J. Hydrol. Hydromech., 72, 2024, 1, 113–130
https://doi.org/10.2478/johh-2023-0043

sciendo Journal of Hydrology and Hydromechanics Institute of Hydrology SAS

Uncertainty analysis of discharge coefficient predicted for rectangular side weir using machine learning methods

Seyed Morteza Seyedian[1*], Ozgur Kisi[2,3]

[1] Department of Range and Watershed Management, Gonbad Kavous University, Gonbad Kavous, Iran.
[2] Department of Civil Engineering, Technical University of Lübeck, 23562, Lübeck, Germany. E-mail: ozgur.kisi@th-luebeck.de
[3] Department of Civil Engineering, Ilia State University, 0162, Tbilisi, Georgia.
[*] Corresponding author. E-mails: s.m.seyedian@gmail.com; seyedian@gonbad.ac.ir

Abstract: The present study used three machine learning models, including Least Square Support Vector Regression (LSSVR) and two non-parametric models, namely, Quantile Regression Forest (QRF) and Gaussian Process Regression (GPR), to quantify uncertainty and precisely predict the side weir discharge coefficient (Cd) in rectangular channels. So, 15 input structures were examined to develop the models. The results revealed that the machine learning models used in the study offered better accuracy compared to the classical equations. While the LSSVR and QRF models provided a good prediction performance, the GPR slightly outperformed them. The best input structure that was developed included all four dimensionless parameters. Sensitivity analysis was conducted to identify the effective parameters. To evaluate the uncertainty in the predictions, the LSSVR, QRF, and GPR were used to generate prediction intervals (PI), which quantify the uncertainty coupled with point prediction. Among the implemented models, the GPR and LSSVR models provided more reliable results based on PI width and the percentage of observed data covered by PI. According to point prediction and uncertainty analysis, it was concluded that the GPR model had a lower uncertainty and could be successfully used to predict Cd.

Keywords: Machine learning; Prediction intervals; Sensitivity analysis; Side weir discharge coefficient; Uncertainty analysis.

INTRODUCTION

Side weirs are the most important and common devices used in water resource management, flow distribution and control, sewerage, flood control, and urban runoff applications to divert water from the main channel to the lateral channel (Abbasi et al., 2021; Haddadi and Rahimpour, 2012; Kilic and Emin Emiroglu, 2022). Determining the lateral flow in the side weir is crucial for water management, water resource projects, water use (industry, agriculture), and water level control (Pospíšilík and Zachoval, 2023; Říha and Zachoval, 2015; Salmasi et al., 2021). Figure 1 illustrates the schematic of a side weir.

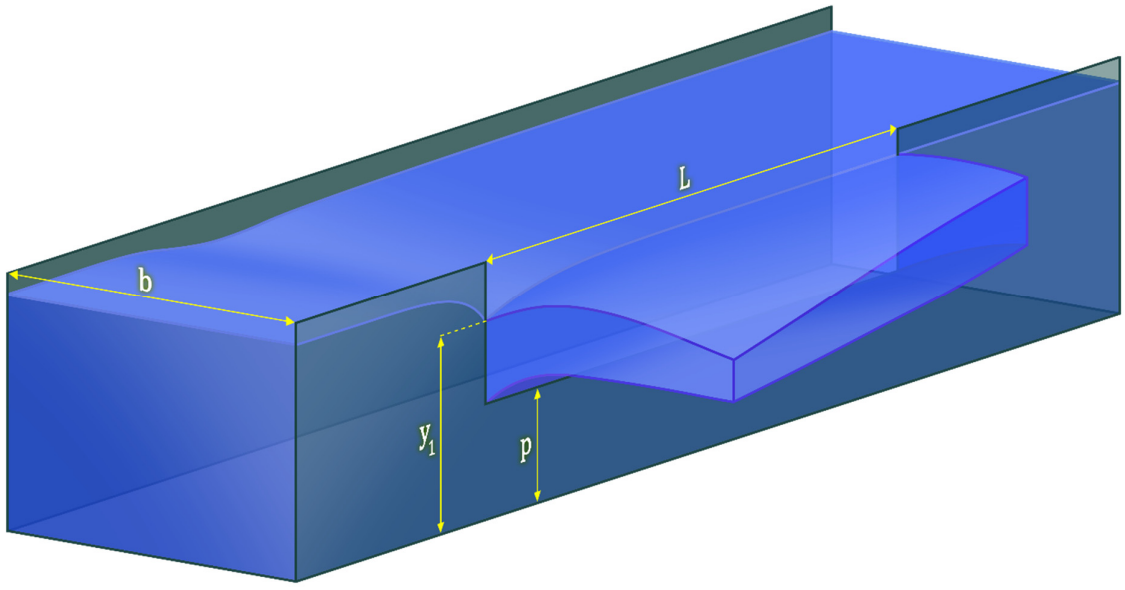


Fig. 1. Schematic representation of flow over a rectangular sharp-crested side weir.

A review of the literature shows that the flow and hydraulics characteristics of rectangular side weirs have widely been studied numerically, theoretically, and experimentally by many researchers (Granata et al., 2013; Hager, 1987; Maranzoni et al., 2017).

Owing to the complexity of hydraulic characteristics of side weirs and human and laboratory equipment errors, the existing equations (regression-based) cannot accurately estimate the discharge coefficient (Cd), so machine learning has been used to predict it (Parsaie and Haghiabi, 2015; Seyedian et al., 2014). Since experimental methods are subject to uncertainty and a limited number of tests yield to the empirical relationships, there is still a need for methods that consider different hydraulic and geometric parameters for the target parameter prediction (Ebtehaj et al., 2018).

Over the past decade, numerous machine learning and soft computing methods have been used in civil engineering, water resources science, hydrology, and hydraulics to model complex phenomena (Seyedian and Rouhani, 2015; Yadav et al., 2022). Out of the many available machine learning techniques, we chose Least Square Support Vector Regression (LSSVR), Quantile Regression Forest (QRF), and Gaussian Process Regression (GPR) in this study due to their computational efficiency, ease of training, and ability to provide Prediction Intervals (PI). Despite extensive research on these models and their applications, there is a lack of information concerning prediction uncertainty in the literature.

The LSSVR was used as a powerful tool to solve regression analysis problems in various fields (Suykens et al., 2002). This technique has been used in many studies for function approximation between variables and predictors (Kisi and Ozkan, 2017; Prayogo and Susanto, 2018; Yi et al., 2018). Liao et al. (2019) showed that the LSSVR was an appropriate tool for reducing the computational burden.

Olyaie et al. (2019) simulation indicated that the LSSVR could predict the Cd of the piano key (PK) weir accurately enough. The LSSVR was used to model the Cd of curved labyrinth overflows (Hu et al., 2021). They found that it could be a suitable tool for predicting Cd. Zounemat-Kermani et al. (2019) examined the precision of the LSSVR in estimating discharge passing triangular arced labyrinth weirs. The outcomes indicated a good agreement between the model estimates and observed discharge. Roushangar and Akhgar (2020) applied the LSSVR to model the discharge coefficient of stepped spillways. According to the results, it performed well in modelling discharge coefficients at stepped spillways.

Another machine learning technique that has many applications in various fields of water engineering sciences is Gaussian Process Regression (GPR) which is an efficient technique. The GPR is a non-parametric powerful probabilistic modelling instrument that enables observations in continuous spaces or time (Bonakdari et al., 2019; Williams and Rasmussen, 2006). Akbari et al. (2019) studied the proficiency of Machine Learning (ML) and nonlinear and multilinear regression models for the Cd of PK and showed that the GPR model surpasses other ML models in predicting the Cd of PK weir. Karbasi et al. (2021) showed that the GPR provided higher accuracy in predicting the side orifice discharge coefficient. The modelling results of a Cd radial gate indicated that the GPR attained acceptable predictable performance (Tao et al., 2022). Nourani et al. (2021) estimated the Cd over broad-crested weirs using the GPR.

Random forests were also presented as a machine learning technique by Breiman (2001) and proven to be a very powerful and popular tool for regression analysis. A generalization of random forests is Quantile Regression Forests (QRF) which can derive conditional quantiles (Meinshausen and Ridgeway, 2006). The QRF has been used in different areas (Ahmed and Lin, 2021; Bhuiyan et al., 2018; Francke et al., 2008). In a QRF model, prediction intervals (PI) are created based on the spread of the dependent variable.

There is no doubt that the accuracy of any measured variable that contributes to determining dependent laboratory data has a straight effect due to its own bias (Borghei et al., 2013). There are some measured values that can cause certain uncertainty in the experimental value (Coleman and Steele, 2009; Johnson and Ayyub, 1996). Borghei et al. (2013), Johnson and Ayyub (1996), and Říha and Zachoval (2014) examined the uncertainty analysis of the Cd. The Cd uncertainty could be estimated by Monte Carlo (MC), bootstrap method (BM), and analytical method (Mohammed and Golijanek-Jędrzejczyk, 2020; Parsaie and Haghiabi, 2021). Gholami et al. (2018) performed uncertainty analysis to quantitatively evaluate the Cd models. They created PI using standard deviation.

Because of the uncertainty inherent in experimental methods, uncertainty analysis is essential (Ebtehaj et al., 2018) and consequently, it remains crucial to conduct testing and comparisons of various techniques for quantifying prediction uncertainty. In recent years, advancements in machine learning models and optimization methods have resulted in better Cd prediction. However, these models are still unable to accurately predict uncertainties. Although the Cd of side weirs has extensively been studied in theory and in the laboratory, PI is usually ignored by most studies and only limited research has been conducted on uncertainty (Parsaie and Haghiabi, 2021). To the best of the authors' knowledge, the GPR, LSSVR, and QRF have not been reported in quantifying side weir uncertainty, and this is the first effort to use them to determine the uncertainty of innovation in the field of hydraulics structures. In this study, the discharge coefficient was predicted using three methods: GPR, LSSVR, and QRF. A sensitivity analysis was performed to determine the effective parameters. To validate the given schemes, the results were compared with empirical equations. Then, for the first time, the discharge coefficient uncertainty was evaluated using these three methods. This paper is intended to provide more understanding of the Cd uncertainty.

Based on the abovementioned explanation, the main contribution of this study can be expressed as (i) developing GPR, LSSVR, and QRF models for the prediction of Cd and (ii) quantifying the discharge coefficient uncertainty using the three models mentioned.

MATERIAL AND METHODS

Data collection

The study used the datasets presented by Bagheri et al. (2014) and Emiroglu et al. (2011). The experimental setup of Emiroglu et al. (2011) was implemented at Firat University in Elazig, Turkey. The rectangular channel was 12 m long. The channel depth and width were 0.5 m and 0.5 m, respectively. The primary channel flow depth was controlled by a sluice gate. Steel plates were used for the rectangular weir, which was installed on the right wall of the primary channel at the same level. An electromagnetic flowmeter was used to measure discharge in the primary channel, and the discharge over the weir was passed into a secondary channel that was calibrated with a standard rectangular weir. Subcritical flow conditions were carried out in all experiments.

Using sharp-crested rectangular weirs, Bagheri et al. (2014) conducted several experiments. The sluice gate controlled the downstream discharge and the flow depth in the primary channel.

Electromagnetic flowmeters were used to measure the discharge in the upstream primary channels.

The statistical description of the experimental data collected for the present study is shown in Table 1 where Fr, y_1/L, p/y_1, L/b, and Cd denote the Froude number, ratios of upstream flow depth to weir length, weir height to upstream flow depth, dimensionless length of a weir, and coefficient of discharge, respectively.

Figure 2 presents the histogram distribution for each variable and the matrix of correlations between the variables used in Cd prediction. In general, there is a weak linear correlation between the variables and Cd, so it is difficult to make precision predictions.

Input structure

Using the data that was described, the LSSVR, QRF, and GPR were employed to model the Cd of rectangular side weirs. Data were separated into two portions for training and testing purposes. Approximately 75% of the data was used for training and 25% for simulation tests.

Proper input parameter selection is essential for the development and application of machine learning models. Physical processes can be simulated accurately by selecting the variables that control the phenomenon (Bowden et al., 2005). To explore the effect of each dimensionless parameter influencing side weir Cd prediction (sensitivity analysis), four different models (M2-M5) were defined. Also, to determine the best combination of dimensionless parameters to predict the Cd with the best accuracy, 10 other models (M6-M15) were presented. As shown in Figure 3, all 15 models were defined for the LSSVR, QRF, and GPR.

Table 1. Statistical parameters for the present study.

