Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 89 No 7, 1816 doi: 10.2166/wst.2024.102

Discharge coefficient of vertical sluice gates with broad crested weir under free-submerged orifice flows using best subset regression

Zhuoying Cang [a], Dongdong Jia [a,b,*], Jinyang Wang [a], Jun Yang [a], Youzhi Hao [a] and Xiaona Chen [a]

[a] Key Laboratory of Port, Waterway and Sedimentation Engineering of Ministry of Transport, Nanjing Hydraulic Research Institute, Nanjing 210024, China
[b] Yangtze Institute for Conservation and Development, Nanjing 210024, China
*Corresponding author. E-mail: ddjia@nhri.cn

ZC, 0009-0009-6427-2674

ABSTRACT

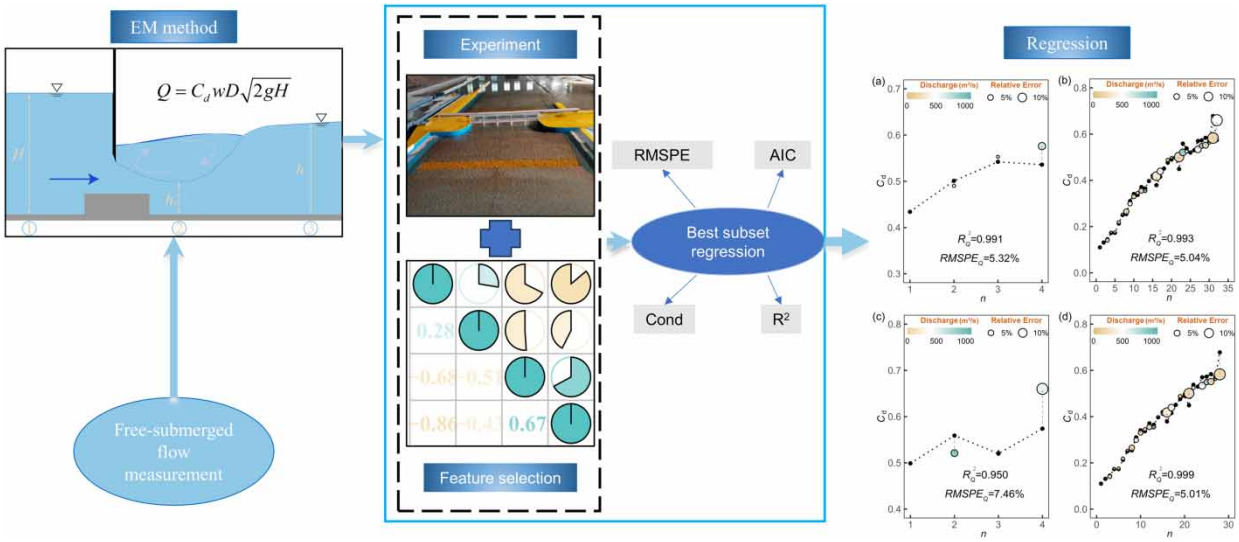
Accurate calculation of flow discharge for sluice gates is essential in irrigation, water supply, and structure safety. The measurement of discharge with the requirement of distinguishing flow regimes is not conducive to application. In this study, a novel approach that considers both free and submerged flow was proposed. The energy–momentum method was employed to derive the coefficient of discharge. Subsequently, the discharge coefficient was determined through the experiment which was performed on the physical model of a vertical sluice gate with a broad-crested weir. Feature engineering, incorporating dimensional analysis, feature construction, and correlation-based selection were performed. The best subset regression method was employed to develop regression equations of the discharge coefficient with the generated features. The derived formula was applied to compute the discharge coefficient in the vertical sluice gate and determine the flow discharge. The accuracy of adopted method was assessed by comparing it with recent studies on submerged flow, and the results demonstrate that the developed approach achieves a high level of accuracy in calculating flow discharge. The coefficient of determination for the calculated flow rate is 0.993, and the root mean square percentage error is 5.04%.

Key words: best subset regression, broad-crested weir, discharge coefficient, flow discharge calculation, free-submerged orifice flow, vertical sluice gate

HIGHLIGHTS

- The feature combination with efficient interpretability of free-submerged flow discharge coefficient was proposed.
- The computational equation applicable to free-submerged flow regimes based on the EM method was established with high accuracy.
- The process of feature selection and calculation is not complicated and has potential for automated calibration.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Gates and weirs are commonly employed devices for regulating water levels, controlling and measuring flow rates in the open channel. The accuracy of flow measurement plays a crucial role in gate operation of the hydraulic structure within open channel or estuary. Moreover, it has a profound impact on water supply, irrigation, flood control, sedimentation transport, navigation, even the safety of hydraulic structures. (Donnelly *et al.* 2022, 2024; Noori *et al.* 2022). An approach extensively utilized to determine flow discharge is the classical energy–momentum (EM) method, which is applicable for free or submerged flow through the gate. The basis of methodology can be traced back to Wóycicki's doctoral thesis, and gained widespread recognition after Henderson's publication (Wóycicki 1931; Henderson 1966). Some scholars also established stage–discharge relationships by employing an analysis that incorporated the theorem of dimensional analysis and incomplete self-similarity theory (Ferro 2000). Ferro (2018) also used the momentum and mass equations to establish the stage–discharge equation for the sharp-crested gate. This equation is characterized by a momentum coefficient, which can be estimated empirically by the ratio between the orifice height and the upstream water depth. Theoretically, the discharge coefficient for both free and submerged flow can be mathematically represented by the equation derived from the EM method. The discharge coefficient (C_d) can be effectively characterized by considering a multitude of factors, such as the energy loss coefficient (k) and contraction coefficient (C_c). These aforementioned parameters assume a crucial role in understanding and predicting flow behaviour. The coefficient of energy loss is commonly acknowledged to be influenced by hydraulic characteristics, like boundary layer and turbulent flow. Meanwhile, the contraction coefficient is closely associated with geometric features, such as channels and gates (Belaud *et al.* 2009; Cassan & Belaud 2012; Castro-Orgaz *et al.* 2013). The C_c value of submerged flow is comparable to that of free flow when at a small opening, irrespective of the submergence. However, this scenario does not hold true for a large opening. In a large gate opening, the contraction coefficient can exceed 0.6 when the flow is adequately submerged. These conditions typically result in large deviations in the predictions and actual measurement of discharge (Belaud *et al.* 2009). In a gate with broad-crested weir, the variations in the flow capacity can be attributed to the interaction between the flow over the weir and the wall jets. Belaud *et al.* (2012) proposed the incorporation of the Boussinesq and Coriolis coefficient into the equation in order to enhance the accuracy of the discharge coefficient prediction, and Bijankhan *et al.*(2017) demonstrated that Coriolis and pressure losses coefficient can improve the discharge calculation. With the improvement of computer technology, contraction and energy loss coefficient can be determined in the data through constrained optimization algorithms. This approach eliminates the need to assume constant values for unknown coefficients empirically in formulas (Habibzadeh *et al.* 2011). In general, the discharge coefficient is widely recognized as a key parameter that reflects various characteristics of the incoming flow within a channel, such as turbulence, viscosity, non-uniform velocity distribution, and velocity head (Bijankhan *et al.* 2017). Therefore, various factors can be utilized as initial features to establish a discharge coefficient calibration model, such as channel width, upstream and downstream water

levels, gate opening, Reynolds number, or Manning's roughness coefficient. The model can be established by directly calibrating the discharge coefficient or calculating the discharge coefficient through calibrating the contraction and energy loss coefficients.

