Assessing Confidence in Performance Assessments Using an Evidence Support Logic Methodology: An Application of TESLA – 9484

Michael Egan, Alan Paulley
Quintessa Limited
Birchwood Park, Warrington WA3 6AE, United Kingdom

Linda Lehman, John Lowe
CH2M Hill Hanford Group
Richland, Washington

Elizabeth Rochette
Washington State Department of Ecology
Richland, Washington

Steve Baker
Umtanum Enterprises
Richland, Washington

ABSTRACT

The assessment of uncertainties and their implications is a key requirement when undertaking performance assessment (PA) of radioactive waste facilities. Decisions based on the outcome of such assessments become translated into judgments about confidence in the information they provide. This confidence, in turn, depends on uncertainties in the underlying evidence. Even if there is a large amount of information supporting an assessment, it may be only partially relevant, incomplete or less than completely reliable. In order to develop a measure of confidence in the outcome, sources of uncertainty need to be identified and adequately addressed in the development of the PA, or in any overarching strategic decision-making processes.

This paper describes a trial application of the technique of Evidence Support Logic (ESL), which has been designed for application in support of 'high stakes' decisions, where important aspects of system performance are subject to uncertainty. The aims of ESL are to identify the amount of uncertainty or conflict associated with evidence relating to a particular decision, and to guide understanding of how evidence combines to support confidence in judgments. Elicitation techniques are used to enable participants in the process to develop a logical hypothesis model that best represents the relationships between different sources of evidence to the proposition under examination. The aim is to identify key areas of subjectivity and other sources of potential bias in the use of evidence (whether for or against the proposition) to support judgments of confidence. Propagation algorithms are used to investigate the overall implications of the logic according to the strength of the underlying evidence and associated uncertainties.

INTRODUCTION

The Need for Systematic Assessments

The Hanford site, located in south central Washington State, occupies 560 square miles and has been used extensively for producing defense materials by the United States Department of Energy (USDOE) and its predecessors the US Atomic Energy Commission and the US Energy Research and Development Administration. Starting in the 1940s, Hanford site operations were primarily for the production of nuclear weapons materials. In the 1960s operations were expanded to producing electricity from a dual-purpose reactor, conducting diverse research projects, and managing wastes. In the late 1960s, the site's original mission ended. This mission left a large inventory of radioactive and mixed waste stored in buried single and double shelled tanks throughout the Hanford site central plateau region, i.e. the '200 Areas'. Numerous past practice wastes sites also exist where millions of curies were discharged into the soil through cribs and ditches.

Today the site's missions are environmental restoration, energy-related research and technology development. As a part of the environmental restoration mission, USDOE is proceeding with plans to permanently dispose of the waste

stored on site. As a result of the large size of the Hanford site and the large number of waste sites (approximately 700 in the 200 Areas alone), coupled with the timescales over which safe disposal has to be demonstrated, Hanford cleanup activities are frequently based on judgment, guided by Performance Assessment (PA) studies, with limited support from direct measurements.

Cleanup and closure activities at Hanford are governed by numerous regulations including the Resource Conservation and Recovery Act (RCRA), as well as the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA), which both require public review and comment. Frequently, comments from stakeholder groups indicate a lack of confidence in the Hanford decision-making process. , It has been recognized that a process is needed whereby all stakeholders and regulators may openly discuss controversial topics with USDOE and their contractors in a meaningful way.

Among the areas of criticism arising from regulator and stakeholder reviews of completed PAs is a shortage of information regarding uncertainties in the parameters used in the assessment models and in the analysis of overall system performance. Performance Assessments at Hanford have typically been done in a deterministic fashion. Key sensitivities to dose have been explored and uncertainty quantification has begun, however, a robust approach to addressing uncertainty in PAs is still being studied.

With parts of the USDOE Complex under NRC consultation, some sites have enthusiastically embraced probabilistic performance assessments and a probabilistic approach to addressing uncertainty. USDOE Head Quarters (HQ) is also urging consistency and systematic approaches throughout the complex. While this probabilistic approach has merit, other approaches are being studied by Hanford contractors to decide which methods are most satisfactory for application at Hanford.

Potential Role of Evidence Support Logic

Evidence Support Logic (ESL) has been developed as a deliberative process for addressing questions of confidence in decisions based on uncertain evidence. The aim of ESL is to identify the amount of uncertainty or conflict associated with evidence relating to a particular decision, and to guide understanding of how the evidence combines to support confidence judgments.

To the extent that stakeholder input is incorporated and influences cleanup or closure decisions, ESL has been recognized as a possible tool for addressing some of the regulatory and stakeholder needs when decisions are based on judgment. It was therefore decided to fund an application of ESL to examine its potential for developing consensus on key parameters in the C Tank Farm Performance Assessment at Hanford, with the aim of promoting a more transparent exchange between stakeholders, regulatory agencies and the USDOE. Because the application was designed to test the usefulness of the ESL approach for possible application within the USDOE Complex, the trial was designed as an exercise only, with no formal implications for existing PA studies or decision making.

THEORETICAL BACKGROUND TO EVIDENCE SUPPORT LOGIC

Decisions on complex issues are typically informed by a wide range of factors, drawing on different lines of evidence and multiple sources of information. Making sense of this information involves judgments about the quality and reliability of the evidence and the extent to which it supports a given proposition.

A familiar example is the way in which evidence is taken into account in a legal case; the jurors examine all the evidence presented to them and then make their judgment about the extent to which this supports the case made by the trial lawyers. Less familiar are ‘high stakes’ regulatory and investment decisions, where the choice to authorize a development to proceed (or not) can also be a highly contentious arena of indirect, and sometimes conflicting, data.

Such evidence can come from a wide variety of sources, including field data, modeling results and expert judgment. While there may be a large amount of information relating to the decision at hand, it may be only partially relevant, incomplete or uncertain. Moreover, the range of available evidence may give an indistinct picture, with no clear

indication of how to best target resources in order to improve understanding and confidence. Disputed interpretations may arise, perhaps because some practitioners appear to be biased by excessive reliance on a particular source of evidence in the face of contradictory, or perhaps more equivocal, evidence from elsewhere. In order to provide a justified interpretation of the available evidence, which can be audited from start to finish, it is therefore necessary to examine and make explicit judgments regarding the quality of the information and how different lines of evidence combine in relation to support for, or contradiction of, a given case.

The technique of Evidence Support Logic was originally developed to support the evaluation of hydrocarbon extraction prospects in the oil and gas industry. The methodology was originally described by researchers at Bristol University in the UK [1-5] and subsequently adapted by Bowden [6], primarily for application in the field of modeling interpretation. The process involves systematically breaking down the proposition under consideration (e.g. "*this field is worth developing for hydrocarbon production*") into a logical hypothesis model whose elements expose judgments relating to confidence in the available evidence. More recently, the approach has been refined and encoded in the TESLA software tool [7], for application to a variety of decision-support contexts, including the evaluation of potential waste disposal sites [8], geochemical data review [9] and the development of PA models for carbon dioxide sequestration [10].