	Fr	p/y_1	L/b	y_1/L	Cd
Minimum	0.09	0.23	0.30	0.09	0.09
Average	0.39	0.68	1.38	0.63	0.54
Maximum	0.83	0.91	3.00	2.88	1.75
Standard deviation	0.18	0.18	0.95	0.61	0.26
Coefficient of variation	0.45	0.26	0.69	0.96	0.47
Skewness	0.48	–0.62	0.66	1.55	1.70

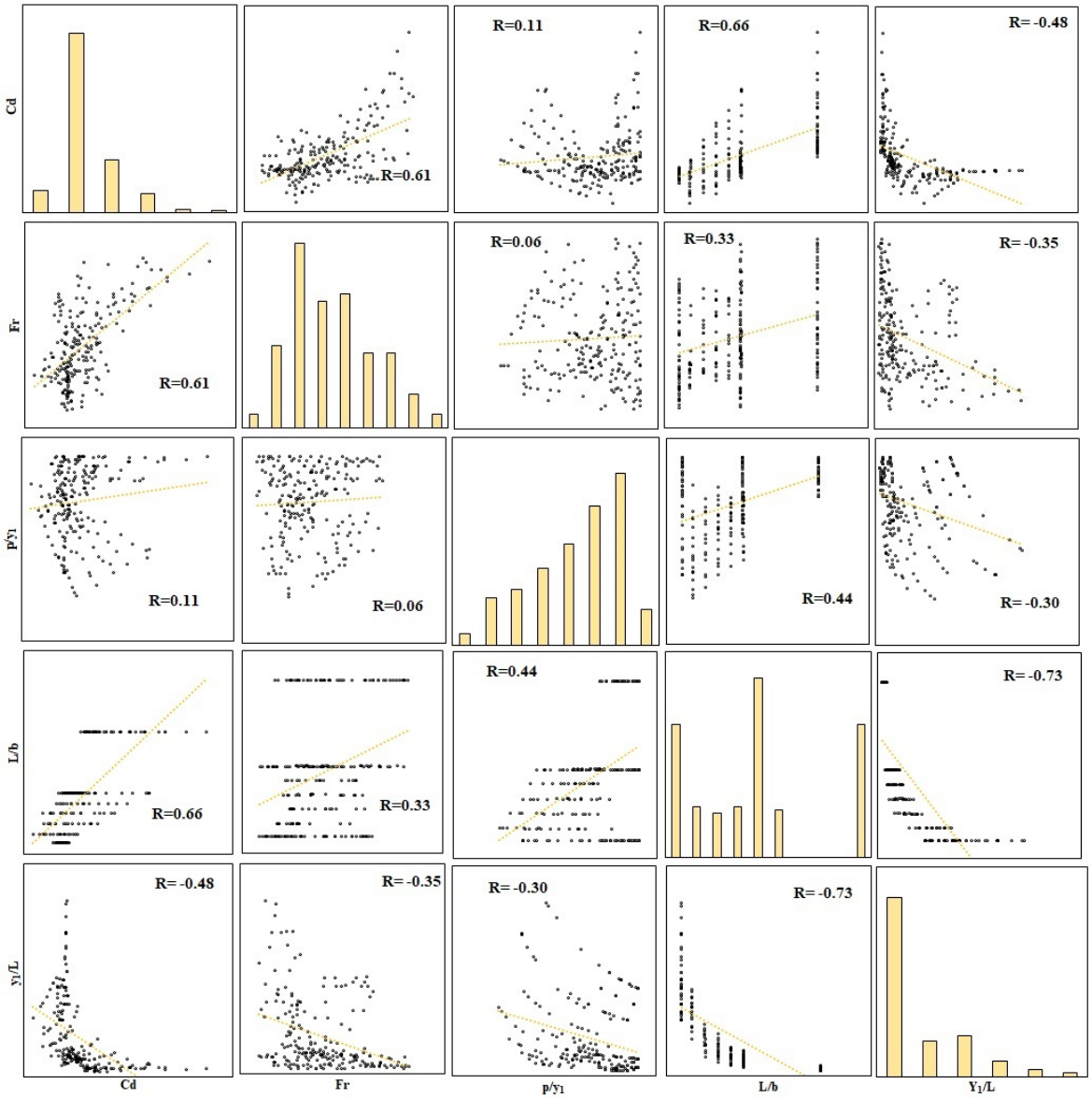


Fig. 2. Histogram and correlation matrix between all variables.

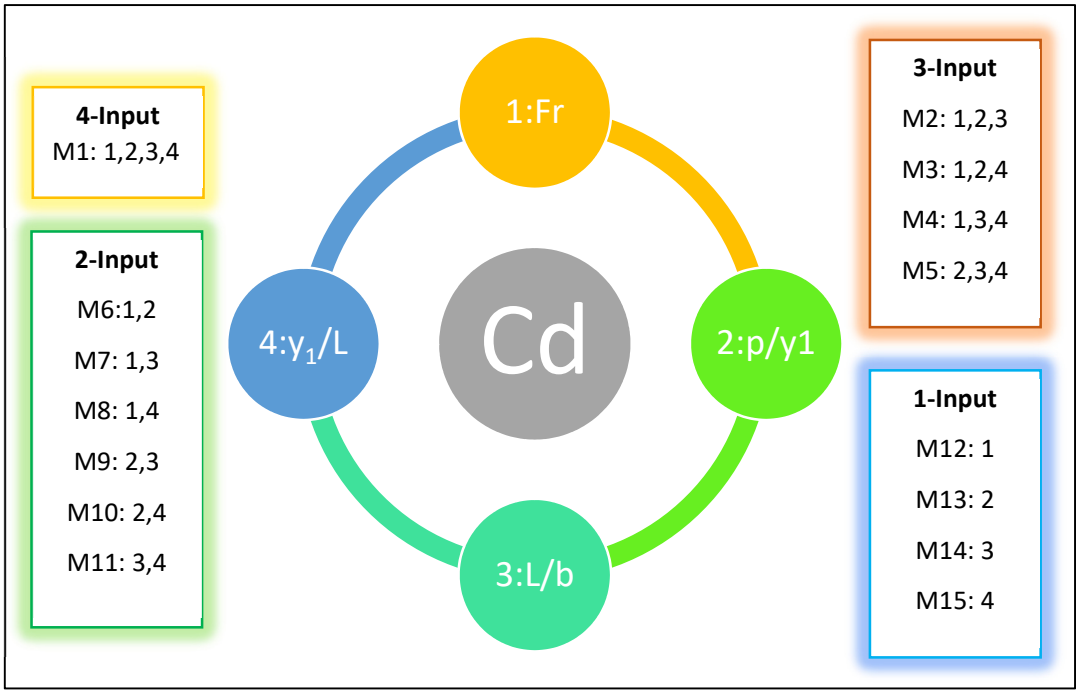


Fig. 3. Input structure for machine learning models.

Dimensional analysis

Dimensional analysis was performed to estimate a function for the side weir discharge coefficient. It is possible to express the discharge coefficient of side weirs as a function of the side weir (V), the upstream depth of the flow (y_1), the width of the primary channel (b), the dynamic viscosity of the water (μ), the acceleration gravity (g), the weir crest length (L), the density of the water (ρ), the channel slope (S_0), the deviation of the angle of flow (ψ), and the side weir height (p).

Dimensionless parameters that impact the Cd were derived through the application of the Buckingham π theorem. The dimensionless parameters can be obtained as follows:

$$C_d = f(p, L, b, g, V, y_1, \rho, \mu, \omega, S_0) \quad (1)$$

According to the parameters affecting the discharge coefficient and the research conducted in this field (Borghei et al., 1999; Ebtehaj et al., 2015; Subramanya and Awasthy, 1972), the discharge coefficient is presented as follows:

$$C_d = f\left(Fr = \frac{V}{\sqrt{g y_1}}, \frac{p}{y_1}, \frac{y_1}{L}, \frac{L}{b}, \omega, S_0\right) \quad (2)$$

New dimensionless parameters affecting the prediction of the discharge coefficient can be obtained as follows:

$$C_d = f\left(Fr_1, \frac{p}{y_1}, \frac{y_1}{L}, \frac{L}{b}\right) \quad (3)$$

Least squares support vector regression

Cortes and Vapnik (1995) proposed the Support Vector Machine (SVM) model based on statistical learning theory and the principle of structure risk minimization. Later, Suykens and Vandewalle (1999) introduced the Least Squares Support Vector Machine (LSSVM) method as an alternative to the SVM. Unlike the SVM, the LSSVM employs linear equations instead of a second-degree programming problem to find solutions. This approach reduces the complexity compared to the standard SVM by utilizing the least-squares optimization method rather than the second-order method. The LSSVM maps inputs from lower dimensions to higher dimensions to transform nonlinear relationships between inputs and outputs into linear ones. This is particularly useful for solving nonlinear problems and miniaturization (Anandhi et al., 2008). As a result, the LSSVM has a higher calculation precision compared to the classic SVM. In the LSSVM model, the linear least-square system is used as the loss function, and the inequality constraints are modified to equality constraints.

Assume that the relationship between the response variables and the independent variable is based on the following function:

$$y(x) = w^T \varphi(x) + b \quad (4)$$

where $x \in R^n$, $y \in R$, and φ represent the high-dimensional space of features as a feature map. To calculate Eq. (4), the following optimization problem is formulated as follows:

$$\text{minimize} \quad J_p(w, \zeta) = \frac{1}{2}(w^T w + \gamma \sum_{l=1}^{k} \zeta_l^2) \quad (5)$$

s.t. $y_l = w^T \varphi(x_l) + b + \zeta_l, l = 1, \ldots, k$

where $\gamma > 0$ and $\zeta_l \in R$ are the regularization constant and the error variables, respectively.

The Lagrangian function for Eq. (5) using Karush-Kuhn-Tucker conditions is given by the following linear system.

$$\begin{bmatrix} 0 & 1_l^T \\ 1_l & \psi + \frac{1}{\gamma} I_k \end{bmatrix} \begin{bmatrix} b \\ r \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \quad (6)$$

where

$$r = [r_1, \dots, r_k]^T,\ Y = [Y_1, \dots, Y_k]^T, 1_\iota = [1, \dots, 1]^T \quad (7)$$

$$\psi_{kl} = \varphi(X_k)^T \varphi(X_l) = K(X_k, X_l) \quad (8)$$

The resulting LSSVR regression equation is obtained as:

$$\hat{m}(x) = \sum_{l=1}^{k} \hat{r}_l K(X_k, X_l) + \hat{b} \quad (9)$$

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and d is the number of dimensions. Many methods are used to estimate $\hat{n}\ of\ n$, one of which is linear smoother. In this method, the LSSVR estimates Eq. (9) as follows:

$$\hat{n}(x) = \sum_{l=1}^{n} l_l(x) Y_l \quad (10)$$

where $x \in \mathbb{R}^d$ and $L(x) = (l_1(x), \dots, l_k(x))^T \in \mathbb{R}^n$. Also, one can use Eq. (11) to estimate confidence intervals (Brabanter et al., 2011).

$$\frac{\hat{n}(x) - P[\hat{n}(x)|X=x]}{\sqrt{V[\hat{n}(x)|X=x]}} \xrightarrow{D} \mathcal{N}(0,1) \quad (11)$$

Distribution convergence is indicated by $\xrightarrow{D}$. There is conditional variance in the model $\mathrm{V}[\hat{n}(x)|X=x] = \sum_{l=1}^{k} l_l(x)^2 \sigma^2(x_l)$ and conditional mean $P[\hat{n}(x)|X=x] = \sum_{l=1}^{k} l_l(x) n(x_l)$. Also, $\sigma^2(x) = \mathrm{V}[Y|X=x] = P[Y^2|X=x] - \{P[Y|X=x]\}^2$. Confidence intervals can be approximated as follows:

$$\hat{n}(x) \pm z_1 - {}^{r}/_{2} \sqrt{V[\hat{n}(x)|X=x]} \quad (12)$$

where $z_1 - {}^{r}/_{2}$ represents $\left(1 - {}^{r}/_{2}\right)$th standard normal distribution of quantile.

Quantile regression forests

In this research, a non-parametric tree-based regression method called Quantile Regression Forests (QRF) was used to predict the discharge coefficient using dimensionless parameters. The QRF is highly capable of data handling and can save the trained model for future predictions (Nateghi et al., 2014). The QRF is derived from random forest regression. The QRF uses a bagged version of decision trees, which reduces the variance and avoids overfitting, which improves the accuracy and stability of the model. The non-parametric approach is employed by QRF to assess conditional quantiles of variables' high-dimensional predictors.

Suppose Y is a response variable (observed value) and X is a predictor variable or covariate, possibly high-dimensional. The primary objective of conducting regression analysis is to determine the correlation between two variables, X and Y. A tree ensemble is generated by random forest using n independent observations $(Y_i, X_i), i = 1, \dots, n$. A bagged version of the training data is used for each tree. In the case of new data X = x, the single tree prediction $T(\theta)$ equals the observed value in the leaf $l(x, \theta)$. When the observation (X_i) belongs to the leaf, the weight vector $w_i(x, \theta)$ will have a positive value; otherwise, it will be equal to zero. The total weight is equal to 1, so

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_j \in R_{l(x,\theta)}\}} \quad (13)$$

in which $R_{l(x,\theta)}$ is a rectangular subspace.

Single tree prediction for X = x based on the weighted average of observations $Y_i, i = 1, \dots, n$ is done as follows:

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta)\, Y_i \quad (14)$$

Conditional mean $E(Y|X=x)$ is estimated by averaging k single trees. If $w_i(x)$ is the average of $w_i(\theta)$ in this set of trees:

$$w_i(x) = k^{-1} \sum_{t=1}^{k} w_i(x, \theta_t) \quad (15)$$

Moreover, the final forecast will be in the form of the following relationship.