With advances in computational efficiency, researchers have effectively harnessed machine learning in investigation of the discharge coefficient. The computer algorithm has led to substantial enhancements in calibration accuracy compared to traditional regression methods. Multivariate Adaptive Regression Splines (MARS), Artificial Neural Networks (ANN), Symbolic Regression (SR), and various other algorithms have been employed in the study of discharge calculation for gates (Vaheddoost *et al.* 2021; Shakouri *et al.* 2023). Researchers have also compared the predictive accuracy using the same sets of data with different algorithms, such as Support Vector Machines (SVM), ANNs, and Generalized Regression Neural Networks (GRNNs). The objective of these studies is to determine the algorithm that produces the most accurate predictions in discharge calculations (Salmasi *et al.* 2021; Khosravi *et al.* 2022). Research findings on the optimal algorithm for discharge calculation exhibit variations due to factors such as feature selection, parameter tuning, comparison criteria, and algorithm implementation. As researchers attempt to enhance the accuracy of models, they often encounter a trade-off between accuracy and complexity in the formula used. Some studies propose formulas with numerous terms and intricate mapping relationships, occasionally involving nested functions. It is crucial to consider the practical implications of such complex formulations, as they may not be readily applicable to engineering scenarios. Currently, most methods for discharge calculation require the distinction of free or submerged flow regimes, and there is no highly accurate unified calculation method available. However, the advancement of algorithms has provided an opportunity to develop a high-precision computational model for calculating discharge in both free and submerged orifice flow.

The objective of this study is to propose an accurate calculation method for measuring the discharge of free-submerged orifice flow with a broad-crested weir. The prototype of the experimental model is located in a tidal river, which means that the gate flow is influenced by the outer river. Therefore, a considerable portion of the measured data belongs to submerged flow. Besides, it is necessary to fully consider the features that affect flow regime and discharge capability. Therefore, features related to contraction coefficient and energy loss coefficient were extracted. These features were subsequently utilized to construct a high-dimensional feature space based on polynomial theory. In order to streamline the feature space, correlation analysis was employed. Finally, the optimal outcome was determined through the implementation of the best subset regression.

2. MATERIALS AND METHODS

2.1. Experimental setup

The experiment was conducted at the Tie Xin Qiao Test Base, a testing site of Nanjing Hydraulic Research Institute. The experimental prototype is a segment of the sluice gate located in the eastern extension channel of the Da Lu Xian in Shanghai. The sluice gate is designed as the vertical flat gate and the prototype segment was represented by a 1:60 undistorted model, which means the vertical scale is 1:60 same as horizontal scale. The model design meets the requirements of gravity similarity, resistance similarity, continuity similarity and time similarity. Therefore, the velocity scale is 7.75, roughness scale is 1.98, time scale is 7.75 and flow scale is 1:27,885. The prototype of river channel has a roughness coefficient ranging from 0.022 to 0.026. Consequently, the corresponding model roughness coefficient is approximately from 0.011 to 0.013. The surface of the river channel is covered with cement mortar and the gates are made of organic glass panels. As depicted in Figure 1, the prototype gate possesses a clear width of 5×18 m, accompanied by apron and anti-scouring grooves on both the upstream and downstream sides. Furthermore, a stilling basin is located in the downstream, and a broad-crested weir is installed at the gate's base. The elevations mentioned are as follows: the weir crest at -1.5 m, the upstream riverbed at -2.5 m, and the downstream riverbed at -2.0 m.

The experiments comprise measurements of inflow discharge and water levels at various locations upstream and downstream. A triangular weir was utilized to accurately measure the inflow discharge in the experimental setup. The water was pumped out and regulated through the stilling well before passing through the triangular weir. The controlling method of intake flow in the model ensured the precision of the experiment and enhanced the accuracy of the discharge measurements. A total of 24 measurement points were positioned along both the upstream and downstream. Each measurement point was equipped with 0.1 mm precision. Three readings were taken at each point, and the value which exhibited stability was recorded as the water level corresponding to the point. The data collected in the experiment encompassed a

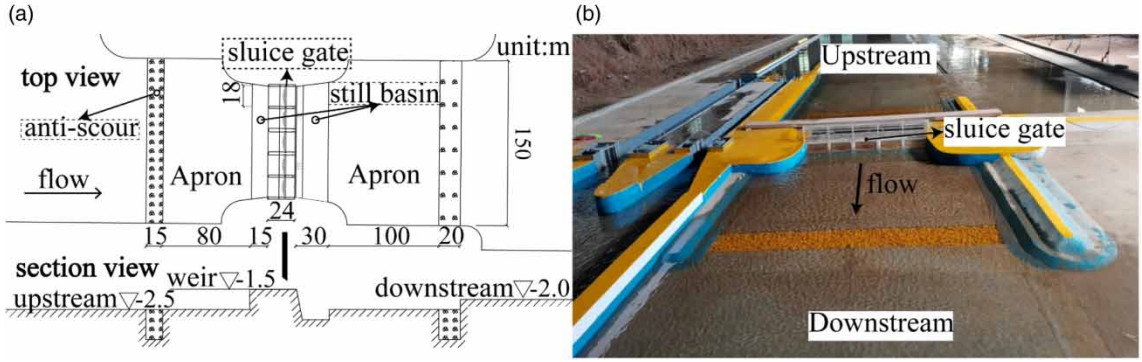


Figure 1 | Layout of the experimental model: (a) sketch of the prototype gate and (b) experimental setup.

wide range of flow rates, gate openings, and tailwater levels. Its aim was to simulate various conditions that could potentially occur during the operation of project. Each experimental group consisted of four sets of tailwater levels, four sets of flow rates, and five sets of gate openings. The tailwater levels for each set are 0.39, 2.20, 3.56, and 4.95 m. The corresponding inflow rates for each set are 200, 500, 875, and 1,100 m^3/s. The gate openings for each set are 0.3 m, 0.5 m, 1.0 m, 1.5 m, and fully open. A total of 80 experiments were collected, and 36 sets were identified as suitable for analysing the orifice flow (Table 1).

