry important physical interpretations in sediment transport processes—where thresholds in vegetation density, flume position, and flow velocity can lead to abrupt transitions in particle mobility, deposition, or bypass.

Ultimately, such SHAP-based diagnostics provide a robust framework not only for comparing model behavior but also for linking algorithmic outcomes to mechanistic hydrological insights. Overall, this SHAP analysis demonstrates that the longitudinal channel pattern, the extent and arrangement of vegetation cover, turbulence, and flow velocity collectively and with nonlinear effects control the spatial pattern of sediment deposition, with each model emphasizing certain distinct aspects of this process.

4.1.1. Enhanced section on the physical interpretation of SHAP

The SHAP dependence plot for "*Cumulative Vegetation Exposure*" (Figure 8) shows a clear but gradual transition in the effect of this feature on predicted deposition. For cumulative exposure values up to approximately 0.23 m, SHAP values are predominantly positive, indicating that limited vegetation exposure enhances net deposition. As exposure increases beyond 0.25 m, SHAP values become increasingly mixed, and by 0.35 m they are predominantly negative. The approximate center of this transition lies near 0.3 m. This behavior is not a single abrupt threshold but rather a gradual shift from a deposition-enhancing regime ($S > E$) to a deposition-reducing regime ($E > S$), where E is erosion rate from bed ($\text{kg}/\text{m}^2/\text{s}$), and S is deposition rate onto bed ($\text{kg}/\text{m}^2/\text{s}$). The value 0.3 m is therefore used as a convenient approximation of the transition zone.

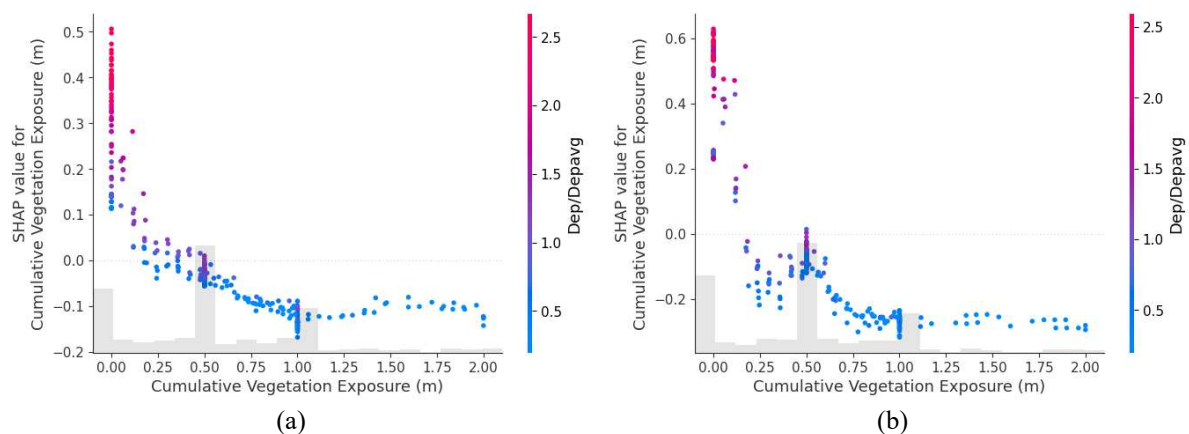


Figure 8. SHAP dependence plots for "*Cumulative Vegetation Exposure*", (a) CatBoost (b) LightGBM

A notable observation from the SHAP analysis is the consistently weak influence of "*local flow velocity*" on the predicted deposition, with mean absolute SHAP values of only 0.049–0.063 and no systematic trend in the dependence plots (Figure 9). Within the present laboratory dataset (rigid emergent cylinders, uniform low velocities), this suggests that once the flow has entered a dense vegetation patch, the velocity is reduced to such an extent that its spatial variations no longer control the net deposition pattern.

This result is qualitatively consistent with earlier experimental observations. For instance, (Bouma et al., 2007) noted that within dense seagrass canopies, local flow velocity alone could not explain sediment trapping patterns, and that patch configuration played a major role. More recently, (Tranmer et al., 2024) showed that sediment supply to the patch interior, rather than the local velocity magnitude, determines whether deposition or erosion dominates. Our SHAP-based findings align with these studies, reinforcing the idea that for densely vegetated patches, models should prioritize geometric and cumulative exposure metrics over local velocity measurements.

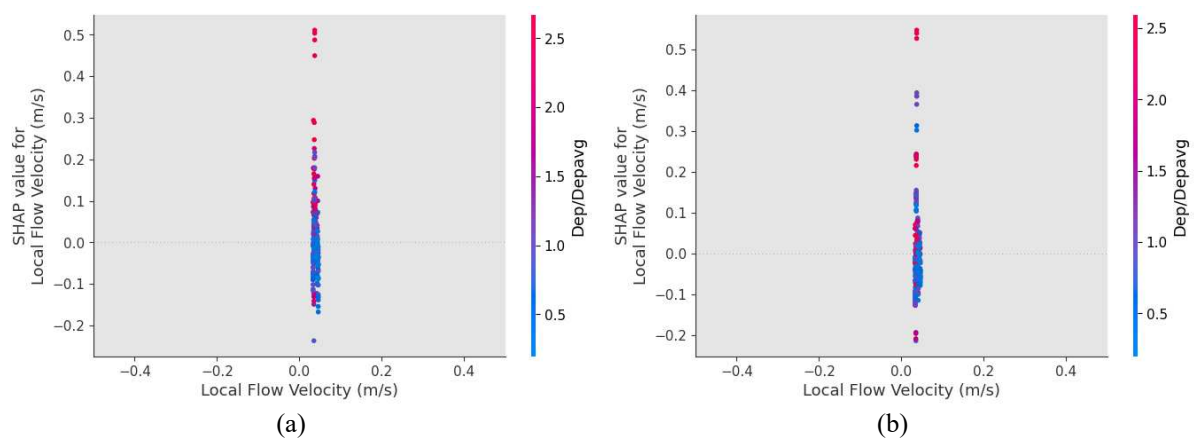


Figure 9. SHAP dependence plots for "*Local Flow Velocity*", (a) CatBoost (b) LightGBM

