Expert judgment-based reliability analysis of the Dutch flood defense system

Rongen, G.; Morales Napoles, O.; Kok, M.

DOI
10.1016/j.ress.2022.108535

Publication date
2022

Document Version
Final published version

Published in
Reliability Engineering and System Safety

Citation (APA)
Rongen, G., Morales Napoles, O., & Kok, M. (2022). Expert judgment-based reliability analysis of the Dutch flood defense system. *Reliability Engineering and System Safety*, *224*, Article 108535. https://doi.org/10.1016/j.ress.2022.108535

Reliability Engineering and System Safety 224 (2022) 108535

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Expert judgment-based reliability analysis of the Dutch flood defense system

G. Rongen [a,b,*], O. Morales-Nápoles [a], M. Kok [a,b]

[a] Civil Engineering and Geosciences, Delft University of Technology, The Netherlands
[b] HKV consultants, The Netherlands

ARTICLE INFO

Keywords:
Flood risk
Reliability analysis
Expert judgment
Cooke's method

ABSTRACT

The Wettelijk Beoordelingsinstrumentarium (WBI) is the legal set of instruments for flood risk analysis in the Netherlands. Often, engineers have the impression that some failure probabilities of flood defenses resulting from these instruments are overestimated. In an effort to better estimate the failure probabilities of dikes along the Dutch river Rhine, this study sets out to assess them with experts and compare them to model results. We used Cooke's method for combining experts' estimates in a structured way and follow two approaches to estimate a system failure probability. In the first approach, experts estimate discharges that lead to at least one dike failure. This gives plausible results; failure probabilities between 1/30 and 1/17.000 in a year. The second approach is based on adjusting existing model-based assessment results, by estimating the model-bias and incorporating additional dependencies. This mostly leads to large, implausible, failure probabilities: Experts tend to give more conservative answers as they are asked for detailed estimates without clear reference values. This results in large uncertainty and consequently (too) high failure probabilities. Our research shows that when applied in a clear frame of reference, structured expert judgments can be successfully used for estimating the reliability of Dutch flood defenses.

1. Introduction

A large part of the Netherlands would regularly flood if it were not protected by dikes. To prevent this, the Dutch have been building and maintaining dikes for centuries. Nevertheless, rivers and sea have flooded the land dozens of times over the past centuries. At present, dikes must have a failure probability lower than 1/100 to 1/30,000 in a year to be considered safe according to law [1]. Such low probabilities correspond to breach events during extreme river discharges or storm events that have never been measured in recent history, causing a lack of empirical data to verify the calculation. Therefore, modeling is needed to estimate this type of extreme flooding events and the possible resulting dike failures.

Using probabilistic methods in the design and assessment of flood defenses is common practice in the Netherlands (see [2] for a historical account and [3] for a recent example of the use of the traditional Dutch standard). The Wettelijk Beoordelingsinstrumentarium (WBI) is the current official set of models and tools for flood risk assessment [4]. The methods give plausible reliability assessments for a small section of dike. However, if applied to all dikes in the Netherlands the combined failure probabilities are overestimated. If such probabilities were correct, the Netherlands would have a dike breach every few years [5], while we know from experience that this is not the case. Parts of the WBI assessment method are therefore likely inaccurate. If this is ignored and the results are treated as correct, major (and probably economically unfeasible) dike improvements throughout the Netherlands would be enforced since the Dutch safety standards is based on an optimal economic safety target [6–8]. Hence, overestimating failure probabilities leads to unnecessary spending of public money.

This study sets out to derive more plausible dike failure probabilities with the aid of experts in the field of flood risk management. We follow two approaches to estimate the system failure probability, both based on estimating fragility functions or fragility curves [9,10]. These are widely used in flood risk assessments [11,12]. The first approach directly collects the experts' estimates of dike failures on the system-level (i.e., the probability that at least one dike fails) to derive more plausible estimate of the system failure probability than currently follows from the WBI. The second, more detailed, approach is based on collecting detailed estimates of failure mechanisms on dike level, combining these with model-based results, and taking load interdependencies into account. From this, we try to determine how well experts can estimate such probabilities. Additionally, it can indicate the causes of the implausible results following from the WBI-method. The two approaches discussed in this research (direct and more detailed) relate to research on risk estimates from fault trees with less and more detail as discussed for example in [13,14].

* Corresponding author at: Civil Engineering and Geosciences, Delft University of Technology, The Netherlands.
E-mail address: g.w.f.rongen@tudelft.nl (G. Rongen).

https://doi.org/10.1016/j.ress.2022.108535
Received 27 December 2021; Received in revised form 1 April 2022; Accepted 15 April 2022
Available online 28 April 2022

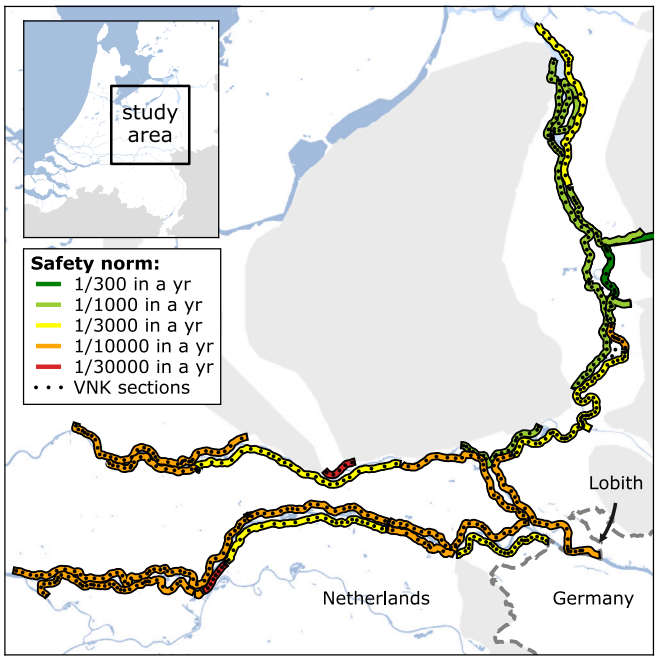

Fig. 1. Overview of the part of the river Rhine considered in this study. The colored lines indicate the dike sections with their norm, while the black dots represent the smaller VNK sections.

To account for interdependencies in hydraulic loads, relevant studies often rely on Monte Carlo-based reliability methods [e.g.,13–16]. In this study, we assume dependence in load and independence in dike strength, a conceivable dependence structure that fits the water system and available data. To ensure scrutability, empirical control, neutrality, and fairness, we applied structured expert judgment with Cooke's method [17,18]. A similar study to estimate reliability of French flood defenses was conducted by [19]. The focus in this study however lies on the system scale rather than the reliability of individual dikes.

The area under investigation is the part of the Dutch river Rhine without tidal or lake level influence. This concerns the river branches Waal, Nederrijn, Lek, Pannerdensch Kanaal and IJssel, as shown in Fig. 1. The system consists of 28 dike trajectories, for which the color indicates the required safety level or norm. This is the maximum allowable probability of failure of the trajectory in a year. The 525 black dots indicate the dike sections as used in *The Flood Risk in the Netherlands* project (VNK2) [20,21].

2. Method for dike safety assessments with expert estimates

First, we briefly explain the dike failure mechanisms considered in this study. Second, we present the current WBI method for assessing the safety of the Dutch hydraulic defense system. Then, we present the approach for incorporating expert judgment in this. Next, Cooke's method for expert judgment, and finally, an overview of the questionnaire is given.

2.1. Failure mechanisms

In this study the failure mechanisms overtopping and overflow, piping, and macro-stability are assessed. These are the most important mechanisms for the river Rhine, based on the failure probabilities calculated in VNK2. [22] gives a description of these failure mechanisms. The WBI has a model for calculating failure probability for each of these mechanisms. The piping failure probability is calculated with the adjusted model of Sellmeijer [23], and the macro-stability failure probability with D-Stability [24]. [25] gives a description of the method for overtopping and overflow, while the critical overtopping discharges are derived from [26].

2.2. Current WBI-model for a dike section's failure probability

A dike in the Netherlands has a safety norm, that is, a maximum allowable failure probability in a one-year period. A *section* consists of several dike stretches, that together must meet this standard. A stretch of dike with comparable properties is a dike *segment*, which can fail according to various failure mechanisms. Ultimately, the combined failure probability of all dike segments and all failure mechanisms, calculated with the WBI, must meet the safety level for the section.

The failure probability of a dike section is a function of the failure probabilities of the separate segments:

$$P_{sc}(Z<0) = \min\left(N \cdot \max_i(P_{sg,i}(Z<0)); 1-\prod_i^n (1-P_{sg,i})\right) \tag{1}$$

In which $P_{sc}(Z<0)$ is the section failure probability, Z is the limit state function (below zero is failure) and $P_{sg,i}(Z<0)$ is the failure probability of segment i. For conciseness, we write $P(Z<0)$ as P_f. Notice that the second argument in the $\min(\cdot)$ function assumes independence between segment failures. N is a measure of the spatial dependence. Considering that spatial correlation generally decreases over distance, a representative length can be derived for which a dike segment is still considered independent. N can then loosely be explained as the number of times that this representative length fits the length of the section. If N is 1, all segments are dependent. If N is large, the segments are combined independently.

