ORNL/TM-2022/2772

Comparative Analysis of Confidence Metrics for Nuclear Criticality Safety

Hany Abdel-Khalik
Jeongwon Seo
Ugur Mertyurek
Goran Arbanas
William Marshall
William Wieselquist

August 2021–June 2022

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2022/2772

Nuclear Energy and Fuel Cycle Division

COMPARATIVE ANALYSIS OF CONFIDENCE METRICS FOR NUCLEAR CRITICALITY SAFETY

Hany Abdel-Khalik[*]
Jeongwon Seo*
Ugur Mertyurek[+]
Goran Arbanas[+]
William Marshall[+]
William Wieselquist[+]

[*]Purdue University
[+]Oak Ridge National Laboratory

August 2021–June 2022

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANSI	American National Standards Institute
CM	calculational margin
ENDF	Evaluated Nuclear Data File
EV	extreme value
GLLS	generalized linear least squares
LTL	lower tolerance limit
ME	modeling error
MOS	margin of subcriticality
NCD	noncovered deviation
ORNL	Oak Ridge National Laboratory
PCM	physics-guided coverage mapping
PDF	probability density function
SVD	singular value decomposition
USL	upper supercriticality limit

EXECUTIVE SUMMARY

Nuclear criticality safety standards provide guidance on the requirements and recommendations to establish confidence in computerized model results used to support operation with fissionable materials. By design, the guidance is not prescriptive, leaving the analysts free to determine how various sources of uncertainties are to be statistically aggregated. This report compares the analyses and key assumptions behind four notable methodologies documented in the nuclear criticality safety literature: the parametric, nonparametric, Whisper, and TSURFER methodologies. Because of the involved use of statistics entangled with heuristic recipes, the results of these methodologies are often difficult to interpret. Also, they are augmented by additional large administrative margins, eliminating the incentive to understand their differences. With the new resurgent wave of advanced nuclear systems focused on economizing operation—including advanced reactors, fuel cycles, and fuel concepts—there is a strong need to develop a clear understanding of uncertainties and their fusion methodologies to reduce uncertainties in a scientifically defensible manner. This report offers a deep dive into the various assumptions of the four noted methodologies, their adequacy, and their limitations, to provide guidance on developing confidence for the emergent nuclear systems. These systems are expected to be challenged by the scarcity of experimental data. To limit the scope, the report focuses on application of these methodologies to criticality safety experiments in which the goal is to calculate a bias, a bias uncertainty, and tolerance limit for k_{eff} to determine an upper subcriticality limit for eigenvalue calculations.

The main conclusions may be summarized as follows:

1. The parametric, nonparametric, and Whisper methodologies primarily rely on the subjective ability of the analyst to select experiments with biases of approximately equal magnitude to the unknown application bias. The use of similarity indices such as 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 magnitudes, a situation that blinds the similarity index. The implication is that all three methodologies could underpredict the true application bias if the norm of the application's sensitivity profile is larger in magnitude than that of the experiments

2. The nonparametric and Whisper methodologies are very sensitive to the experiment(s) with the highest bias and/or uncertainty, so the addition of similar experiments with low uncertainty does not improve confidence in the calculated application bias. The Whisper methodology 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*, and the sensitivity increases 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.

This report recommends the deployment of a new methodology called physics-guided coverage mapping (PCM), which is currently pending intellectual property protection with the US Patent Office, filed by Oak Ridge National Laboratory in September 2021. This methodology meets the following objectives:

1. The method should provide confidence in the calculated bias in a manner that is consistent with the extreme value theorem, e.g., 95/95 representing the coverage and its associated confidence.

Furthermore, the calculated confidence should be verified using numerical experiments with both real and manufactured data

2. Confidence should increase with experiments having high relevance and low measurements uncertainties and should not degrade with experiments having low relevance or high measurement uncertainties

3. The method should include a physics-based scaling algorithm to map biases from the experimental to the application domain to account for the response's sensitivities with respect to model parameters, e.g., cross sections

4. The method should be applicable to nonlinear response variations and non-Gaussian sources of uncertainties.

5. The method should hedge against modeling (i.e., uncharacterized errors) by decreasing the confidence in the calculated bias.

1. INTRODUCTION

Validation of the computer models used to analyze operation with fissionable materials such as spent nuclear fuel requires a scientifically defendable process by which confidence can be established in the model's predictions. This is paramount, because any model contains uncertainties that originate from modeling assumptions or numerical approximations, as well as uncertainties from its input parameters, which propagate to the model predictions. The regulator provides standards in the form of requirements and recommendations indicating on how confidence is to be demonstrated for predictive models. At the highest level of these standards is the golden rule, best captured by the famous Nobel Laureate Richard Feynman, paraphrased here as: "No matter how elegant a theory may look, it is wrong if it does not agree with experiments." This quote cements the dominant role experiments play in model validation, in which the regulator mandates that experimental evidence is fused with simulation results to develop confidence, implying that no confidence could be sought with simulation results only.

Because it is infeasible to develop experiments that cover all possible application conditions—the envisaged model use conditions—the regulator allows a licensee to design a criterion to select a finite set of experimental conditions considered sufficient to cover all application conditions. This criterion involves the use of a metric that measures the relevance[1] of an experiment to the application, with a prefect relevance score assigned to the application itself if employed as an experiment. If the relevance is high, which may require establishing another threshold criterion for what *high* is, then the discrepancy between model predictions and measurements could be applied with confidence, which also needs to be quantified, as a bias to the model predictions at the application conditions. As a simple example: if a highly relevant experiment shows that the code consistently under-predicts k_{eff} by 0.005 for a wide range of experimental conditions, then model predictions should be adjusted by a positive bias of 0.005, and the upper subcriticality limit (USL) should be lowered accordingly, implying that for this simple example, all code predictions above 0.995 would be considered supercritical.

In practice, no experiment has a perfect relevance score, so the analyst must determine another criterion to use for mapping measured discrepancies, referred to hereinafter as the *experimental biases*, from a finite set of experiments to the application conditions, having to incur additional bias to hedge against lower experimental relevance scores. Without a clearly explainable methodology to map biases from the experimental to the application conditions, the analyst must assign additional conservative margins, often done in a heuristic manner, to the experimental biases.

Furthermore, the experimental bias for a given response (e.g., critical eigenvalue) is not a constant value; instead, it is expected to assume a wide range of values because of the various sources of uncertainties in the calculated and the measured responses. This situation is depicted in Figure 1, which shows the measured response value y_m—measured eigenvalue—the corresponding calculated value y_c, the unknown true response value y_{true}, and y_{best} the best estimate after fusing measurements with predictions. The deviation between the true and measured value is attributed to experimental uncertainties, and the deviation between the true and predicted value is caused by uncertainties in the model.

These uncertainties originate from multiple sources which can be classified as either *reducible* (also referred to as *systematic* or *epistemic*) or *irreducible* (*aleatory* or *random*). Reducible uncertainties denote errors resulting from lack of knowledge. The true value of a model parameter is unknown, which can be corrected with additional measurements. Irreducible uncertainties denote inherent randomness that cannot be reduced with additional measurements or better models. This distinction is important because it allows

[1] As will be noted below, other terms have also been used, such as *similarity* and *representativity*.

the licensee to take credit for the reducible sources of uncertainties [1], especially when mapping uncertainties and biases from the experimental to the application conditions.

The distinction between reducible and irreducible uncertainties can sometimes be less clear and could lead to misleading results. For example, as discussed below, the fresh fuel enrichment in a benchmark model should be treated as an aleatory uncertainty, even if an exceptionally accurate measurement is designed to determine its value. This is because in practice the composition is subject to unavoidable random errors from the manufacturing process, rendering a single exceptionally accurate measurement to be nonrepresentative of the true aleatory uncertainty inherent in other fuel pellets.

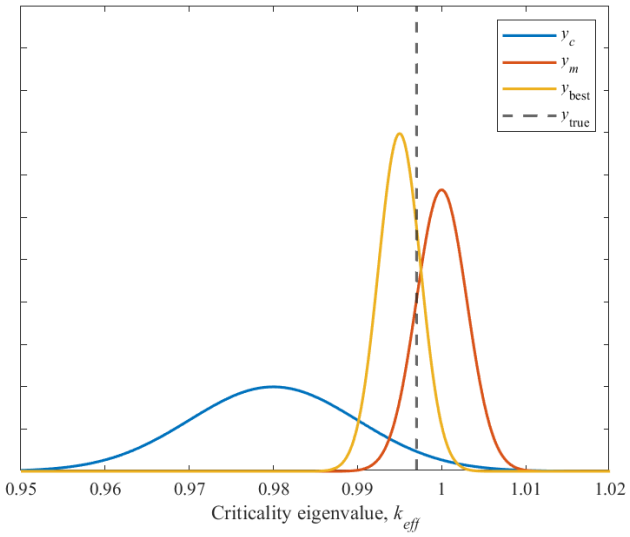


Figure 1. Eigenvalue prediction.

Once all experimental and simulation results have been successfully fused and corrected for the reducible (epistemic) sources of uncertainties, the next step is to quantify all possible remaining deviations between the measurements and the best-estimate code predictions. These deviations result from the irreducible uncertainties, as well as the epistemic uncertainties not covered by the available experiments. As discussed below, 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 fusion 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 this discussion, this PDF will be denoted hereinafter as the PDF of noncovered deviations (NCDs), or NCD PDF, where "noncoverage" implies that the deviations are not explained by the experiments.

As shown below, the mainstream statistical methods assume that the NCD PDF is normal, which reduces the inference problem to the estimation of two features: the mean and standard deviation. Furthermore, because a normal PDF theoretically stretches indefinitely in both directions,[2] the choice of a bias must be based on the selection of an upper limiting value, denoted by tolerance limit, that covers a preset portion of the PDF. If a PDF is perfectly known, then the upper[3] tolerance limit corresponding to a given

[2] This PDF has some of its mass in the positive range and some in the negative range. Negative values occur when the model predictions over-predict the true value of the eigenvalue, which is not interesting to criticality safety because they imply the code is already conservative, hence requiring no additional bias.
[3] The choice between the *upper* or *lower* descriptive depends on the sign of the bias. If the bias is defined as the predicted eigenvalue minus the measured one, then a lower tolerance limit is sought, and vice versa.

coverage p, say $p = 95\%$ can be estimated based on the knowledge of the features.[4] Applying this tolerance limit as a bias to model predictions guarantees that $p\%$ of all model predictions will over-predict the measured values, serving as a measure of confidence for code predictions. There is $(1 - p)\%$ chance that any future model prediction will under-predict the measured value. For the eigenvalue response, this tolerance limit serves as the basis for setting a USL on all code predictions. The USL is often supplemented by an additional administrative margin, as shown in Figure 2.

Although it is outside the scope of this report, it is important to note that the discussion above assumed that the NCD PDF is known exactly. In practice, the confidence is reported using a double hedging approach such as 99%/95%, denoting that with 99% confidence (i.e., not 100% as the previous discussion implies), one establishes that 95% of model predictions will over predict the true response values. This double hedging approach accounts for uncertainties in the estimated features: the standard deviation and the mean, which are calculated based on samples from the PDF [1].

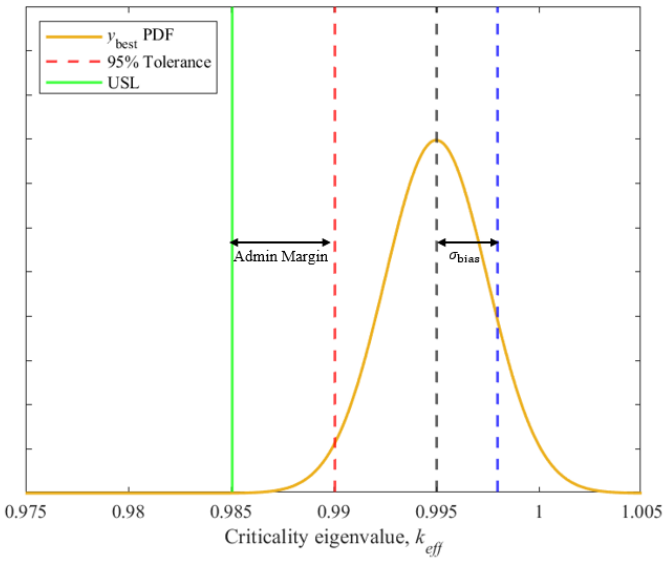


Figure 2. USL NCD PDF and bias tolerance limit.

1.1 UNCERTAINTY SOURCES

This subsection discusses the various sources of uncertainties that control the spread of the NCD PDF, including experimental, benchmark, and calculational uncertainties, as shown in Figure 3. The experimental uncertainties (also referred to as *measurement uncertainties*) originate from the unavoidable errors incurred during the measurement process resulting from the random (aleatory) nature of radiation detection instruments. These uncertainties could also manifest in the form of systematic (epistemic) errors in the experimental setup caused by factors such as equipment misalignment, errors in model specification, or poor calibration. Measurement uncertainty errors are designated as number *4* in Figure 3.

Similarly, benchmark uncertainties, assigned number *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 in which no two fuel pins are expected to have the same exact dimensions and enrichment. If the fuel is irradiated, then only the fuel composition and

[4] The same argument can be applied to other types of non-normal PDFs requiring additional features to be fully characterized, such as the use of higher order moments such as Kurtosis or skewness, for example. This discussion could be generalized to address non-normal PDFs, but this does not directly serve the scope of this work.

geometrical distortions can be estimated using other predictive modeling tools, all of which have their own epistemic and aleatory uncertainties. Alternatively, nondestructive measurement techniques can be used, but these also exhibit their own mixed sources of uncertainties. Furthermore, if the calculational model employed is probabilistic, such as Monte Carlo-based models, then the predicted value is expected to have another random error component that would manifest as an additional term similar in behavior to the random errors resulting from geometry and composition.

Although benchmark uncertainties may be considered errors resulting from the calculational procedure, as in composition and geometry parameters, they may be lumped with the measurement uncertainties for a number of reasons: (1) they cannot be controlled because of their aleatory nature, similar to other experimental conditions (e.g., ambient conditions), (2) they are much smaller than other sources of calculational uncertainties such as nuclear cross-section uncertainties because the benchmark models are carefully designed, and (3) they are independent of other experimental uncertainties. In the remainder of this report, benchmark and experimental uncertainties are denoted as evaluation uncertainties, as indicated in Figure 3.

Calculational uncertainties resulting from modeling assumptions, numerical approximations, and input model parameter uncertainties, all of which can be treated as epistemic. The focus of the current report will be on epistemic parameter uncertainties (assigned number *2* in Figure 3) only for two reasons:

- 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; these two sources are lumped together as solution uncertainties and are assigned number *1* in Figure 3.

- 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.

Before proceeding, it is important to note that while the cross-section model describing the basic interaction between neutrons and target nuclei is probabilistic in nature, cross-section uncertainties are treated as epistemic rather than aleatory. This is because the associated aleatory spread for any given cross section around its true mean value is much smaller in comparison to the deviation between the true mean value and the nominally reported cross-section value in the Evaluated Nuclear Data File (ENDF) [1].

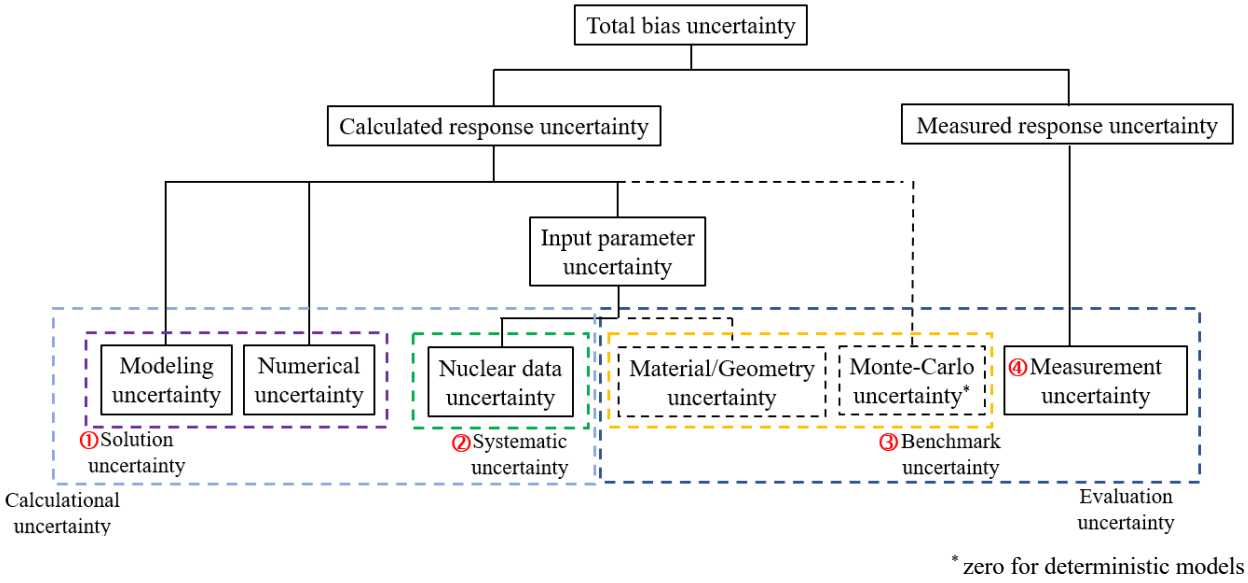


Figure 3. Uncertainty sources and classifications.

