PNNL-15240

Pacific Northwest
National Laboratory
Operated by Battelle for the
U.S. Department of Energy

Review for the Nuclear Regulatory Commission of the Draft NUREG Report: Estimating Loss-of-Coolant Accident (LOCA) Frequencies Through the Elicitation Process

A. J. Brothers

June 2005

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL-15240

Review for the Nuclear Regulatory Commission of the Draft NUREG Report: Estimating Loss-of-Coolant Accident (LOCA) Frequencies Through the Elicitation Process

A. J. Brothers

June 2005

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Abstract

Reviewer comments on the Nuclear Regulatory Commission's draft NUREG report: Estimating Loss-of-Coolant Accident (LOCA) Frequencies Through the Elicitation Process. Briefly describes alternative methods of combining probability judgments, and comments on the process used by the NRC to elicit and combine expert's judgments. Recommendations for sensitivity analysis and suggestions to improve clarity of report are provided.

Contents

Figure

Table

I. General Impression

The process used to elicit frequencies from multiple experts and combine them into overall estimates was carried out in a manner consistent with accepted engineering practice. In particular, the methodology used to obtain frequencies from individual experts was of sound design and execution. Notable aspects of the methodology are the selection and training of experts, the identification of issues, and finding meaningful response modes for small probabilities; all of which lend confidence to the validity of the process. The iterative process, in which the experts were given opportunities to interact with each other and to understand the underlying reasons for differences of opinions, and given the opportunity to revise their opinion afterwards, also lends validity to the process. Training on almanac questions to help avoid overconfidence was also a positive step in the process. The authors show evidence of a good working knowledge of the literature on the psychology of probability elicitation, and the potential pitfalls of well known heuristics and biases. The variables were well defined and broken out into a sufficient level of detail for which meaningful judgments could be given. The use of ratios relative to a base case overcomes the problems people have in making small probability estimates. In sum, the process used to obtain frequencies from individual experts was sound in both the overall approach and in the details.

In addition to the elicitation process, the method for combining expert judgments was a reasonable one. In particular, the overall strategy of aggregating individual expert's responses to obtain an overall estimate of bottom line parameters and then aggregating those to obtain the overall estimates was a good one. As the authors state, this retains the consistency within experts yet captures the variability across experts in the overall estimates. The elicitation process and the general strategy for aggregating and then combining judgments will probably generate little controversy. The actual method used to aggregate individual judgments and of combining the judgments will have the greatest impact on the recommended standards. In particular the combining methodology, while defensible, has generated the most discussion. It is this reviewer's belief that all combining methods are compromises and require judgment in their selection. The method chosen by the authors of the report represents a reasonable approach. The next section presents a discussion of alternative methods for combining probability judgments.

II. Alternative Methods for Combining Probability Judgments

When multiple experts provide different estimates of the frequency of an event, it becomes necessary to somehow reconcile the differences or combine the estimates to obtain a single estimate. Different methods have been proposed for combining individual judgments, and the resulting estimate can differ greatly depending on the method chosen. This is especially true when there is marked variability among the estimates or if there are outliers. Combining methods fall into three broad categories: behaviorally by expert interaction, analytically by a mathematical formula or algorithm, or through a process of Monte Carlo simulation. No one approach is best, or even theoretically justified, for all circumstances, and number of factors should be taken into account when deciding on the approach to be taken. These include how the information will be used, the independence of the estimates, and the extent to which one can assume that the experts are equally knowledgeable in all relevant technical information.