The failure probability of a segment i is calculated as the independent combination of the separate failure mechanisms:

$$\begin{aligned} P_{f,sg} &= 1-\prod_{j=1}^{3}\left(1-P_{f,sg,j}\right) \\ &= 1-(1-P_{f,sg,pip})\cdot(1-P_{f,sg,ms})\cdot(1-P_{f,sg,oo}) \end{aligned} \tag{2}$$

In which j denotes the failure mechanisms, piping (pip), macro-stability (ms), and overtoppping and overflow (oo). The failure probability is typically calculated per event, which for the rivers is a single flood wave. Standard procedure in the Netherlands is to consider 6 months per year for which in each month one flood wave can occur [25], so the maximum exceedance frequency of the limit state is six times per year. Determining the failure probability for a single segment (i.e., one of the parts in Eq. (2)) is challenging. Especially for the geotechnical mechanisms piping and macro-stability, for which the missing knowledge of the subsoil and saturation causes large uncertainty.

The high water levels in the study area are caused by the same high discharge at Lobith [27]. This simplifies the dependencies: Given the peak discharge at Lobith, the peak water levels in the whole of the river system are known to a large degree. We simplify this a bit further, by assuming that the local water levels are completely dependent on the discharge at Lobith.

A comprehensible way of expressing the relation between load and (conditional) failure probability is with a fragility curve, which expresses the failure probability given the load [10]. A load can be, for example, water level, wind speed, discharge, or a combination of these (in which case we would have a fragility (hyper)plane [12]). In case of a peak discharge k, the failure probability can be expressed as:

$$P(Z<0) = \int_{k=0}^{k=\infty} f(k)P(Z<0|k)dk \tag{3}$$

In which $f(k)$ is the probability density of the peak discharge in a year. By integrating the conditional exceedance probabilities (i.e., fragility curves) with the probability density of the water level, the total failure probability in a year is calculated. Examples of fragility curves are given in Fig. 2. The blue lines are curves with large uncertainty about the discharge that leads to failure, the red curves represent small uncertainty

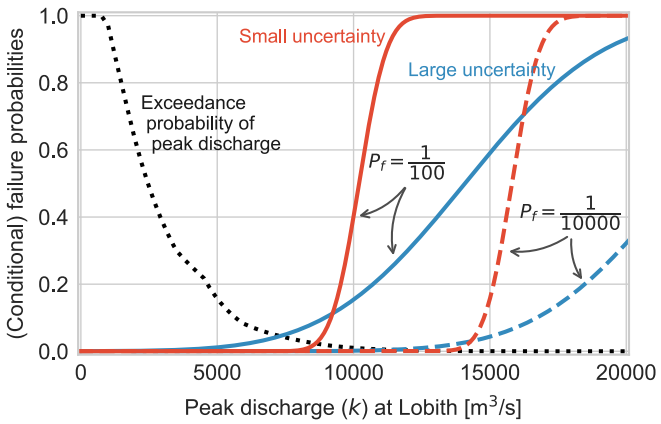

Fig. 2. Examples of fragility curves (the red and blue lines) that represent different failure probabilities and degrees of uncertainty.

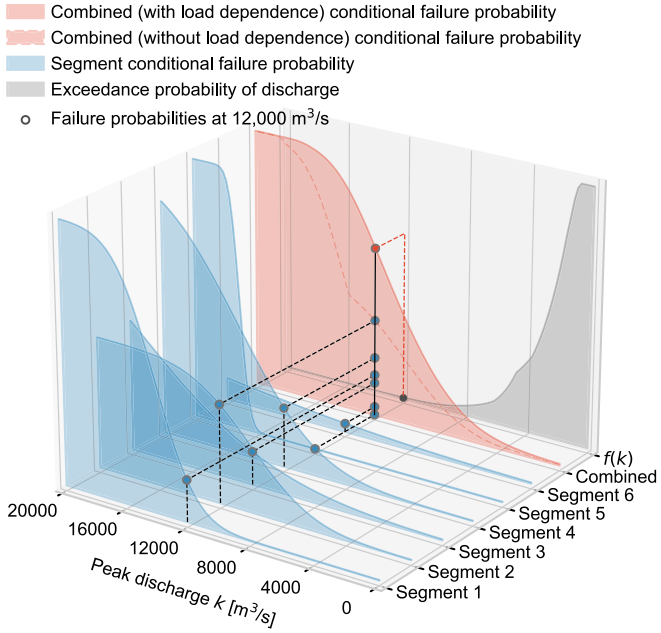

Fig. 3. Illustration of the method for combining failure probabilities conditional to the load. The red solid curve is the result for considering *strength* independent, the dashed red curve for considering *strength* dependent. The total failure probability results from integrating the conditional failure probabilities with the discharge statistics (indicated with the black curve).

on this. Because the exceedance probability of more extreme events often decreases more or less exponentially, the lower tail of the fragility is the most relevant factor that determines the total failure probability. The fragility curve is used extensively in this research because it can be related to experts' uncertainty estimates. The answer to the question: "At what discharge does dike X fail by mechanism Y with a 50% probability?", is similar to, "at what k is $P_X(Z_Y < 0|k) = 0.5$?". This is a point on the curve in Fig. 2. Additionally, by expressing a failure probability in the form of Eq. (3), the dependence of water levels to the discharge at Lobith can be taken into account as well, by combining the failure probabilities conditional to k. This is illustrated in Fig. 3, where the combination of all blue dots at 12,000 m^3/s (conditional failure probabilities) are independently combined to the red dot, which is the combined failure probability conditional to the peak discharge k. The red line itself is also a fragility curve. Using the dependence between water levels along the river (through the peak discharge at Lobith) is not standard procedure in the WBI-method. However, we added it in this study because it results in a more accurate description of dike failure that corresponds better to the fragility curves assessed by the experts.

2.3. Dike assessment with experts

The two approaches we follow for estimating failure probabilities with experts are both based on finding the term $P(Z < 0|k)$ in Eq. (3):

1. In the first approach, we ask experts the *system* conditional failure probability, $P_{sys}(Z < 0|k)$, directly. This means asking the experts at what discharges they believe a dike would fail, with specified probabilities. Integrating with the discharges as shown in Eq. (3) then gives the total failure probability. We used the standard WBI exceedance probabilities of the peak discharges at Lobith in this study; they are not an elicited quantity in this study.
2. In the second approach, we ask experts about several aspects of the safety assessment on a dike scale. These are the bias in the model-based failure probability calculations per failure mechanisms, and the uncertainty in the relation between load and failure probability. These estimates can then be used to modify existing model outcomes and combine this to the term $P_{sys}(Z < 0|k)$ from which the total failure probability is calculated.

The first option is relatively easy for the experts to relate to experience but gives little information about failure probabilities on the dike level. This information is more readily available from the second option, but it is hard for the experts to relate this to a total failure probability of the system. Therefore, we use the first approach to derive an improved (from WBI) probability estimate, and the second approach to assess the values of these estimates and determine what causes potential differences.

For the second approach an estimate of the failure probability for each dike segment (represented by one of the black dots in Fig. 1) is needed. Recall that the study area contains 525 of such segments. Eliciting a failure probability from each expert for each of these segments would be unrealistic within this study. Therefore, we derive the individual failure probabilities from two reference sources:

- For the current dike safety, these are the failure probabilities from VNK2 [20]. This project was finished in 2015 and considered the near future system layout (including the river engineering works 'Room for the River'), making it a useful source for the assessment of dike safety at present.
- The dike safety levels required by law (norms or standards). When a dike is designed by the WBI-method to meet this standard, we know that the failure probability of a trajectory should be lower than the safety standard (when assessed with the same tools with which it was designed.)

The total failure probability can then be calculated by updating them based on expert judgments and combining them into the system failure probability. This is expressed in the following equation:

$$P_{sys}(Z<0) = \int_{k=0}^{\infty} f(k)\left(1 - \prod_{i=1}^{N_{sg}} \prod_{j=1}^{3} \{P_{ij}(Z<0|k-\Delta k_{ij})\}\right) dk \quad (4)$$

in which j denoted one of the three failure mechanisms. Note that the load-reducing effect of the first breach on the next breaches (a relevant topic for dike reliability [13]) is not taken into account, because the system has failed after the first breach. The experts' uncertainty in the discharge at which the dike fails determines how much $P(Z < 0|k)$ varies for different values of the peak discharge k. The estimated bias in the failure mechanism shows whether failure takes place at consistently higher or lower discharges. It is determined by comparing the failure

probability from the experts estimate with the model results. The quantity Δk_{ij} is calibrated such that the difference in failure probability is matched after integrating. Referring back to Fig. 2, the uncertainty expressed by the rate of change in P_{ij} for variations in K, indicates whether the fragility curve behaves more like the red one or blue one in Fig. 2. The bias (Δk_{ij}) indicates whether it is one of the solid curves or dashed curves in the same figure.

In this study, dike properties are assumed to be constant during an event. Therefore, we only need to consider the highest water level in the event to determine the failure probability for piping and stability. For overtopping this is different since wind-generated waves play a key role. The wind conditions are much more volatile than the water levels, so they are considered on a smaller time scale. This time scale is chosen to be 12 h, as explained in [25]. The failure probability given a peak discharge k is then calculated with:

$$\begin{aligned} P(Z_{ov} < 0|k) &= P_{30\mathrm{d}}(H_{ov} > h_{ov}|k) \\ &= 1 - \prod_{t=1}^{N} \left[1 - P_{12\mathrm{h},t}\{H_{ov} > h_{ov}|q(t|k)\}\right] \end{aligned} \tag{5}$$

Where $P_{30\mathrm{d}}(H_{ov} > h_{ov}|k)$ is the exceedance probability of the critical overtopping discharge h_{ov} in the 30-day base duration related to peak discharge k. $P_{12\mathrm{h},t}(H_{ov} > h_{ov}|q(t|k))$ is the exceedance probability of this quantity in a 12-hour window t which depends on the average river discharge during these 12 h $q(t|k)$. The 12-hour average river discharge $q(t|k)$ is a function of a standardized hydrograph shape and peak discharge k. The conditional failure probability in Eq. (5) can be substituted in Eq. (4) as one of the failure mechanisms j. Compensating for a bias is not necessary, as the expert's assessment for the critical overtopping distribution H_{ov} can be estimated directly. For a detailed explanation of this calculation, we refer to [25]. Note that we assume the wind conditions to be spatially uniform in this study.