Building a decision model comprises three main steps:

- Development of the logical hierarchical model; breaking down a single decision statement into a number of underlying hypotheses;
- Parameterization of the model and identification of relevant sources of evidence;
- Propagation of evidence through the model, representing uncertainty using the principles of Interval Probability Theory [1], to provide an assessment of confidence in the top-level proposition.

The feasibility of such a systematic approach is greatly enhanced by the use of software to support model construction, knowledge recording and uncertainty handling. As with all decision support tools and processes, TESLA and ESL do not replace the need for judgment, nor can such methods eliminate subjectivity from the evaluation and interpretation of evidence. However, it is believed that a systematic approach, providing the ability to manipulate the hierarchical structure of a decision as it evolves, can support deliberation between stakeholders, making visible the factors that underpin confidence when dealing with complex judgments.

The Logical Hypothesis Model

The aim of ESL is to identify the amount of uncertainty or conflict associated with evidence relating to a particular decision, and to guide understanding of how the evidence combines to support confidence judgments. Doing this involves analyzing a proposition representing the decision to be made (e.g. "*it is likely to rain in the next three hours*"), rather than the decision itself (e.g. "*shall I take an umbrella with me?*"). The proposition is then broken down via a logical hypothesis model, until a stage is reached where evidence can be gathered for the lowest sub-hypotheses.

Figure 1 illustrates the decomposition of a proposition relating the suitability of a site for detailed investigation as part of its technical evaluation for the siting of a geological waste disposal facility [6]. This shows the initial break down into eight contributory sub-hypotheses that together determine confidence in the main proposition. The general principle at each step in developing the hierarchical model is to undertake a comprehensive top-down analysis of the factors contributing to a hypothesis, until a level of detail is reached at which people are comfortable in providing direct judgments about evidence in terms of the level of support that it provides for or against the sub-hypothesis in question. In the case illustrated here, further development is required under each of the eight sub-hypotheses until a point is reached (the so-called 'leaf' hypotheses of the logic tree) at which success criteria can be identified that define an agreed standard against which the confidence in the evidence can be assessed.

Inevitably, the structure that is adopted in developing a logical hypothesis model is somewhat subjective, in so far as there may be a number of alternative ways of defining a comprehensive top-down hierarchy. Ideally, the input to the model should provided by one or more specialists in the relevant field. Several approaches are available to do this, depending upon the nature of the information, the numbers of the experts and their specialties. The simplest approach is for a single person acting as a facilitator to lead the construction of a hypothesis model in a meeting involving the experts. At each stage, the structure can then be debated until a consensus is reached.

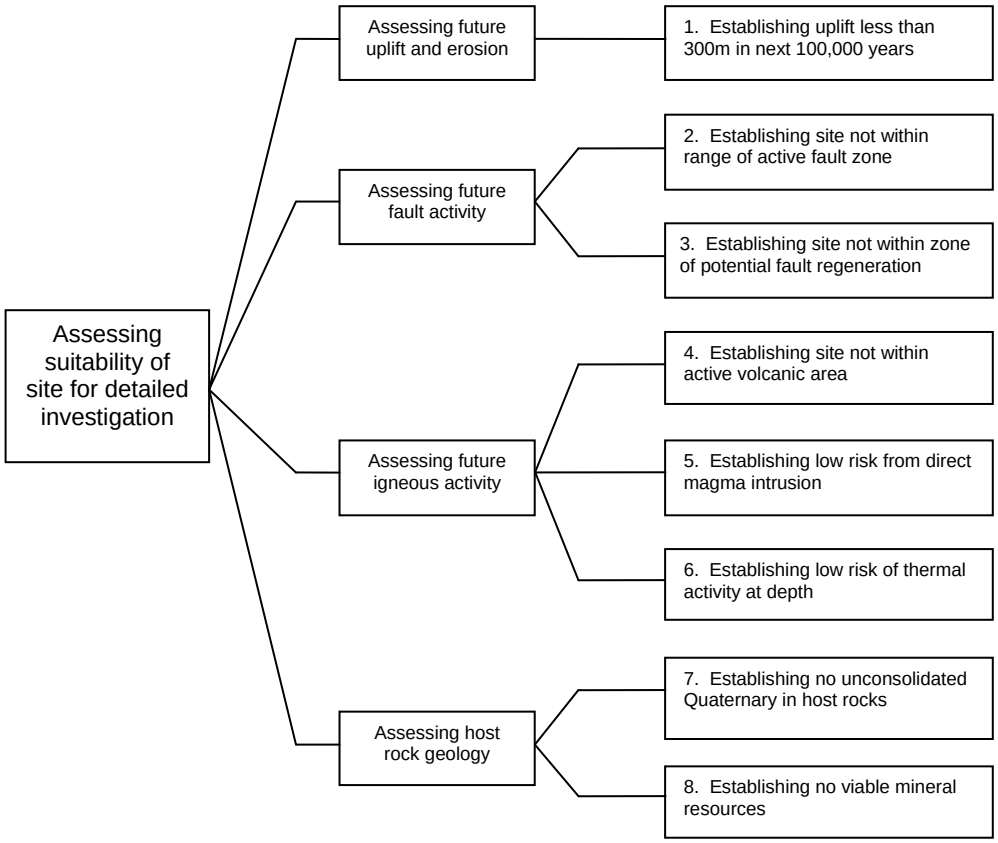

Fig. 1. Simplified example of an evidence-based hypothesis model showing site specific factors for assessing the suitability of a site for inclusion in a detailed investigation programme [6]

Before evidence can be assessed and input to the hypothesis model, it is necessary to describe and parameterize the logic by which that evidence is propagated upwards through the model in order to assess the extent of support for the top-level proposition. The mathematical basis of the ESL methodology is an approach known as Interval Probability Theory (IPT). For present purposes, the aim is simply to outline some major elements of the quantitative methodology, in order to enable a better understanding of the model parameterization process.

Three-value Logic

Classical probability theory follows a two-value logic, whereby evidence must either be in favor of a hypothesis, or against it. This is sometimes described as a 'closed world' perspective, in which evidence 'for' and evidence 'against' are treated as complementary concepts (i.e. $p(A) + p(\text{not } A) = 1$, where $p(A)$ is the probability of event A occurring, or in other words the degree of confidence in evidence supporting the occurrence of A). Three-value logic extends this to allow for a measure of uncertainty as well, recognizing that belief in a proposition may be only partial and that some level of belief concerning the meaning of the evidence may be assigned to an uncommitted state. Uncertainties are handled as 'intervals' that enable the admission of a general level of uncertainty [11], providing a recognition that information may be incomplete and possibly inconsistent (i.e. evidence for + evidence against + uncertainty = 1). A convenient way to represent this three-value logic is the so-called 'Italian flag' representation, in which evidence for a proposition is represented as green, evidence against as red, and residual uncertainty is white [4].

