How do decision-tree-based machine learning techniques compare to hybrid approaches for predicting fluvial dike breach discharge?

Vincent Schmitz[1*], Renaud Vandeghen[2], Sébastien Erpicum[1], Michel Pirotton[1], Pierre Archambeau[1], and Benjamin Dewals[1]

[1]Hydraulics in Environmental and Civil Engineering, Urban and Environmental Engineering, University of Liege, Belgium.

[2]Department of Electrical Engineering and Computer Science, University of Liège, Belgium

[*]Corresponding author: Vincent Schmitz (v.schmitz@uliege.be)

Highlights:

- Machine-learning, analytical and empirical approaches are compared for breach discharge prediction.
- The extremely randomized trees algorithm leads to particularly accurate results.
- A new hybrid approach combines a new analytical model and empirical formulas with machine learning.
- The new analytical model with machine learning performs well outside the training space.

Keywords:

Fluvial dike; breach discharge; side weir; decision-tree; machine learning; analytical model; hybrid model

Acknowledgments

The authors gratefully acknowledge S. Pierard from the Department of Electrical Engineering and Computer Science of the University of Liège for its valuable contribution to the discussions on the implementation of machine learning techniques.

Abstract

The breaching of a fluvial dike can have devastating consequences for flooded areas. Accurate prediction of the breach discharge is crucial for enhancing preventive measures and emergency planning. So far, most studies have relied on empirical formulas developed for simplified configurations, which fail to capture the complexity of a real dike breaching event. In this context, machine learning (ML) models offer promising predictive capabilities. This study focuses on decision-tree-based models, trained on 43 dike breaching laboratory tests, and compares three predictive approaches: (1) direct prediction using ML, (2) direct prediction using empirical formulas developed for simplified configurations, and (3) a novel analytical approach with an empirical parameter computed using ML or empirical regressions.

The extremely randomized trees algorithm demonstrates particularly high accuracy when predicting the breach discharge (approach 1), while the empirical formulas (approach 2) perform poorly. The new analytical model (approach 3) provides intermediate accuracy. Additionally, a novel hybrid approach is proposed, which consists in applying a ML-based corrective term to approaches (2) and (3). This strategy significantly improves the accuracy of the results associated with the empirical formulas (approach 2) and the analytical model (approach 3), both in interpolation and extrapolation, i.e., when tested inside or outside the ML training space. This makes it particularly valuable for predicting the breach discharge beyond the training space, where ML techniques alone are expected to be less effective. The definition of the dike breach invert level, i.e., one of the model inputs, was varied, but it had little influence on the models' performance.

Expanding the experimental dataset by conducting new laboratory or field tests would further enhance the accuracy and reliability of the ML models. Future studies may explore alternative ML models, including physics-guided deep learning algorithms, which, although in their early stages, hold substantial potential for future applications in predicting the breach discharge outside the model training space.

1. Introduction

With the increase in extreme meteorological events, growing urbanization in hinterlands and aging infrastructure, fluvial dikes are becoming more prone to breaches while their potential impact increases substantially (Flynn et al., 2022). Many dikes are decades old and were not originally designed to withstand current environmental and hydrological pressures. As these structures deteriorate over time, their structural integrity becomes increasingly uncertain, highlighting the urgent need for improved risk assessment tools (Ubay-Anongphouth and Alfaro, 2022; Yang et al., 2024). Consequently, developing predictive dike breach models has become critically important for ensuring safe land-use planning and effective emergency response strategies. Among numerical models, spatially non-discretized models offer a good compromise between accuracy and computational speed, allowing rapid assessments in time-sensitive situations (ASCE/EWRI Task Committee, 2011).

While many studies have focused on dam breaching (e.g., Sammen et al. (2017), Marangoz et al. (2024), El Bilali and Taleb (2025), Elalfy et al. (2025), Issakhov et al. (2025) and Zhang et al. (2025)), few have examined dike breaching, despite the significant differences in breaching dynamics. Dike breaching models are essential to accurately reproduce the breach

discharge and the resulting dike erosion. However, the capability of existing experimental and analytical side weir discharge formulas in predicting the flow through a real dike breach remains limited (Schmitz et al., 2024). Those formulas mostly rely on laboratory setups or analytical assumptions that do not represent breaching events faithfully, e.g., prismatic sharp-crested weir geometry (Hager, 1987; Bagheri et al., 2014; Elalfy et al., 2018; Lee, 2019) or narrow constraint hinterland (Ranga Raju et al., 1979; Cheong, 1991; Ibrahim et al., 2022; Wang et al., 2024).

There is no straightforward solution to tackle those limitations as dike breaching events are characterized by complex features, such as 3D flow patterns close to the side breach (Michelazzo et al., 2015; Cheng et al., 2022; Chowdhury et al., 2022) and highly non-uniform breach geometry (Kakinuma et al., 2013; Rifai et al., 2019). Oversimplified experimental setups and strong theoretical assumptions fail at capturing those features.

Within this context, supervised machine learning techniques may be of great help as they are able to replicate complex relationships when properly trained. Specifically, regression algorithms are considered when predicting a continuous target variable, i.e., the dike breach discharge, based on a limited number of features, i.e., upstream flow characteristics, channel and dike geometry, and breach morphology (Haghbin and Sharafati, 2022). Many machine learning techniques have been used in previous works dedicated to the analysis of the flow through a side weir, including linear methods (Jamei et al., 2021), decision-tree-based techniques (Hameed et al., 2021), support vector regressions (Balahang and Ghodsian, 2023, 2024; Li et al., 2024), genetic algorithms (Roushangar et al., 2016; Azimi et al., 2017), and neural networks (Parsaie, 2016; Saffar et al., 2021).

However, most works considered prismatic weirs with idealized geometries, e.g., triangular (Jamei et al., 2021; Balahang and Ghodsian, 2023), rectangular (Parsaie, 2016; Azimi et al., 2017; Hameed et al., 2021; Saffar et al., 2021; Balahang and Ghodsian, 2024), trapezoidal (Roushangar et al., 2016), or semi-circular (Li et al., 2024). Also, they often focused on the determination of the discharge coefficient, whose definition may vary from one experiment to another. Di Bacco and Scorzini (2019) highlighted that this approach could lead to inconsistent datasets and biased regression models.

The choice of a specific machine learning model depends on the data on hand (e.g., type, sample size, distribution, relationships…) and the user's objective (accuracy, speed, interpretability, ease of use…). In this work, multiple linear and decision-tree-based models are considered. Linear methods are intuitive, easy to use and computationally efficient. Methods based on decision trees are also particularly appropriate due to their simple use, i.e., very few hyperparameters, and ability to handle non-linear dependencies. Data collected from experimental tests (Rifai et al., 2019; Schmitz et al., 2021) are used to feed the machine learning models.

