Chapter 2

Structured Elicitation of Expert Judgment for Probabilistic Hazard and Risk Assessment in Volcanic Eruptions

W. P. Aspinall

When a potentially dangerous volcano becomes restless, civil authorities invariably turn to scientific specialists to help them anticipate what the volcano will do next, and to provide them with guidance as to the likely threats. Although it is usually possible to discern the earliest signs of unrest, the science of forecasting the course and timing of eruptions remains inexact.

In this chapter, recent volcanic crises in the Eastern Caribbean are recounted in order to trace the emergence of a need for volcanologists to formalise the way they present scientific advice in such circumstances. The conversation then moves on to the concepts and principles of eliciting expert opinion, and structured elicitation within a mathematical framework, before describing in more detail a specific performance-based procedure for eliciting opinions that relies on proper scoring rules. Ways in which this procedure and its scoring basis have been adapted for use in the recent Montserrat volcanic crisis are discussed, and the purposes for which the formalized procedure has been used during that eruption, in application to hazard and risk management, are described. Finally, a few general observations are offered in respect of the benefits and limitations of using a

structured procedure for eliciting scientific opinion in the unique and special circumstances of a volcanic eruption crisis.

2.1 Volcanological Background

Volcanology began to evolve into a modern, multidisciplinary science at the beginning of the 20th century with fatal explosions at two volcanoes in the Eastern Caribbean (one on St. Vincent and one on Martinique), and at a third in Guatemala, when all three erupted within 6 months of one another in 1902. In a matter of only minutes at each, glowing avalanches (pyroclastic flows - turbulent avalanches of superheated gas and ash) killed about 36,000 people in total. In particular, the devastation of the tropical city of St. Pierre, Martinique, convinced scientists of the time of the need to understand the processes of these fast-moving hot avalanches which could kill thousands of people with so little warning, a phenomenon until then unrecognized and inexplicable. In the case of the eruption of Montagne Pele alone, as many as 29,000 people died when, amongst other factors contributing to the death toll, political priorities took precedence over public concerns.

Then, in 1976, Guadeloupe, another French island in the Eastern Caribbean, went into crisis when abnormal levels of volcanic unrest developed. In a desire to prevent a repetition of the Martinique disaster, and expressly to avoid casualties at almost any cost, the authorities in 1976 initiated a major population evacuation that lasted several months. However, on this occasion the eruption was stillborn (Feulliard *et al.* 1983). At the height of the Guadeloupe crisis, scientists became embroiled in public controversy with each other, with the media and with politicians. Then, after activity diminished, and official anxieties had receded, severe criticisms were levelled at the volcanologists and their advice from many quarters, including other scientists. This led to bitter recriminations amongst the scientists involved and, for volcanology as a science, an unsatisfactory and debilitating loss of credibility.

In such urgent circumstances, striving for scientific consensus invariably proves to be very difficult, if not impossible, and divergences of opinion inevitably develop. The authorities become frustrated with the perceived indecision of the scientists, and journalists have a field day with apparently conflicting views. It was in the light of the unfortunate experiences in Guadeloupe that the suggestion was made (Aspinall and Woo 1994) to consider using a formalized procedure for eliciting volcanologists' opinions during a crisis. The essence of this proposal was to make use of a particular structured procedure, called the "Classical Model", which had been developed for the European Space Agency for risk assessment applications (see

below).

Not long after this suggestion, the Soufrière Hills volcano on the island of Montserrat came to life in July 1995, after 400 years of quiescence, with several steam-venting explosions. Activity then grew gradually more and more violent with the generation of lethal pyroclastic flows and explosive outbursts of incandescent ash, gas and steam. Because the volcano occupies one-third of a small island, only 80 sq km in area, this activity gave rise to major safety concerns for the population. Even though the volcano was to become one of the most closely monitored in the world, with arrays of sophisticated monitoring equipment, the scientists working there still entertained a wide range of opinions about what the volcano might do next. In attempting to furnish good advice to the decision-makers, ideally in the form of consensus, intrinsic scientific difficulties were compounded by the diversity of specialisms and experiences and, as time passed, the fluctuating levels of involvement of individual members of the monitoring team. Thus, the protracted Montserrat crisis has exemplified all the challenges of forecasting volcanic hazards in the face of scientific uncertainty and in the context of safety-critical decision-making. In order to systematize this important aspect of the team's work, a formalized procedure for eliciting expert judgements was adopted for the first time in a volcanic emergency.

2.2 Elicitation of Opinions

Several approaches are available for the elicitation and aggregation of individual experts' judgments, some of which can be denoted as "behavioural", others "mathematical" (Clemen and Winkler 1999). The so-called mathematical methods seek to construct a single 'combined' assessment for each variable or item of interest, one by one, by applying procedures or analytical models that treat each of the separate variables autonomously. The behavioural aggregation methods, on the other hand, try to find homogeneity in information of relevance to the experts' assessments - across all the variables of interest - by getting experts to interact together. Through this interaction, some behavioural approaches, e.g., the expert information approach (Kaplan 1992), aim to obtain a clear agreement among the experts on the ultimate probability density function produced for each and every variable. In other approaches, such as those described by Budnitz *et al.* (1998) or by Keeney and Von Winterfeldt (1989), the interaction process is then followed by some form of elementary mathematical combining of the individual experts' assessments in order to obtain one single (aggregated) probability density function per variable. Typically, these approaches rely on very simple combination schemes, such as ascribing equal weighting to all the participating experts.

The mathematical approaches (with some component of modelling) and the behavioural approaches both seem to provide results that are inferior to simple mathematical combination rules (Clemen and Winkler 1999). Furthermore, a group of experts tends to perform better than a solitary expert. That said, however, it is sometimes found that the best individual in a group can still outperform the group as a whole (Clemen and Winkler 1999). This motivates the adoption of procedures that elicit assessments from individual experts - without interaction between them during the actual elicitation step itself - followed by simple mathematical aggregation in order to obtain a single assessment per variable. In this way, assessments by individual experts are obtained in a neutral approach and given different weights according to each expert's performance and merit. In the present discussion, consideration is given to one performance-based formalized procedure for the elicitation of expert judgements that has emerged from the Delft University of Technology, in The Netherlands. Whilst other approaches to complex decision problems exist, some of which are more rooted in the foundational approach of Bayesian modelling (see, e.g., French 1988; Goldstein and O'Hagan, 1996; O'Hagan, 1998), the Delft methodology is highlighted here because it has been applied extensively as a decision-support tool in many safety-critical risk assessment situations, and has now been used intensively in a volcanic eruption crisis.

2.3 Introduction to the Delft Procedure

Following Cooke (1991), over the last 10 years the Delft University of Technology has developed several methods and tools to support the formal application of expert judgement (see also Cooke and Goossens 2004), including the computer software EXCALIBR (Cooke and Solomatine 1992) for conducting numerical analysis of elicitations. Applications have included consequence assessments for both chemical substances and nuclear accidents, and case histories exist now in several other fields of engineering and scientific interest. The techniques developed by Delft can be used to give quantitative assessments, or qualitative and comparative assessments. The former give rise to assessments of uncertainty in the form of probability distributions, from which nominal values of parameters can be derived for practical applications. The latter lead to rankings of alternatives. The application of these techniques is underpinned by a number of principles, including openness to scrutiny (all data and all processing tools are amenable to peer review and results must be reproducible by competent reviewers), fairness (experts are not pre-judged), neutrality (methods of elicitation and processing should not bias results), and performance control (quantitative assessments are subjected to empirical quality controls).

The overall goal of these formal methods is to achieve rational consensus in the resulting assessments. This requires that the participants and stakeholders 'buy into' or 'take ownership' of the process by which the results are reached, and that the process itself optimizes the participants' performance, as measured by some valid functional criterion. Performance criteria are based on control assessments, that is, assessments of uncertain quantities closely resembling the variables of interest, for which true values (e.g., from experiments or observation) are known, or become known *post hoc*. Criteria for analysing control assessments are closely related to standard statistical methods, and are applied both to expert assessments, and to the combinations of expert assessments. The use of empirical control assessments is a distinctive feature of the Delft methods, and the underlying methodology is described in "A Procedure Guide for Structured Expert Judgement", published by the European Commission as EUR 18820 (Cooke and Goossens 2000).

