Development of an approach to evaluate the failure probabilities of river levees based on expert judgement: Application to a case study

Michel Hathout, Laurent Peyras, Claudio Carvajal, Youssef Diab, Marc Vuillet

HAL Id: hal-03172827
https://hal.inrae.fr/hal-03172827

Submitted on 18 Mar 2021

CIWEM Chartered Institution of Water and Environmental Management   Journal of Flood Risk Management   WILEY

Development of an approach to evaluate the failure probabilities of river levees based on expert judgement: Application to a case study

Michel Hathout[1,2] 🄳 | Laurent Peyras[2] 🄳 | Claudio Carvajal[2] | Youssef Diab[1] | Marc Vuillet[1]

[1]Lab'Urba, EIVP, Université Paris-Est, Paris, France

[2]INRAE, Aix-Marseille University, RECOVER, France

Correspondence
Laurent Peyras, INRAE, 3275 route Cézanne, 13182 Aix en Provence, France.
Email: laurent.peyras@inrae.fr

Abstract

Studies carried out to analyse the risks of levees must include an evaluation of the probabilities of occurrence of different failure mechanisms. The probabilistic quantitative evaluation of these mechanisms remains difficult due to often insufficient data, the natural variability of the materials, structures that are very long, and the unavailability of mechanical models for certain failure mechanisms. This makes it necessary to call for expert judgement to evaluate the probabilities of failure. However, expert judgement generally has qualitative and subjective dimensions, and it includes biases that are liable to impair the capacities of an expert to elicit their evaluations. This article proposes an approach to processing expert judgement that includes the modalities of **I**ndividual **e**xpert **E**licitation, **C**alibration, **A**ggregation, and **D**ebiasing of expert judgement (**IeCAD**). This IeCAD approach has been developed for river levees in view to correcting biased expert evaluations in the case of evaluating the failure probability of structures.

KEYWORDS

bias, calibration, debiasing, elicitation, expert judgement, failure probability, levees

1 | INTRODUCTION

River levees are structures that are raised above the natural level of the land in view to protecting naturally floodable areas (Peyras et al., 2015). Evaluating the reliability of levees is a major challenge for the managers of these structures in order to predict a risk of failure (Kolen, Slomp, & Jonkman, 2013). Moreover, many regulations require the evaluation of the reliability of structures in a probabilistic framework, in order to demonstrate levee failure risks (Ciria et al., 2013).

Many levees are old structures and there is limited knowledge on their initial construction and further reinforcements over time, and information on their behaviour is seldom available (Ciria et al., 2013; Tourment, 2018). So evaluating the reliability of river levees in a probabilistic framework is difficult due to the scarcity of available data, related to the material composition of the levee throughout its length, their geotechnical properties, hydraulic and mechanical behaviour of levees as leakages, pore pressures, and displacements. Contrary to other hydraulic structures such as dams for which large

amounts of data linked to their construction and monitoring are available (Mouyeaux et al., 2018), the lack of information on the composition of levees often makes difficult their evaluation using probabilistic approaches.

Several studies have proposed methods to quantify the reliability of levees. The International Levee Handbook (Ciria et al., 2013) and the FloodProBE project (FloodProbe, 2012) highlight existing methods for analysing the reliability of levees: (a) Expert assessment based on previous experiences using index-based methods; (b) Index-based methods in which a number of performance features are assessed to determine reliability; (c) Empirical models for levee failure mechanisms where rules can be established to assess performance; and (d) Physical models to determine the reliability of a levee based solely on physics (FloodProbe, 2012).

According to this classification, different levee management organisations have developed methods and tools to conduct levee assessments. Serre, Peyras, Curt, Boissier, and Diab (2007); Serre, Peyras, Tourment, and Diab (2008) developed a method to assess levees based on data from visual inspections. Vuillet, Peyras, Carvajal, Serre, and Diab (2013) and Peyras et al. (2015) developed a levee assessment method aimed at estimating the performance of levees using probability distributions for levee performance indicators. Reliability analysis tools can be used to determine the overall reliability of a levee, such as the Risk Assessment for Strategic Planning method in the UK (Gouldby, Sayers, Mulet-Marti, Hassan, & Benwell, 2008), and the flood early warning system in the Netherlands (Knoeff, Vastenburg, Van den Ham, & Lopez de la Cruz, 2011).

Due to complex levee failure mechanisms (Simm et al., 2012), the different sources of data for levee assessments (Van der Meij et al., 2012), and the uncertainty on the data available (Vuillet et al., 2013), a high level of expertise is required to determine the probabilities of the levee failure mechanisms (Mériaux & Royet, 2007). The objective of this study is to propose a probabilistic approach to evaluate the reliability of levees based on expert judgement. The study proposes specific elicitation, calibration, and debiasing methods to process expert evaluations and thus reduce the biases intrinsic to expert judgement. In the end, the objective is to propose methods based on expert judgement for estimating the levee failure probability (the estimated or the subjective probability of failure), constituting an expert estimate of the actual levee failure probability.

In the scientific literature, expert judgement is the expression of an opinion based on knowledge and experience that the expert makes in response to a question (Ortiz et al., 1991). Despite their competences and in-depth knowledge, the opinions expressed by the expert contain the biases specific to opinions and judgements (O'Hagan et al., 2006). Indeed, there is always a transformation that can cause a difference between the knowledge held and the way it is presented and expressed (Bateson, 1979). The cognitive literature gives a definition of biases linked to expert judgement (Yachanin & Tweney, 1982): biases are distortions of expert reasoning that impair the validity of its inferences and conclusions (Kahneman, Slovic, & Tversky, 1982).

The methods of treating expert judgement are mainly dealt with in the statistical literature, with applications in economics (Lichtendahl Jr, Grushka-Cockayne, & Winkler, 2013), insurance (Goulet, Jacques, & Pigeon, 2009), energy (Shlyakhter, Kammen, Brodio, & Wilson, 1994), and industry (Clemen & Lichtendahl, 2002; Cooke, 1991). They have been developed to reduce biases when employing expert judgement and consist of three main modalities for dealing with bias: (a) *elicitation of expert opinion*, by focusing on developing the best modalities for collecting opinions in view to eliciting the most precise expert judgement possible (Vick, 2002), (Soll & Klayman, 2003), (Koehler & Harvey, 2004); (b) *calibration of expert judgements* (Cooke, 1991), including the weighting of opinions as a function of their pertinence, and their aggregation (Lichtendahl Jr et al., 2013); (c) *debiasing expert judgement*, aimed at applying mathematical corrections to expert evaluations (Clemen & Lichtendahl, 2002).

Regarding civil engineering, there are few works that focus on treating expert judgement and reducing bias. Peyras, Royet, and Boissier (2006) suggested employing expert judgement to measure the occurrence of function loss. Nonetheless, the subjective probabilities provided by the experts were not subjected to any specific protocol relating to their collection or treatment. Vuillet et al. (2013) proposed modalities for eliciting expert judgement aimed at reducing biases, though they did not lead to the explicit reduction of cognitive biases.

This article first presents the development of a protocol for eliciting expert opinions in the form of failure mechanisms probabilities. Then, a calibration and aggregation model is proposed for the opinions expressed by a panel of experts. Lastly, this article presents the development of an approach used to debias the opinion of the expert panel. The approach developed is applied to the case of an existing levee.

2 | APPROACH TO EVALUATING FAILURE PROBABILITIES BY EXPERT JUDGEMENT

2.1 | Principles—general overview

The principle of the approach developed to evaluate the failure probabilities of levees is based on the comparison between:

- The *calibration variables* whose true values are known through numerical calculation, and which permit evaluating the capacity of the experts to provide a pertinent and precise estimation, and then calibrate and debias their evaluation,
- The *variables of interest* whose values are searched by expert judgement.

The calibration variables are determined by expert judgement on the one hand, and by numerical calculation on the other. The comparison between the expert and numerical evaluation allows determining the calibration weighting given to the expert opinions of each expert of the panel and the correction coefficients given to the expert biases.

The proposed IeCAD (**I**ndividual **e**licitation, **C**alibration, **A**ggregation, and **D**ebiasing) approach comprises three steps of treating expert judgement (Figure 1):

- *The elicitation by expert judgement* of the failure probabilities of levee failure mechanisms;
- *The calibration and aggregation* of the opinions of different experts in view to aggregating them;
- *The debiaising of expert opinions* in order to process the biases of over- and under-confidence liable to impair the calibrated and aggregated expert evaluations:

2.2 | Identification of the calibration variables and the variables of interest

In our study, the *variables of interest* are the failure probability of cross-sections of an existing levee with respect to the sliding, internal erosion, and scouring failure mechanisms. Concerning the sliding mechanism, the probability of failure can be assessed by a reliability analysis. Concerning internal erosion and scouring, these

both failure mechanisms are subject to many stability criteria that can be found in the literature according to the type of erosion, materials, and levee profiles (Ciria et al., 2013; Vrijiling, Vrijling, 2001; Apel and Thieken, Apel, Annegret, Thieken, Merz, & Blöschl, 2006; Mazzoleni, Barontini, Ranzi, & Brandimarte, 2014; Vorogushyn, Merz, Lindenschmidt, & Apel, 2010). So for internal erosion and scouring mechanisms, it is difficult to make formulation choices given the large number of physical laws and parameters potentially combinable as well as uncertainties resulting from these modelling assumptions and from the natural variability of the properties of the material composition of the levee throughout its length. In the end, there is not a really state of the art consensus for the limit state conditions for internal erosion and scouring in the national regulations, standards and professional recommendations. Failure probabilities of levee with respect to the internal erosion and scouring failure mechanisms cannot be obtained easily with mechanical and modelling approaches, as evidenced by the important research carried out on the subject such as in the ICOLD regional club "European Working Group on Internal Erosion of Dams, Dikes and Levees and their Foundations" (https://internal-erosion.irstea.fr/). So, we are looking for the levee failure probability with respect to internal erosion and scouring mechanisms using expert judgement, and the variables of interest that we consider in our study are the failure probability of levee cross-sections with respect to these mechanisms.

