ELSEVIER

PII: S0951-8320(96)00060-9

Reliability Engineering and System Safety 53 (1996) 127–138

0951-8320/96/$15.00

On the use of the analytic hierarchy process in the aggregation of expert judgments

E. Zio*

Nuclear Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

(Received 10 March 1996; accepted 17 April 1996)

Expert judgments are involved in many aspects of scientific research, either formally or informally. In order to combine the different opinions elicited, simple aggregation methods have often been used with the result that expert biases, interexpert dependencies and other factors which might affect the judgments of the experts are often ignored. A more comprehensive approach, based on the analytic hierarchy process, is proposed in this paper to account for the large variety of factors influencing the experts. A structured hierarchy is constructed to decompose the overall problem in the elementary factors that can influence the expert's judgments. The importance of the different elements of the hierarchy is then assessed by pairwise comparison. The overall approach is simple, presents a systematic character and offers a good degree of flexibility. The approach provides the decision maker with a tool to quantitatively measure the significance of the judgments provided by the different experts involved in the elicitation. The resulting values can be used as weights in an aggregation scheme such as, for example, the simple weighted averaging scheme. Two applications of the approach are presented with reference to case studies of formal expert judgment elicitation previously analyzed in literature: the elicitation of the pressure increment in the containment building of the Sequoyah nuclear power plant following reactor vessel breach, and the prediction of the future changes in precipitation in the vicinity of Yucca Mountain. © 1996 Elsevier Science Limited.

1 INTRODUCTION

It is a matter of fact that expert judgments are widely used in various aspects of many engineering applications. For example, judgments are made to determine the need for a particular study or research in a complex technical problem, to select the appropriate models and tools for the analysis, and to collect and subsequently interpret the relevant data. Often these judgments are informal and implicitly understood, particularly when there is widespread agreement upon them. A typical example of informal use is the original reactor safety study[1] in which the probability distributions were assigned to many events and quantities based on available data coupled with informal expert judgments. This feature of the analysis was a major issue in the criticism of the study.[2]

In many cases, however, a nonuniformity of opinions occurs so that an explicit treatment of expert judgments is in demand. Indeed, after the reactor safety study, formal consideration of expert judgments has become almost mandatory in any probabilistic risk assessment, PRA. These assessments usually deal with highly rare events and processes for which little statistical or experimental evidence is available. Typical examples of these studies are the safety analysis of nuclear power plants and the performance assessment of high level radioactive waste disposals. In these studies the uncertainties related to the relevant phenomena are such that the use of expert judgments is inevitable, due to the space and time scales required for the analysis and given the present state of knowledge.

A formal expert judgment analysis has, as a primary objective, to estimate the possible values of a given uncertain parameter and properly represent the uncertainty associated with it, in light of the state of knowledge currently available in the scientific community. It is of fundamental importance for the consistency of the entire process that the latter aspect of the analysis, that of a proper representation of the

* Permanent address: Polytechnic of Milan, Department of Nuclear Engineering, Via Ponzio 34/3, 20133 Milan, Italy.

state-of-knowledge uncertainty, be carefully accounted for. It is this aspect, and the subjectivity associated with it, that makes the issue of expert judgments so delicate and controversial, as no closed form solution to the problem exists.

The elicitation and formal use of expert judgments, therefore, inevitably raise several complex issues,[3,4] the most important and perhaps controversial being the selection of experts, the calibration of experts and the aggregation of their judgments.

The selection of experts is a fundamental step in an exercise to use expert judgments and it is itself a matter of judgment. In general, experts should be chosen so that as many differing opinions and assumptions as possible are represented. This is expected to warrant a more realistic assessment of the existing uncertainty. Moreover, experts should have little conflict of interest with the results of the analysis, so as to minimize motivational biases, and they should be independent and well-calibrated.

Experts are often biased and this may lead them to express a response that does not correspond to their true knowledge and beliefs. There are several sources of biases which have been examined and somewhat categorized.[5,6] The assessment and correction of these expert biases by the decision maker is called *calibration*.[7,8] However, for the same reason why expert judgments are used in PRAs, the rarity of events, the calibration of experts is a very difficult task.

In the Bayesian interpretation of probability, as a measure of the 'internal' state of knowledge of an assessor with regards to a given event, different assessors would ideally give the same representation of uncertainty for the given event, provided that they share the same evidence and background knowledge. Unfortunately, in practice, this ideal situation never occurs as experts are obviously characterized by very diverse states of knowledge and inevitably base their assessments on different data and assumptions. This, in turn, leads to estimates which often vary greatly. A problem, then, arises whenever there is a need to combine the diverse estimates provided by the experts into a set of summary measures to be used in decision-making tasks or for further application in the risk assessment.

A variety of aggregation techniques have been proposed, ranging from the easy-to-use methods such as simple averaging[9,10] to the more complex techniques such as the classical model[7] and Bayesian aggregation.[3,10,11] A review of the various aggregation methods used in practice can be found in the work by Cooke,[7] Apostolakis,[12] Lindley[13] and Meyer & Booker.[14]

The goal of this paper is to present a simple, systematic way of expressing the degree of confidence that one can associate to the experts' estimates. This is done quantitatively in the form of numerical weights which can then be used in the aggregation step. The method relies on the analytic hierarchy process, AHP,[15] which is an approach often employed in the field of decision analysis.

The AHP is herein used to develop a hierarchical structure for the evaluation of different experts with regards to the most relevant aspects of the expert judgmental process and performance. The structure is flexible so that allowance can be given to a large variety of contributors which are believed to influence the experts' judgments as well as to other types of measures of experts' performance such as, for example, the experts' *calibration* and *informativeness* of the classical model.[7]

Once the hierarchy is established, it is possible to systematically assign pairwise comparison values for the importance of the different element of the hierarchy and quantitatively resolve the structure to provide the ranking weights for all experts considered.