1.2 EXPERIMENTAL RELEVANCE

Experimental relevance is a key requirement for reducing uncertainties. As noted earlier, the regulator allows the licensee to seek an inference technique to reduce the impact of epistemic uncertainties on the calculated bias. If such inference is not completed, then it would be necessary to propagate the cross-section uncertainties, often resulting in a widely spread PDF for the calculated response with high standard deviation. However, reducing cross-section uncertainties is a challenging endeavor because the number of cross sections is much larger than the number of available experiments, forcing the inference problem to be under-determined: that is, the number of unknowns is larger than the number of equations. This implies that it is not possible to explicitly correct for all cross-section epistemic errors, which promotes the analyst to employ the concept of experimental relevance.

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 that are expected to be different for each experiment, as well as the application. This is depicted graphically in 2D (for two cross sections only) in Figure 4, where the blue arrow represents the unknown cross sections' epistemic error vector, the red arrow shows the gradient of the eigenvalue for a given experiment, and the yellow bar is the cross-section component that controls the bias. As illustrated below, the bias is simply the inner product between the gradient vector and the cross-section error vector. For illustration purposes, it is assumed that 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 n components are the first-order derivatives of a given response with respect to n cross-sections, thus representing the uncertain epistemic model parameters. Since the cross-section true error vector is unknown, only the impact on the response

can be assessed by analyzing all possible cross-section variations within their prior uncertainties, as described by a covariance matrix.[5]

If a single experiment is available,[6] then it is impossible to guess what bias should be applied to the application given the observed deviation between the experiment's measured and predicted response. This is demonstrated in Figure 5 for two scenarios, one in which the cross-section error causes zero error in the application response while causing a maximal value for the selected experimental response (Scenario I), and the second in which the errors for the application and experiment have approximately the same magnitude.

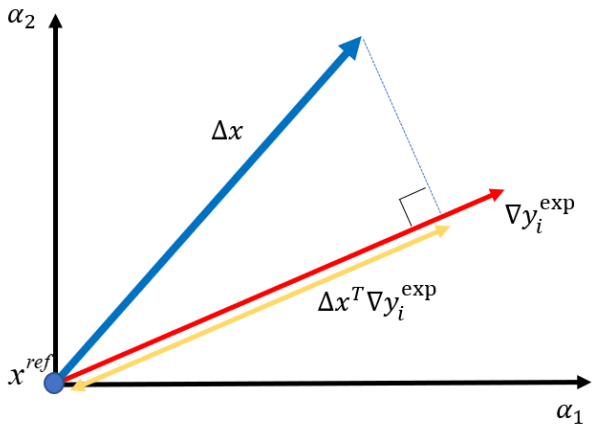


Figure 4. Illustration of bias calculation.

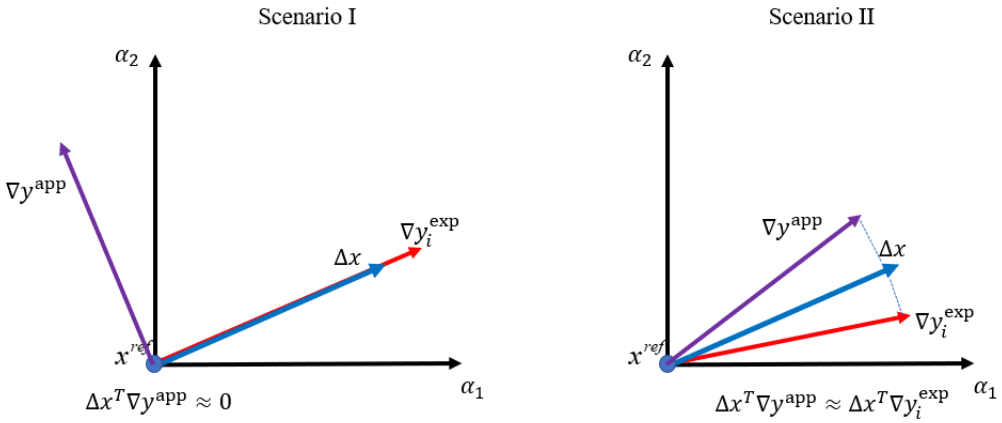


Figure 5. Plausible scenarios for calculating bias.

[5] This is a standard analysis referred to as *uncertainty propagation* or *quantification*, where cross-section variations are randomly sampled, and the code is executed, allowing for estimation of the standard deviation for the responses of interest. Deterministically, the gradient vector can also be used with the covariance matrix to directly calculate the response's standard deviation using the rule of quadrature error propagation, called the *sandwich equation* in the nuclear engineering literature.

[6] The discussion refers to the simple 2D case, which also emulates the real case, with experimental responses being much less than the number of uncertain cross sections.

Extending the idea to higher dimensions and maintaining that the number of cross sections is always much larger than the number of experiments, the resulting distribution of experimental biases (i.e., deviations between measured and predicted responses) 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 for which the biases are expected to be very close to the application bias (an example is demonstrated in Scenario II of Figure 5), which is possible with experimental gradients pointing approximately in the same direction as the application gradient,[7] thus representing the basic idea of similarity indices that are to measure experimental relevance. 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

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

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 Eqs. (1) and (2) is that each experiment allows the analyst to estimate the component of the cross-section error vector along the gradient of that experiment.[8] When the number of experiments is equal to or higher than the number of cross sections, it may be possible to correct for the entire cross-section error vector without knowing the application gradient, thus implying that the corrected cross sections can be used for any other application. This is possible when the gradients from all the experiments provide coverage for the entire cross-section space: that is, they have n independent components along the n dimensions of the cross-section space.

Mathematically, this is described as follows: given a matrix that aggregates as columns the gradients from all the experiments, its rank must be no less than n. If the rank r is less than n, then $n - r$ components of the cross-section error vector cannot be inferred from the experimental measurements. The r dimensional subspace spanned by the experimental gradients will be referred to hereinafter as the *experimentally covered subspace*, whereas the remaining $n - r$ dimensional subspace will be denoted as the *noncovered subspace*. This explains why experimental relevance is needed for realistic problems with extremely high dimensional parameter space (e.g., nuclear cross-sections) that is much higher than the number of available experiments. In this case, it is impossible to estimate the impact on the application if the experiments are not selected to be relevant. This is a key observation that will be recalled when discussing the parametric approach in the following section.

The concept of experimental relevance based on the use of gradients has been widely adopted in the neutronic community because the responses vary nearly linearly with cross-section variations within the

[7] This idea is valid as long as the experimental and application gradients have approximately the same magnitude which is not verified by the parametric, nonparametric, and Whisper methodologies, see later discussion.
[8] Mathematically, this is true if the cross-section covariance matrix is the Identity matrix. Appendix A shows how the same arguments apply in a transformed space, where the covariance matrix is the Identity matrix.

range of cross-section uncertainties. Extension of this idea for nonlinear dependencies is possible but is outside the scope of this report. Mathematically, the relevance c_k is described as follows:

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

where additional weighting by the prior covariance matrix $\boldsymbol{C_\alpha}$ is needed to (1) de-emphasize directions that have very low uncertainty and low sensitivity and (2) to emphasize directions with strong sensitivities and high uncertainties expected to impact the observed deviations between measurements and predictions.[9] The relevance score is sometimes referred to as the *similarity index* by US researchers or as *representativity factors* by European researchers.

Note that the relevance 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 or magnitudes. This must be considered when combining experimental biases to calculate the application bias, implying that simple averaging of biases from equally relevant experiments will be adequate only if the experiments have the same exact gradient norms. To our knowledge, this effect is not accounted for by three of the methodologies studied in this report: the parametric, nonparametric, and Whisper methodologies. The resulting impact will be assessed in the numerical section of this report.

Lastly, as noted above, it is infeasible to select 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. Figure 6 shows the relationship between the application gradient and the experimental gradients aggregating in a subspace. In this simple 3D example, the application has a component that is orthogonal to the experimentally covered subspace, so with a reliable inference technique, the cross-section error vector components along the covered subspace can be estimated, and the experiments contain no information about the error component along the non-covered subspace.

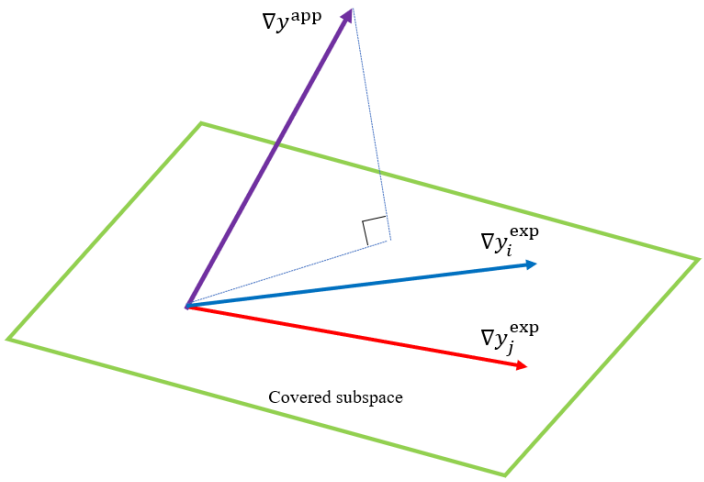


Figure 6. Relationship between application gradient and experimentally covered subspace.

[9] The discussion so far assumes the covariance matrix of the cross sections is the identity matrix; this is done intentionally to simplify the discussion. For the general case of a full covariance matrix, it can be shown that the same ideas apply, but in a transformed coordinate system, where the new coordinate system is indexed such that the lower indexed components are associated with both high sensitivity and high uncertainty. This is demonstrated in Appendix A.

With no experimental coverage for the orthogonal component, basic uncertainty propagation must be used to estimate the impact of cross-sections uncertainties on the responses of interest. This implies that whereas 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 noncovered subspace. To further reduce response uncertainties, additional experiments must be sensitive to new directions along the noncovered subspace. For realistic inference problems, the noncovered subspace is much higher in dimensionality than the covered subspace because of 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 noncovered cross-section subspace. This is another important observation that will be recall in future discussions of the various methodologies used for bias mapping from the experimental to application conditions.

1.3 USL CALCULATIONS

The discussion above sets the stage to introduce the four methodologies surveyed in this report for determining a code's USL: the parametric [2], nonparametric [2], Whisper [3], and TSURFER [4] methodologies. Before reviewing these methodologies, we recall the definitions of *bias*, *bias uncertainty*, *calculational margin* (CM), and *margin of subcriticality* (MOS) from American National Standards Institute (ANSI)/ANS-8.24-2017 [5]. The bias is defined as the systematic difference between the calculated k-eigenvalue and the benchmark k-effective (in our notations k_c and k_m, respectively), and the total bias uncertainty accounts for the combined effects of uncertainties in the calculated and measured responses as shown in Figure 3. The bias and bias uncertainty are employed to calculate the CM and MOS. 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 in criticality literature, *lower tolerance limit* (LTL) is closely related to the concept of CM; it is defined in terms of the bias PDF, which is 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 demonstrated later in the discussion, most methodologies define the CM in the same manner, so 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, and numerical and modeling errors. Furthermore, 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 noncovered epistemic uncertainties resulting from cross sections. In this report, it is assumed that the solution—the modeling and numerical uncertainties (assigned number *1* in Figure 3)—are treated separately via the MOS term.

The bias and bias uncertainties may be viewed as two fundamental quantities based on which CM and MOS can be calculated: the CM and MOS are functions of the bias and bias uncertainties, whose forms are not mandated by the standards but are left to the analyst to determine. As discussed above, a key component of model validation is to infer the NCD PDF 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 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*, in which the PDF is known, and the objective is to generate samples from the PDF. The

sampling and inference problems are the equivalents of the forward and inverse problems in applied mathematics [1].

Generally, the inference problem may be solved in two notably different approaches, the so-called *parametric* and *nonparametric approaches*. The parametric approach, as the name suggests, relies on knowing the PDF type, which allows the tolerance limit to be parametrized in terms of the PDF's features: 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 [1]. This represents the basic idea of the parametric approach, as well as the TSURFER methodology.

In the nonparametric 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 *extreme*) of the original PDF, implying that a majority of its samples would be higher than the 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 nonparametric approach.

Table 1 lists the sources of uncertainties (Figure 3) captured by the CM and MOS for each methodology. This table indicates that all methodologies employ CM to capture the epistemic cross-section 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 are given in later sections.

Table 1. Uncertainties employed for USL calculation

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

2. PARAMETRIC METHODOLOGY

Consider conducting N experiments, each with a different gradient vector, and assume that 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[10] 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

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

[10] In some renditions, the bias is standardized by the measured or calculated value, but this subtlety is discarded here, as it does not add much value to the discussion.

where

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

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

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

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

where

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

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

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 nonconservative bias adjustment parameter $\Delta_m = \max\{0, \beta_p\}$ is introduced to avoid nonconservative 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 Kiedrowski et al. [3], citing an Trumble and Kimball [2], which does not derive nor cite a statistical justification for these equations. Instead, they are listed without proof. The goal in this work is to explain the origin of these equations and to judge their adequacy for the bias calculation.

Starting with the mean bias equation, Eq. (4), the following observations can be made.

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: 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. However, this PDF reflects the distribution of experiments selected by the analyst; they are not random. If indeed the experiments are selected randomly with gradients that are randomly pointing in the cross-section space, then the resulting PDF will simply have a zero mean because 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 scores, then the resulting PDF will have a mean value 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 in which the mean is entirely noninforming about the true application bias, to maximally informing when all experiments have perfect relevance scores.

3. Assuming that all experiments have similar aleatory spread—$\sigma_{e_i} = \text{constant}$—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.[11] Thus, this averaging could have unpredictable results. Consider for example a situation in which the selected experiments have near perfect relevance scores to the application, but 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 larger than 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. The formula used for the standard deviation for the bias, Eq. (5), takes advantage of a famous theorem from statistics, the *variance decomposition theorem* or the *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 this 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). One then calculates the average of the variances of the individual datasets, denoted by the mean of the variances. It can be shown 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 the present context, each dataset represents the PDF of the bias from an individual experiment, and the superset is the ill-defined PDF (as discussed in observation #2) 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, then they will underestimate the true bias uncertainty for the application, and if they are very different, then they could overestimate the

[11] This can be easily assessed by analyzing the prior eigenvalue uncertainties for the various experiments, often showing a wide spread. Note that the propagated uncertainty provides a covariance-weighted metric of the gradient norm per the sandwich equation.

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. The objective of this formula is to calculate the average standard deviation as the inverse of the average confidence, which is different from direct calculations of the average variance.[12] 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 the objective is to estimate the mean value of a given distribution inferred from multiple samples from the distribution (see Appendix B for example). The definition in Eq. (7) resembles the Bayesian update formula but it contains an additional N factor. The work by Trumble and Kimball [2], in which this formula was originally proposed, does not provide a justification; although a classical textbook is cited [6] that does not contain this formula but instead contains the Bayesian update formula. The Bayesian formula is designed to increase confidence in the estimated mean as more samples are added. Equation (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-section uncertainties, thereby resulting in large enough bias uncertainty and raising 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. The pooled variance formula also helps to effectively capture the evaluation uncertainties.

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-section uncertainties. The resulting application bias and bias uncertainty will be smaller, rendering them under-conservative for the application conditions. With scarce relevant experiments—a situation that is common with first-of-a-kind nuclear systems—the licensor will require additional conservative margin which, from a licensee's perspective, limits design freedom and reduces system economy, a situation that is undesirable because it does not provide a venue for taking credit for the reducible 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 solution uncertainties (1) are captured under the MOS.

[12] This situation is encountered often when averaging quantities using different PDFs. For example, it is well-known that the kinetic energy calculated based on the average speed of a thermal neutron using a $p(v)$, a PDF of the neutron speed, is not mathematically equal to the average energy calculated from $p(E)$, a PDF of the neutron kinetic energy.

3. NONPARAMETRIC METHODOLOGY

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

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

The USL for the nonparametric methodology is

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

The basic nonparametric 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: it cannot be parametrized in terms of the PDF's features like the mean and standard deviation. The nonparametric approach solves this inference problem by employing a sampling-based approach to construct a related EV PDF.

To help illustrate the meaning of the EV PDF, we first graphically demonstrate how it may be constructed. Assume first that the original PDF type is known to be a gamma distribution,[13] but its features are unknown, and one is interested in estimating its upper tolerance limit corresponding to 95% coverage. The approach is to first select an order, such as *k*, which implies the need to generate *k* samples from the original PDF, and to take their extreme maximum value representing a single sample of the k^{th} order EV PDF.[14] This process may be repeated indefinitely with many samples until the k^{th} order EV PDF is formed, as shown in Figure 7. If the type of the original PDF distribution is known, then 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 order with more samples from the original PDF. If the goal is to find an upper tolerance limit for the original PDF (e.g., with 95% coverage, shown as the gray vertical dashed line), then an EV PDF with a mass is mostly above this limit should be found. Note that the word "mostly" is used because it is impossible to find an EV PDF whose entire mass is above the tolerance limit because the EV PDF is expected to have a tail stretching to the smallest values attained by the original PDF: negative infinity for a normal distribution. Thus, the minimum *k* must be found that renders a preset portion of the EV PDF above the sought tolerance limit. Assume for example that the k^{th} order EV PDF shown in purple has 3% of its area below the 95% upper tolerance limit (the gray dashed vertical line) for the original PDF. Therefore, if a single EV sample is generated (obtained by sampling *k* samples from the original PDF and taking the maximum), then there will be 97% chance that the EV sample will be higher than the 95% tolerance limit. Thus, it can be stated 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, confidence in the upper limit could be increased, never reaching 100%.

[13] A *gamma distribution* is defined by two parameters with long tail and positive values only, as shown in the curve in Figure 7.

[14] Although the literature defines "k^{th} order" as the statistics seeking k^{th} smallest or k^{th} largest (order) value of given PDF(s), in this report, "k^{th} order" means the extreme value of *k* samples from PDF(s).

