Water Supply

IWA PUBLISHING

Water Supply Vol 23 No 1, 304 doi: 10.2166/ws.2022.421

Predictive explicit expressions from data-driven models for estimation of scour depth below ski-jump bucket spillways

Reza Shafagh Loron[a], Mehrshad Samadi[b,*] and Abolfazl Shamsai[a]

[a] School of Civil Engineering, Sharif University of Technology, P.O. Box 11155-9313, Tehran, Iran
[b] School of Civil Engineering, Iran University of Science and Technology (IUST), Narmak, P.O. Box 16765-163, Tehran, Iran
*Corresponding author. E-mail: mehrshad1364@gmail.com, mehrshad_samadi@alumni.iust.ac.ir

MS, 0000-0003-3867-0264

ABSTRACT

Scour depth estimation is an essential factor in water-related engineering problems. Scouring below spillways may endanger a dam's stability and even lead to dam destruction. As a result, it has undesirable environmental effects due to dam failure. Hence, reliable and accurate scour depth estimation below spillways is an exciting topic for researchers. For this purpose, the published and reliable prototype data related to scour depth below ski jump bucket spillways (D_s) was used to develop data-driven models. This study employed two widely used decision tree (DT) methods, including the M5 model tree (M5MT) and the classification and regression tree (CART), and also multivariate adaptive regression splines (MARS) for the estimation of (D_s). The proposed methods provided explicit and clear equations with straightforward applications for estimating scour depth. For the quantitative assessments of the developed formulas, three common statistical metrics, namely root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (CC), were used. Moreover, comparison results with previous approaches existing in the literature indicated the efficacy of the suggested methods. The obtained results revealed that the MARS technique was the best approach for the estimation of scour depth.

Key words: CART, data-driven models, MARS, model tree, scour depth, spillways

HIGHLIGHTS

- Predictive equations were developed to estimate the scour depth below ski-jump bucket spillways.
- White-box data-driven models were evaluated in this study.
- The MARS model provided more accurate results when compared to decision tree methods and empirical formulas for scour depth prediction.
- Field measurements were used in this study.

1. INTRODUCTION

The scouring phenomenon of spillways, bridges, piers, culverts, and other hydraulic structures is a critical issue and one of the most interesting of hydraulic engineering problems (Samadi *et al.* 2014; Malik *et al.* 2021; Chou & Nguyen 2022; Daneshfaraz *et al.* 2022; Kartal & Emiroglu 2022). The scouring below spillways can jeopardize a dam's safety. The destruction of a dam causes severe damage to the economy, environment, and human life downstream. Therefore, scour below spillways should be monitored and its quantity measured. Scour depth modeling below spillways is one of the most significant challenges in hydraulic engineering research. Estimating scour depth is essential in dam design and assessing its operational safety. Due to the non-linear behavior and stochastic nature of the scouring process, various data-driven methods have been proposed for modeling scour depth (Homaei & Najafzadeh 2020; Pandey *et al.* 2020; Ahmadianfar *et al.* 2022; Devi & Kumar 2022; Homaei & Najafzadeh 2022; Nimbalkar *et al.* 2022; Rathod & Manekar 2022). Moreover, data-driven models are widely and successfully used for modeling water-related problems (Mojaradi *et al.* 2018; Ghasemi *et al.* 2022).

A literature review indicated that the applications of data-driven approaches for modeling scour downstream of the ski-jump bucket spillways can be classified into two study groups. Researchers have used experimental results and field measurements to model scour in both these groups of studies. It should be noted that little field data is published and available in the

literature due to the difficulty of measuring scour depth downstream of dams. Nevertheless, Azmathullah *et al.* (2006) highlighted the importance of field measurements in accurately modeling scour depth. They recommended the application of prototype data for scour depth estimation to more accurately represent natural circumstances than experimental work conducted under controlled conditions and influenced by scale effects. Therefore, developing data-driven approaches seems necessary for predicting scour depth using field measurements.

Regarding experimental work, the following studies have been conducted using data-driven models to estimate and model the scour downstream of ski-jump bucket spillways. Azmathullah *et al.* (2005) implemented an artificial neural network (ANN) to predict scour hole characteristics. They indicated the outperformance of ANN compared to traditional regression methods. Agarwal *et al.* (2010) employed locally weighted projection regression (LWPR) and found that LWPR was more efficient than ANN. Goyal & Ojha (2011) indicated favorable efficiency for support vector machine (SVM) and M5 model tree (M5MT) compared to ANN. Najafzadeh *et al.* (2014) indicated that a combination of the group method of data handling (GMDH) approach with back-propagation (BP) algorithms was more accurate compared to ANN, genetic programming, adaptive neural fuzzy inference system (ANFIS), and conventional equations. Noori *et al.* (2017) modeled the dimensions of a scour hole using the granular computing (GrC) method and demonstrated the potential of the GrC method. Nou *et al.* (2021) combined the ANFIS with particle swarm optimization (PSO) algorithms and demonstrated the superiority of the ANFIS-PSO approach over the stand-alone ANFIS method. Sun *et al.* (2021) used a combination of support vector regression (SVR) with fruitfly optimization algorithms (FOAs). Their proposed method improved the accuracy of scour depth prediction compared with a stand-alone SVR method.

Concerning field measurements, the following studies have been made to estimate the scour downstream of ski-jump bucket spillways using data-driven approaches. Azmathullah *et al.* (2006) and Azmathullah *et al.* (2008a) found that ANN and ANFIS could better predict the depth of scour than traditional formulas. Guven & Azamathulla (2012) provided mathematical expressions using gene expression programming (GEP) to estimate normalized scour depth. Sammen *et al.* (2020) introduced the hybridization of ANN with Harris hawks optimization (ANN-HHO), PSO (ANN-PSO), and genetic algorithm (ANN-GA) to predict normalized scour depth. They illustrated the efficiency of ANN-HHO compared to the ANN-PSO and ANN-GA models.

The literature review indicated that data-driven methods developed using prototype data were fewer than those developed using experimental data. To the authors' knowledge there was no published study conducted using multivariate adaptive regression splines (MARS) and decision tree (DT) approaches using prototype data and nondimensional parameters to estimate scour depth below ski jump bucket spillways. Therefore, this study used field measurements of scour depth below ski jump bucket spillways to develop MARS and two well-known DT approaches, including M5MT and classification and regression tree (CART) algorithms. This research investigated the proposed models' effectiveness and compared their results with existing previous approaches using statistical analysis and graphical evaluation. It is worth mentioning that the main features of the proposed methods derive explicit equations to predict scour depth compared to black-box data-driven methods such as the ANN model. The suggested predictive formulas are beneficial for practical engineering in real-world applications and reduce potential safety risks.

2. MATERIALS AND METHODS

This section presents a description of field measurements used for scour depth estimation. In addition, a brief overview of the data-driven methods used to model scour depth, such as MARS, CART, and M5MT, is also presented.