The objective of this study is to evaluate the ability of multiple linear regressions and decision-tree-based machine learning methods to predict the breach discharge of fluvial dikes, using hydrodynamic data from the main channel and geometric parameters of the breach. These machine learning models are benchmarked against traditional empirical formulas developed for simplified configurations. Their predictive capability is assessed both in interpolation and extrapolation, i.e., when tested within and beyond the training data space. In addition, a new semi-analytical model is introduced. Based on analytical developments and incorporating a single

empirical parameter, this model estimates the breach discharge, with the parameter value determined either through machine learning models or empirical regressions derived from the experimental dataset. Finally, the study explores the influence of feature selection in ML models, particularly the role of variables describing the breach geometry.

Section 2 presents the machine learning models and datasets used in this work. Section 3 introduces a new semi-analytical model for the determination of the breach discharge and compares the breach discharge predicted by the different machine learning models with experimental measurements. Section 4 discusses the performance and limitations of the machine learning models, the new semi-analytical model, and the empirical formulas. It also introduces a new hybrid approach, in which an ML-based corrective term is applied to either the empirical formulas or the analytical model, significantly improving their predictive accuracy. This section also investigates the influence of the breach geometry definition on the numerical results. Finally, conclusions are drawn in Section 5.

2. Methods and data

2.1. Machine learning techniques

In this work, two types of machine learning (ML) techniques were used to predict the dike breach discharge directly, or indirectly through a new analytical model (Section 3). ML methods were selected based on their ease of interpretation, suitability for problems with a small number of features, as is the case here (Section 2.2), and their limited number of hyperparameters. Thanks to this limited number of hyperparameters, the selected methods do not require any validation dataset, allowing a larger portion of the data to be used for testing and thus enabling

a more robust evaluation of model performance. Specifically, we focused on multiple linear regression (MLR) and decision-tree-based regression methods readily available in *scikit-learn* library, i.e., a machine learning library developed in Python (Pedregosa et al., 2011).

The MLR techniques aim at modelling a target value as a linear combination of the features. The main advantages of this method are its simple interpretation and its limited computation cost. However, it fails to reproduce non-linear phenomena.

The decision tree regression technique predicts the target variable by relying on simple decision rules, i.e., boolean logic. In this case, the target variable takes a constant value as long as the boolean results of the decision rules are not modified. This leads to a piecewise constant approximation of the target variable. This algorithm mainly relies on three hyperparameters:

- the minimum number of training points required to build a leaf, i.e., the minimum number of training points that fulfil all conditions of an entire decision chain;
- the minimum number of training points required to split an internal node;
- the tree maximum depth, i.e., the maximum number of decision rules contained in a decision chain.

Although this method is simple to use and to interpret, it tends to overfit the training data and returns discontinuous predictions for the target variable.

To tackle those limitations, more advanced algorithms were developed, e.g., random forests (Breiman, 2001) and extremely randomized trees (Geurts et al., 2006). Random forests rely on multiple decision trees, each built from a bootstrap sample of the training dataset, i.e., a

sample drawn with replacement from the training dataset. The final prediction corresponds to the average prediction of all individual trees. In extremely randomized trees, the whole training dataset is used to build each individual tree, i.e., no bootstrapping, but randomness is added when generating the decision rules. Instead of selecting the most discriminative threshold for each individual feature when defining a new decision rule, random thresholds are assigned to each feature. This step is repeated *k* times for the considered split, with *k* the number of features. Finally, the best of these *k* randomly generated threshold sets is selected as the decision rule. The procedure is thus used each time a decision rule is generated in each decision tree. The final prediction is obtained by averaging the prediction of each individual decision tree. Random forests and extremely randomized trees considerably reduce sensitivity to the training dataset. In Section 3, the predictive capability of the different machine learning techniques is compared.

When building ML models, experimental data are required. Those are generally divided into three datasets, called training, validation, and test datasets. The training dataset is used to fit the model parameters so that it minimizes the predictive error on the target variable. The validation dataset is then used to select the best model hyperparameters, such as the tree structures in decision-tree-based models. Finally, the test dataset serves as a benchmark for evaluating the model's ability to predict the target variable. To allow for a relevant and fair evaluation of the ML model performance, the composition of each set should depend on the experimental data specificities, as detailed in Section 2.2.

2.2. Datasets generation

Data used in this work were collected from laboratory experiments. Most of them were previously presented and discussed by Rifai et al. (2019) and Schmitz et al. (2021). In total, 43 tests are considered here. The laboratory setup and tests features are thoroughly presented in Text S1 and Table S1 of the Online Resource.

As mentioned earlier, the target variable in this work is the dike breach discharge, Q_b. Its value is known to mainly depend on the main channel hydrodynamic state and the shapes of the main channel, dike and breach (Schmitz et al., 2021, 2023). Within this context, five representative non-dimensional features were selected. These were identified in previous experimental campaigns as the most influential parameters affecting the breach discharge, among those that varied in the 43 tested configurations:

$$X = \left[z_{r,adim} = \frac{z_r}{w_{r,FS}}; \mathrm{F}; S_d; L_{k,adim} = \frac{L_k}{w_{r,FS}}; B_{top,adim} = \frac{B_{top}}{w_{r,FS}} \right], \tag{1.1}$$

with z_r the water level in the main channel, $w_{r,FS}$ the main channel width at the free surface, F the Froude number upstream from the breach, S_d the dike slope on the floodplain side, L_k the dike crest width, and B_{top} the breach top width. The breach top width was selected to represent the breach geometry as it is the only breach feature recorded during all experimental campaigns used in this work. During a breaching event, all parameters in Eq. (1.1) are time-dependent, except S_d and L_k.

In all tests, the breach top width was extracted at intervals that range between 1 and 60 seconds, depending on the test phase. In contrast, the hydrodynamic variables, i.e., the water level and the breach discharge, were obtained every 0.02 to 0.1 second, depending on the

experiment. A linear variation was assumed between two successive geometry reconstructions to obtain a breach width associated to each recorded value of the hydrodynamic variables. Additionally, each experimental data point is assumed to be independent of the value of the parameters measured in the past, i.e., each data point is considered as an independent instance of a relation between the output and the input parameters. That way, several thousand data points were obtained for each laboratory experiment, leading to about 1.5 10^6 experimental data points overall.