2.4. Structured expert judgment with Cooke's method

In this research we use Cooke's method for structured expert judgment. Cooke's model is a method for eliciting and combining expert judgments based on empirical control, with the aim to reach rational consensus. This approach is extensively described in [18], here we discuss some of the basic elements of the method.

Cooke's method assigns a weight to each participating expert. The expert makes an uncertainty estimate for each question by estimating several percentiles. Often these are the 5th, 50th and 95th percentile. In this study we also elicit the 1st and 25th percentile for several questions because the failure probabilities during frequent events dominate the total failure probability. The expert weight is calculated by multiplying the *calibration* and *information* scores if the calibration score is above a particular threshold. The calibration score is calculated from the questions for which the answer is known by the researcher, but not by the participants at the moment of the elicitation. These are referred to as seed or calibration variables. Calibration is a measure of the statistical accuracy of the expert. The information score is calculated from all questions and is a measure of the degree of uncertainty of the expert's answer.

The weighted combination of the uncertainty estimates of individual experts are called decision makers (DMs). For each expert e and each item i, a probability density function (PDF) $f_{e,i}$ is calculated from the estimated percentiles. The DM is then the weighted sum of all experts PDFs, or:

$$DM(i) = \sum_{e} w(e) f_{e,i} \Big/ \sum_{e} w(e) \tag{6}$$

This specific DM is called the global DM, because the weight factor is based on all seed variables. Another variant is the DM based on item weights, by changing $w(e)$ in Eq. (6) for $w(e,i)$. In this case, experts that answer more confidently on a certain item, have a larger weight in the decision maker for this answer, provided their calibration score is sufficiently large. Since the DM is a mixture of all experts PDFs it also has an uncertainty estimate itself, and therefore also a weight. This weight can be optimized by excluding experts based on the calibration score threshold. Finally, it is also possible to give all experts equal weight or let the group assign weights according to other criteria.

The Experts' assessments were processed with ANDURYL v1.2 [28]. ANDURYL is a modern implementation of EXCALIBUR, a program designed for processing expert judgments within the framework of Cooke's model. While the mathematical base is the same (i.e., Cooke's method), ANDURYL aims to stimulate further developments and usage, by using Python and making the code publicly available. In this study in particular, items 13–22 ask for 2 more percentiles than items 1–12 (see Table A.2). Dealing with such cases in EXCALIBUR is not possible. Another example of the limitations of EXCALIBUR is Fig. 4 where up to 4 experts or items are excluded at the time to investigate robustness while EXCALIBUR supports only leaving one out at the time. For a detailed comparison, the reader is referred to [28–30].

2.5. Questionnaire, experts, and practical set-up

The conducted questionnaire contains 22 questions. Table A.2 in Appendix shows these questions. The questionnaire itself also had a detailed problem description per question. The 22 questions are composed as follows: First, items 1 to 10 are the seed items from which the calibration scores are calculated. These items are based on data that existed at the moment of elicitation, or shortly after but before processing the results (i.e., a short-term prediction). Second, items 11 and 12 are used to determine $P_{12\mathrm{h}}(h_{ov} > H_{ov})$ from Eq. (5). Then, 13 to 16 are used to determine the bias for the piping and macro-stability models (Δk_{ij} in Eq. (4)). Next, items 17 to 20 are used to determine the uncertainty in the load at which failure occurs (the rate of increase in $P(Z < 0|k)$ for changing k). Finally, items 21 and 22 are used to 'directly' assess the failure probability on system level.

13 experts participated in the elicitation. Table A.3 in the supplementary information lists their names, professional interests, and expertise. Most experts are flood risk advisor, affiliated to national and regional government water authorities, universities, research institutes, consultancy firms and independent consultants. Their expertise ranges from geotechnical engineering (related for example to piping and stability), to revetments (overtopping), and hydraulic loads (discharges). The questionnaire was refined based on two dry runs with two different experts that have a similar background as the 13 that took part in the elicitation. The expert session was organized with video conferencing, due to COVID-19 pandemic restrictions. Two half-day sessions were organized, with two occasions for each session. The first session contained a presentation of the study topic and Cooke's method, followed by answering the questionnaire individually. Experts were able to ask questions during the elicitation, which were then discussed plenary. The second session was used to present and discuss the results. Experts had the opportunity to change their assessments in case they misinterpreted the question, to get their best assessment on the target items. Expert D used this possibility to change two seed items as well. Because the expert had seen the answers (realizations) at this point, the answering was changed to 'not answered,' even though the questions were clearly wrongly interpreted. This led to slightly higher calibration scores for the other experts as well, as the minimum number of answered seed questions by all is used in calculating the calibration score (see [18] for details, N is 2 less). Expert F chose to withdraw from the expert session before filling in the target questions. This expert is therefore removed from the results, leaving twelve experts' assessments. An overview of the experts' and DMs' estimates for all items, is shown in Figs. A.13 and A.14.

Table 1
Calibration and information scores form Cooke's method. The bottom three rows show the decision makers.

	Calibration score	Information score		Weight
		All	Calibr.	
Exp A	0.000304	0.923	0.696	0.000212
Exp B	0.057	1.000	0.980	0.0559
Exp C	0.00248	1.390	1.040	0.00258
Exp D	0.664	0.503	0.564	0.374
Exp E	0.64	0.664	0.777	0.497
Exp G	0.117	0.954	1.090	0.127
Exp H	0.0368	1.400	1.160	0.0428
Exp J	0.121	1.490	1.400	0.169
Exp K	0.0196	1.270	1.840	0.0362
Exp L	0.117	1.710	1.720	0.201
Exp M	1.02×10^{-5}	1.680	1.610	1.65×10^{-5}
Exp N	1.35×10^{-6}	1.760	2.030	2.73×10^{-6}
Global	0.64	0.365	0.399	0.255
Item	0.64	0.464	0.479	0.306
Equal	0.571	0.313	0.302	0.172

3. Expert estimates and resulting failure probabilities

We present the results in four parts. The first is about the expert scores for Cooke's method, which indicate the experts' performance and quality of the elicitation. The second part shows the experts' estimates of the failure probability at the system level, while the third part shows the estimates on a dike level. From this, we calculate the system failure probability, which is compared with the estimated failure probability in the fourth part.

3.1. Results for Cooke's method

The elicitation results for Cooke's method are shown in Table 1. It shows the score for statistical accuracy (calibration score) and the information score. Looking at the calibration scores, expert E and D have a high score (≥ 0.5), experts J, G, L and B a significant score (≥ 0.05) and the rest a score below the significance level. The information scores show less variation, as is usual with Cooke's method. Note that the information scores for all items (column 2) and the seed questions (column 3) are similar, indicating that experts have answered similarly for both categories of questions. Note that experts that have a high calibration score tend to have a lower information score, and vice versa. The weight is the product of information and calibration score, and therefore favors experts with a high calibration score.

We calculated three decision makers: global, item and equal weights. The calibration scores for the three decision makers are all high. The information scores for the decision makers are lower than those of individual experts. This is because the DM estimates are a mixture of all experts, to the extent of the weight factors. Therefore, it is often a much wider distribution than for the individual estimates. This results in a relatively low weight for the DM. When optimizing the significance level, only expert D with the highest calibration score remains. This is both the case for the global and item weights. As this does not provide additional information, the global and item DMs are calculated for all experts without optimization.

The robustness of the results for excluding experts and items are calculated for the Global DM. The variation in information and calibration scores for removing up to four items (experts) is shown in Fig. 4. Fig. 4 shows that the results are stable for excluding items. The DM is most sensitive to item 7 about piping discharge, or the judgments of expert D. Removing either, leads to the largest reduction in DM weights. Notice however that even when excluding four items or experts at the same time, the calibration score of the DM remains greater than 0.18.

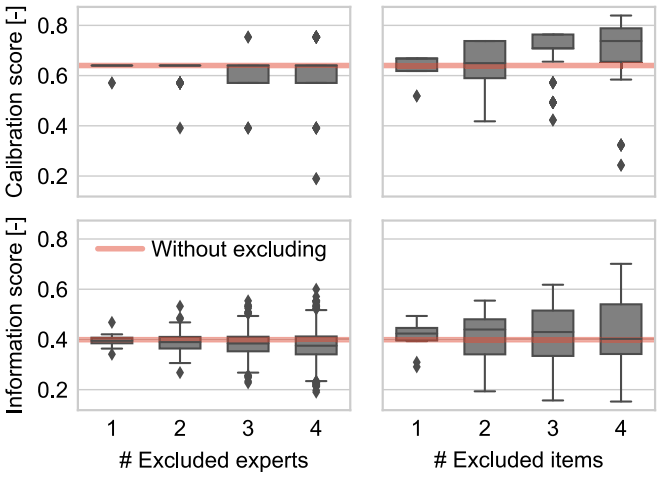

Fig. 4. Sensitivity of calibration (top) and information score (bottom) for excluding experts (left) and items (right).

