Whisper: Sensitivity/Uncertainty-Based Computational Methods and Software for Determining Baseline Upper Subcritical Limits

Brian C. Kiedrowski,
Department of Nuclear Engineering and Radiological Sciences
University of Michigan
2355 Bonisteel Boulevard
Ann Arbor, MI 48109 USA

Forrest B. Brown, Jeremy L. Conlin, Jeffrey A. Favorite, Albert C. Kahler, Alyssa R. Kersting, D. Kent Parsons, Jessie L. Walker
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545 USA

Corresponding Author: Brian C. Kiedrowski (bckiedro@umich.edu)

Page Count: 99
Table Count: 4
Figure Count: 24

Whisper: Sensitivity/Uncertainty-Based Computational Methods and Software for Determining Baseline Upper Subcritical Limits

Brian C. Kiedrowski

University of Michigan

Footnote for the editor/typesetter to include: This work was performed while the first author was an employee at Los Alamos National Laboratory.

Forrest B. Brown, Jeremy L. Conlin, Jeffrey A. Favorite, Albert C. Kahler, Alyssa R. Kersting, D. Kent Parsons, Jessie L. Walker

Los Alamos National Laboratory

Abstract – Nuclear criticality safety analysis using computational methods such as Monte Carlo must establish, for a defined area of applicability, an upper subcritical limit (USL), a calculated multiplication factor k that can be treated as actually subcritical, derived from a calculational margin (combination of bias and bias uncertainty) and a margin of subcriticality. Whisper, a non-parametric, extreme-value method based on sensitivity/uncertainty-based techniques and the associated software are presented. Whisper uses benchmark critical experiments, nuclear data sensitivities from the continuous-energy Monte Carlo transport software MCNP, and nuclear covariance data to set a baseline USL. Comparisons with a traditional parametric approach for validation, which requires benchmark data to be normally distributed, show that Whisper typically obtains similar or more conservative calculational margins; comparisons with a rank-order, non-parametric approach show that Whisper obtains less stringent calculational margins.

Keywords – criticality safety ; validation ; bias ; statistics

I INTRODUCTION

American National Standards Institute/American Nuclear Society (ANSI/ANS) national standards 8.1 [1] and 8.24 [2] require that nuclear criticality safety analysts determine through validation what value of the multiplication factor k predicted by software can be treated as subcritical, i.e., an upper subcritical limit (USL). Performing a validation study for an application requires that the analyst determine the predictive capability of a method, software, and nuclear data for computing k. The discrepancy or deviation of a calculated k from its expected value, i.e., from a benchmark critical experiment, is called the bias. Since the computational methods and the benchmark experiments have associated uncertainties, these must also be factored into determining the USL. The other factor that is included in the USL is an additional margin of subcriticality (MOS) that considers other uncertainties and considerations about the application, its sensitivity of k to changes in process conditions and the suitability of the validation, and the method, software, data, etc.

The process of performing a validation study is traditionally labor intensive and therefore typically done only when necessary, e.g., when an analyst encounters a new application outside the area of applicability (AOA) of the existing validation or when updating software and nuclear data. Because of the level of effort and cost to perform a validation study, many organizations perform them much less frequently than the rate that new versions of software and nuclear data are released. Since a validation study is required to use a particular version of transport software and nuclear data libraries for criticality safety work, many organizations end up using long outdated and unsupported versions.

For example, the Nuclear Criticality Safety division at Los Alamos National Laboratory (LANL) used the same version of MCNP [3] for over a decade (MCNP5-1.25) with nuclear data libraries that were decades old, until, in 2014, a validation study was performed that allowed the division to update to MCNP6.1 and ENDF/B-VII.1 nuclear data[1] [4, 5], the

[1]Throughout this paper, the phrase "ENDF/B-VII.1 nuclear data" is shorthand for data from the ENDF/B-VII.1 nuclear data library that has been processed by NJOY99.393 into a format that is readable by and distributed with MCNP6.1.

current versions at the time. In the time period between the two updates, several bug fixes and enhancements were made that impact criticality calculations, e.g., fission source convergence diagnostics. The work outlined in this paper, i.e., the associated methods and software, was done to not only perform the validation study allowing the current update of software and nuclear data versions, but to facilitate future validation studies so that LANL criticality safety analysts may readily use new versions as they are released.

Sensitivity/uncertainty (S/U) techniques have been used to guide the selection of benchmarks for criticality safety validation for well over a decade with multigroup deterministic or Monte Carlo methods [6, 7]. Recently, continuous-energy Monte Carlo sensitivity methods [8, 9, 10, 11] have been developed and integrated into production software such as MCNP and SCALE [12, 13], with the advantage being that assessing the effect of multigroup cross section generation on the sensitivity coefficient, i.e., determining the implicit sensitivity coefficient, ceases to be a concern.