Once the global dataset has been generated, it must be divided into three subsets: training, validation, and test sets. In the present case, no hyperparameters need to be fixed in the MLR algorithm, and only a few are required in the decision-tree-based models. For the latter, the hyperparameters values selected in this work are generally accepted default values, as listed in Table S2 of the Online Resource. This set leads to satisfactory results while limiting overfitting (Pedregosa et al., 2011). Consequently, no validation dataset was generated in this work. Still, dispatching experimental data points between the training and the test sets should be done carefully to avoid biased evaluation due to too similar training and test data. Data clustering helps gathering alike data subsets to avoid training and testing to be performed on data from the same subset. In this work, this should be done by avoiding training and testing ML models on data points emanating from identical or too similar laboratory experiments, i.e., data points encompassed in alike feature spaces. The laboratory tests could be differentiated by considering three main features, namely the main channel Froude number at overtopping initiation, F_{OT}, the non-dimensional dike crest width at overtopping initiation, $L_{k,adim}$, and the dike slope on the floodplain side, S_d. Figure S2 of the Online Resource shows that 21 clusters were identified using

the OPTICS algorithm (Ankerst et al., 1999) with a minimum cluster membership of two. Each cluster is labeled with a number between -9 and 11. Degenerated clusters containing only one experimental test were labeled with a red negative number. As a result, the test dataset corresponds to one cluster, or degenerated cluster, while the rest of the data forms the training dataset. Once the training dataset is defined, the features are standardized based on this dataset, i.e., mean subtraction and division by the standard deviation (Shanker et al., 1996).

To obtain a global performance score for each ML technique, the mean absolute relative error (MARE) on the target value is computed when successively considering each cluster as the test dataset, so that

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|X_{exp}^{i} - X_{num}^{i}\right|, \tag{1.2}$$

with n the number of data points in the considered test dataset, X_{num}^{i} the value of the target variable obtained from the ML model fed with features associated to data point i, and X_{exp}^{i} the associated experimental value. The averaged value of MARE obtained for all the different test datasets reflects the global performance of the considered ML model.

3. Results

In this section, the performance of the four ML techniques presented in Section 2.1 is compared for both the direct evaluation of the breach discharge and its indirect evaluation via a new analytical model. This model, based on energy conservation, assumes critical flow across the breach section. Following the approach of Schmitz et al. (2023), we introduce a parameter α,which represents the fraction of the breach width that conveys most of the breach discharge

and should be considered the breach's critical section. A detailed presentation of this model is provided in Text S2 of the Online Resource.

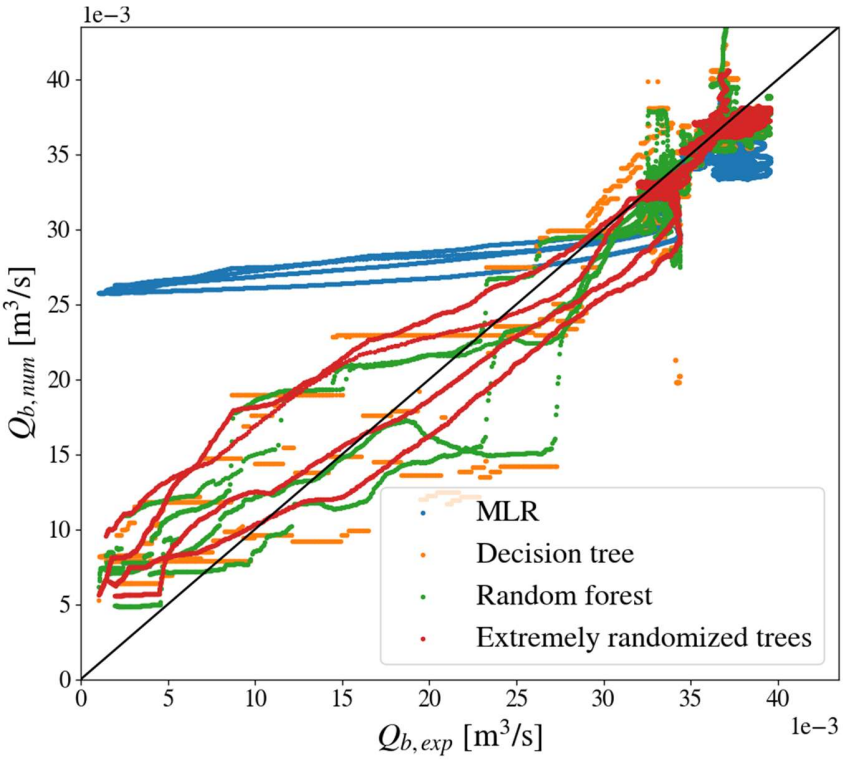
As an illustration, Fig. 1 presents the results generated when using cluster #5 as the test dataset (Table S3 of the Online Resource) and the target variable being Q_b (Fig. 1a) or α (Fig. 1b). Table S4 of the Online Resource provides the averaged value of MARE on Q_b, α and $Q_b(\alpha)$, as defined in Section 2.2.

Overall, all ML techniques lead to more accurate results when directly predicting Q_b (Fig. ***1***a) rather than indirectly through α (Fig. 1b). In all cases, the discrepancies with the experimental data are particularly large when the breach discharge is small and α is large, i.e., shortly after overtopping initiation when the breach is still narrow (Figure S4 of the Online Resource). During this period, the breach geometry and the flow conditions evolve quickly. Consequently, a slight temporal mismatch between hydraulic and breach morphologic variables may have a significant impact on the models' predictive capabilities. Also, much fewer data points can be collected during this highly transient period, i.e., less training points are available, which induces weaker model performance. Although large discrepancies seem to appear in Fig. 1a and 1b, the overall models performances are mostly influenced by the accuracy on larger values of the breach discharge, which are mostly represented in the dataset as they represent most of a typical breaching event.

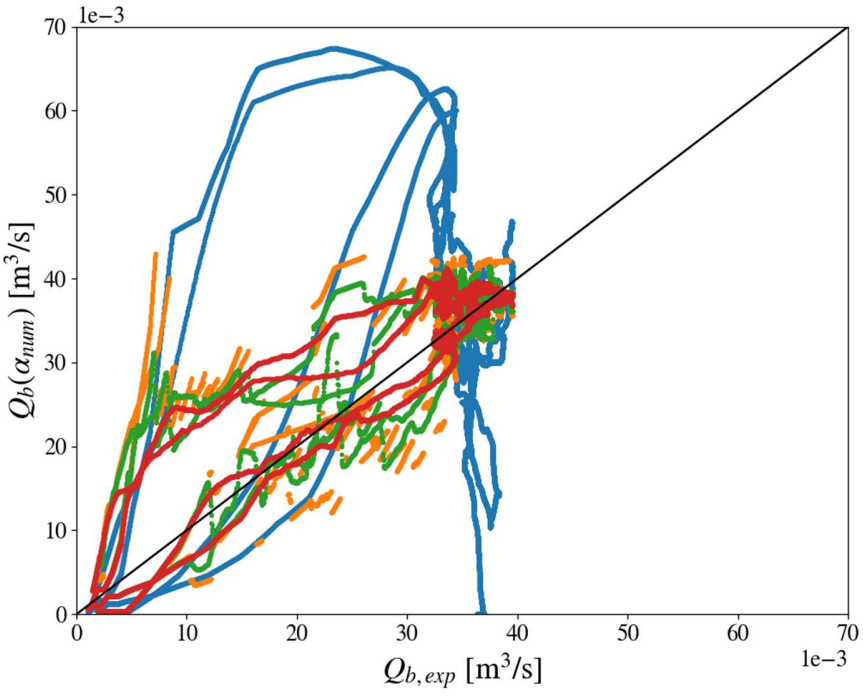
The MLR model appears to perform poorly over the whole tested range, with an averaged MARE of about 13% when predicting Q_b, and above 30% for $Q_b(\alpha)$ (Table S4 of the Online Resource). Fig. ***2*** highlights the particularly high variability of the model performance when

varying the test dataset. The dike breaching process being highly non-linear, a linear model was indeed expected to provide low fidelity results.

