Environ. Water Eng., 2025, 11(3), 316-325 https://doi.org/10.22034/ewe.2024.433409.1905

EWE

Environment and Water Engineering

Homepage: www.jewe.ir

ISSN: 2476-3683

Prediction of air concentration in stepped spillways using data-oriented methods

Kiyoumars Roushangar[1], Hamidreza Abbaszadeh[1], Reza Saadatjoo[1], and Aydin Panahi[1]

[1]Department of Civil Engineering, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran

ARTICLE INFO

Paper Type: Research Paper

Received: 31 December 2023
Revised: 25 January 2024
Accepted: 27 January 2024
Published: 04 February 2024

Keywords
Gaussian Process Regression
Sensitivity Analysis
Stepped Spillway
Support Vector Machine

***Corresponding author:**
K. Roushangar
✉ kroshangar@yahoo.com

ABSTRACT

The high flow velocity and the pressure reduction in the spillways cause damage to the spillways. In the present research, the application of Gaussian Process Regression (GPR) and Support Vector Machine (SVM) was investigated for predicting air concentration in stepped spillways. For this purpose, a comprehensive set of available experimental data obtained from hydraulic models of stepped spillways in the modeling process was utilized. Input models were defined based on various combinations of measured parameters. In predicting the air concentration in the stepped spillway under natural aeration conditions, parameters of discharge (q_w), the ratio of flow depth (normal to spillway step) to channel width (Z/W), the ratio of longitudinal distance from the beginning of the step to the length of the step (x/L), and the ratio of distance from the midpoint line of the spillway step to the step width ($Y=2y/w$) had a significant impact. The results obtained demonstrate the high capability of both methods in estimating the required air concentration on spillways. The results revealed that the Radial Basis Function (RBF) kernel performs favorably. The R^2, DC, and RMSE for the GPR were 0.79, 0.79, and 0.12, respectively, and in the SVM were 0.86, 0.86, and 0.098, respectively.

Highlights

- SVM and GPR accurately predict air concentration in spillways.
- RBF kernel in SVM outperforms other kernels for air concentration modeling.
- Key variables: flow depth-to-width ratio and longitudinal distance ratio.
- Optimal models achieve high R^2 (>0.86) and low RMSE (<0.12).
- Sensitivity analysis highlights critical parameters for spillway design.

How to cite this paper:
Roushangar, K., Saadatjoo, R., Abbaszadeh, H., & Panahi, A. (2025). Prediction of air concentration in stepped spillways using data-oriented methods. *Environment and Water Engineering*, *11*(3), 316-325. https://doi.org/10.22034/ewe.2024.433409.1905

1. Introduction

One of the most critical hydraulic structures ensuring dam safety during flood inflow and its subsequent discharge downstream is the spillway. Due to the long-standing use of hydraulic modeling in spillway design and the considerable accuracy of the results obtained, engineers have continued to rely on physical models and experimental testing to achieve more economical and safer spillway designs. In recent decades, with advancements in technology, the construction of storage dams and associated hydraulic structures has significantly increased. This trend, coupled with the rising number of high dams, has emphasized the need for safer and more cost-effective spillway designs, drawing increased attention from hydraulic engineers. In high-head spillways, the combination of high flow velocities and pressure drops substantially raises the risk of cavitation, potentially leading to severe damage. To mitigate this, specific structures known as chute aerators must be installed in locations where natural aeration from the free surface is insufficient to protect the flow boundaries.

One of the secondary effects of cavitation in many hydraulic structures is noise and vibration, which can affect everything from drive machinery to large valves in industrial and hydraulic systems such as spillways (Süme et al., 2024). Rahmeyer (1981) examined cavitation-induced noise, highlighting its disruptive nature. He emphasized that noise should be considered a critical factor in the design and operation of hydraulic structures, as it may impose limitations. Franc and Michel (2006) found that sustained cavitation decreases the performance of hydraulic structures. Local

roughness on boundaries such as concrete surfaces in spillway channels can cause flow separation and lead to pressure drops. Research by Pfister and Hager (2010) demonstrated that the highest air concentration occurs at the spillway floor up to the jet impact point. Their findings also indicated that air concentration decreases beyond the impact point, creating a significant aeration gradient in that region. Pfister (2011) investigated the hydraulic characteristics of two-phase flow in aerated spillways. Changes in aeration coefficient, jet length, and air concentration distribution were found to depend on variables such as deflector angle and sub pressure within the cavity. The study showed that steeper deflectors performed better in introducing air into the flow. Moreover, an increase in cavity sub pressure led to reductions in both jet length and aeration coefficient, as well as lower air concentrations on the spillway floor. Wu et al., (2011) studied the effects of geometric variables of chute aerators on the amount of water recirculating into the downstream cavity formed beyond the aerator ramp. At low Froude numbers and small chute slopes, the return flow caused by jet impact on the chute floor could fill the downstream cavity, potentially initiating cavitation. The results indicated that the amount of return water is related to the jet length, which increases with greater ramp height or steeper slope. Chakib (2013) examined the initiation point of natural aeration from the free surface in stepped spillways. The findings showed that air entrainment, which reduces wall friction, increases flow velocity, and shifts the initiation point of aeration downstream with increasing flow rate. Salmasi et al. (2021) explored various physical models of stepped spillways and the effects of intense water-air mixing on dissolved oxygen concentrations. The results revealed that as discharge increases and all steps become submerged, the influence of steps on turbulence and aeration diminishes. However, for discharges greater than 0.15 m²/s (per unit channel width), steeper spillways improved aeration efficiency. Raza et al., (2021) numerically investigated the effect of stepped spillway slope on air bubbles and their initiation points. They found that the length of the non-aerated flow region increases when the slope transitions from steep to mild. Ghaderi and Abbasi (2022) studied the impact of step geometry modifications including simultaneous changes to both the step surface and edges on flow patterns, aeration onset, hydraulic jump characteristics downstream, and energy dissipation across different spillway models. They concluded that flow interaction with obstacles influences the aeration onset point, shifting it upstream on the stepped spillway. One of the most critical aspects of cavitation analysis is determining the distribution of air concentration. Existing empirical relationships for estimating air concentration are generally limited to specific geometries and thus have restricted applicability. Additionally, constructing physical models of hydraulic structures is expensive and poses challenges in scaling data from laboratory to prototype conditions. Therefore, approaches that can provide reliable estimations of air concentration distribution are of great importance (Roushangar et al., 2024a).