The resources required for an expert judgment study vary greatly, depending on the size and complexity of the case. Studies undertaken thus far have used as few as four and as many as fifty experts. The amount of expert time required for making the assessments depends on the subject and may vary from a few hours to as much as a week, for each participating expert. A trained uncertainty analyst (or 'facilitator') is essential for defining the issues, guiding the progress of the elicitation and processing the results. In the past, the total man-power time required for such studies has varied between one man-month and one man-year, although in certain special applications (e.g. volcano monitoring) the commitment may be condensed into shorter intervals. Other factors determining resource commitments are travel, training given to experts in subjective probability assessments, and the level of supporting documentation produced. However, post-elicitation processing and presentation of results are greatly facilitated by software support, such as that provided by the EXCALIBR program.

2.4 Structured Expert Judgment

Expert judgement has always played a large role in science and engineering. Increasingly, expert judgement is being recognized as a particular form of scientific data, and formal methods are emerging for treating it as such. This section gives a brief overview of methods for utilizing expert judgement in a structured manner - for more complete summaries see Granger Morgan and Henrion (1990), Cooke (1991), or Meyer and Booker (2001).

In the realms of science and engineering, technical expertise is generally kept separate from value judgements. 'Engineering judgement' is often used to bridge the gap between hard technical evidence and mathemati-

cal rules on the one hand and the unknown, or unknowable, characteristics of a technical system on the other. Numerical statements or evaluations, that are tantamount to data, have to be derived which are suited to the practical problem at hand, and engineers are usually able to provide these, essentially subjective, data through insights from engineering models and from experience. The same is true for scientific 'expert judgement': models and experience largely inform the subjective experts' assessments, which is why certain specialists acquire recognized expertise in certain subject fields. Skipp and Woo (1993) take the conversation further, however: they argue that expert judgment should be distinguished from engineering judgement on the grounds that the former is, and must be, clearly anchored in a formal probabilistic framework, whereas the latter often lacks that attribute.

2.4.1 Point Value Estimations

With many elicitation procedures, most notably the earlier Delphi method (Helmer, 1966), experts are asked to speculate as to the values of unknown quantities - their answers are single point estimates. When these unknown values become known through observation, the observed values can be compared back to the estimates, and adjustments made accordingly. There are several reasons why this type of assessment is no longer in widespread use, which Cooke and Goossens (2004) summarise. First, any comparison of observed values and estimates must make use of some scale on which the values are measured, and the method of comparison must incorporate the same properties of that scale. In other cases, values are fixed only as regards rank order (i.e. on an ordinal scale); a series of values may contain the same information as the series of logarithms of values, etc. To be meaningful, the measurement of discrepancy between observed and estimated values must have the same invariance properties as the relevant scales on which the values are measured. In other words, the meanings of descriptors like 'close' and 'far away' are scale dependent, which makes it very difficult to combine scores for different variables measured on different scales.

A second and, in the present context, critical disadvantage with point estimates is that they give no indication of uncertainty. Expert judgement is typically applied when there is substantial uncertainty regarding true values and, in such cases, it is almost always essential to have some picture of the extent of the uncertainty present in the assessments.

A third disadvantage is that methods for processing and combining judgements are typically derived from methods for processing and combining actual physical measurements. This has the effect of treating expert assessments as if they were physical measurements in the normal sense, which they are not. On the positive side, however, point estimates are easy

to obtain and can be gathered quickly - thus, these types of assessment will always have some place in the world of the expert, if only in the realm of the "quick and easy". More detailed psychometric evaluations of Delphi methods are given by Brockhoff (1975), and for a review of the mathematical probity of the approach, see Cooke (1991).

2.4.2 Discrete Event Probabilities

As discussed by Cooke and Goossens (2004), an uncertain event is one that either occurs or does not occur, though we do not know which *a priori*: the archetypal example is "will it rain tomorrow?". Experts are often asked to assess the probability of occurrence of such events, with the assessment usually taking the form of a single point value in the interval [0,1), with a separate value for each uncertain event. The assessment of discrete event probabilities needs to be distinguished from the assessment of 'limit relative frequencies of occurrence' in a potentially infinite class of experiments (the so-called reference class). The variable 'limit relative frequency of rain in days for which the average temperature is 20 degrees Celsius' is not a discrete event. This is not something that either occurs or does not occur; rather this variable can take any value in [0,1), and under suitable assumptions the value of this variable can be measured approximately by observing large finite populations. If 'limit relative frequency of occurrence' is replaced by 'probability', then careless formulations can easily introduce confusion and misleading outcomes. Misunderstanding is avoided by carefully specifying the reference class whenever discrete event probabilities are not intended.

Methods for processing expert assessments of discrete event probabilities are similar in concept to methods for processing assessments of distributions of random variables. For an early review of methods and experiments see Kahneman *et al.* (1982); for a discussion of performance evaluation see Cooke (1991).

2.4.3 Continuous Uncertain Quantities

When it comes to modelling and other applications involving uncertainty analysis, concern is mostly with random variables taking values in some continuous range. Strictly speaking, the notion of a random variable is defined with respect to a specific probability measure in probability space, hence the term 'random variable' entails a distribution. Therefore the term 'uncertain quantity' is usually preferred - an uncertain quantity assumes a unique real value, but it is not certain what this value is: the uncertainty is described by a subjective probability distribution.

In the present context, specific interest focuses on cases in which the uncertain quantity can assume values in a continuous range. An expert is confronted with an uncertain quantity, say X, and is asked to specify information about his subjective distribution over the possible values of X. The assessment may take a number of different forms. The expert may specify his cumulative distribution function, or his density or mass function (whichever is appropriate). Alternatively, the analyst may require only partial information about the distribution. This partial information might be the mean and standard deviation, say, or it might be values for a number of quantiles of the distribution. For r in $[0, 1)$, the rth quantile is the smallest number x_r such that the expert's probability for the event $X \leq x_r$ is equal to r. The 50% quantile is the median of the distribution. Typically, only the 5%, 50% and 95% quantiles are requested, and distributions are fitted to the elicited quantiles.

When expert judgement is cast in the form of distributions of uncertain quantities, the issues of conditionalization and dependence are important. When uncertainty is quantified in an uncertainty analysis, it is always uncertainty *conditional on something*. Thus it is essential to make clear the background information conditional on which the uncertainty is to be assessed. For this reason, the facilitator should ensure that a clear 'case structure' is always provided. Failure to specify background information can lead to different experts conditionalizing their uncertainties in different ways or on different assumptions, and this can introduce unnecessary 'noise' or scatter into the elicitation process.

The background information will not specify values of all relevant variables. Obviously relevant but unspecified variables should be identified, though an exhaustive list of relevant variables is rarely possible. Uncertainty caused by unknown values of unspecified variables must somehow be 'folded into' the uncertainty of the target variables. This is an essential task, and responsibility, of the experts when developing their assessments. Variables whose values are not specified in the background information can cause latent dependencies in the uncertainties of target variables.

Dependence in uncertainty analysis is an active research topic, and methods for dealing with dependence are still very much under development. It is sufficient here to say that the analyst must identify ahead of time those groups of variables for which significant dependencies may be expected, and must query experts about these via their subjective distributions for the variables in question. Methods for doing this are discussed in Cooke and Goossens (2000), and Kraan and Cooke (2000).

2.5 Performance-based Measures of Expertise

For deriving uncertainty distributions over model parameters from expert judgements, the so-called Classical Model has been developed in Delft (Bedford and Cooke, 2001). Other methods to elicit expert judgements are available, for instance for seismological risk issues (Budnitz *et al.* 1998) and for nuclear safety applications (USNRC 1990). The European Union recently finalized a benchmark study of various expert judgement methods (Cojazzi *et al.* 2000). In a joint study by the European Community and the US Nuclear Regulatory Commission the benefits of the latter method - the so-called NUREG-1150 method (Hora and Iman 1989) - have been adopted, incorporating many elements of the Classical Model (Goossens and Harper 1998).