On the contrary, decision-tree-based ML techniques provide much better results. Among them, the extremely randomized trees model performs the best for all predicted variables, while having the smallest performance variability (Fig. 2). Discontinuities are observed with the basic decision-tree model. These discontinuities are still present, but to a smaller extent, when using the random forest model. Finally, they almost completely fade when the extremely randomized trees model is selected. The latter model presents the most physical behavior as the breach discharge should vary continuously. Consequently, the extremely randomized trees model should be preferred in the present case as it reduces the error on the breach discharge while avoiding non-physical discontinuities in the results. This model will be the only one considered when discussing results in Section 4.



(a)



(b)

Fig. 1 Results predicted by the ML models for (a) the breach discharge, Q_b, (b) and (b) the breach discharge computed from α, $Q_b(\alpha)$. In each case, the test dataset is cluster #5, which contains data from four laboratory experiments.

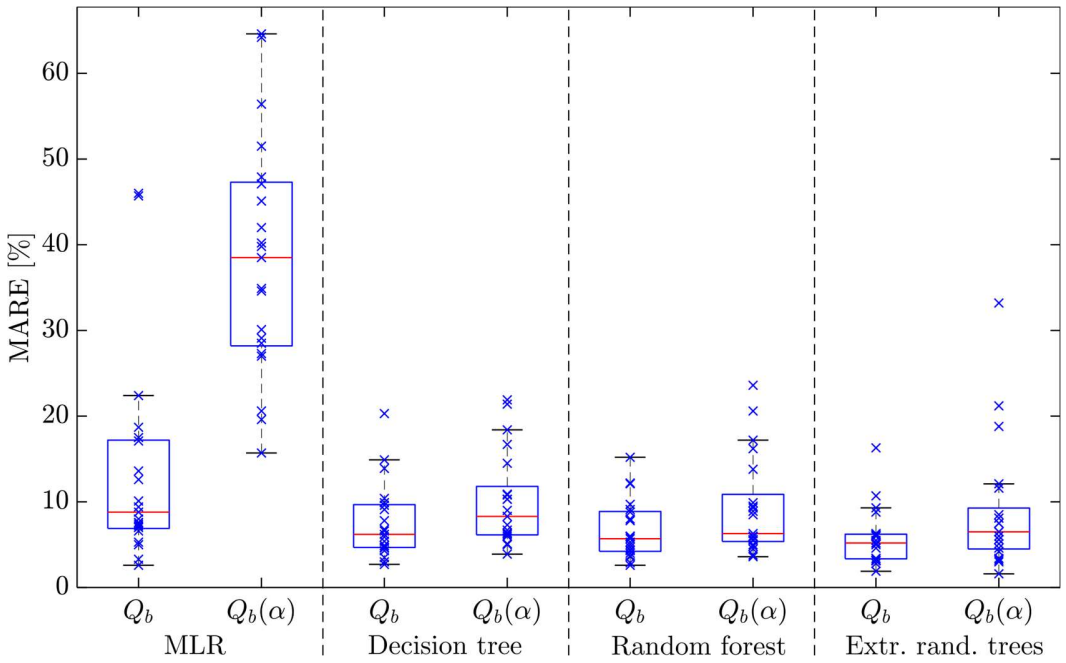


Fig. 2 Boxplot of the MARE on the breach discharge associated to each ML technique and each target value when varying the test dataset

4. Discussion

In this section, the performance of the extremely randomized trees model is compared against empirical formulas for Q_b and α. Interpolation and extrapolation capabilities of the ML model are assessed individually, i.e., by considering the test dataset to be inside or outside of the training dataset space. Finally, the impact of the breach geometry definition on the results is evaluated.

4.1. Comparison between machine learning and empirical formulas in interpolation and extrapolation

Decision-tree-based ML models provide constant predictions when evaluated outside their training space. In this case, their performance is expected to degrade. Features associated with real dike configurations are diverse, but covering the entire feature space in the training dataset is hardly feasible due to the considerable amount of experimental data that should be collected. This limitation highlights the need for good predictive capabilities in both interpolation and extrapolation, i.e., inside and outside the training space, respectively. Within this context, empirical formulas or analytical models are interesting tools as they usually capture the general physical trends, and their results keep evolving outside their fitting space.

Many empirical or semi-empirical formulas have been proposed in literature for the definition of the lateral breach discharge coefficient, C_D, (e.g., Hager (1987), Jalili and Borghei (1996), Bagheri et al. (2014),…) which is directly involved in the expression of the breach discharge as

$$Q_b = \frac{2}{3} C_D \sqrt{2g\left(h - z_b\right)^3} L_s , \tag{1.3}$$

with g the gravitational acceleration [m/s^2], h the flow depth in the main channel upstream of the side opening [m], z_b the crest height of the side weir [m], and L_s the length of the side weir [m].

A thorough evaluation of the predictive capabilities of eleven empirical and semi-empirical formulas was conducted by Schmitz et al. (2024) for various side breach configurations.

Formulas proposed by Jalili and Borghei (1996) and Singh et al. (1994) turned to be the most accurate when considering realistic fluvial dike breaches (Table S5 of the Online Resource).

As stated earlier, the breach top width is the only information about the breach geometry that was measured in all experimental cases. To obtain the breach invert, z_b, the breach is assumed to be perfectly trapezoidal with its side slopes equal to the repose angle of the dike wet material. Additionally, erosion is assumed to be uniform over the entire breach area, which allows for the direct computation of z_b based on B_{top}. The length of the side weir, L_s, is chosen equal to the breach bottom width.

