Expert Judgment in the Uncertainty Analysis of Dike Ring Failure Frequency

Roger M. Cooke
Department of Mathematics
Delft University of Technology
Delft,
The Netherlands

Karen A. Slijkhuis
Ministerie Verkeer en Waterstaat
Bouwdienst
Utrecht

Abstract

The Netherlands are protected by a system of dike rings. A dike ring consists of some 50 dike sections, each section varying between 0.5 and 2 km in length. It has recently been decided to base the design philosophy in The Netherlands on the reliability of dike rings. A dike ring fails when one of its sections fails. Since dike sections are subject to the same environment, there will be significant dependencies in the failure probabilities of dike sections. Because of the new features entailed by the dike ring safety concept, an uncertainty analysis was performed to assess the uncertainty in the predictions of dike ring failure frequency. This analysis made extensive use of structured expert judgment. This chapter reports on the expert judgment methods and results.

Keywords

Dike Ring, structural reliability, expert judgment, uncertainty analysis, sensitivity analysis.

1. Introduction

The Netherlands are situated on the delta of three of Europe's biggest rivers, the Rhine, the Meuse and the Scheldt. Large parts of the country lie lower than the water levels that may occur on the North Sea, the large rivers and the IJsselmeer. Consequently, most of the country is protected by flood defenses.

Prior to 1953 the standard approach to the design of flood defenses was based on the highest recorded water level. In relation to this water level a safety margin of 1 meter

was maintained. In 1953 a major breach occurred killing more than 1800 people in the southwest of the country. Afterwards, an econometric analysis was undertaken by the Delta Committee in which the safety standards were based on weighing the costs of the construction of flood defenses and the possible damage cause by floods. This analysis led to a design probability of flooding of 8×10^{-6} per year. Given the technical capabilities at that time, this safety concept could not be implemented completely. In particular, the probability of a flood defense collapsing, and therefore the probability of flooding, proved to be very difficult to estimate.

For this reason a simplified concept was chosen at the time, based on design loads. The basic assumption is that every individual section of a dike has to be high enough to safely withstand a given extreme water level and the associated wave impact.

The current safety standards are laid down in the Flood Protection Act. This act foresees a change to a new safety concept based on the probability of flooding in certain dike ring areas. A dike ring area is an area that is completely surrounded by water and therefore surrounded by flood defenses. In the past few years models for determining the inundation probability for dike rings have been developed (Slijkhuis 1998, van der Meer, 1997).

Because of the many new features involved in the dike ring safety concept, it was decided to perform an uncertainty analysis on the prediction of failure frequency for one illustrative dike ring, the so called Hoeksche Waard. This involved the following steps:

1. Freezing the structural reliability model for dike section failure.
2. Performing an ‘in-house’ quantification of uncertainty, with dependence, of all input variables.
3. Identifying those input variables whose uncertainty is anticipated to have significant impact on the uncertainty of the dike ring failure frequency.
4. Assessing the uncertainty of the selected variables with structured expert judgment.
5. Propagating the resulting uncertainties through the structural reliability model so as to obtain an uncertainty distribution on the failure frequency of the dike ring.

The motivations and goals for performing an uncertainty analysis, as opposed to a traditional reliability analysis, are set forth in the next section.

The following section expands on uncertainty analysis. Section 3 discusses the structured expert judgment methodology. Section 4 discusses the dike ring study, and section 5 presents results. A final section gathers conclusions.

2. Uncertainty Analysis

In contrast to more standard reliability analyses, the goal of this study is not to predict a failure frequency. Rather, the goal is to determine the *uncertainty* in the failure frequency for a given dike ring. The reason for shifting the focus to uncertainty in this way is because the primary source of our information in this case is expert judgment, not field data. We appeal to expert judgment because there is not sufficient data, and hence substantial uncertainty. The goal of using expert judgment, as discussed in the next section, is to obtain a rationally defensible quantification of this uncertainty.

The first step proved surprisingly difficult. The structural reliability model for a dike ring is large, involving some 300 input variables whose values are not known with certainty. Moreover the model is under continual development and it is very difficult to freeze a model which is known to become outdated before the uncertainty analysis is completed.

Since there are a large number of uncertain input variables in these dike ring models, it is not possible to subject all of these variables to a structured expert judgment elicitation. Instead, we must restrict the expert elicitation to those variables which are judgment most important in driving the uncertainty of the dike ring failure frequency. This requires an initial ‘in house’ quantification of uncertainties (Step 2) and a selection of important variables (Step 3). The techniques used in Step 3 are very much in development at the moment (see e.g. Saltelli et al. 2000), and will not be treated here.

The present chapter focuses on Step 4. The reasons for this are twofold. First, many of the questions put to the experts are of general interest, and do not require lengthy exposition of the structural reliability model. These include:

- Frequencies of exceedence of extreme water levels of the North Sea
- Sea level rise
- Frequencies of exceedence of extreme discharges of the Rhine river
- Influence of climate change and human intervention in the Rhine discharge

Second, this was the major effort of the study, and the application of the structured approach led to significant insights.

3. Expert judgment method

The methodology for expert judgment has been presented in Cooke (1991) and applied in many risk and reliability studies. See Goossens et al (1998) for a recent overview. In particular, this method was used in the study of failure of gas pipelines described in this volume. This section briefly describes the expert judgment method. It is based on Frijters et al (1998), and Cooke and Goossens (2000a,b).

The goal of applying structured expert judgment is to enhance rational consensus. Rational consensus is distinguished from 'political consensus' in that it does not appeal to a "one-man-one-vote" method for combining the views of several experts. Instead, views are combined via weighted averaging, where the weights are based on performance measures, and satisfy a proper scoring rule constraint (Cooke 1991). This model for combining expert judgments bears the name "classical model" because of a strong analogy with classical hypothesis testing. We restrict discussion to the case where experts assess their uncertainty for quantities taking values in a continuous range. There are two measures of performance, calibration and information. These are presented briefly below, for more detail see Cooke (1991).