This allows for the possibility of an automated process that selects relevant benchmarks for a specific application being analyzed computationally, and therefore the calculational margin (bias and bias uncertainty) may be determined. Determining an appropriate MOS is ultimately the responsibility of the analyst, but it is also possible to automate the quantification of the effect of variability and uncertainties because of the nuclear data libraries considering the set of available benchmarks for the validation. The goal is to make validation a routine part of criticality safety evaluations where computational analysis via Monte Carlo is required (a similar method could also be developed for deterministic codes as well, but that is beyond the scope of this paper). Doing this mitigates the issue of using outdated versions of transport software and nuclear data libraries; in principle, a new version can be swapped into the normal workflow of performing a criticality safety evaluation.

The software that was developed for this purpose is named Whisper, and its computational and statistical methods, described in this paper, are termed the Whisper methodology. The software package consists of a main computational analysis package written in Fortran

2003/2008 and a few utility scripts to help with automating the running of MCNP for the creation of nuclear data sensitivity profiles.

As stated earlier, Whisper and the associated methods have already been used at LANL for a validation study involving plutonium systems, and there is associated user documentation for the software [14] distributed with Whisper and a publicly available technical report demonstrating its use for real-world applications [15]. The primary focus of this paper is to document and explain the Whisper methodology in greater detail than is appropriate in either user documentation or a nuclear criticality safety validation report and to offer illustrative analyses of a variety of hypothetical applications with comparisons to other validation methods to show that Whisper produces reasonable USLs.

This paper is structured as follows: First, the basic concepts of validation in criticality safety and two standard approaches for computing the calculational margin (CM) are reviewed in Sec. II. Next, the Whisper methodology is detailed in Sec. III; this discussion includes how the relevant quantities to compute a USL are calculated, how benchmark weight factors may be defined, how unknown benchmark uncertainties may be estimated, how benchmarks of low quality are identified and removed from the validation, and the workflow of the Whisper software and how it may fit in with criticality safety analysis. After that, aspects of what constitutes an acceptable benchmark suite for Whisper are discussed, and details of the benchmark suite distributed with Whisper and used for the results in this paper and in Ref. [15] are summarized in Sec. IV. Finally, example results are given in Sec. V for four hypothetical test cases: a determination of water-reflected plutonium critical mass at various moderation levels, the storage of fresh, low-enriched uranium (LEU) lattices in water, the storage of containers of mixed LEU/power-grade Pu solutions, and an analysis of molten salt reactor (MSR) fuel.

II VALIDATION IN CRITICALITY SAFETY

When performing criticality safety with computational methods, typically the k-eigenvalue form of the neutron transport equation is solved. This equation can be written with operator notation as follows:

$$\hat{\mathbf{\Omega}} \cdot \nabla\psi + \Sigma_t\psi = \mathcal{K}\psi + \frac{1}{k}\mathcal{F}\psi. \tag{1}$$

Here ψ is the neutron angular flux representing the number of neutrons per area, per energy, per solid angle, per time in a differential phase space element in volume, energy, and direction. $\hat{\mathbf{\Omega}}$ is the direction unit vector, Σ_t is the total interaction cross section, $\mathcal{K}$ is the integral scattering operator, and $\mathcal{F}$ is the integral fission operator. The terms have been grouped such that losses with respect to a differential element of phase space are on the left, and gains are on the right.

Typically, the gains and the losses do not balance. A factor of $1/k$ is placed upon the fission term and the value of k is found that causes the losses and gains to balance. The quantity k is purely a mathematical factor to balance an equation and should not be construed as a physical parameter or having any a priori connections with the actual physical processes in the system being analyzed. While it is possible to map the behavior of k as a function of varying physical parameters for a specific system, such results have no generality. In terms of nuclear criticality safety, k has nothing to do with upset conditions or their likelihood, and absent such a detailed study of process conditions, there is no value of k that can be universally assigned that is "safe" with regard to the prevention of criticality.

The methods described in this paper provide a baseline USL considering only the critical experiment benchmarks, transport software, and nuclear data, which are only a part of the overall analysis that goes into a criticality safety evaluation. It does not and cannot replace the role of the analyst toward setting appropriate limits, controls, etc. that consider the process in its entirety, balancing economics, human factors, material controls, safety culture, etc. and not just the computational aspects of analysis. These issues are important for the

nuclear criticality safety analyst, but are beyond the scope of this paper.

Computational methods are not the only means of performing criticality safety analysis to demonstrate subcriticality. Rather, they supplement other approaches such as experiments, standards, single-parameter subcritical limits, handbooks, and hand calculations. Computational tools are used when the results of these simpler methods require control limits that are unacceptable; these unacceptable control limits are because of the conservatisms built into or used when applying the simpler methods. Conversely, computational methods lose their usefulness when the USL becomes too low because of the lack of available benchmark data. In this case, the computational methods need to be supplemented by the other techniques.

