Gene expression models for prediction of dam breach parameters

Ahmed M. A. Sattar

ABSTRACT

Data from a large database of 140 dam failure cases are used with gene expression programming (GEP) to develop new empirical formulae of physical meaning for prediction of non-dimensional key dam breach parameters. The GEP models are trained on 75% of the data set and validated on the remaining 25%. Parametric and error analyses are conducted to confirm the robustness of the developed relations. Moreover, uncertainty analyses using the Monte Carlo technique is performed to check for the output uncertainty of key dam breach parameters and the contribution of various input parameters to the overall output uncertainty. It is found that uncertainties of 20 to 40% are calculated for the developed GEP models with reservoir shape factor and dam erodibility being main influential predictors.

Key words | dam breach, dam breach parameters, gene expression programming, Monte Carlo simulation, risk, uncertainty

Ahmed M. A. Sattar
Department of Irrigation and Hydraulics,
Faculty of Engineering,
Cairo University,
Giza,
Egypt
Currently at:
German University,
Cairo,
Egypt
E-mail: *ahmoudy77@yahoo.com*

LIST OF NOTATION

B_{avg}	breach average width
B_b	breach bottom width
B_t	breach top width
B_{ub}	width of error band
B^*_{avg}	non-dimensional average breach width
c_j	constants used in developed GEP models
DC	core wall dams
D_e	dam erodibility
D_f	dam failure mode
D_t	dam type
EAs	evolutionary algorithms
e_{ij}	error in prediction
$\bar{e}$	mean prediction error
FD	concrete faced dams
f_j	GEP model fitness function
g	gravitational acceleration
GEP	gene expression programming
GP	genetic programming
h	head of gene
h_b	breach height
H_d	dam height
HD	homogenous dam
HE	high erodible dams
H_r	reference dam height of 15 m
H_w	water depth above breach invert
h^*_b	non-dimensional breach height
H^*_d	non-dimensional dam height
i,j	counter indices
k, k'	gradients of the regression line through the origin
LE	low erodible dams
m	number or random variables
m', n'	coefficient of determination of the regression line through the origin
m_{Vi}	mean value of parameter i of all the random samples
MAD	mean absolute deviation
MCS	Monte Carlo simulation method
ME	medium erodible dams
MLSR	multivariable-least-squares-regression
MNR	multiparameter nonlinear regression
n	number of dam failure cases
N	number of Monte Carlo simulations

doi: 10.2166/hydro.2013.084

O	overtopping failure mode
P	piping failure mode
PDF	probability density function
P_j	value predicted by model
$\bar{P}$	mean value of model predictions
Q_p	breach peak flow
Q_p^*	non-dimensional breach peak flow
R	correlation coefficient
R^2	coefficient of determination
RRSE	relative squared error
RS	reservoir shape factor
RSE	root squared error
R_m	cross validation measures
R_0^2	squared correlation coefficient through the origin between predicted and observed values
$R_0'^2$	squared correlation coefficient through the origin between observed and predicted values
S_e	standard deviation of prediction errors
S_{V_i}	sensitivity coefficient of each parameter i
t	tail length of gene
t_f	breach failure time
t_r	reference breach failure time of 1 hour
t_f^*	dimensionless breach failure time
T_j	value observed
$T_{1,2,3,4,5}$	time of breach formation phases
$\bar{T}$	mean value of observed cases
t_f^*	non-dimensional breach failure time
v_i,	random chosen sample of parameter i
Δv_i	$v_i - m_{v_i}$
V_d	reservoir capacity
V_w	volume of water above breach invert
$V_{i,\text{true}}$	true value of parameter i
w_i	regression coefficients
X_i	predictors or breach parameters
Y_i	non-dimensional main breach parameter
Z	breach side slope
ZD	zone-filled dams
δ_{Vi},	true uncertainty of parameter i
$\sigma_{\delta_y}^2$	overall variance of the model output
$\sigma_{\delta_{v_i}}^2$	variance of the calculated difference δ_{Vi}
ε	maximum allowable system error
$\varphi_{\alpha/2}$	standard normal deviate corresponding to a two sided confidence level

INTRODUCTION

The two primary tasks in the study of dam failure are the prediction of the dam breach outflow hydrograph and the routing of that hydrograph through the downstream valley. The routing of dam-break waves is a well-developed research. Good progress has been made in this field, including dry-bed effects (Guo *et al.* 2008), the development of high-resolution schemes (Lin *et al.* 2003) and dam-break flows over mobile bed (Wu & Wang 2007). The greater source of uncertainty in most situations is the prediction of the dam breach outflow hydrograph. In dam failure risk engineering, it is of utmost importance to predict the dam breach outflow hydrograph and its timing relative to events in the failure process and route it through the downstream valley to determine consequences and possible inundations. Costa (1988) reported that the average number of fatalities per dam failure is 19 times greater when there is inadequate or no warning. Warning time is the summation of the breach initiation time, breach formation time, and flood wave travel time from the dam to a population center. Predicting the breach outflow hydrograph depends on the dam breaching process and can be done by identifying key breaching parameters and linking them to dam and reservoir parameters. The US Bureau of Reclamation (1988) grouped outflow hydrograph prediction models based on previous breach key parameters as predictor models, parametric models, and physically based models. However, the current most widely used methods are parametric, which make use of breach parameter estimates derived from regression equations (Wahl *et al.* 2008). Parametric models estimate failure time and breach average width using regression-based methods from past case studies; then simulate breach growth as a linear process and compute the outflow hydrograph from basic hydraulic principles. All available parametric prediction equations are based on regression analysis for selected cases of actual dam failures. All of these prediction equations have major shortcomings: (1) the database of dam failures used to develop these relations is lacking data from failures of large dams with about 75% of the cases having a dam height less than 15 m (Wahl 1998); (2) the present regression-based equations did not distinguish between

different failure modes (Wahl 1998), except for Froehlich (1995a) who considered overtopping failure modes separately from other modes; (3) they did not include parameters reflecting the embankment erodibility or dam type (Morris *et al.* 2009), except for MacDonald & Langridge-Monopolis (1984) who related failure time to the volume of eroded material; and (4) they did not provide adequate representation of modeled data and have large prediction errors. Best prediction models amongst available models had large uncertainties reaching ±1/3 order of magnitude for breach width, and ±2/3 order of magnitude for failure time with prediction errors as high as 675% (Wahl 1998, 2004).

Recently, evolutionary algorithms have been devised by various researchers as a superior alternative for regression analysis and artificial neural networks, for finding relations between various parameters and producing a higher R-squared value and less mean prediction error. Gene expression programming (GEP) involves computer programs of different sizes and shapes encoded in linear chromosomes of fixed length. GEP chromosomes are composed of multiple genes with each gene encoding a smaller sub-program (Ferreira 2001). Applications of evolutionary algorithms, especially GEP in water and environmental engineering are fewer than for the other soft computing tools of artificial neural networks. They are restricted to finding functions in relatively fewer sub-areas including scour prediction downstream of hydraulic structures (Guven & Gunal 2008), daily reference evapotranspiration prediction (Guven *et al.* 2008), predicting sediment transport in sewer pipe systems (Ghani & Azamathulla 2011), developing reservoir operation rules (Fallah *et al.* 2012), rainfall-runoff modeling (Nourani *et al.* 2012), and prediction of bridge pier scour (Azamathulla 2011).

This paper aims at applying GEP to develop new predictive empirical models for prediction of dam breach main parameters. At first, a brief overview on the dam breach process and resulting hydrograph is presented in addition to presenting the parameters of the reservoir and breach included in the collected database. Afterwards, GEP is applied to the database and five models are developed. The developed models are analyzed parametrically to test their physical behavior in addition to validation analyses. Finally, error and uncertainty analysis are performed on the developed models to assess their reliability and robustness.

DAM BREACH OUTFLOW HYDROGRAPH

A dam breach is the failure in embankment, either due to erosion of embankment material or structural failure allowing the release of flood water over/through the embankment in an uncontrolled manner. This failure typically takes the shape of a hole or a gap in the body of the embankment through which flood water exits. The trapezoidal opening is the typical breach shape happening in almost all dam failure cases (Morris *et al.* 2008) as shown in Figure 1. Geometric parameters describing the breach are: breach height measured from the dam crest down to the breach invert (h_b), top width (B_t), bottom width (B_b), average width (B_{avg}), and breach side slope (Z:1).

The rate and size of breach formation depends mainly on dam embankment type and material. Embankment dams are made of compacted earth with the two main types of earthfill and rockfill dams relying on their weight to trap water behind. The embankment type plays an important role in the breach process with the simplest design being homogenous dams (HD) that are constructed of more or less uniform earth materials. The other type of dams is the zoned-fill dam (ZD), which has distinct zones of dissimilar material; typical construction of a ZD uses a local shell with watertight clay core with filter and drain zones preserving the integrity of the shell zone (Dyer *et al.* 2007). Some embankment designs advance from the simple and multilayer embankments to include additional structural measures for better erosion and seepage

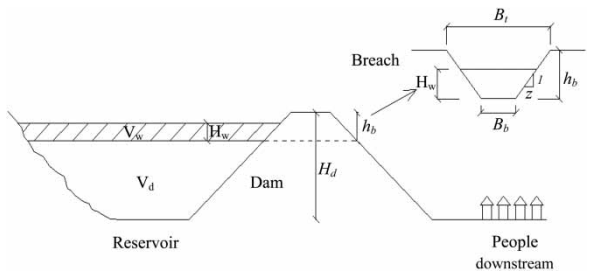

Figure 1 | Geometric parameters for typical dam breach.

protection. This includes the concrete faced dam (FD), which has concrete slabs on its upstream and dams with corewalls (DC) that contain a low-permeability wall built vertically or inclined towards the dam upstream offering an impervious wall to prevent leakage and seepage through the dam body. Failures of HD and ZD are mostly attributed to piping, while DC and FD have the dominant failure mode of overtopping. It is found that over 60% of the cases of dam failure are either caused by overtopping or piping (Singh 1996).

Morris *et al.* (2009) provided detailed explanation on the effect of embankment construction material categories on the breaching process. Construction material categories include non-cohesive fill, cohesive fill, and rock fill. Non-cohesive fill allows quicker erosion to occur to the embankment where removal of material is through surface erosion due to overtopping, which begins at a point where the tractive shear stress exceeds a critical resistance that keeps the material in place. Cohesive fills cause the rate and process of erosion to vary significantly where head cutting dominates with steps created in the eroded face towards the upstream due to overtopping with no seepage occurring due to the low permeability. For rock fills, a surface slip takes place quickly due to seepage or piping on the downstream slope and granular material is removed rapidly in layers (Singh 1996).