In this study, Support Vector Machine (SVM) and Gaussian Process Regression (GPR) methods were employed to estimate the target variables. Various models were developed using different sets of input variables to predict air concentration, and the influential parameters in each case were identified. This research is an extension of previous studies and aims to demonstrate the applicability of SVM and GPR methods in estimating the distribution of air concentration. Additionally, it evaluates the effectiveness of these approaches using available experimental data in this field.

2. Materials and Methods

2.1 Support Vector Machine

Support Vector Machine (SVM), as a supervised learning method, was first introduced by Vapnik (1995) for classification and prediction tasks. SVM is an efficient learning system based on the theory of constrained optimization, utilizing the principle of structural risk minimization to achieve a globally optimal solution. In SVM regression models, a function related to the dependent variable Y, which itself is a function of several independent variables X is estimated. Similar to other regression problems, it is assumed that the relationship between the independent and dependent variables can be represented by an algebraic function $f(x)$ plus a noise term (allowable error ε), as described in Eqs. 1 and 2 (Norouzi et al., 2021).

$$A = f(x) = W^T\phi(x) + b \quad (1)$$

$$Y = f(x) + noise \quad (2)$$

where, W denotes the weight vector, b is the bias term, and φ is the kernel function, the objective is to determine the functional form of $f(x)$. This is achieved by training the SVM model using a set of samples (training set). The regression SVM function can be expressed in the form of Eq. 3.

$$f(x) = \sum_{i=1}^{N} \bar{a}_i \emptyset(X_i)^T \emptyset(X) + b \quad (3)$$

where a_i denotes the average Lagrange multipliers. Since computing $\varphi(X)$ in its feature space may be highly complex, the common practice in SVM regression is to select an appropriate kernel function. The choice of kernel in SVM depends on the size of the training dataset and the dimensionality of the feature vector. In other words, a kernel function should be selected based on these factors to ensure effective training for the given input space. In practice, four common types of kernels are typically employed: linear, polynomial, sigmoid, and radial basis function (RBF) kernels, as represented in Eqs. 4 to 7 (Hassanzadeh & Abbaszadeh, 2023; Abbaszadeh et al., 2024).

$$K(X_i, X_j) = (X_i, X_j) \quad (4)$$

$$K(X_i, X_j) = \left(1 + (X_i, X_j)\right)^d \quad (5)$$

$$K(X_i, X_j) = tanh\left(-a(X_i, X_j) + C\right) \quad (6)$$

$$K(X_i, X_j) = exp(-\|X - X_i\|^2/\sigma^2 \quad) \quad (7)$$

where, $K(X_i, X_j)$ is the covariance or kernel function calculated at the points X_i and X_j. a, C, d and σ represent the kernel functions. d is the degree of the polynomial and C is a positive integer that determines the penalty when a model training error occurs.

2.2 Gaussian process regression

In probability theory, a Gaussian Process (GPR) is a stochastic process defined as a collection of random variables, any finite

subset of which follows a multivariate Gaussian (normal) distribution (Rasmussen & Williams, 2006). In this process, the random variables are typically indexed by time or space. Therefore, if $\{X_i;\ t \in T\}$ is a stochastic sequence, then every finite collection $X_{t1}, \ldots, {}_{tk}=(X_{t1},\ldots,X_{tk})$ has a joint multivariate Gaussian distribution. In other words, in a Gaussian process, any linear combination of a finite number of variables in the process is normally distributed. The index set T is generally an infinite set representing time or spatial domains. In this study, a custom-written program in MATLAB using the M-file format was utilized for modeling the data based on the Gaussian Process Regression approach.

2.3 Air concentration distribution models

To evaluate the performance of Support Vector Machine (SVM) and Gaussian Process Regression (GPR) methods in predicting the distribution of air concentration in spillways, experimental data from Toombes (2002), comprising 3,232 data points, were utilized. These data were obtained from a physical model of a stepped spillway constructed at the University of Queensland. Toombes' investigation focused on two groups of steps specifically the second and ninth steps where the concentration of undissolved air was measured at time intervals of 5 and 30 seconds, respectively. A schematic view of the model is shown in Fig. 1, and the range of hydraulic and geometric characteristics of the model is presented in Table 1.

Fig. 1 Schematic of hydraulic model (Toombes 2002)

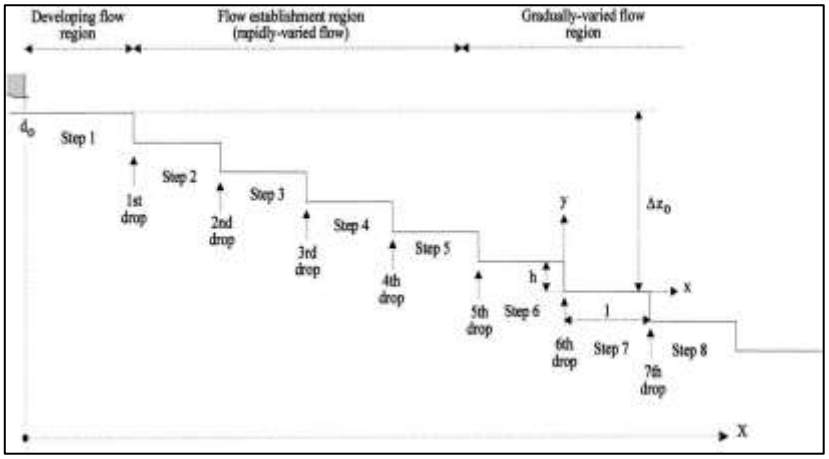


Table 1 Hydraulic and geometric characteristics of the models (Toombes 2002)

