Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 7, 1572 doi: 10.2166/hydro.2024.297

Benchmarking the performance and uncertainty of machine learning models in estimating scour depth at sluice outlets

Xuan-Hien Le [a,b,*], Le Thi Thu Hien [b], Hung Viet Ho [b] and Giha Lee [a]

[a] Department of Advanced Science and Technology Convergence, Kyungpook National University, 2559 Gyeongsang-daero, Sangju 37224, South Korea
[b] Faculty of Water Resources Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi 10000, Vietnam
*Corresponding author. E-mail: hienlx@knu.ac.kr

X-HL, 0000-0002-0947-0805; LTT, 0000-0001-7874-7380; GL, 0000-0002-7560-818X

ABSTRACT

This study investigates the performance of six machine learning (ML) models – Random Forest (RF), Adaptive Boosting (ADA), CatBoost (CAT), Support Vector Machine (SVM), Lasso Regression (LAS), and Artificial Neural Network (ANN) – against traditional empirical formulas for estimating maximum scour depth after sluice gates. Our findings indicate that ML models generally outperform empirical formulas, with correlation coefficients (CORR) ranging from 0.882 to 0.944 for ML models compared with 0.835–0.847 for empirical methods. Notably, ANN exhibited the highest performance, followed closely by CAT, with a CORR of 0.936. RF, ADA, and SVM performed competitive metrics around 0.928. Variable importance assessments highlighted the dimensionless densimetric Froude number (F_d) as significantly influential, particularly in RF, CAT, and LAS models. Furthermore, SHAP value analysis provided insights into each predictor's impact on model outputs. Uncertainty assessment through Monte Carlo (MC) and Bootstrap (BS) methods, with 1,000 iterations, indicated ML's capability to produce reliable uncertainty maps. ANN leads in performance with higher mean values and lower standard deviations, followed by CAT. MC results trend towards optimistic predictions compared with BS, as reflected in median values and interquartile ranges. This analysis underscores the efficacy of ML models in providing precise and reliable scour depth predictions.

Key words: machine learning, Monte Carlo simulation, scour depth estimation, SHAP values, sluice gate, uncertainty quantification

HIGHLIGHTS

- Benchmarked six ML models against empirical formulas for estimating scour depth.
- ML algorithms performed superior performance with CORR [0.882–0.944].
- CAT, ANN, and RF models excelled in precision and accuracy.
- Evaluated predictor importance using permutation and SHAP values.
- Assessed uncertainty in predictions by Monte Carlo and Bootstrap methods.

1. INTRODUCTION

Scouring downstream of sluice outlets remains a primary concern in hydraulic engineering due to its implications for structural damage and the alteration of hydrodynamic conditions within water bodies. The phenomenon of scour pertains to the erosion or displacement of sedimentary materials, such as sand and rocks, by water flow (Verma & Goel 2005; Yeganeh-Bakhtiary *et al.* 2020). Specifically, scour at sluice outlets can imperil the structural integrity of the outlets and proximate infrastructure, potentially resulting in operational complications and even failures (Yousif *et al.* 2019). Attempts have been made to counteract the erosive force of water, such as introducing non-erodible aprons downstream (Aamir & Ahmad 2022). However, depending on its magnitude, a consequential scour hole can critically threaten the foundational stability of gates. This emphasizes the importance of precise scour depth estimation and management strategies.

The maximum scour depth typically reaches the equilibrium stage when no grain movement occurs within the scour hole (Chatterjee *et al.* 1994; Fitri *et al.* 2019). Estimating maximum scour depth at sluice gate outlets is inherently intricate, given the myriad of variables governing the scour processes (Sarathi *et al.* 2008). Influences such as soil properties, initial conditions, and hydrodynamic flow characteristics play pivotal roles in the stability of hydraulic constructions (Najafzadeh

et al. 2017). In particular, some researchers study the effect of auxiliary work like wing walls on this parameter (Le *et al.* 2022) or the roughness of the apron (Aamir *et al.* 2022). Consequently, predicting maximum scour depth, especially given the wide range of possible outcomes, becomes a difficult task requiring in-depth understanding and rigorous modelling approaches.

Over the years, researchers have employed many methods to scour depth estimation at hydraulic structures encompassing empirical equations, physical observations, and hydraulic models (Mutlu Sumer 2007), each of which attempts to predict the scour dynamics based on parameters such as velocity, particle size, and outlet design (Mostaani & Azimi 2022). Despite the varied approaches, each method has presented its challenges. Numerical models, for instance, often leveraging the Navier-Stokes equations coupled with sediment transport formulations, have proven valuable for simulating scour evolution (Olsen & Kjellesvig 1998). Yet, their practical application sometimes struggles with computational demands and reliability (Le *et al.* 2022). Empirical equations generally developed through thorough experimental data analysis and understanding scour depth influencers have found broad utility (Hamidifar *et al.* 2011). However, they sometimes falter in reliability due to their dependency on a limited scope of experimental data and the intrinsic complexity of scour phenomena (Najafzadeh *et al.* 2017).

In light of these challenges and rapid advancements in computational capacities and data analytics, there has been a paradigm shift towards embracing artificial intelligence techniques, particularly ML, to address these challenges (Sharafati *et al.* 2020; Kartal *et al.* 2023). This transition is driven by the inherent ability of ML models to discern intricate and non-linear interrelationships among a multitude of variables, a task that traditional methodologies often grapple with (Sreedhara *et al.* 2021; Le *et al.* 2023). ML algorithms, especially the Gene Expression Programming (GEP), Group Method of Data Handling (GMDH) networks, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), have emerged as indispensable tools in maximum scour depth estimation (Najafzadeh 2015). Studies such as those of Najafzadeh *et al.* (2017) have juxtaposed several techniques, including the GEP, Evolutionary Polynomial Regression, and Model Tree (MT), illuminating the superior predictive accuracy of MT over traditional empirical equations. Similarly, Abd El-Hady Rady (2020) affirmed the genetic programming algorithm's prowess, noting its superior performance over adaptive neuro-fuzzy inference system (ANFIS) models and empirical equations. Qaderi *et al.* (2021) accentuated the aptitude of ANFIS, observing its outperformance over common algorithms such as GEP, GMDH, SVM, and ANN. Parsaie *et al.* (2019) added another layer to this discourse by presenting SVM's slight edge over ANFIS and ANN in maximum scour depth prediction. Despite promising accuracy, the performance of these methods is inherently dependent on the datasets in which they are trained, emphasizing the pivotal role of data quality and volume in model outcomes (Aamir & Ahmad 2016). This inherent dependency underscores the burgeoning significance of quantifying and interpreting uncertainty in ML predictions, especially within the niche domain of hydrodynamic scouring phenomena.

Uncertainty quantification techniques range from probabilistic methods such as Monte Carlo (MC) simulations (Han *et al.* 2011) to non-parametric ones such as bootstrapping (Efron & Tibshirani 1994). Such traditional methods have carved a niche in hydrodynamic studies, offering the dual benefit of assessing variability and conferring confidence in predictions. The recent literature also emphasizes adopting these methodologies to bolster prediction reliability (Palmer *et al.* 2022), especially in risk assessments (Grana *et al.* 2020). However, despite these advancements, the field of uncertainty quantification is still developing (Ustimenko *et al.* 2020). As Hüllermeier & Waegeman (2021) noted, understanding the nuanced distinctions between aleatoric and epistemic uncertainties in machine learning offers a deeper insight into the limits and potential of predictive models. Furthermore, Abed *et al.* (2023) emphasized the increasing role of artificial intelligence in environmental modelling, pointing to the expansive potential of these technologies in enhancing methodological approaches in hydrodynamic studies. Nonetheless, comprehensive studies on uncertainty quantification in the context of scour depth predictions at sluice outlets are scarce. The literature tends to focus on the capabilities of individual ML models rather than exploring comparative or ensemble approaches that might enhance predictive accuracy and reliability (Rezaie-Balf 2019). The present research gap underscores the significance of this research, which endeavours to benchmark the various ML models and quantify the uncertainty in their predictions of maximum scour depth.

In response to these identified gaps, this study endeavours to assess the six ML models and two empirical formulas comprehensively. The evaluation highlights their respective and comparative prediction performances for estimating maximum scour depth at sluice outlets. This approach compares their effectiveness and delves deeper into each model's capabilities through an integrated methodological framework. Furthermore, the research explores the incorporation of advanced interpretability techniques, such as the importance of permutation feature and SHAP (SHapley Additive exPlanations) values, which enhance the transparency and understanding of how different predictors influence the model outputs. These tools

are vital for dissecting the complex dynamics of the predictive models and refining their accuracy. In addition to interpretability, this study also intensively applies MC simulations and Bootstrap (BS) techniques to thoroughly quantify the uncertainty in the predictions provided by these models. By generating a multitude of predictive outcomes through these techniques, the study assesses the reliability and variability of the model forecasts, offering a robust statistical basis to evaluate their predictive confidence.

