Geomechanics and Engineering, *Vol. 31, No. 6 (2022) 637-651*
https://doi.org/10.12989/gae.2022.31.6.637

Prediction of earthquake-induced crest settlement of embankment dams using gene expression programming

Evren Seyrek*[1] and Sadettin Topçu[2a]

[1]*Civil Engineering Department, Engineering Faculty,Kütahya Dumlupınar University, Evliya Çelebi Campus, Kütahya, Turkey*
[2]*Department of Construction Technology, Vocational School of Technical Science, Kütahya Dumlupınar University, Campus of Germiyan, Kütahya, Turkey*

(*Received May 5, 2022, Revised December 4, 2022, Accepted December 15, 2022*)

Abstract. The seismic design of embankment dams requires more comprehensive studies to understand the behaviour of dams. Deformations primarily control this behaviour occur during or after earthquake loading. Dam failures and incidents show that the impacts of deformations should be reviewed for existing and new embankment dams. Overtopping erosion failure can occur if crest deformations exceed the freeboard at the time of the deformations. Therefore, crest settlement is one of the most critical deformations. This study developed empirical formulas using Gene Expression Programming (GEP) based on 88 cases. In the analyses, dam height (H_d), alluvium thickness (H_a), the magnitude-acceleration-factor (MAF) values developed based on earthquake magnitude (Mw) and peak ground acceleration (PGA) within this study have been chosen as variables. Results show that GEP models developed in the paper are remarkably robust and accessible tools to predict earthquake-induced crest settlement of embankment dams and perform superior to the existing formulation. Also, dam engineering professionals can use them practically because the variables of prediction equations are easily accessible after the earthquake.

Keywords: crest settlement; earthquake; embankment dam; freeboard; GEP

1. Introduction

Dams are hydraulic structures that regulate river flows and control the water stored in the reservoir for different purposes such as irrigation, water supply, irrigation, hydropower, etc. With the increase in the world's population, water needs are increasing. Dams are designed to meet this requirement. There are many dams in different parts of the world classified as small or large dams. ICOLD stated that a dam with a height of 15 meters or greater from lowest foundation to crest or a dam between 5 meters and 15 meters impounding more than 3 million cubic meters is classified as a large dam (ICOLD 2011).

Large dams have significant benefits to society and the economy; on the other hand, they can entail acknowledgeable risks. Seyrek and Tosun (2011) stated that dams constructed near urban areas pose a high-risk potential for downstream life and property. Therefore, the safety of dams is a crucial subject in dam engineering and requires more comprehensive studies to understand the behaviour of dams during different loading conditions (Tosun 2015). In recent years, dam safety has drawn increasing attention because floods resulting from dam failures can lead to catastrophic loss of life and property, especially in densely populated areas (Zhang *et al*. 2009).

*Corresponding author, Professor
E-mail: evren.seyrek@dpu.edu.tr
[a]Ph.D.
E-mail: sadettin.topcu@dpu.edu.tr

When regions of the world with different topographic, hydrogeological and seismic characteristics are examined, different dams such as earth-fill dams, rock-fill dams, gravity dams, arch dams, and tailing waste dams and others are observed. Embankment dams are primarily preferred globally as they have a wide range of economy and applicability compared to other types of dams and the project criteria are flexible. The majority of existing dams are embankment dams totaling 78% of all dams in the world; 65% are earthfill, 13% are rockfill dams, and other types of dams make up the remaining (ICOLD 2022).

Dam failures are mostly observed in embankment dams. According to Zhang *et al.* (2009), dam failures observed in earthfill and rock-fill dams constitute 69.4% of reported cases. Different effects can cause dam failure, such as overtopping, piping, sliding, earthquakes, poor design options, and poor planning. Zhang *et al.* (2009) have stated that several causes are often involved in a single failure, and these causes are interrelated with each other. Failure due to overtopping and piping is the most common dam failure mode of reported cases (Foster *et al*. 2000). Historically, few dams have been significantly damaged by earthquakes. On a worldwide basis, only about a dozen dams are known to have failed entirely due to an earthquake; these dams were primarily tailings or hydraulic fill dams or relatively small embankments (USCOLD 1992). Many embankment dams are exposed to earthquake shaking each year, but either the damage caused by the earthquake has not been extensive enough, or in the rare cases where damage was extensive, many of the reservoirs were, by chance, low at the time of the earthquake, so uncontrolled releases did not happen (USBR 2019).

ISSN: 2005-307X (Print), 2092-6219 (Online)

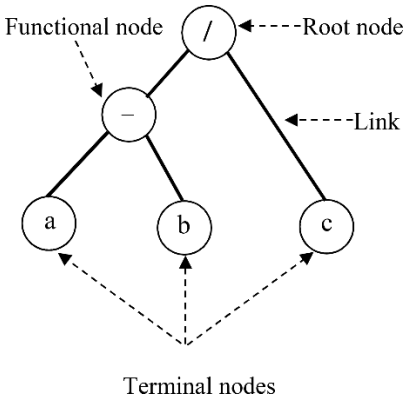

Fig. 1 Representation of a chromosome consisting of one gene with an ETs

One of the rare cases of dam failures during or after a large earthquake occurred in Lower San Fernando Dam in California. This first example failed due to the liquefaction phenomenon under the earthquake loading conditions (Seed *et al.* 1975). The upstream slope of Lower San Fernando Dam failed during the 1971 San Fernando earthquake. Intact blocks of embankment material moved tens of feet on liquefied hydraulic fill shell material.

Although very few dams collapsed due to earthquakes in the literature, dam accidents due to earthquakes are frequently encountered (USCOLD 1992, Pells and Fell 2002, Fell *et al.* 2005). Reported cases show that embankment dams can settle, deform laterally and longitudinally, and exhibit longitudinal and transverse cracking due to earthquake loading. This cracking mechanism can lead to internal erosion and piping failure of dams (Pells and Fell 2002). During seismic loading, pore pressure can increase, resulting in a decrease in shearing resistance. If enough reduction occurs, over a sufficient extent, large deformations can result. Overtopping erosion failure can occur if crest deformations exceed the freeboard at the time of the deformations.

OMNR (2011) stated that settlement of the dam crest could be one of the contributing factors for overtopping, and the freeboard of embankment dams should be considered during assessment. Settlement may occur due to consolidation of the dam or foundation materials under static loading or may be induced by seismic activity. Therefore the impacts of dam and foundation deformations should be reviewed while evaluating new or existing dam structures for earthquake loading. These deformations are frequently determined by using softwares that can perform 2D or 3D analyses based on finite element methods in last years (Karabulut and Geniş 2019, Hu and Huang 2019, Karalar and Cavusli 2019, Nasiri *et al*. 2020, Karalar and Cavusli 2021). Another tools to predict earthquake-induced deformations are simplified methods, empirical and semi-empirical formulas or soft computing approaches based on observed data from case histories.

The simplified method was firstly proposed by Newmark (1965). This method assumes that the deformation mode in response to earthquake shaking will be the sliding of rigid blocks (masses) of dam and foundation material. Sliding is assumed to occur whenever the base acceleration exceeds the yield acceleration, which is the horizontal seismic acceleration that results in a factor of safety of exactly 1.0 using conventional slope stability analysis (FEMA 2005). Several analytical methods have been proposed to simplify or modify the Newmark method (Makdisi and Seed 1978, Rathje and Bray 1999, Bray and Travasarou 2007, Saygili and Rathje 2008). These simplified methods are simple, inexpensive, and substantially less time-consuming than the complicated stress–deformation approaches. They are especially recommended as a screening tool to identify critical dams (Kan *et al.* 2017).

In recent years, the developments in equipping dams with seismic instruments and the increase in the actual numerical data obtained at the dam site have increased the interest of researchers in earthquake-induced dam crest settlement. Swaisgood (2003) has reviewed 69 case histories and proposed a relationship to predict crest settlement of embankment dams using regression techniques. In addition to dam incidents, new relation based on earthquake magnitude, peak ground acceleration, and dam height has been proposed by Swaisgood (2014).
Singh *et al.* (2007) have used 122 published case histories on the performance of earth dams and embankments during past earthquakes to compare observed and estimated permanent deformations from simplified methods. They also developed a relationship between permanent deformation and the ratio of yield acceleration and peak horizontal ground acceleration.

In 2015, data sets of observations from 152 published case histories were used to predict earthquake-induced deformation of the earth dams and embankments using Artificial Neural Network (ANN) (Barkhordari and Entezari 2015). In this study, dam type, dam height, magnitude of earthquake, peak ground acceleration, predominant period of the earthquake ground motion, fundamental period of the structure and yield acceleration values have been used to establish the model.

Javdanian *et al.* (2020) have shown that yield acceleration, maximum horizontal earthquake acceleration, fundamental period dam body, predominant period of earthquake, earthquake magnitude) and dam height is the most influential parameter that affects the seismic crest settlement. Support Vector Regression (SVR) and Multiple

Linear Regression (MLR) models have been developed to estimate the earthquake-induced settlement of embankment dams using 151 real-world cases.