2.2. Discharge calculation

The flow calculation method employed in this study is based on the EM method. The equation utilized for quantifying the flow rate through the gate orifice for both the free and submerged conditions can be expressed as follows (Vaheddoost *et al.* 2021):

$$Q = C_d w D \sqrt{2gH} \tag{1}$$

The discharge coefficient (C_d) can be expressed using contraction coefficient (C_c) and energy loss coefficient (k) and other dimensionless terms:

$$C_d = C_c \frac{(t_1 + t_2 - ((t_1 + t_2)^2 - t_2^2 t_3)^{0.5})^{0.5}}{t_2^2} \tag{2}$$

$$t_1 = 2\left(C_c^2 \frac{w^2}{Hh} - C_c \frac{w}{H}\right) \tag{3}$$

$$t_2 = 1 + k - C_c^2 \left(\frac{w}{H}\right)^2 \tag{4}$$

$$t_3 = 1 - \frac{h^2}{H^2} \tag{5}$$

It is seen from Equations (1)–(5) that the calculation of flow discharge depends on the discharge coefficient, gate opening, net width of the gate and upstream head over the weir, and the effective parameters on the discharge coefficient consist of the contraction coefficient (C_c), the energy loss coefficient (k) and other dimensionless variables like w^2/Hh, w/h and h^2/H^2. The determination of the contraction and energy loss coefficients can be achieved by employing constrained optimization algorithms. Consequently, two calibration methods for the discharge coefficient emerge: one involves the direct calibration of the discharge coefficient, and the other entails the separate calibration of contraction and energy loss coefficients, calculating the discharge coefficient using Equation (2). This study adopts the former method to develop a calculation model of the discharge coefficient in both free and submerged flow conditions, considering the simplification of calculations and ease of utilization.

Table 1 | Description of experiments and runs

Run	Q (m^3/s)	U_0 (m/s)	H_0 (m)	h_0 (m)	H (m)	h (m)	D (m)	w (m)	h_t (m)
S-1	162	0.225	3.31	2.26	4.81	3.76	90	0.5	2.2
S-2	187	0.320	2.39	0.89	3.89	2.39	90	0.5	0.84
S-3	200	0.219	4.59	1.41	6.09	2.91	90	0.3	/
S-4	200	0.267	3.5	1.79	5	3.29	90	0.5	/
S-5	200	0.333	2.5	2.08	4	3.58	90	1	/
S-6	200	0.357	2.24	2.08	3.74	3.58	90	1.5	/
S-7	200	0.241	4.04	3.54	5.54	5.04	90	1	3.56
S-8	200	0.251	3.82	3.61	5.32	5.11	90	1.5	3.56
S-9	200	0.256	3.71	3.61	5.21	5.11	90	2	3.56
S-10	200	0.267	3.49	1.43	4.99	2.93	90	0.5	2.2
S-11	200	0.341	2.41	1.95	3.91	3.45	90	1	2.2
S-12	200	0.357	2.23	2.03	3.73	3.53	90	1.5	2.2
S-13	200	0.365	2.15	2.07	3.65	3.57	90	2	2.2
S-14	200	0.452	1.45	0.07	2.95	1.57	90	0.5	0.39
S-15	200	0.617	0.66	0.35	2.16	1.85	90	1	0.39
S-16	351	0.434	3.89	2.21	5.39	3.71	90	1	2.2
S-17	463	0.744	2.65	0.89	4.15	2.39	90	1	0.84
S-18	500	0.543	4.64	3.13	6.14	4.63	90	1.5	3.56
S-19	500	0.613	3.94	1.54	5.44	3.04	90	1	2.2
S-20	500	0.751	2.94	1.87	4.44	3.37	90	1.5	2.2
S-21	500	0.815	2.59	2.09	4.09	3.59	90	2	2.2
S-22	500	1.339	0.99	0.43	2.49	1.93	90	1.5	0.39
S-23	585	0.700	4.07	2.23	5.57	3.73	90	1.5	2.2
S-24	631	0.918	3.08	0.94	4.58	2.44	90	1.3	0.84
S-25	875	0.893	5.03	1.79	6.53	3.29	90	1.5	2.2
S-26	875	1.162	3.52	2.18	5.02	3.68	90	2	2.2
S-27	875	0.854	5.33	3.23	6.83	4.73	90	2	3.56
S-28	1,100	1.155	4.85	3.63	6.35	5.13	90	3	3.56
F-1	500	0.697	3.28	−0.1	4.78	1.4	90	1	0.39
F-2	875	0.852	5.35	−0.1	6.85	1.4	90	1.5	0.39
F-3	875	1.311	2.95	0.33	4.45	1.83	90	2	0.39
F-4	1,100	1.497	3.4	0.68	4.9	2.18	90	2.5	0.39
VS-1	461	1.229	1	0.49	2.5	1.99	90	1.46	0.39
VS-2	653.3	0.831	3.74	2.24	5.24	3.74	90	1.65	2.2
VF-1	653.3	1.015	2.79	0.94	4.29	2.44	90	1.46	0.84
VF-2	673	0.963	3.16	0.5	4.66	2	90	1.46	0.39

Note: Runs S1-S28 and F1-F4 correspond to submerged and free flow, respectively; VS1-VS2 and VF1-VF2 correspond to submerged and free flow used for validation, respectively. Q is discharge; U_0 is upstream velocity; w is gate opening; H_0 is upstream water level; h_0 is downstream water level; H is the upstream head over the weir; h is the downstream head over the weir; D is the net width of the gate; h_t is tailgate water level.

2.3. Regression analysis

The regression analysis method involves extracting initial features from the physical processes of flow through the sluice gate, constructing of high-dimensional features, utilizing of correlation analysis, and best subset regression. Correlation analysis was chosen to streamline the feature space, ultimately enhancing reliability of the model and reducing the computing

time. The final step entails constructing an optimal discharge coefficient model using the best subset regression approach. The stepwise procedure is outlined as follows:

(1) The selection of the initial features is based on the physical significance of the discharge coefficient equation, specifically focusing on the energy loss coefficient and the contraction coefficient. The Pi-theorem was employed to establish dimensionless features. Furthermore, correlation analysis and collinearity diagnosis were performed to identify relationships among the features and streamline the feature space.
(2) The dimensionless features were then utilized to construct high-dimensional features, which include higher-order and interaction terms. Subsequently, the correlation coefficients were computed between new features and the discharge coefficient. The discharge coefficient is the dependent variable, while the other features are independent variables. If the absolute value of the correlation coefficient between features is greater than 0.9, the feature which has the higher correlation with the discharge coefficient would be retained, while the remaining features would be removed.
(3) The regression model was derived using a best subset regression, which involves considering all possible combinations of predictor variables. Various criteria such as the Akaike information criterion and R^2 number were employed to assess the quality of each model. These criteria were computed for each model, and one with the most favourable values was selected as the optimal discharge coefficient model. Furthermore, computed discharge values and residuals were compared with the findings of previous studies to evaluate the accuracy and practicality of the proposed method.