The breach process has been analyzed by various researchers for defining breach stages and development, but no universal definition is available. A widely adopted one divides the breach process into two stages: one is the breach initiation and the second is the breach development or formation; for example, Wahl (1998) and Morris *et al.* (2009). Morris *et al.* (2009) related the breaching process to the breach flood hydrograph and behavior of embankment material to best describe the stages of breaching. Figure 2 shows a typical breach hydrograph. Maximum breach flow occurs during the breach formation phase (T_3 to T_5) with Q_p and time of occurrence T_4, being functions of the available storage volume behind the breach and embankment design and material. According to Wurbs (1987), the peak flow Q_p is also a function of the dam type. For large dams, peak discharge occurs with maximum breach width and depth, while in small dams, peak discharge occurs before full development of the breach due to the significant change in reservoir level during the breach formation stage. Singh (1996) defined the failure time t_f, as the time period for both stages of the embankment breach process; i.e., from T_1 to T_5. Fread (1988) defined the time of failure as the time between dam face upstream breaching until full formation of the breach. In case of overtopping failures, dam face upstream breaching starts after the downstream face has eroded away and the resulting breach has progressed back across the width of the dam crest. Therefore, the breach failure time t_f in the current study is taken as the breach formation time, i.e., from T_3 to T_5. This is the failure time definition adopted by all researchers in the field (e.g., Singh 1996; Wahl 1998; Morris *et al.* 2009; Xu & Zhang 2009; Wahl 2010, among others). Moreover, all available data on failure time are measured and recorded based on this definition.

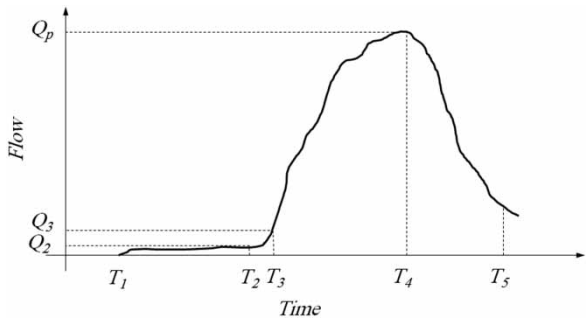

Figure 2 | Generic breach hydrograph (after Morris *et al.* 2009).

Inundation maps for the downstream valley are produced by routing the breach outflow hydrograph through the downstream valley, where population areas at risk are identified. The importance of predicting the breach outflow hydrograph is more apparent when the population at risk is located close to the dam, where peak attenuation and other flood routing effects have not yet taken place (Wahl 2010). In the context of breach outflow hydrograph prediction, various modeling strategies and approaches are found in the literature ranging from simple regression to process-based erosion and hydraulic models; e.g., BREACH model, Fread (1988), and Temple *et al.* (2005). In addition to breach peak outflow Q_p, Wahl (2010) identified the breach average width B_{avg} and failure time t_f as being the two parameters of most interest to fully describe the breach outflow hydrograph.

DAM FAILURE DATABASE

While there are numerous case studies for dam failures with associated breach process and parameters throughout the world, only a few cases are well documented with high quality data. This forced several researchers to present various empirical predictive models for breach parameters based on a compilation of a small number of case studies ranging from six (Hagen 1982) to 63 (Froehlich 1995a) and recently 75 cases (Xu & Zhang 2009), most of them for relatively small dams. In this study, a database of 140 earthfill and rockfill embankment dam failures is collected from various sources: Singh (1996), Wahl (1998), Zhumadian Water Resources Authority (1997), Henan Water Resources Authority (2005), and Xu & Zhang (2009). All of the 140 cases have sufficient information for breach parameter predictive models as shown in Table 1.

Singh (1996) suggested a dam classification based on size, classifying as small, medium, large, and major dams. In the classification, large and major dams have height more than 15 m and/or reservoir capacity more than 3×10^6 m^3. Figure 3 shows the frequency distribution of dam failure case studies according to the Singh (1996) classification. It is observed that 40% of the dam failure cases are for large dams higher than 15 m and 51% of the cases are for large dams with reservoir capacity exceeding 3×10^6 m^3. Thus, the selected database can serve as a good basis for predictive model development compared to previous databases. On the other hand, the selected database (shown in Table 1) contains another important classification of dams according to their erodibility, where dams are classified into high HE, medium ME, and low erodible LE, as suggested by Xu & Zhang (2009). Dam material erosion is the predominant mechanism for breaching in earthfill and/or rockfill dams, where water flows over the dam causing overtopping and external erosion or through the dam leading to internal erosion or piping. In Table 1, the dams are classified according to their failure mode D_f, which is either piping failure P, or overtopping failure O. The availability of such data as dam erodibility and dam failure type allows the formulation of more general predictive models incorporating various important breach parameters and dam material specifications.

The database contains detailed information on many parameters of dam failure; this includes geometric, hydraulic and geomorphic parameters. Geometric parameters for the dam include: dam height (H_d), volume of water above breach invert (V_w) and water depth above breach invert (H_w), while geometric parameters for the breach include: breach shape, height (h_b) and average width (B_{avg}). While, hydraulic parameters of breach include: failure mode (D_f), peak outflow rate (Q_p), and failure time (t_f). Finally, the geomorphic parameters included in the database include dam type (D_t), and erodibility (D_e). Table 2 presents descriptive statistical analysis for the dam failure dataset used in GEP model development.

EVOLUTIONARY ALGORITHMS

Evolutionary algorithms (EAs) are a class of problem-solving techniques based on the Darwinian theory of evolution and involve searching within a population of solutions. A possible and acceptable solution, i.e., a member of the population, is called an individual. Each iteration of an EA includes a competitive selection that weeds out poor solutions through the evaluation of a fitness value that indicates the quality of the individual as a solution to the problem. GEP was invented by Ferreira in 1999, and is the natural development of EAs. The great insight of GEP was the invention of chromosomes capable of representing any expression tree (ET); GEP surpasses genetic programming (GP) (Ferreira 2001). In GEP, complex relations are encoded in simpler, linear structures of a fixed length called chromosomes. The chromosomes consist of a linear symbolic string of a fixed length composed of one or more genes. To express the genetic information encoded in the gene, Ferreira (2001) used ET representations. Due to the simple rules that determine the structure of the ET, it is possible to infer the gene composition given the ET and vice versa using the unequivocal Karva language. The Karva language represents genes in a sequence that begins with a start codon, continues with amino acid codons, and ends with a termination codon. Consider, for example, the following algebraic expression (also referred to as a program):

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

Table 1 | Dam failure case studies utilized in this study

#	Name	D_t	D_e	D_f	H_d (m)	V_w (x10^{6}m^3)	H_w (m)	h_b (m)	B_{avg} (m)	Q_p (m^3/s)	t_f (hr)
1	Apishapa, USA	HD	HE	P	34.1	22.2	28	31.1	86.5	6,850	2.5
2	Baimiku, China	HD	ME	O	8	0.2	8	8	–	–	–
3	Baldwin Hills, USA	HD	–	P	71	0.91	12.2	21.3	25	1,130	1.3
4	Banqia, China	DC	HE	O	24.5	607.5	31	29.5	291	78,100	5.5
5	Bayi, China	HD	ME	P	30	23	28	30	40	5,000	–
6	Bearwallo Lake, USA	HD	–	P	–	0.0493	5.79	6.4	12.2	–	–
7	Bradfield, UK	HD	–	P	28.96	–	–	–	–	1,150	0.5
8	Break Neck, USA	–	–	–	–	–	–	7	30.5	–	3
9	Buckhaven, USA	–	–	O	–	0.0247	6.1	6.1	4.72	–	–
10	Buffalo Creek, USA	HD	–	P	14.02	0.484	14.02	14	125	1,420	0.5
11	Bullock Draw, USA	HD	–	P	5.79	0.74	3.05	5.79	12.5	–	–
12	Butler, USA	HD	–	O	–	2.38	7.16	7.16	62.5	810	–
13	Canyon, USA	–	–	O	6.1	–	–	–	–	–	0.1
14	Castlewood, USA	DC	ME	O	21.3	6.17	21.6	21.3	44.2	3,570	–
15	Caulk Lake, USA	–	–	P	–	0.698	11.1	12.2	35.1	–	–
16	Cheaha Creek, USA	ZD	–	O	7	–	–	–	–	–	5.5
17	Clearwater Lakel, USA	HD	–	O	–	0.466	4.05	3.78	22.8	–	–
18	Chenying, China	HD	ME	O	12	5	12	12	–	1,200	1.83
19	Coedty, UK	DC	HE	O	11	0.311	11	11	42.7	–	–
20	Cougar Creek, Canada	–	–	–	–	0.0298	11.1	10.4	–	–	–
21	Dalizhuang, China	HD	ME	O	12	0.6	12	12	–	–	–
22	Danghe, China	DC	LE	O	46	10.7	24.5	25	58	2,500	3
23	Davis, USA	FD	ME	P	11.9	58	11.58	11.9	18.3	510	–
24	DMAD, USA	HD	–	–	8.8	19.7	–	–	–	793	–
25	Dells, USA	–	HE	O	18.3	13	18.3	18.3	–	5,440	0.67
26	Dongchuankou, China	HD	HE	O	31	27	31	31	–	21,000	–
27	Dushan, China	HD	ME	O	17.7	0.67	17.7	17.7	–	–	–
28	Elk, USA	DC	ME	O	9.1	1.18	9.44	9.14	36.6	–	0.83
29	Emerym, USA	–	–	P	–	0.425	6.55	8.23	10.8	–	–
30	Erlangmiao, China	HD	ME	O	12.1	0.196	9	9	18.8	–	–
31	Erindale, Canada	DC	–	O	10.67	–	–	4.6	–		0.5
32	Euclides de Cunha, Brazil	HD	–	O	53.04	–	58.22	53	–	1,020	7.3
33	Fengzhuang, China	HD	ME	O	10	0.625	8	8	35	–	–
34	Fogelman, USA	–	–	P	–	0.493	11.1	12.6	7.62	–	–
35	Frankfurt, Germany	HD	LE	P	9.8	0.352	8.23	9.75	6.9	79	2.5
36	Fred Burr, USA	HD	–	–	10.4	0.75	10.2	10.4	–	654	–
37	French, USA	HD	HE	P	12.2	3.87	8.53	14.2	27.4	929	0.58
38	Frenchman, USA	HD	ME	P	12.5	16	10.8	12.5	54.6	1,420	3
39	Frias, Argentina	FD	ME	O	15	0.25	15	15	–	400	0.25
40	Goose Creek, USA	HD	–	–	6.1	10.6	1.37	4.1	26.4	565	0.5

(*continued*)

