Methods for the evaluation and synthesis of multiple sources of information applied to nuclear computer codes

S. Destercke *,[1] E. Chojnacki

Institut de Radioprotection et de Sûreté Nucléaire
13115 St-Paul Lez Durance, France

Abstract

This work is devoted to methods used to evaluate and synthesize multiple sources giving information about a variable which true value is not precisely known. We first recall probabilistic and possibilistic approaches to solve the problem. Each approach offers a formal setting to evaluate, synthesize and analyze information coming from multiple sources. They are then applied to the results of uncertainty studies performed in the framework of BEMUSE project.

Key words: aggregation, uncertainty, possibility theory, data analysis, probability, validation

1 Introduction

Best estimates computer codes are increasingly used in nuclear industry for the accident management procedures and have been planned to be used for the licensing procedures. Contrary to conservative codes which are supposed to give penalizing results, best estimate codes attempt to calculate accidental transients in a realistic way. It becomes therefore of prime importance, in particular for technical organizations in charge of safety assessment (e.g. IRSN), to know the uncertainty on the results of such codes. Thus statistical methods

* Corresponding author
Email addresses: sebastien.destercke@irsn.fr (S. Destercke), eric.chojnacki@irsn.fr (E. Chojnacki).
[1] Address : Institut de Radioprotection et de Sûreté Nucléaire, Bat 702, 13115 St-Paul Lez Durance, France Tel: +33 4 42 19 97 02 Fax: +33 4 42 19 91 66

have been developed to take into account the uncertainties coming from the input data and models parameters.

Studies performed from these methods (e.g. UMS, BEMUSE) allowed to point out two important issues:

(1) the difficulty and the need to build synthetic representations of input uncertainty
(2) the need to compare, analyze and synthesize results of uncertainty studies

It appears that both issues can be viewed as problems of information fusion in presence of multiple sources. Concerning the former issue, the evaluation of input uncertainty comes either from experts or from experimental results. It is therefore desirable to take these multiple sources into account during the modeling of input uncertainty. In the same way, by looking at the result of each uncertainty study as a single source of information, their analysis, comparison and synthesis can be viewed as an information fusion problem.

In this paper, we propose to recall the principles of the probabilistic method and to show the advantages of extending these principles to the frame of other uncertainty theories (here, possibility theory). A practical application (related to the second issue mentioned above) based on the results obtained during BEMUSE benchmark is described.

The rest of the paper is divided in two main sections. Section 2 introduces the problems of information modeling, evaluation and aggregation, as well as how probability and possibility theories can be used to solve these problems. Section 3 then explains how these methods have been applied to the BEMUSE program, and comments the various results.

2 Methodology

Most of the formal approaches proposing to handle information provided by multiple sources consists in three main steps : modeling the information provided by each source, evaluating the sources by the quality of the provided information and finally synthesizing this information. Results can then be used either to build a final uncertainty model of the studied variable or to analyze the different sources and the relations between them.

This section details each of these steps, both for the probabilistic and the possibilistic approaches.

The probabilistic approach is explained in ((Bedford and Cooke 2001),Ch 10) and is extensively studied and motivated in (Cooke 1991). It is mathematically

well-founded and its practical interest has been confirmed by many applications over the years. Nevertheless, it can be argued that using it when available information is scarce or not fully reliable (as is often the case when multiple sources are needed) forces one to introduce additional assumptions that are not justified by available knowledge (Ferson and Ginzburg 1996).

Possibility theory (Dubois and Prade 1988), on the other hand, offers a simple and convenient formal alternative that explicitly takes account of imprecision and scarcity in the available information. As we shall see, the possibilistic approach also offers more aggregation operators than the probabilistic one. This last point amount to a greater flexibility in the information treatment. The possibilistic approach presented in this paper heavily draws on the methodology presented in (Sandri et al. 1995).

For sake of brevity, we only introduce in this paper the basic ideas of the methods as well as the formulas that will be used in the application. The methods and the case studied here are the most commonly encountered in practical applications, and are likely to be directly applicable in numerous situations. Otherwise, we send back readers to works referenced in this paper for generalizations and more complex propositions.

2.1 Modeling information

When multiple sources give information about a badly known value (due to lack of experimental values, of imperfect observation, ...), this information consists most of the time in some (imprecise) characteristics or parameters of an unknown probability distribution. Usually, the provided information is not sufficient to define a single probability distribution, and there is potentially an infinity of probability distributions corresponding to the information. The usual procedure is to single out a distribution that fits the given information and minimize some information measure (typically, the entropy). This procedure is based on the legitime informal argument that, if one have to add information, it should be the least possible amount. An alternative is to consider other models than probability distributions (such as possibility distribution) that explicitly takes account of this imprecision and incompleteness in the information given by the source. This way, the model becomes less precise, but no information is added to the one we have.

2.1.1 probabilistic modeling

Partial probabilistic information can come in many different ways: the value of some percentiles, characteristics of a parameterized family of distribution (e.g. mean and variance of a normal distribution) to which the random variable is

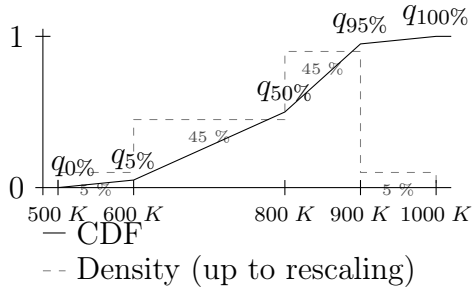

Figure 1. Example of probabilistic modeling

supposed to belong, information about the mode, the mean, the median, or even qualitative comparisons (see (Walley 1991), ch. 4. for a review).