The *calibration variables* must be of the same nature as the variables of interest to permit calibration and debiasing (Cooke, 1991). We adopted for the calibration variables the failure probability of the cross-sections of levees with respect to the sliding failure mechanism. Indeed, the sliding mechanism has a precise formulation of its limit conditions and so permits calculating a failure probability using quantitative approaches (Peyras, Merckle, Royet, Bacconnet, & Ducroux, 2010).
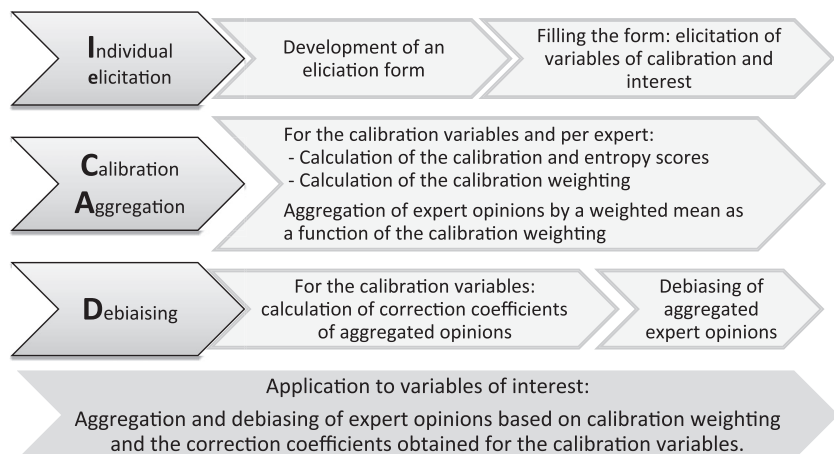
FIGURE 1 Explanatory chart of the **I**ndividual **e**xpert **E**licitation, **C**alibration, **A**ggregation, and **D**ebiasing (IeCAD) approach developed
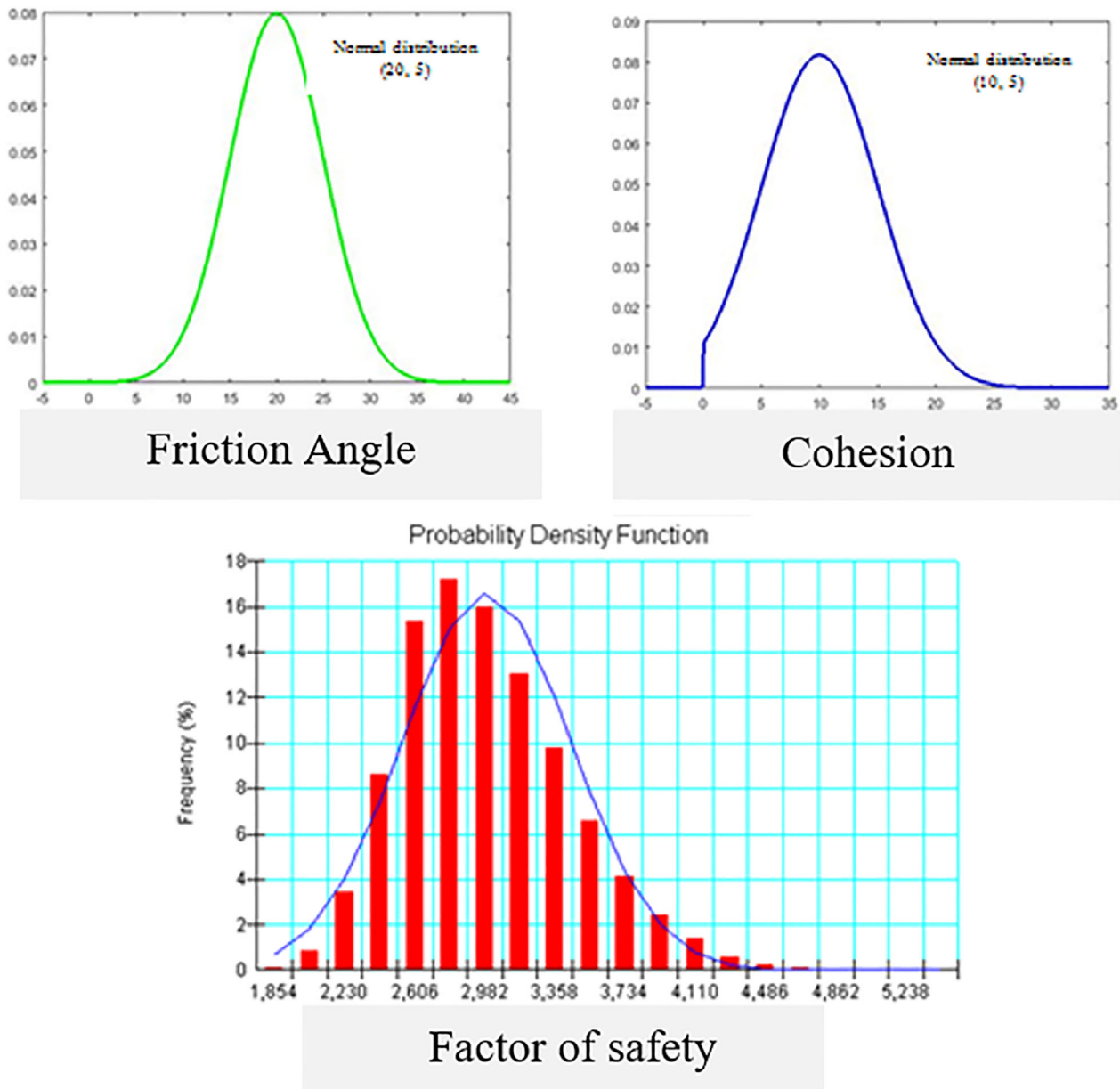
FIGURE 2 Example of a probabilistic model of the resistive properties of materials for a levee cross-section

To determine the calibration variables by numerical calculation, we search the failure probability of the sliding failure mechanism using a mechanical-probabilistic model based on the limit equilibrium and in which the resistive properties of the materials are modelled by probabilistic laws (Mouyeaux et al., 2018). It is then possible to use Monte Carlo simulations to obtain a probabilistic distribution of the safety factor and calculate the failure probability of the cross-section studied. Figure 2 illustrates the probabilistic distributions for friction and cohesion angle on a levee cross-section. With Monte Carlo simulations and considering a Morgenstern-Price method for sliding criteria, the probabilistic distribution of the safety factor obtained (Figure 2), and then the failure probability of the
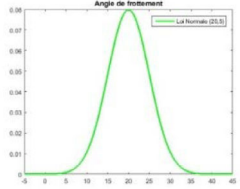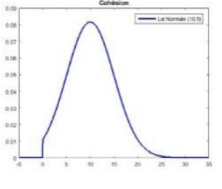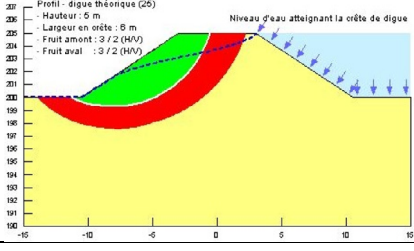
levee cross-section can be evaluated considering the integral of safety factor distribution less than 1.

For the sliding failure mechanism, we study inner slope instability in considering deep sliding circle into the embankment and levee foundations, in order to study slope instability conducting to the complete levee failure (see the slope circle considered in Figure 3).

2.3 | Approach to elicit expert opinions for river levees

Implementing the approach to eliciting expert opinions starts with identifying a panel with several experts whose

Sheet 01	Cross-section – Theoretical levee 1		
Question			
Given all the information available, what is the probability P_f that the levee will fail by **sliding of the downstream slope** if a flood whose level reaches the crest of the levee occurs?			
Information			
Geometric data	- Height = 5 m – Altitude at crest = 205 m NGF – Width of crest = 6 m - Upstream batter = 3/2 (H/V) – Downstream batter = 3/2(H/V)		
Hydraulic data	- Flood water level reaching the crest of the levee - Hydraulic gradient level in steady state		
Geotechnical data	Body and foundation - Sandy-loam material – Permeability: $K = 10^{-6}\ m/s$ – Specific gravity: $\gamma_h = 18\ kN/m^3$		
Laws of probability of input data			
	Angle of friction Law of probability: log normal distribution Mean: 20° - Standard deviation: 5°	Cohesion Law of probability: log normal dist. truncated at 0 Mean: 10 kPa – Standard deviation: 5 kPa, Min = 0 kPa	
Deterministic analysis	Calculated SF Downstream slope 1,3		
Response			
Probabilistic analysis	Probability of sliding failure (P$_f$)	P_f : quantile 5%	
		P_f : quantile 95%	
		P_f: Most likely value (q50%)	

FIGURE 3 Example of a levee cross-section sheet taken from the questionnaire form

activities are directly linked to river levees (Koehler & Harvey, 2004).