This paper aims to develop new empirical models for predicting earthquake-induced crest settlements of embankment dams. At first, a database was established with a detailed literature survey excluding cases susceptible to liquefaction problems. Eighty-eight cases from different parts of the world have been analyzed. Two empirical formulations named GEP-I and GEP-II are enhanced to predict crest settlements by using Gene Expression Programing (GEP), which is mainly used to solve complex and multivariable engineering problems. Additionally, formulations derived from GEP-I and GEP-II have been compared with Swaisgood's (2014) model. The analyses show that the proposed models can be used to predict earthquake-induced crest settlements of embankment dams for rapid evaluation of critical dams.

2. Gene Expression Programming (GEP)

Genetic algorithms (GAs) are a class of evolutionary algorithms popularized by John Holland and his team in the 1970s. Genetic algorithms are an intuitive approach that considers gene structures in living and methods in the biological evolution process (Goldberg 1989). Individuals are chromosomes, genes are input contained in chromosomes, and a population is a society consisting of chromosomes in these algorithms. Genetic Programming (GP) is a branch of genetic algorithm belonging to the family of evolutionary algorithms and was first proposed by Koza (1992). Genetic programming is like genetic algorithms, yet, genetic programming analyzes using parse tree (expression tree) structure rather than through bite strip (comprehension) as in genetic algorithm (Pourzangbar 2012). Gene Expression Programming (GEP) using GP is a new technique developed by Ferreira (2001) for generating computer models from learned or discovered information. The fundamental difference between GEP from GP is the structure of individuals. The individuals are expression trees (ETs) in GP. They are nonlinear entities of different sizes and shapes. The individuals are encoded as fixed-length linear sequences (chromosomes consisting of one or more genes) in GEP as in GAs, then expressed as nonlinear entities of different sizes and shapes. Let the algebraic notation in Eq. (1) be given below as an example.

$$\frac{a-b}{c} \tag{1}$$

a, b and c are the terminals formed by variables in the model. -, ÷ are functions that determine the formal organization of terminals. Nodes and links provide the interaction of this function and terminal with each other. The principal function of the gene organization is the root node. This algebraic representation above is shown in Fig. 1.

Depending on the complexity of the problem, these ETs can also be encoded as sub-expression trees (sub-ETs) where more than one gene is considered. The phenotypic gene information in ETs is displayed in the Karva language

0	1	2	3	4
÷	–	a	b	c

Fig. 2 Karva expression of a gene

as top-to-bottom and left-to-right genotype gene information, respectively, expressed in Fig. 2. This genotype gene information is also known as an Open Reading Frame (ORF).

Genes processed by genetic operators in GEP are organized from head and tail. The head consists of functions and terminals, as in the gene in the example above. The tail consists of terminals only. The length of the tail, t, be calculated from t=h.(n-1)+1. h: Head length; n: Number of the variable. Accordingly, the tail length of the gene given in the example is t=11, and it is presented in bold (See Fig. 3).

Chromosome number and structure should be determined in GEP. The number of genes and length of the head (head size) that make up the structure of the chromosomes should be determined according to the difficulty and complexity of the problem. The number of variables and their contribution to the solution is essential for creating a high-performance model. Generally, a number of variables increases, number of genes and length of the head also tend to increase. Likewise, the functions that make up the head of the gene and the linking functions in the case of using more than one gene should be defined following the problem. The linking functions that connect genes may not be the same in the models created for every issue. It may vary depending on the nature of the problem and the outcome being discussed. It is recommended to use basic mathematical operators (+, -, x, /) to create simple and pellucid models (Sattar 2014).

GEP generates the population by randomly creating individuals from functions and terminals where the chromosome structure is ready. The first individual created in the population is the parental chromosome. The best offspring chromosomes created by applying genetic operators to this parent chromosome are passed on to future generations. According to the fitness criterion, the best offspring chromosomes are selected, where objective functions are used, depending on the target model. The fitness criterion, fi, is given in Eq. (2), where RMSE (Root Mean Square Error) is used, which is considered in obtaining the regression function. The fitness criteria range from 0 to 1000. If RMSE=0, f_i=1000, and perfect fit is achieved. Thus, a new population is created by allowing the best-fit individuals (chromosomes) to continue their lives.

$$f_i = 1000.\frac{1}{1+RMSE} \tag{2}$$

The evolutionary process in the genetic algorithm is terminated based on some convergence criterion. The maximum number of generations is defined, or the process can be completed when multiple generations can create without changing the best fitness value.

Many genetic operators exist in the GEP technique (Ferreira 2001). The replication operator is used to store good genes on the chromosome that affect the outcome for

0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5
÷	–	a	b	c	**b**	**c**	**b**	**c**	**c**	**a**	**b**	**a**	**a**	**b**	**a**

Fig. 3 Schematic view of a gene that consists of head and tail

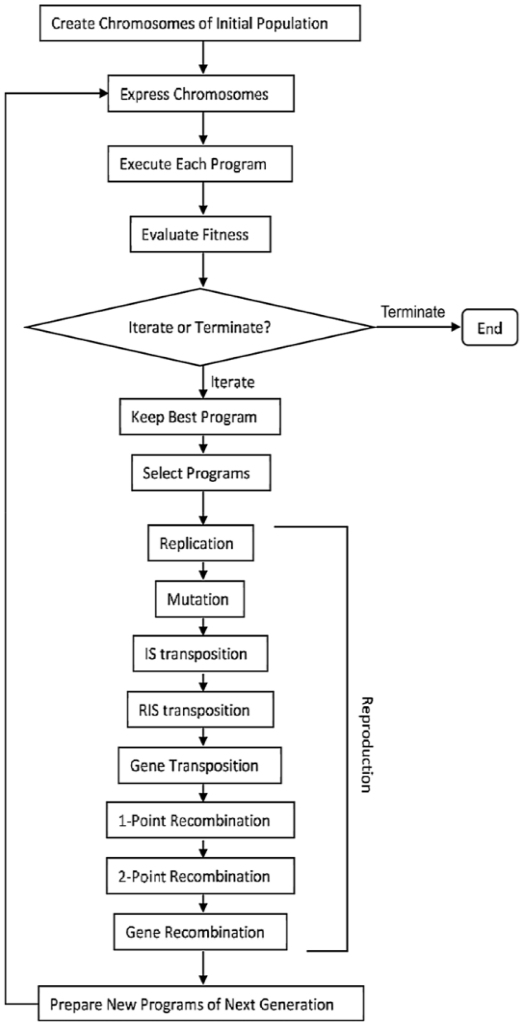

Fig. 4 The flowchart of gene expression programming (GEP) algorithm (Ferreira 2001)

later generations. The symbol in the head of genes can be replaced with another with the mutation operator, but the change in the tail part is limited to terminals only. The structural organization of the chromosomes as described above is preserved with the mutation. A transposition operator replaces a certain number of gene information in chromosomes or copies other gene information instead. The transposition occurs in three different ways: the transposition of insertion sequential genes (IS Transposition), the transposition of root insertion sequence (RIS Transposition), and the gene transfer (Gene Transposition). The IS Transposition replaces a certain number of randomly selected gene information in the head of the chromosome with the same number of gene information in another part of the same chromosome. The RIS Transposition occurs when a certain number of randomly selected gene information is placed in different places in the head of the chromosome. Gene transfer is achieved by copying a group of genes on the chromosome and substituting them. The recombination operator is used in the reproduction of chromosomes by rearranging a single point (1-point Recombination), two points (2-point Recombination), and a certain number of genes (Gene Recombination). Genes are replaced at a randomly chosen spot between two chromosomes during the crossover, forming new offspring with 1-Point Recombination. With 2-point Recombination, genes are replaced at two randomly selected points between two chromosomes during the crossover, creating new offspring. By Gene Recombination, randomly selected genes are recombined to form new generations. The flowchart of the GEP is summarized in Fig. 4.

In recent years, GAs, GP and GEP, which mimic the process of biological evolution, have been used to predict the necessary design parameters in the field of geotechnics. This artificial intelligence (AI) approach can eliminate the uncertainties seen in the findings of experimental investigations on non-homogeneous soils. In particular,