Experiment	q_w (m^3/s)	Step	W (m)	Y (-)	X (m)	Z (mm)	C
1	0.15	2	0.5	0	0-0.24	0.75-176.70	0-1
2	0.15	2	0.5	0.7	0-0.24	0.75-171.70	0-1
3	0.15	2	0.5	0.99	0-0.24	0.75-176.70	0-1
4	0.15	2	0.5	0	0-0.24	0.75-211.70	0-1
5	0.13	2	0.5	0	0-0.24	1.70-211.70	0-1

Table 2 The defined models

Scenario 1 (natural aeration)		Scenario 2 (artificial aeration)	
Model	Input parameters	Model	Input parameters
S02 (I)	q_w	S09 (I)	q_w
S02 (II)	q_w, Z/W	S09 (II)	q_w, Z/W
S02 (III)	q_w, Y	S09 (III)	q_w, x/L
S02 (IV)	q_w, x/L	S09 (IV)	q_w, Z/W, x/L
S02 (V)	q_w, Z/W, Y		
S02 (VI)	q_w, Z/W, x/L		
S02 (VII)	q_w, x/L, Y		
S02 (VIII)	q_w, Z/W, x/L, Y		

To define the input models, various combinations of measured variables were considered, including: q_w: unit discharge, Z/W: the ratio of flow depth (perpendicular to the spillway step) to channel width, x/L: the ratio of the longitudinal distance from the beginning of the step to the step length, and $Y=2y/w$: the ratio of the lateral distance from the centerline of the spillway step to the step width. The defined models used to estimate the air concentration distribution along the second and ninth steps of the spillway are presented in Table 2.

2.4 Normalization of the data

One of the essential steps in data preprocessing is normalization. Applying certain preprocessing techniques to both input and target variables can significantly improve the performance of machine learning models. Generally, using raw data may reduce the speed and accuracy of the model. Therefore, normalization is especially beneficial when the input variables exhibit wide variation, as it facilitates more efficient and effective model training. In this study, the normalization process was carried out using Eq. 8, as proposed by Roushangar et al. (2023, 2024b):

$$x_n = 0.1 + 0.9 \times \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

where X_{min} and X_{max} denote the minimum and maximum observed values, respectively, and X_n represents the

normalized data. To achieve more accurate and reliable results, the training process was repeated several times, and a data split strategy was adopted in which 75% of the data was used for training and the remaining 25% for testing.

2.5 Statistical indicators

To evaluate the performance of the applied methods in estimating the distribution of air concentration, several statistical indicators were employed, including the correlation coefficient (R), the coefficient of determination (DC), the root mean square error (RMSE), and the Kling-Gupta Efficiency index (KGE) (Amini et al., 2021; Abbaszadeh et al., 2023a,b). In general, the optimal model was identified as the one exhibiting the lowest RMSE values and the highest values of KGE, DC, and R ideally approaching one indicating both high accuracy and strong agreement between predicted and observed data (Daneshfaraz et al., 2022a,b; Norouzi et al., 2023).

$$R = \frac{\sum_{i=1}^{N}(C_{mi} - \overline{C}_m) \times (C_{pi} - \overline{C}_p)}{\sqrt{\sum_{i=1}^{N}(C_{mi} - \overline{C}_m)^2 \times \sum_{i=1}^{N}(C_{pi} - \overline{C}_p)^2}} \quad (9)$$

$$DC = 1 - \frac{\sum_{i=1}^{N}(C_m - C_p)^2}{\sum_{i=1}^{N}(C_m - \overline{C}_p)^2} \quad (10)$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(C_m - C_p)^2}{N}} \quad (11)$$

$$KGE = 1 - \sqrt{(R-1)^2 + \left(\frac{\overline{X_{Cal}}}{\overline{X_{Obs}}} - 1\right)^2 + \left(\frac{\sigma_{Cal}/\overline{X_{Cal}}}{\sigma_{Cal}/\overline{X_{Obs}}} - 1\right)^2} \quad (12)$$

where C_m denotes the measured air concentration distribution, $\overline{C_m}$ represents the mean of the measured air concentration distribution, C_p is the predicted air concentration distribution, $\overline{C_P}$ is the mean of the predicted air concentration distribution, and N indicates the number of data points. In Eq. 12, β is the ratio of the mean of the predicted data to the mean of the observed data, and γ represents the ratio of the standard deviation of the predicted values to the standard deviation of the observed values (202 et al., 2023).

3. Results and Discussion

In this study, the impact of different kernel functions on the performance of models S02 (VIII) and S09 (IV) in predicting the air concentration distribution at two distinct locations—the second and ninth steps of a stepped spillway was investigated. The models were developed using the Support Vector Machine (SVM) approach, and various kernel functions were evaluated. The results of this evaluation indicate that the Radial Basis Function (RBF) kernel exhibited the best performance in estimating the air concentration distribution. Detailed results are presented in Table 3. These findings suggest that within the studied range, the use of the RBF kernel for modeling and predicting air concentration distribution is optimal, yielding superior results compared to other kernel functions. This insight can contribute to improving the accuracy and predictive capability of similar models in fields related to air quality and environmental studies. Similar observations have been reported in previous research by Pal et al., (2014), Parsaie et al. (2018), and Hassanzadeh and Abbaszadeh (2023), where the RBF kernel demonstrated higher accuracy than Polynomial, Linear, and Sigmoid kernels in predicting target variables for various hydraulic models.