2.1. Field measurements and existing approaches

A limited number of field measurements of scour depth below spillways have been reported in the literature. This study used published data of scour below a ski-jump bucket spillway reported by Azamathulla *et al.* (2008b). They reported the head between the upper water level (reservoir level) and the tailwater level, H_1, (m), discharge intensity, q, ($m^3/s/m$), and depth of scour, D_s, (m) of 82 field measurements of various dams.

A literature review indicated that some traditional formulas are suggested for predicting scour depth below spillways. Table 1 shows some traditional formulas for calculating D_s (Mason & Arumugam 1985; Azamathulla *et al.* 2008b; Kumar & Sreeja 2012; Azamathulla 2013; Khatsuria 2013).

Table 1 | Various proposed formulas for the estimation of scour depth below spillways, as reported by various researchers (Mason & Arumugam 1985; Azamathulla et al. 2008b; Kumar & Sreeja 2012; Azamathulla 2013; and Khatsuria 2013)

Approach	Formula
Veronese-(B) (1937)	$D_s = 1.90q^{0.54}H_1^{0.225}$
Wu (1973)	$\frac{D_s}{H_1} = 2.11(Fr_1)^{0.51}$
Martins-(B) (1975)	$\frac{D_s}{H_1} = 2.976(Fr_1)^{0.6}$
Taraimovich (1978)	$D_s = 0.633q^{0.67}H_1^{0.25}$
Sofrelec (1980)	$D_s = 2.3q^{0.6}H_1^{0.1}$
Incyth (1982)	$D_s = 1.413q^{0.5}H_1^{0.25}$
CWPRS (1986)	$\frac{D_s}{H_1} = 2.7656(Fr_1)^{0.6224}$
Azmathullah *et al.* (2006)	$D_s = 1.42q^{0.44}H_1^{0.3}$
Kumar & Sreeja (2012)	$D_s = 1.16q^{0.7}H_1^{0.25}$

where $Fr_1 = \frac{q}{\sqrt{gH_1^3}}$ is defined as the Froude number, and g is acceleration due to gravitation.

It is worth mentioning that Guven & Azamathulla (2012) used the GEP approach, which is a robust white-box data-driven method that provided a mathematical expression for the estimation of normalized scour depth (D_s/H_1):

$$\frac{D_s}{H_1} = (Fr_1^2 e^{(0.861/Fr_1)^{Fr_1}})^{0.5} - 0.362(10^{Fr_1} - 2.734)^{-1} + (0.895Fr_1^{0.5} - 0.024) \quad (1)$$

2.2. Multivariate adaptive regression splines (MARS)

Friedman (1991) developed the concept of multivariate adaptive regression splines (MARS). The main advantage of the MARS method, which generates a flexible mathematical formula using piecewise linear regression models, is that it does not require hypotheses concerning the relation between input and output variables (Parsaie *et al.* 2018). The MARS approach uses linear functions that can model nonlinear systems with a reduced degree of complexity in formulating the problem. MARS estimates an output parameter using the linear combination of many basis functions (BFs). A BF illustrates the relationship between inputs and outputs. MARS constructs an explicit equation to determine the output parameter (y) in the following general form (Sihag *et al.* 2021):

$$y = B_0 + \sum_{m=1}^{M} B_m \beta_m(x) \quad (2)$$

where B_0 is a constant value, x is the input variable, B_m is the corresponding coefficient of each BF, and M is the total number of BFs. β_m is the mth BF which is defined as follows (Yonesi *et al.* 2022):

$$\beta_m(x) = \max(0, c - x) \quad \text{or} \quad \beta_m(x) = \max(0, x - c) \quad (3)$$

The MARS model is constructed in two steps: forward and backward. In the forward step, all possible BFs are added to a MARS model that may result in an overfitted model. Afterward, the BFs of less importance are eliminated in the backward step concerning the generalized cross-validation (GCV) criterion.

2.3. Decision trees (DTs)

Decision tree (DT) algorithms provide a set of logical rules that are used for classification and regression issues. The main concept of DTs for solving a complex problem is to divide the input domain of the problem into several subdomains and create a specialized model for each sub-domain (Enayati *et al.* 2022; Torabi *et al.* 2022). This work reduces the degree of complexity of the problem with the combination of local models and enhances the predictive capability and accuracy of the model. The result of a DT is expressed as a hierarchical inverse tree-like structure with split rules into internal and terminal

nodes and provides the predictive models in each terminal node. A DT is divided into two or more groups at each internal node based on the specific DT algorithm. A binary DT creates two branches in each internal node. To produce a DT, an inference method or division condition is used during the tree development process. The model's division criterion involves calculating the standard deviation of the class values entering the node as an error value and calculating the predicted reduction in this error as the test result for each feature in that node.

A CART is a binary DT that can be used for classification and regression problems (Breiman 1984). Each internal node classifies the data into two groups by a simple if-then rule based on a single variable (Kamranzad *et al.* 2013). In each class, the response factor must optimize homogeneity while minimizing total deviation. Another popular binary DT is the M5 model tree (M5MT). The M5MT was introduced by Quinlan (1992) as a DT model used exclusively for regression problems and numerical prediction. The outcome of M5MT is the extraction of knowledge from a tree structure in the form of if-then rules, considering splitting variables, the range of splitting variables, and multivariate linear regression models on the leaves.

The results of M5MT are based on its providing linear regression models at terminal nodes for the estimation of the output parameter (Sihag *et al.* 2022; Singh *et al.* 2022). The input variable within the internal node of the tree is selected based on the feature that results in the greatest possible decrease in predicted error when measured against the standard deviation of the output parameter (Khosravi *et al.* 2022). The result of a terminal node can be expressed as:

$$O = w_0 + w_1 x_1 + w_2 x_2 + \ldots \tag{4}$$

where O is the output parameter, $w_1, w_2, \ldots$ are the coefficients of the multiple linear regression model, and $x_1, x_2, \ldots$ are the input variables that contribute to the prediction of the output parameter.

In summary, M5MT and CART are popular and widely used binary DTs for predictive purposes. M5MT and CART have several advantages, such as simple and understandable construction, reduced computing costs, and visual depiction. The major difference between M5MT and CART is that M5MT generates multivariate linear functions in terminal nodes while CART provides constant numerical values. The rules of DTs are clear and easy to use for everyone. More details about M5MT and CART algorithms can be found in Wang & Witten (1997) and Breiman (1984).

3. MODEL DEVELOPMENT FOR ESTIMATION OF SCOUR DEPTH

Training and testing datasets are essential for the construction of data-driven models. Hence, 80% of the data set was used for training, and 20% remained for the testing set. Additionally, non-dimensionless parameters were utilized in the creation of the proposed models. Therefore, $q/\sqrt{gH_1^3}$ and (D_s/H_1) were considered input and output variables, respectively. It is worth mentioning that the earlier studies conducted by Guven & Azamathulla (2012) and Sammen *et al.* (2020) employed dimensionless parameters to predict scour depth using field data. The statistical characteristic values of input and output variables are listed in Table 2.

