Book of Extended Abstracts
41st IAHR World Congress
22-27 June 2025
Singapore

COMPARISON OF MACHINE LEARNING TECHNIQUES AND EMPIRICAL FORMULAS FOR THE PREDICTION OF THE DISCHARGE THROUGH A FLUVIAL DIKE BREACH

Vincent SCHMITZ[1], Sébastien PIERARD[2], Renaud VANDEGHEN[2], Sébastien ERPICUM[1], Michel PIROTTON[1], Pierre ARCHAMBEAU[1], Benjamin DEWALS[1]

[1] Hydraulics in Environmental and Civil Engineering, Urban and Environmental Engineering, University of Liege, Belgium
email: v.schmitz@uliege.be
email: s.erpicum@uliege.be
email: michel.pirotton@uliege.be
email: pierre.archambeau@uliege.be
email: b.dewals@uliege.be

[2] Department of Electrical Engineering and Computer Science, University of Liège, Belgium
email: s.pierard@uliege.be
email: r.vandeghen@uliege.be

ABSTRACT

The breaching of a fluvial dike is a complex phenomenon involving 3D flow patterns and a complex breach geometry. Oversimplifications inherent to traditional empirical and analytical approaches lead to inaccurate predictions of the breach discharge. Machine learning models are interesting tools as they can replicate complex relationships when properly trained. This study assesses the performance of a decision-tree-based model, specifically the extremely randomized trees method, using experimental data from previous works. This model is evaluated in both interpolation and extrapolation, i.e., when the model is evaluated inside or outside the training set space. It performs well in both cases, although results slightly degrade in extrapolation. It is then compared to classical empirical formulas. The latter provide low fidelity results in this case. A corrective term computed using machine learning is then coupled with the empirical formulas, which significantly improve their accuracy. Overall, the extremely randomized trees method yields satisfactory results when directly evaluating the dike breach discharge or when coupled with an empirical formula. Future work could expand the training set by exploring additional configurations, further increasing the reliability of the model.

Keywords: *Fluvial dike; Dike breaching; Breach discharge; Machine learning; Decision tree; Empirical formulas*

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). As a result, predictive dike breach models are crucial for safe land-use planning and emergency response. Existing experimental and analytical side weir discharge formulas struggle to accurately predict the discharge through a real dike breach due to oversimplified experimental setups and theoretical assumptions, respectively (Schmitz et al., 2024).

The complexity of dike breaching events, which involve 3D flow patterns (Neary et al., 1999; Michelazzo et al., 2015; Cheng et al., 2022; Chowdhury et al., 2022) and non-uniform breach geometries (Rifai et al., 2019), challenges these models. Machine learning techniques may be of great help as they are able to replicate complex relationships when properly trained. Within this context, decision-tree-based regression models are particularly suitable due to their simplicity of use, i.e., very few hyperparameters, and ability to handle non-

linear dependencies. In this work, experimental data from previous studies (Rifai et al., 2019; Schmitz et al., 2021) are used to train these models.

This work aims to evaluate the predictive capabilities of the extremely randomized tree method, i.e., a decision-tree-based model, for dike breach discharge compared to classical empirical formulas. The models are tested for both interpolation (within the training set space) and extrapolation (beyond the training set space).

The remainder of the paper is organized as follows: the experimental data and the machine learning model are detailed in Section 2. In Section 3, results are presented and discussed to assess the performance of each modelling approach in both interpolation and extrapolation. Finally, conclusions are drawn in Section 4.