As a parallel, the approach can be regarded as somewhat similar to the *performance shaping factors* approach often used in human reliability analysis to assess human error rates by formally individuating those factors which can occur in a situation requiring human intervention, and which can influence the performance of an operator.[16]

The proposed approach is applied for the aggregation of expert judgments in two case studies presented in literature, namely the estimation of the pressure increment in the Sequoyah nuclear power plant containment building due to reactor vessel breach[17] and the prediction of future change in precipitation at the Yucca Mountain site for a high level radioactive waste repository.[18]

In Section 2 we outline the fundamental concepts underpinning the analytic hierarchy process. Section 3 presents an example of a possible hierarchy for use in the evaluation of the degree of confidence to be placed by the decision maker on the experts' estimates. In Section 4 such hierarchy and the AHP are used to determine the weights for the weighted aggregation of the estimates of the pressure increment in the Sequoyah containment and the prediction of the future change in precipitation at Yucca Mountain, at the 7,500 years time mark. Some concluding remarks end the paper.

2 THE ANALYTIC HIERARCHY PROCESS

In this section we briefly describe the basic concepts of the Analytic Hierarchy Process. What follows does not pretend to be an exhaustive and rigorous treatment of the subject but rather a somewhat concise 'snapshot' of the approach. For further details on the subject, the interested reader should consult Ref. 15.

We begin our presentation by referring to a simple example of a three-level hierarchy with three alternatives at the bottom, as shown in Fig. 1. The analytic hierarchy process is then based on the following major steps:

1. define precisely the top goal of the hierarchy and place it at the top level of the hierarchy;
2. build downward the hierarchy in different levels by putting in each level those factors directly influencing the elements of the level just above and directly influenced by the elements of the level just below. Directed arrows are placed to specify the interconnections between the elements;
3. at the bottom of the hierarchy place the available alternatives to be compared;
4. for each entry of each level build a pairwise comparison matrix to assess the importance of the influence of the relevant entries of the level below in relation to the element under analysis. In other words, given an element k in level s, all entries of the level below, $s-1$, which affect k are compared in a pairwise fashion in terms of their relevance to k. The proper question in the pairwise comparison is of the form: 'Considering entries X and Y of level $s-1$, how much more important is entry X compared to entry Y with respect to their influence on element k of level s?' The pairwise comparisons can be performed in a qualitative fashion and then translated into a numerical scale, or directly into a certain numerical scale. Typically, the scale of integer numbers from 1 to 9 is used and the values a_{ij} obtained from the comparisons are organized in a square matrix. For example, performing the comparison of element A with element B, the scale is the following:

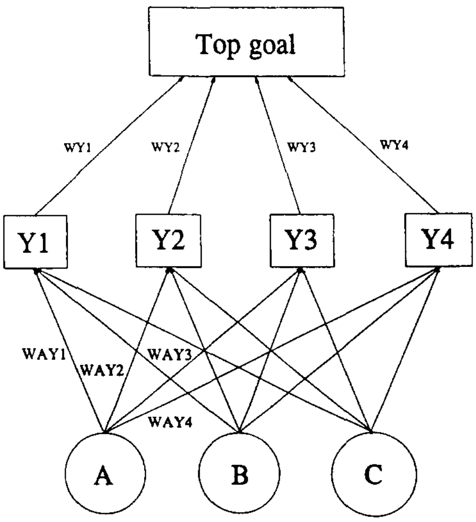

Fig. 1. A simple three-levels hierarchy.

1 = A and B equally important
3 = A slightly more important than B
5 = A strongly more important than B
7 = A very strongly more important than B
9 = A absolutely more important than B.

By definition an element is equally important when compared to itself so the principal diagonal of the matrix is filled with ones. The appropriate reciprocals, 1/3, 1/5,.....,1/9 are inserted where the reverse comparison, B vs A, is made. The numbers 2, 4, 6, 8 and their reciprocals can be used to facilitate expressing judgments for intermediate situations. In other words, the expert is allowed to resort to the use of a measure of 4, for example, when making a comparison of A and B which he believes cannot exactly be expressed by 3 nor 5;

5. for each element k in level s, determine the *potency* (*strength, priority, weight*) $w_{i(s-1),k(s)}$ with which each element i in level $s-1$ affects element k. The priorities $\{w_{i(s-1),k(s)}\}$ of the entries i in level $(s-1)$, relative to their importance for an element k in the next level (s) can be determined by solving an eigenvector problem. More precisely, it can be shown that given the matrix of pairwise comparisons for the element of interest, the principal eigenvector provides the vector of priorities, when normalized, and the maximum eigenvalue is a measure of consistency of the comparisons entered in the matrix. For complete consistency, the maximum eigenvalue, λ_{max} should be equal to the order of the matrix, n. The level of consistency of a given pairwise comparison matrix can be measured by a parameter called consistency ratio, CR, defined as the ratio of the consistency index $\mathrm{CI} = (\lambda_{max} - n)/(n-1)$ and the random index RI, which is the statistically averaged consistency index of randomly generated matrices of order n with entries artificially forced to be consistent. A consistency ratio of 0.10 or less is considered acceptable.[15] Further clarification regarding the meaning of consistency in this case can be obtained with the aid of a simple example. Suppose that in doing pairwise comparisons of three elements A, B, C, the following relations are obtained: A is 4 *times* more important than B; A is 8 *times* more important than C. Then, in order for our matrix of comparisons to be consistent we expect that the judgment relating B and C will state that B is *twice* as important as C. Although this matter might seem trivial, in practice it is quite common to encounter inconsistencies particularly when the order of the matrix is large;
6. in case of large inconsistencies in a matrix, revise

its entries by redoing the judgments on the individual pairwise comparisons or by forcing the values a_{ij} to be mathematically consistent by setting them equal to w_i/w_j, where $w_i = w_{i(s-1)k(s)}$, $w_j = w_{j(s-1)k(s)}$ are the priority values of elements i and j of level $s-1$ in regards to their relevance to element k of level s immediately above. Indeed, in the ideal case of the comparisons being the results of exact physical measurements w_i, $i = 1, 2, \ldots n$, the relation between the matrix entries and the weights is simply $a_{ij} = w_i/w_j$, for $i,j = 1, 2, \ldots n$. Note that usually it is sufficient to revise the values for those entries a_{ij} for which $|a_{ij} - w_i/w_j|$ is maximum or for all the entries of the row for which the root mean square deviation of a_{ij} and w_i/w_j is maximum. For more details on the revision process see the work by Saaty;[15]

7. once all the priority vectors are available, multiply them appropriately through the branches of the hierarchy (just like in a probability tree) to determine the overall weights of the bottom-level alternatives with regards to the previously defined top goal. For example, if $w_{AY} = \{w_{AY1}, w_{AY2}, w_{AY3}, \ldots\}$ is the vector $(1 \times n)$ of strengths of alternative A at the bottom level of the hierarchy in Fig. 1 and $w_Y = \{w_{Y1}, w_{Y2}, w_{Y3}, \ldots\}^T$ is the vector $(n \times 1)$ of priorities for the elements at the second level, then we multiply $w_A = w_{AY} \cdot w_Y$ to get the weight of alternative A measuring its relative importance with regards to the top goal.