With those caveats stated, in terms of the role of computation in criticality safety, the analyst is tasked with ensuring that the application or process being analyzed stays subcritical for all representative computational models used to describe process conditions. Mathematically, this means that the gains are less than the losses in the transport equation, i.e., $k < 1$, for the computational models being analyzed.

When using computational methods, the analyst must ensure that the value of k, a mathematical quantity, being predicted by a computational method actually corresponds to a subcritical configuration in reality. The discrepancy between computational results and reality is referred to as a bias, and it is up to the criticality safety analyst to quantify it. Sources of bias range from errors and approximations in the method or software, inadequacy of the computational model, and inaccuracies of the nuclear data used in the simulation, which is typically the dominant source. To ascertain the bias, the criticality safety analysts should assess the performance of the method's ability to predict k of relevant (i.e., having similar neutronic properties as the application being analyzed) critical experiment benchmarks. Since all experimental measurements carry uncertainty, as does the process of representing a physical system with a computational model, the bias also has an associated uncertainty that must be quantified as well. Taken together, the bias and bias uncertainty define a quantity called the calculational margin. Furthermore, taking credit for bias where

the software systematically predicts a k higher than a reference value for the AOA, referred to in this paper as a non-conservative bias, is prohibited by ANSI/ANS-8.24 unless there is "an understanding of the cause(s) of such bias."

The ANSI/ANS-8.24 national standard requires an additional margin in addition to the calculational margin to ensure that the simulated system is actually subcritical; this is called the margin of subcriticality (MOS). This paper outlines techniques for offering a baseline of this additional margin. Unlike the calculational margin, which is a more mathematically defined quantity, the MOS incorporates aspects that are related to not only the software and nuclear data but aspects of the process being analyzed. This paper only addresses the former, software and nuclear data, and leaves aspects of the process up to the analyst.

Once these two margins are known, the USL is defined as

$$\text{USL} = 1 - \text{CM} - \text{MOS}. \tag{2}$$

The computed k for an application model, k_A, plus its uncertainty, σ_A, at some chosen confidence level must be less than the the USL to assert that the application model is subcritical. Typically only statistical uncertainty from the Monte Carlo process is included in σ_A, but other uncertainties (e.g., impact of manufacturing tolerances, effects of temperature fluctuations, etc.) may be included as needed. The confidence level is chosen by the analyst (or, more typically, by the analyst's institution or regulator) and this determines a multiplicative factor n_σ applied to the computed model uncertainty σ_A. The factor n_σ is the number of standard deviations to achieve some confidence level, e.g., $n_\sigma = 2.6$ for the 99% confidence level of a normally distributed quantity—the mean k_A from the Monte Carlo calculation should be normally distributed if the simulated fission source distribution reached convergence prior to recording estimates of k_A, the neutronically relevant regions of the problem were adequately sampled, and enough samples (i.e., active cycles in MCNP parlance) of k_A were performed.

Define the amount that the application exceeds the USL for an application A as

$$\delta_A = (k_A + n_\sigma \sigma_A) - \mathrm{USL}_A. \tag{3}$$

If $\delta_A < 0$, then the analyst can be confident that the model for the application A is actually subcritical; if $\delta_A \geq 0$, then the analyst cannot. Conversely, $\delta_A = 0$ does not imply that the application is critical to any degree of confidence, and neither does $\delta_A > 0$ mean that the application is necessarily supercritical. In either case ($\delta_A = 0$ or $\delta_A > 0$), the criticality of the application model is ambiguous.

II.A Standard Approaches for Determining the Calculational Margin

A standard approach, which is outlined in Ref. [16], is briefly reviewed here for comparison to the Whisper methodology. First, the literature is reviewed and benchmarks that are similar (either qualitatively or quantitatively based on physical, neutronic properties) to the application case being analyzed are selected. Let N be the number of benchmarks in a set, and once this set is known, the bias and bias uncertainty may be determined.

Two standard methods are presented here. The first is referred to as the parametric method, which requires that the benchmark data be normally distributed. The second is the non-parametric or rank-order approach, which has no restriction on how the benchmark data are distributed.

Normality of the benchmark data is assessed using statistical tests. A popular approach in criticality safety and other fields that use statistical analysis is the Shapiro-Wilk normality test. Alternatively, the χ^2 test may be used when the sample size is large (Ref. [16] recommends greater than 50 benchmarks).

II.A.1 Parametric method

The parametric method may be used if the k of the benchmarks in the set are normally distributed.

Define $\tilde{k}_i$ as the scaled multiplication factor of the ith benchmark that is the calculated k divided by the benchmark k. The latter is often, but not always, unity. This assumes the bias in the calculation is unaffected by the scaling, which is valid for small differences in k.

Let σ_i be the uncertainty for the benchmark, which is the square root of the sum-in-quadrature of the benchmark and calculational (Monte Carlo statistical) uncertainties. The weight factor for each benchmark is the inverse of the variance,

$$w_i = \frac{1}{\sigma_i^2}, \tag{4}$$

and W is the sum of all w_i.