3.1 Calibration.

The term calibration was introduced by psychologists (Kahneman et al 1982) to denote a correspondence between subjective probabilities and observed relative

frequencies. This idea has fostered an extensive literature and can be operationalized in several ways. In the version considered here, the classical model treats an expert as a classical statistical hypothesis, and measures calibration as the degree to which this hypothesis is supported by observe data, in the sense of a simple significance test.

More precisely, an expert states n fixed quantiles for his/her subjective distribution for each of several uncertain quantities taking values in a continuous range. There are n+1 'inter-quantile intervals' into which the realizations (actual values) may fall. Let

$$p = (p_{1,\ldots}p_{n+1}) \qquad (1)$$

denote the theoretical probability vector associated with these intervals. Thus, if the expert assesses the 5%, 25%, 50%, 75% and 95% quantiles for the uncertain quantities, then n = 5 and p = (5%, 20%, 25%, 25%, 20%, 5%). The expert believes there is a 20% probability that the realization falls between his/her 5% and 25% quantiles, etc.

In an expert judgment study, experts are asked to assess their uncertainty for variables for which the realizations are known post hoc. These variables are chosen to resemble the quantities of interest, and/or to draw on the sort of expertise which is required for the assessment of the variables of interest. They are called "calibration" or "seed" variables.

Suppose we have such quantile assessments for N seed variables. Let

$$s = (s_1,\ldots s_{n+1}) \qquad (2)$$

denote the empirical probability vector of relative frequencies with which the realizations fall in the inter quantile intervals. Thus s_2 = (#realizations strictly above the 5% quantile and less than or equal to the 25% quantile)/N, etc.

Under the hypothesis that the realizations may be regarded as independent samples from a multinomial distribution with probability vector p, the quantity[1]

$$2NI(s,p) = 2N\sum_{i=1..N} s_i \ln(s_i/p_i) \qquad (3)$$

is asymptotically Chi-square distributed with n degrees of freedom and large values are significant. Thus, if χ_n is the cumulative distribution function for a Chi-square variable with n degrees of freedom, then

$$CAL = 1 - \chi_n(2NI(s,p)) \qquad (4)$$

is the upper tail probability, and is asymptotically equal to the probability of seeing a disagreement no larger than I(s,p) on N realizations, under the hypothesis that the realizations are drawn independently from p.

We take CAL as a measure of the expert's calibration. Low values (near zero) correspond to poor calibration. This arises when the difference between s and p cannot plausibly be the result of mere statistical fluctuation. For example, if N = 10, and we find that 8 of the realizations fall below their respective 5% quantile or above their respective 95% quantile, then we could not plausibly believe that the probability for such events was really 5%, as the expert maintains. This would correspond to an expert giving 'overconfident' assessments. Similarly, if 8 of the 10 realizations fell below their 50% quantiles, then this would indicate a 'median bias. In both cases, the value of CAL would be low. High values of CAL indicate good calibration.

It is well to emphasize that we are not testing or rejecting hypotheses here. Rather, we are using the standard goodness of fit scores to measure an expert's calibration.

3.2 Information

Loosely, information measures the degree to which a distribution is concentrated. This loose notion may be operationalized in many ways. For a discussion of the pro's and con's of various measures, see Cooke (1991). We shall measure information as

[1] I(s,p) is called the relative Shannon information of s with respect to p. For all s,p with $p_i > 0$, i =

Shannon's relative information with respect to a user-selected background measure. The background measure will be taken as the uniform measure over a finite 'intrinsic range'. For a given uncertain quantity and a given set of expert assessments, the intrinsic range is defined as the smallest interval containing all the experts' quantiles and the realization, if available, augmented above and below by K%. The overshoot term K is chosen by default to be 10, and sensitivity to the choice of K must always be checked (see Table 2 below).

To implement this measure, we must associate a probability density with each expert's assessment for each uncertain quantity. When the experts have given their assessments in the form of quantiles, as above, we select that density which has minimal Shannon information with respect to the background measure and which complies with the expert's quantile assessments. When the uniform background measure is used, the minimum information density is constant between the assessed quantiles, and the mass between quantiles i-1 and i is just p_i. If $f_{k,j}$ denotes the density for expert k and uncertain quantity j, then Shannon's relative information with respect to the uniform measure on the intrinsic range I_j is:

$$I(f_{k,j}, U_j) = \int_{u \in I_j} f_{k,j}(u) \ln(f_{k,j}(u)) du + \ln(|I_j|) \qquad (5)$$

where $|I_j|$ is the length of I_j. For each expert, an information score for all variables is obtained by summing the information scores for each variable. This corresponds to the information in the expert's joint distribution relative to the product of the background measures under the assumption that the expert's distributions are independent. Roughly speaking, with the uniform background measure, more informative distributions are gotten by choosing quantiles which are closer together whereas less informative distributions result when the quantiles are farther apart.

The calibration measure CAL is a"fast" function. With, say, 10 realizations we may typically see differences of several orders of magnitude in a set of, say 10 experts. Information on the other hand is a "slow" function. Differences typically lie within a

1,…N, we have $I(s,p) \geq 0$ and $I(s,p) = 0$ if and only if s=p (see Kullback 1959).

factor 3. In the performance based combination schemes discussed below, this feature means that calibration dominates strongly over information. Information serves to modulate between more or less equally well calibrated experts. The use of the calibration score in forming performance based combinations is a distinctive feature of the classical model and implements the principle of empirical control discussed above.

3.3 Combination

Experts give their uncertainty assessments on query variables in the form of, say, 5%, 25%, 50%, 75% and 95% quantiles. An important step is the combination of all experts‘ assessments into one combined uncertainty assessment on each query variable. The three combination schemes considered here are examples of "linear pooling"; that is, the combined distributions are weighted sums of the individual experts' distributions, with non-negative weights adding to one. Different combination schemes are distinguished by the method according to which the weights are assigned to densities. These schemes are designated "decision makers". Three decision makers are described briefly below.