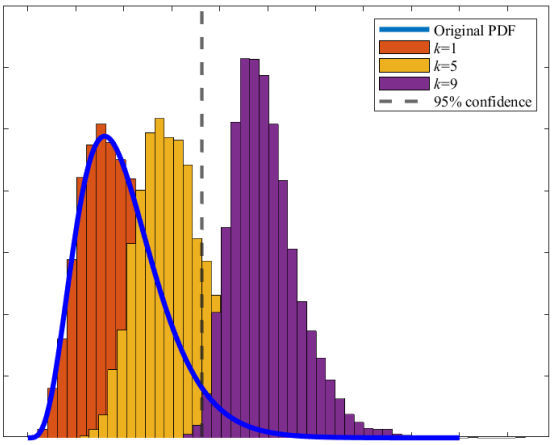


Figure 7. Extreme value statistics example.

This simple example can be easily generalized when the type of the original PDF is unknown. 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. Because 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 $p = 95\%$, a 95%/95% upper tolerance limit can be determined for any distribution as long as independent samples can be drawn from the same distribution. If the original PDF is unknown, then 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 nonparametric approach does not require estimation of the original PDF's features—the mean value and standard deviation for a normal distribution. Instead, the tolerance limit can be determined directly. Furthermore, if the features can be readily estimated, then it would be moot to attempt the nonparametric approach because there are known formulas and/or tables for determining the tolerance limit as a function of the features. If the nonparametric approach is used, then the same results would be obtained as with the parametric approach because in this case the type of the original PDF is known.

Also note that all samples must be independent and generated from 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 this case, this means generating samples from the same distribution. This requirement is important for ensuring (1) that the EV PDF samples can be related to the tolerance limit of the original PDF generating the samples, and (2) that the EV PDF progressively moves to the right with higher orders. For example, consider Figure 8, in which the objective is to generate the third 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, which is represented by the top two plots for PDFs with low overlap, the EV PDF will be heavily biased by the third PDF, the most extreme of the three, which essentially reduces to sampling only the third PDF. However, as the PDFs get closer to each other, as shown in the two bottom plots, the EV PDF begins to shift toward 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 in the discussion of the Whisper methodology below.

Next, we discuss how the nonparametric methodology has been applied for criticality safety applications. If the objective is to estimate $p\%$ upper tolerance with $q\%$ confidence for the application bias, then the goal is to determine N the minimum number of experiments. A straightforward application of nonparametric methodology must assume that there exists an unknown PDF from which biases are sampled, thus allowing the sought tolerance limit to be determined. If this assumption is acceptable, then the following equation can be solved for N: $q = 1 - p^N$. However, this assumption is problematic, as discussed above, because this hypothetical PDF is ill defined. As discussed in the previous section, there is no PDF that describes all possible experiments because the analyst must select which experiments to include and determine how relevant they are to the application. The situation is different from its common use in other engineering applications such as manufacturing, in which a tolerance limit is estimated 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%. Because of the inherent uncertainties in the process, the fuel pellet enrichments are expected to have a PDF with a mean value of 4% and some spread. The distribution of the fuel pellet enrichment describes a PDF for which a tolerance limit can be calculated by sampling N pellets.

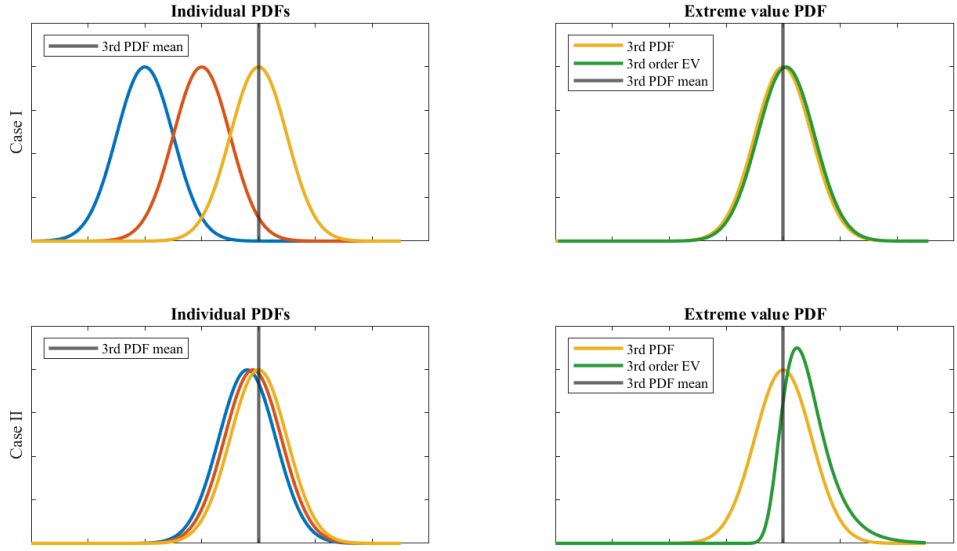


Figure 8. Extreme value statistics with different PDFs.

Applying the nonparametric methodology to samples that are not iid (i.e., generated from different experiments) challenges the basic assumption of the EV PDF. For example, assume that the analyst selects experiments with relevance scores above a minimum threshold such as 0.85. In this case, the spread of the resulting PDF will be determined by the spread of the relevance score, as well as the spread of the norms of the experimental gradients: see the earlier discussion of Eq. (1). Thus, for this approach to be effective, the norm of the application gradient must be in the same order of magnitude as that of the experiments. If the norm 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, then the resulting PDF would be wider, thus conservatively impacting the calculated tolerance for the application bias. In response to this nonstandard use of the EV theorem, the nonparametric 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: the error term $\varrho\sigma_p$ which appears in Eq. (10).

The basic nonparametric 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, the analyst must ensure that the application gradient is similar in magnitude to the experiments, which is possible with expert judgment. If an experimental relevance score could be employed, then the calculated tolerance would be closer to the true value for the application bias, thus allowing the analyst to drop the additional conservative term $\varrho\sigma_p$. If no knowledge about the application is included, then 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, this approach does not allow the analyst to take credit for the irreducible sources of uncertainties.

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

4. WHISPER METHODOLOGY

The Whisper methodology was developed by Los Alamos National Laboratory researchers [3]. It is promoted to provide the following features:

1. it hybridizes the use of parametric and nonparametric methodologies
2. It relies on the concept of EV theorem and uses calculated tolerance to set the CM.
3. 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.
4. It employs TSURFER-based approach to determine the noncovered uncertainties used to set the MOS.

The full implementation may be found in the paper by Kiedrowski et al.[3], but a brief overview of the steps is given here. First, the methodology generates an EV-like PDF which is used to calculate a tolerance value m, covering preset area q, for example, 95%, under the PDF. Whisper's EV-like PDF and the associated tolerance are explained below. 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: margin for software error (in our notation, the solution uncertainties, source 1), margin for the noncovered nuclear cross-section uncertainties, and margin for the application.[15] The expert opinion is that the margin for software is set to be 0.005, and the margin for non covered cross sections is calculated by the generalized linear least squares (GLLS) methodology, such that

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

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

[15] The application margin is purely heuristic and is not statistically justified in the Whisper methodology.

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

Whisper, which is markedly different from the basic nonparametric methodology, generates an EV-like PDF using samples that are not iid, because the samples are generated from different PDFs. Each PDF represents one experiment, with the PDF assumed to be known: in the normal case, the experimental bias sets the PDF's mean value, and the evaluation uncertainty sets the PDF's standard deviation. Then the methodology calculates an EV-like PDF of N^{th} order, with N being the number of experiments. Because the original PDFs are fully characterized, Whisper explicitly constructs the N^{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.[16] Finally, it defines the tolerance limit m as the value that covers a preset portion of the EV-like PDF.

Whisper employs a linear heuristic methodology to diminish the number of samples generated from low-relevance experiments by assigning a weight w that varies linearly with the relevance score. For each experiment, only $w\%$ of the generated samples are used to construct the EV-like PDF. To understand what this means, consider one of the low relevance experiments with $w = 0.1$, for example, and consider that a total number of one million samples was generated from its associated PDF. Of these one million samples, only 50,000 samples were generated in the tail end of the PDF. These 50,000 samples will have the most impact on the EVs generated from all the experiments. With $w = 0.1$, Whisper reduces these samples to 5,000. The premise is that the 5,000 samples will contribute less to the tail end of the EV-like PDF, hence allowing the impact of that low relevance experiment on the calculated tolerance limit to be reduced. A numerical experiment is presented to demonstrate the effectiveness of this premise.

The linear heuristic methodology also limits 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, thus raising the tolerance, as seen in Figure 8. This is counterintuitive, because one should develop higher confidence in the bias when an increasing number of experiments provides similar bias results—a basic premise of any statistical inference methodology. Whisper sets a maximum threshold on the weights to ensure that the calculated tolerance does not increase indefinitely with the number of experiments. The selected function for the required weight is

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

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

[16] The generalization to non-normal PDFs is not relevant, so the discussion will focus on normal PDFs only.

$$w_i = \max\left\{0, \frac{c_{k,i} - c_{k,acc}}{c_{k,max} - c_{k,acc}}\right\}. \quad (17)$$

A numerical experiment is conducted to explain the impact of this weighting procedure.

Finally, Whisper performs a TSURFER-based approach to calculate an estimate of the application bias and bias uncertainty, and it employs the bias uncertainty to calculate the residual uncertainty in the application response which is used to set the MOS. The residual response uncertainty represents the part that is resulting from the noncovered cross-section subspace, as discussed above in Section 0. Because Whisper's MOS calculation is very similar to that of TSURFER, this discussion will be deferred to the TSURFER methodology section, and the current discussion will be limited to the CM calculations. The objective is to report as MOS the irreducible uncertainties from both the noncovered cross-section subspace, as well as the evaluation uncertainties.

The following observations are made about the Whisper methodology.

1. The nonparametric methodology's real power is that it can create an EV PDF that progressively moves toward 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 toward the tail end of the original PDF. As demonstrated above, this logic does not apply when samples are taken from different PDFs, thus losing the ability to compare samples from the same PDF. If the PDFs have low overlap (see Figure 8), then the EVs will be dominated by the PDF with the highest values. 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 with a mass concentrated above an upper tolerance limit that is already fixed, albeit unknown by the original PDF. If the original PDF were known, then there would be no need to calculate an EV PDF because the tolerance limit would be fully determined by the original PDF. The nonparametric approach allows for estimation of 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 blue PDF in Figure 9, and another with lower relevance and higher average bias. Each experiment represents a collection of closely grouped experiments with approximately the same bias and spread. For simplicity, assume that the weights for the two groups of experiments are 1.0 and 0.5, respectively. With one million samples generated, the EV PDF will have 50% of its samples generated from the high-relevance PDF(s), and the other 50% will be generated 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 approximately 5%. The area above the same bar under the Whisper-weighted PDF will be slightly lower than 5%. Therefore, 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. Numerical results are presented to show the impact of the weighting on moving the tolerance limit.

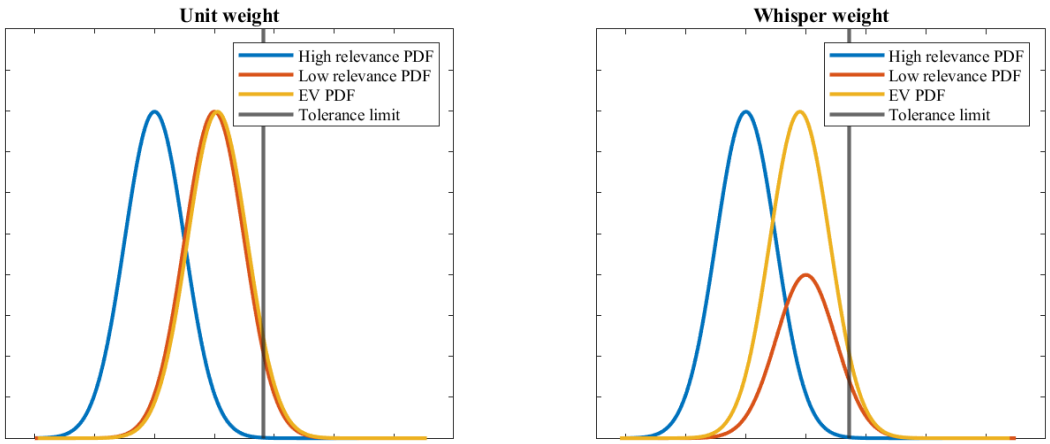


Figure 9. Impact of relevance on tolerance limit.

4. If one employs experiments with perfect (or very high) relevance scores, then the differences in their biases will be mainly determined by the magnitude of their gradients. As explained above, if all the experiments have the same relevance but different biases, then 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, then 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 as based on the analyst's best judgment of selecting experiments with sensitivities of the same magnitude as those of the application.

5. Assuming that 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-section uncertainties only. Thus, if one conducts the same experiment twice, with one being of unreasonably high uncertainty, then 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.

6. The MOS quantifies the irreducible, noncovered, uncertainties via a TSURFER-based cross-section adjustment procedure. This procedure, as discussed in the next section, employs the aleatory evaluation uncertainties, implying that the residual uncertainties will be contaminated by the evaluation uncertainties which are already accounted for in the CM definition. If the evaluation uncertainties are significantly smaller than cross-sections uncertainties, then this double-counting will not noticeably impact the calculated tolerance limit.

7. Finally, perhaps the most complex argument against using experimental biases directly without a proper mapping analysis, which will be elaborated on in the next section, is that the measured biases do not guarantee they will over-predict the true application bias. Because the biases have wide variability, in practical scenarios, any EV-based approach will hedge against this problem by selecting the maximum bias value, which is effectively the case with both the basic nonparametric approach and the Whisper methodology, as demonstrated in the numerical section. While this EV-

based approach is sufficiently conservative from a safety viewpoint, it does not allow the licensee to take credit for the reducible uncertainties.

8. The TSURFER-based adjustment performed by Whisper is not intended to reduce uncertainties; instead, it is calculated to determine the MOS term which hedges against the irreducible uncertainties from both cross sections, as well as aleatory evaluation uncertainties.

Whisper employs CM to account for the systematic bias from cross-section 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 nonparametric methodology, it can be confidently argued that Whisper 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 irreducible evaluation uncertainties, sources (3) and (4), and the noncovered cross-section 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 account for a portion of these sources which were already accounted for in the CM. While acceptable from a safety point of view, this double counting cannot be traced to a statistical justification.

5. TSURFER METHODOLOGY

The TSURFER methodology is parametric; it assumes a parametric 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 reducible uncertainties by solving a mathematical minimization problem to find optimal adjustments for the cross sections. The premise is that one can correct for the cross-section errors that belong to the covered subspace. The residual uncertainties resulting from the noncovered 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 document by Williams et al. [4], and a brief discussion on the GLLS procedure can also be found in Appendix C.

The CM or LTL for the TSURFER methodology can be described by the bias β_T and its uncertainty $\sigma_{k'}$ which are evaluated by the GLLS procedure such that

$$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'}, \tag{18}$$

And the USL is calculated with the MOS by

$$\begin{aligned} \mathrm{USL}_T &= 1 - \mathrm{CM}_T - \mathrm{MOS}_T \\ &= 1 + \beta_T - \varrho\sigma_{k'} - 0.005. \end{aligned} \tag{19}$$

Because both TSURFER and Whisper employ the idea of cross-section adjustments, albeit for different goals, a more detailed account of how this is achieved is presented here. The focus is not on exposing the

mathematical details of the minimization problem; instead, 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, the lack of a formal verification procedure for the calculated application bias, and the impact of modeling errors. This report dives deeply into the mechanics of cross-section adjustments to provide the insight needed to guide future work focused on first-of-a-kind nuclear systems. This presentation is novel and provides one of the key contributions of this work: 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 above, the regulatory process does not mandate a specific procedure to perform model validation, but it does require 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. This represents the basic idea behind correcting for the reducible sources of uncertainties.

Cross sections' prior uncertainties are typically high and incomplete, resulting in high uncertainties for the quantities of interest, such as eigenvalue. However, the experiments 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 very high, it is infeasible to build experiments that can be used to correct for all sources of cross-section 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. However, this is 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-posed inverse problem presents a formidable challenge for the optimization search, resulting in the so-called *error compensation phenomenon* in which the cross sections are incorrectly over- or under-adjusted as compared to their true unknown errors, leading to the same responses residual, the post-adjustment deviations between measured and predicted responses.

Therefore, it is possible to find a theoretically infinite number of cross-section adjustments that give rise to the same level of agreement between measured and predicted responses, making it difficult to determine which adjustments would be applicable to the application. This is because the application model is not included in the determination of the optimal adjustments, so one cannot know a priori whether the adjusted cross-sections would improve or deteriorate the model predictions for the application conditions. Although regularization techniques have been developed to render the optimization search well-posed, achieved by selecting the adjustments that render a unique solution by enforcing some mathematical criterion, such as the minimum distance from the best-known prior cross-section values, these regularization techniques are blind to the application conditions and hence cannot guarantee that the adjustments will improve the model predictions for the application conditions.