II.A. Combining Analytically

There is a vast literature on combining individual experts' probability judgments in order to obtain an overall judgment. This literature identifies a variety of combining methods and criteria by which to judge their effectiveness. There is no single method that is applicable in all circumstances or even criteria which are universally agreed to by which to judge the performance of the methods. There are two perspectives on analytical combining methods: the axiomatic and the Bayesian (Genest & Zidek, 1986). In the axiomatic approach, the emphasis is on the functional form of the combining rule and determining which functional forms are appropriate given the underlying assumptions. In the Bayesian approach, the experts' probabilities are viewed as data to be aggregated using likelihoods. Depending on the independence assumptions, a Bayesian combining rule can take on a variety of functional forms, including a linear weighted average, and may or may not satisfy a given set of criteria for judging the reasonableness of the rules (Clemen and Winkler, 1986). Axiomatic approaches are also not without their issues. The set of assumptions embodied in the axioms may or may not be appropriate to the circumstances, and it may be hard to decide whether or not a particular set of axioms apply in a given situation. As an example, consider the Unanimity Principle discussed by Clemen and Winkler (1990). According to the unanimity principle, if all the experts agree that the probability of an event is a particular value, then the combined judgment should also be this value. This seems like a reasonable and desirable property for a combining method to have. Both arithmetic averages and geometric means satisfy the unanimity principle and they also satisfy the more general Compromise Principle, which states that the combined estimate should be in the range of the individual estimates. Yet, as Clemen and Winkler point out, one can think of situations in which this would not be a desirable property, and there are combining methods that don't satisfy this principle that would be appropriate to use in those situations. Bayesian methods may or may not satisfy these two principles depending on the underlying conditional independence assumptions. For example, if one assumes independence of the experts conditional on the event whose probability is being estimated, then Bayesian methods don't uniformly satisfy the Compromise principle and it may be perfectly reasonable for it to be violated. The situation discussed in Clemen and Winkler involves predicting the probability of rain in which the prior belief is 0.5, and the expert weather forecasters all agree that there is a 0.55 chance of rain. For two experts, assuming conditional independence, the posterior probability is 0.60. If there are ten experts the posterior probability is 0.90. This is a violation of the unanimity principle, and because the posterior estimate is outside the range of the original estimates; it is also a violation of the

compromise principle. However, the use of a Bayesian posterior as a combining rule is reasonable for the circumstance in which the weather forecasters are considered independent sources of data because as the data accumulates in favor of rain the probability of rain should be an increasing function of the amount of data. On the other hand if the weather forecasters all base their estimates on the same data source, independence may not be justified and the combining rule could be inappropriate. Clemen and Winkler show that if one makes different assumptions of conditional independence, then Bayesian combining methods will uniformly satisfy the unanimity and compromise principles. What is worth noting here is that both the methods and criteria are situational specific and categorical statements of the universal applicability of a particular method need to be examined closely with regard to whether the underlying assumptions are reasonable for a particular application[1].

A number of analytical combining methods have been advocated. These include arithmetic averages of means, medians, logs of odds ratios, etc. Each of these methods has its advocates and theoretical and practical justifications. There is no single correct answer that fits all circumstances. One consideration in a choice of a method is the extent to which the combined measure of central tendency is representative of majority opinion while still giving some consideration to outliers. Also to be considered is whether the combined distribution should reflect both the inherent uncertainty in the phenomenon itself and the variability in the opinions across expert. Theoretical justifications for combining rule need to consider the practical implications. Theoretical arguments may depend on a number of assumptions that may be difficult to verify in practice. To blindly follow a specific combining algorithm because it seems to have the theoretical high ground without consideration of the implications in a particular application can lead to overly conservative estimates that result in additional resource requirements that might best be utilized elsewhere.

There are two special considerations in the NUREG that should be taken into account when selecting a combining method. These are that the frequencies being estimated are very small, and the fact that there are outliers that differ by several orders of magnitude with the majority opinion. If a desirable property of the combining method is that it reflect the majority opinion while giving due consideration to the outliers, then in these circumstances, a simple arithmetic mean, while some might argue that it is theoretically justified, may give too much weight to the outlier depending on its relative value. The outlier is just one opinion and the combined judgment should not be governed by a single individual, but should be reflective of the group opinion with some adjustment for the minority opinion. One thing to keep in mind with regard to the justification for using an arithmetic average is that the assumptions used to make a strong case for arithmetic average may not be met. In particular, while all the experts may indeed be experts, the nature and the extent of their expertise may vary, and consequently the amount of relevance with regard to a particular judgment may also vary. Thus the assumption that one is sampling from experts with the implicit additional assumption that their expertise is equally valid or relevant may not in fact be true. For this reason, some might advocate tossing out outliers, or using an "Olympic" scoring in system, in which the top and bottom judgments are tossed out. When the

[1] A theoretical aside is whether a group probability judgment is even a meaningful concept. Morris (1986) has called into question the legitimacy of an "aggregate opinion". The Bayesian perspective is that probabilities are subjective beliefs, and according to Morris only individuals can have beliefs and not groups. Consequently, there is no such thing as a joint state of information. Shafer (1986) on the other hand has pointed out that individuals can be as divided in their thinking as groups, and by analogy argues for the legitimacy of a group probability estimates. So at the heart of combining methods are fundamental philosophical differences as to what it actually means to combine probability estimates as well as what method to use and what are the appropriate assumptions.

