Available online at www.sciencedirect.com
SCIENCE DIRECT®
Reliability Engineering and System Safety 85 (2004) 355–369
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

Summary from the epistemic uncertainty workshop: consensus amid diversity

Scott Ferson[a,*], Cliff A. Joslyn[b], Jon C. Helton[c], William L. Oberkampf[d], Kari Sentz[b]

[a]*Applied Biomathematics, 100 North Country Road, Setauket, NY 11733, USA*
[b]*Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[c]*Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287, USA*
[d]*Sandia National Laboratories, Albuquerque, NM 87185-0828, USA*

Abstract

The 'Epistemic Uncertainty Workshop' sponsored by Sandia National Laboratories was held in Albuquerque, New Mexico, on 6–7 August 2002. The workshop was organized around a set of Challenge Problems involving both epistemic and aleatory uncertainty that the workshop participants were invited to solve and discuss. This concluding article in a special issue of *Reliability Engineering and System Safety* based on the workshop discusses the intent of the Challenge Problems, summarizes some discussions from the workshop, and provides a technical comparison among the papers in this special issue. The Challenge Problems were computationally simple models that were intended as vehicles for the illustration and comparison of conceptual and numerical techniques for use in analyses that involve: (i) epistemic uncertainty, (ii) aggregation of multiple characterizations of epistemic uncertainty, (iii) combination of epistemic and aleatory uncertainty, and (iv) models with repeated parameters. There was considerable diversity of opinion at the workshop about both methods and fundamental issues, and yet substantial consensus about what the answers to the problems were, and even about how each of the four issues should be addressed. Among the technical approaches advanced were probability theory, Dempster–Shafer evidence theory, random sets, sets of probability measures, imprecise coherent probabilities, coherent lower previsions, probability boxes, possibility theory, fuzzy sets, joint distribution tableaux, polynomial chaos expansions, and info-gap models. Although some participants maintained that a purely probabilistic approach is fully capable of accounting for all forms of uncertainty, most agreed that the treatment of epistemic uncertainty introduces important considerations and that the issues underlying the Challenge Problems are legitimate and significant. Topics identified as meriting additional research include elicitation of uncertainty representations, aggregation of multiple uncertainty representations, dependence and independence, model uncertainty, solution of black-box problems, efficient sampling strategies for computation, and communication of analysis results.
Published by Elsevier Ltd.

Keywords: Epistemic uncertainty; Challenge Problems; Aleatory uncertainty; Aggregation; Repeated parameters; Interval uncertainty; Independence in imprecise probability models

1. Introduction

This paper summarizes the results of the Epistemic Uncertainty Workshop [1] held 6–7 August 2002 in Albuquerque, New Mexico. It reviews the purpose of the Challenge Problems [2] around which the workshop was organized, answers some of the questions and concerns voiced by participants during the workshop about why the Challenge Problems were structured as they were, and recounts some of the most interesting points of dispute and consensus arising from the discussions there. It also synoptically compares the quantitative answers to the Challenge Problems offered in the various papers of this special issue of *Reliability Engineering and System Safety*. In most cases, these papers represent the completion in published form of the presentations made at the workshop. In a few cases, papers were contributed by researchers who were not able to participate in the workshop itself. This paper was not seen by the authors who contributed to this special issue before they wrote their papers.

In Section 2, we review the origin and intentions behind the Challenge Problems, and describe several technical issues that were considered during their development and during the workshop. In Section 3, we summarize the proceedings from the workshop, which appear in this special issue, and provide preliminary quantitative comparisons of the answers to the Challenge Problems provided by

* Corresponding author. Fax: +1-631-751-3435.
E-mail address: scott@ramas.com (S. Ferson).

0951-8320/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.ress.2004.03.023

the various papers. A brief summary discussion and outline of future research needs is presented in Section 4.

2. The challenge problems

Over the last 3 years, the Epistemic Uncertainty Project at Sandia National Laboratories has focused on the question of how epistemic and aleatory uncertainty [3–7] should be handled within probabilistic modeling and quantitative risk analyses. This project investigated the use of both probabilistic and non-traditional forms of uncertainty quantification for modeling the performance of complex engineering systems. We identified a number of technical questions about which we could find little or no consensus across the disparate communities of risk analysts, modelers and information theorists engaged in the quantitative analysis and simulation of such systems. In order of importance, these questions were:

(1) How should epistemic uncertainty about a quantity be represented?
(2) How can epistemic and aleatory uncertainty about a quantity be combined and propagated in calculations?
(3) How should multiple estimates of uncertain quantities be aggregated before calculation?
(4) How should the technical issue of repeated uncertain parameters be handled in practical calculations?
(5) How might various approaches be adapted for use in practical calculations based on sampling strategies?

As a part of the Sandia Epistemic Uncertainty Project, a set of Challenge Problems [2] was devised for the purpose of sparking discussion about the variety of important practical issues involving epistemic and aleatory uncertainty embodied in these questions.

The Challenge Problems are a series of simple but numerically explicit problems expressed in clear and context-free terms that, it is hoped, reduce very complex analytical issues to purely mathematical and computational ones. Challenge Problem A addresses the evaluation of the simple arithmetic expression $(a + b)^a$, where a and b are characterized with different degrees of information. There are six subproblems to Challenge Problem A, representing situations ranging from very poor states of information about the inputs to full information about the distributions of the inputs. Challenge Problem B involves a somewhat more complicated mathematical expression that calculates the steady-state magnification factor for a hypothetical mass-spring-damper system from inputs about which there is variously good or poor information.

Of course, the Challenge Problems as constructed are mere models for the kinds of problems encountered in real engineering modeling, and thus they neglect many important issues in uncertainty analysis. Participants at the workshop asked, for instance, why independence was assumed among all variables when functional relationships and stochastic dependence among variables and common-mode or common-cause failures may actually be the rule rather than the exception in many real-world engineering applications [8]. Other participants noted that a major issue that had been neglected in the design of the Challenge Problems was that of model uncertainty [9,10], i.e. the uncertainty in the mapping of inputs to outputs. In many high-profile problems in risk analysis, ranging from modeling global climate change to failure rate assessment in abnormal environments, it is the process models themselves that are perhaps the most uncertain component of a quantitative assessment.

The reason that the Challenge Problem set ignored these and other issues was so they could focus on the even more fundamental issues of representing and propagating epistemic uncertainty. We completely agree that accounting for dependence among variables and model uncertainty are crucial issues that deserve special and concentrated attention. It might be reasonable to consider another set of challenge problems to address these issues in the future.

Discussions about the Challenge Problems continue online at the 'Submitted Comments' section of Sandia National Laboratories' Epistemic Uncertainty Project website (http://www.sandia.gov/epistemic/).

2.1. Motivation of the challenge problems

The intent behind the Challenge Problems was to pose some simple questions that could serve as archetypes of the kinds of problems that analysts actually encounter. Given a simple setting and a clear question, it might be possible for analysts to agree about the appropriateness of an answer, irrespective of their beliefs about the interpretation of probability or their convictions about its universality or completeness. Moreover, we were concerned that complicated or computationally demanding problems would discourage individuals from attempting their solution and thus from participating in the exercise. If people with radically different points of view could agree on the answer to a given problem, then the debates that divide them can be considered purely philosophical. On the other hand, if the different perspectives lead to radically different *answers* for a risk assessment that represent a significant disparity in predictions and perhaps lead to different final decisions, then analysts have to face the issues under debate directly and decide which approach should be used. While designing the Challenge Problems, we genuinely did not know which outcome the workshop would witness.

A few of the participants at the workshop complained that the Challenge Problems were unrealistic to the point of being nothing more than toy problems without even pedagogical value. They argued that the sterility of the context in which the problems were posed both robbed them of realism and made them impossible to solve because of the severing of the essential connection between the analyst

and the engineer (or subject-matter specialist) by which information is marshaled and accessed. This point of view maintains that an analyst can, should and actually *must* have a constant dialog with the engineers or subject-matter experts on whose knowledge the assessment is based.

