Comparative Analysis of Standard and Advanced USL Methodologies for Nuclear Criticality Safety

Jeongwon Seo*[1], Hany S. Abdel-Khalik[1], Ugur Mertyurek[2],
Goran Arbanas[2], William Marshall[2], William Wieselquist[2]

[1]Purdue University, School of Nuclear Engineering
[2]Oak Ridge National Laboratory

*Mailing address: 205 Gates Rd, West Lafayette, IN, 47906, United States

*Telephone number: +1) 765-337-3766

*E-mail: seo71@purdue.edu

Number of pages: 44

Number of tables: 5

Number of figures: 18

Comparative Analysis of Standard and Advanced USL Methodologies for Nuclear Criticality Safety

Jeongwon Seo*[1], Hany S. Abdel-Khalik[1], Ugur Mertyurek[2], Goran Arbanas[2], William Marshall[2], William Wieselquist[2]

*Corresponding Author: seo71@purdue.edu

[1]Purdue University, School of Nuclear Engineering
[2]Oak Ridge National Laboratory

Abstract:

American National Standards Institute/American Nuclear Society national standards 8.1 and 8.24 provide guidance on the requirements and recommendations for establishing confidence in the results of computerized models used to support operation with fissionable materials. By design the guidance is not prescriptive, leaving freedom to the analysts to determine how the various sources of uncertainties are to be statistically aggregated. Due to the involved use of statistics entangled with heuristic recipes, the resulting safety margins are often difficult to interpret. Also, these technical margins are augmented by additional administrative margins, which are required to ensure compliance with safety standards or regulations, eliminating the incentive to understand their differences. With the new resurgent wave of advanced nuclear systems, e.g., advanced reactors, fuel cycles, and fuel concepts, focused on economizing operation, there is a strong need to develop a clear understanding of uncertainties and their consolidation methods to reduce them in manners that can be scientifically defended. In response, the current studies compare the analyses behind four notable methodologies for upper subcriticality limit estimation that have been documented in the nuclear criticality safety literature: the parametric, non-parametric, Whisper, and TSURFER methodologies. Specifically, the work offers a deep dive into the various assumptions of the noted methodologies, their adequacy, and their limitations to provide guidance on developing confidence for the emergent nuclear systems, expected to be challenged by the scarcity of experimental data. To limit the scope, the current work will focus on the application of these methodologies to criticality safety experiments, where the goal is to calculate a bias, a bias uncertainty, and tolerance limit for k_{eff} in support of determining an upper subcriticality limit for nuclear criticality safety.

Keywords: Similarity Index, Uncertainty Quantification, Upper Subcritical Limit, Model Validation, Criticality Safety.

I. INTRODUCTION

In the context of nuclear criticality safety, American National Standards Institute/American Nuclear Society (ANSI/ANS) national standards 8.1 [1] and 8.24 mandate [2] that analysts establish safety margins that are reasonable and justifiable. In order to ensure reliability of the models employed for a given application, it is essential to validate its simulated results using similar/relevant benchmark experiments. However, model validation is challenged when relevant experiments are lacking, as is often the case with first-of-a-kind nuclear systems. Moreover, the resulting safety margins may be difficult to interpret due to the involved use of statistics. Therefore, this research endeavors to investigate different consolidation methodologies, which are documented in the nuclear criticality safety literature, by means of comparative analysis. This study aims to assess their assumptions, adequacy, and limitations, and thereby provides guidance on developing confidence for the emergent nuclear systems, expected to be challenged by the scarcity of experimental data.

The premise of model validation is to select a finite set of experimental conditions which are considered sufficient to cover application conditions. This criterion involves the use of a metric that measures the *relevance* of an experiment (also referred to as similarity or representativity by different researchers [3], [4]) to the application. For example, a prefect relevance score may be assigned to the application itself if employed as an experiment. Since the relevance is typically measured as an integral quantity valued between -1.0 and 1.0 (or between 0.0 and 1.0 in an absolute sense), the experimental conditions with relevance scores close to 1.0 are preferable. Among many relevance metrics proposed by different researchers, c_k similarity index has been widely used in the neutronics community, since it represents the correlation on the response space merged with the sensitivity/uncertainty (S/U) techniques and perturbation theory [3], [5], [6].

Since no experiment has a perfect relevance score in practice, another criterion must be determined to use for mapping observed discrepancies, referred to hereinafter as the *experimental biases*. From a finite set of experiments to the application conditions, analysts have to incur additional bias to hedge against lower experimental relevance scores. Without a justifiable mapping methodology of the experimental biases to the application conditions, the analyst must assign additional conservative margins, often done in a heuristic manner.

Furthermore, the experimental bias for a given response, e.g., critical eigenvalue, power history, void fraction, isotopic concentrations, etc., is expected to assume a wide range of values due to the various sources of uncertainties in the consolidation process. This situation is depicted in Figure 1, which shows the measured response value y_m, the corresponding calculated value y_c, the unknown true response value y_{true}, and y_{best} the best estimate after fusing code-simulated results with their associated measurements. The deviation between the true and measured value is attributed to experimental uncertainties, and that between the true and predicted value is due to uncertainties in the model.

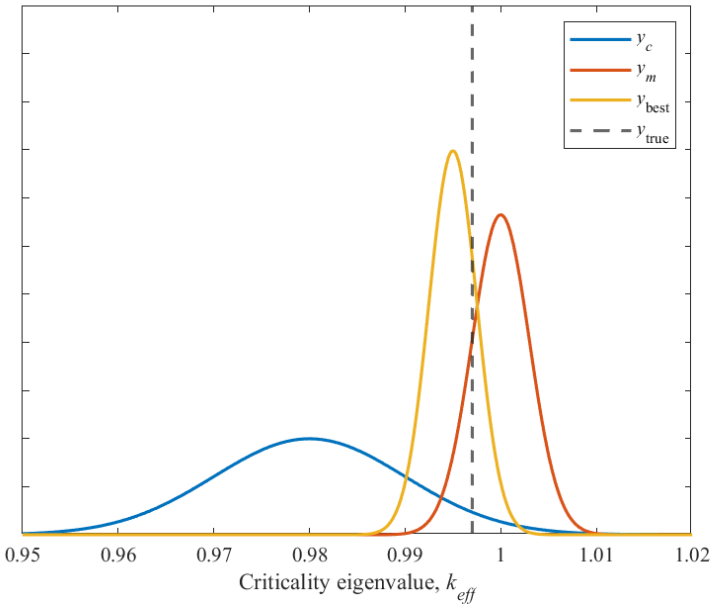


Figure 1. Eigenvalue prediction

Although the uncertainties originating from multiple sources can be classified in various ways, this work focuses on the criterion that classifies these uncertainties into two groups by their reducibility, i.e., *epistemic* (also referred to as reducible or systematic) and *aleatory* (irreducible or random). The distinction between epistemic and aleatory uncertainties can sometimes be less clear and could lead to misleading results. For example, the measurement uncertainty is often treated as aleatory because of the random nature of measurement equipment. Nonetheless, it could be still possible to improve this uncertainty by using more sensitive equipment, cross-validation with other groups of researchers or facilities, recalibration of equipment, etc.

Assuming one has fused all experimental and simulation results and successfully corrected for the epistemic sources of uncertainties, the next step in model validation is to quantify all possible remaining deviations between the measurements and the best-estimate code predictions. These deviations are a result of the aleatory uncertainties as well as a part of the epistemic uncertainties that is not covered by the available experiments. These deviations can be described by a probability density function (PDF) for the variable $dy = y_m - y_{\text{best}}$, representing the errors that could not be reduced by the experimental/analytic consolidation process. Estimating this PDF denotes the core objective of model validation as it is required to properly set safety limits and identify the domain of model validation. For the sake of the effective discussion, this PDF will be denoted hereinafter as the PDF of non-covered deviations, or NCD PDF, where "non-coverage" implies that the deviations are not explained by the experiments.

The mainstream statistical methods assume that the NCD PDF is normal, which reduces the inference problem to the estimation of two statistical features, i.e., the mean and standard deviation. Further, because a normal PDF theoretically stretches indefinitely in both directions, the choice of a bias must be based on the selection of an upper limiting value (or lower, depending on the sign

of the bias)[1], denoted by tolerance limit, that covers a preset portion of the PDF. If a PDF is perfectly known, one can estimate the upper/lower tolerance limit that corresponds to a given coverage level p, say $p = 95\%$. This tolerance limit represents the range within which the true value is expected to lie with a probability of $p\%$. Said differently, there remains a probability of $(1-p)\%$ that the true value falls outside of this coverage. For the eigenvalue response, this tolerance limit serves as the basis for setting an upper subcriticality limit (USL) on all code predictions, often supplemented by an additional administrative margin as shown in Figure 2.

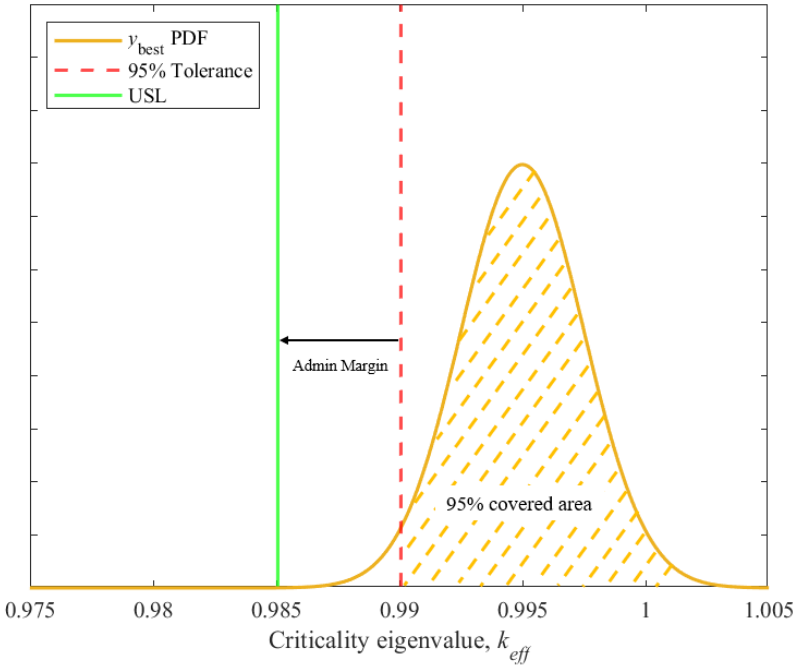


Figure 2. USL non-covered deviations PDF, and tolerance limit

Although outside the scope of this work, it is also important to note that the previous discussion assumed that one knows the NCD PDF perfectly. In practice, the confidence is reported using a double-hedging approach, e.g., 99%/95%, denoting that with 99% confidence the true value would fall into this coverage which is expected to contain 95% of the data population. This double-hedging is required to account for uncertainties in the estimated features, i.e., the standard deviation and the mean, which are calculated based on the limited number of samples from the PDF. Details on this double-hedging approach may be found in an earlier publication [7] and an NRC technical report [8].

This study is structured as follows: First, the basic concepts of validation in criticality safety, including various types of uncertainty sources in experimental biases, the concept of relevance, and USL estimation in Section II. Next, Section III investigates performance of the noted methodologies by conducting a numerical experiment with simplified toy models. Concluding remarks and further research are summarized in Section IV. Appendix I-IV provides a brief

[1] In order to ensure safety, the practice of estimating safety margins typically involves the consideration of only underestimated k_{eff}, with overestimated k_{eff} being excluded. Therefore, the tolerance limit used in this work is limited to a single-sided (or one-sided) confidence interval.

summary of the four noted methodologies, and discusses their assumptions, adequacy, and additional observations made.

II. BACKGROUND AND RELEVANCE

As the bias and bias uncertainty are key elements to determine USL, this section presents a brief background on three relevant topics: a) classification of various sources of uncertainty; b) the concept of relevance employed for benchmark selection, and c) an overview of the USL estimation process. The material in this section may be found in the literature, however, compiled here to help set the stage for the following discussions.

II.A. Uncertainty Classification

This subsection discusses the various sources of uncertainties that control the spread of the noted NCD PDF, including experimental, benchmark, and calculational uncertainties as shown in Figure 3. Regarding the measurement uncertainties assigned number ④, they originate from the unavoidable errors from the measurement process, e.g., those resulting from the aleatory nature of radiation detection instruments. They could also manifest in the form of epistemic errors in the experimental setup due to, for example, equipment misalignment, errors in model specification, poor calibration, etc.

Similarly, benchmark uncertainties, assigned ③ in Figure 3, contain both aleatory and epistemic sources of errors. For example, aleatory errors originate from model parameters that specify geometry and composition resulting from the manufacturing process. Further, if the calculational model employed is probabilistic, e.g., Monte Carlo-based, then the predicted value is expected to have another random error component. The benchmark uncertainties may be lumped with the measurement uncertainties for a number of reasons: a) they cannot be controlled[2], due to their aleatory nature, similar to other experimental conditions, e.g., ambient conditions; b) since the benchmark models are carefully designed, these uncertainties are much smaller than other sources of calculational uncertainties such as nuclear cross-sections uncertainties; and c) they are independent of other experimental uncertainties. In the remainder of the text, we will denote both benchmark and experimental uncertainties as evaluation uncertainties as noted in Figure 3.

The calculational uncertainties, resulting from modeling assumptions, numerical approximations, and input model parameter uncertainties, all can be treated as epistemic. The current study focuses on epistemic parameter uncertainties assigned number ②for two reasons. First, recent advances in high fidelity simulation have provided a clear venue for reducing the first two sources, allowing analysts to set a fixed upper limit on their contributions, akin to an administrative margin. For the sake of the effective discussion, these two sources are lumped together as solution uncertainties and are assigned number ①. Second, model parameters, i.e., nuclear cross-sections, continue to be the major source of uncertainty in neutronic calculations, representing the primary driver in criticality safety calculations [9]–[12].

[2] Although Monte-Carlo uncertainty technically can be controlled by using more cycles and particles, the use of a high-fidelity model in this study has resulted in its uncertainty of approximately 10 pcm. This level of uncertainty is regarded as sufficiently accurate, and any further improvement in accuracy may entail a considerable computational burden.

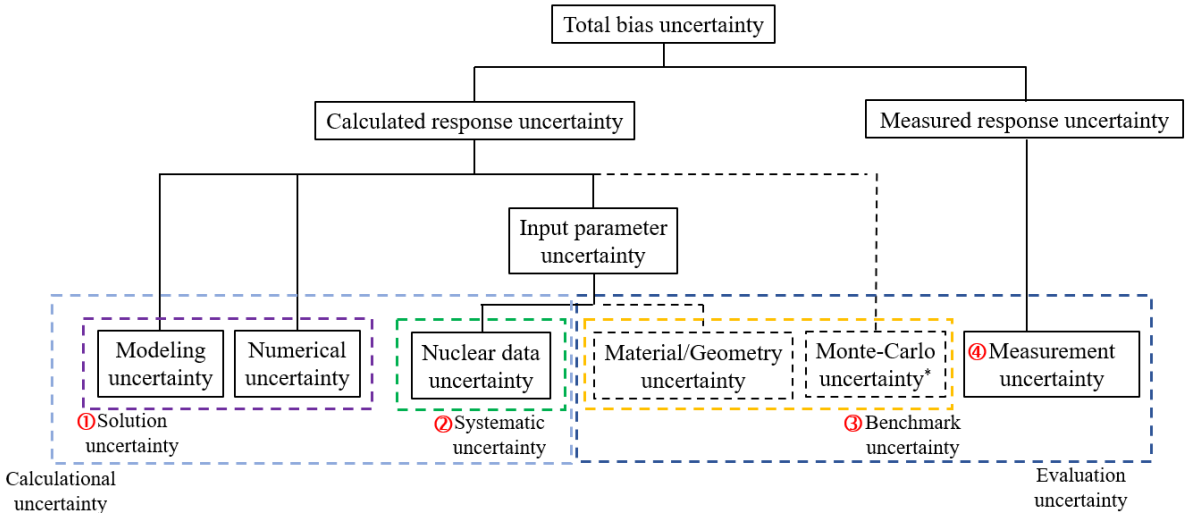


Figure 3. Uncertainty sources and classification

II.B. Experimental Relevance

Experimental relevance is a key requirement to reduce uncertainties. As pointed out earlier, the regulator allows the licensee to seek an inference technique to reduce the impact of epistemic uncertainties on the calculated biases. If such inference is not completed, one would be forced to propagate the cross-sections uncertainties, often resulting in a widely spread PDF for the calculated response, i.e., with high standard deviation. Reducing cross-sections uncertainties is however a challenging endeavor, since the number of cross-sections is much larger than the number of available experiments, i.e., under-determined problem.