The MARS algorithm provided the simple linear equations for the estimation of D_S/H_1 as follows:

$$\begin{aligned} \frac{D_S}{H_1} &= 1.13584 + 1.21025 \times BF_1 - 4.92422 \times BF_2 \\ BF_1 &= \max\,(0, Fr_1 - 0.106425) \\ BF_2 &= \max\,(0, 0.186907 - Fr_1) \end{aligned} \tag{5}$$

As can be seen, the simple mathematical expressions were obtained with two BFs for predicting scour depth.

Table 2 | The statistical values of training, testing, and all the data sets used for the development of data-driven models

	Train dataset			Test dataset			All dataset		
Parameter	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**
Fr_1	0.0040	4.4699	0.1869	0.0088	0.1850	0.0696	0.0040	4.4699	0.1669
D_s/H_1	0.0572	6.3500	0.7315	0.1687	1.2936	0.6323	0.0572	6.3500	0.7146

It is worth noting that the number of BFs and the value of the GCV criterion are important in developing the MARS model in order to find the MARS equation for the estimation of D_s/H_1. As stated previously, the MARS model is developed in two steps. For generating the MARS equation, 12 BFs were considered in the first step, and in the second stage (the pruning stage), 10 BFs were removed. As a result, the final MARS equation with 2 BFs was obtained for the estimation of D_s/H_1. Furthermore, the value of the GCV criterion for the MARS equation was equal to 0.07617.

It is noticeable that the MARS equation (Equation (5)) is similar to a compact DT because, concerning the value of Fr_1, the MARS equation can be converted to three simple if-then rules, which are obtained as follows:

$$\begin{aligned}
&\text{Rule 1: If} \quad Fr_1 \le 0.106425 \rightarrow \frac{D_S}{H_1} = 1.13584 - 4.92422 \times (0.0186907 - Fr_1)\\
&\text{Rule 2: If} \quad 0.106425 < Fr_1 \le 0.186907 \rightarrow \frac{D_S}{H_1} = 1.13584 + 1.21025 \times (Fr_1 - 0.106425) - 4.92422 \times (0.186907 - Fr_1)\\
&\text{Rule 3: If} \quad Fr_1 > 0.186907 \rightarrow \frac{D_S}{H_1} = 1.13584 + 1.21025 \times (Fr_1 - 0.106425)
\end{aligned} \tag{6}$$

Since the data set chosen for this paper doesn't have any categorical data, CART created a regression tree for predicting scour depth. The CART algorithm was used to make a simple regression tree, as shown in Figure 1.

As seen in Figure 1, two branches have divided the domain of the problem into two terminal nodes, which contain constant numerical values, and this has terminated the growth of the CART tree with 2 rules. This is an important point: the Least Squared Deviation (LSD) impurity measure is employed for splitting rules and goodness of fit criteria. In addition, each terminal node's estimated category is the weighted average of the target values for records in the node.

Eventually, a simple regression tree generated by the CART method was constructed with two branches and two terminal nodes. As previously stated, a regression tree by the CART method provides constant numerical values for the estimation of D_S/H_1. The equations related to the CART tree can be expressed as follows:

$$\begin{aligned}
&\text{Rule 1: If} \quad Fr_1 \le 0.139 \rightarrow \frac{D_S}{H_1} = 0.404\\
&\text{Rule 2: If} \quad Fr_1 > 0.139 \rightarrow \frac{D_S}{H_1} = 1.569
\end{aligned} \tag{7}$$

Regarding the value of Fr_1, the appropriate rule was selected and immediately computed D_s/H_1 without the need to conduct any mathematical calculations. Finally, the M5MT approach presented a regression tree, as illustrated in Figure 2.

As seen in Figure 2, the M5MT divided the input domain of the problem into two sub-subdomains and provided two linear regression models at two terminal nodes for the estimation of D_s/H_1. M5MT constructs a regression tree by recursive splitting based on treating the standard deviation of the class values that reach a node as a measure of the error at the node. In addition, the M5MT method employs a pruning procedure to avoid overfitting the obtained linear models. After pruning, M5 used a smoothing procedure to compensate for discontinuities that occurred due to the pruning procedure. Therefore, the smoothed and pruned M5MT was generated using training data for the estimation of D_s/H_1. The coefficients of linear models in M5MT are obtained using the least-squares method.

The related rules of Figure 2 are as follows:

$$\begin{aligned}
&\text{LM\# 1: If} \quad Fr_1 \le 0.139 \rightarrow \frac{D_S}{H_1} = 2.7519 \times Fr_1 + 0.315\\
&\text{LM\# 2: If} \quad Fr_1 > 0.139 \rightarrow \frac{D_S}{H_1} = 1.3184 \times Fr_1 + 0.7578
\end{aligned} \tag{8}$$

As seen, compared to constant numerical values provided by the CART method in terminal nodes, the M5 DT presented multivariate linear models in terminal nodes, which increased the M5 models' flexibility for scour depth estimation. As previously discussed, regression trees and M5MT are used to solve regression problems. However, the main difference between M5MT and regression trees is that the leaves of the regression trees have a constant value. In contrast, M5MT provides linear models in their leaves, which can predict numeric values for a given data sample.

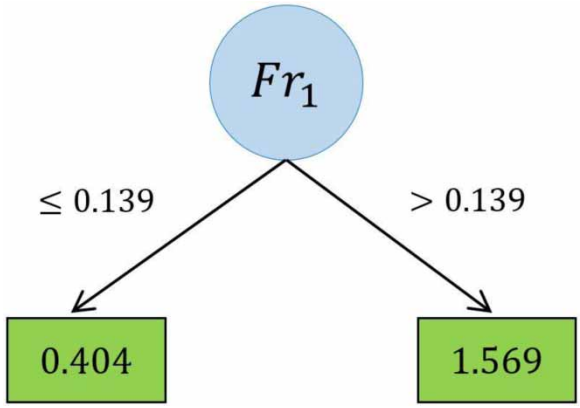

Figure 1 | The regression tree generated by CART for the estimation of $\frac{D_S}{H_1}$.

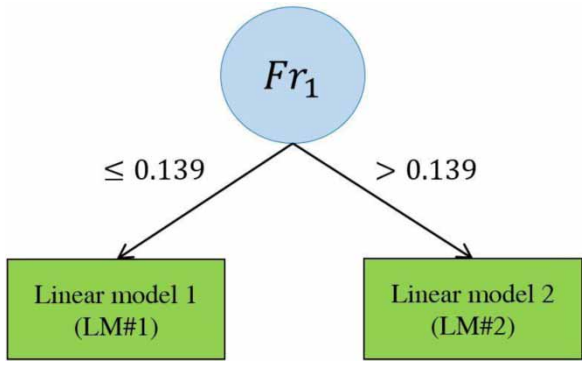

Figure 2 | The regression tree created by M5MT for the estimation of $\frac{D_S}{H_1}$.