In practice, however, there are surely practical situations in which communication with the empiricist is inconvenient or impossible (such as when information must be collected from papers and reports or when the informant is unavailable). Our defense of the Challenge Problems is that they are intended to serve a pedagogical value in isolating crucial issues involved in the representation and propagation of epistemic uncertainty. We did not see any practical way to introduce the iterative refinements typical of elicitations into the problems. Despite their many shortcomings as representative models of how a real analysis is carried out, the problems do seem appropriate for their limited purpose. If an uncertainty propagation method *cannot* produce answers to these simple problems, then it seems likely that the method is not general enough for the problems that real analyses pose. Moreover, if different analysts do not arrive at similar or at least consistent answers to these simple problems, then the uncertainty analysis community is more fractionated than is generally recognized.

2.2. The engineer's beliefs or believing the engineer

Many probabilists consider it the job of the analyst to elicit information from the engineer, and think this process is not done until a complete probabilistic model has been specified. Formal elicitation [11,12] has proved extremely valuable in gleaning essential information that would otherwise be lost in the depths of an engineer's memory, or the bowels of a laboratory's filing cabinets.

However, it is not uncommon for an engineer, at some point, to disclaim further information about some component or process, especially one that has not been the subject of careful or extensive (or sometimes *any*) study. The engineer may answer an elicitation with 'I don't know', which can sometimes lead to the peculiar situation in which the analyst insists the engineer knows more than he or she professes to know.

We believe that, although perhaps often a useful elicitation device, this insistence may not always be reasonable. As R.A. Fisher used to quip, it is evidently easier for the practitioner of natural science to recognize the difference between knowing and not knowing than this seems to be for the more abstract mathematician. This notion of probabilistic omniscience is *not* a necessary feature of the probabilistic approach, and indeed, there are many fields (notably, robust statistics [13]) that reject it outright.

When deprived of an engineer's knowledge, the elicitation process usually devolves to finding what the engineer *believes* about the component or process, as opposed to what he may *know* from specific empirical study. Although there may be useful knowledge embedded in an engineer's subjective beliefs, there may not be an unlimited amount of such knowledge. Is it ever reasonable to believe an engineer when he or she claims not to know? It does not seem heretical to ask about the limits of what can be asserted in an assessment *based on empirical knowledge*.

Many probabilists argue that such a quest is naïve in itself, because the model construction process is inherently subjective. Yet an analysis that eschews subjective belief and traffics only in information objectively established by measurement, census, experiment and theoretical constraints should also, it seems, be of *some* interest in a quantitative assessment. Such analyses are likely to involve interval characterizations of uncertainty with little, if any, structure within the intervals. The non-traditional methods that have been brought to bear on the Challenge Problems may develop into tools appropriate for such analyses. This exercise represented by the Challenge Problems, far from being merely academic, reflects a sincere desire to understand how to best respect and numerically capture the varied statements and protestations of the engineers and empiricists on which analysts depend, and how to propagate their knowledge through mathematical models.

2.3. Aggregation

This section reviews the background of the first of five technical issues addressed by the Challenge Problems that are commonly involved in computational problems involving epistemic uncertainty and that arouse on-going debate and some considerable confusion among analysts.

Analysts commonly encounter the situation where multiple estimates of the value of some quantity or variable are available. In a sense, this is a problem of too much information because it means the analyst must decide how to combine this information before proceeding with the analysis. In some cases, these estimates may be independent of one another, and they may be dependent in other cases. Sometimes, in fact, the epistemic relationship among the estimates may be unclear. The separate estimates can arise from completely different sources and thus have different mathematical representations (e.g. a collection of sample data, a single theoretical interval, semi-quantitative expert opinion or anecdote, or a probability distribution from another model). In such situations, not only is there competing information, but it may be expressed in conflicting formats. The question for the analyst is how disparate information embodied in different representations can be aggregated into a single representation for use in subsequent calculations. In their design, the Challenge Problems emphasize this issue.

Before and at the workshop, the aggregation question provoked considerable discussion. For instance, Vicki Bier of the University of Wisconsin at Madison argued that simple intervals whether derived empirically or from expert

opinion might not necessarily circumscribe all possible values a variable could take. Roger Cooke of the Technische Universiteit Delft complained that divorcing the Challenge Problems from a motivating context prevents an analyst from inferring which approach should be used for aggregation. He also suggested that it is not clear that the aggregation should precede calculations, because one can sometimes obtain different answers depending on whether one aggregates multiple estimates then computes, or computes with the several estimates separately and then aggregates the results. Further research is clearly necessary to understand the proper methods and uses of aggregation in uncertainty analyses. It seems likely that this is a problem without a universally applicable solution.

2.4. Repeated parameters

Many uncertainty calculi are known to be sensitive to uncertain parameters or variables that appear multiple times in one expression. For instance, in expressions involving intervals that appear multiple times, naively applying interval analysis [14,15] can yield an unnecessarily wide answer. In the first Challenge Problem, for example, when the inputs $a = [0.1, 1]$ and $b = [0, 1]$ are propagated through the function $(a + b)^a$, the repetition of the parameter a causes the answer to be [0.1, 2], even though it can be demonstrated through other means that the range of real values that are possible in this case is only [0.692, 2]. In contrast, naively applying the method of discrete probability distributions [16–18] leads to results with smaller variances and lighter tails than are really justified. In both situations, repeated parameters are a problem because they can introduce their uncertainty more than once into a calculation.

The underlying problem is that the perfect dependency [19] between the different occurrences of the quantity is not appropriately represented in the calculation. This same issue arises not only in interval analysis and the propagation of discrete probability distributions, but also in fuzzy arithmetic [20], probabilistic logic [21], and most uncertainty propagation methods. In the case of interval analysis and fuzzy arithmetic, the existence of repeated parameters can cause the calculations to overstate the uncertainty of the results, making them suboptimal and misleading. In the case of probabilistic methods that are sensitive to repeated parameters, the uncertainty can be underestimated, which could be an even worse error in risk analysis.

In sharp contrast, Monte Carlo simulation correctly accounts for repeated parameters because of the way it instantiates random values. Indeed, early researchers recognized this as perhaps the primary advantage of Monte Carlo methods for uncertainty propagation over other probabilistic methods such as discrete approximations [16], transformations [22], or other analytical approaches [17]. The problem can reappear, however, in stepwise Monte Carlo simulations in which simulations are chained together so that the output of one simulation is used as input to another simulation, or components of the simulations are performed separately by different teams [23,24].

We believe that any method of propagating uncertainty through calculations is not mature enough for serious use unless it has practical strategies to overcome the difficulties of repeated parameters. The algebraic expression $(a + b)^a$ and the mass-spring-damper problem [2] were selected because they have repeated parameters that cannot be removed by algebraic rearrangement. During the workshop, most of the presenters used methods that appropriately account for the repeated parameter. One presenter initially chose a different interpretation of the repetition of a, but later came to agreement with others on the subject.

2.5. Intervals as observations

The Challenge Problems were designed to emphasize the commonness of epistemic uncertainty, especially as represented by intervals. We had not at first considered this a terribly controversial thing to do, but before and during the workshop, it became clear that some probabilists consider the idea that uncertainty could legitimately be manifest as intervals both radical and untenable. Tony O'Hagan of University of Sheffield wrote before the workshop in the 'Submitted Comments' section of the website http://www.sandia.gov/epistemic/#, "I do not believe there is any real situation in which one could only know that a parameter lies in an interval," and concluded that problems that feature such uncertainty were 'non-problems'.

The discussion on this topic has been extensive both before and during the workshop, and we only want to give a flavor of it here. It seems to us that real-world situations can produce uncertainties that are appropriate to characterize as intervals. Because of the controversy, we briefly mention four kinds of such situations.

(1) Sometimes a quantity may not have been studied at all, and the only real information about it comes from theoretical constraints. Physical limits may be used to circumscribe possible ranges of quantities even when no empirical information about them is available. Typically, these intervals are rather wide, and they are often considered vacuous statements because they represent no empirical information. Nevertheless, expressions that involve them may not be vacuous. The uncertainty they induce in computations depends on the uncertainties of other quantities and how all the quantities are combined together in a mathematical expression.

(2) Some monitoring plans call for periodic inspection of components. In such cases, a component may be in good working order at one inspection, but not at the next. When did the component fail? It seems entirely reasonable to hold that there was a window of time between the last two inspections during which the component failed and that the natural mathematical representation of the failure point is an interval. Such inferences based on temporal observations

occur in many domains and circumstances. For instance, forensic analysis of crime sequencing is often based on synthesis of the testimonies of multiple witnesses that yield an interval 'window of opportunity'.