With this formalism, evidence for and evidence against can be evaluated independently, each ranging from 0 to 1, with uncertainty taking a value from –1 to 1. An uncertainty of 1 implies that there is no evidence at all on which to

base a judgment, whereas a negative value indicates a situation in which the assessed evidence for and against total more than 1; that is to say, a situation in which the evidence appears to be in conflict.

Model Parameterization

At every branch within the hierarchical hypothesis model, each sub-hypothesis is assigned a *sufficiency* that dictates how much weight should be given to it when determining the combined evidence of its parent hypothesis. With the approach to ESL encoded in TESLA, two separate values for this parameter may be assigned, representing the designated sufficiency of the sub-hypothesis both for and against the success of the upper level hypothesis. In effect, when determining the sufficiency of a sub-hypothesis, consideration is given to its overall relevance to making a judgment about the dependability of the higher level hypothesis. This is equivalent to asking the question:

If it were assumed that this sub-hypothesis alone was completely dependable – what is the likelihood that the higher level proposition would also be completely dependable?

The sufficiency parameter can therefore take a value between 0 (insufficient) and 1 (completely sufficient) – a greater level of sufficiency results in evidence values associated with the sub-hypothesis being propagated more strongly up the hierarchy.

Within each set of 'sibling' sub-hypotheses at a given node in the logical model, there is a chance that some of the contributing information may be overlapping, or shared. This is reflected in the *dependency* parameter, which describes the degree of commonality that is understood to exist between contributing hypotheses. The role of dependency in the quantitative propagation of evidence through the hypothesis model is to avoid double-counting of the support provided by shared lines of evidence. It therefore effectively provides a subtractive element to the propagation algorithm. In eliciting dependency, the question is asked:

How much shared information to the sub-hypotheses use in contributing to the dependability of the parent?

Parameter values for sufficiency and dependency are necessarily assigned by judgment. However, making such judgments explicit in a workshop setting provides an important means for developing common understanding of the logical basis for understanding how diverse pieces of evidence fit together.

In a workshop context, the elicitation process can be based on a mapping process that converts linguistic responses, such as 'Very High', 'High', 'Intermediate', 'Low' and 'Very Low', to a numerical value for computational purposes. As with all linguistic to numerical conversions, it is important that all those contributing to the evaluation understand the conversion factors that have been used and have the opportunity to modify, revise or refine their initial judgment if the conversion appears to misrepresent their intention.

One final concept is required in order to identify those sub-hypotheses that are considered essential to the success of the parent hypothesis – that is, any hypotheses which, upon failure, will necessarily lead to failure of the hypothesis above. The Boolean operator *necessity* is used to indicate such a sub-hypothesis: it is set at TRUE if the hypothesis is necessary and FALSE otherwise. In terms of propagation, a threshold value for confidence in the evidence against provided by a necessary hypothesis is set (typically >0.5) and, if the assessed value is greater than this, then a confidence value of at least this size will be propagated to the next level in the hypothesis hierarchy. Bowden [6] provides some hypothetical examples to illustrate the impact of the different combinations of logical operators on the propagation of evidence.

Within TESLA, there is the option to structure a logic model such that *ALL* sub-hypotheses at a given node are considered necessary to confidence in the parent. This is logical equivalent of the 'weakest link' argument, in which no significant benefit is gained from combining sources of evidence, since the weakest level of evidence it taken to dominate the overall propagation of confidence. Alternatively, it is possible to develop a logical structure in which, at a given node, *ANY* of the sub-hypotheses can be used to support (or refute) the proposition, but not all in combination. In this case, therefore, it is the largest degree of confidence in the evidence that determines the propagated value.

Evaluating Evidence

There are many potential contributions to residual uncertainty in the treatment of evidence; in essence, the assignment of a level of belief to an uncommitted state (i.e. neither for nor against) should reflect "*anything we are not sure of*". This incorporates not only the awareness that exists in relation to uncertainties in the system under review and its behavior, but also a measure of degree of belief in that understanding. Contributions to the uncommitted state in the probability triplet therefore include:

- Incomplete knowledge of the leaf hypotheses in the logical hierarchy - we don't understand all the processes involved;
- Incomplete characterization of the system - we don't have all the data;
- Uncertain quality - we have the data but we're not sure of their reliability for use as evidence;
- Uncertain meaning - we have data but we're unsure what they mean;
- Conflict - we have relevant data from different sources which don't agree; and
- Variability - we have data but they don't give us a unique answer.

Bowden [6] discusses a classification scheme for uncertainty to account for such contributions, based on the work of several authors [2, 12, 13].

Judgments on evidence can be elicited in several ways; the standard approach followed in TESLA involves separate elicitation of 'evidence for' and 'evidence against' the dependability of each leaf hypothesis. Because this depends on expert judgment, a helpful approach can be to make use of qualitative linguistic judgments that are subsequently mapped to a numerical scale via a utility function.

In the application of ESL, when judgments are made about evidential support, two main steps are followed in characterizing uncertainty. First, specialists are asked about the overall adequacy of the knowledge base on which they are being asked to make judgments. Then, for the evidence that is available, a judgment is made about its 'face value' in support of (or against) the hypothesis in question, modified by belief regarding the quality of that evidence.

In seeking first to understand the adequacy of the knowledge base, it is possible to invite expert judgment in relation to the following questions:

(i) How much information would you ideally wish to have in order to be confident in providing a judgment of evidence in support of, or against, the proposition?
(ii) In relation to the hypothetical ideal, how much information do you actually have on which to base a judgment?

On the basis of this analysis, a quantitative estimate can be made of the 'uncertainty due to lack of knowledge' for the particular sub-hypothesis under consideration. Consideration of adequacy of the evidence provides for the possibility of drawing greater confidence from the results a detailed investigation programme than that from a less mature knowledge base.

The evidence for and against each leaf hypothesis are then also elicited by expert judgment, based on an evaluation of the information available. Elicitation is carried out in two stages in which judgments are made first of the face value of the evidence and then a further judgment is made of the quality of the evidence. In other words, two questions are asked:

(iii) Assuming that the information is of high quality and trustworthy, what support does it give to the dependability of the hypothesis?
(iv) How much faith do you have that the information on which you have based your judgment is of high quality and is trustworthy?