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i \quad (16)$$

For $X = x$, the definition of the conditional distribution function of Y can be expressed as follows.

$$R(y|X=x) = P(Y \le y|X=x) = E\left(1_{\{Y \le y\}} \middle| X = x\right) \quad (17)$$

where Y is observations of the response variable, X refers to the predictor variable or covariate, and $E\left(1_{\{Y \le y\}} \middle| X = x\right)$ is the conditional mean which is estimated by the weighted mean over the observation of $1_{\{Y \le y\}}$ (Meinshausen and Ridgeway, 2006).

According to QRF, Y has the following conditional distribution function for a given $X = x$.

$$\hat{R}(y|X=x) = \sum_{i=1}^{n} w_i(x)\, 1_{\{Y_i \le y\}} \quad (18)$$

For any α value, the prediction intervals can be constructed using quantile regression. In this study, a 95% prediction interval is used, which is defined as:

$$I(x)[Q_{0.025}(x), Q_{0.975}(x)] \quad (19)$$

Gaussian process regression

The Gaussian Process Regression (GPR) is a non-parametric kernel-based model with high computational efficiency and precision (Liu et al., 2016). This method is highly capable of modelling complex nonlinear issues.

The Gaussian Process Regression (GPR) utilizes several random variables, some of which exhibit Gaussian distributions. This extends the concept of the Gaussian distribution (Zhao et al., 2011). The GPR is a way of establishing prior distributions in function space, which is a natural generalization of the Gaussian distribution whose mean and covariance are represented by a vector and matrix, respectively. Although the Gaussian distribution is defined over vectors, it is applied to functions. The GPR is a suitable method that allows for the development of flexible classification and regression models, without being limited to simple parametric forms for regression or class probability functions. One significant advantage of using GPR is the availability of a diverse range of covariance functions. These functions have varying degrees of smoothness and continuous structures, allowing modellers to select the most suitable function for their specific application. Gaussian processes are important in statistical modelling because of their normal properties. They allow for the specification of distributions across functions that have one or more input variables. For instance, in a regression model with Gaussian errors, the use of matrix calculations can help deduce outcomes that are suitable for datasets with sample sizes larger than one thousand.

By considering x and y as the input and output domains, respectively, from which n pairs (x_i, y_i) are drawn identically and independently distributed. The main assumption of GPR is that y is given by:

$$y = f(x) + \varepsilon \quad (20)$$

The observation error, ε, can be characterized by a zero-mean value distribution ($\mu(x) = 0$), and variance σ^2 and $f(x)$ are the GPR function values (Williams and Rasmussen, 2006).

The joint distribution is determined by the kernel function as follows:

$$\begin{bmatrix} y \\ y_t \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x,x) + \sigma_N^2 I_N & k(x_t,x)^T \\ k(x_t,x) & k(x_t,x_t) \end{bmatrix}\right) \quad (21)$$

where $k(x_t,x) = [k(x_t,x_1), k(x_t,x_2), \ldots, k(x_t,x_N)]$ (22)

$x = [x_1, x_2, \ldots, x_N]^T$ is the training input matrix, $y = [y_1, y_2, \ldots, y_N]^T$ is the training output vector, x_t is the test input, and y_t is the test output dataset. The predictor distribution over y_t is expressed as Eq. (23).

$$P(y_t|x,y,x_t) \sim N(\bar{y}_t, cov(y_t)) \quad (23)$$

where $\bar{y}_t = k(x_t,x)[k(x,x) + \sigma_N^2 I_N]^{-1} y$ (24)

$$cov(y_t) = k(x_t,x_t) - k(x_t,x)[k(x,x) + \sigma_N^2 I_N]^{-1} k(x_t,x)^T \quad (25)$$

In the present study, squared exponential kernels were used that can be expressed as:

$$k(x_i,x_j) = \sigma_f^2 exp\left(-\frac{1}{2}\frac{|x_i - x_j|^2}{\gamma^2}\right) \quad (26)$$

where σ_f is the signal standard deviation and γ is the length scale for predictors. The values of $\eta = \{l, \sigma_f^2, \sigma_l^2\}$ (hyper-parameters) are calculated by maximizing the log-likelihood function as follows (Momeni et al., 2020; Schulz et al., 2018).

$$L(\eta) = \log P(y|x,\eta) =$$
$$= -\frac{1}{2} y^T (k(x,x)^{-1} y) - \frac{1}{2} log|k(x,x)| - \frac{n}{2}\log 2\pi \quad (27)$$

When the specified conditions are satisfied, the training process will be terminated.

Existing equations

A combination of Fr, p/y_1, L/b, and L/y_1 parameters was considered as input parameters to model Cd according to Emiroglu et al. (2011) (Eq. 28). Eqs. (29–30) show the models proposed by Subramanya and Awasthy (1972), and Cheong (1991) used only the Froude number (Fr) parameter to predict Cd.

$$c_d = \left[0.836 + \left(-0.035 + 0.39\left(\frac{p}{y_1}\right)^{12.69} + 0.158\left(\frac{L}{b}\right)^{0.59} + 0.049\left(\frac{L}{y_1}\right)^{0.42} + 0.244 Fr_1^{2.125}\right)^{3.018}\right]^{5.36} \quad (28)$$

$$c_d = 0.611\sqrt{1 - \left(\frac{3Fr_1^2}{Fr_1^2+2}\right)} \quad (29)$$

$$c_d = 0.45 - 0.221 Fr_1^2 \quad (30)$$

Model performance evaluation

A comparison of the proposed methods for predicting the Cd of side weirs is presented in this study using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Relative Absolute Error (RAE) (Eqs. (31) to (33)). The RMSE shows the standard deviation of the difference between the predicted samples and the experimental ones. The major advantage of the RMSE and RAE is that they represent the error on the same scale as the output variable. These indices can be determined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(C_{doi} - C_{dpi})^2}{\sum_{i=1}^{n}(C_{doi} - \overline{C_{do}})^2} \quad (31)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(C_{dpi} - C_{doi})^2}{n}} \quad (32)$$

$$RAE = \frac{\sum_{i=1}^{n}|C_{dpi} - C_{doi}|}{\sum_{i=1}^{n}|C_{doi} - \overline{C_{do}}|} \quad (33)$$

The measured values are represented by O, P is the predicted value obtained by the machine learning models used, $\overline{O}$ is the average of observation values, and n is the number of data samples. The closer the values of RMSE and RAE to 0 and R^2 to 1, the higher the performance of the machine learning models.

Containing ratio

Containing Ratio (CR) is the ratio of observation data that lies within the prediction bounds to total observation data (Xiong et al., 2009). It is clear that having a higher CR for the estimated prediction bounds is preferable as it indicates more observation data contained within the prediction bounds. The CR range is between zero and one, and CR = 1 expresses perfect simulation.

Average relative bandwidth

To facilitate prediction bounds, a dimensionless index is represented by Average Relative Bandwidth (RB). RB range is between 0 and ∞. The lower the value, the more ideal it would be.

$$RB = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(Cd_i^u - Cd_i^l\right)}{Cd_i} \quad (34)$$

where Cd_i (i = 1,2, …, n) represents the number of observed Cd and Cd_i^u and Cd_i^l show the upper bounds and lower bounds of the Cd, respectively.

Mean prediction bandwidth

Mean Prediction Bandwidth (MPB) is another index that shows bandwidth.

$$MPB = \frac{1}{n}\sum_{i=1}^{n}(s_i^u - s_i^l) \quad (35)$$

The range of MPB is between zero and ∞. The closer to zero, the better.

Figure 4 depicts the flowchart for prediction and uncertainty analysis using machine learning models.

RESULTS AND DISCUSSION

Effect of the dimensionless parameters on Cd

To predict the side weir discharge coefficient, this section evaluates three machine learning models (LSSVR, QRF, and GPR). As mentioned in the data collection section, 216 data points were used in this research. Some fundamental strategies are used to evaluate these models, such as the sensitivity test, input structure, uncertainty analysis, and comparison with empirical equations. The quantitative statistical outcomes of the LSSVR, QRF, and GPR models are presented in Figure 5 for the training and test phases.

Moreover, Figure 6 depicts the Taylor diagram of the observed and predicted Cd for the GPR, LSSVR, and QRF. The Taylor diagram gives information about the correlation, root-mean-square-difference (RMSD), and standard deviation on a single plot (Taylor, 2001).

The RMSD measures the disagreement between two datasets, whereas the correlation coefficient determines the extent of the linear relationship between them (Taylor, 2001). The predicted values obtained by the M1 scenario are closer to the target point (observed), demonstrating accurate efficiency.

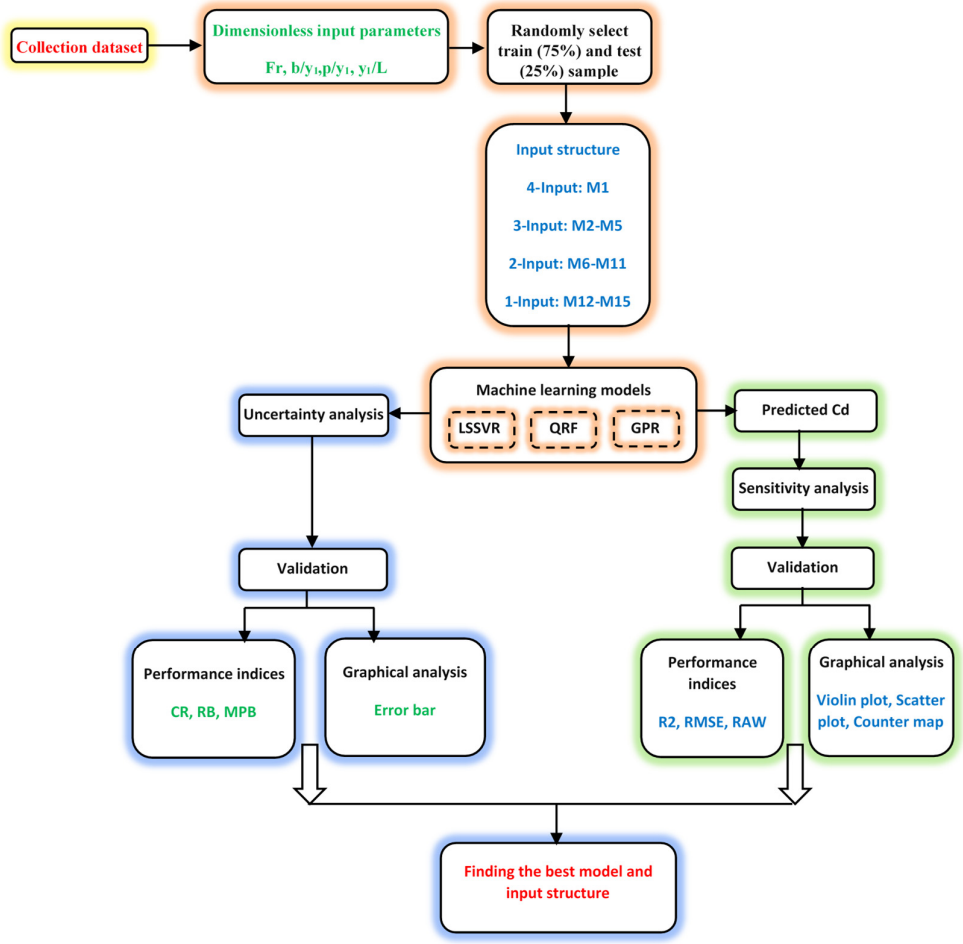
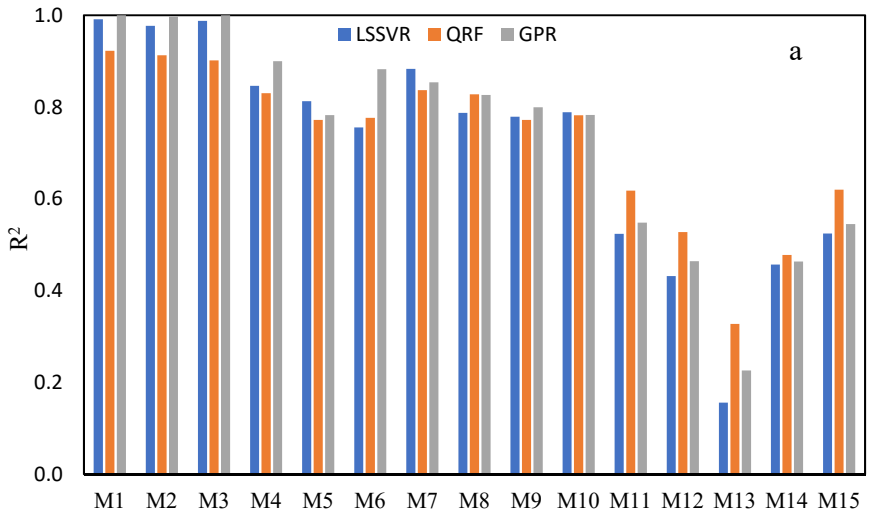
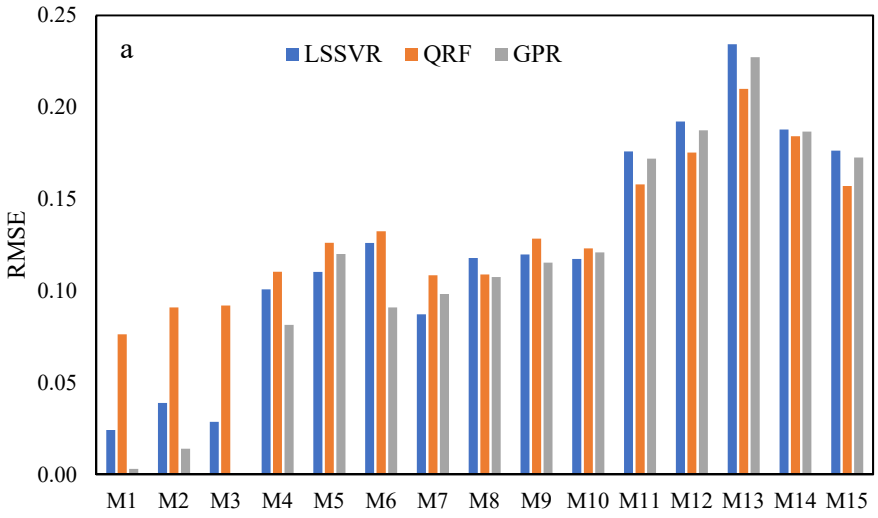
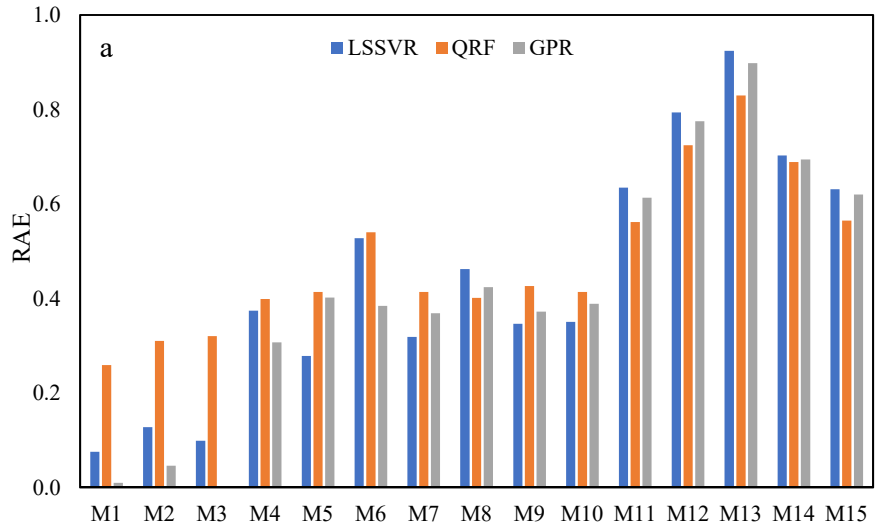
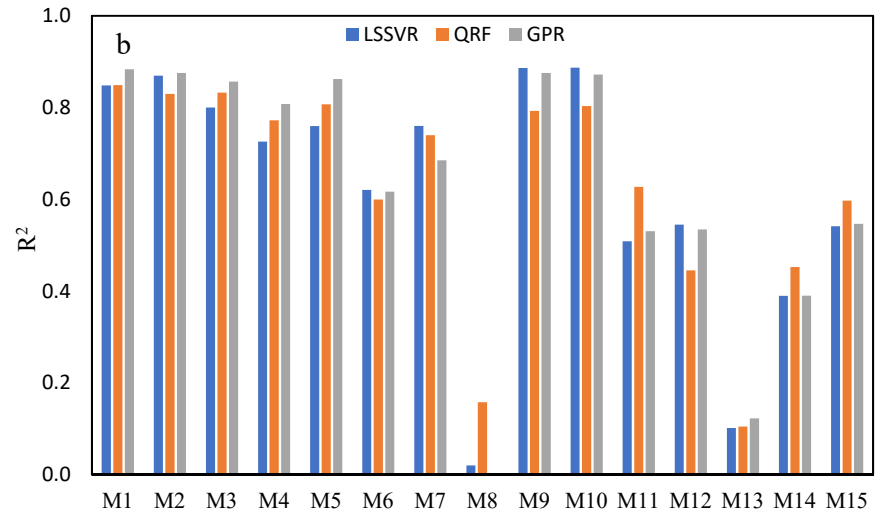