The name 'Classical model' derives from an analogy between calibration measurement and classical statistical hypothesis testing, and the approach provides a basis for performance-based linear pooling or weighted averaging of subjective opinions from a group of experts. The weights are derived from the experts' calibration and information performances, as measured against so-called 'seed' variables. These are variables from the experts' field of specialization whose values become known, or are 'realised', *post hoc*. Questioning experts about their beliefs with respect to seed variables serves *a threefold purpose*: *i*) to quantify their performance as subjective probability assessors, *ii*) to enable performance-optimized combinations of expert distributions to be generated, and *iii*) to evaluate and hopefully validate the resulting combination of expert judgements.

The methodology implementing the Classical Model contains three alternative weighting schemes for aggregating the distributions elicited from the experts. These weighting schemes are denoted as: 'equal weighting', 'global weighting', and 'item weighting', and are distinguished by the ways in which weights are assigned to the uncertainty assessments of each expert. The 'equal weighting' aggregation scheme assigns equal weight to each expert. If N experts have assessed a given set of variables, the weights for each density are $1/N$; hence for variable i in this set the (equal weights) decision maker's CDF is given by:

$$F_{ewdm,j} = \left(\frac{1}{N}\right) \sum_{j=1}^{N} f_{j,i} \tag{2.1}$$

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

'Global' and 'item-based' weighting techniques are termed performance-

based weighting techniques because weights are developed based on an expert's performance against a set of seed variables. Global weights are determined, per expert, by combining the expert's calibration score with his information score to provide an overall 'expert weight'. The calibration score is determined for each expert by their performance in assessing a set of seed variables. The expert's information score is related to the width of their expressed uncertainty band, and by the location of their median choice relative to the seed realization. As with global weights, item weights are determined by the expert's calibration score, but whereas global weights are determined by expert only, item weights are determined jointly by expert and by variable in a way that is sensitive to the expert's informativeness for each separate variable.

As just mentioned, the performance-based expert weight uses two quantitative measures of performance: 'calibration' and 'informativeness'. Calibration measures the statistical likelihood that a set of empirical results corresponds, in some statistical sense, with the experts' assessments (for more detail, see Cooke, 1991, and Bedford and Cooke, 2001, from which sources the essentials of the following discussion are drawn).

At the heart of Cooke's "classical" model is the following statistical concept: given a set of known (or knowable) seed items, test the hypothesis H_0: "This expert is well calibrated". For a set of seed questions, and for each expert, this leads to a likelihood of acceptance at some defined significance level. The steps to finding where this confidence level falls for each expert start by first getting him to provide quantile information for his uncertainty distributions on each of the several seed variables that make up the set of calibration questions. If, say, an expert gives 90% confidence bands for a large number of such variables, then it might be anticipated that about 10% of actual realizations will actually fall outside his chosen bands. Thus, for an expert assessing 20 variables for which realizations become known *post hoc*, three or four outcomes falling outside the relevant bands would be no cause for alarm, as this can be interpreted as sampling fluctuation - in other words, the underlying hypothesis would receive support from the performance evidence. If, however, ten or more of the 20 variables fell outside the individual expert's bands, it would be difficult to assert that so many outliers could result from chance fluctuations, and it would be more reasonable to infer that the expert either grossly mislocates his bands, defines them too narrowly, or both.

Suppose, for illustration, the expert's uncertainty distribution on a particular variable is partitioned into four intervals: 0%-5%; 5%-50%; 50%-95%; and 95%-100%. If he or she is very well calibrated, then about 5% of the realizations of the seed questions might fall in to the expert's intervals represented by 0%-5%, about 45% of realizations should coincide with

his interval 5%-50%, and so on. If it is observed on the basis of N calibration variables that $s_1 N$ realizations fell into the 0%-5% interval, $s_2 N$ realizations fell into the 5%-50% interval, etc., then the expert's density is $(s_1, ...s_4)$, and his distribution can be compared with the hypothesized density of $(p_l, ...p_4) = \{0.05, 0.45, 0.45, 0.05\}$, stipulated by the chosen quantiles. Quantifying the expert's discrepancies against the realizations can be used to derive an information or informativeness metric for his performance.

In practice, the individual's information score can be estimated relative to a density function that is uniform, or log-uniform, over some intrinsic range for the variable in question (the log form is usually adopted where experts' value spreads can span several orders of magnitude). In most applications, it is usual to elicit from the experts just their 5%, 50% and 95% quantiles (although others can be used), to represent their expected value and their 90% credibility bounds about this value. The latter need not be symmetric and, commonly, are not. The analyst then defines a slightly wider 'intrinsic range' for each variable by appending small, identical overshoots at each end of the smallest interval such that the nominal 100% range for the variable in question encloses all the experts' quantiles and the relevant realization. The sizes of these supplementary bounding limits are normally decided by the analyst (typically, 10% is added to the range on either side), and this provides approximate representations of the tails of the histogrammic distributions, beyond the 5%ile and 95%ile values that have been elicited.

The resulting probability densities are then associated with the experts' assessments for each target variable in such a way that, *for each expert*, *a*) the ascribed densities agree with the expert's quantile assessments, and *b*) the densities are minimally informative with respect to the background measure.The expert's informativeness is scored per target variable by computing the relative information of the expert's density for that variable with respect to the background measure, using the relation:

$$I_j(s_j, p) = \frac{1}{n} \sum_{i=1}^{n} s_i \ln \left(\frac{s_i}{p_i} \right) \tag{2.2}$$

where s_i is the distribution obtained from the expert on each of the seed variables, and p_i is the background reference density function for each seed, scaled appropriately for the item in question.

Thus in general terms, item-based informativeness represents the degree to which an expert's uncertainty distribution for each variable is concentrated, relative to a selected background measure for that item, and the overall information score for each expert is the average of his information scores over all variables. This is proportional to the information in the expert's joint distribution relative to the joint background measure, under

the assumption of independence between variables. Here, independence in the experts' distributions means that, after seeing a realization for one or more variables, the experts would not revise their distributions for other variables. Scoring calibration and information under the assumption of independence reflects the notion that expert learning is not a primary goal of the process.

The individual's information score is always a positive number, arranged such that increasing values indicating greater information relative to the background measure. Since the intrinsic range depends on the totality of the experts' assessments, this range can change as experts are added or removed, which may exert some influence on the information scores of the remaining experts, although generally this is small. As just noted, information scores are always positive and, *ceteris paribus*, experts with high information scores are favoured over experts with low informativeness.

Turning to calibration performance, statistical likelihood is invoked to measure the degree to which evidence about the expert's beliefs supports the corresponding hypothesis of 'good' calibration. The basis for this is defined in Cooke's formulation by:

$$C_j = 1 - \chi^2_R(2 * N * I(s_j, p) * Power) \tag{2.3}$$

where j denotes the expert, R is number of quantiles, N is the number of seed variables used in calibration, and $I(s_j, p)$ is a measure of information (see above).

Here, C_j corresponds to the asymptotic probability, under the hypothesis, of seeing a discrepancy between s and p at least as great as $I(s_j, p)$ and, for N large, C_j is taken to be approximately chi-squared distributed. The number of degrees of freedom is related to the number of quantiles used to define the expert's distribution (as mentioned above, this is usually three). If the expert provides a judgement distribution s, that is equal to some hypothesized background distribution p, then he (or she) achieves the best possible calibration score of 1.

Thus, in words, the expert's calibration score is the probability that any divergence between his probabilities and the corresponding distributions from observed values of the seed variables might have arisen by chance. A low score (near zero) means that it is likely, in a statistical sense, that the expert's evaluations are 'wrong'. Similarly, a high score (near one, but greater than, say, 0.5) means that the expert's evaluations are statistically well-supported by realizations of the set of seed variables.