2. MATERIALS AND METHODS

2.1. Overview of empirical equations

Scouring downstream of sluices is an intricate phenomenon influenced by numerous hydraulic, geometric, and sedimentary factors. The formation and extent of the scour hole (as depicted in Figure 1) are mainly affected by parameters such as bed sediment characteristics, the flow conditions both upstream and downstream of the sluice gate, sluice gate opening size, and apron length (Farooq & Ghumman 2019).

Central to understanding the scour phenomenon is the equilibrium or maximum scour depth (d_s), which is a crucial metric that describes the shape of the scour. The determination of d_s is affected by parameters such as the input velocity (V), tailwater depth (d_t), the open height of the sluice gate (a), apron length (L), apron roughness (Lim & Yu 2002; Dey & Westrich 2003), and bed material properties, including soil density (ρ_s), its geometric standard deviation (σ_g), mean grain size (D_{50}), and the soil type (Najafzadeh & Lim 2015). Further compounding the determination is the gravitational acceleration (g), water density (ρ_w), and kinematic viscosity (ν). The connection between the scour depth and efficient variables has been recognized as:

$$d_s = f(V, a, L, d_t, D_{50}, \sigma_g, \rho_s, \rho_w, g, \nu) \tag{1}$$

In scour modelling, it has been demonstrated that the utilization of dimensionless parameters in ML techniques has the potential to improve the predictive capacity of scouring models compared with using dimensional parameters (Guven & Gunal 2008; Azamathulla & Ghani 2011). To simplify the relationships that govern the scouring process, the Buckingham π theorem is employed. This theorem is a fundamental principle in dimensional analysis, aiding in the reduction of complex physical phenomena to dimensionless relationships. It categorizes variables into repeating and non-repeating groups to derive dimensionless parameters, known as π terms. In this context, the repeating variables considered are a, V, and ρ_w. The eight non-repeating variables are L, d_t, D_{50}, ρ_s, d_s, g, ν, and σ_g. From these variables, eight dimensionless π terms are formulated as follows:

$$\pi_1 = F;\ \pi_2 = F_d;\ \pi_3 = D_{50}/a;\ \pi_4 = d_t/a;\ \pi_5 = L/a;\ \pi_6 = \sigma_g;\ \pi_7 = V \cdot a/\nu,\ \text{and}\ \pi_8 = d_s/a$$

Here, F represents the Froude number (see Equation (2)), F_d represents the densimetric Froude number (see Equation (3)), π_7 represents the Reynolds number kept much higher than the threshold value for a turbulent flow in a fully rough zone (Aamir & Ahmad 2022).

It is noted that although the Reynolds number (π_7) is crucial for identifying the flow regime, it is found to have an insignificant effect on maximum scour depth in turbulent conditions. In addition, Aamir & Ahmad (2019) performed a test to determine the significance of each π term in predicting maximum scour depth, concluding that σ_g is a negligible quantity. Therefore, instead of the 10 variables in Equation (1), the relative maximum scour depth (d_s/a) can be determined by five π terms in the following function:

$$F = \frac{V}{\sqrt{g \cdot a}} \tag{2}$$

$$F_d = \frac{V}{\sqrt{\left(\frac{\rho_s}{\rho_w} - 1\right) g \cdot D_{50}}} \tag{3}$$

$$\frac{d_s}{a} = f\left(\frac{L}{a}, \frac{D_{50}}{a}\frac{d_t}{a}, F, F_d\right) \tag{4}$$

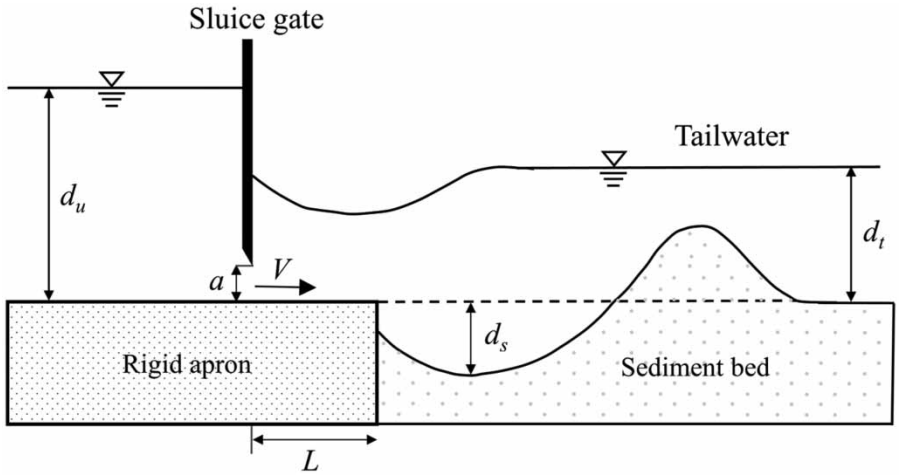

Figure 1 | Schematic of sediment geometry after sluice.

From Equation (4), researchers have developed many empirical formulas over the years. The equations that have been proposed for this prediction are briefly summarized in Table 1.

Table 1 illustrates that these formulas frequently simplify the underlying complexities if there is insufficient input data. For instance, many formulations consider d_s to be a function of a single parameter, as seen in Equations (5) and (6), or focus on jet water characteristics and specific soil, as denoted by Equations (7) and (8). However, a challenge arises in parameters like $F_{d(95)}$, which often prove difficult to obtain, putting them impracticable in practical scenarios. Another observation from Table 1 is the neglect of tailwater depth (d_t) in Equations (5)–(8). This is in contrast with the argument made by Dey & Sarkar (2006) that a maximal scour depth decreases as d_t increases up to the critical tailwater depth. Furthermore, the authors mentioned that an increase in sediment size (D_{50}) is correlated with a decrease in maximum scour depth, which is not found in Equations (5) and (6) where the impact of sediment characteristics area has been ignored. The value of the d_s/a decreases as the L/a increases, which is not discovered in Equations (5), (6), and (8).

Table 1 | Empirical formulas for estimating maximum scour depth

Researcher	Equation	
Chatterjee *et al.* (1994)	$\frac{d_s}{a} = 0.775F$	(5)
Aderibigbe & Rajaratnam (1998)	$\frac{d_s}{a} = -6.11 + 3.35F_{d(95)}$ where $F_{d(95)}$ is a F_d based on D_{95}	(6)
Lim & Yu (2002)	$\frac{d_s}{a} = 1.04\sigma_g^{-0.69}F_d^{1.47}\left(\frac{D_{50}}{a}\right)^{0.33}e^{-0.04\beta}\left(\frac{L}{a}\right)^{1.4}$ where $\beta = \sigma_g^{-0.5}F_d^{-0.35}\left(\frac{D_{50}}{a}\right)^{-0.5}$	(7)
Hopfinger *et al.* (2004)	$\frac{d_s}{a} = 0.43\left(\frac{D_{50}}{a}\right)^{0.11}F_d^{1.1} - 0.2$	(8)
Dey & Sarkar (2006)	$\frac{d_s}{a} = 2.59F_d^{0.94}\left(\frac{L}{a}\right)^{-0.37}\left(\frac{d_t}{a}\right)^{0.16}\left(\frac{D_{50}}{a}\right)^{0.25}$	(9)
Aamir *et al.* (2022)	$\frac{d_s}{a} = 2.35F_d^{1.80}\left(\frac{L}{a}\right)^{-0.76}\left(\frac{d_t}{a}\right)^{0.25}\left(\frac{D_{50}}{a}\right)^{0.49}\left(\frac{k_s}{e\,a}\right)^{-2.79}$ where k_s is the roughness	(10)

According to the findings of Aamir & Ahmad (2019), empirical equations, including those that rely on complex multiple linear regressions, can fail to estimate maximal scour depth accurately on occasion. It is important to note that many of these formulas were developed and calibrated using particular experimental datasets and conditions. Thus, their efficacy may fluctuate when applied to different datasets. This emphasizes the significance attributed to the initial experimental conditions and datasets in the process of formula determination. With advances in technology and computational methods, there has been a shift towards using data-driven techniques, such as ML-based methods, to boost the predictability of maximum scour depth.

2.2. Data collection

For this research, our focus was concentrated on experimental data sourced from the seminal works of Dey & Sarkar (2006) (hereafter Dey_2006 for short) and the more recent findings of Aamir *et al.* (2022) (hereafter Aamir_2022 for short). Both researchers investigated the local scour caused by the 2D submerged water jet after the sluice gate. This intentional data selection was premised on two empirical equations from these studies, encapsulated in Equations (9) and (10). Adapting our study to these specific equations allows for a consistent and systematic comparative analysis, which serves as a reasonable yardstick against which the performance of other prediction models can be evaluated. The data in Table 2 show the variety of test runs with detailed metrics, such as apron length, gate opening, tailwater depth, and the water jet's Froude number located behind the sluice gate. This study only looks at smooth, rigid aprons; their roughness is insignificant. The condition of a submerged hydraulic jump was maintained throughout all experiments in which the tailwater depth was greater than the conjugate depth of a free hydraulic jump (Aamir *et al.* 2022).