Fig. 4. The flowchart for the prediction and uncertainty analysis of Cd.




a
LSSVR
QRF
GPR
RMSE
0.25
0.20
0.15
0.10
0.05
0.00
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15


a
LSSVR
QRF
GPR
RAE
1.0
0.8
0.6
0.4
0.2
0.0
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15


b
LSSVR
QRF
GPR
R2
1.0
0.8
0.6
0.4
0.2
0.0
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15

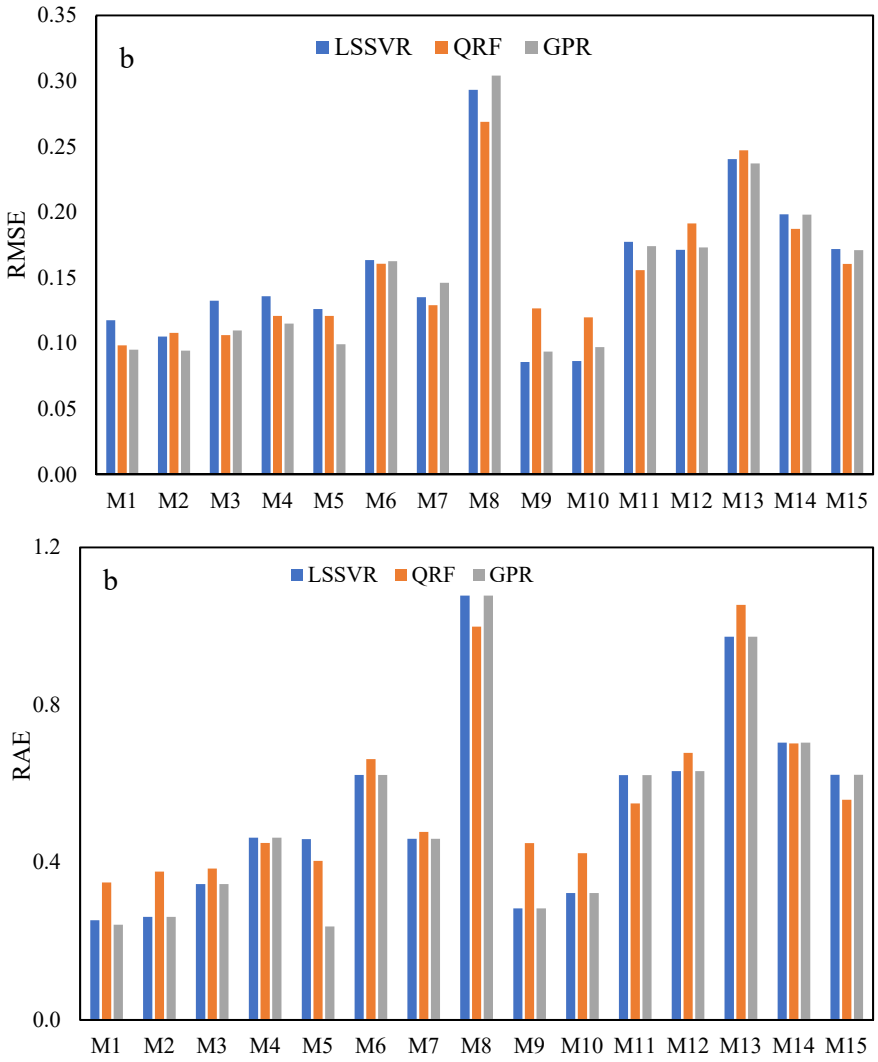


Fig. 5. The comparison of the models' results: a) training and b) testing.

There is a superior capability associated with the M1 input for all models. For example, in the LSSVR, QRF, and GPR, the model with M1 input leads to (R^2 = 0.848, RMSE = 0.118, RAE = 0.253), (R^2 = 0.849, RMSE = 0.098, RAE = 0.349), and (R^2 = 0.883, RMSE = 0.095, RAE = 0.242), respectively. All criteria show that the GPR has superior prediction accuracy.

Based on the comparison of the LSSVR, QRF, and GPR results based on Figure 5, M1 and M2 have no significant differences in prediction accuracy. In particular, compared to the LSSVR model, the GPR seems to perform slightly better in terms of efficiency. The GPR has lower prediction error values and a higher level of agreement in M2 input (R^2 = 0.875, RMSE = 0.094, RAE = 0.262). A desirable prediction was made by the LSSVR and QRF models, although slightly lower than the GPR, the results still yield acceptable statistical metrics of (R^2 = 0.870, RMSE = 0.105, RAE = 0.262) and (R^2 = 0.830, RMSE = 0.108, RAE = 0.376), respectively.

Overall, for the 3-input structure (M2-M5), as compared with the LSSVR model, these models had a better correlation. These results confirm the superior estimation capability of the GPR and QRF models compared to the LSSVR models in point prediction. In the 2-input structure, M9 and M10 achieved a better result in test phase using the LSSVR with average (R^2 = 0.887, RMSE = 0.095, RAE = 0.303), QRF with average (R^2 = 0.798, RMSE = 0.124, RAE = 0.436), and GPR with average (R^2 = 0.873, RMSE = 0.095, RAE = 0.303). The accuracy is relatively acceptable even with two and three input variables.

Among the single inputs (M12-M15), the Fr and y_1/L have the highest accuracy, but the combinations of these two variables (M8) have poorer results in the 2-input structure. The combination of Fr and y_1/L input parameters provides poorer accuracy than the 1-input structure.

In some machine learning models, adding more input parameters increases complexity, which may reduce the model performance (Bonakdari et al., 2015; Cartwright, 2015), but in this research, all models provided accurate results with 3 and 4-input parameters.

To select the most efficient models, visual assessment is very important. In the testing phase, a scatterplot was used to compare the predicted and measured Cd (Figure 7) in M1. The performance of prediction models is mostly evaluated with a scatterplot. A notable feature is that it offers details about the diversion of the data points from their original observation.

Figure 7 illustrates that all models with the M1 scenario are capable of producing a suitable result. A good correlation was found between the model and the observed data, as shown in

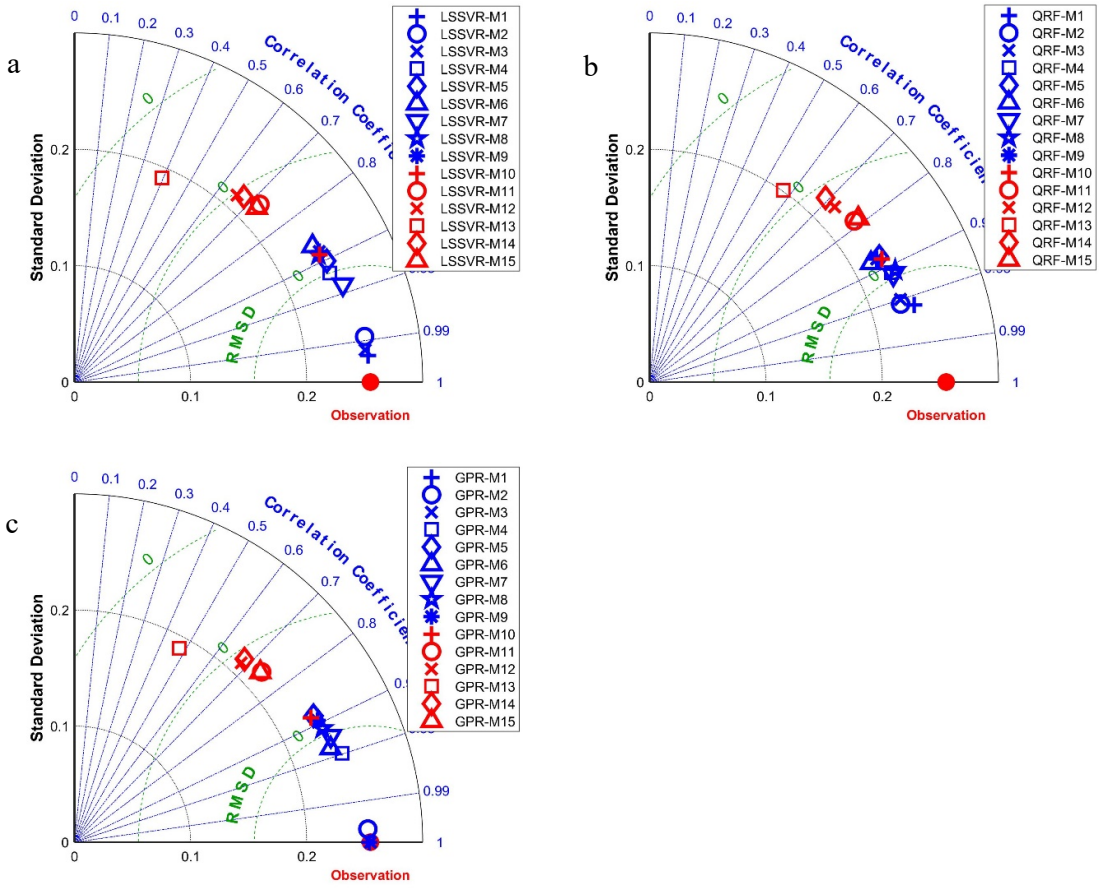


Fig. 6. The Taylor diagram for (a) LSSVR, (b) QRF, and (c) GPR.