Regarding Figures 1 and 2, the CART and M5 models have similar structures, and the splitting value for Fr_1 was 0.139. The value of the splitting criterion for Fr_1 is established by optimizing the training data set to improve the estimation and minimize the estimation error for the training data, but they do not necessarily have a physical significance. This issue was highlighted and expressed by previous researchers (Bhattacharya *et al.* 2007; Bonakdar & Etemad-Shahidi 2011; Samadi *et al.* 2014).

4. PERFORMANCE CRITERIA OF DEVELOPMENT MODELS

Three common statistical indicators are utilized to evaluate scour depth prediction formulas: correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE). These statistical indices are as follows:

$$CC = \frac{\sum_{i=1}^{i=n}[((D_s/H_1)_i^{mess} - \overline{(D_s/H_1)^{mes}})((D_s/H_1)_i^{pre} - \overline{(D_s/H_1)^{pre}})]}{\sqrt{\sum_{i=1}^{i=n}((D_s/H_1)_i^{mess} - \overline{(D_s/H_1)^{mes}})^2 \cdot \sqrt{\sum_{i=1}^{i=n}((D_s/H_1)_i^{pre} - \overline{(D_s/H_1)^{pre}})^2}}} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}((D_s/H_1)_i^{mes} - (D_s/H_1)_i^{pre})^2} \quad (10)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{i=n}|(D_s/H_1)_i^{mes} - (D_s/H_1)_i^{pre}| \quad (11)$$

where $(D_s/H_1)_i^{mes}$ and $(D_s/H_1)_i^{pre}$ are measured and predicted scour depth. In addition, $\overline{(D_s/H_1)^{mes}}$ and $\overline{(D_s/H_1)^{pre}}$ are averaged of measured and predicted scour depth. The total amount of data is denoted by n.

5. RESULTS AND DISCUSSION

The statistical measurement values of data-driven models are tabulated in Table 3.

For the training dataset, the CC, RMSE, and MAE values of MARS were 0.9584, 0.2545, and 0.1761, respectively. The CC, RMSE, and MAE values of M5MT were 0.9531, 0.2716, and 0.1887. The CC, RMSE, and MAE values of CART were 0.6896, 0.6631, and 0.2925, respectively. Therefore, MARS outperformed M5MT and CART in the training stage. In addition, the CC, RMSE, and MAE values of MARS were 0.7734, 0.2374, and 0.1429 more precise than M5MT with CC = 0.7579, RMSE = 0.2643, and MAE = 0.1853, and the CART method with CC = 0.6246, RMSE = 0.3413, and MAE = 0.2450 for testing data sets. Therefore, based on the values of statistical indices for training and testing data sets, the MARS model's performance was obviously better than the M5MT and CART algorithms.

MARS was similar to the compact DT method in that it provided three if-then rules for estimating scour depth with regard to the value of Fr_1. However, both DT algorithms, i.e., M5MT and CART models, had similar structures with the same splitting criterion (i.e., 0.139). Moreover, M5MT and CART provided two if-then rules for scour depth predictions. However, compared to constant numerical values presented by CART in terminal nodes, M5MT provided multivariate linear models that increased the generalizability of M5MT. This issue improved the power prediction of M5MT compared to CART. The decision rules were obtained from the DT methods that employed a single variable, i.e., Fr_1, for scour depth prediction. The appropriate rule was selected based on the value of Fr_1 and simply computed scour depth. As observed, these rules were easy to use for computing scour depth.

However hand, MARS provided more rules (three rules) and caused more flexibility and generalizability for the estimation of scour depth, while M5MT and CART generated two rules for scour depth predictions. Overall, the MARS equation is superior to decision rules obtained from M5MT and CART methods. Furthermore, the MARS model, as the best predictive formula, was compared to earlier robust data-driven models reported by Guven & Azamathulla (2012) and Sammen *et al.* (2020). Table 4 summarizes the values of statistical indices of the MARS, ANN-HHO, and GEP for the estimation of D_s/H_1.

Table 3 | The values of the statistical measurements of developed models for the estimation of $\frac{D_S}{H_1}$

Model	CC	RMSE	MAE
MARS (Train)	0.9584	0.2545	0.1761
MARS (Test)	0.7734	0.2374	0.1429
M5MT (Train)	0.9531	0.2716	0.1887
M5MT (Test)	0.7579	0.2643	0.1853
CART (Train)	0.6896	0.6631	0.2925
CART (Test)	0.6246	0.3413	0.2450

Table 4 | The proposed MARS model was compared to ANN-HHO presented by Sammen *et al.* (2020) and GEP presented by Guven & Azamathulla (2012) in the training and testing stages for the estimation of D_s/H_1

Approach	Category	CC	RMSE	MAE
MARS (Present study)	Training	0.9584	0.2545	0.1761
MARS (Present study)	Testing	0.7734	0.2374	0.1429
ANN-HHO (Sammen *et al.* 2020)	Training	0.9557	0.2626	0.1791
ANN-HHO (Sammen *et al.* 2020)	Testing	0.7765	0.2538	0.1760
GEP (Guven & Azamathulla 2012)	Training	0.9564	0.3606	0.1957
GEP (Guven & Azamathulla 2012)	Testing	0.7813	0.2582	0.1826

As observed in Table 4, in the training phase, the values of CC, RMSE, and MAE of the MARS approach were 0.9584, 0.2545, and 0.1761, respectively, followed by ANN-HHO with CC = 0.9557, RMSE = 0.2626, and MAE = 0.1791, and GEP with CC = 0.9564, RMSE = 0.3606, and MAE = 0.1957. Similarly, regarding error values, MARS has the best performance with RMSE = 0.2374 and MAE = 0.1429 compared to ANN-HHO with RMSE = 0.2538 and MAE = 0.1760 and GEP with RMSE = 0.2582 and MAE = 0.1826 in the testing phase. So, it can be concluded that the MARS approach is the best predictive data-driven model for estimating D_s/H_1.

In addition, compared to ANN-HHO, the MARS method provided simple mathematical expressions that easily and quickly replaced the value of Fr_1 in the MARS equation and estimated scour depth without needing any software or computer programming knowledge. In the following, the proposed data-driven methods were compared with traditional formulas for estimating scour depth. The values of statistical indices are computed and presented in Table 5.

The error values indicated that the Incyth formula for all datasets had the minimum RMSE and MAE values among the traditional formulas. Comparing the values of statistical indices of the Incyth formula (CC = 0.9415, RMSE = 0.2873, and MAE = 0.2002) with the MARS method (CC = 0.9526, RMSE = 0.2517, and MAE = 0.1705) revealed the best performance of MARS for estimation of scour depth. Further, another proposed data-driven approach, i.e., M5MT, with CC = 0.9463, RMSE = 0.2703, and MAE = 0.1881, was slightly better than the Incyth formula.

