**Improving Prediction of Dam Failure Peak Outflow using Neuroevolution Combined with *K*-means Clustering**

Amir Hossein Eghbali, Assistant Professor, Department of Civil Engineering, Islamshahr Branch, Islamic Azad University, Tehran, Iran

Kourosh Behzadian\*, Senior Lecturer, Ph.D., C.Eng, School of Computing and Engineering, University of West London, London, UK, W5 5RF

Farhad Hooshyaripor, Research Assistant, Environmental Science Research Centre, Islamshahr Branch, Islamic Azad University, Tehran, Iran

Raziyeh Farmani, Associate Professor, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK, EX4 4QF

Andrew P Duncan, Associate Research Fellow, Ph.D., College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK, EX4 4QF

\* Corresponding author, Email: kourosh.behzadian@uwl.ac.uk, Tel: +44 (0) 20 8231 2466

**Improving Prediction of Dam Failure Peak Outflow using Neuroevolution Combined with *K*-means Clustering**

**Abstract**

Estimation of peak outflow resulting from dam failure is of paramount importance for flood risk analysis. This paper presents a new hybrid clustering model based on Artificial Neural Networks and Genetic Algorithm (ANN-GA) for improving predictions of peak outflow from breached embankment dams. The input layer of the ANN-based model comprises height and volume of water behind the breach at failure time plus a new parameter of ‘cluster number’. The cluster number is obtained from partitioning the input data set using *K*-means clustering technique. The model is demonstrated using the data samples collected from the literature and compared with three benchmark models by using cross-validation method. The benchmark models consist of a conventional regression model and two ANN models trained by non-linear techniques. Results indicate that the suggested model is able to estimate the peak outflows more accurately especially for big flood events. The best prediction for the current database was obtained from a five-clustered ANN-GA model. The uncertainty analysis shows the five-clustered ANN-GA model has the lowest prediction error and the smallest uncertainty.

**Keywords:** Artificial neural networks; dam failure; genetic algorithm;hybrid model; *K*-means clustering.

# Introduction

Embankment failures usually happen during hydrologic flash floods when the flood discharge greatly exceeds the maximum capacity of spillways or due to uncontrolled seepage through the embankment body and progressive internal erosion. Due to the potential risk of dam failure and the corresponding hazards to inhabited areas downstream, analysis of dam failure and the consequent damage resulting from the flood wave is of paramount importance to researchers, engineers and insurers. Because of numerous loss of life and the high level of economic damage in some past failure cases, dam failure can be considered to be one of the most catastrophic phenomena in the world. A brief review of some historical dam failure cases reveals the considerable range of hazards and subsequent financial and human losses. For instance, the South Fork Dam in Pennsylvania, USA, failed in 1889 due to overtopping in which over 2200 people died. It also resulted in significant property losses (Singh and Scarlatos, 1988). Breaching dikes in the Netherlands in 1953 due to a heavy coastal storm surge has been one of the biggest natural disasters in Dutch history, which caused thousands of lost lives and a direct economic loss of about 14% of the Dutch GDP (Huisman et al. 1998). Hence, finding reliable methods for assessing the dam failure and its consequences are critical for analysis of both flood risk and dam safety.

Estimation of the probable flood under uncertain hydrologic conditions and routing the flood wave through downstream rivers could provide invaluable information for decision makers. However, the accuracy of assessment of flood outflow and corresponding damage are heavily reliant on the appropriate calculation of the outflow hydrograph caused by a dam break (Wahl 2010). In addition, dam failure risk assessment (DFRA) outlines two primary tasks: 1) to analyse the feasible dam failure scenarios and 2) to compute a more realistic flood wave. These tasks involve: 1) prediction of the reservoir outflow hydrograph and 2) routing of this boundary condition through the tail-water areas (Pilotti et al. 2010). The current study has addressed the first task of DFRA focusing on estimation of peak-discharge from embankment dam failures. This can then be used in flood routing methods to estimate peak flow rates at locations downstream from a breached embankment dam (SCS, 1981, Costa, 1985), Barker and Schaefer (2007), and Environment Agency 2014). More specifically, the estimated outflow hydrograph feeds the mathematical routing models as the main input parameter to produce water levels and flow velocities at downstream locations (Thornton et al., 2011). The hydraulic routing of large floods is a well-established science while modelling the prediction of an outflow hydrograph is a complex task due to the sources of uncertainty involved (Wahl 2010). Each of these methods requires an accurate estimation of the maximum outflow rate from the reservoir. In other words, an effective estimation of the peak outflow as the main parameter of a dam failure outflow hydrograph requires the specifications of the dam such as breach geometry, dam geometry and dam materials. Improvement of the quality of this estimation has been addressed by multiple research works (e.g. Pierce et al. 2010; Thornton et al. 2011; Gupta and Singh 2012; Froehlich 2016).

Application of data mining techniques such as statistical methods for prediction of the peak outflow has received a lot of attention due to the desire to reduce mathematical complexity. Development of data mining methods is highly dependent on the data samples and their quality. While the number of samples from historic dam failures is rather small due to it being a relatively rare phenomenon, the existing data samples from real breached embankment dams are of high value (Thornton et al., 2011). This fact has not forced researchers’ to abandon attempts to develop predictive models of dam peak outflows. A number of researchers have developed such prediction models using traditional statistical regression methods and artificial intelligence techniques (e.g. neural networks and genetic programming). These predictive models have been used to estimate a variety of outflow parameters such as breach width and depth, flood hydrograph peak, and time to peak. Application of various multivariate regression analyses has been a common approach in recent years for prediction of peak discharge such as Froehlich (1995) with 32 data samples, Froehlich (2008) with 74 data samples, Pierce at al. (2010) and Thornton et al. (2011) with 87 data samples. The available historic data sets have also been expanded using statistical methods such as the copula technique to generate synthetic samples (Hooshyaripor et al. 2014).

In addition, the lack of availability of various geometric elements in the historic databases points to likely inconsistency within and between the databases which may result in significant uncertainties in the statistical analyses (Hanson et al. 2005). Wahl (2004) showed that the uncertainty involved in the estimates by a number of regression models could range from ±0.32 to ±1. Overall, in spite of numerous historic data samples, predictions may not be sufficiently accurate. This can be attributable to some factors such as complexity of the dam-break phenomenon and limitations of the commonly used statistical methods (Pierce et al. 2010).

Artificial intelligence (AI) techniques have been widely used for improved accuracy of approximation of unknown functions. To overcome some of the above shortcomings, these were applied by hydraulic researchers (Babaeyan Amini et al. 2011; Hooshyaripor et al. 2014; Hakimzadeh et al. 2014). For instance, Babaeyan Amini et al (2011) used the assembled data set by Wahl (1998) to predict peak outflow from breached embankments using Artificial Neural Network (ANN) and Genetic Algorithm (GA) methods. GA is a widely used evolutionary algorithm in many engineering disciplines with a successful application in flood management strategies (Javadi et al. 2005). Nourani et al. (2012) applied ANN with 24 experimental samples comprising 7 variables to investigate peak outflow. Sattar (2014) used 51 historical samples for peak outflow prediction, 63 data samples for dam breach width prediction, and 36 data samples for failure time prediction with Gene Expression Programming (GEP). Hakimzadeh et al. (2014) also applied Genetic Programming (GP) to those 24 experimental samples, which were used by Nourani et al. 2012. Hooshyaripor et al. (2014) showed that a better performance can be achieved by using an ANN model when compared with linear regression analysis when a richer database is used. Advantages of the ANN models over linear statistical methods can be explained by factors such as their data-driven nature, model–free form of predictions, tolerance to data errors, and lower uncertainty for prediction (Hooshyaripor et al. 2015).

The ANN models need to be trained before use for prediction purposes. During the ANN training, the optimal ANN parameters including values of weights and biases are identified. Back propagation (BP) based algorithms are historically the most widely used techniques for ANN training for optimising the ANN parameters (Varshney et al. 2014). However, these algorithms are hindered by inconsistent and unpredictable performances (Subramanian and Hung 1990). In addition, the abilities of gradient-based search techniques such as BP are generally limited and questionable when searching for globally optimal solutions. Global search techniques have been proposed as a potential solution to this limitation. (Subramanian and Hung, 1990). Various alternative techniques for optimising ANNs’ parameters have been used, such as nonlinear programming techniques and evolutionary algorithms. The combination of ANNs with evolutionary algorithms (also known as Neuroevolution) has been demonstrated to strengthen the model performance for other water systems applications ([Rivero](http://link.springer.com/search?facet-author=%22Daniel+Rivero%22) et al. 2009; Behzadian et al. 2009; Mulia et al. 2013). Therefore, neuroevolution has been chosen in the present study for performance improvement of the predictive ANN model for dam failure peak outflow by integrating it with GA for efficient training.