The major advantage of the hierarchical structure is that it allows for a detailed, structured and systematic decomposition of the overall problem into its fundamental components and interdependencies, with a large degree of flexibility. On the other hand, the pairwise comparison approach represents a simple and intuitive way of expressing judgments on the relative importance of the different constituents of the hierarchy. From a practical point of view, it is worth pointing out that a user's friendly software package, EXPERT CHOICE, has been developed by T. L. Saaty for the application of the AHP method. This program allows to build the hierarchical structure and the associated matrices with great ease and automatically solve for the priority vectors at all levels, thus rendering the implementation of the approach very manageable.

3 BUILDING A HIERARCHY FOR THE DEGREE OF CONFIDENCE IN EXPERTS' ESTIMATES

In this section, we present a possible hierarchy structure to be employed by the decision maker for the evaluation of the degree of confidence that can be associated to the experts' judgments. The hierarchy does not pretend to be an exhaustive, complete and, most of all, unique representation of the performance evaluation process. Rather it is kept to a simple form, but yet explanatory, with the objective of setting the focus on the general approach.

We consider a four level hierarchy with three possible expert alternatives in the bottom level, as shown in Fig. 2. On the top level we place the top goal of the process, i.e., the evaluation of the confidence in the experts' judgments by the decision maker. On the second level we place the relevant issues of the expert's judgmental process which are believed to influence the top goal: 1) Expert's personal knowledge, PK; 2) Expert's sources of information, SI; 3) Expert's unbiasedness, UNB; 4) Expert's relative independence, IND; 5) Expert's personal interest in the study, PI; 6) Expert's past experience, PE; 7) Expert's performance measure, PM. These items are believed to contribute directly to the overall ranking of expert judgments by the decision maker.

Each of these contributors is further decomposed in more elementary constituents which represent the entries for the next level of the hierarchy.

The expert's personal knowledge (PK) is specified as the knowledge in the general area of study (GAK), the specific problem (SPK), the specific site (SSK) and the probability and statistics field (PSK), the latter topic being closely related to the ability of the expert in providing judgments in the form of probability distributions.

The sources of information (SI) used by the expert are categorized according to their reliability (RSI), specificity to the problem (SSI), objectivity (OSI), and variety (VSI).

The degree of unbiasedness (UNB) is broken down into two major subsets relative to the 'freedom' of the expert from internal biases (IB) and external biases (EB), respectively.

The internal biases subset is further decomposed into anchoring to an initial value (ANB), availability with which related instances are recalled to mind (AVB) and overconfidence (OCB), for which the elicited probability distributions tend to have a narrower range than it would be justified by the expert's knowledge and true beliefs.

The subset of external biases accounts for those influences coming from institutional affiliation (IAB), educational background (EBB) and political environment (PEB).

The issue of independence (IND) is intended as the degree of independence of the judgments of the various experts among each other. There are many sources of dependence amongst experts, such as similarities in training and experience. It is important to understand the experts' assumptions before an

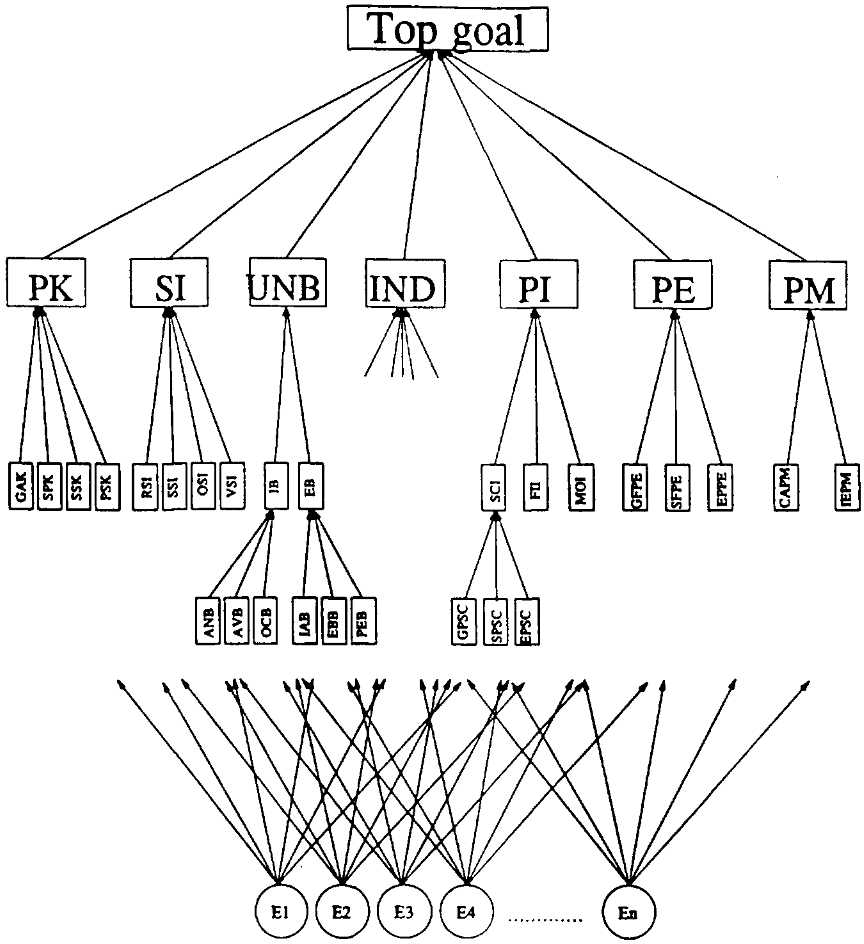

Fig. 2. Hierarchical tree for the evaluation of the degree of confidence on the expert's estimates.

assessment of dependence is attempted and particular care must be exercised in this process since highly-dependent experts are unlikely to lead to an increase in the assessed values.