For simplicity, we consider in this paper that the information is given in term of percentiles of the unknown probability distribution (typically, the $5\%, 50\%$ and 95% percentiles). If X is the unknown variable, and P a probability measure, then the $k\%$ percentile, denoted $q_{k\%}$, is the deterministic value x s.t. $P(X \leq x) = k\%$. This choice is explained by the fact that percentiles are the commonest type of information encountered in applications where few experimental data or evidences are available. Choosing the probability distribution that add a minimal amount of information then comes down to make a linear interpolation between each percentiles. If $B + 1$ percentiles values have been given (including $q_{0\%}$ and $q_{100\%}$), then the corresponding probability density $p = (p_1, \ldots, p_B)$ is an histogram made of B interpercentiles (an interpercentile being the difference between two successive $q_{k\%}$ values given by the source. If more than one sources give information about a variable, we will note q_l and q_u the lower and upper bounds of the intrinsic range of variation (i.e. the minimum and maximal values taken by the variable)

Figure 1 represents the cumulative distribution function (CDF) corresponding to the clad temperature of a reactor where information given by the source is $q_{0\%} = 500\ K$, $q_{5\%} = 600\ K$,$q_{50\%} = 800\ K$,$q_{95\%} = 900\ K$, $q_{100\%} = 1000\ K$. The corresponding probability density $p = (0.05, 0.45, 0.45, 0.05)$ is pictured in dashed lines.

2.1.2 Possibilistic modeling

A possibility distribution over the reals is formally defined as a mapping $\pi : \Re \rightarrow [0, 1]$. Formally, it is equivalent to the membership function of a fuzzy set (Zadeh 1978).

From this distribution can be defined two set-functions that describe the likelihood of events:

- Possibility measure: $\Pi(A) = \sup_{x \in A} \pi(x)$.

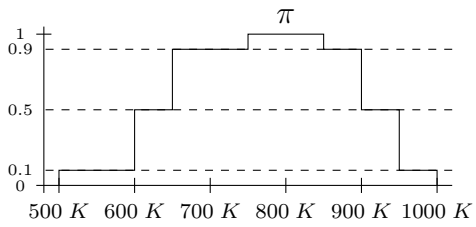

Figure 2. Example of possibilistic modeling

- Necessity measure: $N(A) = 1 - \Pi(A^c)$.

Possibility degree of an event expresses the extent to which this event is plausible, i.e., consistent with our knowledge. Necessity degrees express the certainty of events, by duality.

Given a possibility distribution π, to every value $\alpha \in [0, 1]$ we can associate an α-cut A_α which is the set $A_\alpha = \{x|\pi(x) \geq \alpha\}$. a possibility distribution can then be viewed as a collection of nested sets that are α-cuts (i.e. $A_1 \subseteq A_\alpha \subseteq A_\beta \subseteq A_0$ where $\alpha \geq \beta$).

A possibility distribution can also be interpreted as the simplest model describing a family of probability distributions (Dubois and Prade 1992), and is thus a model of partial probabilistic information. In this latter case, each α-cut can be seen as a confidence interval having a confidence level $1 - \alpha$, since we have $P(A_\alpha) \geq N(A_\alpha)$ $(N(A_\alpha) = 1 - \alpha)$ where P is the imprecisely known probability measure. Possibility distributions are natural candidates to model incomplete probabilistic information given in term of confidence intervals, but can also be used when the information concerns other characteristics (e.g. mean, percentiles, mode) of the unknown probability distribution (see (Baudrit and Dubois 2006,Dubois et al. 2004) for detailed discussions).

Figure 2 represents a possibility distribution corresponding to the clad temperature of a reactor where information given by the source consists of four intervals $[750\ K, 850\ K], [650\ K, 900\ K], [600\ K, 950\ K], [500\ K, 1000\ K]$ which have respective confidence levels of $10\%, 50\%, 90\%$ and 100%.

2.2 Evaluating the sources

Once sources have given information about the value of a variable, it is desirable to evaluate the quality of the information delivered by the sources, in order to know which are the "best" sources. In the approaches used here, this evaluation is done by two criteria:

- Informativeness (Inf): evaluate the precision of the information given by the sources, by using a measure of comparison between the model built from

the source information and the model corresponding to ignorance in the considered theory. Of course, the more precise the source, the more useful it is.

- Calibration (Cal): evaluate, by using so-called seed variables, how good the evaluations of the sources are. Seed-variables are variables which are in nature close (i.e. concern similar physical phenomenum) to the unknown variables and which exact value are unknown to the sources but are (or will be) known to the assessors. Computing a calibration score then consists of comparing the information given by the sources with the true values of the seed-variables. A "good" source is then a source which information is in concordance with the observed values of the seed variables. Here, we consider that the seed variable values are precisely known [2] .

A global score rating the source is then computed from these two criteria. Such a score should follow common sense rules such as fairness (the more the sources is precise and close to observed values, the greater its score) , pertinence (scores computation should be based only on observed values) and soundness (scores between sources should be directly comparable). Let us also note that the definition of suitable seed-variables must be done with great care, and is feasible in most situations (see (Cooke 1991) for a detailed discussion).

Once the global score of each source is determined, it is used either to discriminate sources between themselves or directly in the aggregation procedure.

2.2.1 Probabilistic evaluation

Informativeness In probability theory, the model commonly used to represent a state of ignorance regarding a particular variable is the uniform distribution, which will be denoted u. Let $p = (p_1, \ldots, p_B)$ be the probability distribution derived from the source information. Then, the informativeness of a source s for this variable is

$$I(p, u) = \sum_{i=1}^{B} p_i \log\left(\frac{p_i}{u_i}\right) \tag{1}$$

where $I(p, u)$ is the Kullbach-Leibler (KL) divergence (sometimes called distance, although it is not a metric) of u from p. If we take back the example from figure, then we have $p = (0.05, 0.45, 0.45, 0.05)$ and $u = (0.20, 0.40, 0.20, 0.20)$. If sources provide information for more than one seed variable, the global informativeness score of a source s is the mean of its informativeness over all seed variable. Let N be the number of seed variables and let $I(p_j, u_j)$ denote

[2] Extension when the true value of the seed variables is itself imprecisely known can be found in ((Kraan 2002), Ch. 3) and in (Sandri et al. 1995), respectively for the probabilistic and possibilistic approaches.

the informativeness score for the j^{th} seed variable. The global informativeness score of source s is then

$$Inf_p(s) = \frac{1}{N}\sum_{j=1}^{N} I(p_j, u_j) \tag{2}$$

Let us note that the uniform distribution is always taken over the entire range $[q_l, q_u]$ (even if a source has given values $q_{0\%} \geq q_l$ and $q_{100\%} \leq q_u$).

