Nuclear Engineering and Technology 53 (2021) 2523–2533

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

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

Original Article

Uncertainty quantification and propagation with probability boxes

L. Duran-Vinuesa [a, *], D. Cuervo [b]

[a] Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Industriales, C/ de José Gutiérrez Abascal, 2, 28006, Madrid, Spain
[b] Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Navales, Avda. de La Memoria, 4, 28004, Madrid, Spain

ARTICLE INFO

Article history:
Received 5 October 2020
Received in revised form
21 January 2021
Accepted 7 February 2021
Available online 14 February 2021

Keywords:
Uncertainty-quantification
Uncertainty-propagation
Probability-boxes
BEPU
Wilks

ABSTRACT

In the last decade, the best estimate plus uncertainty methodologies in nuclear technology and nuclear power plant design have become a trending topic in the nuclear field. Since BEPU was allowed for licensing purposes by the most important regulator bodies, different uncertainty assessment methods have become popular, overall non-parametric methods. While non-parametric tolerance regions can be well stated and used in uncertainty quantification for licensing purposes, the propagation of the uncertainty through different codes (multi-scale, multiphysics) in cascade needs a better depiction of uncertainty than the one provided by the tolerance regions or a probability distribution. An alternative method based on the parametric or distributional probability boxes is used to perform uncertainty quantification and propagation regarding statistic uncertainty from one code to another. This method is sample-size independent and allows well-defined tolerance intervals for uncertainty quantification, manageable for uncertainty propagation. This work characterizes the distributional p-boxes behavior on uncertainty quantification and uncertainty propagation through nested random sampling.

1. Introduction

The first formulation of the 10-CFR-50 of the United States Nuclear Regulatory Commission (USNRC) in 1970 established general design criteria to assist in the preparation of the Nuclear Power Plant (NPP) licensing applications [1]. Deterministic Safety Analysis (DSA) mission is to demonstrate the NPP design compliance with that rules and was tackled from a conservative perspective in Appendix K to 10-CFR-50.46 applied to the Emergency Core Cooling System (ECCS) design, using conservative codes and input data. In the 80 s decade, with undergoing research on phenomena modeling as well as a better understanding of the underlying physics together with the development of the computational capabilities, the Best Estimate (BE) computer codes use arouse. BE codes can demonstrate safety margins compliance combined with conservative or realistic input data and boundary conditions [2].

During the 80s, the USNRC researched the development of a methodology to use realistic and physically-based analysis methods. BE calculations were first mentioned with regulatory purposes in 1989, accompanied by a complementary uncertainty analysis in the Regulatory Guide (RG) 1.157 – *Best-Estimate Calculations of ECCS Performance* [3]. The USNRC Code Scaling Applicability and Uncertainty (CSAU) methodology [4] was published the same year. Best Estimate Plus Uncertainty (BEPU) concept was born and developed in the 90s, but practical licensing applications started in the 2000s. A comprehensive historical revision of the development and advances in BEPU methodologies can be found in Ref. [5] as well as the Atucha-II NPP BEPU approach in Chapter 15 of the Final Safety Analysis Report (FSAR) [6]. The BEPU approach has found its application to the DSA with licensing purposes not only endorsed by Appendix K to10-CFR-50.46 and RG 1.157 for BE Loss of Coolant Accident (LOCA) analysis. RG 1.203 – *Transient and Accident Analysis Methods* published in 2005 that allows BE based evaluation models for a subset of transient events described in FSAR Chapter 15 and RG 1.70 – *Standard Format and Content of Safety Analysis Reports for Nuclear Power Plants* [7]. BE codes can be used if biases are accounted for, and some of them may not require a complete uncertainty analysis, *"However, in most cases, the Standard Review Plan (SRP) guidance is to use 'suitably conservative' input parameters"* [8].

Though BEPU is limited to the analysis of accidents, future research lines on BEPU face the FSAR adaptation to apply BEPU methods to each step of FSAR. Thus, creating an evolved BEPU-FSAR defining a proposal for improving nuclear reactors safety connecting the radiation protection As Low As Reasonably Achievable

* Corresponding author.
E-mail addresses: luisfelipe.duran@upm.es (L. Duran-Vinuesa), d.cuervo@upm.es (D. Cuervo).

https://doi.org/10.1016/j.net.2021.02.010

(ALARA) criterion for radiation exposure, Extended Safety Margins (ESM), and Independent Assessment (IA) towards new safety-barriers [9,10].

A photograph of uncertainty methodologies used worldwide can be found in Ref. [11,12] that can be classified according to their nature: deterministic or stochastic, with the latter being the most popular and widely used by the most relevant companies and regulator-bodies [13], and the one treated in this work. Before the uncertainty analysis, a Phenomena Identification and Ranking Table (PIRT) is deployed to identify the crucial parameters involved in the safety analysis and in a second step, for the selected phenomena modeling inputs and parameters, the State of Knowledge is to be characterized, based on existing literature or experimental data, for instance, utilizing statistical techniques like Bayesian Inverse Uncertainty Quantification (IUQ) [14].

These uncertainties are propagated through the code by any random sampling technique and large number of code-runs depending on the methodology. This process is called stochastic uncertainty propagation. The uncertainty analysis proceeds with the Uncertainty Quantification (UQ) method for one variable or a group of them, and the subsequent sensitivity analysis to investigate, for instance, the main contributor to the selected output uncertainty.

On the one hand, UQ methods can be based on non-parametric methods: Wilks' theory of Order Statistics [15,16] and Hutson fractional statistics [17] among other examples [18,19] that are compared in the work of Saez-Villanueva et al. [20]. These methods share the needless to assume the output variable obeys a concrete probability distribution. On the other hand, Dempster-Shafer structures, Theory of Evidence, and probability boxes (p-boxes) [21] suggest a mean to perform UQ based on distributional or parametric p-boxes construction.

For stochastic UQ, Wilks' formula is the trend because of its simplicity and the reduced sample sizes needed. However, reduced sample sizes may affect the quality of the sensitivity analysis. Higher orders of Wilks' formula that impose high fixed sample-sizes can guarantee the sensitivity coefficients convergence and a (95,95) tolerance interval. The fixed sample sizes' main drawback is that the size must increase for higher orders and become prohibitive for some calculations (multiphysics coupling, core burn-up simulations, and transients). It would be desirable to obtain (95/95) tolerance regions for any sample size that guarantees other statistics or the sensitivity analysis convergence. In this sense, Wilks' formula may result in a rigid one.

The illustrated method allows obtaining well-defined and conservative statistic regions for any sample-size in a flexible manner, showing a comparison against the non-parametric Wilks' centered formula [22] for a core thermal-hydraulics study case taken from the Uncertainty in Analysis in Modelling Benchmark [23]. Comparisons against other UQ methods can be found in recent work [43]. The benchmark aims to create a state-of-the art report on the current status and needs of uncertainty analysis, creating a roadmap for the development and validation of the methods required for uncertainty analysis. The benchmark's second phase (core phase) consists of uncertainty propagation on standalone physics simulations: bundle thermal-hydraulics among them. The selected case is a numerical exercise based on the Peach Bottom II BWR fuel assembly, corresponding to the Exercise III-1a, through which, core boundary conditions and modelling parameters uncertainties are propagated. In the present work, the maximum outer clad temperature of the fuel assembly is regarded to make a step-by-step depiction of the UQ method.

The benchmark in-depth analysis and results will be treated separately in future works that will address further details, applications and analysis of the results, comparing against available experimental data. Subsequently, the present work is devoted to establishing the basis of an uncertainty quantification method in the context of nuclear engineering safety analysis and BEPU methodologies.

This work treats the method's underlying theory and conceptual interpretation and performs a comparative assessment between the proposed probability-boxes technique and Wilks' formula in terms of robustness and conservatism. In Section 2, the statistical background and definitions are given together with the tolerance interval formulation. Section 3 presents the theoretical formulation of the distributional p-boxes method proposed, and Section 4 presents the analysis methods and results obtained. Finally, conclusions are derived from the analysis in Section 5.