From a theoretical perspective, although the adjustment procedure is blind to the application conditions, it guarantees that within the limit of infinite measurements, the adjustments will converge to the true cross-section errors. However, this guarantee is not practically valuable because realistically, even with a rich experimental program, the number of experimental responses is very small compared to the number of uncertain cross sections. To overcome this fundamental difficulty, engineering practitioners require that the adjustment procedure employ only experiments that are judged to be relevant to the application. High relevance implies that both the experiment and the application have a similar dependence on the uncertain cross sections. To explain this, consider two responses y^{app} and y^{exp}, representing respectively the responses, such as eigenvalues, of the application and a single experiment. Both depend on three common cross sections x_1, x_2, and x_3 with equal prior uncertainties.[17] A linear-based relevance requires that the two responses' first order derivatives satisfy the following approximate relationship:

$$\begin{bmatrix} \frac{\partial y^{\text{app}}}{\partial x_1} \\ \frac{\partial y^{\text{app}}}{\partial x_2} \\ \frac{\partial y^{\text{app}}}{\partial x_3} \end{bmatrix} \approx \alpha \begin{bmatrix} \frac{\partial y^{\text{exp}}}{\partial x_1} \\ \frac{\partial y^{\text{exp}}}{\partial x_2} \\ \frac{\partial y^{\text{exp}}}{\partial x_3} \end{bmatrix} \rightarrow \nabla y^{\text{app}} \approx \alpha \nabla y^{\text{exp}} \tag{20}$$

This expression implies that the relative ratio of any two cross-section sensitivities is approximately the same for both the experiment and the application. Mathematically, this means that the gradient of the application ∇y^{app} and experiment ∇y^{exp} are approximately pointing in the same direction. If this relationship is satisfied, then the cross-section variations will have a similar impact on both the application and experimental responses, so the discrepancies observed in the experimental domain could be mapped with confidence to the application domain. To illustrate this further, one can plot the expected variations of the application and experimental responses for a wide range of cross-section variations. Each sample is denoted by a lower bracketed subscript (i), as follows:[18]

$$\Delta y_{(i)}^{\text{app}} \approx \frac{\partial y_{(i)}^{\text{app}}}{\partial x_1} \Delta x_{1,(i)} + \frac{\partial y_{(i)}^{\text{app}}}{\partial x_2} \Delta x_{2,(i)} + \frac{\partial y_{(i)}^{\text{app}}}{\partial x_3} \Delta x_{3,(i)} = \nabla y_{(i)}^{\text{app}\,T} \Delta x_{(i)} \tag{21}$$

$$\Delta y_{(i)}^{\text{exp}} \approx \frac{\partial y_{(i)}^{\text{exp}}}{\partial x_1} \Delta x_{1,(i)} + \frac{\partial y_{(i)}^{\text{exp}}}{\partial x_2} \Delta x_{2,(i)} + \frac{\partial y_{(i)}^{\text{exp}}}{\partial x_3} \Delta x_{3,(i)} = \nabla y_{(i)}^{\text{exp}\,T} \Delta x_{(i)}, \tag{22}$$

where Δx is a three-component vector, a compact representation of the three cross-section variations. In this case, the variations in both responses can be expected to be highly correlated, reaching perfect correlation if the linearity assumption is satisfied.

Returning to the ill-posed inverse problem solution, assume that it produces cross-section adjustments that satisfy the following equation:

[17] The general case of nonequal correlated uncertainties is considered in this report but omitted here to simplify the introductory remarks.

[18] To avoid confusion with earlier discussion, brackets are used for the subscript to denote sampled calculational values obtained by running the simulation with different perturbations. In an earlier discussion, the subscript was used to denote an experiment.

$$y_m^{\exp} - y_c^{\exp} \cong \frac{\partial y^{\exp}}{\partial x_1}\Delta x_1^a + \frac{\partial y^{\exp}}{\partial x_2}\Delta x_2^a + \frac{\partial y^{\exp}}{\partial x_3}\Delta x_3^a = \frac{\partial y^{\exp}}{\partial x_1}\Delta x_1^t + \frac{\partial y^{\exp}}{\partial x_2}\Delta x_2^t + \frac{\partial y^{\exp}}{\partial x_3}\Delta x_3^t, \quad (23)$$

or in vector notations

$$y_m^{\exp} - y_c^{\exp} \cong \nabla y^{\exp T}\Delta x^a = \nabla y^{\exp T}\Delta x^t, \quad (24)$$

where $y_m^{\exp}$ is the measured value for the response, $y_c^{\exp}$ is the corresponding predicted value with no adjustments using the reference cross-section values, Δx_i^a is the adjusted value for the i^{th} cross section, and Δx_i^t is the true error. This equation implies that the adjustment procedure was able to find cross-section adjustments Δx^a to explain the observed deviations between the reference and measured responses, which are attributed to the true unknown errors Δx^t. Because of the error-compensation phenomenon, these adjustments are not exactly equal to the true parameter errors. The implication is that the adjusted cross sections will introduce a residual error, $\Delta x^e = \Delta x^t - \Delta x^a$, denoted hereinafter as the adjusted cross-section residual error vector. This residual error vector belongs to the noncovered subspace defined above.

This situation is depicted in Figure 10 for two cases with different true cross-section errors Δx^t. In both cases, x^{ref} denotes the reference values for the cross sections, also referred to as *prior values*, x^t is the unknown true values, and x^a is the adjusted values. The Δx^a denotes the cross-section adjustments, and Δx^e is the adjusted cross-sections residual errors. Note that the two gradient vectors for the experiment and application are not pointing in the same direction because it is difficult to construct an experiment that is an exact duplicate of the application. In this simple example, one can show that Δx^a will be pointing in the same direction as the $\nabla y^{\exp}$ when all parameters have the same prior uncertainty, a special case used here for illustration. The horizontal blue double arrow denotes the true error in the application response, assuming a normalization of $\nabla y^{\text{app}T}\nabla y^{\text{app}} = 1$. The orange line segment denotes the adjustment of the reference application response, ideally to be the same as the blue double-arrow, leaving the red line segment as the residual true error for the application response, obtained with the adjusted cross sections. In a perfect adjustment scenario, the red line segment should ideally go to zero. This does not imply that Δx^e is zero, which occurs only if the application gradient is orthogonal to Δx^e. This is intuitively meaningful because all cross-section errors that are orthogonal to the experiment gradient will have no impact on its predicted value, so it is not important to capture them. Thus, the concern of any adjustment procedure with limited measurements should be focused on how much of an impact Δx^e has on the application's response of interest, which is given by $\nabla y^{\text{app}T}\Delta x^e$. Note that $\nabla y^{\text{app}T}\Delta x^t$ describes the true pre-adjustment error in the application response, which is unknown, because no experimental measurements have been made for the application. However, the goal of model validation is to ensure that this discrepancy can be bounded statistically based on the available measurements.

Consider now the two different cases, each with different true cross-section errors: in Case A the errors cause a small error in the application response, whereas in Case B, the cross-section errors have a larger impact on the application response, as visually compared by the length of the blue double arrows. In Case A, the Δx^e has a significant impact on the application's response, causing its post-adjustment predictions to be inferior to the reference predictions (the red line segment is longer than the blue double arrow). In Case B, the impact is much smaller, allowing the adjustments (orange double arrow) to explain the majority of the error in the reference prediction. It is impossible to know which scenario is at play

because the true cross-section errors are unknown. The implication is that the adjustment procedure does not allow one to know whether the adjusted parameters will indeed improve the predictions of the application response. As noted earlier, in the limit of infinite measurements, the true cross-section errors can be captured, and this becomes a non-issue.

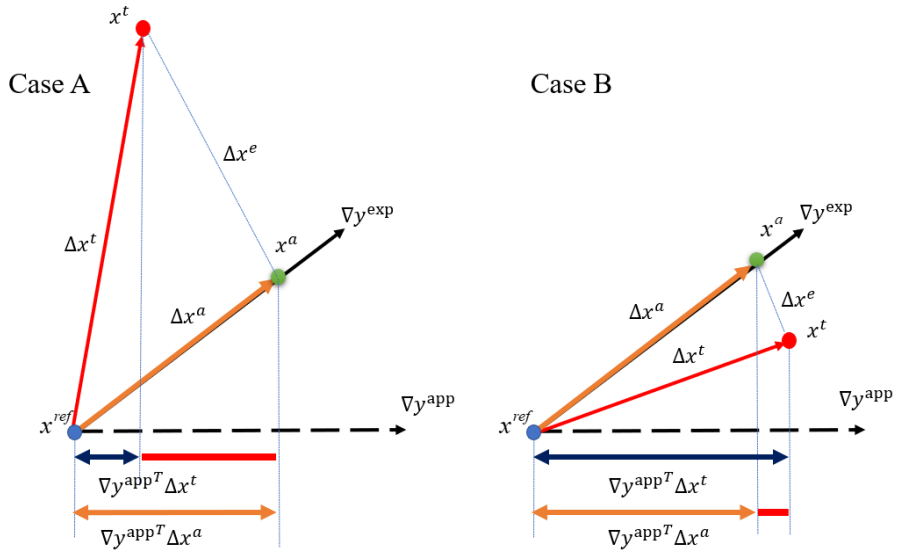


Figure 10. Gradient-based adjustment and error.

It can be observed that Δx^e has no impact on the experimental responses because it is orthogonal to the gradient. One can easily show that

$$\nabla y^{\exp^T}\Delta x^e = \frac{\partial y^{\exp}}{\partial x_1}\Delta x_1^e + \frac{\partial y^{\exp}}{\partial x_2}\Delta x_2^e + \frac{\partial y^{\exp}}{\partial x_3}\Delta x_3^e \cong 0, \tag{25}$$

meaning that Δx^e has little or no impact (with linearity) on the estimated experimental response. This describes the key challenge of any adjustment technique that is a low or zero residual for the experimental responses—small deviations between measured and post-adjustment predicted responses—does not guarantee an improved prediction for the application response. This is because the true cross-section errors are unknown, and the adjustment procedure is blind to the application conditions. If the experiment has a high relevance score in the sense of Eq. (3), then the adjusted cross-section residual errors Δx^e will also have little or no impact on the application response. This represents the basic idea behind experiment(s) selection via relevance measures as given by Eq. (3).

In practice, the experiment is never exactly representative of the application, implying that the application gradient ∇y^{app} will always have a non-zero component that is orthogonal to that of the experiment gradient $\nabla y^{\exp}$ (Figure 6):

$$\nabla y^{\text{app}} = \nabla y^{\text{app}\,\parallel} + \nabla y^{\text{app}\,\perp}, \text{ such that } \cos(\nabla y^{\text{app}\,\parallel}, \nabla y^{\text{app}\,\perp}) = 0 \text{ and } \cos(\nabla y^{\text{app}\,\parallel}, \nabla y^{\exp}) = 1.0. \tag{26}$$

Hence,

$$\nabla y^{\text{app}^T}\Delta x^e = \nabla y^{\text{app}\,\parallel^T}\Delta x^e + \nabla y^{\text{app}\,\perp^T}\Delta x^e = \nabla y^{\text{app}\,\perp^T}\Delta x^e. \tag{27}$$

This equation describes the core idea behind the concept of relevance, that is, with Δx^e being orthogonal to the experiment(s) gradient, i.e., $\nabla y^{\exp T} \Delta x^e = 0 = \nabla y^{\text{app}\,\parallel T} \Delta x^e$, the actual adjustments Δx^a are restricted to the same subspace spanned by the experiment(s) gradients, denoted by the covered subspace. In fact, for the general case of *N* experiments[19], one can show that

$$\Delta x^a = \beta_1 \nabla y_1^{\exp} + \beta_2 \nabla y_2^{\exp} + \cdots + \beta_N \nabla y_N^{\exp}. \tag{28}$$

This equation implies that each experiment provides information about the true cross-section errors' vector along its own gradient.[20] If one has enough experiments with *n* independent gradients covering the *n*-dimensional cross-section space, then the adjustment procedure will be able to recover the true cross-section errors. However, this is not the case, as explained earlier, so at best, the experiments would be able to recover information about *N* directions, assuming all experiments have *N* independent gradients. The remaining *n*-*N* directions, forming the non-covered subspace, will be contaminated by the residual errors vector Δx^e, whose impact on the application response, i.e., $\nabla y^{\text{app}\,\perp T} \Delta x^e$, remains unknown. To hedge against this, the prior cross-section uncertainties belonging to the noncovered subspace are propagated to the response, serving to set the tolerance limit for the residual noncovered uncertainties.

If the application's gradient belongs to the subspace spanned by the experiment(s) gradients,

$$\nabla y^{\text{app}} = \sum_{i=1}^{N} \eta_i \nabla y_i^{\exp}, \tag{29}$$

then $\nabla y^{\text{app}\,T} \Delta x^e = 0$, implying the adjusted parameters residual errors will have little or no impact on the application response. This means that cross-section uncertainties along the noncovered subspace will have little or no impact on the application bias.

Unlike the previous three methodologies, the TSURFER methodology takes credit for the reducible 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, it employs the application's gradient to calculate the corresponding bias. This is fundamentally different from the three previous methodologies, which employ the experimental bias directly as an application bias, lacking a formal approach to perform the needed mapping, and leaving the analyst to heuristically add margin to characterize lack of knowledge about the uncertainties resulting from the mapping. Finally, if the experiments employed are not relevant, then TSURFER cannot guarantee that the application bias will improve the predictions, representing a key challenge for any inference technique to reduce uncertainties with limited measurements.

[19] This assumes that parameters have equal and independent prior uncertainties; the covariance matrix is proportional to the identity matrix. For the sake of the current discussion, this is not an important distinction. For the general case, the covariance matrix for the parameters is used to weigh the sensitivities (Appendix C).
[20] For the general case with nonidentity covariance matrix, the gradient is transformed by the Chelosky factor of the covariance matrix (Appendix A).

6. NUMERICAL EXPERIMENTS

To help compare the various methodologies, several numerical experiments are conducted. The first is based on a toy model in which the true bias values are known, allowing the mechanics of the various methodologies to be revealed and the adequacy of their assumptions to be tested. Next, a second experiment is conducted in a similar manner using real nuclear criticality benchmark models. Then, a validation approach is employed to compare the four methodologies using a large suite of benchmark models. This is followed by an analytical benchmark experiment prepared by International Working Sub-Group 11. The experiment is used to analyze the impact of c_k sorting on the various methodologies and the underdetermined nature of the Bayesian inference analysis. Finally, an initial demonstration of the PCM methodology is provided to hedge against modeling errors for the TSURFER methodology.

6.1 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, and possible Monte Carlo calculational uncertainties. Because 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

$$k_m = a^T x + k_c + \epsilon, \tag{30}$$

or in matrix form for 40 different experiments,

$$\begin{bmatrix} k_{m_1} \\ k_{m_2} \\ \vdots \\ k_{m_{40}} \end{bmatrix} = \begin{bmatrix} a_1^{(1)} x^{(1)} + a_1^{(2)} x^{(2)} + k_{c_1} + \epsilon_1 \\ a_2^{(1)} x^{(1)} + a_2^{(2)} x^{(2)} + k_{c_2} + \epsilon_2 \\ \vdots \\ a_{40}^{(1)} x^{(1)} + a_{40}^{(2)} x^{(2)} + k_{c_{40}} + \epsilon_{40} \end{bmatrix}, \tag{31}$$

where k_{m_i} and k_{c_i} represent 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, with a correlation matrix $R \in \mathbb{R}^{2\times 2}$ selected as

$$R = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix}, \tag{32}$$

where the off-diagonal terms imply a positive correlation, as shown in Figure 11. 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: the gradient of the response with respect to the input parameters.

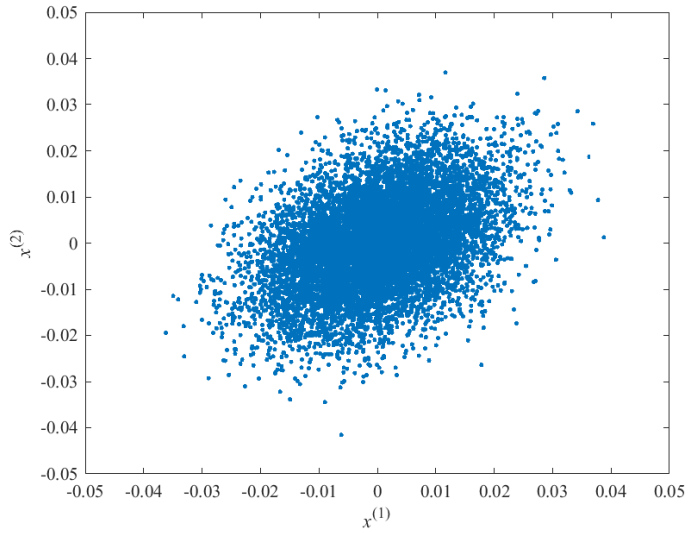


Figure 11. Toy model parameters prior uncertainty.

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 in which the true parameter values are used to generate the mean value of the measurements. Specifically, the true values for the parameters $x_{\text{true}} = \left[x_{\text{true}}^{(1)} \quad x_{\text{true}}^{(2)}\right] = [0.0084 \quad -0.0021]$ are selected from the pool of random samples shown in Figure 11, and the reference values of the experiments, k_{c_i}, are back-calculated as

$$k_{c_i} = 1.0 - a_i^{(1)} x_{\text{true}}^{(1)} + a_i^{(2)} x_{\text{true}}^{(2)}. \tag{33}$$

The application response's calculated value is modeled as

$$k_c^{\text{app}} = 0.9856 + 1.6151 x^{(1)} - 0.4038 x^{(2)}. \tag{34}$$

This results in an estimated value of 1.0 using the true values of the input parameters, and it produces a response uncertainty of 1,500 pcm. This model yields a true bias of 1,440 pcm, meaning that the model underestimates the true value of k_{eff} by approximately one standard deviation of the prior uncertainty.