outlier is greater than the group judgments, as is generally the case in the NUREG report, a less drastic means of ensuring that the combined judgment reflects the group opinion is to use geometric means or the average of log odds (which in this case, because of the small probabilities involved, amounts to the same thing.) This results in a group judgment that reflects the majority opinion while making some adjustment for the opinion of outliers. This seems like the most reasonable approach for the NUREG. To quote Robert Winkler: "My own feeling is that different combining rules are suitable for different situations, and any search for a single, all purpose, "objective" combining procedure is futile… Since there is no single combining procedure for all seasons, a subjective element cannot be avoided… Some situations will engender more agreement than others on the combining rule that seems most appropriate, but an "objective" rule is an unattainable goal." (Winkler, 1986) Consequently, it would seem to me that the approach taken should represent a reasonable compromise for the situation given. This is what the authors of the NUREG have done in this reviewer's opinion.

II.B. Combining Using Monte Carlo Simulation

Monte Carlo simulation combines entire distributions of expert judgments in a way that captures both the experts' uncertainty about the phenomena and the variability between experts. It is sometimes debated whether cumulative distributions functions should be averaged vertically or horizontally. I think a fundamental issue, which is recognized by the authors of the paper we are reviewing, is estimating the variance of both the phenomena itself and the variance in opinion. As a simple example, if two experts agree on the variance but only differ on means, then if the distributions are combined using the average of the means and the variances, the combined distribution will be narrower than if the variability of the experts is an additional component of the variance in the combined distribution. The former says the experts disagree about the means, but they agree there is only so much variability. The later says the experts disagree about the values and the combined variance captures uncertainty within the phenomena as well as the variance within experts. The appropriate method will depend on the circumstances.

II.C. Combining Behaviorally

Combining estimates behaviorally consists of some type of group interaction to arrive at an agreed upon estimate. The interactions can be face to discussion or rounds of revised judgments based on anonymous reasons for the judgments of the other experts as in the Delphi procedure. Some advocate behavioral approaches to the exclusion of analytical combining methods and when consensus in unattainable recommend sensitivity analysis in lieu of a combined estimate (Morgan and Henrion, 1990). The perspective of the authors of the NUREG report is to celebrate the diversity of opinion rather than promote discussion among the experts with the intention of developing a consensus view. They recruited experts with a variety of perspectives with the intention of calculating a combined estimate reflective of the variety within the group. On the positive side, this is less likely to lead to a "group think" in which one few people are able to sway the group into a consensus opinion that doesn't reflect the diversity of opinion. On the negative side more discussion has the potential for additional insights beyond what any individual would generate on their own. While "group think" is a negative, there is value seeing to what extent a group mind can be created from the panel of experts. However, the literature is confusing and often contradictory. Some authors recommend exploring group interaction, while others claim that it is better to just average the experts' estimates without any interaction. Possibly more could have been done from a behavioral perspective to create more consistency

among the experts. One additional consideration is that if "group think" is bad, in that one or two individuals dominate the consensus estimate, then it would seem to be equally bad to combine the estimates analytically in way that allows one or to outliers to dominate the estimate.

III. Calibration and Overconfidence

The literature on overconfidence is not always consistent. A review in von Winterfeldt and Edwards (1986) suggests that changing the response mode (e.g., NUREG asks for ratios) is one way of overcoming overconfidence. They also claim that experts with substantive knowledge and expertise in estimating probabilities are not overconfident. The NUREG may not suffer from overconfidence for both of these reasons. The response mode has been changed from direct estimates of probabilities to ratios relative to a base case. This would seem to help to alleviate the problem of overconfidence, but raises some issues concerning validity. So it may not be appropriate to adjust the experts' confidence limits rather than take them at face value. Just because the experts showed overconfidence in the almanac questions does not necessarily imply that they would not be well calibrated in providing estimates for which they have domain knowledge. There is also evidence that the use of probing questions during the elicitation process, such as, asking them how they would account for some named value outside their stated range to explore the limits, can be effective to overcome tendencies for overconfidence.

IV. Specific Combining Measures: Geometric versus Arithmetic Means

The review generated considerable discussion as to whether the combined measure of central tendency should be based on geometric means or simple averaging. While simple averaging has a lot to be said for it in terms of getting results as good or better than more complicated combining methods in many situations and being technical defensible, the assumptions that make it theoretically appealing may not be met in this particular case. Foremost among them is the extent to which all the panelists are really "experts", or the extent to which their expertise is relevant for a specific estimate. In addition to the assumptions not necessarily being met, the other consideration is whether it is reasonable to use averages within this context. Given the diversity of opinion and the small probabilities, simple averages may not be appropriate in that it results in the small numbers being overwhelmed by larger outliers. It is my opinion that having combined means that are representative of all panelists, with due consideration given to the minority view, overrides other considerations.