2. A brief conceptual background on Figure of Merit and tolerance regions

According to a general definition, a Figure of Merit (FOM) is the quantity used to measure a system's performance, useful in comparing the quality of things or methods. For UQ purposes, the FOM can be defined as the Tolerance Region (TR) of the selected output variable, namely the “Minimum Departure from Nucleate Boiling Ratio” (MDNBR) - either “Minimum Departure from Nucleate Boiling” or “Minimum Nucleate Boiling Ratio".

Let $\boldsymbol{X}$ be a continuous random variable and $\boldsymbol{S}$ to be a random sample of $\boldsymbol{X}$. A two-sided TR (Eqn. (1)) is an interval $TR(S) = [L, U]$ such that at least a portionβ of $\boldsymbol{X}$ is contained in $\boldsymbol{TR}(\boldsymbol{S})$ (aka coverture, or first-order probability) with at least γ probability (aka confidence level, or second-order likelihood) [24,25].

$$P_S\{P_X\{X \in [L,U]\} \geq \beta\} \geq \gamma \approx P_S\{X_\beta \subset [L,U]\} \geq \gamma \tag{1}$$

This definition of TR gives a range of values that the variable $\boldsymbol{X}$ may take, accomplishing a well-defined statistic criterion (β, γ). For one-sided TRs, Eqn. (2) represents a well-defined condition to achieve. Namely, a unilateral (95,95) TR, also known as “tolerance limit”, is required by some regulatory bodies for the acceptance of BEPU based safety analysis [26] and hence, its importance.

$$P_S\{P_X\{X \leq U\} \geq \beta\} \geq \gamma \approx P_S\{U \geq X_\beta\} \geq \gamma \tag{2}$$

For two-sided TRs, there are infinite two-sided β regions drawn into a Probability Density Function (PDF) that can be approached by two-sided TRs in Eqn. (3). A more precise definition is that the region of $\boldsymbol{X}$ to enclose is the centered β region. Eqn. (3) includes this amendment for centered two-sided tolerance regions.

$$P_S\left\{L \leq X_{(1-\beta)/2} \cap U \geq X_{(1+\beta)/2}\right\} \geq \gamma \tag{3}$$

These definitions of TR are strongly connected to the conceptual interpretation of the uncertainty sources. While coverage β is a measure of the uncertainty caused by the inherent variability of the variable $\boldsymbol{X}$ and the lack of knowledge on its determination, the confidence level γ is linked to the epistemic uncertainty or lack of knowledge of $\boldsymbol{X}$, including statistic uncertainty when $\boldsymbol{X}$ is calculated from random sampling techniques. In uncertainty propagation, this relationship can be misleading because the code's output aleatory uncertainty embeds the inputs' variability and its epistemic uncertainty (i.e., the measurement error) together.

The output epistemic uncertainty will only be caused by the sample size's finitude, so-called statistic uncertainty [25]. Through the propagation of uncertainties, epistemic uncertainties are embedded in a cascade from one to another code.

Wilks' sample-sizes obtained for different orders in the work of Hohn and Connolly [22], extended by the author, are shown in

Table 1
Sample sizes to build non-parametric TRs based on Wilks' theory of order statistics.

Criterion	One-sided			Non-centered two-sided			Centered two-sided		
	1st order	2nd order	3rd order	1st order	2nd order	3rd order	1st order	2nd order	3rd order
90/90	22	38	52	38	65	91	58	93	124
95/95	59	93	124	93	153	208	146	220	286
99/99	459	662	838	662	1001	1307	1057	1483	1851

Table 1 for various Wilks' theory formulations (one-sided, non-centered two-sided, and centered two-sided). In the next section, the parametric method to calculate TRs according to Eqn. (3) is detailed. Note that Wilks' TRs only give bounding values, $[L, U]$, for the two-sided case.

Employing Wilks' formula is impossible to separate epistemic from aleatory uncertainty due to the limited information that the TR gives in this sense, and difficulties are found to propagate the output uncertainty through other codes. The main advantage of p-boxes is that epistemic and aleatory uncertainty can be treated separately. Indeed, distributional p-boxes offer more information about the output uncertainty on the imprecise probabilities space, for instance, its probability distribution. Thus, it is easier to propagate this uncertainty to downward codes in the chain of uncertainty propagation in multi-physics or multi-scale using classical Monte-Carlo nested sampling or more advanced optimized techniques [27].

3. A parametric UQ method based on probability boxes

The goal of the method is to obtain TR from the p-boxes. This analysis method is based on the hypothesis that the chosen variable belongs to a probability distribution. It is also essential to consider the random origin of the sample: statistical inference applied to uncertainty quantification must be performed only with a random sampling method and take special considerations when stratified Latin Hypercube Sampling (LHS) is used [28,29].

First-order probability could be calculated by integrating the centered β portion of the Cumulative Density Function (CDF) if the sample size would tend to the infinity. As sample-sizes are finite, PDF's parameters are determined with a concrete certainty level or confidence level, $(1-\alpha)=\gamma$, regarding statistic uncertainty. The p-box can be drawn by evaluating all possible combinations of Confidence Interval (CI) limits on the CDF: the maximum and minimum bounds enclosing the true CDF [30]. From the p-box, a TR can be calculated as stated in Eqn. (3).

The method consists of four essential steps to obtain $(\beta,\ \gamma)$ parametric TR from p-boxes:

a. Select the best probabilistic model or Cumulative Density Function (CDF) and the Maximum Likelihood Estimate (MLE) for its parameters
b. Parameters' confidence intervals CI calculation
c. P-box construction technique from the CDF and parameters' CI
d. $(\beta,\ \gamma)$ TR calculation

2.1. Parametric family selection based on statistical inference

The first step is to fit the best probabilistic model to the sample employing the Likelihood Function (LF). Let $\boldsymbol{X}$ be a Simple Random Variable and $S(X)$ a Simple Random Sample of $\boldsymbol{X}$ so that $S(X)=\{x_1, x_2, \ldots, x_n\}$. The function $L_i$defined for a probabilistic model, f^*, depending on the vector of parameters $\theta=\{\theta_1, \theta_2, \ldots, \theta_\pi\}$ in Eqn. (4) means the probability of one element x_i to be observed for a given θ_0. The probability of occurrence of the sample $S(X)$ is given by Eqn. (5), which is the definition of Likelihood Function, $L(\theta)$. For numerical reasons, the Log-Likelihood Function, $\mathscr{L}(\theta)$, is defined in Eqn. (6).

$$L_i(\theta_0, x_i)=P(X=x_i|\theta=\theta_0\ )=f_X^*(x_i;\theta_0) \tag{4}$$

$$L(\theta)=K\prod_{i=1}^{m}L_i(\theta, x_i) \tag{5}$$

$$\mathscr{L}(\theta)=\ln[L(\theta)]=K'\sum_{i=1}^{m}\mathscr{L}_i(\theta, x_i) \tag{6}$$

The $\mathscr{L}(\theta)$ maximization problem is addressed for each model f^*. Eqn. (7) is solved and the maximum likelihood estimator (MLE), $\widehat{\theta}$, is obtained as the value that maximizes $\mathscr{L}(\theta)$. MLE is found for a total amount of 8 models in this work. The model with the lower Akaike Information Criterion (AIC) in Eq. (8) value will represent the best the original sample distribution accounting a penalty for the model's complexity by its number of parameters, k. It is adequate a χ^2-Goodness-of-Fit (GOF) [31], Anderson-Darling GOF [32], or Kolmogorov-Smirnov GOF [33] to support the results.

$$\frac{\partial\mathscr{L}(\theta)}{\partial\theta_i}=0\ \rightarrow\mathscr{L}(\widehat{\theta})=\max(\mathscr{L}(\theta)) \tag{7}$$

$$AIC=2k-2\mathscr{L}(\widehat{\theta}) \tag{8}$$

Let us take an example to illustrate the optimization process. The selected variable is the maximum outer clad temperature in [K] of the Peach Bottom II BWR fuel assembly, taken from the Uncertainty in Analysis and Modelling Benchmark, Phase II, Exercise III-1a [23]. $\mathscr{L}(\widehat{\theta})$ is calculated for 8 models, and the three best fits are shown in Table 2. The normal distribution fits the best the sample with the highest $\mathscr{L}(\widehat{\theta})$ value, endorsed by the χ^2-GOF test that means the null hypothesis cannot be rejected with a p-value of about 92.98%, far from the minimum 5% for a 95% confidence. Fig. 1 plots $\mathscr{L}(\theta)$ for the Normal distribution parameters, $[\mu,\ \sigma]$, and the maximum likelihood estimator, $\widehat{\theta}=[\widehat{\mu},\ \widehat{\sigma}]=[568.68,\ 0.19]$.