The mean multiplication factor $\bar{k}$ is determined from a simple weighted average:

$$\bar{k} = \frac{1}{W} \sum_{i=1}^{N} w_i \tilde{k}_i. \tag{5}$$

The bias β is then

$$\beta = \bar{k} - 1. \tag{6}$$

When the bias is negative, the calculation tends to predict values of k that are lower than reference values. Likewise, a positive bias means that the calculation tends to predict values of k that are higher than the reference values.

Recall that taking credit for non-conservative bias is usually not permitted. For this reason, a non-conservative bias adjustment parameter,

$$\Delta_m = \max\{0, \beta\}, \tag{7}$$

will normally be added to the calculational margin.

Next, the bias uncertainty must be estimated. This is typically done with a quantity called the pooled variance, σ_β^2, which is the quadrature sum of the weighted variance in k about the mean, s_k^2 and the average variance of k, $\bar{\sigma}_k^2$:

$$\sigma_\beta = \sqrt{s_k^2 + \bar{\sigma}_k^2}. \tag{8}$$

The weighted variance in k about the mean is the weighted standard deviation,

$$s_k^2 = \frac{1}{W}\frac{N}{N-1}\sum_{i=1}^{N} w_i \left(\tilde{k}_i - \bar{k}\right)^2, \tag{9}$$

and the average variance of k is

$$\bar{\sigma}_k^2 = N\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)^{-1} = \frac{N}{W}. \tag{10}$$

The bias uncertainty σ_β must then be multiplied by the single-sided tolerance factor κ to obtain the interval that guarantees a certain percentage p of the benchmarks will be bounded to some confidence level q. The factor κ may be determined by

$$\kappa = \frac{t\left[q, N-1, z(p)\sqrt{N}\right]}{\sqrt{N}}, \tag{11}$$

where t is the inverse of the cumulative distribution function (CDF) of the noncentral t-distribution with probability q, degrees of freedom $N-1$, and noncentrality parameter $z(p)\sqrt{N}$; z is the inverse of the standard normal CDF with probability p. For the results presented in this paper with this method, q and p are 0.99.

Because the single-sided tolerance factor from Eq. (11) requires the inverse of the noncentral t-distribution, which is not available in many statistical analysis or spreadsheet soft-

ware packages, often an asymptotic form [17] is used,

$$\tilde{\kappa} = \frac{z(p) + \sqrt{z^2(p) - ab}}{a}, \tag{12}$$

where

$$a = 1 - \frac{z^2(q)}{2\,(N-1)} \tag{13}$$

and

$$b = z^2(p) - \frac{z^2(q)}{N}. \tag{14}$$

The asymptotic single-sided tolerance factor $\tilde{\kappa}$ from Eq. (12) is a good approximation of κ from Eq. (11) when N is large (i.e., $N > 20$), but deviates significantly for small sample sizes. The results presented in this paper use the rigorous κ from Eq. (11).

The calculational margin is defined as the sum of the bias, bias uncertainty times the single-sided tolerance factor, and a non-conservative bias adjustment parameter:

$$\mathrm{CM} = -\beta + \kappa\sigma_\beta + \Delta_m. \tag{15}$$

If the benchmarks are normally distributed, then the parametric method described can not only be used, but trending of the calculational margin on physical parameters (e.g., uranium enrichment, spectral indices, etc.) may be done as well. Since trending with this method is not performed in this paper, the discussion of how to perform the analysis is not included in this paper, but a description may be found in Ref. [16].

II.A.2 Non-parametric, rank-order method

If the benchmark k are found to not be normally distributed, then the calculational margin must be determined non-parametrically. One standard method in criticality safety is discussed in Ref. [16], which is a rank-order approach that determines a confidence level that a fraction of the population of benchmarks is above the smallest, i.e., worst case, $\tilde{k}$.

The bias for the non-parametric, rank-order method is

$$\beta = \min\left\{\tilde{k}_i\right\} - 1. \tag{16}$$

The bias uncertainty is the total uncertainty, i.e., benchmark and statistical uncertainties summed in quadrature, at some confidence level, i.e., multiplied by some factor n_σ, of the benchmark corresponding to the minimum $\tilde{k}$.

The baseline calculational margin is the combination of the bias and bias uncertainty of the benchmark with the minimum $\tilde{k}$ and an additional non-parametric margin m_{NP}. As before, no credit is normally taken for non-conservative bias by applying the non-conservative bias adjustment parameter Δ_m.

To determine m_{NP}, first the non-parametric confidence level C_{NP} is obtained by

$$C_{NP} = 1 - p^N, \tag{17}$$

where p is the desired population fraction, typically 0.95 for this method. Based on the value of C_{NP}, Ref. [16] recommends values for the non-parametric margin m_{NP} ranging from 0 to 0.05 for $p = 0.95$. These values are reproduced in Table I.