Table 1 | continued

#	Name	D_t	D_e	D_f	H_d (m)	V_w (x10^{6}m^3)	H_w (m)	h_b (m)	B_{avg} (m)	Q_p (m^3/s)	t_f (hr)
41	Gouhou, China	FD	LE	P	71	3.18	44	48	99.5	2,050	2.33
42	Grand, USA	DC	ME	O	7.6	0.255	7.5	7.5	10.7	–	0.5
43	Granite Creek, USA	–	–	–	–	–	–	–	–	1,841	–
44	Hass Pond, USA	–	–	P	–	0.0234	2.99	3.96	10.7	–	–
45	Hart, USA	HD	HE	P	–	6.35	10.7	10.8	73.9	–	–
46	Hatchtown, USA	ZD	–	P	192	14.8	16.8	18.3	151	3,080	3
47	Hatfield, USA	–	HE	O	6.8	12.3	6.8	6.8	91.5	3,400	2
48	Hebron, USA	HD	–	P	11.58	–	12.9	15.3	45.7	–	2.25
49	Hell, USA	–	ME	P	67.1	30.6	35.1	56.4	121	7,360	–
50	Herrin, USA	ZD	–	O	–	–	10.7	10.7	47.2	–	–
51	Horse, USA	FD	ME	P	12.2	12.8	7.01	12.8	73.1	3,890	3
52	Hougou, China	HD	ME	O	8	0.24	8	8	20	–	–
53	Hutchinson Lake, USA	HD	–	O	–	1.17	4.42	3.75	33.4	–	–
54	Huoshishan, China	HD	HE	O	13	0.22	16	16	30	–	–
55	Huqitang, China	HD	LE	P	9.9	0.424	5.1	9	7.5	50	4
56	Iowa Beef, USA	HD	–	P	4.57	0.333	4.42	4.57	16.8	–	–
57	Ireland No.5, USA	HD	–	P	–	0.16	3.81	5.18	13.5	110	0.5
58	Jacobs Creek, USA	–	–	P	–	0.423	20.1	21.3	17.5	–	–
59	Jiahezi, China	HD	HE	P	18	42	12	18	–	–	–
60	Johnston City, USA	HD	–	P	4.27	0.575	3.05	5.18	8.23	–	–
61	Johnstown, USA	ZD	ME	O	38.1	18.9	24.6	24.4	94.5	8,500	–
62	Kaddam, India	HD	–	O	12.5	–	–	15.2	137.2	–	1
63	Kelly, USA	HD	HE	O	11.6	0.777	11.3	12.8	27.3	680	0.5
64	Kendall Lake, USA	HD	–	O	5.49	–	–	–	–	–	–
65	Kodaganar, India	HD	ME	O	11.5	12.3	11.5	11.5	–	1,280	–
66	Kraftsmens Lake, USA	HD	–	O	–	0.177	3.66	3.2	14.5	–	–
67	La Fruta, USA	HD	–	P	–	78.9	7.9	14	58.8	–	–
68	Lake, USA	HD	ME	P	15.2	0.789	14	17.1	18.9	–	1
69	Lake, USA	HD	ME	P	13	4.09	6.25	8.69	39.2	290	3
70	Lake Avalon, USA	HD	–	P	14.5	31.5	13.7	14.6	130	2,320	2
71	Lake Barcoft, USA	HD	–	O	21.03	–	–	11	–	–	1
72	Lake Genevieve, USA	HD	–	P	–	0.68	6.71	7.92	16.8	–	–
73	Lake Philema, USA	HD	–	O	–	4.78	9	8.53	47.2	–	–
74	Lambert Lake, USA	HD	–	P	–	0.296	12.8	14.3	7.62	–	–
75	Laurel Run, USA	HD	–	O	12.8	0.555	14.1	13.7	35.1	1,050	–
76	Lawn Lake, USA	HD	–	P	7.9	0.798	6.71	7.62	22.2	510	–
77	Lijiaju, China	HD	ME	O	25	1.14	25	25	–	2,950	–
78	Lily Lake, USA	HD	–	P	–	0.0925	3.35	3.66	10.8	71	–
79	Little, USA	HD	HE	P	26.2	1.36	22.9	27.1	29.6	1,330	0.33
80	Liujiatai, China	DC	ME	O	35.9	40.54	35.9	35.9	–	28,000	–

(*continued*)

Table 1 | continued

#	Name	D_t	D_e	D_f	H_d (m)	V_w (x10^{6}m^3)	H_w (m)	h_b (m)	B_{avg} (m)	Q_p (m^3/s)	t_f (hr)
81	Longtun, China	DC	HE	O	9.5	30	9.5	9.5	–	–	–
82	Long Branch, USA	–	–	P	–	0.284	3.17	3.66	9.14	–	–
83	Lower Latham, USA	HD	–	P	–	7.08	5.79	7.01	79.2	340	–
84	Lower, USA	DC	ME	O	41.2	49.3	39.6	39.6	133	–	–
85	Lower, USA	HD	HE	O	11.3	19.6	11.3	11.3	67	1,800	–
86	Lyman, USA	ZD	HE	P	19.8	35.8	16.2	19.8	97	–	–
87	Lynde, USA	DC	ME	P	12.5	2.88	11.6	12.5	30.5	–	–
88	Mahe, China	HD	HE	O	19.5	23.4	19.5	19.5	–	4,950	–
89	Machhu II, India	HD	–		60.05	–	–	60	–	–	2
90	Mammoth, USA	DC	ME	O	21.3	13.6	21.3	21.3	–	2,520	3
91	Martin, USA	–	HE	P	10.4	136	8.53	12.8	186	3,115	–
92	Melville, USA	ZD	–	P	10.97	24.7	7.92	9.75	32.8	–	–
93	Merimac Lake, USA	HD	–	O	–	0.0696	3.44	3.05	14.2	–	–
94	Mill River, USA	–	–	–	13.1	2.5	–	–	–	1,645	–
95	Nanaksagar, India	–	–	–	15.85	–	–	16	46	9,700	12
96	Nahzille, USA	HD	–	O	5.49	–	–	5.03	6.71	–	–
97	North Branch, USA	HD	–	–	5.5	0.0222	5.49	–	–	29.4	–
98	Niujiaoyu, China	DC	LE	P	10	0.144	7.2	7.2	13	–	3
99	Oakford Park, USA	DC	–	O	6.1	–	–	4.6	–		1
100	Oros, Brazil	ZD	LE	O	35.4	660	35.8	35.5	165	9,630	–
101	Otter, USA	HD	ME	P	6.1	0.109	5	6.1	9.3	–	–
102	Otto Run, USA	HD	–	–	5.8	0.0074	5.79	–	–	60	–
103	Pierce Reservoir, USA	HD	–	P	–	4.07	8.08	8.69	30.5	–	1
104	Potato, USA	HD	ME	O	–	0.105	7.77	7.77	16.5	–	–
105	Prospect, USA	HD	HE	P	–	3.54	1.68	4.42	88.4	116	–
106	Puddingstone, USA	HD	–	O	–	0.617	15.2	15.2	–	480	0.25
107	Qielinggou, China	HD	HE	O	18	0.7	18	18	–	2,000	0.17
108	Quail, USA	HD	ME	P	24	30.8	16.7	21.3	70	3,110	–
109	Rainbow Lake, USA	HD	–	O	–	6.78	10	9.54	38.9	–	–
110	Renegade Resort, USA	–	–	O	–	0.0139	3.66	3.66	2.29	–	–
111	Rito, USA	HD	HE	P	7.3	0.0247	4.57	7.32	13.3	–	–
112	Salles Oliveira, Brazil	HD	–	–	35.05	71.5	38.4	35	168	7,200	2
113	Sandy Run, USA	HD	–	O	8.53	0.0567	8.53	–	–	435	–
114	Schaeffer, USA	DC	HE	O	30.5	4.44	30.5	30.5	137	4,500	0.5
115	Scott Farm, Canada	–	–	P	–	0.086	10.4	11.9	15	–	–
116	Shangliuzhuang, China	HD	ME	O	14	0.11	14	14	–	–	–
117	Shanhu, China	HD	HE	P	11.5	1.78	12.5	13	41	–	–
118	Sheep, USA	HD	ME	P	17.1	0.91	14.02	17.1	22	–	–
119	Sherbume, USA	DC	–	P	10.36	–	–	–	–	960	2
120	Shilongshan, China	HD	ME	O	14	2.06	14	14	–	–	–

(*continued*)

Table 1 | continued

#	Name	D_t	D_e	D_f	H_d (m)	V_w (x10^{6}m^3)	H_w (m)	h_b (m)	B_{avg} (m)	Q_p (m^3/s)	t_f (hr)
121	Shimantan, China	HD	HE	O	25	117	27.4	25.8	367	30,000	5.5
122	Sinker Creek, USA	HD	–	P	21.34	3.33	21.34	21.3	70.6	–	2
123	South Fork, USA	HD	–	–	1.8	0.0037	1.83	–	–	122	–
124	Spring, USA	HD	HE	P	5.5	0.136	5.49	5.49	14.5	–	–
125	Statham, USA	HD	ME	O	5.5	0.564	5.55	5.12	21	–	–
126	Swift, USA	FD	ME	O	57.6	37	47.85	57.6	225	24,947	–
127	Teton, USA	ZD	ME	P	93	310	77.4	86.9	151	65,120	4
128	Tiemusi, China	HD	HE	O	12	0.11	12	12	–	–	–
129	Tongshuyuan, China	HD	ME	O	13	0.4	10	10	–	–	–
130	Trial, USA	–	ME	P	–	1.48	5.18	5.18	21	–	–
131	Trout Lake, USA	–	–	O	–	0.493	8.53	8.53	26.2	–	–
132	Upper, USA	–	ME	O	5.2	0.222	5.18	5.18	16.5	–	–
133	Wanshangang, China	HD	ME	O	13	1.5	12	12	40	–	–
134	Wheatland No.1, USA	HD	–	P	13.6	11.6	12.2	13.7	35.4	–	1.5
135	Wilkinson, USA	DC	HE	P	3.2	0.533	3.57	3.72	29	–	–
136	Winston, USA	DC	LE	O	7.3	0.662	6.4	6.1	19.8	–	5
137	Yuanmen, China	HD	HE	O	19.2	6.4	19.2	19.2	–	–	0.5
138	Zhonghuaju, China	HD	HE	O	16	0.14	16	16	–	–	0.4
139	Zhugou, China	DC	HE	O	23.5	18.43	23.5	23.5	135	11,200	0.43
140	Zuocun, China	DC	HE	O	35	40	35	35	–	23,600	1

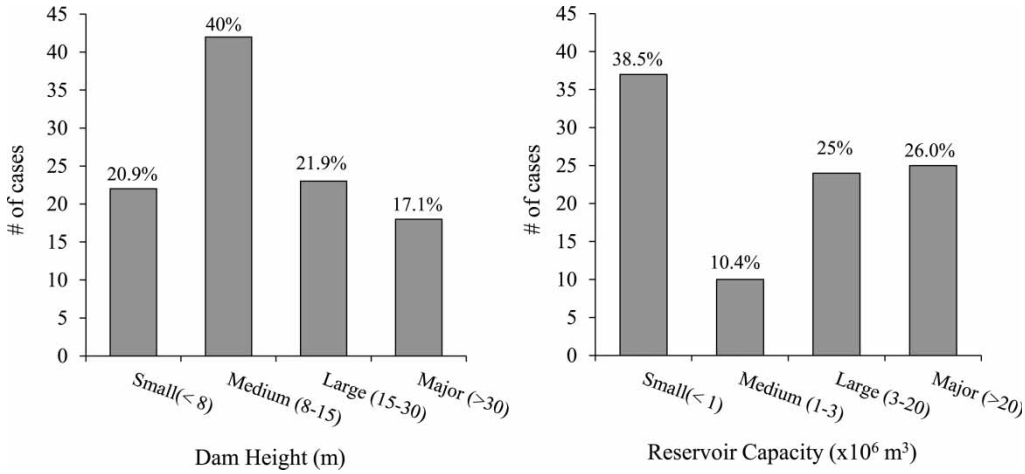

Figure 3 | Frequency distribution of dam failure cases according to dam size.