2.3. ML models

2.3.1. Random Forest (RF)

RF is a method for ensemble learning that generates the mode (classification) or mean (regression) prediction of the individual trees for an unexplored dataset using a multitude of decision trees (DTs) constructed during training (Tin Kam 1998). Each tree is constructed using a different bootstrap sample, and a random subset of features is considered at each split in the tree. Owing to its capacity to handle large data sets with higher dimensionality, it has been an essential tool in various scientific domains (Habib *et al.* 2023). RF possesses an inherent capability to measure the importance of individual features, making it invaluable for choosing features in complex hydrodynamic predictions. The overall prediction is obtained by averaging the predictions generated by each individual tree:

$$F_{\mathrm{RF}}(x) = \frac{1}{T}\sum_{i=1}^{T} f_i(x) \tag{11}$$

where the number of trees is denoted by T; and the prediction of the i-th tree is denoted by $f_i(x)$.

2.3.2. Adaptive Boosting (ADA)

ADA, short for 'Adaptive Boosting', is an iterative ensemble method primarily used for increasing the performance of weak classifiers. The core principle behind ADA is to weigh each sample in the dataset based on the errors of the previous iteration. Each iteration's weight adjustments ensure that the subsequent weak learner (DT) focuses more on previously misclassified samples. This iterative adjustment continues until the error converges to a minimum value or a certain number of trees are attained. Due to its adaptive nature, ADA has been effective in domains with intricate boundaries between classes or non-linear regression tasks (Freund & Schapire 1997). Its potential to identify complicated non-linear interactions between input features and maximum scour depth can be pivotal for accurate predictions. The ADA regression function could be

Table 2 | Overview of experimental data

Investigator	Number of runs	D_{50}/a	L/a	d_t/a	F	F_d	d_s/a
Dey_2006	225	0.02–0.4	26.7–55	6.6–13.9	2.4–4.9	3.3–22.1	1.5–8.2
Aamir_2022	126	0.02–1.3	20–100	6.7–40	1.5–12.1	2.7–25.3	0.3–20.2

expressed as:

$$F_{\mathrm{ADA}}(x) = \sum_{i=1}^{T} \alpha_i \cdot f_i(x) \tag{12}$$

where α_i denotes the weight assigned to the i-th tree, which is computed using the error of that tree.

2.3.3. CatBoost (CAT)

CAT is a gradient-boosting algorithm that can inherently handle categorical data without requiring explicit one-hot or label encoding (Prokhorenkova *et al.* 2018). Its primary advantage is its ability to handle categorical variables without manual preprocessing by transforming them into numerical values using various techniques like one-hot encoding and mean encoding. CAT also provides built-in support for handling missing values. For the regression problem, the CAT equation can be described as:

$$F_{\mathrm{CAT}}(x) = \sum_{i=1}^{T} \eta_i \cdot f_i(x) \tag{13}$$

where η_i represents the learning rate multiplied by the contribution of the i-th tree.

2.3.4. Support Vector Machine (SVM)

Originally designed for classification, SVM can be adapted for regression tasks through Support Vector Regression. The primary idea underlying SVM is to identify a hyperplane that best fits the data, ensuring that errors do not exceed a specified threshold while also keeping the hyperplane as flat as possible (Noori *et al.* 2022). A prominent advantage of SVM is its capability to work in a transformed feature space via the kernel trick, effectively handling non-linear relationships (Cortes & Vapnik 1995). For regression tasks, the simplified representation of the SVM could be depicted as:

$$F_{\mathrm{SVM}}(x) = \omega \cdot \phi(x) + b \tag{14}$$

where b denotes the bias; ω represents the weight vector; and $\phi(x)$ indicates the transformation of input x through the kernel function.

2.3.5. Lasso Regression (LAS)

LAS, a linear regression variant, incorporates L1 regularization, which can shrink some coefficients to zero, acting as a feature selector. Especially in situations where there is multicollinearity between predictor variables, LAS offers stable solutions by penalizing the absolute size of the coefficients (Tibshirani 1996). LAS seeks to minimize the sum of squares of residuals, provided that the sum of the absolute values of the coefficients does not exceed a constant. Mathematically, this objective function is:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{15}$$

where λ denotes the regularization parameter; β_0 and β_j denote coefficients; y and x are the response and predictors variables; p and n are the number of predictors and observations.

2.3.6. Artificial Neural Network (ANN)

ANNs are computational structures inspired by biological neural networks, capable of representing intricate and non-linear mappings between inputs and outputs (Le *et al.* 2021). Comprising interconnected layers of nodes (or 'neurons'), ANNs can learn and represent almost any function given enough depth and data. In maximum scour depth prediction, ANN can adaptively learn from the data without relying on a pre-determined functional form and can capture the inherent complexities and

non-linearities of hydraulic processes. In a single-layer ANN, the output for an input vector x is computed as follows:

$$F_{\mathrm{ANN}}(x) = \sigma\left(\sum_{i=1}^{n} \omega_i x_i + b\right) \tag{16}$$

where ω denotes the weight vector; n represents the number of input nodes; and σ denotes the activation function.

2.4. Uncertainty analysis methods

2.4.1. Permuation importance

Permutation feature importance provides a metric to discern the significance of each feature (or variable) used by an ML algorithm. By shuffling the values of a particular variable and measuring the subsequent decrease in model performance (typically, accuracy or error rate), one can ascertain the importance of that feature in the model's predictions (Breiman 2001). The rationale behind this is that if a predictor is vital for the model, randomly altering its values would result in a notable drop in the model's performance. Conversely, insignificant features would have a negligible impact on the performance metric. Altmann *et al.* (2010) further elucidated that permutation importance can offer insight into the model's decision-making process and can be especially effective when benchmarking multiple ML algorithms, ensuring a uniform evaluation metric.

2.4.2. SHAP values

SHAP values are a state-of-the-art model interpretability method that ascribes each variable as an essential value for a specific prediction (Rodríguez-Pérez & Bajorath 2020). Originating from cooperative game theory, SHAP values mathematically guarantee a unique and consistent attribution for each feature, ensuring fairness and avoiding potential biases (Lundberg & Lee 2017). Within the context of scour depth prediction, SHAP values can unveil the influence of individual predictors and the intricate non-linear relationships and interactions among predictors, thereby furnishing a nuanced understanding of how each feature contributes, either positively or negatively, to the predicted maximum scour depth.

2.4.3. Uncertainty quantification

Quantifying the uncertainty in model predictions is of paramount importance, especially in scenarios with high-stakes outcomes, such as estimating maximum scour depths at sluice outlets. Two prevalent methods, MC simulation and BS, are employed to understand this uncertainty.

Rooted in probabilistic theory, MC methods involve running numerous simulations with random input values within a predefined distribution to approximate the output's expected distribution (Brownlee 2019). By repeatedly sampling and simulating, MC furnishes an explicit representation of the prediction's uncertainty, capturing both its variability and sensitivity to changes in the input values (Han *et al.* 2011). While the technique's inherent simplicity makes it attractive, its effectiveness hinges on the availability of a well-defined probability distribution for each input parameter and requires extensive computational resources due to repeated simulations. In maximum scour depth prediction, MC simulations can provide a probability distribution of the predicted scour depth rather than a single deterministic value.

Another widely acknowledged method for uncertainty quantification is BS. Efron & Tibshirani (1994) presented BS as a resampling technique where random samples (with replacement) are drawn from the dataset and are used to gauge the variability or confidence intervals of an estimator. Recent studies have indicated bootstrapping's viability in enhancing the robustness of ML predictions in water resources (Palmer *et al.* 2022). The key advantage of BS lies in its non-parametric nature, requiring no assumptions about the data's distribution, making it an ideal choice for complex, non-linear datasets (Gewerc 2020), as frequently encountered in hydraulic studies. In the current analysis, both MC and BS methods were employed to understand the uncertainty bounds of the ML models, a step crucial in establishing the credibility of ML predictions for maximum scour depth estimations.

2.5. Model design and hyperparameter tuning

Hyperparameter tuning emerges as a crucial procedure when enhancing the capabilities of ML models to estimate the maximum scour depth accurately. The grid search technique is a well-regarded method in hyperparameter optimization owing to its rigorous and exhaustive exploration of potential parameter combinations. Such meticulousness is especially suited to datasets of moderate sizes, as it facilitates the evaluation of the ML model across every permutation of hyperparameters within a

Table 3 | Descriptive summary of the hyperparameters and their ranges

Algorithm	Hyperparameter	Value range	Optimal value
RF	n_estimators	[50, 100, 200]	50
	max_depth	[None, 10, 20, 30]	10
	min_samples_leaf	[1, 2, 4]	1
	min_samples_split	[2, 3, 4, 5]	5
	bootstrap	[True, False]	True
ADA	n_estimators	[50, 100, 200]	200
	learning_rate	[0.001, 0.01, 0.1, 0.5, 1.0]	0.01
	loss	[linear, square, exponential]	Linear
	base_estimators	DecisionTreeRegressor(max_depth)	4
CAT	learning_rate	[0.01, 0.05, 0.1, 0.5, 1]	0.01
	iterations	[100, 500, 1,000]	1,000
	depth	[3, 5, 7]	3
SVM	kernel	[linear, poly, rbf, sigmoid]	rbf
	C	[0.01, 0.1, 1, 10, 100]	10
	epsilon	[0.01, 0.1, 1]	1
LAS	alpha	[0.001, 0.01, 0.1, 1, 10, 100, 1,000]	0.1
ANN	drop_rate	[0.1, 0.2, 0.25, 0.3]	0.25
	hidden layer	[2, 3]	3
	number of units	[128, 64]	128, 64, 64[a]
	loss	mse, mae	mse

[a]The number of units per respective hidden layer for ANN is 128, 64, 64.

predetermined grid. The widely recognized Python library, scikit-learn, offers efficient functionalities for the effective application of grid search procedures (Pedregosa *et al.* 2011).