Question (iii) is broadly equivalent to the evaluation of sufficiency in parameterization of the hierarchical relationships between hypotheses in the logical model. Question (iv) extends the evaluation of confidence further to an appraisal of the quality of the available evidence in order to modify its face value. The overall assessment of confidence in the evidence then needs to take into account both the net value of the evidence that exists and the previously estimated uncertainty due to lack of knowledge.

Propagation of Evidential Judgments

The mathematical algorithms used in ESL to propagate evidential judgments are founded on standard interval probability theory [1]. For a detailed presentation of the relevant algorithm, the reader is referred to the summary in documentation of the TESLA software tool [7], within which the propagation of confidence in 'evidence for' is treated independently from that for 'evidence against', but using the same underlying equations.

Assigned values of *sufficiency* and *dependency* are directly incorporated as parameters in the propagation algorithm. The role of the *necessity* parameter, as well as those of *ANY* and *ALL* parameters (see above), are taken into account via heuristics that govern the whole propagation process. The principle of these heuristics is to make sure that, where relevant, the confidence values associated with evidence that supports (or refutes) specific sub-hypotheses are directly propagated up the logic tree at the node in question, with no account being taken of other lines of evidence.

Visualization and Analysis

Two main approaches to visualizing outputs from an ESL analysis have been devised for use within TESLA: the Evidence-Ratio Plot and the sensitivity (or Tornado) plot [6].

The **Evidence-ratio Plot** (Figure 2) provides a visual presentation of the distribution of confidence in evidence relating to each leaf hypothesis in the logical hierarchical model (or sub-element of the model). The horizontal axis indicates the percentage uncertainty in the evidence, or (in other words) that fraction of the total available belief which is assigned to an uncommitted state. An increasing negative value along the horizontal axis represents the existence of increasing conflict in the evidence.

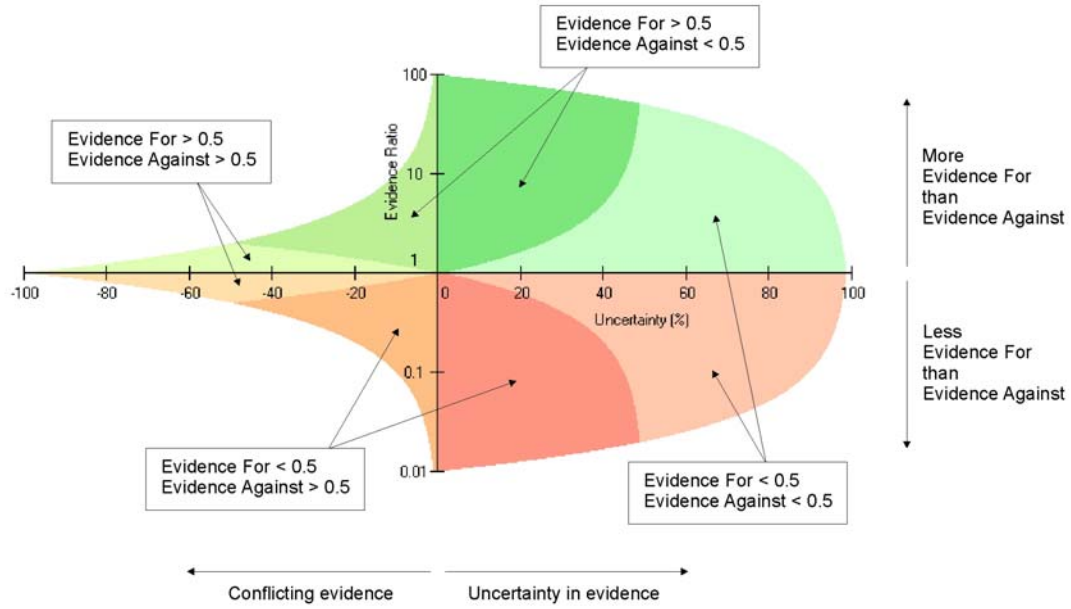

Fig. 2. Regions of the Evidence-ratio Plot

The vertical axis indicates the ratio of evidence for to evidence against associated with each leaf hypothesis. For presentational purposes, in order to avoid division by zero (or division of zero) in this calculation, any evidence values of zero are converted to a minimum value of 0.01. This results in a possible range of between 0.01 and 100. The values are then plotted using a logarithmic scale on the vertical ratio axis.

Values plotted above the horizontal axis represent a favorable balance of evidence, indicating support for the hypothesis under consideration; those below the line represent an unfavorable balance of evidence and hence a lack of support for the hypothesis. Regions representing greater than 50% evidence for and against respectively are

shaded on Figure 2, providing a visual guide to the extent of support that is judged to exist. For situations where there are with high levels of conflict (to the left of the vertical axis), the evidence ratio will still be a significant confidence measure but conflict resolution is likely to take precedence over further data gathering as a means of reducing uncertainty.

It can be informative to plot confidence in the main proposition associated with the ESL model on the same diagram as that associated with each leaf hypothesis in the logical model. This can provide a strong visual indicator of the potential implications of bias. For example, in situations where an 'outlier' piece of evidence is strongly (or even exclusively) relied upon in order to justify an overall conclusion, the end result will inevitably be skewed towards that location on the Evidence-ratio Plot. It may also be skewed towards the vertical axis, indicating a lack of awareness of the inherent uncertainties associated with judgments based on that evidence. By contrast, where full account is taken of the balance of evidence, including the possible weight of contradictory evidence and residual uncertainty, the 'true' evidential support for the top-level proposition can be clearly visualized.

The **tornado plot** (or sensitivity plot) identifies those regions where small changes in confidence in the underlying evidence values (i.e. reducing the uncertainty) have the greatest impact on the overall result. The derivation of the Tornado Plot is essentially a first-order, second moment differential calculation, and is implemented in TESLA by temporarily incrementing by a marginal amount the evidence values of each hypothesis in turn, noting the change in evidence values of the top hypothesis. The impact is thus defined by the ratio of the change in confidence associated with the main proposition to the change in evidence for the leaf hypothesis.

The impact on overall confidence associated with each piece of leaf hypothesis evidence is converted to a percentage value and plotted as a horizontal bar, colored green to describe evidence for, and red to describe evidence against. The hypotheses are then plotted in descending order of total impact, thereby giving the whole plot its tornado-like appearance from which it takes its name. An example is illustrated in Figure 3.