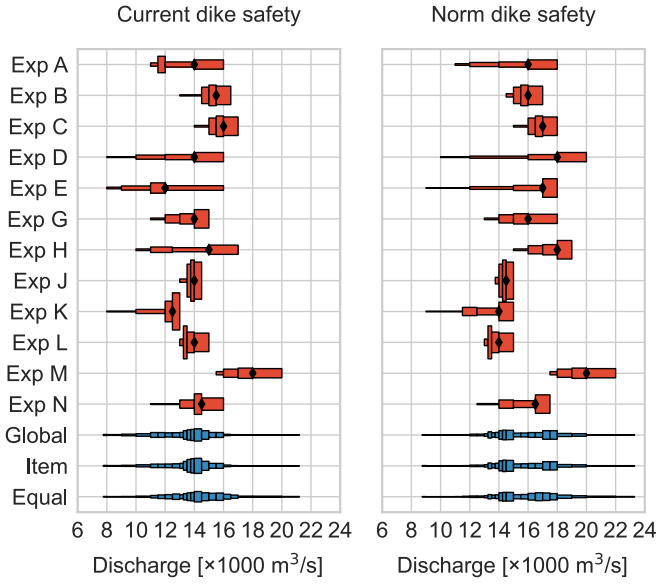

Fig. 5. PDFs of the discharge that causes at least a single dike failure in the system, under current dike conditions (left) and norm safety conditions (right). The experts' estimates are shown in red, the DMs' in blue.

3.2. System failure probability

The experts estimated the discharge for which at least one dike in the system fails. This question is part of the first approach, in which we estimate the system failure probability directly. Reference values for discharges, such as the 1995 extreme river discharge (around 12,000 m^3/s) and the design discharge of Room for the River (around 16,000 m^3/s) were discussed during the expert elicitation. Fig. 5 shows the experts' and DMs' estimates. They are expressed as probability densities in between the assessed percentiles (i.e., 1, 5, 25, 50, 95). The thicker the square, the higher the probability density in that range. The black diamond indicates the 50th percentile. As experts assessed more percentiles below the 50th, the distributions tend to be more asymmetrical. The DM results show even more percentiles (are more irregular) because they are a weighted combination of more than one expert. Therefore, they are drawn in greater detail. For the current situation, most experts estimate at least one dike to fail at a discharge in between 12,000 and 16,000 m^3/s, while for the dikes matching the norm, this is in between 14,000 and 18,000 m^3/s.

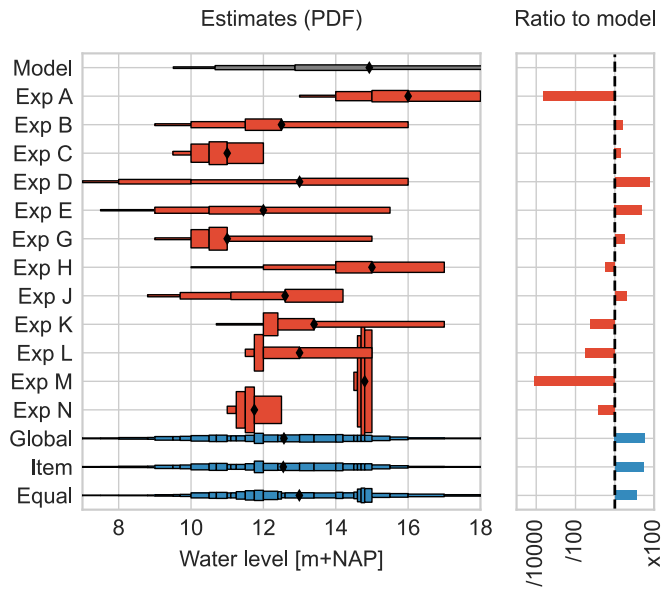

Fig. 6. PDFs of the water level at which the considered schematization is expected to fail due to piping, according to the model (gray), the experts' (red) and the DMs' (blue). On the right the ratio between expert or DM, and the model outcome is shown.

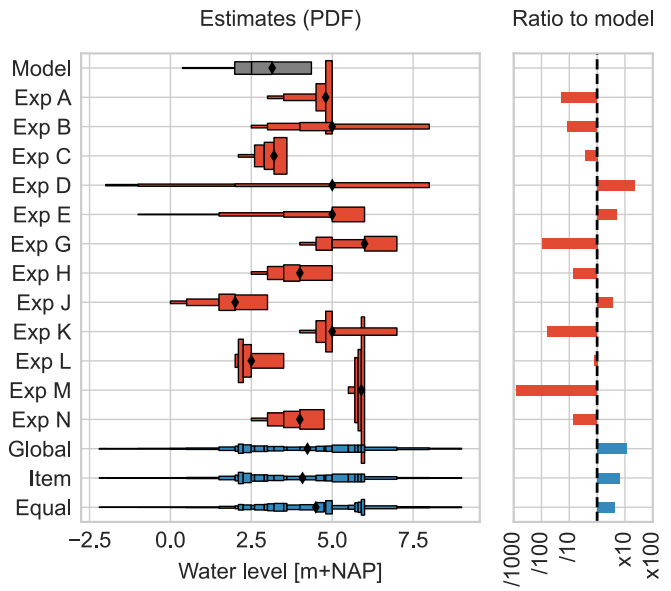

Fig. 7. PDFs of the water level at which the considered schematization is expected to fail due to stability, according to the model (gray), the experts' (red) and the DMs' (blue). On the right the ratio between expert or DM, and the model outcome is shown.

3.3. Failure mechanism factors

The experts also estimated the bias in the models for the relevant failure mechanisms (i.e., the Δk_{ij} in Eq. (4)). Together with the estimates of the general uncertainty 3.3.4, these are used for the second approach, in which the failure probability is calculated. Experts assessed the conditions that lead to the failure definition from WBI, as well as the conditions for reaching a breach. The former expresses the bias in the model, while the latter expresses the bias in the probability of flooding (assumed to be the direct result of a breach). Two important assumptions for answering the questions are: (i) that the high water level lasts for 5 days, and (ii) no emergency measures (e.g., sandbags) are used to prevent failure.

3.3.1. Piping

We chose a single, representative, schematization to estimate the bias in the 'adjusted Sellmeijer' model [23]. For this, we selected a schematization with average characteristics from the VNK2-dataset. Then, the failure probability was calculated, through the experts estimating the water level at which they expect a breach to occur with a 1, 5, 25, 50, and 95% chance (implicitly constructing a fragility curve). The failure probability is calculated by combining this with the water level statistics. The results of this are shown in Fig. 6. The model result is a continuous distribution of the water level at which the dike is expected to fail. However, for comparison to the experts' estimates, it is drawn for the same percentiles. The right side of the figure shows the factor between the failure probability from the adjusted Sellmeijer model, 9.31×10^{-4}, and the failure probability from the expert. Note that the two experts with highest calibration scores, expert D and E, estimate a 10 to 100 times higher failure probability. This is due to the uncertainty in their estimates, which leads to a substantial conditional failure probability at water levels that occur often. The Global and Item DM are similar to D's and E's estimates, as these experts have by far the largest contribution to the DM. Other experts, like M, only contribute to any discernible extent within the equal weights DM.

3.3.2. Macro-stability

The bias for the D-stability model (for macro-stability assessments) is determined in an equivalent way to the piping bias. The results are shown in Fig. 7. The model for this failure mechanism presents less variation in failure-critical water level compared to the expert assessments. Most experts estimate the model to be conservative, which surprisingly was not the case for piping. Again, experts D and E estimate a more conservative outcome than the model, and again this results in assigning a significant failure probability to water levels that occur often. The estimates for the macro-stability failure water level show less variation than those for piping. The plenary discussion showed that it was easier to narrow the relevant water level range for macro-stability than it was for piping, since experts find it easier to indicate referential points with a small or large failure probability. However, the ability to do so depends largely on the background of the expert.

3.3.3. Overtopping

In addition to the water level at the dike, waves can cause failure due to overflow and overtopping. Therefore, instead of the water levels, the distribution of the critical overtopping discharge is elicited $P(H_{ov} > h_{ov})$, which was then used to determine the dike failure probability directly. Fig. 8 shows the results. The right graph gives the difference in failure probability compared to the model results, for all dike sections. According to eight of the experts, the failure probabilities from the WBI are too unconservative, for four experts they are conservative. The difference between the failure probabilities is small, most factors are between 1/5 and 2. That is because the overtopping or overflow discharge increases rapidly with rising water levels: the difference between a low and high discharge estimate is whether the dike fails at a water level just below the crest (with some wave attack), or at a water level just above the crest. In the study area, the exceedance frequencies of these two situations are not that far apart.

3.3.4. General uncertainty in water level at failure

To calculate the total system failure probability in the second approach, we need an overall estimate of the distribution of failure probabilities conditional on the discharge (i.e., the $P(Z < 0|k)$-part in Eq. (4)). The results of the experts' assessments for this are shown in Fig. 9. During high water, a discharge (at Lobith) difference of 1,000 m^3/s typically leads to a 25 cm difference in local water level. Extrapolating this, a range of 4 meters gives a 16,000 m^3/s discharge difference. This is a broad range but note that it spans the range between a very small (1%) and large failure (95%) probability.