Figure 7. It is visible from the scatterplots that almost all data points of the models fall within ±25% relative error bands.

Discrepancy ratio

Figure 8 depicts the Discrepancy Ratio (DR) of the LSSVR, QRF, and GPR for the M1 input structure in the test phase. The DR is obtained by dividing the predicted discharge coefficient with the observed discharge coefficient ($DR = Cd_{Predicted}/Cd_{Observed}$). This diagram is shown to visually evaluate the efficiency of model. Model performance improves as more data are closer to the 1 range. Obviously, DR of 84%, 87%, and 76% of the data predicted by the LSSVR, GPR, and QRF, respectively are in the range of 1 ± 0.1. For the LSSVR, QRF, and GPR, the mean of DR is 1.02, 1.03, and 1.00, respectively. In all models, only 4% of the data (two data points) is less than 0.9, while for the LSSVR, QRF, and GPR models, 14%, 20%, and 9% of the data are greater than 1.1, respectively. Therefore, it is clear that the QRF tends to slightly overestimate Cd. Consequently, the GPR exhibits no significant bias and performs well in estimating the coefficient of discharge.

Violin plot

Graphical analysis techniques are key to selecting the best model for the prediction of Cd. For this purpose, violin plots are used as an advanced graphical tool to explain the similarity between predicted and observed Cd values. Statistically, a violin plot shows the distribution shape of data by integrating a kernel density plot with a box plot. Similar information can be obtained by five number summary (maximum (Q_4), third quartile (Q_3), median (Q_2), first quartile (Q_1), and minimum (Q_0)). In this section, the violin plot was employed to evaluate the model performance in predicting Cd (Figure 9).

The medians of the observed data, LSSVR, QRF, and GPR are 0.46, 0.46, 0.48, and 0.47, respectively. In addition, the interquartile range (IQR) is an important element that can be found in the violin plot. IQR = Q3 – Q1 is the distance between the upper and lower quartiles. The IQR of the observed data, LSSVR, QRF, and GPR is 0.20, 0.19, 0.28, and 0.22, respectively.

Figure 9 clearly shows that the Cd predicted using GPR resembles the observed Cd. The results of the GPR model are in better agreement with the observed values than those of the LSSVR and QRF models. In addition, there is an acceptable similarity between the distributions of the observed and predicted Cd by GPR. It is noticeable that the GPR can accurately predict the higher values of the Cd. In general, all three models showed suitable prediction performance. The best agreement was observed between the GPR and the observed data, so it was recognized as the superior model.

Sensitivity analysis

Several hydraulic and physical variables have considerable effects on Cd. A sensitivity analysis is performed to determine how dimensionless parameters affect model performance. To conduct an accurate simulation, identifying the parameters that have an impact on the Cd is extremely important.it is crucial to identify the parameters that influence the value of Cd (Tao et al., 2022).

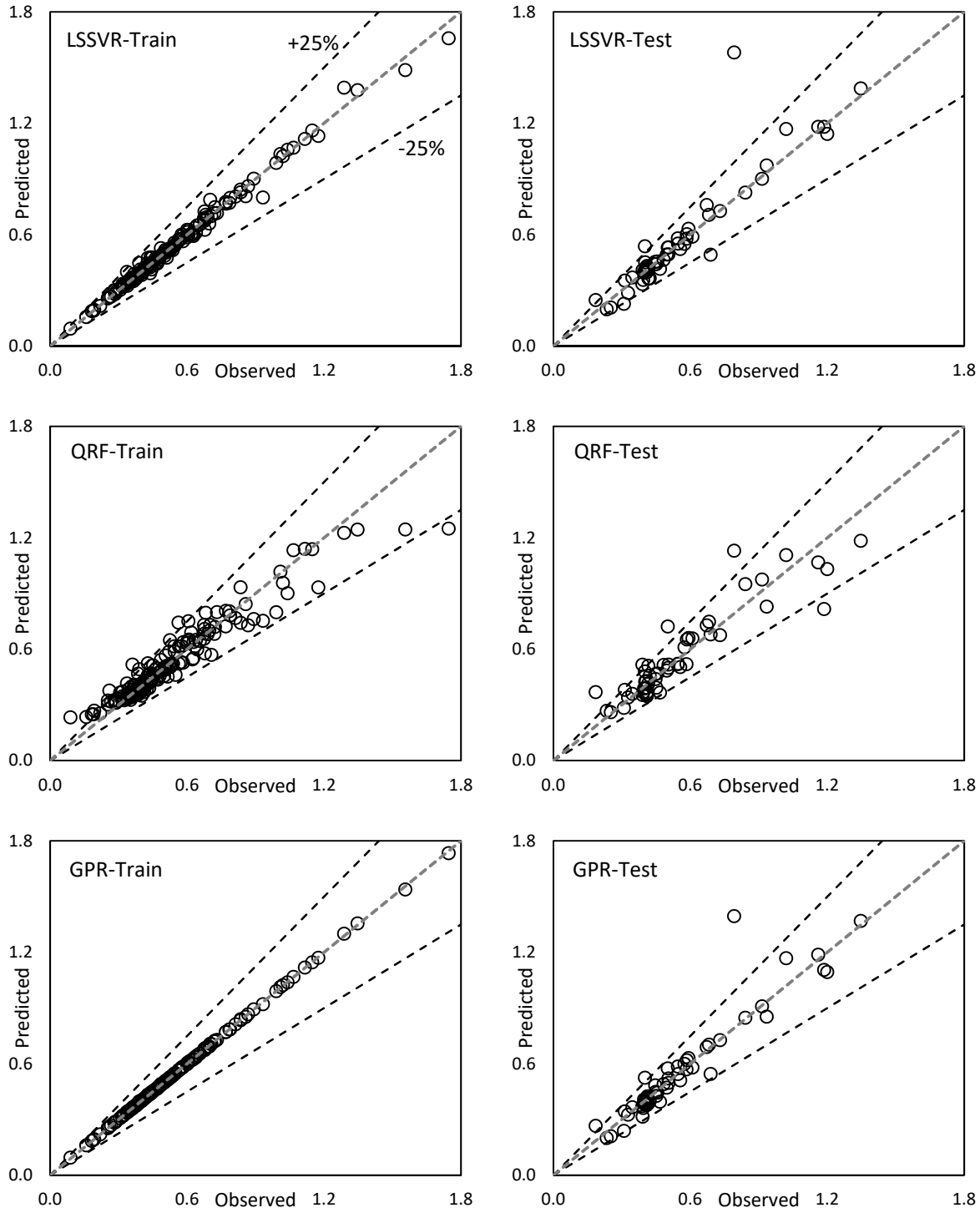


Fig. 7. The scatterplots of LSSVR, QRF and GPR in M1 scenario.

All the input parameters from M1 were sequentially removed and the statistical effects of removing them were evaluated. The M2, M3, M4, and M5 inputs were created by the elimination of y_1/L, L/b, p/y_1, and Fr, respectively (Figure 3). Using this procedure, each scenario (M2–M5) was compared with its original situation (M1) to demonstrate the impact of each parameter. The results are depicted in Figure 5 for the LSSVR, QFR, and GPR.

In the LSSVR and QRF, the elimination of p/y_1 and Fr shows significant influences on Cd. Also, in the GPR, the elimination of p/y_1 and L/b exerts a significant effect. Borghei et al. (1999); Jalili and Borghei (1996) considered the effect of L/b and p/y_1 on Cd. Additionally, Agaccioglu and Yüksel (1998), Emiroglu et al. (2011), Hussain et al. (2021), and Kaya et al. (2011) found that Cd values tend to increase with increasing L/b values.

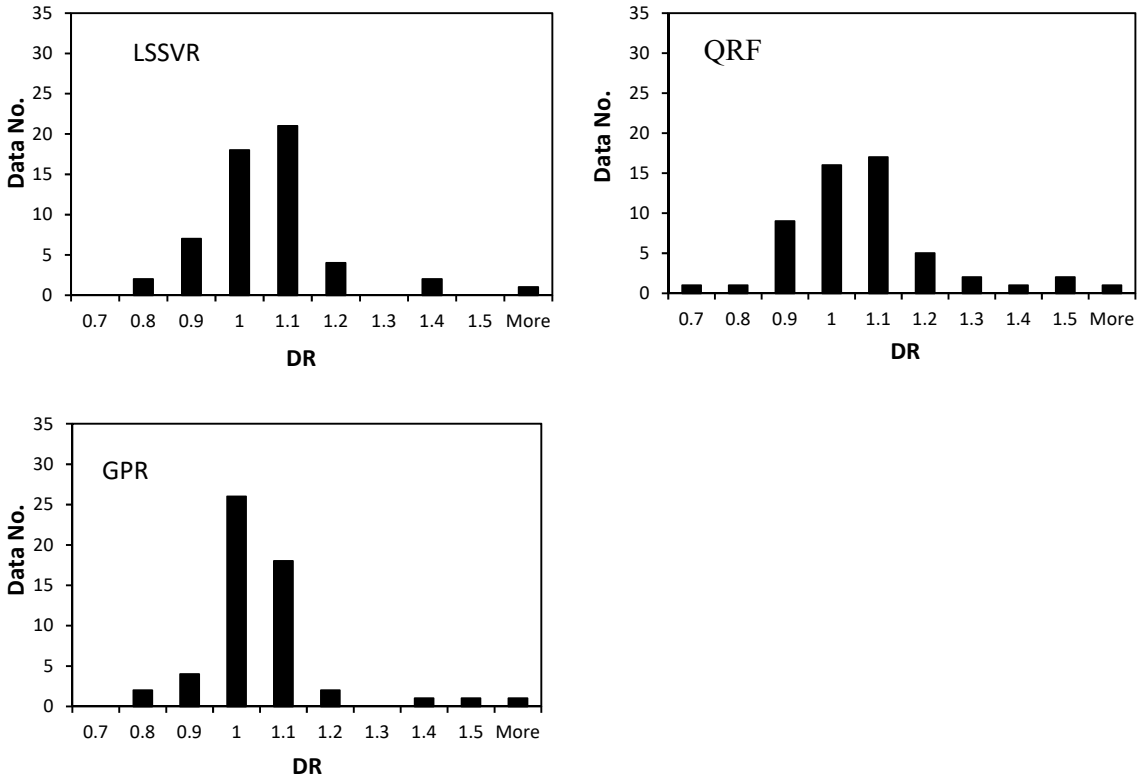


Fig. 8. The discharge coefficient discrepancy ratios (DR) in testing data.

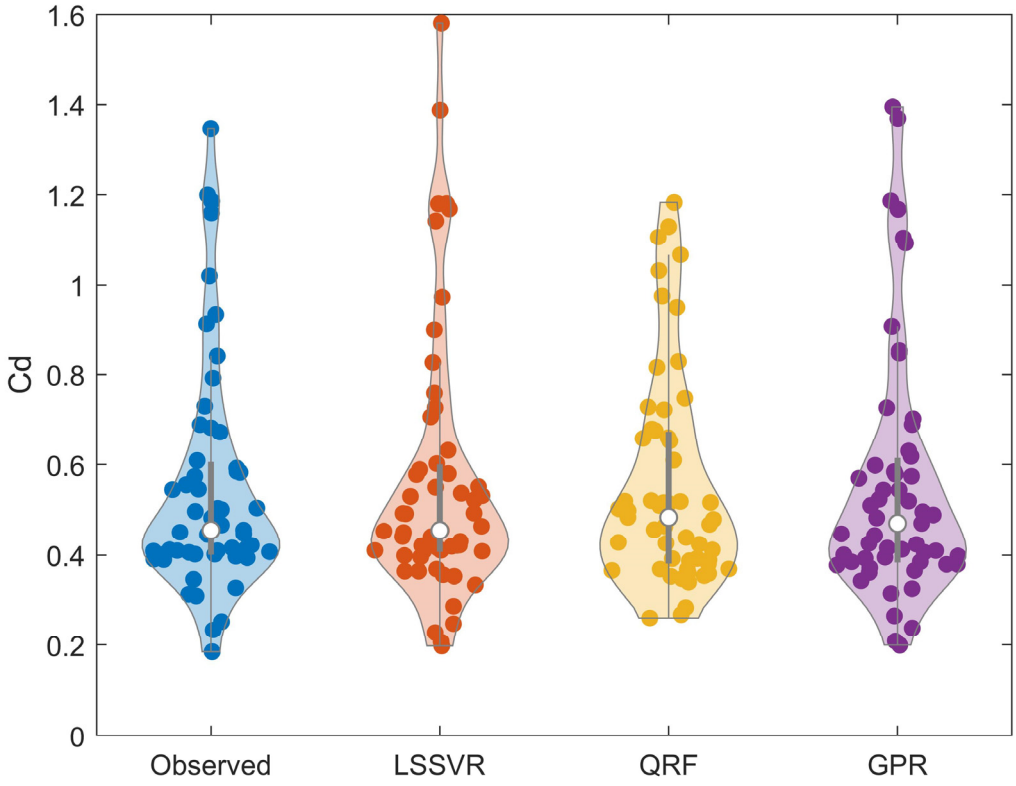


Fig. 9. The violin plot for a comparison between observed and predicted test data for the M1 input structure.