The expert's personal interests (PI) are divided in scientific (SCI), financial (FII) and moral (MOI). For the latter two we subjectively decided that the confidence placed on the experts judgments would be higher for those experts having lower financial and moral interests involved. Therefore, in the subsequent matrix construction via pairwise comparison, those experts showing lower financial and moral interests will get higher values. We believed it worthwhile to address this issue at this point for it is very important to clearly specify in which direction should the comparisons and the associated values go, in order not to distort the analysis.

The topic of scientific interest was further decomposed into interest in the general problem (GPSC), specific problem (SPSC) and elicitation process (EPSC). The expert's past experience (PE) can be classified in terms of the contribution to the general field of research (GFPE), the specific field (SFPE) and the involvement in previous elicitation processes (EPPE).

Under the topic of expert's performance measure (PM) we considered the probability assessment ability of the expert, i.e., calibration (CAPM), and the relative amount of information of the expert, measured by the information entropy (IEPM) as suggested in the classical model.[7] Note here that, again, a high value of entropy signifies lack of information in an expert's subjective probability assessment so that attention must be given to the direction of the comparison values to make sure that a highly entropic expert gets a small relative weight.

Finally, the bottom layer of the hierarchy contains the alternative experts participating in the elicitation.

In the construction of the hierarchy, it is important that exhaustiveness and completeness of the influencing factors in each level be sought, up to the desired level of resolution, and that the meaning of each element be precisely defined together with its direction of influence.

Once again, we wish to emphasize that the definition and decomposition of the structure is flexible and subjective to the decision maker, and that no absolute hierarchy can be defined but rather modifications are needed for the specific cases.

4 APPLICATIONS

The hierarchical structure presented in the previous section has been used in a simulated exercise in which the degree of confidence in the experts' judgments is evaluated with regards to two case studies presented in literature.

Three fictitious experts, *A*, *B* and *C*, have been considered. Pairwise comparisons were made between the elements of each level of the hierarchy; the associated matrices were built and evaluated to provide the priority weight vectors. No formal exercise was attempted when assigning values to the pairwise comparisons, since reference was made to our set of three fictitious experts; rather, values were assigned with a certain subjectivity and freedom, although particular care was given to the achievement of consistency and to the logical sense of the results.

In a real practical case, the decision maker should try to benefit from the available tools to evaluate the different elements of the hierarchy and to sustain his assignment of values to the pairwise comparisons. In some circumstances, such as the evaluation of the expert's performance in the elicitation process, rigorous mathematical tools have been proposed and used.[7,12,13] In other circumstances, the comparison is more subjective and the decision maker should consider consulting the normative experts and the generalists, who have actively participated in the elicitation process and can greatly aid in the matrix building process.

The priorities were propagated through the branches of the hierarchy to obtain the weight ranking of the confidence in the estimates provided by the three experts. The entire implementation of the process was made by means of the Expert Choice package of T. L. Saaty.

Revision of some of the pairwise comparisons was made necessary for achieving an improved consistency. Again, the revision task proved to be easy with the aid of the computer software, which automatically identifies the most inconsistent entries and even suggests the proper modifications for maximum consistency. It is the opinion of the authors, however, that a better revision is obtained by actually reprocessing the inconsistent comparisons instead of just blindly using the value suggested by the computer.

Figure 3 reports, for example, the matrix of pairwise comparisons for the first level, containing those factors believed to directly influence the top goal. Numerical manipulation of this matrix leads to a consistency ratio of 0.112 which was reduced in a second iteration to an acceptable value of 0.083 after revision of two of the comparisons by the authors.

	PK	SI	UNB	PM	PI	IND	PE
PK	1	3	3	3	8	6	9
SI	1/3	1	2	3	6	3	7
UNB	1/3	1/2	1	2	2	2	4
PM	1/3	1/3	1/2	1	1/2	3	5
PI	1/8	1/6	1/2	2	1	3	3
IND	1/6	1/3	1/2	1/3	1/3	1	1/2
PE	1/9	1/7	1/4	1/5	1/3	2	1

Consistency Ratio CR = 0.083

	Weight *w*
PK :Expert's personal knowledge	0.389
SI :Expert' sources of information	0.231
UNB :Expert's degree of unbiasedness	0.124
IND :Expert's level of independence from the other experts	0.042
PI :Expert's personal interest in the study	0.085
PE :Expert's past experience	0.036
PM :Expert's performance measures	0.092

Fig. 3. Matrix of pairwise comparisons for the top goal and associated priorities.

As expected, the resulting vector of priorities turned out to be somewhat reflective of the subjective opinion of the authors acting as the decision maker. The experts' personal knowledge (PK), sources of information (SI) and degree of unbiasedness (UNB) were found to be ranked as the most important contributors to the evaluation of the degree of confidence in the experts' judgments. The personal knowledge of the expert and the sources of information available to him constitute the basis for his or her estimates and therefore play an important role, as shown by their weight of 0.389 and 0.231, respectively. For these estimates to be meaningful there must be no bias in the expert's judgment and, indeed, the corresponding attribute of unbiasedness also receives a significant weight of 0.124. The experts ability to perform his or her judgmental task as well as the personal interest in the study are weighed somewhat less because both attributes can be improved during the overall elicitation process by appropriate discussion and training. Finally, both the experts' independence from each other and past experience are considered lower-weight attributes with regards to the aggregation process, and receive weights of 0.042 and 0.036, respectively. Other matrices which required careful revision by the authors, in order to achieve the desired level of consistency, are those for the experts' sources of information (SI) and personal interest (PI). For all other matrices, consistency was readily obtained.

In the unbiasedness (UNB) comparison matrix, the internal (IB) and external (EB) biases were considered to have the same importance. Also, no difference in the degree of independence (IND) was assumed to exist among the experts whose judgments were believed to be totally independent. Clearly, this is just a subjective assumption made by the authors for this example case. In practical cases, dependence will most likely exist and shall be assessed. In such cases, the entries to the matrix UNB will no longer be all ones and consequently the priority weights will no longer be equal to 1/3 for all experts. The same comments hold true also for the matrix of moral interest (MOI) which was assumed to make no difference among the experts.

The overall result of the AHP process gave higher confidence to the estimates given by expert C with a priority of 0.377, followed by A with a priority of 0.371 and B with 0.251. The reasons for this are easily traceable through the branches of the hierarchy and the priority vectors. The overall consistency index turned out to be 0.06 and therefore acceptable.