Obviously, the higher the accuracy of the input hydrograph, the more precise outputs would result in the routing models which significantly affect the ultimate risk management plan. Nowadays, data mining techniques are widely used by many researchers for estimation of the key parameters (e.g. peak value) of input hydrograph. Drawing upon the knowledge of the previous data mining models, this paper aims to enhance the quality of estimation of dam failure outflows by introducing a new neuroevolution approach in a data-driven model combined with a clustering method. The suggested model uses *K*-means clustering approach for dividing the dam failure database to a specific number of clusters with similar attributes. This improves the prediction accuracy. It also provides a more reliable way of training ANN by using a Genetic Algorithm and thus achieving global optimum rather than local optimum which is common in the previously developed ANN prediction models. Applications of data driven models have proved that imbalances within datasets can be alleviated by using an appropriate data clustering technique such as *K*-means clustering or fuzzy c-mean techniques (Hammouda and Karray, 2000; Arthur and Vassilvitskii, 2007; Kim and Seo, 2015). Data samples are partitioned into a number of clusters by using the *K*-means clustering method (Arthur and Vassilvitskii, 2007). The prediction performance of the developed model is then compared with a number of previously developed models as benchmarks.

# Methodology

Artificial Neural Networks (ANNs), genetic algorithm (GA) and *K*-means clustering have been used here as the core tools for prediction of peak outflow (*Qp*) of a breached dam. More specifically, the ANN curve fitting and GA optimisation tools in MATLAB® (R2014b) platform were combined together with the data, clustered by using the *K*-means clustering function in MATLAB (R2014b). The basic principle of the data mining approach mainly used in the literature for estimation of peak outflow is described below and then followed by describing the benchmark models and the hybrid ANN-GA model.

##  Prediction of peak outflows

Height and volume of a dam reservoir at failure time have been recognised as the main explanatory factors in most of the historic dam breach cases (Nourani et al., 2012). Since the main objective of this study is estimation of peak outflow (*Qp*) of a breached dam, these factors are included; i.e. water volume above the breach invert (*Vw*) and water depth above the breach invert (*Hw*) as shown in Fig. 1. Development of nonlinear regression relations fitted to the historic data is a well-established technique used by many researchers (Pierce at al., 2010). The most frequently used relation appears to be the one developed by Froehlich (1995) which was confirmed by Wahl (2004) as one of the best empirical relations that ever been developed. The general form of this relation which has been used here as an empirical approach is expressed in Eq. (1):

 (1)

where *Qp*= predicted peak outflow (m3/s); *Vw*= reservoir volume at the time of failure (m3); *Hw* = height of water in the reservoir at the time of failure (m); and *a*, *b*, and *c*= constant coefficients. These coefficients can be obtained by fitting observed and predicted variables to a training data set. The performance can then also be evaluated using a test data set.

##  Benchmark models

To compare the performance of the suggested ANN-GA model with other developed models, three conventional methods are used here as benchmark: (1) nonlinear multivariate regression models derived from statistical analyses (MVR hereafter), (2) traditional ANN trained with Levenberg-Marquardt algorithm, (ANN-LM hereafter); and (3) ANN trained with Generalized Reduced Gradient (GRG) method, (ANN-GRG hereafter). The general form of the empirical relation proposed by Froehlich (1995) in Eq. (1) can be viewed as one of the best forms of the MVR model and is adopted here as the benchmark MVR model to predict the peak failure outflow. ANNs are model-free universal function estimators which can be trained to learn correlated patterns between input data set and corresponding target values (Cybenko 1989). A standard ANN architecture composed of three layers: input layer, hidden layer and output layer is used here. After 10 times iteration of the standard ANN model runs with different number of neurons (between 1 and 4) in the hidden layer, no improvement was observed for hidden layers greater than one and hence one hidden layer was selected. In addition, other similar studies show that ANN models with one hidden layer are sufficient to approximate a complex nonlinear function (Cybenko, 1989). Other applications of ANN models in hydrology and flood predictions have resulted in the best satisfactory state using a single hidden layer (Nourani et al. 2012; Noori and Hooshyaripor 2014; Hooshyaripor et al. 2015). The input layer provides input data for the prediction model and includes two conventional variables of *Vw* and *Hw* in the benchmark ANN models. The output layer consists of a single neuron that represents the prediction of the peak outflow.

The ANN-LM is used here as a benchmark model, which is trained by using the non-linear Levenberg-Marquardt (LM) algorithm (Noori et al. 2010). It is based on a feed-forward neural network with a single hidden layer and is set up with two input variables (*Vw* and *Hw*) and one output variable (*Qp*). The LM algorithm gradually updates the ANN weights and biases as shown here in Eq. (2) for updating weights:

 (2)

where ; *J*= the Jacobian matrix of the error vector *ei*­evaluated in *W*; diag= the diagonal matrix consisting of the diagonal elements of *JTJ*; and = the gradient of the *ei* value. The vector error *ei* is the error of the network for *i*th data sample (*ei* = *yip* - *yio*). The parameter *λ* is the damping factor adjusted during every iteration.

ANN-GRG is used as a second benchmark with the same input and output layers but employs the GRG Algorithm to train the ANN. GRG is a nonlinear program to compute a least square errors solution developed by Lasdon et al. (1974). The GRG algorithm can be categorised as a nonlinear extension of the simplex method. The algorithm solves systems of nonlinear equations at each step to maintain feasibility by selecting a search direction and then a line search for each iteration (Lasdon et al. 1978).

##  Hybrid ANN-GA combined with K-means clustering

The novel ANN-GA model is proposed, in which conventional ANN training techniques used for the benchmarks are replaced with the GA as a global optimisation tool to find the best values for the ANN’s weights and biases, aims to overcome limitations of the gradient descent based methods (e.g. error back-propagation) which suffer from the possibility of being trapped in suboptimal (i.e. local) error minima and thus cause the algorithm to converge prematurely (Vishwakarma, 2012). In other words, The ANN-GA model uses an evolutionary algorithm (i.e. genetic algorithm) for training the ANN. This evolutionary algorithm is more robust than the nonlinear methods in the conventional ANN for achieving near optimal parameters in the training process. More specifically, the GA enables global optimal solutions to be located by using a population of candidate solutions to explore the objective space concurrently as well as not relying on a gradient-based approach. Instead, the population proceeds through consecutive generations in which fitter solutions are selected as "parents" to be used to combine and produce offspring solutions for the next generation using crossover and mutation operators.

Due to significant differences between the peak outflow values found in the existing database collected from various case studies (i.e. values with different orders of magnitude) (Wahl 1998), clustering of the input data set using the *K*-means method (Hammouda and Karray, 2000) is proposed. More specifically, a cluster identifier number (*Nc*) obtained from clustering input data samples (i.e. *Vw* and *Hw*) is proposed as a new input variable which represents the group to which that sample belongs (Fig. 2). The *K*-means clustering method is a process of partitioning a data set into a specific number of *K* groups based on the proximity to the corresponding cluster centres (Kim and Seo, 2015). The algorithm used here to identify the cluster centres minimises the objective function of dissimilarity measure which is considered as the Euclidean distance (Arthur and Vassilvitskii, 2007). Hence, the objective function of *K*-means clustering can be expressed as:

|  |  |
| --- | --- |
|  | (3) |

where *x* = (*x1, x2, …, xn*) = set of inputs where each observation can be multi-dimensional real vector; *K* (≤ *n*) = the number of clusters; *S* = {*S1, S2, …, Sk*} = set of clustered data; *μi* =the mean of points in *Si*. The *K*-means clustering method relies on a pre-specified number of clusters as key feature. Therefore, the *K*-means clustering function for different numbers of clusters need to be analysed to partition the input data set (i.e. *Vw* and *Hw*) to identify the most appropriate number of clusters.

Fig. 2 illustrates the ANN architecture which includes the number of layers and neurons. The above-mentioned additional input feature (not included in the benchmark models): the cluster number to which each data sample belongs (*Nc*) is included in the input layer. The normalised model output (*yp*) is calculated, as defined in Eq. (4), based on the normalised input variables (*xi*), constant trained ANN parameters and two transfer functions, i.e. tan-sigmoid and Purelin (linear) which are used in the hidden and output layers, respectively:

|  |  |
| --- | --- |
|  | (4) |

where *xi*=*i*th input variable; *Nin*=the number of input variables (here *Nin=3* given three input variables of *Vw*, *Hw* and *Nc*); *M*=the number of neurons in the hidden layer; =weight of *i*th input variable and *j*th hidden neuron; =weight for the output layer's input from the *j*th hidden neuron; = bias of *j*th hidden neuron; = bias for output neuron. To guarantee that the ANN is able to approximate any continuous function, the number of neurons (*M*) in the hidden layer is determined to be 4 based on a trial and error, so as to be smaller than *2Nin+1* (i.e. 7)as suggested by Hecht-Nielsen (1987). Given the known values for *M*, *Nin* and *Nout* (i.e. 1 for one output variable of *Qp*), the total number of decision variables equals *M×(Nin+Nout+1)+Nout* (i.e. *Nvar=21)* including *M×(Nin+Nout)* weights and *M+ Nout* biases.