2.2. Confidence intervals calculation

The second step is to describe the CI of the parameters that define the probabilistic model regarding the estimator's statistic

Table 2
Maximum LLF and χ^2-GOF test results for the three best models fitted.

Probabilistic Model	$\mathscr{L}(\widehat{\theta})$	χ^2-GOF test	P-value (%)
Normal	212.5989	0	92.98
Nakagami	212.5961	0	92.86
Birnbaum-Saunders	212.5906	0	92.79

Table 3
Confidence numerical estimation for some probability distribution.

Distribution	Beta	Nakagami	BB-Saunders	Rician
C_{CC} (%)	97.15	100	99.34	99.43
Distribution	**Rayleigh**	**Normal**	**GE-Value**	**Logistic**
C_{CC} (%)	96.12	98.83	99.95	98.60

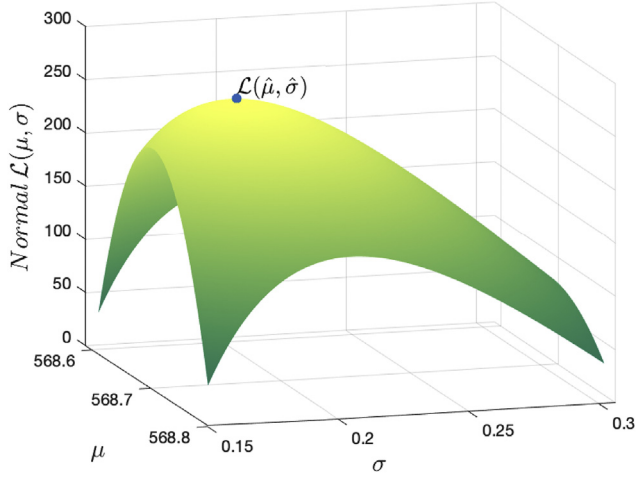

Fig. 1. Normal log-likelihood function $\mathscr{L}(\theta)$ and calculated MLE $\widehat{\theta}$

uncertainty. CIs can be obtained by different methods: the Log-likelihood Ratio (LLR) test and the Wald test, amongst others [34]. Wald's method validity is limited to the condition that $\mathscr{L}(\theta)$ can be well approximated by a quadratic function, and a comparison against LLR-test is conducted in the work of Pawitan [35]; thus, the LLR test is chosen.

LLR test is based on the Profile Likelihood, $R(\theta)$: the normalized likelihood function projection onto each parameter dimension [24], see Eqn. (9). The statistic LLR_i for composite testing hypotheses is defined in Eqn. (10) and a test that rejects the null hypothesis that $\theta_i = \widehat{\theta}_i$ provides the CI at α significance level [36], where π is equal to the number of free distribution parameters (degrees of freedom). The CI for θ_i then verify the condition in Eqn. (11), as depicted in Fig. 2.

$$R(\theta_i) = \max_{\forall N-i}\left[\frac{L\left(\theta_i, \widehat{\theta}_2, \ldots, \widehat{\theta}_N\right)}{L(\widehat{\theta})}\right] \tag{9}$$

$$LLR(\theta_i) = -2\ln[R(\theta_i)] > \chi^2_{(1-\alpha,\pi)} \tag{10}$$

$$R(\theta_i) > \exp\left[-\frac{\chi^2_{(1-\alpha,\pi)}}{2}\right] \tag{11}$$

Fig. 2 shows the projection of this normalized $\mathscr{L}(\theta)$ onto the μparameter dimension, cut by the LLR test condition that defines de CI for the example in step 1. The significance level of the CI must be addressed carefully to preserve rigor. Given a collection of m sets $\{A_1, A_2, \ldots, A_m\}$ and being $1-\alpha_i$ the probability associated with A_i, the intersection likelihood for a joint significance level α_J, expressed as follows, $1-\alpha_J = P(\cap_{i=1}^m A_i)$, depends on the relation between sets A_i. A conservative bound for the joint significance level is provided by Bonferroni's inequality [37] in Eqn. (12) to

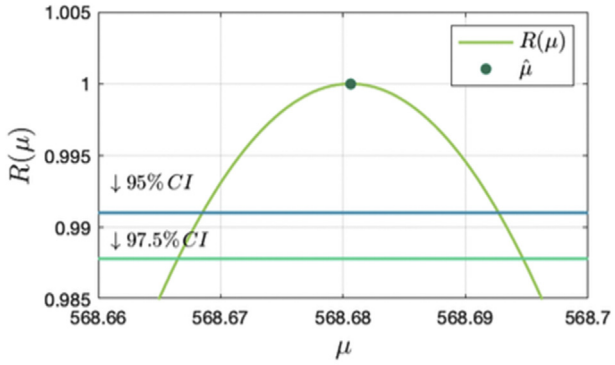

Fig. 2. Representation of the LLR test for $\hat{\mu}$ and estimator CI calculation.

comply with a joint criterion for multiple FOMs. For instance, two simultaneous (95,95) TRs at a significance level $\frac{\alpha_i}{2} = 0.025$, throw a joint confidence level $1-\alpha_J = 0.95$.

$$P(\cap_{i=1}^m A_i) = 1-\alpha_J \geq 1-\sum_{i=1}^m \overline{P}(A_i) = 1-\sum_{i=1}^m \alpha_i \tag{12}$$

2.3. P-box construction method

The next step involves the concept of p-boxes as established by Ferson et al. [21]: p-boxes are probability regions confined by two non-decreasing functions $\left[\underline{F}, \overline{F}\right]$. These functions depict a real and imprecisely known probability distribution F and meet the following properties:

$$\overline{F}(x) \subseteqq [0,1]; \underline{F}(x) \subseteqq [0,1]; \forall\, x \in \mathbb{R} \tag{13}$$

$$\overline{F}(x) > F(x) > \underline{F}(x) \tag{14}$$

$$\overline{F}(x) = 1-\underline{P}(X > x); \underline{F}(x) = \underline{P}(X \leq x) \tag{15}$$

For a Normal CDF depending on two parameters, four CDFs from all combinations of CIs can be used to build a 95% confidence p-box

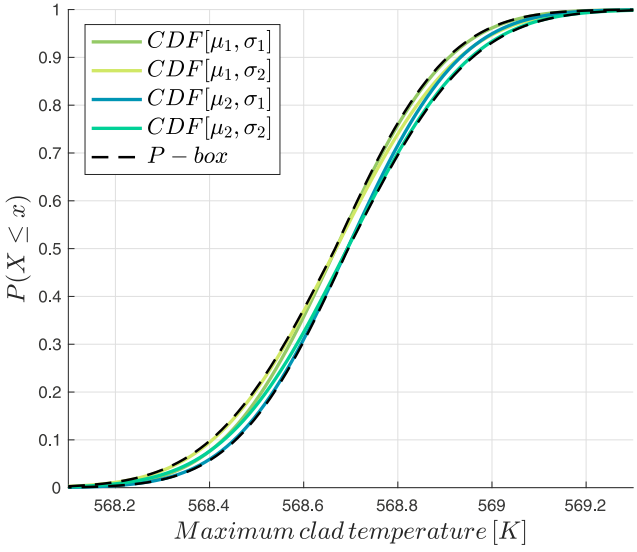

Fig. 3. Probability box composition scheme.