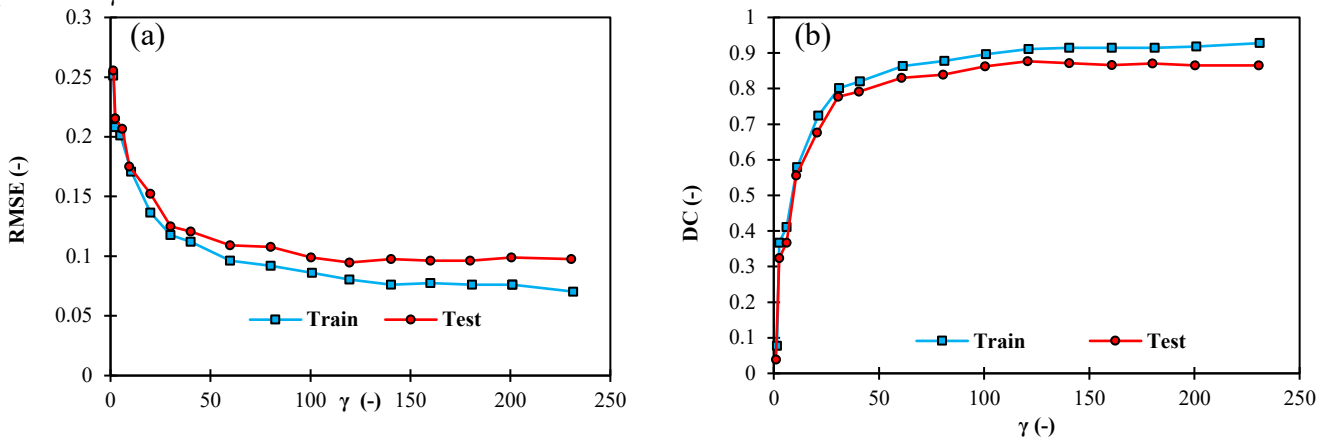
Table 3 Statistical parameters of SVM method with different kernel functions - S02 (VIII) and S09 (IV) models

Model	Kernel	Train				Test			
		R^2	DC	RMSE	KGE	R^2	DC	RMSE	KGE
S02 (VIII)	Linear	0.0125	0.0845	0.3325	0.121	0.0064	0.0627	0.3374	0.120
	Polynomial	0.0827	0.1461	0.2577	0.275	0.0963	0.1266	0.2541	0.272
	RBF	0.9201	0.9181	0.0751	0.945	0.8614	0.8598	0.0979	0.955
	Sigmoid	0.0714	0.0637	0.2707	0.058	0.0078	0.0449	0.2674	0.045
S09 (IV)	Linear	0.1421	0.1445	0.3127	0.385	0.1250	0.2524	0.3047	0.386
	Polynomial	0.0842	0.1861	0.2740	0.298	0.1054	0.0029	0.2716	0.295
	RBF	0.9871	0.9870	0.0313	0.984	0.9722	0.9718	0.0455	0.985
	Sigmoid	0.0484	0.0825	0.2845	0.245	0.0312	0.0875	0.2858	0.225

3.1 Scenario 1

Table 4 presents the results obtained from the Support Vector Machine (SVM) and Gaussian Process Regression (GPR) methods for estimating the distribution of air concentration along the second step of the stepped spillway (2,093 data points). These results demonstrate the efficiency and accuracy of each method in predicting the air concentration distribution. Comparing the outcomes of these two approaches facilitates the selection of the optimal model for air concentration modeling and prediction. Furthermore, Fig. 2 illustrates the variation of the coefficient of determination (DC) and root mean square error (RMSE) against different values of the parameter γ for the SVM method. This graph highlights the influence of the γ parameter on model accuracy and prediction quality. According to Fig. 2, the optimal value of γ in the SVM model is 200, which was consequently adopted as the best-performing parameter in this study.

Fig. 2 The graph of: a) DC and b) RMSE changes against the parameter γ under scenario 1



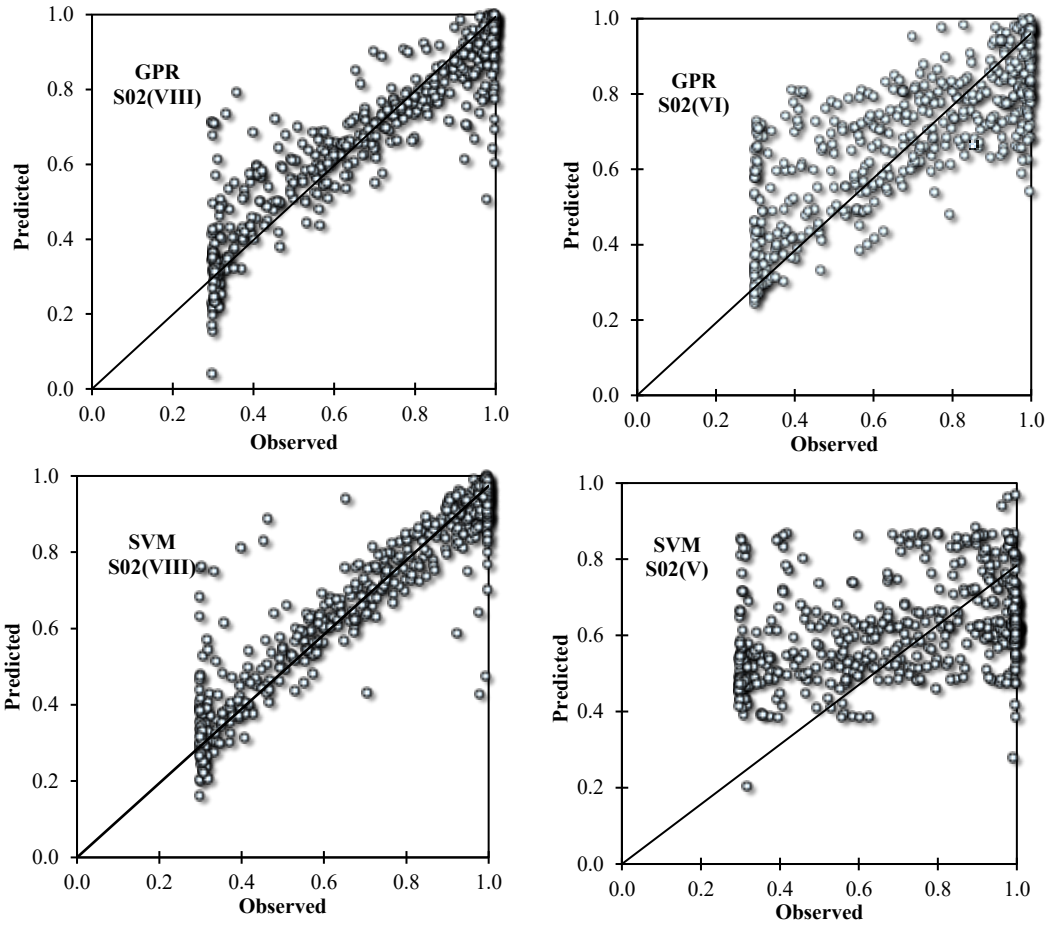
According to the results presented in Table 4, model S02 (VIII), incorporating four input variables, yielded higher values of R^2 and DC and a lower RMSE compared to the other models. These results indicate the superior performance of this model in the process of simulating air concentration distribution. Therefore, model S02 (VIII) is identified as the optimal model for this modeling task. A closer examination of the outputs from various models reveals that using a single input variable (q_w) leads to inadequate modeling results. This finding underscores the critical importance of appropriate selection and tuning of input variables in scientific modeling. Attention to these aspects is essential for achieving accurate and reliable modeling outcomes (Norouzi et al., 2021).