(3) Physicists, chemists and engineers are taught to report the plus-or-minus uncertainties associated with the calibration of measuring devices. Such uncertainties represent another important example of epistemic uncertainty in the form of intervals. Certainly analog instruments require reporting the position of a gauge or needle between two landmark values. A measurement from a device with a digital readout is associated with an interval based on the number of reported digits. For example, the value 12.64 is associated by rounding to the interval [12.635, 12.645]. There is certainly no statistical justification for assuming that the correct value is any closer to the middle of this range than it is to either endpoint (although one might argue that such limits are not absolute in the sense that the true measurement has a zero chance of lying outside of them). Note that this uncertainty from the intrinsic error of measurements is independent of sampling error that results in a statistical distribution of point measurements. In statistical practice, the uncertainty represented by these measurement intervals is frequently ignored. In some situations, it is probably reasonable to neglect such uncertainty, but this is clearly not always the case. For example, the uncertainty should not be ignored when measuring devices have an inherently wide imprecision, when observations are determined indirectly through other measured quantities, or when pooling multiple data streams from devices that are not calibrated in the same way. In such situations, an interval seems an entirely reasonable way to account for uncertainty of this kind [25].

(4) In chemical or purity quantifications, there are sometimes 'non-detects' in which the laboratory procedure can only say that the concentration of a substance is below a certain amount known as the detection limit. Again, the uncertainty here is in the form of an interval, and there is no reason whatever to think that the true concentration is more likely to be in any particular part of the interval. For situations in which such uncertainty cannot be neglected in an assessment, it would seem to be essential to have a way to handle these interval uncertainties in calculations.

O'Hagan also suggested that an engineer would be unlikely to know the shape of a distribution function and yet be unable to specify its parameters except by intervals. It seems to us, however, that this situation arises commonly. In a nuclear power plant risk analysis, for instance, it is common to assume that the failure-free operating time for a certain kind of component is characterized by a Weibull distribution, and yet analysts may find it necessary to determine the distribution's parameters by fitting. This is a case in which aleatory and epistemic uncertainty are intimately intertwined.

2.6. Independence

In probability theory, there are several ways to define the concept of independence between random variables. For random variables X and Y characterized by the joint distribution H with marginals F and G such that $P(X \leq x) = F(x)$, $P(Y \leq y) = G(y)$ and $P(X \leq x, Y \leq y) = H(x, y)$, independence between X and Y implies, and is implied by, each of the following conditions:

(i) $H(x, y) = F(x)G(y)$, for all values x and y,
(ii) $P(X \in I, Y \in J) = P(X \in I)P(Y \in J)$, for any subsets I, J of the real line,
(iii) $h(x, y) = f(x)g(y)$, for all values x and y,
(iv) $E(w(X)z(Y)) = E(w(X))E(z(Y))$, for arbitrary functions w and z, and
(v) $\phi_{X,Y}(t, s) = \phi_X(t)\phi_Y(s)$, for arbitrary t and s,

where P is the probability, f, g and h the density analogs of F, G and H, respectively, E the expectation, and ϕ the Fourier transform (characteristic function) in terms of arguments t and s. When probabilities are precise these various definitions of independence between random variables are all *equivalent*. Each definition implies all the others. Therefore, there is a single concept of independence that simultaneously embodies all of these possible definitions in a traditional probabilistic representation of uncertainty.

There is a decidedly different story in the context of imprecise probabilities and related formalisms in generalized information theory such as Dempster–Shafer evidence theory, random set theory, probability bounds analysis, and possibility theory. Here, the special case of independence, which is unique in probability theory, disintegrates into several different cases when probabilities are imprecise. Couso et al. [26] pointed out that, for imprecise probabilities, the various possible definitions of independence are no longer equivalent to each other. In fact, the different definitions induce distinct concepts of independence for imprecise probabilities. Several concepts [26] might be called independence, of which four seem especially germane for risk analyses:

- *Repetition independence* is when there is stochastic independence (in the traditional sense) between random variables that are identically distributed, although their distribution may be imprecisely known. Repetition independence is thus the analog in the context of imprecise probabilities of the constraint in probability theory that variables are independent and identically distributed (iid).
- *Strong independence*, on the other hand, is the complete absence of any relationship between random variables. Variables X and Y are strongly independent if the set of possible joint distributions is the largest set such that each joint distribution $H(x, y) = F(x)G(y)$, where F is

one of the possible distribution functions characterizing X and G is one of the possible distribution functions characterizing Y. Variables X and Y would be characterized as strongly independent if (i) X and Y result from random experiments, each governed by a unique albeit possibly unknown probability distribution, (ii) the random experiments are stochastically independent (in the traditional sense), and (iii) there is no known relationship between the variables that would preclude some possible combinations of the possible marginal distributions.

- *Epistemic independence* arises when an analyst's uncertainty about either of two outcomes of a random experiment does not change when some information about the outcome of one of them becomes known. Random variables X and Y are epistemically independent if the conditional probability of each given the other is equal to its unconditional probability, so that $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. In the context of imprecise probabilities, epistemic independence is defined in terms of lower bounds on expectations such that $\underline{E}(f(X)|Y) = \underline{E}(f(X))$ and $\underline{E}(f(Y)|X) = \underline{E}(f(Y))$ for all functions f where $\underline{E}(Z)$ denotes the infimum of all expectations of Z over all possible probability distributions that could characterize Z.
- *Random-set independence* is the kind of independence embodied in Cartesian products [27] between Dempster–Shafer structures. For instance, Dempster–Shafer structures X and Y with mass functions m_X and m_Y respectively are random-set independent if the Dempster–Shafer structure for their joint distribution has mass function $m(A_1 \times A_2) = m_X(A_1)m_Y(A_2)$ when A_1 is a focal element of X and A_2 is a focal element of Y, with $m(A) = 0$ for all subsets not of the form $A = A_1 \times A_2$.

Couso et al. [26] reviewed these definitions and gave simple examples of each of these definitions (and more). Couso et al. [28] gave examples of how the definitions could influence numerical calculations. Fetz [29] illustrated the consequences of the various independence definitions in a probabilistic assessment for an engineering system. Cozman and Walley [30] explored the properties of epistemic irrelevance and epistemic independence.

The Challenge Problems [2] assumed all quantities to be 'independent' of one another, which was explained to mean that 'knowledge about the value of one parameter implies nothing about the value of the other'. This corresponds to epistemic independence, although the correspondence was not explicitly stated in the Challenge Problems.

3. Summary of the workshop proceedings

There was a considerable diversity of opinions expressed at the workshop. Opinions diverged on everything from technical details about the best computational techniques to issues about interpretation and strategies. Despite this wealth of disagreement, it is fair to say there was also considerable consensus about what the answers to the problems were, and even about how some of the five focal questions listed in Section 2 should be addressed. This consensus can be seen by comparing the papers written by workshop participants assembled in this special issue. For instance, the similarities are striking among the graphical depictions of the answers to the Challenge Problems given by the various authors. Moreover, the sense at the workshop was that these answers are 'right' in the sense that they present what can be justifiably concluded while respecting what is unknown. Although some participants steadfastly maintained that a purely probabilistic approach was fully capable of accounting for all forms of uncertainty, most participants agreed that the formal quantification of epistemic uncertainty introduces new considerations and that the implicit questions underlying the problem set are legitimate and important.

Of course, it is possible that the self-selection of attendees guaranteed such a consensus at the workshop, and it would be foolish to argue that the idea is dominant in risk analysis. We suspect, however, that the appropriate treatment of epistemic uncertainty is an emerging and important challenge that will receive increased attention in the coming years.

In the following several sections, we summarize the approaches used by the various authors of papers in this special issue and describe the results of preliminary but quantitative comparisons of their answers to the Challenge Problems.

3.1. Contributed papers

Each of the 22 contributed papers in this special issue described a specific approach to the problems of representation and calculation in the presence of aleatory and epistemic uncertainty.

Klir [31] discussed the framework he calls generalized information theory, surveyed the various theories about imprecise probabilities and how they fit within this framework, and considered the unifying principles that apply to all of these theories. Helton et al. [32] described how that the Challenge Problems could be solved by various approaches, but emphasized that these approaches yield results that must be considered within their own respective contexts. Fetz and Oberguggenberger [33] showed that the answers to the Challenge Problems can differ substantially under different kinds of independence among the variables. Hall and Lawry [34] proposed an iterative rescaling method to construct a random set representation of uncertainty that approximates the upper and lower cumulative probability distributions.