Table 2 | Descriptive statistics for dam failure database

Parameter	H_d (m)	V_d (x10^{6}m^3)	V_w (x10^{6}m^3)	H_w (m)	h_b (m)	B_{avg} (m)	Q_p (m^3/s)	t_f (hr)
Minimum	1.8	0.02	0.004	1.37	3.05	2.29	29	0.10
Maximum	192	650	660.000	77.40	86.90	367	78,100	12
Mean	20.2	37.84	23.090	14.09	15.54	55.10	6,420	2.10
Median	12.7	3.16	1.140	11.05	12.00	31.65	1,722	1.66
Standard deviation	23.4	105.40	86.000	11.87	13.12	61.92	13,671	2.14
Kurtosis	28.2	18.50	42.290	7.36	8.20	7.75	16	7.86
Skewedness	4.5	4.20	6.330	2.30	2.48	2.46	4	2.34

where a, b and c are the set of terminals or the variables used in the problem; and $\times, \div$ are the rules (functions) that determine the spatial organization of the terminals. The expression can be represented as an ET (see Figure 4).

The Karva expression of the gene inferred from the above ET is as follows:

$$\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \\ \div & \times & a & b & c \end{array} \quad (2)$$

This Karva expression is a straightforward reading of the ET from left to right and from top to bottom. In this case, the gene starts at '÷' (position 0) and terminates at 'c' (position 4). Equation (2) represents the head of the gene, which contains symbols that represent both functions and terminals. In GEP, the gene consists of a head and a tail. The tail of the gene is a junk sequence of terminals that are extremely important because they allow the modification of the next generations of genes using any genetic operator without restrictions. GEP randomly selects the junk sequence. For each problem, the length of the head h is chosen, whereas the length of the tail t is a function of the head and of the number of variables n: $t = h\ (n-1)\ +1$. Considering the above gene with $h = 5$ and $n = 3$, then $t = 11$. Thus, the length of the gene is $5 + 11 = 16$, as shown below with the tail in bold:

$$\begin{array}{cccccccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 \\ \div & \times & a & b & c & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{b} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{a} & \boldsymbol{a} \end{array} \quad (3)$$

In the above Karva expression, the terminal point of the gene shifts 11 positions to the right (position 16). This Karva expression represents a chromosome that contains only one gene. However, GEP chromosomes are usually composed of more than one gene. Genes interact with one another using genetic operators, thus forming a more complex generation of multigenic chromosomes. For example, a multigenic chromosome with a length of 32 that is composed of two genes from Equation (3) will have the following form:

Gene 0:

$$\begin{array}{cccccccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 \\ \div & \times & a & b & c & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{b} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{a} & \boldsymbol{a} \end{array}$$

Gene 1:

$$\begin{array}{cccccccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 \\ \div & \times & a & b & c & \boldsymbol{c} & \boldsymbol{b} & \boldsymbol{c} & \boldsymbol{c} & \boldsymbol{c} & \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{b} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} \end{array} \quad (4)$$

The above genes encode two separate building blocks that can evolve independently. Therefore, they are more effective than single-genic chromosomes and enable a better representation of complex relations.

For each problem, the chromosome architecture must be determined before the initiation of the solution. The

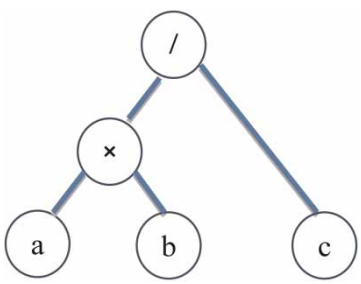

Figure 4 | ET representation of a gene.

architecture includes the number of genes and the length of the head. Moreover, the set of functions must be determined. These functions are the essence of evolution in GEP; they allow modifications without restrictions within the same gene or between various genes in a multigenic chromosome. The choice of an appropriate function set is not the same for every problem and depends primarily on the program performance with the chosen arguments. However, a professional approach would be to initially use the basic mathematical operators ($+$, $-$, $\times$, $/$) to enable the production of simple models.

Following the initial problem set, the GEP creates a random distribution of functions and terminals in the chromosome genes. The first created individual (program) is random and called the 'parent', e.g., the programs shown in Equations (3) and (4). The parents are made to yield 'offspring' through the implementation of high-performing genetic operators. Each individual contributes its own genetic information to the creation of new offspring adapted to the environment with greater fitness and with a higher chance of survival. In the application of GEP to function finding, the goal is to find offspring that are within a certain error of the correct value of the function. Therefore, GEP uses an evolutionary strategy to find the best fittest offspring without halting the evolution of the next generations. In these cases, the parent is usually unfit, but its modified descendants progressively approach a perfect solution. If the error used is the root relative squared error (RRSE), then the fitness function of a program f_i is:

$$f_i = 1000.\frac{1}{1+\mathrm{RRSE}_i} \tag{5}$$

The fitness function ranges from 0 to 1000, with 1000 corresponding to a perfect fit. The root relative square error RRSE_i of an individual program i (i-th offspring) is defined by the following equation:

$$\mathrm{RRSE}_i = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{(ij)}-T_j\right)^2}{\sum_{j=1}^{n}\left(T_j-\bar{T}\right)^2}} \tag{6}$$

where $P_{(ij)}$ is the value predicted by the program i for fitness case j, T_j is the target value for fitness case j, $\bar{T}=\left(\sum_{j=1}^{n}T_j\right)/n$, and n is the number of samples. For a perfect fit, $\mathrm{RRSE}_i=0$ thus, the RRSE_i ranges from 0 to infinity, with zero corresponding to the ideal.

Ferreira (2001) described seven mutations: point mutation, two transpositions, two-point and one-point recombination, transposition, and gene recombination. The most efficient operator in GEP is mutation, which causes populations of individuals to adapt efficiently, thus allowing for the evolution of strong solutions to all problems. Ferreira (2001) recommended using a mutation rate equivalent to two one-point mutations per chromosome. Mutation can occur anywhere in the chromosome and in the head; any symbol can change into another (function or terminal). For example, consider the multigenic chromosome described in Equation (4) as the first random individual (program) created by the GEP. Mutation can change the elements in positions 0 and 1 in gene 0 to $\times$ and $\div$ and the element in position 3 in gene 1 to *a* yielding the following offspring:

Gene 0:

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

Gene 1:

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

(7)

Moreover, GEP includes another process called transposition. GEP randomly selects a sequence in a gene and inserts it into any position in the head of the gene. Consider the chromosome in Equation (7), GEP might randomly choose to insert the sequence '*ab*' from gene 0 into position 3 of gene 1. A cut is made, and the block '*ab*' is copied into the insertion site, yielding the following chromosome:

Gene 0:

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

Gene 1:

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

(8)

The other genetic operations are less important and all depend on the value assigned for the mutation rate. For complete details on GEP and the related genetic operations, interested readers can refer to Ferreira (2001). GEP fitting for experimental data is performed using the commercial nonlinear data-mining software GenXProTools (www.gepsoft.com). The model is run with 20,000 generations, 30,40,50 chromosomes, 3,4 head size, 0.005–0.05 mutation rates and three random numerical constants.

Analysis procedure for GEP model development

The following procedure was used to develop the GEP prediction models:

1. An initial set of control variables (dam failure cases) is chosen as terminals for GEP.
2. The initial work environment is set for GEP by defining the chromosome architecture (number of genes, head size, functions) and mutation rates.
3. GEP randomly formulates the chromosomes of the parent program and implements genetic operators to yield many first-generation offspring.
4. GEP uses the fitness criteria specified in Equation (2) to find the fittest offspring. This offspring represents the solution to the problem in the first generation.
5. GEP considers the selected offspring the new parent and implements genetic operators to yield many second-generation offspring.
6. GEP evolution continues as per steps 3, 4 and 5 until the specified program fitness is met. While the GEP model fitness indicator f_i has no specific range, Ferreira (2001) suggested that 600–800 would be a good range indicating models that yield good predictions. The final GEP model (the fittest offspring of generation i) is scored on a set of performance indicators. These indicators are the square of the Pearson product moment correlation coefficient (R^2), the model fitness (fi), and the relative square error (RSE). A very good model has high R^2 and fi values, and low RSE values, as suggested by Smith *et al.* (1986). The indicators R and RSE are calculated by the following equations, respectively:

$$R_i = \frac{\frac{1}{n}\sum_{j=1}^{n}(T_j - \bar{T})(P_{(ij)} - \bar{P})}{\sqrt{\sum_{j=1}^{n}(T_j - \bar{T})^2/n}\sqrt{\sum_{j=1}^{n}(P_{(ij)} - \bar{P})^2/n}} \tag{9}$$

$$\mathrm{RSE}_i = \frac{\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}{\sum_{j=1}^{n}(T_j - \bar{T})^2} \tag{10}$$

where $\bar{P} = \left(\sum_{j=1}^{n} P_j\right)/n$

7. Steps 1 to 7 are repeated with a different set of control variables to produce another GEP model.

Developed GEP models

In this section, GEP was utilized to develop prediction equations for key breaching parameters essential to construct the breach outflow hydrograph. The complete dam failure dataset is used in model development with the number of cases varying according to the predictor combination attempted. The dam failure dataset has been partitioned into two datasets, one for training (75% of dataset) the GEP-based models and the other for testing the developed GEP-based models (25% of dataset). Both training and testing datasets have similar statistical parameters such as mean and standard deviation.

Key breach parameters Q_p, B_{avg}, t_f, and h_b have been expressed by past investigators in terms of only reservoir parameters and sometimes breach parameters (Wahl 2004). In current work, other important parameters such as: failure mode D_f, dam type D_t, and dam erodibility D_e are included as additional predictors. For generality, key breach parameters are expressed in a non-dimensional form Y_i such that $Q_p^* = Q_p/\sqrt{gV_w^{5/3}}$, $B_{\mathrm{avg}}^* = B_{\mathrm{avg}}/h_b$, $t_f^* = t_f/t_r$ and $h_b^* = h_b/H_d$. For mathematical rigor, reservoir parameters can be represented in non-dimensional form as: the reservoir shape coefficient, $\mathrm{RS} = V_w^{1/3}/H_w$, and the relative dam height, $H_d^* = H_d/H_r$. Hr is set to 15 m and tr is set to 1 hour (Xu & Zhang 2009). Numerical values have been assigned for the discrete variables De, Dt, and Df. For Dt, 4 is given for HD, 3 for DC, 2 for ZD and 1 for FD. For De, 3 is assigned for HE, 2 for ME and 1 for LE. For Df, 1.1 denotes piping failure, and 1.2 overtopping failure. These chosen values for discrete variables have mainly been chosen for

simplicity. However, choosing any other value before developing the GEP model is possible and has been found to have no effect on the physical behavior of the developed equations, despite having a different arrangement of parameters. In the current study, the non-dimensional key breaching parameters *Yi* are considered as a function of:

$$Y_i = f(X_i) = f\left(D_t, D_e, D_f, RS, H_d^*\right) \tag{11}$$

Table 3 shows the sensitivity analysis of various combinations of control variables *Xi* on $Y_i = Q_p^*, B_{\text{avg}}^*$ and t_f^*. Regardless of the chosen combination of variables, a significant improvement over the multivariable-least-squares-regression (MLSR) in data fitting is shown for all developed GEP models, both MLSR and GEP equations have been developed and tested using the same number of failure cases. This is very obvious for Q_p^* and B_{avg}^*, where R^2 of MLSR was 0.039 and 0.11, versus 0.892 and 0.750, for GEP1 and GEP6 models, respectively. This big difference in the R^2 is a major strength and advantage of GEP over traditional regression analyses. GEP models are capable of mapping nonlinear data behaviors in simple equations with the help of individual encoded chromosomes as linear strings of fixed length, which are afterwards expressed as nonlinear entities of different sizes and shapes (Ferreira 2001). Thus, GEP models are more flexible and efficient for fitting data of various relationships. While different cases are used to develop various GEP models, the prediction errors and model fitness shown in Table 3 are used as fair comparison basis as suggested by Xu & Zhang (2009).