Equal weight decision maker

The equal weight decision maker results by assigning equal weight to each density. If E experts have assessed a given set of variables, the weights for each density are 1/E; hence for variable i in this set the decision maker's density is given by:

$$f_{eddm,i} = \left(\frac{1}{E}\right) \sum_{j=1\cdots E} f_{j,i} \qquad (6)$$

where $f_{j,i}$ is the density associated with expert j's assessment for variable i.

Global weight decision maker

The global weight decision maker uses performance based weights which are defined, per expert, by the product of expert's calibration score and his(her) overall information score on seed variables, and by an optimization routine described below

(see, Cooke 1991 for details). For expert j, the same weight is used for all variables assessed. Hence, for variable i the global weight decision maker's density is:

$$f_{gwdm,i} = \frac{\sum_{j=1\cdots E} w_j f_{j,i}}{\sum_{j=1\cdots E} w_j} \qquad (7)$$

These weights satisfy a "proper scoring rule" constraint. That is, under suitable assumptions, an expert achieves his (her) maximal expected weight, in the long run, by and only by stating quantiles which correspond to his(her) true beliefs (see Cooke 1991).

Item weight decision maker

As with global weights, item weights are performance based weights which satisfy a proper scoring rule constraint, and are based on calibration and informativeness, with an optimization routine described below. Whereas global weights use an overall measure of informativeness, item weights are determined per expert and per variable in a way which is sensitive to the expert's informativeness for each variable. This enables an expert to increase or decrease his(her) weight for each variable by choosing a more or less informative distribution for that variable. For the item weight decision maker, the weights depend on the expert and on the item. Hence, the item weight decision maker's density for variable i is:

$$f_{iwdm,i} = \frac{\sum_{j=1\cdots E} w_{j,i} f_{j,i}}{\sum_{j=1\cdots E} w_{j,i}} \qquad (8)$$

3.4 Optimization

The proper scoring rule (Cooke 1991) constraint entails that an expert should be unweighted if his/her calibration score falls below a certain minimum, $\alpha > 0$. The value of α is determined by optimization. That is, for each possible value of α a

certain group of experts will be unweighted, namely those whose calibration score is less than α. The weights of the remaining experts will be normalized to sum to unity. For each value of α we thus define a decision maker dm_α computed as a weighted linear combination of the experts whose calibration score exceeds α. dm_α is scored with respect to calibration and information. The weight which this dm_α would receive if he were added as a "virtual expert" is called the "virtual weight" of dm_α. The value of α for which the virtual weight of dm_α is greatest is chosen as the cut-off value for determining the unweighted expert.

3.5 Validation

When seed variables are available, we can use these variables to score and compare different possible combinations of the experts' distributions, or as we shall say, different decision makers. In particular, we can measure the performance of the global and item weight decision makers with respect to calibration and information, and compare this to the equal weight decision maker, and to the experts themselves. This is done in the following section.

4. The Dike ring expert judgment study

17 experts participated in this expert judgment study. They are all associated with Dutch universities or governmental institutes. The experts were acquainted with the issues, study objectives and methods beforehand, and were elicited individually. A typical elicitation took 3 to 4 hours. Each expert gave 5%, 25%, 50%, 75% and 95% quantiles for 40 uncertain quantities, concerning:

- Frequencies per year of exceedence of extreme water levels of the North Sea
- Sea level rise
- Frequencies of exceedence of extreme discharges of the Rhine river
- Influence of climate change and human intervention in the Rhine discharge
- The significant wave height
- The significant wave period[2]

[2] Significant wave height and wave period are defined as the 67% quantile of the occurring wave height resp. wave period distribution.

- The model term for Zwendl[3]
- The model factor for critical discharge
- The model factor for occurring discharge[4]
- The dependence between model factor for critical discharge between dike sections
- The dependence between model factor for occurring discharge between dike sections

The issue of dependence and its assessment requires more exposition than is possible here, (see Cooke and Goossens 2000b).

Model factors and model terms are used to capture an experts' uncertainty with regard to model predictions, and are defined as the ratio of the realization to the model prediction. A model term is defined as the difference between the realization and the model prediction. An example of the elicitation for the model term Zwendl is given in the appendix.

Seed variables

The model factors and model terms afford the possibility of calibrating the experts' assessments. With some effort, realizations were found from historical and experimental records, and compared with model computations. We are interested in the relation between model predictions and realizations in extreme situations, say water levels higher than 4 meters above normal. We cannot find such realizations in the available data or in controlled experiments. However, we can find 'sub extreme' realizations (water levels 2.5 meters above normal). For these we can compute post hoc the model predictions. Comparing these two, we obtain realizations for the model factors and model terms. This is done in the current study and resulted in 47 seed variables. It is significant that this had *not* been done for the models in question prior

---

[3] Zwendl is a computer code which computes local water levels at estuary measuring stations as a function of inter alia North Sea storm profile and Rhine discharge profile. This model was calibrated extensively during its development. A simplified version used in the current dike ring model considers only a 'standard storm profile' characterised by peak discharge and peak North Sea surcharge. The current study brought to light the fact that this simplified version was never calibrated.

[4] The critical discharge for a dike section is the maximal flux of water [liters /meter second] which a dike section can withstand without failing. The occurring discharge is the actual occurring discharge over the crown of a dike section.

to the current study. These realizations and predictions were not known to the experts (nor to the analysts) at the time of the elicitation.

Figure 1 gives a “range graph” for the experts’ assessments of the significant wave height, and gives the results of 7 realizations.