##  Assessment of performance indicators

After identifying the model’s parameters using a training data set, the overall model performance is validated for a test data set which is ‘unseen’ data during model training. Conventional model validation involves in its simplest form, dividing a database into two subsets, for example, 70% training and 30% test observations (Garthwaite and Jolliffe, 2002). However, if sufficient data are unavailable or there are no appropriate spread of data when partitioning into separate training and test sets, the model error (e.g. root mean square error) on the test data set in the conventional validation method may not be a true representation of the model performance (e.g. the error in training performance may be much larger than the test performance error). To overcome this drawback and present true prediction performance, the cross-validation method can be used in which all data samples participate in the evaluation of the test set (Grossman et al. 2010).

Given the small size of observed data for dam breaks, the *m*-fold cross-validation method (Kohavi, 1995; Vasios et al. 2004) is used here for assessment of the predictive ability of the analysing models. The *m*-fold cross-validation method is an extension of conventional model validation in which, instead of dividing the database into two subsets, it is divided into *m* subsets whose size are as nearly equal as possible. One subset is selected as the test set and the union of the remaining *m-1* subsets form the training set. Then, the model is repeatedly re-trained and its performance is evaluated *m* times, each time using a different data fold as the test set (Stone, 1974; Hjorth, 1993). The overall performance of the *m* validated models is calculated by averaging all *m* individual performance values. The value of *m* between 3 and 20 is often used (Hjorth, 1993); here *m* is assumed to be10 as suggested by Kohavi (1995), in which the union of 9 data-folds (i.e. 90% of data) is allocated for training and the one remaining fold (i.e. 10% of data) is used for test. The process is repeated ten times with a different test data fold in each case. The model performance is evaluated based on various statistics obtained from data samples for each test data subset. Five statistical indices which have been commonly employed in hydrologic and water models are used here for evaluation of the results: Root Mean Square Error (*RMSE*), Relative Square Error (*RSE*), coefficient of determination (*R2*), Nash-Sutcliffe efficiency (*NSE*), and RMSE-observations standard deviation ratio (*RSR*) with the following mathematical equations (Seibert, 2001; Hooshyaripor et al. 2015; Sattar, 2014; Moriasi et al. 2007; Behzadian and Kapelan, 2015):

|  |  |
| --- | --- |
|  | (5) |
|  | (6) |
|  | (7) |
|  | (8) |
|  | (9) |

where *yio*= observed variable for test sample *i*; *yip*= predicted variable for test sample *i*; = mean observed value for test sample*s*; = mean predicted value for test sample*s*; and *n* = the number of test data samples. Note that the *RMSE* measure is also used here as the fitness function for training all the ANN-based models. The NSE measure is sensitive to extreme values and might yield sub-optimal results when the dataset contains large outliers in it. NSE values between 0.0 and 1.0 are generally viewed as acceptable levels of performance with the optimal NSE value of 1 (Moriasi et al. 2007). RSR also varies from the optimal value of 0 to a large positive value. The lower the RSR, the better is the model simulation performance (Legates and McCabe 1999).

# Data collection and sample data set

The current database used to develop the data-driven model includes real-scale samples of recorded dam failures that are limited in the world due to the fact that the dam failure per se is a rare phenomenon (Wrachien and Mambretti 2009). In addition, data collection under dam breaching conditions is a risky operation and whatever data is collected is of high value (Gupta and Singh 2012). Having said this, there are a number of data-driven models which have been developed based on between 20 to 108 cases of dam failures with different parameters (e.g. Pierce et al. 2010; Thornton et al. 2011; Froehlich 2008 and 2016). These databases which are of great value have been used frequently in research works to develop empirical relationships for prediction of different aspects of dam failure (e.g. geometry of the breach, failure time and characteristics of the produced flood wave) based on the physical characteristics of dam reservoirs. For instance, Thornton et al (2011) proposed a multivariate regression model for prediction of peak discharge through breached dam embankments based on the data from 87 dam breach cases. Other empirical models for prediction of peak discharge have been developed based on 32 data samples by Froehlich (1995), 74 data samples by Froehlich (2008), and 87 data samples by Pierce at al. (2010). Examples of developing other data-driven models are Genetic Programming (GP) using 24 experimental samples by Hakimzadeh et al. (2014); Artificial Neural Network (ANN) with 24 experimental samples by Nourani et al. (2012); Fuzzy Logic using 69 historical datasets for prediction of dam breach width by Elmazoghi (2013) and Gene Expression Programming (GEP) with 51 historical samples by Sattar (2014).

Valuable information is available from various real embankment dam failures that have been documented by previous researchers. Wahl (1998) assembled a database comprising of 108 embankment failures from investigation of different case studies. During 1980s and 1990s, several researchers compiled valuable databases of well documented case studies in order to develop predictive relations for breach peak outflows (e.g. Singh and Snorrason, 1984; Froehlich, 1995; Xu and Zhang, 2009). Sattar (2014) collected a database of 140 embankment dam failures from a variety of sources containing various hydraulic, geometric, and geomorphic parameters. The database although is the most comprehensive in quantity, contains missing data for some parameters in some of the cases presented. However, a database needed to be prepared, in which all the main parameters of interest were available. Drawing upon all available databases in the literature, the current study collected all data samples from the cases in which the required parameters of dam failure (i.e. *Vw*, *Hw* and *Qp*) were available. Consequently, a database of 92 dam failures (Appendix I) was compiled from the previous studies of Singh and Scarlatos (1988), Wahl (1998), Taher-shamsi et al. (2003), Xu and Zhang (2009), Pierce et al. (2010) and Sattar (2014). The model input parameters of *Vw* and *Hw* vary from 0.0037mcm to 660mcm and from 1.37m to 77.4m, with average of 30.67cms and 15.7m, respectively. Similarly, the output parameter *Qp* ranges between 2.12 m3/s and 78100 m3/s with average of 4690.5 m3/s. According to the classification of dam size (Singh, 1996), 50% of these cases with available dam height are categorised as large dams. Thus a good combination of different dam sizes can be found in the database.

Based on the cross-validation technique outlined above, the database including 92 samples was first divided randomly into 10 subsets (the statistics of peak outflow for these subsets are given in Table 1). Then, union of 9 randomly selected subsets without replacement constituted 10 different training data sets (i.e. 90% of data) and the remaining 10% of data in each time was used as a test data set. Thus, each model was built 10 times, each time with a different training data set including 82 or 83 dam failures and evaluated with a different test data set including 10 or 9 dam failures, respectively. The test data sets included 9 dam failures for eight of the models trained with 83 dam failures; 10 dam failures were used for testing the two remaining models trained with 82 dam failures. Thus, all 92 dam failures from the database were used as unseen data during one of the test stages for the 10-fold cross-validation modelling of the prediction of dam failure. This guaranteed that all test data samples taken into account for evaluation were unseen.

# Results and discussion

During the training process of the ANN-GA model, the ANN parameters were first optimised. The optimisation parameters in GA have to be set first before performing the optimisation runs as they can have major impact on the optimisation speed in finding optimal solutions and their quality. There is no systematic way to identify the optimal values of these parameters. The GA is first rigorously analysed in a number of trial runs with a combination of parameters in the recommended ranges to identify the optimal parameter values. The following GA parameters were identified after a maximum of 5 trial runs for each GA parameter : population size of 200; roulette selection operator; single point crossover and mutation by gene operators both with uniform function and a probability of 0.7 and 0.05, respectively. The stopping criteria were set as convergence to the best fitness value of a generation to *RMSE* less than 10-7 or the maximum number of 1000 generations having been reached. Also, note that in order to avoid unduly dominating effects of trapping in local minima, especially for non-linear algorithms, the model runs were iterated maximum 10 times with different number of neurons in the hidden layer to attain the best results. The obtained results show that the best model performance can be achieved by using 4 neurons in the hidden layer and therefore network structure of 3-4-1 shown in Fig. 2 was considered as the structure for the ANN-GA models.

All the aforementioned models were only analysed based on two input variables of *Vw* and *Hw*. To assess the performance of the four models, each model was built 10 times using different training data sets. For instance, Fig. 3 shows the scatter of estimated and observed values in the four analysed models for the first data subset. It is apparent in Fig. 3a and 3b that estimations in all the models are far better for the large peak outflows than small values which have been overestimated compared to the observed values. This can be attributed to the RMSE indicator in all models which minimise the overall error for all predictions including both low and high values. The apparent difference of the prediction estimations for these two data groups (i.e. the low and high values) can also be influenced by two main factors: the significant difference in 1) the order of magnitude and 2) the number of the data. As a result, the prediction errors obtained from the overestimations of high values have been balanced on the underestimations of prediction errors in low values. The impact of these two features also seems to be the main reason for large underestimations of low values compared to small overestimations of high values.