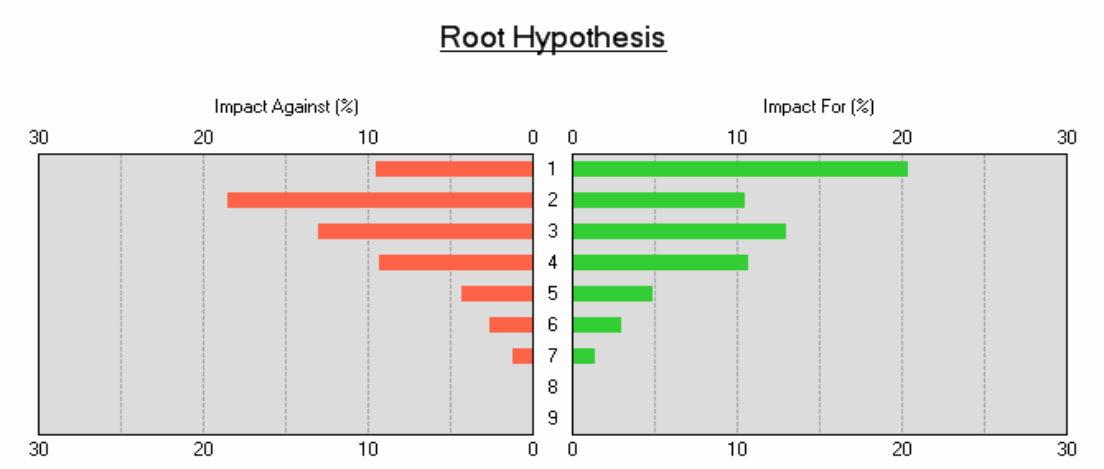

Fig.3. Example Tornado Plot

WORKSHOP ORGANIZATION

In the trial application of ESL conducted at Hanford, it was decided that the methodology should be exercised in a simulated workshop setting, under conditions where no regulatory decision rested on the workshop outcomes. It was agreed that the exercise should apply ESL to the problem of determining confidence in the infiltration rates used within PA for an evolving surface barrier with defined characteristics and design lifetime of 500 years, which could (at least theoretically) be adopted as part of a tank farm closure system. This topic was chosen because stakeholder and regulatory reviews of the recent Single Shell Tank PA for the C Tank Farm questioned assumptions relating to the definition of infiltration rates through degraded barriers. Since the infiltration rate directly affects the projected releases of residual contamination from the tank farm facilities, it is recognized as a sensitive variable. The PA had

to provide justification of infiltration rates pre-Hanford, rates through an intact barrier (cover system) and through the degraded barrier over a time period of 10,000 years, the length of time performance is calculated.

Eight specialists with specific experience relating to the issue at hand, but with differing views on assumptions made in recent PA studies, were invited to serve as the 'expert panel' for the exercise. These included regulatory officials, representatives of local and regional stakeholder groups, including Tribal nations, USDOE employees and advisors and subject-matter experts from National Laboratories. The workshop was independently facilitated by one of the authors of this paper (Michael Egan).

The trial workshop was conducted in Richland, WA, over a two-day period (10-11 June 2008). Because the exercise was preceded by a presentation and discussion on the theoretical basis for ESL, the actual amount of time dedicated to the topic itself in fact proved to be rather less than two whole days. This constraint on time arose because of the need to strike a balance between the desire to provide a thorough 'realistic' test of the methodology and that of opening the workshop to as wide a group as practicable.

A number of people (some 30 or so in total) from a range of interested parties were also invited so participate in the workshop as an 'audience'. These included regulators and local stakeholders, Hanford site contractors and representatives from other USDOE sites. For this particular exercise, which was conducted on a somewhat compressed timescale, it was found that members of the audience were usefully able to act as a resource for panel members in their deliberations on the evidence (e.g. by pointing to data provided in specific reports and analyses). In addition, from the perspective of overall evaluation of ESL as a tool for decision support, audience members were also invited to provide reflections on the process and ESL methodology as the workshop proceeded.

The results from the elicitation process were captured and analyzed using the TESLA software [7].

WORKSHOP OUTCOMES

The workshop represented a novel approach in that stakeholders, regulators and other interested parties were encouraged to work together in developing a common logical structure representing the role of different lines of evidence in support of one of the key assumptions made in post-closure PA. This logical model was then used to explore questions of confidence in those assumptions and to identify where any particular key areas of controversy or uncertainty may lie.

In what follows, the outcomes of the workshop are described under two sub-headings: (i) the analysis that was undertaken by the panel, using the ESL process; and (ii) reflections on specific issues raised by the workshop, including comments provided by panel and audience members.

ESL Results

A condensed version of the final logic tree developed at the workshop, showing only its higher levels, is reproduced at Figure 4. This shows that demonstration of acceptable surface barrier performance was considered by the panel to be composed of evidence that the performance criteria can be achieved by the design of the barrier coupled with evidence that the barrier can actually be constructed and operated to meet those specifications. Specific topics were then considered under each of these main lines of evidence.

An important aspect of the way in which main proposition was defined and the corresponding logical hierarchy was developed for this particular case was the use of *ALL* and *ANY* parameters (see above). Specifically, the logic model suggests that, in order for there to be confidence that barrier performance criteria can be achieved, all of the identified lines of evidence had to be sufficiently strong. Hence, no benefit (in terms of confidence) is gained from combining different lines of evidence, except at the deepest levels in the logic model, where the detailed evidence is evaluated; the weakest line of evidence in favor of the proposition is that which is propagated to the highest level. In this particular case, there were one or two places where complete uncertainty was assigned to the available evidence, which meant that, ultimately, 'zero confidence in favor' was propagated to the highest level.

Likewise, because of the way the logic model is structured and parameterized, the strongest line of evidence against confidence in satisfactory performance dominates 'confidence in evidence against' the top proposition, regardless of the weakness of other lines of evidence against. Because it had been agreed during somewhat hurried discussion at a late stage of the workshop that 100% 'conflict' in evidence could be assigned at one 'leaf' sub-hypothesis in the tree (relating to the likelihood that human intervention would result in unacceptable performance of the barrier over its design lifetime), this meant that '100% confidence in evidence against' was propagated to the highest level.

