A new framework for multi-hazards risk aggregation

Tasneem Bani-Mustafa, Zhiguo Zeng, Enrico Zio, Dominique Vasseur

A new framework for multi-hazards risk aggregation

Tasneem Bani-Mustafa [(1)], Zhiguo Zeng [(1)], Enrico Zio [(2) (3) (4)], Dominique Vasseur [(5)]

[(1)] *Chair on System Science and the Energetic Challenge, EDF Foundation*
Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay,
3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France

[(2)] *MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France*

[(3)] *Energy Department, Politecnico di Milano, Via Giuseppe La Masa 34, Milan, 20156, Italy*

[(4)] *Eminent Scholar, Department of Nuclear Engineering, College of Engineering, Kyung Hee University, Republic of Korea*

[(5)] *EDF R&D, PERICLES (Performance et prévention des Risques Industriels du parc par la simulation et Les Etudes) EDF Lab Paris Saclay - 7 Bd Gaspard Monge, 91120 Palaiseau, France*

Abstract

In this paper, we develop a new method for Multi-Hazards Risk Aggregation (MHRA). A hierarchical framework is first developed for evaluating the trustworthiness of the risk assessment. The evaluation is based on two main attributes (criteria), i.e., the strength of knowledge supporting the assessment and the fidelity of the risk assessment model. These two attributes are further broken down into sub-attributes and, finally, leaf attributes. The trustworthiness is calculated using a weighted average of the leaf attributes, in which the weights are calculated using the Dempster Shafer Theory-Analytical Hierarchy Process (DST-AHP). Risk aggregation is, then, performed by a "weighted posterior" method, considering the level of trustworthiness. An application to the risk aggregation of two hazard groups in Nuclear Power Plants (NPP) is illustrated.

Keywords

Quantitative Risk Assessment (QRA), Risk-Informed Decision Making, Trustworthiness in Risk Assessment, Multi-Hazards Risk Aggregation (MHRA), Strength of Knowledge (SoK), Nuclear Power Plants (NPP)

Acronyms

AHP: Analytical Hierarchy Process
DST: Dempster Shafer Theory
DM: Decision Making
DST-AHP: Dempster Shafer Theory-Analytical Hierarchy Process
EDF: Electricité De France
EFWS: Emergency Feedwater System
IAEA: International Atomic Energy Agency
SoK: Strength of Knowledge
MHRA: Multi-Hazards Risk Aggregation
NPP: Nuclear Power Plants
PRA: Probabilistic Risk Assessment
RIDM: Risk-Informed Decision-Making
USNRC: United-States Nuclear Regulatory Commission

1. Introduction

In Risk-Informed Decision-Making (RIDM), risk metrics are first calculated through Multi-Hazards Risk Aggregation (MHRA) by combining all the relevant information on risk from different contributors (hazard groups) and, then, used to support Decision-Making (DM) (EPRI, 2015). A fundamental criticism of the current practice is that the aggregation is conducted by a simple arithmetic summation of the risk metrics from different hazard groups, without considering the heterogeneity in the degrees of maturity and realism of the risk analysis for each hazard group (EPRI, 2015). For example, in Nuclear Power Plants (NPP), the Probabilistic Risk Assessment (PRA) for internal events has been developed for many years and considered relatively mature compared to external events (EPRI, 2015) or to fire (Siu *et al.*, 2015). Simply adding up the risk indexes can be misleading because it does not consider any information on the trust in the risk indexes calculated for each hazard group. This is a real problem as the results of the PRAs to be aggregated often involve different hazard groups with different levels of realism and trustworthiness.

Various factors contributing to the trustworthiness of risk analysis have been discussed in the literature, including the strength of background knowledge, conservatism, plausibility and realism of assumptions, uncertainty, level of sophistication and details in the analysis, value-ladenness of the assessors, experience, number of approximations and assumptions made in the analysis, etc. (EPRI, 2012), (EPRI, 2015). Communicating these factors to the decision maker can better inform decision making (Flage and Aven, 2009), (EPRI, 2012), (Aven, 2013b), (EPRI, 2015), (Veland and Aven, 2015). For this, some experts propose a broad representation of risk that highlights uncertainties rather than probability (Flage and Aven, 2009), (Aven, 2013b), (Aven and Krohn, 2014). In Aven (2013a), the risk is described in terms of events, consequences, uncertainty (A, C, U) and a structure is presented for linking the elements of a Data-Information-Knowledge-Wisdom hierarchy to this perspective. In (Flage and Aven, 2009), the authors apply the concept of uncertainty as the main component of risk, whereas the probability is regarded as an epistemic-based expression of uncertainty. Their argument is that for decision making purposes, a broad and comprehensive representation of risk is required to cover the events, consequences, predictions, uncertainty, probability, sensitivity, and knowledge. In addition, they propose a simple and practical method to classify uncertainty factors and evaluate the background knowledge given the following criteria: the inter-alia assumptions and presuppositions (solidity of assumptions), historical field data (availability of reliable data), understanding of phenomena, and agreement among experts.

Some attempts are also found in the literature that focus on treating the uncertain assumptions as an implication

of new risk perspectives. Aven (2013b) proposed a method for assessing the assumption deviation risk by three elements: (i) the degree of the expected deviation of the assumption from reality and its consequences (ii) a measure of uncertainty of the deviation and consequences; (iii) the knowledge on which the assumptions are based. Berner and Flage (2016) summarize four approaches for treating uncertain assumptions: (i) law of total expectation; (ii) interval probability; (iii) crude strength of knowledge and sensitivity categorization; (iv) assumption deviation risk. In this work, they extend the method in Berner and Flage (2015) that evaluates the assumption deviation risk based on three criteria: belief in the deviation from the assumption, sensitivity of the risk index and its dependency on the assumption, and SoK on which the assumptions are made. Six settings are identified for the corresponding scenarios resulting given the three criteria. Guidance for treating the uncertainty related to the deviation of assumptions is given for each setting. Finally, an application of Numeral Unit Spread Assessment Pedigree (NUSAP) is proposed for analyzing the strength, importance, and potential value-ladenness of assumptions through a pedigree diagram (Van Der Sluijs *et al.*, 2005), (Boone *et al.*, 2010), (Kloprogge *et al.,* 2011), (De Jong *et al.,* 2012). The pedigree diagram uses seven criteria for evaluating the quality of assumptions: (i) plausibility; (ii) inter-subjectivity peers; (iii) inter-subjectivity stakeholders; (iv) choice space; (v) influence of situational limitations; (vi) sensitivity to view and interests of the analyst (vii) and influence on results (Van Der Sluijs *et al.*, 2005), (Boone *et al.*, 2010), (Kloprogge *et al.,* 2011), (De Jong *et al.,* 2012).

In addition, some attempts are found in the literature for directly evaluating the trustworthiness and other relevant quantities. In Bani-Mustafa *et al.* (2017), the trustworthiness of risk assessment models is evaluated through a hierarchical tree that covers the different factors including modeling fidelity, SoK, number of approximations, amount and quality of data, quality of assumptions, number of model parameters, etc. Trustworthiness is also measured in the literature in terms of maturity and credibility. For example, in Model and Simulation (M&S) and information system, a capability maturity model is used to assess the maturity of a software development process in the light of its quality, reliability, and trustworthiness (Paulk *et al.*, 1993). A predictive capability maturity model has been developed to assess the maturity of M&S efforts through evaluating the representation and geometric fidelity, physics and material model fidelity, code and solution verification, model validation and uncertainty quantification, and sensitivity analysis (Oberkampf *et al.*, 2007). In (Zeng *et al.*, 2016), a hierarchical framework has been developed to assess the maturity and prediction capability of a prognostic method for maintenance decision making purposes. The hierarchical tree covers different attributes that are believed to affect the prediction capability of prognostic methods and the trustworthiness of the results. In (Nasa, 2013), a framework is proposed for assessing the credibility of M&S

through eight criteria: (i) verification; (ii) validation; (iii) input pedigree; (iv) results uncertainty (v) results robustness; (vi) use history; (vii) M&S management; (viii) people qualification. In (Bani-Mustafa *et al.*, 2017), the trust of the model is evaluated based on the level of maturity of the risk assessment model through four main criteria: (i) uncertainty; (ii) knowledge; (iii) conservatism; (iv) sensitivity. Also, the quality of M&S is assured by the American Society of Mechanical Engineers (ASME) through verification and validation (Schwer, 2009). Verification is concerned with evaluating the accuracy of the computational model in representing the conceptual and mathematical model, and validation is concerned with evaluating the accuracy of the model in representing reality (Schwer, 2009).

As seen from the discussions above, there are a number of works concerned with the realism and trustworthiness of risk assessment. These works, however, discuss the contributors to trustworthiness separately: different frameworks cover different aspects of the trustworthiness based on different terminologies. A unified and complete framework that covers all the factors contributing to trustworthiness is lacking. Besides, the current state of the art only focuses on the evaluation of trustworthiness but does not consider how to integrate the trustworthiness into the results of risk assessment, neither does it show how to aggregate the risk of different contributors with different levels of trustworthiness.

In this work, we define the trustworthiness of risk assessment as a metric that reflects the degree of confidence in the background knowledge that supports the PRA, as well as in the suitability, comprehensiveness and completeness of the PRA model formulation and implementation in a way that reflects, to the best possible, reality. With this, the objective is, then, to provide a new approach for MHRA considering trustworthiness. Compared to the existing works, the contributions of the current work include:

(i) a unified framework is developed for the evaluation of trustworthiness in risk assessment;

(ii) a method is developed to integrate the trustworthiness in the result of the risk assessment of a single hazard group;

(iii) an approach is developed for MHRA considering the trustworthiness of risk assessment.

The rest of this paper is organized as follows. In Section 2, we present a hierarchical framework for assessing the trustworthiness of PRA models and in Section 3 we show how to apply it in practice. In Section 4, we show how to aggregate the risks considering trustworthiness. Section 5 applies the developed methods to a case study from the nuclear industry. Finally, in Section 6, we conclude this paper and discuss the potential future work.

2. A hierarchical framework for assessing the trustworthiness of a risk model

As illustrated previously, various factors have been discussed in the literature in relation to the trustworthiness

of risk assessment. In this paper, we only focus on some of the most relevant factors. For example conservatism, uncertainty, level of sophistication and details in the analysis, experience, number of approximations and assumptions made in the analysis are identified in (EPRI, 2012) and (EPRI, 2015) as fundamental factors that influence the realism and trustworthiness of a risk analysis. Background knowledge that supports the risk assessment is also widely accepted as an essential contributor to the trustworthiness (Flage and Aven, 2009), (Aven, 2013a), (Aven, 2013b), (EPRI, 2012), (EPRI, 2015), (Bani-Mustafa *et al.,* 2018). The assumptions that are inevitably made because of incomplete knowledge or for simplifying the analysis (Kloprogge *et al.,* 2011) are considered crucial for the suitability of risk representation and hence, the trustworthiness of its analysis (Boone *et al.*, 2010), (Kloprogge *et al.,* 2011), (De Jong *et al.,* 2012), (Berner and Flage, 2016). The conservatism is also identified as a pivotal contributor to the realism, maturity, and trustworthiness of risk assessment (Aven, 2016), (Bani-Mustafa *et al.*, 2017). Sensitivity analysis is also needed for a comprehensive description of risk (Flage and Aven, 2009), (Bani-Mustafa *et al.*, 2017). Other factors for evaluating the credibility of M&S include verification, validation, input pedigree, result uncertainty, result robustness, use history, M&S management and people qualification (Nasa, 2013).

The factors mentioned above are included in the trustworthiness assessment framework proposed in this paper. Other relevant factors are also considered, for a complete representation of trustworthiness. The trustworthiness of risk assessment is defined in this paper as the degree of confidence that the background knowledge is strong enough to support the PRA and that the PRA model is suitable, correctly and robustly made to make the best use of the available knowledge in order to reflect to the best, reality. Obviously, the background knowledge that supports a risk assessment affects significantly the trustworthiness of its results (Flage and Aven, 2009), (Aven, 2013a), (Aven, 2013b), (Bani-Mustafa *et al.,* 2018). However, having a strong background knowledge is not sufficient to ensure the trustworthiness in the results: the fidelity of the modeling should be also verified. This gives rise to the need of a technically adequate and mature model that is known for its high quality and representativeness of reality (Oberkampf *et al.*, 2007), (Nasa, 2013), (Zeng *et al.*, 2016). In addition, the modeling process should follow a high quality and thorough application procedure, in order to have trustworthy risk analysis results (IAEA, 2006), (Oberkampf *et al.*, 2007), (Schwer, 2009), (Nasa, 2013), (Zeng *et al.*, 2016). Hence, the suitability of the selected model and the quality of its application are considered as relevant attributes in the proposed framework. In fact, since the risk metrics are calculated as a result of modeling and simulation, it is intuitive to understand that the trustworthiness of the risk assessment results can be affected by: the suitability of the selected model, the comprehensiveness and correctness of the application of the model, as well as the background knowledge that supports the modeling and analysis. Besides,

having results that are highly sensitive to changes in the input is an indication that the assessment is less trustworthy, as the results might be dramatically affected by even a small change in the input parameters and assumptions (Flage and Aven, 2009), (Bani-Mustafa *et al.*, 2017). Accordingly, the robustness of the results is regarded as another factor that affects the trustworthiness of risk analysis. In this framework, we use the acronym SoK to represent the strength of the background knowledge that supports the risk assessment and the term "modeling fidelity" to represent the suitability of the selected model, the quality of its application and the robustness of the results, as shown in Figure 1. These two top-level attributes are further decomposed into more tangible sub-attributes.

It should be noted that in general, knowledge includes explicit knowledge, which can be documented and transferred directly, and implicit knowledge, which is possessed by individuals and cannot be documented or transferred directly. The SoK defined in Figure 1 only concerns the explicit knowledge, whereas implicit knowledge is mostly related to the construction and application of the model. Hence, implicit knowledge is viewed as related to the modeling fidelity. The background knowledge is evaluated in Flage and Aven (2009) considering: (i) availability of reliable data; (ii) phenomenological understanding; (iii) quality and plausibility of assumptions; (iv) agreement among peers. In Bani-Mustafa *et al.* (2018), the background knowledge is evaluated by (i) the solidity of assumptions; (ii) the availability of reliable data; (iii) the understanding of phenomena. Each attribute is further broken down into more tangible sub-attributes that define it. For example, the reliability of data is evaluated by its completeness, consistency, validity, accuracy, and timeliness (Bani-Mustafa *et al.,* 2018).

The quality of assumption is evaluated in the literature by different factors. For example, in an application of Numeral Unit Spread Assessment Pedigree (NUSAP), the quality of assumptions is evaluated by (i) plausibility; (ii) inter-subjectivity peers; (iii) inter-subjectivity stakeholders; (iv) choice space; (v) influence situational limitations; (vi) sensitivity to view and interests of the analyst (vii) and influence on results (Van Der Sluijs *et al.*, 2005), (Boone *et al.*, 2010), (Kloprogge *et al.,* 2011). In this paper, we group these factors into three main categories (Bani-Mustafa *et al., 2018*): (i) quality of assumptions; (ii) value-ladenness; (iii) sensitivity. Value ladenness is, in turn, considered as an independent variable that affects the quality of the assumptions and is evaluated using seven main criteria (i) the personal knowledge; (ii) the sources of information; (iii) the non-biasedness; (iv) the relative independence; (v) the past experience; (vi) the performance measure; (vii) the agreement among peers (Zio, 1996), (Bani-Mustafa *et al.,* 2018).

Nevertheless, some of the SoK attributes are more related to the implicit knowledge and affect the construction and formulation of the modeling process and, hence, they are considered under modeling fidelity and not under SoK.

For example, the quality and solidity of assumptions are more related to modeling fidelity, since they affect the formulation of the model. Also, since assumptions are made by experts and inevitably affected by their subjectivity, agreement among peers is considered as a sub-attribute under solidity of assumptions.

In this paper, only the availability of reliable data and phenomenological understanding from (Flage and Aven, 2009) are considered for evaluating the SoK. As said earlier, the quality and solidity of assumptions are treated under modeling fidelity. Finally, we add another attribute to cover the data and information related directly to the known hazards. The known potential hazards attributes are next broken down into three sub-attributes that cover: the number of documented known hazards, the accident analysis report and the expert's knowledge about the hazards. The data and phenomenological understanding attributes are further broken into sub-attributes and leaf attributes (illustrated in Figure 1) according to the framework proposed in (Bani-Mustafa *et al.,* 2018).

Other factors related to the suitability of the model and quality of application are also found in the literature. Examples of these factors are: conservatism, level of sophistication and details in the analysis, experience, number of approximations and assumptions made in the analysis, sensitivity, results robustness, use history, level of details and verification (Paté-Cornell, 1996), (Flage and Aven, 2009), (EPRI, 2012), (Nasa, 2013), (EPRI, 2015), (Aven, 2016), (Bani-Mustafa *et al.*, 2017). These attributes are allocated in the hierarchy according to their relevance to the modeling fidelity and categorized in three groups, i.e., suitability of selected model, quality of the application and robustness of the results, whereas other attributes have been added to complement the overall framework for the trustworthiness of the risk assessment. The overall hierarchical framework is presented in Figure 1, and detailed definitions of the attributes, sub-attributes and “leaf” attributes are given in Table 1-4.

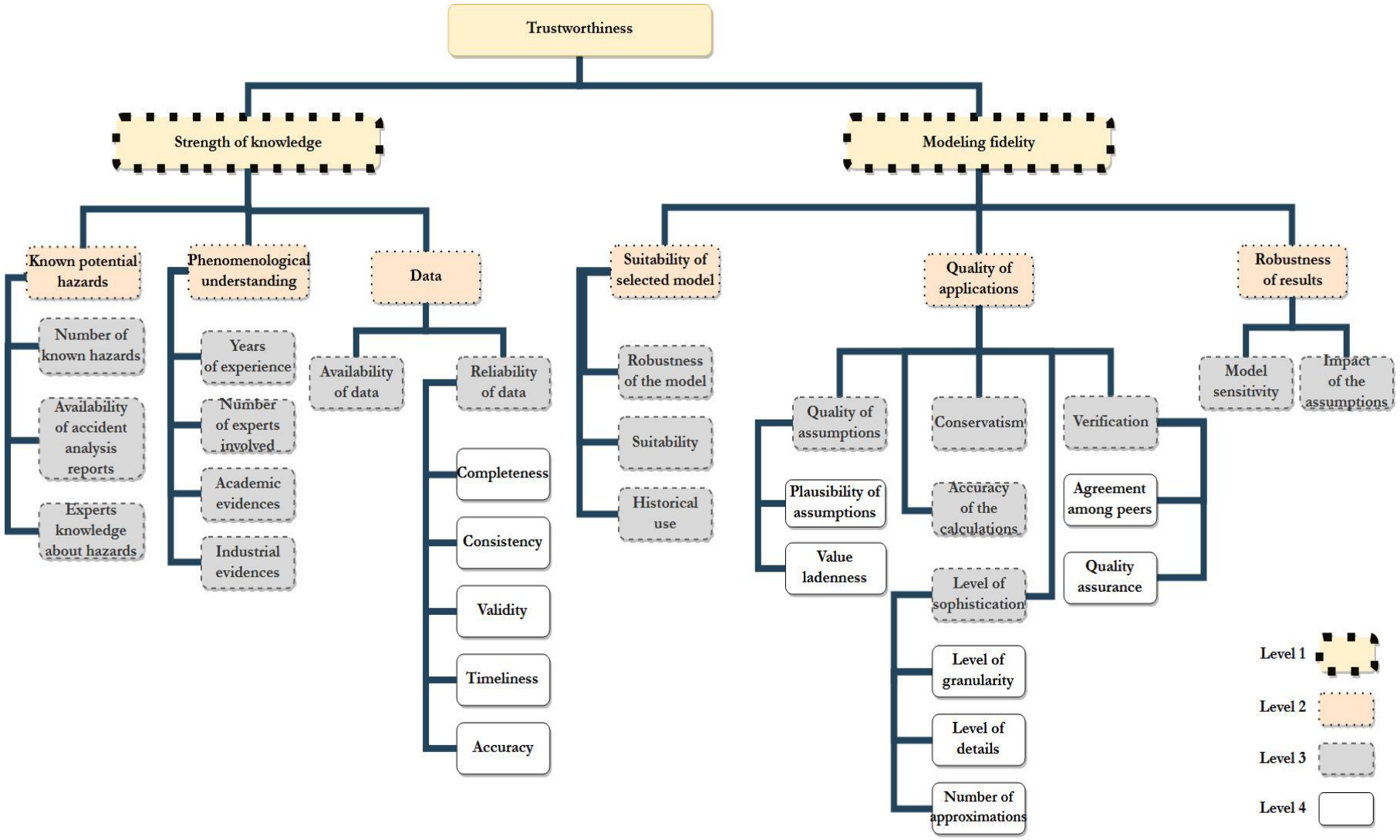

Figure 1 Hierarchical tree for trustworthiness evaluation

1

Table 1 Definition of trustworthiness attributes (Level 1)

Attribute	**Definition**
Modeling fidelity ($MF = T_1$)	The degree of confidence that the selected PRA model is technically adequate for describing the problem of interest and that the model is implemented in a trustable way so that the results can reasonably represent reality, relative to the decision making involved
The strength of knowledge ($SoK = T_2$)	The amount of high-quality explicit knowledge that is available to support the PRA

Table 2 Definition of trustworthiness attributes (Level 2)

Attribute	**Definition**
Robustness of the results ($RoR = T_{1,1}$)	The capability of the PRA results to remain unaffected by small variations in model parameters or model assumptions
Suitability of the model ($SoM = T_{1,2}$)	The technical adequacy of the tool, maturity and ability to model the problem of interest
Quality of application ($QAp = T_{1,3}$)	The degree to which the analysis is implemented with the minimum required levels of details and modeling adequacy that have the degree of quality, suitable for supporting the application of interest
Knowledge of potential hazards and accident evolution processes ($PoH = T_{2,1}$)	The availability of documentation and knowledge of abnormal events, accidents and their evolutions, from similar systems
Phenomenological understanding ($Ph = T_{2,2}$)	The knowledge that supports the comprehension of the system functionality and the related phenomena
Data ($D = T_{2,3}$)	The amount and quality of data needed for estimating the model parameters

2

Table 3 Definition of trustworthiness attributes (Level 3)

Attribute	**Definition**

Model sensitivity $(MS = T_{1,1,1})$	The degree to which the model output varies when one or several parameters change
Impact of assumptions $(IoA = T_{1,1,2})$	The degree to which the model output varies when one or several assumptions change
Robustness of the model $(RoM = T_{1,2,1})$	The capability of the model to keep its performance when applied to a different problem settings
Suitability of the model for the problem $(S = T_{1,2,2})$	The ability to capture all the important details and characterizations of the problem of interest
Historical use $(HU = T_{1,2,3})$	The degree of confidence gained in this method by the long historical usage
Conservatism $(Cv = T_{1,3,1})$	The intentional acts for overestimating the risk by making conservative assumptions out of cautiousness
The accuracy of calculations $(AcC = T_{1,3,2})$	The degree of the voluntarily accepted error in the calculation, e.g., significant figures, simulation errors, and cutoff errors
Quality of assumptions $(QoA = T_{1,3,3})$	The degree to which the assumption is valid, representing reality and supporting the model
Verification $(Vr = T_{1,3,4})$	The degree of assurance that the analysis maintains the requirements of quality control standards and obtains the acceptance from different analysts
Level of sophistication $(LoS = T_{1,3,5})$	The degree of treatment of the problem, and amount of effort and details invested in the problem given its requirement (requirement and complexity)
Number of known hazards $(NH = T_{2,1,1})$	The documented experience on known hazards that might affect the system of interest
Availability of accident analysis reports $(NH = T_{2,1,2})$	The availability of technical reports that cover thoroughly the different sequences of any abnormal activity, incident or accident in the time frame and the progressions of each phase
Experts knowledge about the hazard $(NH = T_{2,1,3})$	The undocumented experience possessed by experts on known hazards
Years of experience $(YE = T_{2,2,1})$	The amount of experience (measured in years) regarding a specific phenomenon
Number of experts involved $(NE = T_{2,2,2})$	The number of experts who are explicitly or implicitly involved in understanding the phenomena and the risk analysis

Academic studies on the phenomena ($AE = T_{2,2,3}$)	The number of academic resources, i.e., articles, books, etc., available about the phenomena of interest
Industrial evidence and applications on the phenomena ($IE = T_{2,2,4}$)	The number of industrial applications and reports related to the specific phenomena or events of interest
Amount of available data ($AD = T_{2,3,1}$)	The amount of data that are needed to evaluate the model parameters
Reliability of data ($RD = T_{2,3,2}$)	The degree to which the properties of data satisfy the requirements of risk analysis

1

2 Table 4 Definition of trustworthiness attributes (Level 4)

Attribute	**Definition**
The plausibility of assumptions ($Pl = T_{1,3,3,1}$)	The degree of realism of the statements made in the analysis, in cases of lack of knowledge or to facilitate the problem solution
Value ladenness of assessors ($VL = T_{1,3,3,2}$)	The experts' degree of objectivity, professionalism, skills and competencies, past fulfillment of assigned missions and level of achievement
Agreement among peers ($Ag = T_{1,3,4,1}$)	The degree of resemblance between the peers on the analysis and assumptions made, if they were asked to perform the analysis separately
Quality assurance ($QA = T_{1,3,4,2}$)	The degree of following the standards in the process of implementing the analysis
Level of granularity ($LoG = T_{1,3,5,1}$)	The depth of analysis and subdivision of the problem constituting elements
Number of approximations ($NoA = T_{1,3,5,2}$)	The intentional simplifications made to facilitate the modeling
Level of details ($LoD = T_{1,3,5,3}$)	The degree with which the important contributing factors are captured in the modeling compared to the requirement of the analysis (e.g., the dependency among components)
Completeness ($LoD = T_{2,3,2,1}$)	The degree to which the collected data contain the needed information for the risk modeling and assessment
Consistency ($LoD = T_{2,3,2,2}$)	The degree of homogeneity of data from different data sources
Validity ($LoD = T_{2,3,2,3}$)	The degree to which the data are collected from a standard collection process and satisfy the syntax of its definition (documentation related)

Timeliness $(LoD = T_{2,3,2,4})$	The degree to which data correctly reflect the reality of an object or event
Accuracy $(LoD = T_{2,3,2,5})$	The degree to which data are up-to-date and represent reality for the required point in time

3. Evaluation of the level of trustworthiness

In this section, a bottom-up method for evaluating the level of trustworthiness is developed in Section 3.1. Then, a combination of Dempster Shafer Theory (DST) and Analytical Hierarchy Process (AHP) are used in Section 3.2 to determine the weights of the attributes/sub-attributes in the method proposed in Section 3.1.

3.1. Evaluation of the trustworthiness

In this framework, five levels of trustworthiness are defined with their corresponding settings:

1. Strongly untrustworthy $(T = 1)$: represents the minimum level of trustworthiness and, therefore, the decision maker has the lowest confidence in the result of the PRA. The analysis is made based on weak knowledge and/or nonrealistic analysis, leading to an estimated value that might be far from the real one. Further analysis and justifications need to be implemented on the risk analysis to enhance its trustworthiness. Otherwise, the risk assessment is not considered representative and one should not rely on its results to support any kind of decision making.

2. Untrustworthy $(T = 2)$: represents a low level of trustworthiness and, therefore, the decision maker has low confidence in the results of the PRA. At this level, the analysis is made based on relatively weak knowledge and/or nonrealistic analysis, leading to unrealistically estimated risk values. Further analysis and justifications need to be implemented on the risk analysis to enhance its trustworthiness. The decision maker can use the results with caution and only as a support for decision making.

3. Moderately trustworthy $(T = 3)$: represents a moderate level of trustworthiness and, therefore, the decision maker has an acceptable level of confidence in the results of the PRA. The analysis is made based on relatively moderate knowledge and/or relatively realistic analysis. The decision maker can rely cautiously on the model output to make the decision.

4. Trustworthy $(T = 4)$: represents a high level of trustworthiness and, therefore, the decision maker has quite high confidence in the results of the PRA. The analysis is made on a relatively high level of knowledge and realistic analysis. The decision maker can rely confidently on the models output to make decisions.

5. Highly trustworthy ($T = 5$): represents the maximum level of trustworthiness. At this level, the PRA model outputs accurately predict the risk index with a proper characterization of parametric uncertainty. The decision maker can rely on the models output to support decision making involving severe consequences, e.g., loss of human lives.

In practice, the trustworthiness of risk assessment might be between two of the five levels defined above: for example, $T = 2.6$ means that the level of trustworthiness is between untrustworthy and moderately trustworthy.

In this paper, the level of trustworthiness of risk assessment is evaluated using a weighted average of the "leaf" attributes in Figure 1.

$$T = \sum_i^n W_i \cdot A_i \quad (1)$$

where W_i is the weight of the leaf attribute that measures its relative contribution to the trustworthiness of risk assessment; A_i is the trustworthiness score for the *i*-th leaf attribute, evaluated based on the scoring guidelines presented in the Appendixes; n is the number of the leaf attributes (in Figure 1, we have $n = 27$). The weights W_i are determined based on Dempster Shafer-Analytical Hierarchy Process (DST-AHP) (Dezert *et al.,* 2010), as discussed in Section. 3.2.

3.2. Dempster Shafer Theory - Analytical Hierarchy Process (DST-AHP) for trustworthiness attributes weight evaluation

The weights of the different attributes in Figure1 can be determined using the AHP method to compare their relative importance with respect to the trustworthiness of risk assessment (Saaty, 2008). AHP is used because it can decrease the complexity of the comparison process, as it allows comparing only two criteria at a time, rather than comparing all the criteria simultaneously, which could be very difficult in complex problems. It should be noted that since there are no alternatives to be compared in this framework, pairwise comparison matrixes of AHP are only used for deriving the attributes (criteria) weights.

To consider the fact that experts are subjective, not fully reliable and might have conflicting viewpoints, as well as considering the incomplete knowledge of the experts, Dempster-Shafer-Analytical Hierarchy Process (DST-AHP) is used. This allows combining multiple sources of uncertain, fuzzy and highly conflicting pieces of evidence with different levels of reliability (Dezert *et al.*, 2010), (Jiao *et al.*, 2016). In this method, the assessors are asked to identify the focal sets that comprise of a single or group of criteria. The experts determine the criteria contained in the focal sets in such a way that they are able to compare them (the focal sets), given their knowledge. Then, pairwise

comparison matrices are constructed for the focal sets. Using focal sets instead of single criteria allows taking into account the partial uncertainty between possible criteria. The basic belief assignments (BBA) of the corresponding focal sets are derived from the pairwise comparison matrices. The BBAs from different experts are combined using the Dempster fusion rule. The weights for each criterion are assumed to be BBA of the corresponding focal element (single criterion), and are derived based on the maximum belief-plausibility principle in Dempster-Shafer theory, or on the maximum subjective probability obtained by probabilistic transformations using the transferable belief model (Dezert *et al.*, 2010), (Dezert and Tacnet, 2011), (Jiao *et al.*, 2016). Again, note that in this work, this method is applied only to derive the relative weights of the criteria, rather than using it to rank alternatives. Similar ideas have been used in Tayyebi *et al.* (2010), Ennaceur *et al.* (2011). The procedure for calculating the weights of the leaf attributes based on DST-AHP is presented below.

I. Constructing pairwise comparison matrices

First, the experts are asked to construct pairwise comparison matrices (also known as knowledge matrices) to compare the relative importance of the attributes and sub-attributes in the same level of the hierarchy with respect to their parent attribute. For example, the pairwise comparison matrix for the attribute modeling fidelity (T_1) is a 3×3 matrix that compares the relative importance of the modeling' fidelity daughter attributes:

$$\begin{array}{c|ccc} & T_{1,1} & T_{1,2} & T_{1,3} \\ T_{1,1} & & & \\ T_{1,2} & & & \\ T_{1,3} & & & \end{array} = \begin{bmatrix} 1 & MF_{12} & MF_{13} \\ MF_{21} & 1 & MF_{23} \\ MF_{31} & MF_{32} & 1 \end{bmatrix}$$

where the columns correspond to the pairwise comparisons of the daughter attributes: robustness of the results ($T_{1,1}$), suitability of the selected model ($T_{1,2}$), and quality of the application ($T_{1,3}$), respectively. The element MF_{ij} is assigned by assessing the relative importance of attribute i to attribute j following the scoring protocols in (Saaty, 2008). For example, the element MF_{12} is assigned by comparing the relative importance of $T_{1,1}$ *to* $T_{1,2}$.

Compared to conventional AHP comparison matrices, the expert is free to choose, based on his/her belief, the elements of the pairwise comparison matrix. These elements can be focal elements that represent a single criteria, e.g., $\{A\}$ or a distinct group of criteria, e.g., $\{A, B\}$ that are comparable favorably (to the best of expert's knowledge) to the universal set that contains all the criteria, which allows accounting for the uncertainty in the judgment (Beynon, Cosker and Marshall, 2001), (Ennaceur, Elouedi and Lefevre, 2011), (Jiao *et al.*, 2016). For example, the expert can choose a focal set of $\{SoM, QAp\}$ if he/she believes that it can be compared favorably to the universal set $\{SoM, QAp, RoR\}$; i.e., the set of $\{SoM, QAp\}$ can be compared to $\{SoM, QAp, RoR\}$ (the sub-attributes *SoM,*

QAp, RoR were defined in Table 1-4). Then, the expert is asked to fill the pairwise comparison matrices to represent his/her belief in the relative importance of a given set (of one or multiple attributes) compared to the others. Favoring the universal set $\{SoM, QAp, RoR\}$ over $\{SoM, QAp\}$, means that the universal set contains an element that is not contained in the other set, and at the same time it is more important than the elements of the other set, i.e., RoR is more important than SoM and QAp. Finally, as in the conventional AHP method, the consistencies of the matrixes need to be tested and the assessors are asked to update their results if the consistency is lower than the required value (Saaty and Vargas, 2012).

II. Computing the weights

In this step, the weights are derived using the conventional AHP technique, according to which the normalized principal eigenvector of the matrix represents the weights. A good approximation for solving the eigenvector problem in case of high consistency is to normalize the columns of the matrix and, then, average the rows for obtaining the weights. For more details on AHP and deriving the weights from pairwise comparison matrices, the reader might refer to (Saaty, 2013). Please note that, as mentioned earlier, the weights derived from the pairwise comparison matrices are assumed to be the BBA of the associated focal sets.

III. Reliability discounting

Usually, multiple experts are involved in evaluating the weights. Each expert is regarded as an evidence source. Reliability of an evidence source represents its ability to provide correct measures of the considered problem (Jiao *et al.*, 2016). Shafer's reliability discounting is often used to consider the reliability of the source information in DST-AHP (Shafer, 1976):

$$m_\delta(A) = \begin{cases} \delta \cdot m(A) & \forall A \subseteq \Theta,\ A \neq \Theta \\ (1-\delta) + (\delta) \cdot m(\Theta), & A = \Theta \end{cases},\ \delta \in [0,1] \tag{2}$$

where Θ represents the complete set of criteria, A is the focal element in the power set 2^Θ, $m(A)$ is the BBA for A, $m_\delta(A)$ is the discounted BBA, δ is the reliability factor. A value of $\delta = 1$ means that the source is fully reliable and a value of $\delta = 0$ means that the source is fully unreliable. The reliability factor of the experts is determined by the decision maker, based on their previous knowledge and experience.

IV. Combination of experts opinions

Next, Dempster's rule of combination (Shafer, 1976) is used to combine two independent pieces of evidence assigned by different experts. The discounted BBAs from different experts are combined by (Jiao *et al.*, 2016):

$$m_{1,2}^{\delta}(C) = \left(m_1^{\delta} \oplus m_2^{\delta}\right)(C) = \begin{cases} 0 & C = \phi, \\ \frac{1}{1-K} \cdot \sum_{A \cap B = C \neq \phi} m_1^{\delta}(A) \cdot m_2^{\delta}(B) & C \neq \phi, \end{cases} \quad (3)$$

where $m_{1,2}^{\delta}(C)$ is the new BBA resulting from the combination of the two discounted BBA $m_1^{\delta}(A)$ and $m_2^{\delta}(B)$ of the two experts. K is the conflict factor in the opinions of experts and given by:

$$K = \sum_{A \cap B = \phi} m_1^{\delta}(A) \cdot m_2^{\delta}(B) \quad (4)$$

V. Pignistic probability transformation

The belief functions resulted from the discounting and combination are defined for focal sets (might contain one or multiple leaf attributes). To obtain the weights of each leaf attribute, the masses ($m_{1,2}^{\delta}(C)$) assigned to the focal sets need to be transformed into masses for the basic elements. In this paper, the transferable belief model proposed by (Smets and Kennes, 1994) is used for the transformation. In this method, the masses $m_{1,2}^{\delta}(C)$ on the credal level are converted to the pignistic level using the insufficient reason principle (Smets and Kennes, 1994), (Aregui and Denœux, 2008):

$$w(x) = \sum_{C \subseteq \Theta, C \neq \phi} \frac{m(C)}{1 - m(\phi)} \frac{1_{C(x)}}{|C|}, \forall x \in \Theta \quad (5)$$

where $w(x)$ denotes the belief assignment of a single element (x) on the pignistic level, 1_C is the indicator function of C: $1_C = 1, if\ x \in C\ and\ 0\ otherwise.$ $|A|$ is the length of A (the number of elements in the focal set). The mass functions obtained from the pignistic probability transformation represent the relative "believed weights" of the attributes.

After obtaining the local weights of the leaf attributes with respect to their parent attribute, the global weights with respect to the top-level attribute, i.e., the trustworthiness, need to be determined. This can be done by multiplying the weight of the daughter attribute by the weights of the upper parent attributes in each level. For example, the "global weight" of the historical use with respect to the trustworthiness, denoted by $W_{global}(HU)$, is calculated by:

$$W_{global}(HU) = w(HU) \times w(SoM) \times w(MF)$$

where $w(HU), w(SoM)\ and\ w(MF)$ are the local weights of the historical use, the suitability of the model, and the modeling fidelity. For simplicity reasons, hereafter the global weights for the leaf attributes are denoted by W_i and in the framework of Figure 1, we have $i = 1,2,\cdots,27$.

4. Evaluation of the risk considering trustworthiness levels

In this section, the "weighted posterior" method (Groen and Mosleh, 1999) is used for integrating the risk index with the trustworthiness of the PRA for a single hazard group (Section 4.1). In Section 4.2, a structured methodology

is developed for determining the weights in the Bayesian "weighted posterior" model. Finally, MHRA considering the level of trustworthiness is discussed in Section 4.3.

4.1. Evaluation of the risk of a single hazard group

After evaluating the level of trustworthiness for the PRA of a given hazard group, the next question is how to integrate the estimated risk from the PRA with the level of trustworthiness. In this paper, we develop a Bayesian averaging model for integrating the trustworthiness based on the "weighted posterior" method (Groen and Mosleh, 1999). Let us consider two scenarios: the risk assessment is trustable, denoted by E_T, and its complement, i.e., the risk assessment is not trustable (E_{NT}). The risk after the integration can, then, be calculated as:

$$Risk|T = P(E_T) \cdot \text{Risk}|E_T + (1 - P(E_T)) \cdot \text{Risk}|E_{NT} \quad (6)$$

where $Risk|T$ is the estimation of risk after considering the trustworthiness of the PRA; $P(E_T)$ is the subjective probability that E_T will occur and is dependent on the trustworthiness of the risk assessment; $\text{Risk}|E_T$ is the estimated risk from the PRA. Due to the presence of epistemic (parametric) uncertainty in the analysis, $\text{Risk}|E_T$ is often expressed as a subjective probability distribution of the risk index. $\text{Risk}|E_{NT}$ is an alternate distribution of the risk when the decision maker thinks the PRA is not trustable. In this paper, we assume $\text{Risk}|E_{NT}$ is a uniform distribution in [0,1], indicating no preference on the value of the risk index. Similar models have been used in literature to consider unexpected events in risk analysis (Kaplan and Garrick, 1981). For example, Kazemi and Mosleh (2012) developed a similar model to calculate the default risk in similar scenarios considering the unexpected events.

The following steps summarize how to use Eq. (6) to evaluate the risk given the trustworthiness of the risk assessment:

i. The risk distribution $\text{Risk}|E_T$ is evaluated for each hazard group using conventional PRA considering the parametric uncertainty propagation.
ii. The level of trustworthiness of PRA of the corresponding hazard group is assessed, using the procedures in Section 3.
iii. The subjective probability of trusting the PRA is determined by the detailed procedures described in Section 4.2.
iv. The level of trustworthiness is integrated in the risk using Eq. (6).

4.2. Determining the probability of trusting the PRA

The probability $P(E_T)$ in Eq. (6), which represents the decision maker's belief that the risk assessment results

are correct and accurate, needs to be elicited from the decision makers. The elicitation process needs to be organized and structured to ensure the quality of the elicitation.

Different methods can be found in the literature for the assessment of a single probability using experts elicitation, such as probability wheels, lotteries betting, etc. (Jenkinson, 2005). In this work, we choose the "certainty equivalent gambles" for the elicitation. Before presenting the procedure for this method, some general recommendations need to be followed to ensure the quality of the elicitation process (Jenkinson, 2005):

i. Background and preparation: uncertain events need to be defined clearly.
ii. Identification and recruitment of experts: The experts who are conducting the elicitation are chosen carefully with low-value ladenness, and a preference of being both substantively and normatively skilled.
iii. Motivating experts: the purpose and use of the work need to be explained to the experts, to motivate them for the elicitation.
iv. Structuring and decomposition: the dependencies and functional relationships need to be first identified by the client and agreed on and modified by the experts if necessary.
v. Probability and assessment training: the experts need to be trained to elicit probabilities.
vi. Probability elicitation and verification: the expert needs to elicit the probabilities paying caution to zero values, cognitive biases, etc. After making the elicitation, the expert needs to make a summary of the elicitation and verify its adequacy.

Then, a "certainty equivalent gamble" is designed to elicit the probability of trust:

i. The elicitor informs the decision maker about the definition of the different levels of trustworthiness and their physical meaning, based on the definitions in Section 3.1.
ii. The decision maker is asked to compare two scenarios: (1) he/she participates in a gamble (given the information from the PRA model) where he/she wins \$1,000 if an accident occurs and \$0 if the accident does not occur; (2) he/she wins $\$x$ for sure.
iii. The experts exchange information between them and discuss.
iv. Suppose that a PRA was conducted and predicted that the consequences occur for sure, and the trustworthiness of the PRA is one of the five levels defined in Section 3.1. Then, for each level of trustworthiness, the elicitor varies the value of x until the decision maker feels indifferent between the two scenarios.

v. The probability of trust at the current level of trustworthiness is, then, calculated by:

$$p = \frac{x}{1000} \tag{7}$$

where 1000 here represents the $1000 that the expert gains if the accident occurs (the model prediction is correct).

vi. The elicitor fits a suitable function to the five data points, in order to determine the probability of trust for trustworthiness levels between the defined levels. The shape of the fitted function should be determined based on the assessors' behavior towards taking risk in trusting a low fidelity PRA:

- A convex function should be chosen if the assessor is risk-averse, meaning that the decision maker trusts only the PRA with high levels of trustworthiness.
- A linear function is chosen if the assessor is risk neutral.
- A concave function is chosen if the assessor is risk-prone, meaning that although a PRA might not have a very high level of trustworthiness, the decision maker is willing to assign a high probability of trust to it.

The risk assessor can eventually use this function to estimate the probabilities of trust for each hazard group.

4.3. MHRA considering trustworthiness levels

The main steps for MHRA considering trustworthiness are presented in Figure 2. Trustworthiness in the PRA of each single group is evaluated and integrated into the risk estimate for the corresponding hazard group first. After the integration, the risk is expressed as a subjective distribution on the probability that a given consequence will occur. Then, the estimated risk from different hazard groups is aggregated. This step can be done by simply adding the risk distributions from different hazard groups, as shown in Eq. (8), where $Risk_{total}$ is the total risk considering the level of trustworthiness; $(Risk_i|\mathrm{T})$ is the risk from the hazard group i given the level of trustworthiness; n is the number of hazard groups. Monte-Carlo simulations can be used to approximate the distribution of $Risk_{total}$.

$$Risk_{total} = \sum_{i=1}^{n}(Risk_i|\mathrm{T}) \quad (8)$$

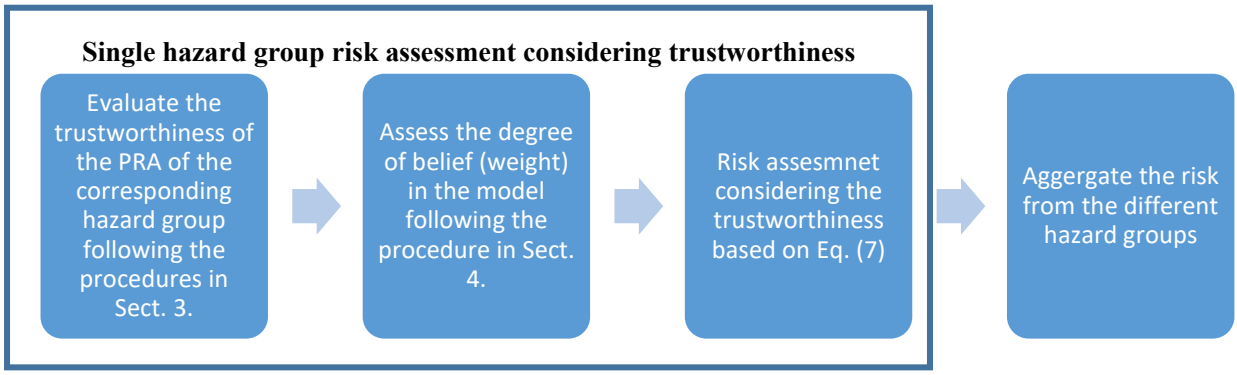

Figure 2 Main steps for MHRA considering the trustworthiness of the PRA

5. Case study

In this section, we apply the developed framework to a case study for two hazard groups in the nuclear industry: the external flooding and internal events hazard groups. The PRA models of the two hazard groups were developed and provided by Electricité De France (EDF) (Bani-Mustafa *et al.,* 2018). The level of trustworthiness is, then, assessed for each hazard group (Section 5.2). The risk distributions from each hazard group are, then, recalculated considering the level of trustworthiness. Finally, the risk is aggregated from the two hazard groups (Section 5.3).

5.1. Description of the PRA model

The two hazard groups considered in this framework are external flooding and internal events. The external flooding refers to the overflow of water that is caused by naturally induced hazards such as river overflows, tsunamis, dam failures and snow melts (IAEA, 2003), (IAEA, 2011). The internal events refer to any undesired event that originates within the NPP and can cause initiating events that might lead to abnormal states and eventually, a core meltdown (EPRI, 2015). Examples of internal events include structural failures, safety systems operation and maintenance errors, etc. (IAEA, 2009).

In risk analysis of NPP in general, risk analysis of different hazard groups are performed on the basis of PRA models for the internal events, which considered as relatively mature and realistic compared to other hazards groups. In external flooding hazard group, it is usually difficult to assess the probability of the flood hazard, especially that no reliable method is available for such a kind of analysis. For example, statistical models might be used to extrapolate external flooding frequencies from historical data. However, Only limited data are usually available on flooding and corresponds to few hundreds of years (usually 100-150 years). These data are used to extrapolate the flooding frequencies on different time interval, which, especially in extreme cases (where no data are available), would result

in large uncertainty (EPRI, 2015).

One of the most challenging point in external flooding risk analysis that the frequency and severity of each flood is site-specific, which would reduce the applicability of data from other sites. Also, the response and of NPP staff, in cases of floods, cannot be assessed easily as there are many factors that could affect their actions. However, it should be noted that it is highly recommended by regulatory bodies to perform some deterministic approaches for analyzing the floods hazards (EPRI, 2015).

In this case study, the risk analysis is provided by EDF (Bani-Mustafa *et al.,* 2018), in which bow-tie models are used to assess the probability of core damage frequency (CDF) (in 1/reactor·year). These large and complex models are of the order of hundreds to thousand basic events and several hundreds of minimal cutsets considering the different hazards that could lead to loss of system and consequently core damage.

Let's take the external flooding hazard group as an example. In this hazard group, the external flooding is considered as the hazard leading to the initiation of events within the plant that would possibly result in a core damage. The model is constructed considering the different equipment and systems that could be affected by water flooding in the NPP at different water heights. The different scenarios of the water arrival at the platform (with different heights) are built and propagated to understand their effect on the core. In this study, the probability of losing an equipment is calculated assuming that the equipment is directly lost once the water reaches the bottom of the equipment. In other words, the probability of losing an equipment equals to the probability that the level of water at the platform reaches the bottom of equipment.

The probability of having different water levels due to floods using a combined hydraulic/hydrologic method. First, data regarding the topography, hydrological and physical characteristic of the river basin were collected from the site of the NPP of interest. These data, allows calculating the water flowrate needed to obtain a specific water height at the platform of the NPP. Then, the data of the millennial flowrates of the river, were used to extrapolate to calculate the "return period" (average time needed for a river flood to occur) and then, extrapolate it to assess the frequencies of river flowrates on which no data are available. In other words, the data regarding the flowrate frequencies and the physical and hydrological nature of the basin, allow evaluating the frequencies of having given heights of water at the platform of the reactor. Therefore, it allows calculating the probability of equipment failures due to water flooding. Other intermediate events are also presented in the PRA models to represent the propagation of the initiating events and the different possible responses from the safeguard systems or the operators till the reactor core meltdown (Bani-Mustafa *et al.*, 2018).

In the original work of EDF, the uncertainty propagation was implemented, but only the mean values of the probability distributions of the risk were considered in MHRA and used for comparison to the safety criteria. However, due to confidentiality reasons, real values cannot be presented. Instead, we disguise the risk distribution, considering also the parametric uncertainty for illustration purposes, as shown in Figure 3.

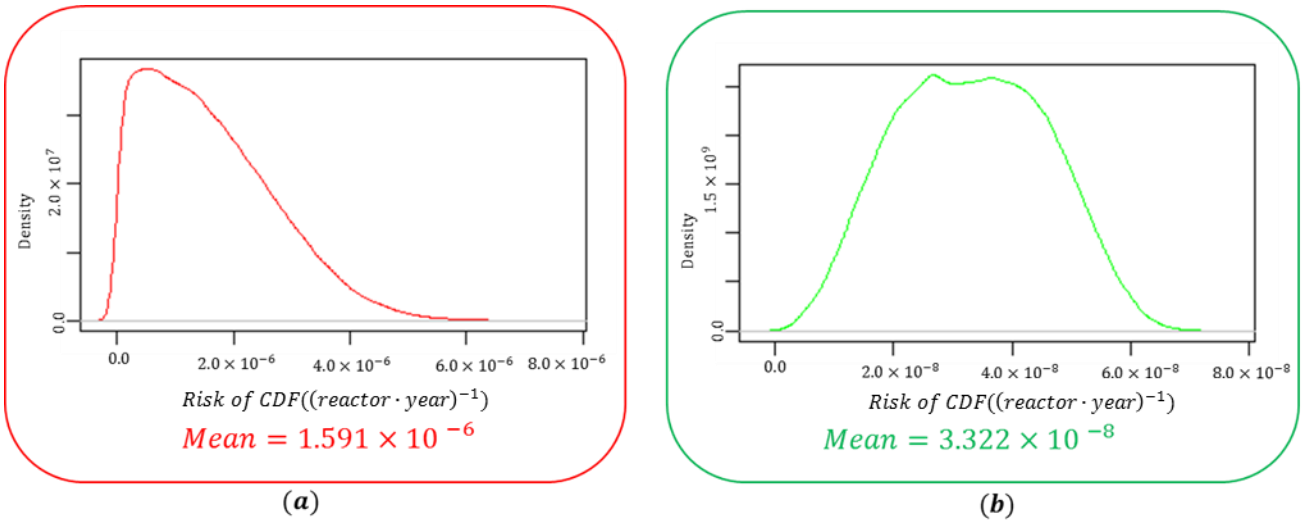

Figure 3 Probability distribution of the risk considering parametric uncertainty: (a) external flooding risk, (b) internal events

5.2. Evaluation of level of trustworthiness

5.2.1. Evaluation of the attributes weights

As illustrated in Section 3, the first step for evaluating the level of trustworthiness is to determine the relative importances (weights) of the trustworthiness attributes. The weights of the attributes are evaluated using the DST-AHP technique. Here, for explanation purposes, the sub-attribute "modeling fidelity" (T_1) is taken as an example to illustrate how to obtain local weights through pairwise comparisons and DTS-AHP.

I. Constructing pairwise comparison matrices

As shown in Section 3, the first step in the DST-AHP technique is to construct the pairwise comparison matrix. Take the daughter attributes of modeling fidelity as an example. In this example, a 4×4 pairwise comparison matrix is constructed in Table 5.

Table 5 Pairwise comparison matrix (knowledge matrix) for comparing modeling fidelity "daughter" attributes

Modeling fidelity	$\{T_{1,1}\}$	$\{T_{1,2}\}$	$\{T_{1,3}\}$	$\Theta = \{T_{1,1}, T_{1,2}, T_{1,3}\}$
$\{T_{1,1}\}$	1	0	0	1/2

$\{T_{1,2}\}$	0	1	0	5/2
$\{T_{1,3}\}$	0	0	1	4
$\{T_{1,1}, T_{1,2}, T_{1,3}\}$	2	2/5	1/4	1

Please note that the zeros that appear in the matrix indicate that there is no need to compare the individual criteria directly: they are compared indirectly through comparing the individual criteria to the universal set Θ (Dezert *et al.*, 2010).
$T_{1,1}$ represents the Quality of application, $T_{1,2}$ represents the Suitability of the model, $T_{1,3}$ represents the robustness of the results

In this matrix, the expert has considered four groups of focal sets: three for individual criteria and one containing all the criteria in order to consider the uncertainty in the evaluation. Choosing focal sets like this means that to the best of their knowledge, the experts believe that the aforementioned focal sets can be favorably compared to the universal set Θ.

II. Computing the weights

In the previous example, the expert was asked to fill the pairwise comparison matrix to express his/her preference of a criterion over another. In this step, the weights of the focal sets are derived using the conventional AHP technique, where the normalized principal eigenvector of the matrix represents the weights. This can be directly done by normalizing each column in the matrix individually and, then, averaging the elements in each row to obtain that weight.

Table 6 Normalized pairwise comparison matrix (knowledge matrix) of modeling fidelity "daughter" attributes

Modeling fidelity	$\{T_{1,1}\}$	$\{T_{1,2}\}$	$\{T_{1,3}\}$	$\Theta = \{T_{1,1}, T_{1,2}, T_{1,3}\}$	Weight (BBA)
$\{T_{1,1}\}$	0.33	0	0	0.06	0.10
$\{T_{1,2}\}$	0	0.71	0	0.31	0.26
$\{T_{1.3}\}$	0	0	0.8	0.5	0.32
$\{T_{1,1}, T_{1,2}, T_{1,3}\}$	0.67	0.29	0.2	0.13	0.32

III. Reliability discounting

After computing the BBA for each expert matrix, the weights need to be discounted based on the reliability of each expert. For illustration purposes, the reliability δ of the expert who made the assessment is assumed to be 0.60. From Eq. (2), the discounted weights are found as the following:

$$m_{0.60}(T_{1,1}) = 0.6 \times 0.10 = 0.06$$

Similarly, for $m_{0.60}(T_{1,2}) = 0.16$, & $m_{0.60}(T_{1,3}) = 0.19$.

Finally, $m_{0.60}(\Theta)$ is found as the following:

$$m_{0.60}(\Theta) = (1 - 0.60) + 0.6 \times 0.32 = 0.59$$

Please note that the BBAs (weights) sum to one before and after the discounting.

IV. Combination of experts opinions

In this case study, three experts have been invited to evaluate the weights; their assigned BBAs are summarized in Table 7 (the BBAs are calculated following the steps in Section 3.2).

Table 7 Discounted basic belief assignment from the three experts

Focal sets of the criteria	Expert 1 $m_\delta(A)$	Expert 2 $m_\delta(A)$	Expert 3 $m_\delta(A)$
$\{T_{1,1}\}$	0.06	0.16	0.02
$\{T_{1,2}\}$	0.16	0.24	0.38
$\{T_{1,3}\}$	0.19	0.24	0.46
$\{T_{1,1}, T_{1,2}, T_{1,3}\}$	0.59	0.36	0.14

The combination of the experts judgments is conducted sequentially. Table 8 shows the procedures for combining the judgments of the first two experts.

Table 8 Dempster's rule of combination matrix

Expert 1 \ Expert 2	$m_\delta(T_{1,1})$	$m_\delta(T_{1,2})$	$m_\delta(T_{1,3})$	$m_\delta(T_{1,1}, T_{1,2}, T_{1,3})$
$m_\delta(T_{1,1})$	$m_\delta(T_{1,1})_1$	ϕ_1	ϕ_2	$m_\delta(T_{1,1})_2$
$m_\delta(T_{1,2})$	ϕ_3	$m_\delta(T_{1,2})_1$	ϕ_4	$m_\delta(T_{1,2})_2$
$m_\delta(T_{1,3})$	ϕ_5	ϕ_6	$m_\delta(T_{1,3})_1$	$m_\delta(T_{1,3})_2$
$m_\delta(T_{1,1}, T_{1,2}, T_{1,3})$	$m_\delta(T_{1,1})_2$	$m_\delta(T_{1,3})_2$	$m_\delta(T_{1,3})_2$	$m_\delta(T_{1,1}, T_{1,2}, T_{1,3})_1$

*Please note that the element ij in the Table represent the multiplication of the elements $1j \times i1$, e.g., $m_\delta(T_{1,1}) \times m_\delta(T_{1,1}) = m_\delta(T_{1.1})_1$; $m_\delta(T_{1,1}) \times m_\delta(T_{1,1}, T_{1,2}, T_{1,3}) = m_\delta(T_{1.1})_2$

From Eq. (4), $K = 0{,}17$.

From Eq. (3):

$$m_{1,2}^{\delta}(T_{1,3}) = \frac{0{,}26}{1-0.17} = 0.31$$

The same steps are repeated for the other mass functions and presented in Table 9. Finally, the new results obtained from the combination of the two experts are further recombined with the BBAs from the third matrix. The results are presented in Table 9.

Table 9 Mass function combinations from the experts

Focal sets of the criteria	Combined mass from experts 1 and 2	Combined mass from experts 1, 2 and 3
	$m_{\delta}(A)$	
$m_{1,2}^{\delta}(T_{1,1})$	0.15	0.05
$m_{1,2}^{\delta}(T_{1,2})$	0.29	0.40
$m_{1,2}^{\delta}(T_{1,3})$	0.31	0.49
$m_{1,2}^{\delta}(T_{1,1}, T_{1,2}, T_{1,3})$	0.25	0.06

V. Pignistic probability transformation

Then, the pignistic mass function is found by Eq. (5):

$$w_{1,2,3}^{\delta}(T_{1,1}) = m_{1,2,3}^{\delta}(T_{1,1}) + \frac{m_{1,2,3}^{\delta}(T_{1,1}, T_{1,2}, T_{1,3})}{3} = 0.05 + \frac{0.06}{3} = 0.07$$

The steps are repeated for the other mass functions and found to be:

$$w_{1,2,3}^{\delta}(T_{1,2}) = 0.42$$

$$w_{1,2,3}^{\delta}(T_{1,3}) = 0.51$$

Note that the three mass functions on the pignistic level sum to one. These pignistic mass functions represent the relative "believed weights" of the three criteria under modeling fidelity after the reliability discounting and transformation. The same steps are repeated for all the criteria. Then, the weights need to be evaluated with respect to the top-level goal: the trustworthiness. As illustrated previously, this can be done easily by multiplying the weight of the daughter attribute by the weight of the upper parent attributes in each level. For simplicity reasons, only the weights of the "leaf" attribute with respect to the top level attribute i.e., trustworthiness, are presented in Tables 10 and 11 (see Section 5.2.2). Note that the weights of the 27 leaf-attributes with respect to the top goal sum to one $\sum_{i=1}^{27} W_i = 1$.

5.2.2. Evaluation of the attributes scores

The next step is to evaluate the attributes score for the hazard group, given the scoring guidelines in Appendixes A-B. Some information regarding the risk assessment process is extracted from the PRA report to support the trustworthiness assessment:

- The heights (water levels) at the plant's platform at which the water can lead to a failure of a specific element were defined.
- The water flowrate that would result in a given water height at the NPP platform in a defined interval of time was predicted.
- The flow-rate was multiplied by a safety factor of 130%.
- The "return period" for each flowrate was obtained from the data of the millennial flooding flowrate of the river of interest and the data were extrapolated to assess the frequencies of extreme flowrates.
- The river flooding is considered as a predictable phenomenon and the probability of failure of transition into the emergency state (i.e., normal shutdown and cooling with steam generator, residual heat removal system, etc.) is assumed to be the intrinsic probability of failure.
- It is assumed that river overflow is the only source of external flooding.
- A combined hydraulic/hydrologic method is adopted, given the special hydrological and physical characteristics of the basin.
- It is assumed that once the water reaches the bottom of the equipment, the equipment fails.
- It is assumed that failing to close the valves (ensuring the volumetric protection sealing-water proofing) causes the total loss of Emergency Feedwater System (EFWS).
- It is assumed that clogging inevitably occurs if the flooding occurs.
- The analysis and model calculation for this hazard group is taken with a specific cutoff error of 10^{-14}.

Based on the excerptions from the report, it can be seen that:

- In this example, the risk analysis and assessment steps follow the IAEA recommendations.
- The calculation of flowrates and flow frequencies are calculated using solid deterministic models. However, extrapolation of the data to obtain the frequencies of floods with extreme flowrates is still doubtful.
- The river overflow is a predictable phenomenon and does not happen suddenly. However, the river overflow is not the only source of flooding. For example, a rupture in the river dikes might also lead to

sudden, unpredictable flooding.

- The application of a combined hydraulic/hydrologic method on the flooding studies of nuclear sites allows a more realistic evaluation of the flooding level and to estimate more precisely the return periods.
- The assumption that the water will fail the equipment directly if it touches its bottom level is conservative.
- Feedback data show that clogging due to river flooding has occurred before in the nuclear industry (see, for example, USNRC General Electric Advanced Technology Manual for more information (NRC, 2011)). However, claiming that each flooding would surely lead to clogging is still questionable and needs to be studied in details, taking into account the different influencing parameters (hydraulic, geometrical and topographical properties) of the area (see (Gschnitz*er et al.*, 2017)).
- In case of failing to close the valves ensuring the volumetric protection, the probability that water will go back through the drainage system is not identified and assumed to be one ($P = 1$), though there are no relevant calculations. Moreover, once the water enters the physical protection locations, the safety-related equipment is assumed to be lost. Both assumptions are conservative to increase the safety margin.

Based on the above observations, the leaf attributes in Figure 1 can be evaluated. For example, quality assurance attribute is evaluated to be five ($T_{1,3,4,2} = 5$), since the PRA is conducted following the IAEA recommendations. The accuracy of the calculation is evaluated to be five ($T_{1,3,2} = 5$), since the cutoff error is apparently very low. The combined hydraulic/hydrologic models used for the flooding studies are able to capture the special hydrological and physical characteristics of the basin, which makes them suitable for the study. Hence, a score of four ($T_{1,2,2} = 4$) is given for the suitability of the model. The assumptions presented above are mostly conservative and unrealistic. Therefore, a score of one ($T_{1,3,3,1} = 1$) is given for the plausibility of the assumptions. The other attributes are scored in the same way. The results are represented in Tables 10 and 11. The level of trustworthiness for the external flooding is, then, calculated by Eq. (1): $T_{ext} = \sum_{i=1}^{27} W_i \cdot A_i = 3.260$.

Table 10 level-3 leaf attributes weights W and scores S for external flooding hazard group

Att	***MS***	***IoA***	***RM***	***S***	***HU***	***Cv***	***AoC***	***NH***	***AR***	***EK***	***YE***	***NE***	***Ac***	***In***	***AD***
W	0.012	0.026	0.025	0.158	0.070	0.025	0.012	0.022	0.032	0.054	0.034	0.017	0.105	0.105	0.065
Score	2	2	3	4	3	4	5	2	2	3	3	4	3	3	3

Table 11 level-4 leaf attributes weights W and scores S for external flooding hazard group

Att	***Pl***	***VL***	***Ag***	***QA***	***LoG***	***NoA***	***LoD***	***C***	***Co***	***V***	***T***	***Ac***
W	0.037	0.029	0.025	0.066	0.006	0.005	0.004	0.017	0.011	0.009	0.011	0.017
Score	1	4	4	5	4	4	4	3	3	3	3	3

The trustworthiness for internal events hazard group (T_{int}) was calculated in the same way and, the result is $T_{int} = 4.414$. These results confirm the expectations that the PRA for internal events is considered relatively mature and well established (EPRI, 2015) in contrast to the PRA of external hazards, which is considered less mature with several limitations (EPRI, 2012).

5.3. Risk assessment considering the level of trustworthiness

5.3.1. Determining the probability of trust in the PRA results

In this step, the decision maker is asked to assign a probability that represents the belief that the risk assessment model output is correct (hereafter called probability of trust), based on the certainty equivalent approach presented in Section 4.2. In this example, we assume that the decision maker exerts a risk-prone behavior and generates the results in Table 12. The data in Table 12 are extrapolated and fitted to a function, as shown in Figure 4.

Table 12 Probability of trust given the level of trustworthiness

Trustworthiness	**Probability of trust**
1	0.05
2	0.50
3	0.75
4	0.90
5	1.00

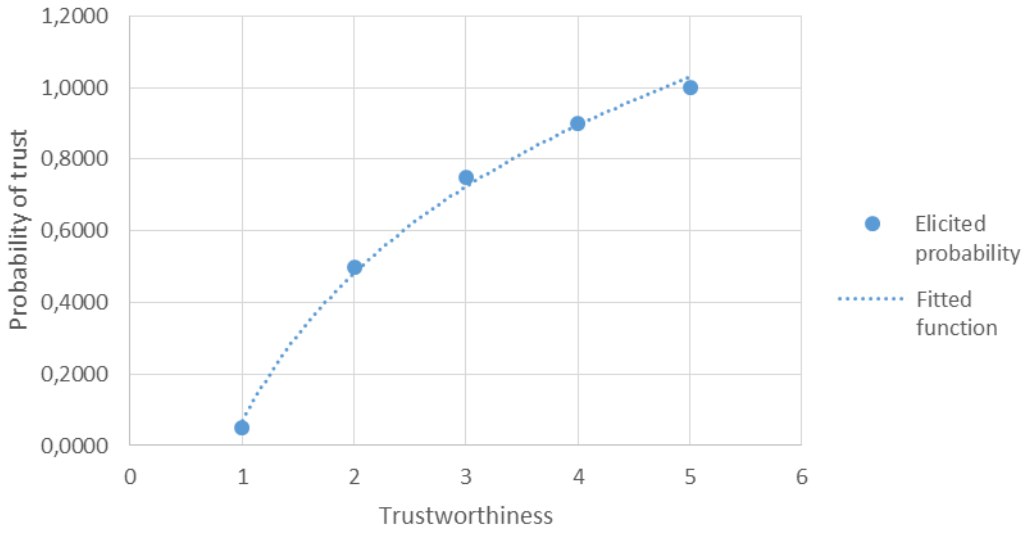

Figure 4 Fitted probability of trust in the PRA given the trustworthiness

Then, the probability that the decision maker trusts each hazard group PRA given their trustworthiness is calculated from the fitted model in Figure 4. The probability of trust for the external flooding p_{ext} is found to be $p_{ext} = 0.783$. The probability of trust for the internal events p_{int} is found to be $p_{int} = 0.957$.

5.3.2. Risk assessment of a single hazard group considering the level of trustworthiness

The level of trustworthiness is integrated with the PRA results for both hazard groups following Eq. (6). The results are presented in Figures 5 and 6, respectively. As illustrated in Figure 5, the mean risk value considering the trustworthiness is $1.088 \times 10^{-1}\ (reactor \cdot year)^{-1}$ for external flooding compared to $1.589 \times 10^{-6}\ (reactor \cdot year)^{-1}$ without considering the level of trustworthiness. For internal events, the mean risk value is $2.149 \times 10^{-2}\ (reactor \cdot year)^{-1}$ considering the trustworthiness compared to $3.322 \times 10^{-8}\ (reactor \cdot year)^{-1}$ without considering it for internal events, as illustrated in Figure 6. It can be seen from the Figures that considering the level of trustworthiness will lead to a larger spread out of the probability distribution of the risk. For further explanation, let's take Figure 5 as an example. In panel **(a)**, which represents the risk analysis considering the parametric uncertainty propagation, the spread-out of the risk distribution is limited to the risk interval $[4.626 \times 10^{-11}, 7.738 \times 10^{-6}]$. On the other hand, the interval of the risk distribution increases to $[3.019 \times 10^{-6}, 2.169 \times 10^{-1}]$ when the level of trustworthiness is considered in the risk analysis (see panel **(b)**). This comes out as a result of accounting for the disbelief in the risk analysis that reflects the ignorance about the real value of risk. Hence, the spread of the risk distribution becomes wider, leading to a higher mean value of the risk In other words, real values of risk can fall in reality in ranges of risk wider than that obtained by the initial analysis and

does not consider the level of trustworthiness.

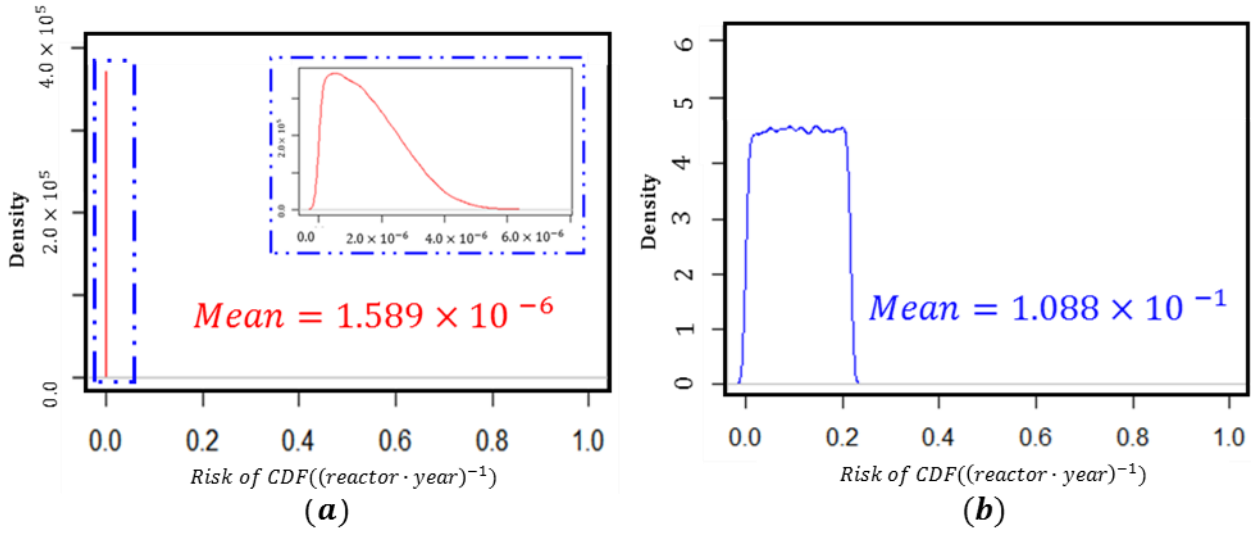

Figure 5 Updated risk estimates after considering the level of trustworthiness for external flooding (a) original risk estimate from the PRA, (b) Risk estimates after integrating the level of trustworthiness

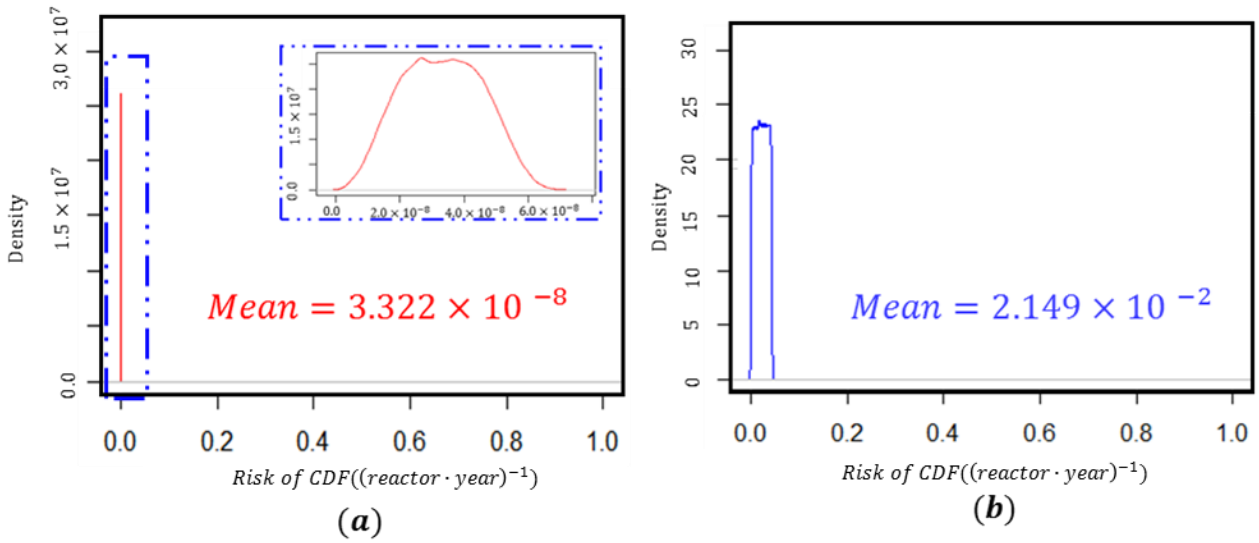

Figure 6 Updated risk estimates after considering the level of trustworthiness for internal events (a) original risk estimate from the PRA, (b) Risk estimates after integrating the level of trustworthiness

5.3.3. Multi-Hazards risk aggregation

Finally, the overall risk given the level of trustworthiness can be calculated using Eq. (8). The results are presented in Figure 7. The empirical probability density function of the risk is evaluated through a Monte-Carlo simulation of 10^5 samples. As a comparison, the MHRA is also conducted using the conventional methods by adding the risk indexes from the two hazard groups directly, without considering the trustworthiness, as shown in

Figure 7 (a). The mean value of the total risk from the two hazard groups considering the level of trustworthiness is found to be 1.303×10^{-1} $(reactor \cdot year)^{-1}$ compared to 1.622×10^{-6} $(reactor \cdot year)^{-1}$ without considering the level of trustworthiness. As discussed earlier, the aggregation of the risks from the two hazard groups needs to consider the different levels of trustworthiness to yield a mathematically appropriate process and a physically meaningful results. In fact, considering the level of trustworthiness in the analysis means that we are accounting for the disbelief, shortcoming, and lack of knowledge in the analysis, which leads to a broader spread-out of the distributions and a larger risk interval. The increase of the interval, in which the risk can fall, represents in fact a more realistic risk analysis as it accounts for the ignorance in the model. The increase in the spread out of probability distribution of risk leads to a higher mean value of risk, as it takes into account the fact that the PRA models of the two hazard groups are based on different levels of trustworthiness.

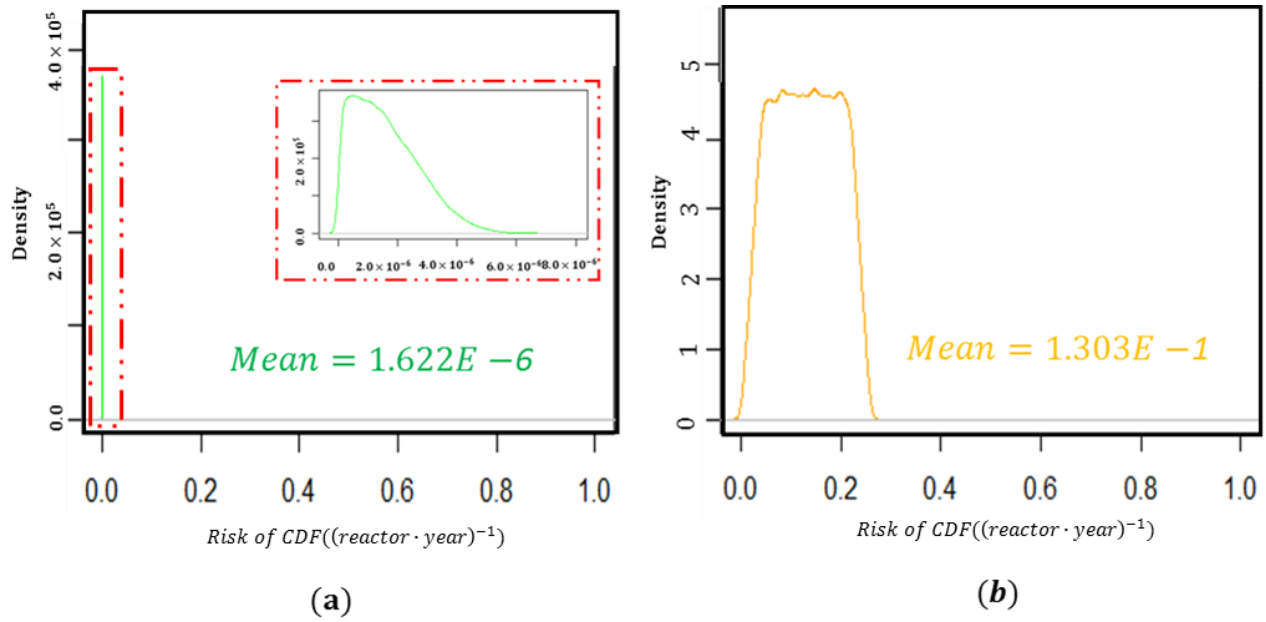

Figure 7 Results of the MHRA, (a) conventional aggregation, (b) considering the level of trustworthiness

6. Discussion and conclusion

In this paper, we have presented a framework for Multi-hazards Risk Aggregation (MHRA) considering trustworthiness. A framework for evaluating the level of trustworthiness is first developed. The framework consists of two main attributes, i.e., strength of knowledge and modeling fidelity. The strength of knowledge attribute covers the explicit knowledge that can be documented, transferred or explained. The modeling fidelity attribute covers the suitability of the tool and the model construction process. The two attributes are broken down into sub-attributes and,

finally, leaf attributes. The total trustworthiness is calculated using a weighted average of the attributes, where the weights are calculated using DST-AHP method. Where AHP method is used to calculate the relative weights of the attributes using experts elicitations, whereas DST method is used to account for the uncertainty in their elicitations.

A MHRA method is, then, developed to aggregate the risk from different hazard groups with different levels of trustworthiness, based on a "weighted posterior" method. An application to a case study of a NPP shows that the developed method allows aggregating risk estimates with different degrees of maturity and realism from different risk contributors.

The current framework represents a systematic way for enhancing the risk assessment and representing a mathematically more appropriate risk aggregation process. This is done by considering the different levels of realism on which the risk analyses of the aggregated hazard groups are based and integrating it in the risk analysis. From a practical point of view, the framework is developed in systematic and practical, procedural steps that facilitate the application of the framework to real life cases. In addition, it represents an illuminating point to better inform risk-based decision making, as it represents the degree of realism of the analysis.

However, a possible of the framework that DST is used only to account for the uncertainty in the experts' elicitations of the relative weights of the attributes and not the scores. Therefore, further studies need to be conducted to integrate DST method to also account for the uncertainty in the evaluation of the attributes scores of each given model.

Also, another possible possible limitation of the framework that the conventional safety criteria cannot be directly applied to the new risk estimates considering the trustworthiness. Therefore, future work is further needed for developing new safety criteria that corresponds with the new risk estimates that consider the trustworthiness to better inform the decision maker.

References

Aregui, A. and Denœux, T. (2008) 'Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities', *International Journal of Approximate Reasoning*. Elsevier, 49(3), pp. 575–594.

Aven, T. (2013a) 'A conceptual framework for linking risk and the elements of the data-information-knowledge-wisdom (DIKW) hierarchy', *Reliability Engineering and System Safety*, 111, pp. 30–36. doi: 10.1016/j.ress.2012.09.014.

Aven, T. (2013b) 'Practical implications of the new risk perspectives', *Reliability Engineering and System Safety*, 115, pp. 136–145. doi: 10.1016/j.ress.2013.02.020.

Aven, T. (2016) 'On the use of conservatism in risk assessments', *Reliability Engineering and System Safety*. Elsevier,

146, pp. 33–38. doi: 10.1016/j.ress.2015.10.011.

Bani-mustafa, T. *et al.* (2018) 'Strength of Knowledge Assessment for Risk Informed Decision Making', in *Esrel*. Trondheim.

Bani-Mustafa, T., Zeng, Z., *et al.* (2017) 'A framework for multi-hazards risk aggregation considering risk model maturity levels', in *System Reliability and Safety (ICSRS), 2017 2nd International Conference on*. IEEE, pp. 429–433.

Bani-Mustafa, T., Pedroni, N., *et al.* (2017) 'A hierarchical tree-based decision making approach for assessing the trustworthiness of risk assessment models', in *PSA (ANS)*. American Nuclear Society (ANS).

Berner, C. and Flage, R. (2016) 'Strengthening quantitative risk assessments by systematic treatment of uncertain assumptions', *Reliability Engineering and System Safety*, 151, pp. 46–59. doi: 10.1016/j.ress.2015.10.009.

Beynon, M., Cosker, D. and Marshall, D. (2001) 'An expert system for multi-criteria decision making using Dempster Shafer theory', *Expert Systems with Applications*, 20(4), pp. 357–367. doi: https://doi.org/10.1016/S0957-4174(01)00020-3.

Boone, I. *et al.* (2010) 'NUSAP: a method to evaluate the quality of assumptions in quantitative microbial risk assessment', *Journal of Risk Research*. Taylor & Francis, 13(3), pp. 337–352. doi: 10.1080/13669870903564574.

Dezert, J. *et al.* (2010) 'Multi-criteria decision making based on DSmT-AHP', in *BELIEF 2010: Workshop on the Theory of Belief Functions*. Belief Functions and Applications Society (BFAS), p. 8–p.

Dezert, J. and Tacnet, J.-M. (2011) 'Evidential reasoning for multi-criteria analysis based on DSmT-AHP', *Advances and Applications of DSmT for Information Fusion*, p. 95.

Ennaceur, A., Elouedi, Z. and Lefevre, E. (2011) 'Handling partial preferences in the belief AHP method: Application to life cycle assessment', in *Congress of the Italian Association for Artificial Intelligence*. Springer, pp. 395–400.

EPRI (2012) *Practical Guidance on the Use of Probabilistic Risk Assessment in Risk-Informed Applications with a Focus on the treatment of Uncertanity*. Palo Alto, California.

EPRI (2015) *An Approach to Risk Aggregation for RisK-Informed Decision-Making*. Palo Alto, California.

Flage, R. and Aven, T. (2009) 'Expressing and communicating uncertainty in relation to quantitative risk analysis', *Reliability: Theory & Applications*. Интернет-сообщество Gnedenko Forum, 4(2–1 (13)).

Groen, F. and Mosleh, A. (1999) 'Behavior of weighted likelihood and weighted posterior methods for treatment of uncertain data', in *Proc. ESREL*.

Gschnitzer, T. *et al.* (2017) 'Towards a robust assessment of bridge clogging processes in flood risk management', *Geomorphology*, 279, pp. 128–140. doi: https://doi.org/10.1016/j.geomorph.2016.11.002.

IAEA (1991) *Data Collection and Record Keeping for the Management of Nuclear Power Plant Ageing*. Edited by IAEA.

IAEA (2003) *External Events Excluding Earthquakes in the Design of Nuclear Power Plants*.

IAEA (2006) *Determining the Quality of Probabilistic Safety Assessment (PSA) for Applications in Nuclear Power Plants*. Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY. Available at: http://www-pub.iaea.org/books/IAEABooks/7546/Determining-the-Quality-of-Probabilistic-Safety-Assessment-PSA-for-Applications-in-Nuclear-Power-Plants.

IAEA (2011) 'IAEA-Publication8635'.

IAEA Safety Standards Series (2009) *Deterministic Safety Analysis for Nuclear Power Plants*.

Jenkinson, D. (2005) *The elicitation of probabilities: A review of the statistical literature*. Citeseer.

Jiao, L. *et al.* (2016) 'Combining sources of evidence with reliability and importance for decision making', *Central European Journal of Operations Research*. Springer, 24(1), pp. 87–106.

De Jong, A., Wardekker, J. A. and Van der Sluijs, J. P. (2012) 'Assumptions in quantitative analyses of health risks of overhead power lines', *Environmental science & policy*. Elsevier, 16, pp. 114–121.

Kaplan, S. and Garrick, B. J. (1981) 'On the quantitative definition of risk', *Risk analysis*. Wiley Online Library, 1(1), pp. 11–27.

Kazemi, R. and Mosleh, A. (2012) 'Improving default risk prediction using Bayesian model uncertainty techniques', *Risk Analysis: An International Journal*. Wiley Online Library, 32(11), pp. 1888–1900.

Kloprogge, P., Van der Sluijs, J. P. and Petersen, A. C. (2011) 'A method for the analysis of assumptions in model-based environmental assessments', *Environmental Modelling and Software*. Elsevier Ltd, 26(3), pp. 289–301. doi: 10.1016/j.envsoft.2009.06.009.

Nasa (2013) 'STANDARD FOR MODELS AND SIMULATIONS-NASA-STD-7009', (I), pp. 7–11.

NRC, U. S. (2011) *General Electric Advanced Technology Manual Chapter 4.8 Service Water System Problems*.

Oberkampf, W. L., Pilch, M. and Trucano, T. G. (2007) 'Predictive capability maturity model for computational modeling and simulation', *cfwebprod.sandia.gov*. Available at: https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/Oberkampf-Pilch-Trucano-SAND2007-5948.pdf%5Cnfile:///Users/markchilenski/Documents/Papers/2007/cfwebprod.sandia.gov%0A/Oberkampf/cfwebprod.sandia.gov%0A 2007 Oberkampf.pdf%5Cnpapers://31a1b09a-25a9-4e20-879d-4.

Paté-Cornell, M. E. (1996) 'Uncertainties in risk analysis: Six levels of treatment', *Reliability Engineering & System*

Safety, 54(2), pp. 95–111. doi: 10.1016/S0951-8320(96)00067-1.

Paulk, M. C. *et al.* (1993) 'Capability Maturity Model for Software, Version 1.1', *Software, IEEE*, 98(February), pp. 1–26. doi: 10.1.1.93.1801.

Popek, E. P. (2017) *Sampling and analysis of environmental chemical pollutants: a complete guide*. Elsevier.

Saaty, T. L. (2008) 'Decision making with the analytic hierarchy process', *International Journal of Services Sciences*, 1(1), p. 83. doi: 10.1504/IJSSCI.2008.017590.

Saaty, T. L. (2013) 'Analytic hierarchy process', in *Encyclopedia of operations research and management science*. Springer, pp. 52–64.

Saaty, T. L. and Vargas, L. G. (2012) *Models, methods, concepts & applications of the analytic hierarchy process*. Springer Science & Business Media.

Schwer, L. E. (2009) 'Guide for Verification and Validation in Computational Solid Mechanics'. IASMiRT.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton university press.

Siu, N. *et al.* (2015) 'FIRE PRA MATURITY AND REALISM: A DISCUSSION AND SUGGESTIONS FOR IMPROVEMENT'.

Van Der Sluijs, J. P. *et al.* (2005) 'Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System', *Risk Analysis*. Wiley Online Library, 25(2), pp. 481–492. doi: 10.1111/j.1539-6924.2005.00604.x.

Smets, P. and Kennes, R. (1994) 'The transferable belief model', *Artificial intelligence*. Elsevier, 66(2), pp. 191–234.

Tayyebi, A. H. *et al.* (2010) 'Combining multi criteria decision making and Dempster Shafer theory for landfill site selection', *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 38(8), pp. 1073–1078.

Veland, H. and Aven, T. (2015) 'Improving the risk assessments of critical operations to better reflect uncertainties and the unforeseen', *Safety Science*, 79, pp. 206–212. doi: http://dx.doi.org/10.1016/j.ssci.2015.06.012.

Zeng, Z. *et al.* (2016) 'A hierarchical decision-making framework for the assessment of the prediction capability of prognostic methods', *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*. SAGE Publications, 231(1), pp. 36–52. doi: 10.1177/1748006X16683321.

Zio, E. (1996) 'On the use of the analytic hierarchy process in the aggregation of expert judgments', *Reliability Engineering and System Safety*, 53(2), pp. 127–138. doi: 10.1016/0951-8320(96)00060-9.

Appendix A: Evaluation guidelines for leaf attributes under modeling fidelity (T_1)

Appendix A.1: Attributes under "robustness of the results attributes"

Table A.1.1 Scoring guidelines for robustness of the results

Score / Attribute	1	3	5
Model sensitivity T_{111}	$T_{111} = 1$ if the ensemble of model parameters greatly influence the final result	$T_{111} = 3$ if the ensemble of model parameters moderately influence the results	$T_{111} = 5$ if the ensemble of model parameters have little or no impact on the results of risk analysis
Impact of the assumptions T_{112}	$T_{112} = 1$ if the assumption greatly influences the results of risk analysis	$T_{112} = 3$ if the assumption moderately influences the results of risk analysis	$T_{112} = 5$ if the assumption has little or no impact on the results of risk analysis

Appendix A.2: Attributes under "suitability of the selected model"

Table A.2.1 Scoring guidelines for suitability of the selected model

Score / Attribute	1	3	5
Robustness of the model T_{121}	$T_{111} = 1$ if the model doesn't show the capability of performing under different settings or when exerting, deliberately, some variations in the assumptions and parameters	$T_{111} = 3$ if the model show the capability of performing moderately under different settings or small deliberate variations in the assumptions and parameters	$T_{111} = 5$ if the model show the capability of performing under different settings or when exerting, deliberately, large variations in the assumptions and parameters

Suitability of the tool T_{122}	$T_{122} = 1$ if the selected model is not usually used for achieving objectives similar to the required ones or it is not suitable for the problem settings and cannot capture all the important aspects of the problem	$T_{122} = 3$ if the selected model is usually used for achieving objectives similar to the required ones or it is suitable for the problem settings but doesn't capture entirely the important aspects of the problem	$T_{122} = 5$ if the selected model is usually used for achieving objectives similar to the required ones and it is suitable for the problem settings in a way that captures entirely the important aspects of the problem in a way that makes it suitable to represent reality
Historical use T_{123}	$T_{123} = 1$ if the selected tool is new or has never proved its successful use before, or if it is a new version of the tool that is quite different from the old one	$T_{123} = 3$ if the selected tool is a new updated version of a tool that has proved its successful use before	$T_{123} = 5$ if the selected tool is quite common tool that has proved its successful use in different problem settings, or if it is a slightly updated version of an old common one that proved it successful use

Appendix A.3: Attributes under "quality of application"

Conservatism:

In this setting, the conservatism is evaluated in the light of three criteria: (i) types of risk index estimates (best judgment, true value with a high confidence and true value with a low confidence); (ii) context of decision making; (iii) the effect of conservatism on the perception of the problem compared to best or true estimates or true and

consequently decision making assumptions and parameters. Figure A.1-3 illustrate the different score for each corresponding scenario.

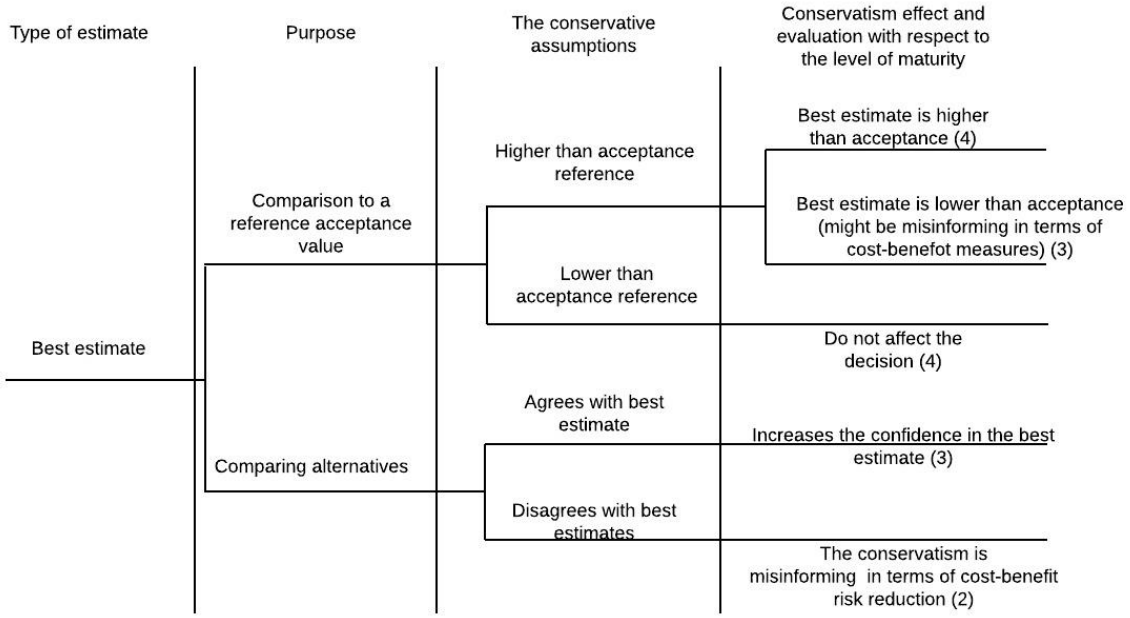

Figure A.3.1 Evaluation of the conservatism in the light of the level of maturity (conservatism VS Best estimate)

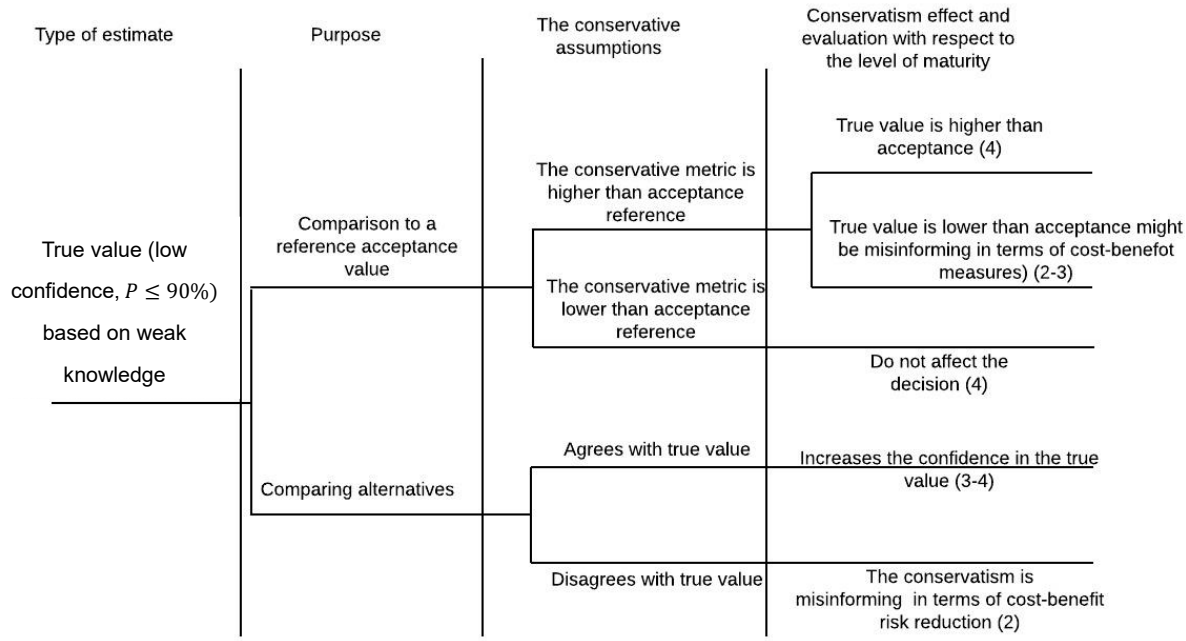

Figure A.3.2 Evaluation of the conservatism in the light of the level of maturity (conservatism VS True value/weak knowledge)

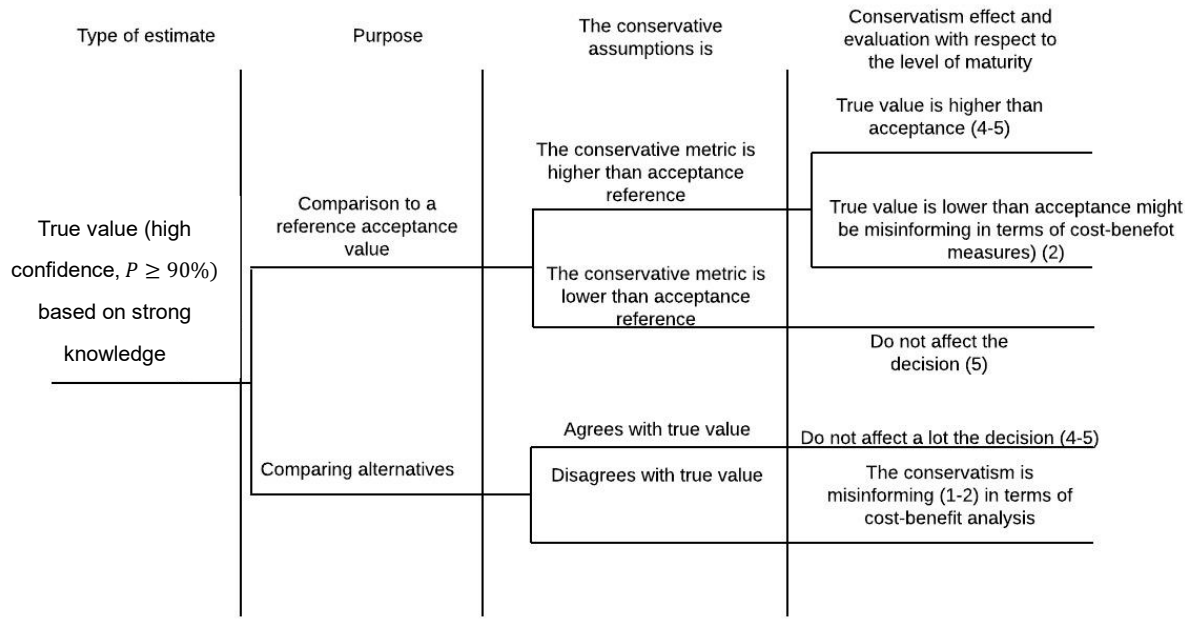

Figure A.3.3 Evaluation of the conservatism in the light of the level of maturity (conservatism VS True value/strong knowledge)

Table A.3.1 Scoring guidelines for the quality of the application

Attribute \ Score	1	3	5
The accuracy of the calculation T_{132}	$K_{131} = 1$ if the setting of accuracy is chosen to be low and high degree of error is accepted in the calculations. For example, the cutoff error (the chosen value of parameters at which lower values are ignored) is set to be large, and a low number of trials are performed	$K_{131} = 3$ if the setting of accuracy is chosen to be acceptable with a tolerable degree of errors. For example, the cutoff error is set to be quite low and a sufficient number of trials are performed	$K_{131} = 5$ if the setting of accuracy is chosen to be high and errors are conservatively accepted in the calculations. For example, the cutoff error is set at to be small, and a high number of trials are performed

Table A.3.2 Scoring guidelines for quality of assumptions *(*Boone *et al.,2010)*

Attribute \ Score	1	3	5
Plausibility of assumptions T_{1331}	$K_{1331} = 1$ if the assumption is not realistic (over conservative or over optimistic), or the available information is not sufficient for assessing the quality of the assumptions	$K_{1331} = 3$ if the assumption is based on existing simple models and extrapolated data	$K_{1331} = 5$ if the assumption is plausible: it is grounded on well-established theory or abundant experience on similar systems, and verified by peer review

Note: If multiple assumptions are involved in the assessment, the final score for T_{1331} is obtained by averaging the scores of all the assumptions.

Table A.3.3 Scoring guidelines for the value-ladenness of the assessors

Attribute \ Score	1	3	5
Personal knowledge (educational background) T_{13321}	$T_{13321} = 1$ if all of the experts hold academic degrees from other domains	$T_{13321} = 3$ if less than two thirds of the experts hold academic degrees in the same field	$T_{13321} = 5$ if over two thirds of the experts hold academic degrees in the same field
Sources of information T_{13322}	$T_{13322} = 1$ if experts can only access academic information source or only industrial information source	$T_{13322} = 3$ if experts can access fully industrial information source and partially academic information source	$T_{13322} = 3$ if experts can fully access both academic and industrial information sources
Unbiasedness and plausibility	$T_{13323} = 1$ if the expert team is very	$T_{13323} = 3$ if the expert team is slightly	$T_{13323} = 5$ if as a team, the experts are

T_{13323}	conservative or optimistic	conservative/optimistic	unbiased: the biases of the experts can compensate one another
Relative independence T_{13324}	$T_{13324} = 1$ if over three quarters of the experts are highly influenced by mangers and stakeholders	$T_{13324} = 3$ if less than one quarter of experts might be influenced by the mangers and stakeholders	$T_{13324} = 5$ if all experts' decisions are highly independent
Past experience T_{13325}	$T_{13325} = 1$ if the experts' experience is less than 5 years	$T_{13325} = 3$ if the experts' experience is between 10-15 years	$T_{13325} = 5$ if the experts' experience is more than 20 years
Performance measure T_{13326}	$T_{13326} = 1$ if the performance of the experts are not evaluated by external peers	$T_{13326} = 3$ if the external peers generally acknowledge the experts' performance but raise some slight concerns	$T_{13326} = 5$ if the external peers endorse the experts' performance and approve them
*Please note the value-ladenness score is calculated by averaging the scores over all the attributes in this table.			

1 Table A.3.4 Scoring guidelines for leaf attributes under verification

Score / Attribute	1	3	5
Agreement among peers T_{1341}	$T_{1341} = 1$ if some experts hold strongly conflicting views on the assumptions	$T_{1341} = 3$ if some experts questions on the assumptions, but do not have strongly conflicting views	$T_{1341} = 1$ if most of the experts agree on the assumptions
Quality assurance T_{1342}	$T_{1341} = 1$ if the analysis does not follow	$T_{1341} = 3$ if the analysis follows moderately the	$T_{1341} = 5$ if the analysis follows

	the quality standards and recommendations set by the PSA community e.g., ASME standards, NRC regulatory guides, IAEA recommendations	quality standards and recommendations set by the PSA community e.g., ASME standards, NRC regulatory guides, IAEA recommendations	entirely and conservatively the quality standards and recommendations set by the PSA community e.g., ASME standards, NRC regulatory guides, IAEA recommendations

1

2 Table A.3.5 Scoring guidelines for leaf attributes under the level of sophistication

Score / Attribute	1	3	5
Level of granularity T_{1351}	$T_{1341} = 1$ if the level of analysis is performed abstractly and coarsely on the level of systems or level the level of large components	$T_{1341} = 3$ if the analysis is performed in to a sufficiently fine level that regards the small components of a system or a small factors of a problem	$T_{1341} = 1$ if the level of analysis is zoomed in to the level of component's small constituting parts e.g., considering the small constituting parts of a manual (i.e., valve, the body, bonnet, ports etc.) when building the physical model for calculating the failure rate of a manual valve

Number of approximations T_{1352}	$T_{1342} = 1$ if there is a large number of approximations and the aggregate of the approximations affects significantly the output	$T_{1342} = 3$ if there is a moderate number of approximations or the aggregate of the approximations affects moderately the output	$T_{1342} = 5$ if there is a low number of approximations and the aggregate of the approximations does not affect, or affects insignificantly the output
Level of details T_{1353}	$T_{1353} = 1$ if most of the relevant contributing factors (including those that are not evident in the model construction requirements) that affect the estimates are not captured in modeling process compared to a complete realistic modeling e.g., the dependency among components in calculating the failure of a given component, environmental and thermal effect on components, level of the PH	$T_{1353} = 3$ if most of the relevant contributing factors (including those that are not evident in the model construction requirements) that estimates are captured in the modeling process compared to a complete realistic modeling e.g., considering the dependency among components in calculating the failure of a given component, environmental and thermal effect on components, level of the PH	$T_{1353} = 3$ if all relevant contributing factors (including those that are not evident in the model construction requirements) that affect the estimates are captured in modeling process compared to a complete realistic modeling e.g., considering the dependency among components in calculating the failure of a given component, environmental and thermal effect on

			components, level of the PH

Appendix B: Evaluation guidelines for the strength of knowledge (T_2) leaf attributes

Appendix B.1: Attributes under "Known potential hazards"

Table B.1.1 Scoring guidelines for leaf attributes under known potential hazards

Score / Attribute	1	3	5
Number of known hazards T_{211}	$T_{211} = 1$ if there is only a few number of known relevant hazards that are considered in the analysis	$T_{211} = 3$ if there is a moderate number of known relevant hazards that are considered in the analysis	$T_{211} = 5$ if there is a high number of known relevant hazards that are considered in the analysis
Availability of accident reports T_{212}	$T_{212} = 1$ if there is no past experience and technical reports that explain and cover in details the timing, causes and different sequences of abnormal activities, incident or accident	$T_{212} = 3$ if there is only a few past experience and technical reports that explain and cover in details the timing, causes and different sequences of abnormal activities, incident or accident, or if there is abundancy of reports that covers accidents without details	$T_{212} = 5$ if there is abundancy of past experience and technical reports that explain and cover in details the timing, causes and different sequences of abnormal activities, incident or accident
Experts knowledge about hazards T_{213}	$T_{213} = 1$ if the expert has a low experience in such a type of analysis and	$T_{213} = 3$ if the expert has a moderate degree of experience in such a type	$T_{213} = 5$ if the expert has a high degree of experience in such a

	hazards, as well as other types of problem, in a way that prevents him from imagining new unknown types of hazards	of analysis and hazards, as well as other types of problem, in a way that allows him to imagine new unknown types of hazards	type of analysis and hazards, as well as other types of problem, in a way that allows him to imagine most of the unknown types of hazards

Appendix B.2: Attributes under "phenomenological understanding"

Table B.2.1 Scoring guidelines for phenomenological understandings' leaf attributes

Score / Attribute	1	3	5
Years of experience (human experience on the phenomenon) T_{221}	$T_{221} = 1$ if the phenomenon is new to a human being, and no theories about the phenomenon have been developed yet or the theories are incapable to explain well the phenomenon (e.g., black holes)	$T_{221} = 3$ if the phenomenon has been investigated for moderate years of experience with few theories that are consistent with preexisting ones but still, do not explain holistically the phenomena (e.g., nuclear physics)	$T_{221} = 5$ if the phenomenon has been investigated for a long time and well-established theories have been developed to explain the phenomenon, which have been proved by many evidences (e.g., classical physics)
Number of experts involved in the analysis T_{222}	$T_{222} = 1$ if there is no experts related to this domain (the assessors involved are not expert in	$T_{222} = 3$ if there is a moderate number of experts of acceptable reliability (two experts)	$T_{222} = 5$ if there is a sufficient number of highly reliable experts (more than two

	this domain) or the experts are unreliable	or a low number of experts of high reliability	experts)
Academic studies on the phenomena (measured by the number of articles and books published on the subject) T_{223}	$T_{223} = 1$ if no or limited published articles supports the understanding of the phenomenon (e.g., Einstein electromagnetic waves)	$T_{223} = 3$ if a moderate amount of the published articles supports the understanding of the phenomenon (e.g., nuclear energy)	$T_{223} = 5$ if a large amount of the published articles supports the understanding of the phenomenon (e.g., kinetic energy)
Industrial pieces of evidence and applications on the phenomena (measured by the number of applications available on this subject) T_{224}	$T_{224} = 1$ if no or few industrial applications and reports support the understanding of the phenomenon (e.g., autonomous vehicles)	$T_{224} = 3$ moderate amount of industrial applications and reports support the understanding of the phenomenon (e.g., machine learning)	$T_{224} = 5$ if la arge amount of industrial applications and reports support the understanding of the phenomenon (e.g., airplanes)

Appendix B.3: Evaluation guidelines for leaf attributes under "Data"

Amount of data T_{231} is measured by a numerical metric, Years of Experience (YoE), defined by the number of related events recorded during a specific period.

YoE =length of the data collection period (in years) × sample size of the data

The amount of data is scored based on the criteria in Table B.3.1.

Table B.3.1 Scoring guidelines for Amount of available data

Value of YoE	Score
< 50	1
50-199	2
200-499	3

500-999	4
>1000	5

Completeness of data refers to the degree to which the collected data contains the needed information. For components and systems, data completeness is characterized by the following criteria (IAEA, 1991):

1. The data should contain baseline information, which covers the design data and conditions of a component at its initial state.
2. The data should contain the operating history, which covers the service conditions of systems and components including transient and failure data.
3. The data should contain the maintenance history data, which covers the components monitoring and maintenance data.

For more details on how each of the previous attributes is identified, see (IAEA, 1991). However, it should be noted that the completeness features are defined differently depending on the problem. For example, data required for quantifying to a component failure frequency is different from that for quantifying a natural event. General scoring guidelines for evaluating T_{2321} are given, based on the degree to which criteria are satisfied, as shown in Table B.3.2.

Table B.3.2 scoring guidelines for data reliability

Score / Attribute	1	3	5
Completeness T_{2321}	$T_{2321} = 1$ if the data fail to contain the necessary information required in developing the risk assessment model (in the light of the completeness characteristics defined above)	$T_{2321} = 3$ if the data contain to an acceptable degree the necessary information required in developing the risk assessment model (in the light of the completeness characteristics defined above)	$T_{2321} = 5$ if the data contain all the necessary information required in developing the risk assessment model (in the light of the completeness characteristics defined above)

The validity of data is evaluated by the following criteria:

1. The integrity of data is carefully managed.
2. Databases are well organized and formatted in a common way, and easily retrieved and manipulated.
3. Data should be collected and entered in the database by well-trained maintenance personnel, and modern computer techniques should be used for data storage, retrieval, and manipulation.
4. The data collection and entering process should include an appropriate quality control mechanism.

Based on the four criteria the evaluation guidelines of T_{2323} can be defined in Table B.3.3.

Table B.3.3 scoring guidelines for data validity

Score / Attribute	1	3	5
Validity T_{2323}	$T_{2323} = 1$ if none of the validity criteria (illustrated above) is fulfilled	$T_{2323} = 3$ if the validity criteria (illustrated above) are partially fulfilled	$T_{2323} = 5$ if all of the validity criteria (illustrated above) are fulfilled

Accuracy measures how close the estimated or measured value is compared to the true value. Accuracy is determined by random and systematic errors in the measurements (Popek, 2017). Since the data involved in nuclear PRA are mostly related to the number of failures or degradations and are usually collected digitally from different sources, systematic errors in the data are very small. This means that the accuracy of data is primarily determined by random errors. Since the error margin of the confidence interval is widely accepted as a good indicator of the random errors, it can be used as a measure of the data accuracy. Error factor may be defined based on the upper and lower bounds of confidence interval:

$$error\ factor = \sqrt{\frac{U_l}{L_l}}$$

where U_l and L_l are the upper and the lower bounds of confidence intervals. The accuracy of data is, then, scored based on the value of error factors, following the guidelines in Table B.3.4 scoring guidelines for data reliability

Table B.3.4 scoring guidelines for data validity

Score / Attribute	1	3	5
Accuracy T_{2325}	$T_{2325} = 1$ if the error factor is greater than 10	$T_{2325} = 3$ if the error factor is between 2-10	$T_{2325} = 5$ if the error factor is less or equal to 2

The rest of the "leaf" attributes of the reliability of data are evaluated following the guidelines in Table B.3.5.

Table B.3.5 scoring guidelines for data reliability

Score / Attribute	1	3	5
Consistency T_{2322}	$T_{2322} = 1$ if the data are not from the same type of power plant, or have different characteristics compared to the system under investigation, e.g., different component or model	$T_{2322} = 3$ if the data are from the same power plant with the same type of component and the same characteristics of the system under investigation but from different manufacturers	$T_{2322} = 5$ if the data are from the same power plant with the same type of components and the components have the same characteristics and the same manufacturer
Timeliness T_{2324}	$T_{2324} = 1$ if the data has never been updated	$T_{2324} = 3$ if the data has been updated a few years ago (10 years and more)	$T_{2324} = 5$ if the data are up-to-date and are updated routinely