The good results of M5MT revealed that dividing the input domain into two sub-domains and fitting linear regression models increased the performance of M5MT for the estimation of scour depth. So, the nonlinearity of scour depth can be modeled by splitting two multiple linear regression functions as much as possible. However, another DT method, i.e., the CART method, had the weakest performance for the estimation of D_s/H_1. The CART model provided constant numerical values for the estimation of D_s/H_1 and failed to model the nonlinearity behavior that exists between input and output variables. Nevertheless, the CART method presented the simplest expressions, which can be useful for quickly estimating scour depth. The scatter plots and results of the MARS, M5MT, and CART methods for training and testing data sets are shown in Figures 3–8.

As seen in the scatter plots and outcomes of the proposed data-driven models, it was clearly observed that the MARS model has the best performance in estimating D_s/H_1 in the training and testing stages. The graphical evaluation results confirm the accuracy of MARS for the prediction of D_s/H_1. Finally, some examples of outcomes of the proposed models, including MARS, M5MT, and CART, are provided in Table 6.

As observed in Table 6, the values of D_s/H_1 predicted by the MARS model were the closest results to measurements of D_s/H_1. It is worth mentioning that previous studies conducted by Samadi & Jabbar (2012), Samadi *et al.* (2015), Haghiabi (2017), Rezaie-Balf (2019), Samadi *et al.* (2020), Samadi *et al.* (2021), and Najafzadeh & Oliveto (2022) have shown the potential and capability of the MARS model for estimating scour depth.

Table 5 | Comparing MARS, M5MT, and CART models with traditional formulas (reported by various researchers, including by Mason & Arumugam 1985; Azamathulla *et al.* 2008b; Kumar & Sreeja 2012; Azamathulla 2013 and Khatsuria 2013) for the estimation of D_s/H_1 for all datasets

Approach	CC	RMSE	MAE
MARS (Present study)	0.9526	0.2517	0.1705
M5MT (Present study)	0.9463	0.2703	0.1881
CART (Present study)	0.6746	0.6201	0.2844
Veronese-(B) (1937)	0.9392	0.5332	0.3880
Wu (1973)	0.9426	0.3427	0.2064
Martins-(B) (1975)	0.9479	0.3097	0.2098
Taraimovich (1978)	0.8871	0.4348	0.2404
Sofrelec (1980)	0.9479	0.8268	0.4583
Incyth (1982)	0.9415	0.2873	0.2002
CWPRS (1986)	0.9480	0.2928	0.2014
Azmathullah *et al.* (2006)	0.9416	0.3261	0.2027
Kumar & Sreeja (2012)	0.8723	0.7890	0.5825

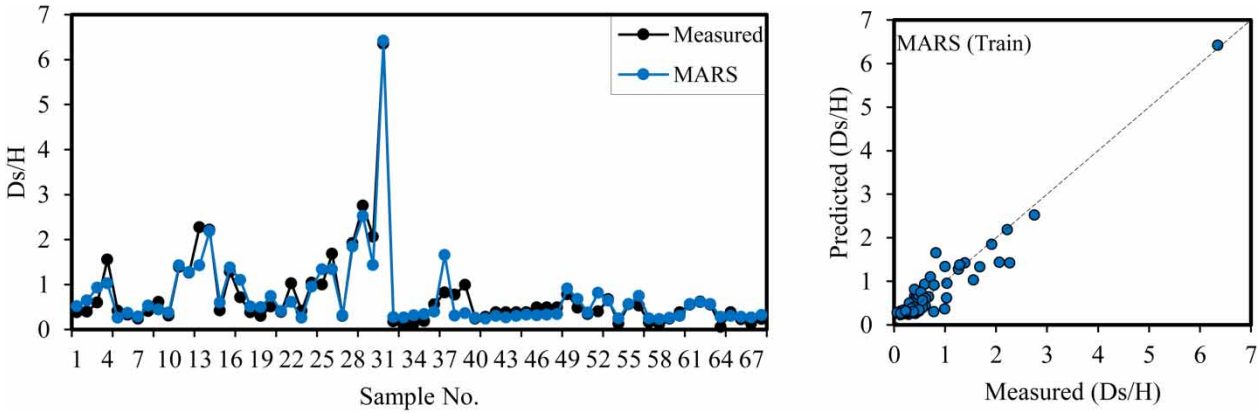

Figure 3 | Comparison of the estimation of D_S/H_1 values using the MARS method versus measured values in the training data set.

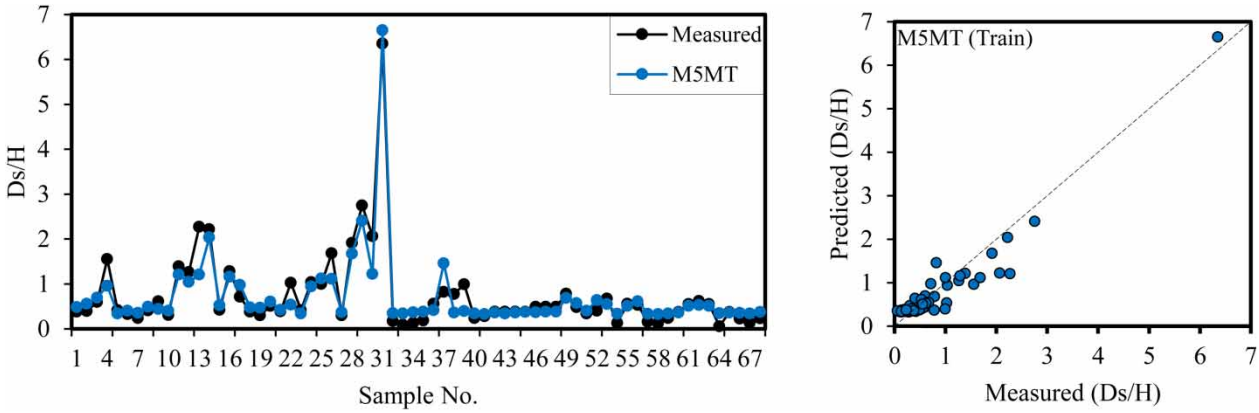

Figure 4 | Comparison of the estimation of D_S/H_1 values using the M5MT method versus measured values in the training data set.