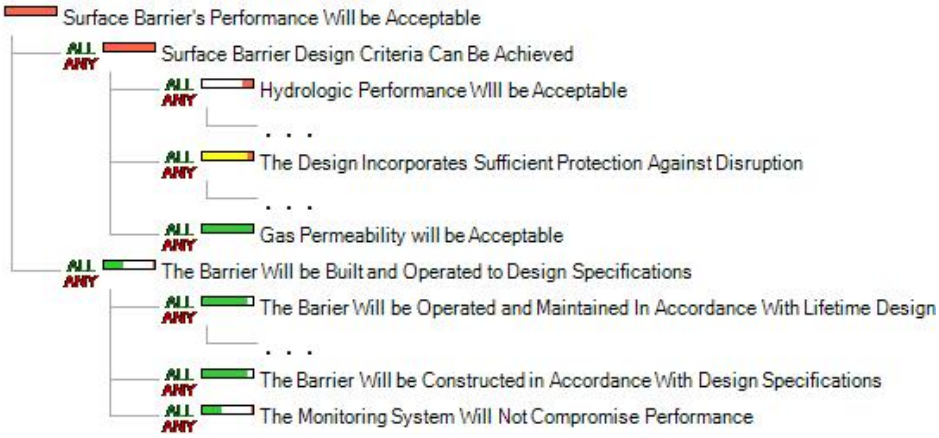

Fig. 4. Condensed ESL Model resulting from the Workshop (showing top-level nodes only)

The overall result of these assumptions was that the workshop outcome showed 100% confidence in evidence against acceptable barrier performance, with no corresponding confidence in favor. This is further illustrated by the evidence ratio plot reproduced in Figure 5. Here, the principal factors governing the confidence assigned to the top-level proposition (shown as item 1) are: 100% conflict associated with evidence relating to human disturbance (item 21) and 100% uncertainty associated with evidence relating to degradation by subsidence and natural disruptive events.

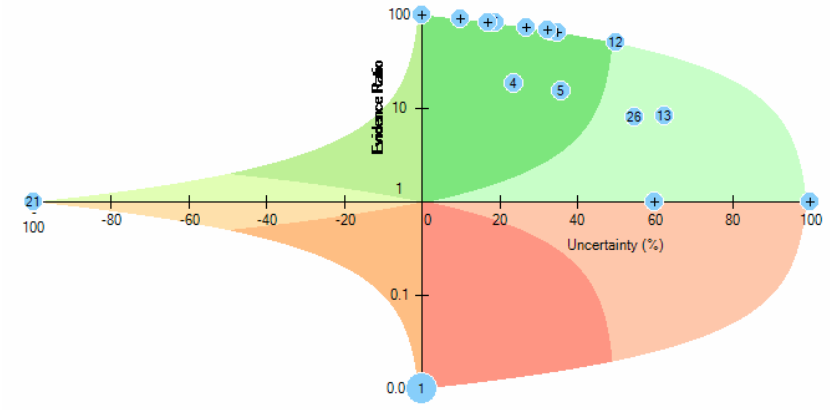

Fig. 5. Evidence Ratio Plot for the Main Proposition (Item 1) according to the Workshop ESL Model

The corresponding 'Tornado' plot was also dominated by the same key sources – those giving the highest confidence in evidence against the top-level proposition and the lowest evidence in favor.

Unavoidable time constraints meant that it was not possible to explore these issues further during the course of the workshop itself. Immediately following the event, however, the authors and others explored some of the critical

issues further, and the outcome of these further reflections were incorporated in a note distributed to all the original workshop participants. Specifically, it was argued that:

- The design of the barrier incorporates a measure of protection against seismic shock;
- The site itself was not susceptible to major natural threats of disruption on a timescale of 500 years;
- Experiments on a prototype barrier had demonstrated no significant subsidence or evidence of the impact on such processes on its capacity to minimize infiltration;
- Other caps, properly constructed elsewhere, have not sown evidence of subsidence having an impact on infiltration; and
- Submission of closure plans involving the construction of such a barrier would necessarily involve consideration of the potential threats associated with void spaces underneath the barrier.

Taken together, these factors (all of which had been mentioned in discussion at the workshop) were suggested to represent evidence that, at least on face value, should be considered to give at least some support the argument that seismic events and subsidence will not significantly affect the barrier's performance. It could therefore be considered reasonable to consider revising those evidence judgments made at the workshop where zero confidence in evidence in favor had been assigned.

There had been substantial debate during the workshop regarding the importance of human actions and their capability to disturb the cap, both in terms of the impact of this on hydrologic performance and the potential implications for intrusion into residual wastes below the ground surface. In attempting to bring this debate to a conclusion, the facilitator suggested to the panel that the force of this debate could perhaps best be represented by a '100% conflict' assignment (i.e. 100% evidence in favor and 100% evidence against). On reflection, however, spurred by the results showing how the ESL model forced the evidence judgments to be combined, this was considered to be a bad call.

The point of principle on this issue is that, in fact, there is in fact very little *evidence* at all that can strictly be applied to judgments about the likelihood and scale of human intrusion into a barrier. While there is evidence of archaeologists disturbing former burial mounds (in the USA and elsewhere), the design of any barrier and the context within which it would be deployed (if it were to be implemented as part of the closure plans for the Tank Farms) is clearly not the same as in past situations. Rather than there being 100% conflict in the evidence, it might be truer to suggest that there is in fact much closer to 100% uncertainty. The conflict that was apparent in the panel's debate on this issue, could perhaps be characterized as being more a question of 'belief' than one about 'evidence'. Indeed, it is difficult to conceive of how any directly relevant evidence could be gathered on this particular issue.

It would be inappropriate to report the detailed implications of these post-workshop considerations as part of the results of the workshop. However, it is relevant to note that they can potentially have a significant impact on the outcome of the ESL study, underlining the importance of embedding such activities in a wider iterative process.

Reflections on the Process

The nature of any systematic process is that it requires a careful step-by-step approach. This can sometimes be frustrating, especially if participants feel that it does not allow them the freedom to move more freely through the issues at stake. ESL seeks to develop structure to thought processes regarding the use of evidence – working towards common understanding through focused deliberation on those issues. It was notable that panel members found the development of underpinning logic a thought-provoking process, and that it promoted considerable deliberation among members of the panel.

Because such discussion focuses on a 'top-down' examination of how different lines of evidence relate to one another in developing confidence, rather than discussing the evidence itself, this deliberative process was potentially frustrating to some audience members, who were necessarily less actively involved in the discussion. Nevertheless, the process opened up important lines of discussion regarding the different viewpoints held by panel members that might otherwise have been hidden, or less explicit, had the discussion been focused from the outset simply on the available information database. Developing the logical hypothesis model was time consuming (taking to the end of the first day of the workshop), but ultimately provided a common framework for moving forward.

Following the event, members of the panel and audience for the workshop were invited to complete a questionnaire, designed to elicit feedback on the ESL process, its benefits and limitations. Whilst it would be improper to consider this an 'opinion poll' on the potential value of ESL for future application at Hanford, or within the wider USDOE program, the constructive nature of many of the comments received indicated to the organizers that there would be potential value from pursuing this technique for further application in areas where common ground is sought on contentious issues relating to confidence in PA and safety case development. It is not possible to reproduce all the questionnaire responses here; however, a summary of the feedback, drawing on quotations from the respondents, is provided in Table I below.