Kozine and Utkin [35] argued that a theory of imprecise coherent probabilities is appropriate and necessary to solve the Challenge Problems, but suggested that this approach would benefit from the addition of numbers to characterize

the credibility of disparate, potentially conflicting statistical evidence. De Cooman and Troffaes [36] reviewed Walley's theory [37] of coherent lower previsions, including different concepts of independence among variables, and showed why it is a good model for solving uncertainty problems that possibly involve conflicting information and suggested that result of aggregation should be independent of the order of combination.

Ferson and Hajagos [38] claimed that a single technology based on the bounds around cumulative distribution functions is sufficient to solve all the problems rigorously and sometimes optimally. Berleant and Zhang [39] used their distribution envelope determination method, implemented with mathematical programming techniques, to compute bounding envelopes around distribution functions arising from uncertainty about input distributions or uncertainty about their interdependence. Tonon [40] used random sets theory to address the spring-mass-damper problem and observed in this application that it is preferable computationally to increase the fineness of the discretization rather than invoking a global optimization tool to ensure the constituent intervals are perfectly computed.

Red-Horse and Benjamin [41] used a method called polynomial chaos expansion (PCE), a function-analytic approximation technique (unrelated to chaos in the dynamical systems sense), to provide p-box-like bounds on the mean, standard deviation and 95th percentile for the output of the mass-spring-damper problem. Li and Hyman [42] developed their method based on a probability distribution variable (PDV) as a new data type, compared it to a random interval approach, and provided answers to the problems as p-boxes and PDV results. Rutherford [43] used a response surface method coupled with a coarse-grained Dempster–Shafer approach to answer many of the Challenge Problems. Ayyub [44] reviewed the taxonomies for knowledge and uncertainty, the limits of knowledge construction, and the differences between recognized and unrecognized ignorance.

O'Hagan and Oakley [45] argued that intervals are unrealistic models of uncertainty and that probability distributions can model any realistic kind of uncertainty. Ben-Haim [46] reviewed his info gap theory and argued that the Challenge Problems would only have meaning, and could only be solved within, the context of decision problems. Kreinovich and Ferson [47] pointed out that a sampling strategy based on Cauchy deviates can propagate interval uncertainty though black box models, and that the Cauchy deviate approach becomes efficient when the number of variables is large.

Agarwal et al. [48] gave an overview of the appearance and treatment of uncertainty in multidisciplinary systems, considered the problem of optimum engineering design under uncertainty and, using evidence theory, solved problems in multidisciplinary system analysis that are extensions of the Challenge Problems. Soundappan et al. [49] compared Bayesian methods with evidence theory using Dempster's rule of combination and measures of information such as non-specificity and strife for solving algebraic problems similar to the Challenge Problems.

Cooke [50] derogated the Challenge Problems as being hopelessly unrealistic, emphasized the importance of operational definitions for alternative mathematical representations of uncertainty, and said that subjectively estimated probabilities from experts would be sufficient for any real problem. Bier [51] pointed out that the endpoints of intervals may not be hard and fast, and that a subtler model of the range may be necessary. Booker and McNamara [52] argued that the elicitations from experts can and should be interactive, with experts seeing the consequences of their estimates on the final answer. Yager [53] showed how formal decision analysis can be conducted within Dempster–Shafer theory.

3.2. Who offered answers to the Challenge Problems

Only Helton et al. [32], Ferson and Hajagos [38], and Berleant and Zhang [39] provided answers to all of the Challenge Problems. Fetz and Oberguggenberger [33], Kozine and Utkin [35], De Cooman and Troffaes [36], and Li and Hyman [42] provided answers to all six subproblems of Challenge Problem A. Other authors, including Klir [31], Hall and Lawry [34], Rutherford [43], and Ayyub [44], provided answers to some of the arithmetic subproblems of Challenge Problem A. Tonon [40] and Red-Horse and Benjamin [41] provided answers to Challenge Problem B, the mass-spring-damper problem.

Several authors elected not to provide any answers to the Challenge Problems. Some authors, such as Kreinovich and Ferson [47], Agarwal et al. [48], Soundappan et al. [49], Cooke [50], Bier [51], and Yager [53], used the forum to focus on related issues and did not address the problems themselves. Ben-Haim [46] declined on principle to provide answers to the Challenge Problems. He argued that modeling of uncertainty cannot usefully be separated from the decisions for which the uncertainty model was constructed. (He also used different numerical values describing the inputs.) Cooke [50] characterized the Challenge Problems as "bizarre" and chose not to address them. O'Hagan and Oakley [45] argued that the Challenge Problems were profoundly unrealistic, but nevertheless did address some of the problems and ventured to answer one of them. However, they added additional information to the problem statement, which impedes the comparison of their result with those of other authors. Booker and McNamara [52] likewise considered the addition of further information, but, because this information was added in (hypothetical) subsequent rounds of expert elicitation, their results are partially comparable to those of other authors.

3.3. Distributional answers to the Challenge Problems

There were two kinds of numerical results given as answers for the Challenge Problems. Some authors gave

answers characterizing the *distribution* of the quantity (whether the quantity $y = (a + b)^a$ for Challenge Problem A, or the steady-state magnification factor D_s for Challenge Problem B). Other authors gave answers that did not characterize the entire distribution, but rather only the *expected value* (or mean, average value, or expectation) of the distribution. These two kinds of answers are obviously quite different. The following section discusses the distributional answers. Section 3.4 below discusses the answers given in terms of the expected values.

3.3.1. Challenge Problem A: $(ab)^a$

Challenge Problem A consisted of six subproblems 1–6, some of which themselves had subproblems. They all concerned the evaluation of the expression $(a + b)^a$, where different kinds and degrees of information about the quantities a and b were available.

In Challenge Problem 1, the available information about a and b were single intervals for each. Almost all of the papers that offered an answer to the Challenge Problem 1 agreed that it was best represented by the interval [0.692, 2]. Indeed, several disparate approaches all gave the same answer in this case. This consensus arose apparently in the agreement among the authors that, when the inputs are simple intervals, the solutions to the problem should reduce to a familiar interval analysis [14]. In some cases, the answer was expressed explicitly as a simple interval of real numbers. In other cases, it was expressed as a degenerate version of a more complicated structure such as a probability box or a Dempster–Shafer structure that is mathematically equivalent to the interval. Helton et al. [32] gave multiple answers to Challenge Problem 1. Not all were mathematically equivalent to intervals, but all at least had the same support (range) as the interval [0.692, 2].

The paper by Ayyub [44] and the paper by Booker and MacNamara [52] disagreed with this consensus. They both gave the answer to Challenge Problem 1 as [0.79, 2] (Booker and MacNamara's [52] 'first iteration'). This appears to be due to their use of endpoint sampling to solve the interval analysis problem. That is, they found the lower bound as the minimum of the four values $\{(a_1 + b_1)^{a_1}, (a_1 + b_2)^{a_1}, (a_2 + b_1)^{a_2}, (a_2 + b_2)^{a_2}\}$. They found the upper bound as the maximum of the same four values. This approach can be called the vertex method because it considers only the corners of the input space. The upper bound found this way conforms with the standard result from interval analysis in this case, but the lower bound, of course, does not. Because the vertex method does not yield the full range when the function is non-monotonic, it can underestimate the uncertainty in the result.

There was likewise a strong consensus about the answers to the Challenge Problems 2a, 2b and 2c. Most of the authors who characterized the distributions of the answers, including Klir [31], Helton et al. [32], Fetz and Oberguggenberger [33], Ferson and Hajagos [38], Berleant and Zhang [39], and Rutherford [43] (2b only) quantitatively agreed on the answers to these problems, although they displayed them in various ways. Klir [31] gave them as a numerical list of intervals. Most of the authors gave the answers only graphically, although Berleant and Zhang [39] gave some of the answers both ways. Some authors displayed them in cumulative form; some displayed the results in exceedance (complementary cumulative) form. Helton et al. [32] gave complementary cumulative plausibility and belief functions (and probability distributions). Fetz and Oberguggenberger [33] displayed both exceedance p-boxes and the constituting random intervals themselves. Such differences are not important quantitatively because each format can, in principle, be converted to any other.