These data were obtained from measurements of wave height distributions at the measurement station Marollegat over six months in situations where the model predictions could be calculated post hoc. We see that most experts placed their median value for this model factor at 1, indicating that the probability of over prediction was equal to the probability of under prediction. Expert 14 believed that the model for significant wave height almost certainly under predicts the actual significant wave height. Eyeballing this data, we might say that the experts are a bit under confident, that is, too many realizations fall within the 25% and 75% quantiles. There might be a slight tendency for the model to under predict, but the tendency seems small relative to the experts’ uncertainty.

A different picture emerges with the model term for the local water level model Zwendl shown in Figure 2:

This data are from the measuring station Dordrecht. When a high local water level was reached, the parameters for calculating Zwendl’s predicted water level were recovered post hoc. Similar pictures emerged from data from other measuring stations (Raknoord and Goidschalxoord).

All experts except numbers 1 and 12 placed their median value at zero. Only one data point is less than zero, and many fall outside the experts’ 95% quantiles. This picture suggests that the model under predicts local water levels to a degree which is significant relative to the experts’ uncertainty. This result was rather surprising. Most engineers assumed that the model Zwendl had been properly calibrated, and therefore put their median value at zero. It turned out that the original model was calibrated almost 30 years ago, according to methods and standards no longer current, and the simplified model actually in use had never been calibrated.

5. Results

Before presenting the results, it is useful to get a picture of how well the experts agree among themselves. For a given item this can be measured as the Shannon relative information of an expert's distribution relative to the equal weight combination of all experts' distributions. Averaging these relative informations over all items, we get the data in Table 1.

The values for all variables range between 1.148 and 0.389, with average value over all experts of 0.62. This value should be compared to the experts' own information relative to the background measure (Table 2 column 4) and to the result of performing robustness analysis on the selection of experts (see below Table 3 column 5).

Table 2 gives scoring results for the experts, and various decision makers.

The item weight decision maker shown in the bold bordered cells exhibits the best performance (as judged by unnormalized weight[5]) and is chosen for this study. The values in the shaded areas are given as comparisons to the item weight values. The unnormalized weights are the global weights w_j in (7). Note that column 6 is just the product of columns 2 and 5.

Column 4 of Table 2 should be compared to column 2 of Table 1. We see that the experts are much more informative relative to the uniform distribution over the intrinsic range, than they are with respect to the equal weight dm. This indicates that the experts' 90% confidence bands display considerable overlap. If there were little overlap, the information scores in Table 1 would be closer to those in Table 2.

The number of seed variables, 47, is quite large; and since 6 or 7 realizations are typically available for a single measuring station, it is doubtful whether the

[5] If the size of the intrinsic range is changed, then the information scores also change, hence the item weight DM should be compared with the glob al weight DM and the equal weight DM with the same intrinsic ranges.

realizations should be considered as independent samples as required by the calibration measure. We can account for this by treating each expert's empirical probability vector s as if it were generated by a smaller number of samples. In short, we replace N in (3) by a smaller number, called the effective sample size. A effective sample size of 9 results from considering the measurements in distinct groups (Figures 1 and 2 each represent one group) as one effective measurement. Although not really defensible, this does establish a lower bound to the effective sample size. The calibration scores in the third column may be compared with many other studies involving a comparable number of seed variables, and it emerges that these experts, as a group, are comparatively well calibrated. Considering all 47 seed variables as independent leads to the scores in the second column. Although the expert performance would be much worse in this case[6], the global and item weight decision makers still reflect good calibration. To render the intuitive meaning of these scores we could say: *The hypothesis that the realizations are drawn independently from a distribution whose quantiles agree with those of the item weight decision maker would be rejected with these 47 realizations at the 0.4 significance level (which of course is much too high for rejection in a standard hypothesis test).*

Another way to assess the decision makers' performances is given in the last column. This is the normalized global weight after adding the decision maker to the pool of experts, and scoring him as another 'virtual expert'[7]. For the 17 experts, the values in the last column are the global weights the experts would receive if the item weight dm were added as a virtual expert. These are called "vitrual weights". The item weight dm has a virtual weight of 0.54237, higher than the combined weight of all experts. The virtual weights of the other decision makers are computed in the same way, but the corresponding weights of the 17 experts are not shown.

The last three rows of Table 2 show the three decision makers with 47 effective seed variables, and where the overshoot term K for the intrinsic range is changed from the default value 10 to 50. In other words, the intrinsic range for each variable is the

[6] We should reflect that *every* statistical hypothesis is wrong and would eventually be rejected with sufficient data. No expert will be *perfectly* calibrated, and imperfections, with sufficient data, will eventually produce a poor goodness of fit. The ratio's of scores is important when used in the performance based weights, but absolute scores are useful for comparing performance across studies.

smallest interval containing all assessed values, augmented by 50% above and below. All experts' information scores increase, as the background measure to which they are compared has become more diffuse (the increased scores are not shown). The equal weight dm's information score changes from 0.955 to 1.218, reflecting the fact that information is a slow function. Whereas the equal weight combination scheme is not affected by changing the intrinsic range, the performance based decision makers are affected slightly, as the experts' weights change slightly. The virtual weight of the item weight decision maker drops from 0.54237 to 0.5314.

The equal weight decision maker is less well calibrated and less informative than the performance based decision makers. Such a pattern emerges frequently, but not always, in such studies.

From the robustness analyses performed on this expert data, we present the results for robustness against choice of experts. Table 3 shows the results of removing the experts one at a time and recomputing the item weight dm. The second and third columns give the information scores over all items, and seed items only, for the 'perturbed dm' resulting from the removal of the expert in column 1. The fourth column gives the perturbed dm's calibration. The last two columns give the relative information of the perturbed expert with respect to the original expert. Column 5 may be compared to column 2 of Table 1.

The average of column 5; 0.026, should be compared with the average of column 2 in table 1; 0.62. This shows that the differences between the dm's caused by omitting a single expert are much smaller than the differences among the experts themselves. In this sense the robustness against choice of expert is quite acceptable. However, it will be noticed that loss of expert 10 would cause a significant degradation of calibration. Apparently this expert is very influential on the dm, and the good performance of this dm is not as robust in this respect as we would like. Of course, loss of robustness is always an issue when we optimize performance.