4.2. Linking SHAP insights to governing equations

Flow and sediment transport governing equations, typically written as:

$$\frac{\partial h}{\partial t} + \frac{\partial(hu)}{\partial x} = 0 \quad (7)$$

$$\frac{\partial(hu)}{\partial t} + \frac{\partial(hu^2 + gh^2/2)}{\partial x} = -gh \frac{\partial z_b}{\partial x} - \frac{\tau_b}{\rho} \quad (8)$$

$$\frac{\partial(hC)}{\partial t} + \frac{\partial(huC)}{\partial x} = \frac{\partial}{\partial x} \left(Dh \frac{\partial C}{\partial x} \right) + E - S \quad (9)$$

capture the essential physics of flow and sediment motion. Equation (7) represents mass conservation (continuity) for water, equation (8) represents momentum conservation, and equation (9) represents sediment mass conservation, where h is water depth (m), t is time (s), u is flow velocity (m/s), x is longitudinal coordinate (m), g is gravitational acceleration (m/s^2), z_b is bed elevation (m), and τ_b is bed shear stress (N/m^2), ρ is water density (kg/m^3), C is sediment concentration (kg/m^3), and D is dispersion coefficient (m^2/s).

The mechanistic insights derived from SHAP analysis can be directly interpreted within the framework of the governing equations for flow and sediment transport (Eqs. 7-9). This connection not only validates the physical consistency of the machine learning model but also illuminates how the learned feature effects emerge from the underlying conservation principles. The net deposition rate (Dep/Dep_{avg}), which is the target variable in this study, is directly proportional to the difference ($S - E$) under quasi-steady conditions. Each of the key features identified by SHAP can be linked to specific terms in Eqs. 7-9:

The nonlinear threshold behavior observed for "*Cumulative Vegetation Exposure*" (Figure 8) reflects the evolving competition between deposition (S) and erosion (E) as flow progresses through vegetation patches. For exposure lengths below approximately 0.3 m, the positive SHAP values indicate that $S > E$, leading to net deposition enhancement. This can be explained through the momentum equation (Eq. 8): vegetation drag reduces local velocity (u) and bed shear stress (τ_b), which directly suppresses the erosion term E (typically parameterized as a

function of τ_b). Simultaneously, reduced velocity increases the residence time of suspended particles, enhancing the settling flux S (which scales with particle settling velocity and near-bed concentration).

As "*Cumulative Vegetation Exposure*" exceeds 0.3 m, the SHAP values become negative, indicating a shift to $E > S$ and net deposition reduction. This transition occurs because vegetation-generated turbulence—particularly shear-layer turbulence at the top of the canopy—propagates downward and increases near-bed turbulent kinetic energy. This elevated turbulence enhances E by increasing instantaneous bed shear stress and may also suppress S by maintaining particles in suspension. The 0.3 m threshold thus represents the distance required for canopy-scale turbulence to become fully developed and influence the bed region—a length scale consistent with theoretical predictions for turbulent adjustment behind roughness elements (Nepf, 2012).

The "*Upstream Turbulence Index*" (ordinal values 0, 1, and 2) encodes the cumulative effect of upstream patch-generated turbulence on the local flow field. Its consistently negative SHAP values (indicating reduced deposition) align with its direct influence on the erosion term E . Turbulence enhances vertical mixing and increases the instantaneous bed shear stress, promoting particle resuspension. This feature effectively serves as a surrogate for the turbulence-related terms that are not explicitly resolved in the one-dimensional advection-diffusion equation but are known from higher-order closures to affect near-bed transport.

The positive SHAP values for "*Distance to Vegetation*" can be linked to the advective transport term $\partial(huC)/\partial x$ in Eq. 9. Immediately downstream of vegetation patches, flow deceleration creates a region of reduced transport capacity, promoting settling. As distance from the patch increases, the flow gradually re-adjusts toward open-channel conditions, and the deposition enhancement decays—a pattern captured by the SHAP dependence.

The weak influence of "*Local Flow Velocity*" on deposition, revealed by the SHAP dependence plots (Figure 9), provides fundamental insight into the hydrodynamic regime within dense vegetation. In the quasi-stagnant interior regions of patches, the advective term $\partial(huC)/\partial x$ becomes negligible compared to the local balance between erosion and deposition. Under such conditions, the system is governed primarily by local turbulent exchange (E and S) rather than by mean flow advection—explaining why velocity itself does not emerge as a dominant predictor. This finding underscores that in densely vegetated channels, patch-scale metrics (cumulative exposure, drag index) are more informative than local velocity for predicting deposition patterns.

The "*Vegetation Drag Index*" (proportional to $a_v \cdot U_0 \cdot H$) directly represents the sink term in the momentum equation (Eq. 8) due to vegetation resistance. Its positive SHAP values at low drag indices (enhancing deposition) reflect the dominance of flow deceleration, while its negative values at high drag indices (reducing deposition) capture the transition to turbulence-dominated regimes. This feature thus encapsulates the dual role of vegetation: reducing mean flow (promoting S) while generating turbulence (promoting E).

This direct linkage between SHAP interpretations and the governing equations demonstrates that the machine learning model has not merely learned statistical correlations, but has captured the underlying physical mechanisms encoded in the conservation laws. The consistency between data-driven insights and process-based theory reinforces the validity of both approaches and highlights the value of interpretable ML as a tool for mechanistic discovery in complex environmental systems.