Each of the experts of the panel is questioned individually on the failure probabilities of several levee cross-sections to be evaluated with respect to different failure mechanisms. The experts are given enough information and time to fully understand the issue before being questioned. At this purpose, a questionnaire form used to collect the expert judgement was developed to this end. For each calibration variable and variable of interest to be elicited, the form contains three main items (Figure 3): the question, the information, and the response:

- The question asked to the experts concerns the sliding failure mechanism: "Given all the information available, what is the probability P_f that the levee will fail due to sliding of the downstream slope if a flood occurs that reaches the crest of the levee?." Each expert estimates the probability P_f related to the sliding failure

probability in the hypothetic situation where the flood reaches the crest of the levee. These estimates (variables of interest) will be after compared to the calculated probabilities (calibration variables) obtained in the identical hydraulic situation.

- The information are the geometric, hydraulic, and geotechnical data, the standard cross-section to be evaluated, the probability laws of the geotechnical characteristics of the cross-section, the data from the deterministic analysis represented by the safety factor SF related to the sliding failure mechanism.
- The response of the expert elicitations in the form of an uncertainty interval and a most likely value:
- An uncertainty interval [quantile 5%, quantile 95%]. Civil engineers are used to working with such quantiles since they are the same in semi-probabilistic methods such as Eurocodes (Vuillet et al., 2013),
- The most likely value contained in the uncertainty interval elicited.

2.4 | Expert opinion calibration and aggregation approach for river levees

2.4.1 | Calibration of expert opinions

The approach to calibrating the panel experts' opinions is based on Cooke's (1991) model. It consists in evaluating and weighting the panel experts' opinions in relation to calibration variables whose real values are known. Cooke's (1991) model allows calculating an individual calibration weight w_e for each expert and the relative calibration weight w'_e in relation to all the experts of the panel, since the closer this expert's elicitations are to reality, the higher the calibration weight assigned to them is.

The calibration weight is determined on the basis of the calibration score C_e and the entropy score K_e according to the following formulas (Cooke, 1991):

$$w_e = C_e . K_e \text{ and } w'_e = \frac{C_e . K_e}{\sum\limits_{e=1}^{z} C_e . K_e}, \qquad (1)$$

where w_e is the calibration weight of the expert e, w'_e is the calibration weight relative to the expert e in a panel of experts, C_e is the expert e's calibration score e, K_e is the expert e's entropy score e, and z is the number of experts.

The calibration score C_e permits evaluating the accuracy of the information given by the expert for the calibration variables, by comparing between a distribution obtained by the expert and a distribution calculated numerically. It is determined by an error probability evaluated with the χ^2 statistical test (interdependence test). The calibration score C_e is evaluated by the following formula (Cooke, 1991):

$$C_e = 1 - P(2*n*I_e(c,p)) = 1 - \mathcal{X}^2(2*n*I_e(c,p)), \qquad (2)$$

where $P()$ is the probability of a random variable following a χ^2 law; $\mathcal{X}^2()$ is the distribution function of a random variable following the χ^2 law; n is the number of calibration variables; c is the calibration vector representing the portion of true values included within each inter-quantile interval $c = \{c_1, c_2, ..., c_j\}$; p is the theoretical probabilities vector representing each inter-quantile interval $p = \{p_1, p_2, ..., p_j\}$; $I_e(c,p) = \sum\limits_{i=1}^{j} c_i \ln\left(\frac{c_i}{p_i}\right)$: relative information between the theoretical probabilities p and the calibration vector c.

The entropy score K_e permits measuring the quantity of information contained in the probabilistic distributions given by the experts. It is based on a measure of the distances between the vector of subjective probabilities s and the vector of theoretical probabilities p. The entropy score K_e is evaluated by the following equation (Cooke, 1991):

$$K_e = \frac{1}{n} \sum_{i=1}^{n} I_{e,i}(p,s), \qquad (3)$$

where s is the vector of subjective probability representing the subjective probabilities of each inter-quantile interval. The probabilities of s are determined by the percentage of each inter-quantile interval relative to the probabilistic scale of the calibration variable $s = \left\{ \frac{q_{5\%} - q_{0\%}}{q_{100\%} - q_{0\%}}, \frac{q_{50\%} - q_{5\%}}{q_{100\%} - q_{0\%}}, \frac{q_{95\%} - q_{50\%}}{q_{100\%} - q_{0\%}}, \frac{q_{100\%} - q_{95\%}}{q_{100\%} - q_{0\%}} \right\}$; and $I_e(p,s) = \sum\limits_{i=1}^{j} p_i \ln\left(\frac{p_i}{s_i}\right)$ is the relative information between the vector of subjective probabilities s and the vector of theoretical probabilities p.

2.4.2 | Aggregation of expert opinions

We propose applying the aggregation of the panel experts' opinions by way of the median of the quantiles. The aim is to conserve the initial form of the expert elicitations given in a probabilistic format, with a value considered as the most likely, and an uncertainty interval (Lichtendahl Jr et al., 2013). This approach corresponds to the weighted sum of the calibrated quantiles of the panel experts' opinions. The aggregated quantiles ($q_{5\%}$, $q_{50\%}$, $q_{95\%}$) are evaluated by the following formulas:

$$q_{5\%} = \sum w'_e . q_{5\%(e)}; q_{50\%} = \sum w'_e . q_{50\%(e)}; q_{95\%} = \sum w'_e . q_{95\%(e)}, \qquad (4)$$

where $q_{5\%}$, $q_{50\%}$, $q_{95\%}$ are the aggregated quantile corresponding to a probability, respectively of 5, 50, and 95%.

At the end of the expert opinion calibration and aggregation phase, we now have a single evaluation of the expert panel that takes into account the relative weight of each expert in the panel. This calibrated and aggregated evaluation is the response of the expert panel to the question of evaluating the failure probabilities of levee for a given failure mechanism.

2.5 | Expert judgement debiasing approach for river levees

The aim is to apply mathematical corrections to the panel's expert evaluations, in order to get as close as

possible to the calibration variables. The objective is to quantitatively identify the trend of the expert panel's opinions for under- or over-confidence which impair the estimation of the most likely value and the uncertainty interval, and then to apply a mathematical correction. We propose to implement an expert opinion debiasing approach based on the model of Clemen and Lichtendahl (2002), consisting in determining three correction coefficients by iterative calculation.

The first coefficient β is intended to correct the most likely value corresponding to the 50% quantile:

$$R_i^* = \beta R_i, \tag{5}$$

where R_i^* is the most likely debiased value; β is the correction coefficient of the most likely value; R_i is the calibrated and aggregated expert elicitation of the most likely value (q50%).

The second coefficient α_L is aimed at correcting the value elicited of the lower bound of the uncertainty interval, corresponding to the 5% quantile:

$$L_i^* = (\beta - \alpha_L) R_i + \alpha_L L_i, \tag{6}$$

where L_i^* is the debiased lower bound of the uncertainty interval; α_L is the correction coefficient of the lower bound of the uncertainty interval; L_i is the calibrated; and aggregated expert elicitation of the lower bound of the uncertainty interval (q5%).

The purpose of the third coefficient α_U is to correct the elicited value of the upper bound of the uncertainty interval, corresponding to the 95% quantile:

$$U_i^* = (\beta - \alpha_U) R_i + \alpha_U U_i, \tag{7}$$

where U_i^* is the debiased upper bound of the uncertainty interval; α_U is the correction coefficient of the upper bound of the uncertainty interval; U_i is the calibrated; and aggregated elicitation of the upper bound of the uncertainty interval (q95%).

3 | APPLICATION AND RESULTS OF THE APPROACH DEVELOPED

The levee studied is an earth-fill levee 5,500 m long raised to protect a town in France (Figure 4). The height of the levee varies from 1 to 6 m (Figure 5).

The approach developed was implemented by a panel of six engineers having different professional backgrounds

and experiences (geotechnics, river hydraulics, civil engineering, and hydrology). The calibration variables (variables no. 1–30) correspond to failure probabilities associated with the sliding mechanism for 30 cross-sections: these variables are obtained by numerical mechanical-probabilistic calculation and permit a robust statistical analysis.

The study was applied to 30 variables of interest (variables no. 31–60; Table 1) obtained by expert judgement:

- Variables of interest no. 31–40 correspond to failure probabilities of the sliding mechanism. For these 10 variables of interest, we have expert judgement evaluation but also numerical calculation evaluation, so we will be able to compare the results from IeCAD approach to the true values of the variables,
- Variables of interest no. 41–60 correspond to failure probabilities with respect to internal erosion and scouring mechanisms.

## 3.1 | Application and results of the expert opinion elicitation phase

Figure 6 presents the elicitations of expert no. 1 relating to failure probabilities (P_f) regarding three failure mechanisms. The number of the variable related to its cross section is given in the vertical axis, the expert values (quantiles 5, 50, 95%) are the black squares and the true values obtained by calculations are the grey dots.

Table 2 shows the distribution the true known values of the calibration variables in the interquantile intervals elicited by the six experts for the cross-sections no. 1 to 30:

The percentage of the true values contained in the range [5%, 95%] varies from 40% (17% + 23%) for expert no. 3 to 66% (33% + 33%) for expert no.2, which is considerably lower than the target percentage of 90% (90% corresponds to the interquantile interval between quantiles 5 and 95%). The distribution of the median of the elicitations of expert opinions contains 55% of the true values in the uncertainty interval [5%, 95%], which is also considerably lower than the target percentage of 90%. This means that a large number of true values lie outside the uncertainty intervals elicited by the panel experts, reflecting a trend towards overconfidence or underconfidence in the expert judgement.