Where numerical agreement could be checked, the results from all the papers agreed in the first two or three decimals places, with a few exceptions that appear to be either minor typographical errors or small differences in approximations. The answers agree not only in terms of the support (range) over which the quantity can vary, but also in terms of the internal uncertainty structure as to the location and magnitude of jumps in probability values. The consensus arises from the fact that all the authors used a stochastic mixture [54] operation to aggregate the estimate for the b parameter. A stochastic mixture is formed by averaging probability values for each of the possible values of the quantity.

Ayyub's answers [44] to Challenge Problems 2a, 2b and 2c, also given as a list of intervals, are close to the same numerical values as given by the other authors, and, when displayed graphically, are visually indistinguishable from the answers given by the consensus. However, they are numerically different to a degree that does not appear to be from differences in the approximation method; the numbers often differ in the second decimal place. It seems clear that these discrepancies are due to Ayyub's use of the vertex method to solve the underlying interval calculations.

Helton et al. [32], Fetz and Oberguggenberger [33], Ferson and Hajagos [38], Berleant and Zhang [39], and Li and Hyman [42] provided distributional answers to Challenge Problems 3a, 3b and 3c. Again, there is strong consensus among the answers obtained, both in terms of the total range and the internal structure of uncertainty. Although numerical comparisons are not convenient, the respective graphs are mostly indistinguishable from one another among the various papers. One discrepancy is that the answers given by Helton et al. [32] do not quite reach the value of $y = 2$, although they should. The reason for this slight underestimation of uncertainty is traceable to their use of a sampling strategy to make interval calculations, which will almost always miss the minimum and maximum possible values. Such a difference could be significant in a risk analysis where probabilities of extreme behaviors are of interest, such as, for instance, if the value 2 represented a flashpoint or other critical threshold. In practice, a sampling strategy would typically include some sort of importance

sampling procedure to assure that outcomes having low probability but high consequences are accounted for.

There was marked diversity in the displays given for problems 3a, 3b and 3c, and several authors displayed multiple answers for each of the problems under different assumptions. For instance, Helton et al. [32] also displayed probability distributions. Fetz and Oberguggenberger [33] displayed answers obtained under different assumptions about what independence means. It is difficult to compare their results to those of other authors because of the complexity of their graphs. Because Fetz and Oberguggenberger [33] assumed *strong independence* between a and b, their results could, in principle, be somewhat narrower than those of the other authors who assumed only *random-set independence*. Li and Hyman [42] also displayed results of a 'refined p-box' approach that produced results that closely resemble probability distributions.

Ayyub [44] also offered answers to Challenge Problems 3a and 3b. Although he used a tableau methodology that superficially seems similar to that used by Berleant and Zhang [39] and by Ferson and Hajagos [38], Ayyub's [44] approach was substantially different from the approaches used by the other authors. For instance, his use of the vertex method produced sometimes sharply narrower interval ranges. Even more significantly in this case, Ayyub [44] did not estimate the mass associated with these intervals by multiplying the respective masses associated with the two intervals ranges for a and b. Instead, he employed a different operation motivated by a fuzzy sets approach to belief and plausibility. Although a graphical comparison of Ayyub's results with those of the other authors would have been interesting, his results were only given numerically. Klir [31] also offered an answer for 3a, but he used inputs slightly different from those prescribed by Oberkampf et al. [2], so the results are not comparable. Booker and MacNamara [52] offered an answer for problem 3c, but it is difficult to compare it to answers given by other authors for similar reasons.

Most authors are also apparently in quantitative agreement about the answers to the remaining arithmetic Challenge Problems (4, 5a, 5b, 5c, and 6). For instance, Fetz and Oberguggenberger [33], Ferson and Hajagos [38], Berleant and Zhang [39], Li and Hyman [42], and Rutherford [43] give essentially the same answer for problem 4, whether they call the result a p-box or cumulative belief and plausibility functions. Some of the displays are noticeably smoother or coarser step functions than the displays of the other authors. For instance, compare the answer to problem 4 given by Rutherford [43] (his Fig. 17) with that given by Li and Hyman [42] (their Fig. 5.12). Evidently this difference is a result of the number of discretization steps used to approximately represent the specified inputs. Similar agreements, notwithstanding minor discrepancies of smoothness, emerge for the results for the other Challenge Problems among these authors. Rutherford [43] did not provide answers to problems 5a, 5b, or 5c.

Helton et al. [32] did not present answers to problems 4, 5a, 5b, 5c, or 6 in terms of bounds on the distribution functions, but they did display 'multiple aleatoric distributions' as collections of possible distribution functions for the uncertain variable y. The collections were consistent with the answers given by the other authors in the sense that all the distributions were entirely inside the respective bounds.

The answers of Hall and Lawry [34] for problems 4, 5a, 5b, and 5c have very noticeable differences from those of the other authors. Although they used the same aggregation operation, they apparently used a different discretization scheme. Their results are quantitatively different in terms of the internal uncertainty structures, and, in particular, are considerably fatter in the right tail of the distributions, although the locations of the answers on the y-axis conform with those observed by the other authors.

3.3.2. Challenge problem B: mass-spring-damper

Helton et al. [32], Ferson and Hajagos [38], Berleant and Zhang [39], Tonon [40], and Red-Horse and Benjamin [41] offered distributional answers to Challenge Problem B, the mass-spring-damper problem. The graphed answers are visually very similar and suggest a substantial quantitative consensus. There are, however, noticeable differences in both the left and right bounds. The answers given by Berleant and Zhang [39] and Tonon [40] suggest that no values can be lower than one and that there must be some values greater than two (i.e. the probability of $D_s < 2$ must be smaller than one). The answer given by Ferson and Hajagos [38] is slightly wider for the left bound. Their answer suggests that some values can be below one, and that it is possible that all values are below two (i.e. the probability of $D_s > 2$ could be as small as zero).

The answer of Red-Horse and Benjamin [41] is more different still from those of the other authors. They gave two answers as bounds on distribution functions (power and linear forms), and both are substantially wider in terms of uncertainty than the bounds described by other authors, although their supports appear to be the same. The reason for the discrepancy may be that the other authors aggregated the multiple estimates for the individual input variables *before* computing D_s, but Red-Horse and Benjamin computed D_s for each possible combination of inputs and aggregated the resulting estimates of D_s.

Helton et al. [32] did not present an answer to Challenge Problem B in terms of bounds on the distribution function for the variable D_s (although they did for the expected value of that variable). Instead, they gave multiple aleatoric distributions for the variable, which are entirely consistent with the bounds given by the other authors because they all lie inside the bounds. This collection of distributions does seem to suggest, however, that the probability of values larger than four is very small, whereas the answers given by other authors suggest it might be as large as 0.1 or so. However, this collection only included 50 distributions.

With additional iterations, their approach gives a collection of distributions that tends to cover the envelopes given by the other authors. For instance, when the number of distributions simulated approaches 1000, the results suggest that values less than one or greater than seven are possible, and that the probability of values being greater than two can be as small as zero.

Among the authors who provided quantitative answers to the Challenge Problems in terms of belief and plausibility functions or bounds on distribution functions, there is very strong consensus. There are small differences due to the inefficiency of sampling, or the number of discretization levels employed. There are, however, answers given by some authors that deviated more substantially from the consensus. As we will see in Section 3.4, there is not as much quantitative agreement when the answers are not expressed as bounds on distribution functions.

3.4. Bounds on expected values

This section summarizes the results that were expressed in terms of the *expected value* of the answer (rather than its entire distribution). Five papers (Helton et al. [32]; Kozine and Utkin [35]; De Cooman and Troffaes [36]; Ferson and Hajagos [38]; and Red-Horse and Benjamin [41]) expressed their answers to some or all of the Challenge Problems in terms of bounds on the expected value of the output variable as shown in Table 1. The approach of De Cooman and Troffaes [36] and that of Kozine and Utkin [35] were both based on imprecise probabilities and particularly Walley's theory of coherent lower previsions. The Dempster–Shafer approach of Helton et al. [32] was a similar but weaker (i.e. more conservative) version of this approach in the sense that it can yield results that are wider (but no tighter) than can be justified by other methods. The probability bounds approach of Ferson and Hajagos [38] was also a similar but still weaker analysis.