To explain this, first note that the systematic bias resulting from the cross-sections uncertainties is not a universal constant value; instead, it changes based on the sensitivities of the response with respect to the cross-sections. Therefore, the estimated sensitivities are expected to be different for each experiment as well as the application [13]. As will be shown later, the bias is simply the inner product between the gradient vector and the cross-section error vector. For illustration purposes, we assume the norm of the gradient is unity, making the projection equal to the bias.

From calculus, the gradient points in the direction of maximum change and its magnitude measures the rate of change along that direction; it is an n dimensional vector whose components are the first-order derivatives of a given response with respect to n cross-sections. Since the cross-section true error vector is unknown, one can only assess the impact on the response by analyzing all possible cross-section variations within their prior uncertainties.

Simply increasing the dimension of the experimental domain to extend this idea would still fail to infer the correct bias needed for the application. This is because each experiment's bias is determined by the inner product of its own gradient with the cross-section error vector. Taking the average of these biases does not determine application bias, because the experimental gradients are essentially blind to the application gradient. Hence, it is important to select experiments whose

experimental gradients pointing approximately in the same direction as the application gradient. Note that the error vector cannot be explicitly determined due to the under-determined nature of the inference problem. Mathematically, the observed deviation between each experiment and its associated calculated value is approximately given by:

$$y_{m_i} - y_{c_i} = \Delta x^T \nabla y_i^{\text{exp}} \tag{1}$$

and the sought bias for the application is given by:

$$\Delta x^T \nabla y^{\text{app}} \tag{2}$$

These equations imply that the ratio of any experiment's bias and the application bias is approximately equal to the ratio of the norms of the experiment and application gradient vectors. This relationship is exact (under the linearity assumption) if the unknown components of the cross-section error vectors along both gradients are the same, which is possible if the two gradients are pointing in the same direction.

Another important consequence of the above equation is that each experiment allows the analyst to estimate the component of the cross-section error vector along the gradient of that experiment. When the number of experiments is equal to or higher than the number of cross-sections, one may be able to correct for the entire cross-section error vector without having to know the application gradient. This is possible when the gradients from all the experiments provide coverage for the entire cross-section space, i.e., they have n independent components along the n dimensions of the cross-section space.

The concept of experimental similarity based on the S/U techniques has been widely adopted in the neutronic community because the responses vary nearly linearly within the range of cross-sections uncertainties. Mathematically, the c_k similarity is described as follows:

$$c_k = \frac{\nabla y^{\text{exp}T} \boldsymbol{C_\alpha} \nabla y^{\text{app}}}{\sqrt{\nabla y^{\text{exp}T} \boldsymbol{C_\alpha} \nabla y^{\text{exp}}} \sqrt{\nabla y^{\text{app}T} \boldsymbol{C_\alpha} \nabla y^{\text{app}}}} \tag{3}$$

where $\boldsymbol{C_\alpha}$ is the prior covariance matrix**,** acting as a weighting structure.

This index is widely used to determine if a critical experiment is similar to an application. According to Oak Ridge National Laboratory (ORNL) criticality safety validation experience and a previous work experience with the SCALE S/U tools, the generally accepted criterion states that an experiment can sufficiently represent an application model with the c_k value larger than 0.9, while critical experiments with c_k values of between 0.8 and 0.9 can be considered only marginally similar, and use of experiments with c_k values less than 0.8 are discouraged [14].

Note that the similarity expression in Eq. (3) is standardized, meaning that two experiments with the same relevance could have different response deviations because their gradients have different norms, i.e., magnitude. Thus, one must account for that when combining experimental biases to calculate the application bias, instead of simply averaging the experimental biases from equally relevant experiments. The simple averaging will be adequate only if the experiments have the same exact gradient norms which is not accounted for by three of the methodologies studied in this work,

i.e., the parametric, non-parametric, and Whisper methodologies. Lastly, as mentioned earlier, it is infeasible to pick an experiment that has a perfect relevance score, hence the NCD PDF, describing the deviations between measured and best-estimate predictions, must be inflated to account for non-perfect relevance.

With no experimental coverage for the orthogonal component, one must rely on basic uncertainty propagation to estimate the impact of cross-sections uncertainties on the responses of interest. This implies that while the experiments can reduce the cross-sections epistemic uncertainties along the covered subspace, they fail to provide any inference on the components belonging to the non-covered subspace. To further reduce response uncertainties, additional experiments must be sensitive to new directions along the non-covered subspace. For realistic inference problems, the non-covered subspace is much higher in dimensionality than the covered subspace due to the infeasibility of conducting many experiments. The implication is that the NCD PDF should not be directly employed to calculate the application bias; instead, it must be inflated to account for the prior parameter uncertainties belonging to the non-covered cross-section subspace. This is another important observation that we will recall in future discussions of the various methodologies used for bias mapping from the experimental to application conditions.

II.C. USL Calculation

The previous discussion has helped set the stage for introducing the four different methodologies surveyed in this study for determining a code's USL. These methodologies are referred to as the parametric [15], non-parametric [15], Whisper [16], and TSURFER [17] methodologies. Before reviewing these methodologies, we recall the definitions of calculational margin (CM), and margin of subcriticality (MOS) from the American National Standard ANSI/ANS-8.24-2017 [2]. The CM is defined as an allowance for the bias and bias uncertainty plus considerations of uncertainties related to interpolation, extrapolation, and trending of the bias. The MOS is an allowance beyond CM to ensure subcriticality.

A closely related term is often used in the criticality literature is the lower tolerance limit (LTL). The LTL is closely related to the concept of CM; it is defined in terms of the bias PDF, assumed to be negative, implying that the code calculations under-predict the true value of k_{eff}. The estimated bias is thus a negative number with a spread that describes the uncertainty in its estimated value. The LTL is defined as a single-sided lower limit for the bias PDF. As will be shown later in the discussion, most methodologies define the CM in the same way, and hence the two terms are essentially the same for most methodologies.

Note that the definitions of bias and bias uncertainties are more prescriptive than the CM and MOS. The bias is clearly defined as the systematic deviation resulting from epistemic uncertainty sources, such as cross-section errors, systematic measurement errors, numerical and modeling errors. And the bias uncertainty results from the aleatory nature of the measurement, the benchmark model parameters, e.g., geometry and composition, the probabilistic nature of the calculations, if any, as well as the non-covered epistemic uncertainties resulting from cross-sections. Throughout this work, the solutional uncertainties, i.e., modeling and numerical, uncertainties -- assigned number ① in Figure 3 -- will be treated separately via the MOS term.

The bias and bias uncertainties may be viewed as two fundamental quantities, based on which the CM and MOS can be calculated, i.e., the CM and MOS are functions of the bias and bias uncertainties. Recalling from the previous discussion that a key component of model validation is to set a tolerance limit that covers a certain preset portion of the NCD PDF. Before discussing how this is done, it is important to note that the goal here is to rely on using samples of experimental biases to identify the NCD PDF. This is a well-known problem in statistics called the inference problem. The other more commonly known problem is the sampling problem, where one knows the PDF and is interested in generating samples from it. The sampling and inference problems are the equivalent of the forward and inverse problems in applied mathematics, details on the mechanics of both problems may be found in an earlier publication [18].

Generally, the inference problem may be solved in two notably different approaches, the so-called parametric and non-parametric approaches. The parametric approach as the name suggests relies on knowing the type of the PDF, allowing one to parametrize the tolerance limit in terms of the PDF's features, e.g., the mean value and standard deviation for a normally distributed PDF. With the features determined, the tolerance limit can be seamlessly calculated with an allowance made for uncertainties in the estimated features. This represents the basic idea of the parametric approach as well as the TSURFER methodology [19].

In the non-parametric approach, the tolerance limit is related directly to the samples by first employing a sampling approach to construct another related PDF, called the extreme value (EV) PDF of k^{th} order. The EV PDF has the majority of its mass concentrated at the tail end -- hence the name extreme -- of the original PDF, implying that a majority of its samples would be higher than sought tolerance limit for the original PDF. This is always possible by increasing the order of the EV PDF as will be discussed later. This is the basic idea of the non-parametric approach.

Before diving into the details, Table 1 lists the sources of uncertainties, see Figure 3, captured by the CM and MOS for each of the methodologies. This table implies that all methodologies employ CM to capture the epistemic cross-sections uncertainties ②, the benchmark uncertainties ③, as well as the measurements uncertainties ④; and the MOS captures the solution uncertainties ①. The Whisper methodology, however employs additional margins for the first three sources under the MOS. Details on how this is performed will be given in later sections.

Table 1. Uncertainties employed for USL calculation

	USL calculation	
	CM calculation	MOS calculation
Parametric/ Non-parametric	(② +)* ③ +④	①
Whisper	(② +)* ③ +④	① + ② + ③ + ④
TSURFER	② + ③ +④	①

* Implicit effect of this uncertainty

III. NUMERICAL EXPERIMENTS

To help compare the various methodologies, USL is estimated focusing two cases: 1) a toy model where the true bias values are known, to reveal the mechanics of the various methodologies and to test the adequacy of their assumptions, and 2) nuclear criticality benchmark models to validate the performance of the four methodologies for actual nuclear engineering applications. The detailed USL estimation process and the observations based on the key assumptions behind the noted four methodologies are discussed through Appendix I-IV.

III.A. USL Calculations with a Toy Model

The toy problem includes two correlated input variables, representing the cross-sections, and one aleatory term aggregating the composition, geometry, measurement uncertainties, possible Monte Carlo calculational uncertainties (since these sources are independent, it is not of primary importance to separate them into different terms for the sake of this study). All the reference values and the range of variations for the epistemic and aleatory parameters are selected to be similar in magnitude to the uncertainties encountered in typical neutronic criticality problems. The resulting response errors and variations are manufactured to be in the ballpark of reported eigenvalue uncertainties. The benchmark model is given by:

$$k_m = a^T x + k_c + \epsilon$$

or in matrix form for 40 different experiments,

$$\begin{bmatrix} k_{m1} \\ k_{m2} \\ \vdots \\ k_{m40} \end{bmatrix} = \begin{bmatrix} a_1^{(1)} x^{(1)} + a_1^{(2)} x^{(2)} + k_{c1} + \epsilon_1 \\ a_2^{(1)} x^{(1)} + a_2^{(2)} x^{(2)} + k_{c2} + \epsilon_2 \\ \vdots \\ a_{40}^{(1)} x^{(1)} + a_{40}^{(2)} x^{(2)} + k_{c40} + \epsilon_{40} \end{bmatrix}$$

where k_{mi} and k_{ci} represents the measured and the reference calculated responses of the i^{th} model, $a_i^{(j)}$ is the j^{th} coefficient, and ϵ_i is an aleatory error term which cannot be explained by the input parameters $x^{(1)}$ and $x^{(2)}$ representing the cross-sections in this toy model. Note that the reference values for the input parameters are assumed to be zero, and their uncertainties are assumed to follow a normal distribution with zero mean and 1% standard deviation. An example correlation coefficient of 0.4 is selected for the parameters, for which the corresponding correlation matrix $R \in \mathbb{R}^{2\times 2}$ can be written as:

$$R = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix}$$

where the off-diagonal terms imply a positive correlation. The sensitivity coefficients for each model, $a_i^{(j)}$s, are selected such that the input parameters uncertainties lead to 1-2% change in the responses. Each row of coefficients emulates the concept of a sensitivity profile, i.e., the gradient of the response with respect to the input parameters.

The measured responses are assumed to follow a normal distribution with a unity mean and standard deviation of 150 pcm. To help evaluate the performance of the various methodologies, a

virtual approach is devised wherein the true parameter values are used to generate the mean value of the measurements. Specifically, the true values for the parameters $x_{\text{true}} = \begin{bmatrix} x_{\text{true}}^{(1)} & x_{\text{true}}^{(2)} \end{bmatrix} = [0.0084 \quad -0.0021]$ are selected from the pool of random samples shown in Figure 4, and the reference values of the experiments, k_{c_i}, are back-calculated as,

$$k_{c_i} = 1.0 - a_i^{(1)} x_{\text{true}}^{(1)} + a_i^{(2)} x_{\text{true}}^{(2)}$$

The application response's calculated value is modeled as:

$$k_c^{\text{app}} = 0.9856 + 1.6151 x^{(1)} - 0.4038 x^{(2)}$$

This results in an estimated value of 1.0 using the true values of the input parameters, and produces a response uncertainty of 1500 pcm. This model yields a true bias of 1440 pcm meaning that the model underestimates the true value of k_{eff} by approximately one standard deviation of the prior uncertainty.

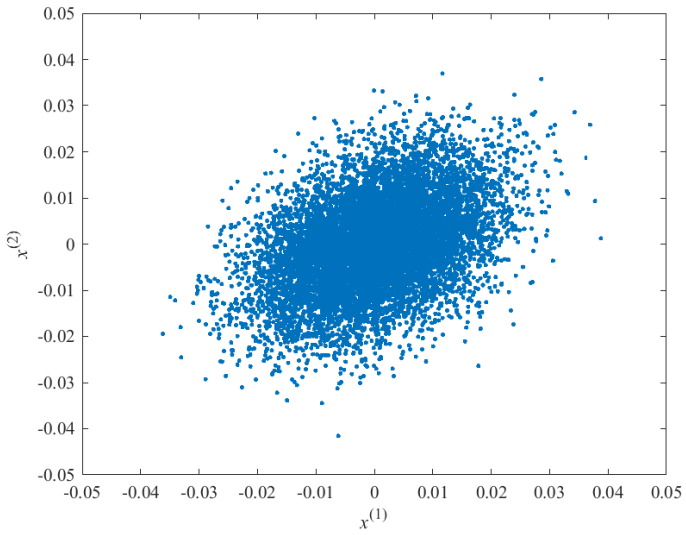


Figure 4. Toy model's parameters prior uncertainty

The error term, ϵ, representing the benchmark uncertainties, is randomly sampled to have standard deviation of 200 pcm, leading to an evaluation uncertainty σ_{ei} of:

$$\sigma_{ei} = \sqrt{\sigma_{\epsilon_i}^2 + \sigma_{m_i}^2} = 250 \text{ pcm}$$

Recall that the evaluation uncertainty aggregates both the benchmark uncertainty and the measurement uncertainty.