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

$$\sigma_{e_i} = \sqrt{\sigma_{\epsilon_i}^2 + \sigma_{m_i}^2} = 250 \text{ pcm}. \tag{35}$$

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 equivalent to the propagated cross-section uncertainty),

$$\sigma_{s_i} = \sqrt{\left(a_i^{(1)}\sigma_{x^{(1)}}\right)^2 + \left(a_i^{(2)}\sigma_{x^{(2)}}\right)^2 + 2a_i^{(1)}a_i^{(2)}\mathrm{Cov}\left(\sigma_{x^{(1)}}, \sigma_{x^{(2)}}\right)}. \quad (36)$$

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 illustrated in Figure 12.

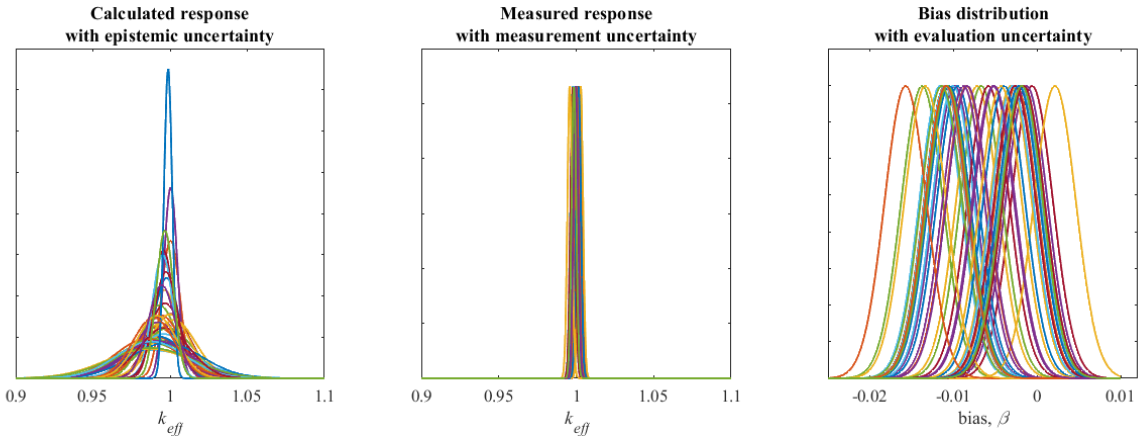


Figure 12. Calculated response, measured responses, and bias distributions.

Table 2 details information about the generated toy models, their c_k values, and Whisper weights.

Table 2. Toy model responses, biases with associated uncertainties, and weights

Model Index	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
Toy #1	0.9981	0.0015	1.0006	0.0020	0.0100	1.0006	0.0025	0.3284	0.3391
Toy #2	0.9998	0.0015	0.9884	0.0020	0.0200	0.9884	0.0025	0.8357	0.8632
Toy #3	0.9998	0.0015	0.9954	0.0020	0.0190	0.9954	0.0025	0.5897	0.6091
Toy #4	1.0013	0.0015	0.9895	0.0020	0.0192	0.9895	0.0025	0.8174	0.8443
Toy #5	1.0009	0.0015	0.9856	0.0020	0.0253	0.9856	0.0025	0.8304	0.8578
Toy #6	1.0000	0.0015	0.9940	0.0020	0.0217	0.9940	0.0025	0.6175	0.6379
Toy #7	0.9998	0.0015	0.9934	0.0020	0.0157	0.9934	0.0025	0.7289	0.7528
Toy #8	1.0009	0.0015	1.0026	0.0020	0.0191	1.0026	0.0025	0.2602	0.2687
Toy #9	1.0008	0.0015	0.9963	0.0020	0.0078	0.9963	0.0025	0.7701	0.7949
Toy #10	1.0008	0.0015	0.9948	0.0020	0.0215	0.9948	0.0025	0.5924	0.6119
Toy #11	1.0007	0.0015	0.9904	0.0020	0.0191	0.9904	0.0025	0.7845	0.8104
Toy #12	1.0009	0.0015	0.9883	0.0020	0.0215	0.9883	0.0025	0.8117	0.8384
Toy #13	0.9988	0.0015	0.9963	0.0020	0.0177	0.9963	0.0025	0.5646	0.5832
Toy #14	1.0010	0.0015	1.0011	0.0020	0.0113	1.0011	0.0025	0.2968	0.3065
Toy #15	0.9995	0.0015	0.9969	0.0020	0.0122	0.9969	0.0025	0.5995	0.6191
Toy #16	1.0000	0.0015	0.9848	0.0020	0.0185	0.9848	0.0025	0.9681	1.0000

Table 2. Toy model responses, biases with associated uncertainties, and weights (continued)

Model Index	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
Toy #17	1.0008	0.0015	0.9968	0.0020	0.0054	0.9968	0.0025	0.8420	0.8683
Toy #18	1.0010	0.0015	0.9954	0.0020	0.0201	0.9954	0.0025	0.5788	0.5979
Toy #19	0.9985	0.0015	0.9866	0.0020	0.0254	0.9866	0.0025	0.8019	0.8283
Toy #20	1.0006	0.0015	1.0025	0.0020	0.0144	1.0025	0.0025	0.2239	0.2312
Toy #21	0.9987	0.0015	0.9940	0.0020	0.0270	0.9940	0.0025	0.5755	0.5945
Toy #22	0.9996	0.0015	0.9870	0.0020	0.0282	0.9870	0.0025	0.7570	0.7820
Toy #23	1.0009	0.0015	0.9964	0.0020	0.0052	0.9964	0.0025	0.9051	0.9333
Toy #24	0.9996	0.0015	0.9921	0.0020	0.0149	0.9921	0.0025	0.8019	0.8282
Toy #25	1.0004	0.0015	0.9873	0.0020	0.0260	0.9873	0.0025	0.7759	0.8015
Toy #26	1.0003	0.0015	0.9988	0.0020	0.0064	0.9988	0.0025	0.5507	0.5683
Toy #27	1.0013	0.0015	0.9953	0.0020	0.0260	0.9953	0.0025	0.5406	0.5584
Toy #28	1.0008	0.0015	0.9914	0.0020	0.0292	0.9914	0.0025	0.6317	0.6525
Toy #29	0.9991	0.0015	0.9924	0.0020	0.0261	0.9924	0.0025	0.6300	0.6508
Toy #30	0.9998	0.0015	0.9942	0.0020	0.0161	0.9942	0.0025	0.6838	0.7062
Toy #31	0.9998	0.0015	1.0001	0.0020	0.0193	1.0001	0.0025	0.3802	0.3927
Toy #32	0.9991	0.0015	0.9923	0.0020	0.0114	0.9923	0.0025	0.8938	0.9230
Toy #33	1.0002	0.0015	0.9951	0.0020	0.0117	0.9951	0.0025	0.7276	0.7514
Toy #34	1.0003	0.0015	0.9940	0.0020	0.0161	0.9940	0.0025	0.6930	0.7158
Toy #35	1.0011	0.0015	0.9970	0.0020	0.0195	0.9970	0.0025	0.5187	0.5357
Toy #36	1.0009	0.0015	0.9880	0.0020	0.0230	0.9880	0.0025	0.7969	0.8232
Toy #37	0.9984	0.0015	0.9877	0.0020	0.0170	0.9877	0.0025	0.9210	0.9513
Toy #38	0.9991	0.0015	0.9962	0.0020	0.0113	0.9962	0.0025	0.6651	0.6868
Toy #39	1.0005	0.0015	0.9918	0.0020	0.0156	0.9918	0.0025	0.8009	0.8272
Toy #40	0.9984	0.0015	0.9967	0.0020	0.0099	0.9967	0.0025	0.6664	0.6880

The USLs calculated by each methodology with 95% confidence are listed in Table 3. Note that the nonparametric margin m_{np} is assumed to be zero because it is an ad hoc parameter that cannot be statistically justified, resulting in an additional conservativism for the calculated bias. Table 3 indicates that the Whisper methodology as well as the nonparametric methodology evaluates USL more conservatively than the parametric methodologies. 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 3. Toy model USL results for 95% confidence

	CM	MOS	USL($= 1.0 - \text{CM} - \text{MOS}$)
Parametric	1,516 pcm	500 pcm	0.9798
Nonparametric	2,409 pcm	500 pcm	0.9709
Whisper	2,023 pcm	678 pcm	0.9730
TSURFER	1,618 pcm	500 pcm	0.9788

Because the true application response is known for the toy model, we can quantify the how far the USLs of the different methodologies are from the true application response. Focusing on the CM only, because it is calculated based on the bias, its value for the various methodologies is compared without the MOS, because the choice of the latter is more arbitrary, as it includes the effects of unknown modeling uncertainties. Figure 13 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 the Bayes theorem. The goal here is to estimate an 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 1,440 pcm which is the same as the mean value of the TSURFER posteriori PDF. The spread of the posteriori PDF is the result of the aleatory uncertainty from the benchmark model and the measurements. Based on this PDF, TSURFER calculated an LTL at 1,618 pcm which covers 95% of the PDF, as follows:

$$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'} = 1440 + 1.65 \times 108 = 1618 \text{ pcm}. \tag{37}$$

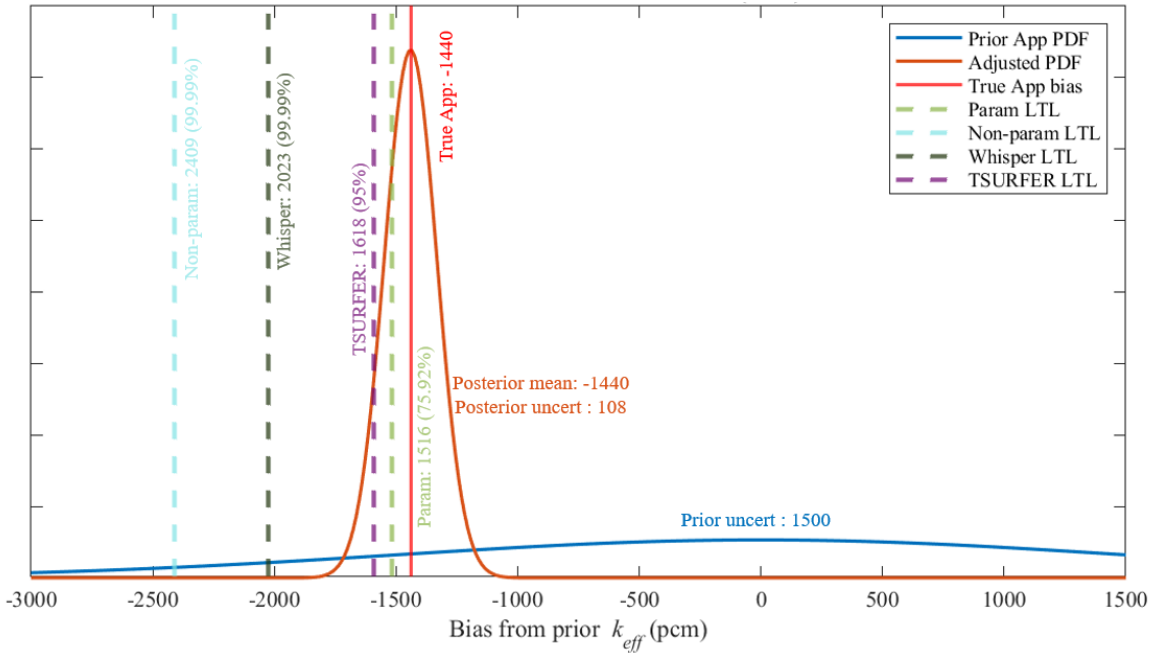


Figure 13. Toy model application PDF and estimated LTLs: bias and 95% LTL.

For the parametric methodology, the inverse-variance weighted average β_p is -636 pcm, setting the nonconservative parameter to 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 as

$$\mathrm{CM}_p = -\beta_p + \varrho\sigma_p + \Delta_m = 636 + 1.65 \times 535 + 0 = 1516 \text{ pcm.} \tag{38}$$

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

For the nonparametric methodology, only the minimum negative bias of -1,529 is employed. (recall the nonparametric margin is assumed to be zero for this analysis because it is heuristically determined, and it will result in even more conservative CM value), yielding CM value of

$$\mathrm{CM}_{np} = -\min\{k_{c_i} - k_{m_i}\} + \varrho\sigma_p + m_{np} + \Delta_m = 1529 + 1.65 \times 535 + 0 + 0 = 2409 \text{ pcm.} \tag{39}$$

Recall that Whisper builds an EV-like PDF by generating samples from the various bias PDFs shown in Figure 12 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 12. 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. Because these two PDFs are heavily overlapped, their samples may be approximately considered iid. They are effectively being sampled from the same PDF, so their maximum will be equivalent to the generation of a second-order EV PDF. For a normal distribution, the 95% confidence interval for 95% coverage using a second-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 nonparametric bias and the evaluation uncertainty with the EV multiplication factor such that

$$\mathrm{CM}_w = m + \Delta_m \approx -\min\{k_{c_i} - k_{m_i}\} + \nu\sigma_e + \Delta_m = 1529 + 1.95 \times 250 + 0 = 2017 \text{ pcm,} \tag{40}$$

which approximates the actual value calculated by Whisper. Both the nonparametric and Whisper CM values produce LTL values providing nearly 100% coverage of the posteriori PDF, which is much higher than the 95% coverage reported by the two methodologies.

In the example above, although the analysis includes 40 experiments, only two experiments that correspond to the two left-most biased PDFs in the right plot of Figure 12 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 14.

This trend shows that as the number of overlapping bias PDFs increases, the multiplier value will also monotonically increase. This behavior is undesirable because 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 such as 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) through (17), establishing 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 a large number of highly relevant experiments.

The previous discussion was motivated by the observation that the most biased PDFs are those controlling the Whisper CM value. To further validate this observation, the calculation of the CM is repeated 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. The results of this numerical experiment are shown in Figure 15. Results indicate that the EV PDF obtained with the two most biased PDFs is very similar to the EV PDF obtained with all PDFs. Also, the tolerance limit obtained with the Whisper weighting is nearly identical to the one with unit weights.

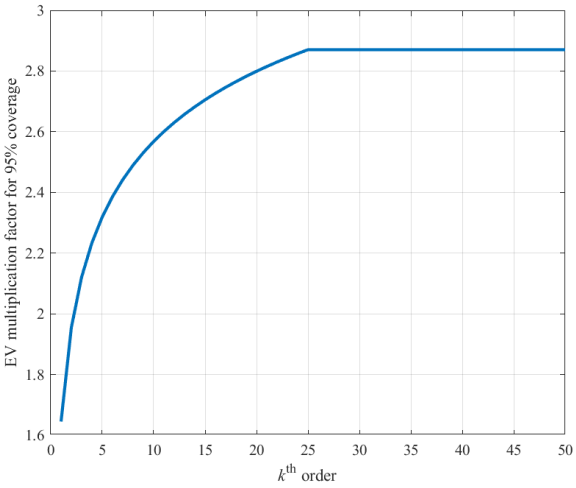
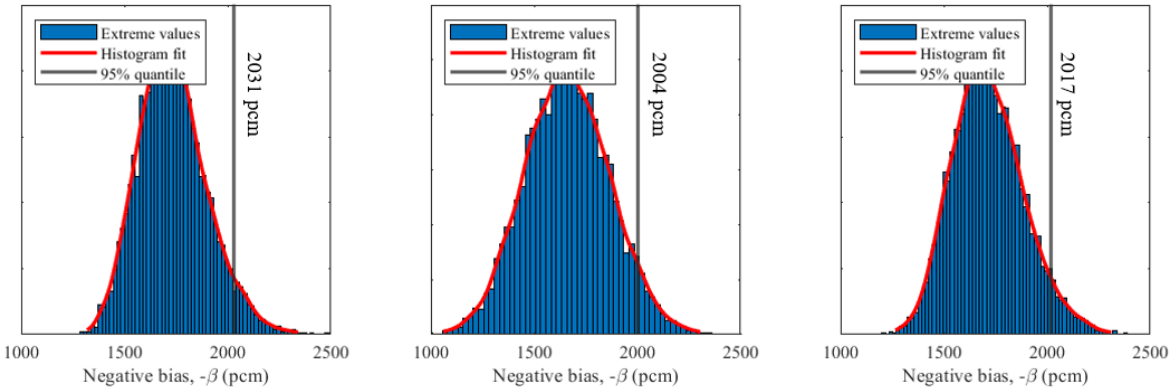


Figure 14. k^{th} order EV multiplier value.



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

Figure 15. Impact of model and weight selection on extreme value.

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

$$\mathrm{MOS}_d = \varrho\sigma_{k'} = 1.65 \times 108 = 178 \text{ pcm}, \tag{41}$$

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

6.2 USL CALCULATIONS WITH PU-SOLUTION BENCHMARKS

This section describes various USL calculation methodologies using a suite of 29 Pu-solution benchmarks containing 15 g/L. The MIX-SOL-THERM-002-001 benchmark is arbitrarily selected as the application model with calculated k_{eff} of 1.0015. 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. Table 4 provides detailed information about this set of the benchmarks, including c_k, Whisper weights, and the measurement uncertainties.

Table 4. 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 16 plots the calculated responses with their epistemic cross-section uncertainties and the measured responses with their evaluation uncertainties. The maximum and minimum cross-section uncertainties are 873 and 860 pcm, respectively, and the evaluation uncertainty is 510 pcm for all except the last three benchmarks, whose evaluation uncertainty is 400 pcm because of lower reported measurement uncertainties.