For Q_p^*, GEP1 and GEP5 had the highest fitting with $R^2 > 0.87$ for training data and $R^2 > 0.79$ for testing data. This R^2 value for the developed GEP models is better than for the regression model where $R^2 = 0.77$, reported by Xu & Zhang (2009), who used multiparameter nonlinear regression (MNR) based on about 70% of the current dam failure dataset. Moreover, GEP models had the highest values of fitness ranging from 729 to 751 and lowest RSE values of 0.109. The GEP1 model contains the dam height, which has been reported as an important predictor for

Table 3 | Sensitivity analysis for the effect of various combinations of control variables X_i on $Y_i = Q^*_p$, B^*_{avg} and t_f^*

				MLSR	GEP					
		# of case studies		R^2	R^2		RSE		Fitness	
Model	Control variables $Y_i = f(X_i)$	Training	Testing		Train	Test	Train	Test	Train	Test
GEP-1	$Q_{\text{p}}^* = f(\text{RS}, H_{\text{d}}^*)$	43	8	0.039	0.892	0.791	0.109	0.302	751	645
GEP-2	$Q_{\text{p}}^* = f(\text{RS}, D_{\text{e}})$	32	6	0.372	0.674	0.621	0.331	0.451	634	575
GEP-3	$Q_{\text{p}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{t}})$	26	8	0.451	0.739	0.613	0.261	0.392	661	520
GEP-4	$Q_{\text{p}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{f}})$	26	8	0.456	0.794	0.699	0.209	0.491	685	538
GEP-5	$Q_{\text{p}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{f}}, D_{\text{t}})$	27	8	0.463	0.867	0.830	0.137	0.484	729	589
GEP-6	$B_{\text{avg}}^* = f(\text{RS}, h_{\text{b}}^*)$	51	12	0.112	0.750	0.730	0.220	0.412	640	480
GEP-7	$B_{\text{avg}}^* = f(\text{RS}, D_{\text{e}})$	40	13	0.717	0.698	0.650	0.308	0.621	642	515
GEP-8	$B_{\text{avg}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{t}})$	36	12	0.640	0.721	0.690	0.278	0.596	654	485
GEP-9	$B_{\text{avg}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{f}})$	36	12	0.649	0.848	0.850	0.155	0.447	716	600
GEP-10	$B_{\text{avg}}^* = f(\text{RS}, h_{\text{b}}^*, D_{\text{e}}, D_{\text{f}}, D_{\text{t}})$	36	12	0.675	0.856	0.806	0.147	0.468	722	593
GEP-11	$t_{\text{f}}^* = f(\text{RS}, h_{\text{b}}^*)$	27	9	0.040	0.577	0.550	0.426	0.695	604	480
GEP-12	$t_{\text{f}}^* = f(\text{RS}, D_{\text{e}})$	21	7	0.410	0.699	0.710	0.301	0.360	645	520
GEP-13	$t_{\text{f}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{t}})$	21	7	0.396	0.602	0.588	0.398	0.521	612	560
GEP-14	$t_{\text{f}}^* = f(\text{RS}, D_{\text{e}}, D_{\text{f}})$	23	7	0.398	0.636	0.621	0.364	0.425	623	583
GEP-15	$t_{\text{f}}^* = f(\text{RS}, h_{\text{b}}^*, D_{\text{e}}, D_{\text{f}}, D_{\text{t}})$	23	7	0.417	0.838	0.808	0.162	0.183	713	680

peak flow and is written as:

$$Q_p^* = \frac{c_1}{\mathrm{RS}} c_2^{H_d^*} + \frac{c_3}{\mathrm{RSH}_d^*} + \frac{c_4}{\mathrm{RS}} c_5^{H_d^*} \quad (12)$$

where $c_1 = -0.15$, $c_2 = 0.13$, $c_3 = 0.069$, $c_4 = -0.16$ and $c_5 = 0.12$. While GEP5 can be written as:

$$Q_p^* = \frac{c_1}{\mathrm{RS}^2} (D_t D_e + D_f RS)(D_f + c_2)(D_f + c_3) \quad (13)$$

where $c_1 = 0.083$, $c_2 = -0.87$ and $c_3 = -0.80$. Similarly, two GEP models (GEP9 and GEP10) for prediction of B_{avg}^* are found to have the highest R^2 of 0.85 for training and 0.80 for testing. This is higher than the $R^2 = 0.67$ reported by Xu & Zhang (2009) using MNR. GEP model fitness had an average value of 720 for training data and 600 for testing data. GEP9 can be written as:

$$B_{\mathrm{avg}}^* = c_1 D_e D_f^4 RS^{c_2} \quad (14)$$

where $c_1 = 0.096$ and $c_2 = 0.85$. On the other hand, GEP10 included important parameters shown in previous studies for prediction of average breach width, and can be written as:

$$B_{\mathrm{avg}}^* = (D_f - 1)\left(\frac{RS}{c_1} + \frac{D_f}{h_b^*}\right)(D_e - D_f + c_2) \quad (15)$$

where $c_1 = 1.91$ and $c_2 = 1.72$. However, combinations of parameters for prediction of failure time only showed that one model, GEP15 had the highest R^2 of 0.84 for training data and 0.80 for testing data, for other models R^2 was less than 0.80 and thus none were selected. GEP model fitness is 713 for training data and 680 for testing data. GEP15 can be written as:

$$t_f^* = \sin\left(D_e^{c_1 \mathrm{RS}}\right) + \mathrm{RS}^{1/2} - D_e^{c_2} + \sin^2\left(D_e^{\mathrm{RS}}\right) \quad (16)$$

where $c_1 = 0.55$ and $c_2 = 0.48$.

Dam erodibility D_e and dam failure mode D_f appeared as important predictors in all chosen GEP models except for GEP1. Similarly, the reservoir shape factor $RS = V_w^{1/3}/H_w$, appeared in all developed GEP models. The appearance of RS, D_e, and D_f in all selected models implies that these variables are good predictors for the breach outflow hydrograph and more influential than other control variables. Other control variables appeared in selected GEP models but with less influence; e.g., dam height H_d^* and breach height h_b^*. These control variables can be deemed as fair predictors for the breach outflow hydrograph. While D_t appeared only in one GEP model with little influence, suggesting that D_t is a poor predictor for breach outflow hydrograph parameters as agreed by Wahl (2004). For selected GEP models, RS, D_e, and D_f are considered good predictors and highly influential variables for peak flow rate, average breach width and failure time prediction. Except for the dam erodibility factor that has not been included in the majority of available prediction equations, the importance and influence of these parameters is similar in existing equations for breach parameters prediction (collected and summarized in Wahl 2004). In all available equations for dam breach parameter predictions, H_w appeared as an influential parameter in 50% of available peak flow prediction equations, H_w in 33% of available failure time prediction equations, and V_w, H_w in about 65% of available average breach width prediction equations.

External validation for developed GEP models

Tropsha *et al.* (2003) recommended external validation criteria for checking models based on their performance in testing data subsets. At least, one of the gradients of the regression line through the origin for predicted versus observed values, $k = \sum_{i=1}^{n}(T_i \times P_i)/P_i^2$, or observed versus predicted values, $k' = \sum_{i=1}^{n}(T_i \times P_i)/T_i^2$, should be close to 1. Also the coefficient of determination for the regression line through the origins, $m' = (R^2 - R_0^2)/R^2$ and $n' = (R^2 - R_0^2)/R^2$, should be lower than 0.1. Moreover, the condition of cross validation should satisfy (Gandomi *et al.* 2011):

$$R_m = R^2 \times \left(1 - \sqrt{|R^2 - R_O^2|}\right) > 0.5 \quad (17)$$

where $R_0^2 = 1 - \sum_{i=1}^{n} P_i^2(1-k)^2 / \sum_{i=1}^{n}(P_i - \bar{P})^2$ is the squared correlation coefficient through the origin between predicted and observed values and $R_0'^2 = 1 - \sum_{i=1}^{n} T_i^2(1-k')^2 / \sum_{i=1}^{n}(T_i - \bar{T})^2$ is the squared correlation coefficient through the origin between observed

and predicted values. The validation criteria and relevant performance of the developed GEP models are shown in Table 4. Models are considered valid for prediction if they satisfy some or all of the required conditions. As observed, all GEP models satisfied all related validation criteria, thus they have a good prediction power and are not chance correlations.

PARAMETRIC ANALYSIS

In this section, a parametric analysis is performed to find the influence of breach parameters, dam erodibility, and failure mode on prediction behavior for developed GEP models. This is performed for further verifications of the developed GEP models, since the examination of how well predicted values agree with the physical behavior of the problem studied, determines the robustness of the equations (Kuo *et al*. 2009). Figures 5 and 6 present the predicted values for Q_p, t_f and B_{avg} obtained by the selected GEP models as a function of H_w, V_w, D_e, and D_f.

Results of analyses in Figure 5 show that the predictions by selected GEP models are in agreement with those of Froehlich (1995a, b), which are most frequently used in the literature. Peak flow rate and average breach width are shown to depend largely on the reservoir storage volume and water depth above the breach bottom. Peak flow increases with the increase in H_w and V_w with a constant or slightly increasing rate. Chinnarasri *et al*. (2004) explained that the increase in H_w and V_w results in a high flow velocity through the breach, wide breach opening and large outflow through the breach. In contrast, when H_w and V_w decrease, the water depth above the breach is also low resulting in low-flow velocity, and a small outflow through the breach. It is noted that predictions of average breach width versus H_w by GEP models did not agree with the Von Thun & Gillette (1990) regression relation but agreed well with the relation recently developed by Xu & Zhang (2009). The developed GEP models had the highest strength in representing measured average breach width with R^2 of 0.85, compared with 0.65 for Xu & Zhang (2009), and 0.49 for the Von Thun & Gillette (1990) model using same number of cases. Wahl (2004) reported that the Von Thun & Gillette (1990) relation can produce acceptable results when reservoir storage is relatively small, which is not the case for half of the utilized database in this study. Moreover, recent developed relations for prediction of average breach width suggested that H_w might not be as good a predictor as V_w (e.g., Froehlich (2008) and Xu & Zhang (2009)). On the other hand, GEP predictions for failure time deviated from Von Thun & Gillette (1990) relations. This might be due to the fact that no available equation can offer a good prediction for time of failure (Attallah 2002) and all available failure time prediction equations have the highest uncertainty band amongst equations for prediction of other breach parameters and highest prediction interval reaching to 40 orders of magnitude (Wahl 2004) for the Von Thun & Gillette (1990) equation.