4.3. Generality and scaling of the threshold behavior

The identification of a threshold at approximately 0.3 m "*Cumulative Vegetation Exposure*" raises a critical question: Is this a case-specific artifact of the experimental conditions, or does

it reflect a fundamental property of vegetated flow-sediment systems? We argue that while the approximate numerical value (0.3 m) is likely scale-dependent, the existence of a threshold behavior—a transition from deposition-promoting to deposition-inhibiting regimes with increasing cumulative vegetation exposure—represents a generalizable mechanistic principle in vegetated channels.

The threshold phenomenon can be understood through the lens of competing length scales in vegetated flows. Theoretical and experimental studies have established that flow adjustment behind a vegetation patch leading edge occurs over a characteristic length scale, L_{adj} , which scales with the drag length scale:

$$L_{adj} \sim \frac{1}{C_D \cdot a_v} \quad (10)$$

where C_D is the drag coefficient and a_v is vegetation density (Luhar et al., 2008; Nepf, 2012; Rominger & Nepf, 2011). Within this adjustment region, mean velocity decreases and turbulent kinetic energy increases, eventually reaching an equilibrium state. The competition between mean flow deceleration (which promotes deposition by reducing bed shear stress) and turbulence enhancement (which promotes erosion by increasing instantaneous shear stress) evolves along this adjustment length. The threshold observed at 0.3 m in this study corresponds to the point where turbulence effects begin to dominate over mean flow deceleration—a point that should exist in any vegetated flow, though its dimensional value will scale with vegetation density and flow conditions.

Using the range of drag coefficients reported for emergent rigid cylinders ($C_D = 0.5$ to 1.5) (Cheng, 2013; Tanino & Nepf, 2008) and the vegetation densities in our experiments ($a_v = 2.78$ m^{-1} for sparse cases and $a_v = 6.25$ m^{-1} for dense cases), the predicted adjustment length spans:

- For dense vegetation ($a_v = 6.25$ m^{-1}): $L_{adj} = 0.107$ to 0.320 m

- For sparse vegetation ($a_v = 2.78 \text{ m}^{-1}$): $L_{adj} = 0.241 \text{ to } 0.719 \text{ m}$

The dimensional threshold of "*Cumulative Vegetation Exposure*" identified in this study (0.3 m) falls within both of these theoretical ranges (0.107–0.320 m for dense, 0.241–0.719 m for sparse). Moreover, the overall theoretical envelope (0.107 to 0.719 m) comfortably contains the observed threshold. This consistency provides strong evidence that the 0.3 m threshold represents the physical adjustment length at which vegetation-generated turbulence becomes fully developed and begins to dominate over mean flow deceleration.

Crucially, while the approximate dimensional threshold (0.3 m) lies between the predicted bounds for the two density extremes, it does not coincide exactly with either. This is expected because the threshold was derived from a global model trained on all experimental runs (both dense and sparse patches together). The 0.3 m value can therefore be interpreted as an effective or average adjustment length across the range of vegetation densities tested. In physical terms, this suggests that the transition from deposition-promoting to deposition-inhibiting behavior occurs when cumulative vegetation exposure reaches a length scale comparable to the penetration depth of canopy-generated turbulence, which for mixed or intermediate densities is approximately 0.3 m under the present hydraulic conditions.

Thus, the dimensional threshold is not an arbitrary experimental artifact but a physically meaningful quantity that scales with $1/(C_D a_v)$. The agreement between the observed threshold (0.3 m) and the theoretical range (0.107 m to 0.719 m) validates the mechanistic interpretation that the competition between flow deceleration and turbulence generation determines the net deposition balance.

The existence of such thresholds is supported by independent experimental observations. Studies on sediment trapping by vegetation patches have consistently reported that deposition efficiency increases with patch length up to a point, beyond which additional length yields

diminishing returns or even reduced trapping (Yager & Schmeeckle, 2013; Zong & Nepf, 2011). These observations align with the concept of an optimal patch length that maximizes net deposition—a length that corresponds to the adjustment scale identified here. The present study advances this understanding by quantifying the threshold directly from data and linking it to competing physical mechanisms (mean flow deceleration versus turbulence generation).

In summary, the threshold identified in this study at 0.3 m cumulative vegetation exposure should be viewed not as a fixed numerical value, but as a manifestation of a fundamental mechanistic transition—the point at which vegetation-induced turbulence begins to dominate over mean flow deceleration in controlling sediment dynamics. This transition is expected to occur in any vegetated channel, with the dimensional threshold scaling inversely with vegetation density and potentially with flow depth. By reframing the finding in dimensionless terms and linking it to established theoretical length scales, this study provides a conceptually generalizable insight that can inform both process-based modeling and nature-based design across a range of scales and conditions.

4.4. Comparison with classical sediment transport equations

Direct application of classical bedload formulas (Meyer-Peter & Müller, 1948; Van Rijn, 1984) to the present dataset is not feasible, because they require the bed slope which is not reported in the original experimental study. Moreover, these equations were developed for bare-bed conditions and predict bedload transport, whereas our target variable (Dep/Dep_{avg}) represents net deposition of suspended sediment influenced by vegetation-induced turbulence.

Nevertheless, the literature uniformly reports that classical equations perform poorly in vegetated channels. (J. Q. Yang & Nepf, 2018) showed that classical shear-stress-based sediment transport models, developed without accounting for vegetation, underestimate transport rates by orders of magnitude in vegetated regions. (X. Wang et al., 2024) confirmed

that unmodified classical equations significantly underestimate or overestimate sediment transport in the presence of vegetation. Building on this, (Zhao & Nepf, 2021) found that the stress-based MPM formula underpredicts measured bedload transport rates in emergent vegetation by up to three orders of magnitude. Similarly, (J. Zhang et al., 2025) demonstrated that a TKE-based predictive model achieves a much stronger fit to measured data ($R^2 = 0.84$) than conventional shear stress-based formulas in emergent vegetation flows.