With two correlated parameters, the prior epistemic uncertainty, σ_{s_i}, is calculated as (this is corresponding to the propagated cross-sections uncertainty),

$$\sigma_{s_i} = \sqrt{\left(a_i^{(1)} \sigma_{x^{(1)}}\right)^2 + \left(a_i^{(2)} \sigma_{x^{(2)}}\right)^2 + 2 a_i^{(1)} a_i^{(2)} \text{Cov}\left(\sigma_{x^{(1)}}, \sigma_{x^{(2)}}\right)}$$

The representative calculated responses with their associated prior epistemic uncertainties, the measured responses with their uncertainties, and the biases of each benchmark with the evaluation uncertainties are graphically illustrated in Figure 5.

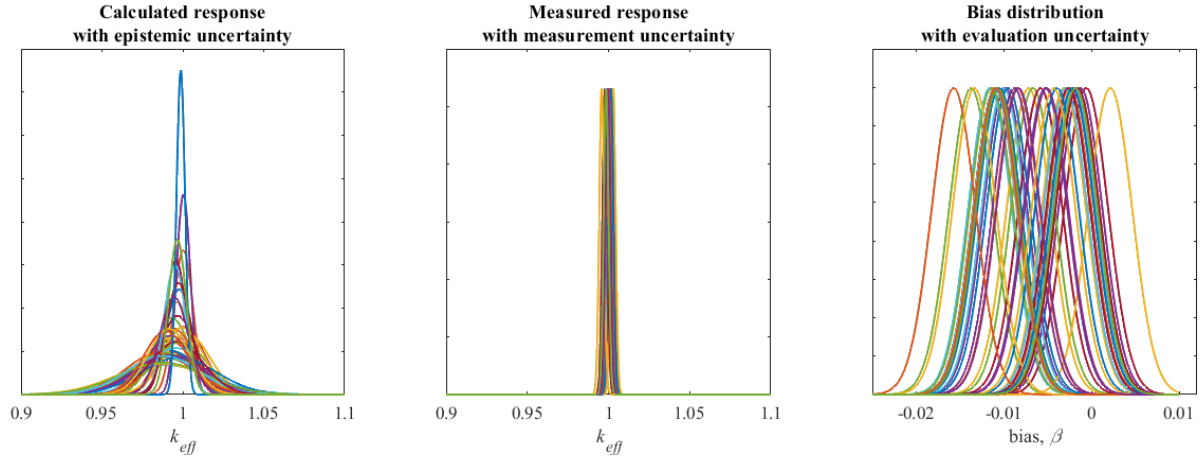


Figure 5. Calculated response, measured responses, and bias distributions

The USLs calculated by each methodology with 95% confidence are listed in Table 2. Note that the non-parametric margin m_{np} is assumed to be zero since it is an adhoc parameter that cannot be statistically justified, resulting in an additional conservative for the calculated bias. Table 2 indicates that the Whisper methodology evaluates USL more conservatively than the other methodologies, while the TSURFER methodology provides the highest USL. The detailed CM and Whisper MOS calculations for 95% confidence, i.e., with coverage parameter $\varrho \cong 1.65$ for a normal distribution, are as follows:

Table 2. Toy model USL results for 95% confidence

	CM	MOS	USL($= 1.0 - \text{CM} - \text{MOS}$)
Parametric	1516 pcm	500 pcm	0.9798
Non-parametric	2409 pcm	500 pcm	0.9709
Whisper	2023 pcm	678 pcm	0.9730
TSURFER	1618 pcm	500 pcm	0.9788

Since the true application response is known for the toy model, we can quantify how far the USLs of the different methodologies are from the true application response. Focusing on the CM only since it is calculated based on the bias, we compare its value for the various methodologies without the MOS, since the choice of the latter is more arbitrary as it includes the effects of unknown modeling uncertainties. Figure 6 shows the results in the form of a PDF for the calculated bias. The blue wide-spread PDF denotes the prior knowledge about the application bias. The red PDF denotes the best-estimate knowledge after fusing the experimental and calculated values. This PDF is the one calculated by TSURFER and represents the true posteriori PDF according to Bayes theorem. The goal here is to estimate LTL for this PDF such that 95% of the values are above the LTL. The implication is that one could assert with 95% confidence that the true value of the bias will not be less than the LTL value. Based on this LTL value, the USL is calculated. For this toy

model, the true application bias is given by 1440 pcm which is the same as the mean value of the TSURFER posteriori PDF. The spread of the posteriori PDF is due to the aleatory uncertainty from the benchmark model and the measurements as explained earlier. Based on this PDF, TSURFER calculated an LTL at 1618 pcm which covers 95% of the PDF, as follows

$$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'} = 1440 + 1.65 \times 108 = 1618 \text{ pcm}$$

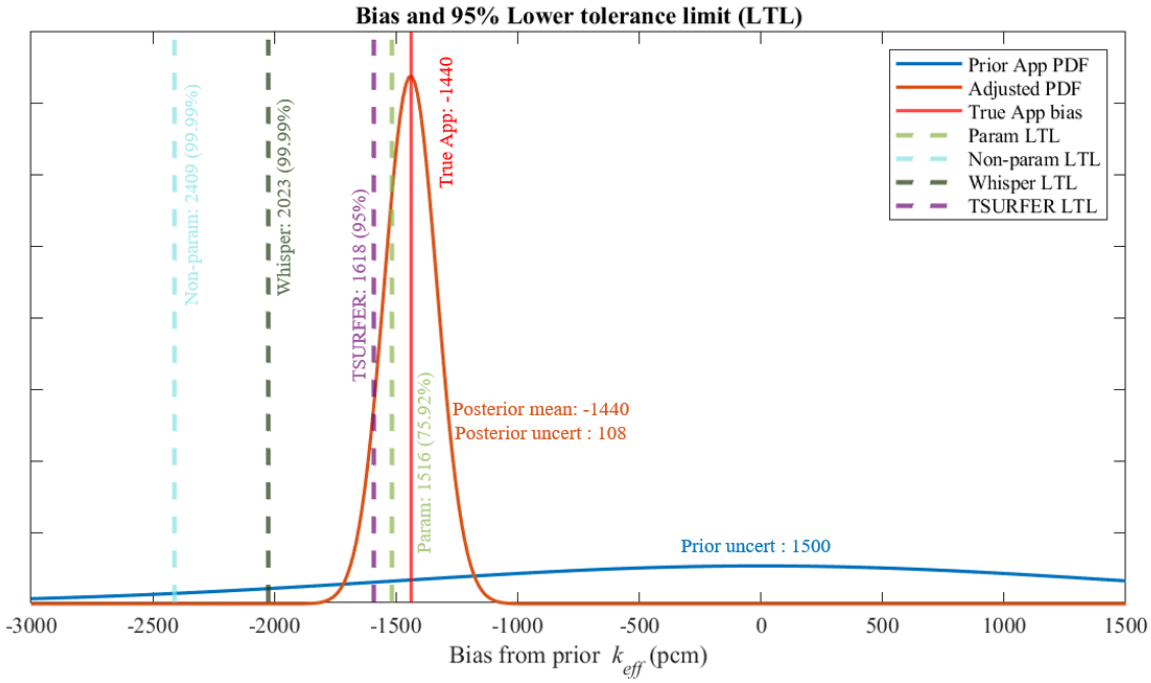


Figure 6. Toy model application PDF and estimated LTLs

For the parametric methodology, the inverse-variance weighted average β_p is -636 pcm, setting the non-conservative parameter to be zero. The pooled variance σ_p consists of two parts, one accounting for the weighted standard deviation and the other for the spread of the calculated responses. The evaluation uncertainty for all the models is selected to be 250 pcm, yielding the same value for the weighted standard deviation. By adding the impact of the response spread, the pooled variance increases to 535 pcm which is approximately two times larger than the evaluation uncertainty. Therefore, the final parametric CM is calculated such as

$$\mathrm{CM}_p = -\beta_p + \varrho\sigma_p + \Delta_m = 636 + 1.65 \times 535 + 0 = 1516 \text{ pcm}$$

which is slightly lower than the true value for the 95% LTL, giving a coverage of 76%.

For the non-parametric methodology, only the minimum negative bias of -1529 is employed (recall the non-parametric margin is assumed to be zero for this analysis since it is heuristically determined, also it will result in even more conservative CM value), yielding CM value of:

$$\mathrm{CM}_{np} = -\min\{k_{c_i} - k_{m_i}\} + \varrho\sigma_p + m_{np} + \Delta_m = 1529 + 1.65 \times 535 + 0 + 0 = 2409 \text{ pcm}$$

Recall that Whisper builds an EV-like PDF by generating samples from the various bias PDFs, shown in Figure 5, and taking their maximum. In the toy problem, this EV-like PDF will be heavily

influenced by the two most negatively biased PDFs in the right graph of Figure 5. This follows because most of the samples generated from the other PDFs will be less than the samples generated from the two most biased PDFs. Since these two PDFs are heavily overlapped, their samples may be approximately considered iid, i.e., they are effectively being sampled from the same PDF, and hence their maximum will be equivalent to the generation of a 2nd order EV PDF. For a normal distribution, the 95% confidence interval for 95% coverage using a 2nd order EV PDF is given by 1.95. Thus, for this toy model, the EV multiplier ν is approximately given by 1.95. The Whisper CM is approximated by the non-parametric bias and the evaluation uncertainty with the EV multiplication factor such that

$$\mathrm{CM}_w = m + \Delta_m \approx -\min\left\{k_{c_i} - k_{m_i}\right\} + \nu\sigma_e + \Delta_m = 1529 + 1.95 \times 250 + 0 = 2017 \text{ pcm}$$

which approximates the actual value calculated by Whisper. We note that both the non-parametric and Whisper CM values produce LTL values that provides nearly 100% coverage of the posteriori PDF, which is much higher than the 95% coverage reported by the two methodologies.

In the above example, although the analysis includes 40 experiments, only two experiments that correspond to the two left-most biased PDFs in the right plot of Figure 5 have influenced the final CM value, resulting in a multiplier of $\nu = 1.95$. This begs the question of how the multiplier value would change with an increasing number of overlapping experiments, a situation that is expected when the analyst employs a large database of experiments. Therefore, the Whisper CM is re-evaluated assuming that k experiments have overlapped, and this is repeated with increasing the value of k; the results are shown in Figure 7.

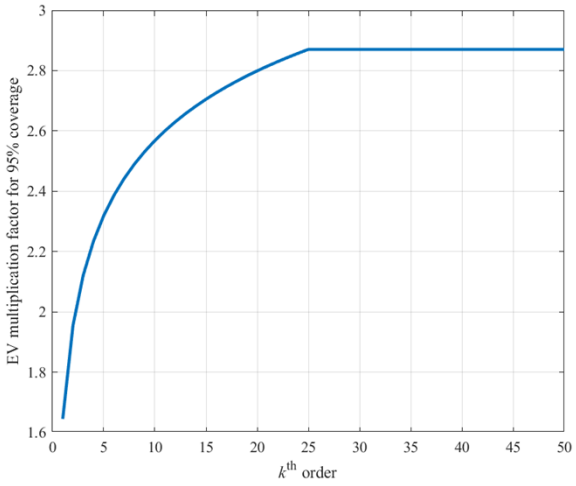


Figure 7. k^{th} order EV multiplier value

This trend shows that as the number of overlapping bias PDFs increases, the multiplier value will also monotonically increase. This behavior is undesirable because it implies one would have less confidence as the number of experiments with similar bias results are included in the analysis. According to basic statistical inference techniques, e.g., Bayesian inference, the confidence should increase when similar measurements are assimilated. To limit this increase, Whisper employs a heuristic weighting procedure, as denoted by Eqs. (15)-(17), which establishes an upper limit on the multiplier value. Specifically, at exactly 25 experiments, the multiplier is affixed to a value of

2.87 which corresponds to 99.79% one-sided tolerance limit of a standard normal distribution. This implies that the reported confidence of 95% would be lower than the actual confidence when the experiments biases are very similar, a situation that is very common when including large number of highly relevant experiments.

The previous discussion was motivated by our observation that the most biased PDFs are the ones controlling the Whisper CM value. To further validate this observation, we repeat the calculation of the CM for three different cases. In the first case, all the bias PDFs are assumed to have the same weights. The second case zeros all the weights except for the two most biased PDFs and the third uses the standard c_k-based weights as employed by Whisper.

Before concluding this section, we recall that the MOS calculations were not explicitly mentioned in the toy problem because three of the methodologies employ a fixed value as an additional margin that hedges against unknown modeling uncertainties, and only Whisper provides a procedure for estimating the additional MOS which is given by:

$$\mathrm{MOS}_d = \varrho\sigma_{k'} = 1.65 \times 108 = 178 \text{ pcm}$$

where the 108 represents the spread of the posteriori TSURFER PDF.

III.B. USL Calculations with Pu-Solution Benchmarks

This section studies various USL calculation methodologies using a suite of 29 Pu-solution benchmarks containing 15 g/L. While selecting their application, two key factors are considered: 1) whether an application has distinct features, e.g., fuel type, geometry, material compositions including fuel enrichment, etc., from those of the experiments, 2) whether the c_k values estimated with this application are sufficiently high, e.g., greater than 0.85. Given these criteria, the MIX-SOL-THERM-002-001 benchmark with calculated k_{eff} of 1.0015 is selected as the application model, which is distinguishable from the experiments in fuel type and maintains high c_k values ranging from 0.85 to 0.92.

The benchmark uncertainties were estimated via a Monte Carlo approach, by sampling the composition and geometry parameters within a small margin of uncertainty, 0.5% - 1.0%. The resulting eigenvalue uncertainties were in the range of 160-250 pcm, which varied according to the experiment and the assumed composition and geometry uncertainties. To simplify the treatment, a fixed value of 200 pcm is assumed to represent the evaluation uncertainties for all experiments, including the Monte Carlo uncertainties. The detailed information about this set of the benchmarks including c_k, Whisper weights, and the measurement uncertainties can be found in Table 3.