An essential aspect of hyperparameter tuning is the use of cross-validation. This study adopts the grid search methodology, which incorporates 5-fold cross-validation. This approach substantially augments the model's resilience, ensuring its predictions remain consistent and reliable when subjected to previously unobserved data. In Table 3, the particular hyperparameters and their respective ranges for each ML algorithm are exhaustively detailed.

For the current investigation, the dataset comprises 351 laboratory-derived samples. These samples elucidate the multifaceted nature of maximum scour depths encountered downstream of sluice gates. The maximum scour depth ratio to the open height of the sluice gate (d_s/a) is conceptualized based on five instrumental variables: d_t/a, L/a, D_{50}/a, F, and F_d. The study guarantees a thorough assimilation of parameters influencing maximum scour depths by encompassing this breadth of variables. Considering the dataset dimensions, a strategic division was executed, allocating 70% of the samples (246 samples) for training purposes while the remaining 30% (about 105 samples) for validation and testing. This distribution assures that the model undergoes intensive training while retaining a significant chunk of data for unbiased performance evaluation.

2.6. Evaluation metrics

Ensuring accurate and reliable assessment of scour depth prediction models requires incorporating a careful selection of performance metrics. Therefore, this study takes advantage of various statistical criteria, each of which scrutinizes diverse aspects of the model predictions compared with the observed values. In essence, these metrics evaluate the predictive accuracy, bias, and overall reliability of the models, thereby facilitating a comprehensive understanding of their respective capabilities.

The following statistical criteria were chosen for model evaluation: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), CORR, and Nash–Sutcliffe Efficiency (NSE). The RMSE and MAE, both in their unique ways, gauge the magnitude of the estimation error, offering insights into model precision and bias, respectively. SMAPE provides a percentage error between the estimated and observed values, offering a scale-independent error metric. CORR reflects the linear relationship, while NSE represents the model's predictive accuracy and is especially appreciated for its ability to delineate the proportion of the variance in the measured data that the model captures.

Supplementary Table 4 briefly summarizes these evaluation metrics, their respective mathematical formulations, and interpretative insights.

3. RESULTS

3.1. Performance of methods: benchmarking

The performance of the methods, including six ML algorithms and two empirical formulas, in estimating the maximum scour depth was quantified and presented in Supplementary Table 5. Subsequently, Figures 2 and 3 translate these statistics into visual representations to intuitively comprehend the comparative performance among the proposed methods.

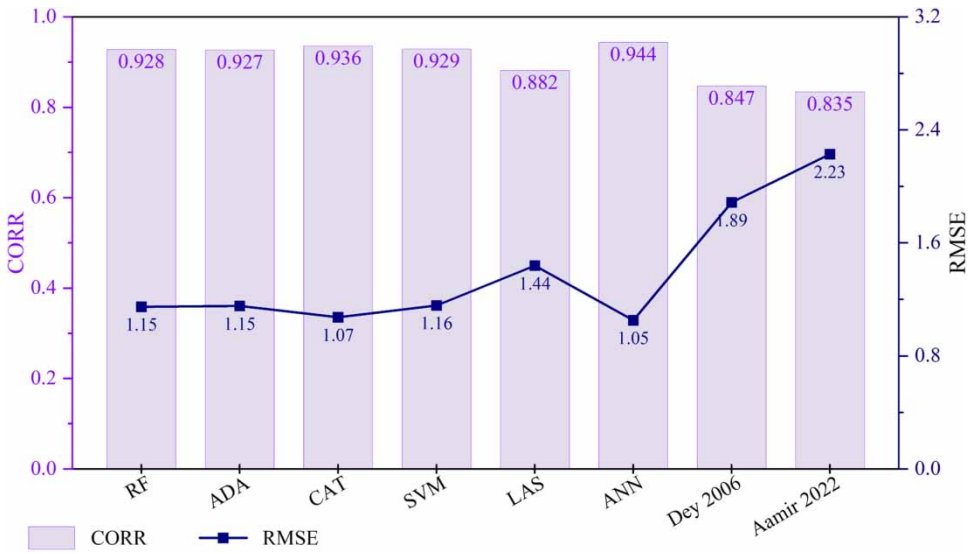

Figure 2 | Comparison of CORR and RMSE for various methods.

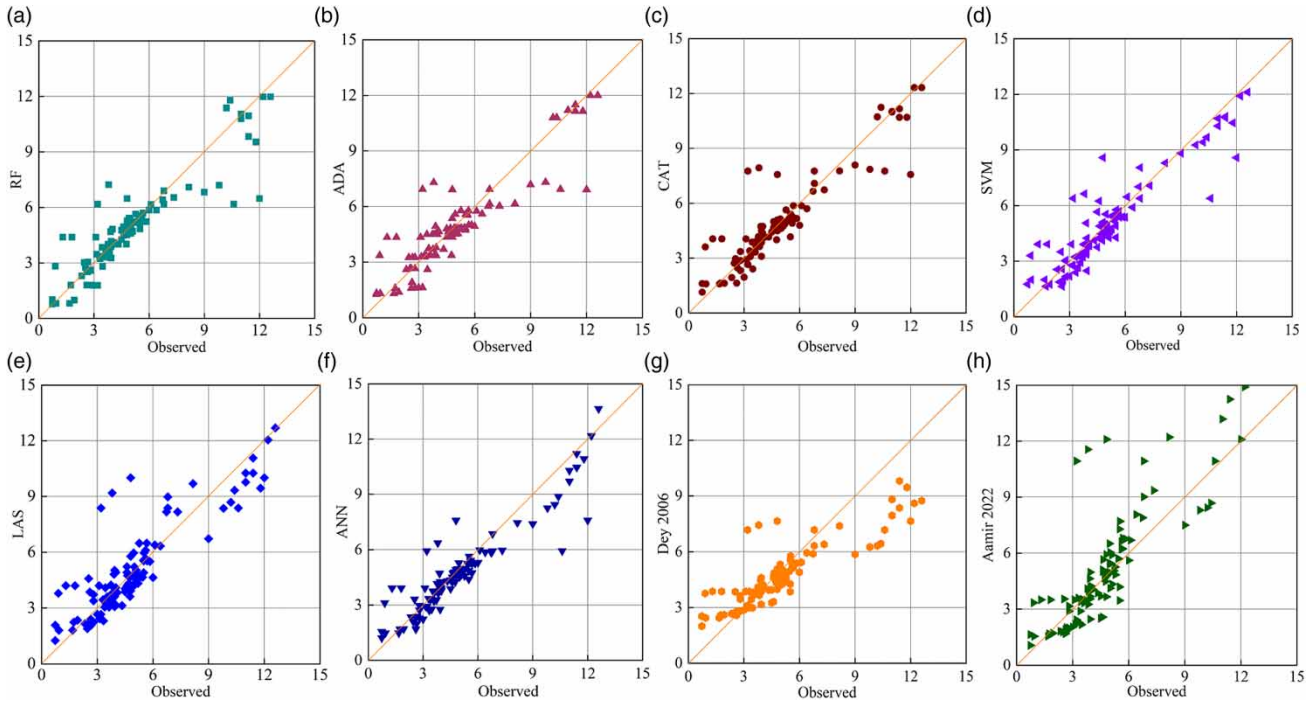

Figure 3 | Scatter plots compare between experimental data and predicted results.

Regarding RMSE and MAE, which offer insights into the dispersion and deviation from observed values, ML algorithms prominently outshone the empirical methods. Specifically, the CAT registered the least deviation from observed values with an RMSE of 1.07 and an MAE of 0.63. ANN and RF followed suit, both mirroring each other in MAE values at 0.66, with the ANN slightly taking the lead in RMSE at 1.05. In contrast, the empirical formula proposed by Aamir_2022 presented the most significant deviation with an MAE and RMSE of 1.33 and 2.23, respectively. Although Dey_2006 presented a marginally better performance than Aamir_2022, the accuracy of the majority of ML models generally overshadowed the empirical equations.

For the SMAPE criterion, the errors of the six ML algorithms fluctuate in the range (7.5–10.3%), which is smaller than the fluctuations of the two empirical formulas (10.4% of Dey_2006 and 12.0% by Aamir_2022). Among these, CAT, RF, and ANN are the leading models with an error level of about 7.6%, significantly lower than ADA (9.1%) and LAS (10.3%). This disparity emphasizes the robustness of ML algorithms, especially CAT, RF, and ANN, in mirroring observed values with decreasing bias.