2. Method

2.1. Experimental data

Data used in this work were collected from laboratory experiments previously presented by Rifai et al. (2019) and Schmitz et al. (2021). In total, 43 tests are considered here (Table 1). A detailed description of the experimental setup and data acquisition systems are available in Rifai et al. (2019) and Schmitz et al. (2021). The experimental setup consists of a horizontal trapezoidal straight main channel with a 3-m-long trapezoidal erodible dike along its right side. A horizontal floodplain was present beside the dike, at the same elevation as the non-erodible riverbed. A calibrated perforated plate was placed at the downstream end of the main channel such that, at the beginning of each test, the water level corresponded to the dike crest elevation. To trigger breaching, an initial notch was created in the dike crest. The water level in the main channel was recorded at a frequency of 10 to 50 Hz, depending on the test. The evolution of the breach top width was recorded in all experiments with a resolution of 1 cm (Rifai et al., 2020) and at intervals that range between 1 and 60 seconds, depending on the test phase.

Table 1. Experimental tests features. Froude numbers, F_{OT}, are computed in the main channel at the overtopping onset. S_d and S_u stand for the dike slope on the floodplain and the main channel side, respectively, while L_k is the dike crest width. The three tests corresponding to underlined values of the inlet discharge Q_{in} were repeated twice. The test corresponding to the bold underlined Q_{in} value was repeated three times.

S_d (-)	S_u (-)	L_k (m)	Q_{in} (l/s)			F_{OT}		
2.0	2.0	0.15	25	40	55	0.083	0.133	0.183
1.5	1.5	0.15	25	40	55	0.071	0.114	0.157
2.0	1.5	0.15	25	40	55	0.071	0.114	0.157
1.5	2.0	0.15	25	40	55	0.083	0.133	0.183
2.5	2.0	0.15	25	40	55	0.083	0.133	0.183
3.0	2.0	0.15	25	40	55	0.083	0.133	0.183
3.0	2.0	0.30	25	-	55	0.083	-	0.183
2.0	2.0	0.00	25	40	55	0.083	0.133	0.183
2.0	2.0	0.30	25	40	55	0.083	0.133	0.183
2.0	2.0	0.60	-	-	55	-	-	0.183
2.0	2.0	0.1	20	21	28	0.066	0.07	0.093
			30	31	**40**	0.100	0.103	**0.133**
			41	47	50	0.136	0.156	0.166
			51	55		0.169	0.182	

2.2. Generation of the machine learning model

Among decision-tree-based machine learning methods, the basic decision tree algorithm is simple to use and to interpret, but it tends to overfit the training data and returns highly discontinuous predictions for the target variable. To tackle those limitations, more advanced algorithms were developed, notably the extremely randomized trees method (Geurts et al., 2006). This method considerably reduces sensitivity to the training data set by adding randomness in the way node splits are defined. In this work, 1000 trees with a maximum depth of 15 were considered. The rest of the hyperparameters was set to the values recommended in the *scikit-learn* documentation, i.e., a machine learning library developed in Python (Pedregosa et al., 2011).

The breach discharge is known to highly depend on the main channel hydrodynamic state and the main channel, dike and breach shapes (Schmitz et al., 2021, 2023). Here, five features for the machine learning model were identified:

$$X = \left[z_{r,adim} = \frac{z_r}{w_{r,FS}}; \mathrm{F}; S_d; L_{k,adim} = \frac{L_k}{w_{r,FS}}; B_{top,adim} = \frac{B_{top}}{w_{r,FS}} \right] \quad (1)$$

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.

As the recording frequency of the breach geometry and the hydrodynamic variables did not match, a linear variation was assumed between two successive breach width recordings to obtain a breach width associated to each measured 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 target value and the input parameters. That way, about 1.5×10^6 experimental data points were obtained overall.

Here, no validation set is required as the value of the hyperparameters was fixed according to *scikit-learn* recommendations. 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. To avoid training and testing the machine learning model on data points emanating from identical or too similar laboratory experiments, i.e., data points encompassed in alike feature spaces, clustering was applied to the experimental tests using the OPTICS algorithm (Ankerst et al., 1999). The laboratory tests were segregated based on the main channel Froude number at overtopping initiation (when the water level reaches the dike crest), F_{OT}, the non-dimensional dike crest width at overtopping initiation, $L_{k,adim}$, and the dike slope on the floodplain side, S_d. The test set then corresponds to one cluster, while the rest of the data forms the training set.