The authors of the NUREG report use Geometric Means. Geometric means are a traditional method of combining judgments. Geometric means especially make sense for small probabilities. While von Winterfeldt and Edwards (1986) recommend simple averages in general, they specify that for small probabilities, the average should be based on log odds. This recommendation is consistent with the method used in the NUREG, and with intuition that all the expert's opinions should be reflected in the results.

The geometric mean will be closer to the group consensus when an outlier is greater than the group estimates. This was the situation in the majority of instances where there was an outlier. For these situations the geometric mean is more conservative in preserving the group judgments. The use of the smaller group probability estimate resulting from the geometric mean would save needless costs assuming the outlier was an anomaly and didn't represent the real "truth". On the other hand, in those situations where the outlier is less than distribution of group judgments, the use of an arithmetic mean preserves a higher probability estimate thus assuring a more conservative design based on the group opinion. Which type of average is closer to the group distribution and less influenced by outliers depends on the relationship between the outlier and the group distribution. Table 1 and Figure 1 show this relationship for a sample set of data. In Table 1, the five judgments are all close to E-6 and the outliers are either one or two orders of magnitude above or below this value. The table compares the arithmetic and geometric means. This information is shown visually in Figure 1. In each panel in the figure, the five experts (E1-E5) are the same and the outlier is varied to be one or two magnitudes either above or below the group judgments. As seen in both the table and the figure, the arithmetic mean will always be larger than the geometric mean. Consequently, which lies closer to the group distribution depends on the direction of the outliers. Because most outliers in the report were above the group distribution, one can argue that the geometric mean better preserves the group judgments in the majority of cases.

Table 1. Relationship between Geometric and Arithmetic Means as Function of Outlier

	Outlier Relationship in Orders of Magnitude				
	OneBelow	TwoBelow	OneAbove	TwoAbove	Yvalue
E1	1.00E-06	1.00E-06	1.00E-06	1.00E-06	0.00E+00
E2	1.10E-06	1.10E-06	1.10E-06	1.10E-06	4.00E-04
E3	9.90E-07	9.90E-07	9.90E-07	9.90E-07	-4.00E-04
E4	1.20E-06	1.20E-06	1.20E-06	1.20E-06	8.00E-04
E5	9.80E-07	9.80E-07	9.80E-07	9.80E-07	-8.00E-04
Out	1.00E-07	1.00E-08	1.00E-05	1.00E-04	0.00E+00
Avg	8.50E+00	8.80E-07	2.55E-06	1.75E-05	0.00E+00
GM	7.10E-07	4.84E-07	1.53E-06	2.25E-06	0.00E+00