Table 4 The results of SVM and GPR methods to estimate air concentration under scenario 1

Model	Method	Train				Test			
		R^2	DC	RMSE	KGE	R^2	DC	RMSE	KGE
S02 (I)	SVM	0.0036	0.0237	0.2707	0.050	0.0035	0.0149	0.2674	0.058
	GPR	0.0030	0.0085	0.2625	0.058	0.0031	0.0017	0.2619	0.065
S02 (II)	SVM	0.2717	0.2193	0.2319	0.514	0.3129	0.2729	0.2231	0.558
	GPR	0.3737	0.3737	0.2077	0.584	0.4009	0.4001	0.2026	0.612
S02 (III)	SVM	0.0086	0.0842	0.3879	0.014	0.0000	0.0150	0.3850	0.0
	GPR	0.0111	0.0011	0.2633	0.050	0.0012	0.0050	0.2623	0.054
S02 (IV)	SVM	0.0020	0.0598	0.2702	0.044	0.0000	0.0060	0.2752	0.0
	GPR	0.0338	0.0337	0.2580	0.185	0.0298	0.0284	0.2579	0.164
S02 (V)	SVM	0.1991	0.0937	0.2499	0.450	0.1612	0.0172	0.2594	0.415
	GPR	0.4099	0.4100	0.2016	0.654	0.4068	0.4064	0.2016	0.645
S02 (VI)	SVM	0.4846	0.4820	0.1889	0.712	0.4429	0.4351	0.1966	0.668
	GPR	0.6668	0.6668	0.1515	0.820	0.6668	0.6668	0.1510	0.821
S02 (VII)	SVM	0.0008	0.0762	0.2723	0.027	0.0033	0.0902	0.2732	0.055
	GPR	0.0794	0.0793	0.2519	0.265	0.0275	0.0115	0.2601	0.175
S02 (VIII)	SVM	0.9201	0.9181	0.0751	0.955	0.8614	0.8598	0.0979	0.945
	GPR	0.8828	0.8827	0.0899	0.941	0.8001	0.7898	0.1200	0.905

Among models incorporating three input variables, the inclusion of the flow depth-to-channel width ratio (Z/W) alongside the flow discharge (q_w) significantly enhanced modeling accuracy, yielding R^2=0.401, DC=0.400, and RMSE=0.202. Under these conditions, model S02 (II) based on Gaussian Process Regression (GPR) outperformed the Support Vector Machine (SVM) model by approximately 47%, establishing it as the superior model. These results underscore the positive impact of the flow depth-to-channel width ratio on improving model performance in predicting the air concentration distribution. Furthermore, among three-variable input models, incorporating the longitudinal distance ratio from the step origin to step length (x/L) alongside Z/W and q_w also led to notable improvements in modeling accuracy. The scatter plots for the two best-performing models, illustrated in Fig. 3, reflect these improvements and reductions in error, demonstrating the beneficial combined effect of these three variables on the accuracy of the selected models.

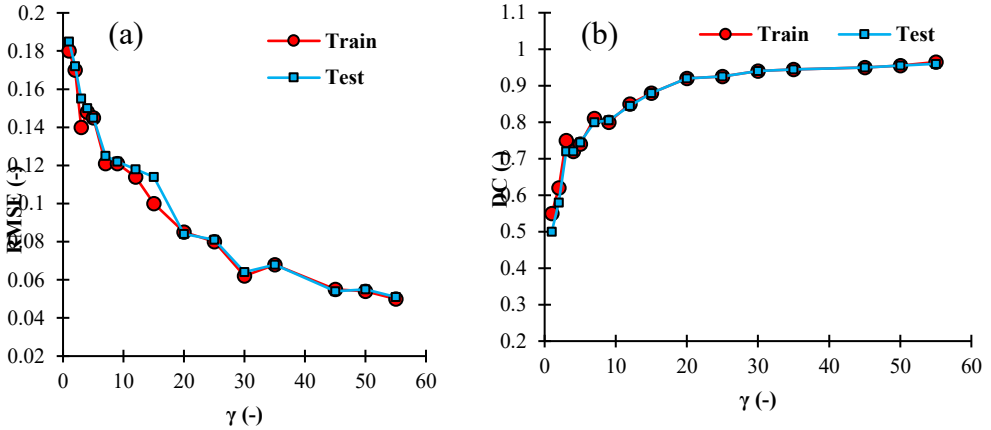
Fig. 3 Scatter plot between observed and predicted data under scenario 1 (Test data set)



3.2 Results under scenario 2

According to Fig. 4, which depicts the variations of the coefficient of determination (DC) and root mean square error (RMSE) across different values of the parameter γ for the best-performing model, the optimal γ value for the Support Vector Machine (SVM) is identified as 50. This finding indicates that the SVM, by utilizing lower values of γ resulting in reduced model complexity and shorter training time is capable of accurately modeling the required air concentration distribution with high precision. Specifically, with the optimal γ value, the model achieves R^2=0.96, DC=0.96, and RMSE=0.056. These results emphasize that appropriate selection of model parameters, particularly γ, can significantly enhance prediction accuracy and improve the SVM's ability to model more complex problems.