In contrast, the CatBoost model developed in this study, using only observable features (flow velocity, water depth, vegetation geometry), achieves $R^2 = 0.953$ and $MAPE = 17.967\%$ on the unseen dataset (Table 10). The substantial gap between published performance of classical and modified equations ($R^2 = 0.84$) and the present ML model ($R^2 \approx 0.95$) highlights the advantage of vegetation-specific, data-driven approaches.

5. Conclusion

This study developed an interpretable ensemble machine-learning framework for predicting sediment deposition in open channels with discontinuous emergent vegetation. Using a dataset of 498 observations obtained from 11 controlled flume experiments, five ensemble algorithms were evaluated through Bayesian hyperparameter optimization and five-fold cross-validation. Beyond predictive accuracy, the framework integrated SHAP analysis to identify the physical mechanisms governing sediment deposition and to improve the interpretability of data-driven predictions.

Among the evaluated models, CatBoost demonstrated the most robust and consistent predictive performance, achieving the highest accuracy and generalization capability on both testing and unseen data. The stability of its predictions across repeated runs, together with cross-validation, learning-curve analysis, and permutation testing, indicates that reliable model generalization can be achieved despite the relatively modest dataset size. These findings suggest that ensemble

learning methods, particularly CatBoost, provide an efficient alternative to computationally intensive numerical simulations and extensive laboratory experimentation for predicting sediment deposition under different vegetation configurations.

Beyond prediction, the principal scientific contribution of this study lies in revealing physically meaningful relationships between vegetation characteristics and sediment deposition. SHAP analysis identified cumulative vegetation exposure as the dominant controlling variable and revealed a distinct nonlinear transition near a cumulative exposure length of approximately 0.3 m. This threshold indicates a shift from vegetation-enhanced deposition to deposition reduction caused by increasingly developed vegetation-induced turbulence. In contrast, local flow velocity exhibited only a weak influence within vegetation patches, suggesting that patch-scale descriptors become more important than local hydraulic variables under quasi-stagnant canopy conditions. These findings not only provide a quantitative interpretation of vegetation–sediment interactions but also generate experimentally testable hypotheses that extend previous qualitative observations and strengthen the integration of explainable artificial intelligence with process-based understanding.

The proposed framework also has important practical implications for river engineering and ecological management. Once trained, the CatBoost model enables rapid evaluation of alternative vegetation layouts while SHAP explanations identify the variables responsible for each prediction, providing transparent support for engineering decisions. Such capability can assist practitioners in designing vegetation arrangements that either enhance sediment retention for bank stabilization and habitat restoration or reduce excessive deposition to preserve channel conveyance and flood capacity. More broadly, this study demonstrates how interpretable machine learning can bridge the gap between predictive modeling and physical understanding, supporting more transparent and sustainable river management strategies.

The present framework should nevertheless be interpreted within the limits of the experimental conditions, including rigid emergent vegetation, uniform sediment properties, and controlled laboratory flows. Additional validation under natural river conditions involving flexible vegetation, heterogeneous sediments, and unsteady hydraulic regimes is required before operational implementation. Future research should therefore focus on expanding the experimental database, incorporating more physically based descriptors of turbulence and vegetation dynamics, and developing hybrid physics-informed machine-learning models that further improve both predictive capability and physical interpretability. Overall, this study demonstrates that interpretable ensemble machine learning can serve not only as an accurate predictive tool but also as a mechanism-discovery framework for advancing the understanding and management of sediment transport in vegetated channels.

Acknowledgements

The authors gratefully acknowledge the significant contributions of the School of Civil Engineering at the University of Tehran and the School of Civil and Environmental Engineering at Tarbiat Modares University to this research. The authors also sincerely appreciate the valuable support and contributions provided by the Department of Civil Engineering at Hakim Sabzevari University and the Department of Civil Engineering at Ferdowsi University of Mashhad.

CRedit authorship contribution statement

R. Arabshahi: Data curing, Conceptualization, Methodology, Formal analysis, Software, Writing original draft, and Visualization; **M.A. Bahrabadi:** Formal analysis, Software, Writing original draft; **S.A.A. Salehi Neyshabouri:** Conceptualization, Supervision, Validation, Writing, Review and Editing; **R. Javidi Sabbaghian:** Supervision, Writing, Review and Editing; **M.A. Banihashemi:** Supervision, Writing, Review and Editing.

710 **Informed consent**

711 Informed consent was obtained from all individual participants included in the study.

712 **Ethical approval**

713 This article does not contain any studies with human participants or animals performed by
714 any of the authors.

715 **Funding**

716 The authors declare that no funding is associated with this study.

717 **Declaration of generative AI and AI-assisted technologies**

718 During the preparation of this work, the authors used AI-powered tools in order to check
719 grammar and enhance the academic quality of the text which was primarily written by the
720 authors. After using these tools, the authors reviewed and edited the content as needed and take
721 full responsibility for the content of the published article.

722 **Declaration of Interest Statement**

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

725 **Data Availability**

726 The datasets used and/or analyzed in this study are available from the corresponding author
727 upon reasonable request.

728 **References**

729 Abba, S. I., Nourani, V., & Elkiran, G. (2019). Multi-parametric modeling of water treatment
730 plant using AI-based non-linear ensemble. *Journal of Water Supply: Research and*
731 *Technology—AQUA*, 68(7), 547–561.

732 Achite, M., Yaseen, Z. M., Heddami, S., Malik, A., & Kisi, O. (2022). Advanced machine
 733 learning models development for suspended sediment prediction: comparative analysis
 734 study. *Geocarto International*, 37(21), 6116–6140.

735 Al Mamlook, R. E., Abdulhameed, T. Z., Hasan, R., Al-Shaikhli, H. I., Mohammed, I., &
 736 Tabatabai, S. (2020). Utilizing machine learning models to predict the car crash injury
 737 severity among elderly drivers. *2020 IEEE International Conference on Electro*
 738 *Information Technology (EIT)*, 105–111.

739 Basumatary, M. M., Maji, S., & Kumar, B. (2025). SHAP-based interpretation of machine
 740 learning models in bed load transport prediction. *Modeling Earth Systems and*
 741 *Environment*, 12(1), 65. <https://doi.org/10.1007/s40808-025-02679-3>

742 Bergstra, J., Yamins, D., & Cox, D. D. (2013). *Hyperopt: A Python Library for Optimizing*
 743 *the Hyperparameters of Machine Learning Algorithms*.

744 Bezak, N., Lebar, K., Bai, Y., & Rusjan, S. (2025). *Using Machine Learning to Predict*
 745 *Suspended Sediment Transport under Climate Change*.

746 Bouma, T. J., van Duren, L. A., Temmerman, S., Claverie, T., Blanco-Garcia, A., Ysebaert,
 747 T., & Herman, P. M. J. (2007). Spatial flow and sedimentation patterns within patches of
 748 epibenthic structures: Combining field, flume and modelling experiments. *Continental*
 749 *Shelf Research*, 27(8), 1020–1045.
 750 <https://doi.org/https://doi.org/10.1016/j.csr.2005.12.019>

751 Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of*
 752 *the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data*
 753 *Mining*, 785–794.

754 Chen, T., Shi, X., Wong, Y. D., & Yu, X. (2020). Predicting lane-changing risk level based

755 on vehicles' space-series features: A pre-emptive learning approach. *Transportation*
 756 *Research Part C: Emerging Technologies*, 116, 102646.
 757 <https://doi.org/https://doi.org/10.1016/j.trc.2020.102646>

758 Chen, Z., Luo, F., Yang, Y., Tao, A., Zhang, C., & Zheng, J. (2026). Predicting suspended
 759 sediment carrying capacity in vegetated flow based on diffusion theory and energy
 760 theory : From theory to modeling ☆. *Journal of Hydrology*, 664(PC), 134570.
 761 <https://doi.org/10.1016/j.jhydrol.2025.134570>

762 Cheng, N.-S. (2013). Calculation of drag coefficient for arrays of emergent circular cylinders
 763 with pseudofluid model. *Journal of Hydraulic Engineering*, 139(6), 602–611.

764 Ding, Y., Zhu, Y., Feng, J., Zhang, P., & Cheng, Z. (2020). Interpretable spatio-temporal
 765 attention LSTM model for flood forecasting. *Neurocomputing*, 403, 348–359.

766 El Bilali, A., Abdeslam, T., Ayoub, N., Lamane, H., Ezzaouini, M. A., & Elbeltagi, A.
 767 (2023). An interpretable machine learning approach based on DNN, SVR, Extra Tree,
 768 and XGBoost models for predicting daily pan evaporation. *Journal of Environmental*
 769 *Management*, 327, 116890.

770 Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The*
 771 *Annals of Statistics*, 29(5), 1189–1232. <http://www.jstor.org/stable/2699986>

772 Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data*
 773 *Analysis*, 38(4), 367–378. [https://doi.org/https://doi.org/10.1016/S0167-9473\(01\)00065-](https://doi.org/https://doi.org/10.1016/S0167-9473(01)00065-2)
 774 2

775 Gao, J., Kennedy, D. M., & McSweeney, S. (2023). Patterns of vegetation expansion during
 776 dune stabilization at the decadal scale. *Earth Surface Processes and Landforms*, 48(15),
 777 3059–3073.

778 Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., & Wolff, E. (2018).
779 Very high resolution object-based land use–land cover urban classification using
780 extreme gradient boosting. *IEEE Geoscience and Remote Sensing Letters*, 15(4), 607–
781 611.

782 Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. “
783 O’Reilly Media, Inc.”

784 Ghisalberti, M., & Nepf, H. M. (2002). Mixing layers and coherent structures in vegetated
785 aquatic flows. *Journal of Geophysical Research: Oceans*, 107(C2), 3–11.
786 <https://doi.org/https://doi.org/10.1029/2001JC000871>

787 Gui, Z., Shan, Y., Li, C., Sun, S., & Liu, C. (2024). Turbulence and sediment deposition in a
788 channel with floating vegetation. *Journal of Hydrology*, 641(March), 131829.
789 <https://doi.org/10.1016/j.jhydrol.2024.131829>

790 Habib, M. A., Abolfathi, S., O’Sullivan, J. J., & Salauddin, M. (2024). Efficient data-driven
791 machine learning models for scour depth predictions at sloping sea defences. *Frontiers*
792 *in Built Environment*, 10, 1343398.

793 Haddadchi, A., & Amir A, D. (2012). Assessment of bed-load predictors based on sampling
794 in a gravel bed river. *Journal of Hydrodynamics*, 24(1), 145–151.

795 Heddham, S. (2021a). 3 - *New formulation for predicting soil moisture content using only soil*
796 *temperature as predictor: multivariate adaptive regression splines versus random forest,*
797 *multilayer perceptron neural network, M5Tree, and multiple linear regression* (P.
798 Samui, H. Bonakdari, & R. B. T.-W. E. M. and M. T. Deo (eds.); pp. 45–62). Elsevier.
799 <https://doi.org/https://doi.org/10.1016/B978-0-12-820644-7.00027-X>

800 Heddham, S. (2021b). *Intelligent Data Analytics Approaches for Predicting Dissolved Oxygen*

801 *Concentration in River: Extremely Randomized Tree Versus Random Forest, MLPNN*
 802 *and MLR BT - Intelligent Data Analytics for Decision-Support Systems in Hazard*
 803 *Mitigation: Theory and Prac* (R. C. Deo, P. Samui, O. Kisi, & Z. M. Yaseen (eds.); pp.
 804 89–107). Springer Singapore. https://doi.org/10.1007/978-981-15-5772-9_5

805 Hori, M., Suzuki, T., Monthum, Y., Srisombat, T., Tanaka, Y., Nakaoka, M., & Mukai, H.
 806 (2009). High seagrass diversity and canopy-height increase associated fish diversity and
 807 abundance. *Marine Biology*, 156, 1447–1458.

808 Huai, W., Li, S., Katul, G. G., Liu, M., & Yang, Z. (2021). Flow dynamics and sediment
 809 transport in vegetated rivers: A review. *Journal of Hydrodynamics*, 33(3), 400–420.

810 Huai, W., Yang, L., Wang, W.-J., Guo, Y., Wang, T., & Cheng, Y. (2019). Predicting the
 811 vertical low suspended sediment concentration in vegetated flow using a random
 812 displacement model. *Journal of Hydrology*, 578, 124101.

813 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Statistical Learning BT - An*
 814 *Introduction to Statistical Learning: with Applications in R* (G. James, D. Witten, T.
 815 Hastie, & R. Tibshirani (eds.); pp. 15–57). Springer US. [https://doi.org/10.1007/978-1-](https://doi.org/10.1007/978-1-0716-1418-1_2)
 816 [0716-1418-1_2](https://doi.org/10.1007/978-1-0716-1418-1_2)

817 Jia, F., Wang, W., Han, Y., Du, J., Zhang, Y., Liu, Z., & Gao, H. (2025). *Predictive*
 818 *framework of vegetation resistance in channel flow*. 1–18.

819 Khosravi, K., Attar, N., Bateni, S. M., Jun, C., Kim, D., Safari, M. J. S., Heddami, S.,
 820 Farooque, A., & Abolfathi, S. (2024). Daily river flow simulation using ensemble
 821 disjoint aggregating M5-Prime model. *Heliyon*, 10(20).

822 Kim, H. S., Kimura, I., & Shimizu, Y. (2015). Bed morphological changes around a finite
 823 patch of vegetation. *Earth Surface Processes and Landforms*, 40(3), 375–388.

824 Kisi, O., & Yaseen, Z. M. (2019). The potential of hybrid evolutionary fuzzy intelligence
825 model for suspended sediment concentration prediction. *Catena*, 174, 11–23.

826 Kitsikoudis, V., Yagci, O., Kirca, V. S. O., & Kellecioglu, D. (2016). Experimental
827 investigation of channel flow through idealized isolated tree-like vegetation.
828 *Environmental Fluid Mechanics*, 16, 1283–1308.

829 Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability
830 problem in machine learning. *Philosophy & Technology*, 33(3), 487–502.

831 Kumar, P., & Sharma, A. (2025a). Experimental analysis of vegetation heterogeneity on flow
832 characteristics and bed morphology in partially vegetated channels. *Earth Surface*
833 *Processes and Landforms*, 50(13), e70174.
834 <https://doi.org/https://doi.org/10.1002/esp.70174>

835 Kumar, P., & Sharma, A. (2025b). Impact of emergent vegetation on three-dimensional
836 turbulent flow properties and bed morphology in a partially vegetated channel.
837 *International Journal of Sediment Research*, 40(2), 286–311.
838 <https://doi.org/https://doi.org/10.1016/j.ijsrc.2024.10.006>

839 Kumar, S., Agarwal, M., & Rathnayake, U. (2026). Standalone and novel hybrid ML models
840 for estimating stream velocities in a river channel with vegetation. *Evolving Earth*,
841 4(February), 100115. <https://doi.org/10.1016/j.eve.2026.100115>

842 Kutner, M. H. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin.
843 <https://books.google.com/books?id=0xqCAAAACAAJ>

844 Li, S. (2024). *Machine - Assisted Physical Closure for Coarse Suspended Sediments in*
845 *Vegetated Turbulent Channel Flows*. <https://doi.org/10.1029/2024GL110475>

846 Li, S., Qu, Y., Zheng, T., & Gentine, P. (2024). Machine-assisted physical closure for coarse

847 suspended sediments in vegetated turbulent channel flows. *Geophysical Research*
848 *Letters*, 51(20), e2024GL110475.

849 Liu, C., Shan, Y., & Nepf, H. (2021). Impact of stem size on turbulence and sediment
850 resuspension under unidirectional flow. *Water Resources Research*, 57(3),
851 e2020WR028620.

852 Liu, X., Wang, S., Zhang, B., Wu, X., Huang, E., & Luo, M. (2024). The effect of emergent
853 vegetation on flow, bedload transport, and bed morphology in a diffluence-confluence
854 unit on a meandering channel. *Journal of Hydrology*, 638, 131488.

855 Liu, Y., Yang, Z., Huai, W., & Liu, X. (2025). Feedbacks of flow fields and sediment
856 deposition to discontinuous vegetation patches: impacts of density and characteristic
857 length of patches. *Journal of Hydrology and Hydromechanics*, 73(1), 84–94.

858 López, F., & García, M. H. (2001). Mean flow and turbulence structure of open-channel flow
859 through non-emergent vegetation. *Journal of Hydraulic Engineering*, 127(5), 392–402.

860 Luhar, M., Rominger, J., & Nepf, H. (2008). *Interaction between flow , transport and*
861 *vegetation spatial structure*. 423–439. <https://doi.org/10.1007/s10652-008-9080-9>

862 Lund, J. W., Groten, J. T., Karwan, D. L., & Babcock, C. (2022). Using machine learning to
863 improve predictions and provide insight into fluvial sediment transport. *Hydrological*
864 *Processes*, 36(8), e14648.

865 Meddage, D. P. P., Ekanayake, I. U., Herath, S., Gobirahavan, R., & Muttill, N. (2022).
866 *Predicting Bulk Average Velocity with Rigid Vegetation in Open Channels Using Tree-*
867 *Based Machine Learning : A Novel Approach Using Explainable Artificial Intelligence*.

868 Meyer-Peter, E., & Müller, R. (1948). Formulas for bed-load transport. Rept 2nd Meeting Int.
869 *Assoc. Hydraul. Struct. Res*, 39–64.

870 Mir, A. A., Patel, M., Albalawi, F., Bajaj, M., & Tuka, M. B. (2024). A comparative
871 ensemble approach to bedload prediction using metaheuristic machine learning.
872 *Scientific Reports*, 14(1), 25725.

873 Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable Machine Learning – A Brief*
874 *History, State-of-the-Art and Challenges BT - ECML PKDD 2020 Workshops*. 417–431.

875 Nepf, H. M. (1999). Drag, turbulence, and diffusion in flow through emergent vegetation.
876 *Water Resources Research*, 35(2), 479–489.
877 <https://doi.org/https://doi.org/10.1029/1998WR900069>

878 Nepf, H. M. (2012). Flow and transport in regions with aquatic vegetation. *Annual Review of*
879 *Fluid Mechanics*, 44(1), 123–142.

880 Noori, R., Ghiasi, B., Salehi, S., Esmaeili Bidhendi, M., Raeisi, A., Partani, S., Meysami, R.,
881 Mahdian, M., Hosseinzadeh, M., & Abolfathi, S. (2022). An efficient data driven-based
882 model for prediction of the total sediment load in rivers. *Hydrology*, 9(2), 36.

883 O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors.
884 *Quality & Quantity*, 41(5), 673–690. <https://doi.org/10.1007/s11135-006-9018-6>

885 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost:
886 unbiased boosting with categorical features. *Advances in Neural Information Processing*
887 *Systems*, 31.

888 Rahaeimehr, R., Babakhani, Z., Fakharzadeh Moghadam, O., Mir Nasir, S., Safaei, P.,
889 Abdollahi, M. A. A., Bakhtiari Haftlang, M., Baghaei-Shiva, G., Salimian, N., Taheri
890 Hosseinkhani, N., Abedi, S., Alinasab, N., Damiri, M., Amiri, F., Arabshahi, R.,
891 Montazeri, S., Fattahiamin, S., Mahbod, A., Bahadori, N., ... Moradi, D. (2025).
892 Application of AI in Research and Data Science. *Kindle*, 5(1 SE-Academic Text Books),

893 1–362. <http://preferpub.org/index.php/kindle/article/view/Book60>

894 Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using
 895 shapley values: application to compound potency and multi-target activity predictions.
 896 *Journal of Computer-Aided Molecular Design*, 34(10), 1013–1026.

897 Rominger, J. T., & Nepf, H. M. (2011). Flow adjustment and interior flow associated with a
 898 rectangular porous obstruction. *Journal of Fluid Mechanics*, 680, 636–659.

899 S. Band, S., Qasem, S. N., Mansor, Z., Pai, H.-T., Mehdizadeh, S., Gupta, B. B., & Mosavi,
 900 A. (2026). Hybrid machine learning and deep learning models for river suspended
 901 sediment load forecasting. *Engineering Applications of Computational Fluid Mechanics*,
 902 20(1), 2591799. <https://doi.org/10.1080/19942060.2025.2591799>

903 Sahoo, G. K., Sahoo, A., Samantara, S., Satapathy, D. P., & Satapathy, S. C. (2022).
 904 Application of adaptive neuro-fuzzy inference system and salp swarm algorithm for
 905 suspended sediment load prediction. In *Intelligent System Design: Proceedings of*
 906 *INDIA 2022* (pp. 339–347). Springer.

907 Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in applying machine
 908 learning models for hydrological inference: A case study for flooding events across
 909 Germany. *Water Resources Research*, 56(5), e2019WR025924.

910 Shapley, L. S. (1953). *Contributions to the Theory of Games, Volume II* (H. W. Kuhn & A.
 911 W. Tucker (eds.); pp. 307–318). Princeton University Press.
 912 <https://doi.org/doi:10.1515/9781400881970-018>

913 Shi, P., Bai, L., Zhao, Z., Dong, J., Li, Z., Min, Z., Cui, L., & Li, P. (2023). Vegetation
 914 position impacts soil carbon losses on the slope of the Loess Plateau of China. *Catena*,
 915 222, 106875.

916 Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine
 917 learning algorithms. *Advances in Neural Information Processing Systems*, 25.

918 Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual
 919 predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647–
 920 665. <https://doi.org/10.1007/s10115-013-0679-x>

921 Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit
 922 environment and water management: a survey of methods, applications, and future
 923 directions. *Environmental Research Letters*, 14(7), 73001.

924 Tanino, Y., & Nepf, H. M. (2008). Laboratory investigation of mean drag in a random array
 925 of rigid, emergent cylinders. *Journal of Hydraulic Engineering*, 134(1), 34–41.

926 Tranmer, A. W., Ji, U., Ahn, M., Jung, S. H., & Yager, E. M. (2024). Characterizing erosion
 927 and deposition in and around riparian vegetation patches: Complex flow hydraulics,
 928 sediment supply, and morphodynamic feedbacks. *Water Resources Research*, 60(2),
 929 e2023WR034859.