With the two measures as a basis, the overall individual expert weights in the Classical model are taken to be proportional to the product of the individual's calibration score (i.e. statistical likelihood) and his informativeness score, where, as noted above, the latter is estimated from all variables

jointly, that is, from both seeds and target variables. Thus, the individuals are awarded weights:

$$W_j = C_j * I_j(S_j, P) \tag{2.4}$$

and these W_j can be normalised across all the experts in a group to obtain relative weights.

"Good" expertise therefore corresponds to good calibration coupled with good informativeness. Relatively speaking, the latter factor is said to be a "slow" function in that large changes in quantile assessments produce fairly small changes in information score whereas a calibration score, on the other hand, is a "fast" function. For example, with a dataset of twenty seed variables and ten experts, and full scoring optimization, calibration scores can typically vary between experts by over four orders of magnitude, whereas information scores seldom vary by more than a factor of three.

A number of analytical elaborations can be associated with the assumptions and numerical procedures underpinning this approach (see Cooke, 1991), one of which needs to be mentioned in the present context. This concerns the power of the statistical hypothesis test, and the estimation of the corresponding chi-squared distribution value. When implementing the basic model for continuous variables, the chi-squared distribution typically has to be computed to four significant places. This entails that calibration scores down to 10^{-4} can distinguished, and that 10^{-4} is the greatest possible ratio of highest-to-lowest calibration scores. As the number of seed realizations increases, individual calibration scores tend to be reduced. Hence, by increasing number of realizations without limit, sooner or later every expert who is not perfectly calibrated will receive the lowest computable calibration score. As this juncture is approached, calibration scores are no longer distinguishable, and individual weights will depend only on information scores.

If it is desired to restore the dominance of calibration over information, then there are two mathematical techniques available for this purpose. First, the accuracy of the numerical routines could be extended, enabling lower calibration scores to be distinguished. However, in real circumstances this is a spurious refinement, of questionable meaningfulness. Alternatively, the power of the test could be reduced by reducing the 'granularity' of the calibration, in other words by replacing the number of seed variables N by some N', where $N' < N$. This ratio is called the power of the calibration test, and N' is the effective number of realizations at power level N'/N. Adopting the second approach draws attention to the fact that the degree to which calibration scores are distinguished, one from another, is strictly a modelling parameter, the value of which can (and ought to be) determined by the analyst. By deciding to fix the numerical accuracy of the test routine, the analyst effectively limits the overall ratio of calibration scores that can be

obtained across the group. If low scores are being caused by a large number of realizations, then the model may be improved by choosing an optimal power level for the circumstances. This issue will be returned to, below, in the context of how pragmatic decisions have been made about choice of model power level for a number of applications of the model. At the very least, failure to recognize the implications of such numerical limitations may lead to poor model performance, and poor decision support.

When it comes to using the scheme for decision support, once choices have been made by the analyst on all issues related to the model, and once the experts have been scored and weighted individually, their opinions can then be pooled together to form a combined expert (or performance-based Decision Maker DM). In fact, the net calibration and informativeness of the DM can also be measured using the same concepts, and the 'opinions' of this entity can be attached to the group, as if it were a synthetic expert. For more detail and discussion, see Cooke *et al.* (1988), Cooke (1991) and Bedford and Cooke (2001); for a diagrammatic representation of the process of pooling experts to obtain a distribution over a target variable, see Figure 2.1.

Thus, in the Delft Classical model, calibration and informativeness are combined to yield an overall score for each expert with the following properties:

1. Calibration predominates over informativeness, and informativeness serves to modulate between more or less equally well-calibrated experts.
2. The score is a long run proper scoring rule, that is, an expert achieves his or her maximal expected score, in the long run, by and only by stating his true beliefs. Hence, the weighting procedure, when regarded as a reward scheme, does not cause the experts to give biased assessments at variance with their real beliefs, in accord with the principle of neutrality.
3. Calibration is scored as 'statistical likelihood with a cut-off'. The measure of individual expertise is associated with a statistical hypothesis, and the seed variables enable measurement of the degree to which that hypothesis is supported by observed data. If this likelihood score is below a certain cut-off point, the expert is 'unweighted'. The use of a cut-off is motivated by property (2) above (however, whereas the theory of proper scoring rules says that there must be such a cut off, it does not say where or at what value the cut-off should be placed).
4. The cut-off value for (un)weighting individuals is determined either by numerically optimizing the calibration and information performance of different combinations of experts, or by constraining it arbitrarily, in some other way (see below).

In applying these concepts, a fundamental assumption of the Classical model (as well as Bayesian models) is that the future performance of experts can be judged on the basis of past performance, as reflected by their scores measured against the set of seed variables. The use of seed variables in this way enables an element of empirical control to be exercised on any related schemes for combining opinions, not just those that optimize performance on seed variables. Therefore, choosing good seed variables is of particular importance in practical terms, and of general philosophical interest, anyway - see Goossens *et al.* (1998) for background and discussion.

In the case of the Montserrat application, a small set of calibration questions was drawn up, in some urgency, that were designed principally to test the participants' reasoning in relation to volcanic hazards and hazard assessment, not in relation to their scientific knowledge or scholarship *per se*. Thus, the calibration exercise sought to adduce judgement expertise in the restricted context of volcanic hazards by requiring the individuals to estimate values and confidence bounds for seed questions that relied on some simple facts drawn from experience as realizations. For example, two such realizations were the percentage of deaths caused globally by pyroclastic flows, and the approximate published economic cost of one well-known evacuation episode.

Once seed questions have been prepared, target variable questions composed, and the experts queried for their opinions on all these items, the next step is to process their responses within the Classical model framework. The numerical procedures needed to do this in practice have been implemented in the software package EXCALIBR (Cooke and Solomatine 1992), which incorporates a capability for computing the fully-optimised scoring scheme and the corresponding results and outcomes. In many applications, however, the owners of an elicitation exercise often have concerns about the extent to which the performance-based calibration is valid or reasonable for their particular problem - and this was true of the Montserrat case. Such concerns can lead to a need in some circumstances to constrain the EXCALIBR decision-maker optimization. The question of how to select the cut-off level for un-weighting individual experts (i.e. deciding when they should receive zero weight) has to be addressed quite frequently in real case studies, and some alternative approaches to the issue are described briefly in the next section.

2.6 Adjusting the Weighting Threshold

As noted above, the cut-off value for un-weighting experts is determined either by optimizing numerically the calibration and information performance of the combined, "synthetic expert" against proper scoring rules, or

by providing a constraining criterion, based on other considerations. In a recent dam safety study in Britain (Brown and Aspinall, 2004), for instance, the outcomes derived from an expert elicitation exercise were obtained after fixing the calibration power and significance level parameters of the hypothesis test so as to (a) ensure that all experts obtain some positive, non-zero weight, and (b) that the ratio between the highest and lowest weights was not too extreme. After discussion with the owners of the survey, the span between the best and poorest performances was fixed, pragmatically, to be no more than two orders of magnitude (i.e. the highest weighting being a factor of 100 times the lowest, or less). This approach, in which the weights of individuals are factored before pooling the whole group, quite strongly moderates the performance optimization of the synthetic decision-maker, and hence curtails the potential weight given to that entity as a virtual expert.

In the dam study case, additional analyses were conducted for the purpose of enhancing the synthetic decision-maker's performance in some realistic sense (but not maximizing it absolutely), such that the harshness of rejection of low-weighted real experts was limited. This was achieved by tuning both the power level of the model and the related significance level setting, which together determine the confidence level for hypothesis rejection upon which the calibration score is based. There is a wide range of possible combinations of settings for these two model parameters and, in the case of the dam study, it was decided that, whatever selections were made, a majority of the group (i.e. for no less than six of the eleven experts) must retain non-zero weights. Supplementary analysis runs were undertaken, therefore, to examine how the elicitation results might change if this position was adopted. The calibration power and significance level were each increased incrementally to allow the analysis to give more weight to the synthetic expert, until the minimum size of a majority quorum, mentioned above, was reached.