Fig. 4 The graph of: a) DC, and b) RMSE changes against the γ under scenario 2



According to Table 5, the results of the Support Vector Machine and Gaussian Process Regression methods for the ninth step of the stepped spillway (1,139 data points) are presented. The comparison of models S09 (II) and S09 (III) shows that substituting the variable Z/W in place of x/L leads to a notable improvement in results. In this scenario, both

SVM and GPR exhibit satisfactory performance using three input variables.

Table 5 The results of SVM and GPR methods to estimate air concentration under scenario 2

Model	Method	Train				Test			
		R^2	DC	RMSE	KGE	R^2	DC	RMSE	KGE
S09 (I)	SVM	0.0006	0.0258	0.4177	0.050	0.0061	0.0854	0.4039	0.085
	GPR	0.0006	0.0214	0.2748	0.051	0.0061	0.0011	0.2711	0.084
S09 (II)	SVM	0.5426	0.4517	0.2035	0.745	0.5030	0.4580	0.2081	0.725
	GPR	0.5746	0.5746	0.1793	0.758	0.5224	0.5225	0.1874	0.723
S09 (III)	SVM	0.0006	0.0458	0.3023	0.025	0.0034	0.0585	0.2996	0.088
	GPR	0.0447	0.0447	0.2686	0.225	0.0045	0.0650	0.2737	0.145
S09 (IV)	SVM	0.9612	0.9611	0.0542	0.984	0.9572	0.9568	0.0564	0.985
	GPR	0.9374	0.9374	0.0688	0.968	0.9289	0.9280	0.0728	0.965

The scatter plots for the two best-performing models are presented in Fig. 5. Analysis of the results obtained from various hydraulic models under different scenarios indicates that the process of modeling air concentration in stepped spillways yields results comparable to those reported by Toombes (2002).

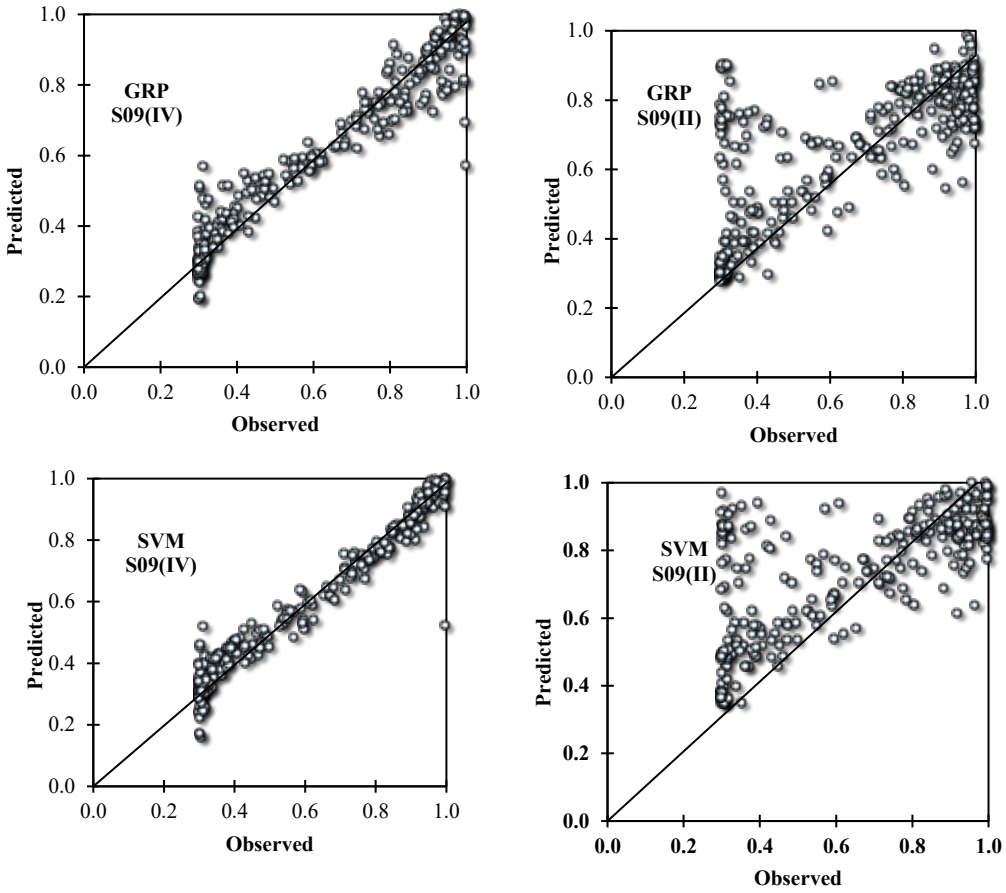


Fig. 5 Scatter plot between observed and predicted data under scenario 2 (Test data set)

3.3 Sensitivity analysis of the optimal model

To investigate the influence of different variables on estimating the distribution of air concentration, sensitivity analysis was conducted. This involved removing individual variables from the best-performing model, retraining the model using Support Vector Machine (SVM) and Gaussian Process Regression (GPR), and then evaluating the model's performance metrics (R^2, DC, RMSE, and KGE) to assess the impact of the excluded variables. The sensitivity test results for the best model under Scenarios 1 and 2 are presented in Table 6. Based on the sensitivity analysis results, the variable x/L is identified as highly significant in predicting the air concentration distribution along the spillway. Conversely, removing the variable q_w has a minor effect on the final results. Similarly, in Scenario 2 (Table 6), both x/L and Z/W are important variables for estimating air concentration along the spillway, while exclusion of q_w again has a negligible impact. Fig. 6 illustrates the effect of removing each variable on the model's performance.