Calibration Suppose a source s has given the same series of percentiles for N different seed variables. We thus have probability distributions that have the same interpercentiles and can be viewed as one distribution $p = (p_1, \ldots, p_B)$ (the values associated to the fixed percentiles can be different for each seed variable). Now, let us assume that for $r_1 N$ seed variables, realizations fall into interpercentile p_1, into p_2 for $r_2 N$ realizations, etc. $r = (r_1, \ldots, r_B)$ is then the empirical density derived from observations of seed variables. Again, we use the KL divergence of p from r to measure the "surprise" of learning r when p is thought to be the right answer, we thus compute

$$I(r, p) = \sum_{i=1}^{B} r_i \log\left(\frac{r_i}{p_i}\right) \tag{3}$$

which reach its minimal value 0 if and only if $r = p$. Since it is known that the value $2 * N * I(r, p)$ converge to a chi-squared distribution with $B - 1$ degrees of freedom as N gets larger (i.e. $P(2 * N * I(r, p) \leq x) \rightarrow \chi^2_{B-1}(x)$), the calibration of a source s is defined as

$$Cal_p(s) = 1 - \chi^2_{B-1}(2 * N * I(r, p)) \tag{4}$$

which can be interpreted as the probability of the source to be "right". In practice, N (the number of seed variables) should be around 10 (or more) to be sure to have good results. This way of computing the calibration score assumes that a source is well calibrated if $p_i\%$ of the realizations of seed variables fall into the interpercentile p_i of the N built probability distributions. Although these assumptions are perfectly justified in a probabilistic framework, they can be challenged if one accepts to use other uncertainty models (as already pointed out in (Sandri et al. 1995)). This will be illustrated and discussed in the application.

Figure 3 illustrates the process of calibration. In this figure, 2 sources each gives their percentiles $q_{0\%}, q_{5\%}, q_{50\%}, q_{95\%}, q_{100\%}$ for two different seed variables. So, the common interpercentile distribution is $p = (0.05, 0.45, 0.45, 0.05)$. If we note r^i the empirical distribution of source i given the realisations of seed variables, we have $r^1 = (0, 0.5, 0.5, 0)$ and $r^2 = (0.5, 0, 0.5, 0)$. Given these seed variables and the information given by the sources, the first source is

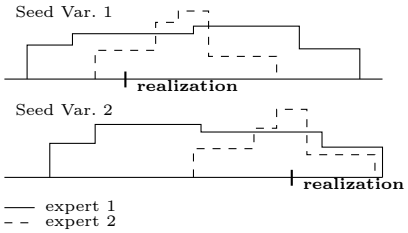

Figure 3. Example of calibration

thus judged better calibrated than the second (as the empirical distribution for source 1 is closer to p). Let us also note that, in term of informativeness, source 2 has a better score than source 1.

Global score If M sources give information about N seed variables, the global score of the m^{th} source s_m is given by

$$W_p(s_m) = C_\kappa \times Inf_p(s_m) \times Cal_p(s_m) \tag{5}$$

where $Inf_p(s_m), Cal_p(s_m)$ are respectively the informativeness and calibration scores introduced above, and C_κ is the function

$$C_\kappa = \begin{cases} 1 \text{ if } Cal_p(s_m) \geq \kappa \\ 0 \text{ if } Cal_p(s_m) < \kappa \end{cases} \tag{6}$$

where level κ makes sure that sources judged too badly calibrated are rejected. κ insures that sources cannot get high scores by giving very precise but badly calibrated information about the seed variables. In practice, κ can be set to a fixed value or be tuned so that the combined probability distribution would get a maximized score on seed variables (see (Cooke 1991) for more details).

Scores of the different sources are then normalized so that they sum up to one. The final score $w_p(s_m)$ of the m^{th} source is

$$w_p(s_m) = \frac{W(s_m)}{\sum_{i=1}^{M} W(s_i)} \tag{7}$$

2.2.2 Possibilistic evaluation

Let us first recall that the cardinality $|A|$ of an interval $A = [\underline{a}, \overline{a}]$, equal to $\overline{a} - \underline{a}$, is a measure of the imprecision of this interval.

Informativeness Given the range q_l, q_u, the possibilistic model corresponding to complete ignorance is the possibilistic distribution π_{ign} s.t. $\pi_{ign}(x) = 1$

if $x \in [q_l, q_u]$, 0 otherwise (let us note that this model is formally equivalent to the interval $[q_l, q_u]$, while the model used for ignorance in probability is NOT equivalent to this interval).

Let us now consider a source s which information about a seed variable is modeled by possibility distribution π_s. the cardinality $|\pi_s|$ of this distribution, which reads

$$|\pi_s| = \int_{q_l}^{q_u} \pi(x)dx$$

Which is simply is the area under the distribution, and also a measure of the imprecision of the information given by s. Informativeness of s can then be computed as

$$I(\pi, s) = \frac{|\pi_{ign}| - |\pi_s|}{|\pi_{ign}|} \tag{8}$$

where the denominator is a normalization factor. Equation (8) has value 1 iff π_s is reduced to a precise value and 0 iff $\pi_s = \pi_{ign}$ (note that $|\pi_{ign}| = q_u - q_l$). If source s give information about N seed variables, then its global informativeness score is

$$Inf_\pi(s) = \frac{1}{N} \sum_{j=1}^{N} I(\pi_m, s) \tag{9}$$

where $I(\pi_m, s)$ is equation (8) computed for the possibility distribution corresponding to the m^{th} seed variable.

Calibration Let X be a seed variable, and x^* the true (known) value of this variable. Then, if source s information correspond to the possibilistic model π_s, the calibration score is simply the value $\pi_s(x^*)$, since this value measures to which extent x^* is judged plausible as the true value of X by source s. If the source gives information about N seed variables, the global calibration score is

$$Cal_\pi(s) = \frac{1}{N} \sum_{j=1}^{N} \pi_{s,n}(x_n^*) \tag{10}$$

where x_n^* is the observed value of the n^{th} seed variable, and $\pi_{s,n}$ is the possibility corresponding to the information given by s for the n^{th} seed variable.

Global score As for the probabilistic approach, the global score of the m^{th} source among M sources is given by is given by

$$W_\pi(s_m) = C_\kappa \times Inf_\pi(s_m) \times Cal_\pi(s_m) \tag{11}$$

2.3 Synthesizing the information

As said before, aggregating the information given by multiple sources can have two purposes: either build a reliable synthetic model that will be used in further processing of the information, or analyze the available information and the relations between the sources (conflict, concordance, ...). Basically, aggregation operators follows three main kinds of behavior:

- Conjunctive behavior: conjunctive operators for uncertainty models are the equivalent of intersection for usual intervals. The benefit of conjunctive operators is that the resulting model is more precise than any of the model provided by the multiple sources. Nevertheless, conjunctive operators make the assumption that all the sources are reliable, and can result in poor reliable models or even in an empty result (i.e. in cases similar to disjoint intervals) if this assumption is not verified. In this sense, using conjunctive operators correspond to an optimistic or adventurous attitude.
- Disjunctive behavior: disjunctive operators for uncertainty models are the equivalent of union for usual intervals. The result of disjunctive operators is usually an imprecise, but highly reliable model. Nevertheless, the result will often be too imprecise to be really useful. Disjunctive operators make the cautious assumption that **at least** one source is reliable. In this sense, they correspond to a cautious or pessimistic attitude.
- Compromise behavior: compromise operators are between a conjunctive and a disjunctive behavior. The most commonly used of such operators is the arithmetic mean, which can be associated to a statistical counting where each model given by a source is considered as an experiment. Compromise operators are often use to resolve the conflict between sources while trying at the same time to gain a maximum of informativeness.

For a more complete characterization of aggregation operators, see (Bloch 1996). Let us note that if the aim of the aggregation is to analyze the information and the sources having delivered it (the purpose followed in this paper), operators from which one can extract meaningful information are to be privileged, while it won't forcefully be the case (although it is often desirable) when the aim is to build a synthetic representation. We now give some details about the basic operators used in probability and possibility theory.

2.3.1 probabilistic synthesis

Desirable properties and usefulness of various aggregation operators of probability distribution has been the subject of many research and discussions over the past (see (Clemen and Winkler 1999) for a recent review). However, it is now commonly accepted that the (weighted) arithmetic mean is the most

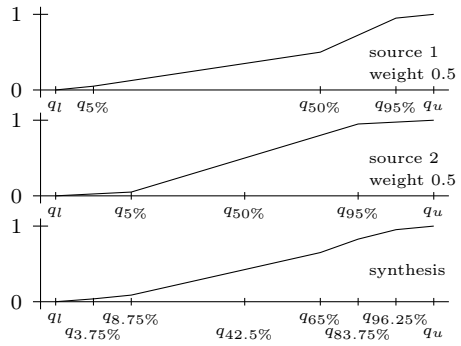

Figure 4. Probabilistic Aggregation Illustration

sensible direct [3] aggregation operator to be used with probability distributions (again, see (Cooke 1991) for a review and (McConway 1981) for a more detailed discussion). If M sources give their advice about a variable, the arithmetic mean reads

$$p_{mean} = \sum_{i=1}^{m} w_i p_i \tag{12}$$

where w_i and p_i are respectively the weight and the probability distribution obtained from the information given by the i^{th} source.

Usual conjunctive [4] and disjunctive operators are not applicable in the probabilistic approach (union of two probability distributions is a set of distributions, not a unique distribution, while the intersection is always empty, except when all distributions are equal). Figure 4 illustrates the fusion of distributions coming from 2 sources which have equal weights.

2.3.2 possibilistic synthesis

There are numerous aggregation operators in possibility theory (see (Dubois and Prade 2001) (Oussalah et al. 2003) for reviews). Although choosing one particular operator is not always easy, their multiplicity bring more flexibility in the aggregation of the information given by multiple sources, and often permit a finer analysis of the available information. In this paper, we restrict ourselves to the three most commonly used conjunctive, disjunctive and compromise operators, which are the one used in the subsequent application.

Conjunctive operator given m sources, each of them giving information modeled by a possibility distribution π_i $i = 1, \ldots, n$, the conjunction is the

[3] Bayesian approaches won't be used here, because they're more closely related to the updating rather than to the aggregation of information

[4] We must mention that the geometric mean can be interpreted, to some extent, as a conjunctive operator

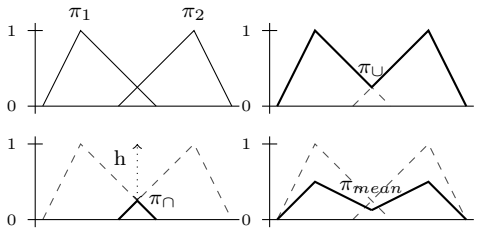

Figure 5. Possibilistic aggregation: illustration

minimum of all these distributions. Its result is denoted $\pi_\cap$ and reads

$$\pi_\cap(x) = \min_{i=1,\ldots,n} (\pi_i(x)) \quad \forall x \tag{13}$$

Arithmetic mean The compromise operator that will be used here is the equivalent of the arithmetic mean used in probability theory (and can thus be compared to it in term of behavior). Its result is denoted π_{mean} and reads

$$\pi_{mean}(x) = \sum_{i=1}^{n} w_i \, (\pi_i(x)) \quad \forall x \tag{14}$$

Disjunctive operator The disjunction operator correspond to the maximum [5] taken over all distributions. Its result is denoted $\pi_\cup$ and reads

$$\pi_\cup(x) = \max_{i=1,\ldots,n} (\pi_i(x)) \quad \forall x \tag{15}$$

Figure 5 illustrate the aggregation of two possibility distributions (sources have equal weights). Let us note that the result of the conjunction and of the arithmetic mean are not forcefully normalized (i.e. there is no value x s.t. $\pi_{res} = 1$). In this last case, the height $h = \max_{x \in [q_l, q_u]}(\pi_{res}(x))$ of the resulting distribution π_{res} can be interpreted as a measure of the conflict between the sources resulting from the aggregation.

Let us note that the methodologies described above are general and can be equally applied to input uncertainty models and to results of uncertainty studies. In the sequel, we concentrate on the second case.

[5] The minimum and maximum are respectively part of a wider family of operators called T-norm and T-conorm.

3 Application to BEMUSE program

Evaluating nuclear power plant performance during transient conditions is a very important issue in thermal-hydraulic research since nuclear energy was used to produce electricity.

During the years, a huge amount of experimental data has been produced from very simple loops and from Integral Test Facilities. A lot of computer codes have also been developed and made available to the nuclear community in order to simulate variables of interest during transient conditions. It is important to evaluate the predicting reliability of such codes by comparing their results to experimental data obtained in small scale facilities. Combining the codes with uncertainty analysis allows for a more realistic modeling of the parameter knowledge, and can thus be helpful to make better predictions. Nevertheless, the final results of such uncertainty analysis can be difficult to compare and to analyze, and the agreement level of such results with experimental data difficult to assess. Hopefully, techniques introduced in section 2 are designed to do such operations, and can thus help the analyst in his task.

To show the usefulness and potential applications of the methodology, we will apply them to the results of the BEMUSE (Best Estimate Methods - Uncertainty and Sensitivity Evaluation) programme (OCDE 2007) performed by the NEA (Nuclear Energy Agency). Our study will focus on the results of the first step of the programme, in which ten participants from nine organisations were brought together in order to compare their respective uncertainty analysis with experimental data coming from the experiment L2-5 performed on the loss-of-fluid test (LOFT) facility. Although most participants (9 out of 10) used similar methodologies to complete their uncertainty evaluation, their results were quite different, due to the fact that different codes were used and that the number, models and physical nature of inputs were different for each participant.

The ten participant of the BEMUSE programme, as well as the code they used and their organization are summarized in table 1. In the first step of BEMUSE programme, the L2-5 experiment has been chosen to apply uncertainty methodologies on a large break loss-of-coolant accident (LB-LOCA transient) performed on an integral test facility.

The L2-5 experiment has been completed on 16 June 1982 in the LOFT facility at INEL (Idaho National Engineering Laboratory). This facility simulated the major components and the system responses of a commercial PWR during a loss-of-coolant accident (LOCA). The core was a semi-scale one with an active height of 1.70m. The experimental assembly included five major subsystems which were instrumented with measurement devices.

Participant	Used code
CEA	CATHARE
GRS	ATHLET
IRSN	CATHARE
KAERI	MARS
KINS	RELAP5
NRI1	RELAP5
NRI2	ATHLET
PSI	TRACE
UNIPI	RELAP5
UPC	RELAP5

Table 1
Participants of BEMUSE programme

The experiment L2-5 itself simulated a guillotine rupture of an inlet pipe in a pressurized water reactor with a true nuclear core. the experiment was initiated (after operating the reactor at 36.0 MW for 40 effective full power hours to build up a fission decay product inventory) by opening two quick-opening blowdown valves upstream a blowdown suppression tank simulating the reactor containment behavior.

As an output of their uncertainty analysis, each participant had to provided lower bounds, reference values and upper bounds for four scalar output parameters as well as the time trends of two output parameters (maximum cladding temperature and upper plenum pressure). For each of these output parameters, experimental values are available (thus, they can be taken as so-called seed variables to assess sources predictive quality). In this paper, we have only considered the four scalar output parameters. These four scalar output parameters are:

(1) The first Peak Cladding Temperature (1PCT) during the blowdown phase
(2) The second peak cladding temperature (2PCT) during the reflood phase
(3) The Time of accumulator injection (T_{inj})
(4) The Time of complete quenching (T_q)

Table 1 summarizes the values given by the participants for the lower bounds, reference calculation and upper bounds for each output. Obtained experimental values are also recalled.

	1PCT (K)			2PCT (K)			T_{inj} (s)			T_q (s)		
	Low	Ref	Up	Low	Ref	Up	Low	Ref	Up	Low	Ref	Up
CEA	919	1107	1255	674	993	1176	14.8	16.2	16.8	30	69.7	98
GRS	969	1058	1107	955	1143	1171	14	15.6	17.6	62.9	80.5	103.3
IRSN	872	1069	1233	805	1014	1152	15.8	16.8	17.3	41.9	50	120
KAERI	759	1040	1217	598	1024	1197	12.7	13.5	16.6	60.9	73.2	100
KINS	626	1063	1097	608	1068	1108	13.1	13.8	13.8	47.7	66.9	100
NRI1	913	1058	1208	845	1012	1167	13.7	14.7	17.7	51.5	66.9	87.5
NRI2	903	1041	1165	628	970	1177	12.8	15.3	17.8	47.4	62.7	82.6
PSI	961	1026	1100	887	972	1014	15.2	15.6	16.2	55.1	78.5	88.4
UNIPI	992	1099	1197	708	944	1118	8.0	16.0	23.5	41.4	62.0	81.5
UPC	1103	1177	1249	989	1157	1222	12	13.5	16.5	56.5	63.5	66.5
Exp. Val.		1062			1077			16.8			64.9	

Table 2
Scalar output values by participants (Exp. Val. : Experimental value)

3.1 Modeling the information

Except for UNIPI, all the participants obtained the lower and upper bound values so that they were respectively lower and larger than the 5% and 95% percentiles with a 95% confidence level (according to order statistics (Conover 1999)). These lower and upper bounds can thus be considered as conservative evaluations of the 5% and 95% percentiles of the unknown probability distributions. Since they are conservative, we have chosen to take them as "fair" evaluations of, respectively, $q_{1\%}$ and $q_{99\%}$.

Given a particular output, let us call $q_{\min}$ and $q_{\max}$ the minimal and maximal values of the lower and upper bounds of this output, taken over all participants. Then, for each output, we take as $[q_l, q_u]$ the interval $[q_{\min}, q_{\max}]$ increased by 2% (e.g. for 1PCT, $q_{\min} = 626$ (KINS), $q_{\max} = 1255$ (CEA) and $[q_l, q_u] = [620, 1261]$).

According to this information, we take the following models:

Probabilistic model: Since the reference values Ref are often close to the middle of interval $[Low, Up]$, and as nominal values are often associated to the median of the distribution, we have chosen to take, for each participant and output, the following distribution : $(q_{0\%}, q_{1\%}, q_{50\%}, q_{99\%}, q_{100\%}) = (q_l, Low, Ref, Up, q_u)$. For example, the distribution corresponding to the information given by NRI1 for the 2PCT is ($q_{0\%} = 592, q_{1\%} = 845, q_{50\%} =$

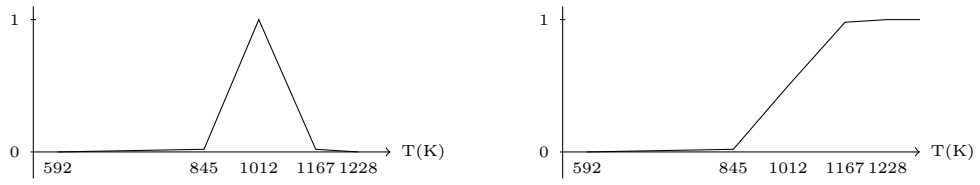

Figure 6. Probability and possibility dist. of NRI1 for the 2PCT

$1012, q_{99\%} = 1167, q_{100\%} = 1228$). The only exception to this rule is the distribution of KINS for T_{inj}, since concentrating 50% of the probability mass on a single value would have no sense. Thus, the distribution of KINS for T_{inj} is $(q_{0\%} = 7.8, q_{1\%} = 13.1, q_{99\%} = 13.8, q_{100\%} = 23.7)$.

Possibilistic model: The interval $[q_l, q_u]$ common to each source is considered as containing with certainty the true unknown value. By giving interval $[Low, Up]$, each source provide a 98% confidence interval, while it is natural to consider the nominal value Ref as the most plausible one. For each source, the possibility distribution that fits this information is s.t. $\pi(q_l) = 0, \pi(Low) = 0.02, \pi(Ref) = 1, \pi(Upp) = 0.02, \pi(q_u) = 0$ (with linear interpolation between each points). When taken as an imprecise probabilistic model, this possibility distribution dominate the probabilistic model (see (Baudrit and Dubois 2006)).

Figure 6 illustrates both models built from the information of NRI2 concerning the second PCT.

3.2 Evaluating the sources

For the evaluation steps, the four scalar parameters were considered as seed variables, as their experimental values are known. Evaluation was then performed according to the methodology described in section 2.2, with the uncertainty models given above. Table summarizes the obtained informativeness, calibration and global scores for both approaches.

Many interesting comments can be made about these results, both from methodological and application standpoints.

3.2.1 Comments on methodology

Global agreement: although there are a few noticeable differences between the results of the two approaches, they globally agree. Indeed, between the first five participants given by the two approaches, four are in common (i.e. IRSN, KINS, NRI1, UNIPI), while the same statement is of course true for

	Prob. approach			Poss. approach		
	Inf.	Cal.	Global	Inf.	Cal.	Global
CEA	8 (0.77)	5 (0.16)	6 (0.12)	8 (0.71)	6 (0.55)	7 (0.40)
GRS	4 (1.23)	1 (0.98)	1 (1.21)	3 (0.84)	7 (0.52)	6 (0.44)
IRSN	5 (0.98)	2 (0.75)	2 (0.73)	6 (0.73)	1 (0.83)	1 (0.60)
KAERI	9 (0.68)	5 (0.16)	7 (0.11)	9 (0.70)	8 (0.48)	8 (0.34)
KINS	3 (1.29)	5 (0.16)	5 (0.21)	7 (0.72)	3 (0.67)	3 (0.49)
NRI1	7 (0.79)	2 (0.75)	3 (0.59)	5 (0.75)	5 (0.63)	4 (0.47)
NRI2	6 (0.79)	8 (0.13)	8 (0.10)	4 (0.78)	2 (0.72)	2 (0.56)
PSI	1 (1.6)	10 (0.004)	10 (0.008)	1 (0.88)	10 (0.25)	10 (0.22)
UNIPI	10 (0.53)	2 (0.75)	4 (0.4)	10 (0.69)	4 (0.67)	5 (0.46)
UPC	2 (1.44)	9 (0.02)	9 (0.025)	2 (0.87)	9 (0.28)	9 (0.24)

Table 3
Results of sources evaluation (Inf.: informativeness ; Cal.: Calibration) by ranks (values)

the last five (i.e. CEA, KAERI, PSI and UPC are common to both rankings). This is not surprising, since even if the models and formulas used by each approach are different, the conceptual methodology is the same for both (i.e. comparing information to non-informative state for informativeness and to known experimental values for calibration).

Preference given to well calibrated sources: a good source is a source which is both informative and well calibrated, or in other words which is useful and reliable. Nevertheless, these two goals are somewhat contradictory, since in general, the more precise a source is, the more chance it has to be wrong, and inversely. In our case, this is well exemplified by UNIPI (poorly informative, well calibrated), PSI and UPC (highly informatives, poorly calibrated). From the results, we see that both approaches tend to privilege well calibrated sources rather than informative sources (for instance, UNIPI has an high rank, even if it the most imprecise, while both UNIPI and UPC have low ranks, even if they are the most informative). This is a good thing, since it is preferable to have reliable information rather than very precise, but wrong information.

Divergences in informativeness scores: In both approaches, informativeness ranking agree together, except for KINS (high and low rank, respectively for the probabilistic and possibilistic approach). This is due to the fact that, although KINS have rather wide and imprecise intervals $[Low, Up]$, the reference value is often very close to one of these two values, thus the corresponding distributions are more dissymmetric than for the other participants. The probabilistic approach tend to focus on this dissymmetry, while the wide span of

$[Low, Up]$ is dominating in the possibilistic approach.

Divergences in calibration scores: although both rankings agree less in the case of calibration scores, we have noticed above that, except for GRS and NRI2, the same sources have high (low) ranks in both approaches. In particular, UPC and PSI are the lowest ranked in each approach, being the only two participants for which two experimental values were outside intervals $[Low, Up]$. The differences, mainly noticeable in the scores of GRS and NRI2, comes from the fact that the formulas used for computing calibration are based on two different notions:

- for the probabilistic approach, the formula assumes that, for a well calibrated source, the experimental distribution should converge to the distribution given by the source. This experimental distribution is updated according to the interpercentiles in which experimental values fall for each seed variable, without taking into account the metric (i.e. where the experimental value actually is in the interprecentile, if it is closer to Ref, to Upp, in the center of the interval, ...). Probabilistic approach assumes that, if the source is well calibrated, the true value will fall $p_1\%$ of the time in the first interpercentiles, $p_2\%$ in the second, etc.
- the possibilistic approach does not assume any kind of convergence between the given distributions and an experimentally built distribution. It rather considers, for each seed variable, how far is the observed value from the nominal value given by the source. It is thus based on metric considerations and not on convergence.

Indeed, if we look at the information given by NRI2 for the four seed variables, each realizations falls into the interval $[Ref, Upp]$ (while being close to Ref), while for GRS, they are evenly divided between $[Low, Ref]$ and $[Ref, Upp]$. Thus, the experimental distributions for NRI2 and GRS are respectively $r^{NRI2} = (0, 0, 1.0, 0)$ and $r^{GRS} = (0, 0.5, 0.5, 0)$, while the source distributions are $p = (0.01, 0.49, 0.49, 0.01)$, which explains why GRS has the highest ranking in the probabilistic approach, and NRI2 one of the lowest. In the possibilistic approach, the high rank of NRI2 comes from the facts that experimental values are close to the given Ref values and that intervals $[Low, Upp]$ are cautious, while remaining no too wide (as shows the informativeness score of NRI2). GRS score is mainly penalized by the narrower (and, hence, more adventurous) intervals $[Low, Upp]$ as well as by the fact that experimental values for T_{Inj} and T_q are respectively close to the given Upp and Low values. Claiming that one approach is always better than another makes poor sense, but since we consider here imprecisely known non-random values for which we have few information, the possibilistic approach seems more fitted to the problem at hand.

The need of seed variables: the recommended number of seed variables

for the probabilistic approach is around ten. This need mainly comes from the fact that the probabilistic approach supposes the convergence between the experimental distribution and the distribution given by a well-calibrated source. In the application, the poor discriminating power of calibration scores (two groups of three participants have equivalent scores) shows some of the defect of taking fewer seed variables. Let us note that the possibilistic approach is meaningful even if only one seed variable is available (of course, the more seed variable we have, the more stable are the score of the sources)

3.2.2 Comments on results

Ranking with respect to the used code: we can observe that the ranking of the participants is poorly correlated with the particular code used to achieve the computations. For instance, GRS and NRI2, as well as IRSN and CEA, have quite different rankings in both approaches, even if they use the same computation code (the same is true for the four participants using RELAP5, whose rankings range from 3 to 9). This indicates that, more than the peculiar used code, it is how it is used that mainly matters.

Coherence with informal observations: in (OCDE 2007), it was observed that only UPC and PSI bounds did not envelop the PCT experimental values (respectively for the first and second PCT), one of the reason being given to explain this was the very narrow uncertainty band considered. This can be found back in the results of table 3, where for both approaches, UPC and PSI get the worst rankings with respect to calibration (exp. values out of bounds) and the best ones with respect to informativeness (narrow uncertainty bands).

Calibration scores of GRS and NRI2: as said above, the two approaches strongly disagree on the calibration scores of GRS and NRI2. It could be argued that, for each approach, the respective low rank of each participant is unfair, as both participants gives globally satisfying information. Nevertheless, this disagreement has some explanation : the possibilistic approach indicates that GRS poorly evaluate T_{Inj} and T_q, while the probabilistic one point out the tendency of NRI2 nominal values to underestimate the experimental results.

Code evaluation: As the various scores are significant by themselves, results can be used to assess the quality of a particular code (managed by a particular user). Calibration indicates how well computed results are in accordance with experimental values, while informativeness gives us indications as to how precise it is. By using only formal mathematics, experimental data and provided information, these methods try to be as objective and as sound as possible in their rankings. Both have clear interpretations and are simple to understand, two features that we consider as advantages.

Figure 7. Application of probabilistic aggregation

3.3 Synthesizing the sources

This section applies the aggregation operators introduced in section 2.3 to the second PCT. Interests and defects of each operator are illustrated, as well as how they cna help to analyze the information and the relations between sources (i.e. the participants).

3.3.1 Probabilistic aggregation

Figure 7 shows the result of aggregating the probability distribution of the various participants. Each arithmetic mean are used with the associated weights, except when specified so on the figure (i.e. all sources with equal weights).

As we see, grouping participants by used codes (left figure) gives poorly calibrated results. CATHARE and RELAP5 users tend to underestimate the experimental value, while ATHLET users tend to overestimate it. Few can be said about the agreement of each code users.

The right figure shows how the scores given to each participant can be used to improve the aggregated distribution, both in term of precision and of quality. Interestingly enough, the best distributions are the one in which all sources are taken into account with their associated scores, and the one considering the four common participants being in the five best scored sources of each approach. Both these two distributions are slightly narrower and more centered around the experimental value than the two others. This shows that using the scores in the aggregation is useful and that the two approaches can help each other in the selection of the best sources. Here again, an eventual conflict between sources is hardly visible. The fact that the arithmetic mean tends to average the resulting distribution is shown in the right figure, in which resulting distributions, although different, remain close to each others. Indeed, we can see here that taking the average is not very discriminative,especially in our case where information given by sources are

3.3.2 Possibilistic aggregation

Figure 8 shows the result of applying the disjunctive operator (i.e. maximum) and the usual compromise operator (i.e. weighted arithmetic mean) to the set of all sources (taking smaller sets of sources do not bring any really useful extra information in these two cases).

These figures well illustrate what was explained in section 2.3.2 : the result

Figure 8. Application of possibilistic aggregation : disjunction (left) and weighted mean (right)

of the disjunction is quite imprecise and the arithmetic mean averages the contribution of all participants, resulting in a smooth distribution which has a peak around 1000 K.

Some interesting and surprising facts can be drawn from these distributions. The fact that, in the distribution resulting from the disjunction, more peaks are below rather than above the experimental value indicates that most sources tends to underestimate it. This is confirmed by the distribution resulting from the arithmetic mean, which peak is slightly below the experimental value.

The relatively high level (~ 0.75) of the left figure shows that the participants are globally agreeing with each other, since the average conflict is low (~ 0.25).

A more surprising characteristic is the "gap" around the experimental value that exhibits the distribution resulting from the disjunction. Indeed, the possibility degree of the experimental value is "only" around 0.8. This gap comes from the fact that the reference value of most participant is not very close from experimental data (this is not the case for the first PCT) and that KINS, which reference value is the closest to the experimental value, also gives a very low upper bound (in fact, the lowest outside of PSI, which is the only participant having an upper bound lower than the experimental value).

Figure 9 shows the result of applying the conjunctive operator (i.e. minimum) to various subgroups of participants. A first remark is that the eventual conflict among each subgroup is here directly visible. For instance, we see that, concerning the second PCT, the information given by both users of CATHARE code are coherent, while the information given by ATHLET users are more conflicting. The higher conflict shown by RELAP5 users is not surprising, since they are more numerous.

The right figure shows that the information given by all sources concerning the second PCT is highly conflicting (conflict ~ 0.9), and thus that the resulting conjunction, although very precise, is judged to be highly unreliable (i.e. the true value has an high chance to be outside it). Inversely, limiting ourselves to the most highly scored participants (either only by possibilistic approach or by both approaches) results in distributions that are reliable (conflict only ~ 0.2). We see that using conjunction with only the most reliable sources results in a distribution well balanced between precision and reliability.

We can also notice that the distribution resulting from the conjunction of ATHLET users distributions is very precise and exhibit a peak very close to the value experimental, and thus that one should take this distribution instead

Figure 9. Application of possibilistic aggregation : conjunction (minimum)

of any of the others shown in figure 9. However, we must not forget that this is because we know the exact value of the pick clad temperature . In practice, building synthetic and reliable models concerns values and variables for which we have scarce or no information. Had we not known the true value, we would have given relatively poor credit to the distribution, due to its relatively low reliability (indeed, the peak of the distribution is induced by the reference values of GRS and NRI2, which are respectively quite larger and lower than the true value!) .

4 Conclusion

Evaluating, synthesizing and analyzing information coming from the result of uncertainty studies performed by computer codes is often a tedious and difficult task. Both probability and possibility theory offer formal methods to do so that are both well-founded and simple to use.

After a brief description, we have applied them to the result of the BEMUSE project to show how they can be practically used. We have shown that they can be useful to build synthetic representation of variables of interest or to analyze the available information. In particular, it has been shown that the scores obtained for each participant was in agreement with informal observations (thus formalizing them) and that their use in the aggregation procedure could help to improve the final result of the aggregation. Similarities as well as dissimilarities between the methods have been underlined. Searching the reasons of the dissimilarities roused some interesting questions and comments as to why some participants had different scores.

It was also shown that using probabilistic modeling and the arithmetic weighted mean to aggregate distributions generally gives a final distribution that is more precise and accurate. Nevertheless, it was also shown that being limited to the arithmetic weighted mean to aggregate distributions also limits the analysis of the available information.

Inversely, even by using the most basic aggregation operators of possibility theory, we were able to derive interesting conclusions about the conflict between sources or about their global evaluations. Moreover, one can directly measure the reliability of a final possibility distribution (whereas such an information is hard to extract from a probability distribution), and thus to recommend it or not as a final synthetic representation of the variable of interest.

On an applicative side, presented methods can help to improve the assess-

ments of computer codes by extracting useful information concerning them. Perspectives are thus to work with experts to deepen the analysis in the most interesting directions. Moreover, presented methods can also be applied to input uncertainty modeling (a critical step that often requires more formalization, as underlined in (OCDE 2007)). Let us also note that the sources involved in the evaluation, aggregation and analysis processes can have heterogenous nature (e.g. experts, experimental data, computer codes, ...) as long as the information can be meaningfully modeled in a common framework.

On a methodological side, it is desirable to develop methods that allow for a finer analysis and treatment of conflicting information. Indeed, treating the conflict only by mean of conjunction, disjunction or arithmetic mean can lead to frustrating results, even in the application considered here, where evaluations given by all sources are quite similar. More complex aggregation methods already exist (see, for example (Delmotte 2007,Dubois and Prade 2001)), but most of them are designed to directly build a synthetic final distribution, without any intention of analyzing available information. This is why IRSN, in association with University Paul Sabatier, develops new aggregation methods designed both to analyze existing information and to build synthetic representations (Destercke et al. 2007). These methods have been integrated to the IRSN software for uncertainty analysis (SUNSET), with which all computations present in this paper were performed.

Bibliography

Baudrit, C., and D. Dubois. 2006. Practical representations of incomplete probabilistic knowledge. *Computational Statistics and Data Analysis* 51(1):86–108.

Bedford, T., and R. Cooke. 2001. *Probabilistic Risk Analysis. Foundations and Methods.* UK: Cambridge University Press.

Bloch, I. 1996. Information combination operators for data fusion : A comparative review with classification. *IEEE Trans. on Syst., Man, and Cybern. A* 26(1):52–67.

Clemen, R., and R. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis* 19(2):187–203.

Conover, W. 1999. *Practical non-parametric statistic.* New York: Wiley. 3rd edition.

Cooke, R. 1991. *Experts in uncertainty.* Oxford, UK: Oxford University Press.

Delmotte, F. 2007. Detection of defective sources in the setting of possibility theory. *Fuzzy Sets and Systems* 158:555–571.

Destercke, S., D. Dubois, and E. Chojnacki. 2007. Possibilistic information fusion using maximal coherent subsets. In *Proc. IEEE Int. Conf. On Fuzzy Systems (FUZZ'IEEE).*

Dubois, D., L. Foulloy, G. Mauris, and H. Prade. 2004. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing* 10:273–297.
Dubois, D., and H. Prade. 1988. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press.
Dubois, D., and H. Prade. 1992. When upper probabilities are possibility measures. *Fuzzy Sets and Systems* 49:65–74.
Dubois, D., and H. Prade. 2001. Possibility theory in information fusion. In G. D. Riccia, H. Lenz, and R. Kruse (Eds.), *Data fusion and Perception*, Vol. CISM Courses and Lectures N 431, 53–76. Berlin: Springer Verlag.
Ferson, S., and L. R. Ginzburg. 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety* 54:133–144.
Kraan, B. C. P. 2002. *Probabilistic Inversion in Uncertainty Analysis and Related Topics*. PhD thesis, Delft Institute of Applied Mathematics.
McConway, K. 1981. Marginalization and linear opinion pools. *Journal of the American Statistical Association* 76(374):410–414.
OCDE. 2007. Bemuse phase iii report : Uncertainty and sensitivity analysis of the loft l2-5 test. Technical Report NEA/NCIS/R(2007)4, NEA, May.
Oussalah, M., H. Maaref, and C. Barret. 2003. From adaptative to progressive combination of possibility distributions. *Fuzzy sets and systems* 139:559–582.
Sandri, S., D. Dubois, and H. Kalfsbeek. 1995. Elicitation, assessment and pooling of expert judgments using possibility theory. *IEEE Trans. on Fuzzy Systems* 3(3):313–335.
Walley, P. 1991. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall.
Zadeh, L. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems* 1:3–28.