Examination of Multiple Predictive Approaches for Estimating Dam Breach Peak Discharges

G. E. Moglen, F.ASCE[1]; K. Hood[2]; and T. V. Hromadka II, M.ASCE[3]

Abstract: A database joining individual earthen dam breach failure studies is assembled and reanalyzed across all aggregate observations. Conventional regression methods are employed along with newer predictive approaches to estimating peak discharges resulting from an earthen dam failure. Goodness of fit is quantified through relative standard error and relative bias. These measures are computed and presented for previous predictive equations. Numerical optimization techniques are used to calibrate power law functions of one, two, and three predictors to estimate peak discharge from the aggregate database. Findings show that equations calibrated from the aggregate database have better goodness-of-fit metrics than those determined from their earlier, individual data sets. Improvement in relative standard error varies from essentially zero to as much as 50%. Two similar innovative techniques are applied to the aggregate database: region of influence (ROI) and k-nearest neighbor (kNN). Both of these approaches identify a subset of most similar observations from the database, given a specific test location. The ROI approach performs poorly in prediction mode, uniformly producing relative standard errors that are greater than the originally calibrated equations and that often exceed 100% of the standard deviation of the observations. Smaller relative standard errors are obtained as ROI size increases, contrary to the spirit of this approach. In contrast, the kNN approach performs well, with best results obtained for a simple numerical average of the k nearest observations. The size of the optimum k neighborhood varied from 3 to 29, with 12 being the median value among the cases examined. Regression equation calibration via logarithmic transformation is briefly explored, and the need to limit predictions to the test space within the convex hull of the observations is discussed. **DOI: 10.1061/(ASCE)HE.1943-5584.0001740.**

Introduction

Dam breach modeling is challenging. Efforts to predict flood magnitudes resulting from breaches are limited by available failure data and by the myriad causes of such failures. The result is that data sets are sparse and observed failures may vary considerably in their behavior because of differing failure mechanics. The challenge to engineers and scientists is to use statistical approaches to make effective estimates from such a limited set of data. The value of such approaches is to provide planning-level estimates of flood magnitudes so that flood inundation extent can be modeled and available to emergency managers.

Historically, regression has served as a favored tool used to generate statistically based estimates from observational data sets. This work employs multiple studies of earthen dam failure, aggregating their individual data sets and examining their predictive equations. We will examine goodness-of-fit measures and equation calibration methods for existing equations and new equations calibrated from the collective, assembled database. One alternative approach to estimating dam failure peak discharges is by physics-based, hydraulic-modeling efforts such as those by the National Weather Service, producing such models as DAMBRK (Fread 1988) and FLDWAV (Fread and Lewis 1988). This approach is well grounded with physically observable and measureable characteristics associated with the dam–channel system. However, compared to regression methods, this approach is much more data intensive, requiring significant user expertise, and may be vulnerable to numerical instabilities. At the other end of the spectrum, alternative statistical approaches have been applied to estimate dam failure peak discharges. A good example is the use of artificial neural networks (e.g., Pektas and Erdik 2014) to address this problem. This approach features flexibility in combining multiple functions to reproduce observations gleaned from a training subset of some overall data set. Performance is then validated based on estimates of observations from a sequestered testing subset. This approach can work well, but the opacity and lack of physical interpretation of the developed results is a weakness.

Traditional regression analysis is not without its limitations, which include assumed independence of predictor variables, homoscedasticity, and the absence of outliers. This study will explore newer approaches that manage or avoid such limitations to estimate flood peaks resulting from earthen dam failures. Specifically, the region of influence (ROI) approach (Burn 1990; Tasker et al. 1996) of regression equation calibration and some k-nearest neighbor (kNN) algorithms (Cover and Hart 1967; Altman 1992) will be explored as they apply to the problem of flood peak estimation from dam failures.

While methods and measures are essential to extending the science, practical application of the database to estimate flood peaks is an important goal. The objective here is to provide an example illustration of the application of the techniques examined in this paper, both traditional and innovative, for comparison so that accuracy and bias can be assessed.

[1]Supervisory Research Hydrologist, Hydrology and Remote Sensing Laboratory, Agricultural Research Service, Beltsville, MD 20910 (corresponding author). Email: glenn.moglen@ars.usda.gov

[2]Instructor, Dept. of Mathematical Sciences, US Military Academy, West Point, NY 10996.

[3]Professor of Mathematics, Dept. of Mathematical Sciences, US Military Academy, West Point, NY 10996.

Note. This manuscript was submitted on January 31, 2018; approved on August 16, 2018; published online on November 27, 2018. Discussion period open until April 27, 2019; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, © ASCE, ISSN 1084-0699.

Background

Linear regression performed via a least-squares technique is among the most commonly used approaches to estimatung a criterion variable, y, given one or more predictor variables, x_j

$$\hat{y} = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + \cdots \tag{1}$$

where $\hat{y}$ = prediction (best estimate) of y for a given set of predictors $\{x_1, x_2, \ldots\}$. The calibration coefficients, $\{c_0, c_1, c_2, \ldots\}$ are determined by minimizing the function F, which is simply the sum of the squared prediction errors across n observations:

$$\min(F) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2}$$

A prediction made using Eq. (1), or variations of this equation discussed subsequently, are most dependable when the specific values of each of the predictor variables are within the range of observations from which the equation was calibrated. Predictions made outside of this range amount to extrapolations from the training data set, and such predictions suffer from the much greater uncertainty associated with extrapolation. When only one predictor variable is used, the simple minimum and maximum of observations of this variable define the bounds over which the equation is most safely applied. When two predictor variables appear in Eq. (1), the graphed set of pairs of predictor variables can be enclosed by a minimum bounding envelope defined by a small subset of these predictors. The term *convex hull* is used to describe this polygonal envelope. The concept of the convex hull is a strictly mathematical construct; however, it has been applied in multiple contexts, including civil engineering applications (e.g., Laton et al. 2007).

Fig. 1 illustrates the concept of a convex hull in the context of the observations of V_w and H_w (defined and described in the next section). The hull shown here is graphed in log-log format because of the wide range of orders of magnitude of especially V_w. The determination of the hull boundaries is dependent on whether it is determined in the arithmetic or log space. The solid line hull shown in Fig. 1 was determined in the log space and is clearly a convex polygon (all angles less than 180°). The convex hull determined in the arithmetic space is generally different than its log-space counterpart, as shown by the dashed lines in Fig. 1. The arithmetic space–determined hull, in fact, results in a polygon that is not strictly convex at all vertices (e.g., see vertex "A") when graphed in the log space. The arithmetic space–determined hull covers a more restrictive area than the log space–determined hull, as will be discussed subsequently. Returning to the discussion of hull dimensionality, if three (or more) predictors are used in Eq. (1), the convex hull becomes three (or more) dimensional, but the same concept as for one or two dimensions still holds; a prediction made based on a set of observations that lies outside the hull is an extrapolation, and a prediction should be made with caution, if at all.

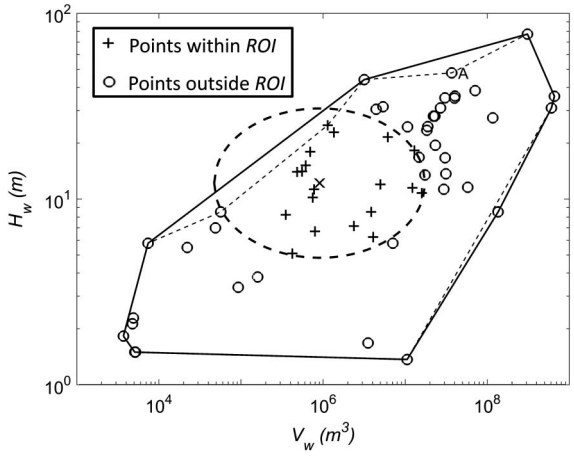

Fig. 1. Convex hull for peak discharge predictors V_w and H_w in log (solid line) and arithmetic (dashed line) spaces. Dashed ellipse shows region of influence for observations closest to to test point at $V_w = 910{,}000$ m^3, $H_w = 12.2$ m.