Alternatively, two new empirical relationships for the calculation of the effective breach width coefficient, α, are proposed in this work. These equations were derived from experimental data (Figure S5 of the Online Resource). The first regression relies solely on the non-dimensional breach top width as its impact on the value of α appeared to be predominant (Schmitz et al., 2023):

$$\alpha_{R1} = \frac{1}{3} + 0.035 B_{top,adim}^{-2.5} . \tag{1.4}$$

The second regression uses three input parameters, namely the non-dimensional breach top width, $B_{top,adim}$, a non-dimensional velocity, $U_{adim} = \frac{U_r}{\sqrt{gB_{top}}}$, and a non-dimensional dike width, $w_{d,adim} = \frac{w_{d,FS}}{B_{top}}$, with $w_{d,FS}$ the dike cross-section width at the water free surface level:

$$\alpha_{R2} = 1.789 w_{d,adim}^{0.242} \left(\frac{U_{adim}}{B_{top,adim}} \right)^{0.225} - 0.5 . \tag{1.5}$$

Finally, a corrective term computed through ML (extremely randomized tree algorithm) may be applied to these experimental formulas, leading to an improved evaluation of the breach discharge, denoted Q_b^+ (Fig. 3). The target variable of this ML model is the difference between the experimental breach discharge and the breach discharge predicted by an empirical formula. The same features as the ones considered for the direct evaluation of Q_b and α are used (Eq. (1.1)). In addition, the empirical value of α is incorporated into the features when evaluating the corrective term related to the breach discharge obtained using Eq. (1.4) or (1.5) with the analytical model, i.e., $Q_b^+(\alpha)$ in Fig. 3.

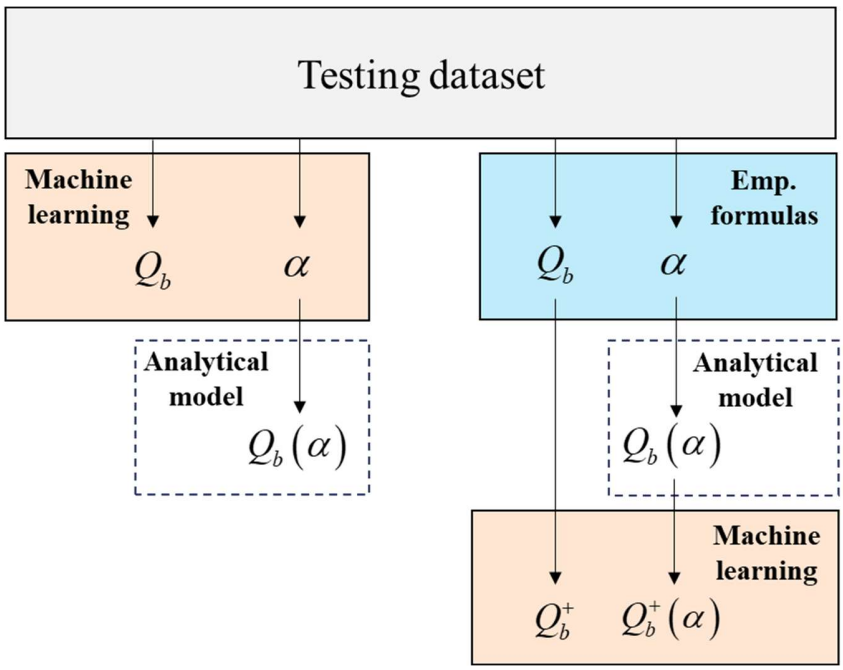


Fig. 3 Flow chart of the different approaches adopted in this work to compute the breach discharge

In the following, the predictive capability of the ML and empirical formulas are assessed in interpolation and extrapolation separately. To this end, Figure S6 of the Online Resource highlights clusters that were considered in extrapolation, i.e., located on the edge of the dataset space, and in interpolation. In total, 10 clusters (25 experimental tests) are considered in

interpolation when individually selected as the test dataset, while 11 clusters are in extrapolation (18 experimental tests).

As expected, Figure S7 and Table S5 of the Online Resource show that results are systematically more accurate in interpolation than in extrapolation when considering the machine learning alone, i.e., the extremely randomized trees algorithm. In interpolation, the averaged MARE on Q_b and $Q_b(\alpha)$ is limited to around 5% in both cases. While the accuracy on both variables drops in extrapolation, the error variability associated to Q_b is more limited than the one on $Q_b(\alpha)$ (Fig. ***4***), making the direct prediction of Q_b more reliable.

In the present work, the concept of extrapolation and interpolation is meaningless for Q_b empirical formulas as they were fitted on a different dataset than the one considered in this work. Their poor and highly variable performance, especially for the formula of Jalili and Borghei (1996) (Fig. 4), suggests that they generalize badly outside their fitting space. The performance of both α regression formulas is largely better in interpolation and extrapolation compared to the empirical formulas for Q_b. However, it decreases by 25% (Eq. (1.4)) to 140% (Eq. (1.5)) in extrapolation compared to the scores in interpolation.

Results are greatly improved when adding a ML-based corrective term to the empirical formulas, i.e., Q_b^+ and $Q_b^+(\alpha)$. In this case, the averaged MARE grows to a smaller extent when evaluated in extrapolation, especially for α regressions, although the performance variability substantially increases (Fig. 4). Overall, very satisfactory results are obtained in both interpolation and extrapolation when using the extremely randomized trees model for the prediction of Q_b and $Q_b(\alpha)$, and when adding a ML-based corrective term to the empirical formula for Q_b derived by Singh et al. (1994) and for Eq. (1.5). Nonetheless, the predictive capability of ML-based models

decreases when evaluated further from their training dataset space. In this case, Eq. (1.5) associated with a ML-based corrective should be favored as the associated analytical model continues evolving based on physical considerations even outside the training space, limiting the impact of a poorly performing ML model.

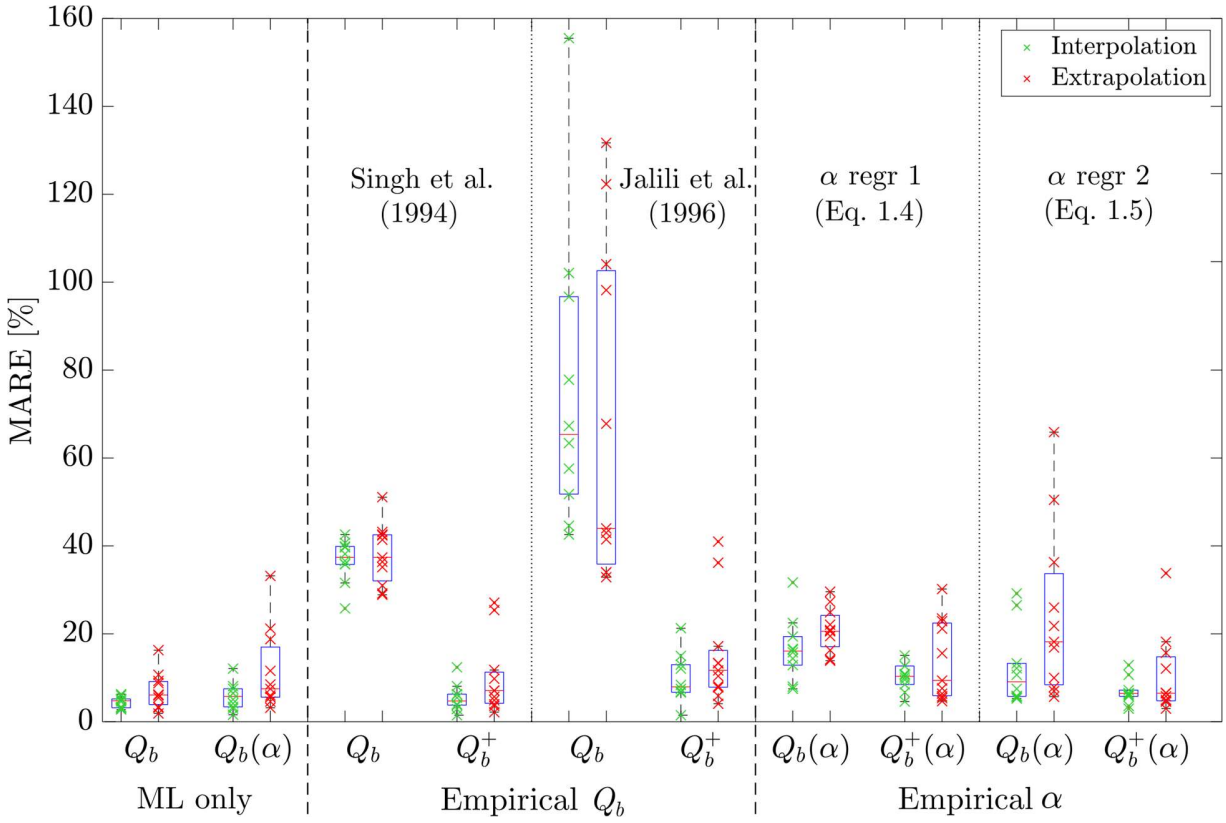


Fig. 4 Boxplot of the MARE values associated with each approach in interpolation and extrapolation

4.2. Impact of the breach geometry definition

Considering different features or adding new ones to the ML models may have a significant impact on the results. So far, the breach invert level, z_b, has been discarded as it was not recorded in all experiments (see Text S1 of the Online Resource). In this section, only tests data for which the 3D breach geometry was recorded are considered (Rifai et al., 2019, 2020).

The performance of the different models is assessed when the experimental breach invert is added to the features of the ML models and considered in the empirical formulas and the analytical model. In this case, the breach is still assumed trapezoidal with its side slopes equal to the repose angle of the dike wet material, but assuming a uniform breach erosion is no longer required. Fig. ***5*** compares the performance of this new approach with the one obtained when z_b is not part of the features of the ML models and approximated based on B_{top} in the analytical model (Text S2 of the Online Resource) and the empirical formulas (Section 4.1). Defining a unique value for z_b is not trivial as the breach bottom is non-uniform (Rifai et al., 2018). The sensitivity of the different models to z_b was assessed by defining it as (a) the minimum breach invert level, (b) the 10th percentile, (c) and the 20th percentile. In all cases, only the breach cross-section parallel to the dike main axis and passing through the center of the crest was considered.

Fig. ***5*** shows that results are systematically improved with the experimental z_b when ML is used at last in the procedure, i.e., when computing Q_b with ML only, Q_b^+ , and $Q_b^+(\alpha)$. This improvement is, however, limited in most cases, while the definition of z_b has almost no impact, i.e., a relative difference of less than 2% on the averaged MARE values is observed in all cases.

Considering the experimental z_b reduces the predictive capability of the new analytical model when ML is not used at the end of the procedure, i.e., when $Q_b(\alpha)$ is computed. In contrast, the predictions from empirical formulas proposed by Singh et al. (1994) and Jalili and Borghei (1996) with no ML-based corrective term are greatly improved when using the experimental breach invert. Here again, its definition has a negligible impact on the results (maximum relative variation of the averaged MARE of 12%).

It can be concluded that considering the experimental z_b is only recommended when using empirical formulas for Q_b on their own. In the other cases, the performance improvement is limited, and might even significantly decrease in some cases, e.g., when predicting $Q_b(\alpha)$. Also, it should be borne in mind that increasing the number of ML features increases the model training time. Finally, the definition of the experimental z_b has a negligible impact on the results.

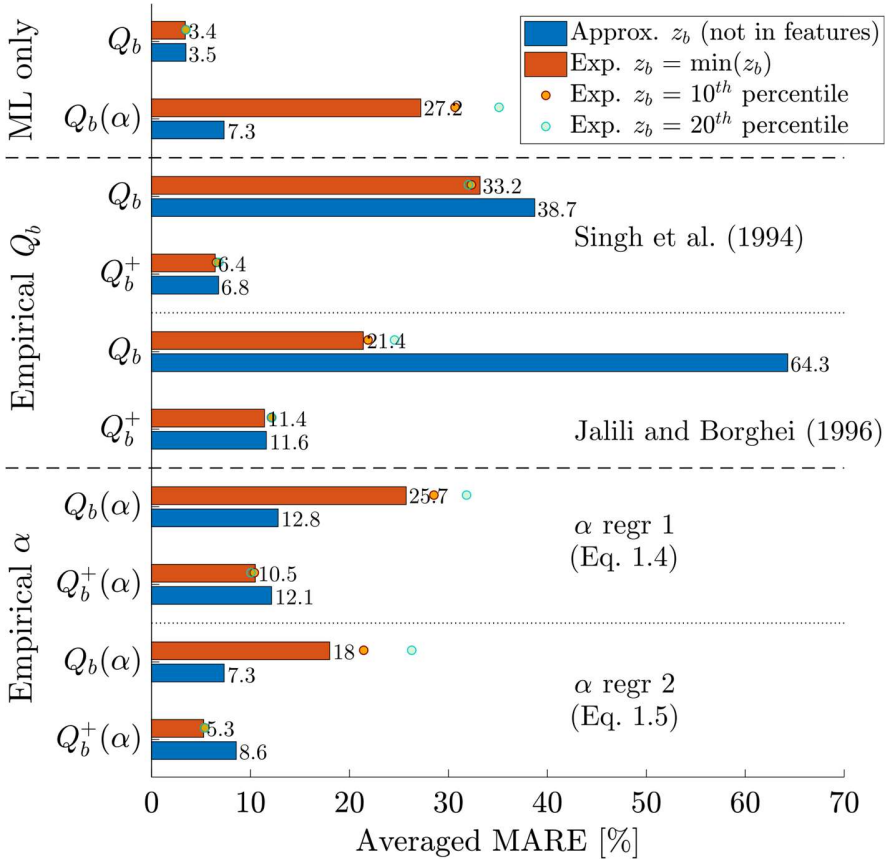


Fig. 5 Impact of the definition of the breach invert on the models' accuracy

5. Conclusions

In this work, two different approaches for the prediction of the discharge through a fluvial dike breach were presented, namely the direct prediction of the discharge Q_b, or its indirect prediction through the parameter α using a new analytical model (Text S2 of the Online Resource). Among the multiple linear regressions and decision-tree-based machine learning techniques presented in this work for the prediction of those parameters (Section 2.1), the extremely randomized trees algorithm provides almost continuous outputs while being very accurate (mean absolute relative error between 5% and 10% on average).

The ML model used to predict Q_b or $Q_b(\alpha)$ was compared to empirical formulas for the prediction of Q_b and α, respectively (Section 4.1). Empirical formulas were less accurate, especially when predicting Q_b. A ML-based corrective term was then added to the different empirical formulas results. It significantly improved the results accuracy, especially when using Singh et al. (1994) formula and the second alpha regression (Eq. (1.5)).

In all cases, the model performance decreased slightly when evaluated in extrapolation, i.e., outside the space of the training dataset, whilst the variability of the model accuracy on the tested set rose (Fig. 4). It is expected that the further the ML models are evaluated from their training space, the lower the model performance. In the case of decision-tree-based ML techniques, the predictions become constant outside the training space. In this context, the use of the second empirical regression for α coupled to a ML-based corrective term is recommended. While offering accurate results, this method keeps evolving outside the training dataset space as it relies on the new analytical model, which uses continuously evolving parameters.

The performance sensitivity to the definition and the addition of the breach invert level, z_b, in the features of the ML models was also investigated (Section 4.2). The use of the experimental value of z_b is only recommended when considering empirical formulas for the prediction of Q_b. In the other cases, its use degrades the results or only slightly improves them in the best case, to the expense of a longer training time of the ML model.

To further improve the performance of the models, expanding the experimental dataset is crucial for increasing the "physical space" explored during training. Notably, some parameters were considered as constant over the entire tested space, e.g., dike height or material median diameter. Incorporating these parameters into the machine learning features and testing configurations with varying values for these parameters would enable the model to better capture the diversity of real-world breach scenarios. Additionally, new data and experimental setups should be comprehensively documented to ensure the consistency and inter-comparability of the database (Di Bacco and Scorzini, 2019).

The use of other ML methods for the prediction of the fluvial dike breach discharge could also be investigated, e.g., neural networks. However, purely data-driven approaches do not obey the governing laws of physical systems (Yin et al., 2021). Within this context, physics-guided deep learning methods have become increasingly popular when analyzing dynamic systems. In these methods, the loss function is reformulated to take into account physical constraints imposed by the user (Ai et al., 2025; Zhan et al., 2025). Although still in their early stages, these methods are promising for improving model extrapolation capabilities by integrating physical principles and thus reducing their reliance on the training dataset space. This emerging research area is evolving quickly and may hold substantial potential for future applications (Wang and Yu, 2023).

References

Ai, C., Ma, Y., Li, Z., Dong, G., 2025. Physics-Informed Neural Networks for Steady-State Weir Flows Using the Serre-Green-Naghdi Equations. Journal of Hydraulic Engineering 151.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure, in: Proceedings 1999 ACM SIGMOD International ConferenceManagement Data, SIGMOD'99. Association for Computing Machinery, Philadelphia, Pennsylvania, USA, pp. 49–60.

ASCE/EWRI Task Committee, 2011. Earthen Embankment Breaching. Journal of Hydraulic Engineering 137, 1549–1564.

Azimi, H., Bonakdari, H., Ebtehaj, I., 2017. A highly efficient gene expression programming model for predicting the discharge coefficient in a side weir along a trapezoidal canal. Irrigation and drainage 66, 655–666.

Bagheri, S., Kabiri-Samani, A.R., Heidarpour, M., 2014. Discharge coefficient of rectangular sharp-crested side weirs, Part I: Traditional weir equation. Flow Measurement and Instrumentation 35, 109–115.

Balahang, S., Ghodsian, M., 2023. Evaluating performance of various methods in predicting triangular sharp-crested side weir discharge. Applied Water Science 13, 171.

Balahang, S., Ghodsian, M., 2024. Enhancing rectangular side weir discharge prediction using stacking technique. Flow Measurement and Instrumentation 97.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Cheng, Y., Song, Y., Liu, C., Wang, W., Hu, X., 2022. Numerical Simulation Research on the Diversion Characteristics of a Trapezoidal Channel. Water 14.

Cheong, H., 1991. Discharge Coefficient of Lateral Diversion from Trapezoidal Channel. Journal of Irrigation and Drainage Engineering 117, 461–475.

Chowdhury, M.K., Konsoer, K.M., Hiatt, M., 2022. Effect of Lateral Outflow on Three-Dimensional Flow Structure in a River Delta. Water Resources Research 58, e2021WR031346.

Di Bacco, M., Scorzini, A.R., 2019. Are We Correctly Using Discharge Coefficients for Side Weirs? Insights from a Numerical Investigation. Water 11.

El Bilali, A., Taleb, A., 2025. A Novel Approach for Predicting peak flow from Breached Dam: Coupling Monte Carlo Simulation, Hydrodynamic Model, and an Interpretable XGBoost Model. Water Resources Management 39, 1177–1194.

Elalfy, E., Czapiga, M.J., Viparelli, E., Imran, J., Chaudhry, M.H., 2025. Modeling Breach Evolution in Noncohesive Earthen Dams by Overtopping. Journal of Hydraulic Engineering 151.

Elalfy, E., Tabrizi, A.A., Chaudhry, M.H., 2018. Numerical and experimental modeling of levee breach including slumping failure of breach sides. Journal of Hydraulic Engineering 144.

Flynn, S., Zamanian, S., Vahedifard, F., Shafieezadeh, A., Schaaf, D., 2022. Data-Driven Model for Estimating the Probability of Riverine Levee Breach Due to Overtopping. Journal of Geotechnical and Geoenvironmental Engineering 148, 04021193.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning 63, 3–42.

Hager, W.H., 1987. Lateral Outflow Over Side Weirs. Journal of Hydraulic Engineering 113, 491–504.

Haghbin, M., Sharafati, A., 2022. A review of studies on estimating the discharge coefficient of flow control structures based on the soft computing models. Flow Measurement and Instrumentation 83, 102119.
Hameed, M.M., AlOmar, M.K., Khaleel, F., Al-Ansari, N., 2021. An Extra Tree Regression Model for Discharge Coefficient Prediction: Novel, Practical Applications in the Hydraulic Sector and Future Research Directions. Mathematical Problems in Engineering 2021.
Ibrahim, I., Riviere, N., Lebouteiller, I., Mignot, E., 2022. Discharge Distribution in Open-Channel T-Shape Bifurcations: Effect of a Reduced Side Branch Width. Journal of Hydraulic Engineering 148, 04022015.
Issakhov, A., Rakhymzhanova, Z., Abylkassymova, A., 2025. Numerical Study of the Water Surface Movement on the Breaching Process of Natural Dams. Water Resources Management 39, 625–643.
Jalili, M.R., Borghei, S.M., 1996. Discussion: Discharge Coefficient of Rectangular Side Weirs. Journal of Irrigation and Drainage Engineering 122, 132–132.
Jamei, M., Ahmadianfar, I., Chu, X., Yaseen, Z.M., 2021. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. Flow Measurement and Instrumentation 77, 101878.
Kakinuma, T., Tobita, D., Yokoyama, H., Takeda, A., 2013. Levee breach observation at Chiyoda experimental flume, in: 12thInternationalSymposium River Sedimentation (ISRS), IRTCES, Kyoto, Japan.
Lee, K., 2019. Simulation of Dam-Breach Outflow Hydrographs Using Water Level Variations. Water Resources Management 33, 3781–3797.
Li, S., Shen, G., Parsaie, A., Li, G., Cao, D., 2024. Discharge modeling and characteristic analysis of semi-circular side weir based on the soft computing method. Journal of Hydroinformatics 26, 175–188.
Marangoz, H.O., Anılan, T., Karasu, S., 2024. Investigating the Non-Linear Effects of Breach Parameters on a Dam Break Study. Water Resources Management 38, 1773–1790.
Michelazzo, G., Oumeraci, H., Paris, E., 2015. Laboratory Study on 3D Flow Structures Induced by Zero-Height Side Weir and Implications for 1D Modeling. Journal of Hydraulic Engineering 141, 04015023.
Parsaie, A., 2016. Predictive modeling the side weir discharge coefficient using neural network. Modeling Earth Systems and Environment 2, 1–11.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Ranga Raju, K.G., Gupta, S.K., Prasad, B., 1979. Side Weir in Rectangular Channel. Journal of the Hydraulics Division 105, 547–554.
Rifai, I., El Kadi Abderrezzak, K., Erpicum, S., Archambeau, P., Violeau, D., Pirotton, M., Dewals, B., 2018. Floodplain Backwater Effect on Overtopping Induced Fluvial Dike Failure. Water Resources Research 54, 9060–9073.
Rifai, I., El Kadi Abderrezzak, K., Erpicum, S., Archambeau, P., Violeau, D., Pirotton, M., Dewals, B., 2019. Flow and detailed 3D morphodynamic data from laboratory experiments of fluvial dike breaching. Scientific data 6, 53.

Rifai, I., Schmitz, V., Erpicum, S., Archambeau, P., Violeau, D., Pirotton, M., Dewals, B., El Kadi Abderrezzak, K., 2020. Continuous Monitoring of Fluvial Dike Breaching by a Laser Profilometry Technique. Water Resources Research 56, e2019WR026941.

Roushangar, K., Khoshkanar, R., Shiri, J., 2016. Predicting trapezoidal and rectangular side weirs discharge coefficient using machine learning methods. ISH Journal of Hydraulic Engineering 22, 254–261.

Saffar, S., Babarsad, M.S., Shooshtari, M.M., Hosein poormohammadi, M., Riazi, R., 2021. Prediction of the discharge of side weir in the converge channels using artificial neural networks. Flow Measurement and Instrumentation 78.

Sammen, S.S., Mohamed, T.A., Ghazali, A.H., El-Shafie, A.H., Sidek, L.M., 2017. Generalized Regression Neural Network for Prediction of Peak Outflow from Dam Breach. Water Resources Management 31, 549–562.

Schmitz, V., Erpicum, S., Abderrezzak, K. El kadi, Rifai, I., Archambeau, P., Pirotton, M., Dewals, B., 2021. Overtopping-Induced Failure of Non-Cohesive Homogeneous Fluvial Dikes: Effect of Dike Geometry on Breach Discharge and Widening. Water Resources Research 57, e2021WR029660.

Schmitz, V., Kitsikoudis, V., Wylock, G., Erpicum, S., Pirotton, M., Archambeau, P., Dewals, B., 2024. Efficient modelling of lateral discharge through a dike breach. Journal of Hydrology 640, 131660.

Schmitz, V., Rifai, I., Kheloui, L., Erpicum, S., Archambeau, P., Violeau, D., Pirotton, M., El Kadi Abderrezzak, K., Dewals, B., 2023. Main channel width effects on overtopping-induced non-cohesive fluvial dike breaching. Journal of Hydraulic Research 61, 601–610.

Shanker, M., Hu, M.Y., Hung, M.S., 1996. Effect of data standardization on neural network training. Omega 24, 385–397.

Singh, R., Manivannan, D., Satyanarayana, T., 1994. Discharge Coefficient of Rectangular Side Weirs. Journal of Irrigation and Drainage Engineering 120, 814–819.

Ubay-Anongphouth, I.O., Alfaro, M., 2022. Delayed instabilities of water-retaining earth structures. Frontiers in Built Environment 8.

Wang, R., Yu, R., 2023. Physics-Guided Deep Learning for Dynamical Systems: A Survey . https://arxiv.org/abs/2107.01272.

Wang, Y., Lv, M., Wang, W., Meng, M., others, 2024. Discharge Formula and Hydraulics of Rectangular Side Weirs in the Small Channel and Field Inlet. Water 16, 713.

Yang, D., Wu, J., Guo, Z., Zeng, X., Zhang, Q., 2024. Safety risk assessment of reservoir dam structure: an empirical study in China. Scientific Reports 14.

Yin, Y., Le Guen, V., Dona, J., Bézenac, E. de, Ayed, I., Thome, N., Gallinari, P., 2021. Augmenting physical models with deep networks for complex dynamics forecasting. Journal of Statistical Mechanics: Theory and Experiment 2021, 124012.

Zhan, C., Zhang, T., Zhang, S., Yang, D., 2025. Solving complex flood wave propagation using split Coefficient-based Physical Informed Neural Network. Journal of Hydrology 654.

Zhang, F., Jia, S., Zhou, X., Wang, L., Zhu, Y., 2025. Modeling flood and breach evolution of the landslide dam due to overtopping. Journal of Hydrology 647.

Statements and Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

The study's conception and design were developed by V. Schmitz, B. Dewals, and R. Vandeghen. The hybrid approach was designed by V. Schmitz and M. Pirotton. Data collection and analysis were performed by V. Schmitz, S. Erpicum, and P. Archambeau. The first draft of the manuscript was written by V. Schmitz and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Not applicable.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.