A similar trend was observed through meticulous analysis when CORR and NSE indices were examined, illuminating the superior performance of ML techniques over empirical formulations, as demonstrated in Figures 2 and 3. These figures interweave the statistical metrics, emphasizing the overarching dominance of ML algorithms, especially ANN and CAT. This superior linear predictive capability is contrasted sharply with the empirical equations, as the visualizations confirm the superior clustering of ML predictions around observed values (see Figure 2). In addition, the scatter plots for ANN and CAT depict a pronounced alignment with the observed data, while empirical methods reveal a more considerable dispersion (see Figure 3).

For the empirical formulas, Dey_2006 still performed better than the Aamir_2022 formula in both criteria, registering CORR and NSE values of 0.847 and 0.687, respectively. Although the difference in linear prediction capabilities between the two formulas was modestly set (at approximately 0.012 for CORR), a clear contrast was revealed in their NSE values, exhibiting a significant difference of 0.124. The ML models demonstrated notable consistency in performance across both CORR and NSE criteria, spanning the range [0.882–0.944] and [0.818–0.903] for each criterion, respectively. Within this range of methods, the ANN model emerged as the superior performer, while the LAS model displayed the poorest performance among ML models. The ANN algorithm is closely followed by the CAT algorithm, which has a CORR value of 0.936 and an NSE value of 0.899. Meanwhile, the RF, ADA, and SVM algorithms maintained competitive performance metrics, exhibiting minor discrepancies in their values, hovering around 0.928 for CORR and 0.883 for NSE.

In general, the comprehensive performance benchmarking of methods for predicting maximum scour depth indicated that ML algorithms consistently demonstrated superior accuracy and efficiency. Notably, the CAT, ANN, and RF models stood out for their precision, closely reflecting observed values with minimal deviations. Meanwhile, SVM, ADA, and LAS models exhibited lower performance in descending order. In contrast, while the empirical formula by Dey_2006 showed relatively better performance than that of Aamir_2022, both were overshadowed by the predictive ability of most ML models.

3.2. Uncertainty analysis of ML models

3.2.1. Insights from variable importance

ML algorithms utilize a variety of predictors to estimate maximum scour depth, with specific predictors having a more significant impact on the output than others. Supplementary Table 6 depicts the importance of the permutation of these critical predictors across six ML models, exhibiting their relative significance in shaping the estimated results. This complex interaction of varying importance is extrapolated graphically in Figure 4, providing a vivid comparative visualization of the influence of each predictor on separate ML algorithms.

A careful analysis in Supplementary Table 6 reveals clear differences in the importance of the variables across the models, which can be broadly classified into three levels based on their influence on the models. The first level, primarily dominated by the F_d variable, was identified as having significant scores in most models, with the highest scores in RF and CAT (respective scores of 0.342 and 0.364). As for the LAS model, F_d indicates its dominance with the highest importance score among all analysed variables, a value of 0.556 compared with 0.444 for d_t/a. In juxtaposition, the ANN assigns relatively less emphasis to F_d (score of 0.170), directing more weight towards d_t/a and F, with respective importance scores of 0.361 and 0.315. This variable, representative of the dimensionless densimetric Froude number, underscores its pivotal role in influencing maximum scour depth predictions.

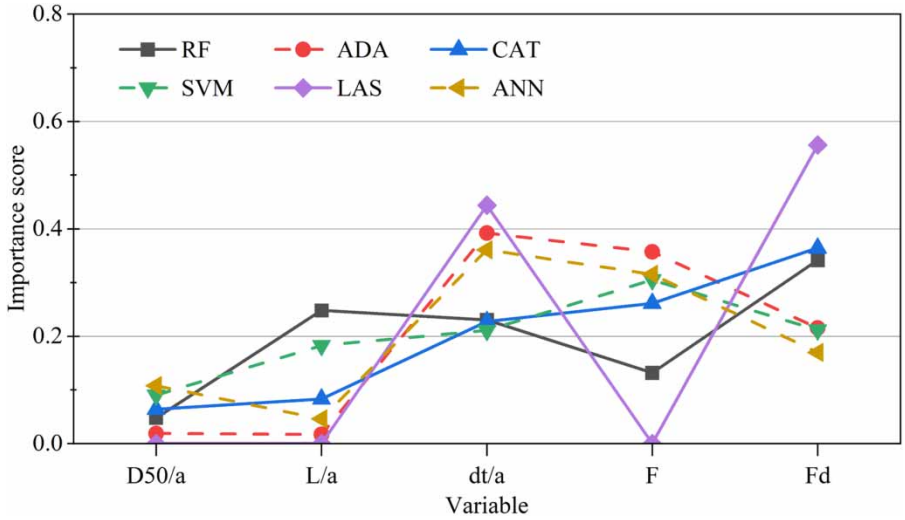

Figure 4 | Predictor importance of methods.

The second level includes the variables d_t/a and F. Both emerged as crucial predictors for ADA and ANN, with significance scores of (0.392 and 0.357) and (0.361 and 0.315), respectively.

Interestingly, F has the highest influence observed in SVM at 0.305, but its importance in LAS is zero. Although slightly lesser, the robust influence of d_t/a was also identified in CAT and SVM, with respective importance scores of 0.227 and 0.211.

The remaining variables, D_{50}/a and L/a, show relatively muted effects across the board. Specifically, L/a appears relatively unimportant in the ADA, LAS, and ANN models, with modest significance scores of less than 0.05. This contrasts RF because L/a achieved an importance score of 0.248, ranking 2nd in importance among the variables compared. For D_{50}/a, most of the recorded values represent the poorest scores for this variable (see Figure 4). The relatively lower significance scores of these predictors, associated with mean grain diameter and characteristic channel length, indicate less direct influence on maximum scour depth estimates in most models, especially LAS, where their impact diminishes to non-existence.

In addition to the variable importance analysis that provides a comprehensive understanding of the influence of each predictor, the SHAP plots in Figure 5 provide a more detailed and interpretative perspective on the variable importance across these six models. SHAP values quantify the average impact of each predictor variable on the model's output magnitude.

In Figure 5, the range of SHAP values denotes the magnitude of the impact a predictor has on the model output. Therefore, a longer bar signifies a more influential variable. In general, variable F is dominant in influencing forecast results in most ML models, except LAS (ranked 3rd). For the RF model, F and L/a variables dominate in terms of influence, closely followed by F. The ADA algorithm showcases a relatively balanced influence among F, L/a, and F_d. The SVM model reflects a similar pattern to the CAT, with F being favoured, albeit with a slightly more substantial influence from F_d and d_t/a. The LAS algorithm exhibits a distinct shift with F_d and d_t/a attracting significant attention, while other variables play a softer role or are even ignored, such as L/a. Lastly, in the ANN, F and D_{50}/a elements exhibit a higher SHAP value than the other predictors.

The contrast between the permutation importance and the SHAP values can be seen in some models. Specifically, while specific predictors like F and F_d maintain consistent importance across multiple models, the relative influence of variables like L/a and d_t/a sees marked fluctuations. This highlights the subtle complexities and nuances between two measures of variable importance. The importance of permutations and SHAP values underscores the multifaceted nature of maximum scour depth prediction and the complex interaction of variables in the predictive power of an ML algorithm.

3.2.2. Uncertainty in predictions

Evaluating uncertainties associated with ML predictions remains essential in determining the reliability and robustness of the algorithms employed. The present study exploited MC and BS techniques to provide insight into this aspect. These methods effectively generated a substantial set of 1,000 estimations on the testing dataset, offering a comprehensive view of the behaviours exhibited by the model under diverse conditions. The performance statistics for the uncertainty estimation methods are systematically presented in Supplementary Table 7 and Figure 6.