2.4. Theory background

2.4.1. Preprocessing

Preprocessing the raw data can improve data quality and enhance a model performance (Habib *et al.* 2023; Yeganeh-Bakhtiary *et al.* 2023). Data preprocessing typically involves data cleaning and feature engineering, aiming to improve the predictive performance of features and provide faster and more cost-effective features that help understand the underlying model generation process. In this study, initial features are extracted based on existing research and physical processes, and dimensionless features were constructed using theorems.

2.4.2. Correlation analysis

The Pearson correlation coefficient is used to assess the linear relationship and directionality between features and to avoid collinearity issues. The formula for the Pearson correlation coefficient r is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{Y})^2}} \tag{6}$$

where x_i, y_i represent the feature; n is the sample size; $\bar{X}$, $\bar{Y}$ represent the sample means of the two variables.

The correlation coefficient, denoted as r, is a measure that ranges from -1 to 1. A value closer to 1 or -1 indicates a stronger linear relationship between the variables, while a value closer to 0 suggests a weak linear relationship. This allows for the identification and removal of inefficient features that exhibit high linear correlation.

2.4.3. High-dimensional feature

In accordance with the polynomial theory, the construction of high-dimensional features involves the inclusion of higher-order terms and interaction terms. Polynomial theory is of great significance in deterministic systems for finding optimal equations and simplifying the structural transformation of nonlinear systems. Also, it is one of the fundamental theories of deep learning (Oh & Pedrycz 2002). Typically, third- or fourth-order iterations are performed in a quadratic form, and the number of iterations required to obtain the optimal solution is related to the data distribution, the dimensionality of the function, and the complexity. The formula is as follows:

$$f = \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_i x_j \tag{7}$$

where x_i, x_j represent features; n is the sample size; b_i, c_{ij} represent polynomial coefficient.

2.4.4. Best subset regression

The best subset regression method is a commonly utilized approach in the field of predictive modeling. It entails the creation of models for every possible combination of features and the subsequent selection of the most effective model based on its predictive performance. However, the computing demands of the method grow exponentially as the number of features increases, approximately $2^n - 1$ times. Due to the exponential increase in computing time for best subset regression with an increasing number of features, it is necessary to perform a correlation analysis to identify and remove inefficient features that exhibit high linear correlation. By doing so, a more efficient regression analysis can be performed on the remaining subset of variables. This will help reduce the computational burden and improve the efficiency of the regression analysis on the entire subset.

2.4.5. Performance evaluation of models

In the process of model production, the Akaike Information Criterion (*AIC*), the condition number (*Cond.No*), coefficient of determination (R^2), and the root mean square percentage error (*RMPSE*) were set as evaluation criteria.

The *AIC* is a criterion established by the Japanese statistician Akaike (Akaikei 1973), based on the concept of information entropy, to evaluate model complexity and goodness of fit. The *AIC* can be expressed as follows:

$$\mathrm{AIC} = 2K - 2\ln(L) \tag{8}$$

where K is the number of parameters to be estimated in the model; L is the likelihood function.

AIC is a measure that penalizes the number of parameters in a model by introducing a penalty term of $2K$. It is used to control the risk of overfitting when the differences in likelihood functions are not significant. A lower *AIC* value indicates a lower likelihood of overfitting. In general, the model with the minimum *AIC* value among multiple models is considered to have the best fit, and this approach is known as the minimum *AIC* estimation method.

Cond.No is an information criterion employed to quantify the presence of multicollinearity within a set of features. A large condition number signifies disparities among the singular values of the coefficient matrix, thereby indicating unstable regression coefficients and a pronounced multicollinearity issue. It is widely acknowledged that a smaller condition number corresponds to a lesser multicollinearity. The condition number can be expressed by:

$$\mathrm{cond}(X) = \left\|A^{-1}\right\| \|A\| \tag{9}$$

where $\|A\|$ is Matrix paradigms for coefficient matrices.

R^2, also known as the coefficient of determination, is a descriptive statistical measure that quantifies the goodness of fit of a model to the observed data. It represents the proportion of the variability in the dependent variable that can be explained by the model. The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{10}$$

where y_i denotes the observed values; $\hat{y}_i$ represents the fitted values; $\bar{y}_i$ represents the mean value of the observed values.

The R^2 statistic serves as a measure of the model's ability to account for the variation in the observed data. The range of R^2 is between 0 and 1. When R^2 is closer to 1, it indicates that the model can effectively explain the variability in the observed data, demonstrating high feature coverage and goodness of fit.

RMPSE is a metric used to measure the prediction error of a model. It is calculated using the following formula:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{y}_i} - 1\right)^2} \cdot 100\% \tag{11}$$

where y_i denotes the observed values; $\hat{y}_i$ represents the fitted values.

2.5. Selection of parameters and dimensional analysis

During the process of feature selection, it is crucial to conduct the evaluations and selections that accurately capture the intrinsic characteristics of the hydraulic system. The model's accuracy, interpretability, and reliability can be significantly enhanced by incorporating relevant features. Therefore, it is necessary to understand the physical role played by the discharge coefficient when water flows through a gate. It is possible to select features closely related to the research objective based on the understanding of physical characteristics, thereby improving the performance of the model. It is important to ensure that these features can be generalized to other scenarios as well.

The hydraulic phenomenon arising from the flow through gate structures exhibits remarkable complexity despite simplicity of the gate structure. It is important to note that the process is not solely influenced by upstream flow conditions and gate geometry, but also connected to downstream flow conditions. Therefore, when considering the initial feature selection for the calculation model of the discharge coefficient under free-submerged flow conditions, it is important to thoroughly consider the hydraulic characteristics of both the upstream and downstream. Additionally, various factors that influence the difference between free and submerged flow regimes must be taken into account. The model can achieve a more comprehensive representation of the features involved by incorporating these considerations, ultimately leading to improved accuracy and applicability. Equations (2)–(5) demonstrate that the discharge coefficient includes various dimensionless terms, such as the energy loss coefficient, contraction coefficient, and other dimensionless parameters. The energy loss coefficient can be explained by various influential factors, including turbulence resulting from irregular fluid motion, flow resistance determined by viscosity, uneven energy dissipation due to non-uniform flow distribution, and the fluid's kinetic energy represented by velocity head. Energy loss primarily arises from the transfer of energy caused by boundary layer friction and the presence of large-scale turbulent structures. Upstream energy loss is attributed to the growth of the bottom boundary layer and turbulence in the recirculation zone. These effects are effectively characterized by Reynolds number at the gate. The loss of energy can also occur due to turbulent shear, entrainment, and jet characteristics in submerged flow condition (Habibzadeh *et al.* 2011; Castro-Orgaz *et al.* 2013). In contrast to free flow, submerged flow demonstrates diminished energy loss, thereby supporting the utilization of Reynolds number downstream as a potential means to characterize these phenomena. Therefore, it is advisable to consider Reynolds number Re_1 at the gate opening and Re_2 in the downstream channel as the initial features of energy loss coefficient k:

$$k = f(H, h, Re_1, Re_2) \tag{12}$$

The calculation method for Re_1 adopts the Equation (13) proposed by Vatankhah & Hoseini (2021), and dynamic viscosity coefficient υ is set to $1.007 \cdot 10^{-6}$ m^2/s:

$$Re_1 = w\sqrt{gH}/\upsilon \tag{13}$$

The contraction coefficient C_c varies with the relative gate opening and submergence (Belaud *et al.* 2009), lip shape, gate type, and water depth of the approaching channel. Consequently, the contraction coefficient may be influenced by geometric features such as channel width b and gate opening w, as well as the flow conditions. In empirical techniques, the downstream depth often carries greater significance. Besides the geometric factors, the contraction coefficient varies with the relative opening of the gate at different submergence conditions. The water level difference between upstream and downstream sections Δh is also utilized due to the impact of downstream in a tidal river. As a result, the statement might be interpreted as:

$$C_c = f(H, h, \Delta h, b, w) \tag{14}$$

Considering the physical parameters in Equations (2)–(5), C_d can be represented as follows:

$$C_d = f(H, h, \Delta h, b, w, Re_1, Re_2) \tag{15}$$

Among them, four terms H, h, b, e are independent features with dimensions [L]. According to the π theorem, one of these terms can be chosen as the repeating variable to construct dimensionless parameters. To better capture the relationship between features and discharge coefficients, we can select the form of dimensionless terms that is similar to those from

Equations (2)–(5). Based on feature selection of previous research (Vaheddoost *et al.* 2021), the final expression can be formulated as follows:

$$C_d = f\left(\frac{w}{H}, \frac{H}{b}, \frac{H}{h}, \frac{w}{h}, \frac{h}{b}, \frac{h}{H}, \frac{w}{\Delta h}, \frac{w}{b}, Re_1, Re_2\right) \tag{16}$$

3. RESULTS

3.1. High-dimensional features and correlation analysis

The correlation coefficients were computed for each feature with respect to the discharge coefficient as illustrated in Figure 2(a). The upper triangle of the heatmap represents the magnitude of correlation using the area of the pipe, while the lower triangle provides the values of the correlation coefficients.

A careful examination with the threshold of 0.9 on the correlation coefficient of features was conducted based on the findings presented in Figure 2(a). The feature that has the highest correlation with the discharge coefficient would be retained, while the remaining features would be removed. Subsequently, features with low correlation to the C_d were selected. As a result of assessment, w/b was excluded and the remaining dimensionless features are h/H, $w/\Delta h$, h/b, H/h, Re_2, w/h, w/H, Re_1, H/b. The higher-order terms and interaction terms were conducted based on the filtering features using Equation (7). A total of 47 features were obtained, correlation coefficients were then computed, and a part of the result is shown in Figure 2(b). In order to optimize the efficacy of the features, shorten the calculation time, ensure model coverage, and mitigate concerns about collinearity, it becomes important to pre-select a specific subset of features. This can be achieved by sequentially retaining features that exhibit the highest correlation with C_d, while simultaneously eliminating other features with correlation coefficient exceeding 0.9. The outcome of feature selection process is depicted in Figure 2(c).

Figure 2(c) illustrates the 19 remaining features: h^2/H^2, h^2/bH, $wh/b\Delta h$, wh/bH,$Re_2/\Delta h$, H/h, wh/H^2, HRe_2/h, Re_2, hRe_1/b, hH/b^2, wRe_2/h, wRe_2/H, w/H, wRe_1/H, H^2/b^2, hRe_2/H, w^2/Hh. Specific features like h^2/H^2 demonstrate a strong correlation of 0.91 with C_d. To enhance the search space optimization, the explanatory method of feature correlation selection was employed, replacing conventional heuristic methods typically utilized in general feature selection algorithms. This approach effectively mitigates the influence of the initial model and the selection order on the algorithm. As a result, the best subset regression method can be employed to conduct regression analysis.

3.2. Calculation of discharge coefficient and flow rate

The best subset regression method was utilized based on the feature selection results shown in Figure 3. *AIC*, *Cond*, R^2, *RMSPE* were calculated for each generated model. The models were then sorted based on *RMSPE* and *AIC* values in ascending order. The top three results of *AIC* and *RMSPE* are shown in Tables 2 and 3, respectively.

According to the results from Tables 2 and 3, No. 1 has the lowest values for both *AIC* and *RMSPE*, where *AIC* is −138.8 and *RMSPE* is 5.3%. Additionally, the R^2 value of No.1 is 0.964, which is the highest among all the combinations. Therefore, considering these factors collectively, No.1 is the optimal feature selection. Hence, the statement of the discharge coefficient C_d might be interpreted as:

$$C_d = f\left(\frac{H}{h}, \frac{wRe_2}{h}, \frac{hRe_1}{b}, \frac{wh}{b\Delta h}, \frac{h^2}{H^2}, \frac{wRe_1}{H}, \frac{wRe_2}{H}\right) \tag{17}$$

The regression equation of discharge coefficient can be expressed as:

$$\begin{aligned} C_d =& 0.083\frac{H}{h} - 0.106e^{-6}\frac{wRe_2}{h} - 0.179e^{-6}\frac{hRe_1}{b} + 0.183\frac{wh}{b\Delta h} \\ & - 0.413\frac{h^2}{H^2} - 0.042e^{-6}\frac{wRe_1}{H} + 0.318e^{-6}\frac{wRe_2}{H} + 0.509 \end{aligned} \tag{18}$$

The validation of Equation (18) and calculation outcome of discharge are shown in Figure 3. In Figure 3, black dots represent the observed values of the discharge coefficient, while coloured dots represent the fitted values of the discharge coefficient. The colour of the dots corresponds to the magnitude of the observed flow rate, and the size of the dots represents