The results produced by this unconventional pooling configuration were not dramatically different from those obtained with overall maximization, although there are notable changes in the results for a few items, and hints of systematic shifts in the central value outcomes in several others. Figure 2.2, for instance, illustrates the impact that fixing the calibration power has on the outcome of one target question for which a parameter evaluation was being sought in the dam erosion modelling study - in this case, the calibration power was set such that at least six of the eleven experts retained positive weightings. The observation that differences in outcomes were generally modest is not surprising, however, if it is pointed out that each of the discounted experts had a low individual performance score, and was not exerting much influence on the joint pooling, anyway. What

is significant, however, is that, as a result, greater authority is given to the synthetic decision-maker than would have been the case in a situation where all experts were allowed non-zero scores. The selective unweighting approach represents a shift towards a more homogeneous collective combination of the views of the most influential experts, and a position where the synthetic decision-maker can then out-score most, if not all, of the individual experts. On this basis, it could be argued that results obtained under this 'constrained optimisation' scheme represent a more robust, and more rational, union of opinions than would be provided by making sure the views of the whole group were utilized with non-zero weightings.

Further analysis of the extent to which individual experts contribute to the synthetic DM is possible by activating EXCALIBR's 'expert robustness' option. This is a facility for re-running the model iteratively, dropping one expert at a time, to show what impact his or her omission has on the DM's calibration score and informativeness. In the dam safety case, a breakdown of the contributions of the six positively-weighted experts indicates that three of them had detectable influences on the outcomes: two influenced (in a positive sense) the DM's calibration score, and another exerted particular pressure on the DM's informativeness score. That said, the other three experts also contributed to characterizing the DM, but to an extent that is much less marked, and very similar, one to another.

The particular expert who influences the DM's informativeness presents an interesting example of how traits can emerge from an expert judgement elicitation under the Classical model rubric: his calibration score was fairly good (but not the best), and for ALL items in the subject questionnaire his informativeness measure is also quite good, but not exceptional. However, he had a particularly effective informativeness score for the seed questions, and this significantly enhances his weight and ranking overall. So, in the robustness trial, dropping this particular expert appears to improve the DM's relative calibration score much more than by dropping any of the other experts (including the lowest weighted!). But, in doing so, the DM's informativeness is reduced significantly, too.

Importantly, what this robustness analysis shows is that the virtual DM was not dominated by any single real expert (as has been found occasionally in other applications). Therefore, it was decided for the dam study that the synthetic decision-maker outputs obtained with the 'constrained optimization' control in place should be used for informing the parameterization of the proposed internal erosion model.

Comparable situations have been encountered in other elicitation exercises - e.g. for volcano monitoring (see below), and civil aviation safety studies - and similar tactics for fixing the weighting/unweighting threshold have been adopted to achieve a pragmatic balance of participating experts,

acceptable to the 'owners' of the studies. However, where the dam study owners wished to ensure that a majority of their panel experts received positive weights, other elicitations have simply adopted the criterion that the synthetic DM achieve a weight equal to, or just exceeding, that of the best individual, without reference to how many experts are unweighted as a consequence. This also provides a basis for obtaining a form of rational consensus. Thus, one of the great strengths of the Delft approach, and the EXCALIBR implementation, is that it allows a wide variety of pooling and scoring schemes to be explored quantitatively within any particular elicitation exercise.

2.7 Expert Opinion Elicitation in the Montserrat Crisis

The Montserrat eruption has been going on for approaching ten years, since the crisis first started in July 1995. In that time, more than seventy scientists and technical support staff have been involved in the monitoring activities at the Montserrat Volcano Observatory. Nearly all of these people have participated in opinion elicitation exercises relating to hazard and risk assessment, at one time or another, and have allowed themselves to be scored against the set of calibration seed questions that was hastily drawn up in the first weeks of the crisis. Figure 2.3 summarizes the spread of individual calibration scores (x-axis) and their associated relative overall weights (y-axis) when calibration and informativeness scores are combined, for three different settings of the power level of the Classical model's calibration test: 1.0; 0.5; 0.33 (with Decision Maker optimization switched off). By effectively reducing the number of degrees of freedom in the problem, lowering the calibration test power reduces the data granularity available to distinguish one expert's performance from another. Another net effect of changing the power level from 1.0 to 0.33 is to shrink the span of scores over the group, in this case from a highest/lowest ratio of 48,000:1 down to about 27:1 (see inset). From a practitioner's perspective, scores that are modulated in this way often furnish an 'expertise profile' of a group that is more tolerable to the stakeholders in a typical elicitation.

The synthetic Decision Maker's performance (filled symbols), as computed by EXCALIBR, maintains a near-constant calibration score under these changes in analysis control, but the DM's overall weighting tends to fall, relative to the best individual experts, as a result of reduced informativeness. Because the confidence bands of all participating individuals are more uniformly treated if the calibration power level is reduced, the DM is obliged to faithfully encapsulate and reflect the wider dispersion of all

the contributing extreme values - put another way, the 'noise' level in the tails of the group's opinions, as a whole, is amplified by reducing calibration power. While this may seem undesirable on some grounds, it has the merit of providing a conservative representation of the true extent, within the expert group, of the scientific uncertainty associated with the problem being addressed.

The EXCALIBR approach has been the basis of many scientific decisions relating to hazard assessment during the Montserrat eruption. One example is the regular elicitation of expert judgements on the probabilities of occurrence of future events at the volcano, and the relative likelihoods of alternative scenarios. Figure 2.4 shows a volcanic event tree from the first few months of the crisis, on which probabilities elicited on three separate occasions are recorded. This is a primitive version of the full volcanic event tree that was elicited, with end branches summarising anticipated relative probabilities of potentially lethal pyroclastic flow and surge (labelled 'PF' on chart) hazards in different areas in simplified form. Pyroclastic flows and surges can be generated either by column collapse in an explosive eruption or by dome collapse, and their hazards are very similar in either case; the risk of fatality for a human being in their path is very high - close to 100% lethality.

When used for public information in this form, one important aspect of the event tree representation of opinions is hidden: for each probability, the elicitation analysis provides a measure of the uncertainty involved, expressed usually by an associated 90% confidence spread (as discussed earlier). Such confidence bands allow the meaningfulness of differences between two (or more) alternative branchings to be appraised by decision-makers. They also allow the extent of scientific uncertainty to be represented for the purposes of simulating an ensemble of possible courses that the eruption might follow. However, showing these uncertainty spreads on an event tree for public consumption usually leads to a 'picture' of scientific complexity that is not helpful to general understanding, and so more detailed information is therefore usually restricted to scientific forums and the development of numerically-based risk assessment models.

In the case of risk assessments for Montserrat, the formal expression of ranges of uncertainties, whether their source is data, models or subjective judgement elicitations, is accomplished by representing them in suitable distributional form within a Monte Carlo analysis framework, implemented in a spreadsheet environment. The risk models that have been developed are structured to link volcanic events and their associated hazards, with their probabilities of occurrence, to estimates of their impacts on the population in terms of potential casualties. The latter factors, too, have major uncertainties associated with them, which need to be incorporated in a full

probabilistic model.

Two examples of risk analyses conducted for Montserrat are shown on Figure 2.5. In these examples, the probability of suffering a given number of casualties from the occurrence of any of the eruptive hazards - pyroclastic flow and surge; ballistics and ashfall - is enumerated (lahars are not included). These volcanic hazards have very different levels of lethality for an exposed person, which is allowed for in the risk assessment model. In the left-hand frame, the results of a series of societal risk estimates from December 1997 to December 1999 are shown, in the form of a so-called Frequency-Number (F-N) plot, conventional in risk analysis work. The curves show how the estimated exposure of the population, in terms of potential multiple casualties, varied with conditions at the volcano (strictly, with the scientists' perceptions of those conditions). The right-hand panel of Figure 5 shows an equivalent F-N plot, for September 2001, and illustrates how an effective reduction of risk could be achieved by moving the boundary of the exclusion zone further from the volcano, relating this at the same time to risk exposures from hurricanes and earthquakes, two other natural hazards that the people of Montserrat live with, long term.