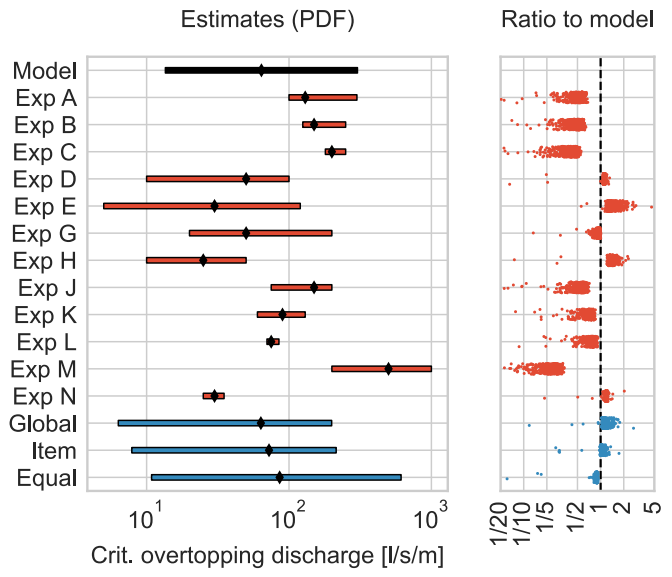

Fig. 8. Experts' estimates for the critical overtopping discharge that causes a breach when exposed for a 6-hour duration. The model result is derived from the log-normal distribution for H_s in [0 - 1] m and an open grass quality. Because of the log-scale, the bar heights are not scaled to the probability density. The ratios on the right are shown as scatters, with each dot representing the ratio for a single dike section.

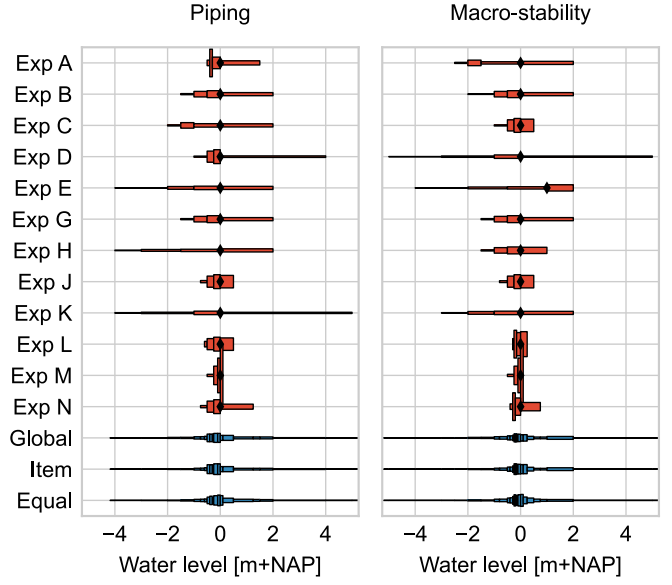

Fig. 9. Uncertainty in the water level at which a dike fails for piping (left) and macro-stability (right), according to the experts (red) or DMs (blue).

The differences between the experts' estimates are, again, large. Experts L, M and *N* give a very small, almost deterministic, range where failure will occur. Experts D, E and K on the contrary are very uncertain about the water levels at which a dike will fail due to piping or stability.

3.4. Comparing the failure probabilities from the two approaches

Recall that we estimated system failure probability in two ways: by integrating the estimates on system level (Section 3.2) with the discharge statistics, and by combining the estimates on dike level (bias, and uncertainty, Section 3.3) with model-based results through Eq. (4). Fig. 10 shows the resulting system failure probabilities for the current dike safety level, and Fig. 11 for the situation in which all dikes match the norm.

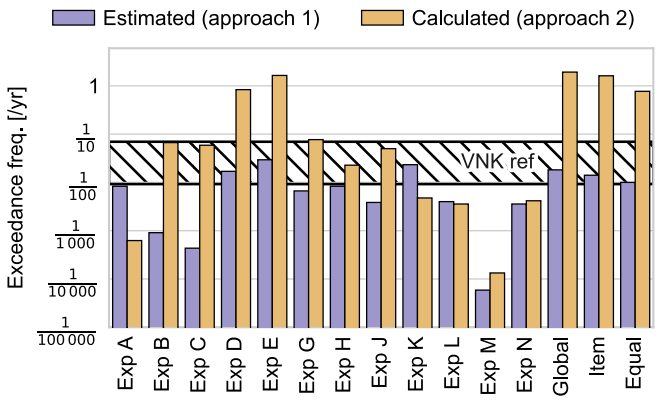

Fig. 10. Estimated and calculated system failure probabilities, for *current dike safety*.

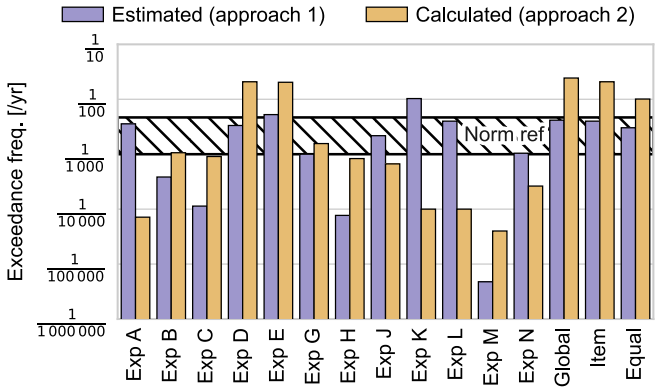

Fig. 11. Estimated and calculated system failure probabilities, *when the dikes match the required safety level.*

Each of the figures contains two bars per expert or DM. The first represents the failure probability from the first approach (i.e., the system failure estimates), the second bar the failure probability for the second approach (i.e., updating model results with estimates). The figures also contain a horizontal hatched area that indicates the range of the reference failure probability. For the current dike safety situation, this is the range between the independent and dependent combination of the VNK2 section failure probabilities: 1/15 to 1/100 in a year. The reference value for the norm safety is the range between the independent and dependent combination of the individual dike safety levels: 1/215 to 1/1000 in a year. Note that we combined the 'signal values' that are about a factor 3 lower than the values displayed in Fig. 1. This is the dike safety level that signals a dike should be reinforced soon [1], so a dike will be reinforced to at least this level (as its reliability constantly decreases due to deterioration [31]).

All experts estimate (approach 1) a slightly smaller to much smaller failure probability, while the calculated answer (approach 2) is significantly higher, ranging from 1/10,000 in a year to larger that once in a year. For the situation in which the dikes have a safety level up to the norm, the estimates are in the range of 1/100 to 1/10,000 in a year, while the calculated answers are in the range 1/10 to 1/17,000 in a year. Looking at the estimated failure probabilities, we see that, for the current dike safety, experts assess the system failure probability to be roughly 1/100 in a year (ranging from 1/30 to 1/30,000). For the norm safety situations, experts estimate the failure probability to be 1/1,000 year (ranging from 1/100 to 1/200,000).

The differences between the two approaches for an expert or DM can be quite large, up to a factor 100 in failure probability. In most cases the calculated failure probability (approach 2) is larger than the estimated one (approach 1), especially for the current dike safety situation. The

main cause of this difference is the uncertainty in the experts' estimates for the failure-critical water level for piping and stability. Uncertain estimates contribute a relatively large probability to frequent events. The exceedance probability for higher water levels decreases rapidly (more or less exponentially). This means higher conditional failure probabilities for low water levels contribute much more to the total failure probability. Experts that estimate a wide range of water levels for the different quantiles, more likely get an implausible system failure probability. Note that the experts did not receive feedback on the implication of their answers during the expert session, as the goal of approach 2 was to compare the results of the different elicitations.

4. Applicability for estimating dike failure probabilities

4.1. Application of Cooke's method

As explained in Section 2.5, the elicitation was organized in two half-day sessions. The first was used for filling in the questionnaire, the second to discuss the results. The questionnaire was not discussed with the participating experts in advance. From the discussion afterwards, we noticed that the questions' underlying assumptions can have a major effect on the estimated uncertainties. Despite the questionnaire being as clear as possible on the assumptions and context, experts will still create their own image of the assessed dike failure. This was most evident in the question of the failure probability of the macro-stability schematization. The effect of their specific image of the failure process is greater than the estimates of uncertainties in, for example, the model parameters.

Discussing the questions together beforehand is common for expert studies using Cooke's method but was not done in this study because of time and COVID constraints. Such a discussion can help to steer the experts into a similar reasoning, likely leading to convergence of their estimates. The risk is that a dominant expert's viewpoint (which could potentially be wrong), can unintentionally get the upper hand in the DM solution through the answers of other (now influenced) experts. Still, such a discussion seems to be preferred for questions that are not a straightforward parameter estimation, or when the answer is difficult to relate to experience.

There is little empirical data for failure of Dutch dikes in their current state. Experts therefore need to think through a number of steps that lead to failure (basically a model) and quantify subsequent steps (indirectly assigning probabilities to them). The more of these steps to go through, the greater the uncertainty in the last answer. In these cases, a plenary discussion could help. When a problem can be interpreted in several ways, experts can remind each other of conditions or ways of reasoning to reach a plausible answer. The questions about the piping and macro-stability schematizations are a definite reference of this. However, as soon as there is a reference, as with the question about the failure probability of the river system, the estimates of the experts are closer together. It is common knowledge among flood risk experts that the Lobith discharge at high water in 1995 was approximately 12,000 m^3/s, and Room for the River was designed at a discharge of 16,000 m^3/s. These are values that may be used as reference values by experts for answering such a question, which, according to the discussion after the elicitation, was done at least by some of the experts.