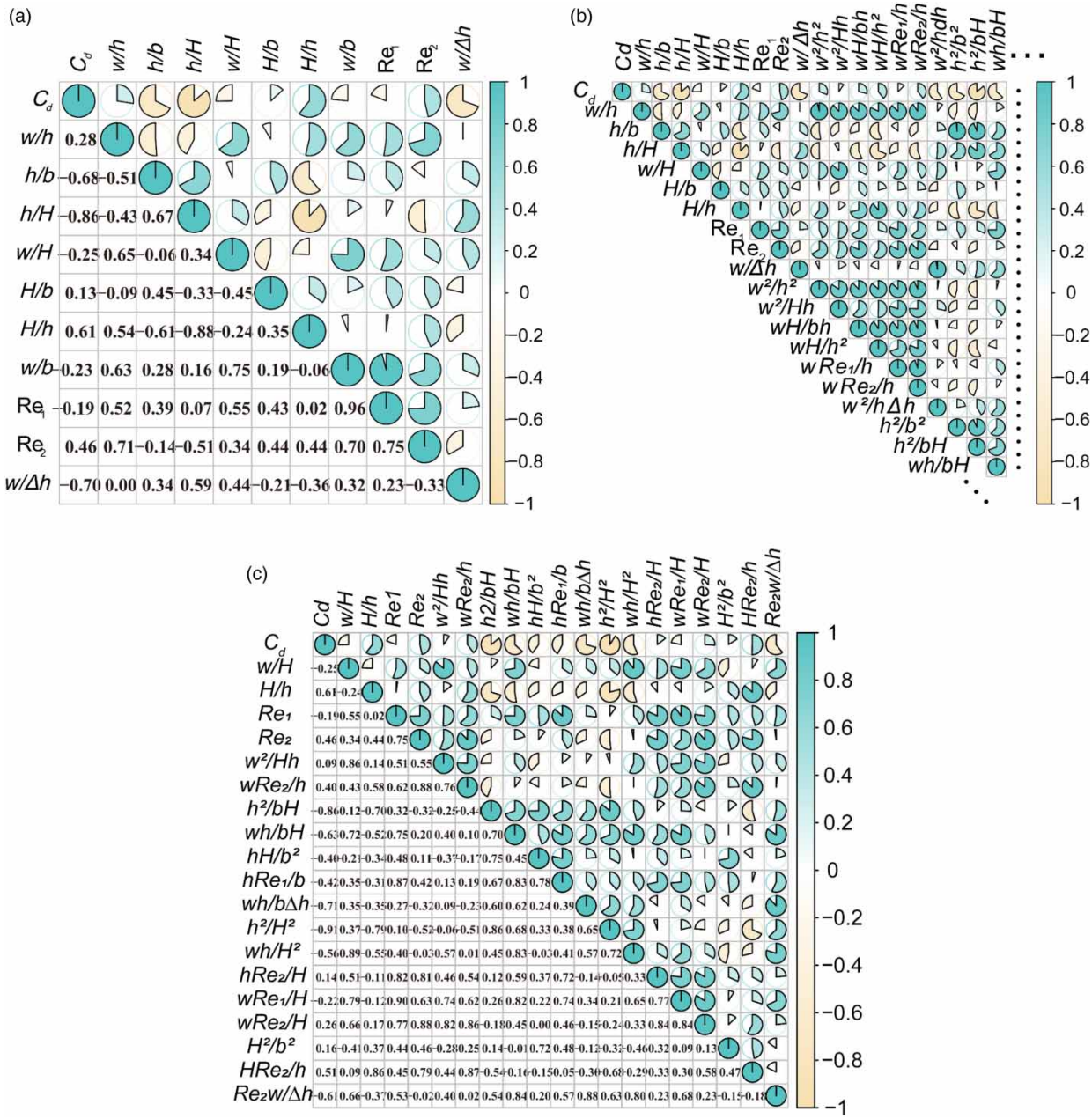


Figure 2 | Heatmap of correlation coefficient: (a) initial features; (b) part of high-dimensional features; and (c) features after selection.

the relative error in flow rate calculation. The left axis scale represents the distribution range of the discharge coefficient. Equation (18) was validated by two sets of free flow (VF-1, VF-2) and submerged flow (VS-1, VS-2) and the discharge validation result is shown in Figure 3(a). The R^2 value for the discharge is 0.991 and *RMSPE* is 5.32%, indicating a high accuracy and good fitting of equation. Other fitted results are shown in Figure 3(b)–3(d). The R^2 for discharge coefficient is 0.964, with *RMSPE* found as 5.28% and the R^2 value in discharge coefficient of free-submerged flow is 0.993 and the *RMSPE* is 5.04%. In the fitted results, 75.0% of the data have an error within 5%, and 91.7% of the data has an error within 10%. It is worth noting that when the discharge coefficient value is relatively large, errors ranging from 5 to 15% occur more frequently. The R^2 for discharge calculation is 0.950, with *RMSPE* found as 7.46% in free flow while the R^2 is 0.999 and *RMSPE* is 5.01% in submerged flow. This outcome indicates that the model can be more suitable to submerged flow than free flow, and the calculation of both flow conditions can obtain high accuracy.

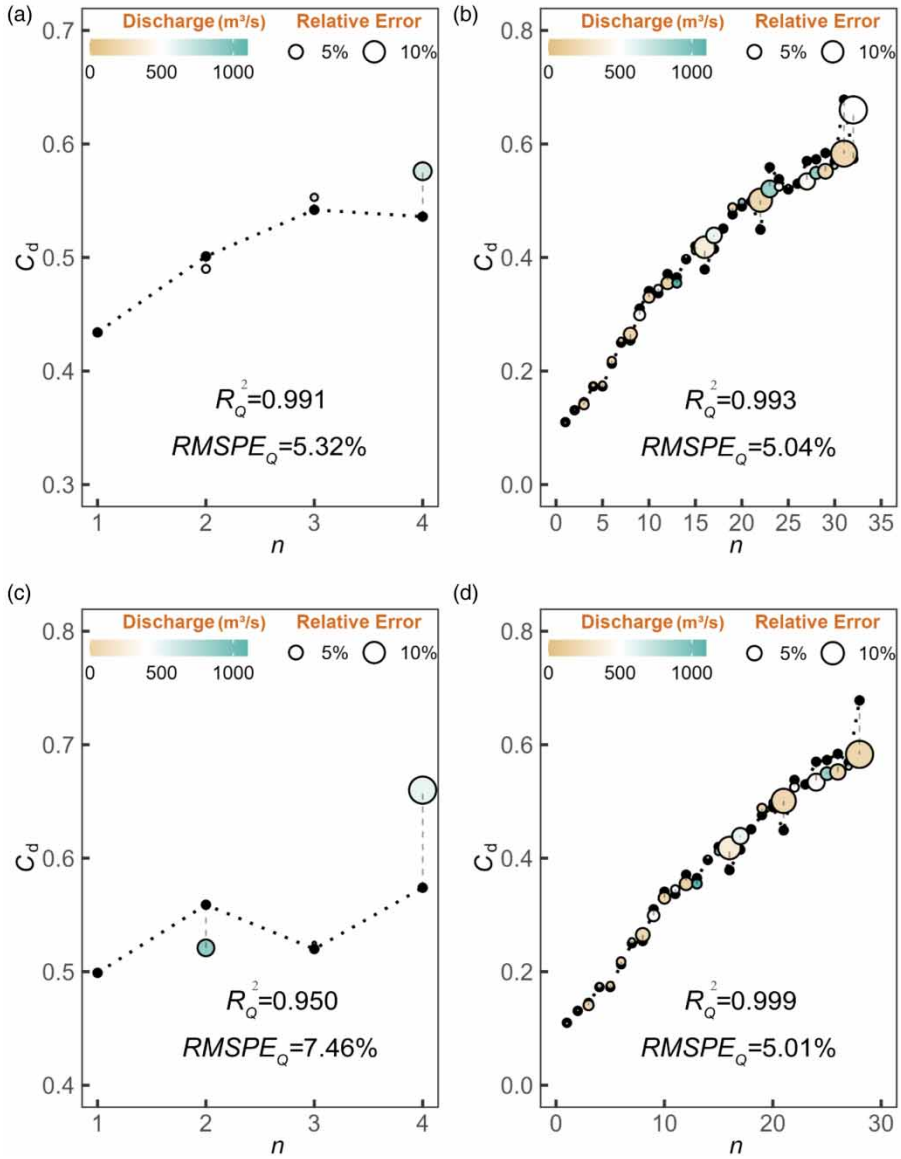


Figure 3 | Discharge calculation outcome: (a) validation; (b) free-submerged flow; (c) free flow; and (d) submerged flow.