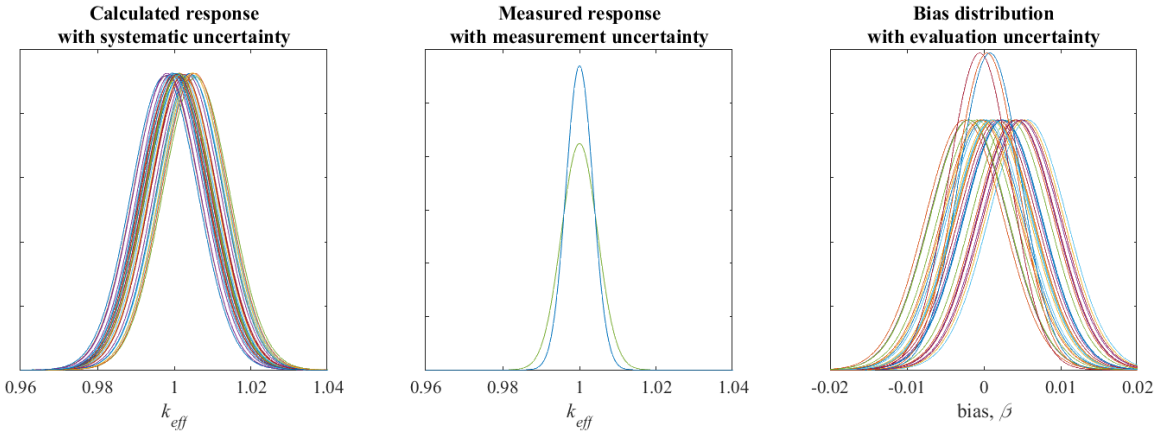


Figure 16. Calculated response, measured responses, and bias distributions.

The USLs for the four methodologies are calculated and provided in Table 5. With the same argument in Section 6.1, the nonparametric margin is set to be zero.

Table 5. Pu-solution USL results for 95% confidence

	CM	**MOS**	**USL(= 1.0 − CM − MOS)**
Parametric	900 pcm	500 pcm	0.9860
Nonparametric	1,153 pcm	500 pcm	0.9835
Whisper	1,448 pcm	1,123 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 nonconservative bias adjustment Δ_m is selected to cancel it out, as discussed above. The parametric CM is calculated as

$$\mathrm{CM}_p = -\beta_p + \varrho\sigma_p + \Delta_m = -139 + 1.65 \times 547 + 139 = 900 \text{ pcm}, \tag{42}$$

the nonparametric methodology CM is given by

$$\mathrm{CM}_{np} = -\min\{k_{c_i} - k_{m_i}\} + \varrho\sigma_p + m_{np} + \Delta_m = 253 + 1.65 \times 977 + 0 + 0 = 1153 \text{ pcm}, \tag{43}$$

and the MOS for the Whisper is calculated as

$$\mathrm{MOS}_d = \varrho\sigma_{k'} = 1.65 \times 373 = 615 \text{ pcm}. \tag{44}$$

Lastly, the TSURFER CM is calculated as

$$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'} + \Delta_m = -78 + 1.65 \times 373 + 78 = 615 \text{ pcm}. \tag{45}$$

Unlike the toy model study, the true application response remains unknown. Nevertheless, if one solely relies on the prior uncertainties as shown in Figure 17, then 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.

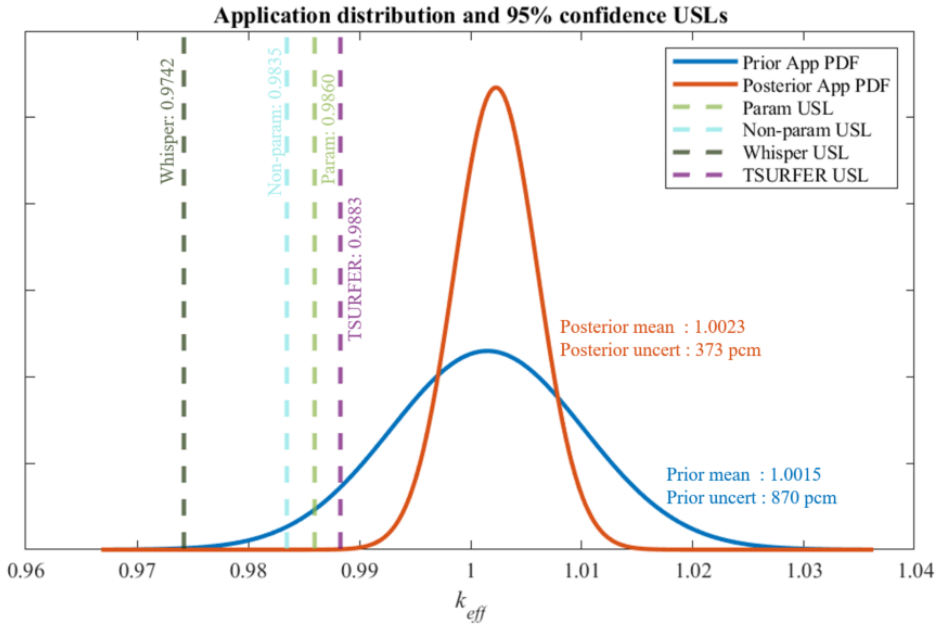
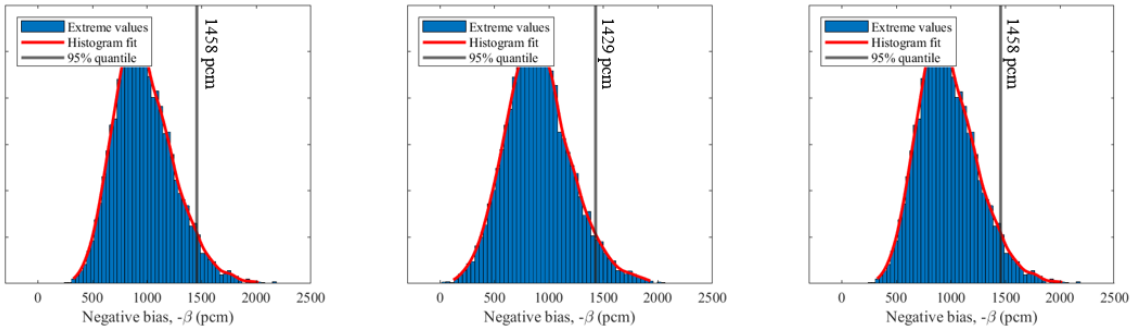


Figure 17. Mix-Sol-Therm benchmark application and estimated USLs.

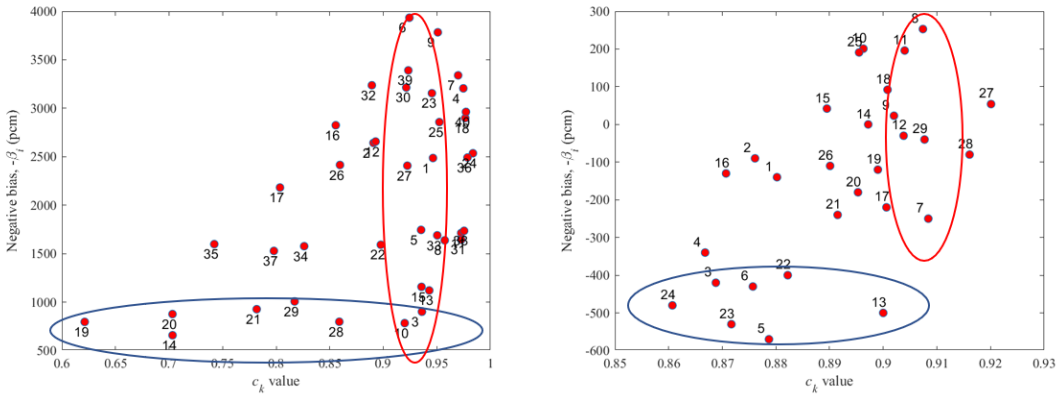
A simple exercise similar to the toy problem was repeated to compare the impact of Whisper weights by comparing three cases with equal weights, limiting the analysis to the most biased ten PDFs, and including all experiments with the c_k-based weights. Figure 18 shows that similar CM values are obtained for the three cases, indicating that the CM values are 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 18. Impact of benchmark and weight selection on extreme value.

Given these results, an additional experiment is performed to determine whether the c_k weighting can effectively reduce the impact of the most negatively biased experiments. The models/benchmarks are grouped in two different ways: the blue groups in both plots of Figure 19 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 19. 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 20 single out one experiment at a time based on the c_k value, and those 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 illustrate the same experiment, but they 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 experiment has the most negative bias and is expected to have the greatest influence on the results. Results indicate that the weighting procedure does have an impact, albeit negligible, on the calculated CM value.

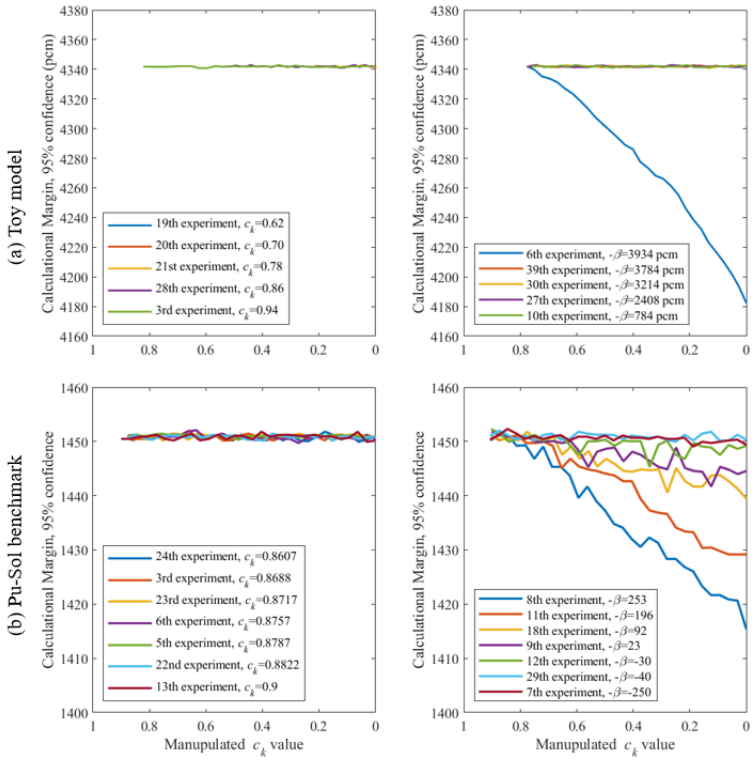


Figure 20. Analysis of Whisper c_k-based weighting.

6.3 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 in which the noted procedure is repeated 62 times, with a different experiment selected as the application in each time. Detailed information about the experiments is given in

Table 6.

Table 6. 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

Table 6. Benchmark models specification (continued)

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
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

Focusing on TSURFER and Whisper methodologies, their resulting CMs and MOSs are compared with the true bias of the selected application, as shown in Figure 21. The TSURFER MOS is set to be a fixed value of 500 pcm, as in the previous analyses. The results show that for both methodologies, the sum of two margins—CM and MOS—is larger than the true application bias represented as the dashed line except for the right-most point of TSURFER.

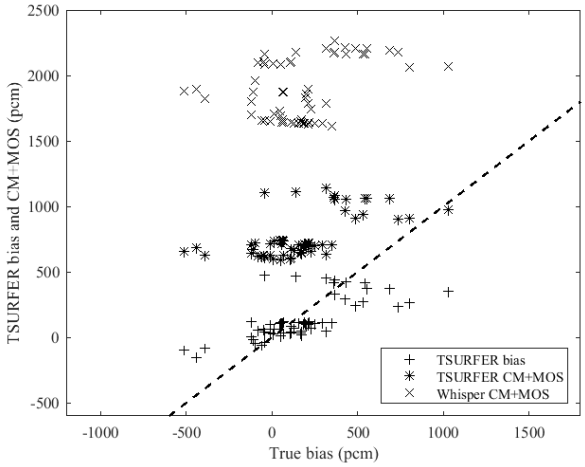


Figure 21. CM and MOS evaluation with different application.

Next, the CM and USL values for the four methodologies in Figure 22 are compared against the true bias. Note that the point on the x-axis represents the selection of a 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—when the code underpredicts the measured value. The applications on the x-axis are included by ranking the Whisper CM values from low to high. This helps separate the Whisper CM values into three distinct groups as discussed below.

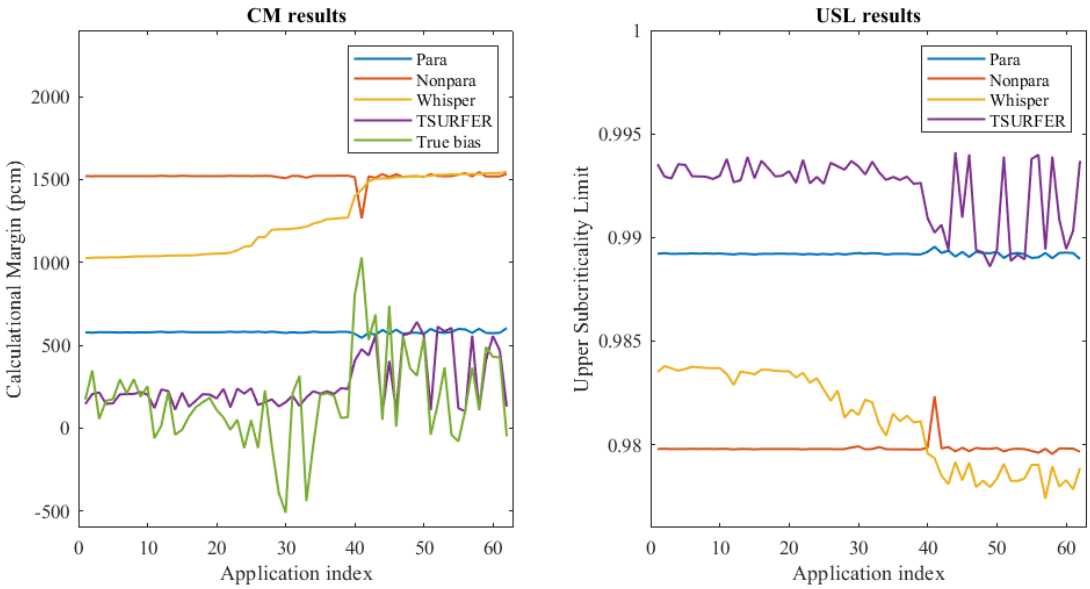


Figure 22. CM and USL with different application.

Note that the parametric and nonparametric CMs on the left do not change significantly, regardless of application selection, because both methodologies evaluate their uncertainties using weighted statistics, which often yields stable results with a large number of experiments. However, the nonparametric CM curve drops once for the HEU-SOL-THERM-001-010 application, which has the highest bias of -1,030 pcm and a prior uncertainty of 1,080 pcm. This drop occurs because the nonparametric bias solely

depends on the most biased experiment. The parametric bias is less impacted by the exclusion of this high bias because it relies on a weighted average formula that is more robust to outliers.

The Whisper CM curve can be divided roughly into three application groups: low, intermediate, and high Whisper CM values. These categories are based on the different cut-off values ($c_{k,acc}$)—per Eqs. (15) through (17)—employed by Whisper to discard experiments. This is accomplished by essentially assigning zero weights to all experiments below a given c_k value, as shown in Figure 23. The low, intermediate, and high CM groups exclude 24, 18, and 7 experiments, respectively, and are assigned zero weights, as shown on the x-axis.

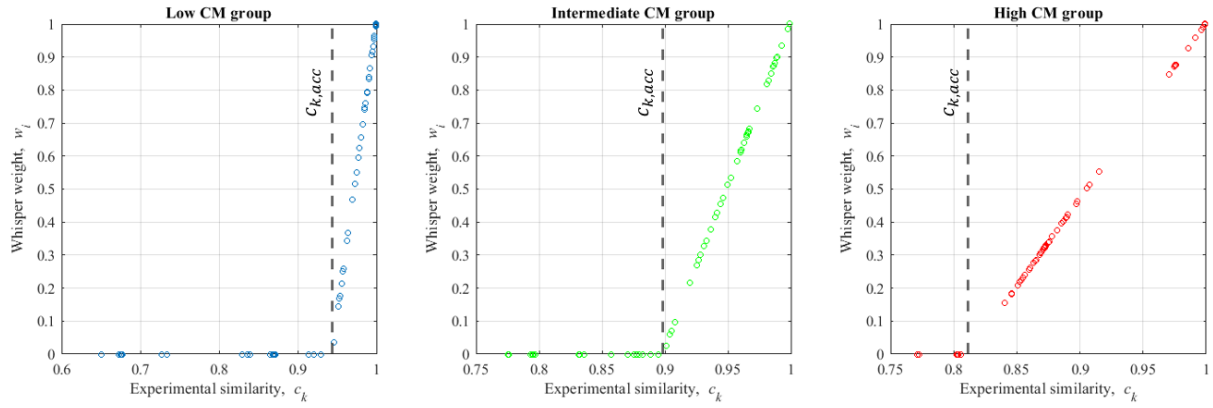


Figure 23. Change in $c_{k,acc}$ with different application selection.

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, as shown in Figure 14. 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. The CM values are fairly constant within each group, indicating a lack of sensitivity to the specific non-zero weight values used by Whisper for each group, an observation supported by earlier numerical experiments. This is addressed in the discussion related to Figure 19.

In the TSURFER results, the CM values are very close to the true bias, but noticeable differences can be observed for one experiment when used as the application. For this case, the nonparametric CM shows a drop of approximately 250 pcm value. TSURFER cannot capture this bias, likely because the modeling errors are not factored into the TSURFER CM calculations. TSURFER assumes that all errors originate from known epistemic sources of uncertainties, such as nuclear cross sections. To hedge against this unknown source of errors, TSURFER employs the MOS as an additional margin. Section 6.5 provides a proposed method, denoted by PCM, which is designed to expand the TSURFER capability to account for modeling errors.