4.2. Experts score and consistency

To find out how experts perform for estimating small probabilities, we followed two approaches to derive a system failure probability. An expert with enough time that thinks through all assumptions and works through the whole calculation, would likely arrive at the same answer following both approaches. In Figs. 10 and 11 we saw however that the differences are large. In Fig. 12 these differences are plotted against the calibration score and information score from Cooke's method. Both

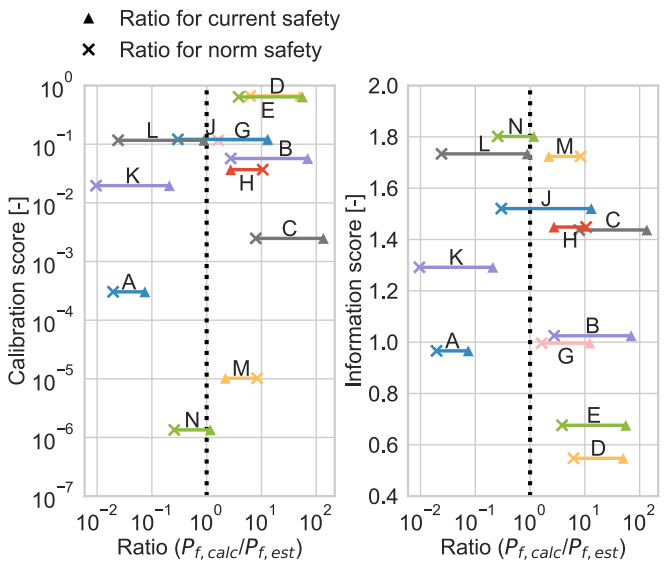

Fig. 12. Relation between failure probability ratio and calibration score (left) and information score (right). A triangle represents the expert's judgement of the factor for the current dike safety, a cross for the safety standard.

figures contain, per expert, two markers with a line between them. A triangle represents the current dike safety situation, and a cross the situation in which the dikes meet the safety standard. Expert *N* estimates a similar probability with the two methods, while, for example, experts K and C, have a larger deviation. We do not observe a clear relation between calibration score (indicating statistical accuracy in the estimation of uncertainty) and consistent answers. Ratios close to 1 we qualify as consistent answering, as this means both approaches lead to the same failure probability result. Consistency is not equal to correctness. However, a positive relation between consistency and calibration score would support the hypothesis that the calibration scores is able to measure experts' statistical accuracy independently of the method used for assessing low probabilities from expert judgments. The comparison between estimated and calculated failure probabilities does not show a clear pruning bias [13]. The fact that (i) we did not ask experts to estimate failure probabilities directly, and (ii) asked experts to estimate system-level probabilities last, might contribute to this.

In Cooke's method, the calibration score influences the DM solution more than the corresponding information score. It generally favors the more uncertain experts: of the 33 studies considered in [29], in 27 studies the rank correlation between the information score and calibration score is less than 0.0. The median rank correlation over all 33 studies is −0.4, indicating an inverse relationship between calibration and information score. Although Cooke's method is not infallible, the use of weight-based DMs has shown to outperform equal weighting [32–34]. In this study however, a large uncertainty (meaning low information, and likely high calibration) dramatically increases the estimated failure probabilities. Consequently, we observe that the 'best' experts as well as the DM's (including the equal weight) estimate (perhaps too) high failure probabilities.

4.3. Model versus experts

This study presents expert judgment as an alternative to a model-based approach. Reality is however less binary. A model-based approach for the Dutch flood defense system mostly comes down to an engineer (or a group of engineers) using a model for a safety assessment. The engineers however make several choices while doing so. First, what model to choose, and subsequently, which parameter values to use. These choices can be passive, for example when based on

a standard procedure, but it is still a choice to adopt these. At the same time, during an elicitation, experts can use models to substantiate their estimates. When viewed from this perspective, the difference between a model-based approach and expert judgment is mainly the focus: Expert judgment clearly puts this on the experts, who determine what model and parameter values to use. Standard procedure and defaults will more likely be questioned in this approach.

5. Conclusion

In this study we applied Cooke's method for elicitation of expert judgments to estimate system failure probability for the Dutch river Rhine. To do so, we followed two different approaches: A direct approach, in which experts estimate the discharge at which at least one dike fails, and a second approach in which estimates of bias and uncertainty are used to adjust failure probabilities from models. The first approach results in failure probabilities from 1/30 to 1/17,000 in a year. For most experts, these probabilities contrast dramatically to the 1/15 to 1/100 in a year failure probability range resulting from the VNK2 failure probabilities: even under the assumption that no emergency measures are used (e.g., placing sandbags), most experts assess a failure probability smaller than 1/100 in a year for the current dike state. For the situation in which the dikes are reinforced to match the required safety level, experts estimate a failure probability varying between 1/100 and 1/200,000 in a year. These estimates are more in line with the expected 1/215 to 1/1,000 in a year failure probability resulting from combining the safety standards assuming independence along dike trajectories and full dependence, respectively. Thus, our expert pool expresses that in the current state, dikes are safer than the models say, while the estimating a more similar safety for the case where the dikes have been reinforced.

Considering the more detailed results, half the experts estimated a failure probability for the piping schematization that is larger than the model suggests. This is surprising because the experts' shared impression is that the adjusted Sellmeijer model results in failure probabilities that are too high. It would be premature to assume this impression is wrong if we only base that on the single assessed schematization. It is however a surprising result. For macro-stability, most experts think the model is slightly conservative, but the differences are smaller than for piping. More schematizations for piping as well as macro-stability should be assessed to reach a general conclusion on a potential model bias.

The second approach, in which the failure probability was calculated by adjusting model results, gives total system failure frequencies ranging from larger than once in a year to 1/7,500 in a year. The upper limits of these recurrence rates are unlikely high and are caused by the large uncertainty ranges given by some experts: Wide estimates for the failure-critical water level range assign too high failure probabilities to frequent events. Especially for the piping and macro-stability assessments, the lack of reference values leads to wide uncertainty estimates. The direct failure discharge estimates, on the other hand, show more narrow uncertainty estimates, as this conceptualization is easier to relate to reality. The questionnaire was not discussed with the experts before the elicitation. Especially for questions that left some room for interpretation, this absence of discussion and expert interaction allowed quite different, unmodulated interpretations to be articulated.

Experts providing wide uncertainty may sometimes result in a high calibration score from Cooke's method, resulting in a large weight for these experts in the Decision Maker solution. A positive relation between the consistency of each experts estimates and the weight in Cooke's method would support the choice of using a global or item weights DM. However, there seems to be no significant relation.

To conclude, we found that experts estimated plausible probabilities of dike failure on a system level, while they struggled to answer the questions concerning dike sections accurately. Compared to a model-based approach, structured expert judgment has the advantage that uncertainties are made explicit, whereas in a model-based approach these may be hidden, latent or ill-defined.

To obtain reliable and defensible estimates of event and exceedance probabilities using expert judgment, for hazards such as dike failure, it is desirable to establish a clear and, if possible, agreed framework of technical definitions, empirical observations, other data and modeling assumptions and results, and established knowledge. Our study of expert judgement-based failure probability estimation for a system of river dikes in the Netherlands, explored some key aspects of this challenge, and how, quantitatively, they influenced our findings. Importantly, these point at certain aspects that merit further investigation by expert elicitation.

CRediT authorship contribution statement

G. Rongen: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **O. Morales-Nápoles:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. **M. Kok:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Guus Rongen reports a relatonship with HKV Consultants that includes: employment. Matthijs Kok reports a relatonship with HKV Consultants that includes: employment.

Acknowledgments

We would like to thank the experts who participated in the study and helped preparing with the preparation. We would also like to thank the TKI-committee: Bas Kolen, Annemargreet de Leeuw, and Durk Riedstra, for their help setting up the study. Finally, we thank the HKV colleagues, whose input has been useful in formulating the questions correctly. In particular, their expertise on the geotechnical details.

This research was funded by the TKI project EMU-FD, The Netherlands. This research project is funded by Rijkswaterstaat, The Netherlands, Deltares, The Netherlands and HKV consultants, The Netherlands.

Appendix. Supplementary information

This appendix contains the supplementary information for the study:

- An overview of the questionnaire, with the questions, realizations, and elicited percentiles, is shown in Table A.2.
- Table A.3 lists 12 experts that participated in the study, their affiliation, and their professional interest. One of the experts wished to remain anonymous, bringing the total number of experts to the 13 mentioned in Section 2.5. The experts are ordered alphabetically based on their first name. This order holds no relation to the expert letters used throughout this article.
- A full overview of the expert estimates for the seed questions is given in Fig. A.13, while Fig. A.14 gives an overview for the target questions.