Table 2 | AIC ranking results

No.	AIC Features	Cond.	R^2	RMSPE%
1	−138.8 $H/h, wRe_2/h, hRe_1/b, wh/b\Delta h, h^2/H^2, wRe_1/H, wRe_2/H$	96.1	0.964	5.3
2	−136.9 $wRe_2/h, hRe_1/b, h^2/H^2, wRe_2/H, HRe_2/h$	99.8	0.957	10.5
3	−136.4 $Re_2, wRe_2/h, h^2/H^2, wRe_1/H, wRe_2/H, HRe_2/h$	97.7	0.960	8.4

Table 3 | RMSPE ranking results

No.	AIC	Cond	R^2	RMSPE%
	Features			
1	−138.8	96.1	0.964	5.3
	$H/h, wRe_2/h, hRe_1/b, wh/b\Delta h, h^2/H^2, wRe_1/H, wRe_2/H$			
2	−134.9	99.1	0.960	7.0
	$H/h, wRe_2/h, hRe_2/H, wh/b\Delta h, h^2/H^2, wRe_1/H, wRe_2/H$			
3	−125.5	83.9	0.945	7.0
	$wRe_2/h, hRe_1/b, wh/b\Delta h, h^2/H^2, wRe_1/H, wRe_2/H$			

The calculation method for free flow has reached a high level of maturity, whereas research pertaining to submerged flow is still in progress. In order to evaluate the accuracy of the free-submerged flow model proposed, the present study compares recent research findings on submerged flow with the outcomes obtained from the free-submerged calculation, which is presented in Table 4. Table 4 presents the research results of different calibration methods for submerged flow in the past two years, Shakouri *et al.*(2023) utilized the experimental data from Sepúlveda and Bijankhan *et al.* in their study (Toepfer 2008; Bijankhan *et al.* 2017; Shakouri *et al.* 2023). They employed SR as the algorithm method, and the calibration approach in Scenario 1 (Sc.1) was as follows: values of k and C_c were computed using a constrained optimal algorithm, and the calibration models of k and C_c were established using SR, respectively. The C_d was then calculated using Equation (2), and the flow rate was determined using Equation (1). The results indicate that the R^2 is 0.992 and *RMSPE* is 5.71% in the discharge calculation. For Scenario 2, the method was direct calibrating the C_d coefficient. The flow rate was then calculated using Equation (1). The final results produced a value of 0.991 for R^2 and a value of 5.82% for *RMSPE*. Vaheddoost *et al.* (2021) utilized the experimental data from Sepúlveda and Rajaratnam and employed the adaptive spline method for k calibration (Rajaratnam & Subramanya 1967; Toepfer 2008; Vaheddoost *et al.* 2021). Based on the calibrated k values, the value of C_c was calculated using an empirical relationship formula, followed by computation of the discharge coefficient and flow rate. The R^2 was found to be 0.986, and the *RMSPE* was 3.10%.

In this study, a direct calibration of discharge coefficient in both free and submerged flow was employed. This method resulted in a value of 0.993 for R^2, which is the highest among the four methods considered, and the *RMSPE* is 5.04%, which is slightly lower than the results obtained for the second scenario in Vaheddoost's study. Overall, the proposed model in this study demonstrates relative stability in terms of error accuracy, with a few data points exhibiting relative errors that exceeded 10%. The current state of affairs presents opportunities for further advancements. In fact, the calculation of submerged flow obtained a value of 0.999 for R^2 and 5.01% for *RMSPE*. The result demonstrates that the model is quite suitable to submerged gate flow with a broad-crested weir. Although the R^2 of free flow is only 0.950, the *RMSPE* is 7.46%

Table 4 | Comparison of discharge calculation methods in recent two years

Flow condition	Method	R^2	*RMSPE, %*	Details
Free-submerged	This study	0.993	5.04	C_d: Calibration Q: Equation (1)
Free		0.950	7.46	
Submerged		0.999	5.01	
Submerged	Shakouri *et al.* (2023) Sc.1	0.992	5.71	k, C_c: Calibration C_d: Equation (2) Q: Equation (1)
Submerged	Shakouri *et al.* (2023) Sc.4	0.991	5.82	C_d: Calibration Q: Equation (1)
Submerged	Vaheddoost *et al.* (2021) Sc.1	0.986	3.10	k, C_c: Calibration C_d: Equation (1) Q: Equation (1)

within the permitted error extent. Besides, the number of terms in Equation (18) is not large, and the mapping relationships are relatively simple. The process of feature selection and calculation is not complicated and has the potential for automated calibration. Therefore, it can be concluded that the proposed method for calculating the free-submerged flow discharge through the gate with a broad-crested weir is not only conducive to practical applications but also demonstrates a high level of computational accuracy.

4. DISCUSSION

The discharge coefficient is determined by initial features h/H, $w/\Delta h$, h/b, H/h, Re_2, w/h, w/H, Re_1, H/b which were then utilized to construct high-dimensional features. As a result of knowing C_d function, the flow discharge can be determined by Equation (1) in the vertical sluice gate with broad-crested weir for both submerged and free flow circumstances. The evolution and innovation of algorithms have greatly contributed to the improvement of accuracy in flow calculations. Simultaneously, the subcritical phenomena and air entrainment regarding flow through gates has remained an enduring and classical area of study in modern hydraulic problems. However, the generalization ability of model merits more research in dimensionless parameter or functional relationship.

During the process of comparing with research results from the past two years, Equations (18) is not suitable to data collected by Shakouri *et al.* (2023) and Vaheddoost *et al.* (2021), and similarly, the discharge coefficient equations produced by Shakouri *et al.* (2023) and Vaheddoost *et al.* (2021) are not suitable to experimental data in this paper. One of the reasons is a significant disparity in the magnitude of the ratio of gate opening and channel width (w/b). The disparity arises from the differences in geometric features between irrigation project and shipping hubs, or other hydraulic structure. Therefore, the method of constructing dimensionless parameters deserves further improvement to extract the geometric features of different types of hydraulic engineering.

The algorithm and formula form need to be carefully chosen from the perspective of application, and the EM method requires further improvement. The formula of C_d expressed by C_c and k is quite complicated in the EM method and the EM method has its own condition of application. The calculation process for C_d is overly complicated based on the algorithm-derived calibration of C_c and k. The multitude terms and intricate mapping relationship would limit the generalization ability and reliability of the formula, although it can achieve a relatively high level of accuracy in discharge prediction.