The priorities thereby obtained were then used in the aggregation of the judgments of three experts for two case studies concerning the elicitation of the pressure increment in the Sequoyah containment building after a reactor vessel breach and the future change in precipitation at Yucca Mountain, respectively.

We wish to emphasize that in our analysis of these case studies the weight ranking of the experts is purely fictitious and there is no intention to express our opinion on the actual participants involved in the real study.

4.1 Case study 1: pressure rise in the Sequoyah containment due to reactor vessel breach

The PRA studies conducted on pressurized water reactors have predicted that accident sequences leading to core melt with the primary system still at high pressure can be important contributors to the total plant risk.[17] In such sequences, the molten corium may melt through the bottom of the vessel while the system is still pressurized. This corium could then be ejected from the vessel, under the high pressure, and be dispersed into the containment thus threatening its integrity. This phenomenon, called direct containment heating, is considered to be one of the most important safety issues in the area of severe accident analysis for many nuclear power plants, and yet it involves the combination of many poorly understood physical and chemical processes.

In Ref. 17, three experts were consulted to provide estimates of pressure increment in the Sequoyah nuclear power plant containment building due to reactor vessel breach. Several initial physical states were postulated and the experts were asked to provide their predictions for many of them. Our study refers to *Case* 1*a* which the experts identified as a base case and used for extrapolating pressure increment estimates for other cases. *Case* 1*a* refers to a physical state in which the Reactor Cooling System pressure is 2500 psia, the cavity is dry, the sprays are not operating and there is ice remaining in the ice condenser. Table 1 shows the expert-generated percentiles for the pressure increment in bars as given in Ref. 17. While experts' assumptions were also reported, they were not used in any way to rank order the degree of confidence in the experts' estimates and the aggregation in that study consisted in a simple equal weights average.

In our analysis we pretended to have looked at the assumptions made by the experts all through the judgmental process and at other background information regarding the experts, and that through the AHP we have generated the following weights for the experts as a measure of the degree of confidence in their judgments: for expert A we have $w_A = 0.371$; for B, $w_B = 0.251$; for C, $w_C = 0.377$. These weights have then been applied to a weighted averaging scheme. The results are reported in Table 1, together with a comparison with a simple aggregation scheme in which the estimates of all experts receive the same weight.

In this case, since both the estimates provided by the experts and the associated weights do not differ substantially, the equal-weights and the weighted averaging schemes give very similar results, as shown in Fig. 4. In practice, in this case it would seem more reasonable to pursue the aggregation by means of simple equal-weights averaging since the differences in the experts' weights would be difficult to defend in light of the many uncertainties in the assessment. However, this might not always occur, as is the case when the estimates provided by the experts cover a very wide range and/or the weights assigned as a measure of confidence in the estimates differ substantially.

As a further investigation we considered another case in which the AHP analysis of the experts'

Table 1. Experts' percentiles estimates of the pressure rise in bars and aggregated result for case study 1

	5%	25%	50%	75%	95%	W
A	3.60	4.00	4.50	5.46	6.22	0.371
B	1.6	—	4.0	—	10.8	0.251
C	3.0	4.0	5.0	6.5	8.0	0.377
Simple	2.73	4.00	4.5	5.98	8.34	—
Weighted	2.868	4.00	4.56	5.976	8.034	—
ε (%)	5.13	0	1.30	−0.06	−3.66	—

ε = relative deviation = (weighted − simple)/simple.

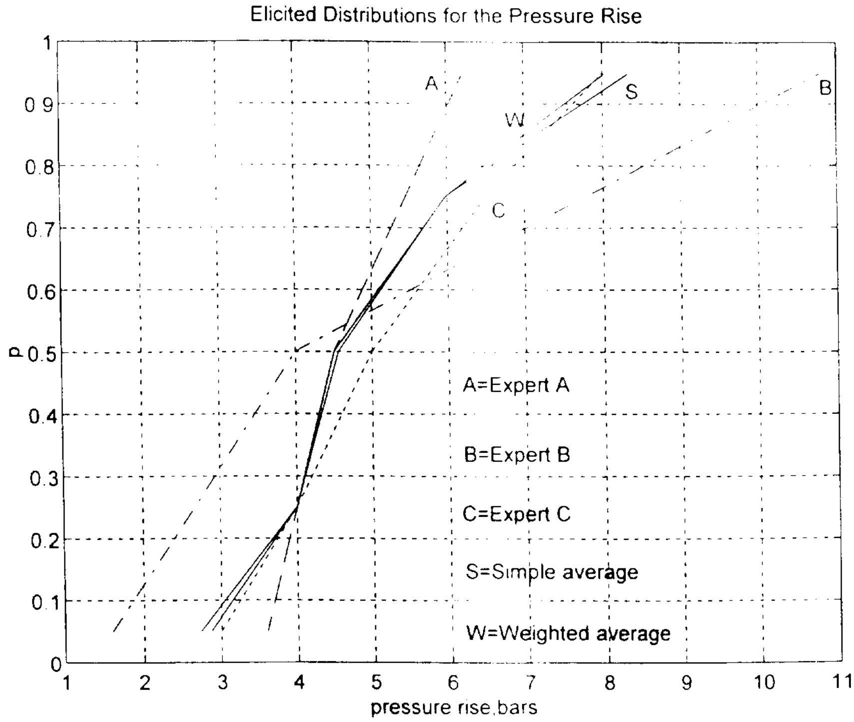

Fig. 4. Estimates of pressure rise in the Sequoyah containment building (Table 1).

estimation process resulted in the following weights: $w_A = 0.421$, $w_B = 0.006$, $w_C = 0.573$. This is shown to happen when Expert B has very poor personal knowledge of both the problem and the elicitation process and uses inappropriate sources of information. These two factors constitute more than 60% of the overall degree of confidence in the expert's judgment so that low weights in these issues result in a low overall weight. Expert B therefore comes out strongly discredited. In this situation, the weighted aggregation of the experts' judgments lead to results which differ more from the simple equal-weights average, as shown in Table 2 and Fig. 5. This is obviously due to the fact that Expert B, which gives an estimated distribution of pressure rise which differs substantially from those provided by Experts A and C, receives a negligible weight in the weighed average.