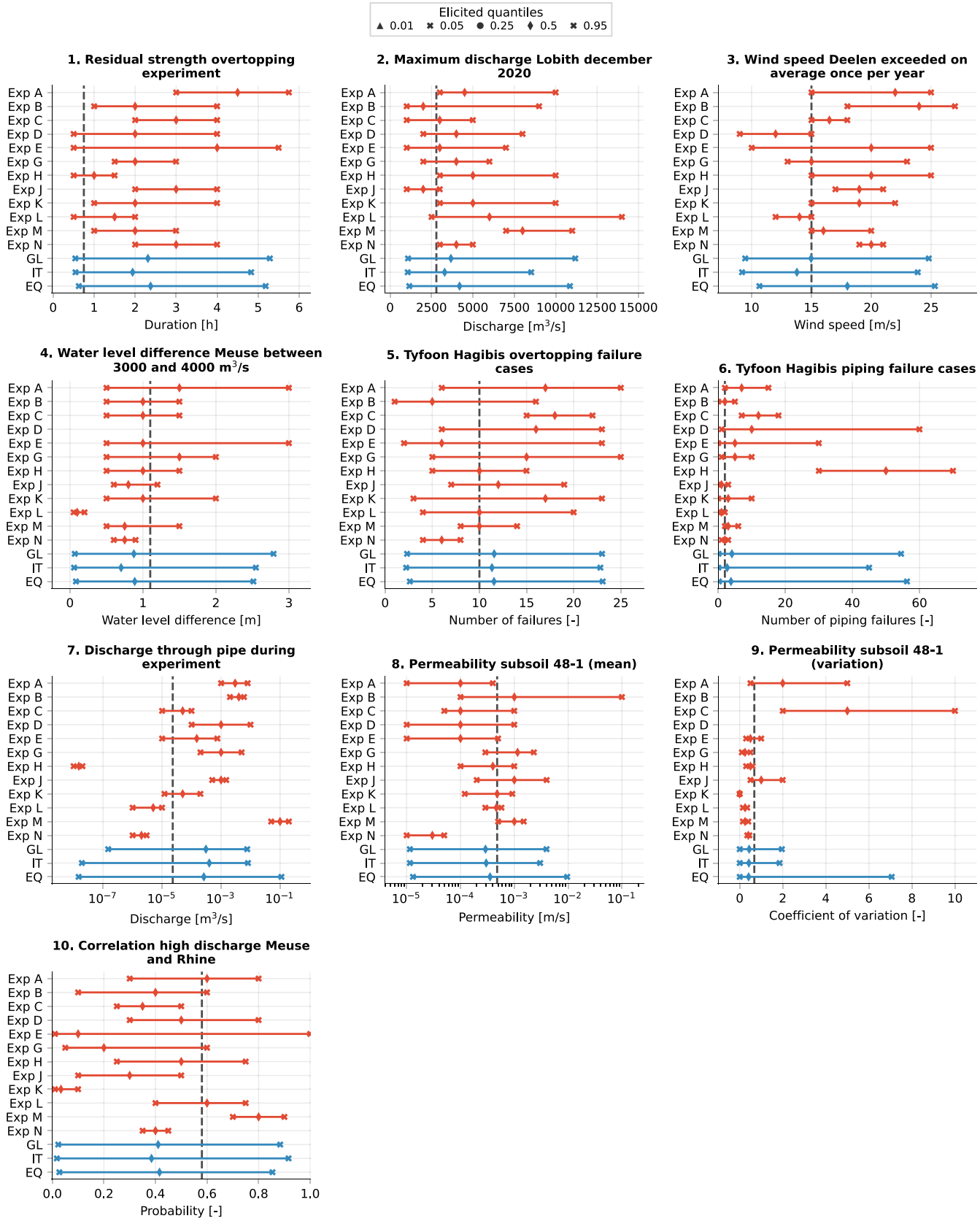

Fig. A.13. Expert and DM estimates for seed questions. The realization is indicated with the dashed vertical line.

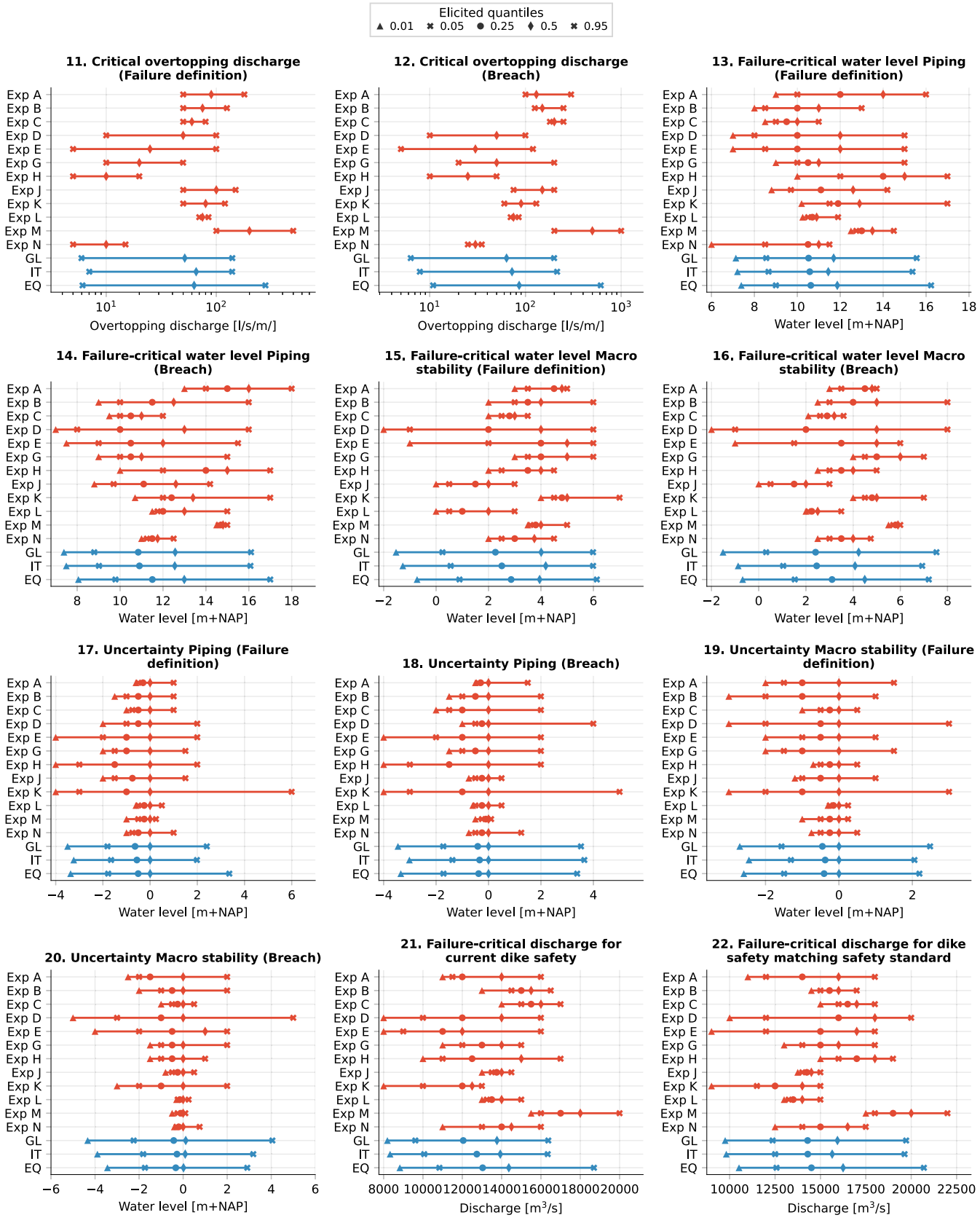

Fig. A.14. Expert and DM estimates for target questions.

Table A.2
Overview of questions in the questionnaire. The questions are adjusted such that they are easier to comprehend without the context described in the questionnaire.

Item	Question	Realization	Elicited percentiles
1.	Looking at the damage from the overtopping experiment in the photo: how long (in hours) did it take between the first visible damage and the damage in the photo?	0.75 h	5, 50, 95
2.	What is the highest discharge (m^3/s) that will occur at Lobith (where the Rhine flows into the Netherlands) in December 2020? *Note that the questionnaire was deducted before this month*	2794 m^3/s	5, 50, 95
3.	What is the wind speed (m/s) that is exceeded on average once a year at Deelen?	15.0 m/s	5, 50, 95
4.	What is the average difference on the Maas (between Venlo and Den Bosch) between the water level at a discharge at Borgharen of 3000 and 4000 m^3/s?	1.1 m	5, 50, 95
5.	Considering the provided information about the flood that followed the heavy rainfall during Typhoon Hagibis, how many of these 29 overflown dikes have failed?	10	5, 50, 95
6.	How many of the 142 dikes that failed in total during the floods after Typhoon Hagibis, where due to the failure mechanism piping?	2	5, 50, 95
7.	Considering the characteristics of the described piping experiment, what is the flow through the well when the critical gradient is reached?	2.3e−05 m^3/s	5, 50, 95
8.	What is the mean (μ) permeability (k) of the subsoil under the dikes of section 48-1?	0.00048 m/s	5, 50, 95
9.	What is the mean coefficient of variation ($V = \sigma/\mu$) of the permeability (k) of the subsoil under the dikes of section 48-1?	0.679	5, 50, 95
10.	Given that a Rhine discharge of 6,000 m^3/s is exceeded. What is the probability that the Meuse discharge of 1,500 m^3/s is exceeded within a period of 10 days before or after the moment the Rhine discharge has been exceeded?	0.58	5, 50, 95
11.	At what overtopping discharge (l/s/m) do you expect these specific wave conditions to erode the 50 cm clay layer? In other words, what do you expect the overtopping discharge to be?		5, 50, 95
12.	At what overtopping discharge (l/s/m) do you expect a breach in the dike (dike opening) to occur, again after 6 h of wave attack?		5, 50, 95
13.	For the given schematization, at what river water level (m+NAP) do you expect piping (an unstable pipe) to occur?		1, 5, 25, 50, 95
14.	For the given schematization, at what river water level (m+NAP) do you expect a breach to occur as a result of a sand-carrying pipe?		1, 5, 25, 50, 95
15.	For the given schematization, at what river water level (m+NAP) do you expect a deformation die to instability of the inner slope, with an entry point in the crest?		1, 5, 25, 50, 95
16.	For the given schematization, at what river water level (m+NAP) do you expect a breach to occur as a result of an instability?		1, 5, 25, 50, 95
17.	For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect piping to start?		1, 5, 25, 50, 95
18.	For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect a breach due to piping?		1, 5, 25, 50, 95
19.	For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect macro-stability to cause a shearing of the inner slope with an entry point in the crest?		1, 5, 25, 50, 95
20.	For river dikes in general, at what water level (relative to 0 m+NAP) do you expect a breach due to macro-stability?		1, 5, 25, 50, 95
21.	At which peak discharge (m^3/s) do you expect at least one dike in the river system to fail for the current dike safety situation?		1, 5, 25, 50, 95
22.	At which peak discharge (m^3/s) do you expect at least one dike to fail in the river system when all dikes meet the required safety level?		1, 5, 25, 50, 95

Table A.3
List of experts with their affiliation and professional interests.