Ebtehaj (2015) showed that Fr and p/y_1 parameters had similar effects on Cd. For instance, if Fr is omitted in the LSSVR and QRF, RAE ascends from 0.253 to 0.459 and 0.349 to 0.404 and R^2 descends from 0.848 to 0.760 and 0.849 to 0.807, respectively. It is therefore concluded that the Fr of a sharp-crested rectangular weir is highly influential on the Cd, as shown in previous studies (Azamathulla et al., 2016; Borghei et al., 1999; Ranga Raju Kittur et al., 1979). The exclusion of y_1/L in the LSSVR and GPR and the exclusion of y_1/L and L/b in the QRF seem to have little effect.

In another scenario, two important parameters (M6: Fr and p/y1 for the LSSVR and QRF, and M9: P/y1 and L/b for the GPR) were used to predict discharge coefficients. For three models, this idea has been tested by modelling with these two effective parameters. The results are shown in Figure 5. The effective contour plot for each model was constructed by plotting the most influential input variables against the predicted discharge coefficients (Figure 10). It should be mentioned that the important parameters in each model considered included Fr and p/y1 for the LSSVR and QRF and L/b and p/y1 for the GPR. Figure 10 shows that the maximum Cd values were in ranges of Fr>0.75 and p/y1>0.65 in the LSSVR, in the ranges of Fr>0.75 and p/y1>0.30 in the QRF, and in the ranges of 1.3<L/b<1.8 and 0.3<p/y1<0.5, and 2.5<L/b and 0.85<p/y1<0.5 in the GPR.

Comparison with classical equations

Extensive literature is available for the prediction of Cd. The existing models were compared with classical equations to assess their precision (Eqs. 28–30). These models were selected for comparison in the present study. The comparison between the observed Cd of rectangular side weir and those predicted by the suggested classical equations are shown in Figure 10, and the statistical performance indices are presented in Table 2.

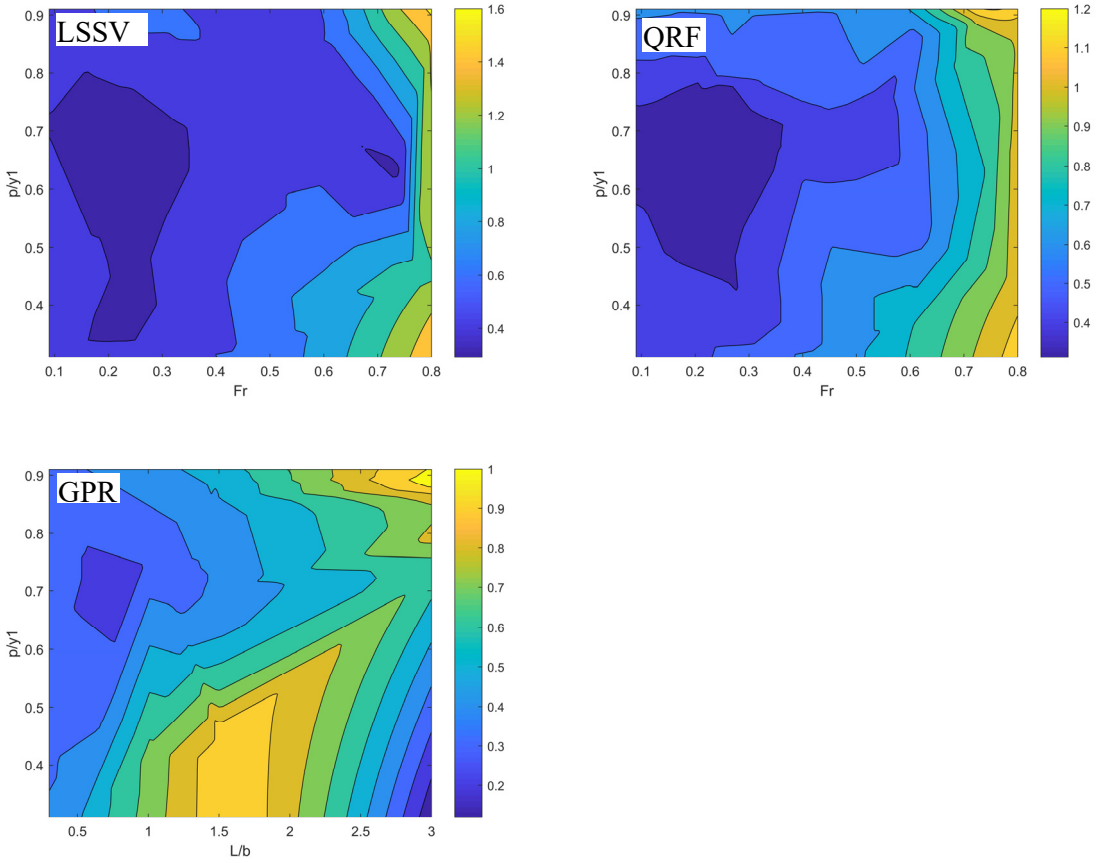


Fig. 10. The variations of the most important parameters against predicted Cd.

Table 2. A comparison of GPR with existing equations using statistical indices.

Equation No.		Train			Test		
		R^2	RMSE	RAE	R^2	RMSE	RAE
28	Emiroglu et al. 2011	**0.70**	**0.14**	**0.46**	**0.54**	**0.18**	**0.51**
29	Subramanya et al.	**0.38**	**0.30**	**1.16**	**0.57**	**0.32**	**1.24**
30	Cheong	**0.38**	**0.30**	**1.05**	**0.57**	**0.32**	**1.06**
	GPR-M1	**0.97**	**0.03**	**0.08**	**0.88**	**0.10**	**0.24**

Comparing Eqs. (29–30), which predict Cd based on Fr, the results are similar. The evaluation of Eq. (28), which considers L/b, L/y_1, p/y_1, and Froude number, shows that Eq. (28) provides the most accurate predictions for Cd as compared to Eqs. (29) and (30). Comparing the classical equations reveals that the L/b, L/y_1, and p/y_1 parameters (Eq. 22) reduce the error by nearly half the error reported by Cheong (1991); Subramanya and Awasthy (1972). According to Table 2, the GPR ($R^2 = 0.88$, RMSE = 0.10, RAE = 0.24) significantly outperforms the empirical equations.

Figure 11 shows that the results produced by Eq. (29) and Eq. (30) are less dispersed than those of Eq. (28). According to Figure 11, (Subramanya and Awasthy, 1972) and (Cheong, 1991) equations yield similar results such that as Cd increases, the prediction accuracy decreases. However, Eqs. (29) and (30) fail to provide good predictions even for small Cd values. Figure 11 also displays that the results of all classical equations have significant errors when Cd exceeds 0.6.

In conclusion, according to the information presented in Table 2, which displays the values of prediction errors using statistical indices, and Figure 11, a significant improvement is shown over classical equations that present weak performance in Cd prediction.

Uncertainty

Model coverage probabilities are quantitatively evaluated by using uncertainty analysis. The uncertainty criteria for all models in testing data are presented in Table 3.

Figure 12 illustrates 95% prediction intervals (PI) for the prediction of Cd applying the LSSVR, QRF, and GPR models in the M1 input structures during the testing phase.

To validate the uncertainty quantification provided by the PI, we computed the containing ratio (CR) of the observed data that fall within these intervals. An accurate uncertainty quantification should present that CRs are similar to the probability of the 95% PI. Across all draws of 55 observations, we would expect 5% of the data points, on average, to fall outside the PI (95% would be in the PI). In this case, we found a coverage ratio of 0.84, 0.98, and 0.93 for the LSSVR, QRF, and GPR in the M1 input structure, respectively, which meant that 46, 54, and 51 of 55 observation data fell within 95% PI. From Figure 12 and Table 3, we find that there is a good match between 95% confidence intervals and the results obtained by the QRF and GPR models, and most of the observed Cd data fall within the PI in M1 input. 95% PI of the LSSVR model is found to underestimate uncertainty (CR = 0.84). It is also interesting to note that the width of the PI is larger in the high value of Cd in the QRF. However, a slight underestimation of the uncertainty can be seen in the GPR (CR = 0.93) where the prediction intervals are narrower than those in the QRF. The values bracketed by PI, and the bandwidth values, indicate that the GPR model can predict Cd in M1 input with smaller uncertainties.

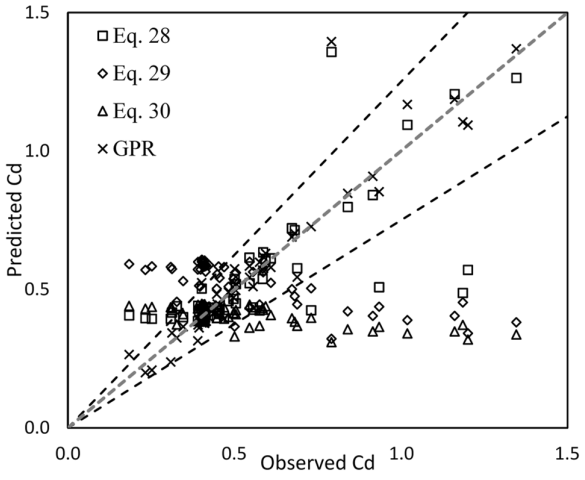


Fig. 11. A comparison of the classical equation with GPR-M1 in the test phase.

Table 3. Uncertainty performance indices.

Input structure	LSSVR			QRF			GPR		
	CR	RB	MPB	CR	RB	MPB	CR	RB	MPB
M1	84	0.44	0.21	98	0.90	0.50	93	0.57	0.26
M2	93	0.48	0.25	100	1.28	0.67	89	0.60	0.27
M3	82	0.49	0.24	100	1.24	0.65	96	0.64	0.29
M4	96	0.96	0.48	98	1.09	0.59	96	0.96	0.45
M5	95	1.03	0.53	98	1.22	0.63	98	1.26	0.58
M6	96	1.10	0.57	98	1.45	0.75	93	1.25	0.59
M7	93	0.87	0.42	96	1.10	0.59	95	1.08	0.50
M8	78	0.96	0.45	84	0.97	0.52	82	1.27	0.62
M9	96	0.70	0.40	95	1.34	0.68	98	1.23	0.57
M10	96	1.05	0.54	96	1.33	0.67	98	1.34	0.62
M11	96	1.25	0.67	98	1.15	0.60	95	1.58	0.74
M12	100	1.47	0.77	95	1.21	0.62	95	1.69	0.79
M13	96	1.80	0.89	85	1.47	0.73	93	1.98	0.93
M14	96	1.29	0.67	95	1.00	0.54	93	1.62	0.75
M15	96	1.26	0.67	91	0.94	0.85	84	0.80	0.37

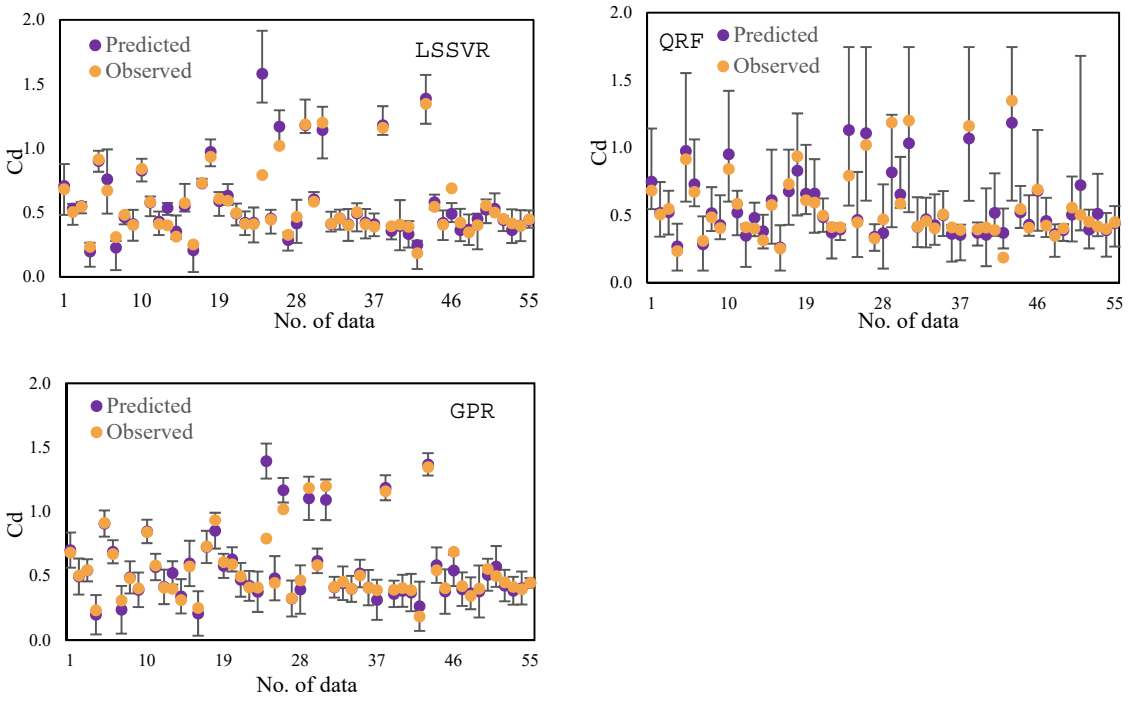


Fig. 12. The observed and predicted Cd in the M1 input for the testing data. The error bars indicate 95% PI.