Decision-tree-based machine learning models provide constant predictions when evaluated outside their training space, i.e., the target variable gets always the same value, leading to lower accuracy. Therefore, their predictive capability should be evaluated in interpolation and extrapolation separately. Clusters were partitioned in two different groups: those considered in extrapolation, i.e., located on the edge of the data set space, and those in interpolation.

To assess the predictive capability of the numerical model, the mean absolute relative error (MARE) on the breach discharge is computed when successively considering each cluster as the test set, so that

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|Q^i_{b,exp} - Q^i_{b,num}\right| \tag{2}$$

with n the number of data points in the considered test set, $Q^i_{b,num}$ the breach discharge value obtained from the machine learning model fed with features associated to data point i, and $Q^i_{b,\exp}$ the associated experimental value. The averaged value of MARE obtained for all the different test sets in interpolation or extrapolation illustrates the global performance of the machine learning model.

3. Results and discussion

Empirical formulas for the breach discharge usually focus on the determination of the discharge coefficient, C_D, involved in the following formula:

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

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]. Among empirical side weir formulas, the ones proposed by Jalili and Borghei (1996) and Singh et al. (1994) are particularly relevant when considering realistic fluvial dike breaches (Schmitz et al., 2024). 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 trapezoidal with its side slopes equal to the repose angle of the dike wet material. Additionally, the 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 is chosen equal to the breach bottom width.

Additionally, a corrective term computed through the extremely randomized tree algorithm may be applied to these experimental formulas, leading to an improved evaluation of the breach discharge, denoted Q_b+. The target variable of this machine learning 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 are considered. Table 2 summarizes the MARE obtained with the extremely randomized trees

model used for the direct evaluation of Q_b, with both empirical formulas, and with each empirical formula combined to a machine learning-based corrective term.

Table 2. Averaged mean absolute relative error (MARE) [%] for each model when evaluated in interpolation and extrapolation.

Numerical approaches	Q_b		Q_b^+	
	Interp.	Extrap.	Interp.	Extrap.
Extr. randomized trees	4.5	7	-	-
Singh et al. (1994)	36.7	38.1	5.5	9.9
Jalili and Borghei (1996)	75.9	68.5	9.9	15.4

As expected, the performance of the different models decreases when evaluated in extrapolation. Notably, both empirical formulas lead to poor fidelity predictions. However, results are greatly improved when adding a machine learning-based corrective term to the empirical formulas. Overall, very satisfactory results are obtained in both interpolation and extrapolation when using the extremely randomized trees model alone and when adding a machine learning-based corrective term to the empirical formula derived by Singh et al. (1994).

4. Conclusion

The extremely randomized trees algorithm was compared to empirical formulas for the prediction of Q_b. Empirical formulas were less accurate. A machine learning-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.

In all cases, the model performance decreased slightly when evaluated in extrapolation, i.e., outside the space of the training set. It is expected that the further the ML models are evaluated from their training space, the lower the model performance. Therefore, the training set space should be extended in future works to cover a broader range of dike configurations. Also, other parameters should be varied, such as the dike height, material median diameter, or cohesion. Finally, physics-guided deep learning methods might be promising to incorporate physics in the learning process of machine learning models (Wang and Yu, 2023).

References

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.

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

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.

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.

Jalili, M.R., Borghei, S.M., 1996. Discussion: Discharge Coefficient of Rectangular Side Weirs. Journal of Irrigation and Drainage Engineering 122, 132–132.

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.

Neary, V.S., Sotiropoulos, F., Odgaard, A.J., 1999. Three-Dimensional Numerical Model of Lateral-Intake Inflows. Journal of Hydraulic Engineering 125, 126–140.

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.

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.

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.

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

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