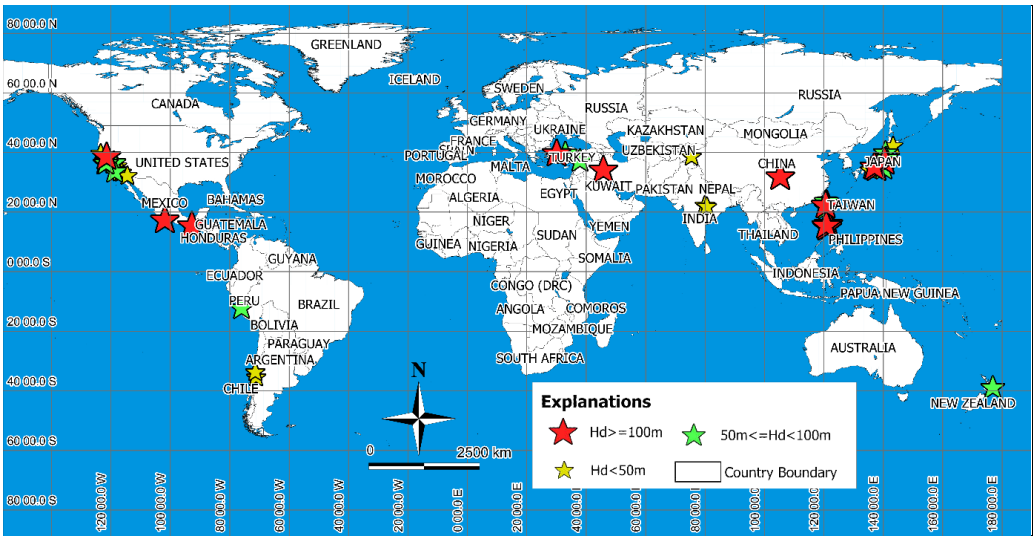

Fig. 5 Location of dams having earthquake-induced crest settlements

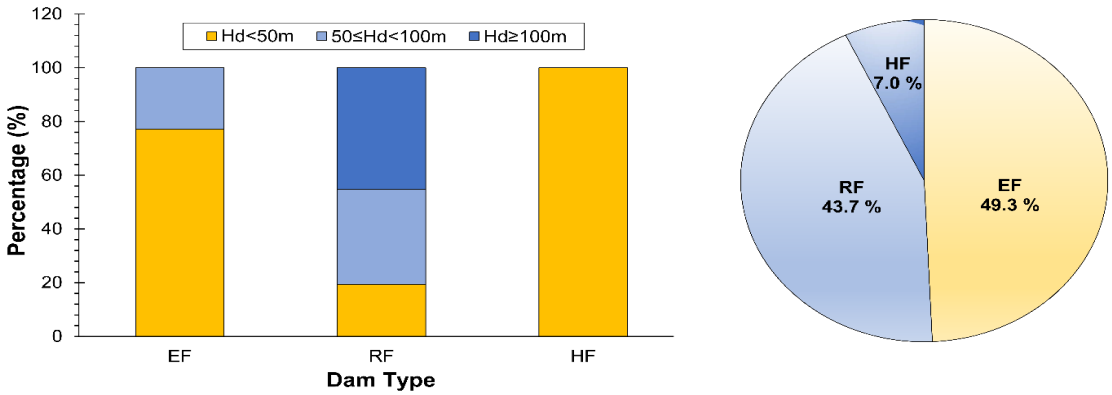

Fig. 6 Distribution of dams based on height and type

GEP (www.gepsoft.com) has a simple user interface and is easy to use, allowing researchers to opt for GEP. In addition, it is known that the models created with GEP give better results than models made with other AI approaches (ANN, ANFIS). GEP has been used for different topics of civil engineering and geotechnical engineering: mechanical properties of SFRHSC with metakaolin and ground pumice (Sarıdemir *et al.* 2017), swell pressure and unconfined compression strength of expansive soils (Jalal *et al*. 2021), compaction parameters of coarse-grained soils with fines content (Sivrikaya *et al.* 2013), soil liquefaction potential (Kayadelen 2011), compression index for fine-grained soil (Mohammedzadeh *et al.* 2019), soil-water characteristic curve parameters of unsaturated soil (Johari and Nejad 2015), soil deformation modules using plate load test results (Mollahasani *et al.* 2011), the axial capacity of driven piles in cohesive soil (Alkroosh and Nikraz 2012), California Bearing Ratio (CBR) of soil (Taskiran 2010), collapse potential of soils (Uysal 2020), effective internal friction angle of soils (Kayadelen *et al*. 2009), effective stress parameter of unsaturated soils (Johari *et al*. 2013) and maximum lateral displacement of retaining wall (Johari *et al.* 2016).

3. Database for GEP modelling

It is known that the success of the models is wholly related to realistic and accurate data. Therefore, initially, an extensive literature review was performed. Secondly, dam incident cases causing dam crest settlement during seismic loading were collected from different sources. Observed earthquake-induced crest settlement values reported in multiple kinds of research were checked, and then an optimized database was established. The database includes 88 cases in different parts of the world. Most of the data were gathered from Pells and Fell (2002)(n=53). Twelve data were obtained from Sing and Roy (2009). The number of data collected from Swaisgood (2014) was 11. The remaining 12 data were obtained from different studies (Sakamoto *et al*. 1998, Chern and Tsai 2001, Yousif *et al.* 2019, Adamo *et al.* 2020, Yamanaka *et al.* 2020). Since database of this study has only cases of embankment dams, it is possible to state that database of this study is reliable although it is limited. The database of the study by Swaisgood (2014) which is frequently used by dam engineering professionals includes accidents belonging to

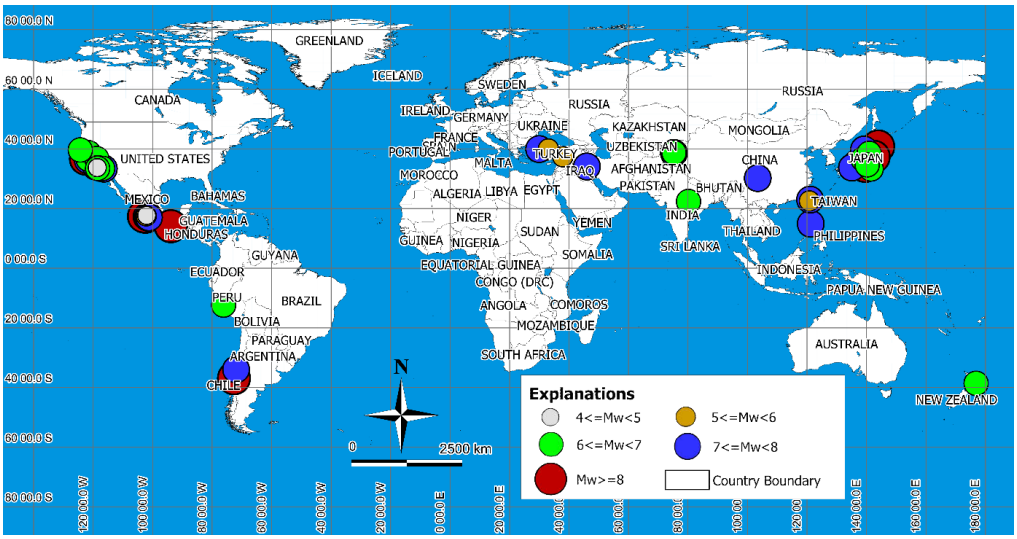

Fig. 7 Epicenter distribution of earthquakes causing to crest settlement

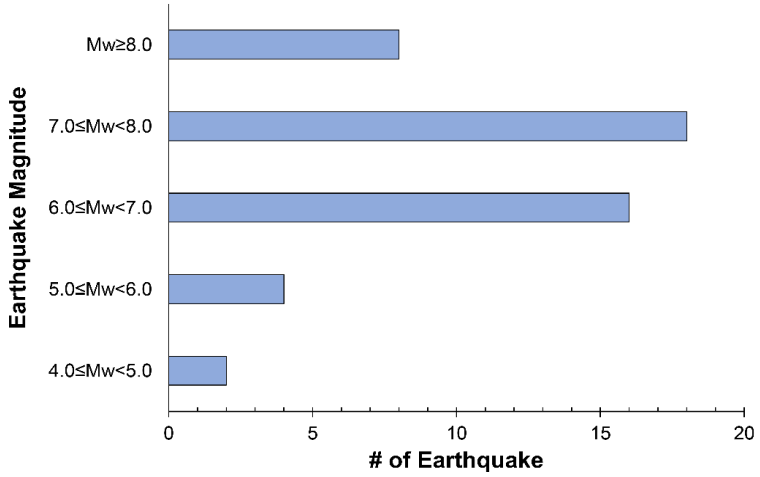

Fig. 8 Distribution of earthquake magnitude causing crest settlement

concrete faced rockfill dams as well as earthfill and rockfill dams. In addition, it is seen that the distribution of input parameters to be given in the following section of the paper represent a wide range, not a narrow range. Therefore the authors believe that database used for GEP model established in this study is reliable and sufficient for rapid risk evaluation after earthquake and pre-design.

Locations of dam sites belonging to 88 dam accidents are given in Fig. 5. As seen in Fig. (5), most of the embankment dams are located in USA and Japan, with a number of 28 and 19, respectively.

Dams discussed in this study are earthfill (EF), rockfill (RF) and hydraulic fill (HF) dam types. EF and RF dams constitute 93 per cent of the database (See Fig. 6). While approximately 78% of EF dams are less than 50 m high, almost 80% of RF dams are higher than 50 m. All dams higher than 100 m in the database are of RF dam type.

Epicenters of past earthquakes causing crest settlement are given in Fig. (7). It is seen that earthquake epicenters are located in regions with high seismic activity, such as Mexico, Japan and USA. 1989 Loma Prieta earthquake (M=7.1), 1906 San Francisco earthquake (M=8.3), 1995 Kobe earthquake (M=7.1), 1999 Chi-Chi earthquake (M=7.7), and 1999 Kocaeli (M=7.8) earthquakes included in this study are important historical earthquakes in the past. Especially Tohoku earthquake (M=9.0) occurred on 11.03.2011, and the Chili earthquake (M=8.8) occurred on 27.02.2010, which are important in terms of their magnitudes.

When the magnitudes of the earthquakes in the database are examined, approximately 71% are between 6.0 and 8.0. Earthquakes greater than 8.0 constitute 17% of all earthquakes (See Fig. 8).