Over the course of the last nine years, there have been many such hazard and risk assessments conducted in Montserrat, and at the heart of most of them has been the use of the Delft expert elicitation procedure. Further descriptions of the evolution of the structured elicitation procedure adopted during the eruption crisis, the variety of applications and the different purposes they have been used for, can be found in Aspinall and Cooke (1998), and in Aspinall *et al.* (2002).

2.8 Concluding Remarks

For the first time, a structured procedure for eliciting scientific opinions has been used during a volcanic eruption as an aid to decision-support. In making use of the Delft methods in the Montserrat crisis management situation, a number of principal pros and cons can be identified:

Disadvantages

At present, a major drawback to using the approach in a volcanic crisis is that the concept and principles of subjective probability are not familiar to many scientists (and this is true of many other disciplines). There is, therefore, reluctance among some volcanologists to embrace what are seen as novel techniques for eliciting and pooling expert opinion, especially in urgent crisis situations.

In addition, the means of quantifying an individual's expertise calibration is, perhaps, more difficult to justify in the context of volcanological hazard assessment than in some other more precisely-defined disciplines, such as meteorology. As discussed above, however, reservations on this score can be ameliorated to some extent by constraining the optimization procedure to produce a group performance profile that is acceptable to the stakeholders in the exercise.

Finally, on a practical note, the execution of a formalized and structured procedure for the elicitation of expert opinion calls for the presence of a specialist 'facilitator' to ensure efficiency, correct implementation, and impartiality. In a volcanic crisis, this represents yet another, additional resource requirement, and one that is not easily fulfilled.

Advantages

The procedure allows an inclusive approach to be adopted to the challenge of combining multiple opinions. In the case of Montserrat, this meant that the whole monitoring team could be involved in the decision-making process. This included local technical support staff, for instance, many of who were more fully and, in some ways, more directly involved in confronting the eruption than were some of the visiting scientists. In a traditional hierarchical team structure, it is likely that their views (and those of other junior scientists) would have been disregarded in any meaningful sense.

Furthermore, the approach is un-biased in that the way individuals are polled semi-confidentially encourages each to express his or her true opinion. When the issues involve critical life-or-death decisions, many people can have reservations about expressing their true scientific beliefs in an open forum - even among their peers, let alone in front of politicians, the media or the public. Adopting a formalized elicitation procedure effectively creates an insulating buffer between the scientists and their scientific concerns, and the recipients of the advice that the scientists are required to provide. This arrangement cultivates an ethos of neutrality within the scientific team, and serves to de-personalise the manner of providing guidance to third parties. In other eruption crises, a scientific culture in which self-promoting, opinionated, publicity-hungry individual experts can flourish and exercise undue influence has often led to dubious, singular or dangerous advice being given to, and acted upon, by authorities.

In proper scientific terms, the relevant decision problem(s) can be treated exhaustively: all sources of uncertainty can be identified, and then treated fully and explicitly. A virtue of having a structured procedure like that of the Delft scheme is that, when and where uncertainties are recognized, their existence is preserved formally, for future scrutiny, and the

treatment they have been given is recorded and auditable.

Although some regard it as a weakness in the methodology, the EXCALIBR procedure can occasionally turn up results that are apparently incoherent or implausible. When this happens, it is usually highlighting discrepancies in reasoning, differences in experience, or inconsistencies in interpretation of data or observations. In these circumstances, the procedure acts as a useful diagnostic device, allowing such shortcomings to be identified and addressed.

Another important attribute of the approach for volcanic crisis management is the transparency of the elicitation process. For instance, the approach used in Montserrat accords with new British government guidelines for the provision of scientific advice and with requirements, when assessing science-based issues, to pool as wide a range of expertise as is feasible.

Overall, the introduction and use of formalized expert judgment elicitations during the Montserrat eruption has been favourably received by all the volcanologists involved, and the experience gained during this crisis suggests benefits significantly outweigh any disadvantages.

Further Reading

For an extensive elucidation of the issues surrounding the elicitation of expert opinion and for a full description of the theory and basis for the structured performance-based technique used in the present paper, see Cooke (1991). Expert judgement in the context of decision-making during volcanic crises is one of many related subjects touched upon in an excellent treatise by Woo (1999, Chapter 6) on the wider application of mathematics and probability theory to natural disasters. Information and general accounts of the eruption of the Soufrière Hills volcano, Montserrat, can be found, inter alia, in: Aspinall *et al.* (1998); Young *et al.* (1998); Francis *et al.* (2000), and, more comprehensively, in Druitt and Kokelaar (2002).

Acknowledgements

The author is greatly indebted to his colleagues Dr. Gordon Woo, Prof. Roger Cooke and Prof. Steve Sparks FRS for extensive discussions, sound advice and endless encouragement. He has also benefited from the contributions of colleagues at the Montserrat Volcano Observatory, and from collaborations with the late Capt. John Savage at British Airways, and Alan Brown of Kellogg Brown and Root. Support is gratefully acknowledged from a Benjamin Meaker Visiting Professorship in the Institute for Advanced Studies, Bristol University, from EU Project EXPLORIS Contract no. EVR1-CT-2002-40026, and from the Environmental Mathematics and Statistics Programme funded jointly by NERC/EPSRC, UK, for attendance at the EMS workshop "Statistics in Volcanology". Two

anonymous reviewers kindly provided very helpful comments that improved this contribution.

Bibliography

[1] Aspinall W. & Cooke R.M. 1998. Expert judgement and the Montserrat Volcano eruption. *In:* Mosleh, Ali & Bari, Robert A. (eds) *Proceedings of the 4th International Conference on Probabilistic Safety Assessment and Management PSAM4, September 13th -18th, 1998, New York City, USA*. Springer, 2113-2118.

[2] Aspinall W.P., Lynch L.L., Robertson R.E.A., Rowley K., Sparks R.S.J., Voight B. and Young S.R. 1998. The Soufrière Hills eruption, Montserrat, British West Indies: introduction to Special Section, Part 1. *Geophysical Research Letters*, **25**, 3387-3388.

[3] Aspinall, W.P., Loughlin, S.C., Michael, F.V., Miller, A.D., Norton, G.E., Rowley, K.C., Sparks, R.S.J. & Young, S.R. 2002. The Montserrat Volcano Observatory: its evolution, organisation, role and activities. *In:* Druitt, T.H. & Kokelaar, B.P. (eds) *The eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999*. Geological Society, London, Memoir **21**, 71-91.

[4] Aspinall W.P. & Woo G. 1994. An impartial decision-making procedure using expert judgement to assess volcanic hazards. Accademia Nazionale dei Lincei - British Council Symposium Large Explosive Eruptions, Rome, 24-25 May 1993: Atti dei Convegni Lincei, **112**, 211-220.

[5] Bedford, T.J. & Cooke, R.M. 2001. *Probabilistic Risk Analysis, Foundations and Methods*. Cambridge University Press, Cambridge.

[6] Brockhoff, K. 1975. The performance of forecasting groups in computer dialogue and face to face discussions. *In:* Linstone, H. & Turoff, M. (eds) *The Delphi Method, Techniques and Applications*. Addison Wesley, Reading Mass, 291-321.

[7] Brown A.J. & Aspinall W.P. 2004. Use of expert opinion elicitation to quantify the internal erosion process in dams. *In: Procedings of The*

13th Biennial British Dams Society Conference. University of Kent, Canterbury.

[8] Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornell, C.A. & Morris, P.A. 1998. Use of technical expert panels: applications to probabilistic seismic hazard analysis. *Risk Analysis*, **18(4)**, 463-69.

[9] Clemen, R.T. & Winkler, R.L. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis*, **19(2)**, 187-203.

[10] Cojazzi, G., Fogli, D., Grassini, G. & Coe, I.M. 2000. Benchmarking structured expert judgement methodologies for the assessment of hydrogen combustion in a generic evolutionary PWR. *In:* Kondo, S. & Furuta, K. (eds) *PSAM2 - Probabilistic Safety Assessment and Management 2*, Universal Academy Press, Tokyo, 1151-57.