Fig. 4 shows the performance of the analysed models with respect to the average of the three indicators (RMSE, RSE and R2) in 10 instances of test data subsets obtained from the cross-validation technique. The results show that the performance of the ANN-GA model is better than other models for all three indicators for both training and test data (i.e. lower values for RMSE and RSE and higher values for R2). Average value of RMSE for the ANN-GA model is less than 5000 m3/s while the value of RMSE for other models are between around 6000 and 8000 m3/s (i.e. between 16% and 38% improvement for the ANN-GA model). The same rate of improvement can be seen for other indicators. Comparison of the two nonlinear models in which a gradient-based optimisation is used for training (i.e. ANN-LM and ANN-GRG) shows that the LM is able to predict the peak outflow better in respect of all indicators. Some negative peak outflows were also observed for some cases in the ANN-LM model; these are assumed to be equal to zero. Such issue was easily removed in the ANN-GRG and ANN-GA models by adding a new constraint for avoiding negative prediction in the training phase. Therefore, the ANN-LM performance for predicting peak outflows of small dams would be slightly inappropriate. By comparing training and test indicators between ANN-GRG and MVR models, it can be seen that the prediction errors for training data in ANN-GRG are slightly better than the MVR model; whilst the prediction errors in test data increase in the ANN-GRG. Such problem can be linked to trapping of the GRG algorithm in local optima. This issue was also solved in the GA algorithm by finding global optima and thus prediction errors of the ANN-GA model were the minimum in both training and test data subsets.

Due to highly variable values for the peak outflow, further improvement of the ANN-GA model was analysed by adding the data clustering number as a new input variable in the ANN structure. Therefore, appropriate number of clusters is identified by partitioning the variables of the input data set (i.e. *Vw* and *Hw*) into different numbers of clusters (from 2 to 7 clusters) using *K*-means clustering technique (Arthur and Vassilvitskii, 2007). Fig. 5 illustrates an instance of the clustering results with 5 clusters for the input variables of *Vw* and *Hw* with associated cluster centre. As can be seen, the first three clusters (i.e. clusters #1, #2, and #3 in Fig. 5) have separated a few data points; each containing a group of dam failures with large water volume while the last two clusters (i.e. clusters #4, and #5 in Fig. 5) encompass the majority of the input data related to especially small water volume dam failures. The single data point in cluster 2 relates to the failure of the Teton dam in [Idaho](https://en.wikipedia.org/wiki/Idaho), United States. The specification of this dam is unique within the existing database. More specifically, the dam is the highest (i.e. 77m) and its capacity (i.e. 310 mcm) is the third in the database. Compared to the other two dams with the largest capacity (i.e. 607.5 and 660 mcm), the height of the Teton dam is almost double the height of the largest dams (i.e. 31 and 35.8 m). This unique specification has led to the creation of a cluster with only one member in five data clustering. In addition, these data points should be kept in the database (i.e. not removed as outlier) because there is limited number of data available and also the special attributes of some data points can be reflected as a unique cluster number. As a result, this can better help the suggested clustering-based model recognise different specifications and predict their peak failure flows special more accurately than the previously developed models. Moreover, the single data point in cluster 2, when it is used for training the model, is not used for predicting any new data point in the same cluster anymore. However, it will be out of other clusters and thus avoid compromising the prediction of data points in other clusters because each cluster number represents similar characteristics of dams.

Also note that the suggested neuroevolution approach develops only one predictive model based on the training data points in all clusters (i.e. 90% of all data samples) not based on the data points in one cluster. More specifically, the *k*-means clustering technique adds a new attribute (i.e. cluster number) to each data sample that is used as a new independent input in the ANN-GA model. The results in the paper show that the new attribute has been very efficient for improving the prediction accuracy. This can be due to the fact that cluster number for each data sample is indicative of a specific range for physical characteristics of the dam failures. In other words, all this will make the predicative model more intelligent and will help identify the failure peak outflow more accurately. In particular, the single data point in cluster#2 with its unique characteristics (i.e. relatively large dam, especially in height) is regarded by the model as a unique cluster number and hence the model is prevented from being compromised by this data point which can lead to some level of inaccuracy in the model prediction.

The cross-validation technique was used to evaluate the performance of the ANN-GA dam failure prediction for each cluster number. These results were then compared with those without data cluster to identify the best number of clusters. Fig. 6 shows the scatter of estimated and observed values of the ANN-GA model for different numbers of clusters within the cross-validation process. All the model predictions in Fig. 6 still suffer from overestimating for small peak outflows. However, the model predictions without the data cluster input feature are almost skewed towards overestimation for moderate and big peak outflows, whilst the problem has been relatively balanced in the clustering-based models. This can be attributed to the fact that clustering of different sizes of the dam failure events can provide fair predictions around the real values, through allowing the ANN hidden units to specialise their responses for different clusters.

Comparison of the statistics for the 5 aforementioned indicators can help to identify the optimum number of clusters. Fig. 7 shows the average of indicators in cross-validation technique for clustered and non-clustered predictions of the ANN-GA models. The model performance in both training and test data sets has improved by adding clustering data. More specifically, although the performances of the clustered models vary in different indicators, the best performance (i.e. the least errors and highest correlation coefficient) in the test step belongs to the five-clustered ANN-GA model. This type of ANN-GA model is able to improve 17% in RMSE (from 5260 to 4370) of test data predictions and 7.5% in R2 (from 0.80 to 0.86), NSE from -0.4 to 0.1, and 19% in RSR (from 1.18 to 0.95) compared to non-clustered model. The details of performance indicators of the five-clustered ANN-GA model for 10 data-folds cross-validation are also given in Table 2. It should also be noted that the model performance for the test phase for other numbers of cluster (i.e. 2, 3 and 7) had a considerably deteriorated accuracy for the predictions. Therefore, inappropriate selection of cluster number can result in a relatively poor performance of the dam failure predictions. The performance for other numbers of clusters (i.e. 4 and 6) contained both improvements and deteriorations - depending on indicator. Thus it was concluded that models with 4 and 6 clusters yielded relatively the same results as the ANN-GA model without clustering.

Fig. 8 presents the frequency percentage of the failure peak outflow and the interval RMSE of the predictions for the entire ‘unseen’ database for the three models (i.e. the MVR, non-clustered ANN-GA and five-clustered ANN-GA models). Fig. 8 also shows the performance of the suggested model and other models based on RMSE within different ranges of peak outflows predictions. As it can be seen, the performance of all analysed models for most ranges including small and medium peak outflows (i.e. ranging from 100 m3/s and less to 40,000 m3/s) are more or less in the same order of accuracy but the performance of the suggested model (i.e. five-clustered ANN-GA) is significantly better (i.e. low RMSE) than others for the intervals of large peak outflows (i.e. between 40,000 and 80,000 m3/s). The improved accuracy of the suggested model can be attributed to clustering of input data and considering the influence of their clusters on the prediction of peak outflow. As this interval only accounted for a small percentage of the entire data, the result showed that the 5-clustered ANN-GA models were able to recognise these events efficiently and thus provide more accurate predictions with lower errors. Although the prediction accuracy of both ANN based models were lower than the MVR model for intervals with low peak outflows, its significance could be overlooked for intervals with large peak outflows especially for the five-clustered ANN-GA model. Overall, the MVR model was comparable with the non-clustered ANN-GA model. However, addition of cluster number to the input feature set considerably improved the accuracy performance of the ANN-GA model for most of the intervals. Further, to analyse and confirm the performance of the suggested model, uncertainty analysis on the prediction values was investigated using the methodology presented by Wahl (2004) and Pierce et al. (2010). The uncertainty of predictions was calculated using two measures: 1) mean prediction error in logarithmic scale (*ē*) and the 95% confidence limits around the mean predicted value (±*2Se*), which can be calculated by Eqs. (10) and (11), respectively:

 (10)

 (11)

where *yo* and *yp* = observed and predicted values, respectively; *σe*=standard deviation of the prediction errors after excluding outliers. Negative and positive mean values in the confidence limits indicate the underestimation and overestimation of the predictors over the observed values, respectively. Fig. 9 illustrates the result of uncertainty analysis for only different forms of the clustered and non-clustered ANN-GA models since the prediction error of other methods were relatively higher than the suggested models. As can be seen, the lowest prediction error and smallest uncertainty was achieved again in the five clustered ANN-GA model with a mean prediction error of 0.199 and width of uncertainty of 0.84. In most of the cases the uncertainty indicators show improved performance for the models, which used clustered inputs.