The results obtained in this study are limited to a single experimental dataset. To establish a more reliable model, it is suggested that new experiments be conducted to expand the range of the data. Application of big data, reasonable feature engineering and evolutionary machine learning algorithms are recommended as a future research direction in terms of modeling.

5. CONCLUSIONS

The accurate calculation of flow rate is crucial for a broad range of applications, and the sluice gate with broad-crested weir is one of the most commonly used structures in flow control and regulation. While there are well-established methods for calculating the flow rate under free-flow conditions and extensive research on submerged flow conditions, there is a lack of methods that can be applied to both free and submerged flow conditions. In this study, we conducted physical model experiments and developed a discharge coefficient calculation model using the EM method, construction of high-dimensional features, correlation analysis, and best subset regression. The analysis yielded the following conclusions:

(1) This paper provides a discharge measurement model for both free and submerged flow in sluice gates with a broad-crested weir, which is based on the direct calibration of discharge coefficient. The EM method, high-dimensional feature construction, correlation analysis, and best subset regression were utilized in the model.
(2) The downstream channel Reynolds number and the water level difference between upstream and downstream were introduced in the model as initial features to enhance the information coverage for explaining the discharge coefficient. The improvement of feature selection is reflected in the final calculation equation, which means the two dimensionless parameters can be considered as effective features in the future research.
(3) The results of the model indicate a high level of accuracy. The R^2 for discharge coefficient is 0.964 and the *RMSPE* is 5.30%. In discharge calculation, the R^2 reaches 0.993 and the *RMSPE* is 5.04% for free-submerged flow. Considering the single flow regime, the calculation of submerged flow obtained a value of 0.999 for R^2 and 5.01% for *RMSPE*,

while the R^2 of free flow is 0.950 and the *RMSPE* is 7.46%. The error of free flow prediction is not good as submerged flow but is still in permitted error extent.

(4) The process of feature selection and calculation is not complicated and has potential for automated calibration. The accuracy of the model is comparable to recent studies on submerged flow in the past few years.

(5) The results obtained in this study are constrained to a single data set. More experiments on free flow can be conducted to establish a more reliable free-submerged flow model. New data sets about different types of hydraulic engineering with different geometric features can be collected to enhance the reliability and generalization ability of the model. Besides, feature engineering and evolutionary algorithms are recommended as the directions for future research.

FUNDING

The work presented in this paper is financially supported by the National Key R&D Program of China (Grant No. 2023YFC3209501) and National Natural Science Foundation of China (Nos U2040215 and 52079080).

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.

REFERENCES

Akaikei, H. 1973 Information theory and an extension of maximum likelihood principle. In: *Proc. 2nd Int. Symp. on Information Theory*, pp. 267–281.

Belaud, G., Cassan, L. & Baume, J.-P. 2009 Calculation of contraction coefficient under sluice gates and application to discharge measurement. *Journal of Hydraulic Engineering* **135** (12), 1086–1091.

Belaud, G., Cassan, L. & Baume, J.-P. 2012 Contraction and correction coefficients for energy-momentum balance under sluice gates. In: *Presented at the World Environmental and Water Resources Congress 2012*. American Society of Civil Engineers, pp. 2116–2127.

Bijankhan, M., Kouchakzadeh, S. & Belaud, G. 2017 Application of the submerged experimental velocity profiles for the sluice gate's stage-discharge relationship. *Flow Measurement and Instrumentation* **54**, 97–108.

Cassan, L. & Belaud, G. 2012 Experimental and numerical investigation of flow under sluice gates. *Journal of Hydraulic Engineering* **138** (4), 367–373.

Castro-Orgaz, O., Mateos, L. & Dey, S. 2013 Revisiting the energy-momentum method for rating vertical sluice gates under submerged flow conditions. *Journal of Irrigation and Drainage Engineering* **139** (4), 325–335.

Donnelly, J., Abolfathi, S., Pearson, J., Chatrabgoun, O. & Daneshkhah, A. 2022 Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model. *Water Research* **225**, 119100.

Donnelly, J., Daneshkhah, A. & Abolfathi, S. 2024 Physics-informed neural networks as surrogate models of hydrodynamic simulators. *Science of The Total Environment* **912**, 168814.

Ferro, V. 2000 Simultaneous flow over and under a gate. *Journal of Irrigation and Drainage Engineering* **126** (3), 190–193.

Ferro, V. 2018 Testing the stage-discharge relationship of a sharp crested sluice gate deduced by the momentum equation for a free-flow condition. *Flow Measurement and Instrumentation* **63**, 14–17.

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

Habibzadeh, A., Vatankhah, A. R. & Rajaratnam, N. 2011 Role of energy loss on discharge characteristics of sluice gates. *Journal of Hydraulic Engineering* **137** (9), 1079–1084.

Henderson, F. M. 1966 *Open Channel Flow*. Macmillan, New York.

Khosravi, K., Khozani, Z. S., Melesse, A. M. & Crookston, B. M. 2022 Intelligent flow discharge computation in a rectangular channel with free overfall condition. *Neural Computing & Applications* **34** (15), 12601–12616.

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

Oh, S.-K. & Pedrycz, W. 2002 The design of self-organizing polynomial neural networks. *Information Sciences* **141** (3), 237–258.

Rajaratnam, N. & Subramanya, K. 1967 Flow equation for the sluice gate. *Journal of the Irrigation and Drainage Division* **93** (3), 167–186.

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 Science & Technology: Water Supply* **21** (1), 232–248.

Shakouri, B., Ismail, I. & Safari, M. J. S. 2023 Energy loss and contraction coefficients-based vertical sluice gate's discharge coefficient under submerged flow using symbolic regression. *Environmental Science and Pollution Research* **30** (31), 76853–76866.

Toepfer, C. S. 2008 *Instrumentation, Model Identification and Control of an Experimental Irrigation Canal* (Doctoral Thesis). *TDX (Tesis Doctorals en Xarxa)*, Universitat Politècnica de Catalunya.

Vaheddoost, B., Safari, M. J. S. & Ilkhanipour Zeynali, R. 2021 Discharge coefficient for vertical sluice gate under submerged condition using contraction and energy loss coefficients. *Flow Measurement and Instrumentation* **80**, 102007.

Vatankhah, A. R. & Hoseini, P. 2021 Discharge equation for round gates in turnout pipes: Dimensional analysis and theoretical approaches. *Journal of Irrigation and Drainage Engineering* **147** (1), 06020015.

Wóycicki, K. 1931 *Wassersprung, Deckwalze und Ausfluss Unter Einer Schütze. Doctoral Thesis*, ETH Zurich.

Yeganeh-Bakhtiary, A., EyvazOghli, H., Shabakhty, N. & Abolfathi, S. 2023 Machine learning prediction of wave characteristics: Comparison between semi-empirical approaches and DT model. *Ocean Engineering* **286**, 115583.

First received 14 December 2023; accepted in revised form 18 March 2024. Available online 29 March 2024