4. Developed GEP models

The success of the GEP model to be developed is closely related to the selection of the independent variables that affect the crest settlement. However, determination of the relevant parameters for settlement is not easy. Because the behaviour of embankment dams under earthquake loading is quite complex. Previous studies show that dam height, earthquake magnitude and peak ground acceleration

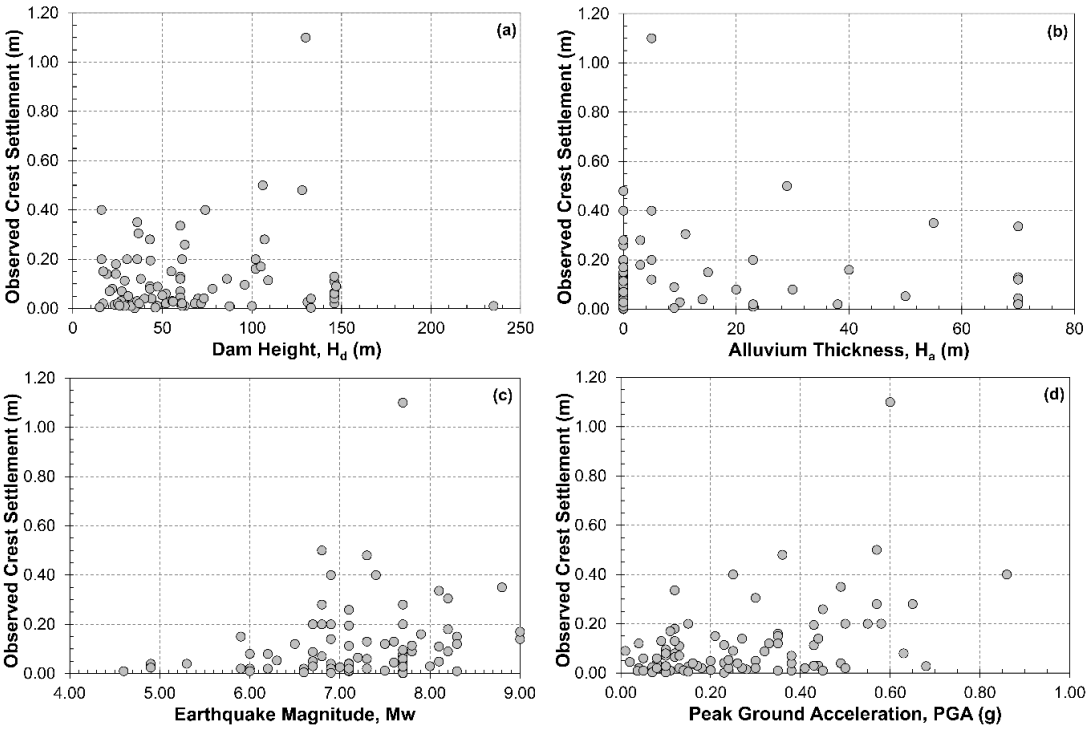

Fig. 9 Distribution of crest settlement with dam height, alluvium thickness, earthquake magnitude and peak ground acceleration values

values are the dominant parameters of settlement behaviour of embankment dams. Besides, crest settlements of embankment dams subjected to the earthquake may be caused directly by the insufficiency of compaction in the embankment and the settlement in the foundation soil under the dam.

Some researchers used different variables, such as the fundamental period dam body and the predominant earthquake period except (Barkhordari and Entezari 2015, Javdanian *et al.* 2020). In addition to the model's success to be established within the scope of this study, it is also vital that the professionals perform a rapid assessment after an earthquake. Swaisgood (2003) stated that several other independent variables such as dam type, distance from seismic source to the dam site, the ratio of crest length to dam height, embankment slope angles and reservoir water level at the time of the earthquake have a minimal effect on crest settlement. It has been observed that there is limited information in the databases included in the studies on earthquake-induced crest settlements of embankment dams. During the literature research, it has been determined that there are differences even in the height information of the same dam used for estimation of crest settlement values. For this reason, a comprehensive research has been carried out on the characteristics of the dams considered. As a result of the research, it was concluded that the dam height and the thickness of the alluvium under the dam foundation are the most reliable data and other characteristics such as dam slopes, crest width and length could not be obtained for all dams studied. Therefore, a database of dams with reliable information on dam height and alluvial thickness was created in model studies.

In addition to dam height (H_d) and alluvium thickness (H_a), earthquake magnitude (M_w) and peak ground acceleration (PGA) have been selected as independent variables to predict crest settlement in this study. Parameters such as damping ratio and dynamic modulus values are known to be effective in the dynamic behavior of dam structures, due to the problems during the determination of the characteristics of the dam structures explained above, the input parameter including material property could not selected for earthfill and rockfill dams. In addition, selecting the material parameters based on estimation is not suitable for reliable database. With this approach it would be accepted that each earthfill and rockfill dam would have the same material properties and this situation would increase the uncertainty on the results. Therefore it was not preferred to add such material properties as input parameters to the model. PGA values have been identified as measured or estimated values on the dam foundation. Fig. 9 shows the correlation of four input parameters compared to the dependent parameter. As seen in Fig. 9, the selected variables alone cannot explain the earthquake-induced settlements in the embankment dams.

After detailed statistical studies to determine the sensitivity of variables with each other or with settlement value, it was decided to combine the M and PGA. This

Table 1 Input and output parameters for GEP models

Model	Input Parameter	Output Parameter
GEP-I	H_d, MAF	S_{crest}
GEP-II	H_d, H_a, MAF	

Table 2 Limit values of input and output variables employed in GEP models for training and testing process(n=88)

Variable Type	Variable Name	Min.	Max.	Average	Standard Deviation
INPUT	Dam height, H_d (m)	15.00	235.00	65.68	43.64
	Alluvium thickness, H_a (m)*	0.00	70.00	9.60	19.80
	Magnitude-Acceleration-Factor (MAF)	1.50	31.34	8.05	5.94
OUTPUT	Crest settlement, S_{crest} (m)	0.001	1.100	0.113	0.154

* for only GEP-II model

Table 3 Parameters employed in the GEP models

Definitions	GEP-I	GEP-II
Function set	+,-,/,x,Log,Sin,Cos,Pow, Inv,Abs	+,-,/,x,Sin,Cos,Min2,Pow, Exp,Inv
Fitness Function	RMSE	
Number of genes	4	
Number of chromosomes	50	
Head size	8	
Linking function	Multiplication	
Mutation	0.00138	
Inversion	0.00546	
One and two-point recombination	0.00277	
Gene recombination	0.00277	
Gene transposition	0.00277	
Random chromosomes	0.0026	

parameter was named magnitude-acceleration-factor (MAF). After the regression analysis, Eq. (3) given in below was obtained.

$$\mathrm{MAF} = e^{\sqrt{2xPGAxMw}} \tag{3}$$

After selecting variables, GEP (www.gepsoft.com) was utilized to develop a prediction equation for the vertical settlement of embankment dams subjected to earthquakes. For this purpose, GEP-based empirical GEP-I and GEP-II forecast models have been developed. The input parameters used in these GEP models are given in Table 1. Unlike the GEP-I model, alluvium thickness at the base of the embankment dam was also considered in the GEP-II model. So, GEP Models developed with the specified input parameters offer flexibility in alluvium thickness information to designers who will use equations.

Eighty-eight carefully selected data from settlement cases reported as a result of earthquakes occurring worldwide were used in GEP models. In all GEP models, 80% of this dataset consisting of 88 cases was considered a training dataset and 20% a testing dataset. The training and testing datasets have similar statistical parameters, such as mean and standard deviation. Limit values of input and output variables in GEP models for the training and testing process are given in Table 2.

In GEP models, the fitness function in the Regression Function module was selected as RMSE (Root Mean Square Error) to develop empirical equations. The parameters employed in GEP-I and GEP -II models are presented in Table 3. The number of genes, the number of chromosomes, head size, and linking function were the same in the attempts to achieve the best fitness in both GEP models. In addition, due to the difference in input parameters, the function sets employed in GEP models were not the same. The rates of genetic operators such as mutation, transposition, and recombination were selected automatically.