The final calculational margin for the non-parametric approach is the same as in Eq. (15) except that it is further increased by the non-parametric margin m_{NP}.

As will be seen, the Whisper methodology, being based on the worst-case discrepancy in k, is similar to the non-parametric, rank-order method. The standard non-parametric method is simply a rank-order approach that does not take into account the discrepancy in the benchmark k and their uncertainties other than that of the worst case. It is also not clear how weighting of the benchmarks to account for the degree of relevance to the application being analyzed can be performed. The Whisper methodology considers these.

One aspect that was glossed over in this discussion of either the standard parametric or non-parametric methods is the selection of relevant benchmarks. This often represents the

largest up-front cost in terms of time for the criticality safety analyst and often, in practice, makes validation an unpalatable exercise for an institution to undertake. The Whisper methodology may also help circumvent this barrier.

II.A.3 Discussion of other approaches and tools

The approaches discussed in Secs. II.A.1 and II.A.2 are certainly not the only available methods for determining a baseline USL. These methods are detailed because the results of those two methods are compared with those from Whisper. Before proceeding with describing Whisper, it is instructive to provide a list, which is certainly not exhaustive, of other approaches and tools for criticality safety validation.

An overview of some methods and comparisons of the results from those may be found in Ref. [18], which includes the two methods [therein called the Washington Safety Management Solutions (WSMS) method] discussed in this paper. Also included in this discussion are the two methods available in the USLSTATS software [19] maintained by ORNL and the four methods [20] included included in the USLSA software [21] from General Electric (GE). Trending analysis is also possible with various methods and software tools such as PARANAL [22] from GE and MACSENS [23] from Institut de Radioprotection et de Surete Nucleaire (IRSN).

Whisper uses the generalized linear least squares method discussed in Ref. [6] to adjust nuclear data for determining a MOS for nuclear data variability (see Sec. III.B.2). These techniques may also be used to estimate the bias and bias uncertainty for the calculational margin as well. It is perfectly valid to do as such, but Whisper does not employ them for this purpose. Rather a different approach (described in Sec. III) is used based on the extreme-value methods. This avoids potential concerns with how the bias computed from a nuclear data adjustment is randomly distributed and whether a linear regression model is appropriate. How the bias is randomly distributed matters when determining the calculational margin, as the appropriate confidence level or single-sided tolerance factor

to apply to the similarly computed bias uncertainty must be decided. Whisper is a non-parametric approach for determining the calculational margin, and tries to avoid these issues.

III WHISPER METHODOLOGY

The Whisper methodology estimates the calculational margin and minimum MOS to determine a baseline USL. The term "baseline" implies that the USL should be viewed as an upper bound for the actual USL applied in a criticality safety evaluation. The nuclear criticality safety analyst retains the responsibility of ensuring subcriticality, as computational tools or analytical techniques, while necessary in many criticality safety evaluations, cannot do that by themselves. Ultimately, the baseline USL is a guide to help the analyst, and he or she may feel compelled to apply additional margin where the baseline is thought, based upon expert engineering judgment, to be insufficient.

The Whisper methodology for computing the calculational margin is non-parametric; it does not require the biases of the benchmark population to follow a normal (or any other) distribution. It also automatically selects and weights benchmarks from a large validation suite that are most neutronically similar to a particular application. For the MOS, the Whisper methodology includes the uncertainties from nuclear covariance data after an adjustment that considers the benchmarks available in the suite. The Whisper methodology may even do trending on both of these quantities by considering a set of application models where the trending parameter is varied; based on the results from multiple application models, fits of the behavior may be made by the criticality safety analyst.

How the Whisper methodology determines the calculational margin and MOS, and therefore the baseline USL, is discussed next. This is followed by an explanation of how Whisper estimates unknown benchmark uncertainties and uses a nuclear data adjustment to reject poor quality benchmarks from the validation. Then, a brief discussion of the computational workflow that a criticality safety analyst would perform to use the Whisper software is given.

III.A Calculational Margin

The calculational margin includes the biases and bias uncertainties of relevant benchmarks, each having a weight w_i based on the degree of relevance with respect to the particular application being analyzed. There are numerous possible statistical methods for computing these quantities depending upon the definitions made. One conservative definition, which is used by Whisper, is to find the value of a calculational margin that would bound the worst-case (i.e., most negative) bias to some probability or confidence interval of a weighted population. The attractiveness of this definition is that the addition of less relevant benchmarks to the set being used in the validation cannot decrease the calculational margin, and therefore a very broad set of benchmarks may be used for performing the validation without concern that irrelevant benchmarks will somehow non-conservatively affect the result. Conversely, because of the weighting based on the degree of relevance, the addition of irrelevant benchmarks will also not produce an overly conservative calculational margin.