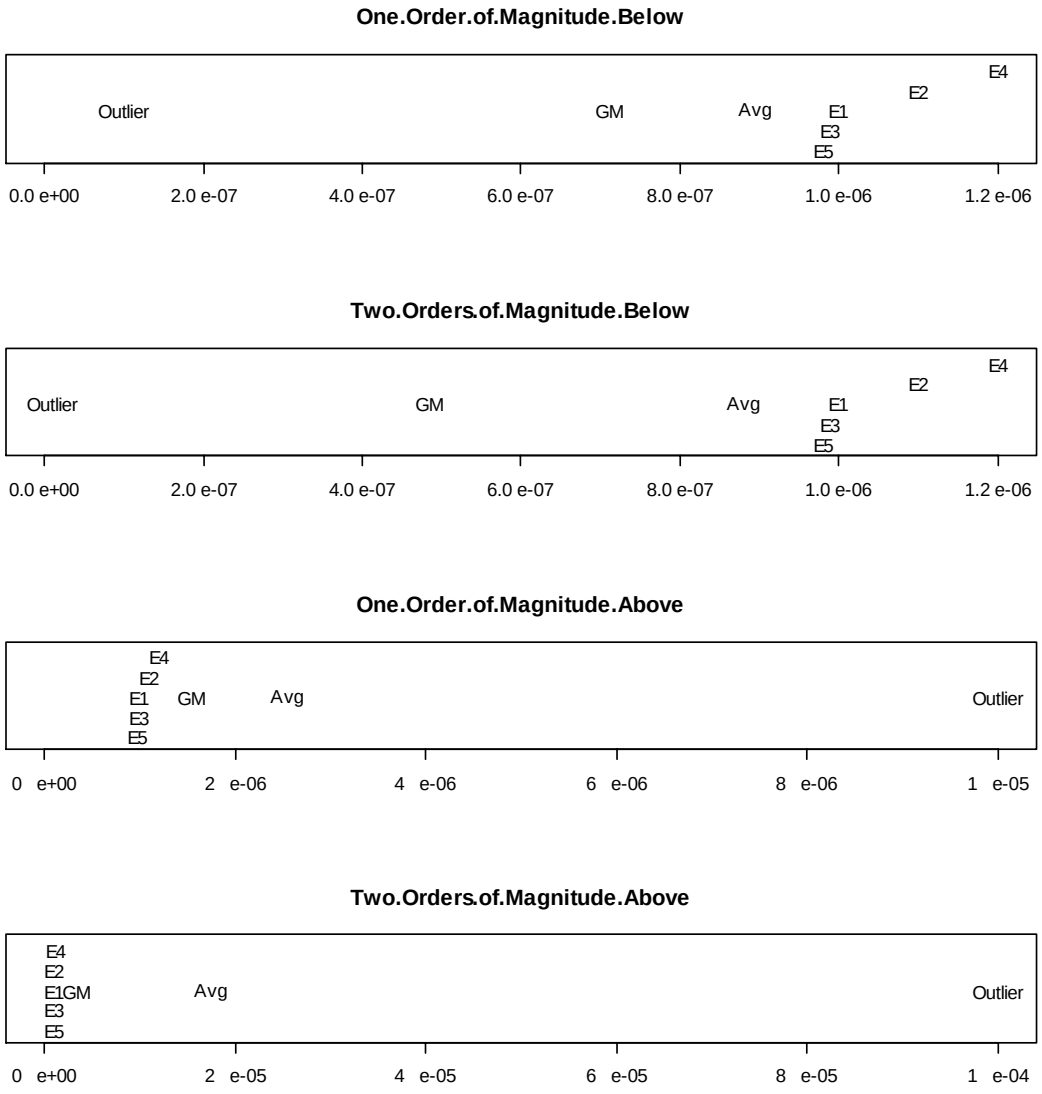

Figure 1. Relationship between Geometric and Arithmetic Means as a Function of Outlier

V. Suggested Sensitivity Analysis and Additional Specific Recommendations

The following recommendations were made to an early version of the draft document. Many of these have been carried out.

It is recommended that the following sensitivity analysis be carried out:

1. Sensitivity Analysis on Functional Form. Rather than exactly specify a split lognormal, find the best fit for the three data points for a variety of possible distributions at the basic response level. These include:
    a. Single Lognormal.
    b. Weibull
    c. Gamma
2. Sensitivity Analysis on means using split lognormal. Use the lower or upper distribution to calculate means depending on which is the more conservative for a particular basic response question.
3. Combining panelist's basic response questions. Use entire distributions as suggested above based on goodness of fit using Monte Carlo simulation to sum these distributions.
4.

Other Recommendations:

1. Exposition. Better clarify what is being summed up in section F.2 bottom of page. In particular more explanation and clarification of the following:
    a. Explain that each "question" is a particular scenario or system that had been identified by the panelist as being relevant.
    b. More explanation of how the panelists decomposed the problem and how it was different for different panelists.
    c. More explanation of how panelists' decomposed answers were combined to obtain their "bottom line estimates".
    d. Better define "bottom line estimates".
    e. Expand flow chart to show process of summation based on decomposition.
2. "Targeted Adjustment" method. While this gave similar results to the Error Factor Adjustment, it strikes me as being too serendipitous. It seems to be in the spirit of fitting existing data rather than an adjustment based on a rule.

VI. References

Clemen, Robert T. and Robert L. Winkler. 1990. "Unanimity and Compromise Among Probability Forecasters." *Management Science*, Vol. 36, No.7.

Genest, Christian and James V. Zidek. 1986. "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science*, Vol. 1. No. 1, 114-148.

Morgan, Granger and Max Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Qualitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.

Morris, Peter A. 1986. "Comments to Genest and Zidek."

Shafer, Glenn. 1986. "Comments to Genest and Zidek."

Von Winterfeldt, Detlof and Ward Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge.

Winkler, Robert L. 1986. "Comment to Genest and Zidek."

Distribution

No. of Copies

OFFSITE

Nuclear Regulatory Commission

Lee Abramson
Carolyn Fairbanks
Charles Green
John Lane
Arthur Solomon
Robert Tregoning

Battelle Memorial Institute

P.M. Scott

No. of Copies

ONSITE

Pacific Northwest National Laboratory

A.J. Brothers
S.R. Doctor
B.A. Pulsipher