[11] Cooke, R.M. 1991. *Experts in Uncertainty*. Oxford University Press, Oxford.

[12] Cooke, R.M. & Goossens, L.H.J. 2000. Procedures Guide for Structured Expert Judgement. Report EUR **18820**, Brussels-Luxembourg.

[13] Cooke, R.M. & Goossens, L.H.J. 2004. Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, **7(6)**, 643-656.

[14] Cooke, R.M., Mendel, M. & Thys, W. 1988. Calibration and information in expert resolution: a classical approach. *Automatica*, **24**, 87-94.

[15] Cooke, R.M. & Solomatine, D. 1992. EXCALIBR Integrated System for Processing Expert Judgements, version 3.0. User's manual, prepared under contract for Directorate-General XII, Delft University of Technology, Delft.

[16] Druitt, T.H. & Kokelaar, B.P. (eds) 2002. *The Eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999*. Geological Society, London, Memoir **21**.

[17] Feuillard, M., Allegre, C.J. *et al.* 1983. The 1975-1977 Crisis of La Soufrière de Guadeloupe (F.W.I.): a still-born magmatic eruption. *Journal of Volcanology and Geothermal Research*, **16**, 317-334.

[18] Francis, P., Neuberg, J. & Sparks, R.S.J. (eds) 2000. *The causes and consequences of eruptions of andesite volcanoes - papers of a Discussion Meeting*. Philosophical Transactions of The Royal Society of London, Series A, **358(1770)**.

[19] French, S. 1988. *Decision Theory: an Introduction to the Mathematics of Rationality*. Ellis Horwood, Chichester.

[20] Goldstein, M. & O'Hagan, A. 1996. Bayes linear sufficiency and systems of expert posterior assessments. *Journal of the Royal Statistical Society B*, **58**, 301-316.

[21] Goossens, L.H.J. & Harper, F.T. 1998. Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. *Journal of Radiological Protection*, **18(4)**, 249-264.

[22] Goossens, L., Cooke, R. & Kraan, B. 1998. Evaluation of weighting schemes for expert judgement studies. *In:* Mosleh, A. & Bari, R. (eds) *Proceedings of the 4th International Conference on Probabilistic Safety Assessment and Management PSAM4*. Springer, New York, 1937-1942.

[23] Granger Morgan, M. & Henrion, M. 1990. *Uncertainty. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.

[24] Helmer, O. 1966. *Social Technology*. Basic Books, New York.

[25] Hora, S. & Iman, R. 1989. Expert opinion in risk analysis: the NUREG-1150 methodology. *Nuclear Science and Engineering*, **102**, 323.

[26] Kahneman, D., Slovic, P. & Tversky, A. (eds) 1982. *Judgement under Uncertainty, Heuristics and Biases*. Cambridge University Press, Cambridge.

[27] Kaplan, S. 1992. 'Expert information' versus 'expert opinions'. Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliability Engineering and System Safety*, **35**, 61-72.

[28] Keeney, R.L. & Von Winterfeldt, D. 1989. On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management*, **36(2)**, 83-86.

[29] Kraan, B.C.P. & Cooke, R.M. 2000. Processing expert judgements in accident consequence modelling. *Radiation Protection Dosimetry*, **90(3)**, 311-315.

[30] Meyer, M.A. & Booker J.M. 2001. *Eliciting and analyzing expert judgment: a practical guide*. ASA-SIAM, Philadelphia/Alexandria.

[31] O'Hagan, A. 1998. Eliciting expert beliefs in substantial practical applications. *The Statistician*, **48(1)**, 21-35.

[32] Skipp, B.O. & Woo G. 1993. A question of judgement: expert or engineering? *In:* Skipp, B.O. (ed) *Risk and Reliability in Ground Engineering*. Thomas Telford, London, 29-39.

[33] USNRC 1990. Severe accident risks: an assessment for five US nuclear power plants. Report **NUREG-1150**, Washington, DC.

[34] Woo, G. 1999. *The Mathematics of Natural Catastrophes*. Imperial College Press, London.

[35] Young, S.R., Sparks R.S.J., Aspinall W.P., Lynch L.L., Miller A.D., Robertson R.E.A., Shepherd J.B. & MVO Staff 1998. Overview of the eruption of Soufrière Hills volcano, Montserrat, 18 July 1995 to December 1997. *Geophysical Research Letters*, **25**, 3389-3392.

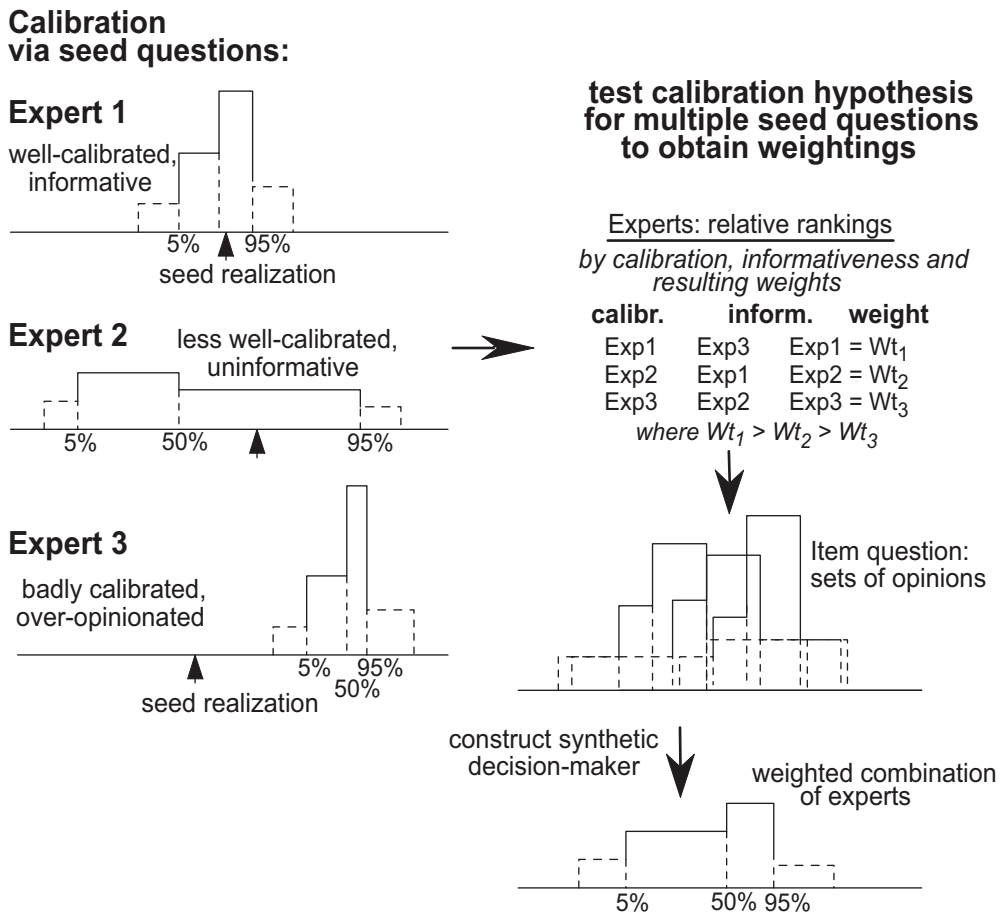

Figure 2.1: Diagram illustrating the basis of the Delft 'classical' expert weighting procedure: performance by a number of experts on a set of seed questions (LH side) leads to individual scoring weights; for the target question, these weights are then used to linearly pool the experts' distributional responses to produce a synthetic decision-maker outcome (see text).

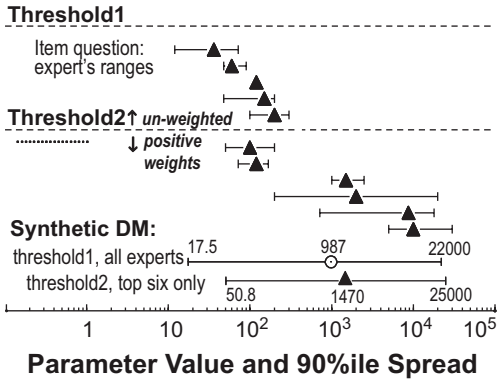

Figure 2.2: Typical experts' responses range graph for an 'item' question, showing the effect on synthetic decision maker results (bottom bars) of fixing the calibration power to 'unweight' the five lowest scoring experts (i.e. those above the Threshold 2 line). Note the high degree of opinionation (overconfidence) exhibited by many of the experts.