In addition to the strict convex hull bounds, the quality and dependability of a calibrated regression equation requires a sufficient number of available observations. There are various rules of thumb (e.g., Green 1991; Harrell 2001) that a minimum of 10 to 20 observations are needed per predictor variable when doing regression analyses. The data set assembled here varies from as few as 17 predictor observations to more than 100 observations. Predictive power can be expected from this data set, but it is also understood that this data set is sparse—requiring appropriate attention to the quality of any calibrated equations and demanding innovative alternative approaches to extract as much information as possible from the data set.

One innovative approach to the traditional linear regression just described is to draw selective subsets from the collected database. This approach is referred to as the region of influence (ROI) approach (Burn 1990; Tasker et al. 1996). The ROI approach is similar in spirit to other machine learning algorithms such as k-nearest neighbor approaches (e.g., Cover and Hart 1967; Altman 1992) and to the lag distance to the sill in geostatistical Kriging (Duricic et al. 2013). The common thread to ROI and other machine learning algorithms is the focus on selective sampling from the overall data set based on a similarity of selected observations to the test "location" where a prediction is to be made. The assumption is that existing observations that are most similar to the test location are the most valuable for prediction at the test location. Thus, this approach draws from the overall data set of observations only those points that are "closest" to the predictor(s) at the test location. The regression equation calibrated from this subset of closest points may be superior to an equation calibrated from the entire data set.

Similarity or closeness of an observation to the test location is determined by calculating the normalized "distance" (D_j) between the test location characteristics ($\tilde{P}$) and those at point, j, in the data set:

$$D_j = \left[\left(\frac{\log(\tilde{P}_1) - \log(x_{1,j})}{\sigma_{\log(x_1)}} \right)^2 + \left(\frac{\log(\tilde{P}_2) - \log(x_{2,j})}{\sigma_{\log(x_2)}} \right)^2 + \cdots + \left(\frac{\log(\tilde{P}_n) - \log(x_{n,j})}{\sigma_{\log(x_n)}} \right)^2 \right]^{1/2} \tag{3}$$

where $x_{i,j}$ = value of observed characteristic i at location j. The standard deviation of the logarithms of characteristic i for the entire observation data set is $\sigma_{\log(x_i)}$, which serves to normalize each of the individual differences between the test location characteristic and

the observation characteristic. The overall mathematical form of Eq. (3) is that of a Euclidean distance equation from geometry, so D_j is often referred to as the distance of observation j from the test location.

The ROI approach generates a tailored regression. The subset of observations with D_j determined in Eq. (3) that are less than some critical distance D_c are those observations that are to be used in the regression analysis. Although not strictly necessary, the same characteristics $x_1 \cdot x_n$ used to determine Euclidean distance are also generally used as predictors from which the regression equation is calibrated. The ROI approach does not produce a single static regression equation that applies in all cases; rather, it is a dynamic approach that produces a regression equation that is unique to the test location and dependent on the closeness, D_j, of predictor variable observations in the overall data set.

Fig. 1 illustrates the ROI concept. The point **x** is arbitrarily chosen here to represent a hypothetical test location's characteristics. The scatter of observations shown in this figure has been filtered so that the nearest 20 points to the test location are graphed using a "+" marker, while those further away are graphed using a "o" marker. The dashed ellipse shows the region within a critical distance, D_c, used to separate the scatter into these distinct groups.

The kNN approach encompasses a broad class of estimation techniques. The common thread in each kNN variant is the use of a distance calculation [such as defined by Eq. (3)]. This distance calculation helps to define a region or neighborhood. The kNN approach differs from ROI in that the estimation at the test location that follows does not generally encompass regression as discussed earlier in Eqs. (1) and (2). The estimation often takes the form of a simple or weighted average from among observations within the selected neighborhood. Two weighted average variants of the kNN approach are examined here.

Regardless of the approach employed to make a prediction, the quality of that prediction is quantified through two goodness-of-fit metrics: relative standard error and bias. The foregoing discussion was written in terms of a generic criterion variable, y. In the remainder of this work, formulas are presented in terms of the specific application of this paper, the prediction of peak discharges from dam failures, Q. The relative standard error, S_e/S_Q, is

$$\frac{S_e}{S_Q} = \frac{\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{Q} - Q_i)^2}} \tag{4}$$

where Q_i = observed peak discharge; $\hat{Q}_i$= predicted peak discharge; $\bar{Q}$ = mean of all observed peak discharges; n = number of observations; and p = number of predictors used in the predictive approach. The denominator of Eq. (4) is simply the standard deviation of the observed flood peaks, and the numerator is the standard error of the predicted flood peaks. The relative standard error is, therefore, simply the absolute standard error normalized by the standard deviation. Because of the squared terms in both the numerator and denominator of Eq. (4), the relative standard error must be positive. In general, the relative standard error is a value between 0 and 1, its magnitude reflecting the improvement in predictive power produced by the peak flow equation being examined compared to simply making a prediction based on the mean of the observations. The closer to 0 the standard error is, the stronger the performance of the predictive equation. The relative standard error can exceed 1, which would be indicative of poor performance by the equation being examined.

The relative bias, B, is

$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i - Q_i)}{\frac{1}{n}\sum_{i=1}^{n} Q_i} \tag{5}$$

Relative bias is a measure of the cumulative sum of all errors, normalized by the mean of the observations. In contrast to the relative standard error, individual terms of the summation in Eq. (5) can be positive or negative, so relative bias can be either positive or negative. A convenient conceptualization of relative bias comes from observing that two errors of equal and opposite signs cancel one another and collectively contribute a net zero to the mean bias. Relative bias is ideally zero and is, by definition, zero in the context of linear regression.

Data and Methods

Database

Efforts to quantify peak flood flows resulting from dam breaches extend well into the past. The earliest dam failure in the assembled database is from 1864 in Bradfield, UK, and the average year of failure is approximately 1950. The database assembled (appendix) contains observations of flood peaks from numerous sources resulting from earthen dam failures quantified in previous studies. Table 1 presents a summary of the database characteristics in terms

Table 1. Summary of dam breach database characteristics

Variable(s)	Description	Unit	Number of observations	Variable mean, median	Variable range	Standard deviation of Q for subset of database
Q	Failure peak discharge	m^3/s	120	3,850; 667	1.3 to 78,100	10,400
H_d	Height of dam	m	78	20.2, 14.3	1.5 to 93	12,600
H_w	Height above breach invert	m	103	14.2, 9.6	1.37 to 77.4	11,200
L	Length	m	19	431, 238	4.9 to 2000	18,700
S	Reservoir storage	m^3	70	4.12×10^7, 1.11×10^7	42,000 to 6.50×10^8	13,100
V_w	Volume stored above breach invert	m^3	105	2.62×10^7, 1.32×10^6	3,700 to 6.6×10^8	11,100
W	Average width	m	31	53.0, 40.5	9.63 to 250	11,600
V_w, H_w	See above for description and units		103	V_w: 2.74×10^7, 1.18×10^6 H_w: 14.3, 10.1	V_w: 3,700 to 6.60×10^8 H_w: 1.37 to 77.4	11,200
V_w, H_w, L	See above for description and units		19	V_w: 5.02×10^7, 5.39×10^6 H_w: 16.3, 13.5 L: 454, 238	V_w: 4,770 to 6.08×10^8 H_w: 1.5 to 47.9 L: 4.9 to 2000	19,600
V_w, H_w, W	See above for description and units		31	V_w: 4.23×10^7, 3.87×10^6 H_w: 15.6, 10.8 W: 51.1, 37.3	V_w: 13,300 to 6.60×10^8 H_w: 1.68 to 77.4 W: 9.63 to 250	11,600