Table 2. Experts' percentiles estimates of the pressure rise in bars and aggregated result for case study 1 (Expert B discredited)

	5%	25%	50%	75%	95%	W
A	3.60	4.00	4.50	5.46	6.22	0.421
B	1.6	—	4.0	—	10.8	0.006
C	3.0	4.0	5.0	6.5	8.0	0.573
Simple	2.733	4.0	4.5	5.98	8.34	—
Weighted	3.244	4.00	4.783	6.023	7.267	—
ε (%)	18.69	0	6.30	0.72	−12.86	—

ε = relative deviation = (Weighted − Simple)/Simple.

4.2 Case study 2: precipitation change at 7500 years in the vicinity of Yucca Mountain

The performance of radioactive waste repositories is influenced by a variety of processes such as groundwater flow and fracture movement, which are in turn affected by conditions in the surface environment. The temporal scale over which the performance assessment of these repositories is required is of the order of 10,000 years. Over such a period of time, the underlying cause of change in surface environment is inferred to be climate change but considerable uncertainty remains over the mechanisms of such change.[19]

In the elicitation of future climate in the Yucca Mountain vicinity,[18] the primary variables considered are changes in annual precipitation and temperature as well as changes in the seasonal variability of precipitation at the site over the course of 10,000 years. These variables have been selected by the generalists to include those believed to most likely impact the performance of the repository. Five specialists have been asked to provide distributions for the main variables and discuss the climatic controls that were inferred to be the principal cause of any expected changes, at different time epochs of 100, 1000, 3000, 5000, 7500 and 10000 years.

In our analysis we considered the judgments of three of the five selected experts, with regards to the change in precipitation at 7500 years. Table 2 reports

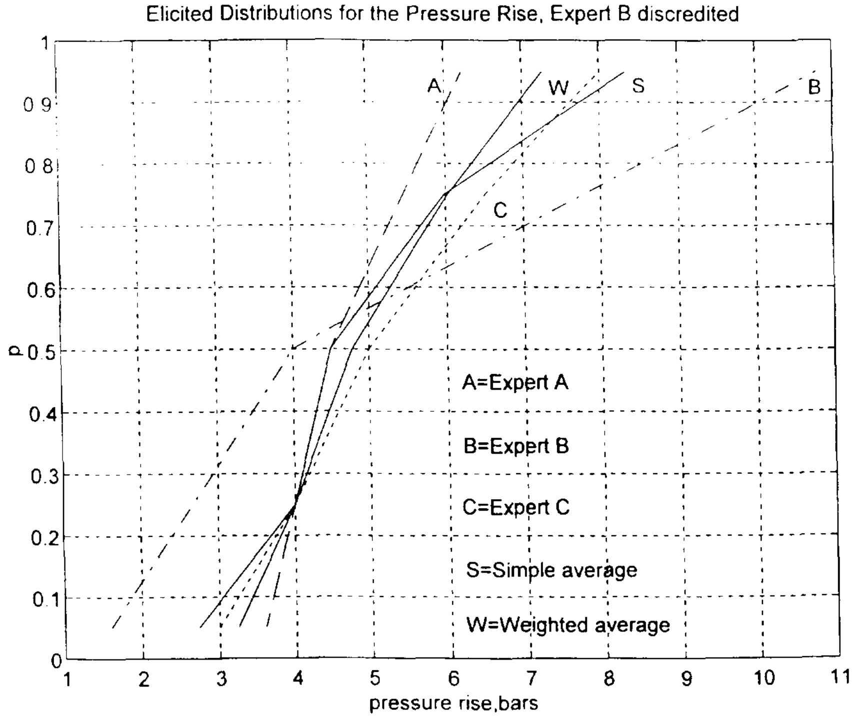

Fig. 5. Same case as in Fig. 4 but with expert B discredited (Table 2).

the values provided by the experts in terms of actual change from the current average annual precipitation of 150 mm. In the same fashion as for case study 1, the distributions provided by the experts were aggregated via a weighted average using the priorities previously assessed with the AHP as weights. The aggregated results are shown in Table 3, together with a comparison with the equal-weights aggregation scheme.

In this case, even when the weights given to the estimates provided by the experts are not substantially different, the aggregated result presents a significant deviation from a simple equal-weights average, as shown in Fig. 6. This is to be expected since the distributions elicited are quite different among each other, covering largely different ranges of the parameter values. In a situation like this, the weights

Table 3. Experts' percentiles estimates of the change in precipitation from the current average annual precipitation in mm. and aggregated result for case study 2

	5%	25%	50%	75%	95%	*W*
A	− 16.9	− 6.15	4.5	22.05	31.45	0.371
B	75.9	101.2	126.5	177.1	278.3	0.251
C	− 70	− 37.5	− 10	− 5	0	0.377
Simple	− 3.68	19.18	40.33	64.72	103.25	—
Weighted	− 13.2	8.98	29.65	50.75	81.52	—
ε (%)	260.1	− 53.2	− 26.5	− 21.6	− 21.0	—

ε = relative deviation = (Weighted − Simple)/Simple.

assigned to the confidence in the experts' estimates play an important role even if their differences are not so strong.

Notice that Expert B appears to be quite an outlier and his elicited, large, positive values of precipitation change drive the aggregated averages. As a further investigation, we used, again, the second AHP process which leads us to discredit the estimates of this expert. Indeed, we obtain the following weights: $w_A = 0.421$, $w_B = 0.006$, $w_C = 0.573$. Once again, the results of Table 4 and Fig. 7 show a large difference in the aggregates between the weighted and simple average methods.

5 CONCLUSIONS

The issue of aggregation is a matter of paramount importance in the use of expert judgments for complex technical problems, particularly when a strategic decision is required as an end result. Many different models have been developed to evaluate the experts' performance and provide a measure of the degree of confidence in the judgments given by them.

In this paper we have presented a simple, systematic approach to aid the evaluation of the confidence in the experts' judgments taking into account, in a formal and explicit manner, several factors that may influence it.