by the maximum and minimum value of the calculated CDFs (see Fig. 3).

$$\overline{F}(x_i) = \max(CDF_1(x_i), \ldots, CDF_m(x_i)) \tag{16}$$

$$\underline{F}(x_i) = \min(CDF_1(x_i), \ldots, CDF_m(x_i)) \tag{17}$$

2.4. Tolerance region calculation and uncertainty quantification

The last step is to calculate the uncertainty or TR by integrating the p-box components $\left[\underline{F}, \overline{F}\right]$. The expression in Eqn. (18) is proposed to generate centered TRs, according to Eqn. (3) for a criterion (β, γ).

$$\Delta \boldsymbol{X} = TR = \left[\overline{F}_{(1-\beta)/2}, \underline{F}_{(1+\beta)/2}\right] \tag{18}$$

3. Validation and performance analysis

The assessment and validation of the distributional p-boxes method are presented in this section. Subsection 4.1 goes through the p-box construction method, excluding the initial inference part. Performance metrics of the method, including inference, are presented in subsection 4.2. The third part is focused on the sensitivity of the p-boxes to fitting parameters and the validity of the method hypothesis of inference. The last part of the analysis investigates the propagation of uncertainty using p-boxes with a Monte-Carlo based algorithm.

Robustness measures the statistical fluctuations that occur when the same method is carried out several times with different random samples of the same size for the same FOM. The variation of the TR limits changes the coverage of the TR each time. A robust method minimizes these statistical fluctuations and tends to offer pretty similar results between realizations. The metrics used to estimate the robustness and conservativeness of the methods are defined in the following lines.

To measure the robustness, a mother sample of $i = 1, .., Z = 100,000$ elements has been generated from the simulation in Section 3 to ensure the four statistical moments (mean, standard deviation, skewness, and kurtosis) convergence under 1%. An amount of $j = 1, \ldots, M = 15,000$ data subsets are obtained by subsampling N number of elements. N equals Wilks' centered formula sample sizes (see Table 1) to compare against (95,95) TRs from Wilks' centered method for different orders. The coverture statistics can be defined, starting with the coverage of each subset M:

$$C_j = \frac{1}{Z}\left(\sum_{i=1}^{Z} I[x_i \in TR_j]\right) \tag{19}$$

$$\begin{cases} I[a] = 1 \;\; if \;\; a = true \\ I[a] = 0 \;\; if \;\; a = false \end{cases}$$

Then, the mean and standard deviation of the coverage can be written as follows in Eqns. (20) and (21):

$$C_\mu = \frac{1}{M}\sum_{j=1}^{M} C_j \tag{20}$$

$$C_\sigma = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(C_j - C_\mu)^2} \tag{21}$$

The coefficient of variation, C_{CV}, measures the method's robustness as the relative standard deviation of the covered portion of the mother sample (Eqn. (22)).

$$C_{CV} = \frac{C_\sigma}{C_\mu}\cdot 100 \tag{22}$$

Confidence is estimated by the coefficient of conservativeness, C_{CC}, defined for a reference region to be covered by the TR. In this case, the centered 95% quantile of the mother sample: $SV_{ref} = [P_{2.5}, P_{97.5}]$.

$$C_{CC} = \frac{1}{M}\left(\sum_{j=1}^{M} I\left[SV_{ref} \subset TR_j\right]\right)\cdot 100 \tag{23}$$

The coefficient of conservativeness means the portion of the M TRs calculated that cover the reference region. As Z and M tend to infinity, the coefficient of conservativeness equals the method's confidence because the reference region approaches the real centered 95% probability of the selected random variable.

3.1. TRs construction method validation

A validation exercise requires to know the real probability regions before the test. Subsequently, the probabilistic model is a priori known. The TR construction method is validated by random sampling for each of the known distributions against the known β probability regions to be covered. Fig. 4 shows the centered 95% probability regions [-1, 1], and the (95,95) TRs for different probability distributions and sample sizes. The confidence estimation for each distribution is shown in Table 3.

The reader can observe two kinds of fluctuations: the first one is the lower and upper limits' asymmetric behavior. The second one

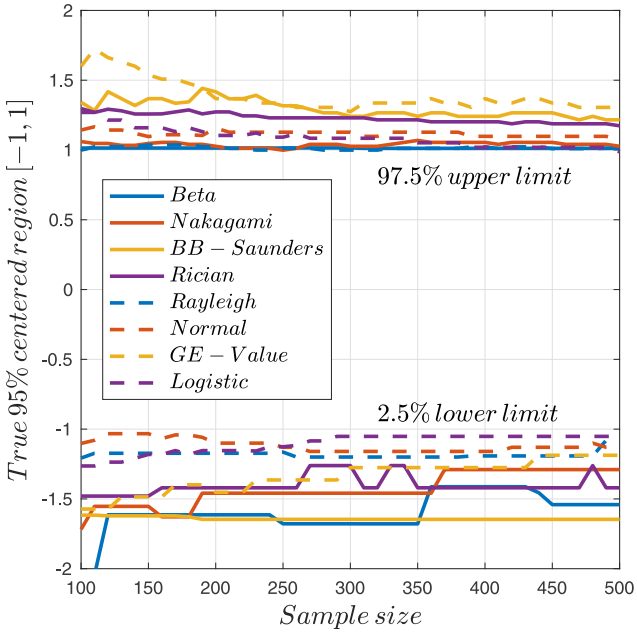

Fig. 4. Normalized (95,95) TRs for different distributions' approach to the real 95% centered probability region.

depends on the probability distribution itself. Though all the distributions were tested for all sample sizes, some of them proved a better fitting. These distributions show minor fluctuations and reduced TRs than those which fit the worse (Subsection 4.3). Moreover, the p-box algorithm may stick the limits to a particular value for those distributions with higher CI (extreme cases). This bias explains the straight limits evolution for the wider TRs given by the worst fits. However, the best models show fewer fluctuations and narrower CIs, and the algorithm performs very well for non-extreme cases, the ones selected to perform UQ. Besides, the treatment of the worst cases suggests an improvement area for the algorithm's future development.

3.2. TR construction method performance and comparison against Wilks method

Let us define the FOM as the maximum outer clad temperature in [K] of the Peach Bottom II BWR fuel assembly, taken from the Uncertainty in Analysis and Modelling Benchmark, Phase II, Exercise III-1a [23]. Metrics defined in the introduction of this Section 4 are obtained for the p-box method and the centered Wilks' method to compare both methods' behavior at different sample sizes (or orders). Table 4 shows the metrics' results.

The same trend can be observed for both methods in Fig. 5 (left) for the mean coverage. Fig. 5 (center) shows similar robustness of around 1%. Wilks' method robustness improves with the increasing sample size while p-boxes robustness seems to remain constant. Both methods show a constant confidence coefficient, higher for the p-boxes method in Fig. 5 (right). This result shows the more conservative character of the p-boxes method.

The maximum clad temperature uncertainty bands are plotted in Fig. 6 as the TR width or ΔT. To reduce statistical noise, Fig. 6 plots the Bootstrap mean estimate of the TR widths and the Bootstrap estimate of the Standard Error of the Mean (SEM), $\widehat{\sigma}_B$ [38]. P-boxes and Wilks' methods follow the same trend with the sample size: as the coverage diminishes, the TR limits get tighter to give smaller TRs and offer similar results (see Table 5).

The reader may realize that the provided uncertainty ranges are narrower than 1 K for some cases. This is an uncertainty range lower than the accuracy of the experimental measures from which the correlations are developed that can be justified because heat transfer parameters uncertainties are not addressed in this numerical exercise, where only boundary conditions, pressure losses and mixing parameters' uncertainties are concerned.