Table 3. Employed benchmarks specification

Benchmark name	Measured, k_m		Calculated, k_c			Bias ($k_c - k_m$)		Weight	
	k_{eff}	σ_m	k_{eff}	σ_ϵ	σ_s	β	σ_e	c_k	w
PU-SOL-THERM-003-001	1.0000	0.0047	1.0014	0.0020	0.0087	0.0014	0.0051	0.8802	0.9566
PU-SOL-THERM-003-002	1.0000	0.0047	1.0009	0.0020	0.0087	0.0009	0.0051	0.8761	0.9522
PU-SOL-THERM-003-003	1.0000	0.0047	1.0042	0.0020	0.0087	0.0042	0.0051	0.8688	0.9442
PU-SOL-THERM-003-004	1.0000	0.0047	1.0034	0.0020	0.0087	0.0034	0.0051	0.8668	0.9421

PU-SOL-THERM-003-007	1.0000	0.0047	1.0057	0.0020	0.0087	0.0057	0.0051	0.8787	0.9550
PU-SOL-THERM-003-008	1.0000	0.0047	1.0043	0.0020	0.0087	0.0043	0.0051	0.8757	0.9517
PU-SOL-THERM-004-001	1.0000	0.0047	1.0025	0.0020	0.0087	0.0025	0.0051	0.9084	0.9873
PU-SOL-THERM-004-002	1.0000	0.0047	0.9975	0.0020	0.0087	-0.0025	0.0051	0.9074	0.9862
PU-SOL-THERM-004-003	1.0000	0.0047	0.9998	0.0020	0.0087	-0.0002	0.0051	0.9020	0.9803
PU-SOL-THERM-004-004	1.0000	0.0047	0.9980	0.0020	0.0087	-0.0020	0.0051	0.8963	0.9741
PU-SOL-THERM-004-005	1.0000	0.0047	0.9980	0.0020	0.0087	-0.0020	0.0051	0.9040	0.9825
PU-SOL-THERM-004-006	1.0000	0.0047	1.0003	0.0020	0.0087	0.0003	0.0051	0.9038	0.9823
PU-SOL-THERM-004-007	1.0000	0.0047	1.0050	0.0020	0.0087	0.0050	0.0051	0.9000	0.9782
PU-SOL-THERM-004-008	1.0000	0.0047	1.0000	0.0020	0.0086	0.0000	0.0051	0.8972	0.9751
PU-SOL-THERM-004-009	1.0000	0.0047	0.9996	0.0020	0.0086	-0.0004	0.0051	0.8895	0.9667
PU-SOL-THERM-004-010	1.0000	0.0047	1.0013	0.0020	0.0086	0.0013	0.0051	0.8707	0.9463
PU-SOL-THERM-004-012	1.0000	0.0047	1.0022	0.0020	0.0086	0.0022	0.0051	0.9006	0.9788
PU-SOL-THERM-004-013	1.0000	0.0047	0.9991	0.0020	0.0086	-0.0009	0.0051	0.9008	0.9790
PU-SOL-THERM-005-001	1.0000	0.0047	1.0012	0.0020	0.0087	0.0012	0.0051	0.8990	0.9771
PU-SOL-THERM-005-002	1.0000	0.0047	1.0018	0.0020	0.0086	0.0018	0.0051	0.8953	0.9730
PU-SOL-THERM-005-003	1.0000	0.0047	1.0024	0.0020	0.0086	0.0024	0.0051	0.8915	0.9689
PU-SOL-THERM-005-004	1.0000	0.0047	1.0040	0.0020	0.0086	0.0040	0.0051	0.8822	0.9588
PU-SOL-THERM-005-005	1.0000	0.0047	1.0053	0.0020	0.0086	0.0053	0.0051	0.8717	0.9474
PU-SOL-THERM-005-006	1.0000	0.0047	1.0048	0.0020	0.0086	0.0048	0.0051	0.8607	0.9354
PU-SOL-THERM-005-008	1.0000	0.0047	0.9981	0.0020	0.0086	-0.0019	0.0051	0.8955	0.9733
PU-SOL-THERM-005-009	1.0000	0.0047	1.0011	0.0020	0.0086	0.0011	0.0051	0.8901	0.9674
PU-SOL-THERM-006-001	1.0000	0.0035	0.9995	0.0020	0.0086	-0.0005	0.0040	0.9201	1.0000
PU-SOL-THERM-006-002	1.0000	0.0035	1.0008	0.0020	0.0086	0.0008	0.0040	0.9161	0.9957
PU-SOL-THERM-006-003	1.0000	0.0035	1.0004	0.0020	0.0086	0.0004	0.0040	0.9077	0.9865

Like the toy problem, Figure 8 plots the calculated responses with their epistemic cross-sections uncertainties and the measured responses with their evaluation uncertainties. The maximum and minimum cross-sections uncertainties are 873 and 860 pcm, respectively; and the evaluation uncertainty is 510 pcm for all the benchmarks except for the last three benchmarks whose evaluation uncertainty is 400 pcm due to lower reported measurement uncertainties.

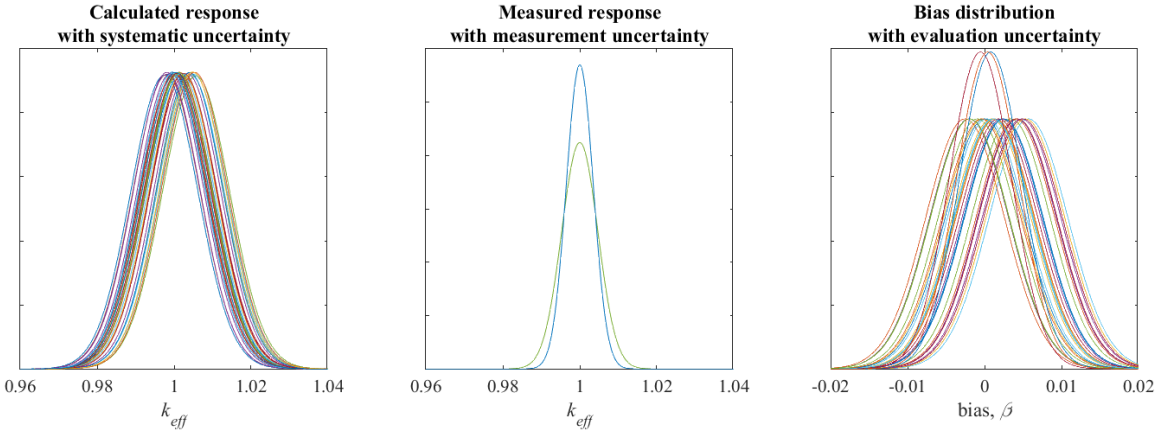


Figure 8. Calculated response, measured responses, and bias distributions

The USLs for the four methodologies are calculated and provided in Table 4. With the same argument in Section III.A, the non-parametric margin is set to be zero.

Table 4. Pu-Solution USL results for 95% confidence

	CM	MOS	USL($= 1.0 - \mathrm{CM} - \mathrm{MOS}$)
Parametric	900 pcm	500 pcm	0.9860
Non-parametric	1153 pcm	500 pcm	0.9835
Whisper	1448 pcm	1123 pcm	0.9743
TSURFER	785 pcm	500 pcm	0.9871

In this case study, the parametric methodology's inverse-variance weighted bias is positive, so the non-conservative bias adjustment Δ_m is selected to cancel it out, as discussed earlier. The parametric CM is calculated as

$$\mathrm{CM}_p = -\beta_p + \varrho\sigma_p + \Delta_m = -139 + 1.65 \times 547 + 139 = 900 \text{ pcm}$$

And the non-parametric methodology CM is given by:

$$\mathrm{CM}_{np} = -\min\{k_{c_i} - k_{m_i}\} + \varrho\sigma_p + m_{np} + \Delta_m = 253 + 1.65 \times 977 + 0 + 0 = 1153 \text{ pcm}$$

And the MOS for the Whisper is calculated as

$$\mathrm{MOS}_d = \varrho\sigma_{k'} = 1.65 \times 373 = 615 \text{ pcm}$$

Lastly, the TSURFER CM is calculated as

$$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'} + \Delta_m = -78 + 1.65 \times 373 + 78 = 615 \text{ pcm}$$

Unlike the toy model study, the true application response remains unknown. Nevertheless, if one solely relies on the prior uncertainties as shown in Figure 9, the Whisper-determined USL is equivalent to a 99.9% confidence because 99.9% of the area under the prior PDF is above the reported USL limit.

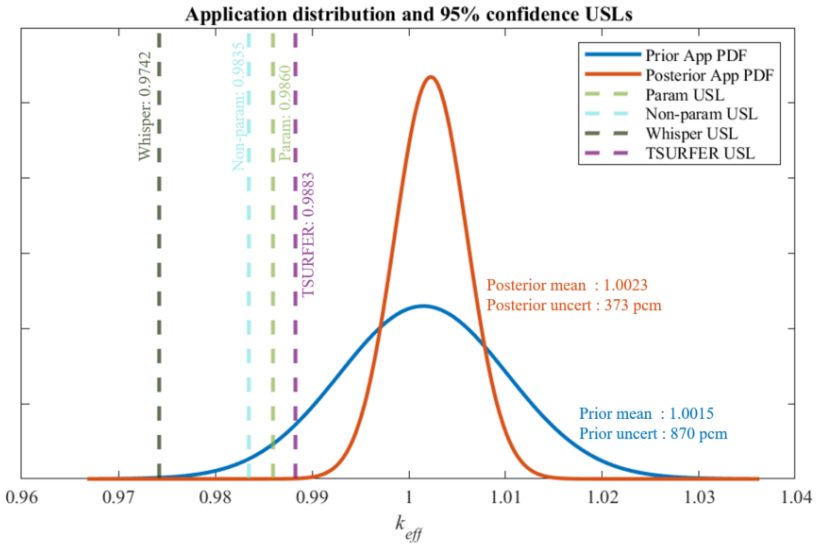
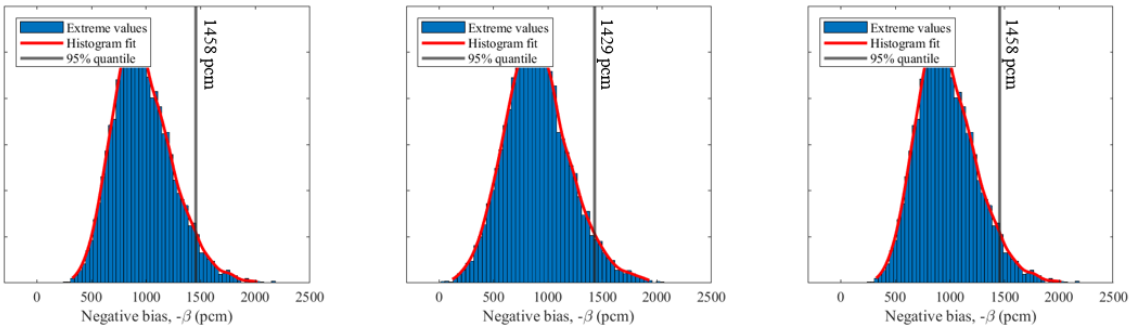


Figure 9. Mix-Sol-Therm benchmark application and estimated USLs

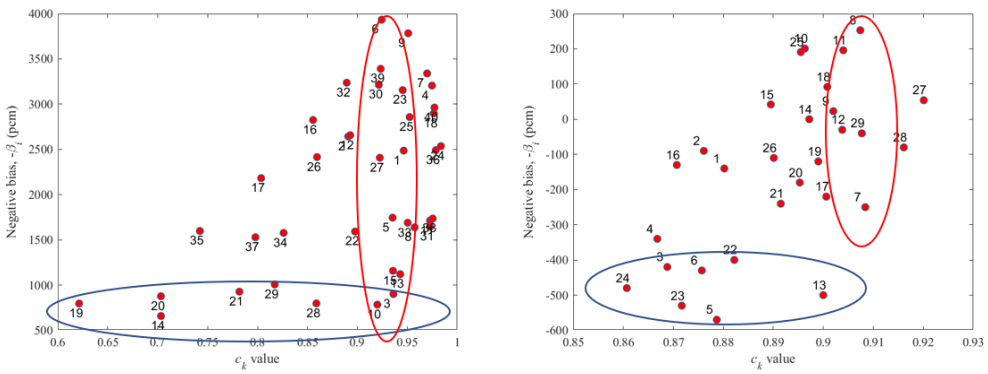
Similar to the toy problem, a simple exercise was repeated to compare the impact of Whisper weights, by comparing three cases with equal weights, limiting to the analysis to the most biased ten PDFs, and including all experiments with the c_k-based weights. Figure 10 shows that similar CM values are obtained for the three cases, indicating that the CM values is weakly sensitive to the Whisper's weighting procedure.



(a) All models with unit weight **(b) Ten most biased with unit weight** **(c) All models with Whisper weights**

Figure 10. Impact of benchmark and weight selection on extreme value

Given these results, we perform an additional experiment to determine whether the c_k weighting can effectively reduce the impact of the most negatively biased experiments. To achieve that, the models/benchmarks are grouped in two different ways: the blue groups in both plots of Figure 11 have a low bias, but different c_k values, and the red groups have different biases (including the most biased ones), but similar c_k values.



(a) Toy model **(b) Pu-Sol benchmark**

Figure 11. Bias and c_k value scatter plot

For each group, one experiment is singled out and its weight is gradually reduced to zero to estimate the impact of the c_k weighting and the most negative biases on the calculated CM values. The two graphs on the left of Figure 12 single out one experiment at a time based on the c_k value, and the ones on the right are based on the most negative bias. For example, the dark red plot on the bottom left graph singles out the 13th experiment with $c_k = 0.9$ and gradually reduces its weight. The graphs on the right perform the same experiment but single out the experiments based on their biases. For example, the blue graph on the bottom right singles out the 8th experiment whose bias is -253 pcm. This is the experiment with the most negative bias and is expected to have the biggest influence on the results. Results indicate that the weighting procedure does have an impact, albeit negligible, on the calculated CM value.

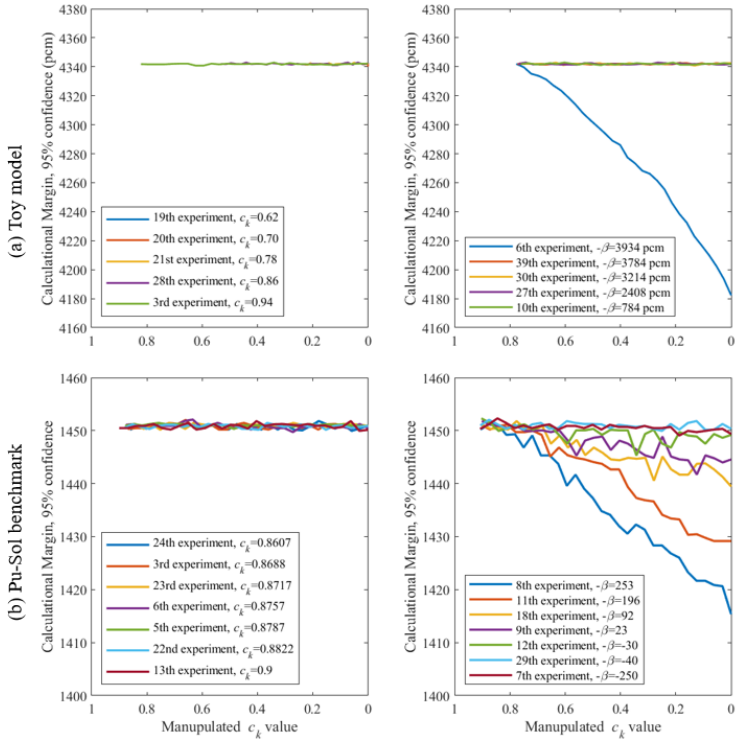


Figure 12. Analysis of Whisper c_k-based weighting

III.C. One-At-a-Time Experiment Validation

To help validate the performance of the four methodologies, a simple numerical experiment is employed taking advantage of the available benchmark models and their reported measurements. A single benchmark model is singled out as the application, and its true bias is compared to the CM and MOS values calculated by the four methodologies. An ideal performance would be one in which the true bias is similar in magnitude to the calculated bias and is upper bounded by the sum of the CM and MOS values. A library of 62 uranium-fueled benchmark experiments is employed, wherein the noted procedure is repeated 62 times, selecting a different experiment as the application in each time. The detailed information about the experiments is in Table 5.