The approach is based on the analytic hierarchy

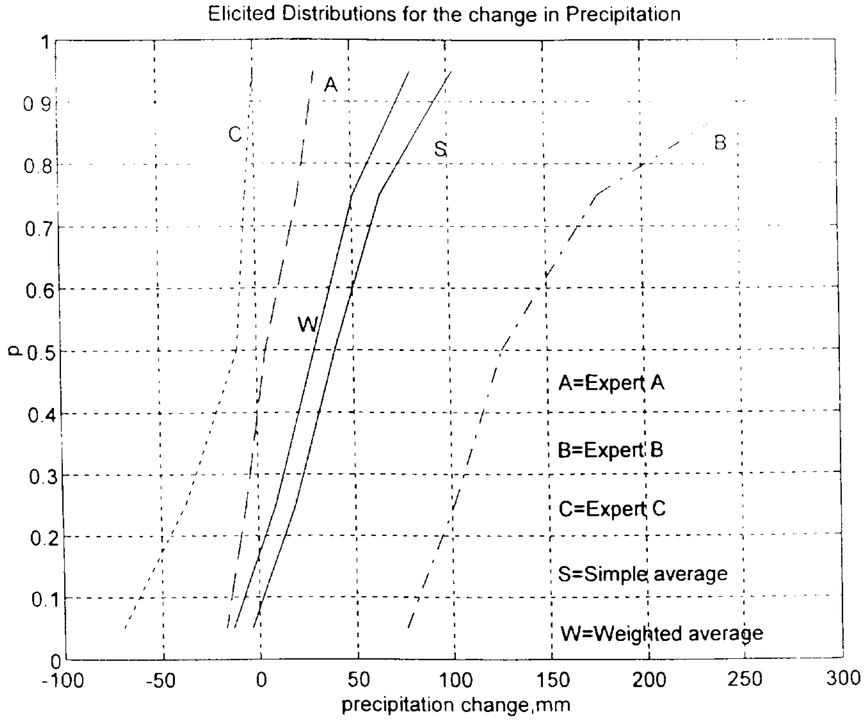

Fig. 6. Estimates of changes in precipitation at Yucca Mountain (Table 3).

process often used in decision analysis. This technique allows us to define a clear picture of the problem and to decompose it into a structured hierarchy of more elementary subsets which are easier to analyze separately. The assessment of the relative importance of the elements of each subset by pairwise comparison is easy to handle and allows us to achieve a high level of consistency. The approach is flexible, explicit and easily traceable.

The hierarchical structure allows for the consideration of a large variety of basic measures which are inferred to be affecting the experts' judgments and which can be assessed in terms of the behavioral or mathematical models as appropriate. These models can provide measures to quantify the different aspects relevant to the degree of confidence that can be placed on the experts' judgments and in this sense the AHP approach here presented should not be considered as substitutive of any of these models but rather as an additional tool to systematically and explicitly account for the different sources of influence.

Table 4. Experts' percentiles estimates of the change in precipitation from the current average annual precipitation in mm. and aggregated result (Expert B discredited)

	5%	25%	50%	75%	95%	W
A	−16.9	−6.15	4.5	22.05	31.45	0.421
B	75.9	101.2	126.5	177.1	278.3	0.006
C	−70	−37.5	−10	−5	0	0.573
Simple	−3.68	19.18	40.33	64.72	103.25	—
Weighted	−46.77	−23.47	−3.08	7.48	14.91	—
ε (%)	1170.9	−222.4	−107.6	−88.4	−85.6	—

ε = relative deviation = (Weighted − Simple)/Simple.

An example of a possible hierarchy for the evaluation of the confidence in the experts' judgments has been presented. The hierarchy was kept to a reasonably simple level and no particular effort was made to make it exhaustive and complete, in order not to lose the focus of the study, i.e., the investigation of the capability of the AHP process as an analytical tool in the context of expert judgments. The hierarchy structure, as here presented, is based on the assumption that the elements in each subset are independent. In some situations this might not hold true so that the AHP process needs to be appropriately modified by treating dependence and independence separately and then combining them.[15]

The associated matrices were built through simple pairwise comparisons and the resulting priority vectors turned out to be reasonable in the sense that they reflect the authors' opinion on the importance of the different factors in the hierarchy. Obviously, the results can be questionable in as far as the assessment is subjective, and therefore different conclusions are expected to be drawn by different decision makers (much in the same manner as different experts provide diverse, subjective opinions).

The results of the AHP process were utilized in two case studies previously examined in the literature: 1) the elicitation of the pressure increment in the

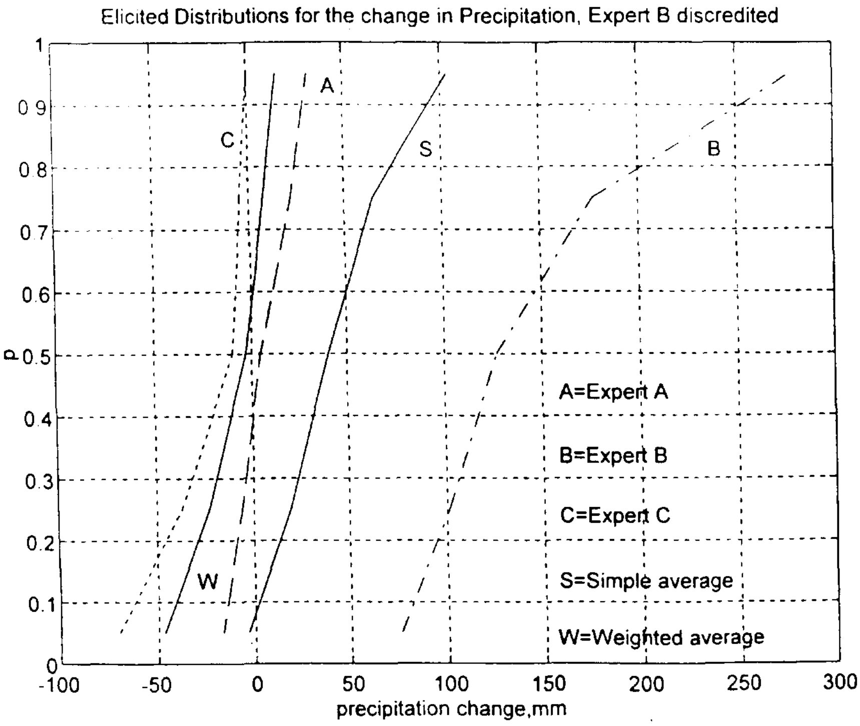

Fig. 7. Same case as in Fig. 6 but with expert B discredited (Table 4).

Sequoyah containment building due to reactor vessel breach and 2) the prediction of future change in precipitation at Yucca Mountain. Once the priorities for the confidence in the experts' judgments are available, their application as weights in an aggregation scheme is straightforward.