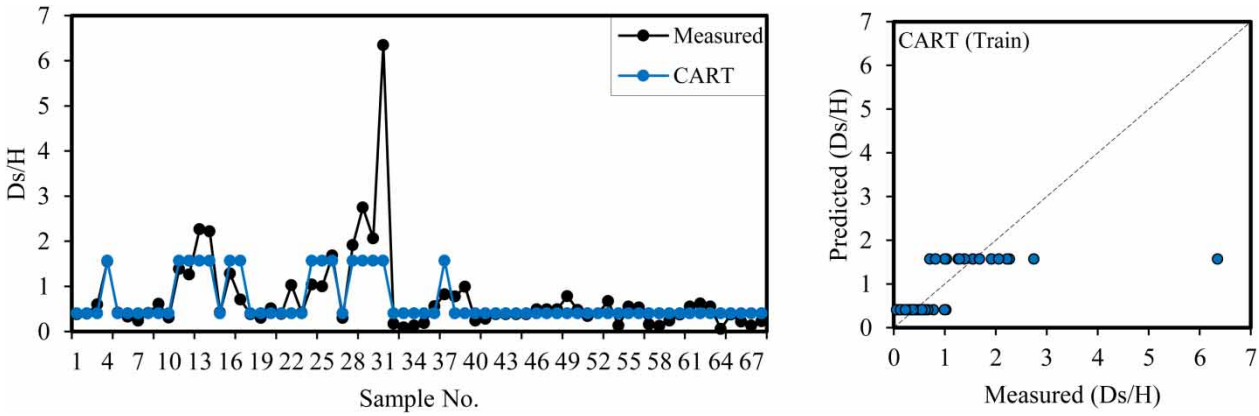

Figure 5 | Comparison of the estimation of D_S/H_1 values using the CART method versus measured values in the training data set.

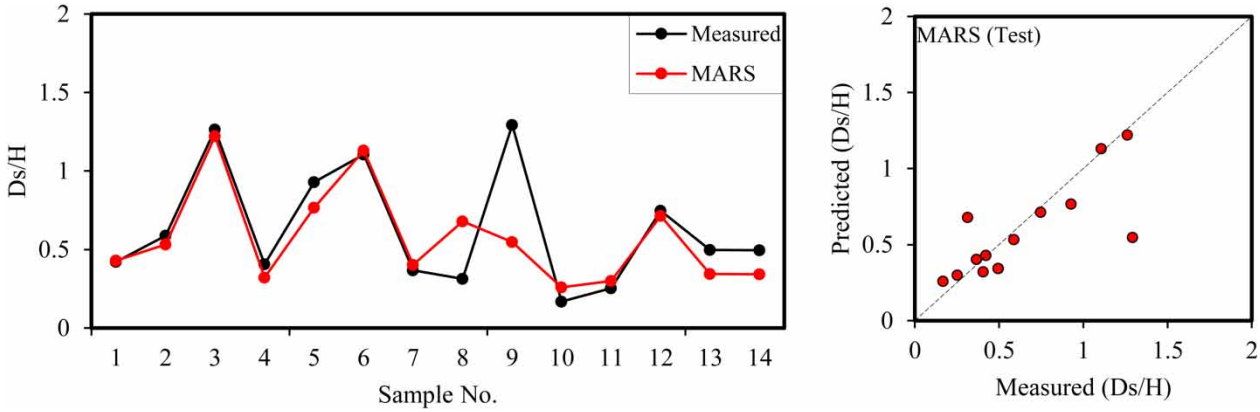

Figure 6 | Comparison of the estimation of D_s/H_1 values using the MARS method versus measured values in the testing data set.

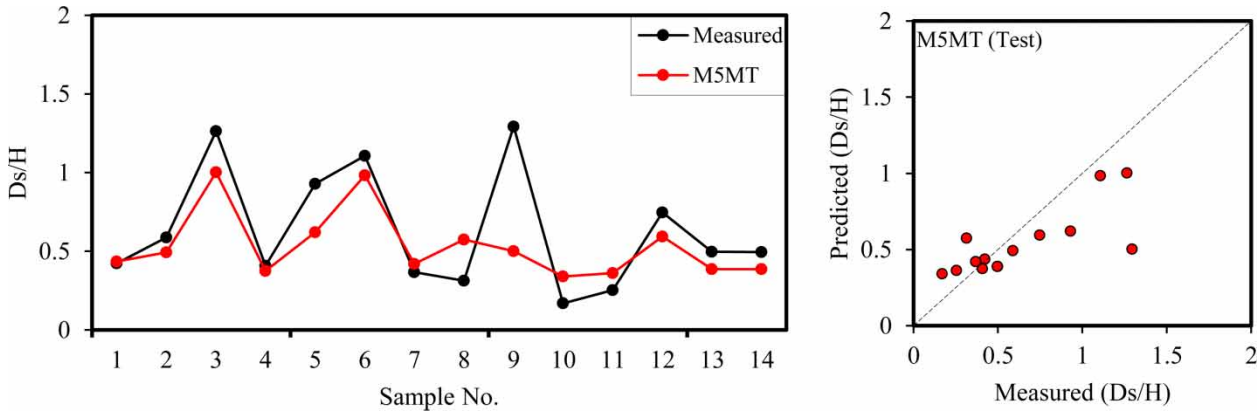

Figure 7 | Comparison of the estimation of D_s/H_1 values using the M5MT method versus measured values in the testing data set.

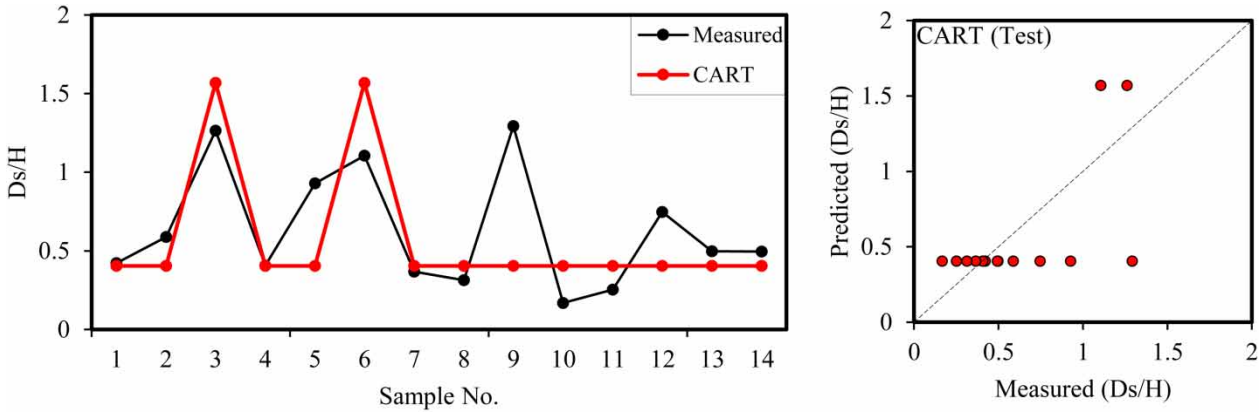

Figure 8 | Comparison of the estimation of D_s/H_1 values using the CART method versus measured values in the testing data set.