Table 5. Benchmark Models Specification

Benchmark name	Measured, k_m		Calculated, k_c			Bias ($k_c - k_m$)	
	k_{eff}	σ_m	k_{eff}	σ_ϵ	σ_s	β	σ_e
HEU-SOL-THERM-013-001	1.0012	0.0026	0.9976	0.0010	0.0078	-0.0037	0.0028
LEU-COMP-THERM-010-013	1.0000	0.0021	0.9977	0.0010	0.0070	-0.0023	0.0023
LEU-COMP-THERM-017-008	1.0000	0.0031	0.9978	0.0010	0.0064	-0.0022	0.0033
LEU-COMP-THERM-002-001	0.9997	0.0020	0.9976	0.0010	0.0078	-0.0021	0.0022
LEU-COMP-THERM-010-008	1.0000	0.0021	0.9979	0.0010	0.0067	-0.0021	0.0023
LEU-COMP-THERM-001-006	0.9998	0.0030	0.9977	0.0010	0.0067	-0.0021	0.0032
LEU-COMP-THERM-002-004	0.9997	0.0020	0.9977	0.0010	0.0075	-0.0020	0.0022
LEU-COMP-THERM-002-005	0.9997	0.0020	0.9978	0.0010	0.0073	-0.0020	0.0022
LEU-COMP-THERM-001-007	0.9998	0.0030	0.9979	0.0010	0.0066	-0.0019	0.0032
LEU-COMP-THERM-001-002	0.9998	0.0031	0.9980	0.0010	0.0068	-0.0018	0.0033
LEU-COMP-THERM-017-012	1.0000	0.0031	0.9982	0.0010	0.0063	-0.0018	0.0033
LEU-COMP-THERM-017-010	1.0000	0.0031	0.9983	0.0010	0.0063	-0.0017	0.0033
LEU-COMP-THERM-017-011	1.0000	0.0031	0.9983	0.0010	0.0063	-0.0017	0.0033
LEU-COMP-THERM-001-004	0.9998	0.0030	0.9982	0.0010	0.0067	-0.0016	0.0032
HEU-SOL-THERM-001-003	1.0000	0.0025	0.9986	0.0010	0.0124	-0.0014	0.0027
LEU-COMP-THERM-017-013	1.0000	0.0031	0.9987	0.0010	0.0064	-0.0013	0.0033
LEU-SOL-THERM-004-001	0.9994	0.0008	0.9983	0.0010	0.0077	-0.0011	0.0013
LEU-COMP-THERM-017-014	1.0000	0.0031	0.9989	0.0010	0.0064	-0.0011	0.0033
LEU-SOL-THERM-004-003	0.9999	0.0009	0.9988	0.0010	0.0074	-0.0011	0.0013
LEU-COMP-THERM-017-003	1.0000	0.0031	0.9993	0.0010	0.0065	-0.0007	0.0033
LEU-COMP-THERM-002-003	0.9997	0.0020	0.9990	0.0010	0.0077	-0.0007	0.0022
LEU-COMP-THERM-017-007	1.0000	0.0031	0.9994	0.0010	0.0063	-0.0006	0.0033
LEU-COMP-THERM-002-002	0.9997	0.0020	0.9991	0.0010	0.0078	-0.0006	0.0022
LEU-COMP-THERM-001-001	0.9998	0.0031	0.9992	0.0010	0.0069	-0.0006	0.0033
LEU-COMP-THERM-017-006	1.0000	0.0031	0.9995	0.0010	0.0062	-0.0005	0.0033
LEU-SOL-THERM-004-007	0.9996	0.0011	0.9991	0.0010	0.0069	-0.0005	0.0015
LEU-COMP-THERM-017-005	1.0000	0.0031	0.9995	0.0010	0.0062	-0.0005	0.0033
LEU-COMP-THERM-010-005	1.0000	0.0021	0.9999	0.0010	0.0060	-0.0001	0.0023
LEU-SOL-THERM-004-006	0.9994	0.0011	0.9993	0.0010	0.0070	-0.0001	0.0015
LEU-COMP-THERM-010-012	1.0000	0.0021	1.0001	0.0010	0.0068	0.0001	0.0023
LEU-COMP-THERM-010-006	1.0000	0.0021	1.0001	0.0010	0.0062	0.0001	0.0023
HEU-SOL-THERM-001-006	1.0000	0.0025	1.0004	0.0010	0.0111	0.0004	0.0027
LEU-SOL-THERM-004-005	0.9999	0.0010	1.0003	0.0010	0.0071	0.0004	0.0014
LEU-COMP-THERM-017-002	1.0000	0.0031	1.0004	0.0010	0.0065	0.0004	0.0033

LEU-SOL-THERM-004-002	0.9999	0.0009	1.0004	0.0010	0.0075	0.0005	0.0013
LEU-COMP-THERM-017-001	1.0000	0.0031	1.0006	0.0010	0.0065	0.0006	0.0033
LEU-SOL-THERM-004-004	0.9999	0.0010	1.0007	0.0010	0.0072	0.0008	0.0014
LEU-COMP-THERM-010-009	1.0000	0.0021	1.0010	0.0010	0.0068	0.0010	0.0023
LEU-COMP-THERM-010-010	1.0000	0.0021	1.0011	0.0010	0.0068	0.0011	0.0023
LEU-COMP-THERM-010-007	1.0000	0.0021	1.0012	0.0010	0.0066	0.0012	0.0023
LEU-COMP-THERM-010-011	1.0000	0.0021	1.0012	0.0010	0.0068	0.0012	0.0023
LEU-COMP-THERM-010-003	1.0000	0.0021	1.0039	0.0010	0.0072	0.0039	0.0023
LEU-COMP-THERM-010-001	1.0000	0.0021	1.0044	0.0010	0.0072	0.0044	0.0023
LEU-COMP-THERM-010-002	1.0000	0.0021	1.0051	0.0010	0.0072	0.0051	0.0023
HEU-SOL-THERM-001-010	1.0000	0.0025	0.9897	0.0010	0.0108	-0.0103	0.0027
HEU-SOL-THERM-001-006	1.0000	0.0021	0.9977	0.0010	0.0070	-0.0023	0.0023
HEU-SOL-THERM-001-005	1.0000	0.0031	0.9978	0.0010	0.0064	-0.0022	0.0033
HEU-SOL-THERM-001-007	0.9997	0.0020	0.9976	0.0010	0.0078	-0.0021	0.0022
HEU-SOL-THERM-001-003	1.0000	0.0021	0.9979	0.0010	0.0067	-0.0021	0.0023
HEU-SOL-THERM-001-008	0.9998	0.0030	0.9977	0.0010	0.0067	-0.0021	0.0032
HEU-SOL-THERM-001-001	0.9997	0.0020	0.9977	0.0010	0.0075	-0.0020	0.0022
HEU-SOL-THERM-001-009	0.9997	0.0020	0.9978	0.0010	0.0073	-0.0020	0.0022
HEU-SOL-THERM-001-004	0.9998	0.0030	0.9979	0.0010	0.0066	-0.0019	0.0032
HEU-SOL-THERM-001-002	0.9998	0.0031	0.9980	0.0010	0.0068	-0.0018	0.0033
LEU-SOL-THERM-004-001	1.0000	0.0031	0.9982	0.0010	0.0063	-0.0018	0.0033
LEU-SOL-THERM-004-002	1.0000	0.0031	0.9983	0.0010	0.0063	-0.0017	0.0033
HEU-SOL-THERM-013-004	1.0000	0.0031	0.9983	0.0010	0.0063	-0.0017	0.0033
HEU-SOL-THERM-013-003	0.9998	0.0030	0.9982	0.0010	0.0067	-0.0016	0.0032
LEU-SOL-THERM-004-003	1.0000	0.0025	0.9986	0.0010	0.0124	-0.0014	0.0027
HEU-SOL-THERM-013-002	1.0000	0.0031	0.9987	0.0010	0.0064	-0.0013	0.0033
LEU-SOL-THERM-004-004	0.9994	0.0008	0.9983	0.0010	0.0077	-0.0011	0.0013
LEU-SOL-THERM-002-002	1.0000	0.0031	0.9989	0.0010	0.0064	-0.0011	0.0033

First, focusing on TSURFER and Whisper methodologies, their resulting CM and MOS are compared with the true bias of the selected application as shown in Figure 13. The TSURFER MOS is set to be a fixed value of 500 pcm as in the previous analyses. The results show that the sum of two margins, i.e., CM and MOS, for both methodologies are larger than the true application bias represented as the dashed line except for the right most point of TSURFER.

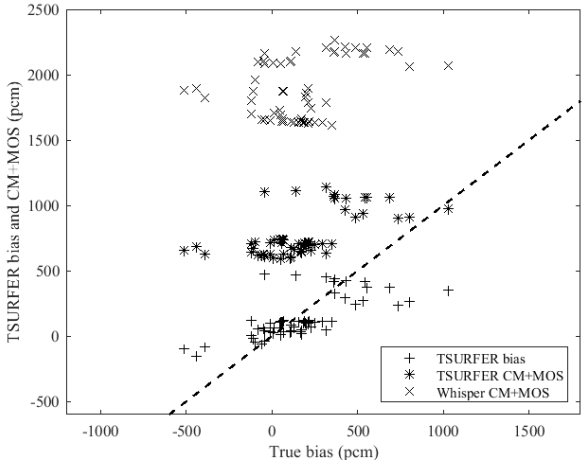


Figure 13. CM and MOS evaluation with different application

Next, the CM and USL values for the four methodologies are compared in Figure 14 against the true bias. Note that in point on the x-axis represents the selection of different experiment as an application, and all other 61 experiments are employed to estimate the CM and USL values. Only the positive bias cases are considered important, that's when the code under predicts the measured value. The order of the applications on the x-axis is selected by ordering the Whisper CM values from low to high.

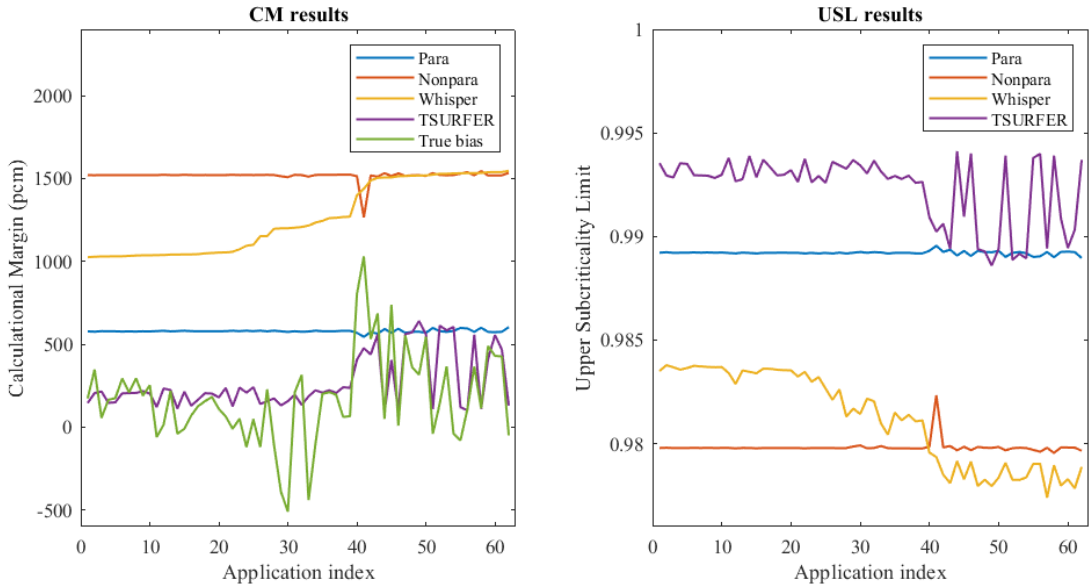


Figure 14. CM and USL with different application

Note that the parametric and non-parametric CMs on the left do not change significantly regardless of application selection, because both methodologies evaluate their uncertainty using weighted statistics which often yields stable results with a large number of experiments. But notice that the non-parametric CM curve drops once for application HEU-SOL-THERM-001-010. This application has the highest bias of -1030 pcm and a prior uncertainty of 1080 pcm. This drop occurs because the non-parametric bias solely depends on the most biased experiment. The parametric bias is less impacted by the exclusion of this high bias since it relies on a weighted average formula which is more robust to outliers.

With regard to TSURFER results, the CM values are very close to the true bias, however noticeable differences can be observed for one experiment when used as the application. For this case, the non-parametric CM shows a drop of approximately 250 pcm value. TSURFER is not able to capture this bias which is likely due to the modeling errors that are not factored into the TSURFER CM calculations. This follows because TSURFER assumes all errors originate from known epistemic sources of uncertainties, e.g., nuclear cross-sections. To hedge against this unknown source of errors, TSURFER employs the MOS as an additional margin.

Finally, when comparing the Whisper and non-parametric USL values, it is observed that Whisper sometimes becomes more conservative than the non-parametric methodology. This occurs after adding the MOS term which provides an additional margin for the non-covered cross-sections uncertainties, represented by the residual uncertainty after performing TSURFER-like cross-

section adjustments. TSURFER accounts for this residual uncertainty in the CM value. As pointed out earlier, this effectively results in double-counting for the residual nuclear data uncertainties, hence the lower USL values.

IV. Conclusion and Further Research

The main conclusions may be summarized as follows: 1) The parametric, non-parametric, and Whisper methodology rely primarily on the subjective ability of the analyst to select experiments that have biases of approximately equal magnitude to the unknown application bias. The use of similarity indices like the c_k metric does not guarantee that the application and a given experiment have the same bias magnitude even if they have perfect similarity. This situation occurs when the sensitivity profiles are pointing in the same direction but with different magnitude, a situation that blinds the similarity index. The implication is that all three methodologies can potentially under-predict the true application bias if the norm of the application's sensitivity profile is larger in magnitude than that of the experiments; 2) The non-parametric and Whisper methodologies are very sensitive to the experiment(s) with the highest bias and/or uncertainty, meaning that the addition of similar experiments with low uncertainty does not help improve the confidence in the calculated application bias. For the Whisper methodology, the bias continuously increases with the number of experiments, implying that the addition of experiments with similar biases/uncertainties reduces rather than increases the confidence in the calculated application bias and its uncertainty. To limit this unbounded bias increase, Whisper employs a heuristic thresholding methodology; 3) The TSURFER methodology is sensitive to the presence of uncharacterized error sources, referred to as modeling errors, with the sensitivity increasing with the similarity index, meaning that TSURFER could under-predict the true application bias if the experiments with high similarity have uncharacterized modeling error sources. Thus, future work will focus on quantifying uncharacterized error sources using cross validation to optimize TSURFER methodology, and extend this idea to more complicated problems, where the dependence between experiment(s) and application is highly nonlinear, that may arise in many nuclear engineering applications due to the complexity of nuclear system.

Acknowledgement

This work was funded in part by Oak Ridge National Laboratory.

Appendix A. Parametric Methodology

Consider conducting N experiments, each with a different gradient vector, and assume the application gradient is not included in the analysis. Each experiment records a measured value of k_{m_i}, a corresponding calculated value of k_{c_i}, and their evaluation uncertainty σ_{e_i}. Let the bias be given as $\beta_i = k_{c_i} - k_{m_i}$. Thus, each experiment defines its own PDF of expected deviations between measured and predicted values, defined as a normal distribution with mean value β_i, denoted as the experimental bias, and uncertainty given by the standard deviation σ_{e_i}. As reported in the literature, the parametric approach calculates the application bias β_p as:

$$\beta_p = \overline{k} - \overline{m} \tag{4}$$

where

$$\overline{k} = \sum_{i=1}^{N} \left(\frac{k_{c_i}}{\sigma_{e_i}^2} \right) / \sum_{i=1}^{N} \left(\frac{1}{\sigma_{e_i}^2} \right)$$
$$\overline{m} = \sum_{i=1}^{N} \left(\frac{k_{m_i}}{\sigma_{e_i}^2} \right) / \sum_{i=1}^{N} \left(\frac{1}{\sigma_{e_i}^2} \right)$$

And the pooled variance σ_p^2 is defined as s^2, the sum of the weighted variance in k about the mean and $\overline{\sigma}^2$, the average variance such as

$$\sigma_p^2 = s^2 + \overline{\sigma}^2 \tag{5}$$

where

$$s^2 = \frac{N}{N-1} \sum_{i=1}^{N} \left(\frac{\beta_i - \beta_p}{\sigma_{e_i}} \right)^2 / \sum_{i=1}^{N} \left(\frac{1}{\sigma_{e_i}^2} \right) \tag{6}$$

$$\overline{\sigma}^2 = N \left(\sum_{i=1}^{N} \left(\frac{1}{\sigma_{e_i}^2} \right) \right)^{-1} \tag{7}$$

The CM is calculated as the sum of the bias and its uncertainty multiplied by the one-sided tolerance factor ϱ such as

$$\mathrm{CM}_p = -\beta_p + \varrho \sigma_p + \Delta_m \tag{8}$$

where non-conservative bias adjustment parameter $\Delta_m = \max\{0, \beta_p\}$ is introduced to avoid non-conservative bias. Finally, the USL for the parametric methodology is given by:

$$\begin{aligned} \mathrm{USL}_p &= 1.0 - \mathrm{CM}_p - \mathrm{MOS}_p \\ &= 1.0 + \beta_p - \varrho \sigma_p - \Delta_m - 0.005 \end{aligned} \tag{9}$$

These equations are provided in [16], citing an older reference [15], which does not derive nor cite a statistical justification for these equations, instead they are listed without proof. The goal of this work is to explain the origin of these equations and judge their adequacy for the bias calculation.