When comparing the Whisper and nonparametric USL values, it can be observed that Whisper sometimes becomes more conservative than the nonparametric methodology. This occurs after adding the MOS term, which provides an additional margin for the non-covered cross-section uncertainties, represented by the residual uncertainty after performing TSURFER-like cross-section adjustments. TSURFER accounts for this residual uncertainty in the CM value. This effectively results in double-counting for the residual nuclear data uncertainties, hence the lower USL values.

6.4 CM VERIFICATION WITH ANALYTICAL CRITICALITY SAFETY BENCHMARKS

The analytical criticality safety problem (hereinafter the *analytical problem*) is employed to measure the effect of individual benchmarks as measured in terms of their c_k values on the CM calculations and the impact of noncovered subspace. This problem occurs when the number of experiments is smaller than the number of parameters. The analytical problem includes 10 independent input parameters with one constant input parameter representing the cross-sections, and 9 benchmark models with one application. The calculated responses can be represented by the multiplication of the sensitivity profiles of the benchmarks and the application by the input variables, such as

$$k_c = Sx \tag{46}$$

where $S \in \mathbb{R}^{9\times10}$ is the sensitivity profile matrix, and $x \in \mathbb{R}^{10}$ is a vector of the input parameters.

Figure 24 presents the resulting calculated responses with their associated systematic uncertainties, the measured responses with the measurement uncertainties, and the biases of each benchmark with the evaluation.

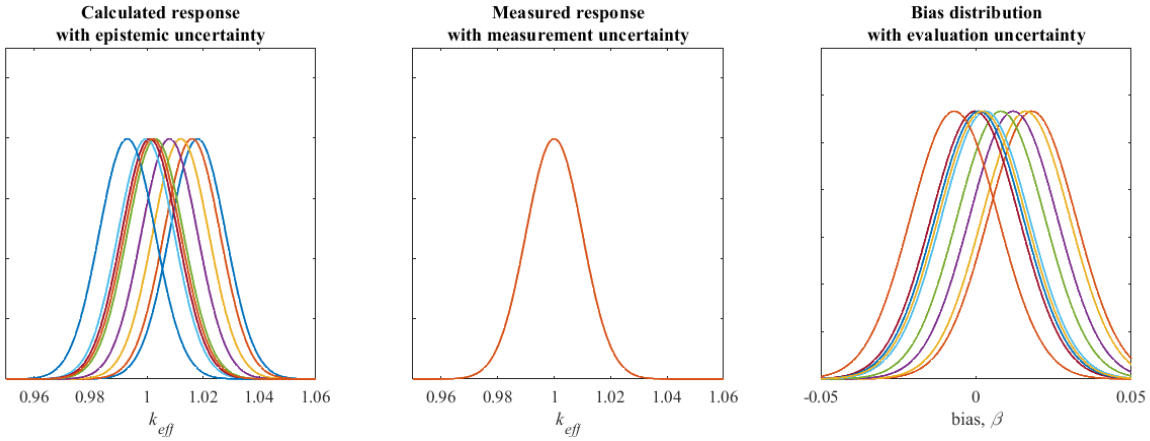


Figure 24. Calculated response, measured response, and bias distribution.

The measurement PDFs are all overlapping because all measurements are identical, representing critical conditions with identical uncertainties. The detailed benchmark specifications with their c_k values and Whisper weights are provided in Table 7.

Table 7. Employed benchmarks specification

Benchmark name	Measured, k_m		Calculated, k_c		Bias ($k_c - k_m$)		Weight	
	k_{eff}	σ_m	k_{eff}	σ_s	β	σ_e	c_k	w
Analytical Benchmark #1	1.0000	0.0100	1.0180	0.0100	0.0180	0.0141	0.9900	1.0000
Analytical Benchmark #2	1.0000	0.0100	1.0160	0.0100	0.0160	0.0141	0.9000	0.9091
Analytical Benchmark #3	1.0000	0.0100	1.0120	0.0100	0.0120	0.0141	0.8000	0.8081
Analytical Benchmark #4	1.0000	0.0100	1.0080	0.0100	0.0080	0.0141	0.7000	0.7071
Analytical Benchmark #5	1.0000	0.0100	1.0030	0.0100	0.0030	0.0141	0.6000	0.6061
Analytical Benchmark #6	1.0000	0.0100	0.9997	0.0100	-0.0003	0.0141	0.5000	0.5051
Analytical Benchmark #7	1.0000	0.0100	1.0010	0.0100	0.0010	0.0141	0.4000	0.4041
Analytical Benchmark #8	1.0000	0.0100	0.9930	0.0100	-0.0070	0.0141	0.3000	0.3030
Analytical Benchmark #9	1.0000	0.0100	1.0020	0.0100	0.0020	0.0141	0.2000	0.2020

The CMs are calculated using the four methodologies with two different sorting metrics by adding one benchmark experiment at a time in a descending and ascending order of c_k. The results are shown in Figure 25. The left plot of Figure 25 shows the CM results by including experiments in a descending order, the most relevant experiments are assimilated with the measurements early on. Interestingly, the CMs are showing a climbing trend, implying that the addition of less relevant experiments erodes the confidence earlier established by the higher relevant experiments. The reason for that may be understood by analyzing the biases in Table 7 and the c_k values of each analytical benchmark and plotting them as shown in Figure 26.

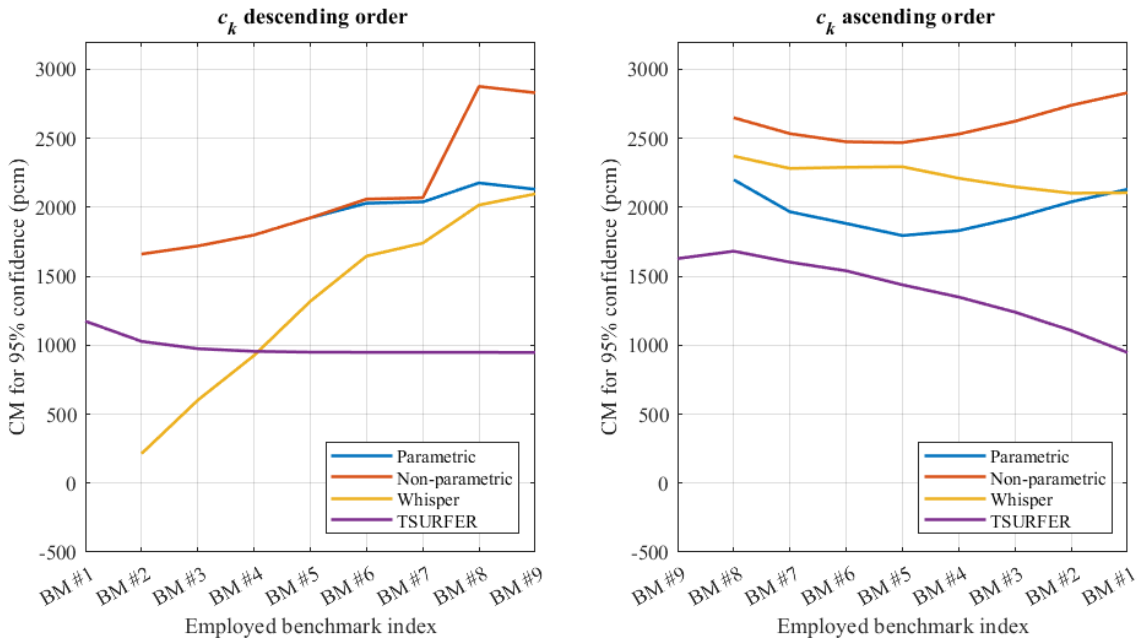


Figure 25. CM evaluation by various methodologies and different sorting metrics.

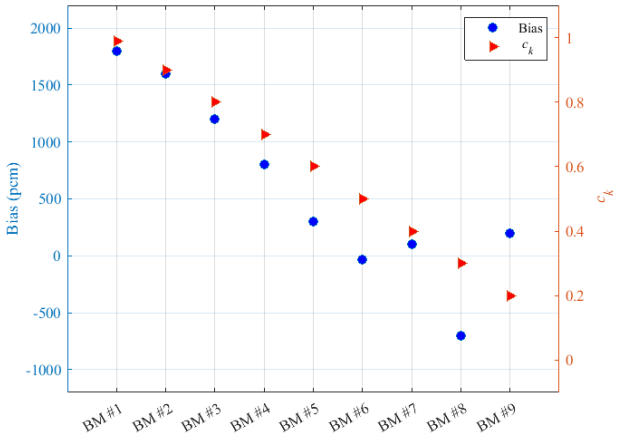


Figure 26. Bias and c_k of analytical benchmarks.

Most of the experiments have a positive bias with similar uncertainty, implying that the PDFs with the most coverage over the negative bias values will have the most influence on the calculated CM. Specifically, experiment #8 will have the most impact, followed by experiment #6, and then experiment #7. In descending order, each new bias PDF has increasingly more coverage in the negative bias region; hence the increase in the CMs value for the parametric, the nonparametric, and the Whisper methodologies. The TSURFER-based CMs have a declining trend, unlike those of the other methodologies, because although the coverage in the negative bias region increases, the benchmark c_k values have a declining trend, which eventually leads to a smaller impact on the TSURFER CMs. This is a desirable trend, as it confirms consistency with basic statistical inference expectations, so one should have more confidence as additional experimental evidence is included in the assimilating procedure.

In ascending order, the experiments with the highest PDF area in the negative bias region contribute to the calculated CMs, and the addition of more relevant experiments introduces little impact on the calculated CMs for the other three methodologies. The slight reduction noticed to the result of the reduction in the weights. For TSURFER, the descending order produces reasonable results, as the addition of less relevant experiments has an increasingly lower impact on the calculated CMs. In the ascending order, the earlier experiments with low relevance are not capable of reducing the bias: hence the higher CM values. As more relevant experiments are added, the slope of the purple curve increases, indicating that more confidence can be gained with the increase in experimental relevance.

Table 8 details the calculations of the resulting CM values employing all the experiments. Note that the Whisper CM approximation is not approximated here because the most biased PDF, corresponding to experiment #8, has a very low weight of 0.3, so the exact value is reported directly.

Table 8. Analytical benchmark CM calculation

Methodology	CM calculation
Parametric	$\mathrm{CM}_p = -\beta_p + \varrho\sigma_p + \Delta_m = -586 + 1.65 \times 1295 + 586 = 2131$ pcm
Nonparametric	$\mathrm{CM}_{np} = -\min\{k_{c_i} - k_{m_i}\} + \varrho\sigma_p + m_{np} + \Delta_m$ $= 700 + 1.65 \times 1295 + 0 + 0 = 2831$ pcm
Whisper	$\mathrm{CM}_w = m + \Delta_m = 2078$ pcm
TSURFER	$\mathrm{CM}_T = -\beta_T + \varrho\sigma_{k'} + \Delta_m = -1082 + 1.65 \times 576 + 1082 = 948$ pcm

Figure 27 shows the resulting LTLs (negative sign of CMs) and USLs with the application prior and posterior PDFs. The calculated k_{eff} values for most of the benchmark experiments are higher than the measured k_{eff} positive-biased PDFs, and the measurement uncertainties are substantially high, being of the same magnitude as the prior uncertainties. The uncertainty spread for these three methodologies is mostly determined by the evaluation uncertainty, or the measurement uncertainty in this analysis. Therefore, the CM values obtained using the four methodologies are mainly impacted by the high measurement uncertainties rather than the bias, which is canceled out by the nonconservative bias adjustment parameter Δ_m. As noted above, Whisper effectively double counts for the residual parameter uncertainty using the MOS margin. This is evident when comparing the USL and CM values.

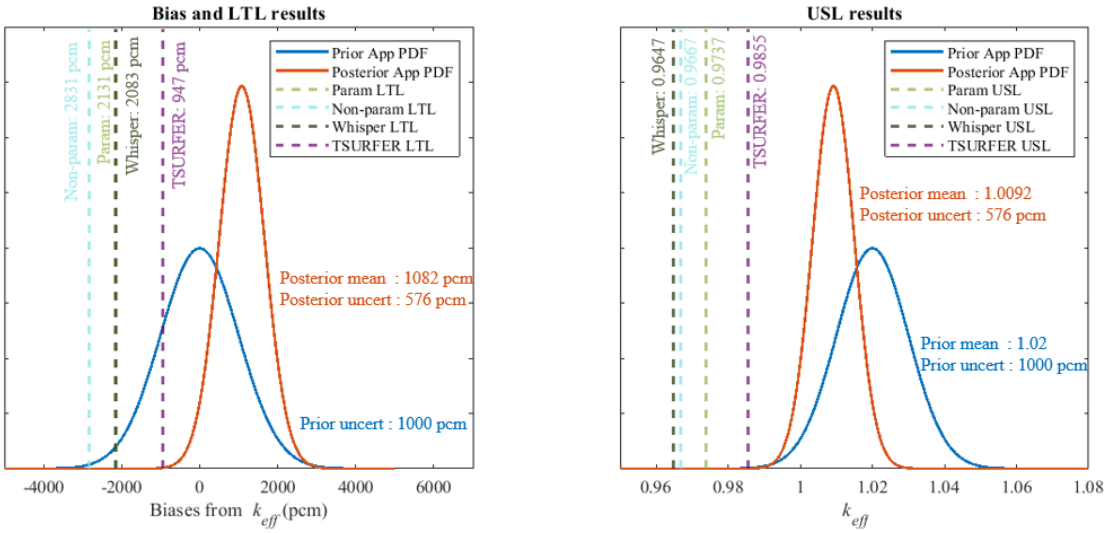


Figure 27. CM and USL results for analytical benchmarks.

The next numerical experiment focuses on a key challenge in any inference procedure: the inability of an experiment to cover the entire uncertainty space of the parameters, the nuclear cross sections. It is difficult to assess the impact of the noncovered directions in the uncertainty space on the application response, especially when the application is sensitive to some of these directions. It is thus an intuitive requirement to include aspects of the application model in the inference procedure. This may be done directly via a quantitative approach such as the TSURFER methodology, which includes the sensitivity profile of the application in the CM calculations, or it may be done indirectly via the similarity metric, as in the Whisper methodology. The other two methodologies do not include the application in the analysis: they rely solely on expert judgment to select the most representative experiments.

Since the MOS relies on an analytical benchmark developed by Subgroup 11 (as discussed in Section 6.4), the focus here is only on the calculations of the CM to understand the impact of the noncovered uncertainties. For simplicity, we assume that the USL is given by,

$$\text{USL} = 1 - \text{CM}, \tag{47}$$

because the MOS is often treated as a fixed value, or it could be used to double-count for the residual uncertainties. The nonconservative bias adjustment parameter margin is set to zero, ignoring that the bias could be positive. These assumptions are adequate because the main goal is to show whether the various methodologies can account for the noncovered parameter subspace.

In the SG11 benchmark, the reference k_{eff} is calculated by the multiplication of the sensitivity profiles by the reference nuclear data. The calculated k_{eff} for the experiments and the application can be written as

$$k_i^{\exp} = S_i^{\exp} \times x_{ref}, \; i = 1,2, \ldots,9 \text{ for the experiments, and} \tag{48}$$

$$k^{\text{app}} = S^{\text{app}} \times x_{ref} \text{ for the application,} \tag{49}$$

where the length of each sensitivity vector is 10, $S \in \mathbb{R}^{10}$. Because the number of benchmarks is less than the number of uncertain parameters, or the dimension of the sensitivity vectors, the matrix of the experimental sensitivity profiles is underdetermined, with at most a rank of 9. Therefore, the null space of this matrix is expected to be of dimension 1 or more, implying that there is at least one direction in the parameter space that is not covered by any of the experiments. If this direction has an error component, then it is not possible to estimate its impact on the application responses using the experimental data. Thus, numerical experiments are designed with a known error component along the null space to estimate its true impact on the application response and to compare that to the calculated CM values by the various methodologies.

The null space, spanned by a single vector V_{10}, may be calculated using the singular value decomposition (SVD):

$$S^{\exp} = U\Sigma V^T, \tag{50}$$

$$V = \begin{bmatrix} | & | & & | \\ V_1 & V_2 & \cdots & V_{10} \\ | & | & & | \end{bmatrix}, \tag{51}$$

where $S^{\exp} \in \mathbb{R}^{9\times10}$, $U \in \mathbb{R}^{9\times9}$, $\Sigma \in \mathbb{R}^{9\times10}$, and $V \in \mathbb{R}^{10\times10}$. To ensure that the introduced error component does not change the experimental biases provided by the benchmark, the true cross-section error is selected as

$$x_{true} = x_{ref} + \Delta\widetilde{x}, \tag{52}$$

$$x_{GLLS} = x_{ref}. \tag{53}$$

It is assumed that the true cross-section error has two components: one that can be obtained using the GLLS solution, and the other that lives in the null space of the sensitivity matrix. Because the application's sensitivity vector is not explicitly included in the analysis of the first three methodologies, it is assumed that the application sensitivity vector is scaled by different multipliers while maintaining its direction to ensure that its c_k values with the various experiments remain the same.

Specifically, the application sensitivity vector is scaled by two different multipliers, 0.5 and 2.0, respectively, and the corresponding impact on the response (k_{t05} and k_{t20}) is calculated, as well as the estimated GLLS solutions (k_{G05} and k_{G20}),

$$k_{G05}^{\text{app}} = S^{\text{app}} x_{ref} + 0.5 \times S^{\text{app}} \Delta x, \tag{54}$$

$$k_{t05}^{\text{app}} = S^{\text{app}} x_{ref} + 0.5 \times S^{\text{app}} \Delta \tilde{x}, \tag{55}$$

$$k_{G20}^{\text{app}} = S^{\text{app}} x_{ref} + 2.0 \times S^{\text{app}} \Delta x, \text{and} \tag{56}$$

$$k_{t20}^{\text{app}} = S^{\text{app}} x_{ref} + 2.0 \times S^{\text{app}} \Delta \tilde{x}, \tag{57}$$

where the first term of the right-hand side of each equation is equivalent to the reference k_{eff} of the application $S^{\text{app}} x_{ref} = k_{ref}^{\text{app}}$. The first term is kept constant to emulate the reference k_{eff} value that does not change because of a change in the sensitivity values and that only scales the second term to emulate the increase in the first-order derivatives.