Despite the similarities of their approaches, their respective answers to the Challenge Problems in these papers differ to a surprising degree. Table 1 assembles side by side the numerical answers about the expected value of the output variable from the five papers for the various problems. The entries denoted as '(graphical)' represent answers that are not given as interval ranges but as rather more complicated structures (these entries are discussed below).

There is some quantitative agreement by the various authors on a few of the expected values. Within expressed significant digits, all three papers that gave an answer for Challenge Problem 1 agreed that is the interval [0.69, 2]. Note here that the bounds on the expected value $E(y)$ are the same as the bounds on the variable y itself obtained in other papers.

After problem 1, however, the papers display little agreement. In some cases, the numerical differences are rather large. This disagreement is apparently partially due to differences in how multiple estimates were aggregated before they were combined arithmetically. De Cooman and Troffaes [36] took the interval data in the Challenge Problems as rigorous and used intersections of closed convex sets of probability distributions to aggregate multiple estimates (or the convex hull of their union when those intersections were empty). Ferson and Hajagos [38] used stochastic mixture (i.e. linear opinion pooling) to aggregate the multiple estimates. Kozine and Utkin [35] developed a novel approach for aggregation that generalizes various possible methods and requires the analyst to specify numerical credibilities that reflect differential confidence of the analyst in the various estimates. To illustrate the potential flexibility of their approach, they introduced credibilities in their solutions to problems 2, 3, and 5 that did not appear in the statement of the Challenge Problems. This certainly accounts for part of the observed quantitative differences in their answers compared to those of the other authors.

Surprisingly, there was also disagreement about the answers to Challenge Problems 4 and 6. Because these problems did not have multiple estimates for any of

Table 1
Comparison of bounds on expected values

	Helton et al. [32]	Kozine and Utkin [35]	De Cooman and Troffaes [36]	Ferson and Hajagos [38]	Red-Horse and Benjamin [41]
1	–	[0.69, 2.0]	[0.692201, 2.0]	[0.692, 2]	–
2a	–	[0.93, 1.84]	[0.956196, 1.8]	[0.84, 1.89]	–
2b	–	[0.93, 1.76]	[0.956196, 1.7]	[0.82, 1.85]	–
2c	–	[0.93, 1.52]	[0.692201, 2.0]	[0.83, 1.73]	–
3a	–	[0.944, 1.473]	[1.04881, 1.2016]	[0.83, 1.56]	–
3b	–	[0.964, 1.418]	[1.04881, 1.1156]	[0.82, 1.44]	–
3c	–	[1.187, 1.242]	[0.692201, 2.0]	[0.946, 1.25]	–
4	[1, 3.7]	[0.859, 1.108]	[1.00966, 4.08022]	[0.9944, 4.416]	–
5a	(Graphical)	[1.45, 2.824]	[1.54027, 2.19107]	[1.05, 3.79]	–
5b	(Graphical)	[1.373, 2.607]	[1.54027, 1.81496]	[1.03, 3.48]	–
5c	(Graphical)	[1.802, 2.298]	[1.00966, 4.08022]	[1.12, 2.94]	–
6	[1.05, 3]	[1.019, 2.776]	[1.05939, 2.86825]	[1.052, 2.89]	–
B	(Graphical)	–	–	[1.17, 3.72]	(Graphical)

the inputs, there could not be any differences arising from the choices about aggregation methods. There are at least four possible sources for the observed discrepancies among the answers:

1. *Nesting.* Given the relative natures of the various approaches employed by the different authors, one might anticipate that their answers would be nested in a particular way. For instance, the intervals given by Helton et al. [32] should enclose those given by either Kozine and Utkin [35] or De Cooman and Troffaes [36]. Likewise, the results given by Ferson and Hajagos [38] should enclose those given by Helton et al. [32]. In fact, very little such nesting of results can be observed in the table above.
2. *Differences in truncation.* It seems possible and perhaps likely that the discrepancies among these answers could in part be the result of different choices about whether or where the distributions were truncated to finite ranges.
3. *Numerical approximation error.* Of course, the observed discrepancies may be attributable to numerical approximation error in the calculation algorithm and the accumulated machine round-off error in its computer implementation of each of the various approaches (e.g. in the linear programming algorithms). If this were the case, then the discrepancies would in principle evaporate under additional computational effort. However, assuming that the authors reported significant digits so that they reflect the reliability of the respective answers, approximation error does not seem to be tenable as the full explanation of the observed discrepancies.
4. *Different representations of independence.* Finally, the disagreement about the answers may also be due to different assumptions about the independence between the variables. If this were the case, one would expect that the answers would be nested in some way, which is not apparent in the answers given in the table.

The answers given by Helton et al. [32] (in their Figs. 16 and 17) for the three subproblems of Challenge Problem 5 were probability distributions, Dempster–Shafer structures and possibility distributions. They all have supports of [1, 3.6], so they are quantitatively comparable to the other answers listed in the table, but the structures are non-degenerate and therefore structurally different from the intervals given by all the other authors. Whereas most authors gave a set of possible values for the expected value, Helton et al. [32] gave distributions of one kind or another for the expected value. The recognition of this *qualitative* difference between answers is an especially interesting outcome of the Albuquerque workshop. It points to an issue even more fundamental than numerical consensus and discrepancies. Its resolution will involve developing an agreement about what the answer should look like.

Only three papers gave characterizations of the expected value of D_s. Helton et al. [32] (in their Fig. 19) described a Dempster–Shafer structure over the range [1.44, 2.86], which was computed with a sampling strategy. Red-Horse and Benjamin [41] (in their Fig. 7) gave a structure that was similarly structured and computed but had the support [1.4, 3.6]. Ferson and Hajagos [38] gave the simple interval [1.17, 3.72], which was based on moment propagation. It is almost sure that their moment propagation would overestimate the uncertainty of the answer, whereas, depending partially on the number of replications, a sampling strategy could tend to underestimate the uncertainty.

Overall, there was considerably less quantitative agreement among authors about their estimates of the expected values for the quantities in the Challenge Problems than there was about the distributions of the quantities. Although most authors expressed their estimates of the expectations as intervals, some authors used more complex uncertainty structures. Among the answers given as intervals, the bounds on the expected value depend to a noticeable degree on the method for aggregation used to solve the problem. Beyond that, there also appear to be unexplained differences in the calculated numerical values offered by the various authors. The source and meaning of these residual disagreements expressed in the answers about the expected values have not been fully determined.

3.5. Synopsis of the various approaches

Table 2 summarizes some of the most important features of the various approaches used in the papers. The second column of the table specifies which of the Challenge Problems were solved in each paper. Several papers, such as Red-Horse and Benjamin [41], Ben-Haim [46], and O'Hagan and Oakley [45], either *partially* solved some of the Challenge Problems or solved *other* problems that were related to the Challenge Problems. These efforts are not reflected in the column, which only mentions a problem if a complete and final answer was given that can be quantitatively compared with answers given in the other papers.

The third column indicates how the uncertain quantities were represented in each paper. Generally this also indicates the theoretical approach employed by the authors. In several cases, there were multiple approaches employed in a single paper. In such cases, multiple answers were typically computed for each problem. Sometimes, however, the different approaches led to the same answer for a particular problem.

The fourth column of the table indicates how each paper dealt with the technical problem of aggregation described in Section 2.3. Many of the papers employed multiple different aggregation approaches. In this column, the term *mixture* indicates that the 'equal credibility' specified in the Challenge Problems [2] of multiple estimates for a quantity was modeled by equal probability using a stochastic mixture model in which the distribution functions are (vertically) averaged together with equal weights.