Table I. Summary of Participant Feedback following the ESL Trial Workshop

Did the ESL Workshop provide a useful process to discuss issues? • *Yes. It may be interesting to have different stakeholder groups create evidence ratio plots separately and then bring to them together to discuss only the differences [and the reasons for them].* • *Most certainly. For me it clarified even more robustly the necessity of defining the issues as a community of affected individuals.* • *Yes. It was useful, but ESL is only one of several tools that can accomplish the same or similar objectives.* • *Yes and No. Yes as it allowed some discussion about the various parameters of barrier design, construction and maintenance in a systematic process… However, two days were not enough… the facilitator often cut [discussions on stakeholder concerns] short to 'make it through the exercise'.*
Did the process bring transparency to aspects of uncertainty and/or confidence? • *No, as not enough time was allow to discuss, review and resolve the information pertaining to the issues at hand… The process demonstrated how subjective inputs to the model could be.* • *The tool is too rough to be able to address uncertainty except in very broad way. The tool is workable at [a more detailed] level, it seems, but it would be a very large effort.* • *If you mean by transparency that it exposed various aspects of uncertainty and the confidence in evidence, then yes. But again, not uniquely so.* • *Transparency is improved by the discussion. Confidence is only improved if all parties are allowed to voice their judgments and influence the decision.* • *Yes, and I think it makes some individuals and contractors very uncomfortable. This process requires that <u>all participants</u> be able to back up their claims with data, with integrity, with transparency.*
Did the workshop provide improvement over what has been done in the past? • *Yes, in that, if the method is adapted for use at Hanford, it would have the interested parties involved in the decision-making process much earlier in the sequence of events.* • *Yes. Often decisions are made by only USDOE and the contractors without, or in spite of, input from regulatory agencies and stakeholders. This process put issues on the table for open debate.* • *Yes, one of the challenges we face is not just to discuss uncertainties, but also to discuss why we have enough confidence to move to the next step in a technical program given those uncertainties. This tool is the first one I have seen that allows one to make some semi-quantitative evaluation of confidence.* • *At this point, I am not sure. I think it will depend on USDOE's path forward.*
Do you have any suggestions to make this process better? • *It might work well to have individual groups go through the process at each level, then have all parties engage in the process together… strong facilitation is needed in the Hanford environment.* • *The exercise demonstrated confusion about confidence in evidence and uncertainty in the data/evidence … this distinction must be made clear to all participants.* • *The process was designed to be inclusive, especially of regulators and stakeholders. [However] there were some regulatory concerns that were 'cut off', because of schedule constraints … this process should not be schedule constrained.* • *A more nuanced classification of the relative importance of an item in the influence-model might allow a very negative piece of evidence to remain in the model without destroying the overall confidence in the system.*

What kind of ground rules would make the process more successful?

- *Maintain respect and tolerance for differing opinions and interpretations. Put a "bookmark" in areas where major conflict seems to occur; the process can move forward, and then return to more problematic areas.*
- *Less time spent on the exercise and more time spent on the discussion and resolution of concerns.*
- *Assure that all participants are heard.*
- *If this had been an actual determination of confidence in a barrier, I would certainly have done it "in-house" first, using a facilitator (independent of the proponent) to create the technically correct relationships and influences into the model. The reason to involve the facilitator in this internal effort is to allow him/her to then explain the model to any external stakeholder groups and obtain (1) their agreement to the model as it has been constructed and their confidence levels, or (2) their suggested changes to the model and their confidence levels. Then a facilitated meeting would be held to openly compare and contrast results, and delve into only the differences.*

What, if any, are the limitations of using this process?

- *By selecting the participants, the process has the potential to bias the results – much like the selection of a jury in a trial.*
- *If not enough folks [from USDOE, its contractors, regulators or other stakeholders] are involved, the results could still prove subjective and unjustified in the public's mind, and thus unaccepted.*
- *The process of reaching consensus may be thought to take too long to pursue. Also, if the selection of participants is too narrow, there will be no confidence in the outcome.*
- *It would be labor-intensive and thus expensive to follow this process, but in order to make a safety case something like this process needs to be used, and this is a nifty way to keep the technical discussions focused on confidence, a quality often overlooked.*
- *It takes time. It is a 'messy' democratic process and therefore not always 'cost efficient'. Note that I did not say 'cost effective'. It flies in the face of political or other types of manipulation – which I find is a bonus, but others may not.*

Would you like to see ESL applied again?

- *Yes. It is worth some further investigation.*
- *No, I believe that the Kepner/Tregoe software covering decision making, uncertainty, root cause analyses, etc. is far better. It has been applied by many world-class companies to solve extremely difficult problems.*
- *Yes. [The process could be applied to] various aspects of the groundwater flow, fate and transport models and conceptual models at Hanford; same as applied to vadose zone; treatment methods for carbon tetrachloride plume.*
- *I would like to see it applied internally, to 'check it out' within my project by trying it on a specific sub-system-level issue. A small, very focused effort to start with. We would need a facilitator to start us on this process. If that looks useful and teaches us something, then we would expand its use, perhaps using an internal, but trained, facilitator who has learned from the first effort.*

Other comments?

- *More background information and reading materials before the workshop would have allowed participants to come better prepared with a better understanding of how the process works.*
- *[The process] does not need to be used [solely] to establish understanding with stakeholders. It can also be effectively used as an internal tool only, at various stages of a technical program, as a way to see if over time there is an improvement of confidence and whether resources are being used to address areas where it appears there can be significant payoff in terms of enhancing confidence.*
- *I would suggest that if this exercise is to be useful in the design of site barriers, then a parallel exercise on barrier monitoring and performance criteria be implemented.*
- *I hope that the invitation we received for this workshop is not an empty gesture. It is important that the Hanford leadership understand the value of having us all at the table, even if it is 'messy'.*

A potential problem with arranging a single, experimental workshop is that people may not have confidence that their issues are being given due regard, because there is no wider context in which to situate that process. The facilitator was conscious that constraints on the time given to the exercise meant that discussions on potentially important concerns, given wider questions of confidence in PA studies carried out for the Hanford site, were necessarily curtailed. At times, there was also evident discomfort with the artificiality of the exercise, and the assumptions that had to be introduced in order set some constraints on the discussion, simply in order to make the

analysis possible. Moreover, in the absence of a sufficiently well constrained or clearly defined 'real-life' issue to address as part of the exercise, it was necessary to create one. This is critical because, in order to make judgments about 'confidence in evidence', the discussions must be framed by assumptions regarding the particular situation to which we want that evidence to apply. Attempts were made to deal with such possible concerns in discussions with nominated panel members and others ahead of the workshop; nevertheless, earlier and tighter clarification of the scope of the exercise would have been an advantage.