Figure 28 shows the different results caused by the change in the application sensitivity profiles. The middle plot employs the reference application sensitivity profiles. The USLs for three noted methodologies (the parametric, the nonparametric, and the Whisper methodologies) stay the same, regardless of the sensitivity changes, while the TSURFER USL and the true k_{eff} values change. The TSURFER USL is set according to the discrepancy between the reference and the GLLS result. That is, the larger discrepancy results in the lower USL. For example, the TSURFER USL on the left plot of Figure 28 is set to be the highest among the other USLs, whereas its USL is set to be almost the same as that of the nonparametric methodology on the right plot, which is known as a conservative USL calculation method.

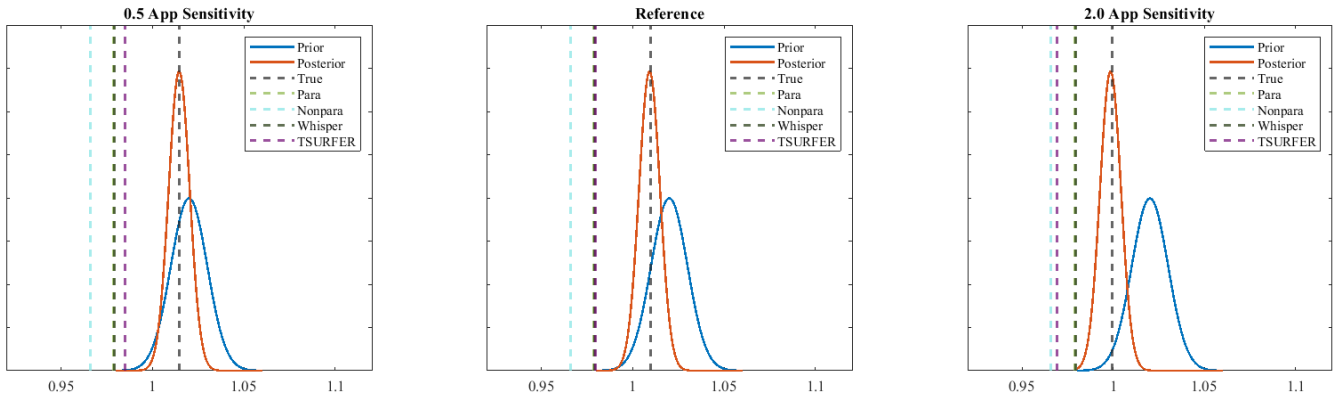


Figure 28. Impact of change in sensitivities on USL calculation.

These results indicate that, except for the TSURFER methodologies, the other methodologies do not have a built-in functionality to account for situations in which the c_k value is blind to the differences in magnitude between the application and experiments.

6.5 ACCOUNTING FOR MODELING ERRORS USING PCM

A key assumption in the TSURFER CM calculation is that the calculated responses' uncertainty originates mainly from the cross-sections uncertainties, implying that other sources, such as modeling errors, could have an undesirable impact on the results if left unaddressed. This assumption is generally acceptable if the unaccounted sources of uncertainties are significantly smaller than those resulting from cross-section uncertainties. However, two experiments could have the same near-perfect similarity to the application, albeit with different biases. In this case, TSURFER will be forced to take a weighted average of the two biases. Thus, there a method must be developed to hedge against the presence of modeling errors. Currently, TSURFER employs a χ^2-based rejection criterion when the differences between measurement and predictions are too large. However, a χ^2 of 1 does not necessarily hedge against this situation because cross-section uncertainties are known to be overly conservative compared to the actual differences between measurements and predictions.

Hence, another method is needed to calculate a more realistic estimate of the modeling errors and include it in the assimilating procedure. Such a method was recently developed in collaboration with Oak Ridge National Laboratory (ORNL), as denoted by PCM. The objective is to calculate an estimate of the modeling errors using physics-based or data-driven approaches. Details on this method may be found in a nonprovisional patent recently filed by ORNL [Mertyurek September 2021].

To demonstrate this method, experiment HEU-SOL-THERM-001-010 is selected as the application (Section 6.3). Recall that this experiment had the highest bias of -1,030 pcm, and its prior uncertainty was 1,080 pcm. Figure 29 analyzes the results of calculating CM and MOS limits using the four different methodologies by including one experiment at a time, ordered according to their bias values, adding first experiments with the highest negative bias (expected to have the highest impact on most of the methodologies as explained earlier), and ending with experiments with the highest positive bias which have no impact because the nonconservative bias adjustment was used.

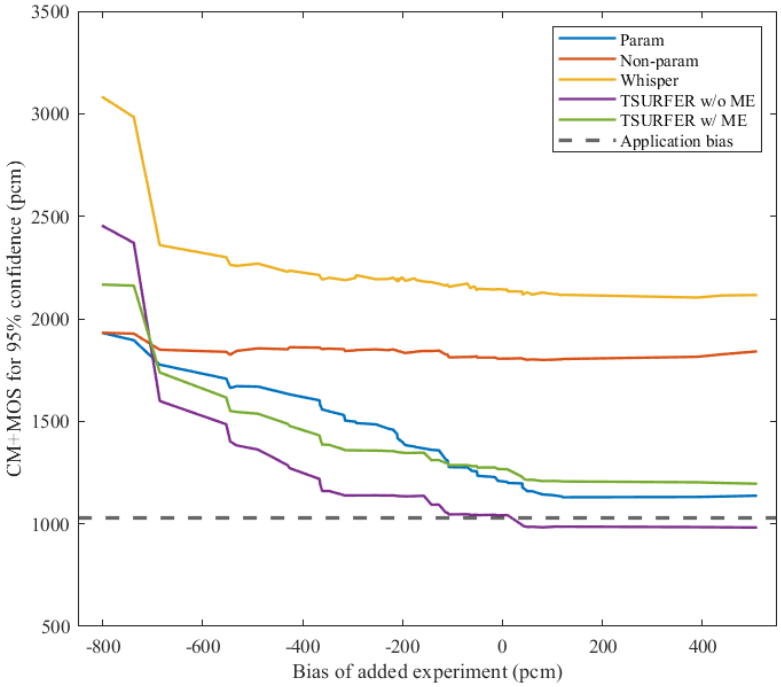


Figure 29. Impact of modeling error on CM+MOS calculation.

The legend of Figure 29 "TSURFER w/o ME" refers to the standard TSURFER application which does not account for modeling error (ME), whereas the "TSURFER w/ ME" refers to TSURFER augmented by the proposed PCM methodology. Note that the TSURFER MOS margin is a fixed amount, whereas PCM directly incorporates modeling errors into the assimilating procedure as an additional source of epistemic uncertainty, i.e., added to that of the cross sections; hence, no additional provision is made to add a fixed MOS value. Several observations may be made here. First, note that the nonparametric CM+MOS margins are fairly constant, as they are determined by the most biased experiment that is added early on during the assimilating procedure. For Whisper, a drop is noticed initially as a result of the drop in its MOS value, which is similar to the trend calculated by TSURFER. Recall that TSURFER's CM is based on the residual errors, and it is used by Whisper as an MOS margin. However, TSURFER uses a fixed value for the MOS. This explains the similar trends with a low number of experiments. As the number of experiments increases, the CM value calculated by TSURFER continues to drop, whereas that of Whisper remains constant, as it is impacted by the most negatively biased values like the nonparametric methodology. Also, notice that early on, the TSURFER results show a quick drop as more experiments with lower bias values are added. This trend is slower with the PCM methodologies because of the additional uncertainties included to hedge against modeling errors.

Figure 30 compares the calculated CM+MOS values to the true biases in a manner similar to that used in Figure 21, except that Figure 30 eliminates the TSURFER bias values to reduce the clutter. The 45-degree line is retained to demonstrate when the true bias exceeds the calculated margins. Results indicate that the PCM augmentation allows TSURFER to calculate upper-bounding margins for all the analyzed experiments while remaining approximately 1% below the Whisper margins.

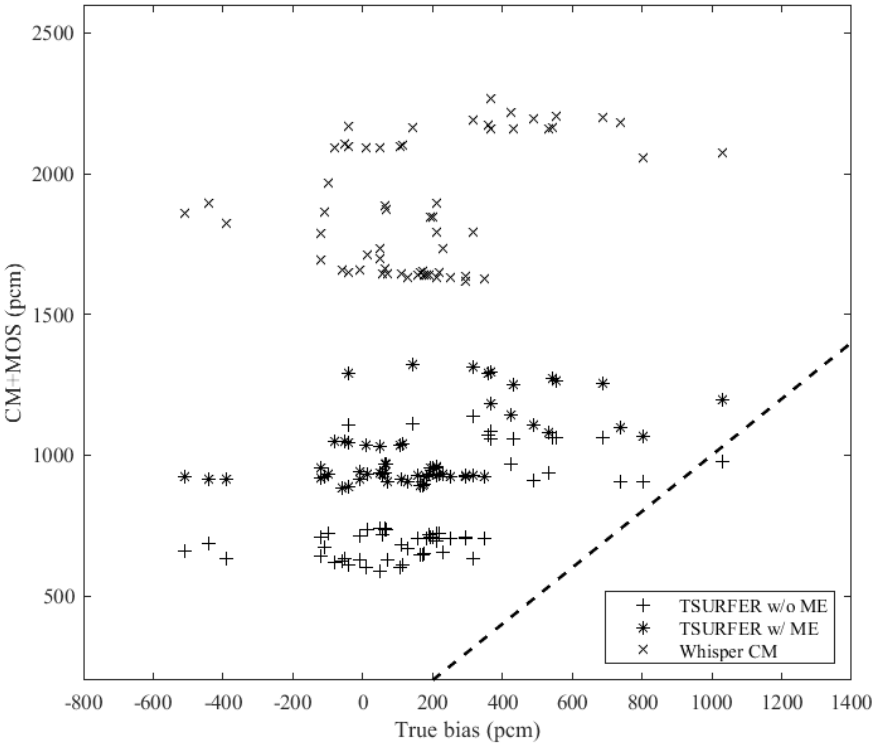


Figure 30. Margin evaluation of PCM methodology.

ACKNOWLEDGMENTS

This work was supported by the Nuclear Criticality Safety Program, funded and managed by the National Nuclear Security Administration for the Department of Energy.

7. REFERENCES

[1] H. Abdel-Khalik, D. Huang, U. Mertyurek, W. Marshall, and W. Wieselquist, "Overview of the Tolerance Limit Calculations with Application to TSURFER," *Energies* 14, 21, 7092, 2021, doi: 10.3390/en14217092.

[2] E. F. Trumble and K. D. Kimball, "Statistical Methods for Accurately Determining Criticality Code Bias," in 1997 ANS topical meeting on criticality safety challenges in the next decade, Lake Chelan, WA, September 7-11, 1997..

[3] B. C. Kiedrowski et al., "Whisper: Sensitibvity/Uncertainty-Based Computational Methods and Software for Determining Baseline Upper Subcritical Limits," *Nucl. Sci. Eng.* 181, 1, 17–47, 2015, doi: 10.13182/NSE14-99.

[4] 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*, ORNL/TM-2005/39,. 2009.

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

[6] P. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill Book Company, 1969.

[7] J. Seo, H. S. Abdel-Khalik, and A. S. Epiney, "ACCRUE—An Integral Index for Measuring Experimental Relevance in Support of Neutronic Model Validation," *Front. Energy Res.*, 9, 1–17, 2021.

APPENDIX A. NON-INTRUSIVE STOCHASTIC APPROACH

Rewrite the relevance c_k as

$$c_k = \frac{\nabla y^{\exp T} \boldsymbol{C_\alpha} \nabla y^{\text{app}}}{\sqrt{\nabla y^{\exp T} \boldsymbol{C_\alpha} \nabla y^{\exp}} \sqrt{\nabla y^{\text{app}T} \boldsymbol{C_\alpha} \nabla y^{\text{app}}}}. \tag{C-1}$$

This equation implies that the relevance is a normalized vector inner product with a weighting (or rotation) matrix. The denominator of the equation serves as a normalization factor which bounds the relevance from -1 to 1. Now, consider the numerator of the equation above.

$$\nabla y^{\exp T} \boldsymbol{C_\alpha} \nabla y^{\text{app}}. \tag{C-2}$$

Because a real symmetric matrix can be diagonalized by an orthogonal matrix, the covariance matrix $\boldsymbol{C_\alpha}$ is diagonalized as

$$\boldsymbol{C_\alpha} = \boldsymbol{U D U^T}, \tag{C-3}$$

where $\boldsymbol{U}$ is an orthonormal matrix representing a rotation, and $\boldsymbol{D}$ is a diagonal matrix representing a scaling.

Then, the numerator of the original equation can be written as

$$\nabla y^{\exp T} \boldsymbol{U \Sigma^2 U^T} \nabla y^{\text{app}} = \left(\nabla y^{\exp T} \boldsymbol{U \Sigma}\right)\left(\nabla y^{\exp T} \boldsymbol{U \Sigma}\right)^T \Rightarrow Z Z^T, \tag{C-4}$$

where Z is $\nabla y^{\exp T} \boldsymbol{U \Sigma}$, the components of which are rotated by $\boldsymbol{U}$ matrix and scaled by $\boldsymbol{\Sigma}$ matrix.

The components of Z represent the rotated sensitivity profile scaled by cross-section uncertainty, so the inner product of Z itself emphasizes the components of higher sensitivity and higher uncertainty.

APPENDIX B. INVERSE-VARIANCE WEIGHTING

In statistics, inverse-variance weighting is a method of aggregating multiple random variables to minimize the variance of the weighted average. Given a sequence of independent observations x_i with variance σ_i^2, the average weighted by arbitrary weights, w_i, which is subject to the constraint, $\sum_i w_i = 1$, can be written as

$$\hat{x} = \sum_i w_i x_i \,. \tag{C-1}$$

The variance of $\hat{x}$ is given by

$$\text{Var}(\hat{x}) = \sum_i w_i^2 \sigma_i^2 \,. \tag{C-2}$$

To minimize the variance of $\hat{x}$ with the constraint $\sum_i w_i = 1$, a Lagrange multiplier w_0 is introduced to enforce the constraint, such that

$$\mathrm{Var}(\hat{x}) = \sum_i w_i^2 \sigma_i^2 - w_0 \left(\sum_i w_i - 1 \right). \tag{C-3}$$

For $k = 1,2, \ldots,$

$$0 = \frac{\partial}{\partial w_k} Var(\hat{x}) = 2 w_k \sigma_k^2 - w_0 \Rightarrow w_k = \frac{w_0}{2\sigma_k^2}. \tag{C-4}$$

Using the constraint $\sum_i w_i = 1,$

$$\sum_i w_i = \frac{w_0}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_n^2} \right) = 1 \tag{C-5}$$

$$\Rightarrow w_0 = 2 \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_n^2} \right)^{-1}. \tag{C-6}$$

Therefore, the individual weights become

$$w_k = \frac{w_0}{2\sigma_k^2} = \frac{1}{\sigma_k^2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_n^2} \right)^{-1} = \frac{1}{\sigma_k^2} \left(\sum_i \frac{1}{\sigma_i^2} \right)^{-1}. \tag{C-7}$$

Inverse-variance weighted mean and its error

According to the obtained inversed variance weighting, the weighted mean is

$$\bar{x} = \frac{\sum_i \frac{1}{\sigma_i^2} x_i}{\sum_i \frac{1}{\sigma_i^2}}, \tag{C-8}$$

and the variance (or error in the weighted mean) is

$$\mathrm{Var}(\bar{x}) = \frac{\sum_i \left(\frac{1}{\sigma_i^2}\right)^2 \sigma_i^2}{\left(\sum_i \frac{1}{\sigma_i^2}\right)^2} = \frac{1}{\left(\sum_i \frac{1}{\sigma_i^2}\right)^2}. \tag{C-9}$$

APPENDIX C. GLLS FORMULTION

The main goal of GLLS is to consolidate knowledge from different sources, such as computer code calculations, observations, or measurements. With the linearity valid, GLLS finds the solution that minimizes the objective function subject to the constraint $k'(\alpha') = m'$, which is also represented by χ^2 with N degree of freedom, such as

$$\chi^2_N = [k' - k]^T \boldsymbol{C_k^{-1}} [k' - k] + [m' - m]^T \boldsymbol{C_m^{-1}} [m' - m], \tag{C-1}$$

where k is a prior calculated k_{eff} vector, k' is an adjusted (posterior) k_{eff} vector, and m is a measurement k_{eff} vector.

The minimizer of this objective function above may be given by

$$\Delta k = -\boldsymbol{C_k}(\boldsymbol{C_k} + \boldsymbol{C_m})^{-1} d, \tag{C-2}$$

where $\Delta k = k' - k$, and d is the discrepancy vector, $d = k - m$.

The posterior covariance matrix for the k_{eff} is given by

$$\boldsymbol{C_{k'}} = \boldsymbol{C_k} - \boldsymbol{C_k}(\boldsymbol{C_k} + \boldsymbol{C_m})^{-1} \boldsymbol{C_k}. \tag{C-3}$$

The diagonal elements of this matrix represent the adjusted uncertainty in $k_{eff.}$