Table 2
Summary of approaches to the Challenge Problems

Author(s)	Problems solved	Representation of inputs	Aggregation method	Computational algorithm	Strategy for repeated parameters	Answer to problem 1
Klir [31]	1,2,3a	Dempster–Shafer structures	Mixture	(Unspecified)	(Unspecified)	[0.69, 2]
Helton et al. [32]	All	Probability distributions, Dempster–Shafer structures, possibility distributions	Mixture	Random sampling	Sampling	Degenerate Dempster–Shafer structure over [0.7, 1.98], probability distribution over same range (both given graphically)
Fetz and Oberguggenberger [33]	1,2,3,4,5,6	Random intervals, sets of probability measures, fuzzy sets	Mixture	Global optimization	Exact evaluation	(Given graphically, but consistent with [0.69, 2])
Hall and Lawry [34]	4,5	Random sets	Mixture, other methods	Optimization (gradient)	Sampling	-
Kozine and Utkin [35]	1,2,3,4,5,6	Imprecise coherent probabilities	(Depends on numerical credibilities)	Mathematical programming	Mathematical programming	[0.69, 2.0]
De Cooman and Troffaes [36]	1,2,3,4,5,6	Coherent lower previsions	Natural extension of pointwise maximum, lower envelope (pointwise minimum)	Independent natural extension	Independent natural extension	Vacuous lower prevision over [0.692201, 2.0]
Ferson and Hajagos [38]	All	P-boxes	Mixture, other methods	Cartesian product of intervals and proababilities	Subinterval reconstitution	[0.692, 2]
Berleant and Zhang [39]	All	Random sets in joint distribution tableaus	Mixture	Interval-based discrete convolution	Systematic sampling	[0.69, 2]
Tonon [40]	B	Random sets	Dempster's rule, mixture	Vertex method	Sampling	–
Red-Horse and Benjamin [41]	B	Families of polynomial chaos expansions	A posteriori mixture	Chaos expansion technique	Sampling	–
Li and Hyman [42]	1,2,3,4,5,6	Probability distribution variables	Mixture	PDV arithemetic	Dependency tracking	[0.6922, 2.0], refined p-box over that range
Rutherford [43]	1,2b,4,6	Basic probability assignments response surface	Mixture, Dempster's rule	'Intelligent' sampling	Sampling	[0.69, 2]
Ayyub [44]	1,2,3a,3b,6	Tableaus	Mixture	Vertex method	Vertex method	[0.7943282, 2.0]
O'Hagan and Oakley [45]	1	Probability distributions	Encourage consensus among experts	Bayesian	–	(Not comparable)
Ben-Haim [46]	None	Info-gap model	Info-gap theory	–	Info-gap theory	–
Kreinovich and Ferson [47]	None	Intervals	–	Cauchy deviate method	Sampling	–
Agarwal et al. [48]	None	Interval variables with known bpas	Dempster's rule	–	–	–
Soundappan et al. [49]	None	Evidence theory, Bayesian probability density functions	Mixture, Bayes' rule	Optimization, probability integration	Optimization formulation	(Not comparable)
Cooke [50]	None	Subjective probability distributions	Mixture	–	–	–
Bier [51]	None	Percentiles of subjective probability distribution	Mixture, copula-based methods	–	–	–
Booker and McNamara [52]	1, 3c	Endpoint intervals	Convex hull of outputs	Vertex method	Vertex method	[0.79, 2.0]
Yager [53]	None	Dempster–Shafer structures, fuzzy measures	Order-weighted averaging	–	–	–

The fifth column describes the computational algorithm employed in each paper and the sixth column indicates the technique used to handle the problem of repeated parameters discussed in Section 2.4. Of course, the phrases used in these columns can only very crudely characterize the methods employed in the individual papers, which should be read for full details. The term *sampling* means that a scalar real value was sampled for each input and that these real numbers were combined arithmetically and propagated through the response function to obtain an output. This approach gives the same value to each instance of a repeated parameter. The observed extreme values computed in this way were presumed to be the largest and smallest possible values. The *vertex method* indicates that only the corners of the input space were considered in the search for the global minimum and maximum. In the case of two variables, these are just the four pairs of endpoints mentioned in Section 3.3.1. For the meanings of other terms used in the table, consult the individual papers.

The last column of Table 2 gives the answer to the first Challenge Problem, which was expressed in terms of intervals. What is shown in the column is whatever the author(s) of each paper offered as the answer. In most case, the answer was explicitly given as an interval. Because this is the simplest problem, it is the only one whose answer can be conveniently characterized in the table. Of course, this problem is not representative of the other Challenge Problems, some of which are much more complicated. The column illustrates the diversity of different answers given by the several authors even for the simplest problem. Nevertheless, the answers also suggest that, emerging through this diversity, there is a quantitative consensus among most of the authors.

4. Discussion and future research needs

Is a consensus emerging about how epistemic uncertainty should be handled in numerical calculations? Is a consensus emerging about how it should be combined with aleatory uncertainty? There is, apparently, a partial consensus among some analysts, at least about several general ideas and specific methodological details. Among the authors who provided quantitative answers to the Challenge Problems in terms of belief and plausibility functions or bounds on distribution functions, there was very strong consensus, although there were small differences due to the inefficiency of sampling, or the number of discretization levels employed, and different algorithmic strategies. There appears to be wide agreement that the more epistemic uncertainty there is about a quantity the less able analysts will be to characterize the quantity with a precise probability distribution.

It is abundantly clear, however, that the consensus on these matters does not include many probabilists. Fundamental disagreement remains about the proper statistical foundation of the new methods. Some researchers criticize them as misguided and unnecessary. They maintain that the methods are poor substitutes for the proper use of probability theory. When such fundamental disputes arise, it is natural (and historically common) that schools of thought develop around the respective positions. These schools of thought foster discourse and progress within groups of like-minded researchers but generally inhibit discussion with outsiders. Over time, the paucity of social and intellectual interactions allows the growth of distinct jargons within the two schools that further suppresses cross-communication. The result is often a disadvantageous fragmentation of the discipline.

The dangers to any discipline of fragmentation are real. It is important to concentrate on developing areas of concordance. We should temper the enthusiasm about any emerging consensus with due attention to the criticisms, and avoid ignoring the difficult interactions or leaving behind the nay-sayers. At the same time, we should strive to understand the dissatisfaction that led to the development of new approaches and acknowledge the limitations and inconveniences of standard methods. There will always be disagreements, but open and inclusive discussion is the path to a healthy scientific/engineering discipline.

Several participants of the Albuquerque workshop expressed interest in a follow-up event on related computational problems involving epistemic uncertainty. Possible topics for future workshops include questions on handling model uncertainty and dependence, or lack of information about dependence, among variables. Several other fundamental questions also need further research. For instance, how can an analyst meaningfully elicit or otherwise acquire representations of uncertainty within the various non-traditional theories considered potentially useful in handling epistemic uncertainty? How can the meaning of the numerical values (e.g. ‘belief’, ‘possibility’, etc.) be standardized across multiple informants and users? How should the results be communicated to decision makers? How can sampling methods be made efficient for computationally difficult problems? How should sensitivity analyses be performed and interpreted in the context of alternative uncertainty representations? One or more workshops could fruitfully be organized around such questions so as to encourage focused but inclusive discussion.

In addition to the fundamental questions, there are other issues concerning the mechanics of calculation that deserve consideration. For instance, it might be useful to hold a workshop that focuses attention on a significantly more realistic and complicated problem set. More complex problems would help to establish that proposed methods would actually be practicable for use in real-world settings. It is not clear, however, that such a workshop could be organized around prescribed challenge problems similar to those considered in the Albuquerque workshop. The more realistic and complicated the problems become, the less

reasonable it becomes to expect participants to solve the problems themselves before or during the workshop.

Another possible workshop topic concerns the related issue of how uncertainty can be handled when the model is a black box whose internal details are not known to the analyst. Many of the methods of uncertainty propagation that are in widespread use are 'intrusive' in the sense that they must be applied to individual binary operations such as a particular addition or multiplication. This greatly limits their applicability to an important class of problems. In many engineering applications, there is a need for techniques that can propagate uncertainty through black box codes that cannot be decomposed into a composition of binary operations. These black boxes are often large, legacy software programs characterized by high input dimensionality, long computation times, and strong disincentives or restrictions against recoding. They often incorporate iterative convergence testing or an interactive re-gridding of the mesh for numerical solutions to partial differential equations that implies that each vector of real-valued inputs potentially induces a different sequence of mathematical operations. Such codes cannot, even in principle, be decomposed into a finite set sequence of binary operations. Most of the non-traditional methods discussed at the workshop, and indeed, most uncertainty methods generally, are intrusive and cannot be applied to such problems. There is a need for research to generalize these methods (or invent new methods) that can be applied to this important class of problems.