Table 6 The results of the sensitivity analysis method for predicting the air concentration distribution of the best model under scenario 1 and 2

Best model	Removed parameter	Method	Train			Test		
			R^2	DC	RMSE	R^2	DC	RMSE
S02 (VIII) (Scenario 1)	q_w	SVM	0.9201	0.9181	0.0751	0.8614	0.8598	0.0979
		GPR	0.8828	0.8827	0.0899	0.8001	0.7898	0.1200
	q_w, Y	SVM	0.4846	0.4820	0.1889	0.4429	0.4351	0.1966
		GPR	0.6668	0.6668	0.1515	0.6668	0.6668	0.1510
	$q_w, \frac{x}{L}, Y$	SVM	0.2716	0.2193	0.2319	0.3129	0.2729	0.2321
		GPR	0.3737	0.3737	0.2077	0.4009	0.4001	0.2026
S09 (IV) (Scenario 2)	q_w	SVM	0.9197	0.9166	0.0794	0.9074	0.9035	0.0842
		GPR	0.9001	0.9005	0.0867	0.8851	0.8839	0.0924
	$q_w, \frac{x}{L}$	SVM	0.3880	0.0896	0.2622	0.3961	0.0465	0.2648
		GPR	0.5618	0.5617	0.1820	0.5332	0.5330	0.1853
	$q_w, \frac{z}{W}$	SVM	0	0.0017	0.2916	0.0005	0.0176	0.2867
		GPR	0.0393	0.0393	0.2694	0.0140	0.0056	0.2704

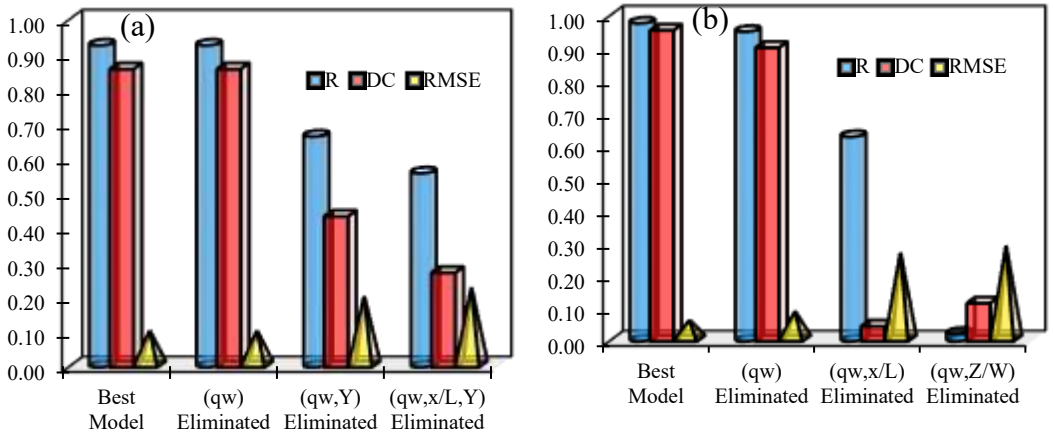


Fig. 6 Values of the sensitivity analysis test errors of the best model: a) Scenario 1, and b) Scenario 2

4. Conclusion

In the present study, an extensive dataset of experimental measurements (3,233 data points) was employed to investigate the estimation of required air concentration in stepped spillways. Multiple models based on various combinations of measured variables were developed, and the influence of each variable on the prediction of aeration concentration was assessed. The overall findings of this research are summarized as follows:

1. Both Support Vector Machine (SVM) and Gaussian Process Regression (GPR) methods demonstrated high accuracy in estimating air concentration in stepped spillways. Among different kernel functions tested within the SVM approach, the Radial Basis Function (RBF) kernel yielded the best results for predicting the required air concentration.

2. For estimating air concentration along the second step of the stepped spillway, based on 2,093 data points, model S02 (VIII) with four input variables q_w, Z/W, x/L, and Y was identified as the best-performing model, achieving R^2=0.79, DC=0.79, and RMSE=0.12 for GPR, and R^2=0.86, DC=0.86, and RMSE=0.098 for SVM.

3. For estimating air concentration along the ninth step of the stepped spillway, based on 1,139 data points, model S09 (IV) with three input variables qw, Z/W, and x/L was selected as the best model, with values of R^2=0.92, DC=0.93, and RMSE=0.073 for GPR, and R^2=0.96, DC=0.96, and RMSE=0.056 for SVM.

4. Both SVM and GPR methods can be regarded as reliable approaches, providing acceptable and desirable results for estimating the required air concentration.

For future research, it is recommended to explore various artificial intelligence and hybrid models and compare their performance against the findings of this study.

Statements and Declarations

Data Availability

The data can be sent on request by the corresponding author via email.

Conflicts of interest

The authors of this paper declared no conflict of interest regarding the authorship or publication of this article.

Author contribution

K. Roushangar and R. Saadatjoo: Conceptualization; K. Roushangar and R. Saadatjoo: Methodology; K. Roushangar, H. Abbaszadeh, and R. Saadatjoo: Investigation; K. Roushangar, H. Abbaszadeh, and R. Saadatjoo: Writing Original Draft; K. Roushangar, H. Abbaszadeh, R. Saadatjoo and A. Panahi: Review-Editing; K. Roushangar: Supervision.

AI Use Declaration

This study did not incorporate artificial intelligence techniques; instead, all analyses and optimizations were conducted using conventional and widely accepted analytical methods.

References

Abbaszadeh, H., Norouzi, R., Sume, V., Kuriqi, A., Daneshfaraz, R., & Abraham, J. (2023a). Sill role effect on the flow characteristics (experimental and regression model analytical). *Fluids*, *8*(8), 235. DOI: 10.3390/fluids8080235

Abbaszadeh, H., Daneshfaraz, R., & Norouzi, R. (2023b). Experimental Investigation of Hydraulic Jump Parameters in Sill Application Mode with Various Synthesis. *Journal of Hydraulic Structures*, *9*(1), 18-42. DOI: 10.22055/jhs.2023.43208.1245

Abbaszadeh, H., Daneshfaraz, R., Sume, V., & Abraham, J. (2024). Experimental investigation and application of soft computing models for predicting flow energy loss in arc-shaped constrictions. *AQUA—Water Infrastructure, Ecosystems and Society*, *73*(3), 637-661. DOI: 10.2166/aqua.2024.010