The mean features of the model with the best performance (i.e. five-clustered ANN-GA), obtained from the above analysis, are presented for future reference. Table 3 provides the values of the weights and biases of the ANN illustrated in Fig. 2 when 5 clusters are used. Table 4 represents the main clustering features for the five-clustered ANN-GA model. The 5 clusters in Table 4 are represented by cluster centre values for both variables of the peak outflow predictor (i.e. *Vw* and *Hw*). Given these two variables for prediction of the peak outflow in a new case, the relevant cluster number can be identified by calculating the Euclidian distance between those variables and each of the corresponding cluster centres in Table 4. The cluster number with the least Euclidian distance is selected as the input value for the cluster number.

# Summary and conclusions

The paper presented a data driven predictive model based on ANN combined with GA for training as an alternative to the both the common gradient based ANN training approaches and other statistical models for prediction of peak outflow from breached embankment dams. Compared to the previous research works, the paper has developed a new (neuroevolution) approach in a data-driven model combined with a clustering technique for a more accurate estimation of peak discharge of a dam failure. The suggested approach can be efficiently used for prediction of reservoir outflow hydrograph outlined by DFRA with higher accuracy especially for large embankment dams that can cause immense destruction and numerous fatalities when they fail. The ANN-GA model was first compared with three models including two ANN-based models (trained using nonlinear Levenberg-Marquardt and Generalized Reduced Gradient algorithms) and multivariate regression (MVR). These models required the height and volume of water behind the breach at failure time as the inputs. The *K*-means clustering technique was applied to the data set to generate cluster number as an additional input feature in the ANN-GA model. The data samples for 92 cases in which all the input-output variables (including height and volume of water behind the breach and peak outflow) were collected from the literature and the performance of the analysed models were evaluated using cross-validation technique. Finally, the uncertainty analysis was applied to investigate the uncertainty of the predictions in the proposed ANN-GA model. Based on the results obtained, the following can be concluded:

1. The results show that the overall performance of the ANN-GA is better than other models with respect to all three indicators for both training and test data (i.e. lower values for RMSE and RSE and higher values for R2).
2. The predictions of the ANN-GA model without clustering are biased towards overestimation for moderate and big peak outflows where there are a few data available. Performance for these is relatively balanced for the clustering-based ANN-GA models.
3. Use of *K*-means clustering dataset pre-processing and adding cluster number as an additional input feature in the ANN-GA model considerably improves the performance of the prediction of the peak dam failure; minimises error and uncertainty and maximises correlation coefficients. The problem with the low peak outflow predictions has been alleviated by data clustering. The five-clustered ANN-GA model resulted in the best performance.
4. All the models have far better estimations for the large peak outflows than for small peak-outflows which have been overestimated compared to the observed values; although the regression based model had a better performance for small values.

A procedure is proposed for estimation of peak outflow for any new embankment dam data sample based on recognition of membership of the most appropriate cluster from the database. However, it should be noted that if the database changes, the analysis should be repeated to identify: 1) the new weights and biases for the ANN-GA model and 2) the appropriate number of clusters which can result in the best performance for the test data set. It is also recommended that the aforementioned approach be developed for prediction of other dam breach parameters using similar clustering techniques.

# List of Notation

The following list of acronyms is used in the paper:

|  |
| --- |
| *Fd*= dam failure mode |
| *Hd*= dam height (m) |
| *Vd*= dam capacity (mcm) |
| *Vw*= water volume above the breach invert (mcm) |
| *Hw*= water depth above the breach invert (m) |
| *Qp*= Peak outflow (m3/s) |
| *Nc*= cluster number |
| *O*= overtopping failure |
| *W*= wave action failure |
| *S*= sliding failure |
| *P*= piping failure |
| *RSE*= relative square error |
| *RMSE*= root mean square error |
| *R2=* coefficient of determination |
| *yo*= observed value of peak outflow |
| *yp*= predicted value of peak outflow |
| *σe*= standard deviation of the prediction errors |

# Acknowledgments

The authors gratefully acknowledge the financial and other supports of this research provided by the Islamic Azad University, Islamshahr branch, Tehran, Iran and also wish to thank the anonymous reviewers for making constructive comments which substantially improved the quality of the paper.

# References

Arthur, D., and Vassilvitskii, S. (2007). “K-means: The Advantages of Careful Seeding. SODA ‘07“ *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. 1027–1035.

Babaeyan Amini, A., Nourani, V., and Hakimzadeh, H. (2011). “Application of artificial intelligence tools to estimate peak outflow from earth dam breach.” *Int. J. Earth Sci. Eng*., 4(6), 243-246.

Barker, B., and Schaefer, M. (2007). “Dam Safety Guidelines-Technical Note 1: Dam break inundation analysis and downstream hazard classification.” *Washington State Dept. of Ecology, Water Resources Program*, Dam Safety Office, Olympia, WA.

Behzadian, K., Kapelan, Z., Savic, D., and Ardeshir, A. (2009). “Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks.” *Environmental Modelling & Software*, 24(4), 530-541.

Behzadian, K. and Kapelan, Z. (2015). “Modelling metabolism based performance of an urban water system using WaterMet2” *Resources, Conservation and Recycling*, Vol. 99, 84–99

Costa, J. E. (1985). “Floods from dam failures.” Rep. No. 85-560, U.S. Geological Survey, Denver.

Cybenko, G. (1989). “Approximation by superposition of a sigmoidal function.” *Math. Cont. Sig. Syst.*, 2, 303-314.

Elmazoghi H. 2013. Fuzzy algorithm for estimating average breach widths of embankment dams, *Nat. Hazards.* 68:229–248.

Environment Agency. (2014). “Reservoir flood maps (RFM) guide: Explanatory note on reservoir flood maps for local resilience forums– version 3.” Rep. No. LIT 6882, Dept. for Environment, Food and Rural Affairs, Bristol, U.K.

Froehlich, D. C. (2016) “Predicting Peak Discharge from Gradually Breached Embankment Dam.” *J. Hydrol. Eng*., DOI: 10.1061/(ASCE)HE.1943-5584.0001424.

Froehlich, D.C. (1995). “Peak outflow from breached embankment dam.” *J. Water Resour. Plann. Manage.*, 121(1), 90–97.

Froehlich, D.C. (2008). “Embankment Dam Breach Parameters and Their Uncertainties.” *J. Hydraul. Eng.*, 134(12), 1708–1721.

Garthwaite, P.H., Jolliffe, I.T., and Jones, B. (2002). “Statistical inference”. *Oxford University Press*.

Grossman, R., Seni, G., Elder, J., Agarwal, N., and Liu, H. (2010). “Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions.” *Morgan & Claypool*.

Gupta, S. & Singh, V. (2012). Discussion of “Enhanced Predictions for Peak Outflow from Breached Embankment Dams” by Christopher I. Thornton, Michael W. Pierce, and Steven R. Abt.” *J. Hydrol. Eng*. 17(3), 463–466.

Hakimzadeh, H., Nourani, V., and Babaeian Amini, A. (2014). “Genetic Programming Simulation of Dam Breach Hydrograph and Peak Outflow Discharge.” *J. Hydrol. Eng*., 19(4), 757-768.

Hammouda, K., and Karray, F. (2000). “*A Comparative Study of Data Clustering Techniques*.” Fakhreddine Karray University of Waterloo, Ontario, Canada.

Hanson, G.J., Cook, K.R., and Hunt, S.L. (2005). “Physical modeling of overtopping erosion and breach formation of cohesive embankments.” *Trans. ASAE*, 48, 1783–1794.

Hecht-Nielsen, R. (1987). “Kolmogorov’s mapping neural network existence theorem.” *Proc. 1st IEEE Int Conf on Neural Networks*, San Diego, 3, 11-14

Hjorth, J.S.U. (1993). *Computer intensive statistical methods: Validation, model selection, and bootstrap*. CRC Press.

Hooshyaripor, F., Tahershamsi, A., and Golian, S. (2014). “Application of copula method and neural networks for predicting peak outflow from breached embankments.” *J. Hydro-Environ. Res*., [8(3](http://www.sciencedirect.com/science/journal/15706443/8/3)), 292–303.

Hooshyaripor, F., Tahershamsi, A., and Behzadian, K. (2015). “Estimation of Peak Outflow in Dam Failure Using Neural Network Approach under Uncertainty Analysis.” *Water Resources*, 42(5), 723-736.

Huisman, P., Cramer, W., and Van, E. G. (1998). Water in the Netherlands, Netherlands Hydrological Society, Netherlands: Delft.

Javadi AA, Farmani R, Tan TP. (2005) An intelligent self-learning Genetic Algorithm; development and engineering applications, Advanced Engineering Informatics, volume 19, pages 255-262, DOI:10.1016/j.aei.2005.07.003

Kim, S.E., and Seo, I.W. (2015). “Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers.” *J. Hydro-Environ. Res*., http://dx.doi.org/10.1016/j.jher.2014.09.006.