The results of case study 1 have shown that when the experts' elicited distributions do not differ excessively, the effect of the weights is significant only if their difference is substantial, i.e., if one or more of the experts can be for some reason discredited (their weights being close to zero) or if some experts are dominant over others (their weights being relatively larger). If the experts are closely weighted, it seems appropriate to apply a simple average method because the uncertainties in the overall assessment render it difficult to sustain small weights' differences. On the other hand, case study 2 presents a situation in which the elicited experts' judgments differ profoundly. In this case, even with values of the weights being all close together, the effect of the weights is quite important. This second case study is quite interesting for expert B appears to be an outlier in the assessment and by fictitiously discrediting his judgments we find very different results. Large differences in the weights assigned to the experts are obviously found to lead to aggregate results that differ largely from those obtained by equally weighting all experts estimates and this can strongly affect the final decision. In this sense, the AHP can be seen to provide a very simple and explicit tool for assigning appropriate weights and allows us to sustain or negate further assumptions in the aggregation step such as the equal-weights averaging.

The analyses with the AHP were performed using the Expert Choice software package of T. L. Saaty which made the implementation quite easy to handle. The software is very user friendly and quite automated. Furthermore, it offers the possibility of performing sensitivity analyses on the results obtained, and that could help the decision maker individuate those factors for which more care in the evaluation should be warranted and for which a deeper examination of the expert's judgmental process might be needed.

In qualitative terms, the approach presented seems to meet the important objective of providing a systematic scheme for individuating, in a clear and traceable manner, if some experts are to be discredited or if all are to be assigned basically equal weights. Moreover, in the case of widely differing elicited distributions, the weights play an important quantitative role in the aggregate measure of interest and the structure of the AHP is able to identify the most fundamental issues affecting the confidence in experts' estimates which might require additional inspection.

We finally note that a similar approach, based on a somewhat modified hierarchical tree, could also be undertaken to tackle the controversial issue of experts' selection in a systematic and explicit manner.

Further investigation in the practical application of the AHP process in the field of expert judgment seems worthwhile, for the AHP provides a systematic, explicit and traceable way of evaluating the degree of confidence that can be placed on the experts' estimates and it is intuitive and easy to handle.

REFERENCES

1. Nuclear Regulatory Commission, Reactor safety study: an assessment of accident risks in U.S. commercial nuclear power plants. *NUREG* 75-014, USNRC, Washington DC, 1975.
2. Lewis, H. W. *et al.*, Risk Assessment Review Group report to the US Nuclear Regulatory Commission. *NUREG/CR*-0400, USNRC, Washington DC, 1979.
3. Bonano, E.J. & Apostolakis, G. E., Theoretical foundations and practical issues for using expert judgments in uncertainty analysis of high level radioactive waste disposal. *Radioactive Waste Management and the Nuclear Fuel Cycle*, **16** (1991) 137–159.
4. Chhibber, S., Apostolakis, G. & Okrent, D., A taxonomy of the use of expert opinions in probabilistic risk assessment. *Reliab. Engng System Safety*, **38** (1993) 27–46.
5. Tversky, A. & Kahneman, D., Judgment under uncertainty: heuristics and biases. *Science*, **185** (1974) 1124.
6. Christensen-Szalanski, J. J. J., Beck, D. E., Christensen-Szalanski, C. M. & Koepsell, T. D., Effects of expertise and experience on risk judgments. *J. Appl. Psych*, **68** (1983) 278.
7. Cooke, R. M., *Experts in Uncertainty: Expert Opinion and Subjective Probability in Science*, Oxford University Press, New York, 1991.
8. Mendel, M. B. & Sheridan, T. B., Filtering information from human experts. *IEEE Trans. Systems. Man. Cyber*, **36** (1989) 6.
9. Wheeler, T. A., Hora, S. C., Cramond, W. R. & Unwin, S. D., Analysis of core damage frequency from internal events: expert judgment elicitation. *NUREG/CR*-4550, Sandia National Laboratories, prepared for USNRC, 1989.
10. Probability encoding of hydrologic parameters for basalt, elicitation of expert opinions from a panel of five consulting hydrologists. *RHO-BW-CR*-145 *P.*, Rockwell Hanford Operations, Rockwell International, Richland, Washington, 1984.
11. Mosleh, A. & Apostolakis, G. E., The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk Analysis*, **6** (1986) 447–461.
12. Apostolakis, G. E., Expert judgment in probabilistic safety assessment. In *Accelerated Life Testing and Experts' Opinions in Reliability*, Proc. Int. School of Physics, Italian Physical Society, Course CII, (eds C. A. Clarotti and D. V. Lindley) North-Holland, New York, NY, 1988.
13. Lindley, D. V., The use of probability statements. In *Accelerated Life Testing and Experts' Opinions in Reliability*, Proc. Int. School of Physics, Italian Physical Society, Course CII, (eds C. A. Clarotti and D. V. Lindley) North-Holland, New York, NY, 1988.
14. Meyer, M. & Booker, J., *Eliciting and Analyzing Expert Judgment: A Practical Guide*, Academic Press, London, 1991.
15. Saaty, T. L., *The analytic hierarchy process*, McGraw-Hill International, New York, 1980.
16. Embrey D. E., Humphreys, P. C., Rosa, E. A., Kirwan, B. & Rea, K., SLIM-MAUD: an approach to assessing human error probabilities using structured expert judgment. *NUREG/CR*-3518, US Nuclear Regulatory Commission, Washington DC, 1984.
17. Harper F. T., *et al.*, Evaluation of severe accident risks: quantification of major input parameters—experts' determination of containment loads and molten core containment issues. *NUREG/CR*-4551, *SAND*86-1309, Vol.2, Rev.1, Part 2, US Nuclear Regulatory Commission, Washington DC, 1991.
18. De Wispelare, A. R., Herren, L. T., Miklas, M. P. & Clenien, R. T., Expert elicitation of future climate in the Yucca Mountain vicinity. *Iterative Performance Assessment Phase* 2.5, prepared for NRC-02-93-005, Center for Nuclear Waste Regulatory Analyses, San Antonio, TX, 1993.
19. Wilmot, R. D., The treatment of climate-driven environmental change and associated uncertainty in post-closure assessments. *Reliab. Engng System Safety*, **42** (1993) 181–200.