[7] It can be shown that adding a dm as a virtual expert and recomputing weights yields a new dm which is identical to the initial dm (see Cooke 1991).

For selected items of interest, Table 4 compares the item weight and the equal weight dm's with the in house assessments, where available, used in step 2 of the uncertainty analysis. The shaded cells indicate significant differences.

In general the numbers in Table 4 are in the same ballpark. But there are important differences. The in house assessments tend to be more concentrated than the combined experts' assessments. For the occurring discharge at a predicted level of 100 litres per meter per second the equal weight dm has a large 95% quantile.

6. Conclusions

The results presented above show that the structured expert judgment approach can be a valuable tool in structural reliability uncertainty analysis. Expert judgment is more than just subjective guessing. It can be subjected to rigorous empirical control and can contribute positively to rational consensus. The extra effort involved in defining seed variables and measuring expert performance, as opposed to simple equal weighting of experts has paid off in this case. The experts themselves and the problem owners appreciate that this same effort contributes toward rational decision making by extending the power of quantitative analyses. Without expert judgment, any quantification of the uncertainty in dike reliability models would be impossible and the discussion would remain stalled with safety concepts which we cannot implement.

Not only do we now have a defensible picture of the uncertainty in dike ring reliability calculations, but we can also judge where additional research effort might be expected to yield the greatest return in reducing this uncertainty and improving predictions.

The most obvious conclusion in this regard is that more effort should be devoted to calibrating the simplified version of the Zwendl model. From the engineer's viewpoint, this is the most significant result of this study. The relatively large uncertainty in the item weight dm for this variable throws some doubt on the reliability calculations for dike rings performed to date. In particular, there is greater probability that the local water levels are under predicted by the deterministic models.

The greatest improvement in these results would be gained by re-calibrating the model Zwendl for local high water levels. We anticipate that this would lead to higher predictions and smaller uncertainty bands. The techniques of using structured expert judgement to quantify uncertainty in combination with uncertainty propagation could be employed to many reliability models. However, the costs and effort involved are significant and thus these techniques are most suitable in problems of high public visibility where rational consensus is important.

Appendix: Example of elicitation question: model term Zwendl

The following is a translation of the text used to elicit the distribution of the model term for Zwendl. Experts were made familiar with the notion of subjective probability distribution and quantiles. The model term is queried in four ways, first unconditional on the actual local water level; then, conditional on a predicted local water level of 2m, 3m and 4m. This was done to see if the uncertainty in the model prediction depended on the predicted value. There was no significant pattern of dependence in the experts' responses. The unconditional assessments were used for calibrating the experts.

In the model for predicting dike ring inundation probability, Zwendl is used to compute locale water levels as a function of deterministic values for the water level at the Meuse mouth and the Rhine discharge at Lobith. In calculating the failure probabilities a model term for Zwendl will be added to the water levels computed with Zwendl. This model term is a variable whose distribution reflects the uncertainty in the Zwendl calculation in the following way:

Model term Zwendl = real local water level – local water level computed with Zwendl.

We would like to quantify the uncertainty in the output of Zwendl for the dike ring Hoeksche Waard. This uncertainty should take into account that Zwendl uses a standard water level profile for the North Sea and single value for the Rhine discharge instead of real temporal profiles.

We would like to quantify your uncertainty for the computation of local water levels using Zwendl. We do this by asking you to state your 5%, 25%, 50%, 75% and 95% quantiles for your subjective uncertainty distributions.

- *For an arbitrary water level what is the difference between the measured local water level and the water level computed with Zwendl (in m)*

5%	*25%*	*50%*	*75%*	*95%*

- *Assume that Zwendl predicts a water level of 2m above N.A.P. (the standard baseline water level) for a given location in the dike ring area of Hoeksche Waard; what is then the difference between the actually measured value of the local water level and the value computed with Zwendl (in m):*

5%	*25%*	*50%*	*75%*	*95%*

- *Assume that Zwendl predicts a water level of 3m above N.A.P. (the standard baseline water level) for a given location in the dike ring area of Hoeksche Waard; what is then the difference between the actually measured value of the local water level and the value computed with Zwendl (in m):*

5%	*25%*	*50%*	*75%*	*95%*

- *Assume that Zwendl predicts a water level of 4m above N.A.P. (the standard baseline water level) for a given location in the dike ring area of Hoeksche Waard; what is then the difference between the actually measured value of the local water level and the value computed with Zwendl (in m):*

5%	*25%*	*50%*	*75%*	*95%*

References

Cooke, R. (1991) *Experts in Uncertainty*, Oxford University Press

Cooke, R.M. and Goossens, L.H.J. (2000 a) "Procedures guide for structured expert judgment in accident consequence modelling" *Radiation Protection Dosimetry* vol. 90 No. 3, pp 303-309.

Cooke R.M. and Goossens, L.H.J (2000 b) *Procedures Guide for Structured Expert Judgment* European Commission, Directorate-General for Research, EUR 18820 EN, Luxembourg.

Frijters, M., Cooke, R. Slijkhuis, K. and van Noortwijk, J. (1999) *Expertmeningen Onzekerheidsanalyse, kwalitatief en kwantitatief verslag*, Ministerie van Verkeer en Waterstaat, Directoraat-Generaal Rijkswaterstaat, Bouwdienst; ONIN 1-990006.

Goossens, L.H.J.,Cooke, R.M. and Kraan, B.C.P. (1998) 'Evaluation of weighting schemes for expert judgment studies' *Proceedints of the 4th Ionternational Conference on Probabilistic Safety Assessment and ManagementI* Springer, New York, 1937-1942.

Kahneman, D., Slovic, P. and Tversky, A. (eds) (1982) *Judgment uncr Unceraínty*, Cambridge University Press.