This definition is now expressed mathematically. The mathematics is expressed in terms of the opposite-signed bias, such that negative bias (i.e., k predicted lower than reference values) leads to positive values. Using the opposite-signed bias simplifies the mathematical descriptions of the method. To begin, suppose that the experimental benchmarks are independent and that benchmark i has an opposite-signed bias X_i, a random variable distributed with CDF $F_i(x)$. Let random variable

$$X = \max\{X_1, \ldots, X_N\}. \tag{18}$$

The CDF for X is the product of the individual CDFs, i.e.,

$$F(x) = \mathbb{P}(X \leq x) = \prod_{i=1}^{N} F_i(x). \tag{19}$$

Here x is selected to satisfy $F(x)$ equal to some probability or one-sided confidence interval.

The corresponding probability density function (PDF) is

$$f(x) = \frac{dF}{dx} = \sum_{i=1}^{N} f_i(x) \prod_{\substack{j=1 \\ j \neq i}}^{N} F_j(x) = F(x) \sum_{i=1}^{N} \frac{f_i(x)}{F_i(x)}. \tag{20}$$

To illustrate Eq. (20) and the resulting distribution, consider the case of two PDFs $f_1(x)$ and $f_2(x)$ that are normally distributed as $\mathcal{N}_1(0,1)$ and $\mathcal{N}_2(\frac{1}{4},\frac{1}{4})$ respectively, and let $f(x) = \max\{x_1, x_2\}$.

Figure 1 gives curves for the density functions $f_1(x)$ and $f_2(x)$ with dashed lines and the maximum distribution $f(x)$ with a solid one. There are a few observations to make about $f(x)$ when the $f_i(x)$ are normally distributed. First, $f(x)$ is not normally distributed, and therefore many of the standard statistical techniques, which assume normality, do not apply to $f(x)$. The mean of $f(x)$ also exceeds both the means of $f_1(x)$ and $f_2(x)$; this property is true for an arbitrary number of normal distributions $f_i(x)$. Another property is that as N increases, the mean of $f(x)$ always increases; this is even true if all the $f_i(x)$ are identical normal distributions, but then the mean of $f(x)$ increases very slowly for large N. $f(x)$ is also not symmetric; it has a positive skewness, and this property makes using $f(x)$ attractive for criticality safety as it always leads to conservative results. A potential drawback is that a single function $f_j(x)$ with a very large mean relative to the other $f_i(x)$ almost completely determines $f(x)$ [where $f(x)$ is approximately $f_j(x)$], which can lead to an overly conservative and constraining result. For this reason, a method for the selection and weighting of the $f_i(x)$ is needed.

Using normal distributions as an example is not merely an illustrative pedagogical choice; for critical experiment benchmarks, the multiplication factor k for benchmark i is usually normally distributed about its quoted mean benchmark value $k_{i,\text{bench}}$ and benchmark uncertainty $\sigma_{i,\text{bench}}$. The computational method predicts a value $k_{i,\text{calc}}$ and, if it is a Monte Carlo

method, a calculated uncertainty of the mean $\sigma_{i,\text{calc}}$. The mean bias β_i is defined as

$$\beta_i = k_{i,\text{calc}} - k_{i,\text{bench}}. \tag{21}$$

The uncertainty in the bias of the ith benchmark is the square root of the sum in quadrature of the benchmark and calculated uncertainties,

$$\sigma_i = \sqrt{\sigma_{i,\text{bench}}^2 + \sigma_{i,\text{calc}}^2}. \tag{22}$$

The CDF for the normal distribution for the opposite-signed bias $-\beta_i$ given weight w_i is

$$F_i(x) = (1 - w_i) + \frac{w_i}{2}\left[1 + \text{erf}\left(\frac{x + \beta_i}{\sqrt{2\sigma_i^2}}\right)\right]. \tag{23}$$

Note the plus sign in the numerator of the error-function argument. The CDF of the normal distribution is x minus the mean; but, recall this distribution describes the opposite-signed bias, and hence the plus sign.

The weighting factor w_i biases the CDF of the normal distribution, and except when $w_i = 1$, $F_i(x)$ may no longer be interpreted as the probability that the opposite-signed bias of benchmark i, a random variable, is less than a number x. Accordingly, the weights w_i used in the $F_i(x)$ of Eq. (19), the extreme value CDF $F(x)$, are scaled such that the largest w_i is always one; this ensures $F(x)$ satisfies the mathematical requirements of being a CDF, even if, strictly speaking, all but one of the individual $F_i(x)$ do not.

Given the benchmark data and a set of weight factors, it is then possible to use Eq. (19) to compute the calculational margin to some confidence level. This is discussed next, followed by an explanation of a method for assigning weight factors and a discussion of a possible approach to handle small sample sizes.