930 Van Rijn, L. C. (1984). Sediment transport, part I: bed load transport. *Journal of Hydraulic*
 931 *Engineering*, 110(10), 1431–1456.

932 Vargas-Luna, A., Duró, G., Crosato, A., & Uijttewaal, W. (2019). Morphological adaptation
 933 of river channels to vegetation establishment: A laboratory study. *Journal of*
 934 *Geophysical Research: Earth Surface*, 124(7), 1981–1995.

935 Viana, C. M., Santos, M., Freire, D., Abrantes, P., & Rocha, J. (2021). Evaluation of the
 936 factors explaining the use of agricultural land: A machine learning and model-agnostic
 937 approach. *Ecological Indicators*, 131, 108200.

938 Wang, J., Liu, J., Sun, Y., Li, J., & Cao, Z. (2024). Flow resistance of emergent rigid

939 vegetation in steady flow. *Journal of Hydrology and Hydromechanics*, 72(2), 207–222.

940 Wang, S., Peng, H., & Liang, S. (2022). Prediction of estuarine water quality using
941 interpretable machine learning approach. *Journal of Hydrology*, 605, 127320.

942 Wang, W.-J., Huai, W.-X., Li, S., Wang, P., Wang, Y.-F., & Zhang, J. (2019). Analytical
943 solutions of velocity profile in flow through submerged vegetation with variable frontal
944 width. *Journal of Hydrology*, 578, 124088.

945 Wang, X., Gualtieri, C., Huai, W., Liu, H., & Yu, S. (2024). An improved formula for bed-
946 load rate in open channel flows with emergent vegetation. *Physics of Fluids*, 36(1),
947 13309. <https://doi.org/10.1063/5.0185667>

948 Yagci, O., Özgür Kirca, V. S., Kitsikoudis, V., Wilson, C. A. M. E., Celik, M. F., & Sertkan,
949 C. (2024). Experimental study on influence of different patterns of an emergent
950 vegetation patch on the flow field and scour/deposition processes in the wake region.
951 *Water Resources Research*, 60(1), e2023WR034978.

952 Yager, E. M., & Schmeeckle, M. W. (2013). The influence of vegetation on turbulence and
953 bed load transport. *Journal of Geophysical Research: Earth Surface*, 118(3), 1585–
954 1601.

955 Yang, C. T. (1996). *Sediment transport: theory and practice*. McGraw-Hill New York.

956 Yang, C. T., Marsooli, R., & Aalami, M. T. (2009). Evaluation of total load sediment
957 transport formulas using ANN. *International Journal of Sediment Research*, 24(3), 274–
958 286.

959 Yang, J. Q., & Nepf, H. M. (2018). A Turbulence-Based Bed-Load Transport Model for Bare
960 and Vegetated Channels. *Geophysical Research Letters*, 45(19), 10,410-428,436.
961 <https://doi.org/https://doi.org/10.1029/2018GL079319>

962 Yang, L., Fang, H., Yang, Z., & Huai, W. (2023). Longitudinal dispersive coefficient in
963 channels with aquatic vegetation: A review. *Journal of Hydrodynamics*, 35(3), 379–395.

964 Yang, L., Huai, W., & Guo, Y. (2021). Stochastic simulation of the suspended sediment
965 deposition in the channel with vegetation and its relevance to turbulent kinetic energy.
966 *Water Resources Research*, 57(9), e2021WR030380.

967 Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., & Qin, H. (2020). Comparison of support
968 vector regression and extreme gradient boosting for decomposition-based data-driven
969 10-day streamflow forecasting. *Journal of Hydrology*, 582, 124293.

970 Yue, Z. J., Wang, N. N., Xu, B. D., Huang, X., Yang, D. M., Xiao, H. B., & Shi, Z. H.
971 (2025). Toward Trustworthy Machine Learning for Daily Sediment Modeling in the
972 Riverine Systems: An Integrated Framework With Enhanced Uncertainty Quantification
973 and Interpretability. *Water Resources Research*, 61(5), e2024WR038650.
974 <https://doi.org/https://doi.org/10.1029/2024WR038650>

975 Zhang, J., Lei, J., Huai, W., & Nepf, H. (2020). Turbulence and particle deposition under
976 steady flow along a submerged seagrass meadow. *Journal of Geophysical Research:*
977 *Oceans*, 125(5), e2019JC015985.

978 Zhang, J., Lu, Y., Wang, W., Li, Z., Li, P., Guo, M., Yu, K., & Wang, H. (2025). Prediction
979 of the bedload transport rate in vegetated rivers based on turbulent kinetic energy.
980 *Journal of Hydroinformatics*, 27(3), 580–599. <https://doi.org/10.2166/hydro.2025.317>

981 Zhang, Z., Huang, J., Duan, S., Huang, Y., Cai, J., & Bian, J. (2022). Use of interpretable
982 machine learning to identify the factors influencing the nonlinear linkage between land
983 use and river water quality in the Chesapeake Bay watershed. *Ecological Indicators*,
984 140, 108977.

985 Zhao, T., & Nepf, H. M. (2021). Turbulence Dictates Bedload Transport in Vegetated
986 Channels Without Dependence on Stem Diameter and Arrangement. *Geophysical*
987 *Research Letters*, 48(21), e2021GL095316.
988 <https://doi.org/https://doi.org/10.1029/2021GL095316>

989 Zong, L., & Nepf, H. (2011). Spatial distribution of deposition within a patch of vegetation.
990 *Water Resources Research*, 47(3).

991 Zounemat-Kermani, M., Stephan, D., Barjenbruch, M., & Hinkelmann, R. (2020). Ensemble
992 data mining modeling in corrosion of concrete sewer: A comparative study of network-
993 based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced*
994 *Engineering Informatics*, 43, 101030.

995