Starting with the mean bias equation, Eq. (4), we make the following observations.

1. The individual experimental biases represent the systematic deviations between the measured and calculated eigenvalue, with the spread of each PDF determined by the aleatory uncertainties resulting from the evaluation procedure, i.e., inclusive of both benchmark uncertainties and measurement uncertainties.

2. The calculation of the mean value in Eq. (4) emulates the Bayesian estimation of the mean of an assumed super distribution for all possible experimental biases. This assumption is not correct because this distribution is not a proper distribution, i.e., it is ill-defined, for the following reasons. Recall the discussion on the systematic bias dependence on the inner product between the cross-section error vector and the experiment gradient. Building a histogram of the experimental biases implies building a PDF that describes the distribution of biases from all conducted (or possible to conduct) experiments. This PDF however reflects the distribution of experiments selected by the analyst, i.e., they are not random. If indeed the experiments are selected randomly, i.e., with gradients that are randomly pointing in the cross-section space, the resulting PDF will simply have a zero mean, since all directions are equally probable to be selected at random. Moreover, this PDF is expected to have a finite range from a maximal negative value when the experiment gradient is opposite in direction to the cross-sections error vector, and passing through zero when the gradient is orthogonal to the error vector, and up to a maximum value when the gradient is parallel to the error vector. The maximum negative and positive limits depend on the norm of the gradients for the selected experiments. If the analyst selects experiments with high relevance score, the resulting PDF will have a mean value that is close to the application bias. Therefore, the shape of this PDF is entirely based on the decisions made by the analyst, implying that the mean value of this PDF will also be heavily impacted by the selected experiments, ranging from a situation where the mean is entirely non-informing about the true application bias to being maximally informing when all experiments have perfect relevance score.

3. Assuming all experiments have similar aleatory spread, i.e., $\sigma_{e_i} = \text{constant}$, the mean value reduces to a simple average formula of all the experiments' biases. As noted earlier, this is acceptable only if all experiments have the same norm for their gradient vectors, which is unlikely to be the case. Thus, this averaging could have unpredictable results. Consider for example a situation where the selected experiments have near perfect relevance score to the application. However, the normed application's gradient has a magnitude that is larger than any of the experiments' gradients. The result is that the true application bias would be bigger the mean bias calculated from the experiments, which is an undesirable scenario. This situation is depicted in the numerical section,

where multiple experiments with nearly equal relevance have a wide range of bias values.

4. Considering the formula used for the standard deviation for the bias, Eq. (5), which takes advantage of a famous theorem from statistics, called variance decomposition theorem or total variance theorem, sometimes referred to by practitioners as pooled variance. This theorem states that the variance may be decomposed into two terms, variance of the means and mean of the variances. This theorem is useful when analyzing a superset of data composed of multiple datasets, each with its own mean and variance, and the goal is to calculate the variance of the superset. The theorem states that one can achieve that by first calculating a superset mean, which represents the mean of all the means of the individual datasets. Next, one calculates the variance of the means of the datasets around the calculated superset mean, denoted by the variance of the means, represented by Eq. (6). Next, one calculates the average of the variances of the individual datasets, denoted by the mean of the variances. One can show that the variance of the means plus the mean of the variances is equal to the variance of all the data in the superset. In our context, each dataset represents the PDF of the bias from an individual experiment, and the superset is the ill-defined PDF of all possible experiments. This definition is problematic because:

    a. The first term, the variance of the means, captures the variance of the experiments selected by the analyst. If these experiments have similar biases, they will underestimate the true bias uncertainty for the application, and if they are very different, they could overestimate the true value. Again, this is all because the hypothesized PDF for which the mean and standard deviation are calculated is ill-defined.

    b. The second term, the mean of the variances, is inconsistent with the formula given by variance decomposition theorem, and its definition cannot be traced to a source in the literature. This formula tries to calculate the average standard deviation as the inverse of the average confidence, which is different from direct calculations of the average variance. In Bayesian statistics, the inverse variance is often denoted as the confidence. The idea of using confidence instead of variance is a direct result of Bayesian updating when one is trying to estimate the mean value of a given distribution, inferred from multiple samples from the distribution [18]. The definition in Eq. (7) resembles the Bayesian update formula but it contains an additional N factor. The reference [15] which originally proposed this formula does not provide a justification for it. In other parts of this reference, a classical textbook is cited [20], which does not contain this formula; instead it contains the Bayesian update formula. The Bayesian formula is designed to increase confidence in the estimated mean as more samples are added. Eq. (7) is problematic because if one of the experiments has very low uncertainty, resulting from extremely careful measurements and benchmarking practices, the resulting variance will approach zero in the limit

of one perfect measurement, and the resulting application bias will be solely determined by this experiment, which may not even have a high relevance score. This means that for the benchmark uncertainties are being effectively treated as epistemic rather than aleatory uncertainties.

Despite these issues, the parametric approach produces conservative results from a safety analysis viewpoint as long as the following two conditions are satisfied: a) the aleatory uncertainties for the different experiments are similar in magnitude, ensuring that the bias is not influenced by a single or few experiments (due to the incorrect use of the confidence rather than variance to calculate the average variance); and b) the selected experiments have a wide range of biases, covering the range of variations from prior cross-sections uncertainties, thereby resulting in large enough bias uncertainty, which raises no red flags about its adequacy for the application conditions. If these two conditions are satisfied, then the parametric approach would calculate a mean bias and a standard deviation that are representative of the epistemic uncertainties resulting from the cross-sections. Due to the pooled variance formula in Eq. (5), it is also effectively capturing the evaluation uncertainties.

A key challenge with the parametric approach is that as analysts transition to using high fidelity simulation tools, the biases for the existing body of benchmark experiments are expected to get smaller, much smaller than the range implied by the prior cross-sections uncertainties. The resulting application bias and bias uncertainty will be smaller, rendering them under-conservative for the application conditions. With a lack of relevant experiments, which is common for first-of-a-kind nuclear systems, the licensor will require additional conservative margin. However, from the practitioner's perspective, excessively large margin may restrict design freedom and lower system economy, which is undesirable because it does not provide a venue for taking credit for the epistemic uncertainties.

Recalling Table 1, the parametric approach effectively accounts for the systematic bias resulting from cross-sections ②, and also the aleatory sources ③ and ④, as long as the aleatory sources have the same magnitudes across the pool of available experiments. The solutions uncertainties ① are captured under the MOS margin.

Appendix B. Non-parametric Methodology

The CM for the non-parametric methodology is the same as in Eq. (8) except that the bias is determined as the minimum bias of all the benchmarks biases, and an additional non-parametric margin m_{np} is heuristically added if the number of benchmarks involved is small. The CM for the non-parametric methodology can be written as:

$$\mathrm{CM}_{np} = -\min\left\{k_{c_i} - k_{m_i}\right\} + \varrho\sigma_p + \Delta_m + m_{np} \tag{10}$$

The USL for the non-parametric methodology is:

$$\begin{aligned} \mathrm{USL}_{np} &= 1.0 - \mathrm{CM}_{np} - \mathrm{MOS}_{np} \\ &= 1.0 + \min\left\{k_{c_i} - k_{m_i}\right\} - \varrho\sigma_p - \Delta_m - m_{np} - 0.005 \end{aligned} \tag{11}$$

The basic non-parametric methodology may be stated as follows: given the ability to randomly generate samples from an unknown PDF, determine the number of samples and a corresponding upper tolerance limit that covers a preset portion of the PDF with preset confidence. Note that this problem statement assumes that the PDF is unknown, i.e., it cannot be parametrized in terms of the PDF's features, like the mean and standard deviation. The non-parametric approach solves this inference problem by employing a sampling-based approach to construct a related extreme value EV PDF. Figure 15 graphically demonstrates how the EV PDF may be constructed.

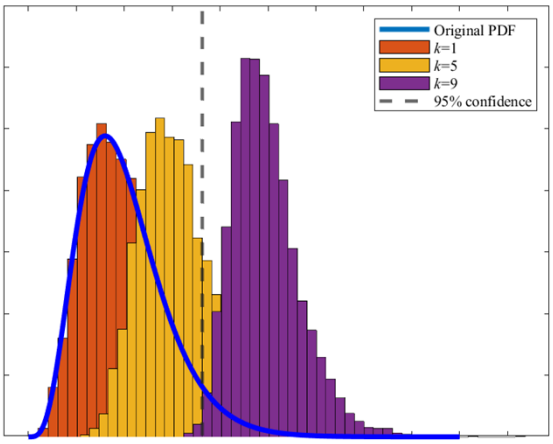


Figure 15. Extreme value statistics example

Assume first that the original PDF type is known to be a gamma distribution but its parameters are unknown, and one is interested in estimating its upper tolerance limit corresponding to 95% coverage. The idea is to first pick an order, say k, which implies the need to generate k samples from the original PDF, and to take their extreme value, i.e., maximum, representing a single sample

of the k^{th} order EV PDF[3]. If the type of the original PDF distribution is known, one can exactly determine the form of the EV PDF either analytically or via exhaustive numerical experiments.

The EV PDF has an interesting behavior; its mass keeps on shifting to the right with increasing the order, i.e., with more samples from the original PDF. If one is interested in finding an upper tolerance limit for the original PDF, say with 95% coverage, shown as the grey vertical dashed line, then one simply needs to find an EV PDF whose mass is mostly above this limit. We say "mostly" because it is impossible to find an EV PDF whose entire mass is above the tolerance limit, simply because the EV PDF is expected to have a tail stretching to the smallest values attained by the original PDF, e.g., negative infinity for a normal distribution. Thus, one needs to find the minimum k that renders a preset portion of the EV PDF above the sought tolerance limit. Assume for example that one finds that the k^{th} order EV PDF shown in purple has 3% of its area below the 95% upper tolerance limit (the grey dashed vertical line) for the original PDF. This means that if one generates a single EV sample (obtained by sampling k samples from the original PDF and taking the maximum), there will be 97% chance that the EV sample will be higher than the 95% tolerance limit. Thus, one can state with 97% confidence that k samples are sufficient to determine a 95% upper tolerance limit for the original PDF. Clearly as the number of affordable samples from the original PDF increases, the confidence in the upper limit could be increased, never reaching 100%.

This simple example can be easily generalized when the type of the original PDF is not known. The mathematical argument would be as follows: first calculate the probability that k samples from the original PDF would be less than the $p\%$ tolerance limit. Since all the samples are independent, this probability is p^k. Then $1 - p^k$ must be the probability that at least one of the samples (i.e., the maximum) is greater than the $p\%$ tolerance limit. Thus, one can state that with $1 - p^k$ confidence the maximum of k independent samples drawn from the original PDF could be used as a $p\%$ upper tolerance limit. The most widely known result of EV PDF is the famous Wilks's formula which states that for $k = 59$ and $p = 95\%$, one can determine a 95%/95% upper tolerance limit for any distribution, as long as one can draw independent samples from the same distribution. Said differently, if the original PDF is not known, the maximum of 59 randomly generated samples, all drawn from the same distribution, would serve as a 95% upper tolerance limit with 95% confidence. The key challenge with this approach is that many samples would be needed to develop high confidence in the tolerance limit.

Note that the basic non-parametric approach does not require one to estimate the original PDF's features, e.g., mean value and standard deviation for a normal distribution; instead the tolerance limit can be determined directly. This also means that if the features can be readily estimated, it would be moot to try the non-parametric approach, simply because there are known formulas and/or tables for determining the tolerance limit as a function of the features. If one proceeds with

[3] While other literature defines "k^{th} order" as the statistics seeking k^{th} smallest or k^{th} largest (order) value of given PDF(s), in this manuscript "k^{th} order" means by the extreme value of k samples from PDF(s).

the non-parametric approach, they would be able to obtain the same results obtained with the parametric approach, since in this case the type of the original PDF is known.

Also note that all samples must be independent and generated from the same PDF, i.e., the original PDF for which a tolerance limit is sought. Mathematically, the samples are denoted as iid samples, short for independent samples from identical distributions, in our case, this means generating samples from the "same" distribution. This requirement is important for two reasons, first to ensure that the EV PDF samples can be related to the tolerance limit of the original PDF generating the samples, and second to ensure that the EV PDF progressively moves to the right with higher orders. To demonstrate, consider Figure 16, where one attempts to generate the 3rd order EV PDF using three different PDFs (i.e., an incorrect application of EV theorem because the samples are no longer iid). In the first case represented by the top two plots for PDFs with low overlap, the EV PDF will be heavily biased by the third PDF, i.e., the most extreme of the three, which essentially reduces to sampling only the third PDF. On the other hand, as the PDFs get closer to each other, as shown in the two bottom plots, the EV PDF will start to shift towards the right, reducing back to the iid case. More importantly, with different PDFs used to generate the samples, it is no longer clear which tolerance limit is being estimated. The relevance of this observation will become clear when we discuss the Whisper methodology.

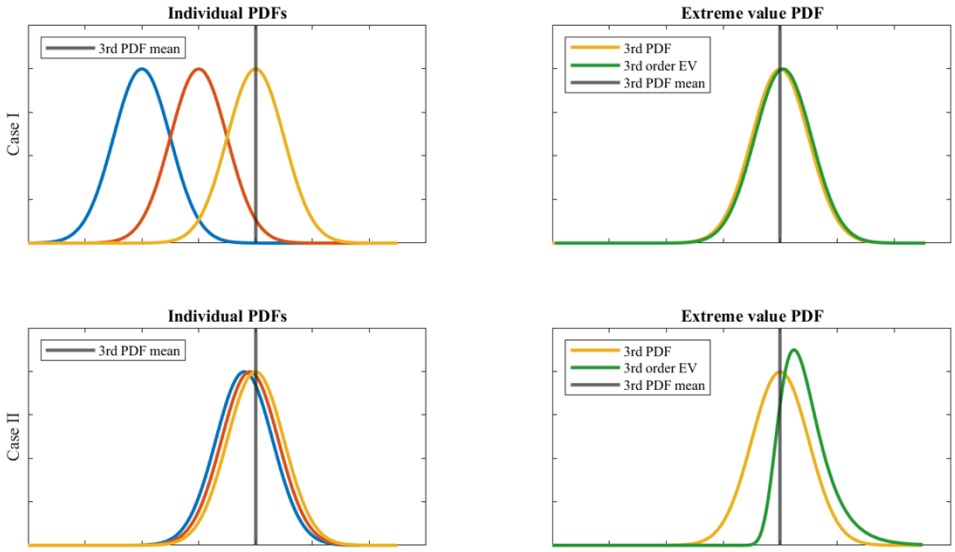


Figure 16. Extreme value statistics with different PDFs

Next, consider that one is interested in estimating $p\%$ upper tolerance with $q\%$ confidence for the application bias. The goal is to determine N the minimum number of experiments to achieve that. A straightforward application of non-parametric methodology requires one to assume that there exists an unknown PDF from which biases are sampled, allowing one to determine the sought tolerance limit. If this assumption is acceptable, one simply solves the following equation for N, $q = 1 - p^N$. This assumption however is problematic as discussed earlier because this hypothetical PDF is ill-defined. As discussed in the previous section, there is no such a PDF that describes all

possible experiments, simply because it is up to the analyst to select what experiments to include and how relevant they are to the application. The situation is different from its common use in other engineering applications, such as manufacturing where one is interested in estimating a tolerance limit for a clearly defined process implying a well-defined PDF. For example, consider an enrichment plant configured to produce low-enriched fuel pellets at a nominal enrichment of 4%. Due to the inherent uncertainties in the process, the fuel pellets' enrichments are expected to have a PDF with a mean value of 4% and some spread. The distribution of the fuel pellets enrichment describes a PDF for which a tolerance limit can be calculated by sampling N pellets.