Kullback, S. (1959) *Information Theory and Statistics* Wilely, New York.

Meer, J.W. van der, (1997) ' Golfoploop en golfoverslag bij dijken (wave run up and overtopping of dikes, in Dutch) Delft, Waterloopkundig Laboratorium, H2458/H3051.

Saltelli, A., Chan, K. and Scott, E.M. (2000) *Sensitivity Analysis*, Wiley, Chichester.

Slijkhuis, K.A.H. (1998) 'Beschrijvingenbundel' (in dutch) Ministerie Verkeer en Waterstat, Directoraat-Generaal Rijkswaterstaat, Bouwdienst Rijkswaterstaat, Utrecht.

Slijkhuis, K.A.H., Frijters, M.P.C., Cooke, R.M. and Vrouwenvelder, A.C.W.M. (1998) “Probability of flooding: an uncertainty analysis”, *Safety and Reliability* (Lyderson, Hansen and Sandtorv (eds), 1419-1425, Balkema, Rotterdam.

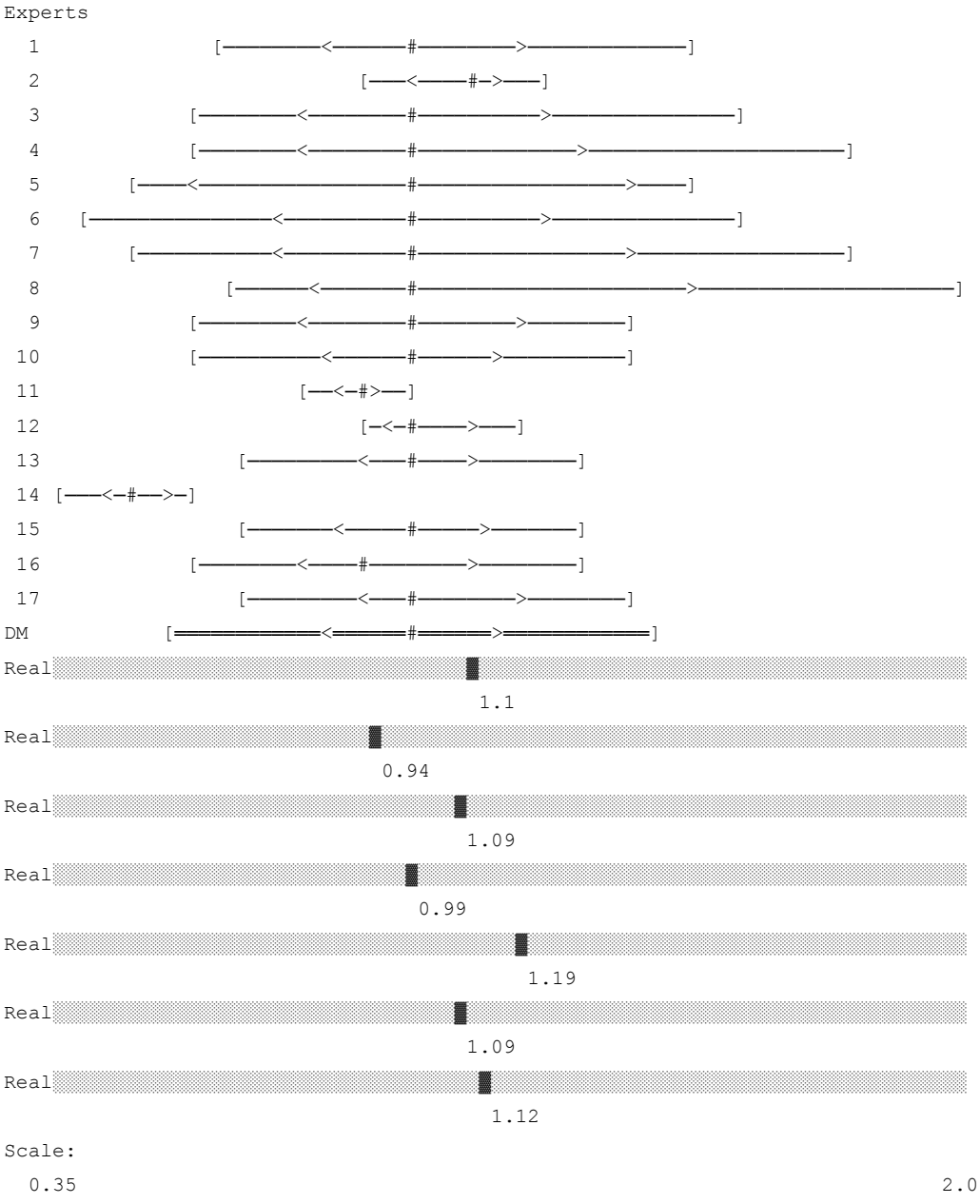

Figure 1. Range graph for model factor for significant wave height and 7 realizations; ‘[‘ and ‘]’ denote 5% and 95% quantiles respectively, ‘<’ and ‘>’ denote 25% and 75% quantiles respectively, ‘#’ denotes the median, ‘dm’ is the item weight decision maker (see Table 2).