In 3-input structures (M2-M5), the width of PI is overestimated in some cases (e.g., the QRF in M2 and M3), while in other cases, it is underestimated (e.g., the LSSVR in M1 and M3), leading to PIs that are either too wide or too narrow. As noted above, the expected proportion of true values in the 95% PI is 0.95. It can be concluded that in the 3-input structures, the LSSVR and GPR models predict uncertainty more accurately than the QRF model. In terms of the point prediction in the 3-input structure, all models predict Cd with high accuracy in the M2 input structure. Therefore, 95% PIs created by models in the M2 input structure are shown in Figure 12. In the M2 input structure, the LSSVR and GPR could create PI with a combination of parameters including Fr, p/y_1, and L/b with smaller uncertainties compared to the QRF model. The QRF overestimates uncertainty (CR = 1, RB = 0.48, MPB = 0.67) whereas the LSSVR slightly underestimates it (CR = 0.93, RB = 1.28, MPB = 0.25).

The 95% PI obtained from models in the 2-input structures can be seen in Table 3. The results show that models underestimate uncertainty in some cases (e.g., the LSSVR, QRF, and GPR in M8) or overestimate in others (e.g., the QRF in M6 and the GPR in M9). In general, poor results indicate that models could not predict the uncertainty with acceptable accuracy; more observed data are bracketed within the 95% PI.

In the 1-input structures (M12–M15), it can be seen that the P/y_1 parameter has higher uncertainty in all models than other parameters. The higher uncertainty of P/y_1 is related to the high value of bandwidth (i.e., 0.89, 0.73, and 0.93 for the LSSVR, QRF, and GPR, respectively).

By removing parameters in input structures (e.g., compare 3-input with 2-input), the predicted bandwidth increases by the models. Compared to the other models, the GPR method with 4 inputs is more suitable and reliable when estimating Cd based on all six prediction and uncertainty performance evaluation criteria.

DISCUSSION

In this study, we aimed to evaluate the precision of three machine learning models for predicting side weir Cd. Accurate prediction of Cd is crucial due to the complex nonlinear relationship between hydraulic and geometric parameters and Cd. We selected the QRF, LSSVR, and GPR based on their remarkable progression in previous studies (Francke et al., 2008; Hu et al., 2021; Tao et al., 2022). It is worth mentioning that the prediction interval for side weir Cd is novel since it has not been extensively studied in the literature.

The results indicate that the GPR can predict Cd with the highest precision. Additionally, the impact of input parameters on prediction accuracy was explored by evaluating ML model precision using various combinations of input parameters. This research not only focused on prediction intervals but also examined the role of input variables in the prediction process.

To verify the superiority of the GPR, the forecasting results were compared with some classical equations according to which the GPR exhibited lower bias than the classical equations.

The predictions produced by the ML models are inadequate and do not allow decision-makers to gauge the accuracy of each prediction. A more viable option is interval prediction, which can measure the level of uncertainty in the Cd and is preferable over conventional point prediction. Therefore, the uncertainty of forecasting results was calculated and analysed for the ML models. By means of interval prediction, decision-makers can receive a quantitative assessment of the potential range of the Cd. The constructed prediction intervals (PIs) produced by various ML models differ in quality due to model parameters. In practical applications, decision-makers anticipate a PI with a coverage probability equivalent to or greater than the given confidence level and a narrow interval width. A PI that has a high coverage probability but an excessively wide width is useless. A comparative analysis reveals that the GPR is significantly more effective than the other compared methods in building high-

quality PIs with both a high coverage probability and narrow width. The GPR can directly build PIs without making any erroneous assumptions, demonstrating high robustness and reliability. This makes it a promising tool for constructing PIs of Cd. On the other hand, the PIs constructed using the QRF model are unsatisfactory in terms of the confidence level and the interval width. We used these methods with no additional calibration procedure. Roy and Larocque (2019) showed that calibrating the QRF would certainly improve its performance, and this is certainly an avenue for future research.

As mentioned in the "Input structure" section, the applied models were constructed using a total of 15 dimensionless parameters (see Figure 3). Figure 5 shows that there are significant differences in the importance of the dimensionless parameters. While p/y_1 was found to be the most influential parameter in the models, the order and significance of the remaining variables differed. The following considerations are presented when analysing these results:

a) The parameters of each model influence the results.

b) The nature of each model for computing the importance of the dimensionless parameters is different.

The calculation results show that the point prediction and prediction interval of the GPR are higher than those of the other models. One of the reasons for the superiority of this model is its generalization properties. By choosing various kernel functions, users have the ability to incorporate prior knowledge and specifications regarding the model's shape. Also, usability and flexibility in implementation can be another advantage of this model. It is worth noting that the outcomes support the aim of preparing a robust ML model and greatly help the application of side weir.

CONCLUSION

In the present study, three machine learning models, the LSSVR and two non-parametric models, the QRF and the GPR, were used to predict side weir Cd. Moreover, the uncertainty and reliability of the models were investigated for Cd prediction. Four hydraulic and geometrical parameters, i.e., Froude number (Fr), ratios of upstream flow depth to length of weir (y_1/L), height of weir to upstream flow depth (p/y_1), and dimensionless length of a weir (L/b), were used as inputs, and the coefficient of discharge (Cd) was defined as model output. The scenarios to develop the models included 15 input structures (M1-M15). According to the statistical indices of the test data, all three ML models performed well in predicting the side weir Cd, and the RMSE-value was equal to 0.118, 0.098, and 0.095 for the LSSVR, QRF, and GPR, respectively. The comparison between ML models and classical equations (regression-based) demonstrated that ML models outperformed in their ability to predict Cd. With negligible difference, the result showed that the GPR (R^2=0.883, RMSE = 0.095, RAE = 0.242) in the test phase of M1 input outperformed both the LSSVR (R^2 = 0.848, RMSE = 0.118, RAE = 0.253) and the QRF (R^2 = 0.849, RMSE = 0.098, RAE = 0.349). Overall, the superiority of the GPR in point prediction over other investigated models was found in all input structures (M1–M15). The outcomes indicated that all implemented machine learning models were statistically valid. Additionally, sensitivity analysis provided insight into the importance of input parameters, and thus p/y_1 and Fr in the LSSVR and QRF, whereas p/y_1 and L/b in the GRP were the most effective parameters in the model's efficiency. A comparison of different developed input structures showed that the models with effective parameters could predict the Cd, but the prediction accuracy was significantly lower than the best structure (M1) in the LSSVR and QRF. Although the use of two parameters reduced the complexity of the simulation, the results were inadequate.

Uncertainty remains an important issue in Cd prediction, but tools used to determine uncertainty must follow the application and development of the latest modelling methods. Owing to the increasing use of non-parametric and efficient methods for modelling, this research compared the performance of two non-parametric models with the LSSVR in predicting uncertainty according to the input parameters. Compared to the other methods, the GPR (CR = 93, RB = 0.57, MPB = 0.26) provided reasonable results in the best input structure (M1) compared to the LSSVR (CR = 84, RB = 0.44, MPB = 0.21) and QRF (CR = 98, RB = 0.90, MPB = 0.50). Overall, the PI estimated using the QRF was generally wider than the PI estimated using the LSSVR and GPR in most input structures. In terms of prediction performance, the GPR method outperformed the other models. Additionally, for uncertainty estimation, the GPR method provided results that were similar in quality to those obtained from the LSSVR method in all input structures. The results showed that the GPR could gain high accuracy in Cd prediction and high-performance PI.

The primary shortcoming of the ML models in this study was their reliance on a black-box approach to predict Cd, which limited their transparency. Additionally, the applied ML models demonstrated certain limitations such as the need for determining internal parameters, ensuring model stability, and other drawbacks that require user expertise. Furthermore, these ML models were limited in their capability of forecasting Cd beyond the training data since they could not well extrapolate the data.

ML is concerned with the amount of data used for training. To identify the optimal train-test ratio, future investigations could explore various data portions, such as 80–20 (80% train and 20% test), 70–30, 65–35, 60–40, and 50–50. There is no established method for determining the best model parameters, which largely relies on user knowledge. Hence, optimization algorithms to select the best parameters could be another area of focus for future studies. Additionally, coupling ML models with pre-processing data techniques may enhance Cd prediction accuracy.

Acknowledgements. Gonbad Kavous University provided partial funding for this research under Grant number 6/01/39.

REFERENCES

Abbasi, S., Fatemi, S., Ghaderi, A., Di Francesco, S., 2021. The effect of geometric parameters of the antivortex on a triangular labyrinth side weir. Water, 13, 1. http://dx.doi.org/10.3390/w13010014

Agaccioglu, H., Yüksel, Y., 1998. Side-weir flow in curved channels. J. Irrig. Drain. Eng., 124, 3, 163–175. http://dx.doi.org/10.1061/(ASCE)0733-9437(1998)124:3(163)

Ahmed, M.H., Lin, L.-S., 2021. Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique. J. Hydrol., 597, 1–12. http://dx.doi.org/10.1016/j.jhydrol.2021.126213

Akbari, M., Salmasi, F., Arvanaghi, H., Karbasi, M., Farsadizadeh, D., 2019. Application of Gaussian process regression model to predict discharge coefficient of gated piano key weir. Water Resour. Manage., 33, 11, 3929–3947. http://dx.doi.org/10.1007/s11269-019-02343-3

Anandhi, A., Srinivas, V.V., Nanjundiah, R.S., Nagesh Kumar, D., 2008. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. International Journal of Climatology, 28, 3, 401–420.