Another distinct feature in Figure 5 is that the increase in embankment material erodibility leads to wider breaches, larger peak flow rates, and shorter failure. Morris *et al*. (2009) stated that the embankment erodibility significantly affects the rate of erosion and hence breach formation. Breach formation through highly erodible embankment material tends to develop through the progressive surface erosion of material, whilst breach formation through less erodible embankment material tends to result in the formation of headcuts (steps). In the case of surface erosion, a gully is created with a steepening slope cutting back at a set angle towards the crest elevation of the upstream edge of the embankment crest, causing a rapid erosion and breach widening allowing greater flow through the breach and speeding up the failure time (Visser 1998; Chinnarasri *et al*. 2004; Temple *et al*. 2005).

Figure 6 shows the impact of dam failure mode, D_f on the average breach width and peak breach flow rate based

Table 4 | External validation statistical measures for developed GEP models (calculated using testing dataset)

Model	R ($R > 0.8$)	K ($0.85 < K < 1.15$)	K' ($0.85 < K' < 1.15$)	m' ($m' < 0.1$)	n' ($n' < 0.1$)	R_m ($R_m > 0.5$)
GEP-1	0.89	1.10	0.84	−0.16	−0.23	0.51
GEP-5	0.91	1.12	0.86	−0.03	−0.01	0.80
GEP-9	0.91	1.13	0.84	−0.09	−0.07	0.60
GEP-10	0.92	1.07	0.82	−0.25	−0.13	0.43
GEP-15	0.95	1.14	0.81	−0.04	−0.03	0.72

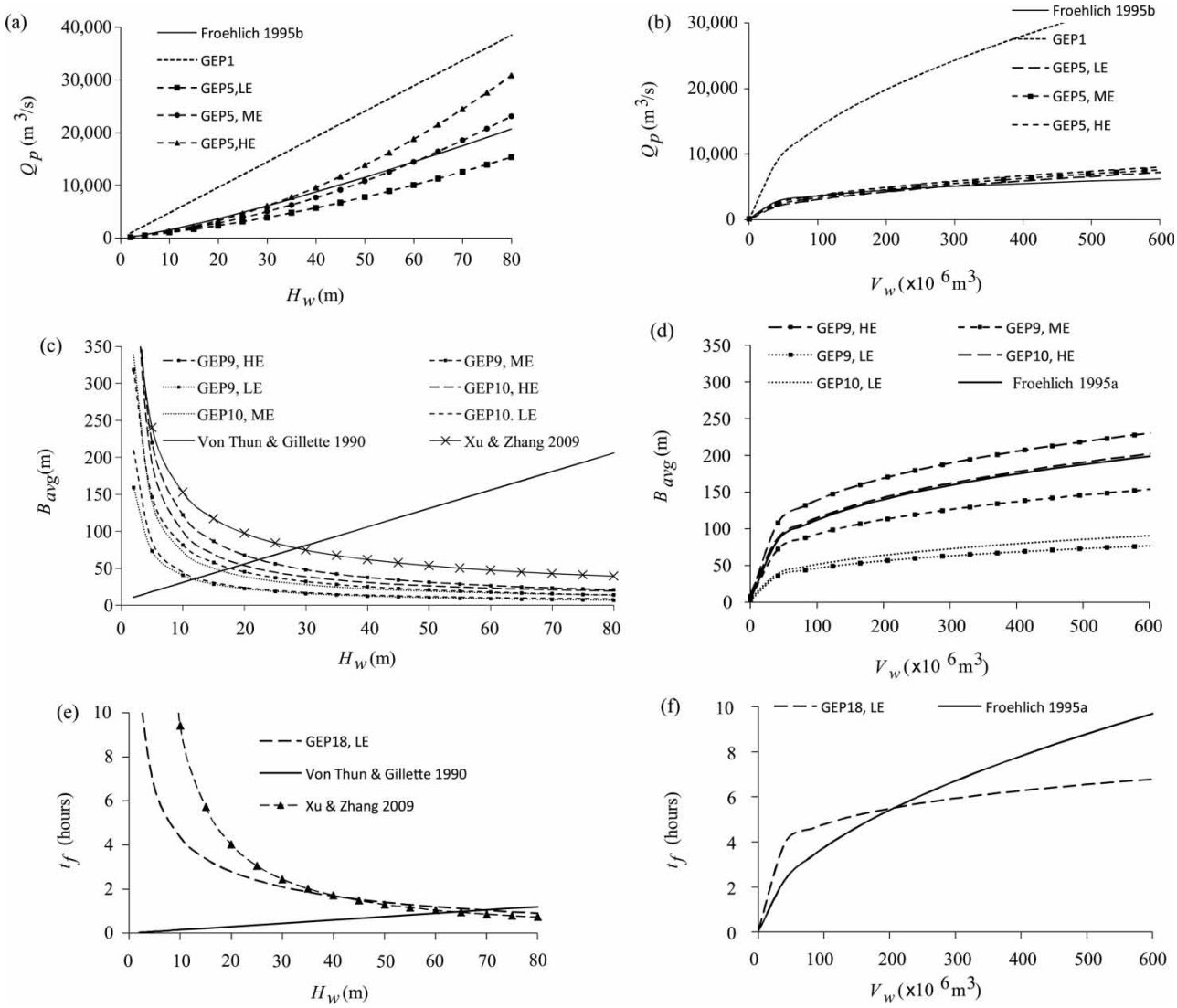

Figure 5 | Parametric analyses for Q_p, t_f and B_{avg} in selected GEP models; where $H_d = 20$ m, $h_b = 17$ m, $D_f = 1.1$, and $D_t = 4$, 2, 1 for Q_p, t_f and B_{avg} respectively ($H_w = 14$ m for charts b, d, f and $V_w = 23.09 \times 10^6$ m^3 for charts a, c, e).

on GEP models 5, 9, and 10. Results show that overtopping failure leads to wider average breach widths and higher breach peak flow rates than piping/seepage failure. Xu & Zhang (2009) reported the same result for their developed regression-based models including a factor for dam failure mode. Dewey & Gillette (1993) attributed the increase in peak flow rate in overtopping failure to the fact that overtopping flow is accompanied by continued inflow floods and overtopping initiates breach over a length of the dam rather than at a localized point as in the case of piping/seepage failures. In addition, in a seepage/piping failure, some water is released via a formed pipe through the dam, and therefore has no contribution to the later overtopping erosion after the collapse of the pipe (Xu & Zhang 2009). Froehlich (1995a) suggested that this is due to the higher rate of erosion induced by the overtopping flow that causes rapid extension of the breach and provided a regression-based equation for average breach width including a factor for dam failure mode. Moreover, Figure 7 presents Q_p^* and B_{avg}^* plotted against reservoir shape factor, *RS*. It is observed that with the increase in reservoir shape factor (large reservoir)

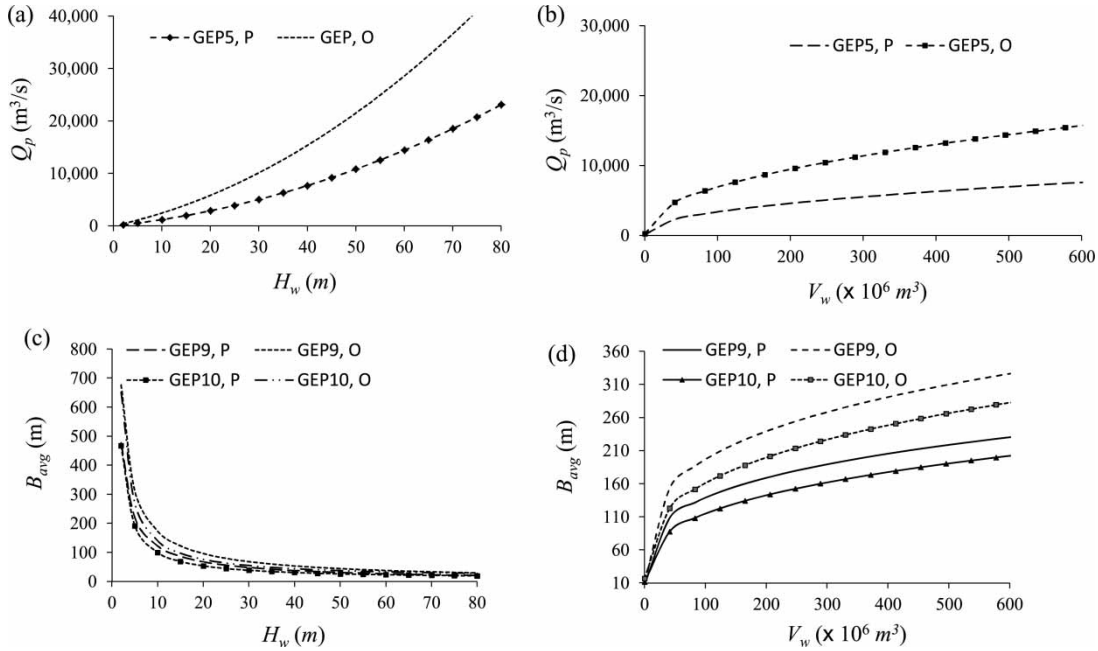

Figure 6 | Parametric analyses for Q_p and B_{avg} in selected GEP models; where $H_d = 20$ m, $h_b = 17$ m, $D_f = 1.1$ for piping and 1.2 for overtopping, $D_e = 3$ and $D_t = 4$ for Q_p ($H_w = 14$ m for charts b, d and $V_w = 23.09 \times 10^6$ m^3 for charts a, c).

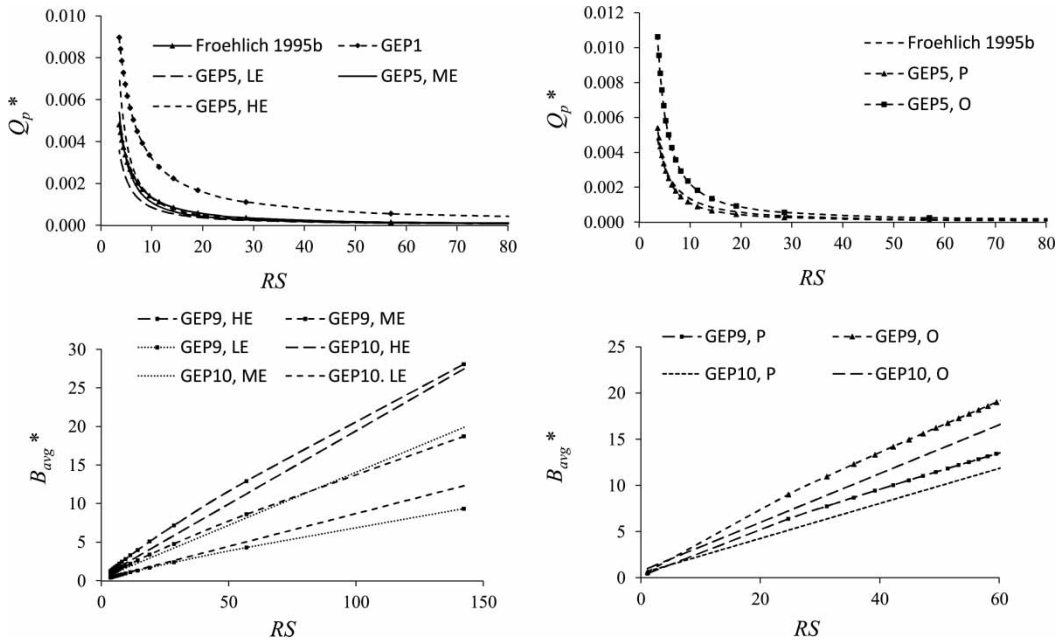

Figure 7 | Parametric analyses for Q_p^* and B_{avg}^* in selected GEP models; where $H_w = 14$ m, $V_w = 23.09 \times 10^6$ m^3, $H_d = 20$ m, $D_f = 1.1$ for piping and 1.2 for overtopping, and $D_t = 4$ and 1 for Q_p and B_{avg} respectively.