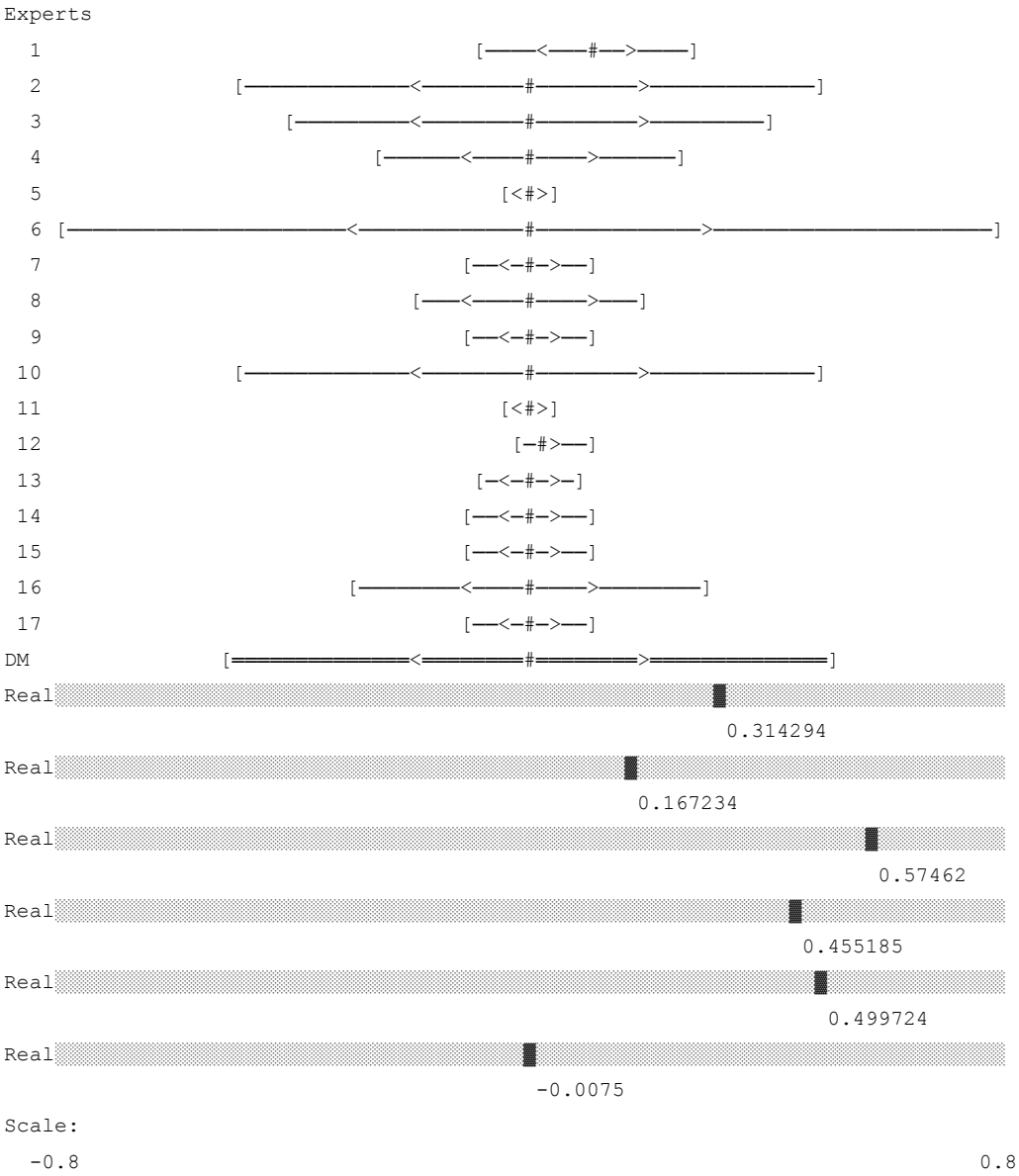

Figure 2. Range graph for model term for local water level model Zwendl and 6 realizations from the measuring station Dordrecht; '[' and ']' denote 5% and 95% quantiles respectively, '<' and '>' denote 25% and 75% quantiles respectively, '#' denotes the median, 'dm' is the item weight decision maker (see Table 2).

Expert nr	Average relative information w.r.t. equal weight DM	
	All variables	Seed variables Only
1	0.729	0.476
2	0.578	0.574
3	0.380	0.328
4	0.687	0.677
5	0.784	0.596
6	0.428	0.368
7	0.523	0.539

8	0.478	0.444
9	0.519	0.510
10	0.427	0.318
11	0.837	0.723
12	1.039	1.093
13	0.601	0.402
14	1.148	0.999
15	0.410	0.375
16	0.489	0.272
17	0.440	0.412

Table 1. Average relative information of experts with respect to the equal weight decision maker, for all variables and for seed variables only.

Expert nr	Calibr. (effective nr seeds = 47)	Calibr. (effective nr seeds = 9)	Mean relative information wrt background measure Total	Seeds	Un-normalized weight	Normalized weight with itemwgt DM
1	0.00010	0.30000	1.777	1.103	0.00011	0.00024
2	0.00010	0.30000	1.533	1.244	0.00012	0.00027
3	0.00010	0.40000	1.146	0.803	0.00008	0.00018
4	0.00010	0.10000	1.457	1.449	0.00014	0.00032
5	0.00010	0.05000	2.007	1.568	0.00016	0.00035
6	0.02500	0.80000	0.797	0.430	0.01075	0.02367
7	0.00010	0.01000	1.065	0.951	0.00010	0.00021
8	0.00010	0.40000	1.353	1.054	0.00011	0.00023
9	0.00010	0.05000	1.498	1.410	0.00014	0.00031
10	0.30000	0.95000	1.171	0.648	0.19433	0.42772
11	0.00010	0.00500	2.152	2.132	0.00021	0.00047
12	0.00010	0.00100	2.619	2.460	0.00025	0.00054
13	0.00010	0.10000	1.884	1.526	0.00015	0.00034
14	0.00010	0.00050	2.402	2.053	0.00021	0.00045
15	0.00010	0.10000	1.247	1.240	0.00012	0.00027
16	0.00100	0.60000	1.381	0.827	0.00083	0.00182
17	0.00010	0.10000	1.168	1.120	0.00011	0.00025
Item wgt DM	0.40000	0.95000	1.039	0.616	0.24642	0.54237
Global wgt DM	0.40000	0.90000	1.000	0.572	0.22865	0.52717
Equal wgt DM	0.05000	0.80000	0.955	0.760	0.03798	0.15444
Item wgt K = 50	0.40000		1.289	0.8527	0.3411	0.5314
Global wgt K = 50	0.30000		1.470	0.9341	0.2802	0.5272
Equal wgt K=50	0.10000		1.218	1.014	0.1014	0.2521