3.3. Tolerance regions sensitivities to p-box construction parameters

This subsection goes through the TR sensitivity to the similarity of the sample to the distributional model used to construct the TR. A crucial hypothesis of the p-boxes method is that the sample obeys a probabilistic model chosen by its best similarity. Despite this, the model will not perfectly fit the sample. Other models can be added to the fitting algorithm to cope with this fact, or at least, the researcher must be aware of the effect of the disparity between the sample and the probabilistic model on the TR width.

Fig. 7 shows the TR's width sensitivity to parameters involved in the fitting process: the maximum log-likelihood of the fit and the p-value of the GOF tests that endorse the fitted distribution. Fig. 7 (a) shows the TR's width trend with the maximum log-likelihood to measure the similarity between the observed sample and the fitted model. The greater the similarity, the tighter the TR will be. This fact suggests that worse models throw conservative TRs. Models with a higher number of parameters deserve special consideration giving wider TRs (i.e. Generalized Extreme Value distribution [39] vs. Normal distribution).

The p-value of a given GOF test is used to fit a probabilistic model to the sample. However, the only information that the p-value gives is the significance level, which is not linked to the distribution similarity to the sample (i.e., one could not select a model based on p-value ranking), as suggested in Fig. 7(b).

The p-value is not correlated to the TR's width. A positive correlation between sample-size and maximum log-likelihood values is shown in Fig. 7(c). Hence, the trends observed in Fig. 6 are justified by the greater degree of similarity reached for higher sample sizes.

3.4. Uncertainty quantification at any sample-size

In the present subsection, one of the advantages of the non-parametric method is tested: its capability to calculate a given $[\beta, \gamma]$ TRs independently on the sample-size. The standard Normal PDF is taken with a $\beta_{centered} = [-1.96,\ 1.96]$ in a numeric exercise to test this feature. TRs for sample sizes from 10 to 500 have been obtained. Fig. 8 (a) shows the TRs limits for this numerical exercise, and (b) the TRs calculated for the maximum outer clad temperature of the study case illustrated.

Sample sizes in the exercise cover from 10 to 500. During the algorithm performance for the smaller sample sizes, the code throws the widest TRs and needs considerably more computational effort. Moreover, the reader should realize that very low sample sizes reduce the quality of the analysis. Several tens of code runs are needed to avoid TRs that double the central region to be covered. However, it seems the only sample size limit is the algorithm's capability to apply statistical inference.

A last concern on the comparison against the Wilks' method is that, though Wilks formula is sample size dependent, Fig. 5 shows the Wilks' estimated confidence level value of 95% while P-boxes confidence level is overestimated. Wilks' sample size derivation includes the imposition of 95% confidence level in the sample size obtention, giving the exact 95% confidence level unlike the p-boxes method, which imposes the 95% confidence level in the TR calculation, but offers overestimated confidence levels. In conclusion, TRs calculated by Wilks formula properly preserve the TR statistic design criterion (β, γ) thus establishing a reference to the comparison but finds a limitation on the sample size that can be mitigated by means of the p-boxes method that has demonstrated its capability to equal Wilks formula TRs in despite of the confidence level overestimation.

Table 4
Coverage statistics and metrics results (Sample Size, SS).

	1st order/SS 146		2nd order/SS 220		3rd order/SS 286		4th order/SS 345	
Method	p-box	Wilks	p-box	Wilks	p-box	Wilks	p-box	Wilks
C_μ	98.68	98.65	98.30	98.21	98.05	97.92	97.83	97.73
C_σ	0.0072	0.0094	0.0074	0.0089	0.0070	0.0083	0.0069	0.0079
C_{CV} (%)	0.7311	0.9517	0.7479	0.9070	0.7130	0.8478	0.7006	0.8116
C_{CC} (%)	99.40	95.24	99.37	95.98	99.56	95.18	99.51	95.05

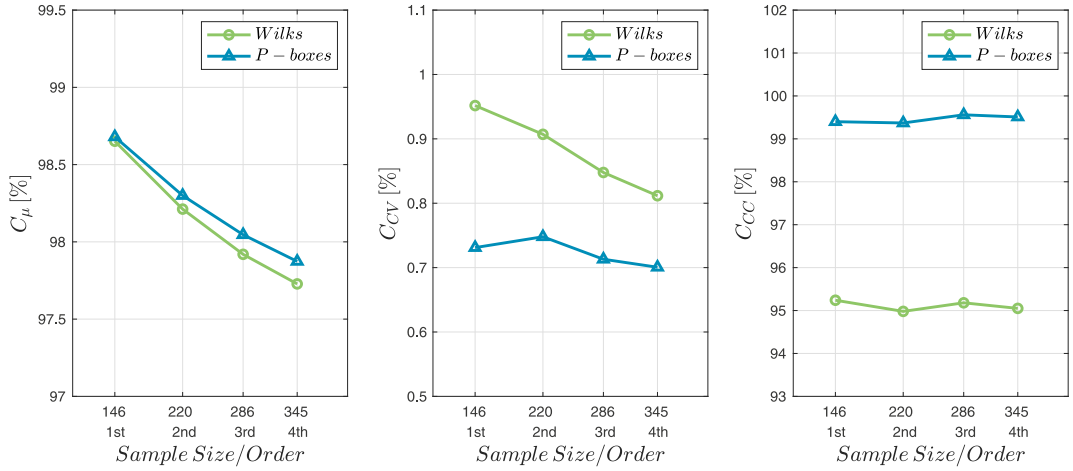

Fig. 5. Metrics comparison for Wilks centered formula and p-boxes method.

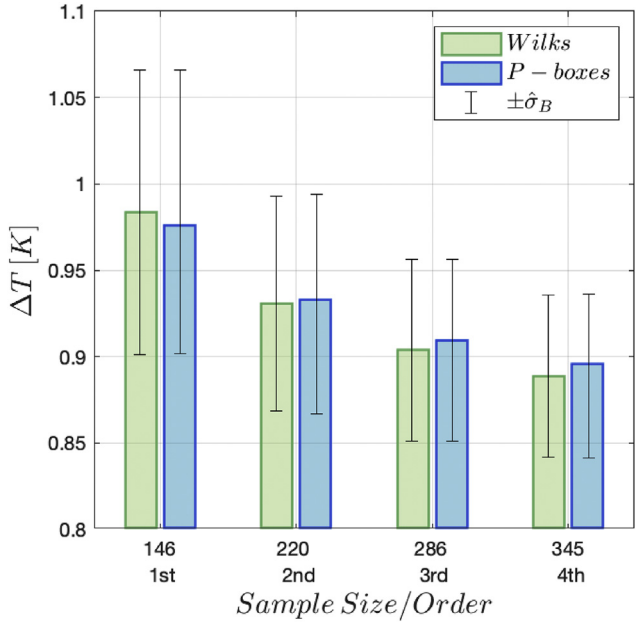

Fig. 6. Bootstrap estimates for maximum clad temperature uncertainty or 95/95 TR widths.

3.5. Stochastic uncertainty propagation with a basic nested sampling algorithm

Another of the strengths of the parametric method is the possibility to preserve statistic uncertainty through uncertainty propagation. When the input parameter comes from another uncertainty propagation exercise, it is usual to find a probability distribution with its parameters CIs (this is, a p-box definition). A Monte-Carlo based nested sampling is used to sample p-boxes: firstly, the distribution parameters are sampled assuming a uniform distribution for each one, and then, the probability distribution with these parameters is sampled. This method can be understood as the code's output uncertainty surrogate that directly generates outputs samples instead of a more complex surrogate for the whole code: this method generates samples of the output variable in its uncertainty range, preserving its probability distribution. Instead of building a surrogate model for the code and make the input perturbations, the probability-box sampling algorithm can generate more output samples than those originals from which the output probability distribution was generated. In this case, the greater sample size generated will not reduce the output aleatory uncertainty, unlike the original model runs or the code's surrogate output. Some p-boxes sampling methods based on MC and Polynomial Chaos Expansion (PCE) are briefly described in Ref. [27,40].