B_{avg}^* increases and Q_p^* decreases. This has been reported in the literature by Wahl (2010) and Chinnarasri *et al.* (2004). This is due to the fact that small reservoirs (with lower *RS*) drain significantly before the breach is fully formed, allowing the peak outflow to occur during breach formation leading to a lower breach opening and higher peak outflow. On the other hand, large reservoirs (with higher *RS*) maintain head until the breach reaches its

maximum size leading to wider breach openings and lower peak flow which occurs only after breach full formation.

ERROR ANALYSIS

The error analysis is concerned with the variation in the output of a model about a mean. It has been established by Corder (1967) as an approach influenced by behaviorism through which applied linguists sought to use the formal distinctions between the learners' first and second languages to predict errors. Error analysis provides indication for the error in model output and gives a general idea of how far from the reported value the true error free value might be. This is performed through calculating mean prediction error and width of certain confidence interval, e.g., a 95% interval. Wahl (2004) provided quantitative assessments for the errors in breach parameter prediction using many of the available prediction relations on 108 documented dam failure case studies. Due to the presence of a wide range of dam sizes in the used datasets, he expressed the errors in breach parameter prediction as a number of log cycles and used the objective outlier exclusion algorithm of Rousseeuw (1998) to exclude truly anomalous data while retaining the data characteristic variability. In this part of the study, the Wahl (2004) approach is adopted to present a clear idea about the prediction error found in GEP-based models and the confidence interval of their prediction in addition to comparing error analysis results with best available prediction equations chosen in the previous section. The analysis is applied to the dataset of 140 dam failure cases used in this study, which includes the data subsets used to derive all previous breach prediction equations. While this could provide some advantages for the GEP models, it provides fair indication for comparison of prediction capability for various equations (Wahl 2004; Xu & Zhang 2009). The outlier-exclusion algorithm is applied to prediction errors, which are defined in log cycles as:

$$e_{ij} = log_{10}(P_{ij}) - log_{10}(T_j) \tag{18}$$

Remaining error data are used to calculate main indicators defined as: mean prediction error ($\bar{e} = \sum_{i=1}^{n} e_{ij}$), width of error band $B_{ub} = \pm 1.96 S_e$ and the 95% confidence band around the predicted value:

$$\left\{P_{ij} \times 10^{-\bar{e}+1.96S_e},\ P_{ij} \times 10^{-\bar{e}+1.96S_e}\right\} \tag{19}$$

where S_e is the standard deviation of prediction errors and P_{ij} is taken as unity.

Table 5 summarizes the results of the error analysis performed on GEP models for prediction of peak outflow rate,

Table 5 | Error analysis estimates for developed GEP models based on log cycles (calculated using all dataset)

		# of case studies				
Parameter	**GEP model**	**Before excluding outliers**	**After excluding outliers**	**Mean prediction error**	**Width of error band**	**95% prediction interval around hypothetical predicted value of 1.0**
Q_p^*	GEP-1	51	50	+0.145	±0.62	0.17–2.97
	GEP-5	35	33	−0.0002	±0.40	0.40–2.53
	Froehlich (1995b)	57	55	−0.071	±0.62	0.28–4.88
	SCS (1981)	57	50	+0.111	±0.67	0.17–3.62
	Costa (1985)	51	50	+0.030	±0.82	0.14–6.18
B_{avg}^*	GEP-9	48	46	−0.029	±0.27	0.57–2.01
	GEP-10	48	41	−0.024	±0.24	0.61–1.84
	Froehlich (1995a)	90	88	−0.004	±0.39	0.41–2.49
	Von Thun & Gillette (1990)	59	56	±0.094	±0.40	0.32–2.03
t_f^*	GEP-15	30	28	+0.107	±0.42	0.30–2.02
	Froehlich (1995a)	41	40	−0.299	±0.75	0.35–11.15
	US Bureau of Reclamation (1988)	37	33	−0.385	±0.87	0.33–17.97

failure time and average breach width, respectively. In general, all developed GEP models outperformed the available selected equations from the literature in all parameters of the error analysis. For Q_p^* GEP models, mean prediction errors of −0.0002 to +0.145 were calculated. Width of error band of ±0.40 to ±0.62 orders of magnitude are calculated, which are less than prediction errors returned by Froehlich (1995b), Costa (1988), and Soil Conservation Services (1981) equations. Moreover, the same GEP models had a 95% prediction interval of about 0.17 to 2.97, which is less than that of Froehlich (1995b) and Costa (1988) intervals that had an upper band of 4.88 and 6.18 respectively. Similar results are calculated for both B_{avg}^* and t_p^* GEP prediction models.

UNCERTAINTY ANALYSIS

In dam failure risk assessment study, the uncertainties of influencing parameters have to be included since they could dramatically influence the outcome (Wahl 2004; Froehlich 2008). The sacristy of real data on dam failure used to construct prediction models imposes uncertainty on model predictions. Despite using in this study a larger data set than previous studies, there is still a certain degree of uncertainty of prediction outcomes, which can usually be reduced by collecting more data about the input parameters. Therefore, this section presents a quantitative assessment for the stochastic character of the developed GEP models and uncertainty in their predictions using the Monte Carlo Simulation (MCS) (Vose 1996; Frey & Li 2001). The uncertainty in key breach parameters Q_p^*, B_{avg}^*, t_f^* described by Equations (12)–(16) is considered due to the uncertainty in the model input parameters, RS, H_d^*, h_b^*, D_e, D_f, and D_t. The model input parameters are uncorrelated and can be considered independent variables where random sampling can be applied without having possible parameter combinations and without affecting the tails of the output distributions (Burmaster & Anderson 1994). Distribution fitting (Verbeeck *et al.* 2006) has been used to determine the probability density function (PDF) of various input parameters utilizing the complete dam failure data set shown in Table 1. Random variables that cannot appear in reality have been excluded by using truncated distributions with minimum and maximum boundaries determined from the available data set. Optimal distributions with highest scores (Vose 1996) in fitting tests are found to be Gen. Extreme Value for *RS* and H_d^*, and Weibull for h_b^*. For discrete variables D_e, D_f and D_t. the Poisson distribution was used to model their variations. The number of Monte Carlo trials needed to achieve a particular level of reliability has been given in Froehlich (2008) as $N = \left(\varphi_{\alpha/2}^2/4\varepsilon^2\right)^m$, where $\varphi_{\alpha/2}$ = standard normal deviate corresponding to a two sided confidence level and equals 1.645 for 90% confidence, m = number of random variables and ε = maximum allowable system error taken as 12% by Froehlich (2008). For MCS analyses, 250,000 runs are chosen such that the allowable system error is 3.7% for GEP models with two variables and 17% for other models with four variables. This number of MCS runs produced a converged variance for GEP models output (Verbeeck *et al.* 2006). For each MCS run, the deterministic GEP model is used to obtain a single outcome. Thus, 250,000 outcomes are calculated for each of the non-dimensional breaching parameters Q_p^*, B_{avg}^*, t_f^* and the uncertainty of these parameters is calculated as: $100 \times \text{MAD}/\text{Median}(P)$, where Mean Absolute Deviation, $\text{MAD} = 1/N \sum_{i=1}^{N} |P_i - \text{Median}(P)|$ (Walker 1931). Results of uncertainty analysis for key breach parameters using the developed GEP models are presented in Table 6. The analysis for uncertainty for Q_p^* resulted in a MAD of 0.00066 and 0.00046, which are 22 and 36% of the median values, for GEP1 and GEP5 respectively. For B_{avg}^*, the MAD of 0.7 was calculated, which is around 35% of the median values for GEP9 and GEP10. While the MAD of t_f^* was 0.441 and 19% of the median for GEP15. Uncertainties in the order of magnitude of 25–35% have been reported in the literature to be acceptable ranges for reliable models (Verbeeck *et al.* 2006).

Once the output uncertainty is determined, the least square linearization technique is used to split the output uncertainty into its sources (complete details can be found in Verbeeck *et al.* 2006). This method is a multiple regression between the parameter deviation from the mean and the output and its main equation is written as:

$$y = w_1 \Delta v_1 + w_2 \Delta v_2 + \ldots + w_n \Delta v_n + b \tag{20}$$

Table 6 | Uncertainty analysis for GEP models outputs from MCS (results are based on 250,000 MCS)

Parameter	Model	Median	MAD	Uncertainty %	Percentile	
					50th	95th
Q_p^*	GEP-1	0.0029	0.00066	22.53	0.00054	0.0055
	GEP-5	0.0013	0.00046	36.10	0.000415	0.0030
B_{avg}^*	GEP-9	2.0310	0.7763	38.21	0.641	4.3347
	GEP-10	2.0080	0.6968	34.69	0.698	4.7631
t_f^*	GEP-15	2.3155	0.441	19.04	0.575	3.621

where y is the main key breach parameter Q_p^* or B_{avg}^* or t_f^*, Δv_i is the difference between v_i, the random chosen sample of parameter $\underline{i}$ and m_{Vi} is the mean value of parameter i of all the random samples. When m Monte Carlo simulations are carried out, Δv_i for each parameter and the model output y are calculated for each simulation. The regression coefficients w_i, are estimated by minimizing the sum of squared errors. Based on the regression coefficients and the variations of the parameter uncertainties, the sensitivity coefficient of each parameter $\underline{i}$ (S_{V_i}) can be approximated by:

$$S_{V_i} = 100 \times w_i^2 \sigma_{\delta_{V_i}}^2 / \sum_{i=1}^{n} w_i^2 \sigma_{\delta_{v_i}}^2 \tag{21}$$

where $\sigma_{\delta_{v_i}}^2$ is the variance of the calculated difference δ_{Vi}. Contribution of various dam and reservoir parameters in the overall uncertainty of the main key breach parameters are shown in Table 7. These results identify the most influential parameters on the non-dimensional key breach parameters as presented by developed GEP models. Similar to results in Figure 5, the dam erodibility is found to be an important predictor for all tested GEP models except for t_f^*. Its contributions ranged from 40 to 70%, which is more than all other predictors appearing in equations. This is followed by the reservoir shape factor RS, which appeared in all equations with influential contribution ranging from 20 to 50% and reached 99% in the case of t_f^*. Other parameters like dam type and dam failure mode appeared with little influence except for D_f in GEP10 which contributed to 40% of the total output uncertainty.