Table 2. Results of scoring experts and decision makers

Excluded Expert	Rel.Inf.to Background		Calibration	Average Rel.Inf.to original DM	
Name	total	Seeds only		Total	Seeds only
None	1.039	0.616	0.4000	0	0
1	1.037	0.616	0.40000	0.001	0.000
2	1.034	0.616	0.40000	0.000	0.000
3	1.039	0.616	0.40000	0.000	0.000
4	0.921	0.616	0.40000	0.007	0.000
5	1.038	0.616	0.40000	0.001	0.000
6	0.968	0.512	0.30000	0.062	0.028
7	1.039	0.616	0.40000	0.000	0.000
8	1.030	0.602	0.40000	0.001	0.002
9	1.039	0.616	0.40000	0.000	0.000
10	0.951	0.625	0.02500	0.350	0.280
11	1.034	0.616	0.40000	0.001	0.000
12	1.037	0.615	0.40000	0.001	0.001
13	1.038	0.616	0.40000	0.000	0.000
14	1.037	0.613	0.40000	0.001	0.001
15	1.020	0.616	0.40000	0.001	0.000
16	1.044	0.616	0.40000	0.008	0.002
17	0.975	0.552	0.40000	0.005	0.002

Table 3. Robustness of the item weight dm against choice of experts.

Frequency yearly maximum North Sea ≥ 4.5m	**5%**	**25%**	**50%**	**75%**	**95%**
in house	0.00015	0.00029	0.00040	0.00068	0.00118
item weights	0.00002	0.00014	0.00053	**0.00184**	**0.00299**
Equal wgts	0.00001	0.00022	0.00056	**0.00162**	**0.00450**

Sea level rise in 100 yr: increment to 2.5m water level	**5%**	**25%**	**50%**	**75%**	**95%**
item weights	-0.003	0.19	0.39	0.60	0.81
equal weights	**-0.28**	0.22	0.36	0.58	0.99

100yr maximum North Sea level	**5%**	**25%**	**50%**	**75%**	**95%**
item weights	3.49	3.89	4.10	4.50	5.00
equal weights	3.12	3.80	4.24	4.66	5.50

Probability yearly maximum Rhine discharge ≥ 16000 m3/s	**5%**	**25%**	**50%**	**75%**	**95%**
item weights	1.10E-05	1.58E-04	6.27E-04	1.57E-03	4.11E-03
equal weights	8.52E-06	1.59E-04	7.06E-04	1.72E-03	5.15E-03

Modelterm Zwendl for prediction = 4m	**5%**	**25%**	**50%**	**75%**	**95%**
in house	-0.25	-0.1	0	0.1	0.25
item weights	**-0.91**	**-0.39**	0.00	**0.39**	**0.94**
equal weights	**-0.61**	-0.15	0.04	0.22	**0.65**
Modelfactor wave height for prediction = 1m	**5%**	**25%**	**50%**	**75%**	**95%**
in house	0.83	0.93	1.00	1.08	1.20
item weights	0.58	0.85	1.00	1.15	1.43
equal weights	0.54	0.85	1.00	1.17	1.57
Modelfactor wave periode for prediction = 4s	**5%**	**25%**	**50%**	**75%**	**95%**
in house	3.34	3.71	4	4.31	4.8
item weights	2.77	3.40	4.49	4.60	5.33
equal weights	2.40	3.54	4.06	4.63	**6.25**
Critical discharge for prediction = 50l/s/m	**5%**	**25%**	**50%**	**75%**	**95%**
in house	23	36	50	69	109
item weights	18	34	49	**118**	**207**
Equal weights	12	34	64	**115**	**185**
Occurring discharge for prediction = 100l/s/m	**5%**	**25%**	**50%**	**75%**	**95%**
in house	46	73	100	137	218
item weights	30	62	100	150	240
equal weights	**23**	70	101	130	**526**

Table 4. Comparison of in house, item weight dm and equal weight dm for items of interest

EXERCISES

1. Suppose that a dike section has a probability of 8×10^{-6} per year of failing. Suppose a dike ring consists of 50 dike sections, and suppose that the failures of different sections are independent. What is the probability per year that the dike ring fails?
2. (Continuation) Instead of assuming independence, suppose that the event that one dike ring fails is the conjunction of two equally probable events, one event concerns only factors local to the given dike section and is independent of similar events at other dike sections; the other event concerns meteorlogical factors factors which affect all dike sections in the same way. What is the probability that the dike ring fails?
3. Two experts have assessed their 5%, 50% and 95% quantiles for 10 continuous variables for which the realizations have been recovered. For the first expert, 3 realizations fall beneath his 5% quantile, 2 fall between the 5% and 50% quantile, 2 fall between the 50% and 95% quantile, and 3 fall above the 95% quantile. For the second expert, 1 realization falls below the 5% quantile, 7 fall between then 5% and 50% quantiles, one falls between the 50% and 95% quantile, and 1 falls above the 95% quantile. Compute the calibration scores for these two experts.
4. Two experts assess their 5%, 50%and 95% quantiles for an unknown relative frequency. For the first expert these quantiles are 0.1, 0.25, 0.4; and for the second expert these quantiles are 0.2, 0.4, 0.6. Use a uniform backgroun d meaure on the interval [0, 1] and compute the Shannon relative information in ech expert's assessment using the minimum information density for each expert, subject to the quantile constraints.
5. (Continuation) Compute the equal weight combination for the two experts in exercise 4, and compute the Shannon relative information of this combination relative to the uniform background measure on [0, 1].
6. (Continuation) Use the calibration scores from exercise 3 and the information scores from exercise 4 to compute the item-weight combination for the variable in exercise 4. What is the Shannon relative information for this combination relative to the uniform background measure on [0, 1]?