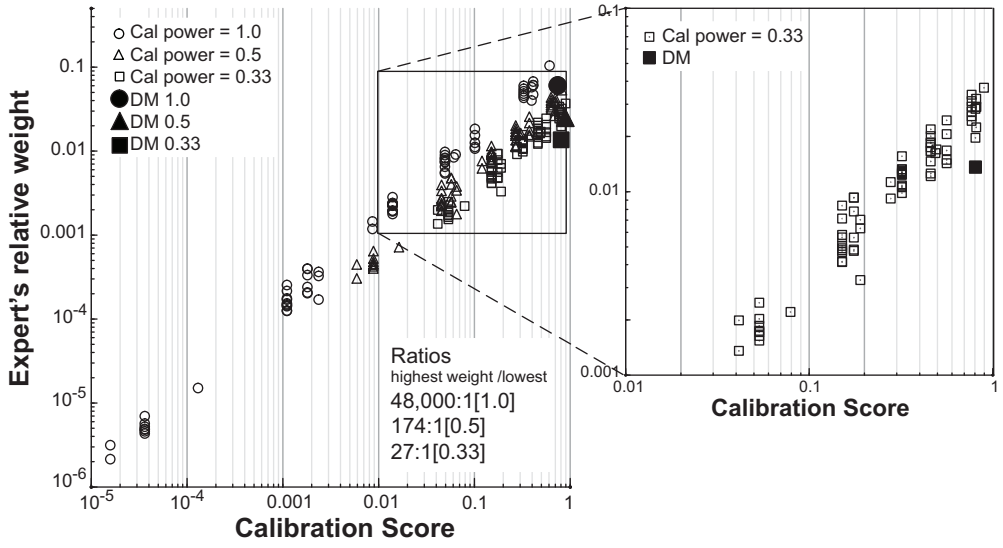

Figure 2.3: Plot showing the calibration scores and overall weights of more than 70 scientists involved in Montserrat, together with the effect on the group profile of adjusting the calibration power. Right-hand panel shows a blow-up of the case when calibration power is reduced by a factor of 3, resulting in a tighter span of weights (see text for discussion of the effect on the DM).

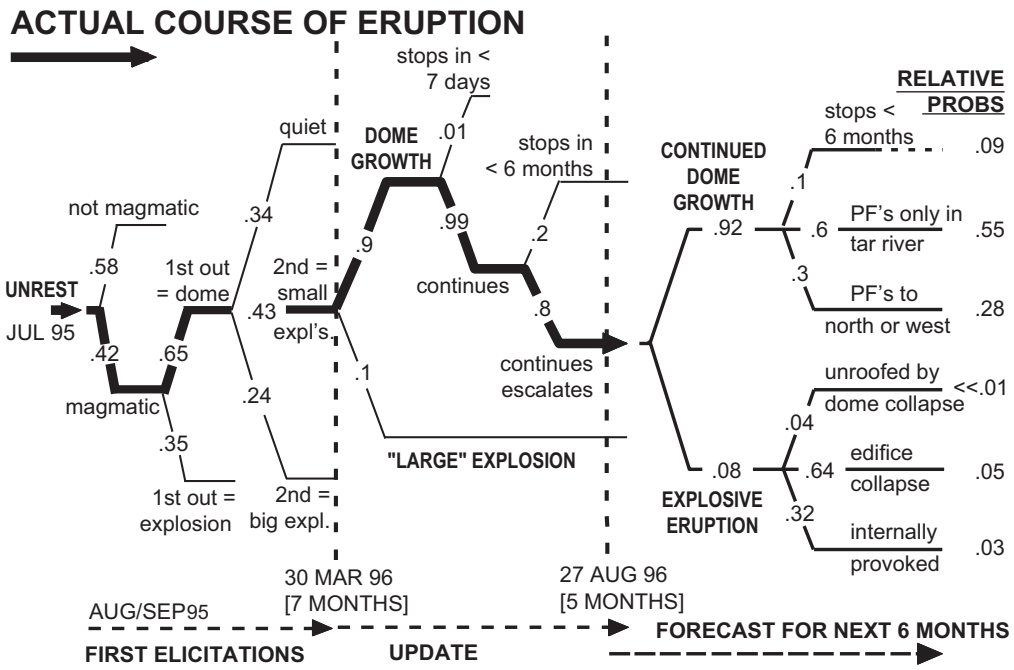

Figure 2.4: Typical simplified volcanic event tree, incorporating probabilities of occurrence of different eruptive events derived from successive expert elicitations. The end branches of this tree summarise the anticipated relative probabilities of potentially lethal pyroclastic flow and surge ('pf') hazards in different areas, in a primitive form. Similar hazards from pyroclastic flow and surges can be generated either by explosive eruption or by dome collapse, and the risk of fatality for a human being in their path is very high - close to 100% lethality.

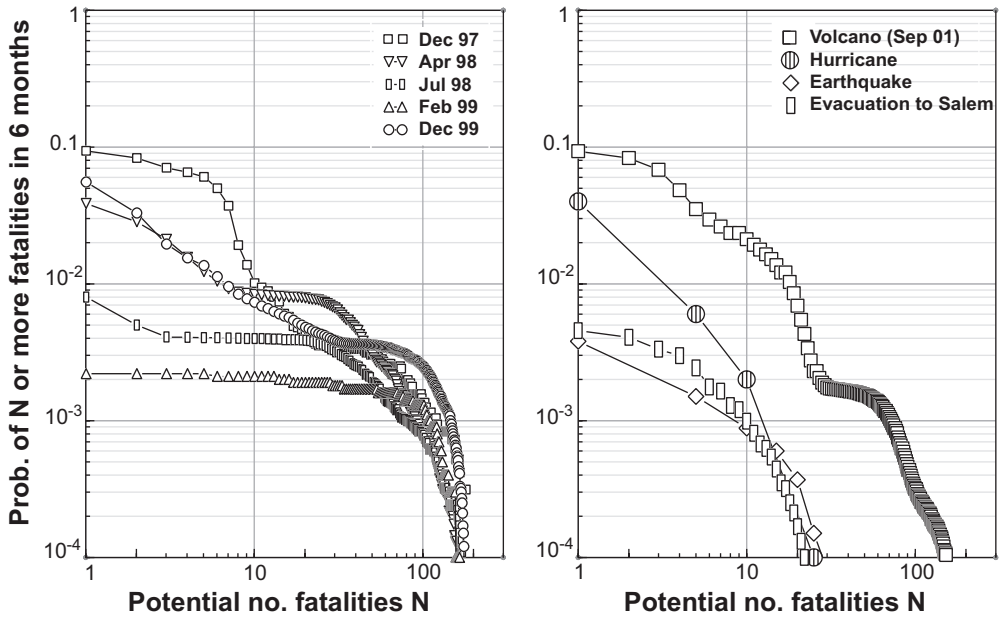

Figure 2.5: Illustrations of probabilistic risk model results for Montserrat, depicted as $F-N$ (frequency-number) population casualty exposure curves. In these examples, the probability of suffering a given number of casualties from any of the eruptive hazards - pyroclastic flow and surge; ballistics and ashfall - is evaluated (lahars are not included here). These volcanic hazards have very different levels of lethality for an exposed person, which is allowed for in the risk assessment. Left-hand panel shows a series of risk curves from Dec 1997 - Dec 1999, and how societal risk exposure changed through time with activity levels at the volcano; the right-hand panel shows the societal risk at Sept 2001, together with long-term risk exposure to hurricane and earthquake on Montserrat, and the effect of increasing the evacuated area to move the boundary further away from the volcano.