Kohavi, R. (1995). “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.” *Proc. of the 14th International Joint Conference on Artificial Intelligence,* 2(12), 1137-1143.

Lasdon, L.S., Fox, R.L., and Ratner, M.W. (1974). “Nonlinear Optimization Using The Generalized Reduced Gradient Method.” *Rev. Fr. Autom. Inf. Rech. Oper.,* 8, 73-103.

Lasdon, L.S., Waren, A.D., Jain, A., and Ratner, M. (1978). “Design and Testing of a generalized reduced gradient code for nonlinear programming.” *ACM Transaction on Mathematical Software*, 4(1), 34-50.

Legates, D.R., and McCabe. G.J. (1999). Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Res.* 35(1): 233-241

MATLAB® R2014b (The MathWork Inc., Massachusetts).

Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D., Veith, T.L. (2007). “Modelevaluation guidelines for systematic quantification of accuracy in watershedsimulations.” *Trans ASABE,* 50(3), 885–900.

Mulia, I.E., Tay, H., Roopsekhar, K., and Tkalich, P. (2013). “Hybrid ANN–GA model for predicting turbidity and chlorophyll-a concentrations.” *J. Hydro-Environ. Res.,* 7(4), 279–299.

Noori, R., Khakpour, A., Omidvar, B., and Farokhnia, A., 2010. “Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic.” [*Expert Systems with Applications*](http://www.sciencedirect.com/science/journal/09574174), 37(8), 5856–5862.

Nourani, V., Hakimzadeh, H., and Babaeyan Amini, A. (2012). “Implementation of artificial neural network technique in the simulation of dam breach hydrograoh.” *J. Hydroinformat*., 14(2), 478–496.

Noori, R. and Hooshyaripor, F., 2014. Effective prediction of scour downstream of ski-jump buckets using artificial neural networks, *Water Resources*, 41(1), 8-18.

Noori, R., Deng, Z., Kiaghadi, A., and Kachoosangi, F. (2015). "How Reliable Are ANN, ANFIS, and SVM Techniques for Predicting Longitudinal Dispersion Coefficient in Natural Rivers?." *J. Hydraul. Eng*., 10.1061/(ASCE)HY.1943-7900.0001062, 04015039.

Pierce, M.W., Thornton, C.I., and Abt, S.R. (2010). “Predicting peak outflow from breached embankment dams.” *J. Hydraul. Eng*., 15(5), 338-349.

Pilotti M., Tomirotti M., Valerio G. and Bacchi B. (2010). Simplified Method for the Characterizationof the Hydrograph following a Sudden Partial Dam Break” *J. of Hydraul. Eng.*, Vol. 136, No. 10, 693–704.

Rivero, D., Dorado, J., Fernandez-Blanco, E., and Pazos, A. (2009). “A genetic algorithm for ANN design, training and simplification.” *Proceedings of the 10th International Work-Conference on Artificial Neural Networks*, Bio-Inspired Systems: Computational and Ambient Intelligence, **1**, June 10-12, 2009, Salamanca, Spain, 391-398.

Sattar, A.M.A. (2014). “Gene expression models for prediction of dam breach parameters*.” J.. Hydroinformat.*, 16(3), 550-571.

SCS (Soil Conservation Service). (1981). “Simplified dam-breach routing procedure.” *U.S. Dept. of Agriculture*, Washington, DC.

Seibert, J., (2001). “On the Need for Benchmarks in Hydrological Modeling”, *Hydrological Processes*. 15: 1063-1064.

Singh, K.P., and Snorrason, A. (1984). “Sensitivity of outflow peaks and flood stages to the selection of dam breach parameters and simulation models.” *J. Hydrol*., 68, 295–310.

Singh, V.P., and Scarlatos, P.D. (1988). “Analysis of gradual earth dam failure.” *J. Hydraul. Eng*., 114(1), 21-42.

Singh, V.P. (1996). *Dam breach modeling technology*, Kluwer Academic, Norwell, Mass.

Subramanian, V., and Hung, M.S. (1990). “A GRG@-based System for Training Neural Networks: Design and Computational Experience.” *ORSA Journal on Computing*, 5(4), 386-394.

Stone, M. (1974). “Cross-validatory choice and assessment of statistical predictions.” *Journal of the Royal Statistical Society*, Series B (Methodological), 111-147.

Taher-shamsi, A., Shetty, A.V., and Ponce, V.M. (2003). “Embankment dam breaching: geometry and peak outflow characteristics.” *Dam Eng*., 14(2), 73-87.

Thornton, C.I., Pierce, M.W. and Abt, S.R. (2011). “Enhanced Predictions for Peak Outflow from Breached Embankment Dams.” *J. Hydrol. Eng.*, 16(1), 81–88.

Varshney, S., Chaurasiya, R., and Tayal, Y. (2014). “Character Recognition by Levenberg-Marquardt (L-M) Algorithm Using Back Propagation ANN.” *International Journal on Recent and Innovation Trends in Computing and Communication,* 2(8), 2368-2372.

Vasios, C.E., Matsopoulos, G.K., Ventouras, E.M., Nikita, K.S., and Uzunoglu, N. (2004). “Cross-validation and Neural Network Architecture Selection for the Classification of Intracranial Current Sources.” *7th Seminar on Neural Network Applications in Electrical Engineering*, NEUREL-2004, University of Belgrade, Yugoslavia.

Vishwakarma, D.D. (2012). “Genetic Algorithm based Weights Optimization of Artificial Neural Network.” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 1(3), 206-211.

Wahl, T.L. (1998). “Prediction of embankment dam breach parameters, A literature review and needs assessment.” *Rep. No. DSO-98-004*, Bureau of Reclamation, U.S. Department of the Interior, Denver, 60.

Wahl, T.L. (2004). “Uncertainty of predictions of embankment dam breach parameters.” *J. Hydraul. Eng*., 130(5), 389–397.

Wahl, T.L. (2010). “Dam breach modeling – an overview of analysis methods.” *Joint Federal Interagency Conference on Sedimentation and Hydrologic Modeling*, June, Las Vegas, NV.

Wrachien D.D and Mambretti S. (2009). “Dam-break Problems, Solutions and Case Studies”, *WIT press*, Southampton, Boston.

Xu, Y., and Zhang, L.M. (2009). “Breaching parameters for earth and rockfill dams.” *J. Geotech. Geoenviron. Eng*., 135(12), 1957-1970.