Ets (Expression trees) consisting of four genes and the inter-gene linking function are presented in Figs. 10 and 11 for GEP-I and GEP-II models, respectively. While d0 and d1 refer to MAF and H_d in the GEP-I model, d0, d1, and d2 refer to H_d, H_a and MAF, respectively, in the GEP-II model. According to GEP models, prediction formulas are given in Eqs. (4) and (5). The constants in these equations that are explicated ETs on the GEP-I and GEP-II models are given in Table 4.

$$\begin{aligned} S_{crest} = {} & \{\mathrm{Log}[\mathrm{Log}(c0 + \mathrm{Sin}(d1 \times c5) - c4)]\} \\ & \times \left\{ \frac{(\frac{d0}{c6}) \times (c8 - d1)}{\frac{\mathrm{Cos}(c9)}{c6}} \right\} \\ & \times \left\{ \mathrm{Sin}(c3) + \left[\frac{1}{d1} \times (c6 \times c7) \right]^{|d0|} \right\} \\ & \times \left\{ \left(\mathrm{Log}(c8)^{\mathrm{Sin}(d0 \times c3 \times c2) - c2} \right) \right\} \end{aligned} \tag{4}$$

Table 4 The constants of the formulation for GEP-I and GEP-II models

Constants	GEP-I				GEP-II			
	Sub-ET1	Sub-ET2	Sub-ET3	Sub-ET4	Sub-ET1	Sub-ET2	Sub-ET3	Sub-ET4
c0	-4.954							
c1								-10.179
c2				-7.747		-9.915		
c3			2.899	3.461	-9.464			
c4	-7.803				-2.101		4.905	
c5	-9.972							
c6		-7.142	7.815					-1.261
c7			1.988				6.404	
c8		-14.977		4.415		-6.894		
c9		-5.091			-1.734		-9.134	

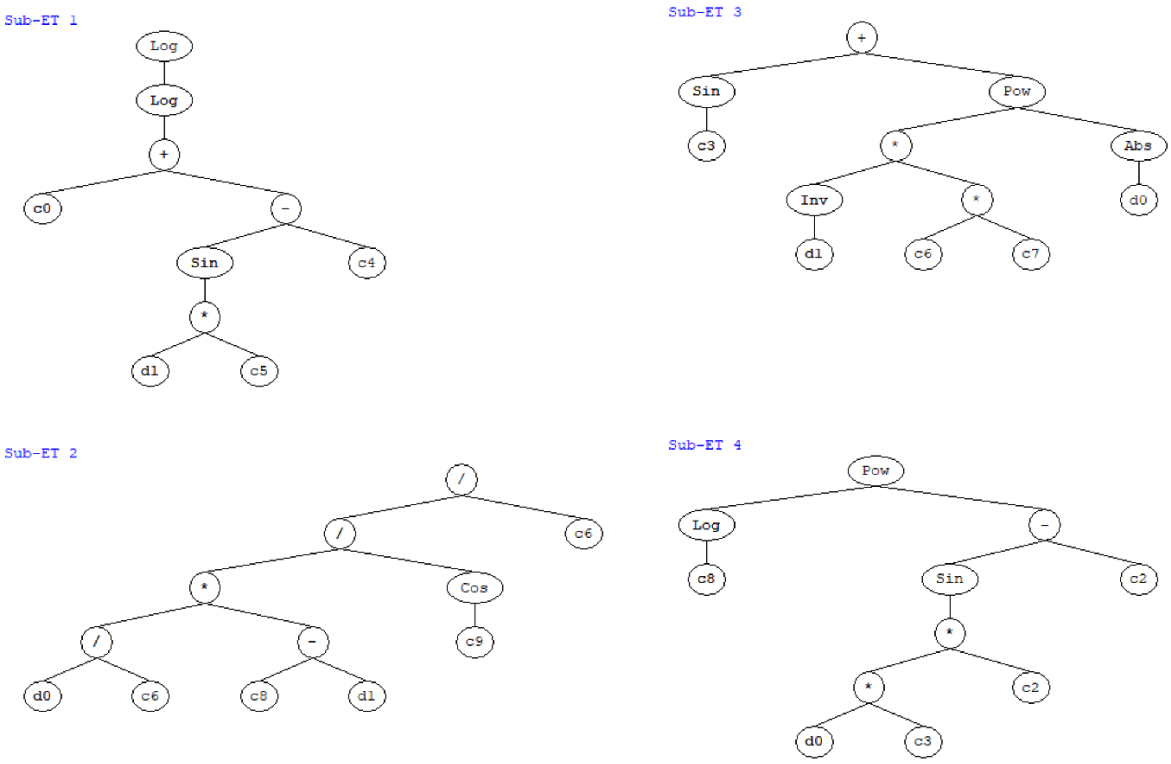

Fig. 10 Expression trees for the GEP-I model

$$S_{crest} = \{c4 + \cos[\min(\sin(d0); (d1 \times c3)) + d2^{c9}]\} \times \left\{d2 - \frac{(\frac{c8}{d0} \times c8)}{c2 + d2}\right\} \times \left\{e^{\left(\frac{1}{(\frac{d2 \times d1}{c4}) - c7}\right) + c9}\right\} \times \left\{\min\left[\left(\frac{d2 + d1}{c6}\right) - (d1 - d0); c1\right] - d0\right\} \quad (5)$$

$$S_{crest} = \log[\log(2.849 - \sin(9.972 \times H_d))] \times \left[-\frac{MAF}{18.836} \times (14.977 + H_d)\right] \times \left[0.240 + \left(\frac{15.534}{H_d}\right)^{MAF}\right] \times 0.645^{7.747 - \sin(26.812 \times MAF)} \quad (6)$$

Crest prediction formulas for GEP-I and GEP-II models were simplified as in Eqs. (6) and (7) using the constants of formulations given in Table 4. These equations were obtained after extensive analysis in GEP and were chosen among different relations. Especially despite the highly non-linear relationship between input data and output data, it should be stated that these prediction equations are user-friendly and can be used with simple calculation tools.

$$S_{crest} = \{\cos[\min(\sin(H_d); (-9.464 \times H_a)) + MAF^{-1.734}] - 2.101\} \times \left[MAF - \frac{47.531}{H_d \times (MAF - 9.915)}\right] \times e^{\frac{4.905}{(MAF \times H_a) - 31.410} - 9.134} \times \min\left(\frac{(1.261 \times H_d) - (2.261 \times H_a) - MAF}{1.261}; (-10.179)\right) - H_d\} \quad (7)$$

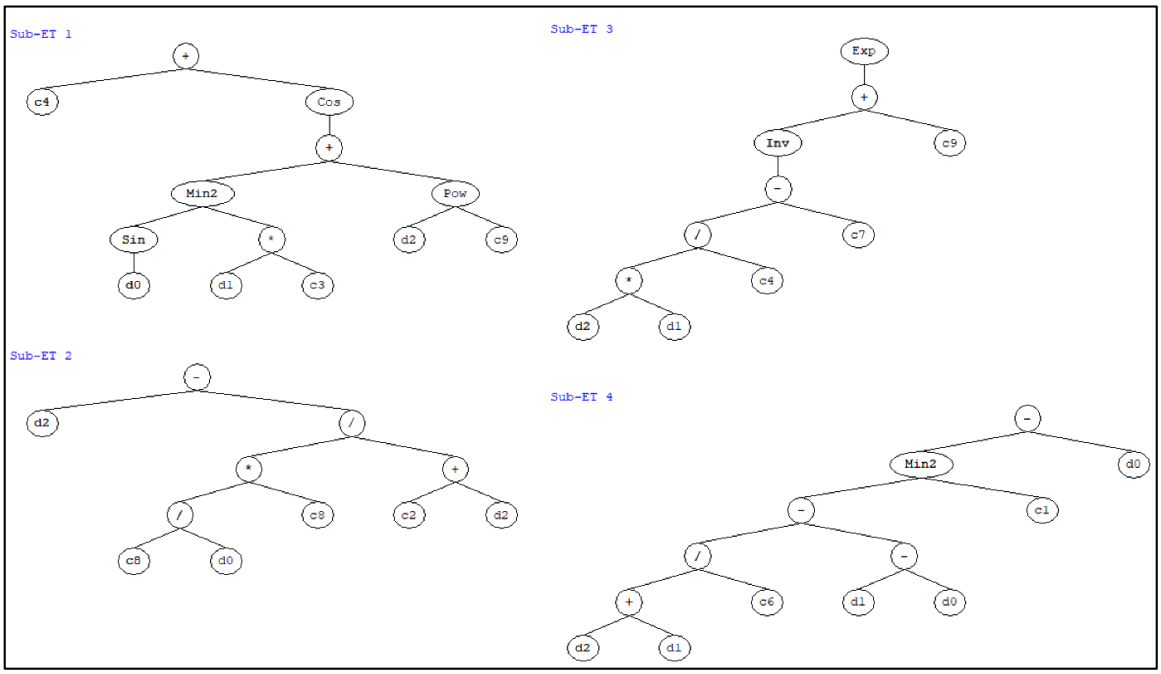

Fig. 11 Expression trees for the GEP-II model

5. Evaluation of the GEP models

5.1 Performance of GEP models

This study presents three statistical parameters (R-square, R2, mean absolute error, MAE and root-mean-squared error, RMSE) in Eqs. (8)-(10) were utilized to evaluate the performance of empirical equations based on GEP models. In these equations, $y_i$is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the average of the observed values, and n is the total number of data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

Also, the performance of model results was compared to the formulation developed by Swaisgood (2014). It was discussed that there were different estimation equations for earthquake-induced dam crest settlements in the introduction section of the paper. The most crucial reason for the selection of Swaisgood (2014) for the comparison is that it is frequently used in different researches and mentioned as the simpler method in the guides of seismic design of embankment dams (FEMA 2005, USBR 2019). Also, the U.S. Army Corps of Engineers (USACE) Risk Management Center (RMC) developed a set of toolboxes for embankment dam crest settlement by using Swaisgood's (2014) formulation (USACE 2022).

Observed and prediction crest settlement values for the training and test data of the GEP models and Swaisgood's (2014) formulation are given in Figs. 12 and 13. Also, the linear least-square fit line and the R^2 values are added to these figures. R^2 values in the training sets are 0.802 and 0.839 for GEP-I and GEP-II models, whereas R^2 values in the test sets are 0.809 and 0.829 for GEP-I and GEP-II models. These figures show that the R^2 values of GEP-I and GEP-II models are higher than the R^2 of Swaisgood (2014) for training and test data. Figs. 12 and 13 exhibit how well the nonlinear relation between parameters is obtained from training and testing results of GEP models.