III.A.1 Computation of bias, bias uncertainty, and calculational margin

The calculational margin is, again, the combination of the bias and bias uncertainty. Perhaps ironically, the calculational margin with the extreme value distribution can be well defined, whereas the definitions of bias and bias uncertainty are somewhat arbitrary. Where the definition of bias becomes important is when following the convention of not taking credit for non-conservative bias. Therefore, the unadjusted calculational margin (where no consideration has been made for non-conservatively crediting bias) is defined first to be the value of m such that the extreme value CDF

$$F(m) = q. \tag{24}$$

Typical values of q are 0.95 or 0.99; for this paper, a value of 0.99 is used throughout. The quantity m corresponds to the calculational margin that bounds the worst-case benchmark bias and bias uncertainty to probability q.

The bias represents, for a given AOA, the systematic deviation of calculated results from their respective reference values. In statistics, bias is typically defined as the expected value or mean of the systematic deviation. The parametric approach in Sec. II.A.1 uses the sample mean as the bias, whereas the non-parametric, rank-order approach in Sec. II.A.2 defines the bias more conservatively by using the largest deviation under-predicting the reference k. Whisper, based on extreme values, combines the two definitions. The opposite-signed bias is defined as the mean of the extreme value PDF, or equivalently

$$\beta = -\int_{-\infty}^{\infty} x f(x) dx = -\int_{-\infty}^{\infty} x F(x) \sum_{i=1}^{N} \frac{f_i(x)}{F_i(x)} dx. \tag{25}$$

Different definitions of bias are certainly permissible and may be debated, but this definition is consistent with the methods presented in Sec. II.A.

If the PDFs, $f_i(x)$, are normally distributed, as they are for the benchmark critical

experiments, there is no known analytic form for β in Eq. (25), so the integral must be approximated numerically. In Whisper, this integration is performed with the trapezoid rule. The integration bounds are determined as the points at which the CDF is below some tolerance ϵ_x and above $1 - \epsilon_x$. The integration points are refined until the result of the integral converges to within some tolerance ϵ. The default values for these tolerances in Whisper, and the ones used to compute the Whisper bias results quoted in this paper, are $\epsilon_x = 1 \times 10^{-6}$ and $\epsilon = 1 \times 10^{-9}$; sensitivity studies were performed showing that smaller values for either of these have a negligible effect on the integration results.

The bias uncertainty follows from how the calculational margin is defined:

$$\sigma_\beta = m + \beta, \tag{26}$$

which is simply a number representing the amount that the unadjusted calculational margin exceeds the opposite-signed bias. Since the extreme value distribution is not normally distributed, such rules as, e.g., 2σ representing a 95% confidence level do not apply. The bias uncertainty is computed by Whisper, but never used directly. This is consistent with ANSI/ANS-8.24, which states, "Individual elements (e.g., bias and bias uncertainty) of the calculational margin need not be computed separately. Methods may be used that combine the elements into the calculational margin."

The final calculational margin is determined by adding the non-conservative adjustment parameter Δ_m from Eq. (7) to m,

$$\text{CM} = m + \Delta_m. \tag{27}$$

III.A.2 Selection of benchmark weights

In performing a validation for a specific application, benchmarks should be selected that have similar neutronic properties that are important to the application's multiplication factor

k. More precisely, for an adequate determination of computational bias of an application, the benchmarks selected should share similar sources of bias in their underlying neutronic properties. The assertion is that the dominant source of computational bias in critical experiment benchmarks (which implies that the experiment and descriptions thereof are accurate and of high quality) is from uncertainties in the nuclear data. This implies that some quantity should be used that considers both what nuclear data matters most toward determining k, i.e., what nuclear data is k most sensitive to, and the uncertainty of that nuclear data. A traditional parameter that has been used in the last several years [6] is the correlation coefficient c_k, which convolves both the nuclear data sensitivity coefficients and covariance data. Based on the property that c_k describes a common source of computational bias, it may be used to assign a weight factor to compute the calculational margin.

To explain c_k, the sensitivity coefficient is introduced. The sensitivity coefficient with respect to the effective multiplication k for some nuclear data x (an isotope, reaction, and energy range) is defined as the ratio of the relative differential change in k to the relative differential change in x and can be written as

$$S_{k,x} = \frac{x}{k}\frac{\partial k}{\partial x}. \tag{28}$$

The sensitivity coefficients may be obtained using direct perturbations or adjoint-based methods. The former approach involves changing each nuclear data x individually by some small fraction, and the relative change in k is found by comparing the perturbed calculated value of k to a reference k. Performing direct perturbations is a robust approach for determining the sensitivity coefficient but is rather cumbersome in practice when there is a large number of nuclear data that must be considered, as there always is in performing a validation study. The alternative uses adjoint-based perturbation theory, which can estimate an arbitrary number of sensitivity coefficients in a single calculation. The sensitivity coefficients are

estimated with the following ratio of adjoint-weighted integrals:

$$S_{k,x} = -\frac{\left\langle \psi^\dagger, \left(\Sigma_x - \mathcal{K}_x - k^{-1}\mathcal{F}_x\right)\psi \right\rangle}{\langle \psi^\dagger, \mathcal{F}\psi \rangle}. \tag{29}$$