Applying the non-parametric methodology to samples that are not iid, i.e., generated from different experiments, challenges the basic assumption of the EV PDF. As an example, assume that the analyst selects experiments that have a relevance score above a minimum threshold, e.g., 0.85. In this case, the spread of the resulting PDF will be determined by both the spread of the relevance score as well as the spread of the norms of the experimental gradients, see the earlier discussion surrounding Eq. (1). Thus, for this approach to be effective, the norm of the application gradient needs to be in the same order of magnitude as that of the experiments. If it is higher, then the calculated tolerance will be under-predicting the real tolerance for the application bias. A better approach would be to scale down the biases by their corresponding gradients' norms. If one lowers the minimum threshold for experimental relevance, the resulting PDF would be wider, thus conservatively impacting the calculated tolerance for the application bias. In response to this non-standard use of the EV theorem, the non-parametric methodology, as used in the nuclear criticality safety literature, includes an additional term to the tolerance limit, representing the variance of the bias from all available experiments, i.e., the error term $\varrho\sigma_p$ appearing in Eq. (10).

It is thus concluded here that the basic non-parametric approach has the following advantages and disadvantages. It allows analysts to estimate an upper tolerance limit for the application bias with minimal knowledge about the various sources of uncertainties. In doing so, one must ensure that the application gradient is similar in magnitude to the experiments, which is possible with expert judgment. If one could employ an experimental relevance score, the calculated tolerance would be closer to the true value for the application bias, allowing one to drop the additional conservative term $\varrho\sigma_p$. If no knowledge about the application is included, the resulting tolerance is determined by the worst experimental bias plus an additional term capturing the variance of the experimental biases $\varrho\sigma_p$. Finally, it does not allow the analyst to take credit for the irreducible sources of uncertainties.

Recalling the sources of uncertainties, it hedges for the epistemic uncertainties, source ②, as it is based on the worst systematic bias, and the evaluation uncertainties, sources ③ and ④, as it employs the pooled variance as an additional term in the CM definition. Recall the pooled variance contains a term that averages the evaluation uncertainties from all experiments. Finally, it accounts for the solution uncertainties, source ① in the MOS term.

Appendix C. Whisper Methodology

The Whisper methodology has been developed by Los Alamos National Laboratory researchers [16]. It is promoted to provide the following features: a) it hybridizes the use of parametric and non-parametric methodologies; b) it relies on the concept of EV theorem and uses calculated tolerance to set the CM; c) it employs a heuristic formula to reduce the number of samples generated from low-relevance experiments in an attempt to reduce their impact on the calculated tolerance limit; and finally d) it employs TSURFER-based approach to determine the non-covered uncertainties which are used to set the MOS.

The full implementation may be found in the following reference [16], however a brief overview of the steps is given here. First, it generates an EV-like PDF which is used to calculate a tolerance value m, covering preset area q, say 95%, under the EV-like PDF. We explain later why we use the qualifier "like" when describing Whisper's EV PDF and the associated tolerance. The CM is determined as:

$$\mathrm{CM}_w = m + \Delta_m \tag{12}$$

The MOS for the Whisper methodology can be represented by a sum of three terms, i.e., margin for software error (in our notation, the solution uncertainties, source ①), margin for the non-covered nuclear cross-sections uncertainties, and margin for the application. As per the expert opinion, the margin for software is set to be 0.005 and the margin for non-covered cross-sections is calculated by the TSURFER methodology, such that

$$\mathrm{MOS}_d = \varrho\sigma_{k'} \tag{13}$$

where $\sigma_{k'}$ is the residual, i.e., non-covered, uncertainty for the application response, resulting from a TSURFER-based adjustment procedure. The USL for the Whisper methodology can be written as

$$\begin{aligned} \mathrm{USL}_w &= 1.0 - \mathrm{CM}_w - \mathrm{MOS}_w \\ &= 1.0 - m - \Delta_m - 0.005 - \varrho\sigma_{k'} \end{aligned} \tag{14}$$

Markedly different from the basic non-parametric methodology, Whisper generates an EV-like PDF using samples that are not iid, this is because they are generated from different PDFs. Each PDF represents one experiment with the PDF assumed known, i.e., in the normal case the experimental bias sets the PDF's mean value and the evaluation uncertainty sets the PDF's standard deviation. Then it calculates an EV-like PDF of k^{th} order, with k being the effective EV order, i.e., the number of the overlapping extreme experiments (Details on how this is performed will be given in the numerical sections). Since the original PDFs are fully characterized, Whisper explicitly constructs the k^{th} order EV-like PDF, which can be done analytically if the original PDFs are normal, or it can be done numerically for general PDFs. Finally, it defines the tolerance limit m as the value that covers a preset portion of the EV-like PDF.

Further, Whisper employs a linear heuristic method for two reasons: first, to diminish the number of samples generated from low relevance experiments by assigning a weight w that varies linearly with the relevance score, meaning that only w% of its generated samples are used to construct the EV-like PDF for each experiment. Another more subtle reason is to limit the impact of the number of experiments with similar biases on the calculated tolerance. When an increasing number of experiments with similar biases are included, the resulting EV-like PDF will continue shifting its mass to more extreme values, raising the tolerance, as shown in Figure 16. This is counter-intuitive as one should develop higher confidence in the bias when an increasing number of experiments provide similar bias results, a basic premise of any statistical inference methodology. Whisper sets a maximum threshold on the weights to ensure the calculated tolerance does not increase indefinitely with the number of experiments. The selected function for the required weight is

$$w_{req} = w_{min} + w_{penalty}(1 - c_{k,max}) \tag{15}$$

where w_{min} and $w_{penalty}$ are heuristic constants that are set to be 25 and 100, respectively, for this analysis, and $c_{k,max}$ is the maximum c_k value of the selected benchmark experiments. The sum of individual weight factors w_i should be the same as the required weight w_{req} calculated in Eq. (15) such that

$$w_{req} = \sum_i w_i \tag{16}$$

and the individual weight factors also satisfy the following linear relation with an appropriately selected acceptance c_k, $c_{k,acc}$, such that

$$w_i = \max\left\{0, \frac{c_{k,i} - c_{k,acc}}{c_{k,max} - c_{k,acc}}\right\} \tag{17}$$

An exemplary numerical test is conducted to explain the impact of this weighting procedure with the benchmark experiments in Section III.C. Figure 17 shows how different cut-off values ($c_{k,acc}$) discard experiments.

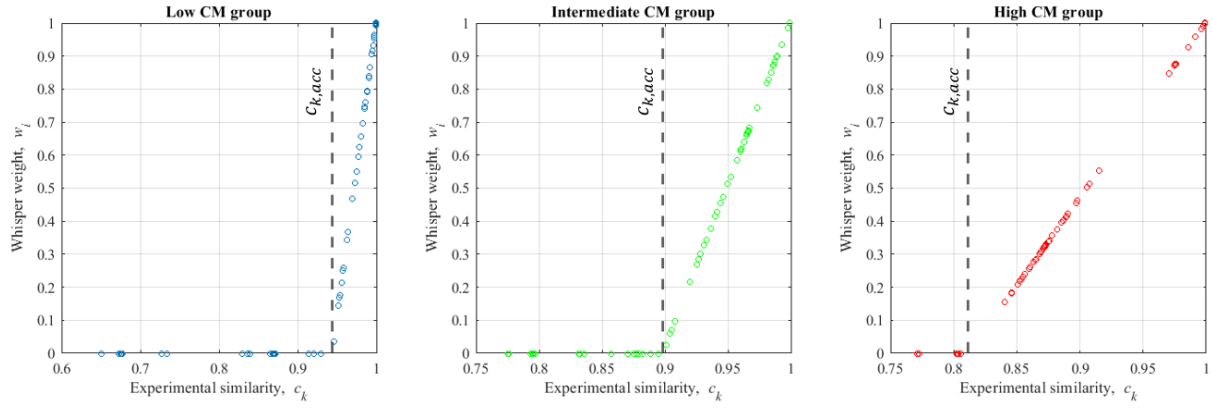


Figure 17. Change in $c_{k,acc}$ with different application selection

These results imply that the $c_{k,acc}$ works as a heuristic cutoff of the linear relationship between the Whisper weights and c_k values. In this analysis, a high value of $c_{k,acc}$ removes the experimental bias PDFs with lower relevance resulting in the reduction of calculated CM. As pointed out earlier, the cutoff procedure is mainly designed to hedge against the monotonic increase in the CM value with the increased number of experiments, see Figure 7. Results indicate that going from a high cutoff value of 0.94 down to a cutoff value of 0.81 reduces the CM value by approximately 500 pcm. Also, notice that within each group, the CM values are fairly constant indicating lack of sensitivity to the specific non-zero weight values used by Whisper for each group, an observation that was supported by earlier numerical experiments.

Finally, Whisper employs the bias uncertainty to calculate the residual uncertainty in the application response which is used to set the MOS. The idea is to report as MOS the aleatory uncertainties from both the non-covered cross-section subspace as well as the evaluation uncertainties.

Focusing on Whisper's CM calculations, the following observations are made.

1. Recall that the non-parametric methodology's real power is that it can create an EV PDF that progressively moves towards the tail end of the original PDF. This is possible if one can generate multiple iid samples from the same PDF and take their maximum values, thus ensuring that the increased samples will push the EVs further towards the tail end of the original PDF. As demonstrated earlier, this logic does not apply when one samples from different PDFs, losing the ability to compare samples from the same PDF. If the PDFs have low overlap (see Figure 16), the EVs will be dominated by the PDF with the highest values, e.g., for normal PDFs, the PDF with the highest standard deviation and/or highest mean value will dominate the EV PDF. This will be demonstrated numerically.

2. When sampling from a single PDF, the goal is to construct an EV PDF whose mass is concentrated above an upper tolerance limit that is already fixed -- albeit unknown -- by the original PDF. If the original PDF was known, one would not need to calculate an EV PDF, because the tolerance limit would be fully determined by the original PDF. Recall that the key power of the non-parametric approach is that allows one to estimate a tolerance limit when the original PDF is unknown.

3. Consider two experiments, one with a very high relevance score and low average bias, represented by the red PDF in Figure 18, and another with lower relevance and higher average bias. One can consider that each experiment represents a group of closely grouped experiments with approximately the same bias and spread. For simplicity, assume the weights for the two groups of experiments are 1.0 and 0.5 respectively. The EV PDF will have 50% of its samples generated from the high relevance PDF(s) and the other 50% from the low relevance PDF(s). This is because 50% of the low relevance samples will be eliminated by the Whisper weighting procedure. The resulting extreme PDF will thus have two modes as shown. Note that each mode is simply a scaled version of the original PDFs. If no weighting is employed, then the EV PDF will simply reduce to the original low

relevance PDF. Consider the tolerance limit corresponding to the case with no weighting, shown as the black vertical bar. The area above this bar is say 5%. The area above the same bar under the Whisper-weighted PDF will be slightly lower than 5%. Hence, to obtain the same confidence, the tolerance limit obtained from the Whisper-weighted EV-like PDF will move slightly to the left to cover the same area.

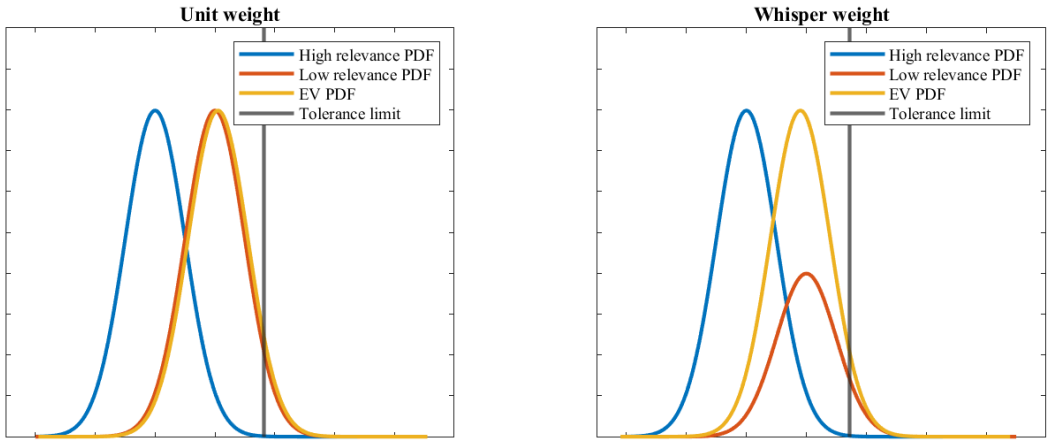


Figure 18. Impact of relevance on tolerance limit

4. With experiments with perfect (or very high) relevance scores employed, the differences in their biases will be mainly determined by the magnitude of their gradients. As explained earlier, if all the experiments have the same relevance but different biases, the application bias will be determined by the experiment with the highest gradient norm. This does not guarantee whether this bias will under or over predict the true application bias without comparing the application gradient norm to the norms of the experiments' gradients. This is not checked by Whisper. Instead, the highest bias is expected to impact the tolerance limit obtained by Whisper. As explained earlier, if the application has a gradient of higher magnitude than the experiment with the highest bias, the calculated bias would under-predict the true bias. The parametric approach hedges for this scenario by employing the pooled variance, which is expected to be big enough, as calculated over many experiments. Whisper does not hedge for this scenario, except based on the analyst's best judgment of selecting experiments with sensitivities of the same magnitude as those of the application.

5. Assuming one employs two experiments with the same relevance score, but with two different evaluation uncertainties, the tolerance limit will be determined by the PDF with the higher evaluation uncertainties. This is because the Whisper weighting employs a relevance score that does not account for the evaluation uncertainties; instead, it is based on the prior cross-sections uncertainties only. Thus, if one conducts the same experiment twice, with one being unreasonably high uncertainty, the Whisper tolerance will be determined by the less accurate measurements which is undesirable from practical considerations. This forces the analyst to design a heuristic criterion to reject experiments before calculating the tolerance limit.

Therefore, Whisper employs CM to account for the systematic bias from cross-sections uncertainties, source ②, because it effectively uses the most conservative bias to set the tolerance limit. In doing so, it does not explicitly account for the difference in magnitude between the experimental and application gradients, however by a) employing the pooled variance's first term; the standard deviation of the biases around their mean value, b) relying on the expert-judgment of the analyst to pick experiments with similar sensitivities to those of the application, and c) the tolerance limit reducing to the most conservative experimental bias like the basic non-parametric methodology, it can be confidently argued that it calculates a conservative estimate of the application bias. It also accounts for the evaluation uncertainties ③ and ④ through the use of pooled variance; the second term, being the mean of the evaluation uncertainties. For MOS, it employs a TSURFER-based procedure to calculate residual uncertainties, which are composed of the aleatory evaluation uncertainties, source ③ and ④, and the non-covered cross-sections uncertainties, a portion of source ②. Because TSURFER relies on the concept of assimilating measurements and predictions to increase confidence, the final uncertainty in the bias will be less than the prior uncertainties in ③, ④ and ②. Thus, the Whisper's MOS will be accounting for a portion of these sources, which were already accounted for in the CM. This double-counting while acceptable from safety point of view, cannot be traced to a statistical justification.