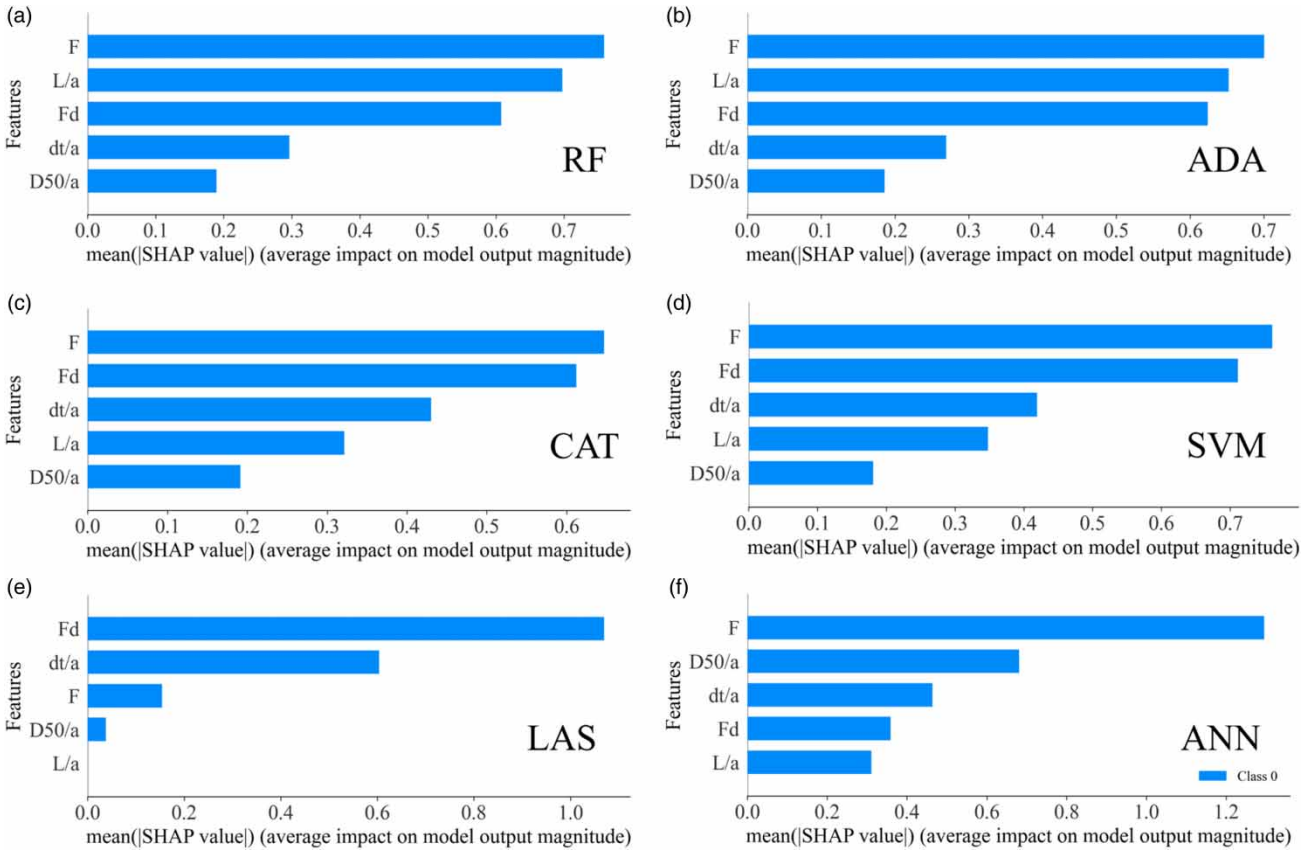

Figure 5 | SHAP plots of methods.

From Supplementary Table 7 and Figure 6, it is clear that the ANN model consistently displays superior performance, marked by a score of 0.938 for MC and 0.9231 for BS. In contrast, the LAS exhibits the worst performance, with respective scores of 0.882 and 0.875. However, the uncertainty of the LAS model, characterized by values of 0.0012 for MC and 0.0088 for BS, is markedly lower.

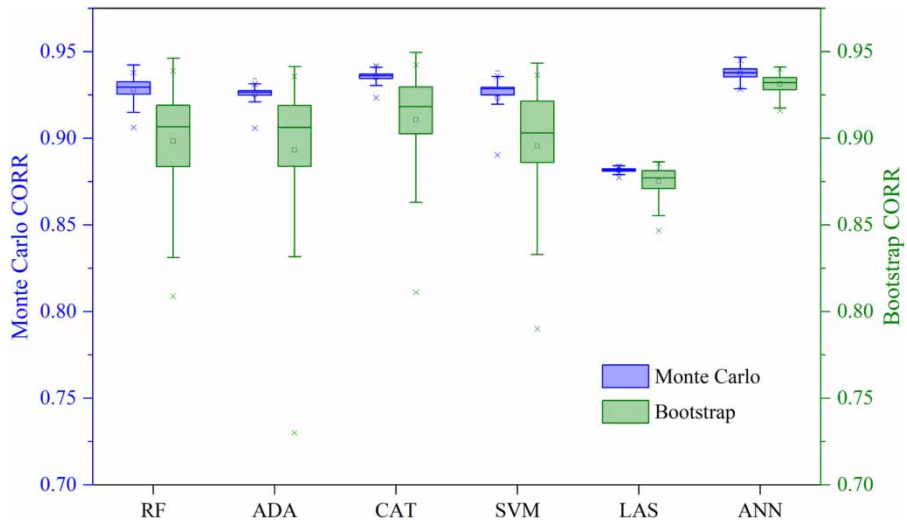

Figure 6 | Box plots of methods for estimating the uncertainty of algorithms.

This indicates that its predictions, though slightly less accurate, are more consistent. The CAT algorithm closely follows the performance, reflecting promising scores of 0.935 and 0.911, respectively. Besides, the standard deviations, which represent the variation in predictions, were relatively low for all models, suggesting consistent predictive capabilities. Notably, the ANN showed exceptionally low standard deviations of around 0.005 for both techniques, reinforcing its stable performance. The upper 95% CI further corroborated these findings, with the ANN model consistently outperforming the others.

Figure 6 depicts several statistical properties of the 1,000 simulations for the six ML algorithms. It can be observed that the median values were close to the mean values from Supplementary Table 7. This indicated a symmetric distribution of prediction values around the median, suggesting minimal skewness in the predictive outcomes. For all models, the median values derived from the MC technique are marginally higher than those from the BS approach, indicating a slight leaning of the MC simulations towards optimistic predictions. For example, the RF has a median of 0.930 and 0.907 for MC and BS, respectively. For the 1st Quartile (Q1) and 3rd Quartile (Q3) values, a narrower interquartile range (IQR) is identified for the LAS and ANN models, suggesting a higher concentration of prediction values within this range. In contrast, RF exhibited a wider spread with IQR values of 0.007 and 0.035 for MC and BS, respectively, indicating greater prediction variability. Additionally, the IQR values of the MC method were mostly smaller than those of the BS method, as shown by the range of [0.001–0.007] compared with the range of [0.007–0.036]. This trend is repeated across models, illustrating a consistent pattern.

The results in the prediction uncertainty analysis of ML models using MC and BS techniques exhibit varying performance levels and internal variability. The ANN stands out with superior mean and standard deviation performance, but this does not overshadow the remarkable results from other algorithms such as RF, ADA, CAT, and SVM. The performance of the CAT closely follows the ANN, with standard deviations of about 0.0044 and 0.0274 for MC and BS, respectively. While the RF, SVM, and ADA models exhibit a competitive performance with a 1,000-prediction mean of approximately 0.925, the standard deviation of ADA (about 0.0052) is slightly smaller than that of RF and SVM (about 0.0072 and 0.0125, respectively). The synthesis of results from MC and BS techniques offers a balanced perspective, emphasizing the strengths and limitations of each approach in capturing the inherent unpredictabilities in maximum scour depth estimation at sluice outlets.

4. DISCUSSION

4.1. Comparative analysis: ML models vs. empirical equations

The field of maximum scour depth prediction is witnessing a gradual transition from long-standing empirical methods to ML algorithms. Historically, empirical equations were derived from observational and experimental data to provide a generalized solution for complex physical phenomena. In this context, the empirical formulas of Dey_2006 and Aamir_2022 were introduced to estimate maximum scour depth. However, as evidenced in the results presented, there appears to be a paradigm shift in the effectiveness of prediction methods.

ML models, especially CAT, ANN, and RF, have showcased a pronounced superiority in statistical metrics. The performance benchmarking emphasized their lower deviation and bias compared with empirical equations. Although the Dey_2006 formula still holds up relatively better than the Aamir_2022 formula, the sheer dominance of ML algorithms, particularly CAT and ANN, cannot be overlooked. The consistent accuracy and efficiency delivered by these ML models hint at the vast potential of data-driven techniques in capturing complex processes, a characteristic that traditional empirical formulas sometimes miss (Muzzammil & Alam 2010; Khosravi *et al.* 2021; Le & Le 2024). Furthermore, as more data becomes available from diverse environments and scenarios, these ML algorithms can continuously learn and improve, a feature not easily attainable with static empirical equations (Sharafati *et al.* 2021).

The distinction between ML models and empirical equations becomes even more profound when the predictive capability is dissected. For instance, the scatter plots of empirical methods divulge a broader dispersion compared with the likes of ANN and CAT. This divergence may be due to the inherent limitations of empirical formulas as they are built on certain assumptions. These assumptions often oversimplify complex real-world scenarios, leading to compromised predictive abilities (Aamir & Ahmad 2016). These findings align with the observations of Sharafati *et al.* (2021), who pointed out that ML models produce more accurate forecasts and display less bias than conventional empirical methods because of their capacity to adapt and learn from data.

In summary, while empirical equations have played a pivotal role in maximum scour depth predictions for years, the emergence of ML offers a paradigm of increased accuracy and adaptability. Additionally, the ability of ML algorithms to incorporate multidimensional factors into their predictions can enable researchers and professionals to integrate a broader

spectrum of parameters, thereby enriching the depth and breadth of scour depth analysis. Such advancements might also pave the way for real-time monitoring and prediction systems, leveraging the real-time learning capability of these algorithms.

4.2. Importance and impact of uncertainty in predictions

The variability in predictions showcased by the various ML models in this study draws attention to the underlying uncertainty levels inherent in their algorithms. One pivotal observation from the results is the disparity between models regarding their permutation importance and SHAP values. The permutation feature importance, while offering a comprehensive view, tends to give a global perspective on predictor significance. On the other hand, SHAP values provide a more granular, instance-specific interpretation of variable importance. Such distinctions emphasize the multifaceted nature of scour depth prediction, wherein each model can process predictor importance differently, thus influencing the final output (Kaur *et al.* 2020). This also underlines the sensitivity of these models to predictors and their interrelationships, which can vary based on the underlying mathematical architecture of each model (Štrumbelj & Kononenko 2014).