http://dx.doi.org/https://doi.org/10.1002/joc.1529
Azamathulla, H.M., Haghiabi, A.H., Parsaie, A., 2016. Prediction of side weir discharge coefficient by support vector machine technique. Water Supply, 16, 4, 1002–1016. http://dx.doi.org/10.2166/ws.2016.014
Bagheri, S., Kabiri-Samani, A.R., Heidarpour, M., 2014. Discharge coefficient of rectangular sharp-crested side weirs. Part II: Domínguez's method. Flow Meas. Instrum., 35, 116–121. http://dx.doi.org/10.1016/j.flowmeasinst.2013.10.006.
Bhuiyan, M.A.E., Nikolopoulos, E.I., Anagnostou, E.N., Quintana-Seguí, P., Barella-Ortiz, A., 2018. A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula. Hydrol. Earth Syst. Sci., 22, 2, 1371–1389. http://dx.doi.org/10.5194/hess-22-1371-2018
Bonakdari, H., Ebtehaj, I., Samui, P., Gharabaghi, B., 2019. Lake Water-Level fluctuations forecasting using Minimax Probability Machine Regression, Relevance Vector Machine, Gaussian Process Regression, and Extreme Learning Machine. Water Resour. Manage., 33, 11, 3965–3984. http://dx.doi.org/10.1007/s11269-019-02346-0
Bonakdari, H., Zaji, A.H., Shamshirband, S., Hashim, R., Petkovic, D., 2015. Sensitivity analysis of the discharge coefficient of a modified triangular side weir by adaptive neuro-fuzzy methodology. Meas., 73, 74–81. http://dx.doi.org/10.1016/j.measurement.2015.05.021
Borghei, S.M., Jalili, M.R., Ghodsian, M., 1999. Discharge coefficient for sharp-crested side weir in subcritical flow. J. Hydraul. Eng., 125, 10, 1051–1056. http://dx.doi.org/10.1061/(ASCE)0733-9429(1999)125:10(1051)
Borghei, S.M., Nekooie, M.A., Sadeghian, H., Jalili Ghazizadeh, M.R., 2013. Triangular labyrinth side weirs with one and two cycles. Proc. Inst. Civ. Eng. Water Manage., 166, 1, 27–42. http://dx.doi.org/10.1680/wama.11.00032
Bowden, G.J., Maier, H.R., Dandy, G.C., 2005. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. J. Hydrol., 301, 1, 93–107. http://dx.doi.org/10.1016/j.jhydrol.2004.06.020
Brabanter, K.D., Brabanter, J.D., Suykens, J.A.K., Moor, B.D., 2011. Approximate confidence and prediction intervals for least squares support vector regression. IEEE Trans. Neural Networks, 22, 1, 110–120. http://dx.doi.org/10.1109/TNN.2010.2087769
Breiman, L., 2001. Random forests. Mach. Learn., 45, 1, 5–32. http://dx.doi.org/10.1023/A:1010933404324
Cartwright, H.M., 2015. Artificial Neural Networks. Springer, New York.
Cheong, H.F., 1991. Discharge coefficient of lateral diversion from trapezoidal channel. J. Irrig. Drain. Eng., 117, 4, 461–475. http://dx.doi.org/10.1061/(ASCE)0733-9437(1991)117:4(461)
Coleman, H.W., Steele, W.G., 2009. Experimentation, Validation, and Uncertainty Analysis for Engineers. Wiley, New York, NY, USA.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn., 20, 3, 273–297. http://dx.doi.org/10.1007/BF00994018
Ebtehaj, I., Bonakdari, H., Gharabaghi, B., 2018. Development of more accurate discharge coefficient prediction equations for rectangular side weirs using adaptive neuro-fuzzy inference system and generalized group method of data handling. Meas., 116, 473–482. http://dx.doi.org/10.1016/j.measurement.2017.11.023
Ebtehaj, I., Bonakdari, H., Zaji, A.H., Azimi, H., Khoshbin, F., 2015. GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp-crested side weirs. Eng. Sci. Technol. Int. J., 18, 4, 746–757. http://dx.doi.org/10.1016/j.jestch.2015.04.012
Emiroglu, M.E., Agaccioglu, H., Kaya, N., 2011. Discharging capacity of rectangular side weirs in straight open channels. Flow Meas. Instrum., 22, 4, 319–330. http://dx.doi.org/10.1016/j.flowmeasinst.2011.04.003
Francke, T., López-Tarazón, J.A., Schröder, B., 2008. Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. Hydrol. Process., 22, 25, 4892–4904. http://dx.doi.org/10.1002/hyp.7110
Gholami, A., Bonakdari, H., Ebtehaj, I., Mohammadian, M., Gharabaghi, B., Khodashenas, S.R., 2018. Uncertainty analysis of intelligent model of hybrid genetic algorithm and particle swarm optimization with ANFIS to predict threshold bank profile shape based on digital laser approach sensing. Meas., 121, 294–303. http://dx.doi.org/10.1016/j.measurement.2018.02.070
Granata, F., de Marinis, G., Gargano, R., Tricarico, C., 2013. Novel approach for side weirs in supercritical flow. J. Irrig. Drain. Eng., 139, 8, 672–679. http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000600
Haddadi, H., Rahimpour, M., 2012. A discharge coefficient for a trapezoidal broad-crested side weir in subcritical flow. Flow Meas. Instrum., 26, 63–67. http://dx.doi.org/10.1016/j.flowmeasinst.2012.04.002
Hager, W., 1987. Lateral outflow over side weirs. J. Hydraul. Eng., 113, 4, 491–504. http://dx.doi.org/10.1061/(ASCE)0733-9429(1987)113:4(491)
Hu, Z., Karami, H., Rezaei, A., DadrasAjirlou, Y., Piran, M.J., Band, S.S., Chau, K.-W., Mosavi, A., 2021. Using soft computing and machine learning algorithms to predict the discharge coefficient of curved labyrinth overflows. Eng. Appl. Comput. Fluid Mech., 15, 1, 1002–1015. http://dx.doi.org/10.1080/19942060.2021.1934546
Hussain, A., Shariq, A., Danish, M., Ansari, M., 2021. Discharge coefficient estimation for rectangular side weir using GEP and GMDH methods. Adv. Comput. Des., 6, 2, 135–151. http://dx.doi.org/10.12989/acd.2021.6.2.135
Jalili, M.R., Borghei, S.M., 1996. Discussion: Discharge coefficient of rectangular side weirs. J. Irrig. Drain. Eng., 122, 2, 132–132. http://dx.doi.org/10.1061/(ASCE)0733-9437(1996)122:2(132)
Johnson, P.A., Ayyub, B.M., 1996. Modeling uncertainty in prediction of pier scour. J. Hydraul. Eng., 122, 2, 66–72. http://dx.doi.org/10.1061/(ASCE)0733-9429(1996)122:2(66)
Karbasi, M., Jamei, M., Ahmadianfar, I., Asadi, A., 2021. Toward the accurate estimation of elliptical side orifice discharge coefficient applying two rigorous kernel-based data-intelligence paradigms. Sci. Rep., 11, 1, 19784. http://dx.doi.org/10.1038/s41598-021-99166-3
Kaya, N., Emiroglu, M.E., Agaccioglu, H., 2011. Discharge coefficient of a semi-elliptical side weir in subcritical flow. Flow Meas. Instrum., 22, 1, 25–32. http://dx.doi.org/10.1016/j.flowmeasinst.2010.11.002
Kilic, Z., Emin Emiroglu, M., 2022. Study of hydraulic characteristics of trapezoidal piano key side weir using different approaches. Water Supply, 22, 8, 6672–6691. http://dx.doi.org/10.2166/ws.2022.264
Kisi, O., Ozkan, C., 2017. A new approach for modeling sediment-discharge relationship: Local weighted linear regression. Water Resour. Manage., 31, 1, 1–23. http://dx.doi.org/10.1007/s11269-016-1481-9
Liao, K.-W., Chien, F.-S., Ju, R.-J., 2019. Safety evaluation of a water-immersed bridge against multiple hazards via machine learning. Appl. Sci., 9, 15, 3116. http://dx.doi.org/10.3390/app9153116
Liu, Y., Guo, J., Wang, Q., Huang, D., 2016. Prediction of filamentous sludge bulking using a state-based Gaussian processes regression model. Sci. Rep., 6, 1, 31303. http://dx.doi.org/10.1038/srep31303
Maranzoni, A., Pilotti, M., Tomirotti, M., 2017. Experimental and numerical analysis of side weir flows in a converging channel. J. Hydraul. Eng., 143, 7, 1–15. http://dx.doi.org/10.1061/(ASCE)HY.1943-7900.0001296
Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. J. Mach. Learn. Res., 7, 6, 983–999.
Mohammed, A.Y., Golijanek-Jędrzejczyk, A., 2020. Estimating the uncertainty of discharge coefficient predicted for oblique side weir using Monte Carlo method. Flow Meas. Instrum., 73, 1–15. http://dx.doi.org/10.1016/j.flowmeasinst.2020.101727

Momeni, E., Dowlatshahi, M.B., Omidinasab, F., Maizir, H., Armaghani, D.J., 2020. Gaussian process regression technique to estimate the pile bearing capacity. Arabian J. Sci. Eng., 45, 10, 8255–8267. http://dx.doi.org/10.1007/s13369-020-04683-4

Nateghi, R., Guikema, S.D., Quiring, S.M., 2014. Forecasting hurricane-induced power outage durations. Nat. Hazard., 74, 3, 1795–1811. http://dx.doi.org/10.1007/s11069-014-1270-9

Nourani, B., Arvanaghi, H., Salmasi, F., 2021. A novel approach for estimation of discharge coefficient in broad-crested weirs based on Harris Hawks Optimization algorithm. Flow Meas. Instrum., 79, 1–13. http://dx.doi.org/10.1016/j.flowmeasinst.2021.101916

Olyaie, E., Banejad, H., Heydari, M., 2019. Estimating discharge coefficient of PK-weir under subcritical conditions based on high-accuracy machine learning approaches. Iran. J. Sci. Technol. Trans. Civ. Eng., 43, 1, 89–101. http://dx.doi.org/10.1007/s40996-018-0150-z

Parsaie, A., Haghiabi, A., 2015. The effect of predicting discharge coefficient by neural network on increasing the numerical modeling accuracy of flow over side weir. Water Resour. Manage., 29, 4, 973–985. http://dx.doi.org/10.1007/s11269-014-0827-4

Parsaie, A., Haghiabi, A.H., 2021. Uncertainty analysis of discharge coefficient of circular crested weirs. Appl. Water Sci., 11, 2, 1–6. http://dx.doi.org/10.1007/s13201-020-01329-6

Pospíšilík, Š., Zachoval, Z., 2023. Discharge coefficient, effective head and limit head in the Kindsvater-Shen formula for small discharges measured by thin-plate weirs with a triangular notch. J. Hydrol. Hydromech., 71, 1, 35–48. http://dx.doi.org/doi:10.2478/johh-2022-0040

Prayogo, D., Susanto, Y.T.T., 2018. Optimizing the prediction accuracy of friction capacity of driven piles in cohesive soil using a novel self-tuning least squares support vector machine. Adv. Civ. Eng., 2018, 1–9. http://dx.doi.org/10.1155/2018/6490169

Ranga Raju Kittur, G., Gupta Sushil, K., Prasad, B., 1979. Side weir in rectangular channel. J. Hydraulics Div., 105, 5, 547–554. http://dx.doi.org/10.1061/JYCEAJ.0005207

Říha, J., Zachoval, Z., 2014. Discharge coefficient of a trapezoidal broad-crested side weir for low approach Froude numbers. J. Hydraul. Eng., 140, 8, 1–6. http://dx.doi.org/10.1061/(ASCE)HY.1943-7900.0000889

Říha, J., Zachoval, Z., 2015. Flow characteristics at trapezoidal broad-crested side weir. J. Hydrol. Hydromech., 63, 2, 164–171. http://dx.doi.org/10.1515/johh-2015-0026

Roushangar, K., Akhgar, S., 2020. Particle swarm optimization-based LS-SVM for hydraulic performance of stepped spillway. ISH J. Hydraul. Eng., 26, 3, 273–282. http://dx.doi.org/10.1080/09715010.2018.1481773

Roy, M.-H., Larocque, D., 2019. Prediction intervals with random forests. Statistical Methods in Medical Research, 29, 1, 205–229. http://dx.doi.org/10.1177/0962280219829885

Salmasi, F., Nouri, M., Sihag, P., Abraham, J., 2021. Application of SVM, ANN, GRNN, RF, GP and RT models for predicting discharge coefficients of oblique sluice gates using experimental data. Water Supply, 21, 1, 232–248. http://dx.doi.org/10.2166/ws.2020.226

Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. J. Math. Psychol., 85, 1–16. http://dx.doi.org/10.1016/j.jmp.2018.03.001

Seyedian, S.M., Ghazizadeh, M.J., Tareghian, R., 2014. Determining side-weir discharge coefficient using Anfis. Proc. Inst. Civ. Eng. Water Manage., 167, 4, 230–237. http://dx.doi.org/10.1680/wama.12.00102

Seyedian, S.M., Rouhani, H., 2015. Assessing ANFIS accuracy in estimation of suspended sediments. Građevinar, 67, 12, 1165–1176. http://dx.doi.org/10.14256/JCE.1210.2015

Subramanya, K., Awasthy, S.C., 1972. Spatially varied flow over side-weirs. J. Hydraulics Div., 98, 1, 1–10. http://dx.doi.org/10.1061/JYCEAJ.0003188

Suykens, J.A.K., De Brabanter, J., Lukas, L., Vandewalle, J., 2002. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing, 48, 1, 85–105. http://dx.doi.org/10.1016/S0925-2312(01)00644-0

Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett., 9, 3, 293–300. http://dx.doi.org/10.1023/A:1018628609742

Tao, H., Jamei, M., Ahmadianfar, I., Khedher, K.M., Farooque, A.A., Yaseen, Z.M., 2022. Discharge coefficient prediction of canal radial gate using neurocomputing models: an investigation of free and submerged flow scenarios. Eng. Appl. Comput. Fluid Mech., 16, 1, 1–19. http://dx.doi.org/10.1080/19942060.2021.2002721

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106, D7, 7183–7192. http://dx.doi.org/10.1029/2000JD900719

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

Xiong, L., Wan, M., Wei, X., O'Connor, K.M., 2009. Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. Hydrol. Sci. J., 54, 5, 852–871. http://dx.doi.org/10.1623/hysj.54.5.852

Yadav, A., Hasan, M.K., Joshi, D., Kumar, V., Aman, A.H., Alhumyani, H., Alzaidi, M.S., Mishra, H., 2022. Optimized scenario for estimating suspended sediment yield using an artificial neural network coupled with a genetic algorithm. Water, 14, 18. http://dx.doi.org/10.3390/w14182815

Yi, T., Zheng, H., Tian, Y., Liu, J.-P., 2018. Intelligent prediction of transmission line project cost based on least squares support vector machine optimized by particle swarm optimization. Math. Probl. Eng., 2018, 1–12. http://dx.doi.org/10.1155/2018/5458696

Zhao, K., Popescu, S., Meng, X., Pang, Y., Agca, M., 2011. Characterizing forest canopy structure with lidar composite metrics and machine learning. Remote Sensing of Environment, 115, 8, 1978–1996. http://dx.doi.org/10.1016/j.rse.2011.04.001

Zounemat-Kermani, M., Golestani Kermani, S., Kiyaninejad, M., Kisi, O., 2019. Evaluating the application of data-driven intelligent methods to estimate discharge over triangular arced labyrinth weir. Flow Meas. Instrum., 68, 101573. http://dx.doi.org/10.1016/j.flowmeasinst.2019.101573

Received 6 March 2023
Accepted 15 May 2023