In this section, a comparative exercise on reproducing samples from p-boxes and samples from the original code results is performed, comparing different sample types:

1. Sample size 10,000, from the original code-runs. (Case O).
2. Sample size 10,000, from a p-box built from 146 original elements. (Case A).
3. Sample size 10,000, from a p-box built from 220 original elements. (Case B).
4. Sample size 10,000, from a p-box built from 286 original elements. (Case C).
5. Sample size 10,000, from a p-box built from 345 original elements. (Case D).
6. Sample size 10,000, from a p-box built from 10,000 original elements. (Case Z).

The mean and standard deviation residuals (Eqn. (24) and Eqn. (25), accordingly) are plotted in Fig. 9 for cases O, A, and D. The convergence is similar for the original model and the p-box model nested sampling.

$$R_{\mu i} = 100 \cdot \frac{\mu_i - \mu_{i-1}}{\mu_{i-1}} [\%] \tag{24}$$

$$R_{\sigma i} = 100 \cdot \frac{\sigma_i - \sigma_{i-1}}{\sigma_{i-1}} [\%] \tag{25}$$

A one-way ANOVA [41,42] has been carried out with 1000 realizations for each sample type by resampling the 10,000 elements original sample. The ANOVA tests the null hypothesis that the

Table 5
Bootstrap two-sided TRs estimates.

Method	Wilks	P-boxes
1st **order SS 146**	[568.17, 569.16]	[568.19, 569.16]
2nd **order SS 220**	[568.19, 569.13]	[568.20, 569.14]
3rd **order SS 286**	[568.21, 569.12]	[568.22, 569.13]
4th **order SS 345**	[568.22, 569.10]	[568.22, 569.11]

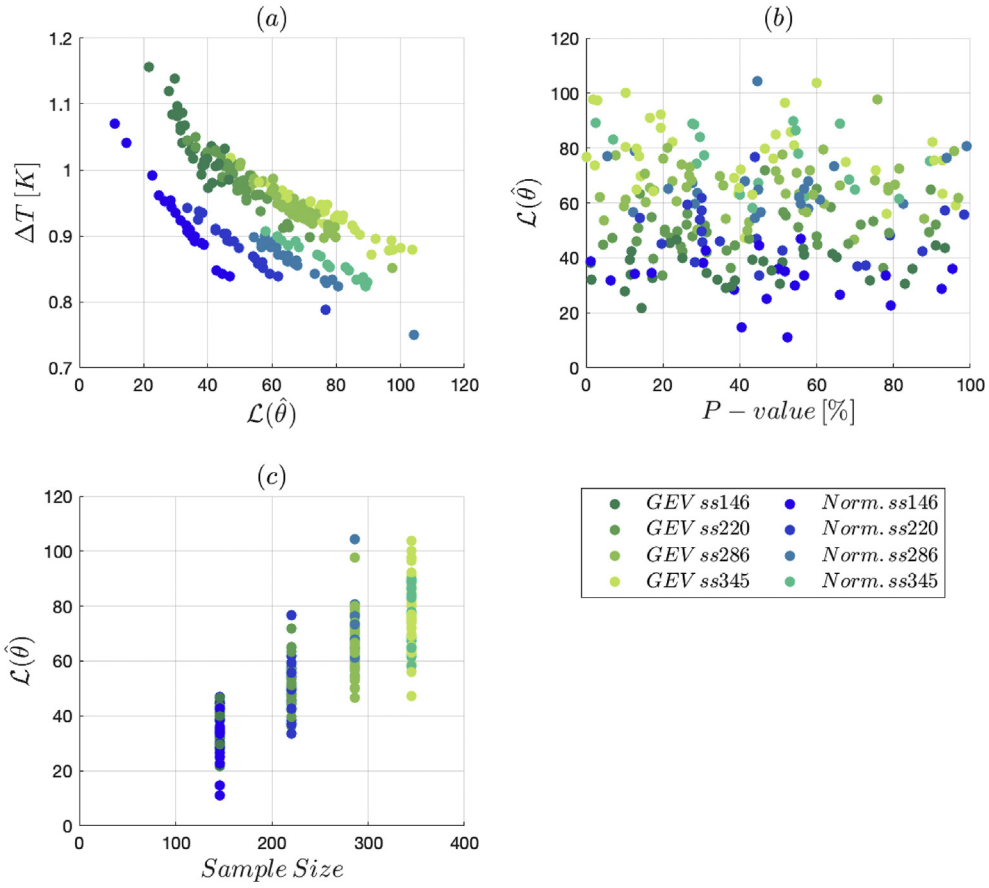

Fig. 7. A sensitivity study on the TRs obtained by the p-boxes method.

different sample types are drawn from populations with the same mean against the alternative hypothesis that the population means are different. Table 6 shows the ANOVA table. The analysis reveals the variability comes from differences between data in each group instead of differences between the groups' mean and the null hypothesis above is accepted with a p-value of 0.25. Subsequently, one could claim that all the sample types come from a population with the same mean at a significance level greater than 5%.

Differences between groups appear regarding the sparsity of the population mean. A box-chart is plotted in Fig. 10 for the mean and standard deviation of each sample type. The original sample shows small statistic uncertainty uniquely caused by its finite sample size.

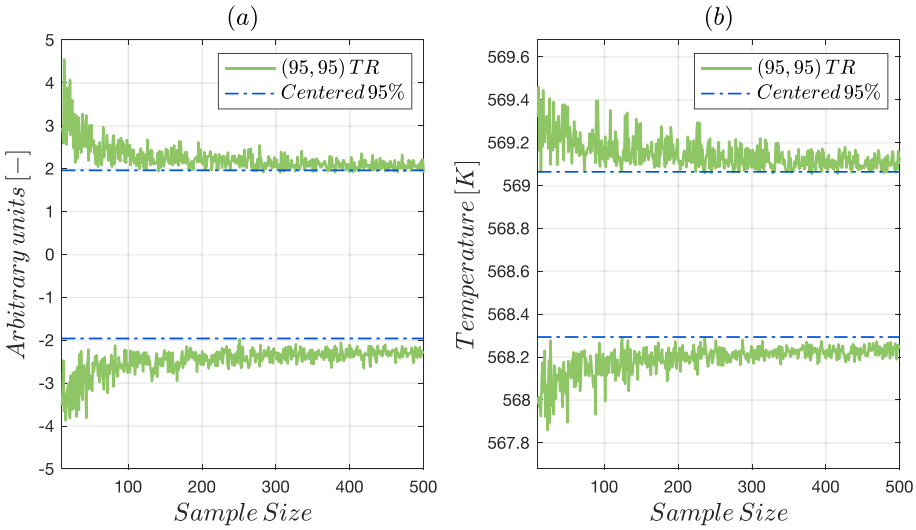

Fig. 8. (95,95) Tolerance Regions for the normal case (a) and the maximum clad temperature of the benchmark case (b).

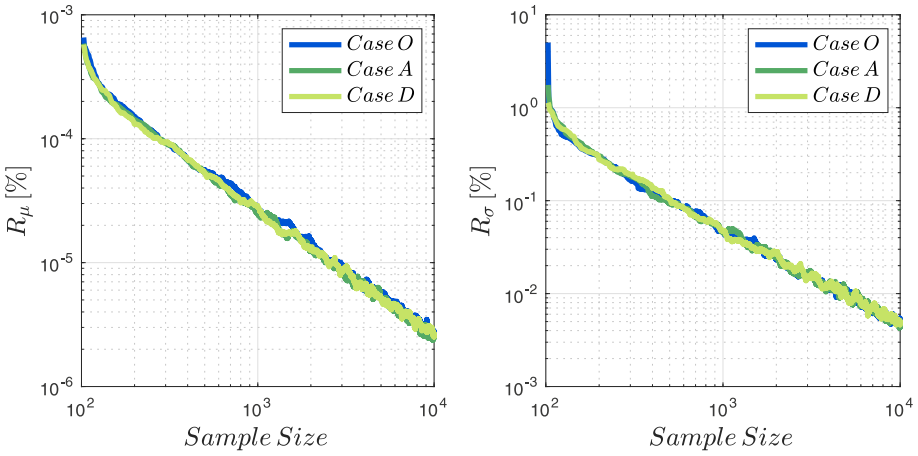

Fig. 9. The moving mean of the convergency rate for the mean and the standard deviation.