Another important aspect to mention is the performance difference across models under uncertainty estimation techniques. Techniques such as MC and BS serve as valuable tools in uncertainty analysis, gauging the robustness and reliability of employed algorithms (Papadopoulos & Yeung 2001). As illustrated, the ANN model emerged as superior in performance, with remarkably consistent predictive capabilities. Such prediction consistency can greatly enhance trust in model outputs, especially when they are used to inform critical decisions. However, it should be noted that high prediction performance does not necessarily guarantee that the model accurately captures real-world complexities (Abdar *et al.* 2021). For instance, despite showing slightly lesser accuracy, the LAS model had lower uncertainty, which implies more consistent predictions. This observation underscores the trade-off between accuracy and consistency, highlighting that higher accuracy does not necessarily translate to more reliable predictions, especially under varied conditions.

A deeper dive into the MC and BS techniques reveals nuanced differences in their capabilities to capture uncertainties. The MC method displayed a subtle tendency towards optimistic predictions (Nguyen *et al.* 2021). In situations where overestimations could lead to wasted resources or other negative consequences, such biases need to be carefully considered. On the other hand, the BS approach provided a slightly wider spread of predictions, indicating a broader scope of possibilities, which might be preferable in applications requiring more conservative estimates. In summary, resolving these uncertainties carefully and accurately will enhance the reliability and trustworthiness of the models and their utility in real-world applications, fostering more informed decision-making processes.

5. CONCLUSIONS

This study advanced the understanding of maximum scour depth prediction behind sluice outlets by evaluating the efficacy of ML models relative to traditional empirical equations. The research notably highlighted how six ML algorithms – RF, ADA, CAT, SVM, LAS, and ANN – outperform the empirical equations of Dey_2006 and Aamir_2022 in predicting maximum scour depth. This achievement underscores the significant potential of integrating ML into hydraulic engineering practices.

The findings demonstrate that variable importance varies across ML models, with the dimensionless densimetric Froude number (F_d) consistently emerging as a pivotal predictor in most models. This research also established that other hydraulic parameters such as d_t/a, F, L/a, and D_{50}/a significantly influence scour depth predictions, reflecting the complex dynamics that govern scouring processes.

The utilization of permutation importance and SHAP values has provided a nuanced view of how predictors impact model outputs, enhancing the interpretability of ML models. While specific predictors maintained consistent importance across models (like F_d or F), the relative influence of others witnessed marked fluctuations, highlighting the multifaceted nature of scour depth prediction and the dynamic interactions between predictors.

Our uncertainty analysis, employing both MC and BS methods, revealed varying performance levels among ML models. The ANN displayed superior performance, marked by higher mean values and minimal standard deviations in the predictions, closely followed by the CAT. Other models like RF, ADA, and SVM also showcased commendable performance. The MC simulations generally leaned slightly towards optimistic predictions compared with the BS approach, as evidenced by marginally higher median values and the differences in their interquartile ranges.

The study, however, recognizes certain limitations. The reliance on specific datasets might affect the generalizability of the findings, and the computational intensity of some ML models could restrict their practical applicability in certain settings. The inherent unpredictability in estimating scour depth at sluice gates requires a combination of empirical knowledge and

advanced computational techniques. The findings of this study advocate for integrating traditional hydraulic understanding with modern ML algorithms to achieve more accurate and reliable maximum scour depth predictions. In light of these findings, future research could further investigate hybrid models that combine the strengths of multiple algorithms to enhance prediction accuracy and reliability.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://github.com/LXHien88/Scour_Depth_ML/.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Aamir, M. & Ahmad, Z. 2016 Review of literature on local scour under plane turbulent wall jets. *Physics of Fluids* **28** (10). doi:10.1063/1.4964659.

Aamir, M. & Ahmad, Z. 2019 Estimation of maximum scour depth downstream of an apron under submerged wall jets. *Journal of Hydroinformatics* **21** (4), 523–540. doi:10.2166/hydro.2019.008.

Aamir, M. & Ahmad, Z. 2022 Effect of apron roughness on flow characteristics and scour depth under submerged wall jets. *Acta Geophysica* **70** (5), 2205–2221. doi:10.1007/s11600-021-00672-9.

Aamir, M., Ahmad, Z., Pandey, M., Khan, M. A., Aldrees, A. & Mohamed, A. 2022 The effect of rough rigid apron on scour downstream of sluice gates. *Water* **14** (14), 2223. doi:10.3390/w14142223.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V. & Nahavandi, S. 2021 A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* **76**, 243–297. doi:10.1016/j.inffus.2021.05.008.

Abd El-Hady Rady, R. 2020 Prediction of local scour around bridge piers: Artificial-intelligence-based modeling versus conventional regression methods. *Applied Water Science* **10** (2), 57. doi:10.1007/s13201-020-1140-4.

Abed, M., Imteaz, M. A. & Ahmed, A. N. 2023 A comprehensive review of artificial intelligence-based methods for predicting pan evaporation rate. *Artificial Intelligence Review* **56** (2), 2861–2892. doi:10.1007/s10462-023-10592-3.

Aderibigbe, O. & Rajaratnam, N. 1998 Effect of sediment gradation on erosion by plane turbulent wall jets. *Journal of Hydraulic Engineering* **124** (10), 1034–1042. doi:10.1061/(ASCE)0733-9429(1998)124:10(1034).

Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. 2010 Permutation importance: A corrected feature importance measure. *Bioinformatics* **26** (10), 1340–1347. doi:10.1093/bioinformatics/btq134.

Azamathulla, H. M. & Ghani, A. A. 2011 ANFIS-based approach for predicting the scour depth at culvert outlets. *Journal of Pipeline Systems Engineering and Practice* **2** (1), 35–40. doi:10.1061/(ASCE)PS.1949-1204.0000066.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32. doi:10.1023/A:1010933404324.

Brownlee, J. 2019 *A Gentle Introduction to Monte Carlo Sampling for Probability*. Available from: https://machinelearningmastery.com/monte-carlo-sampling-for-probability (accessed 9 September 2023).

Chatterjee, S. S., Ghosh, S. N. & Chatterjee, M. 1994 Local scour due to submerged horizontal jet. *Journal of Hydraulic Engineering* **120** (8), 973–992. doi:10.1061/(ASCE)0733-9429(1994)120:8(973).

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20** (3), 273–297. doi:10.1007/BF00994018.

Dey, S. & Sarkar, A. 2006 Scour downstream of an apron due to submerged horizontal jets. *Journal of Hydraulic Engineering* **132** (3), 246–257. doi:10.1061/(ASCE)0733-9429(2006)132:3(246).

Dey, S. & Westrich, B. 2003 Hydraulics of submerged jet subject to change in cohesive bed geometry. *Journal of Hydraulic Engineering* **129** (1), 44–53. doi:10.1061/(ASCE)0733-9429(2003)129:1(44).

Efron, B. & Tibshirani, R. J. 1994 *An Introduction to the Bootstrap*. Chapman and Hall/CRC. doi:10.1201/9780429246593.

Farooq, R. & Ghumman, A. R. 2019 Impact assessment of pier shape and modifications on scouring around bridge pier. *Water* **11** (9), 1761. doi:10.3390/w11091761.

Fitri, A., Hashim, R., Abolfathi, S. & Abdul Maulud, K. N. 2019 Dynamics of sediment transport and erosion-deposition patterns in the locality of a detached low-crested breakwater on a cohesive coast. *Water* **11** (8), 1721. doi:10.3390/w11081721.

Freund, Y. & Schapire, R. E. 1997 A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55** (1), 119–139. doi:10.1006/jcss.1997.1504.

Gewerc, A. 2020 *Bootstrap to Quantify Uncertainty*. Available from: http://www.alangewerc.com/blog/Bootstrap-to-Quantify-Uncertainty (accessed 22 December 2022).

Grana, D., Azevedo, L. & Liu, M. 2020 A comparison of deep machine learning and Monte Carlo methods for facies classification from seismic data. *Geophysics* **85** (4), WA41–WA52. doi:10.1190/geo2019-0405.1.

Guven, A. & Gunal, M. 2008 Genetic programming approach for prediction of local scour downstream of hydraulic structures. *Journal of Irrigation and Drainage Engineering* **134** (2), 241–249. doi:10.1061/(ASCE)0733-9437(2008)134:2(241).

Habib, M. A., O'Sullivan, J. J., Abolfathi, S. & Salauddin, M. 2023 Enhanced wave overtopping simulation at vertical breakwaters using machine learning algorithms. *PLoS ONE* **18** (8), e0289318. doi:10.1371/journal.pone.0289318.

Hamidifar, H., Omid, M. H. & Nasrabadi, M. 2011 Scour downstream of a rough rigid apron. *World Applied Sciences Journal* **14** (8), 1169–1178.

Han, J., Chen, H. & Cao, Y. 2011 Uncertainty evaluation using Monte Carlo method with MATLAB. In *IEEE 2011 10th International Conference on Electronic Measurement & Instruments*, 16–19 August 2011, pp. 282–286. doi:10.1109/ICEMI.2011.6037817.