Name	Affiliation	Specialism
Carlijn Bus	Waterschap Brabantse Delta	Specialized in flood risk assessment of dikes.
Don de Bake	HKV	Senior advisor flood risk. Specialized in flood risk management, dike safety assessment, and dike restoration projects. Policy advisor to the Ministry of Infrastructure and Water Management.
Henk van Hemert	Rijkswaterstaat	over 25 years professional experience in dike projects with a geotechnical focus
Jan Blinde	Deltares	Specialism: flood risk, dike design, dike assessment
Jan Tigchelaar	HKV	Specialized in geotechnics and probability applied to dike failure and flood risk. Advisor in different national and international projects.
Jan-Kees Bossenbroek	Waterschap Hollandse Delta	Flood risk advisor, specialized in applying the flood risk approach and flood defense knowledge in the South-Holland Delta.
Leo van Nieuwenhuijzen	Waterschap Rijn en IJssel	Flood risk advisor, contact point for calamity care in case of imminent flood waves.
Marinus Aalberts	Witteveen+Bos	Senior engineer in flood risk and dike design. Member of Expertise Network for Flood Protection (ENW)
Philippe Schoonen	Waterschap Drents Overijsselse Delta	Technical manager Flood Protection, Coordinator innovation program
R.B. Jongejan PhD	Jongejan Risk Management Consulting BV	Specialized in flood risk analysis and probabilistic design; independent engineering consultant
Stefan van den Berg	Rijkswaterstaat	Flood risk advisor in the execution phase of projects, with a focus on connecting theory and practice.
Wim Kanning	Deltares and Delft University of Technology	Expert in levee safety, geotechnical reliability, and risk.

References

[1] Ministry of Infrastructure and Environment. Regeling veiligheid primaire waterkeringen 2017 no IENM/BSK-2016/283517. 2016, https://wetten.overheid.nl/BWBR0039040/2017-01-01.

[2] Vrijling J. Probabilistic design of water defense systems in The Netherlands. Reliab Eng Syst Saf 2001;74(3):337–44. http://dx.doi.org/10.1016/S0951-8320(01)00082-5, URL https://www.sciencedirect.com/science/article/pii/S0951832001000825.

[3] Torres-Alves GA, Morales-Nápoles O. Reliability analysis of flood defenses: The case of the nezahualcoyotl dike in the aztec city of tenochtitlan. Reliab Eng Syst Saf 2020;203:107057. http://dx.doi.org/10.1016/j.ress.2020.107057, URL https://www.sciencedirect.com/science/article/pii/S0951832020305585.

[4] Slomp R, Knoeff H, Bizzarri A, Bottema M, De Vries W. Probabilistic flood defence assessment tools. In: E3S Web of conferences, vol. 7. 2016, p. 1–14. http://dx.doi.org/10.1051/e3sconf/20160703015.

[5] ILT- Informatiehuis Water. Waterveiligheidsportaal - landelijk veiligheidsbeeld. 2021, Online URL https://waterveiligheidsportaal.nl/#/nss/nss/current. [Accessed 15 January 2021].

[6] Jonkman SN, Jongejan R, Maaskant B. The use of individual and societal risk criteria within the dutch flood safety policy—Nationwide estimates of societal risk and policy applications. Risk Anal Int J 2011;31(2):282–300.

[7] Kind JM. Economically efficient flood protection standards for the N etherlands. J Flood Risk Manag 2014;7(2):103–17.

[8] Dupuits E, Schweckendiek T, Kok M. Economic optimization of coastal flood defense systems. Reliab Eng Syst Saf 2017;159:143–52.

[9] Sayers P, Meadowcroft I. RASP-A hierarchy of risk-based methods and their application. In: Defra Flood and Coastal Management Conference 2005. 2005, p. 1–18.

[10] Van der Meer J, Ter Horst W, Van Velzen E. Calculation of fragility curves for flood defence assets. Flood Risk Manag Res Pract 2009;567–73.

[11] Kok M, Jongejan R, Nieuwjaar M, Tanczos I. Fundamentals of flood protection. Tech. rep, Ministerie van Infrastructuur en Milieu, Expertise Netwerk Waterveiligheid; 2017, p. 144.

[12] Nofal OM, van de Lindt JW, Do TQ. Multi-variate and single-variable flood fragility and loss approaches for buildings. Reliab Eng Syst Saf 2020;202:106971. http://dx.doi.org/10.1016/j.ress.2020.106971, URL https://www.sciencedirect.com/science/article/pii/S0951832019313262.

[13] Dupuits E, Klerk W, Schweckendiek T, de Bruijn K. Impact of including interdependencies between multiple riverine flood defences on the economically optimal flood safety levels. Reliab Eng Syst Saf 2019;191:106475. http://dx.doi.org/10.1016/j.ress.2019.04.028, URL https://www.sciencedirect.com/science/article/pii/S0951832018308676.

[14] Curran A, De Bruijn K, Domeneghetti A, Bianchi F, Kok M, Vorogushyn S, et al. Large-scale stochastic flood hazard analysis applied to the Po River. Nat Hazards 2020;104. http://dx.doi.org/10.1007/s11069-020-04260-w.

[15] Klerk W, Kok M, De Bruijn K, Jonkman S, Van Overloop P. Influence of load interdependencies of flood defences on probabilities and risks at the bovenrijn/ijssel area, The Netherlands. In: Proceeding of the 6th International Conference on Flood Management–ICFM6. 2014, p. 1–13.

[16] Jongejan RB, Diermanse F, Kanning W, Bottema M. Reliability-based partial factors for flood defenses. Reliab Eng Syst Saf 2020;193:106589.

[17] Hokstada P, Øien K, Reinertsen R. Recommendations on the use of expert judgment in safety and reliability engineering studies. Two offshore case studies. Reliab Eng Syst Saf 1998;61(1–2):65–76.

[18] Cooke RM, Goossens LL. TU delft expert judgment data base. Reliab Eng Syst Saf 2008;93(5):657–74. http://dx.doi.org/10.1016/j.ress.2007.03.005.

[19] Hathout M, Vuillet M, Carvajal C, Peyras L, Diab Y. Expert judgments calibration and combination for assessment of river levee failure probability. Reliab Eng Syst Saf 2019;188:377–92. http://dx.doi.org/10.1016/j.ress.2019.03.019, URL https://www.sciencedirect.com/science/article/pii/S0951832018303296.

[20] Projectbureau VNK2. De veiligheid van Nederland in kaart. Tech. Rep. November, Rijkswaterstaat Projectbureau VNK; 2010.

[21] Jongejan R, Maaskant B. Quantifying flood risks in the netherlands. Risk Anal 2015;35(2):252–64.

[22] 't Hart R, de Bruijn H, de Vries G. Fenomenologische beschrijving. Tech. rep., Deltares; 2016, p. 208.

[23] Knoeff H, Sellmeijer J, Lopez J, Luijendijk S. SBW Piping - Hervalidatie piping. Tech. rep., Delft: Deltares; 2009.

[24] Van der Meij R. D-Stability - User Manual. Tech. rep., Deltares; 2019, p. 171.

[25] Geerse C. Hydra-Zoet for the fresh water systems in the Netherlands - Probabilistic model for the assessment of dike heights. Tech. rep., HKV consultants; 2011, p. 167.

[26] van Hoven A. Schematiseringshandleiding grasbekleding. Tech. rep. november, Deltares; 2019.

[27] Chbab H. Basisstochasten WBI-2017 - Statistiek en statistische onzekerheid. Tech. rep., Deltares; 2017, p. 214.

[28] Rongen G, 't Hart CMP, Leontaris G, Morales-Nápoles O. Update (1.2) to ANDURIL and ANDURYL: Performance improvements and a graphical user interface. SoftwareX 2020;12:100497. http://dx.doi.org/10.1016/j.softx.2020.100497, URL http://www.sciencedirect.com/science/article/pii/S2352711020301837.

[29] Leontaris G, Morales-Nápoles O. ANDURIL: A MATLAB toolbox for analysis and decisions with UnceRtaInty: Learning from expert judgments. SoftwareX 2018;7:313–7. http://dx.doi.org/10.1016/j.softx.2018.07.001.

[30] 't Hart CMP, Leontaris G, Morales-Nápoles O. Update (1.1) to ANDURIL — A MATLAB toolbox for analysis and decisions with UnceRtaInty: Learning from expert judgments: ANDURYL. SoftwareX 2019;10:100295. http://dx.doi.org/10.1016/j.softx.2019.100295, URL http://www.sciencedirect.com/science/article/pii/S2352711019302419.

[31] Chen H-P, Mehrabani MB. Reliability analysis and optimum maintenance of coastal flood defences using probabilistic deterioration modelling. Reliab Eng Syst Saf 2019;185:163–74.

[32] Clemen RT. Comments: Comment on Cooke's classical method. Reliab Eng Syst Saf 2008;93(5):760–5. http://dx.doi.org/10.1016/j.ress.2008.02.003.

[33] Eggstaff JW, Mazzuchi TA, Sarkani S. The effect of the number of seed variables on the performance of Cooke' s classical model. Reliab Eng Syst Saf 2014;121:72–82.

[34] Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. Reliab Eng Syst Saf 2017;163:109–20.