# Appendix: The database of failure dam

| No | Name  | *Hd* (m) | *Vd* (106×m3) | *Fd*1 | Above Breach | *Qp* (m3/s) | *Nc* | Reference |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  *Vw* (106×m3) | *Hw**(m)* |
| 1 | Apishapa, US | 34.1 | 22.5 | P | 22.2 | 28 | 6850 | 4 | Xu & Zhang, 2009 |
| 2 | Baldwin Hills, US | 49 | 1.1 | P | 0.91 | 12.2 | 1130 | 5 | Froehlich, 1995 |
| 3 | Banqiao, China | 24.5 | 492 | O | 607.5 | 31 | 78100 | 1 | Xu & Zhang, 2009 |
| 4 | Bayi, China | 30 | 30 | P | 23 | 28 | 5000 | 4 | Xu & Zhang, 2009 |
| 5 | Big Bay Dam, US | 15.6 |   |   | 17.5 | 13.59 | 4160 | 4 | Pierce et al., 2010 |
| 6 | Boystown, US |   |   |   | 0.358 | 8.96 | 65.13 | 5 | Pierce et al., 2010 |
| 7 | Bradfield, UK | 28.96 | 3.2 |   | 3.2 | 28.96 | 1150 | 5 | Singh and Scarlatos, 1988 |
| 8 | Break Neck Run Dam, US | 7 | 0.0493 |   | 0.049 | 7 | 9.2 | 5 | Singh and Scarlatos, 1989 |
| 9 | Buffalo Creek, US | 14 | 0.61 | S | 0.48 | 14.02 | 1420 | 5 | Singh and Scarlatos, 1988 |
| 10 | Butler, Us |   |   | O | 2.38 | 7.16 | 810 | 5 | Wahl, 1998 |
| 11 | Caney Coon Creek, US |   |   |   | 1.32 | 4.57 | 16.99 | 5 | Pierce et al., 2010 |
| 12 | Castlewood, US | 21.3 | 4.23 | O | 6.17 | 21.6 | 3570 | 5 | Xu & Zhang, 2009 |
| 13 | Chenying, China | 12 | 4.25 | O | 5 | 12 | 1200 | 5 | Xu & Zhang, 2009 |
| 14 | Cherokee Sandy, US |   |   |   | 0.444 | 5.18 | 8.5 | 5 | Pierce et al., 2010 |
| 15 | Colonial #4, US |   |   |   | 0.0382 | 9.91 | 14.16 | 5 | Pierce et al., 2010 |
| 16 | Dam Site #8, US |   |   |   | 0.87 | 4.57 | 48.99 | 5 | Pierce et al., 2010 |
| 17 | Danghe, China | 46 | 15.6 | O | 10.7 | 24.5 | 2500 | 5 | Xu & Zhang, 2009 |
| 18 | Davis Reservior, US | 11.9 | 58 | P | 58 | 11.58 | 510 | 4 | Xu & Zhang, 2009 |
| 19 | Dells, US | 18.3 | 13 | O | 13 | 18.3 | 5440 | 5 | Xu & Zhang, 2009 |
| 20 | DMAD, US |   |   |   | 19.7 | 8.8 | 793 | 4 | Pierce et al., 2010 |
| 21 | Dongchuankou, China | 31 | 27 | O | 27 | 31 | 21000 | 4 | Xu & Zhang, 2009 |
| 22 | Eigiau, UK | 10.5 | 4.52 |   | 4.52 | 10.5 | 400 | 5 | Singh and Scarlatos, 1988 |
| 23 | Elk City, US | 9.1 | 0.74 | O | 1.18 | 9.44 | 608.79 | 5 | Taher-shamsi et al., 2003 |
| 24 | Euclides da Cunha Dam, Brazil | 53 | 13.6 | O | 13.6 | 58.22 | 1005.2 | 5 | Taher-shamsi et al., 2003 |
| 25 | Frankfurt, Germany | 9.8 | 0.35 | P | 0.352 | 8.23 | 79 | 5 | Xu & Zhang, 2009 |
| 26 | Fred Burr, US | 10.4 | 0.75 |   | 0.75 | 10.2 | 654 | 5 | Wahl, 1998 |
| 27 | French Landing, US | 12.2 | — | P | 3.87 | 8.53 | 929 | 5 | Xu & Zhang, 2009 |
| 28 | Frenchman Dam, US | 12.5 | 21 | P | 16 | 10.8 | 1420 | 5 | Xu & Zhang, 2009 |
| 29 | Frias, Argentina | 15 | 0.25 | O | 0.25 | 15 | 400 | 5 | Xu & Zhang, 2009 |
| 30 | Goose Creek Dam, US | 6.09 | 10.6 | O | 10.6 | 1.37 | 492.7 | 5 | Taher-shamsi et al., 2003 |
| 31 | Gouhou, China | 71 | 3.3 | P | 3.18 | 44 | 2050 | 5 | Xu & Zhang, 2009 |
| 32 | Grand Rapids, US | 7.6 | 0.22 | O | 0.255 | 7.5 | 7.5 | 5 | Singh and Scarlatos, 1988 |
| 33 | Hatfield, US | 6.8 | 12.3 | O | 12.3 | 6.8 | 3400 | 5 | Xu & Zhang, 2009 |
| 34 | Haymaker, US |   |   |   | 0.37 | 4.88 | 26.9 | 5 | Pierce et al., 2010 |
| 35 | Hell Hole, US | 67.1 |   | P | 30.6 | 35.1 | 7360 | 4 | Xu & Zhang, 2009 |
| 36 | Hemet Dam | 6.09 | 8.63 |   | 8.63 | 6.09 | 1600 | 5 | Taher-shamsi et al., 2003 |
| 37 | Horse Creek, US | 12.2 | 21 | P | 12.8 | 7.01 | 3890 | 5 | Xu & Zhang, 2009 |
| 38 | Horse Creek #2, US |   |   |   | 4.8 | 12.5 | 311.49 | 5 | Pierce et al., 2010 |
| 39 | Huqitang, China | 9.9 | 0.734 | P | 0.424 | 5.1 | 50 | 5 | Xu & Zhang, 2009 |
| 40 | Ireland No. 5, US |   |   | P | 0.16 | 3.81 | 110 | 5 | Froehlich, 1995 |
| 41 | Johnstown, US | 22.86 | 18.9 | O | 18.9 | 22.25 | 7079.2 | 4 | Wahl, 1998 |
| 42 | Kelly Barnes, US | 11.6 | 0.505 | P | 0.777 | 11.3 | 680 | 5 | Xu & Zhang, 2009 |
| 43 | Knife Lake Dam | 6.096 | 9.86 |   | 9.86 | 6.096 | 1098.66 | 5 | Taher-shamsi et al., 2003 |
| 44 | Kodaganar, India | 11.5 | 12.3 | O | 12.3 | 11.5 | 1280 | 5 | Xu & Zhang, 2009 |
| 45 | lake Avalon, US | 14.63 | 7.77 | P | 31.5 | 13.7 | 2321.9 | 4 | Taher-shamsi et al., 2003 |
| 46 | Lake Latonka, US | 13 | 4.59 | P | 4.09 | 6.25 | 290 | 5 | Xu & Zhang, 2009 |
| 47 | Lake Tanglewood, US |   |   |   | 4.85 | 16.76 | 1351 | 5 | Pierce et al., 2010 |
| 48 | Laurel Run, US | 12.8 | 0.379 | O | 0.555 | 14.1 | 1050 | 5 | Froehlich, 1995 |
| 49 | Lawn Lake, US | 7.9 |   | P | 0.798 | 6.71 | 510 | 5 | Wahl, 1998 |
| 50 | Lijiaju, China | 25 | 1.14 | O | 1.14 | 25 | 2950 | 5 | Xu & Zhang, 2009 |
| 51 | Lily Lake, US |   |   | W/P | 0.0925 | 3.35 | 71 | 5 | Froehlich, 1995 |
| 52 | Little Deer Creek, US | 26.2 | 1.73 | P | 1.36 | 22.9 | 1330 | 5 | Xu & Zhang, 2009 |
| 53 | Little Wewoka, US |   |   |   | 0.987 | 9.45 | 42.48 | 5 | Pierce et al., 2010 |
| 54 | Liujiatai, China | 35.9 | 40.54 | O | 40.54 | 35.9 | 28000 | 4 | Xu & Zhang, 2009 |
| 55 | Lower Latham, US |   |   | P | 7.08 | 5.79 | 340 | 5 | Froehlich, 1995 |
| 56 | Lower Reservoir, US |   |   |   | 0.604 | 9.6 | 157.44 | 5 | Pierce et al., 2010 |
| 57 | Lower Two Medicine, US | 11.3 | 19.6 | P | 19.6 | 11.3 | 1800 | 4 | Xu & Zhang, 2009 |
| 58 | Mahe, China | 19.5 | 23.4 | O | 23.4 | 19.5 | 4950 | 4 | Xu & Zhang, 2009 |
| 59 | Mammoth, US | 21.3 | 13.6 | O | 13.6 | 21.3 | 2520 | 5 | Xu & Zhang, 2009 |
| 60 | Martin Cooling Pond Dike, US | 10.4 | 136 | P | 136 | 8.53 | 3115 | 3 | Xu & Zhang, 2009 |
| 61 | Middle Clear Boggy, US |   |   |   | 0.444 | 4.57 | 36.81 | 5 | Pierce et al., 2010 |
| 62 | Mill River, US | 13.1 | 2.5 |   | 2.5 | 13.1 | 1645 | 5 | Wahl, 1998 |
| 63 | Murnion, US |   |   |   | 0.321 | 4.27 | 17.5 | 5 | Pierce et al., 2010 |
| 64 | Nanaksagar Dam, India | 15.85 | 210 |   | 210 | 15.85 | 9709.5 | 3 | Taher-shamsi et al., 2003 |
| 65 | North Branch, US | 5.5 |   |   | 0.022 | 5.49 | 29.5 | 5 | Wahl, 1998 |
| 66 | Oros, Brazil | 35.4 | 650 | O | 660 | 35.8 | 9630 | 1 | Xu & Zhang, 2009 |
| 67 | Otto Run, US | 5.8 |   |   | 0.0074 | 5.79 | 60 | 5 | Singh and Scarlatos, 1988 |
| 68 | Owl Creek, US |   |   |   | 0.12 | 4.88 | 31.15 | 5 | Pierce et al., 2010 |
| 69 | Peter Green, US |   |   |   | 0.0197 | 3.96 | 4.42 | 5 | Pierce et al., 2010 |
| 70 | Prospect, US |   |   | P | 3.54 | 1.68 | 116 | 5 | Xu & Zhang, 2009 |
| 71 | Puddingstone Dam, US | 15.24 | 0.616 | O | 0.617 | 15.2 | 480 | 5 | Froehlich, 1995 |
| 72 | Qielinggou, China | 18 | 0.7 | O | 0.7 | 18 | 2000 | 5 | Xu & Zhang, 2009 |
| 73 | Quail Creek, US | 24 | 50 | P | 30.8 | 16.7 | 3110 | 4 | Xu & Zhang, 2009 |
| 74 | Salles Oliveira, Brazil | 35 | 25.9 | O | 71.5 | 38.4 | 7200 | 4 | Singh and Scarlatos, 1988 |
| 75 | Sandy Run, US | 8.5 | 0.0568 | O | 0.0568 | 8.53 | 435 | 5 | Singh and Scarlatos, 1988 |
| 76 | Schaeffer Reservoir, US | 30.5 | 3.92 | O | 4.44 | 30.5 | 4500 | 5 | Xu & Zhang, 2009 |
| 77 | Shimantan, China | 25 | 94.4 | O | 117 | 27.4 | 30000 | 3 | Xu & Zhang, 2009 |
| 78 | Site Y-30–95, US |   |   |   | 0.142 | 7.47 | 144.42 | 5 | Pierce et al., 2010 |
| 79 | Site Y-36–25, US |   |   |   | 0.0357 | 9.75 | 2.12 | 5 | Pierce et al., 2010 |
| 80 | Site Y-31 A–5, US |   |   |   | 0.386 | 9.45 | 36.98 | 5 | Pierce et al., 2010 |
| 81 | Sinker Creek Dam, US | 21.34 | 3.33 | S | 3.33 | 21.34 | 926 | 5 | Taher-shamsi et al., 2003 |
| 82 | South Fork, US |   |   | O | 18.9 | 24.6 | 8500 | 4 | Froehlich, 1995 |
| 83 | South Fork Tributary, US | 1.8 |   |   | 0.0037 | 1.83 | 122 | 5 | Pierce et al., 2010 |
| 84 | Stevens Dam, US |   |   |   | 0.0789 | 4.27 | 5.92 | 5 | Pierce et al., 2010 |
| 85 | Swift, US | 47.9 | 37 | O | 37 | 47.85 | 24947 | 4 | Xu & Zhang, 2009 |
| 86 | Taum Sauk Reservoir, US |   |   |   | 5.39 | 31.46 | 7743 | 5 | Pierce et al., 2010 |
| 87 | Teton, US | 93 | 356 | P | 310 | 77.4 | 65120 | 2 | Xu & Zhang, 2009 |
| 88 | Upper Clear Boggy, US |   |   |   | 0.863 | 6.1 | 70.79 | 5 | Pierce et al., 2010 |
| 89 | Upper Red Rock, US |   |   |   | 0.247 | 4.57 | 8.5 | 5 | Pierce et al., 2010 |
| 90 | Weatland Number, US | 13.6 | 11.5 | P | 11.6 | 12.2 | 566.34 | 5 | Pierce et al., 2011 |
| 91 | Zhugou, China | 23.5 | 15.4 | O | 18.43 | 23.5 | 11200 | 4 | Xu & Zhang, 2009 |
| 92 | Zuocun, China | 35 | 40 | O | 40 | 35 | 23600 | 4 | Xu & Zhang, 2009 |
| Max. | 93 | 650 |  | 660 | 77.4 | 78100 |  |  |
| Min. | 1.8 | 0.049 |  | 0.0037 | 1.37 | 2.12 |  |  |
| Ave. | 21.6 | 44.15 |  | 30.67 | 15.7 | 4690.5 |  |  |