Hopfinger, E. J., Kurniawan, A., Graf, W. H. & Lemmin, U. 2004 Sediment erosion by Görtler vortices: The scour-hole problem. *Journal of Fluid Mechanics* **520**, 327–342. doi:10.1017/S0022112004001636.

Hüllermeier, E. & Waegeman, W. 2021 Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* **110** (3), 457–506. doi:10.1007/s10994-021-05946-3.

Kartal, V., Emiroglu, M. E., Katipoglu, O. M. & Karakoyun, E. 2023 Prediction of scour hole characteristics caused by water jets using metaheuristic artificial bee colony-optimized neural network and pre-processing techniques. *Journal of Hydroinformatics*. doi:10.2166/hydro.2023.230.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. & Vaughan, J. W. 2020 Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery*, pp. 1–14. doi:10.1145/3313831.3376219.

Khosravi, K., Khozani, Z. S. & Mao, L. 2021 A comparison between advanced hybrid machine learning algorithms and empirical equations applied to abutment scour depth prediction. *Journal of Hydrology* **596**, 126100. doi:10.1016/j.jhydrol.2021.126100.

Le, X.-H. & Le, T. T. H. 2024 Predicting maximum scour depth at sluice outlet: A comparative study of machine learning models and empirical equations. *Environmental Research Communications* **6** (1), 015010. doi:10.1088/2515-7620/ad1f94.

Le, X. H., Nguyen, D. H., Jung, S., Yeon, M. & Lee, G. 2021 Comparison of deep learning techniques for river streamflow forecasting. *IEEE Access* **9**, 71805–71820. doi:10.1109/ACCESS.2021.3077703.

Le, H. T. T., Nguyen, C. V. & Le, D.-H. 2022 Numerical study of sediment scour at meander flume outlet of boxed culvert diversion work. *PLoS ONE* **17** (9), e0275347. doi:10.1371/journal.pone.0275347.

Le, X. H., Eu, S., Choi, C., Nguyen, D. H., Yeon, M. & Lee, G. 2023 Machine learning for high-resolution landslide susceptibility mapping: Case study in Inje County, South Korea. *Frontiers in Earth Science* **11**. doi:10.3389/feart.2023.1268501.

Lim, S.-Y. & Yu, G. 2002 Scouring downstream of sluice gate. In *First International Conference on Scour of Foundations (ICSF-1)*, November 17–20, 2002, pp. 395–409.

Lundberg, S. M. & Lee, S.-I. 2017 A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*. Curran Associates Inc., pp. 4768–4777. doi:10.5555/3295222.3295230.

Mostaani, A. & Azimi, A. H. 2022 Analytical approach for predicting local scour downstream of submerged sluice gate with an apron. *International Journal of Sediment Research* **37** (4), 522–537. doi:10.1016/j.ijsrc.2022.01.003.

Mutlu Sumer, B. 2007 Mathematical modelling of scour: A review. *Journal of Hydraulic Research* **45** (6), 723–735. doi:10.1080/00221686.2007.9521811.

Muzzammil, M. & Alam, J. 2010 ANFIS-based approach to scour depth prediction at abutments in armored beds. *Journal of Hydroinformatics* **13** (4), 699–713. doi:10.2166/hydro.2010.006.

Najafzadeh, M. 2015 Neuro-fuzzy GMDH based particle swarm optimization for prediction of scour depth at downstream of grade control structures. *Engineering Science and Technology, an International Journal* **18** (1), 42–51. doi:10.1016/j.jestch.2014.09.002.

Najafzadeh, M. & Lim, S. Y. 2015 Application of improved neuro-fuzzy GMDH to predict scour depth at sluice gates. *Earth Science Informatics* **8** (1), 187–196. doi:10.1007/s12145-014-0144-8.

Najafzadeh, M., Tafarojnoruz, A. & Lim, S. Y. 2017 Prediction of local scour depth downstream of sluice gates using data-driven models. *ISH Journal of Hydraulic Engineering* **23** (2), 195–202. doi:10.1080/09715010.2017.1286614.

Nguyen, D., Sadeghnejad Barkousaraie, A., Bohara, G., Balagopal, A., McBeth, R., Lin, M.-H. & Jiang, S. 2021 A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. *Physics in Medicine & Biology* **66** (5), 054002. doi:10.1088/1361-6560/abe04f.

Noori, R., Ghiasi, B., Salehi, S., Esmaeili Bidhendi, M., Raeisi, A., Partani, S., Meysami, R., Mahdian, M., Hosseinzadeh, M. & Abolfathi, S. 2022 An efficient data driven-based model for prediction of the total sediment load in rivers. *Hydrology* **9** (2), 36. doi:10.3390/hydrology9020036.

Olsen, N. R. B. & Kjellesvig, H. M. 1998 Three-dimensional numerical flow modeling for estimation of maximum local scour depth. *Journal of Hydraulic Research* **36** (4), 579–590. doi:10.1080/00221689809498610.

Palmer, G., Du, S., Politowicz, A., Emory, J. P., Yang, X., Gautam, A., Gupta, G., Li, Z., Jacobs, R. & Morgan, D. 2022 Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Computational Materials* **8** (1), 115. doi:10.1038/s41524-022-00794-8.

Papadopoulos, C. E. & Yeung, H. 2001 Uncertainty estimation and Monte Carlo simulation method. *Flow Measurement and Instrumentation* **12** (4), 291–298. doi:10.1016/S0955-5986(01)00015-2.

Parsaie, A., Haghiabi, A. H. & Moradinejad, A. 2019 Prediction of scour depth below river pipeline using support vector machine. *KSCE Journal of Civil Engineering* **23** (6), 2503–2513. doi:10.1007/s12205-019-1327-0.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. 2011 Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** (85), 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. 2018 CatBoost: Unbiased boosting with categorical features. *ArXiv*. 10.48550/arXiv.1706.09516.

Qaderi, K., Javadi, F., Madadi, M. R. & Ahmadi, M. M. 2021 A comparative study of solo and hybrid data driven models for predicting bridge pier scour depth. *Marine Georesources & Geotechnology* **39** (5), 589–599. doi:10.1080/1064119X.2020.1735589.

Rezaie-Balf, M. 2019 Multivariate adaptive regression splines model for prediction of local scour depth downstream of an apron under 2D horizontal jets. *Iranian Journal of Science and Technology, Transactions of Civil Engineering* **43** (1), 103–115. doi:10.1007/s40996-018-0151-y.

Rodríguez-Pérez, R. & Bajorath, J. 2020 Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design* **34** (10), 1013–1026. doi:10.1007/s10822-020-00314-0.

Sarathi, P., Faruque, M. A. A. & Balachandar, R. 2008 Influence of tailwater depth, sediment size and densimetric Froude number on scour by submerged square wall jets. *Journal of Hydraulic Research* **46** (2), 158–175. doi:10.1080/00221686.2008.9521853.

Sharafati, A., Tafarojnoruz, A., Shourian, M. & Yaseen, Z. M. 2020 Simulation of the depth scouring downstream sluice gate: The validation of newly developed data-intelligent models. *Journal of Hydro-environment Research* **29**, 20–30. doi:10.1016/j.jher.2019.11.002.

Sharafati, A., Haghbin, M., Motta, D. & Yaseen, Z. M. 2021 The application of soft computing models and empirical formulations for hydraulic structure scouring depth simulation: A comprehensive review, assessment and possible future research direction. *Archives of Computational Methods in Engineering* **28** (2), 423–447. doi:10.1007/s11831-019-09382-4.

Sreedhara, B. M., Patil, A. P., Pushparaj, J., Kuntoji, G. & Naganna, S. R. 2021 Application of gradient tree boosting regressor for the prediction of scour depth around bridge piers. *Journal of Hydroinformatics* **23** (4), 849–863. doi:10.2166/hydro.2021.011.

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

Tibshirani, R. 1996 Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** (1), 267–288.

Tin Kam, H. 1998 The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8), 832–844. doi:10.1109/34.709601.

Ustimenko, A., Prokhorenkova, L. & Malinin, A. 2020 Uncertainty in gradient boosting via ensembles. *ArXiv*. 10.48550/arXiv.2006.10562.

Verma, D. V. S. & Goel, A. 2005 Scour downstream of a sluice gate. *ISH Journal of Hydraulic Engineering* **11** (3), 57–65. doi:10.1080/09715010.2005.10514801.

Yeganeh-Bakhtiary, A., Houshangi, H. & Abolfathi, S. 2020 Lagrangian two-phase flow modeling of scour in front of vertical breakwater. *Coastal Engineering Journal* **62** (2), 252–266. doi:10.1080/21664250.2020.1747140.

Yousif, A. A., Sulaiman, S. O., Diop, L., Ehteram, M., Shahid, S., Al-Ansari, N. & Yaseen, Z. M. 2019 Open channel sluice gate scouring parameters prediction: Different scenarios of dimensional and non-dimensional input parameters. *Water* **11** (2), 353. doi:10.3390/w11020353.

First received 17 December 2023; accepted in revised form 31 May 2024. Available online 10 June 2024