Table 6
ANOVA table.

Source	Sum of squares	Degrees of freedom	Mean Squared Error	F-statistic	P-value
Groups	0.00168	5	0.00034	1.32	0.2543
Error	1.5332	5994	0.00026		
Total	1.53488	5999			

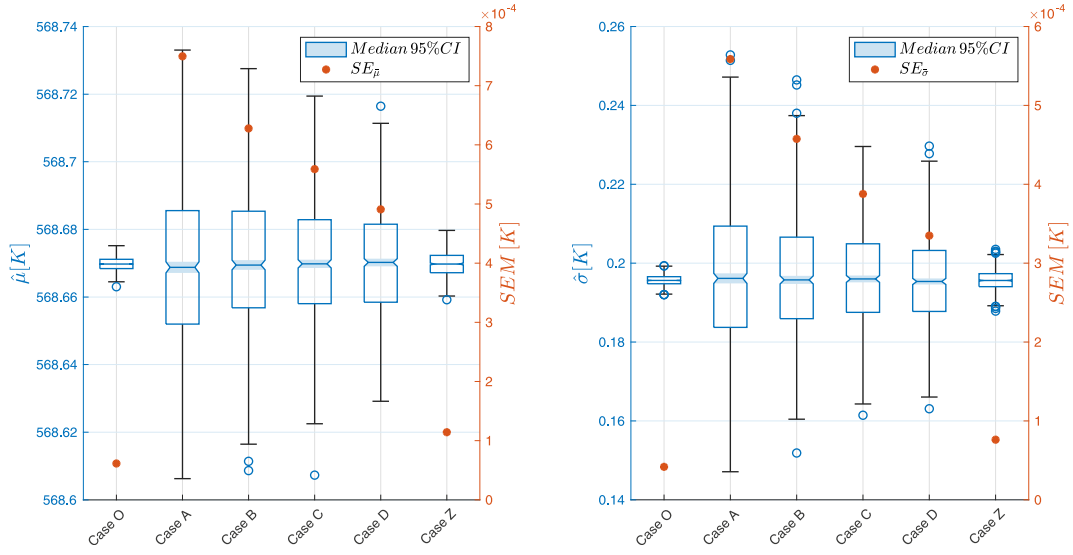

Fig. 10. Mean of each sample type box plot and Standard Error of the Mean.

However, samples taken from p-boxes show a higher interquartile range (cases A, B, C, and D). This variability mostly comes from the reduced sample size from which the p-box was drawn despite the equal sample size between groups, considering epistemic uncertainty.

For the same sample size, a greater sparsity of the mean and standard deviation in case Z suggests another uncertainty source: the p-boxes calculation itself. While sample O is generated from code outputs, Z is obtained from a p-box drawn from the elements in O. Case Z presents a slightly higher interquartile range and SEM, caused by the p-box sampling method itself (i.e., the hypothesis of the uniform distribution of the mean and standard deviation) with a despicable effect in the practical sense. Subsequently, the p-box introduces bias, characterized by the SEM. The bias order of magnitude is ~10^{-4} for all cases because of the great sample size. Between groups, the order of magnitude remains. Consequently, p-boxes show a good performance on uncertainty propagation, preserving the original sample's epistemic uncertainty.

4. Conclusions

An in-depth analysis of the distributional probability boxes (p-boxes) has been carried out to demonstrate their applicability to BEPU methodologies. For uncertainty quantification, a method to

build tolerance regions (TR) from p-boxes is presented and compared against the Wilks' formula theory TRs. Then, the p-boxes method performance has been analyzed with the following results:

1. The TR construction method has demonstrated that valid TRs can be obtained from distributional p-boxes preserving aleatory and epistemic uncertainty separation (Fig. 4).
2. TRs obtained are like those of Wilks in terms of coverture or first-order probability but show a more conservative character in terms of confidence or second-order probability (Fig. 5). In terms of coverture, the p-box method behaves like the Wilks formula: decreasing similarly with the sample size (Fig. 6).
3. The TR construction method's most critical hypothesis is the probabilistic model of the p-boxes in the fitting algorithm. Numerical experiments have demonstrated its capability and flexibility to perform UQ at practically any sample size (Fig. 8) and that the worse models throw the more conservative results (Fig. 7). Even accounting for the worst models, the TR construction algorithm performs acceptably on uncertainty quantification. Models depending on a higher number of parameters, with the same maximum likelihood criterion, will give wider TRs. The p-value of the GOF test should not be used to rank models on the list.

A final statement on the p-boxes and its comparison against Wilks uncertainty quantification method is the main advantage of p-boxes are its capability to compute TRs independent on the sample size but overestimating confidence level while Wilks formula precisely characterizes the confidence level but only for limited sample sizes.

For uncertainty propagation, uncertainty taken from upstream codes may be in the form of a distributional p-box, given by a probability distribution, its parameters, and their confidence intervals. A basic nested-sampling algorithm analysis has demonstrated its validity on uncertainty propagation with the following remarks:

1. The p-box nesting-algorithm does not improve the statistical moments' convergence (Fig. 8) and behaves like the original code uncertainty.
2. The p-box description of the output uncertainty reproduces the original code uncertainty. In this sense, p-boxes performs like a surrogate model for the code to reproduce the output uncertainty. This technique has been demonstrated (Fig. 10).
3. The p-box definition includes statistical uncertainty that will be propagated in the form of the distribution parameters sparsity. The analysis suggests that the nesting method introduces more uncertainty with a despicable effect (Fig. 10). Moreover, the sensitivity analysis shows that bad or inaccurate distribution fits will propagate higher uncertainty, protecting against uncertainty underestimation.

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.

Acknowledgements

The authors would like to express their gratitude to the Programa Propio of the Universidad Politécnica de Madrid, without its founding, this research could not had been possible.