The necessary time constraints associated with the trial were an important element of the experience and the outcome from the workshop. In practice, when ESL is being used to address complex and potentially controversial topics, it is rare to compress everything into a single workshop. For example, there might typically be an initial meeting to clarify objectives, to go through key issues concerned with the decision logic, and to identify the success and failure criteria for confidence in evidence. Often, this will be followed by a period of review and reflection on the outcome, before coming back to populate the logic model with judgments about evidence. Further iteration and reflection on the outcome would then take place. Whilst it was possible to reproduce some of the main elements of the process during the course of the workshop, corners were inevitably cut in some places and this may have influenced some of the judgments that were made. In a 'real' application, time would normally be given to collective learning from the initial results, and to consider whether the judgments and assumptions embedded in the logic model should be revisited in the light of the first provisional outcomes.

Partly because of such constraints, the facilitator was also conscious of 'leading the witness' more than should perhaps ideally be the case. So long as it is necessary to have an eye to the schedule, there is always a potential conflict with providing sufficient space for deliberation, in order that participants reach their own conclusions. It should not necessarily be considered wrong for an independent (and technically informed) facilitator to provide inputs to help things along, provided that this is done appropriately. However, there was at least one occasion during the workshop (discussed above) when this may have led to a fault in the way in which evidence was handled.

RECOMMENDATIONS

In decision making that is based on an open-world perspective on the available evidence (i.e. a recognition that it can be appropriate to assign a level of belief to an uncommitted state), it is not possible to deal with absolute truths or in mathematical terms of accuracy and precision. The ESL methodology works in quantitative mathematical terms, and the top-level result is a measure of overall confidence in the model, or hypothesis, under evaluation. However, care is required in the interpretation of such output, to avoid the GIGO (garbage in/gospel-out) epithet that is traditionally associated with apparently numerically precise output from fuzzy inputs. The primary inputs to the ESL process (logical model parameterization and evidence evaluation) are best understood in terms of linguistic or verbal expressions of subjective judgment, and hence are essentially 'soft' or 'fuzzy', terms. This is a deliberate aspect of the process – as one questionnaire respondent observed, when a site operator is attempting to make a 'case' based on the outcome of PA and other work, the challenge is not just to discuss (or even to reduce) uncertainties, but also to discuss the basis for confidence to move to the next step in a technical program in the light of those uncertainties. In this respect, subjective judgment (and 'fuzzy' assessments of confidence) cannot be avoided; ESL is designed to capture those judgments within a formalized, transparent process.

Feedback from participants in the trial application suggests that there ESL is potentially capable (alongside other decision support tools and techniques) of supporting more wide-ranging discussions to address the question of confidence in decision making informed by PA. It is potentially useful both as an internal program management tool, to ensure that key issue of confidence are being identified and addressed, and in engagement with stakeholders on major areas of controversy. Recommendations for future workshops, applying this technique, include:

- Prior preparation of panel members by carefully defining the elicitation topic prior to the session;
- Clarity regarding the criteria that are relevant to assessing the available evidence, separating fact and value judgments;
- Allowing sufficient time for all relevant issues to be addressed to the satisfaction of participants, including breaking up the discussion into two or more sessions, with time between to allow for the logic model to be reviewed, for information to be processed and for evidence to be systematically compiled; and
- Providing assurance that related concerns not covered by the objectives of a specific workshop will be noted and addressed as part of an ongoing program to address key issues.

An important outcome of this particular exercise was that a number of the panelists expressed concerns in the ability of USDOE to provide institutional controls during the 500 year expected barrier design life. Outstanding policy issues regarding institutional control and land use (and their implications for the acceptability of near-surface disposal of residual wastes) made it difficult for stakeholders and regulators to come to resolution on issues relating to the quality of key aspects of evidence regarding barrier effectiveness. Policy issues such as these can not be solved at the technical level, but must be resolved through other forums to address stakeholder and regulator concerns. This also has implications for the way that PAs are developed and presented. Assumptions embedded in analyses for specific closure plans and designs need to be based on clear justification of the proposed closure strategy (including the extent of proposed waste retrievals) and assumed barrier concept, taking into account both policy considerations and an examination of the potential implications of alternative courses of action.

REFERENCES

1. W. CUI and D.I. BLOCKLEY, "Interval probability theory for evidential support." International Journal of Intelligence Systems, 5, pp.183-192 (1990).

2. L. FOLEY, L. BALL, A. HURST, J. DAVIS, and D. BLOCKLEY, "Fuzziness, incompleteness and randomness: classification of uncertainty in reservoir appraisal", Petroleum Geoscience, 3, pp.203-209 (1997).

3. J.W. HALL, D. I. BLOCKLEY, and J.P. DAVIS, "Uncertain inference using interval probability theory", International Journal of Approximate Reasoning, 19, pp247-264 (1998).

4. D. BLOCKLEY and P. GODFREY, "Doing it Differently – Systems for Rethinking Construction", Thomas Telford, London (2000).

5. J.P. DAVIS and J.W. HALL, "A software-supported process for assembling evidence and handling uncertainty in decision-making", Decision Support Systems, 35, pp.415-433 (2003).

6. R.A. BOWDEN, "Building confidence in geological models", In: Curtis, A. and Wood, R. (Eds.) Geological Prior Information, Geological Society, London, Special Publications (2004).

7. M.J. EGAN, "Evidence Support Logic: A Guide for TESLA Users", Quintessa Limited (available at: http://www.quintessa.org/software/index.html?tesla.html).

8. T. SEO, H. TSUCHI, R. METCALFE, Y. SUYAMA, H. TAKASE, A. BOWDEN, M. TOIDA, M. FURUICHI, A. MATSUMURA, M. YOSHIMURA, and A. HORIO, "A decision making methodology taking into account uncertainties and its possible application for the selection of preliminary investigation areas", Extended abstract for a presentation given at the Distec 2004 International Conference, Berlin, Germany, 26-28 April (2004).

9. M.J. EGAN and R.A. BOWDEN, "Application of evidence support logic to the role of palaeohydrogeology in long-term performance assessment", Quintessa Report for UK Nirex Limited, QRS-1219A-1 (2004).

10. R. METCALFE, P. MAUL, S. BENBOW, C. WATSON, D. HODGKINSON, A. PAULLEY, L. LIMER, R. WALKE, and D. SAVAGE, "A Unified Approach to Performance Assessment of Geological CO_2 Storage", Proc. Ninth International Conference on Greenhouse Gas Technologies, GHGT-9, Washington DC, 16-20 November (2008).

11. E. WALTZ and J. LLINAS, "Multisensor data function", Artech House, London (1990).

12. S.O. FUNTOWICZ and J.R. RAVETZ, "Uncertainty and Quality in Science for Policy", Dordrecht: Kluwer (1990).

13. H. HOFFMAN-REIM and B. WYNN, "In risk assessment, one has to admit ignorance", Nature, 416, p123 (2002).