Acknowledgements

We thank the authors of the contributed papers who patiently answered many of our questions. Any residual mistakes or misunderstandings about their papers are our fault. We also thank Troy Tucker who offered several helpful suggestions. This work was performed for Sandia National Laboratories, which is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Security Administration under contract DE-AC04-94AL-85000.

References

[1] http://www.sandia.gov/epistemic/eup_workshop1.htm
[2] Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S. Challenge Problems: uncertainty in system response given uncertain parameters. Reliab Engng Syst Saf 2004; this issue.
[3] Apostolakis G. The concept of probability in safety assessment of technological systems. Science 1990;250:1359–64.
[4] Helton JC. Treatment of uncertainty in performance assessments for complex systems. Risk Anal 1994;17:483–511.
[5] Hoffman FO, Hammonds JS. Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. Risk Anal 1994;14:707–12.
[6] Paté-Cornell ME. Uncertainties in risk analysis: six levels of treatment. Reliab Engng Syst Saf 1996;54:127–32.
[7] Helton JC. Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. J Stat Comput Simul 1997;57(1–4):3–76.
[8] Vesely WE, Goldberg FF, Roberts NH, Haasl DF. Fault tree handbook. Washington, DC: Nuclear Regulatory Commission; 1981.
[9] Mosleh A, Siu N, Smidts C, Lui C. Proceedings of Workshop I in Avdanced Topics in Risk and Reliability Analysis. Model Uncertainty: Its Characterization and Quantitfication. NUREG/CP-0138, US Nuclear Regulatory Commission, Washington, DC; 1994.
[10] Laskey KB. Model uncertainty: theory and practical implications. IEEE Trans Syst, Man, Cybern—Part A: Syst Hum 1996;26:340–8.
[11] Meyer MA, Booker JM. Eliciting and analyzing expert judgment: a practical guide. New York: Academic Press; 1991.
[12] Cooke RM. Experts in uncertainty. Cambridge: Oxford University Press; 1991.
[13] Huber PJ. Robust statistics. New York: Wiley; 1981.
[14] Moore RE. Interval analysis. Englewood Cliffs, NJ: Prentice-Hall; 1966.
[15] Neumaier A. Interval methods for systems of equations. Cambridge: Cambridge University Press; 1990.
[16] Kaplan S. On the method of discrete probability distributions—application to seismic risk assessment. Risk Anal 1981;1:189–98.
[17] Iman RL, Helton JC. Comparison of uncertainty and sensitivity analysis techniques for computer models. NUREG/CR-3904, SAND84-1461, Sandia National Laboratories, Albuquerque, NM; 1985.
[18] Morgan MG, Henrion M. Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge, UK: Cambridge University Press; 1990.
[19] Hyman JM. FORSIG: an extension of FORTRAN with significance arithmetic. LA-9448-MS, Los Alamos National Laboratory, Los Alamos, NM; 1982. See also the website . Available at http://math.lanl.gov/ams/report2000/significancearithmetic.html.
[20] Manes EG. A class of fuzzy theories. J Math Anal Appl 1982;85:409–51.
[21] Hailperin T. Boole's logic and probability. Amsterdam: North-Holland; 1986.
[22] Springer MD. The algebra of random variables. New York: Wiley; 1979.
[23] Kirchner TB. QS-CALC: an interpreter for uncertainty propagation. Fort Collins, Colorado: Quaternary Software; 1992.
[24] Ferson S. Automated quality assurance checks on model structure in ecological risk assessments. Hum Ecol Risk Assess 1996;2:558–69.
[25] Joslyn C. Measurement of possibilistic histograms from interval data. Int J Gen Syst 1997;26(1–2):9–33.
[26] Couso I, Moral S, Walley P. A survey of concepts of independence for imprecise probabilities. Risk Decis Policy 2000;5:165–81.
[27] Yager RR. Arithmetic and other operations on Dempster–Shafer structures. Int J Man–Mach Stud 1986;25:357–66.
[28] Couso I, Moral S, Walley P. Examples of independence for imprecise probabilities. In: De Cooman G, Cozman FF, Moral S, Walley P, editors. Proceedings of the First International Symposium on Imprecise Probability and Their Applications. Imprecise Probabilities Project, Universiteit Gent; 1999. p. 121–30.
[29] Fetz T. Sets of joint probability measures generated by weighted marginal focal sets. In: De Cooman G, Fine TL, Seidenfeld T, editors. Proceedings of the Second International Symposium on Imprecise Probability and Their Applications. Maastricht: Shaker Publishing; 2001. p. 171–8.
[30] Cozman FG, Walley P. Graphoid properties of epistemic irrelevance and independence. In: De Cooman G, Fine TL, Seidenfeld T, editors.

Proceedings of the Second International Symposium on Imprecise Probability and Their Applications. Maastricht: Shaker Publishing; 2001. p. 112–21.
[31] Klir GJ. Generalized information theory: aims, results, and open problems. Reliab Engng Syst 2004; this issue.
[32] Helton JC, Johnson JD, Oberkampf WL. An exploration of alternative approaches to the representation of uncertainty in model predictions. Reliab Engng Syst 2004; this issue.
[33] Fetz T, Oberguggenberger M. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. Reliab Engng Syst 2004; this issue.
[34] Hall JW, Lawry J. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. Reliab Engng Syst 2004; this issue.
[35] Kozine IO, Utkin LV. An approach to combining unreliable pieces of evidence and their propagation in a system response analysis. Reliab Engng Syst Saf 2004; this issue.
[36] De Cooman G, Troffaes MCM. Coherent lower previsions in systems modeling: products and aggregation rules. Reliab Engng Syst Saf 2004; this issue.
[37] Walley P. Statistical reasoning with imprecise probabilities. London: Chapman and Hall; 1991.
[38] Ferson S, Hajagos J. Arithmetic with uncertain numbers: rigorous and (often) best-possible answers. Reliab Engng Syst Saf 2004; this issue.
[39] Berleant D, Zhang J. Representation and problem solving with distribution envelope determination (DEnv). Reliab Engng Syst Saf 2004; this issue.
[40] Tonon F. Using random set theory to propagate epistemic uncertainty through a mechanical system. Reliab Engng Syst Saf 2004; this issue.
[41] Red-Horse JR, Benjamin AS. A probabilistic approach to uncertainty quantification with limited information. Reliab Engng Syst Saf 2004; this issue.
[42] Li W, Hyman JM. Computer arithmetic for probability distribution variables. Reliab Engng Syst Saf 2004; this issue.
[43] Rutherford B. A response-modeling approach to characterization and propagation of uncertainty specified over intervals. Reliab Engng Syst Saf 2004; this issue.
[44] Ayyub BM. From dissecting ignorance to solving algebraic problems. Reliab Engng Syst Saf 2004; this issue.
[45] O'Hagan A, Oakley JE. Probability is perfect, but we can't elicit it perfectly. Reliab Engng Syst Saf 2004; this issue.
[46] Ben-Haim Y. Uncertainty, probability and information-gaps. Reliab Engng Syst Saf 2004; this issue.
[47] Kreinovich V, Ferson S. A new Cauchy-based black-box technique for uncertainty in risk analysis. Reliab Engng Syst Saf 2004; this issue.
[48] Agarwal H, Renaud JE, Preston EL, Padmanabhan D. Uncertainty quantification using evidence theory in multidisciplinary design optimization. Reliab Engng Syst Saf 2004; this issue.
[49] Soundappan P, Nikolaidis E, Haftka RT, Grandhi R, Canfield R. Comparison of evidence theory and Bayesian theory for uncertainty modeling. Reliab Engng Syst Saf 2004; this issue.
[50] Cooke R. The anatomy of a squizzel: the role of operational definitions in representing uncertainty. Reliab Engng Syst Saf 2004; this issue.
[51] Bier V. Implications of the research on expert overconfidence and dependence. Reliab Engng Syst Saf 2004; this issue.
[52] Booker JM, McNamara LA. Solving black box computation problems using expert knowledge theory and methods. Reliab Engng Syst Saf 2004; this issue.
[53] Yager RR. Uncertainty modeling and decision support. Reliab Engng Syst Saf 2004; this issue.
[54] S. Ferson, V. Kreinovich, L. Ginzburg, K. Sentz, D.S. Myers, Constructing probability boxes and Dempster–Shafer structures. Sandia National Laboratories, Technical Report SAND2002-4015, Albuquerque, NM, 2002. Available at http://www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf.