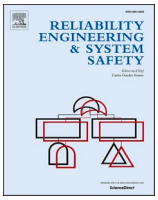
Reliability Engineering and System Safety 206 (2021) 107309

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis

Salvatore F. Greco, Luca Podofillini *, Vinh N. Dang

Risk and Human Reliability Group, Laboratory for Energy Systems Analysis, Paul Scherrer Institute, 5232 Villigen PSI, Switzerland

ARTICLE INFO

Keywords:
Human reliability analysis
Simulator data
Performance variability
SACADA
HuREX
Bayesian inference

ABSTRACT

The models adopted in Human Reliability Analysis (HRA) characterize personnel tasks and performance conditions via categories of task and influencing factors (e.g. task types and Performance Shaping Factors, PSF). These categories cover the variability of the operational tasks and conditions affecting performance, and of the associated Human Error Probability (HEP). However, variability exists as well within such categories, for example because of the different scenarios and plants in which data is collected, as well as of the operating crew differences (within-category and crew-to-crew variability). This paper presents a Bayesian model to mathematically aggregate simulator data to estimate failure probabilities, explicitly accounting for the specific tasks, scenarios, plants and crew behavior variability, within a given "constellation" (i.e. combination) of task and factor categories. The general aim of the proposed work is to provide future HRA with reference data with stronger empirical basis for failure probability values, both for their nominal values as well as for their variability and uncertainty. Numerical applications with both artificially-generated data and real simulator data are provided to demonstrate the effects of modelling variability in HEP estimates, to avoid potential overconfidence and biases. The applicability of the proposed model to ongoing simulator data collection programs is also investigated.

1. Introduction

Human Reliability Analysis (HRA) is the part of Probabilistic Safety Assessment (PSA) addressing the human contribution to the quantification of risk of complex technical systems, typically nuclear power plants, chemical and aerospace systems [1,2]. HRA aims to identify the safety-critical tasks performed by the personnel, to characterize the contextual factors influencing human performance, and to quantify the probability of failures.

To derive the human failure probability values (also referred to as Human Error Probabilities, HEPs), HRA methods characterize the personnel tasks and the factors deemed to influence task performance, the so-called Performance Shaping Factors (PSFs), e.g. adequacy of procedural guidance, of the human-machine interface, time available to accomplish the task, etc. HRA models characterize tasks and factors as categorical elements, with taxonomies and metrics dependent on the method. For instance, the Human Error Assessment and Reduction Technique (HEART, [3,4], newly issued in [5]) identifies nine generic task types (e.g. "complex task requiring high level of comprehension and skill") together with thirty-eight error producing conditions (e.g. "a low signal-noise ratio"). The Technique for Human Error Rate Prediction (THERP, [6]) characterizes tasks at a lower level of decomposition (e.g. "set a rotary control to an incorrect setting", "check/reading digital indicators") and PSFs such as training and stress (e.g. "Very low", "Optimum" stress). A similar use of categorical elements appears in all HRA methods, e.g. [7,8], and [9]. Recently, advanced models such as Bayesian Belief Networks (BBN) have been developed for HRA applications to capture the complex task, PSF, and HEP relationships and to enhance traceability in use of diverse data and judgment [10,11].

Reference data for the task categories and the PSF effects is needed to parametrize a method's quantification model, both for traditional as well as for advanced models. The data is generally obtained by combining empirical data and expert judgment [12]. Since the early developments of HRA, empirical data has been mainly gathered from human factor studies, data collection campaigns in main control room simulators, retrospective analyses of accidents, near misses and operational events [1,13]. An important turning point for HRA came from the International [14] and US HRA [15] Empirical Studies, aimed at assessing the validity of HRA method predictions against data from nuclear power plant main control room simulators. Besides improving HRA practice and methods, these studies resulted in methodological

* Corresponding author.
E-mail address: luca.podofillini@psi.ch (L. Podofillini).

https://doi.org/10.1016/j.ress.2020.107309
Received 7 February 2020; Received in revised form 8 October 2020; Accepted 7 November 2020
Available online 13 November 2020

Nomenclature

E:	evidence of the Bayesian model, expressed as set of pairs $\{(k_{ij}, N_{ij})$.
F:	set of taxonomy categories (e.g. task type and PSF levels/ratings), referred as "constellation".
$f_F(HEP)$:	parametric variability distribution, representing the overall spectrum of variability within a given constellation F.
$f_{c\|t}(p_{c\|t}\|p_t^*, \boldsymbol{\theta}_{c\|t})$:	"crew-to-crew" variability term of $f_F(HEP)$, modelling the variability across the crews performing the specific task/context realization (characterized by the crew-generic error probability value p_t^*) within the constellation F.
$f_t(p_t\|\boldsymbol{\theta}_t)$:	"within-category" variability term of$f_F(HEP)$, modelling the variability across the task/context realizations within the constellation F.
k_F, N_F:	total number of failures and observations for the constellation F ("lumped data").
(k_{ij}, N_{ij}):	number of failures observed on N_{ij} repetitions of the i-th task performed by the j-th crew.$i = 1, 2 \ldots, m$, $j = 1, 2 \ldots, n$)}, where m: total number of tasks in the dataset; n: total number of crews performing the i-th task.
$L(E\|\boldsymbol{\theta})$:	likelihood function of the Bayesian model, i.e. the probability density that evidence E is observed.
$N(\ldots)$:	normal distribution
$p_{c\|t}$:	crew-specific HEP variable.
$P_F(p_{c\|t})$:	estimated HEP variability distribution for the constellation F.
p_t:	task-, context-specific HEP variable (crew-generic).
p_t^*:	specific numerical value (i.e. a realization) of p_t.
t:	index for the task/context realization within the constellation F.
$(z_t, z_{c\|t})$:	normally-distributed auxiliary variables associated to p_t and $p_{c\|t}$.
(α, β):	shape parameters of the beta prior distributions.
(μ_t, σ_F):	parameters of the lognormal variability distribution (mean and standard deviation) used in the numerical application.
$\boldsymbol{\theta}_F$:	set of (unknown) parameters of the variability distribution $f_F(HEP)$.
$\boldsymbol{\theta}_t$:	set of (unknown) parameters of the within-category variability term (subset of $\boldsymbol{\theta}_F$).
$\boldsymbol{\theta}_{c\|t}$:	set of (unknown) parameters of the crew-to-crew variability term (subset of $\boldsymbol{\theta}_F$).
$\pi_0(\boldsymbol{\theta})$:	prior distribution of the Bayesian model, representing the knowledge on the set of parameters, (e.g. $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_{c\|t}$), before collecting the evidence E.
$\pi(\boldsymbol{\theta}\|E)$:	posterior distribution of the Bayesian model, representing the knowledge on the set of parameters, i.e. $\boldsymbol{\theta}_t$or $\boldsymbol{\theta}_{c\|t}$, after collecting the evidence E.

advances in the collection of simulator data for HRA purposes, with important implications on several recent activities [16–18]. Two notable, ongoing, data collection programs are the HUman REliability data Extraction framework, HuREX [16], and the Scenario Authoring, Characterization, And Debriefing Application, SACADA [17]: with their long-term data collection perspective, these are expected to produce a large amount of empirical evidence for new HRA reference data, more representative of recent operational conditions, e.g. reflecting modern interfaces and procedural guidance.

The majority of recent research activities dealing with the use of simulator data for HRA has addressed the development of protocols to collect data: notably, the interpretation of performance outcomes in terms of failure or success and the definition of the types of information on crew performance to collect (Hallbert et al., 2013; [17,19]. Open issues remain for how to use this information to quantify HEP values and how to eventually incorporate them into HRA methods, with various approaches being investigated [20–26].

Similarly to HRA methods, the data collection protocols characterize simulator observations through categories related to taxonomies of tasks (e.g. "entering step in procedure" in [16]), failure mechanisms (e.g. "failure to prioritize" in [17]), contextual factors (e.g. "overloaded status of alarm board" in [17]), and the like. The data associated to these categories is collected from different simulator scenarios, different plants, from crews with different behavioral styles, and different realizations of the contextual factors. Research on quantification of HEP values from the emerging data is ongoing internationally. A number of pioneering works [20,24,26] have shown the advantages of Bayesian inference models in using the collected simulator data to quantify the HEP (and the associated uncertainty) for multiple "constellations" (i.e. combinations) of taxonomy categories, e.g. from the SACADA taxonomy [17]: macrocognitive function "understanding the situation/problem", given the situational factor "information quality" with level "conflicting". These works focused on the relationship between the given task, the set of PSFs and the error probability, and investigated performance variability in simulator tasks under different PSF effects, i.e. "across constellations": with respect to the previous example, e.g. when the "information quality" is "misleading" instead of "conflicting". However, variability in simulator data exists as well within task and PSF categories, i.e. "within the constellation", for instance, due to the different scenarios and plants in which data is collected as well as to operating crew differences (we refer to it as "within-category" and "crew-to-crew" variability, respectively). Such variability requires explicit consideration: the simple approach of lumping all data relevant to a given constellation of categories would focus on the "population average"-HEP of the constellation. However, it may not adequately represent the existing sources of variability, and may possibly lead to overconfident results [13,27,28].

The present paper proposes an inference model to derive HEP estimates from simulator data that explicitly addresses within-category and crew-to-crew variability aspects within a given constellation of task type and PSF categories. The first aspect stems from differences across simulator scenarios and plant-specific realizations of the contextual factors associated to the same categories; the latter from differences across the operating crews, e.g. different problem-solving styles, communication strategies, modality of information sharing, team coordination (e.g. tendency to prioritize tasks). The emerging simulator data is used to inform both the average HEP value as well as the associated variability bounds (hence, the focus on within-category and crew-to-crew variability). The main idea is to produce reference HEP values that can be used to inform HRA methods task type and PSF categories (or PSF multiplier values, depending on the method) as well as anchoring values for parametrizing advanced HRA models, such as BBNs. The parameters of the model are inferred via a Bayesian hierarchical framework, generally applicable to diverse taxonomies of task and PSF categories familiar to the HRA community. Because of the limited data available, most of the established HRA models (e.g. THERP, [6]; the Standardized Plant Analysis Risk–Human reliability, SPAR-H, [7,8]; the Cognitive Reliability and Error Analysis Method, CREAM, [9]) assess data variability by expert judgment: as the running simulator campaigns will produce new data, it becomes important that data variability be formally incorporated in the HEP estimates, decreasing (and eventually replacing) the judgment.

Table 1
Sources of uncertainty and variability in HEP estimates by HRA methods (given a constellation of task type and PSF categories). Note our work addresses the first two items of this table.

Source of uncertainty and variability	Description	Example
Crew characteristics	Inherent performance variability across people and crews, due to different behavioural characteristics, abilities, attitudes, etc.	Both crews A and B perform exactly the same task in the exact same context. Crew A fails, crew B succeeds. Also inherent randomness of certain human behaviour: same person/crew performs the same task under the same performance conditions: sometimes fails, sometimes succeeds.
Contextual factors	Variability (aleatory) across the different realizations of the contextual factors described by the same category of factor taxonomy	Variability within PSF "time pressure" due to variability in time and sequence of events within the same scenario (dynamic change). Variability within "indications of conditions" PSF due to different indications and/or designs, all can be characterized as "misleading"
Assessment of PSF ratings	Uncertainty on the assessed PSF states for the investigated context. Can also manifest as inter-analyst / rater variability.	It is not possible to state with certainty whether "time pressure" during performance should be considered "moderate" or "high", due to inherent imprecision of contextual factor descriptions and different subjective interpretation of the PSF category
Model limitations	Uncertainty (epistemic) due to inherent, fundamental limitations of HRA models	Incompleteness of PSFs to represent a specific context of operations, limitation of underlying cognitive models to fully represent cognitive processes, lack of representation of safety culture, organizational and cross-organizational influences.
Scarcity of data	Uncertainty (epistemic) due to the limited knowledge of human performance in specific combinations of scenario/context of operation	Low-probability events (medium Loss Of Coolant Accident, with High Pressure Injection system failing to operate)

The adoption of variability models is well established in PSA to consider source-to-source variability in parameter estimation problems: plant-to-plant variability in the estimation of component failure rates [29,30] and other reliability measures [31,32]; expert-to-expert variability in the estimation of rare event frequencies [33] and in HRA model construction [34]; combination of statistical data with expert estimates [35] and reliability data [36].

The paper is structured as follows. Section 2 discusses uncertainty and variability aspects in HRA and in simulator data collection. Section 3 presents the developed Bayesian variability model and the underlying modelling assumptions. In Section 4, numerical applications with artificially generated data show the effects of modelling variability in HEP estimates and investigate the data requirements of the proposed model. In addition, an application to real simulator data from two different data sources (Halden project data from [20] and HuREX data from [26]) is presented. The results are further discussed in Section 5, along with insights and recommendations on the applicability of the model. Conclusions are given at closure.

2. Uncertainty and variability aspects in HRA and simulator data for HRA

The results of HRA methods support risk-relevant decisions; an important requirement is to ensure that the uncertainties of HEP estimates are appropriately quantified [37]. The next sections discuss how uncertainty and variability have been treated in existing HRA methods (Section 2.1) and in the analysis of simulator data (Section 2.2).

2.1. Treatment of uncertainty and variability in existing HRA methods

HRA quantification methods aim at representing the relationships between HEPs and PSFs, taking into account as well the interactions among PSFs. Tasks and contexts are typically characterized via constellations of categories (e.g. of task types and PSFs). As the constellation of these category changes, HRA models provide different Human Error Probability (HEP) estimates, representing the spectrum of performance conditions variability. The models produce estimated HEPs and characterize the uncertainty associated with these estimates, in the form of uncertainty distributions or bounds. For a given task type, a set of PSFs ratings yields a specific HEP distribution. Our work deals with the assessment of these distributions, which represent different aspects of uncertainty and variability [6,38], as summarized in Table 1.

Depending on the methods, bounds and distributions are derived in different ways. As discussed in Chapter 7 of the Handbook [6], THERP assumes a lognormal distribution of the HEPs to account for the various sources of uncertainty and variability associated to HEP values (such as those listed in Table 1). For each failure included in its database, THERP provides a nominal HEP (the median of the uncertainty distribution) as well as an Error Factor (EF)[1]. These uncertainty bounds, exclusively derived by expert judgment, are meant to reflect the THERP's analysts "judgment regarding the likelihood of various values of HEPs" (from [6]) associated to a task. Different from THERP, HEART's HEP values and bounds are obtained by aggregating empirical evidence on human performances from diverse information sources in the human factor literature [3,4], and the recently consolidated HEART version from [5]. In particular, for each generic task type, the author used the log-geometric mean of the set of data to derive the HEP central value and the log-standard deviation from the central value to calculate the HEP bounding values (in the form of 5th/95th percentiles). As a further example, the SPAR-H method adopts beta distributions (CNI, Constrained Non-Informative priors, by [39]) to determine uncertainty on HEP because the beta distribution can mimic both normal and lognormal distributions, with the advantage that it is defined from 0 to 1 [7]. As a general conclusion, except for HEART for which uncertainty is derived empirically, expert judgment is the dominant source for all other HRA methods.

2.2. Characterization of uncertainty and variability aspects in simulator data for HRA

The usefulness of simulator studies to inform human reliability models is recognized widely [12,40–43], along with the need for the models to represent the variability of human performance in response to emergency conditions. For example, the Human Cognitive Reliability (HCR) model [44,45] and the Operator Reliability Experiment (ORE) [46] from the early 1980s were aimed at generating time reliability curves based on the variability of operating crew response time to emergency conditions, observed in simulator studies.

More recently, the International [14] and US [15] Empirical Studies

[1] The Error Factor is defined as the square root of the ratio of the upper to lower bound of the uncertainty distribution.

Table 2
Hypothetical simulator data used to inform the categorical elements of a generic HRA model for HEP estimation. Categorical elements taken from the SACADA taxonomy [17]. Note that the table reports only few elements of the rich SACADA context characterization.

Categorical elements of HRA models ("constellation *F*"):					
Task type: understanding the situation/problem					
Information quality: conflicting					
Diagnosis basis: procedure					
Data from specific simulator contexts					
Scenario	**Realization of contextual factors**	**Task realization**	**Plant**	**Crews**	**Failures**
SGTR	One level indication in steam generator stuck low	Transfer to SGTR procedure	A	5	0
SGTR	One level indication in steam generator offset	Transfer to SGTR procedure	B	6	1
SGTR	One level indication in steam generator indicates zero	(…)	(…)	(…)	(…)
			Total	50	3
SLOCA	One indication on pressurizer pressure stuck high	Transfer to SLOCA procedure	A	5	1
SLOCA	One indication on pressurizer pressure indicates zero	Transfer to SLOCA procedure	B	6	2
SLOCA	Offset indication on pressurizer pressure	(…)	(…)	(…)	(…)
			Total	50	7

were carried out to assess strengths and weaknesses of HRA methods, by comparing HRA predictions to observations of real operational crew responding to simulated accidents. Among various lessons learned, significant performance variability was observed. As a result of team dynamics, work processes, communication strategies, sense of urgency, and willingness to take knowledge-based actions, the observed performances differed not only in terms of the rate of progress through the procedures but also in terms of paths through the procedure or even the applied procedures. Subsequent studies on simulator data further analyzed the variability of crew strategies to make decisions and solve conflicts, especially in cases of complex simulated emergencies that involve non-typical conditions with multiple malfunctions [47,48]. These studies provided important insights on the characterization of crew performance, error identification and analysis, and characterizations of procedures and interfaces; capturing this variability is necessary for the design of HRA databases, as well as when analysing specific failure events [47,48].

Recently, two important simulator data collection initiatives have been initiated: SACADA [17] and HuREX [16]. In a similar way as HRA methods, these data collection protocols operate over taxonomies of categorical factors. SACADA characterizes the context via the "situational factors" (e.g. "information quality", with the levels: "missing", "misleading", and "conflicting"), associated to high-level categories of individual and team cognitive functions (namely, "macrocognitive functions", e.g. "monitoring/detecting", "deciding/response planning"). Crew performance in a simulated scenario is evaluated according to a discrete rating classification (e.g. "satisfactory", "unsatisfactory", etc.) and the issues that negatively influenced the performance are classified in terms of both failure modes (e.g. "key alarms not detected or not responded to") and error causes (e.g. "multiple simultaneous alarms"). Similarly to SACADA, the HuREX protocol classifies performance failures in simulator data collection (namely, "unsafe acts") according to a categorical taxonomy based on cognitive activities (e.g. "situation interpreting"), generic task types (e.g. "measuring parameter - reading simple value", "transferring step in procedure"), error modes (e.g. "error of commission"), and contextual information relevant to the simulated scenario (e.g. "procedure conformity", "task familiarity").

An example of collected data tailored to the SACADA taxonomy is given in Table 2 (note that the table reports only few elements of the rich SACADA context characterization). It considers hypothetical data collected on the task type "understanding the situation/problem", where the alarm board of the Human Machine Interface (HMI) shows one status indication conflicting with critical alarms ("information quality: conflicting" in Table 2), the diagnosis of the latter being procedure-driven ("diagnosis basis: procedure" in Table 2). Table 2 includes failure/success data gathered in different plants (therefore with different HMIs, procedures, training programs) from different operating crews performing in two different simulated scenarios: for instance, 50 observations for Small Loss of Coolant Accident (SLOCA) scenario where the operators have to diagnose the SLOCA following a drop in pressurizer pressure, and 50 observations for Steam Generator Tube Rupture (SGTR) scenario where the operators have to diagnose the SGTR based on an anomalous variation of steam generator water level, given that in both situations a conflicting status indication is displayed (e.g. in SGTR scenario, "one level indication in steam generator stuck low" as in Table 2).

As the SACADA and HuREX databases are being populated, on-going research addresses the use of the collected data to inform HRA models. For example, Jung et al. (2020) [26] derive HEP values for the task categories addressed by the HuREX taxonomy [16], e.g. "directing manipulation", "entering step in procedure". Kim et al. (2018) [21] uses logistic regression analysis to estimate the quantitative relationships between PSFs and HEP values from a set of 10000 HuREX observations. In all these works, the relevant HEP values are estimated via a Bayesian update (e.g. the conjugated beta-binomial model): the HEP value associated to each taxonomy category is modeled as a unique value (i.e. the HEP population average), to be estimated based on the simulator data evidence. Returning to the example data in Table 2, when using this data to inform a quantitative HRA model on the considered "constellation" (i.e. combination) of task type and PSF ratings, the lumped-data approach would aggregate all observations as a single piece of evidence of 10 failures over 100 trials. This approach lumps together a number of variability aspects. Indeed, the dataset contains observations of tasks performed in different scenarios and different plants (e.g. "monitoring trend of steam generator level" in SGTR scenario, third column in Table 2), corresponding to different realizations of the associated task type (e.g. "understanding the situation/problem" in Table 2). The context of operation presents specificities that vary from plant to plant: in the example provided in Table 2, the HMI design of the alarm board in plant A is different from the one installed in plant B (e.g. different design and position of the alarms on screen; different number of simultaneous alarms); also, the specific procedural guidance and training program can vary between plant A and B. These plant-specific differences correspond to different realizations (second column in Table 2) of the associated contextual factors (e.g. "information quality: conflicting" and "diagnosis basis: procedure" in Table 2). Then, different crews are involved with crew-specific behavioral styles (e.g. different team dynamics, communication strategies, etc.).

A similar modelling approach with lumped data was adopted in a previous work by Groth et al. (2014) [20], where simulator observations from the US Empirical Study [15] were used in a Bayesian conjugate beta-binomial model with the goal to improve the reference HEP values of the SPAR-H method [7,8].

Concerning SACADA data, a number of feasibility studies have addressed the use of the collected data to inform HRA models [22–25], all based on variants of Bayesian approaches. Azarm et al. (2018) [24] proposes a multi-step methodology to identify critical situational factors for each macrocognitive function addressed by SACADA taxonomy [17] and uses a conjugate beta-binomial model to estimate HEP distributions for different combinations of these factors. Similarly to [20] and [26], the Bayesian estimates in [24] lump the data available for the relevant factor combination. Azarm et al. (2018) [24] acknowledge the presence of residual variability (e.g. plant-to-plant, crew-to-crew), but the authors average it out since the current amount of SACADA data does not allow a

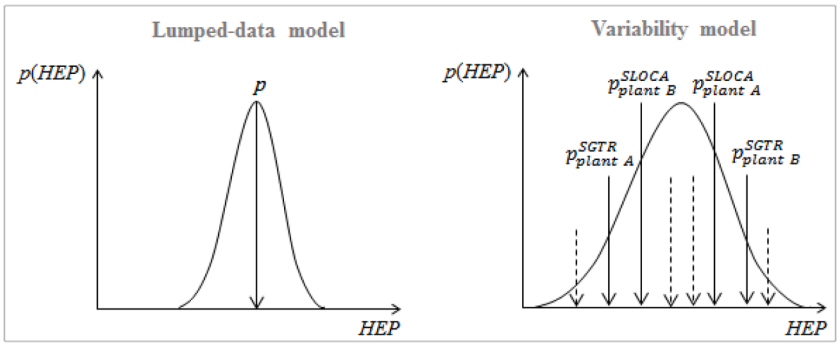


Fig. 1. Simplified comparison between lumped-data and variability models (generic distributions shown). Left: probability density as uncertainty on the HEP population average (lumped-data model). Right: probability density as variability and uncertainty on HEP values variable by source (variability model) (given a constellation of task type and PSF categories).

complete treatment of all sources of uncertainty. Other works adopt more advanced modelling techniques, specifically Bayesian Belief Networks (BBNs), to provide a richer characterization of the task, scenario, and context factors and of their relationships. Nelson & Grantom (2018) [25] use BBNs to model the relationships between situational factors and error modes per each macrocognitive function of SACADA data collection taxonomy, and produce HEP estimates conditional on the set of situational factors. Groth (2018) [23] propose a comprehensive framework combining SACADA data, taxonomies of performance influencing factors, causal BBNs, and Bayesian parameter updating to improve both the qualitative and quantitative basis of HRA models.

The BBN-based approaches [13,49,25] resort to a flexible framework to represent different variability aspects into the conditional probability distributions of the node categories and propagate this information through the BBN model. For instance, crew-to-crew variability nodes could be devised to explicitly represent the influence of different crew behavioral styles on the HEP. This calls for approaches to formally incorporate data variability (crew-to-crew, within-category) into the BBN conditional probability distributions. In this direction, the present work could support the development of empirically-based anchor information (i.e. reference HEP values and associated variability bounds) for multiple constellations of node categories of emerging BBN-based HRA models.

To summarize the above discussion, observations in simulator data collection (for a given constellation of task type and PSF categories) bring two aspects of variability into the HEP estimates: on one hand, the variability stemming from the different realizations of the associated constellation of factors of the HRA model (namely, "within-category" variability); on the other hand, the variability due to the different crew-specific features (namely, "crew-to-crew" variability). As formally presented in the next section, modeling variability entails considering the evidence from different realizations and different crews as multiple pieces of evidence, pertaining to a population of failure probability values. Fig. 1 illustrates the difference between the lumped (left) and the population variability (right) models with reference to the simplified data collection example of Table 2. It is important to note that in the lumped approach, the probability density function associated to the HEP value represents the uncertainty about the assumed unique value of the HEP itself (i.e. the population average). In the population variability approach, the function represents both the variability of the HEP value within the population and the uncertainty about the population parameters. For use in PSA, HEP values need to be plant- and scenario-specific; therefore, from Fig. 1, focusing on the population average, the lumped approach may not represent the intrinsic variability of the sources.

3. A Bayesian variability model for simulator data

This section presents the mathematical model to account for the two variability aspects relevant for HRA data collection from simulators: within the categories of the data collection taxonomy and crew-to-crew. After discussing the underlying modelling assumptions (Section 3.1), the variability model (Section 3.2) is then coupled to a hierarchical Bayesian model (Section 3.3) to infer from data on the parameter of the HEP variability distribution.

3.1. Modelling assumptions

The idea is to build a general quantitative tool, able to mathematically aggregate simulator data from nuclear power plants to estimate failure probabilities (with their variability and uncertainty distributions), for constellations of categorical elements (e.g. task type, set of PSF ratings) of a data collection taxonomy (e.g. SACADA, HuREX). The quantity of interest for the developed model is the HEP value associated to the given constellation, $\boldsymbol{F} = \{F_1, F_2 \ldots, F_\delta\}$:

$$HEP = f(F_1, F_2 \ldots, F_\delta) \quad (1)$$

where $\boldsymbol{F}$ is the set of δ categorical elements used by the taxonomy to represent the simulator data record (e.g. in Table 2, F_1 represents the task type "understanding the situation/problem", F_2 the PSF "information quality: conflicting", and F_3 the PSF "diagnosis basis: procedure"). Each F_i can be expressed as a binary (e.g. present / not present; adequate / not adequate) or a multi-valued (e.g. rating) variable, depending on the particular taxonomy.

Evidence on human performance from simulator data may come in different forms, depending on the aims of the simulator program, its scope, and the intended use of the data. In this study, we focus on data from large-scale simulator programs, in the form of records of failure/successes, while operators perform tasks under a specific combination of PSF states.

The proposed inference model is intended for general application to any HRA model for HEP quantification (the applicability is further discussed in Section 5). The following list briefly restates the key terminology used in Sections 1-2, in order to support the understanding of model development in the remainder of this section:

- "categories": refers to the taxonomy of task types and PSF levels adopted by the given data collection protocol (e.g. SACADA, HuREX) or HRA method. For instance, task type "diagnosis", or PSF "time available" with level "barely adequate";
- "constellation" (set $\boldsymbol{F}$ in this paper): refers to a combination of the aforementioned categories, e.g. $\boldsymbol{F}$: {task type = diagnosis, with PSFs: "time available" = "barely adequate", "diagnosis basis" = "procedure-directed check", etc.}. Generally, HRA models provide HEP estimates as function of these constellations: accordingly, the goal of the proposed model is to infer the HEP uncertainty distribution for a given constellation, from simulator data;

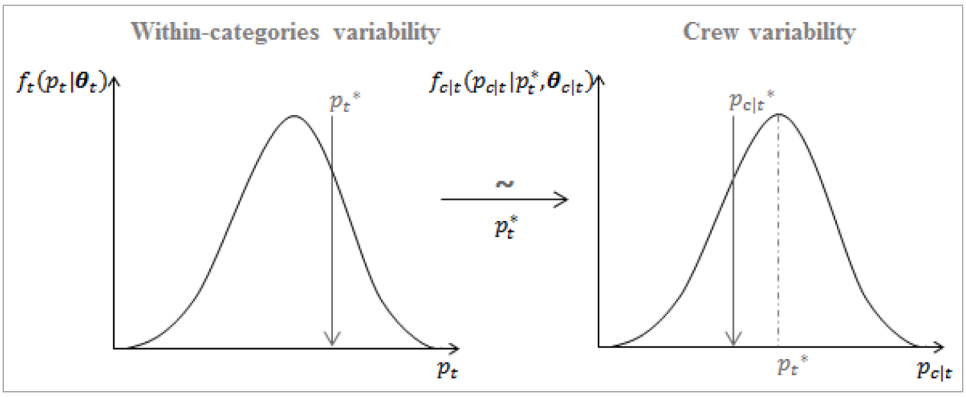


Fig. 2. Sketch of the variability model (generic distributions shown). HEP represented by a population variability distribution, $f_F(HEP|\boldsymbol{\theta}_F)$, combining variability within-category - $f_t(p_t|\boldsymbol{\theta}_t)$, on the left - and crew-to-crew - $f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$, on the right (see eq. 2). The crew-specific HEP variable, $p_{c|t}$, is distributed around the HEP value of a specific realization of the task and PSF constellation (p_t^* in right plot).

- "within-category" variability: refers to variability aspects stemming from the different scenario-specific tasks associated to the same task type (e.g. different realizations of the category "diagnosis"), as well as from the different plant-specific operational contexts associated to the same set of PSF levels (e.g. different realizations of "barely adequate time" for PSF "time available"). Hence the term "within-category", since the same category (i.e. a task type or a PSF level) envelopes different realizations, according to the data collection protocol;
- "crew-to-crew" variability: refers to variability aspects stemming from the different behavioral characteristics (e.g. different problem-solving styles, communication strategies etc.) of the operating crews.

3.2. Variability model for HEP

The core of the variability model is the formulation of the HEP as an inherently variable quantity, represented by a probabilistic variability distribution (the population variability), $HEP \sim f_F(HEP)$. The distribution function, $f_F(HEP)$, is assumed known (e.g. lognormal) and reflects both variability aspects in HEP estimates discussed earlier: within-category as well as crew-to-crew variability. The quantity to infer from evidence is the set of (unknown) parameters of the variability distribution, as opposed to the 'lumped-data' approach, where the unknown quantity is the unique HEP value (the population average).

The variability model, shown in Fig. 2, is based on the following concepts:

- each realization of a constellation of categorical elements of the taxonomy is characterized by a unique HEP, p_t. With reference to Table 2, one such realization is the task of transferring to the SGTR procedure, in case one level indication in the steam generator is stuck low, following the procedures of plant A, for instance with associated HEP p_t^*. Basically, a realization defines the simulator scenario and the specific task to be performed by the crew. In this interpretation, Table 2 includes six realizations of the same constellation "understanding the situation/problem" in case of "conflicting information quality", associated to six different values of p_t^*. Different plants determine different realizations, because, although enveloped by the same constellation, the PSF manifestations may be different (different procedures, different HMI interfaces, and so forth). Variable p_t is continuous, distributed according to a known distribution f_t with vector of unknown parameters $\boldsymbol{\theta}_t$: $p_t \sim f_t(p_t|\boldsymbol{\theta}_t)$. p_t is intended as the failure probability to perform the specific task manifestation in the specific context manifestation, defined by the simulator run design (hence, the pedix t, for "task").
- crew variability manifests as a crew-specific HEP variable $p_{c|t}$, that models the failure probability of a specific crew given the task performed in the specific simulator scenario, i.e. in a realization of the constellation $\boldsymbol{F}$ (e.g. from Table 2, the failure probability of one of the five crews from plant A performing the task "monitoring trend of steam generator level" in the corresponding SGTR scenario). It is assumed that the $p_{c|t}$ is a continuous variable distributed around each p_t^* according to a known distribution $f_{c|t}$, with unknown parameters $\boldsymbol{\theta}_{c|t}$: $p_{c|t} \sim f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$. Crew variability is modeled as variability of HEP values across different crews for the same task.

According to this formulation, the "HEP" variable in eq. (1) is represented by $p_{c|t}$, the probability of failure of a specific crew, given a specific task/context constellation.

Combining within-category and crew variability effects, the variability function $f_F(HEP = p_{c|t})$ can be expressed as:

$$f_F\left(p_{c|t}|\boldsymbol{\theta}_F\right) = f_F\left(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right) = \int f_t\left(p_t^*|\boldsymbol{\theta}_t\right) \cdot f_{c|t}\left(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t}\right) dp_t^* \quad (2)$$

where $\boldsymbol{\theta}_F = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$ is the vector of the unknown parameters of the overall HEP variability distribution.

It is important to stress that the model considers $p_{c|t}$ as a crew-specific HEP value (given the specific task and context realization corresponding to the simulator run). This means that the model foresees that the crew performance of a task in response to a specific simulator run (e.g. one of the scenarios in Table 2) is not deterministic. The probability value $p_{c|t}$ associated to a specific crew represents two aspects. On the one hand, it represents the fact that it is not possible to exactly foresee the crew behavior because of the complexity of the factors involved and of intrinsic limitations of human performance models (i.e. "model limitations" in Table 1). On the other hand, it represents the intrinsic variability of human performance, even in presence of the same crew in response to the same simulator run (e.g. response times, level of attention, alertness of the same person/crew vary over time, "crew characteristic" in Table 1). These two aspects are presented separately to ease the discussion, but of course are closely linked: some crew characteristics are considered as aleatory because of model limitations to foresee them.

Both $p_{c|t}$ and the variability function in eq. 2 reflect the aleatory uncertainty elements from Table 1. Epistemic (state-of-knowledge) uncertainty comes in the uncertainty associated to the parameters of the variability distribution ($\boldsymbol{\theta}_F$). Ideally, as more data is collected, $\boldsymbol{\theta}_F$ would be progressively better estimated, with the epistemic component progressively decreasing, and consequently the expected $p_{c|t}$ distribution would get closer to the true (unique) HEP variability distribution for the constellation $\boldsymbol{F}$; the limiting case, with infinite data available, would be that the expected distribution only represents the inherent variability of the HEP. This aspect highlights a significant difference with the lumped

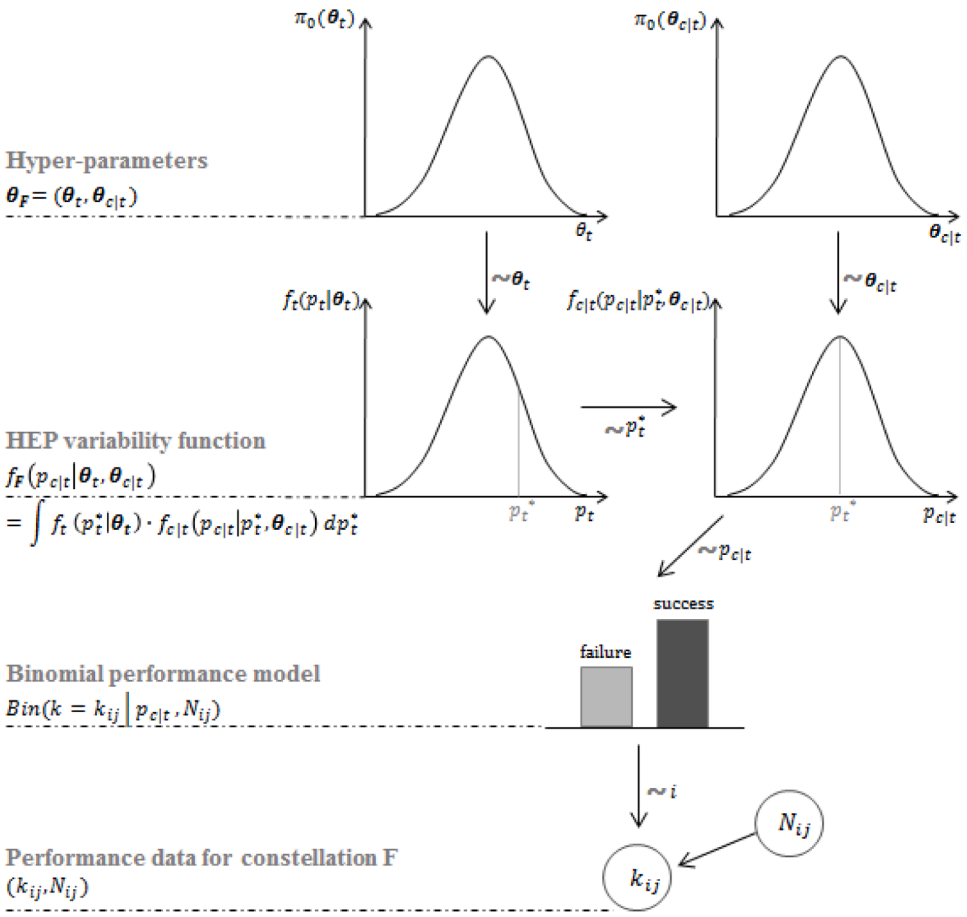


Fig. 3. The Bayesian hierarchical variability model, from top to bottom: $\pi_0(\boldsymbol{\theta_F})$, prior distributions for model parameters $(\boldsymbol{\theta_F})$; $f_F(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$, the HEP variability distribution, where $f_t(p_t|\boldsymbol{\theta}_t)$ models within-category variability and $f_{c|t}(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t})$ models crew-to-crew variability; $Bin(k = k_{ij}|p_{c|t}, N_{ij})$, the binomial distribution of evidence of k_{ij} failures on N_{ij} repetitions of the i-th task by the j-th crew. Generic distributions shown.

approach, where a unique HEP (i.e. the population average) is the unknown parameter. In the lumped configuration, with increasing evidence, the uncertainty distribution will narrow to the unique estimate.

The hierarchical Bayesian model is implemented to update the analyst's degree of belief on the set $\boldsymbol{\theta_F}$ and finally derive the estimated uncertainty distribution of $p_{c|t}$.

3.3. Development of the Bayesian inference model

Fig. 3 gives an overview of the hierarchical Bayesian model. The

Simulator data available for the constellation *F*:
Task type: understanding the situation/problem
Information quality: conflicting
Diagnosis basis: procedure

Scenario	Realization of contextual factors	Task (index *i*)	Plant	Crew (index *j*)	Repetitions (N_{ij})	Failures (k_{ij})
SGTR	One level indication in steam generator stuck low.	Monitoring trend of steam generator level	A	1	1	0
			A	2	1	0
			(...)	(...)	(...)	(...)
			B	1	1	1
			B	2	1	0
sLOCA	One indication on pressurizer pressure stuck high.	Monitor pressurizer pressure	A	1	1	0
			A	2	1	1
			(...)	(...)	(...)	(...)

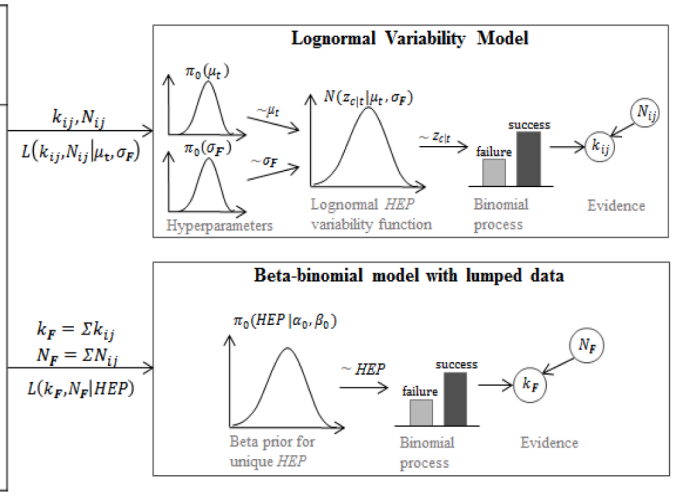


Fig. 4. Overall aggregation framework to compare the variability and the lumped data models. Left: artificial data for the constellation ***F*** based on the example in Table 2.Top right: lognormal variability model, informed by the crew-specific data points (k_{ij}, N_{ij}) and returning as output the posteriors for the HEP variability distribution parameters, i.e. μ_t and σ_F. Bottom right: conjugated beta-binomial model with lumped data (k_F, N_F), giving as output the posterior distribution for the single-value HEP (population average).

general structure of the model is based on the formulation of the Bayes theorem as follows [38,50]:

$$\pi(\boldsymbol{\theta}|E) = A^{-1}L(E|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) \tag{3}$$

where:

- $\boldsymbol{\theta}$ is the set of unknown parameters of the inference problem;
- π_0 and π are the prior and posterior probability functions for $\boldsymbol{\theta}$, modelling the state of knowledge of the analyst on the set of investigated parameters respectively before and after the evidence$E$ is collected (top level in Fig. 3);
- $L(E|\boldsymbol{\theta})$ is the likelihood term, interpreted as the probability density that the evidence is observed (second and third levels in Fig. 3);
- E is the set of evidence from the available information sources (bottom level in Fig. 3);
- $A^{-1} = \int L(E|\theta)\pi_0(\theta)d\theta$, the denominator of eq. 3, normalizes function π to a probability density function.

For the variability model in Section 3.1, $\boldsymbol{\theta_F} = \{\boldsymbol{\theta}_t, \boldsymbol{\theta}_c\}$ is the set of unknown parameters of the parametric variability function $f_{\boldsymbol{F}}(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t})$.

Empirical evidence comes in the form of failure data (i.e. number of failures on number of task repetitions) collected on crew performance on simulator scenarios characterized by the same constellation $\boldsymbol{F}$. It is assumed that data was collected concerning m different task/context realizations within constellation $\boldsymbol{F}$, and n_i crews that performed the i-th task. Evidence E is represented as the set of pairs $\{(k_{ij}, N_{ij}), i = 1, 2 \ldots, m, j = 1, 2 \ldots, n)\}$, where k_{ij} is the number of failures observed on N_{ij} repetitions of the i-th task performed by the j-th crew (Fig. 4, left, columns "Repetitions" and "Failures"). This type of datasets enters the likelihood term of the Bayesian model as evidence to update the prior degree of belief of the analyst on the parameters of the HEP variability model for the constellation $\boldsymbol{F}$ (Fig. 4, right). Note that in the numerical examples, N_{ij} is set equal to 1 (see Fig. 4, column "Repetitions"), recognizing that it would be very difficult to aggregate performances on the exact same task by the exact same crew (this aspect will be further discussed in Section 4 and in Section 5).

The construction of the likelihood term $L(E|\vartheta_t, \vartheta_{c|t})$ requires to express the probability of observing k_{ij} failures on N_{ij} repetitions of the specific i-th task. For the generic piece of simulator evidence, (k_{ij}, N_{ij}), the likelihood term can be written as:

$$L_{ij}\left(k_{ij}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}\right) = \int_{p_{c|t}} f_{\boldsymbol{F}}\left(p_{c|t}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right) Bin\left(k = k_{ij}|p_{c|t}, N_{ij}\right) dp_{c|t} \tag{4}$$

By substituting eq. 2 into eq. 4, the likelihood term becomes:

$$\begin{aligned} &L_{ij}\left(k_{ij}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}\right) \\ &= \int_{p_t}\int_{p_{c|t}} f_t(p_t|\boldsymbol{\theta}_t) f_{c|t}\left(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}\right) Bin\left(k = k_{ij}|p_{c|t}, N_{ij}\right) dp_{c|t}dp_t \end{aligned} \tag{5}$$

where:

- the probability density that the failure probability of the i-th specific task is p_t, i.e. p_t is one realization of the possible within-category variability, modeled by $f_t(p_t|\boldsymbol{\theta}_t)$;
- the probability density that the crew-specific HEP value would manifest as $p_{c|t}$ (i.e. one realization of the possible crew-to-crew variability) is modeled by $f_{c|t}(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t})$. The task-specific HEP, p_t, constitutes the reference probability value around which $p_{c|t}$ is distributed;
- the probability of observing k_{ij} failures in N_{ij} repetitions of the i-th task if the failure probability for the single repetition $p_{c|t}$ is described by the binomial distribution $Bin(k = k_{ij}|p_{c|t}, N_{ij})$.

Each probability value p_t and $p_{c|t}$ is one possible value within their variability; therefore, the expression $f_t(p_t|\boldsymbol{\theta}_t)f_{c|t}(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t})Bin(k = k_{ij}|p_{c|t}, N_{ij})$ is averaged (integrated) on the variability distributions for p_t and $p_{c|t}$.

When the i-th task is performed by n_i crews, the evidence takes the form of the number of failures observed for each crew: (k_{i1}, N_{i1}), (k_{i2}, N_{i2}), … (k_{in_i}, N_{in_i}). The likelihood term L_i relevant to the i-th task becomes:

$$\begin{aligned} &L_i\left(k_{i1}, k_{i2}, \ldots, k_{in_i}|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{i1}, N_{i2}, \ldots, N_{in_i}\right) \\ &= \int_{p_t} f_t(p_t|\boldsymbol{\theta}_t) \prod_{j=1}^{n_i} \int_{p_{c|t}} f_{c|t}\left(p_{c|t}|p_t^*, \boldsymbol{\theta}_{c|t}\right) Bin\left(k = k_{ij}|p_{c|t}, N_{ij}\right) dp_{c|t}dp_t \end{aligned} \tag{6}$$

Note that in the expression above the probability density of observing the evidence (k_{i1}, N_{i1}), (k_{i2}, N_{i2}), … (k_{in_i}, N_{in_i}) given the within-category reference probability p_t is written as:

$$\prod_{j=1}^{n_i} \int_{p_{c|t}} f_{c|t}\left(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}\right) Bin\left(k = k_{ij}|p_{c|t}, N_{ij}\right) dp_{c|t}$$

Since all crews are carrying out the same specific task, the crew-to-crew variability effect is expressed for all crews conditional on the same reference HEP value, p_t. Then, the probability density of observing each (k_{ij}, N_{ij}) is multiplied because, given p_t, each crew's behavior is independent (the effect of the PSFs common for all crews is represented in the variable p_t).

Extending eq. 6 to the entire set of m task realizations in the constellation $\boldsymbol{F}$, the likelihood term is then:

$$\begin{aligned} L\left(E|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right) &= L\left(k_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_j|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}\right) \\ &= \prod_{i=1}^{m} L_i\left(k_{ij}, j = 1, \ldots, n_j|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}, N_{ij}\right) \\ = \prod_{i=1}^{m} \int_{p_t} f_t(p_t|\boldsymbol{\theta}_t) &\prod_{j=1}^{n} \int_{p_{c|t}} f_{c|t}\left(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}\right) Bin\left(k = k_{ij}\middle|p_{c|t}, N_{ij}\right) dp_{c|t}dp_t \end{aligned} \tag{7}$$

Eq. 7 assumes that the failure observations across the different tasks are independent. This implies that crew variability effects on the crew-specific HEP variable, $p_{c|t}$, do not replicate across different tasks: in other words, no systematic effects of crew under-performance (i.e. crew-specific HEP value consistently above average) or over-performance (i.e. crew-specific HEP value consistently below average) are modeled.

The posterior degree of belief on the unknown parameters of the HEP variability distribution for a generic constellation $\boldsymbol{F}$ of task and PSF categories is then expressed as follows:

$$\pi\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}|E\right) = \frac{L\left(E|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right)\pi_0\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right)}{\iint L\left(E|\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right)\pi_0\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}\right)d\boldsymbol{\theta}_t d\boldsymbol{\theta}_{c|t}} \tag{8}$$

where the final formulation can be derived by substituting the likelihood term of eq. 7 in eq. 8.

The posterior probability distribution of eq. 8 can be subsequently used to compute the estimated HEP variability distribution for the constellation $\boldsymbol{F}$, $P_{\boldsymbol{F}}(p_{c|t})$:

$$\begin{aligned} P_{\boldsymbol{F}}\left(p_{c|t}\right) &= \int_{\boldsymbol{\theta_F}} f_{\boldsymbol{F}}\left(p_{c|t}|\boldsymbol{\theta_F}\right)\pi(\boldsymbol{\theta_F}|E)d\boldsymbol{\theta_F} \\ = \int_{\boldsymbol{\theta}_t}\int_{\boldsymbol{\theta}_{c|t}}\int_{p_t} f_t(p_t|\boldsymbol{\theta}_t) &f_{c|t}\left(p_{c|t}|p_t, \boldsymbol{\theta}_{c|t}\right)\pi\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}|E\right)dp_t d\boldsymbol{\theta}_{c|t}d\boldsymbol{\theta}_t \end{aligned} \tag{9}$$

Formally, $P_{\boldsymbol{F}}(p_{c|t})$ is derived by weighting the parametric distribution, adopted as variability model for HEP, by the posterior distribution of the unknown HEP distribution parameters computed by the Bayesian model.

Within this mathematical framework, the incorporation of further empirical evidence can be accomplished in subsequent steps in a traceable and reproducible way. This feature is of key importance, considering that data collection process from simulators is a long-term program. Indeed, the posterior distributions of HEP computed by the

model can be used as prior state of knowledge in future analyses and then updated as new observations become available.

Finally note that the "lumped-data" approaches, e.g. of [20] and [26], entail aggregating the evidence to inform a unique HEP value for the constellation $\boldsymbol{F}$ (i.e. the population average), i.e.:

$$k_{\boldsymbol{F}} = \sum_{i=1}^{m}\sum_{j=1}^{n} k_{ij}, N_{\boldsymbol{F}} = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \tag{10}$$

where $k_{\boldsymbol{F}}$ and $N_{\boldsymbol{F}}$ are respectively the total number of failures and observations aggregated for the constellation $\boldsymbol{F}$ (Fig. 4, bottom right). In [20] and [26], the pair ($k_{\boldsymbol{F}}$; $N_{\boldsymbol{F}}$) enters a conjugate beta-binomial model to update the prior state of knowledge on the population-average HEP, represented by a beta distribution with shape parameters α_0 and β_0. The update with lumped-data,

$$\alpha = \alpha_0 + k_{\boldsymbol{F}},\ \beta = \beta_0 + N_{\boldsymbol{F}} - k_{\boldsymbol{F}} \tag{11}$$

yields the posterior distribution of the beta-binomial model (again a beta distribution, with parameters α and β), representing the final uncertainty on the population-average HEP.

3.4. Use of lognormal probability density functions to represent variability

This section presents the model in case lognormal distributions are used to represent both variability terms in eq. 2, within-category and crew variability, f_t and $f_{c|t}$, respectively (Fig. 4, top right) – this configuration will be used in the applications in Section 4. The adoption of lognormal functions as population variability curves has been a common practice when developing hierarchical Bayesian models for PSA applications [27,29,51].

Considering a generic constellation of categorical elements $\boldsymbol{F}$, in this configuration both variability terms embodied in $f_{\boldsymbol{F}}(p_{c|t}|\vartheta_{\boldsymbol{F}})$ as in eq. 2 (within-category and crew-to-crew variability) are distributed accordingly to lognormal probability density functions, therefore:

$$\begin{aligned}\ln(p_t) = z_t \sim N(z_t|\mu_t,\sigma_t); \ln\left(p_{c|t}\right)\\ = z_{c|t} \sim N\left(z_{c|t}|z_t,\sigma_{c|t}\right)\end{aligned} \tag{12}$$

where z_t and $z_{c|t}$ are the normally-distributed auxiliary variables associated to p_t and $p_{c|t}$, respectively (the letter N is used in eqs. 12-14 and Fig. 4 to denote normal distributions). In this case, the set of unknown parameters to be determined by the Bayesian inference model is then $\boldsymbol{\theta_F} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_{c|t}) = (\mu_t,\ \sigma_t, \sigma_{c|t})$. Subsequently, the likelihood term for the generic piece of simulator evidence (eq. 5) can be expressed as follows:

$$\begin{aligned}&L_{ij}\left(k_{ij}|\mu_t,\sigma_t,\sigma_{c|t},N_{ij}\right)\\ &= \int_{z_t}\int_{z_{c|t}} N(z_t|\mu_t,\sigma_t)N\left(z_{c|t}|z_t,\sigma_{c|t}\right)Bin\left(k=k_{ij}|e^{z_{c|t}},N_{ij}\right)dz_{c|t}dz_t\end{aligned} \tag{13}$$

Rearranging the right-side member of the equation:

$$\begin{aligned}&L_{ij}\left(k_{ij}\middle|\mu_t,\sigma_t,\sigma_{c|t},N_{ij}\right)\\ &= \int_{z_{c|t}} Bin\left(k=k_{ij}|e^{z_{c|t}},N_{ij}\right)\left(\int_{z_t} N(z_t|\mu_t,\sigma_t)N\left(z_{c|t}\middle|z_t,\sigma_{c|t}\right)dz_t\right)dz_{c|t}\\ &= \int_{z_{c|t}} Bin\left(k=k_{ij}|e^{z_{c|t}},N_{ij}\right)N\left(z_{c|t}\middle|\mu_t,\sqrt{\sigma_t^2+\sigma_{c|t}^2}\right)dz_{c|t}\\ &= \int_{z_{c|t}} Bin\left(k=k_{ij}|e^{z_{c|t}},N_{ij}\right)N\left(z_{c|t}|\mu_t,\sigma_{\boldsymbol{F}}\right)dz_{c|t}\end{aligned} \tag{14}$$

The last relationship exploits the fact that the convolution of the two normal distributions of z_t and $z_{c|t}$ is again a normal distribution, with mean μ_t and standard deviation $\sigma_{\boldsymbol{F}} = sqrt(\sigma_t^2 + \sigma_{c|t}^2)$. According to eq. 14, the final set of unknown parameters for the inference problem becomes $\boldsymbol{\theta_F} = (\mu_t,\ \sigma_{\boldsymbol{F}})$, which respectively represent the mean and the standard deviation of the HEP variability distribution in the logarithmic space.

The extension of eq. 14 to the entire set of simulated observations relevant to $\boldsymbol{F}$ (see eq. 6-7), as well as the specialization of the posterior formula to the new set of unknown parameters (see eq. 8), are done as in Section 3.2.

The last step of the Bayesian model development entails the definition of appropriate prior distributions for the parameters of the lognormal variability model, namely $\pi_0(\mu_t)$ and $\pi_0(\sigma_{\boldsymbol{F}})$ (usually referred to in Bayesian literature as the "hyper-priors" of a hierarchical model, see [50]). In the model application presented in Section 4, both diffuse and informative priors are used for the hyper-parameters of the Bayesian model, μ_t and $\sigma_{\boldsymbol{F}}$. For the case of diffuse priors, as suggested in [51] for lognormal variability distributions in lack of information, uniform distributions are adopted for both the natural logarithm of the mean, $\pi_0(log(\mu_t))$, defined between natural log(1E-5) and 0 (corresponding to the upper limit HEP= 1), and the standard deviation, $\pi_0(\sigma_{\boldsymbol{F}})$, defined between 0.1 and 4 (corresponding to error factors of 1.18 and 720.54, respectively). These ranges have been defined to cover values of interest for HRA applications. More information on the development of proper prior distributions can be found in literature [27,50].

For all applications, an algorithm has been developed for the R programming environment [52] for the numerical solution of the various equations. The developed R code is available on request to the authors.

4. Numerical application

After a first comparison of the proposed variability model with a lumped data model (Section 4.1), the present section addresses the model sensitivity to data availability, both in presence of diffuse (Section 4.2.1), as well as of informed priors (Section 4.2.2). Artificial data is used, i.e. data generated with known characteristics (e.g. median, mean, percentiles of the underlying data distributions): this allows investigating the Bayesian update process, for which the known values become target values. An application to simulator data from literature [20,26] is presented later (Section 4.3).

Concerning the generated data, two cases of target HEP variability distribution are considered, both lognormal:

- Case 1: median = 5e-2, mean = 5.46e-2, and error factor = 2
- Case 2: median = 5e-3, mean = 6.25e-3, and error factor = 3

The two cases represent HEP ranges of practical interest for HRA, with relatively high (Case 1) and moderate (Case 2) HEP values. The case of lower HEP values (e.g. median 5e-4 or lower) is not considered in this paper because, as it will become clear later in the result presentation, the use of the proposed model would require a very large amount of simulator data, of questionable practicality.

Each data element is generated by first sampling a possible HEP value from the variability distribution for Case 1 or 2. Recalling from Section 3.2, this HEP value is crew-specific. Then, the realization of the number of observed failures, k_{ij}, on N_{ij} repetitions (by the same crew) is sampled from a Binomial distribution, obtaining the data element (k_{ij}, N_{ij}). Different couples (k_{ij}, N_{ij}) are generated from different HEP values, based on the total number of task realizations relevant to the constellation $\boldsymbol{F}$ assumed to be available from the simulator data collection (referred as $N_{\boldsymbol{F}}$ in Section 3), and constitute the evidence against which the variability model has been tested. For the applications in this paper, N_{ij} is set to 1: each crew performs the same task only once in the dataset. This corresponds to the lowest possible amount of information on the variability in HEP. Ideally, as simulator data is accumulated over the years, evidence on multiple repetitions may be available (for example some simulator scenarios are trained recurrently by the same crew). This aspect will be returned to in the discussion. To investigate the data requirements, different sample sizes are considered, from relatively small sets (e.g. $N_{\boldsymbol{F}} = 10 \div 50$) to larger sets (e.g. $N_{\boldsymbol{F}} = 200 \div 1000$), to reflect possibly different data availability in the long-term. Note that while N_{ij}

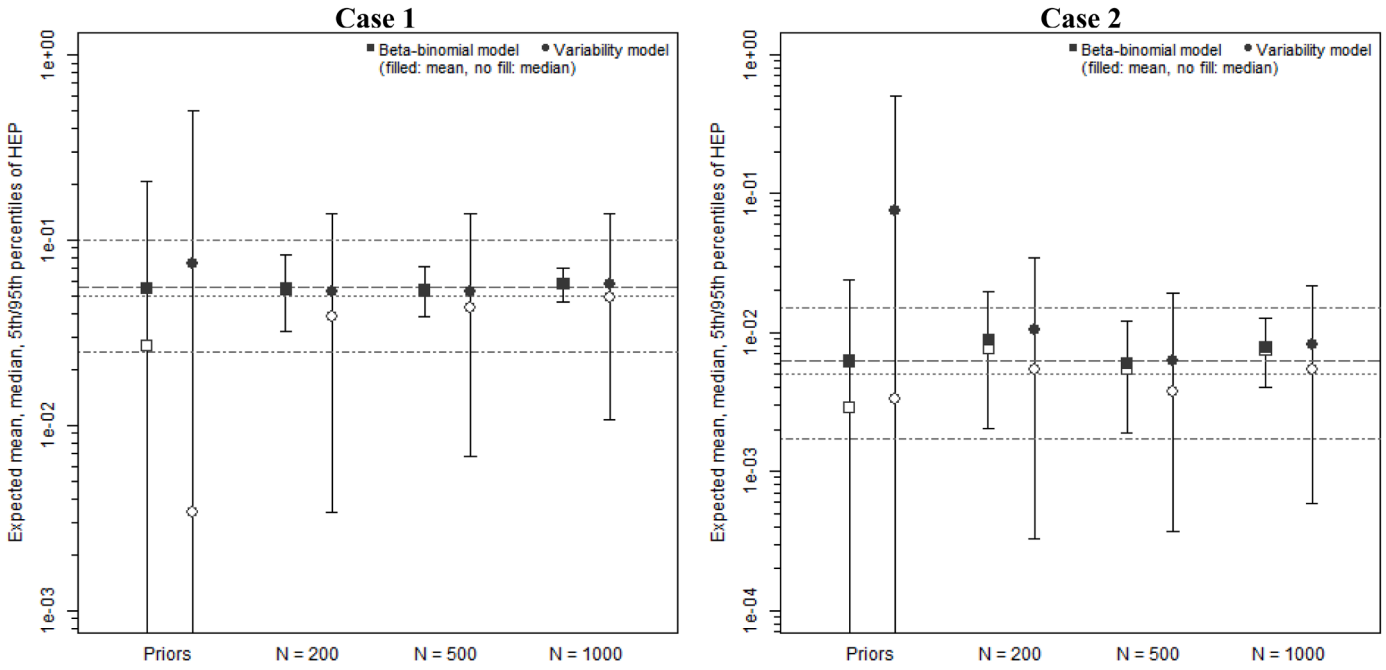


Fig. 5. Expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of $P_F(HEP)$ by the lognormal variability model and the lumped-data beta-binomial model, tested against the same simulator datasets (number of simulated tasks: 200, 500, 1000). Datasets are artificially generated from lognormal HEP variability distribution with: median 5e-2 (dotted line), mean 5.46e-2 (dashed line) and error factor 2 (dot-dashed lines at 5th percentile 2.5e-2 and 95th percentile 1.0e-1) for Case 1 (left); median 5e-3 (dotted line), mean 6.25e-3 (dashed line) and error factor 3 (dot-dashed lines at 5th percentile 1.7e-3 and 95th percentile 1.5e-2) for Case 2 (right).

refers to crew-specific evidence, N_F refers to the whole data accumulated for the constellation $\boldsymbol{F}$ from different plants, crews, as well as realizations of the task types and PSFs defined by $\boldsymbol{F}$: this justifies the possibility to accumulate evidence on the order of 1000 data points for the estimation of the parameters $\boldsymbol{\theta_F}$ of the variability function.

4.1. Variability model vs lumped-data approach

With reference to the two Cases 1 and 2, this section presents the numerical differences between the proposed variability model and a beta-binomial model representative of the lumped-data approach. Fig. 5 and Table 3 show the results. In both Cases 1 and 2, the expected mean, median, 5th and 95th percentiles of the $P_F(p_{c|t})$ estimated by the lognormal variability model are compared with the respective statistics provided by a beta-binomial model, with increasing sample sizes (200, 500, and 1000 observations, in x-axis). Consistently with the variability model, the beta-binomial model (eq. 11) uses a diffuse prior, in particular the CNI prior, as in [20] (with parameters: $\alpha_0 = 0.5$, $\beta_0 = 8.66$ for Case 1; $\alpha_0 = 0.5$, $\beta_0 = 79.5$ for Case 2). Comparing the expected error factors, the beta-binomial model provides a $P_F(p_{c|t})$ that is overly-narrow with respect to the target HEP variability distribution, with values of error factor significantly smaller than the target one (Table 3, with target values of 2 and 3 for Case 1 and Case 2, respectively). On the other hand, the lognormal variability model provides broader $P_F(p_{c|t})$'s, with error factors larger than the target values and tending to decrease to the target error factor with increasing sample sizes. While still larger that the target values, at 1000 observations the error factors reach the values of about 5 (Table 3), which starts to be of practical use for PSA applications (see analysis in the next Section 4.2). Indeed, the larger error factors from the variability model compared to the beta-binomial as well as the decreasing tendency are not surprising: the important point for the practical application of the proposed model is to investigate the model data requirements for practical applications. This will be the goal of the next Section 4.2. Concerning the estimated mean and median, both models tend to converge to the target values, as expected with slower convergence for Case 2.

To show the practical implications if variability is not modelled, assume plant-specific data is collected to infer the plant-specific HEP of a PSA operator action, with plant data from ten operating crews (Table 4). Assume also that data is available from simulator databases on the corresponding constellation (e.g. the case $N_F = 200$, Table 3). The data can be used as prior, then updated by the plant-specific data. Table 4 shows the difference in the posterior estimates depending on whether the prior distribution for the HEP is constructed with the lumped data model (Table 3, "lumped posterior", $N_F = 200$) or the variability model (Table 3, "Var. model posterior", $N_F = 200$). Three hypothetical data outcomes are considered, with increasing number of observed failures across the ten crews (Table 4, first column: 0, 1, and 2 failures). Given the plant-specific nature of the task (i.e. same scenario, same context of operation: no within-category variability in data), the observations from the ten different crews are all treated as "lumped", neglecting the underlying crew-to-crew variability aspects in performance, and entered as unique data point in a simple beta-binomial model. Depending on the data outcome, the posterior distribution may become very different. In general, the variability model is more sensitive to the new data as compared to the lumped one. For the considered example, as the number of observed failures increases, the posterior mean for the variability model moves closer to the frequentist estimate (0.1, 0.2 for the 1 and 2 failure cases, respectively). Intuitively, this is due to the fact that the prior for the variability model represents larger variability of performance conditions and crew behaviours, which may also include those characteristic of the plant under consideration. On the other hand, the lumped data prior is narrowed to the population average, which may represent a biased initial value for the specific plant. Mathematically, as the evidence deviates from the population average, the likelihood of the evidence is multiplied by a smaller likelihood value for the lumped data prior (more peaked) compared to the variability model prior (more diffuse).

Table 3
Comparison between the lognormal variability model and the beta-binomial: numerical results for Cases 1 and 2 (from Fig. 5). Number of simulated tasks: 200, 500, 1000.

Case 1 - target statistics: median = 5e-2, mean = 5.46e-2, and EF = 2						
	Model (pdf)	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
	Lumped (CNI prior)	5.50e-2	2.69e-2	2.36e-4	2.06e-1	29.54
	Variability model (prior)	7.44e-2	3.35e-3	2.01e-5	4.98e-1	157.39
N_F=200, 11 failures	Lumped (posterior)	5.50e-2	5.36e-2	3.18e-2	8.31e-2	1.62
	Var. model (posterior)	5.24e-2	3.85e-2	3.35e-3	1.38e-1	6.43
N_F=500, 27 failures	Lumped (posterior)	5.40e-2	5.34e-2	3.86e-2	7.14e-2	1.36
	Var. model (posterior)	5.29e-2	4.33e-2	6.73e-3	1.38e-1	4.53
N_F=1000, 58 failures	Lumped (posterior)	5.80e-2	5.77e-2	4.64e-2	7.05e-2	1.23
	Var. model (posterior)	5.75e-2	4.86e-2	1.07e-2	1.38e-1	3.59
Case 2 – target statistics: median = 5e-3, mean = 6.25e-3, and EF = 3						
	Model (pdf)	**Mean**	**Median**	**5th perc**	**95th perc**	**EF**
	Lumped (CNI prior)	6.25e-3	2.87e-3	2.48e-5	2.39e-2	31.07
	Variability model (prior)	7.44e-2	3.35e-3	2.01e-5	4.98e-1	157.39
N_F=200, 2 failures	Lumped (posterior)	8.93e-3	7.79e-3	2.06e-3	1.97e-2	3.09
	Var. model (posterior)	1.05e-2	5.34e-3	3.27e-4	3.43e-2	10.24
N_F=500, 3 failures	Lumped (posterior)	6.03e-3	5.48e-3	1.87e-3	1.21e-2	2.54
	Var. model (posterior)	6.25e-3	3.76e-3	3.68e-4	1.92e-2	7.22
N_F=1000, 8 failures	Lumped (posterior)	7.87e-3	7.57e-3	4.02e-3	1.27e-2	1.78
	Var. model (posterior)	8.09e-3	5.34e-3	5.86e-4	2.15e-2	6.06

Table 4
Example of HEP estimation for a plant-specific task: prior distribution from lumped-data model (Table 3, "lumped posterior", N_F = 200) and from the variability model (Table 3, "Var. model posterior", N_F = 200).

	Prior from lumped-data model			**Prior from variability model**			
Evidence	**Mean**	**Median**	**EF**	**Mean**	**Median**	**EF**	**Δ% mean**
Priors	5.50e-2	5.36e-2	1.62	5.24e-2	3.85e-2	6.43	+ 5%
0 failures, 10 trials	5.25e-2	5.11e-2	1.62	3.81e-2	3.00e-2	4.94	+ 38%
1 failures, 10 trials	5.71e-2	5.57e-2	1.58	6.54e-2	5.76e-2	3.09	- 13%
2 failures, 10 trials	5.81e-2	6.03e-2	1.55	9.26e-2	8.53e-2	2.48	- 37%

4.2. Sensitivity to available data

The collection of simulator data is resource-intensive and requires important time and money investments [22]: it becomes important to investigate the amount of data required such that the estimates produced by the model are of practical use (i.e. the associated uncertainties are not too large). In this section, for Case 1 and Case 2, convergence of the posterior statistics is followed as the available sample size increases. The error factor is particularly important for practical applications: too large error factors (e.g. 10, meaning a factor of 100 between the 95th and the 5th percentiles) entail diffuse posterior estimates of limited practical use. The aim of this section is to investigate the sample size required to obtain error factors comparable to those typical for HRA, e.g. around 5. Indeed, this sample size depends on the variability distribution of the HEP to be estimated. As already mentioned, the two cases 1 and 2 are deemed as representative of the range of interest for practical applications: larger HEP values (e.g. ~ 0.1) can be expected to be less problematic to estimate, while smaller values (e.g. below 0.001) may require too large data sizes for being of practical interest (at least with the model presented in this paper).

4.2.1. Diffuse priors

Fig. 6 shows the posterior estimates by the variability model, set up with flat hyper-priors $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ as a function of the sample size N_F (from $N_F = 50$ to 1000) for Cases 1 (Fig. 6, top) and 2 (Fig. 6, bottom). From left to right, the figures report the estimated posterior error factor, mean, and median. For each sample size, 100 datasets are sampled to represent the spread of the posterior estimates (each estimate represented by a dot in the figures).

From Fig. 6, the expected statistics of $P_F(p_{c|t})$ across the different datasets tend to converge to the target statistics as the sample size increases. The expected mean and median, averaged over the Monte Carlo samples, get close to their target values, at $N_F \approx 200$ for Case 1 and at $N_F \approx 250$ Case 2. Indeed, for Case 1 at $N_F \approx 200$, the average expected mean is 5.3e-2, with 50% confidence interval (25th - 75th percentiles) of (4.2e-2, 6.2e-2), and the average expected median is 3.9e-2, with 50% confidence interval of (2.9e-2, 4.9e-2); for Case 2 at $N_F \approx 250$, the average expected mean is 7.3e-3, with 50% confidence interval (25th - 75th percentiles) of (4.6e-3, 8.4e-3), and the average expected median is 3.7e-3, with 50% confidence interval of (2.1e-3, 4.2e-3).

The speed of convergence of the expected error factors is lower compared to the mean and median. For instance, for Case 1, 300 observations are approximately needed to observe an average expected error factor close to 5, i.e. 5.5 at $N_F \approx 300$, with 50% confidence interval (4.8, 6.1). For Case 2, with $N_F \approx 1000$, the average expected error factor is 6.1, with 50% confidence interval (6.1, 6.1, note the 25th and 75th percentiles match because of numerical discretization). Indeed, the speed of convergence to the target values depends on the amount of evidence at disposal. As the HEP values progressively decrease, fewer failure are observed (i.e. Monte Carlo sampled): as anticipated, for cases with lower HEP values (e.g. below 0.001), the model would require an impracticably large data size (e.g. above 10^4 data points).

In conclusion, this sensitivity analysis shows that for constellations $\boldsymbol{F}$ characterized by HEP values in the range ~ $0.1 \div 0.001$, the variability model with diffuse hyper-priors can provide results of practical value for HRA applications with few hundred data points. The latter data requirement are met by the current availability of data points for many constellations $\boldsymbol{F}$ in SACADA [22] and HuREX [26]. When lower HEP values are involved (e.g. HEP ~ 0.001 and below), the adoption of informative prior distributions may be a viable option to decrease the data requirements, as presented in the next Section 4.2.2.

4.2.2. Informative priors

This section investigates how much data requirements can be reduced with informative hyper-priors for both parameters μ_t and σ_F. Case 1 and Case 2 are addressed in Figs. 7, 8 and Figs. 9, 10, respectively. Values are reported in Tables A.1-A.2 in Appendix A.

In Fig. 7, two configurations can be distinguished: only the mean HEP is informed (left plot), both mean and standard deviation are informed (right plot). Both plots show the effect of different

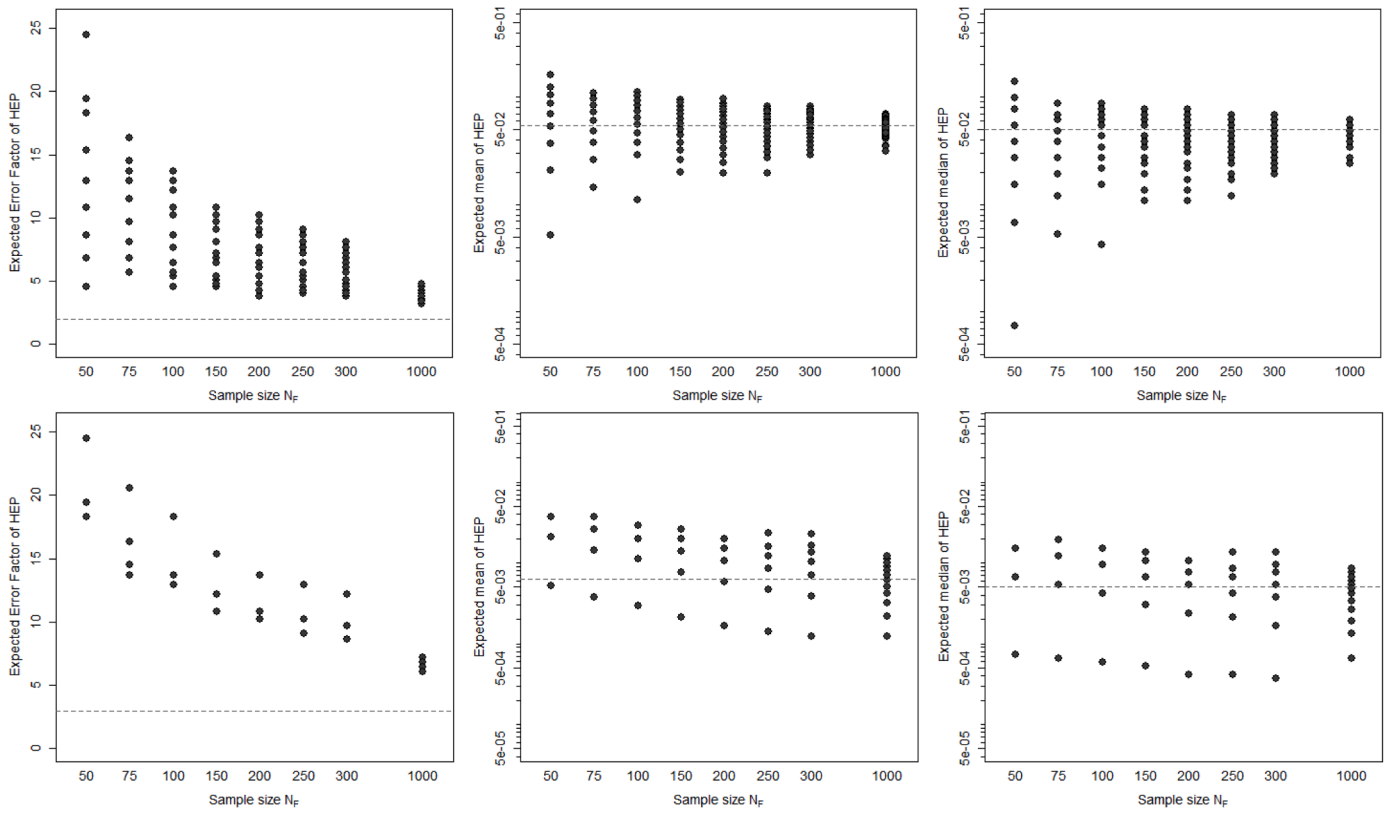


Fig. 6. Data requirements of the lognormal variability model with flat hyper-priors $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$. Top: Case 1 (median = 5e-2, mean = 5.46e-2 and error factor = 2. Bottom: Case 2 (median = 5e-3, mean = 6.24e-3, and error factor = 3. For each sample size (x-axis), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right: expected error factor, mean (log-scale), and median (log-scale) of the $P_F(HEP)$'s returned by the model.

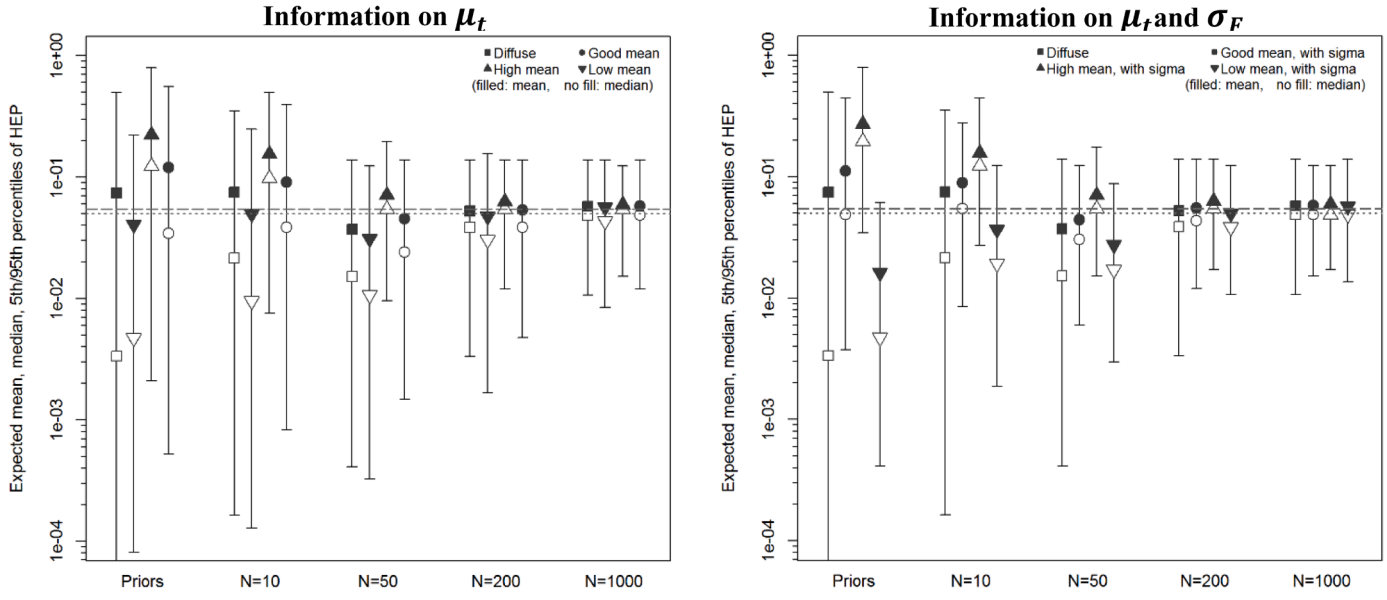


Fig. 7. Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$, and to the sample size, Case 1. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: "Good mean", 5e-3/5e-1; "High mean", 5e-2/1; "Low mean", 5e-4/5e-2. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ are informative. "With sigma" corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor).

combinations for $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ on the posterior HEP estimates as the number of simulator runs increases (in x-axis). The prior information may be available from HRA methods or generic failure databases. The considered prior distributions for the mean, $\pi_0(\mu_t)$, are (left plot):

- "Diffuse": flat distributions for the parameters of the lognormal, mean and standard deviation, same as for Section 4.2.1;
- "Good mean": prior distribution informed around the correct median HEP value for Case 1 (lognormal, with median = 5e-2, 5th percentile = 5e-3, 95th percentile = 5e-1);

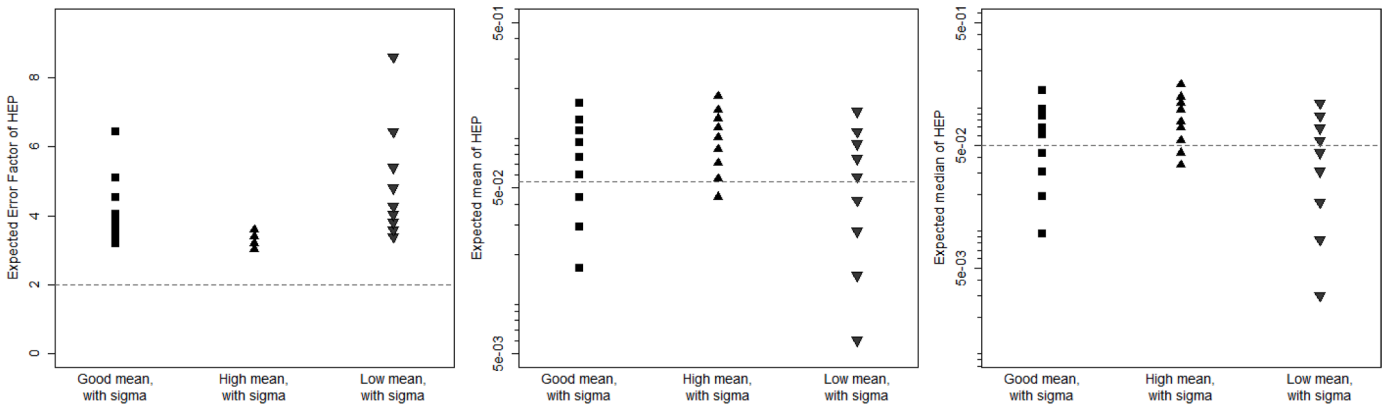


Fig. 8. Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ at $N_F = 50$ (target HEP variability distribution as in Case 1: median = 5e-2, mean = 5.46e-2, and error factor = 2). For each option of informative priors in x-axis (Fig. 7, right plot), 100 datasets (dots) are Monte Carlo-sampled from the target distribution. From left to right, in y-axis: expected error factor, mean (log-scale), and median (log-scale) of the $P_F(HEP)$'s provided by the model for each choice of prior (dotted lines: statistics of the target distribution).

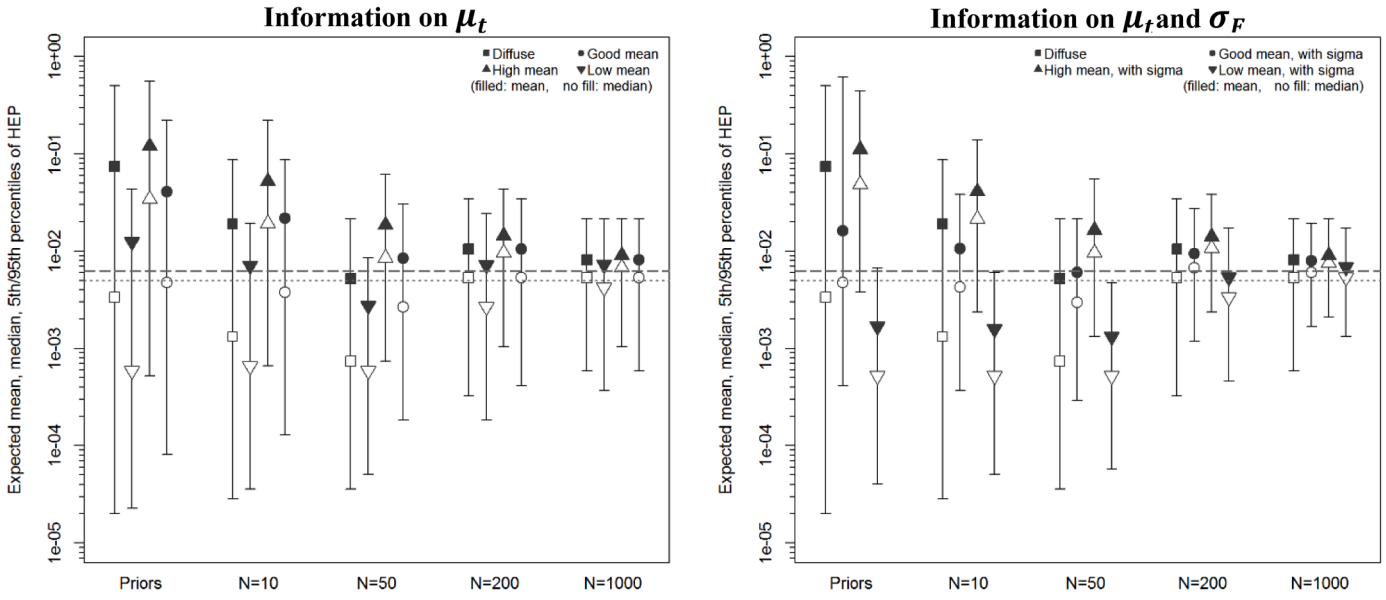


Fig. 9. Sensitivity of the lognormal variability model to the choice of prior distributions for the hyper-parameters, i.e. $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$, and to the sample size, Case 2. Left: only $\pi_0(\mu_t)$ is informative. Upper/lower bounds of the lognormal distributions: "Good mean", 5e-4/5e-2; "High mean", 5e-3/5e-2; "Low mean", 5e-5/5e-3. Right: both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ are informative. "With sigma" corresponds to a normal distribution with bounds 1.5/5 (expressed in terms of error factor).

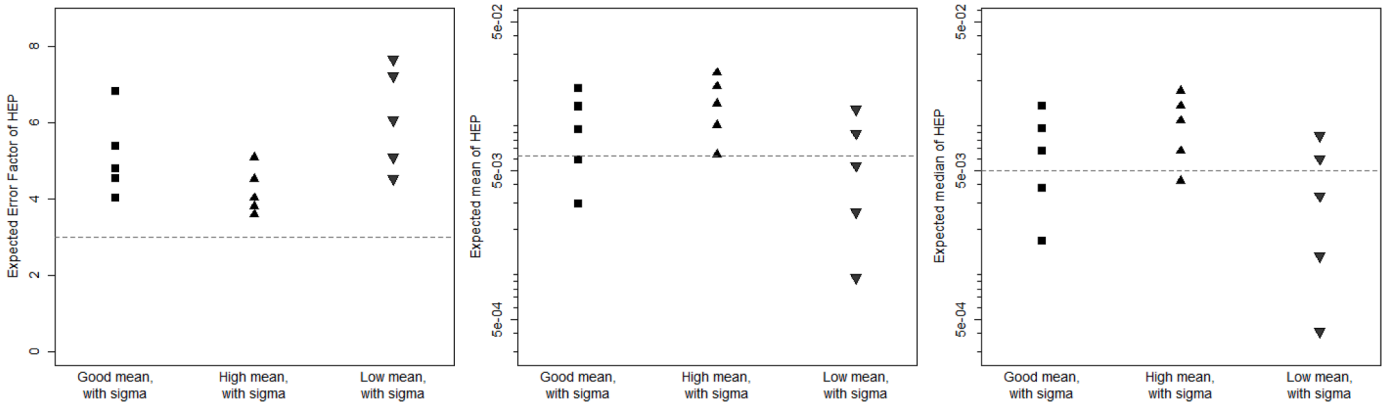


Fig. 10. Behavior of the lognormal variability model with informative priors on both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ at $N_F = 200$ (target HEP variability distribution as in Case 2: median = 5e-3, mean = 6.25e-3, and error factor = 3). Same considerations as in Fig. 8. Dotted lines: statistics of the target distribution.

- "Low mean" and "High mean": prior distributions with median shifted by one order of magnitude below and above the correct median HEP value for Case 1, respectively (for "Low mean": lognormal, with median = 5e-3, 5th percentile = 5e-4, 95th percentile = 5e-2; for "High mean": lognormal, with median = 5e-1, 5th percentile = 5e-2, 95th percentile = 1).

The "Good mean" prior assumes that the information at disposal is correct in the order of magnitude of the HEP range, with two orders of magnitude between the 5th and the 95th percentiles. The "Low mean" and "High mean" priors assume the presence of biases of one order of magnitude.

Additional information on the standard deviation, $\pi_0(\sigma_F)$, is modelled by a normal distribution ("with sigma") with 5th and 95th percentiles corresponding to error factors of 1.5 and 5, respectively (right plot). Limiting values for error factor close to 5 are commonly accepted in establishing confidence intervals for HRA applications [6].

With informative $\pi_0(\mu_t)$ (Fig. 7, left plot), when the information on μ_t is not biased ("Good mean"), it is possible to achieve reasonable approximations of the target mean (5.5e-2) and median (5e-2) already at $N_F = 200$. Indeed, at $N_F = 200$ the "Good mean" error factor is 16% lower than the one obtained with "Diffuse" prior (see Table A.1). In case of biased information on μ_t, sensible overestimation ("High mean") or underestimation ("Low mean") of the expected mean and median can be observed for all datasets, of course tending to decrease with the amount of data available.

Data requirements can be significantly reduced if both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ are informative (Fig. 7, right). If the information on μ_t is not biased ("Good mean, with sigma"), it is possible to obtain good approximations of the expected mean and median, as well as acceptable error factors, already in the range $N_F = 10 \div 50$. For instance, at $N_F = 50$, the model with "Good mean, with sigma" prior returns an expected error factor approximately 4 times lower than the value provided with the "Diffuse" prior, and significantly closer to the target value for Case 1 (error factor = 2). Still at $N_F = 50$, the biased hyper-priors reflect in biased HEP estimates (Table A.1 and Fig. 7, right), however the correct values lie within the 90% confidence bounds (5th and 95th percentiles). As the data set increases, the effect of the prior information is progressively reduced, as shown by the statistics for $N_F = 200$ and 1000, very close to the target values.

To further investigate the possible reduction in data requirements, Fig. 8 further examines the sample size of $N_F = 50$, a size reasonably achievable by current simulator data collection programs aggregating multiple plants. Fig. 8 shows the results for 100 Monte Carlo-sampled datasets relevant to Case 1 at $N_F = 50$. The results confirm that such size is well enough for "Good mean, with sigma": average expected mean of 5.8e-2 (50% confidence: 4.4e-2, 7.7e-2), average expected median of 4.3e-2 (50% confidence: 3.1e-2, 6.1e-2), average expected error factor = 4.4 (50% confidence: 3.8, 4.5). The Monte Carlo samples show that the biased estimates are not usable, because the correct values lie outside the 50% confidence interval: for "High mean, with sigma", average expected mean 8.3e-2 (50% confidence: 7.1e-2, 1.0e-1), average expected median 6.6e-2 (50% confidence: 5.5e-2, 7.7e-1); for "Low mean, with sigma", average expected mean 4.2e-2 (50% confidence: 2.7e-2, 5.8e-2), average median = 2.9e-2 (50% confidence: 1.7e-2, 4.3e-2). It is however important to mention that the potential bias may be relatively easy to identify a posteriori. For example, from the Monte Carlo samples at $N_F = 50$, the expected change in marginal prior medians (see Table A.1) after the evidence is:

- for "Good mean, with sigma" between 24% and 36% of the marginal prior median (= 4.9e-2);
- for "High mean, with sigma" between 72% and 285% of the marginal prior median (= 2e-1);
- for "Low mean, with sigma" between 254% and 796% of the marginal prior median (= 4.8e-3).

Indeed, large deviations of the posterior median from the marginal prior median could be used as indicators of an initial bias.

Fig. 9 and Table A.2 present the results relevant to Case 2 and Fig. 10 further explores the influence of informative priors at $N_F = 200$:

- "Good mean": lognormal, with median = 5e-3, 5th percentile = 5e-4, 95th percentile = 5e-2;
- "Low mean": lognormal, with median = 5e-4, 5th percentile = 5e-5, 95th percentile = 5e-3;
- "High mean": lognormal, with median = 5e-2, 5th percentile = 5e-3, 95th percentile = 5e-1.

Compared to Case 1, Case 2 is characterized by a "weaker" evidence of failure (note that HEP ~ 0.001 in Case 2): this aspect influences the efficiency of informative priors in reducing the data requirements of the model. With informative $\pi_0(\mu_t)$, from the cross-comparison with Case 1 results (left plots in Figs. 7 and 9; Tables A.1 and A.2), the model tends to return significantly higher values of the expected error factor in Case 2: this suggests that informing only μ_t is not sufficient to achieve good approximation of the target mean (6.3e-3) and median (5e-3) with acceptably low N_F (e.g. already at $N_F = 200$ as for Case 1).

When informing both $\pi_0(\mu_t)$ and $\pi_0(\sigma_F)$ without bias ("Good mean, with sigma" in Fig. 9, right), good approximations of the expected mean and median, as well as acceptable error factors, can be achieved in the range $N_F = 50 \div 200$ (note the increased data requirements compared to range $N_F = 10 \div 50$ for Case 1). For instance, at $N_F = 200$, the model with "Good mean, with sigma" prior returns an expected error factor approximately two times lower than the value provided with the "Diffuse" prior and closer to the target value for Case 2 (error factor = 3). Still at $N_F = 200$, however, the biased hyper-priors ("Low mean, with sigma" and "High mean, with sigma") reflect in biased HEP estimates (Table A.2 and Fig. 9, right), however the correct values lie within the 90% confidence bounds. Fig. 10 shows the results for 100 Monte Carlo-sampled datasets relevant to Case 2 at $N_F = 200$. The analysis confirms that "Good mean, with sigma" performs efficiently at this sample size: average expected mean = 7.0e-3 (50% confidence: 5.8e-3, 9.4e-3), average expected median = 4.8e-3 (50% confidence: 3.8e-3, 6.7e-3), and average expected error factor = 5.4 (50% confidence: 4.8, 5.4). On the other hand, for the configurations with biased priors, the correct values of the statistics (target mean = 6.3e-3 and target median = 5e-3) lie outside the 50% confidence interval: for "High mean, with sigma", average expected mean = 1.1e-2 (50% confidence: 1.0e-2, 1.4e-2), average expected median = 8.0e-3 (50% confidence: 6.7e-3, 1.1e-2); for "Low mean, with sigma", average expected mean 3.7e-3 (50% confidence: 2.6e-3, 5.3e-3), average median = 2.2e-3 (50% confidence: 1.3e-3, 3.4e-3). As for Case 1, the potential bias in Case 2 can be easily identified by the observed large deviations of the posterior median from the marginal prior median (see Table A.2) across the different configurations, e.g. at $N_F = 200$:

- for "Good mean, with sigma" between 20% and 39% of the marginal prior median (= 4.8e-3);
- for "High mean, with sigma" between 78% and 86% of the marginal prior median (= 4.9e-2);
- for "Low mean, with sigma" between 150% and 554% of the marginal prior median (=5.2e-4).

In conclusion, the analysis highlighted the following two aspects. First, for a given acceptable level of approximation of the target error factor, unbiased informative priors on both the mean and the standard deviation of HEP distribution are effective in reducing the overall data requirements of the lognormal variability model. Secondly, especially for constellations $\boldsymbol{F}$ characterized by lower orders of magnitude of HEP or limited performance data N_F (or both), biased informative priors have a strong influence on the HEP uncertainty distribution estimated by the model. Following on this, reducing as much as possible the bias in

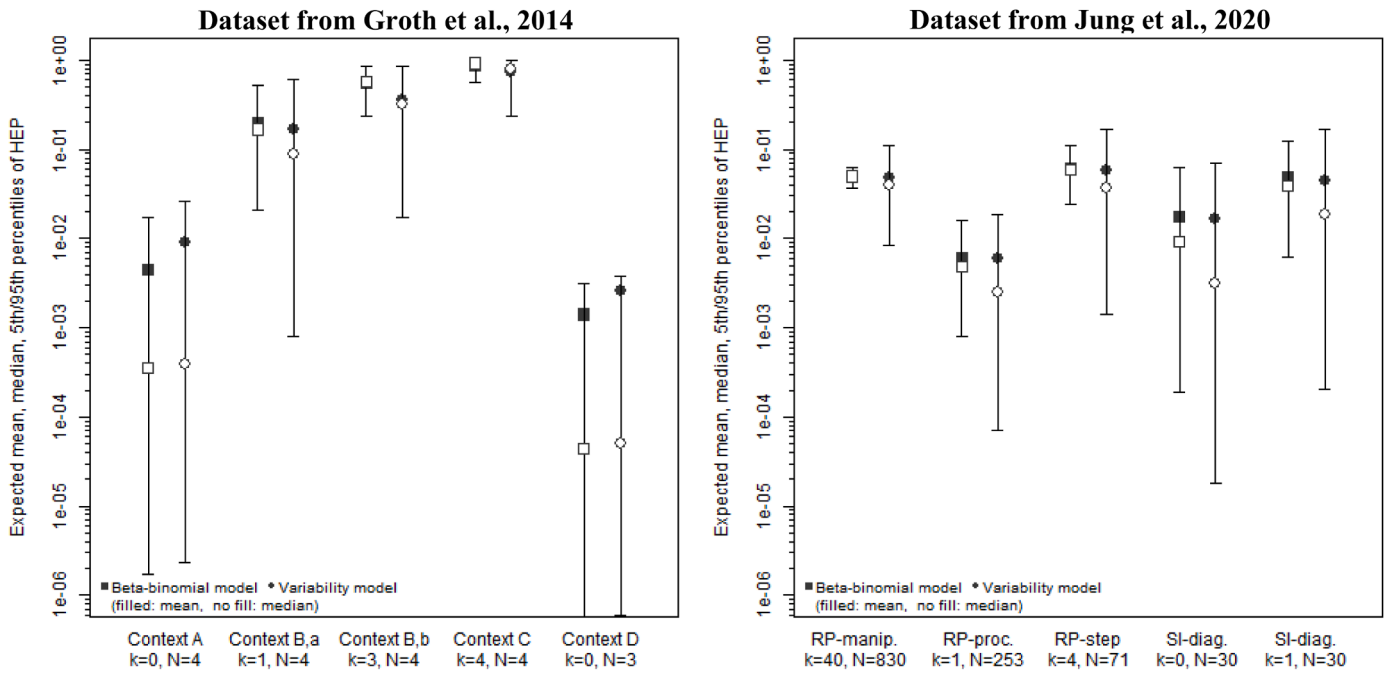


Fig. 11. Results from the application of the lognormal variability model to real simulator data available in literature (datasets in x-axis: left, [20]; right, [26]). On y-axis (in log-scale): expected mean (filled symbols), median (blank symbols), and 5th – 95th percentiles (whiskers) of the $P_F(HEP)$'s estimated by both the lognormal variability model (circles) and the lumped-data beta-binomial model (circles) given the same marginal prior distribution on HEP.

informative priors becomes of key importance. Besides the approach adopted for the purposes of this numerical application, different techniques (e.g. posterior predictive checks) are available in Bayesian literature to assist the analyst in selecting adequate prior distributions and reduce the initial bias [50].

4.3. Application to real simulator data from literature

The proposed variability model is applied to failure data of operating crews in nuclear power plants available in the literature (Halden project data from [20] and HuREX data from [26]). Both [20] and [26] use the simulator data to inform HEPs of constellations of task type and PSF levels. [20] address constellations of SPAR-H PSFs (e.g. "complexity", "stressors"), while [26] addresses the HuREX framework for different combinations of cognitive activities (e.g. "situation interpreting", "execution") and generic task types (e.g. "verifying state of indicator"; "directing manipulation"). In particular, [20] addresses five contexts (for the sake of brevity, only SPAR-H's PSFs with ratings different than "nominal" are reported; see [8] for further information on PSF definitions):

- Context A: Time = extra; Complexity = moderate; Procedures = available but poor.
- Contexts B_a, B_b: Time = barely adequate; Stressors = high; Complexity = moderate; Procedures = available but poor.
- Context C: Time = inadequate; Stressors = high; Complexity = high; Procedures = available but poor; Work processes = poor.
- Context D: Time = extra.

For [26] the following operator activities are considered:

- RP-manipulation: cognitive activity = response planning; task type = directing manipulation.
- RP-procedure: cognitive activity = response planning; task type = transferring procedure.
- RP-step: cognitive activity = response planning; task type = transferring step procedure.
- SI-diagnosis: cognitive activity = situation interpreting; task type = diagnosing.

In [26], the authors adopted a conservative assumption consisting of adding a fictitious recorded failure for all those constellations $\boldsymbol{F}$ where actually no failures have been observed. For instance, this was the case of RP-step dataset. In this application, the latter has been treated in two different configurations: the conservative dataset as used by the authors (with one postulated failure: $k_F = 1$, $N_F = 30$), and the real dataset (with zero failures observed: $k_F = 0$, $N_F = 30$).

Both Groth et al. (2014) [20] and Jung et al. (2020) [26] adopt the lumped approach, with the conjugated beta-binomial model. Concerning the prior, Groth et al. (2014) [20] uses the CNI prior (from [39]), built on the basic HEP provided by SPAR H in correspondence of the context. Jung et al. (2020) [26] adopts the Jeffreys non-informative distribution, a beta distribution with both shape parameters (i.e. α_0 and β_0 in eq. 11) equal to 0.5.

An important difference between the data sets of Groth et al. (2014) [20] and Jung et al. (2020) [26] concerns their size. Groth et al. (2014) [20] addresses rather small data sets, four data points on average, including very challenging tasks. Jung et al. (2020) [26] addresses significantly larger datasets, because of the different granularity of the data collection taxonomy and because of the larger number of crews from which data is collected. This difference allows comparing the performance of the variability and the beta-binomial models (with lumped data) under very different data availability conditions.

The expected statistics (mean, median, and 5th /95th percentiles) of the HEP posterior distributions estimated by both variability and lumped-data models are shown in Fig. 11 (y-axis, in log-scale), for each of the datasets used in the application (x-axis, left: [20]; right: [26]). A summary of the numerical results is given in Table A.3 in Appendix A. Note that the results for lumped-data models in Table A.3 and Fig. 11 are slightly different from the numerical values in [20] and [26], since the prior distributions adopted by these works (the CNI for [20]; the Jeffreys for [26]) were adapted in this application to ensure a fair comparison with the variability model. In particular, for the results to be comparable, the literature models and the variability model should start from the

same expected HEP distribution ($P_F(p_{c|t})$ from eq. 9 for the variability model). To do this, the mean of the lognormal variability model (i.e. μ_t) were assigned the literature priors, i.e. CNI prior for $\pi_0(\mu_t)$ for the comparison with [20]; Jeffreys prior for $\pi_0(\mu_t)$ for the comparison with [26]. Then, the expected HEP distribution from the variability model, i.e. the lognormal parametric distribution weighted by the joint hyperprior $\pi_0(\mu_t, \sigma_F)$, was derived (for $\pi_0(\sigma_F)$, the diffuse prior mentioned in Section 3.4 was used). Finally, the lumped-data priors were re-calculated such that the corresponding expected HEP distribution would fit the one from the variability model.

From the comparison of result, a general tendency can be observed: overall, the lumped-data beta-binomial models tend to return narrower posterior distributions if compared to the variability model. This tendency replicates across all the tested datasets, with a magnitude that depends on the amount of evidence available (i.e. the sample size and the observed failures). In particular, for [20] (Fig. 11, left), the differences in the two models are small for "Contexts A" and "Context D": the corresponding datasets are characterized by few observations and zero failures. As the number of observed failures increases (e.g. "Contexts B_a" and "Context B_b"), the differences between the posteriors become larger, see the expected error factor in "Contexts A/D, B_a and B_b" in Table A.3 (e.g. for "Context B_b", the variability model returns an expected error factor 3.7 times higher than the lumped-data model).

A similar trend can be observed for the data-rich application [26]. The differences in the expected error factors become more evident with progressively increasing the number of observed failures in the dataset (e.g. see the different spreads in the HEP uncertainty distributions from "SI-diagnosis, k=0" to "SI-diagnosis, k=1", in Fig. 11, right); the differences persist at very high numbers of observed failures (e.g. for "RP-manipulation" dataset, the error factor estimated by the lumped-data model is approximately 2.8 times lower than the variability model).

5. Discussion

The application to simulator data in Section 4 has demonstrated the large impact on the estimated HEP distribution of considering the underlying variability in the HRA data. As presented in Section 3, the two models reflect two different interpretations of the target HEP. The variability model considers the HEP as a quantity that is specific for a crew and for a realization of the constellation; correspondingly, the HEP variability reflects the variability of the crews and of the realizations. The beta-binomial model considers the HEP as a unique quantity for a given constellation, aggregating all variability aspects in its value.

It is important to note that there is no right or wrong interpretation of the HEP quantity: it depends on the application at hand. For example, an important HRA issue is to investigate PSF effects across different constellations. The effect on the HEP of changes in one or more elements of the vector $\boldsymbol{F}$ in eq. 1 may be investigated by focusing on the aggregated effect, i.e. on the population average across crews and within-constellation, therefore adopting the typical beta-binomial model. On the other hand, as presented in Section 2, when the estimated HEP is used to inform a given constellation of an HRA model, adopting a variability model becomes important to capture the variability elements discussed in Section 2 and ideally allow for plant-specific HEP values (as demonstrated in Section 4).

The model presented here supports a first investigation of the need for modelling variability. The interpretation of the HEP as a crew-specific quantity strongly limits the possibility to aggregate the data to inform HEP values. As shown by Fig. 4, the data informing the HEP variability distribution are only 0's and 1's because of the constraint that one crew only performs the exact same task only once. An alternative would be to consider the HEP values as dependent on particular crew features or styles (e.g. of communication or decision-making), as opposed to being just crew-specific. This approach would not consider each crew being characterized by a different HEP value: each crew feature or style would be connected with an HEP value. Numerically, this would allow aggregating more evidence on the single HEP realization (the number of task repetitions in Fig. 4 would be per crew feature or style, and not per single crew). On the other hand, this may allow analysis of crew features and styles on the HEP, opening to additional applications to inform crew training. Current work by the authors is addressing definitions of appropriate features and styles as well as the associated adaptations to the model.

As presented in Section 3, the inference model is intended for general application to any HRA model for HEP quantification. The currently available HRA models strongly differ in the task and factors considered and in the granularity of their definition. It can be expected that these aspects are strongly connected with the variability that the model shall be able to represent. For instance, the simulator data used in Section 4.3 (Halden in [20]; HuREX in [26]) correspond to constellations at very different granularity. [20] uses the SPAR-H factor taxonomy on an operator task definition close to what would be used for PSA applications (e.g. "isolate the ruptured steam generator and control pressure"). On the other hand, HuREX in [26] operates at a more microscopic granularity level (e.g. "determine the condition of Adverse Containment", "check if the three Reactor Coolant Pumps should be stopped"). As a working hypothesis, it may be reasonable to assume that the coarser the granularity of the model (more macroscopic tasks), the larger the variability corresponding to the within-category variability. Also, the more the task involves decision-making and communication at the crew level, the more crew variability will be relevant, compared for example to execution-related tasks performed by single persons. Finally, it can be expected that variability would also be larger for HRA models with coarser PSF categories, e.g. binary as opposed to multivalued. With the current interest by the community on empirically estimated HEPs, it may be well important that future studies will address the extent to which variability shall be addressed as well as with the goal of develop guidelines to do it.

HRA research is addressing advanced modelling techniques, in particular Bayesian Belief Networks, to represent the complex relationships among influencing factors as well as to formally incorporate a diversity of data sources. Indeed, within-category variability can be incorporated in these models via appropriate conditional probability distributions. BBNs can incorporate crew-to-crew variability as well, either implicitly, into the BBN internal distributions, as well as explicitly, as dedicated nodes [10,23]. The work presented in this paper can be used to enhance the empirical basis of the BBN distributions, e.g. as anchoring distributions to populate the model relationships via filling algorithms such as those in [49].

6. Conclusions

Due to lack of data, judgments are currently the main source of information to assess the uncertainty and variability in the error probability estimates produced by HRA models. With the on-going large data collection activities, it becomes important that uncertainty and variability be empirically based, along with the associated point estimates.

This paper presents a Bayesian hierarchical model that addresses the HEP variability due to operating crew differences as well as variability within the categories of task type and performance factors. Such models are typically used to consider source-to-source variability of failure probability estimates for hardware components: this paper presents their formulation and use for human failure data from simulators.

Table A.1
Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Fig. 7 (Case 1, target HEP variability distribution with median = 5e-2, mean = 5.46e-2, and error factor = 2).

	Prior distribution	Mean	Median	5th perc	95th perc	EF
No evidence (marginal priors)	Diffuse	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
	Low mean	4.06e-02	4.75e-03	8.11e-05	2.21e-01	52.14
	High mean	2.23e-01	1.23e-01	2.10e-03	7.92e-01	19.40
	Good mean	1.20e-01	3.43e-02	5.21e-04	5.59e-01	32.75
	Low mean, with sigma	1.61e-02	4.75e-03	4.13e-04	6.14e-02	12.19
	High mean, with sigma	2.71e-01	1.96e-01	3.43e-02	7.92e-01	4.81
	Good mean, with sigma	1.11e-01	4.86e-02	3.76e-03	4.43e-01	10.85
N_F=10, 1 failure	Diffuse	7.47e-02	2.15e-02	1.63e-04	3.51e-01	46.42
	Low mean	4.97e-02	9.55e-03	1.29e-04	2.48e-01	43.79
	High mean	1.55e-01	9.77e-02	7.56e-03	4.98e-01	8.11
	Good mean	9.06e-02	3.85e-02	8.30e-04	3.94e-01	21.80
	Low mean, with sigma	3.67e-02	1.92e-02	1.87e-03	1.23e-01	8.11
	High mean, with sigma	1.56e-01	1.23e-01	2.72e-02	4.43e-01	4.04
	Good mean, with sigma	8.91e-02	5.46e-02	8.50e-03	2.78e-01	5.72
N_F =50, 2 failures	Diffuse	3.73e-02	1.52e-02	4.13e-04	1.38e-01	18.31
	Low mean	3.12e-02	1.07e-02	3.27e-04	1.23e-01	19.40
	High mean	7.15e-02	5.46e-02	9.55e-03	1.96e-01	4.53
	Good mean	4.53e-02	2.42e-02	1.48e-03	1.38e-01	9.66
	Low mean, with sigma	2.75e-02	1.71e-02	2.98e-03	8.70e-02	5.40
	High mean, with sigma	7.08e-02	5.46e-02	1.52e-02	1.75e-01	3.39
	Good mean, with sigma	4.42e-02	3.05e-02	5.99e-03	1.23e-01	4.53
N_F =200, 11 failures	Diffuse	5.24e-02	3.85e-02	3.35e-03	1.38e-01	6.43
	Low mean	4.74e-02	3.05e-02	1.67e-03	1.56e-01	9.66
	High mean	6.27e-02	5.46e-02	1.20e-02	1.38e-01	3.39
	Good mean	5.38e-02	3.85e-02	4.75e-03	1.38e-01	5.40
	Low mean, with sigma	4.97e-02	3.85e-02	1.07e-02	1.23e-01	3.39
	High mean, with sigma	6.30e-02	5.46e-02	1.71e-02	1.38e-01	2.85
	Good mean, with sigma	5.50e-02	4.33e-02	1.20e-02	1.38e-01	3.39
N_F =1000, 58 failures	Diffuse	5.75e-02	4.86e-02	1.07e-02	1.38e-01	3.59
	Low mean	5.62e-02	4.33e-02	8.50e-03	1.38e-01	4.04
	High mean	5.95e-02	5.46e-02	1.52e-02	1.23e-01	2.85
	Good mean	5.76e-02	4.86e-02	1.20e-02	1.38e-01	3.39
	Low mean, with sigma	5.67e-02	4.86e-02	1.35e-02	1.38e-01	3.20
	High mean, with sigma	5.96e-02	4.86e-02	1.71e-02	1.23e-01	2.69
	Good mean, with sigma	5.79e-02	4.86e-02	1.52e-02	1.23e-01	2.85

Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Fig. 7 (Case 1, target HEP variability distribution with median = 5e-2, mean = 5.46e-2, and error factor = 2).

The presented case studies demonstrate the significant overconfidence in the HEP estimates if variability is not considered, e.g. if all data is lumped to feed a beta-binomial Bayesian model (as typically done in most HRA applications). Also, this may results in significant biases for plant-specific human error probabilities.

Empirically informing variability requires a large amount of data: therefore, numerical applications have investigated the practical applicability of the proposed model. For moderately high HEP values (in the range of 1e-2), estimates of practical use can be obtained with few hundred, say below 500, data points (i.e. simulator runs). This is already achievable by current simulator programs depending on the constellation of tasks and performance factors. Prior information on the model parameters, e.g. from available HRA methods, can reduce the data requirements. For HEP values in the range of 1e-2, about 50 data points are demonstrated to become enough. For lower HEP values, in the range of 1e-3, estimates of practical use become achievable with few hundred data points. Of course, biases in the prior distributions may result in biases in the posterior estimates. However, this paper has shown that a simple check of the change between the prior and posterior estimates may reveal the presence of the initial bias. Data requirements for further low HEP ranges, i.e. below 1e-3, may be impractical for many operator tasks with the proposed model.

The proposed model treats variability as a continuum. Especially when considering crew-to-crew variability, it may be important to identify relevant crew features that play a role in determining the failure probability. Besides allowing aggregating data from different crews on the basis of their common traits, this may support training of operators on the crew skills that allow lower failure probability values. Work by the authors is ongoing along this direction.

This work is part of a larger effort to derive empirically-based reference HEP values to strengthen the technical basis of HRA methods. The long-term aim is to develop a framework to process diverse data sources, e.g. simulator data, data from existing HRA methods, operational experience data, and evidence from human factor studies. The main thrust is that a mathematical, traceable aggregation of these sources will allow to feed with new data as it becomes available, progressively replacing older evidence that may become outdated because of new advances in plant operation and design.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was funded by the Swiss Federal Nuclear Safety Inspectorate (ENSI), under contract Nr. 101163. The views expressed in this work are solely those of the authors.

Appendix A

Table A.2
Numeric results from sensitivity analysis on choice of priors for the lognormal variability model as shown in Fig. 9 (Case 2, target HEP variability distribution with median = 5e-3, mean = 6.25e-3, and error factor = 3).

	Prior distribution	Mean	Median	5th perc	95th perc	EF
No evidence	Diffuse	7.44e-02	3.35e-03	2.01e-05	4.98e-01	157.39
(marginal priors)	Low mean	1.25e-02	5.86e-04	2.26e-05	4.33e-02	43.79
	High mean	1.20e-01	3.43e-02	5.21e-04	5.59e-01	32.75
	Good mean	4.06e-02	4.75e-03	8.11e-05	2.21e-01	52.14
	Low mean, with sigma	1.68e-03	5.21e-04	4.04e-05	6.73e-03	12.92
	High mean, with sigma	1.11e-01	4.86e-02	3.76e-03	4.43e-01	10.85
	Good mean, with sigma	1.61e-02	4.75e-03	4.13e-04	6.14e-02	12.19
N_F =10, 0 failures	Diffuse	1.91e-02	1.32e-03	2.85e-05	8.70e-02	55.26
	Low mean	7.11e-03	6.58e-04	3.59e-05	1.92e-02	23.10
	High mean	5.23e-02	1.92e-02	6.58e-04	2.21e-01	18.31
	Good mean	2.18e-02	3.76e-03	1.29e-04	8.70e-02	25.95
	Low mean, with sigma	1.57e-03	5.21e-04	5.09e-05	5.99e-03	10.85
	High mean, with sigma	4.11e-02	2.15e-02	2.36e-03	1.38e-01	7.65
	Good mean, with sigma	1.06e-02	4.23e-03	3.68e-04	3.85e-02	10.24
N_F =50, 0 failures	Diffuse	5.22e-03	7.39e-04	3.59e-05	2.15e-02	24.48
	Low mean	2.75e-03	5.86e-04	5.09e-05	8.50e-03	12.92
	High mean	1.85e-02	8.50e-03	7.39e-04	6.14e-02	9.11
	Good mean	8.46e-03	2.66e-03	1.83e-04	3.05e-02	12.92
	Low mean, with sigma	1.32e-03	5.21e-04	5.72e-05	4.75e-03	9.11
	High mean, with sigma	1.65e-02	9.54e-03	1.32e-03	5.46e-02	6.43
	Good mean, with sigma	6.05e-03	2.98e-03	2.92e-04	2.15e-02	8.60
N_F =200, 2 failures	Diffuse	1.05e-02	5.34e-03	3.27e-04	3.43e-02	10.24
	Low mean	7.17e-03	2.66e-03	1.83e-04	2.42e-02	11.50
	High mean	1.45e-02	9.55e-03	1.05e-03	4.33e-02	6.43
	Good mean	1.05e-02	5.34e-03	4.13e-04	3.43e-02	9.11
	Low mean, with sigma	5.36e-03	3.35e-03	4.64e-04	1.71e-02	6.06
	High mean, with sigma	1.40e-02	1.07e-02	2.36e-03	3.85e-02	4.04
	Good mean, with sigma	9.44e-03	6.73e-03	1.18e-03	2.72e-02	4.81
N_F =1000, 8 failures	Diffuse	8.10e-03	5.34e-03	5.86e-04	2.15e-02	6.06
	Low mean	7.23e-03	4.23e-03	3.68e-04	2.15e-02	7.65
	High mean	9.07e-03	6.73e-03	1.05e-03	2.15e-02	4.53
	Good mean	8.10e-03	5.34e-03	5.86e-04	2.15e-02	6.06
	Low mean, with sigma	6.83e-03	5.34e-03	1.32e-03	1.71e-02	3.59
	High mean, with sigma	8.99e-03	7.56e-03	2.10e-03	2.15e-02	3.20
	Good mean, with sigma	7.90e-03	5.99e-03	1.67e-03	1.92e-02	3.39

Table A.3
Numeric results from the application of the lognormal variability model on real simulator data taken from [20] (upper table) and [26] (lower table) shown in Fig. 11.

	Model	Mean	Median	5th perc.	95th perc.	EF
Context A	Beta-binomial	4.46e-03	3.57e-04	1.70e-06	1.74e-02	101.16
N_F =4, 0 failures	Variability model	9.13e-03	3.87e-04	2.35e-06	2.61e-02	105.34
Context B	Beta-binomial	2.00e-01	1.68e-01	2.05e-02	5.23e-01	5.05
N_F =4, 1 failure	Variability model	1.67e-01	8.80e-02	8.03e-04	6.15e-01	7.68
Context B bis	Beta-binomial	5.64e-01	5.67e-01	2.33e-01	8.50e-01	1.91
N_F =4, 3 failures	Variability model	3.54e-01	3.22e-01	1.74e-02	8.50e-01	6.99
Context C	Beta-binomial	8.60e-01	9.22e-01	5.67e-01	1.00e-00	1.33
N_F =4, 4 failures	Variability model	7.36e-01	7.84e-01	2.33e-01	1.00e-00	2.07
Context D	Beta-binomial	1.41e-03	4.35e-05	4.66e-07	3.18e-03	82.62
N_F =3, 0 failure	Variability model	2.62e-03	5.11e-05	5.94e-07	3.74e-03	79.34
RP-manipulation	Beta-binomial	4.87e-02	4.99e-02	3.61e-02	6.37e-02	1.33
N_F =830, 40 failures	Variability model	4.83e-02	3.92e-02	8.41e-03	1.12e-01	3.65
RP-procedure	Beta-binomial	6.09e-03	4.77e-03	8.03e-04	1.61e-02	4.47
N_F =253, 1 failure	Variability model	5.94e-03	2.49e-03	7.07e-05	1.89e-02	16.35
RP-step	Beta-binomial	6.21e-02	5.87e-02	2.41e-02	1.12e-01	2.16
N_F =71, 4 failures	Variability model	5.80e-02	3.61e-02	1.41e-03	1.68e-01	10.91
SI-diagnosis	Beta-binomial	1.73e-02	9.12e-03	1.87e-04	6.37e-02	18.46
N_F =30, 0 failures	Variability model	1.68e-02	3.18e-03	1.78e-05	6.91e-02	62.23
SI-diagnosis	Beta-binomial	4.80e-02	3.92e-02	6.08e-03	1.22e-01	4.47
N_F =30, 1 failures	Variability model	4.46e-02	1.89e-02	2.03e-04	1.68e-01	28.83

References

[1] Kirwan B. A guide to practical Human Reliability Assessment. Boca Raton, FL, USA: CRC press; 1994.
[2] Podofillini L. Human Reliability Analysis. In: Hansson Moller, editor. Handbook of Safety Principles. Holmberg, Carl Rollenhagen: Wiley; 2017.
[3] Williams JC. HEART – A Proposed Method for Assessing and Reducing Human Error. 9th Advance in Reliability Technology Symposium. University of Bradford; 1986. 1986, B3/R/1-B3/R/13.
[4] Williams JC. A data-based method for assessing and reducing human error to improve operational performance. In: Proceedings of the IEEE Fourth Conference on Human Factors and Power Plants; 1988. p. 436–50. 5–9 June.
[5] Williams JC. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. Saf. Reliab. 2015;35 (3):5–25.
[6] Swain AD, Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power plant applications. Washington DC, USA: U.S. Nuclear Regulatory Commission; 1983. NUREG/CR-1278.
[7] Gertman DI, Blackman HS, Marble JL, Byers JC, Smith CL. The SPAR-H Human Reliability Analysis Method 2005. NUREG/CR-6883.

[8] Whaley AM, Kelly DL, Boring RL, Galyean WJ. SPAR-H step-by-step guidance. Idaho Falls, Idaho: Idaho National Labs; 2011. p. 83415. INL/EXT-10-18533.
[9] Hollnagel E. Cognitive Reliability and Error Analysis Method (CREAM). Oxford: Elsevier Science Ltd; 1998.
[10] Groth KM, Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: a methodology and example mode. Proc Inst Mech Eng, Pt O: J Risk Reliab 2012 2012;226(4):361–79.
[11] Mkrtchyan L, Podofillini L, Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. Reliability Engineering and System Safety 2015;139:1–16. 2015.
[12] Spurgin AJ. Human Reliability Assessment – theory and practice. Boca Raton, FL, USA: CRC press; 2010.
[13] Hallbert B, Kolaczkowski A. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study. Washington, D.C.: U.S. Nuclear Regulatory Commission; 2007. p. 1–4. NUREG/CR-6949INL/EXT-06-11670.
[14] Forester J, Dang VN, Bye A, Lois E, Massaiu S, Bromberg H, Ø. Braarud P, Boring R, Männistö I, Liao H, Julius J, Parry G, Nelson P. The International HRA Empirical Study Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127. Washington DC, USA: US Nuclear Regulatory Commission; 2014.
[15] Forester J, Liao H, Dang VN, Bye A, Lois E, Presley M, Marble J, Nowell R, Broberg H, Hildebrandt M, Hallbert B, Morgan T. The US HRA Empirical Study - Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156. Washington DC, USA: US Nuclear Regulatory Commission; 2016.
[16] Park J, Jung W, Kim S, Choi SY, Kim Y, Lee SJ, Yang JE, Dang VN. A guideline to collect HRA data in the simulator of nuclear power plants, KAERI/TR-5206. Republic of Korea: Korea Atomic Energy Research Institute; 2013.
[17] Chang JY, Bley D, Criscione L, Kirwan B, Mosleh A, Madary T, Nowell R, Richards R, Roth EM, Sieben S, Zoulis A. The SACADA database for human reliability and human performance. Reliability Engineering & System Safety 2014;125:117–33.
[18] Liao H, Forester J, Dang VN, Bye A, Chang JY, Lois E. Assessment of HRA method predictions against operating crew performance: Part III: Conclusions and achievements. Reliability Engineering & System Safety 2019;191:106511.
[19] Kim Y, Park J, Jung W. A classification scheme of erroneous behaviors for human error probability estimations based on simulator data. Reliability Engineering & System Safety 2017;163:1–13. ISSN 0951-8320.
[20] Groth KM, Smith CL, Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. Reliability Engineering & System Safety 2014;128(Supplement C):32–40.
[21] Kim Y, Park J, Jung W, Choi SY, Kim S. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. Reliability Engineering & System Safety 2018;173:12–22. ISSN 0951-8320.
[22] Chang JY, Franklin C. SACADA Data for HEP Estimates. Probabilistic Safety Assessment and Management PSAM 2018;14. September 2018.
[23] Groth KM. A framework for using SACADA to enhance the qualitative and quantitative basis of HRA, in 14th Reliability Safety Assessment and Management (PSAM 14). In: 2018: UCLA Meyer & Renee Luskin Conference Center; 2018.
[24] Azarm MA, Kim IS, Marks C, Azarm F. Analyses methods and pilot applications of SACADA database, in 14th Probabilistic Safety Assressment and Management (PSAM 14). In: 2018: UCLA Meyer & Renee Luskin Conference Center; 2018.
[25] Nelson PF, Grantom CR. Methodology for Supporting the Determination of Human Error Probabilities from Simulator Sourced Data, in 14th Probabilistic Safety Assessment and Management (PSAM 14). In: UCLA Meyer & Renee Luskin Conference Center; 2018.
[26] Jung W, Park J, Kim Y, Choi SY, Kim S. HuREX – A framework of HRA data collection from simulators in nuclear power plants. Reliability Engineering & System Safety 2020;194:106235. 2020ISSN 0951-8320.
[27] Siu NO, Kelly DL. Bayesian parameter estimation in probabilistic risk assessment. Reliability Engineering & System Safety 1998;62(1):89–116.
[28] Kelly DL, Smith CL. Bayesian inference in probabilistic risk assessment - The current state of the art. Reliability Engineering & System Safety 2009;94(2):628–43.
[29] Apostolakis G, Kaplan S, Garrick BJ, Duphily RJ. Data specialization for plant specific risk studies. Nuclear Engineering and Design 1980;56(2):321–9.
[30] Kaplan S. On a two-stage Bayesian procedure for determining failure rates. IEEE Trans Power Apparatus Syst 1983;102(1):195–262.
[31] Droguett E, Groen F, Mosleh A. Bayesian assessment of the variability of reliability measures. Pesquisa Operacional 2006;26:109–27.
[32] Yue M, Chu T-L. Estimation of failure rates of digital components using a hierarchical Bayesian method. In: PSAM8 - International Conference on Probabilistic Safety; 2006. May 14-19, 2006.
[33] Mosleh A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. Reliability Engineering & System Safety 1992;38(1):47–57.
[34] Podofillini L, Dang VN. A Bayesian Approach to Treat Expert-Elicited Probabilities in Human Reliability Analysis Model Construction. Reliability Engineering & System Safety 2013;117:52–64.
[35] Droguett EL, Groen F, Mosleh A. The combined use of data and expert estimates in population variability analysis. Reliability Engineering & System Safety 2004;83(3):311–21.
[36] VanDerHorn E, Mahadevan S. Bayesian model updating with summarized statistical and reliability data. Reliability Engineering & System Safety 2018;172:12–24.
[37] Mosleh A, Chang YH. Model-based human reliability analysis: prospects and requirements. Reliability Engineering and System Safety 2004;83:241–53.
[38] Mosleh A, Smith C. The Feasibility Of Employing Bayesian Techniques And Other Mathematical Formalisms In Human Reliability Analysis. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study 2007:5–15. NUREG/CR-6949INL/EXT-06-11670.
[39] Atwood CL. Constrained noninformative priors in risk assessments. Reliab. Eng. Syst. Saf. 1996;53(1):37–46.
[40] Hallbert B, Gertman D, Lois E, Marble J, Blackman H, Byers J. The use of empirical data sources in HRA. Reliability Engineering & System Safety 2004;83:139–43.
[41] NEA/CSNI. Research on Human Factors in New Nuclear Plant Technology. NEA/CSNI/R(2009)7. Nuclear Energy Agency; 2009.
[42] Prvakova S, Dang VN. A review of the current status of HRA data. In: Proceedings of ESREL 2013, European Safety and Reliability Conference; 2013. 29 Sept. – 2 Oct.
[43] Skjerve AB, Bye A. Simulator-based Human Factor Studies across 25 years. London: Springer-Verlag; 2011.
[44] Hannaman GW, Spurgin AJ, Lukic Y. Human cognitive reliability model for PRA analysis. Palo Alto, California: EPRI Electric Power Research Institute; 1984. NUS-4531.
[45] Hannaman GW, Spurgin AJ, Lukic Y. A Model for Assessing Human Cognitive Reliability in PRA studies. In: IEEE Third Conference on Human Factors in Nuclear Power Plants. New YorkUSA: Institute of Electronic and Electrical Engineers; 1985. June 23-27.
[46] Moieni P, Spurgin AJ, Singh A. Advances in Human Reliability Analysis Methodology. Part I: Frameworks. Models and Data. Reliability Engineering and System Safety 1994;44:27–55.
[47] Massaiu S, Holmgren L. Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators. HWR-1121 2014. OECD Halden Reactor Project.
[48] Massaiu S, Holmgren L. The 2013 Resilient Procedure Use Study with Swedish Operators: Final Results. HWR-1216 2017. OECD Halden Reactor Project.
[49] Mkrtchyan L, Podofillini L, Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. Reliability Engineering & System Safety 2016;151:93–112.
[50] Gelman A, Carlin J, Stern H, Rubin D. Bayesian Data Analysis. Second Edition. Chapman and Hall/CRC; 2003.
[51] Kelly DL, Smith CL. Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook. London, UK: Springer-Verlag; 2011.
[52] Venables WN, Smith DM. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Version 3.6.1 (2019-07-05) 2019.