The distribution of the median of the elicitations of expert opinions contains 26% of true values in the interval [0%, 5%], which is higher than target percentage of 5%. This means that a high percentage of the expert evaluations led to overestimated values for the failure
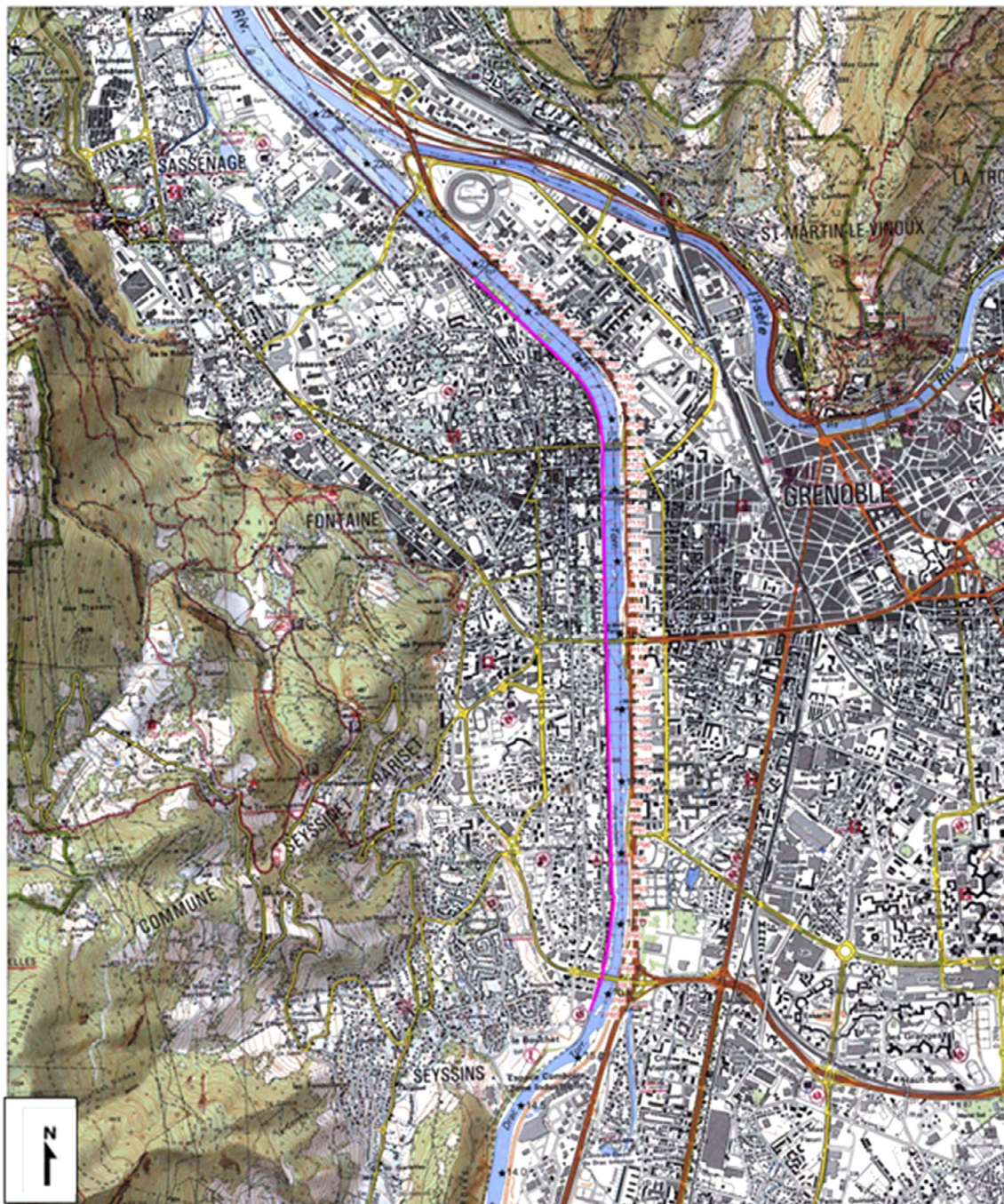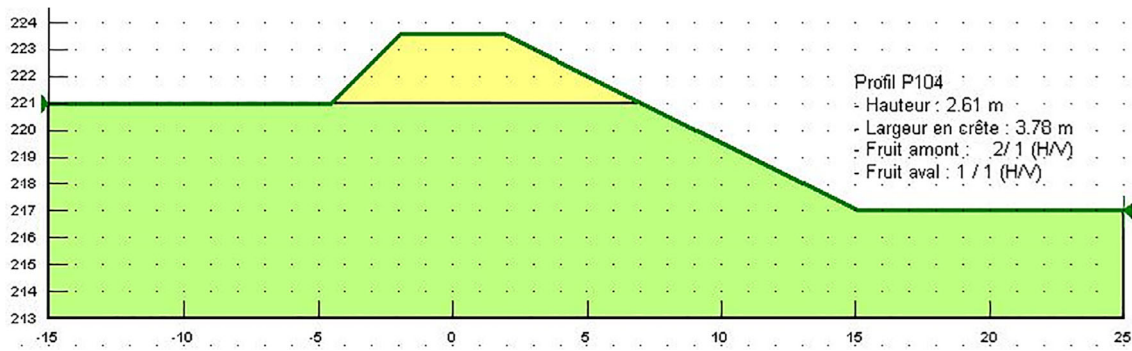
FIGURE 4 Aerial view of the levee studied (scale: 1/25,000 in A3 format)

probability, reflecting a trend towards under-confidence, and expressing caution in the experts' evaluations. These overestimated evaluations tend towards safety, but can also lead to decisions to carry out expensive levee reinforcement works in excess of their real necessity. On the other hand, the distribution of the mean of the elicitations of expert opinions contains 19% of true values in the interval [95%, 100%], which is also higher than the target percentage of 5%. This means that a high percentage of the expert evaluations also gave underestimated failure probability values, reflecting a trend towards overconfidence. These underestimated evaluations do not tend towards safety and lead to overestimating the levee's resistance. The combined presence of large numbers of biases of over- and under-confidence (with nonetheless a greater bias towards underconfidence tending towards safety) demonstrate in a general way the biases intrinsic to expert evaluations.

HATHOUT ET AL.

CIWEM Chartered Institution of
Water and Environmental
Management

Journal of
Flood Risk Management—WILEY 9 of 18

FIGURE 5   Standard cross-section

TABLE 1   Table summarising the calibration variables and the variables of interest used in the approach

	Calibration variables		Variables of interest	
Failure mechanism	Number of variables	No. of variable	Number of variables	No. of variable
Sliding	30	No.1–30	10	No. 31–40
Internal erosion	—	—	10	No. 41–50
Scouring	—	—	10	No. 51–60

3.2 | Application and results of the expert opinion calibration phase

The result of the expert opinion calibration is given with the mean of the calibration scores, the entropy scores and relative calibration weighting (Table 3), obtained from the 30 calibration variables (no. 1–30).

Since elicitation is performed on the $q_{5\%}$, $q_{50\%}$ and $q_{95\%}$ quantiles, the theoretical probability vector p associated with the inter-quantile intervals is p = {0.05, 0.45, 0.45, 0.05}. The calibration vectors c can be obtained directly from Table 2 (for expert 1, c = {0.30, 0.37, 0.23, 0.10}). The calibration score C_e presented in Table 3 is then calculated using the relative information between c and p vectors, $I_e(c,p)$ according Equation 2 (for expert 1: $I_e(c,p)$ = 0.38, which gives a score C_e = 0.79 by using Equation 2).

For the calculation of the entropy score, as the elicited failure probability values are expressed in the format (10^{-x}), we change the variable in order to conserve a constant difference in absolute value between (10^{-x}) and $(10^{-(x-1)})$, which amounts to applying the common logarithm to the elicited probabilities $(\log_{10}(P_f))$. For example, expert 1 elicits the probability values of 10^{-4}, 10^{-2} and 10^{-1} for the quantiles 5, 50 and 95%, respectively, for the first calibration variable. With the variable change to base $\log_{10}$, we thus obtain $q_{5\%}$ = −4, $q_{50\%}$ = −2 and $q_{95\%}$ = −1, which makes it possible to evaluate the subjective probability vector s_1 = {0.24, 0.39, 0.19, 0.18}

and the relative information $I_{e,1}(p,s)$ = 0.31 with Equation 3. The entropy score K_e is finally evaluated with Equation 3, and corresponds to the average of the relative information $I_{e,i}(p,s)$ evaluated for the 30 calibration variables.

The results obtained from the calibration of the expert opinions show that the values of the calibration score (C_e) vary between 3 and 91%. The highest value was obtained by expert no.2, indicating that their elicitations contained the highest number of true calibration values in comparison to the elicitations of the other panel experts. On the contrary, the lowest calibration score was obtained by expert no.3, indicating that their elicitations contained the lowest number of true calibration values.

Regarding the entropy score, the values varied from 0.52 to 1.04. The highest value was obtained by expert no.4, indicating that the uncertainty intervals elicited by expert no. 4 were more precise than those given by the other experts of the panel. On the contrary, the lowest score was obtained by expert no.5, indicating that the uncertainty intervals elicited by this expert were wider than those given by the other experts of the panel.

Finally, the calibration phase allowed assigning a relative calibration weighting to each expert, taking into account both the pertinence and the precision of their evaluations. In our application, it turned out that expert no. 2 had the best relative calibration weighting (40%) in the panel, contrary to expert no.5 (3%).