Table 2. Goodness-of-fit characteristics for equations calibrated in previous studies and full data set regression equations calibrated in this study

Predictor(s)	Relevant studies: n	Equation	S_e/S_Q	Relative bias (m^3/s)
H_d	Singh and Snorrason (1982): 8	$Q = 13.4(H_d)^{1.89}$	0.994	0.145
	This study: 78	$Q = 154(H_d)^{1.20}$	0.885	−0.00018
H_w	Bureau of Reclamation (1982): 21	$Q = 19.13(H_w)^{1.85}$	0.762	0.0066
	This study: 103	$Q = 41.0(H_w)^{1.63}$	0.758	0.00092
L	Pierce et al. (2010): 14	$Q = 0.12(L)^{1.79}$	1.23	0.470
	This study: 19	$Q = 11.2(L)^{1.09}$	0.782	0.0011
S	Singh and Snorrason (1984): 8	$Q = 1.776(S)^{0.47}$	0.819	−0.281
	This study: 70	$Q = 0.345(S)^{0.588}$	0.738	−0.0001
V_w, H_w	MacDonald and Langridge-Monopolis I (1984): 23	$Q = 1.154(V_w \cdot H_w)^{0.412}$	0.767	−0.528
	MacDonald and Langridge-Monopolis II (1984): 23	$Q = 3.85(V_w \cdot H_w)^{0.411}$	0.753	1.26
	Froehlich (1995): 22	$Q = 0.607(V_w^{0.295} \cdot H_w^{1.24})$	0.691	−0.525
	This study: 103	$Q = 0.640(V_w^{0.392} H_w^{0.866})$	0.559	0.00021
V_w, H_w, L	Pierce et al. (2010): 14	$Q = 0.012(V_w^{0.493} H_w^{1.205} L^{0.226})$	0.187	0.0449
	This study: 17	$Q = 0.0338(V_w^{0.512} H_w^{1.08} L^{0.0761})$	0.0883	0.000012
V_w, H_w, W	Pierce et al. (2010): 25	$Q = 0.863(V_w^{0.335} H_w^{1.833} W^{-0.663})$	0.422	−0.0796
	This study: 31	$Q = 1.16(V_w^{0.0419} H_w^{1.89} W^{0.389})$	0.118	−0.0859

of the variables quantified; the number of observations of each variable; the mean, median, and range of observation values for each variable; and the standard deviation of the flood peaks associated with observations of each predictor variable. This latter quantity is provided for context as it indicates the denominator value in Eq. (4).

Numerous efforts have been made to predict earthen dam flood failure peak flows using a regression model approach. Table 2 summarizes a subset of such efforts organized by predictor variable(s) used. Table 2 also reports the relative standard error and relative bias. Since the subset of flood values, Q, associated with each regression equation is somewhat different, a direct comparison of goodness-of-fit measures is not strictly possible, although there is considerable overlap between these subsets. Another matter to clarify is that the "Relevant studies: n" entry in Table 2 shows the number of observations that served to calibrate the equations provided in this table. The goodness-of-fit performance, however, is quantified based on the overall set of observations shown in the appendix.

Regression Calibration Method

As evidenced by the equations presented in Table 2, a power-law formulation is, by far, the dominant mathematical modeling structure that has been used to predict flood flows caused by an earthen dam failure. These formulations take the form

$$\hat{Q} = c_0 \cdot x_1^{c_1} \cdot x_2^{c_2} \dots \tag{6}$$

where $\hat{Q}$ denotes the model prediction of a flood flow, the set of predictor variables is $\{x_1, x_2, \dots\}$, and the set of coefficients $\{c_0, c_1, c_2, \dots\}$ is calibrated to fit the observed data set, analogously to the earlier discussion on the presentation of Eqs. (1) and (2).

A common method for the calibration of power-law relationships is to linearize them through a logarithmic transformation. Taking the logarithm of both sides of Eq. (6)

$$\log(\hat{Q}) = \log(c_0) + c_1 \cdot \log(x_1) + c_2 \cdot \log(x_2) + \cdots \tag{7}$$

which can be rewritten as

$$Y = C_0 + c_1 \cdot X_1 + c_2 \cdot X_2 + \cdots \tag{8}$$

where $Y = \log(\hat{Q})$, and the capitalized quantities are simply the log-transformed version of the original quantities in Eq. (7). Eq. (8) is the standard linear equation presented earlier as Eq. (1).

An alternative to the linearization described in Eqs. (7) and (8) is to calibrate the c_x values through nonlinear numerical optimization. In this approach, an optimizer routine systematically varies the set of c_x values with the goal of minimizing the same objective function, F, presented earlier as Eq. (2). The two alternative methods to calibrate the c_x values in the power law expression yield different sets of calibrated values, c_x. The logarithmic linearization method minimizes the differences between predicted and observed peak discharges in the logarithmic space, while the numerical optimization approach minimizes differences in the arithmetic space. The term *bias* now becomes dependent on perspective. The logarithmic transformed calibration will be unbiased (i.e., errors have zero mean) in the logarithmic space but will be biased in the arithmetic space. The numerically optimized calibration will be essentially unbiased (a very small residual can result from this approach) in the arithmetic space, but biased if examined in the logarithmic space. Studies in the statistical sciences (e.g., Miller 1984), biological sciences (e.g., Sprugel 1983; Xiao et al. 2011), and physical sciences (e.g., Ferguson 1986; Delmas et al. 2015) have raised this issue. Correction factors or other approaches have been suggested to remove the bias resulting from a logarithmic transform based on the nature of the distribution of residuals.

In this work, the smallest relative standard error is valued as the primary objective, and minimized bias as the secondary objective. The rationale for this view is that model accuracy is paramount (resulting in high importance placed on the standard error), and lack of bias is also a desired outcome. Additionally, the arithmetic space is used as the appropriate space in which to determine both the relative standard error and relative bias, since it is the peak discharges that are ultimately of interest, not their logarithms. The reported goodness-of-fit performance in Table 2 for each of the published peak flow equations serves as a baseline against which other approaches examined here can be compared.

Results

New Power-Law Relationship Approach

For each regression equation presented in Table 2, numerical optimization was employed as described earlier to calibrate a version of

these equations specific to the data set presented in the appendix. Calibrated equations are shown in Table 2 along with their goodness-of-fit performance. Table 2 shows that the power-law model relationships calibrated in this study are better in terms of both S_e/S_Q and relative bias for all relationships examined. While this is a good result, it is not a surprising result. The relationships found here were calibrated, in most cases, from a larger data set than the individual relationships also presented in Table 2. The previously calibrated relationships did not have the benefit of the additional observations used here and, thus, are operating in a predictive mode, rather than calibration mode, for a subset of observations used in the determination of the goodness-of-fit measures presented in this table. In many cases, the S_e/S_Q values are greater than 0.7, indicating that the calibrated regression model is unable to reduce the standard error to less than 70% of the standard deviation of the observations. This is reflective of the degree of variability in dam failure peak discharges.

Region of Influence Approach

The ROI approach was explored in the context of the same predictor variable equations as shown in Table 2. In all cases, numerical optimization was used to calibrate the regression equations rather than using the logarithmic transformation of Eq. (7). The ROI approach was explored allowing the k neighborhood size to vary. A rule of thumb of a minimum of 10 observations per predictor was examined by allowing k to vary from the p (where p is the number of predictor variables used) nearest neighbors up to the total number of observations, n_{obs}, in the data set. The solid line in Fig. 2 shows the measure, S_e/S_Q, as a function of k. For the case considered in the figure, $Q = f(V_w, H_w)$, as the neighborhood size increases, the S_e/S_Q measure degrades rapidly, reaching a fairly stable, approximately maximum value for k in the low 20s. When $k = n_{obs}$, the ROI calibration simplifies to the straight traditional regression approach quantified in Table 2.

On first examination, the ROI results point to using a smaller k value to give the best predictive performance. The calculation of S_e/S_Q is made by testing agreement at all n_{obs} observations in the data set. In practice, the application of the ROI technique at an unknown (V_w, H_w) test location would not generally have the benefit of a direct observation at $D = 0$ in Eq. (3). To examine this more typical predictive mode performance of the ROI approach, the analysis algorithm was modified to omit observation, j, for which $D_j = 0$ from the selected subset for regression. The relationship of S_e/S_Q is shown in Fig. 2 as the dashed line. There are two

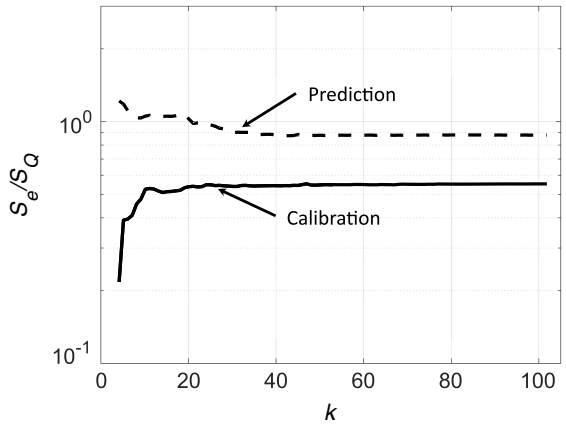

Fig. 2. Relative standard error for ROI application of $\hat{Q} = f(V_w, H_w)$ as a function of neighborhood size, k.

Table 3. Summary of region of influence and k-nearest neighbor approaches as applied to dam breach data set

Predictor(s)	Approach	Calibration		Prediction	
		S_e/S_Q	k	S_e/S_Q	k
H_d	ROI	0.758	10[a]	0.992	10[a]
	kNN(0)	—	—	0.926	25
	kNN(2)	—	—	0.983	30[b]
H_w	ROI	0.687	10[a]	0.881	10[a]
	kNN(0)	—	—	0.883	12
	kNN(2)	—	—	1.10	23
L	ROI	0.766	10[a]	1.37	10[a]
	kNN(0)	—	—	1.03	10
	kNN(2)	—	—	1.39	19[c]
S	ROI	0.716	10[a]	1.26	10[a]
	kNN(0)	—	—	0.804	3
	kNN(2)	—	—	0.851	30[b]
V_w, H_w	ROI	0.538	20[a]	1.03	20[a]
	kNN(0)	—	—	0.682	9
	kNN(2)	—	—	0.893	17
V_w, H_w, L[d]	kNN(0)	—	—	1.17	15
	kNN(2)	—	—	1.17	15
V_w, H_w, W	ROI	0.125	30[a]	0.872	30[a]
	kNN(0)	—	—	1.07	29
	kNN(2)	—	—	1.08	29

[a]S_e/S_Q value is reported at fixed value of $10p$ for all ROI analyses.
[b]Minimum S_e/S_Q value observed at fixed maximum of $k = 30$.
[c]Minimum S_e/S_Q value observed at maximum allowed by data set ($k = 19$).
[d]ROI approach not employed for this model because total number of observations was less than $10p$.

important interpretations of these S_e/S_Q relationships with k: (1) ROI performance is considerably poorer when used in a predictive mode than is suggested by the calibration statistics; (2) predictive mode performance favors larger, rather than smaller, k values for calibration—going against the rationale of the ROI approach that a smaller neighborhood near the test location should be sought in a targeted regression. In the case shown in Fig. 2, predictive mode (dashed line) performance seems to reach a best plateau at around $k = 20$–30 observations. As k continues to increase, further gains in reduced S_e/S_Q are minimal. Thus, the longstanding argument for using the ROI approach is contradicted by the illustrated regression relationship. Table 3 confirms that, for this dam breach data set, similar performance of the ROI approach is observed regardless of the predictor variable examined.

k-Nearest Neighbor Approach

The kNN approach draws on the same distance relationship, Eq. (3), as the ROI approach. These approaches diverge at this point, with ROI using conventional regression methods with the k neighborhood subset. The kNN approach essentially seeks to generate a "surface" that describes the variation in the criterion variable as a function of one or more varying predictor variables. A predictive surface of $\hat{Q}$ is calculated using a weighted average based on the distances calculated using Eq. (3):

$$\hat{Q} = \frac{\sum_{j=1}^{k} \frac{1}{D_j^{\alpha}} \cdot Q_j}{\sum_{j=1}^{k} \frac{1}{D_j^{\alpha}}} \tag{9}$$

Many natural phenomena vary spatially according to an inverse distance squared relationship, which corresponds to a value of $\alpha = 2$ in Eq. (9). Additionally, this study will explore a simpler "straight" average of k observations by setting α to zero.

Table 4. Application of $\hat{Q} = f(V_w, H_w)$ at ($V_w = 910{,}000$ m^3, $H_w = 12.2$ m)

Source of estimate	Equation used	Q (m^3/s)	Residual (m^3/s)
Observed	N/A	1,130	—
MacDonald and Langridge-Monopolis I (1984)	$Q = 1.154[(910,000) \cdot (12.2)]^{0.412}$	922	−208
MacDonald and Langridge-Monopolis II (1984)	$Q = 3.85[(910,000) \cdot (12.2)]^{0.411}$	3,030	1,900
Froehlich (1995)	$Q = 0.607[(910,000)^{0.295} \cdot (12.2)^{1.24}]$	773	−357
Regression power-law model: this study	$Q = 0.640[(910,000)^{0.392} \cdot (12.2)^{0.866}]$	1,210	80
ROI (calibration)	$Q = 0.601[(910,000)^{0.0958} \cdot (12.2)^{2.27}]$	654	−476
ROI (prediction)	$Q = 0.511[(910,000)^{0.100} \cdot (12.2)^{2.30}]$	635	−495
kNN ($\alpha = 0$, $k = 9$)	N/A	570	−560
kNN ($\alpha = 2$, $k = 17$)	N/A	657	−473

The goodness-of-fit performance for these two variations of the kNN approach [kNN(0): $\alpha = 0$ and kNN(2): $\alpha = 2$] are presented in Table 3 allowing for direct comparison to ROI performance. Several observations are readily made: (1) relative standard error, S_e/S_Q, is consistently smaller when $\alpha = 0$ as opposed to $\alpha = 2$; (2) kNN(0) values are better than the ROI equivalent in five of six cases examined (essentially equal in the remaining case); (3) kNN did not perform well ($S_e/S_Q > 1$) for either three predictor analyses: (V_w, H_w, L and V_w, H_w, W), possibly because both of these data sets are relatively small in total observations; (4) focusing on kNN(0), the optimal k-neighborhood size varies widely from 3 to 29, with 12 being the median value among the cases examined; and (5) S_e/S_Q values could not be determined for kNN in calibration mode because $D_j = 0$ in Eq. (3) when evaluated at $k = 1$, thereby yielding perfect agreement between observation and prediction, but this is not helpful in assessing kNN performance.

Application

To provide an additional perspective on the performance of the various estimation approaches, a brief demonstration of each approach is presented in the context of the functional form, $\hat{Q} = f(V_w, H_w)$, at the location $V_w = 910{,}000$ m^3, $H_w = 12.2$ m. This is the same location as the "**x**" shown in Fig. 1. There were three previous study predictive equations for this pair of predictors, and their results, along with the approaches presented in this work, are summarized in Table 4. While the performance of any approach for an arbitrary analysis point is, at best, anecdotal, it appears that the magnitude of the errors (residuals) shown in Table 4 are generally consistent with the overall findings associated with the various approaches presented. Table 4 shows that the regression power-law model calibrated from data in the appendix and shown in Table 2 performs well, yielding the estimate closest to the observed value. Langridge-Monopolis I provides the next closest estimate. The same pattern emerges as before, with the ROI (calibration) estimate being closer to the observed than the ROI (prediction estimate). The kNN(0) and kNN(2) estimates are the lowest of all developed estimates. Finally, note that the range of estimates varies from 50% to 270% of the observed flood magnitude. This range highlights the challenge of making flood estimates in the context of these multiple approaches and the sparseness of the observational database.

Summary and Conclusions

This study examined both traditional and newer innovative approaches to making informed estimates based on earthen dam failure observations that are sparse owing to the rareness of occurrence of such observations. While the equations and estimates reported here are supported by an aggregate database (appendix), caution should be used in interpreting any flood estimate, as highlighted by the large range of flood estimates developed in the presented application.

Table 5 provides a matrix-organized summary of our findings. A set of seven new regression equations was developed and presented to predict floods from earthen dam breaches as a function of several dam and reservoir characteristics. These regression equations, in all cases, produced goodness-of-fit statistics that were superior to those of the previously published equations.

Two innovative prediction approaches were examined: ROI and kNN. These approaches shared the characteristic of asserting neighborhood similarity based on a Euclidean distance metric. The most similar observations were used to draw a selective subset from the overall data set of collected observations. In the case of the ROI approach, the subset was then used to calibrate a formal regression equation specific to the test location in question. In the kNN approach, observations were used in a weighted average scheme to make an estimate at the test location.

Quantifying the performance of the ROI and kNN approaches depends strongly on whether strict calibration statistics or typical prediction statistics are sought. Goodness-of-fit statistics are calculated by quantifying the errors between the predicted and observed criterion variable—flood peak in this case—across all observations. The statistical approach to managing the neighborhood concept of the ROI and kNN approaches is ambiguous. The term *calibration mode* was used when the neighborhood used by the ROI or kNN approach was allowed to include the observation itself in making a prediction for that observation. The term *prediction mode* was used when the observation was omitted from the neighborhood used in its prediction. In calibration mode, the ROI approach performed exceedingly well, favoring the smallest possible k neighborhood size. In prediction mode, the ROI approach did not perform nearly as well, and larger k neighborhood sizes were favored. Because the general application of the ROI approach will be at locations where an observation is unavailable, we assert that the prediction mode approach provides the more realistic assessment of this approach's performance. The kNN approach suffers from the same problem as ROI in its neighborhood definition. In this work, calibration mode estimates were not made since the optimum solution would be at $k = 1$ and the goodness-of-fit performance would be perfect, a meaningless outcome.

For the two innovative approaches examined, the kNN(0) variant was found to perform better than the kNN(2) variant and ROI approach. This result means that taking the arithmetic mean of the observations in a test location's k neighborhood yielded the best results. The optimal size of the k neighborhoods was found to vary from 3 to 29 observations, depending on the predictor variable examined, with a median of 12 observations across the seven functional forms investigated.

Table 5. Summary of approaches explored in this study

Approach (symbol)	Description	Calibration information	Goodness-of-fit details	Variations	Summary findings
Historical	Historical regression equations included here for perspective.	A least-squares linear regression approach is generally coupled with a log-transformed equation.	All observations included in determination of relative standard error.	N/A	Performance is varied, but poorer than for full data set used in this study. This is to be expected since equations were calibrated with a smaller subset of data.
Numerical optimization	Regression approach calibrated by gradient search.	A least-squares objective function is minimized based on all observations (nontransformed).	All observations included in determination of relative standard error.	N/A	Performance is uniformly better than historical regression equations. Calibration by numerical optimization yields better S_e/S_Q than calibration via logarithmic linearization.
Region of influence (ROI)	Regression approach focused on k most similar observations.	Same as numerical optimization, but using only k observations.	Standard error was studied in two ways: calibration mode and prediction mode.	Calibration mode results were helped by the inclusion of the observation test point in the regression itself. Prediction mode results were poorer, but more representative.	Relative standard error in predictive mode was found to improve as k neighborhoods grew in size, contradicting the rationale of the ROI approach.
k-nearest neighbor [kNN(0), kNN(2)]	Machine learning approach similar to ROI but predictions are made based on weighted average of k observations, not formal regression.	k observations are blended using distance-based weighting.	Only leave-one-out (prediction mode) basis is rational for quantifying standard error.	Weighting exponent is zero—equal weight for all observations. Weighting exponent is 2—inverse distance squared weighting.	This approach, in predictive mode, performed better than ROI and kNN(2). Although its physical basis appeals, performance of this approach was not as strong as the equal-weight basis of kNN(0).

Two supporting matters were identified and discussed. The first concerned the concept of a convex regression hull, previously illustrated in Fig. 1. It has long been understood that one should not apply regression equations beyond the bounds of the data used in generating the regression. It is similarly important that the kNN approach not be applied outside the convex hull of the observation data set. The second matter concerned the means used to calibrate power-law regression equations. Numerical optimization, rather than the commonly used logarithmic transform, was used here to calibrate equations of the form shown in Eq. (6) and presented explicitly in Table 2. Although anecdotal, the results presented here clearly show the capacity for numerical optimization to effectively calibrate regression equations that meet or exceed the performance of power-law equations that were calibrated through log-linearization. The authors took the position that any equations calibrated would be used in the real, arithmetic space, not the logarithmic space, and thus numerical optimization was used to develop unbiased equations in the arithmetic space.

The pragmatic reader seeking a best predictive approach to estimating a dam breach peak discharge has much to wade through here. A single "best" approach or model is impossible to identify. The "Application" section illustrated the wide range of possible prediction outcomes for a specific observation and set of predictors. Based on the results presented, defensible guidance is to use both the numerical optimization equation (referred to as "This study" in Table 2) and the kNN(0) approach because these were the top performing methods. The specific numerical optimization or kNN neighborhood would be determined by the predictor variables available. Careful judgment should, of course, be exercised to ensure that the observed predictors lie within the regression hull of observations used to develop a prediction from either approach. Any prediction made should also be tempered by the information and caveats enumerated in Table 5.

Appendix. Dam Failure Data Used in This Study

ID number	Dam location	H_d (m)	H_w (m)	L (m)	S (m^3)	V_w (m^3)	W (m)	Q (m^3/s)	Reference
1	Apishapa, Colorado	34.14	28	—	22,500,000	22,200,000	82.4	6,850	Xu and Zhang (2009)
2	Baldwin Hills, California	71	12.2	198	1,100,000	910,000	59.6	1,130	Froehlich (1995)
3	Banqiao, China	24.5	31	2,000	492,000,000	607,500,000	—	78,100	Fujia and Yumei (1994)
4	Bass Haven Lake, Texas	—	4.9	—	—	641,000	22.9	240	USCOLD (1988)
5	Bayi, China	30	28	—	30,000,000	23,000,000	—	5,000	Xu and Zhang (2009)
6	Big Bay Dam, Mississippi	15.6	13.5	576.07	17,500,000	17,500,000	—	4,160	Yochum et al. (2008)
7	Bila Desna, Czech Republic	—	10.7	—	—	290,000	29.6	320	Jansen (1983)
8	Boystown, Pennsylvania	—	8.96	—	—	358,000	—	65.13	SCS (1986)
9	Bradfield, UK	28.96	—	382	3,200,000	—	50	1,150	Singh and Scarlatos (1988)
10	Break Neck Run, Pennsylvania	7	—	—	49,000	—	86	9.2	Singh and Scarlatos (1988)
11	Buffalo Creek, West Virginia	14.02	14.02	—	484,000	484,000	128	1,420	Singh and Scarlatos (1988)
12	Butler, Arizona	—	7.16	—	—	2,380,000	9.63	810	Wahl (1998)
13	Caney Coon Creek, Oklahoma	—	4.57	—	—	1,320,000	—	16.99	SCS (1986)
14	Castlewood, Colorado	21.34	21.6	—	4,230,000	6,170,000	47.4	3,570	SCS (1986)
15	Reservoir 3, Centralia, Washington	—	5.5	—	—	13,333	10.1	71	Costa (1994)
16	Chenying, China	12	12	—	4,250,000	5,000,000	—	1,200	Xu and Zhang (2009)
17	Cherokee Sandy, Oklahoma	—	5.18	—	—	444,000	—	8.5	SCS (1986)
18	Clinton Lake Dam, Illinois	19.8	—	—	91,540,000	—	—	4,254	Singh and Snorrason (1982)
19	Colonial #4, Pennsylvania	—	9.91	—	—	38,200	—	14.16	SCS (1986)
20	Dam Site #8, Mississippi	—	4.57	—	—	870,000	—	49	SCS (1986)
21	Danghe, China	46	24.5	—	15,600,000	10,700,000	—	2,500	Xu and Zhang (2009)
22	Davis Reservoir, California	11.89	11.58	—	58,000,000	58,000,000	—	510	Xu and Zhang (2009)
23	Dells, Wisconsin	18.3	18.3	—	13,000,000	13,000,000	—	5,440	Xu and Zhang (2009)
24	DMAD, Utah	8.8	—	—	19,700,000	19,700,000	—	793	Pierce et al. (2010)
25	Dongchuankou, China	31	31	—	27,000,000	27,000,000	—	21,000	Xu and Zhang (2009)
26	Eigiau, UK	10.5	10.5	—	4,520,000	4,520,000	—	400	Singh and Scarlatos (1988)
27	Elk City	9.1	9.44	564	—	1,180,000	—	608.79	Taher-shamsi et al. (2003)
28	Euclides de Cunha, Brazil	53.04	58.22	—	13,600,000	13,600,000	—	1,020	Taher-shamsi et al. (2003)
29	Field Test 1-1, Norway	—	6.1	—	—	73,000	—	190	Hassan et al. (2004)
30	Field Test 1-2, Norway	—	5.9	—	—	63,000	—	113	Hassan et al. (2004)
31	Field Test 1-3, Norway	—	5.9	—	—	63,000	—	242	Vaskinn et al. (2004)
32	Field Test 2-2, Norway	—	5	—	—	35,900	—	74	Hassan et al. (2004)
33	Field Test 2-3, Norway	—	6	—	—	67,300	—	174	Vaskinn et al. (2004)
34	Field Test 3-3, Norway	—	4.3	—	—	22,000	—	170	Vaskinn et al. (2004)
35	FP&L Martin Plant, Florida	—	5.09	—	—	125,000,000	27.7	2,750	SFWMD (1980)
36	Frankfurt, Germany	9.75	8.23	—	350,000	352,000	—	79	Xu and Zhang (2009)
37	Fred Burr, Montana	10.4	10.2	—	752,000	750,000	30.8	654	Boner and Stermitz (1967)
38	French Landing, Michigan	12.19	8.53	—	—	3,870,000	34.3	929	Xu and Zhang (2009)
39	Frenchman Creek, Montana	12.5	10.8	—	21,000,000	16,000,000	37.3	1,420	Oltman (1955)
40	Frias, Argentina	15	15	62.2	250,000	250,000	—	400	Xu and Zhang (2009)
41	Goose Creek, South Carolina	6.1	1.37	—	10,600,000	10,600,000	—	565	Taher-shamsi et al. (2003)
42	Gouhou, China	71	44	—	3,300,000	3,180,000	—	2,050	Xu and Zhang (2009)
43	Grand Rapids, Michigan	7.6	7.5	—	220,000	255,000	—	7.5	Singh and Scarlatos (1988)

Appendix. (*Continued.*)

ID number	Dam location	H_d (m)	H_w (m)	L (m)	S (m^3)	V_w (m^3)	W (m)	Q (m^3/s)	Reference
44	Granite Creek, Alaska	—	—	—	—	—	—	1,841	NRC (1983)
45	Hatchtown, Utah	19.2	16.8	238	14,800,000	14,800,000	44.8	3,080	Wahl (1998)
46	Hatfield, Wisconsin	6.8	—	—	12,300,000	—	—	3,400	Xu and Zhang (2009)
47	Haymaker, Montana	—	4.88	—	—	370,000	—	26.9	SCS (1986)
48	Hell Hole, California	67.06	35.1	—	30,600,000	30,600,000	103.2	7,360	Xu and Zhang (2009)
49	Hemet Dam, California	6.09	6.09	—	8,630,000	8,630,000	—	1,600	Taher-shamsi et al. (2003)
50	Horse Creek	12.2	7.01	701	21,000,000	12,800,000	—	3,890	Xu and Zhang (2009)
51	Horse Creek #2, Colorado	—	12.5	—	—	4,800,000	—	311.49	SCS (1986)
52	Huqitang, China	9.9	5.1	—	734,000	424,000	—	50	Xu and Zhang (2009)
53	Ireland No. 5, Colorado	—	3.81	—	—	160,000	18	110	Froehlich (1995)
54	Johnstown (South Fork Dam, Pennsylvania)	38.1	24.6	284	18,900,000	18,900,000	64	8,500	Wahl (1998)
55	Johnstown, Pennsylvania	22.86	22.25	—	18,900,000	18,900,000	—	7,079.20	Wahl (1998)
56	Kelly Barnes, Georgia	11.58	11.3	—	505,000	777,000	19.4	680	Xu and Zhang (2009)
57	Kinkaid Lake Dam, Illinois	28	—	—	96,840,000	—	—	2,011	Singh and Snorrason (1982)
58	Knife Lake Dam, Minnesota	6.096	6.096	—	9,860,000	9,860,000	—	1,098.66	Taher-shamsi et al. (2003)
59	Kodaganar, India	11.5	11.5	—	12,300,000	12,300,000	—	1,280	Xu and Zhang (2009)
60	Lake Avalon, New Mexico	14.5	13.7	—	7,750,000	31,500,000	42.7	2,320	Taher-shamsi et al. (2003)
61	Lake in the Hills Dam No. 1, Illinois	12.2	—	—	740,000	—	—	238	Singh and Snorrason (1982)
62	Lake in the Hills Dam No. 2, Illinois	4.4	—	—	100,000	—	—	321	Singh and Snorrason (1982)
63	Lake Latonka, Pennsylvania	13	6.25	—	1,590,000	4,090,000	28	290	Wahl (1998)
64	Lake Marian Dam, Illinois	15.2	—	—	190,000	—	—	90	Singh and Snorrason (1982)
65	Lake Springfield Dam, Illinois	14.6	—	—	66,000,000	—	—	3,437	Singh and Snorrason (1982)
66	Lake Tanglewood, Texas	—	16.76	—	—	4,850,000	—	1,351	SCS (1986)
67	Laurel Run, Pennsylvania	12.8	14.1	—	385,000	555,000	40.5	1,050	Froehlich (1995)
68	Lawn Lake, Colorado	7.9	6.71	—	—	798,000	14.2	510	Wahl (1998)
69	Lijiaju, China	25	25	—	1,140,000	1,140,000	—	2,950	Xu and Zhang (2009)
70	Lily Lake, Colorado	—	3.35	—	—	92,500	—	71	Froehlich (1995)
71	Little Deer Creek, Utah	26.21	22.9	—	1,730,000	1,360,000	63.1	1,330	Wahl (1998)
72	Little Wewoka, Oklahoma	—	9.45	—	—	987,000	—	42.48	SCS (1986)
73	Liujiatai, China	35.9	35.9	—	40,540,000	40,540,000	—	28,000	Xu and Zhang (2009)
74	Lower Latham, Colorado	—	5.79	—	7,080,000	7,080,000	25.7	340	Froehlich (1995)
75	Lower Reservoir, Maine	—	9.6	—	—	604,000	—	157.44	SCS (1986)
76	Lower Two Medicine, Montana	11.28	11.3	—	19,600,000	29,600,000	—	1,800	Boner and Stermitz (1967)
77	Mahe, China	19.5	19.5	—	23,400,000	23,400,000	—	4,950	Xu and Zhang (2009)
78	Mammoth, Utah	21.3	—	—	13,600,000	—	—	2,520	Xu and Zhang (2009)
79	Martin Cooling Pond Dike, Florida	—	8.53	—	136,000,000	136,000,000	—	3,115	Xu and Zhang (2009)
80	Middle Clear Boggy, Oklahoma	—	4.57	—	—	444,000	—	36.81	SCS (1986)
81	Mill River, Massachusetts	13.1	—	—	2,500,000	2,500,000	—	1,645	Wahl (1998)
82	Murnion, Montana	—	4.27	—	—	321,000	—	17.5	SCS (1986)
83	Nanaksagar, India	15.85	—	—	210,000,000	—	—	9,700	Taher-shamsi et al. (2003)
84	North Branch Tributary, Pennsylvania	5.5	5.49	—	—	22,200	—	29.5	Wahl (1998)
85	Oros, Brazil	35.36	35.8	—	650,000,000	660,000,000	110	9,630	Wahl (1998)
86	Otto Run, Pennsylvania	5.8	5.79	—	—	7,400	—	60	Pierce et al. (2010)
87	Owl Creek, Oklahoma	—	4.88	—	—	120,000	—	31	SCS (1986)
88	Peter Green, New Hampshire	—	3.96	—	—	19,700	—	4	SCS (1986)
89	Pierce Lake Dam, Illinois	14	—	—	3,280,000	—	—	864	Singh and Snorrason (1982)
90	Porter Hill, Oregon	—	5	—	—	15,000	12	30	Costa and O'Connor (1995)
91	Prospect, Colorado	—	1.68	—	—	3,540,000	13.1	116	Wahl (1998)
92	Puddingstone, California	—	15.2	—	—	617,000	—	480	Froehlich (1995)
93	Qielinggou, China	18	18	—	700,000	700,000	—	2,000	Xu and Zhang (2009)
94	Quail Creek, Utah	—	16.7	—	—	30,800,000	56.6	3,110	Wahl (1998)
95	Salles Oliveira, Brazil	35.05	38.4	—	25,900,000	71,500,000	—	7,200	Singh and Scarlatos (1988)
96	Sandy Run, Pennsylvania	8.53	8.53	—	56,800	56,700	—	435	Singh and Scarlatos (1988)
97	Schaeffer, Colorado	30.5	30.5	335	3,920,000	4,440,000	80.8	4,500	Wahl (1998)
98	Sherbourne, New York	10.36	—	91.4	42,000	—	—	960	Singh and Snorrason (1982)
99	Shimantan, China	25	27.4	500	94,400,000	117,000,000	—	30,000	Fujia and Yumei (1994)
100	Sinker Creek Dam, Idaho	21.34	21.34	—	3,330,000	3,330,000	—	926	Taher-shamsi et al. (2003)
101	Site Y-30-95, Mississippi	—	7.47	—	—	142,000	—	144.42	SCS (1986)
102	Site Y-31 A-5, Mississippi	—	9.45	—	—	386,000	—	36.98	SCS (1986)
103	Site Y-36-25, Mississippi	—	9.75	—	—	35,700	—	2.12	SCS (1986)
104	South Fork Tributary, Pennsylvania	1.8	1.83	—	—	3,700	—	122	Pierce et al. (2010)
105	South Fork, Pennsylvania	—	24.6	—	—	18,900,000	—	8,500	Froehlich (1995)

Appendix. (*Continued.*)

ID number	Dam location	H_d (m)	H_w (m)	L (m)	S (m^3)	V_w (m^3)	W (m)	Q (m^3/s)	Reference
106	Stevens Dam, Montana	—	4.27	—	—	78,900	—	5.92	SCS (1986)
107	Swift, Montana	57.61	47.85	226	37,000,000	37,000,000	—	24,947	Singh and Snorrason (1982)
108	Taum Sauk, Missouri	—	31.46	2,000.1	5,390,000	5,390,000	—	7,743	FERC (2006)
109	Teton, Idaho	92.96	77.4	—	356,000,000	310,000,000	250	65,120	Wahl (1998)
110	Upper Clear Boggy, Oklahoma	—	6.1	—	—	863,000	—	70.79	SCS (1986)
111	Upper Red Rock, Oklahoma	—	4.57	—	—	247,000	—	8.5	SCS (1986)
112	USDA-ARS Test #1, Oklahoma	2.29	2.29	7.3	—	4,900	—	6.5	Hanson et al. (2005)
113	USDA-ARS Test #3, Oklahoma	2.29	2.29	7.3	—	4,900	—	1.8	Hanson et al. (2005)
114	USDA-ARS- Test #4, Oklahoma	1.5	1.5	4.9	—	5,090	—	2.3	Hanson et al. (2005)
115	USDA-ARS Test #6, Oklahoma	1.5	1.5	4.9	—	5,190	—	1.3	Hanson et al. (2005)
116	USDA-ARS Test #7, Oklahoma	2.13	2.13	12	—	4,770	—	4.2	Hanson et al. (2005)
117	Weatland Reservoir, Wyoming	13.6	12.2	—	11,500,000	11,600,000	—	566.34	Pierce et al. (2010)
118	Weslake Dam, Illinois	14.6	—	—	2,800,000	—	—	35	Singh and Snorrason (1982)
119	Zhugou, China	23.5	23.5	—	15,400,000	18,430,000	—	11,200	Xu and Zhang (2009)
120	Zuocun, China	35	35	—	40,000,000	40,000,000	—	23,600	Xu and Zhang (2009)

Acknowledgments

This work benefited greatly from the thoughtful input and critique provided by eight reviewers and from early discussions between the first author and Rachel Moglen. The USDA prohibits discrimination in all its programs and activities on the basis of race, color, national origin, age, disability, and, where applicable, sex, marital status, familial status, parental status, religion, sexual orientation, genetic information, political beliefs, reprisal, or because all or part of an individual's income is derived from any public assistance program. (Not all prohibited bases apply to all programs.) Persons with disabilities requiring alternative means for communication of program information (e.g., Braille, large print, audiotape) should contact USDA's TARGET Center at (202) 720-2600 (voice and TDD). To file a complaint of discrimination, write to USDA, Director, Office of Civil Rights, 1400 Independence Avenue, S.W., Washington, DC, 20250-9410, or call (800) 795-3272 (voice) or (202) 720-6382 (TDD). USDA is an equal opportunity provider and employer.

References

Altman, N. S. 1992. "An introduction to kernel and nearest-neighbor nonparametric regression." *Am. Statistician* 46 (3): 175–185. https://doi.org/10.1080/00031305.1992.10475879.

Boner, F. C., and F. Stermitz. 1967. *Floods of June 1964 in northwestern Montana.* Washington, DC: USGS.

Bureau of Reclamation. 1982. *Guidelines for defining inundated areas downstream from bureau of reclamation dams*. Reclamation Planning Instruction No. 82-11. Washington, DC: Bureau of Reclamation.

Burn, D. H. 1990. "Evaluation of regional flood frequency analysis with a region of influence approach." *Water Resour. Res.* 26 (10): 2257–2265. https://doi.org/10.1029/WR026i010p02257.

Costa, J. E. 1994. *Multiple flow processes accompanying a dam-break flood in a small upland watershed, Centralia, Washington*. Water-Resources Investigations Rep. No. 94-4026. Denver: USGS.