Table 6 | The values of D_s/H_1 resulting from the MARS, M5MT, and CART models

Sample No.	Fr_1	D_s/H_1	MARS	M5MT	CART
1	0.044	0.422	0.430	0.435	0.404
2	0.153	1.555	1.028	0.960	1.569
3	0.141	1.038	0.950	0.943	1.569
4	0.185	1.263	1.222	1.002	1.569
5	0.019	0.300	0.311	0.368	0.404
6	0.032	0.344	0.372	0.403	0.404
7	0.071	0.556	0.567	0.512	0.404
8	0.013	0.223	0.282	0.352	0.404

6. SUMMARY AND CONCLUSIONS

The scouring process downstream of a dam is one of the main parameters affecting the dam's stability and adverse environmental effects. Scouring can endanger the safety of a dam and related structures. Therefore, correct and reliable scour depth estimation is one of the most important topics for water and hydraulic engineering. This study used three robust white-box data-driven models based on the two popular decision tree methods (M5MT and CART algorithms) and the MARS method to generate explicit equations for scour depth estimation. Field measurements of the scour depth below ski-jump bucket spillways were used to develop the white-box data-driven models. Concerning statistical assessments of the developed models, it was found that the MARS method is the best predictive model for the estimation of scour depth.

The proposed data-driven approaches provided explicit expressions for scour depth prediction. These equations simply and easily compute the depth of scour below ski-jump bucket spillways. The reasonable and effective performance of the MARS model indicated that this method is a high-potential data-driven method for estimating scour depth, which is critical for hydraulic engineering in the design, construction, and stability of dams. The mathematical expressions provided by the proposed methods are clear and understandable for everyone without needing any prior knowledge about the physics of scour depth or data-driven models. These simple rules, generated by the MARS and DTs algorithms, help engineers quickly and accurately approximate scour depth. The findings of this study appear to support the applicability of the suggested approaches for modeling scour depth.

FUNDING

This study did not receive funding from any sources.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

REFERENCES

Agarwal, M., Goyal, M. & Deo, M. C. 2010 Locally weighted projection regression for predicting hydraulic parameters. *Civil Engineering and Environmental Systems* **27** (1), 71–80.

Ahmadianfar, I., Jamei, M., Karbasi, M., Sharafati, A. & Gharabaghi, B. 2022 A novel boosting ensemble committee-based model for local scour depth around non-uniformly spaced pile groups. *Engineering with Computers* **38** (4), 3439–3461.

Azamathulla, H. M. 2013 Comments on 'Evaluation of selected equations for predicting scour at downstream of ski-jump spillway using laboratory and field data' by Chandan Kumar, P. Sreeja [Engineering Geology 129–130 (2012) 98–103]. *Engineering Geology* **1** (152), 210–211.

Azmathullah, H. M., Deo, M. C. & Deolalikar, P. B. 2005 Neural networks for estimation of scour downstream of a ski-jump bucket. *Journal of Hydraulic Engineering* **131** (10), 898–908.

Azmathullah, H. M. D., Deo, M. C. & Deolalikar, P. B. 2006 Estimation of scour below spillways using neural networks. *Journal of Hydraulic Research* **44** (1), 61–69.

Azamathulla, H. M., Deo, M. C. & Deolalikar, P. B. 2008a Alternative neural networks to estimate the scour below spillways. *Advances in Engineering Software* **39** (8), 689–698.

Azamathulla, H. M., Ghani, A. A., Zakaria, N. A., Lai, S. H., Chang, C. K., Leow, C. S. & Abuhasan, Z. 2008b Genetic programming to predict ski-jump bucket spill-way scour. *Journal of Hydrodynamics, Series B* **20** (4), 477–484.

Bhattacharya, B., Price, R. K. & Solomatine, D. P. 2007 Machine learning approach to modeling sediment transport. *Journal of Hydraulic Engineering* **133** (4), 440–450.

Bonakdar, L. & Etemad-Shahidi, A. 2011 Predicting wave run-up on rubble-mound structures using M5 model tree. *Ocean Engineering* **38** (1), 111–118.

Breiman, L. 1984 *Classification and Regression Trees (First edition)*. Routledge, New York. DOI: https://doi.org/10.1201/9781315139470.

Chou, J. S. & Nguyen, N. M. 2022 Scour depth prediction at bridge piers using metaheuristics-optimized stacking system. *Automation in Construction* **140**, 104297.

Daneshfaraz, R., Abam, M., Heidarpour, M., Abbasi, S., Seifollahi, M. & Abraham, J. 2022 The impact of cables on local scouring of bridge piers using experimental study and ANN, ANFIS algorithms. *Water Supply* **22** (1), 1075–1093.

Devi, G. & Kumar, M. 2022 Estimation of local scour depth around twin piers using gene expression programming (local scour around twin piers). *Water Supply* **22** (6), 5915–5932.

Enayati, M., Bozorg-Haddad, O., Pourgholam-Amiji, M., Zolghadr-Asli, B. & Tahmasebi Nasab, M. 2022 Decision tree (DT): a valuable tool for water resources engineering. In: *Computational Intelligence for Water and Environmental Sciences* (*Studies in Computational Intelligence, vol 1043*). (O. Bozorg-Haddad & B. Zolghadr-Asli, eds). Springer, Singapore, pp. 201–223. DOI: https://doi.org/10.1007/978-981-19-2519-1_10.

Friedman, J. H. 1991 Multivariate adaptive regression splines. *The Annals of Statistics* **19** (1), 1–67.

Ghasemi, M., Hasani Zonoozi, M., Rezania, N. & Saadatpour, M. 2022 Predicting coagulation–flocculation process for turbidity removal from water using graphene oxide: a comparative study on ANN, SVR, ANFIS, and RSM models. *Environmental Science and Pollution Research* **29**, 72839–72852. DOI: https://doi.org/10.1007/s11356-022-20989-2.

Goyal, M. K. & Ojha, C. S. P. 2011 Estimation of scour downstream of a ski-jump bucket using support vector and M5 model tree. *Water Resources Management* **25** (9), 2177–2195.

Guven, A. & Azamathulla, H. M. 2012 Gene-expression programming for flip-bucket spillway scour. *Water Science and Technology* **65** (11), 1982–1987.

Haghiabi, A. H. 2017 Estimation of scour downstream of a ski-jump bucket using the multivariate adaptive regression splines. *Scientia Iranica* **24** (4), 1789–1801.

Homaei, F. & Najafzadeh, M. 2020 A reliability-based probabilistic evaluation of the wave-induced scour depth around marine structure piles. *Ocean Engineering* **196**, 106818.

Homaei, F. & Najafzadeh, M. 2022 Failure analysis of scouring at pile groups exposed to steady-state flow: on the assessment of reliability-based probabilistic methodology. *Ocean Engineering* **266**, 112707.

Kamranzad, B., Jabbari, E. & Samadi, M. 2013 Assessment of soft computing models to estimate wave heights in Anzali port. *Journal Of Marine Engineering* **9** (17), 27–36.