In addition to the R^2 value, MAE and RMSE values are also calculated for GEP models and Swaisgood's (2014) formulation (Table 5). The larger R^2 and smaller values of MAE and RMSE dictate the accuracy and generalization of the models. Although it can be said that R^2 values of GEP models are little low, it is possible to conclude that the model results are successful since the other statistical evaluation parameters RMSE and MAE are quite small. In addition, it is clear that the prediction capability of the proposed GEP models is sufficient, especially despite the uncertainties inherent in the earthquake behavior and the high non-linear relationship between input and output parameters as seen in Fig. 9.With this perspective, the GEP-II model has successful performance in predicting the earthquake-induced crest settlement. GEP –I model follows the GEP-II model in term of prediction performance.

Statistical parameters are given in Table 5 to show overall performance. To see the performance of the models for each training and testing data, an error plot between the actual and predicted values is given in Fig. 14. As seen in this figure, the error values of the GEP models are smaller than the error values calculated according to Swaisgood's (2014) equations. Especially GEP-II model accurately

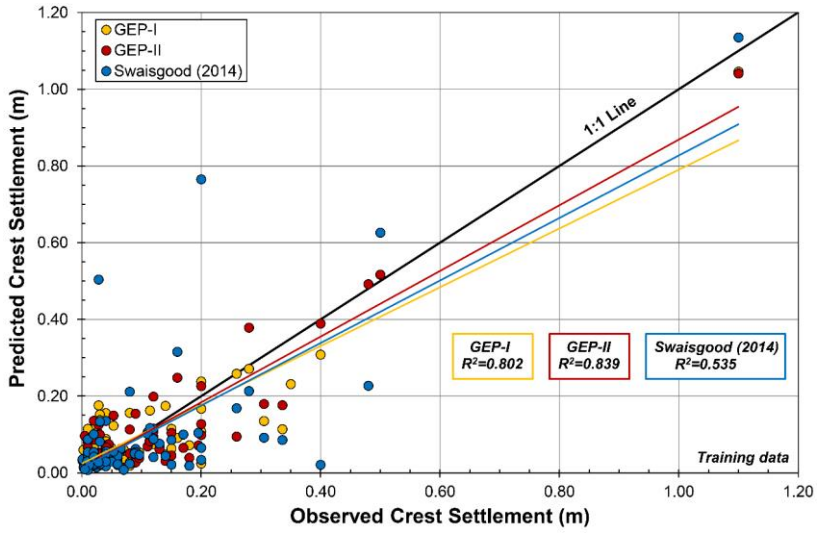

Fig. 12 Comparison of observed and predicted crest settlement values in the training set

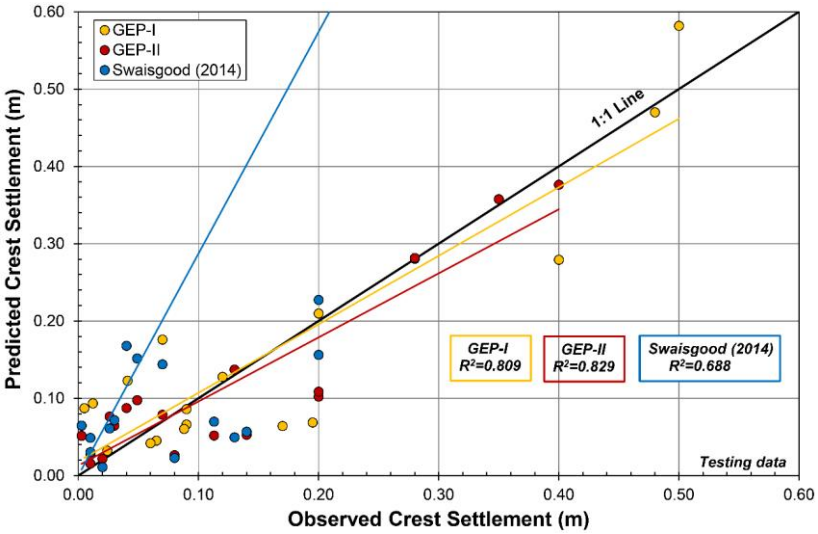

Fig. 13 Comparison of observed and predicted crest settlement values in the testing set

Table 5 Statistical performance of GEP models and previous studies

Model	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
GEP-I	0.802	0.069	0.053	0.809	0.068	0.051
GEP-II	0.839	0.065	0.051	0.829	0.049	0.039
Swaisgood (2014)	0.535	0.127	0.072	0.688	0.422	0.199

captured the target results for both training and test data. The equation of Swaisgood (2014) tends to overestimate the actual data. As a general result of the evaluations, it can be concluded that there is a strong correlation between the predictions of proposed GEP-II formulation and the actual data for training and test sets.

5.2 Sensitivity and parametric study

For this study, sensitivity analysis was performed to determine the contribution of input parameters on earthquake-induced crest settlement value of embankment dams. A simple procedure given in Eqs. (11) and (12)

Table 6 Sensitivity analysis of the input parameters in GEP-II model

Parameter	H_d	H_a	MAF
Sensitivity (%)	25.34	12.02	62.64

$$N_i = f_{max}(x_i) - f_{min}(x_i) \tag{11}$$

$$S_i = \frac{N_i}{\sum_{j=1}^{n} N_j} \times 100 \tag{12}$$

where $f_{min}(xi)$ and $f_{max}(xi)$ are the minimum and maximum values of the predicted output values for the i^{th} input domain while other input values are equal to their mean values. As seen in Table 6, the most influential input variable on the earthquake-induced crest settlement values is MAF parameter calculated based on M and PGA with a contribution of 62.64%. This input parameter is followed by H_d and H_a.

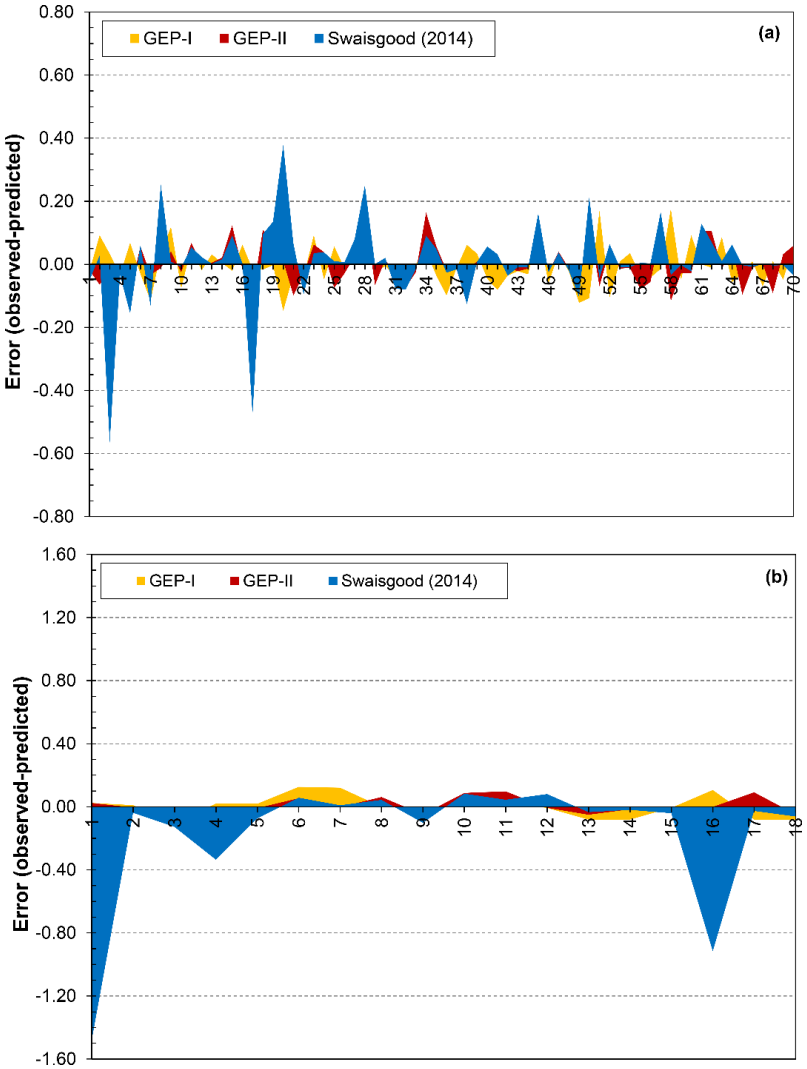

Fig. 14 Error between observed and predicted results of the training set (a) and testing set (b)

In addition to sensitivity analysis, parametric study is an important stage to check and verify the robustness of the developed models. For this purpose, the trends of the output parameter, S_{crest} were evaluated with the input variables. Changes in S_{crest} values were obtained only by changing the value of one variable from minimum to maximum and other inputs were maintained at average values. Fig. 15 illustrates the tendency of the earthquake-induced crest settlement predictions to the variations of the variables, H_d, H_a and MAF. As seen in Fig. 15, S_{crest} increases with increasing H_d, H_a and MAF values. It is clear that parametric analysis show an acceptable trend for the GEP models.