Costa, J. E., and J. E. O'Connor. 1995. "Geomorphically effective floods." In *Natural and anthropogenic influences in fluvial geomorphology*, edited by J. E. Costa, A. J. Miller, K. W. Potter, and P. R. Wilcock. Washington, DC: American Geophysical Union.

Cover, T. M., and P. E. Hart. 1967. "Nearest neighbor pattern classification." *IEEE Trans. Inf. Theory* 13 (1): 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Delmas, M., Y. Gunnel, and M. Calvet. 2015. "A critical appraisal of allometric growth among alpine cirques based on multivariate statistics and spatial analysis." *Geomorphology* 228: 637–652. https://doi.org/10.1016/j.geomorph.2014.10.021.

Duricic, J., T. Erdik, and P. Van Gelder. 2013. "Predicting peak breach discharge due to embankment dam failure." *J. Hydroinf.* 15 (4): 1361–1376. https://doi.org/10.2166/hydro.2013.196.

FERC (Federal Energy Regulatory Commission). 2006. *Report of findings on the overtopping and embankment breach of the Upper Dam—Taum Sauk pumped storage project*. FERC No. 2277. Washington, DC: Taum Sauk Investigation Team.

Ferguson, R. I. 1986. "River loads underestimated by rating curves." *Water Resour. Res.* 22 (1): 74–76. https://doi.org/10.1029/WR022i001p00074.

Fread, D. L. 1988. *The NWS DAMBRK Model: Theoretical background/user documentation*. Silver Spring: HRL-256 Hydrologic Research Laboratory, National Weather Service.

Fread, D. L., and J. M. Lewis. 1988. "FLDWAV: A generalized flood routing model." In *Proc., National Conf. on Hydraulic Engineering*, 668–673. Colorado Springs, CO: ASCE.

Froehlich, D. C. 1995. "Peak outflow from breached embankment dam." *J. Water Resour. Plann. Manage.* 121 (1): 90–97. https://doi.org/10.1061/(ASCE)0733-9496(1995)121:1(90).

Fujia, T., and L. Yumei. 1994. "Reconstruction of Banqiao and Shimantan dams." *Int. J. Hydropower Dams* 1 (4): 49–53.

Green, S. B. 1991. "How many subjects does it take to do a regression analysis?" *Multivariate Behav. Res.* 26 (3): 499–510. https://doi.org/10.1207/s15327906mbr2603_7.

Hanson, G. J., K. R. Cook, and S. L. Hunt. 2005. "Physical modeling of overtopping erosion and breach formation of cohesive embankments." *Trans. ASAE* 48 (5): 1783–1794. https://doi.org/10.13031/2013.20012.

Harrell, F. E. 2001. *Regression modeling strategies*. New York: Springer.

Hassan, M., M. Morris, G. Hanson, and K. Lakhal. 2004. "Breach formation: Laboratory and numerical modeling of breach formation." In *Dam Safety 2004*. Phoenix: Association of State Dam Safety Officials.

Jansen, R. B. 1983. *Dams and public safety*. Denver: US Dept. of the Interior, Bureau of Reclamation.

Laton, W. R., R. J. Whitley, and T. V. Hromadka II. 2007. "A new mathematical technique for identifying potential sources of groundwater contamination." *Hydrogeol. J.* 15 (2): 333–338. https://doi.org/10.1007/s10040-006-0106-4.

MacDonald, T. C., and J. Langridge-Monopolis. 1984. "Breaching characteristics of dam failures." *J. Hydraul. Eng.* 110 (5): 567–586. https://doi.org/10.1061/(ASCE)0733-9429(1984)110:5(567).

Miller, D. M. 1984. "Reducing transformation bias in curve fitting." *Am. Statistician* 38 (2): 124–126.

National Research Council. 1983. *Safety of existing dams: Evaluation and improvement*. Washington, DC: National Academies Press.
Oltman, R. E. 1955. *Floods of April 1952 in the Missouri River basin*. Washington, DC: USGS.
Pektas, A. O., and T. Erdik. 2014. "Peak discharge prediction due to embankment dam break by using sensitivity analysis based ANN." *KSCE J. Civ. Eng.* 18 (6): 1868–1876. https://doi.org/10.1007/s12205-014-0047-8.
Pierce, M. W., C. I. Thornton, and S. R. Abt. 2010. "Predicting peak outflow from breached embankment dams." *J. Hydrol. Eng.* 15 (5): 338–349. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000197.
SCS (Soil Conservation Service). 1986. *A study of predictions of peak discharge from dam breach*. National Bulletin Rep. No. 210-6. Washington, DC.
SFWMD (South Florida Water Management District). 1980. *Embankment failure, Florida Power and Light Company, Martin Plant cooling reservoir*. Interim Final Draft Rep. West Palm Beach, FL: SFWMD.
Singh, K. P., and A. Snorrason. 1982. "Sensitivity of outflow peaks and flood stages to the selection of dam breach parameters and simulation models." Accessed April 9, 2018. https://www.isws.illinois.edu/pubdoc/CR/ISWSCR-289.pdf.
Singh, K. P., and A. Snorrason. 1984. "Sensitivity of outflow peaks and flood stages to the selection of dam breach parameters and simulation models." *J. Hydrol.* 68 (1–4): 295–310. https://doi.org/10.1016/0022-1694(84)90217-8.
Singh, V. P., and P. D. Scarlatos. 1988. "Analysis of gradual earth dam failure." *J. Hydraul. Eng.* 114 (1): 21–42. https://doi.org/10.1061/(ASCE)0733-9429(1988)114:1(21).
Sprugel, D. G. 1983. "Correcting for bias in log-transformed allometric equations." *Ecology* 64 (1): 209–210. https://doi.org/10.2307/1937343.
Taher-shamsi, A., A. V. Shetty, and V. M. Ponce. 2003. "Embankment dam breaching: Geometry and peak outflow characteristics." *Dam Eng.* 14 (2): 73–87.
Tasker, G. D., S. A. Hodge, and C. S. Barks. 1996. "Region of influence regression for estimating the 50-year flood at ungaged sites." *Water Resour. Bull.* 32 (1): 163–170. https://doi.org/10.1111/j.1752-1688.1996.tb03444.x.
Thornton, C. I., M. W. Pierce, and S. R. Abt. 2011. "Enhanced predictions for peak outflow from breached embankment dams." *J. Hydrol. Eng.* 16 (1): 81–88. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000288.
USCOLD (US Committee on Large Dams). 1988. *Lessons from dam incidents*. Reston, VA: ASCE.
Vaskinn, K. A., A. Løvoll, K. Höeg, M. Morris, G. Hanson, and M. A. A. M. Hassan. 2004. "Physical modeling of breach formation—Large scale field tests." In *Dam Safety 2004*. Phoenix: Association of State Dam Safety Officials.
Wahl, T. L. 1998. Prediction of embankment dam breach parameters: A literature review and needs assessment. Rep. No. DSO-98-004. Denver: Bureau of Reclamation, US Dept. of the Interior.
Xiao, X., E. P. White, M. B. Hooten, and S. L. Durham. 2011. "On the use of log-transformation vs. nonlinear regression for analyzing biological power laws." *Ecology* 92 (10): 1887–1894. https://doi.org/10.1890/11-0538.1.
Xu, Y., and L. M. Zhang. 2009. "Breaching parameters for earth and rockfill dams." *J. Geotech. Geoenviron. Eng.* 135 (12): 1957–1970. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000162.
Yochum, S. E., L. A. Goertz, and P. H. Jones. 2008. "Case study of the Big Bay Dam failure: Accuracy and comparison of breach predictions." *J. Hydraul. Eng.* 134 (9): 1285–1293. https://doi.org/10.1061/(ASCE)0733-9429(2008)134:9(1285).