P_f
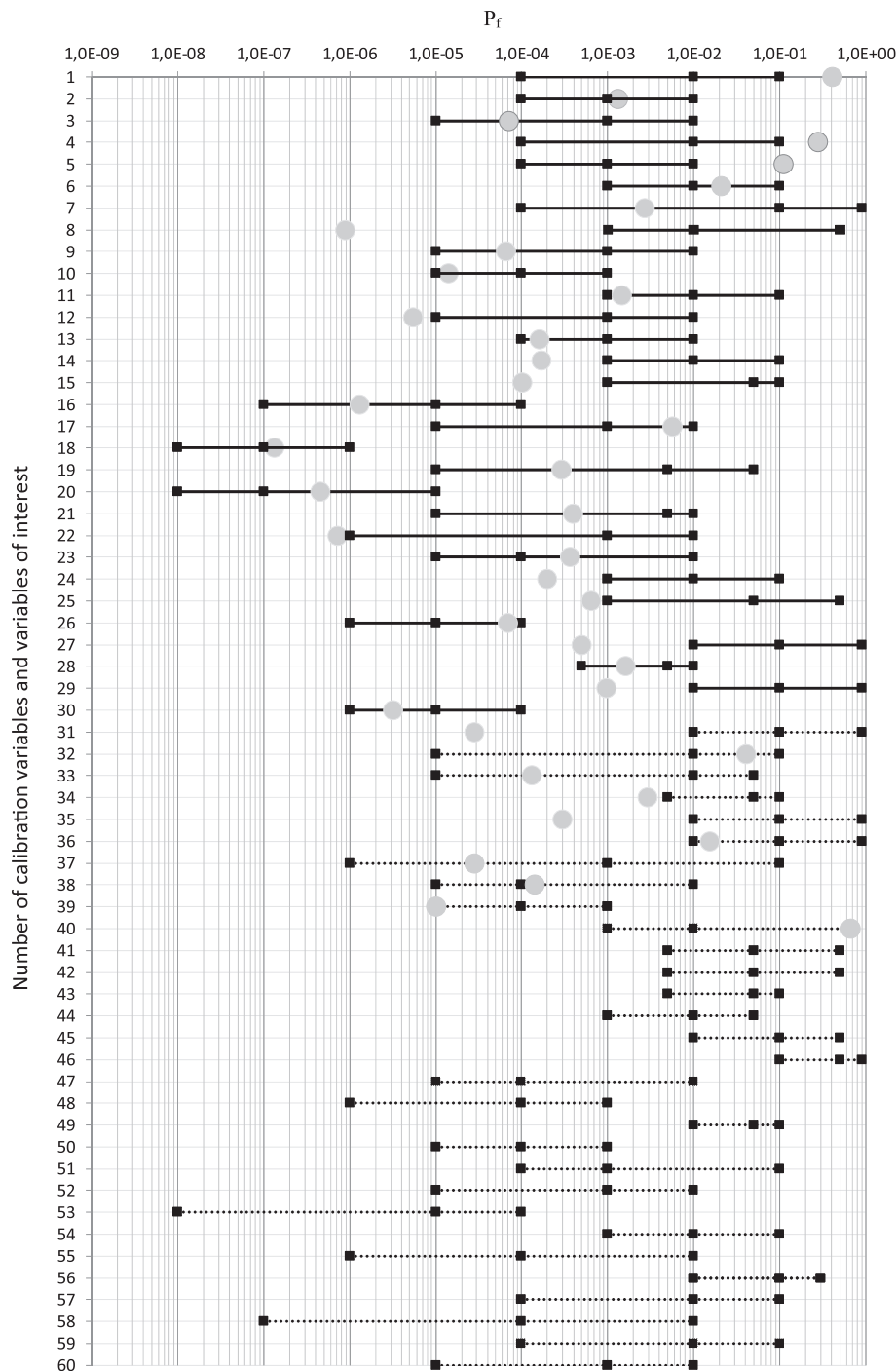
FIGURE 6 Elicitations of expert no. 1 relating to calibration variables

3.3 | Application and results of the expert opinion aggregation phase

The aggregation phase corresponds to a weighted sum performed for each of the quantiles $q_{5\%}$, $q_{50\%}$ and $q_{95\%}$ (using Equation 4) and for each of the variables. For the levee section corresponding to variable n°1, the results obtained are $q_{5\%} = -3.34$, $q_{50\%} = -1.93$ and $q_{95\%} = -0.72$ (in $\log_{10}[P_f]$ scale), which correspond respectively to $q_{5\%} = 4.6\mathrm{E}{-}04$, $q_{50\%} = 1.2\mathrm{E}{-}02$ and $q_{95\%} = 1.9\mathrm{E}{-}01$ (P_f). Figure 7 illustrates the result for the 30 calibration

variables associated with the theoretical levee cross-sections (continuous lines).

Figure 7 can be interpreted using Table 4 below, which presents the distribution of the true values of the calibration variables in the inter-quantile intervals at the end of the expert opinion aggregation phase:

At the end of the calibration and aggregation phase, we observed that the distributions of the calibrated and aggregated expert opinions contained 60% of true values in the uncertainty interval [5%, 95%]. Thus, the aggregation and the calibration of expert opinions led

TABLE 2 Distribution of the true values of the calibration variables (no. 1–30) in the interquantile intervals elicited by the experts and the median of the elicitations
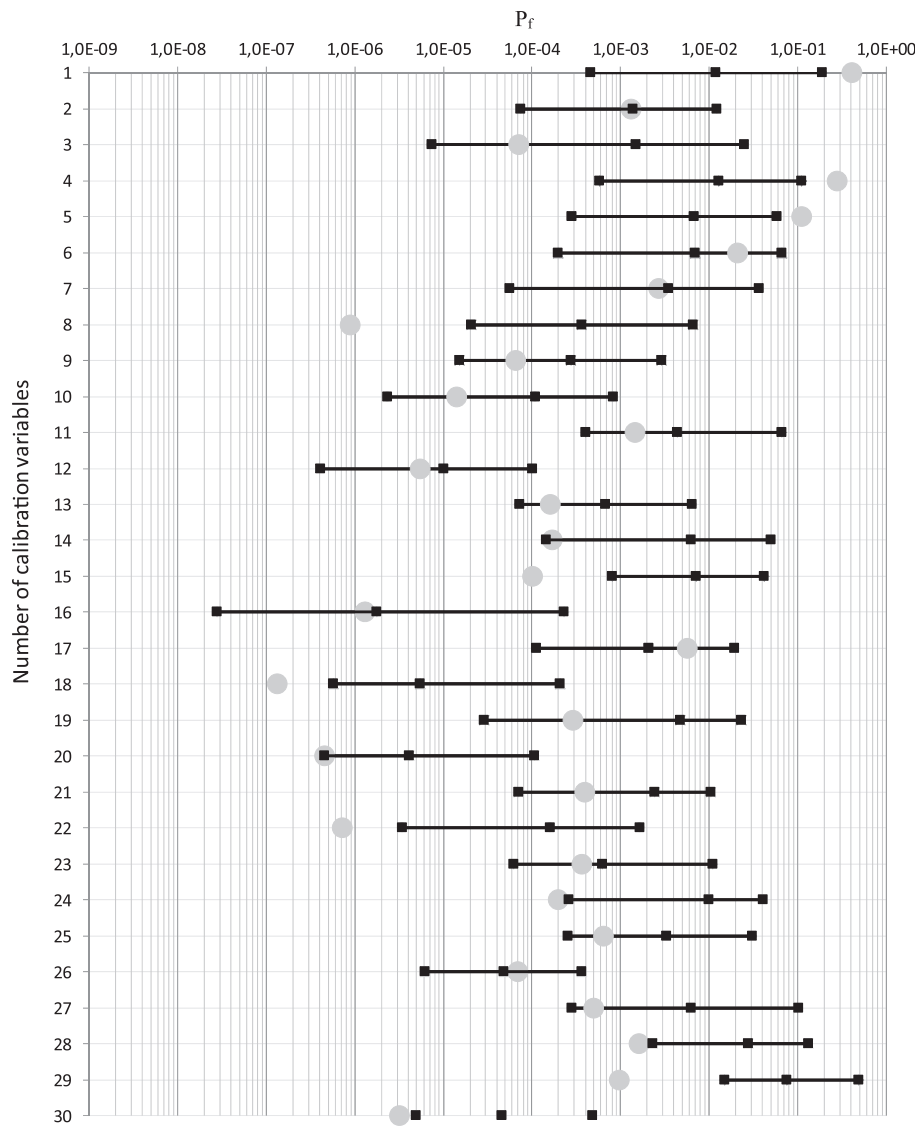
Calibration variable (no. 1–30)						
Inter-quantile interval			[0%, 5%]	[5%, 50%]	[50%, 95%]	[95%, 100%]
Expert no. 1						
(P_f) (no. 1–30)	n		9	11	7	3
	%		30%	37%	23%	10%
Expert no. 2						
(P_f) (no. 1–30)	n		9	10	10	1
	%		30%	33%	33%	3%
Expert no. 3						
(P_f) (no. 1–30)	n		10	5	7	8
	%		33%	17%	23%	27%
Expert no. 4						
(P_f) (no. 1–30)	n		10	8	8	4
	%		33%	27%	27%	13%
Expert no. 5						
(P_f) (no. 1–30)	n		0	4	15	11
	%		0%	13%	50%	37%
Expert no. 6						
(P_f) (no. 1–30)	n		8	8	6	8
	%		27%	27%	20%	27%
Median of elicitations						
$(\mathbf{P_f})$ (no. 1–30)	n		7.67	7.67	8.83	5.83
	%		26%	26%	29%	19%

TABLE 3 Results of the expert opinion calibration phase in comparison to the calibration variables

No. expert	Calibration score C_e	Entropy score $\log_{10}(P_f)$ K_e	Weight of individual calibration w_e	Weight of relative calibration w'_e
1	0.79	0.74	0.59	0.28
2	0.91	0.91	0.83	0.40
3	0.03	0.79	0.02	0.01
4	0.46	1.04	0.48	0.23
5	0.14	0.44	0.06	0.03
6	0.19	0.52	0.10	0.05
			Total	1

to a higher number of true values in the uncertainty interval [5%, 95%] in comparison to the raw elicitations comprising 55% of true values in this interval (Table 2). This calibration and aggregation phase therefore improved the quality of the expert evaluations, despite the fact that the correction did not allow reaching the percentage of the ideal target interval (90%), leaving a still considerable bias of overconfidence or under-confidence following the calibration and aggregation phase.

More specifically, the results obtained following the calibration and aggregation phase showed a substantial reduction of the true values in the overconfidence interval [95; 100%], falling from 19% (Table 2) to 10% (Table 4), which was a significant improvement and thus tending to increase the reliability of the levees. The evolutions in the interval [5%, 95%] were slight, changing from 26% (Table 2) to 30% (Table 4), expressing a small evolution in processing under-confidence during the calibration and aggregation phase.

FIGURE 7 Calibration variables, calibrated and aggregated calibration variables and their uncertainty intervals

Calibrated and aggregated elicitations of calibration variables (no. 1–30)					
Interquartile interval		[0%, 5%]	[5%, 50%]	[50%, 95%]	[95%, 100%]
(P_f) (no. 1–30)	n	9	15	3	3
	%	30%	50%	10%	10%

TABLE 4 Distribution of the true calibration values (no. 1–30) in the inter-quantile intervals

3.4 | Application and results of the expert panel opinion debiasing phase
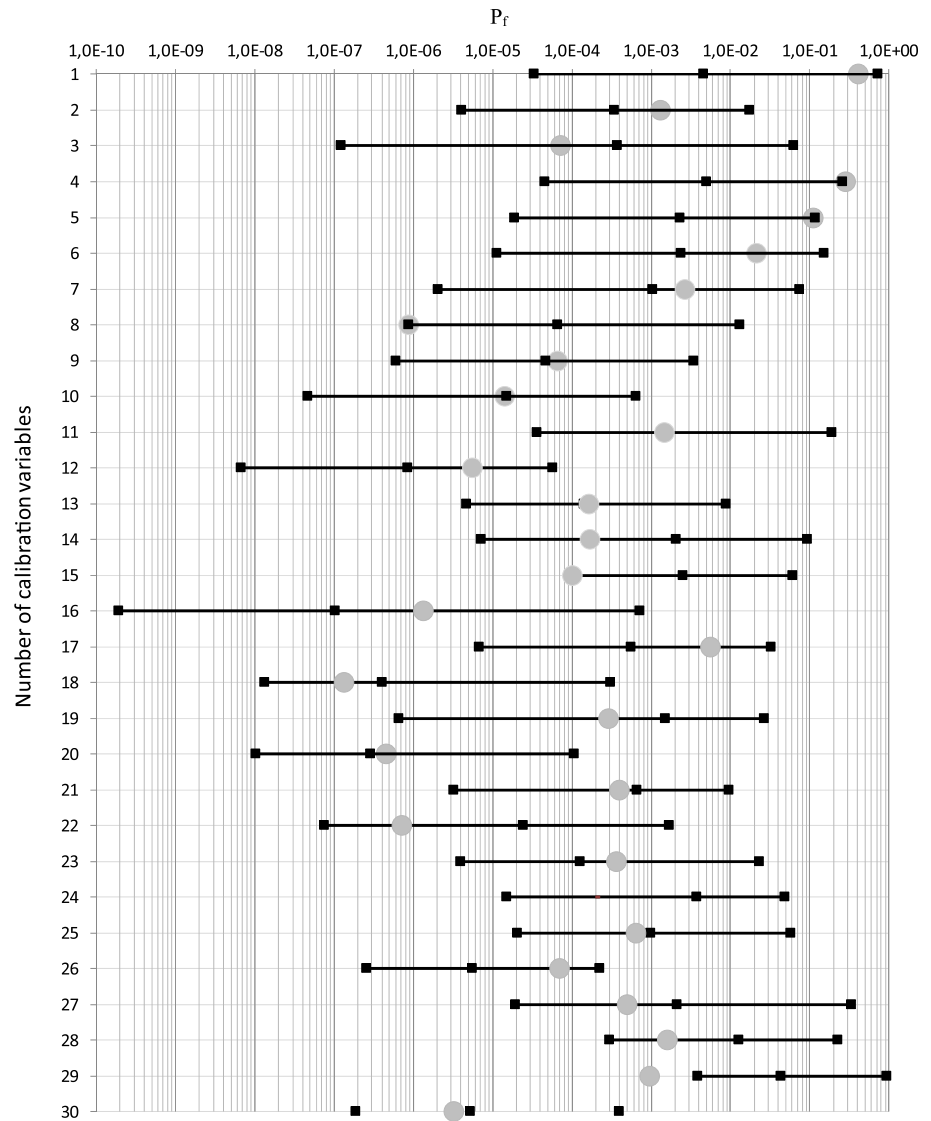
The correction coefficients defined in the debiasing phase were obtained by an iterative calculation, starting with the coefficient β (Equation 5): until obtaining 50% of true values in the interquartile interval [0%,50%] (and therefore 50% in the interquartile interval [50%,100%]). Then, the correction coefficients α_L and α_U (Equations (6) and (7) respectively) are also obtained by iterative calculation: until obtaining respectively 5% of

the true values in the interquantile interval [0%, 5%] and 5% of the true values in the inter-quantile interval [95%, 100%]. The correction coefficients obtained by these iterative calculations are: (a) coefficient $\beta = 1.22$ for the correction of quantile $q_{50\%}$, (b) coefficient $\alpha_L = 1.51$ for the correction of quantile $q_{5\%}$ and (c) coefficient $\alpha_U = 1.83$ for the correction of quantile $q_{95\%}$.

- The correction coefficient ($\beta = 1.22$) shows that panel of experts tended to reduce its estimations regarding the most likely central variables (R_i). Thus, a coefficient β

FIGURE 8 Debiased calibration variables of the panel of experts

must be applied to the central values resulting from the aggregation phase to obtain the value of the 50% quantile of the calibration variable;

- The correction coefficient ($\alpha_L = 1.51$) reflects that the distances estimated between (L_i) and (R_i) resulting from the aggregation phase tend to be lower than those they should have been. Thus, it was necessary to apply a coefficient α_L to obtain the debiased distance between the 5% and the 50% quantile;
- The correction coefficient ($\alpha_U = 1.83$) reflects that the distances estimated between (R_i) and (U_i) resulting from the aggregation phase tend to be lower than they should have been. Thus, a coefficient α_U must be applied to obtain the debiased distance between the 50% and the 95% quantile.

For the levee section corresponding to variable n°1, with these correction coefficient values, and using Equations (5)–(7), the values of the debiased quantiles are

$q_{5\%} = -4.48$, $q_{50\%} = -2.35$ and $q_{95\%} = -0.14$ (in $\log_{10}(P_f)$ scale), which correspond, respectively, to $q_{5\%} = 3.3E{-}05$, $q_{50\%} = 4.4E{-}03$ and $q_{95\%} = 7.2E{-}01$ (P_f). Figure 8 shows the debiased opinions of the expert panel resulting from the application of correction coefficients to the aggregated opinions of the panel of experts.

Figure 8 is interpreted using Table 5 below, which presents the distribution of the true values of the calibration variables in the interquantile intervals following the debiasing phase.

At the end of the debiasing phase, we observed that the distributions of the opinions of the expert panel contained 94% of true values in the uncertainty interval [5%, 95%]. Thus, the debiasing phase of the opinions of the expert panel permitted reaching the percentage of the ideal target.

The results obtained after the debiasing phase show an optimal reduction of true values in the overconfidence interval [0%; 5%], falling from 30% (Table 4) to 3% (Table 5), indicating a considerable improvement in the

Inter-quantile interval		[0%, 5%]	[5%, 50%]	[50%, 95%]	[95%, 100%]	**TABLE 5** Distribution of the true values of the calibration variables (no. 1–30) in the debiased interquantile intervals
(P_f) (no. 1–30)	n	1	14	14	1	
	%	3%	47%	47%	3%	
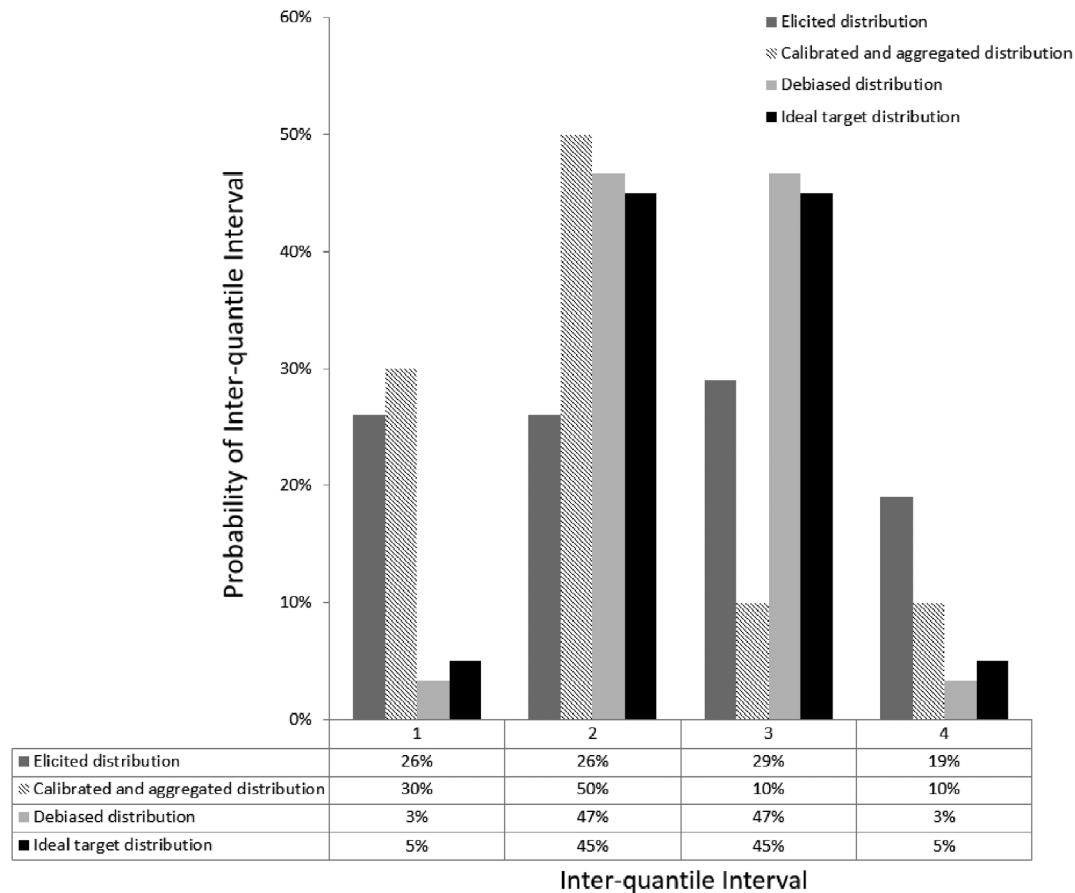
FIGURE 9 Evolution of the quality of information in the sequence of processing the **I**ndividual **e**xpert **E**licitation, **C**alibration, **A**ggregation, and **D**ebiasing (IeCAD) approach

quality of the evaluation of the structures. The evolutions in the interval [95%, 100%] were also considerable, decreasing from 10% (Table 4) to 3% (Table 5), expressing a marked evolution in the treatment of overconfidence bias during the expert opinion debiasing phase, tending towards an improvement in the reliability of the structures.

3.5 | Synthesis of the IeCAD approach to processing expert judgement

Figure 9 summarises the evolution of the quality of information provided by expert judgement in the sequence of the different phases of processing the IeCAD approach:

- interval [5%, 95%]: the distributions of the expert panel opinions contained 55% of true values following the

elicitation phase, then 60% following the calibration phase, then 94% following the debiasing phase;
- interval [0%, 5%]: the distributions of the expert panel opinions contained 26% of true values following the elicitation phase, then 30% following the calibration phase, then 3% following the debiasing phase;
- interval [95%, 100%]: the distributions of the expert panel opinions contained 19% of true values following the elicitation phase, then 10% following the calibration phase, then 3% following the debiasing phase.

3.6 | Application to the variables of interest

The expert judgement calibration and debiasing approach was applied to the variables of interest. Figure 2 presents

TABLE 6 Distribution of the true values of the variables of interest (no. 31–40) in the calibrated and aggregated confidence intervals

Inter-quantile interval		[0%, 5%]	[5%, 50%]	[50%, 95%]	[95%, 100%]
Calibrated and aggregated variables					
(P_f) (no. 31–40)	n	2	5	2	1
	%	20%	50%	20%	10%
Calibrated, aggregated, and debiased variables					
(P_f) (no. 31–40)	n	1	6	3	0
	%	10%	60%	30%	0%

FIGURE 10 Debiased variables of interest associated with their confidence intervals
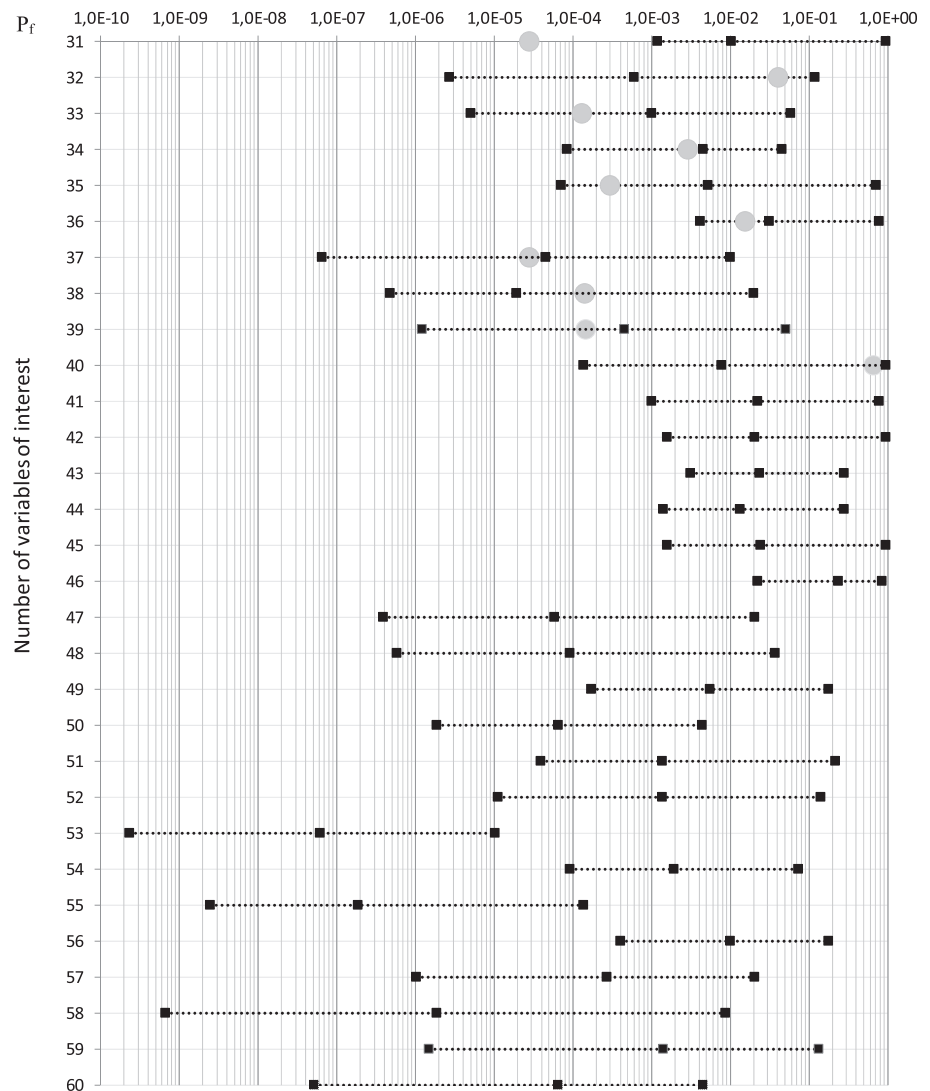
the results on 30 variables of interest corresponding to 10 levee cross-sections studied for three failure mechanisms (sliding, internal erosion, and scouring).

For variables of interest no. 31–40 relating to the sliding mechanism, Table 6 shows the distribution of the true values in the interquantile intervals, calibrated and aggregated on the one hand, and aggregated and debiased on the other.

At the end of the debiasing phase, we observed that the distributions of the debiased expert panel opinions contained 90% of true values in the uncertainty interval [5%, 95%]. Thus, the expert panel opinion debiasing phase permitted reaching the ideal target interval.

In detail, the results obtained after the debiasing phase show a reduction of the true values in the under-confidence interval [0%; 5%], falling from 20%

(Table 6 – calibration and aggregation phase) to 10% (Table 6- debiasing phase), which is a considerable improvement, tending towards increased reliability of the structures. The evolutions in the interval [95%, 100%] are equally considerable, dropping from 10% (Table 6 – calibration and aggregation phase) to 0% (Table 6- debiasing phase), reflecting a substantial change in the treatment of the overconfidence bias during the expert opinion debiasing phase. These evolutions of overconfidence bias are positive since they tend towards greater reliability for the structure.

The correction provided by debiasing therefore appears significant, demonstrating the advantage of applying a full treatment including calibration-aggregation and debiasing to the evaluations performed by the panel of experts for failure probabilities. By way of illustration, Figure 10 shows the treatment of the variables of interest for the internal erosion (no.41–50) and scouring (no.51–60) failure mechanisms for which no limit state models were available.

3.7 | Discussion

The results obtained in the case study show that the calibration-aggregation phase as well as the debiasing phase of the developed approach allow improving the results obtained from expert elicitations.

The quality of the final result is logically conditioned by the quality of the experts. Logically more experienced experts can result in better evaluation of the levee failure probability, statistically. This would be reflected in particular in the width of the uncertainty intervals ([5%, 95%] in the case study): the better the experts, the narrower the interval (leading to a more precise result); on the contrary, with less experienced experts, the interval obtained would be wider (leading to an imprecise result).

In the case study, the panel was composed of six experts, which is a rather high number for a risk analysis study of river levee (Peyras et al., 2012). These experts had different experiences in the concerned disciples (geotechnics, civil engineering, hydraulics) and also different levels of qualification (junior, confirmed and senior). Thus, there is necessarily a dispersion in their assessments.

This article did not study the influence of the number of experts on the results. In the case study, a panel of six experts was selected in order to reproduce a panel of experts as implemented in the river dike risk analysis studies (typically 3–6 experts—see Peyras et al., 2012). However, an increase in the number of experts would not necessarily lead to an improvement in results: for example, it can be expected that the quality of results would be better if the number of experts were reduced by keeping the most qualified experts. Thus, the professional qualification of experts would have a greater influence on the results in relation to the number of experts.

4 | CONCLUSION

Regarding the field of river levees, the lack of data on structures, the uncertainties that impair them and the lack of consensus in the state of the art for the limit states related to internal erosion and scouring failure mechanisms, reduce the opportunity of using statistical and probabilistic methods for determining their failure probability using mechanistic-probabilistic approaches. This makes expert judgement essential for evaluating the failure probabilities of levees with respect to different failure mechanisms.

However, the presence of biases, which will impair the opinions of experts, is a considerable drawback when calling on expert judgement in a risk analysis study. Consequently, our study proposed an approach for evaluating the failure probabilities of levees based on expert judgement and which includes the treatment of the latter.

The approach proposed permits eliciting, calibrating, aggregating and debiasing expert evaluations. The application of this methodological approach provided significant advantages:

- combining expert opinions based on calibration and entropy scores;
- estimating a quantitative uncertainty on the final failure probability obtained using an uncertainty interval;
- identifying the best expert elicitations and assigning a calibration weighting according to their pertinence and precision;
- treating the presence of over- and under-confidence bias quantitatively in order to obtain a final debiased failure probability.

At the end, the methods proposed use the expert judgement for estimating the levee failure probability (the *estimated or the subjective probability of failure*), constituting an expert estimate of the actual levee failure probability.

On the basis of this work, our research will be continued by testing our method to several dike case studies so as to continue validating the robustness of the approach. At the end, we plan to use the IeCAD approach in an operational framework to evaluate the probabilities of failure mechanisms (internal erosion and scouring) in risk analysis studies.

HATHOUT ET AL.

CIWEM Chartered Institution of Water and Environmental Management Journal of **Flood Risk Management**—**WILEY** 17 of 18

Otherwise, the approach developed for river levees can be adapted and applied to other areas where recourse to expert judgement is the only means available for obtaining usable information in the framework of carrying out a reliability analysis. In particular, the approach applied to levees can be extended to other hydraulic structures by making specific adjustments for each category of civil engineering structure, as small dams often poorly documented or large linear channels.

DATA AVAILABILITY STATEMENT

Data available on request from the authors: The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Michel Hathout https://orcid.org/0000-0001-8620-5047
Laurent Peyras https://orcid.org/0000-0002-4720-3572

REFERENCES

Apel, H., Annegret, H., Thieken, A. H., Merz, B., & Blöschl, G. (2006). A probabilistic modelling system for assessing flood risks. *Natural Hazards*, *38*(1–2), 79–100. https://doi.org/10.1007/s11069-005-8603-7

Bateson, G. (1979). *Mind and nature: A necessary unity*. New York, NY: Bantam Books.

CIRIA, MEDE & USACE. (2013). *The international levee handbook*. London, England: Author.

Clemen, R. T., & Lichtendahl, K. C. (2002), Debiasing expert overconfidence: A Bayesian calibration model. PSAM6, San Juan, Puerto Rico.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science* (p. 336). Oxford, NY: Oxford University Press.

FloodProbe. (2012). *Combining information for urban levee assessment*, Report Number: WP03-01-12-24. Retrieved from http://www.floodprobe.eu

Gouldby, B., Sayers, P., Mulet-Marti, J., Hassan, M., & Benwell, D. (2008). A methodology for regional-scale flood risk assessment. *Water Management*, *161*(3), 169–182.

Goulet, V., Jacques, M., & Pigeon, M. (2009). Modeling without data using expert opinion. *The R Journal*, *1*, 31–36.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty, heuristics and Biaises*. Cambridge, England: Cambridge University Press.

Knoeff, J., Vastenburg, E. W., Van den Ham, G. A., & Lopez de la Cruz, J. (2011). *Automated flood risk management*. Proceedings of the 5th International Conference on Flood Management (ICFM5), Japan.

Koehler, J., & Harvey, N. (2004). *Blackwell handbook of judgment & decision making* (p. 675). Oxford, England: Blackwell Publishing.

Kolen, B., Slomp, R., & Jonkman, S. N. (2013). The impacts of storm Xynthia February 27–28, 2010 in France: Lessons for flood risk management. *J Flood Risk Manag*, *6*, 261–278. https://doi.org/10.1111/jfr3.12011

Lichtendahl, K. C., Jr., Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, *59*, 1594–1611.

Mazzoleni, M., Barontini, S., Ranzi, R., & Brandimarte, L. (2014). An innovative probabilistic methodology for evaluating the reliability of discrete levee reaches owing to piping. *Journal of Hydrologic Engineering*, *20*, 04014067. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001055

Mériaux, P. & Royet, P. (2007). *Surveillance, maintenance and diagnosis of flood protection dikes: A Practical handbook for owners and operators*. Paris, France: Editions Quae.

Mouyeaux, A., Carvajal, C., Bressolette, P., Peyras, L., Breul, P., & Bacconnet, C. (2018). Probabilistic stability analysis of an earth dam by SFEM based on field data. *Computers and Geotechnics*, *101*(2018), 34–47. https://doi.org/10.1016/j.compgeo.2018.04.017

O'Hagan, A., Buck, C. E., Daneshkah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncetains judgments: Eliciting experts' probabilities* (p. 338). Chichester, England: John Wiley & Sons.

Ortiz, N. R., Wheeler, T. A., Breeding, R. J., Hora, S., Meyer, M. A., & Keeney, R. L. (1991). Use of expert judgment in NUREG-1150. *Nuclear Engineering and Design*, *126*, 313–331.

Peyras, L., Carvajal, C., Felix, H., Bacconnet, C., Royet, P., Becue, J. P., & Boissier, D. (2012). Probability-based assessment of dam safety using combined risk analysis and reliability methods: Application to hazards studies. *European Journal of Environmental and Civil Engineering*, *16*, 795–817. https://doi.org/10.1080/19648189.2012.672200

Peyras, L., Merckle, S., Royet, P., Bacconnet, C., & Ducroux, A. (2010). Study on a semi-probabilistic method for embankment hydraulic works – Application to sliding mechanism. *European Journal of Environmental and Civil Engineering*, *14*(5), 669–691.

Peyras, L., Royet, P., & Boissier, D. (2006). Dam ageing diagnosis and risk analysis: Development of methods to support expert judgment. *Canadian Geotechnical Journal*, *43*, 169–186.

Peyras, L., Tourment, R., Vuillet, M., Beullac, B., Delaunay, C., & Bambara, G. (2015). Development of an expert-led GIS-based approach for assessing the performance of river levees: The Digsure method and tool. *Journal of Flood Risk Management*, *10*, 393–407. https://doi.org/10.1111/jfr3.12178

Serre, D., Peyras, L., Curt, C., Boissier, D., & Diab, Y. (2007). Evaluation of hydraulic works in civil genius. *Canadian Geotechnical Journal*, *2007*(44), 1298–1313.

Serre, D., Peyras, L., Tourment, R., & Diab, Y. (2008). Levee performance assessment methods integrated in a GIS to support planning maintenance actions. *Journal of Infrastructure Systems—ASCE*, *14*(3), 201–213.

Shlyakhter, A. I., Kammen, D. M., Brodio, C. L., & Wilson, R. (1994). Quantifying the credibility of energy rojections from trends in past data: The U.S. energy sector. *Energy Policy*, 119—130.

Simm, J. D., Wallis M., Smith, P., Deniaud, Y., Tourment, R., Veylon, G., Durand, E., McVicker, J., Hersh-Burdick, R., & Glerum, J. (2012), *The significance of failure modes in the design and management of levees—A perspective from the International Levee Handbook team*. FloodRisk Rotterdam, Netherlands, November 20–22, 2012.

Soll, J. B., & Klayman, J. (2003). Overconfidence in intervalle estimates. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30, 299–314.

Tourment, R. (2018). Small dams and levees—General report, *ICOLD 26th Congress on large dams, GR-Q. 103*, Vienna, July 4–6, 2018. CRC Press/Balkema, pp. 115-159

Van der Meij, R., Van den Ham, G.A., Morris, M., Lhomme, J., Tourment, R., & Maurel, P. (2012). *Combining different data sources for assessment of urban flood defences*. Floodrisk Rotterdam Netherlands, November 20–22, 2012.

Vick, S. G. (2002). *Degrees of belief: Subjective probability and engineering judgment* (p. 455). New York, NY: American Society of Civil Engineers.

Vorogushyn, S., Merz, B., Lindenschmidt, K.-E., & Apel, H. (2010). A new methodology for flood hazard assessment considering dike breaches. *Water Resources Research*, 46, 1–17. https://doi.org/10.1029/2009WR008475

Vrijling, J. K. (2001). Probabilistic design of water defense systems in The Netherlands. *Reliability Engineering and System Safety*, 74, 337–344. https://doi.org/10.1016/S0951-8320(01)00082-5

Vuillet, M., Peyras, L., Carvajal, C., Serre, D., & Diab, Y. (2013). Levees performance evaluation based on subjective probability. *European Journal of Environment and Civil Engineering*, 17(5), 329–349.

Yachanin, S. A., & Tweney, R. D. (1982). The effect of the thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychometric Society*, 19, 87–90.