It is understood that the GEP-II model can be used successfully both in terms of statistical success criteria and in terms of parametric study results showing the accuracy of the parameters. However, as in each prediction equation, the success of the prediction equation proposed in this study is based on selection of the suitable and reliable input parameters. A schematic view of the usage of the prediction model was given in Fig. 16. Determining the PGA value is one of the most important steps in the crest settlement value estimation. At this stage, if there is a strong ground motion station at the dam site, the PGA value can be obtained directly. However, if no record has been taken at the dam site, the PGA value can be calculated with ground motion prediction equations (GMPEs). The success of this process depends on the selection of the reliable GMPEs. The database of the prediction equation to be selected should be suitable in terms of dam site geological conditions, faulting type, seismic source-site distance and earthquake magnitude range.

7. Conclusions

In this research, the GEP algorithm was operated to develop the prediction equations for earthquake-induced dam crest settlement based on 88 cases observed in different parts of the world. Two different GEP model was established, namely GEP-I and GEP-II. Dam height (H_d) and MAF (magnitude-acceleration-factor) were selected as input variables on the GEP-I model. In the GEP-II model, besides these two parameters, alluvium thickness (H_a) was also included to model as a variable parameter. The performance of the models and empirical formula developed by Swaisgood (2014) were evaluated based on

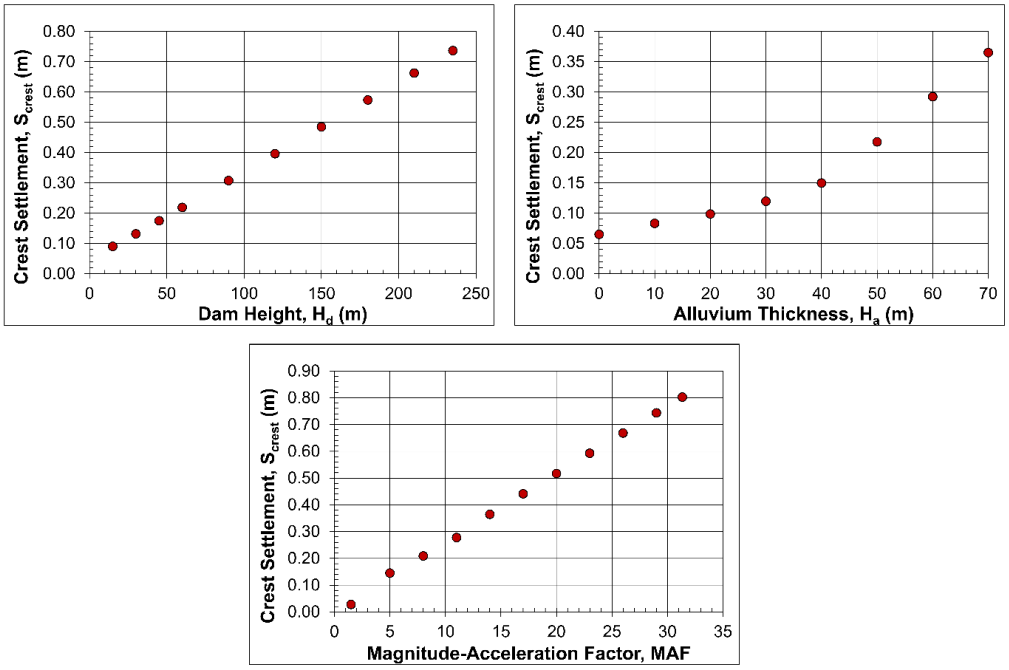

Fig. 15 Parametric analysis of GEP-II model with respect to input parameters

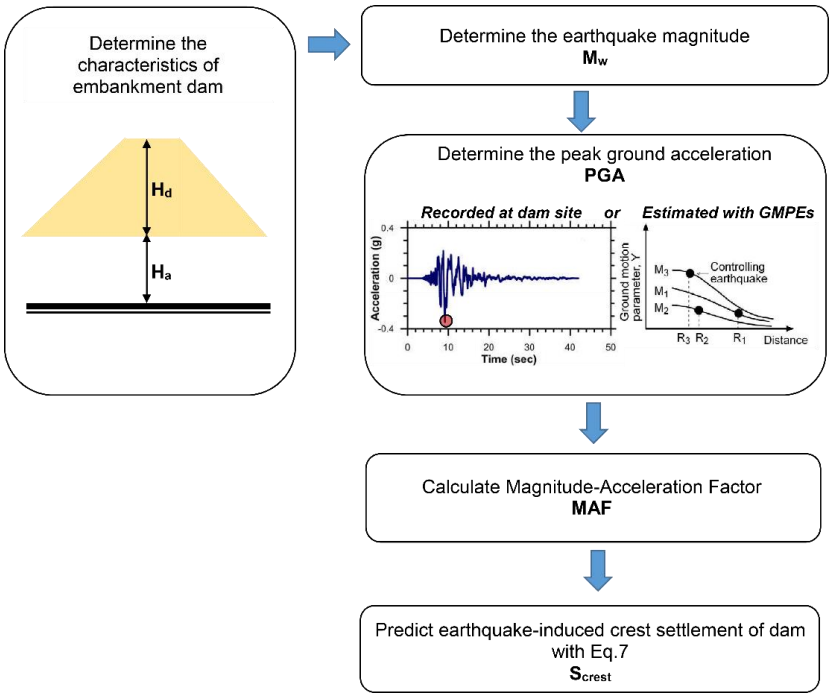

Fig. 16 Schematic view of the procedure of the GEP-II model

statistical results of R2, MAE and MAPE values. The conclusions of the study are summarized below.

- Formulations of the GEP-I and GEP-II models have successful performance compared to the prediction equation by Swaisgood (2014) based on R^2, MAE and MAPE values for both training and test data.
- The formulation of Swaisgood (2014) has the lowest performance in predicting crest settlement for 88 case data. The error graph shows that Swaisgood (2014) overestimated the actual data compared to GEP-I and GEP-II models.
- GEP-II model, which includes alluvial thickness as an input parameter, is the most successful. If the alluvium thickness is not known precisely, the prediction equation based on the GEP-I model can be used
- GEP-II model developed in the paper are robust and accessible tools to predict earthquake-induced crest settlement of embankment dams.
- The fact that the cases in the database used for these GEP models do not contain liquefaction is the most important advantage when compared to the other prediction equations in the literature.

- It is thought that the GEP-II model obtained in this study will be an important alternative for post-earthquake emergency action plan and rapid risk assessment studies, as it is simpler and less time consuming compared to both complex stress strain behavior analyses and Modified Newmark methods.
- Since the variables in both GEP models are easily accessible after the earthquake, they can be used practically by dam engineering professionals.
- In the future, other soft computing methods such as Artificial Neural Network, Fuzzy Logic and Machine Learning can be utilized and results can be compared with GEP models.

References

Adamo, N., Al-Ansari, N., Sissakian, V., Laue, J. and Knutsson, S. (2020), "Dam safety and earthquakes", *J. Earth Sci. Geotech. Eng.*, **10**(6), 79-132.

Alkroosh, I. and Nikraz, H. (2012), "Predicting the axial capacity of driven piles in cohesive soils using intelligent computing", *Eng. Appl. Artif. Intell.*, **25**(3), 618-627. https://doi.org/10.1016/j.engappai.2011.08.009.

Barkhordari, B.K. and Entezari, Z.H. (2015), "Prediction of permanent earthquake induced deformation in earth dams and embankments using artificial neural networks", *Civil Eng. Infrastruct. J.*, **48**(2), 271 -283.

Bray, J.D. and Travasarou, T. (2007), "Simplified procedure for estimating earthquake-induced deviatoric slope displacements", *J. Geotech. Geoenviron. Eng. - ASCE*, **133**(4), 381-392. https://doi.org/10.1061/(ASCE)1090-0241(2007)133:4(381).

Chern, J.C. and Tsai, M.S. (2001), "Seismic safety analysis of earth dam — case history studies", *Proceedings of the* 4th *International Conferences on Recent Advances in Geotechnical Earthquake Engineering and Soil Dynamics*, San Diego, USA, March.

Fell, R., MacGregor, P., Stapledon, D. and Bell, G. (2005), *Geotechnical Engineering of Dams*, CRC Press, London.

FEMA (2005), "Federal guidelines for dam safety:earthquake analyses and design of dams", FEMA-65, Federal Emergency Management Agency.

Ferreira, C. (2001), "Gene expression programming: A new adaptive algorithm for solving problems", *Complex Syst.*, **13**(2), 87-129.

Foster, M., Fell, R. and Spannagle, M. (2000), "The statistics of embankment dam failures and accidents", *Can. Geotech. J.*, ***37***, 1000-1024. https://doi.org/10.1139/t00-030.

Goldberg, D.E. (1989), *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Longman Publishing, Boston, MA, USA.

Hu, H. and Huang, Y. (2019), "A dynamic reliability approach to seismic vulnerability analysis of earth dams", *Geomech. Eng.*, **18(**6), 661-668. https://doi.org/10.12989/gae.2019.18.6.661.

ICOLD (2011), *Constitution*, International Commission on Large Dams, Paris, France.

ICOLD (2022), https://www.icold-cigb.org/

Jalal, F.E., Xu, Y., Iqbal, M., Javed, M.F. and Jamhiri, B. (2021), "Predictive modeling of swell-strength of expansive soils using artificial intelligence approaches: ANN, ANFIS and GEP", *J. Environ. Management*, **289**, 1-17. https://doi.org/10.1016/j.jenvman.2021.112420.