References

[1] USNRC, "10 CFR Part 50 - Domestic Licensing of Production and Utilization Facilities." .
[2] Safety Assessment and Verification for Nuclear Power Plants: Safety Guide, IAEA, Vienna, 2001.
[3] Regulatory Guide 1.157 (Task RS 701-4), Best-Estimate Calculations of Emergency Core Cooling System Performance.," p. 20.
[4] B. Boyack, et al., Quantifying Reactor Safety Margins: Application of Code Scaling, Applicability, and Uncertainty Evaluation Methodology to a Large-Break, Loss-Of-Coolant-Accident, *NRC NUREG-Series Publications/CR-5249*, 1989 iv.
[5] F. D'Auria, C. Camargo, O. Mazzantini, The Best Estimate Plus Uncertainty (BEPU) approach in licensing of current nuclear reactors, Nucl. Eng. Des. 248 (2012) 317–328, https://doi.org/10.1016/j.nucengdes.2012.04.002.
[6] F. D'Auria, O. Mazzantini, The best-estimate plus uncertainty (BEPU) challenge in the licensing of current generation of reactors, in: *Proceedings Of an International Conference On Opportunities And Challenges For Water Cooled Reactors In the 21st Century*, Vienna, Austria, 27-30 *October 2009*, Vienna, 2011, p. 13 [Online]. Available: http://www-pub.iaea.org/MTCD/Publications/PDF/P1500_CD_Web/htm/pdf/topic4/4S08_F.%20D. Auria.
[7] N.R.C. US, Regulatory Guide 1.70, standard Format and content of safety analysis reports for nuclear power plants - LWR edition, Rev 3 (" Dec. 1978).
[8] N.R.C. US, Regulatory Guide 1.203, Trans. Acc. Anal. Methods (2005) 53.
[9] F. D'Auria, N. Debrecin, H. Glaeser, Strengthening nuclear reactor safety and analysis, Nucl. Eng. Des. 324 (Dec. 2017) 209–219, https://doi.org/10.1016/j.nucengdes.2017.09.008.
[10] F. D'Auria, N. Debrecin, H. Glaeser, The technological challenge for current generation nuclear reactors, NUCET 5 (3) (Sep. 2019) 183–199, https://doi.org/10.3897/nucet.5.38117.
[11] IAEA, Best Estimate Safety Analysis for Nuclear Power Plants: Uncertainty Evaluation, Internat Atomic Energy Agency, Vienna, 2008.
[12] L.L. Briggs, Nuclear Engineering Division, "Uncertainty Quantification Approaches for Advanced Reactor Analyses, *Argonne National Lab. (ANL)*, Argonne, IL (United States), 2009, https://doi.org/10.2172/956921.
[13] C. Frepoli, J.P. Yurko, R.H. Szilard, C.L. Smith, R. Youngblood, H. Zhang, 10 CFR 50.46c rulemaking: a novel approach in restating the LOCA problem for PWRs, Nucl. Technol. 196 (2) (Nov. 2016) 187–197, https://doi.org/10.13182/NT16-66.
[14] X. Wu, T. Kozlowski, H. Meidani, K. Shirvan, Inverse uncertainty quantification using the modular Bayesian approach based on Gaussian process, Part 1: Theory, Nucl. Eng. Des. 335 (Aug. 2018) 339–355, https://doi.org/10.1016/j.nucengdes.2018.06.004.
[15] S. Wilks, Determination of sample sizes for setting tolerance limits, Ann. Math. Stat. 12 (1) (1941) 91–96, https://doi.org/10.1214/aoms/1177731788.
[16] S. Wilks, Statistical prediction with special reference to the problem of tolerance limits, Ann. Math. Stat. 13 (4) (1945) 400–409, https://doi.org/10.1214/aoms/1177731537.
[17] Alan D. Hutson, Calculating nonparametric confidence intervals for quantiles using fractional order statistics, J. Appl. Stat. 26 (3) (1999) 343–353, https://doi.org/10.1080/02664769922458.
[18] R. Beran, P. Hall, Interpolated nonparametric prediction intervals and confidence intervals, J. Roy. Stat. Soc. B 55 (3) (1993) 643–652, https://doi.org/10.1111/j.2517-6161.1993.tb01929.x.
[19] E. Zio, F. Di Maio, Bootstrap and order statistics for quantifying thermal-hydraulic code uncertainties in the estimation of safety margins, Sci. Technol. Nucl. Instal. 2008 (2008) 1–9, https://doi.org/10.1155/2008/340164.
[20] F. Sanchez-Saez, A. Sánchez, J. Villanueva, S. Carlos, S. Martorell, Uncertainty analysis of a large break loss of coolant accident in a pressurized water reactor using non-parametric methods 174, *Reliability Engineering & System Safety*, Jun. 2018, pp. 19–28, https://doi.org/10.1016/j.ress.2018.02.005.
[21] S. Ferson, V. Kreinovick, L. Ginzburg, F. Sentz, Constructing Probability Boxes and Dempster-Shafer Structures, *SNL Report*, 2003, p. 809606, https://doi.org/10.2172/809606. SAND2002-4015.
[22] I.S. Hong, A. Connolly, Generalized tolerance limit evaluation method to determine statistically meaningful minimum code simulations, in: Volume 4: Structural Integrity; Next Generation Systems; Safety And Security; Low Level Waste Management And Decommissioning; Near Term Deployment: Plant Designs, Licensing, Construction, Workforce and Public Acceptance, Orlando, Florida, USA, Jan, 2008, pp. 653–660, https://doi.org/10.1115/ICONE16-48448.
[23] J. Hou, et al., BENCHMARK FOR UNCERTAINTY ANALYSIS IN MODELLING (UAM) FOR DESIGN, OPERATION AND SAFETY ANALYSIS OF LWRs, in: Specification and Support Data for the Core Cases (Phase II), 3 Vols, vol. 2, OECD Nuclear Energy Agency, 2019.
[24] G.J. Hahn, W.Q. Meeker, L.A. Escobar, Statistical Intervals: A Guide for Practitioners and Researchers, John Wiley & Sons, 2017.
[25] R. Mendizábal, Contribución al estudio de las metodologías de cálculo realista con incertidumbre (BEPU), dentro del análisis determinista de seguridad de plantas nucleares, " Universidad Politécnica de Madrid, Madrid, 2016.
[26] US NRC, "Regulatory Guide 1.105, Set-points for safety related instrumentation. Rev. 3".
[27] S. Bi, M. Broggi, P. Wei, M. Beer, The Bhattacharyya distance: enriching the P-box in stochastic sensitivity analysis, Mech. Syst. Signal Process. 129 (Aug. 2019) 265–281, https://doi.org/10.1016/j.ymssp.2019.04.035.

[28] L.L. Sharon, in: Sampling: Design and Analysis, second ed., Brooks/Cole, 2009.
[29] D. Grabaskas, R. Denning, T. Aldemir, M. Nakayama, The use of Latin Hypercube sampling for the efficient estimation of confidence intervals, in: *International Congress on Advances In Nuclear Power Plants* 2012, ICAPP *2012* vol. 2, 2012.
[30] C. Baudrit, D. Dubois, Practical representations of incomplete probabilistic knowledge, Comput. Stat. Data Anal. 51 (1) (Nov. 2006) 86–108, https://doi.org/10.1016/j.csda.2006.02.009.
[31] V. Voinov, N. Balakrishnan, M.S. Nikulin, Chi-squared Goodness of Fit Tests with Applications, Elsevier/AP, Amsterdam, 2013.
[32] Anderson–Darling Test, in *the Concise Encyclopedia Of Statistics*, Springer New York, New York, NY, 2008, pp. 12–14.
[33] Kolmogorov–Smirnov Test, in *the Concise Encyclopedia Of Statistics*, Springer New York, New York, NY, 2008, pp. 283–287.
[34] R.F. Engle, Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, in: In *Handbook Of Econometrics*, vol. 2, 1984, pp. 775–826.
[35] Y. Pawitan, A reminder of the fallibility of the Wald statistic: likelihood explanation, Am. Statistician 54 (1) (2000) 54–56, https://doi.org/10.1080/00031305.2000.10474509.
[36] S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. 9 (1) (1938) 60–62, https://doi.org/10.1214/aoms/1177732360.
[37] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, 1936.
[38] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. 1 (1) (Feb. 1986) 54–75, https://doi.org/10.1214/ss/1177013815.
[39] T.G. Bali, The generalized extreme value distribution, Econ. Lett. 79 (3) (Jun. 2003) 423–427, https://doi.org/10.1016/S0165-1765(03)00035-1.
[40] R. Schöbi, B. Sudret, Propagation of uncertainties modelled by parametric P-boxes using sparse polynomial Chaos expansions, in: In *12th Int. Conf. On Applications Of Statistics And Probability In Civil Engineering (ICASP12)*, Canada, Vancouver, 2015, p. 9 [Online]. Available: https://hal.archives-ouvertes.fr/hal-01247151.
[41] C.F. Wu, M. Hamada, Experiments: Planning, Analysis, and Parameter Design Optimization, Wiley, New York, 2000.
[42] M.H. Kutner (Ed.), Applied Linear Statistical Models, fifth ed., McGraw-Hill Irwin, Boston, 2005.
[43] E. Zugazagoitia, C. Queral, K. Fernández-Cosials, J. Gómez, L. Duran-Vinuesa, J. Sanchez-Torrijos, J.M. Posada, Uncertainty and sensitivity analysis of a PWR LOCA sequence using parametric and non-parametric methods, Reliab. Eng. Syst. Saf. 193 (2020), https://doi.org/10.1016/j.ress.2019.106607.