1 Note that O=overtopping; P=piping; S=sliding; W=wave action



Fig 1 Schematic representation of a dam breach and the main geometric parameters for (a) profile of dam reservoir and (b) cross-section of dam breach

|  |
| --- |
| *Vw**Hw**Nc**Qp* *W11**W21**W31* *W34**b1**b2**b3**b4**W’1**W’2**W’3**W’4**b'* |

Fig 2. Architecture of the proposed ANN-GA

|  |  |
| --- | --- |
|  |  |

Fig 3. Estimated versus observed values of peak discharge in a) training and b) test for the first data subset

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig 4. Performance of the analysed model in terms of a) RMSE, b) RSE, and c) R2.

Fig 5. Input data clustering with five clusters and corresponding cluster centres shown as C Clus.

Fig 6. Scatter of observed and estimated values for different test sample sets of cross-validation method in the ANN-GA model a) without clustering and clustered data samples with b) 3, c) 4, d) 5, e) 6, and f) 7 groups.

|  |  |
| --- | --- |
|  |  |
|   |
|  |

Fig 7. Performance indicators of the ANN-GA models for different numbers of clusters



Fig. 8 Comparison of the models with respect to interval RMSE of peak outflow

Fig. 9 uncertainty analysis for prediction error as a) mean prediction error (*ē*) and b) width of uncertainty band (±*2Se*)

Table 1. Statistics of peak outflow for ten data subsets

|  |  |
| --- | --- |
| Parameter | Sample set number |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| No of data | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| mean | 5723.8 | 1979.8 | 3667.0 | 8500.3 | 2139.6 | 9198.2 | 3862.5 | 5415.5 | 2303.7 | 4955.3 |
| max | 24947 | 9709.5 | 23600 | 65120 | 7360 | 78100 | 24947 | 28000 | 8500 | 28000 |
| min | 4.42 | 8.5 | 7.5 | 7.5 | 16.99 | 36.98 | 14.16 | 4.42 | 2.12 | 29.5 |
| standard deviation | 8149.9 | 3031.6 | 7656.1 | 21467.5 | 2294.7 | 24920.3 | 8172.1 | 9188.6 | 3288.1 | 9243.3 |

Table 2. Performance indicators of the five-clustered ANN-GA model for 10 data-folds

|  |  |  |
| --- | --- | --- |
| Subset No. | Train | Test |
| R2 | RMSE | RSE | NSE | RSR | R2 | RMSE | RSE | NSE | RSR |
| 1 | 0.97 | 2235 | 0.03 | 0.97 | 0.17 | 0.82 | 5266 | 0.47 | 0.53 | 0.69 |
| 2 | 0.95 | 2806 | 0.05 | 0.95 | 0.22 | 0.93 | 955 | 0.11 | 0.89 | 0.33 |
| 3 | 0.95 | 2623 | 0.05 | 0.95 | 0.22 | 0.69 | 5385 | 0.56 | 0.44 | 0.75 |
| 4 | 0.92 | 2999 | 0.08 | 0.92 | 0.28 | 0.99 | 978 | 0.002 | 1.00 | 0.04 |
| 5 | 0.96 | 2536 | 0.04 | 0.96 | 0.20 | 0.77 | 4638 | 4.59 | -3.59 | 2.14 |
| 6 | 0.94 | 2103 | 0.06 | 0.94 | 0.24 | 0.95 | 6236 | 0.07 | 0.93 | 0.26 |
| 7 | 0.95 | 2699 | 0.05 | 0.95 | 0.22 | 0.95 | 6230 | 0.65 | 0.35 | 0.81 |
| 8 | 0.96 | 2460 | 0.04 | 0.96 | 0.20 | 0.74 | 4986 | 0.33 | 0.67 | 0.57 |
| 9 | 0.94 | 2902 | 0.05 | 0.95 | 0.22 | 0.76 | 4330 | 1.95 | -0.95 | 1.40 |
| 10 | 0.95 | 2555 | 0.04 | 0.96 | 0.20 | 0.95 | 4737 | 0.29 | 0.71 | 0.54 |
| Average | **0.95** | **2592** | **0.05** | **0.95** | **0.22** | **0.86** | **4374** | **0.90** | **0.10** | **0.75** |

Table 3. The weights and biases of the five-clustered ANN-GA model

|  |  |  |  |
| --- | --- | --- | --- |
| Weights in the first layer | Biases in the first layer | Weights of the second layer | Bias of the second layer |
| W11 | 23.181 | W21 | 17.485 | W31 | 13.436 | b1 | -13.18 | W'1 | 7.951 | b' | 5.38 |
| W12 | -12.855 | W22 | -0.068 | W32 | 5.193 | b2 | 18.632 | W'2 | 3.113 |  |  |
| W13 | 15.546 | W23 | -6.628 | W33 | 16.406 | b3 | 4.743 | W'3 | -0.37 |  |  |
| W14 | 18.235 | W24 | 1.016 | W34 | 10.461 | b4 | 4.248 | W'4 | 1.186 |  |  |

Table 4. Main clustering features of the five-clustered ANN-GA model

|  |  |  |
| --- | --- | --- |
| Cluster No. |  | Variable limit |
| **Variable name** | **Cluster centre** | **minimum** | **maximum** |
| 1 | *Vw* (mcm) | 633.75 | 607.5 | 660 |
| *Hw* (m) | 33 | 31 | 35.8 |
| 2 | *Vw* (mcm) | 310.0 | 310 | 310 |
| *Hw* (m) | 77 | 77 | 77 |
| 3 | *Vw* (mcm) | 154.33 | 117 | 210 |
| *Hw* (m) | 17 | 8.5 | 27.4 |
| 4 | *Vw* (mcm) | 30.48 | 17.5 | 71.5 |
| *Hw* (m) | 25 | 8.8 | 34.4 |
| 5 | *Vw* (mcm) | 3.41 | 0.0037 | 37 |
| *Hw* (m) | 12 | 1.37 | 58.22 |