Kartal, V. & Emiroglu, M. E. 2022 Experimental study of scour morphology from plunging water jets. *Water Supply* **22** (5), 5410–5433.

Khatsuria, R. M. 2013 Comments on 'Evaluation of selected equations for predicting scour at downstream of ski-jump spillway using laboratory and field data' by Chandan Kumar, P. Sreeja [Engineering Geology 129–130 (2012) 98–103]. *Engineering Geology* **155**, 94–95.

Khosravi, K., Golkarian, A., Omidvar, E., Hatamiafkoueieh, J. & Shirali, M. 2022 Snow water equivalent prediction in a mountainous area using hybrid bagging machine learning approaches. *Acta Geophysica*. DOI: https://doi.org/10.1007/s11600-022-00934-0.

Kumar, C. & Sreeja, P. 2012 Evaluation of selected equations for predicting scour at downstream of ski-jump spillway using laboratory and field data. *Engineering Geology* **129**, 98–103.

Malik, A., Singh, S. K. & Kumar, M. 2021 Experimental analysis of scour under circular pier. *Water Supply* **21** (1), 422–430.

Mason, P. J. & Arumugam, K. 1985 Free jet scour below dams and flip buckets. *Journal of Hydraulic Engineering* **111** (2), 220–235.

Mojaradi, B., Alizadeh, S. F. & Samadi, M. 2018 Estimation of water quality index in talar river using gene expression programming and artificial neural networks. *Iranian Journal of Watershed Management Science and Engineering* **12** (41), 61–72.

Najafzadeh, M. & Oliveto, G. 2022 Scour propagation rates around offshore pipelines exposed to currents by applying data-driven models. *Water* **14** (3), 493.

Najafzadeh, M., Barani, G. A. & Hessami-Kermani, M. R. 2014 Group method of data handling to predict scour at downstream of a ski-jump bucket spillway. *Earth Science Informatics* **7** (4), 231–248.

Nimbalkar, P., Rathod, P., Manekar, V. & Bhalerao, A. 2022 Scour model for circular compound bridge pier. *Water Supply* **22** (5), 5111–5125.

Noori, R., Sheikhian, H., Hooshyaripor, F., Naghikhani, A., Adamowski, J. F. & Ghiasi, B. 2017 Granular computing for prediction of scour below spillways. *Water Resources Management* **31** (1), 313–326.

Nou, M., Zolghadr, M., Bajestan, M. S. & Azamathulla, H. M. 2021 Application of ANFIS–PSO hybrid algorithm for predicting the dimensions of the downstream scour hole of ski-jump spillways. *Iranian Journal of Science and Technology, Transactions of Civil Engineering* **45** (3), 1845–1859.

Pandey, M., Zakwan, M., Khan, M. A. & Bhave, S. 2020 Development of scour around a circular pier and its modelling using genetic algorithm. *Water Supply* **20** (8), 3358–3367.

Parsaie, A., Haghiabi, A. H., Saneie, M. & Torabi, H. 2018 Applications of soft computing techniques for prediction of energy dissipation on stepped spillways. *Neural Computing and Applications* **29** (12), 1393–1409.

Quinlan, J. R. 1992 Learning with continuous classes. In: *5th Australian Joint Conference on Artificial Intelligence*. Vol. 92, pp. 343–348.

Rathod, P. & Manekar, V. L. 2022 Comprehensive approach for scour modelling using artificial intelligence. *Marine Georesources & Geotechnology*. DOI: 10.1080/1064119X.2022.2035025.

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.

Samadi, M. & Jabbar, E. 2012 Assessment of regression trees and multivariate adaptive regression splines for prediction of scour depth below the ski-jump bucket spillway. *Journal of Hydraulics* **7** (3), 73–79.

Samadi, M., Jabbari, E. & Azamathulla, H. M. 2014 Assessment of M5′ model tree and classification and regression trees for prediction of scour depth below free overfall spillways. *Neural Computing and Applications* **24** (2), 357–366.

Samadi, M., Jabbari, E., Azamathulla, H. M. & Mojallal, M. 2015 Estimation of scour depth below free overfall spillways using multivariate adaptive regression splines and artificial neural networks. *Engineering Applications of Computational Fluid Mechanics* **9** (1), 291–300.

Samadi, M., Afshar, M. H., Jabbari, E. & Sarkardeh, H. 2020 Application of multivariate adaptive regression splines and classification and regression trees to estimate wave-induced scour depth around pile groups. *Iranian Journal of Science and Technology, Transactions of Civil Engineering* **44** (1), 447–459.

Samadi, M., Afshar, M. H., Jabbari, E. & Sarkardeh, H. 2021 Prediction of current-induced scour depth around pile groups using MARS, CART, and ANN approaches. *Marine Georesources & Geotechnology* **39** (5), 577–588.

Sammen, S. S., Ghorbani, M. A., Malik, A., Tikhamarine, Y., AmirRahmani, M., Al-Ansari, N. & Chau, K. W. 2020 Enhanced artificial neural network with Harris hawks optimization for predicting scour depth downstream of ski-jump spillway. *Applied Sciences* **10** (15), 5160.

Sihag, P., Dursun, O. F., Sammen, S. S., Malik, A. & Chauhan, A. 2021 Prediction of aeration efficiency of parshall and modified venturi flumes: application of soft computing versus regression models. *Water Supply* **21** (8), 4068–4085.

Sihag, P., Singh, B., Said, M. A. B. M. & Azamathulla, H. M. 2022 Prediction of Manning's coefficient of roughness for high-gradient streams using M5P. *Water Supply* **22** (3), 2707–2720.

Singh, B., Ebtehaj, I., Sihag, P. & Bonakdari, H. 2022 An expert system for predicting the infiltration characteristics. *Water Supply* **22** (3), 2847–2862.

Sun, X., Bi, Y., Karami, H., Naini, S., Band, S. S. & Mosavi, A. 2021 Hybrid model of support vector regression and fruitfly optimization algorithm for predicting ski-jump spillway scour geometry. *Engineering Applications of Computational Fluid Mechanics* **15** (1), 272–291.

Torabi, M., Sarkardeh, H. & Mirhosseini, S. M. 2022 Estimating the permeability coefficient of soil using CART and GMDH approaches. *Water Supply* **22** (8), 6756–6764.

Wang, Y. & Witten, I. H. 1997 Induction of model trees for predicting continuous lasses. In *Proceedings of the Poster Papers of the European Conference on Machine Learning*. University of Economics, Faculty of Informatics and Statistics, Prague

Yonesi, H. A., Parsaie, A., Arshia, A. & Shamsi, Z. 2022 Discharge modeling in compound channels with non-prismatic floodplains using GMDH and MARS models. *Water Supply* **22** (4), 4400–4421.

First received 22 September 2022; accepted in revised form 29 November 2022. Available online 8 December 2022