Appendix D. TSURFER Methodology

The TSURFER methodology is parametric; it assumes a Gaussian shape for the PDFs obtained from each experiment. Unlike the standard implementation presented in Section 0, it allows the analyst to take credit for the epistemic uncertainties by solving a mathematical adjustment problem which minimizes L_2 norm of the sum of two terms to find optimal adjustments for the cross-sections. The first term minimizes the L_2 norm of the adjustments of the cross-sections to ensure their consistency with their prior values, and the second term minimizes the discrepancy between the measurements and predictions for the selected experimental responses. The premise is that one can correct for the cross-section errors that belong to the covered subspace. The residual uncertainties resulting from the non-covered subspace are propagated to the response and are used as the basis for calculating the tolerance limit. A key difference between TSURFER and the previous methodologies is that it provides a mathematically justifiable approach to map the biases from the experimental to the application domain, a mapping process that accounts for the differences between the experiments' gradients and the application gradient. The full methodology may be found in a previous publication [17] and a brief discussion on the TSURFER formulation is also summarized below.

Considering that there are M available experiments to predict the application k_{eff} bias, the corresponding prior values, i.e., code-calculated, for both the experiments and the application may be aggregated in a vector $k \in \mathbb{R}^{M+1}$ such that:

$$k = [k_1 \quad k_2 \quad \cdots \quad k_{M+1}]^T$$

where the last element is the prior application k_{eff}. The corresponding measurements for the first M values are designated by another vector $m \in \mathbb{R}^M$. In this formulation, the last element of m is set to the prior value of k_{eff}, assumed to have no corresponding experimental value.

The prior cross-sections uncertainties are described by a multivariate joint Gaussian PDF with a vector of means representing the reference multi-group cross-sections and a covariance matrix given by:

$$\boldsymbol{C_{\alpha\alpha}} = \begin{bmatrix} cov(\alpha_1,\alpha_1) & cov(\alpha_1,\alpha_2) & \cdots & cov(\alpha_1,\alpha_n) \\ cov(\alpha_2,\alpha_1) & cov(\alpha_2,\alpha_2) & \cdots & cov(\alpha_2,\alpha_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\alpha_n,\alpha_1) & cov(\alpha_n,\alpha_2) & \cdots & cov(\alpha_n,\alpha_n) \end{bmatrix} \in \mathbb{R}^{n\times n}$$

The adjusted cross-sections are calculated as the minimizer of the following quadratic form subject to the linearity constraint $k'(\alpha') = m'$:

$$\alpha^* = \underset{\alpha'}{\operatorname{argmin}} [\alpha' - \alpha]^T \boldsymbol{C_{\alpha\alpha}^{-1}} [\alpha' - \alpha] + [m' - m]^T \boldsymbol{C_{mm}^{-1}} [m' - m] \tag{18}$$

where $\boldsymbol{C_{mm}} \in \mathbb{R}^{(M+1)\times(M+1)}$ is the covariance matrix for the measured k_{eff}. The constraint implies that the adjusted cross-sections α' will update the best-estimated values (the components of m') for all M experiments as well as the application. The last element of the vector m' is taken to represent the best-estimate for the application k_{eff} value, and the last component of $m' - m$ is referred to as the application bias.

The objective function in Eq. (18) may be re-written in terms of the calculated and adjusted k_{eff} values as:

$$\chi^2_M = [k' - k]^T \boldsymbol{C_{kk}^{-1}} [k' - k] + [m' - m]^T \boldsymbol{C_{mm}^{-1}} [m' - m]$$

where χ^2_M is the M-degrees of freedom chi-square value describing the discrepancies between the prior and adjusted k_{eff} values. The $\boldsymbol{C_{kk}} \in \mathbb{R}^{(M+1)\times(M+1)}$ matrix denotes the prior covariance matrix for the calculated k_{eff} values given by:

$$\boldsymbol{C_{kk}} = \boldsymbol{S_{k\alpha}} \boldsymbol{C_{\alpha\alpha}} \boldsymbol{S_{k\alpha}^T} \tag{19}$$

where $\boldsymbol{S_{k\alpha}} \in \mathbb{R}^{(M+1)\times n}$ matrix aggregates the first-order sensitivity profiles for all M experiments and the application.

Assuming that the linearization of the constraint $k'(\alpha') = m'$ is sufficiently accurate within the limitations of first-order sensitivity theory, the minimizer of t

he objective function in Eq. (18) may be given by:

$$\Delta k = - \boldsymbol{C_{kk}} (\boldsymbol{C_{kk}} + \boldsymbol{C_{mm}})^{-1} \boldsymbol{d} \tag{20}$$

where $\Delta k = k' - k$ and $d \in \mathbb{R}^{M+1}$ is the discrepancy vector, $d = k - m$

The posterior (i.e., post the consolidation of experimental and prior values) covariance matrix for the k_{eff} values is given by:

$$\boldsymbol{C_{k'k'}} = \boldsymbol{C_{kk}} - \boldsymbol{C_{kk}} (\boldsymbol{C_{kk}} + \boldsymbol{C_{mm}})^{-1} \boldsymbol{C_{kk}} \tag{21}$$

The diagonal elements of this matrix describe the confidence one has in the posterior k_{eff} values.

The CM (also referred to as the LTL) for the TSURFER methodology can be described by the bias β_T that is the last element of Δk in Eq. (20) and its uncertainty $\sigma_{k'}$ that is the square root of the last element of $\boldsymbol{C_{k'k'}}$ in Eq. (21) such that

$$\mathrm{CM}_T = - \beta_T + \varrho \sigma_{k'} \tag{22}$$

And the USL is calculated with the MOS by

$$\begin{aligned} \mathrm{USL}_T &= 1 - \mathrm{CM}_T - \mathrm{MOS}_T \\ &= 1 + \beta_T - \varrho\sigma_{k'} - 0.005 \end{aligned} \tag{23}$$

Since both TSURFER and Whisper employ the idea of cross-section adjustments, albeit for different goals, this work presents a more detailed account of how this is achieved. Instead of focusing on the mathematical details of the minimization problem, the objective is to highlight the key challenges, which remain unaddressed by the nuclear literature, such as the error compensation phenomena, the impact of prior covariance data, the impact of low relevance experiments, and the lack of a formal verification procedure for the calculated application bias, and the impact of modeling errors. To achieve that, the current work dives deeply into the mechanics of cross-section adjustments to help solicit the insight needed to guide future work focused on first-of-a-kind nuclear systems, presenting one of the key contributions of this work. It provides insight on how to interpret the biases calculated which surprisingly could at times degrade rather than improve model predictions; a critically needed discussion that is currently absent from the cross-section adjustment literature.

As mentioned earlier, the regulatory process does not mandate a specific procedure to perform model validation, however it requires that two independent sources of knowledge be consolidated as a basis for establishing confidence in model predictions: 1) the measurements collected from experiments with conditions that are representative of the application, and 2) the model predictions that simulate the same experimental conditions. The premise, as best supported by the Bayes theorem, is that the confidence fused from both sources will be higher than the prior confidence obtained with the simulation only, representing the basic idea behind correcting for the epistemic sources of uncertainties.

Focusing here on cross-sections prior uncertainties, they are typically high and incomplete resulting in high uncertainties for the quantities of interest, e.g., eigenvalue. The experiments, however, are carefully conducted to allow for highly accurate low uncertainty measurements, providing a venue for the analyst to improve model predictions by analyzing the sources of uncertainties responsible for the observed deviations between measured and predicted responses. Because the number of cross-sections is substantially high, it is infeasible to build experiments that can be used to correct for all sources of cross-sections uncertainties. Hence, it is important to devise a methodology to measure the value of an experiment via a relevance score. The goal of these experiments is to regress back the observed deviations to their sources by calculating cross-section adjustments that minimize the deviations.

The search for the optimal cross-section adjustments is cast as an inverse problem that requires an optimization search. A successful search for the optimal adjustments ideally implies the ability to estimate their true values which allows for improved predictions not only for the experimental conditions but also for the application conditions. This is, however, not an easy endeavor because, in most realistic situations, the inverse problem is ill-posed, a situation that arises when the number of cross-sections is much higher than the number of measured experimental responses. The ill-

posedness presents a formidable challenge for the optimization search resulting in the so-called error-compensation phenomenon, where the cross-sections are incorrectly over- or under-adjusted, i.e., as compared to their true unknown errors, leading to the same responses residual, i.e., the post-adjustment deviations between measured and predicted responses.

This makes an inverse problem challenging to choose which cross-section changes would be appropriate for the application because there are theoretically an infinite number of cross-section adjustments that might result in the same degree of agreement between measured and expected responses. This is due to the fact that the application model is not taken into account while determining the best-estimates, making it impossible to predict a priori whether the adjusted cross-sections will result in better or worse model predictions for the application conditions. Although regularization techniques have been developed to render the optimization search well-posed, this can only be done by selecting the adjustments that produce a unique solution by enforcing some mathematical criterion, such as the minimum distance from the best-known prior cross-section values. Because these regularization techniques are blind to the application conditions, they cannot ensure that the adjustments will improve the model predictions for the application conditions.

Note that unlike the previous three methodologies, the TSURFER methodology takes credit for the epistemic sources of uncertainties in the CM calculation, based on a mathematically rigorous approach for mapping the experimental biases to determine the application bias. As noted earlier, each experimental bias is heavily influenced not only by the true cross-section error vector but also by its own gradient norm. The adjustment procedure automatically accounts for the relative strength of each experiment's gradient when calculating the optimal cross-section adjustments, and moreover, employs the application's gradient to calculate the corresponding bias. This is fundamentally different from the three previous methodologies which employs the experimental bias directly as an application bias, lacking a formal approach to perform the needed mapping, leaving the analyst to heuristically add an additional margin to characterize lack of knowledge about the uncertainties resulting from the mapping. Finally, if the experiments employed are not relevant, TSURFER cannot guarantee that the application bias will improve the predictions, representing a key challenge for any inference technique that tries to reduce uncertainties with limited measurements.

References

[1] ANSI/ANS-8.1-1998, "Nuclear Criticality Safety in Operations with Fissionable Material Outside Reactors," La Grange Park, Illinois, 2007.

[2] ANSI/ANS-8.24-2017, "Validation of Neutron Transport Methods for Nuclear Criticality Safety Calculations," La Grange Park, Illinois, 2017.

[3] B. L. Broadhead, B. T. Rearden, C. M. Hopper, J. J. Wagschal, and C. V. Parks, "Sensitivity- and Uncertainty-Based Criticality Safety Validation Techniques," *Nucl. Sci. Eng*., vol. 146, no. 3, pp. 340–366, 2004, doi: 10.13182/NSE03-2.

[4] G. Palmiotti and M. Salvatores, "Developments in sensitivity methodologies and the validation of reactor physics calculations," *Sci. Technol. Nucl. Install*., vol. 2012, 2012, doi: 10.1155/2012/529623.

[5] E. Alhassan et al., "Benchmark Selection Methodology for Reactor Calculations and Nuclear Data Uncertainty Reduction," *Ann. Nucl*. *Energy*, 2015.

[6] B. T. Rearden, "Perturbation Theory Eigenvalue Sensitivity Analysis with Monte Carlo Techniques," *Nucl. Sci. Eng*., vol. 146, no. 3, pp. 367–382, 2004, doi: 10.13182/NSE03-03.

[7] H. Abdel-Khalik, D. Huang, U. Mertyurek, W. Marshall, and W. Wieselquist, "Overview of the Tolerance Limit Calculations with Application to TSURFER," *Energies*, vol. 14, no. 21, p. 7092, 2021, doi: 10.3390/en14217092.

[8] J. C. Dean and R. W. Tayloe, "Guide for Validation of Nuclear Criticality Safety Calculational Methodology, NUREG/CR-6698," 2001, [Online]. Available: https://www.nrc.gov/docs/ML0502/ML050250061.pdf

[9] H. Glaeser, "GRS method for uncertainty and sensitivity evaluation of code results and applications," *Sci. Technol. Nucl. Install*., vol. 2008, 2008, doi: 10.1155/2008/798901.

[10] M. N. Avramova and K. N. Ivanov, "Verification, validation and uncertainty quantification in multi-physics modeling for nuclear reactor design and safety analysis," *Prog. Nucl. Energy*, vol. 52, no. 7, pp. 601–614, 2010, doi: 10.1016/j.pnucene.2010.03.009.

[11] H. S. Abdel-Khalik, Y. Bang, and C. Wang, "Overview of hybrid subspace methods for uncertainty quantification, sensitivity analysis," *Ann. Nucl. Energy*, vol. 52, pp. 28–46, 2013, doi: 10.1016/j.anucene.2012.07.020.

[12] W. A. Wieselquist et al., *SAMPLER: A Module for Statistical Uncertainty Analysis with SCALE Sequences*, ver. 6.2.3., ORNL, ORNL/TM-2005/39, 2018.

[13] B. L. Broadhead, C. M. Hopper, and C. V Parks, "Sensitivity and Uncertainty Analyses Applied to Criticality Safety Validation: Illustrative Applications and Initial Guidance, NUREG/CR-6655," vol. 1 and 2, 1999, [Online]. Available: https://www.nrc.gov/docs/ML0037/ML003726890.pdf

[14] J. M. Scaglione, D. E. Mueller, J. . Wagner, and W. . Marshall, "An Approach for Validating Actinide and Fission Product Burnup Credit Criticality Safety Analyses — Criticality (k eff) Predictions," *Nucl. Tech*., vol. 188, no.1, pp. 266-279, 2014, doi: 10.13182/NT13-151.

[15] E. F. Trumble and K. D. Kimball, "Statistical Methods for Accurately Determing Criticality Code Bias," *in 1997 ANS topical meeting on criticality safety challenges in the next decade*, 1997.

[16] B. C. Kiedrowski et al., "Whisper: Sensitibvity/uncertainty-based computational methods and software for determining baseline upper subcritical limits," *Nucl. Sci. Eng*., vol. 181, no. 1, pp. 17–47, 2015, doi: 10.13182/NSE14-99.

[17] M. L. Williams, B. L. Broadhead, M. A. Jessee, J. J. Wagschal, and R. A. Lefebvre, *TSURFER: An Adjustment Code to Determine Biases and Uncertainties in Nuclear System Responses by Consolidating Differential Data and Benchmark Integral Experiments*, ver 6.2.3., ORNL, ORNL/TM-2005/39, 2018.

[18] D. Huang, J. Seo, S. Magdi, A. Badawi, and H. Abdel-Khalik, "Non-intrusive stochastic approach for nuclear cross-sections adjustment," *Ann. Nucl. Energy*, vol. 155, p. 108162, 2021, doi: 10.1016/j.anucene.2021.108162.

[19] C. M. Perfetti and B. T. Rearden, "Estimating Code Biases for Criticality Safety Applications with Few Relevant Benchmarks," *Nucl. Sci. and Eng*, vol. 193, no. 10. pp. 1090–1128, 2019. doi: 10.1080/00295639.2019.1604048.

[20] P. R. Bevington and D. K. Robinson, *Data Reduction And Error Analysis For The Physical Sciences*, 3rd ed, Boston, MA: McGraw-Hill Book Company, 2003, ISBN 0-07-247227-8.