Amini, A., Hamidi, S., Shirzadi, A., Behmanesh, J., & Akib, S. (2021). Efficiency of artificial neural networks in determining scour depth at composite bridge piers. *International Journal of River Basin Management*, *19*(3), 327-333. DOI: 10.1080/15715124.2020.1742138

Chakib, B. (2013). Numerical Computation of Inception Point Location for Flat-sloped Stepped Spillway. *International Journal of Hydraulic Engineering*, *2*(3), 4752. DOI: 10.5923/j.ijhe.20130203.03

Daneshfaraz, R., Norouzi, R., Abbaszadeh, H., Kuriqi, A., & Di Francesco, S. (2022a). Influence of sill on the hydraulic regime in sluice gates: an experimental and numerical analysis. *Fluids*, *7*(7), 244. DOI: 10.3390/fluids7070244

Daneshfaraz, R., Norouzi, R., Abbaszadeh, H., & Azamathulla, H. M. (2022b). Theoretical and experimental analysis of applicability of sill with different widths on the gate discharge coefficients. *Water Supply*, *22*(10), 7767-7781. DOI: 10.2166/ws.2022.354

Daneshfaraz, R., Norouzi, R., Ebadzadeh, P., Di Francesco, S., & Abraham, J. P. (2023). Experimental study of geometric shape and size of sill effects on the hydraulic performance of sluice gates. *Water*, *15*(2), 314. DOI: 10.3390/w15020314

Franc, J. P., & Michel, J. M. (2006). Fundamentals of cavitation (Vol. 76). Springer science & Business media.

Ghaderi, A. and Abbasi, S. (2022). The Effects of Modifying the Geometric Shapes of steps in Stepped Spillway on Hydraulic Parameters and Energy Dissipation. *Iranian Journal of Soil and Water Research*, *53*(5), 1035-1055. DOI: 10.22059/ijswr.2022.342428.669257

Hassanzadeh, Y., & Abbaszadeh, H. (2023). Investigating Discharge Coefficient of Slide Gate-Sill Combination Using Expert Soft Computing Models. *Journal of Hydraulic Structures*, *9*(1), 63-80. DOI: 10.22055/jhs.2023.43683.1251

Norouzi, R., Sihag, P., Daneshfaraz, R., Abraham, J., & Hasannia, V. (2021). Predicting relative energy dissipation for vertical drops equipped with a horizontal screen using soft computing techniques. *Water Supply*. *21*(8), 4493–4513. DOI: 10.2166/ws.2021.193

Norouzi, R., Ebadzadeh, P., Sume, V., & Daneshfaraz, R. (2023). Upstream vortices of a sluice gate: An experimental and numerical study. *AQUA—Water Infrastructure, Ecosystems and Society*, *72*(10), 1906-1919. DOI: 10.2166/aqua.2023.269

Pfister, M. (2011). Chute Aerators: Steep Deflectors and Cavity Subpressure. *Journal of Hydraulic Engineering*, *137*(10), 1208–1215. DOI: 10.1061/(ASCE)HY.1943-7900.0000436

Pal, M., Singh, N. K., & Tiwari, N. K. (2014). Kernel methods for pier scour modeling using field data. *Journal of Hydroinformatics*, *16*(4), 784-796. DOI: 10.2166/hydro.2013.024

Parsaie, A., Haghiabi, A. H. Saneie, M., & Torabi, H. (2018). Applications of soft computing techniques for prediction of energy dissipation on stepped spillways. *Neural Computing and Applications*, *29*, 1393-1409. DOI: 10.1007/s00521-016-2667-z

Pfister, M., & Hager, W. H. (2010). Chute Aerators: Air Transport Characteristics. *Journal of Hydraulic Engineering*, *136*(6), 352–359. DOI: 10.1061/(ASCE)HY.1943-7900.0000189

Rahmeyer, W. J. (1981). Cavitation damage to hydraulic structures. *American Water Works Association*, *73*(5), 270-274. DOI: 10.1002/j.1551-8833.1981.tb04703.x

Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (Vol. 1, p. 159). Cambridge, MA: MIT press.

Raza, A., Wan, W., & Mehmood, K. (2021). Stepped spillway slope effect on air entrainment and inception point location. *Water*, *13*(10), 1428. DOI: 10.3390/w13101428

Roushangar, K., Goodarzi, S., & Abbaszadeh, H. (2024a). Numerical investigation of the performance of blade groynes on scouring and its effect on hydraulic parameters of sediment and flow. *Environmental Water Engineering*, *10*(1), 121-136. DOI: 10.22034/ewe.2023.388931.1851

Roushangar, K., Shahnazi, S., & Mehrizad, A. (2024b). Data-intelligence approaches for comprehensive assessment of discharge coefficient prediction in cylindrical weirs: Insights from extensive experimental data sets. *Measurement*, *233*, 114673. DOI: 10.1016/j.measurement.2024.114673

Roushangar, K., Shahnazi, S., & Sadaghiani, A. A. (2023). An efficient hybrid grey wolf optimization-based KELM approach for prediction of the discharge coefficient of submerged radial gates. *Soft Computing*, *27*(7), 3623-3640. DOI: 10.1007/s00500-022-07614-7

Salmasi, F., Abraham, J., & Salmasi, A. (2021). Effect of stepped spillways on increasing dissolved oxygen in water, an experimental study. *Journal of Environmental Management*, *299*, 113600. DOI: 10.1016/j.jenvman.2021.113600

Süme, V., Daneshfaraz, R., Kerim, A., Abbaszadeh, H., & Abraham, J. (2024). Investigation of clean energy production in drinking water networks. *Water Resources Management*, *38*, 2189–2208. DOI: 10.1007/s11269-024-03752-9

Toombes L (2002). Experimental Study of Air-Water Flow Properties on Low-gradient Stepped Cascades. Ph.D. thesis, Dept of Civil Engineering, University of Queensland, Brisbane, Australia.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wu, J., Ma, F., & Dai, H. C. (2011). Influence of filling water on air concentration. *Journal of Hydrodynamics*, *23*(5), 601-606. DOI: 10.1016/S1001-6058(10)60155-2