Table 7 | Contribution of uncertain dam and reservoir parameters to the overall uncertainty of the developed GEP models (results are based on 250,000 MCS)

Contribution of parameter uncertainty to output uncertainty

Parameter	Q_p^*		B_{avg}^*		t_f^*
	GEP-1	GEP-5	GEP-9	GEP-10	GEP-15
RS	43.88	38.19	20.25	17.67	99.75
H_{d}^*	56.11	0	0	0	0
h_{b}^*	0	0	0	0.778	0
D_{t}	0	6.58	0	0	0
D_{f}	0	6.20	12.59	43.40	0
D_{e}	0	49.02	67.15	38.14	0.25

CONCLUSIONS

GEP has been used on a large database of dam failure case studies to develop new empirical formulae with physical meaning to predict the main breach hydrograph parameters in non-dimensional form Q_p^*, B_{avg}^*, t_f^*. The new equations have the advantage of being developed from a large database where more than one-half of cases are for large dams and containing information on dam erodibility, failure mode and dam type. The available dataset has been divided into training and testing datasets. Predictive models produced by GEP are found to be superior over those previously produced by MNR and MLSR. New external validation methods have been applied to the developed GEP models based on testing datasets. Results of error analysis show the superiority of the developed GEP models over existing regression-based models. An uncertainty analysis has been performed using MCS to determine the uncertainties associated with the output from the developed GEP models as a result of the combined effect of the input parameters distributions. The uncertainty results showed an average 25% of the median value for Q_p^*, 35% for B_{avg}^*, and 19% for t_f^*. Using parametric and uncertainty analysis

results, the reservoir shape factor, the dam erodibility, and the dam failure mode are found to have large weights and influence on output predictions in all models.

REFERENCES

Attallah, T. A. 2002 A Review on Dams and Breach Parameter Estimation. M.Sc. Thesis, submitted to Virginia Polytechnic Institute and State University, Blacksburg, VA.

Azamathulla, H. 2011 Gene-expression programming to predict scour at a bridge abutment. *Journal of Hydroinformatics* **14** (2), 324–331.

Burmaster, D. E. & Anderson, P. D. 1994 Principles of good practice for the use of Monte Carlo techniques in human health and ecological risk assessments. *Risk Analysis* **14** (4), 477–481.

Chinnarasri, C., Jirakitlerd, S. & Wongwises, S. 2004 Embankment dam breach and its outflow characteristics. *Civil Engineering and Environmental Systems* **21** (4), 247–264.

Corder, S. P. 1967 The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching* **5** (1–4), 160–170.

Costa, J. E. 1985 Floods from dam failures. US Geological Survey, Open-File Rep. No. 85-560, USGS, Denver.

Costa, J. E. 1988 *Floods from Dam Failures. Flood Geomorphology*. John Wiley & Sons, New York.

Dewey, R. L. & Gillette, D. R. 1993 Prediction of embankment dam breaching for hazard assessment. *Proc., ASCE Conf. on Geotechnical Practice in Dam Rehabilitation*, Raleigh, NY, pp. 131–144.

Dyer, M., Utili, S. & Zielinski, M. 2007 Influence of the Desiccation Fine Fissuring on the Stability of Flood Embankments. Report UR11, Strathclyde University, www.floodrisk.org.

Fallah, E., Haddad, O. B. & Marino, M. A. 2012 Developing reservoir operational decision rule by genetic programming. *Journal of Hydroinformatics* **15** (1), 103–119.

Ferreira, C. 2001 Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems* **13** (2), 87–129.

Fread, D. L. 1988 *BREACH: An Erosion Model for Earthen Dam Failures.* National Weather Service, National Oceanic and Atmospheric Administration, Silver Spring, Maryland. Revised 1991.

Frey, H. C. & Li, S. 2001 Quantification of variability and uncertainty in stationary Natural gas-fuelled internal combustion engine NOx and total organic compounds emission factors. *Proc. Ann. Meeting Air and Waste Management Assoc.*, Abstract No.695, A&AWMA, Pittsburgh, PA.

Froehlich, D. C. 1995a Peak outflow from breached embankment dam. *Journal of Water Resources Planning and Management* **121** (1), 90–97.

Froehlich, D. C. 1995b Embankment dam breach parameters revisited. *Proc. 1995 ASCE Conf. on Water Resources Engineering*, New York, pp. 887–891.

Froehlich, D. C. 2008 Embankment dam breach parameters and their uncertainties. *Journal of Hydraulic Engineering* **134** (12), 1708–1721.

Gandomi, A. H., Alavi, A. H., Mirzahosseini, M. R. & Nejad, F. M. 2011 Nonlinear genetic-based models for prediction of flow number of asphalt mixtures. *Journal of Materials in Civil Engineering, ASCE* **23** (3), 248–263.

Ghani, A. A. & Azamathulla, H. 2011 Gene-expression programming for sediment transport in sewer pipe systems. *Journal of Pipeline Systems Engineering and Practice* **2** (3), 102–106.

Guo, W. D., Lai, J. S. & Lin, G. F. 2008 Finite-volume multi-stage schemes for shallow-water flow simulations. *International Journal for Numerical Methods in Fluids* **57** (2), 171–204.

Guven, A., Aytek, A., Yuce, M. I. & Aksoy, H. 2008 Genetic programming-based empirical model for daily reference evapotranspiration estimation. *Clean – Soil, Air, Water* **36** (10–11), 905–912.

Guven, A. & Gunal, M. 2008 Genetic programming approach for prediction of local scour downstream of hydraulic structures. *Journal of Irrigation and Drainage Engineering* **134** (2), 241–249.

Hagen, V. K. 1982 Re-evaluation of design floods and dam safety. *Proceedings, 14th International Congress on Large Dams*, Rio de Janeiro, Brazil, 1, pp. 475–491.

Henan Water Resources Authority 2005 *The August 1975 Catastrophic Flood Disaster in Henan*. Yellow River Water Conservancy, Zhenzhou, China.

Kuo, Y. L., Jaksa, M. B., Lyamin, A. V. & Kaggwa, W. S. 2009 ANN-based model for predicting the bearing capacity of strip footing on multi-layered cohesive soil. *Computers and Geotechnics* **36** (3), 503–516.

Lin, G. F., Lai, J. S. & Guo, W. D. 2003 Finite-volume component-wise TVD schemes for 2D shallow water equations. *Advances in Water Resources* **26** (8), 861–873.

MacDonald, T. C. & Langridge-Monopolis, J. 1984 Breaching characteristics of dam failures. *Journal of Hydraulic Engineering* **110** (5), 567–586.

Morris, M., Hassan, M., Buchholzer, Y. & Davies, T. 2008 HR BREACH: Developing a practical breach model to meet industry needs. *Proc., 28th Annual United States Society on Dams (USSD) Conf.*, Portland, Oregon, pp. 753–766.

Morris, M., Hassan, M., Kortenhaus, A. & Visser, P. 2009 Breaching Processes: A State of the Art Review. Report T06-06-03, FLOODsite project report, www.floodsite.net.

Nourani, V. N., Komasi, M. & Alami, M. T. 2012 Geomorphology-based genetic programming approach for rainfall-runoff modeling. *Journal of Hydroinformatics* **15** (2), 427–445.

Rousseeuw, P. J. 1998 Chapter 17: Robust estimation and identifying outliers. In: *Handbook of Statistical Methods for Engineers and Scientists*, 2nd edn (H. M. Wadsworth, ed.). McGraw-Hill, New York, pp. 17.1–17.15.

Singh, V. P. 1996 *Dam Breach Modeling Technology*. Kluwer Academic Publishers, The Netherlands.

Smith, S. J., Sharpley, A. N., Williams, J. R., Nicks, A. D. & Jones, O. R. 1986 Sediment-nutrient transport in agricultural runoff. *Proc. Fourth Federal Interagency Conf*. **2** (7), 11–20.

Soil Conservation Service (SCS) 1981 Simplified Dam-Breach Routing Procedure. Tech. Release No. 66 (Rev. 1).

Temple, D. M., Hanson, G. J., Nielsen, M. L. & Cook, K. R. 2005 Simplified breach analysis model for homogeneous embankments: Part I, Background and model components. Proc. 25th Annual USSD Conference, Salt Lake City, Utah, pp. 151–161.

Tropsha, A., Gramatica, P. & Gombar, V. K. 2003 The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **22** (1), 69–77.

US Bureau of Reclamation 1988 *Downstream Hazard Classification Guidelines*. ACER Technical Memorandum No. 11, Assistant Commissioner-Engineering and Research, Denver, Colorado.

Verbeeck, H., Samson, R., Verdonck, F. & Raoul, L. 2006 Parameter sensitivity and uncertainty of the forest carbon flux model FORUG: a Monte Carlo analysis. *Tree Physiology* **26** (6), 807–817.

Visser, P. J. 1998 *Breach growth in sand-dikes*. PhD Dissertation, submitted to Delft University of Technology, Delft, The Netherlands.

Von Thun, J. L. & Gillette, D. R. 1990 *Guidance on Breach Parameters*. Internal Memorandum, US Dept of the Interior, Bureau of Reclamation, Denver, CO.

Vose, D. 1996 *Quantitative Risk Analysis: A Guide to Monte Carlo Simulation Modeling*. John Wiley & Sons, New York.

Wahl, T. L. 1998 Prediction of Embankment Dam Breach Parameters – A Literature Review and Needs Assessment. Dam Safety Report No. DSO-98-004, US Dept of the Interior, Bureau of Reclamation, Dam Safety Office.

Wahl, T. L. 2004 Uncertainty of predictions of embankment dam breach parameters. *Journal of Hydraulic Engineering* **130** (5), 389–397.

Wahl, T. L. 2010 Dam breach modeling – an overview of analysis methods. In: Joint Federal Interagency Conference on Sedimentation and Hydrologic Modeling, Las Vegas Nevada, USA.

Wahl, T. L., Hanson, G. J., Courivaud, J. R., Morris, M. W., Kahawita, R., McClenathan, J. T. & Gee, D. M. 2008 Development of next-generation embankment dam breach models. Proc., 28th Annual United States Society on Dams (USSD) Conf., Denver, pp. 767–779.

Walker, H. 1931 *Studies in the History of the Statistical Method*. Williams & Wilkins Co., Baltimore, MD, pp. 24–25.

Wu, W. M. & Wang, S. S. Y. 2007 One-dimensional modeling of dam-break flow over movable beds. *Journal of Hydraulic Engineering* **33** (1), 48–58.

Wurbs, R. A. 1987 Dam-breach flood wave models. *Journal of Hydraulic Engineering* **113** (1), 29–46.

Xu, Y. & Zhang, L. M. 2009 Breaching parameters for earth and rockfill dams. *Journal of Geotechnical and Geoenvironmental Engineering* **135** (12), 1957–1970.

Zhumadian Water Resources Authority ZWRA 1997 *Log of the August 1975 Storm Event in Zhumadian*. ZWRA, Henan, China (in Chinese).

First received 29 June 2013; accepted in revised form 8 August 2013. Available online 9 October 2013