Javdanian, H., Sanayei, H.R. and Shakarami, L. (2020), "A regression-based approach to the prediction of crest settlement of embankment dams under earthquake shaking", *Scientia Iranica*, **27**, 671-681. https://doi.org/10.24200/sci.2018.50483.1716.

Johari, A. and Nejad, A.H. (2015), "Prediction of soil-water characteristic curve using gene expression programming", *Iranian J. Sci. Technol. – T. Civil Eng.*, **39**(1), 143-165.

Johari, A., Habibagahi, G., Nakhaee, M. and Habibagahi, G. (2013), "Prediction of unsaturated soils effective stress parameter using gene expression programming", *Scientia Iranica*, **20**(5), 1433-1444.

Johari, A., Javadi, A. and Najafi, H. (2016), "A genetic-based model to predict maximum lateral displacement of retaining wall in granular soil", *Scientia Iranica*, **23**(1), 54-65. https://doi.org/10.24200/sci.2016.2097.

Kan, M.E., Taiebat, H.A. and Taiebat, M. (2017), "A framework to assess Newmark-type simplified methods for evaluation of earthquake-induced deformation of embankments", *Can. Geotech. J.*, **54**(3), 392-404. https://doi.org/10.1139/cgj-2016-0069.

Karabulut, M. and Genis, M. (2019), "Pseudo seismic and static stability analysis of the Torul Dam", *Geomech. Eng*., **17**(2), 207-214. https://doi.org/10.12989/gae.2019.17.2.207.

Karalar, M. and Cavusli, M. (2019), "Assessing 3D seismic damage performance of a CFR dam considering various reservoir heights", *Earthq. Struct*., **16**(2), 221-234. http://doi.org/10.12989/eas.2019.16.2.221.

Karalar, M. and Cavusli, M. (2021), "Three dimensional seismic deformation-shear strain-swelling performance of America-California Oroville Earth-Fill Dam", *Geomech. Eng*., **24**(5), 443-456. https://doi.org/10.12989/gae.2021.24.5.443.

Kayadelen, C. (2011), "Soil liquefaction modeling by genetic expression programming and neuro-fuzzy", *Exp. Syst. Appl.*, **38**(4), 4080-4087. https://doi.org/10.1016/j.eswa.2010.09.071.

Kayadelen, C., Günaydın, O., Fener, M., Demir, A. and Özvan, A. (2009), "Modeling of the angle of shearing resistance of soils using soft computing systems", *Exp. Syst. Appl.*, **36**(9), 11814-11826. https://doi.org/10.1016/j.eswa.2009.04.008.

Koza, J.R. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection,* The MIT Press, Cambridge, MA.

Makdisi, F.I and Seed H.B. (1978), "Simplified procedure for estimating aam and ambankment earthquake-induced deformations", *J. Geoteh. Eng. Division,* **104**(7), 849-867. https://doi.org/10.1061/AJGEB6.0000668.

Mohammadzadeh, S.D., Kazemi, S.F., Mosavi, A., Nasseralshariati, E. and Tah, J.H. (2019), "Prediction of compression index of fine-grained soils using a gene expression programming model", *Infrastructures*, **4**(2), 1-12. https://doi.org/10.3390/infrastructures4020026.

Mollahasani, A., Alavi, A.H. and Gandomi, A.H. (2011), "Empirical modeling of plate load test moduli of soil via gene expression programming", *Comput. Geotech.*, **38**(2), 281-286. https://doi.org/10.1016/j.compgeo.2010.11.008.

Nasiri, F., Javdanian, H. and Heidari, A. (2020), "Seismic response analysis of embankment dams under decomposed earthquakes", *Geomech. Eng*., **21**(1), 35-51. https://doi.org/10.12989/gae.2020.21.1.035.

Newmark, N.M. (1965), "Effects of earthquakes on dams and embankments", *Geotechnique*, **15**, 139-160. http://dx.doi.org/10.1680/geot.1965.15.2.139.

OMNR (2011), "Geotechnical design and factors of safety", Technical Bulletin, Ontario Ministry of Natural Resources, Canada.

Pells, S. and Fell, R. (2002), "Damage and cracking of embankment dams by earthquakes and the implications for internal erosion and piping", Report No. R-408; School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia.

Pourzangbar, A. (2012), "Determination of the most effective

parameters on scour depth at seawalls using genetic programming (GP)", *Proceedings of the 10th International Conference on Coasts, Ports and Marine Structures (ICOPMASS 2012*), Tehran, Iran, November.

Rathje, E.M. and Bray, J.D. (1999), "An examination of simplified earthquake-induced displacement procedures for earth structures", *Can. Geoteh. J.*, **36**(1), 72-87. https://doi.org/10.1139/t98-076.

Sakamoto, T., Fujisawa, T., Nakamura, A., Yoshida, H. and Iwashita, T. (1998), "Safety estimation of rock-fill dams during major earthquakes", *First U.S.-Japan Workshop on Advanced Research on Earthquake Engineering for Dams,* Vicksburg, USA, November.

Sarıdemir, M., Seercan, M.H. and Çelikten, S. (2017), "Mechanical properties of SFRHSC with metakaolin and ground pumice: Experimental and predictive study", *Steel Compos. Struct.*, **23**(5), 543-555. https://doi.org/10.12989/scs.2017.23.5.543.

Sattar, A.M. (2014), "Gene expression models for prediction of dam breach parameters", *J. Hydroinform.*, **16**(3), 550-571. https://doi.org/10.2166/hydro.2013.084.

Saygili, G. and Rathje, E.M. (2008), "Empirical predictive models for earthquake-induced sliding displacements of slopes", *J. Geotech. Geoenviron. Eng. - ASCE*, **134**(6), 790-803. https://doi.org/10.1061/(ASCE)1090-0241(2008)134:6(790).

Seed, H.B., Lee, K.L., Idriss, I.M. and Makdisi, F.I. (1975), "The slides in the San Fernando Dams during the earthquake of February 9, 1971", *J. Soil Mech. Found. Eng. Division*, **101**(7), 651-688. https://doi.org/10.1061/AJGEB6.0000178.

Seyrek, E. and Tosun, H. (2011), "Deterministic approach to the seismic hazard of dam sites in Kızılırmak basin, Turkey", *Nat. Hazards*, **59**,787-800. https://doi.org/10.1007/s11069-011-9795-7.

Singh, R., Roy, D. and Das, D. (2007), "A correlation for permanent earthquake-induced deformation of earth embankments", *Eng. Geol.*, **90**(3-4), 174-185. https://doi.org/10.1016/j.enggeo.2007.01.002.

Sivrikaya, O., Kayadelen, C. and Cecen, E. (2013), "Prediction of the compaction parameters for coarse-grained soils with fines content by MLA and GEP", *Acta Geotechnica Slovenica*, **10**(2), 29-41.

Swaisgood, J.R. (2003), "Embankment dam deformations caused by earthquakes", *Proceeding of the 7th Pacific Conference on Earthquake Engineering*, Christchurc, New Zealand, February.

Swaisgood, J.R. (2014), "Behavior of embankment dams during earthquake", *J. Dam Saf.*, **12**(2), 35-44.

Taskiran, T. (2010), "Prediction of California bearing ratio (CBR) of fine grained soils by AI methods", *Adv. Eng Softw.*, **41**(6), 886-892. https://doi.org/10.1016/j.advengsoft.2010.01.003.

Tosun, H. (2015), Earthquakes and Dams, In (Ed.), Earthquake Engineering - From Engineering Seismology to Optimal Seismic Design of Engineering Structures, IntechOpen, London, UK.

USACE (2022), RMC Empirical Crest Deformation Toolbox, US Army Corps of Engineers, Institute for Water Resources, Risk Management Center, https://www.rmc.usace.army.mil/Software/RMC/Toolboxes/Seismic-Hazard-Suite/.

USBR (2019), "Seismic risks for embankments", Technical Report: Best Practices in Dam and Levee Safety Risk Analysis-Part IV-Embankments and Foundations, A Joint Publication by U.S. Department of the Interior Bureau of Reclamation and U.S. Army Corps of Engineers, USA.

USCOLD (1992), "Observed performance of dams during earthquakes", Committee on Earthquakes, Denver, CO, USA.

Uysal, F. (2020), "Prediction of collapse potential of soils using gene expression programming and parametric study", *Arabian J. Geosci.*, **13**(19), 1-13. https://doi.org/10.1007/s12517-020-06050-x.

Yamanaka, V.H.A., Fernandez, R.M., Federman, D.K., Elizalde, M. and Aparicio, J. (2020), "Effects of the September 2017 Earthquakes on Mexican Dams", *J. Perform. Constr. Fac.,* **34**(4), 1-21. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001417.

Yousif, O.S.Q., Zaidn, K., Alshkane, Y., Khani, A. and Hama, S.K. (2019), "Performance of Darbandikhan dam during a major earthquake on November 12, 2017", *Proceedings of the 3rd Meeting of EWG Dams and Earthquakes*, Lisbon, Portugal, May.

Zhang , L.M., Xu, Y. and Jia, J.S. (2009), "Analysis of earth dam failures: A database approach", *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, **3**(3), 184-189. https://dx.doi.org/10.1080/17499510902831759.

IC