Here ψ is the angular neutron flux, $\psi^\dagger$ is its adjoint function, Σ_x is the macroscopic interaction cross section for nuclear data x, $\mathcal{K}_x$ is the scattering operator for nuclear data x, $\mathcal{F}_x$ is the fission operator for nuclear data x, and the brackets denote integration over all phase space. When the nuclear data x is not a cross section (e.g., fission ν), $\Sigma_x = 0$, when it does not involve scattering, $\mathcal{K}_x = 0$, and when it does not involve fission, $\mathcal{F}_x = 0$. Estimating the adjoint-weighted integrals with Monte Carlo may be done using various techniques. In terms of automation, the iterated fission probability method is attractive because it is typically robust and involves minimal interaction on the part of the analyst [24].

The sensitivity coefficients are organized into a sensitivity row vector $\mathbf{S}$. The corresponding nuclear data (relative) covariance matrix $\mathbf{C}_{xx}$ is obtained from the nuclear data evaluation library. The covariance in k for systems A and B with respective sensitivity row vectors $\mathbf{S}_A$ and $\mathbf{S}_B$ may be found by

$$\mathrm{Cov}_k(A, B) = \mathbf{S}_A \mathbf{C}_{xx} \mathbf{S}_B^\top. \tag{30}$$

The variance is the special case when the two systems are the same, e.g.,

$$\mathrm{Var}_k(A) = \mathbf{S}_A \mathbf{C}_{xx} \mathbf{S}_A^\top. \tag{31}$$

The correlation coefficient c_k for systems A and B is

$$c_k(A, B) = \frac{\mathrm{Cov}_k(A, B)}{\sqrt{\mathrm{Var}_k(A)}\sqrt{\mathrm{Var}_k(B)}}. \tag{32}$$

As with normal correlation coefficients, a c_k of one implies perfect (linear) correlation, and

A and B can be thought to have identical sources of bias. If c_k is zero, then the two systems are completely dissimilar. Negative c_k indicates linear anticorrelation, meaning that the sources of bias between A and B are common, but change k in opposite directions; having a significantly large negative c_k is uncommon in practice for criticality and is usually treated as zero correlation; Whisper sets negative c_k to zero.

To turn c_k into a weight factor, define the maximum and the acceptance c_k, $c_{k,\max}$ and $c_{k,\text{acc}}$ respectively. The maximum c_k is the largest correlation coefficient in the benchmark population for a given application. The acceptance c_k is chosen to ensure that an adequate sample size (total sample weight) has been chosen, i.e.,

$$\sum_i w_i = w_{\text{req}}, \tag{33}$$

where the sample weight and the required weight w_{req} are to be defined.

Let the weighting factor be given by the linear relation

$$w_i = \max\left\{0, \frac{c_{k,i} - c_{k,\text{acc}}}{c_{k,\max} - c_{k,\text{acc}}}\right\}, \tag{34}$$

which assigns a weight for the most similar benchmark as one, and a benchmark with a c_k right at the acceptance c_k or below as zero. The choice of this linear relation is arbitrary, but is, however, based on a couple of observations. The first is that the correlation coefficient c_k captures linear correlation and is based upon a first-order Taylor series approximation for uncertainty propagation, and therefore a linear relationship on c_k seems appropriate. Secondly, empirical studies performed in the development of this method (not presented in this paper) show that this choice of function for the weight factor, as opposed to other possibilities that were considered such as simply scaling by $c_{k,\max}$, usually results in relatively smooth variations in the calculational margin with respect to variation of physical parameters, e.g., reflector thickness, fissile concentration, etc.

From this definition, a function for an adequate sample weight w_{req} within a population

may be chosen. A logical supposition is that the more similar a benchmark in the set is to the application, the fewer overall are needed for the validation; heuristically, this can be thought of as having a higher confidence that the sample chosen is representative of the application. Conversely, the required weight should increase when there is a lack of a very close match to perform the validation. For these reasons, the following function is selected:

$$w_{\text{req}} = w_{\min} + w_{\text{penalty}} \left(1 - c_{k,\max}\right). \tag{35}$$

Here $w_{\min}$ represents the minimum sample weight that the analyst allows for the validation. A value of 25 is selected for the results in this paper, and is based on statistical "rules of thumb" for adequate sample sizes of unweighted populations. The quantity w_{penalty} represents a penalty factor for not having a benchmark that agrees with the application. The value should be chosen such that the required sample size grows with a lower $c_{k,\max}$, but does not grow so quickly as to encompass the entire benchmark suite. Based upon the studies in this paper, an empirical value of 100 is chosen for w_{penalty}.

The process for generating weight factors is as follows: First, the $c_{k,i}$ are computed for all benchmarks with respect to an application. Based on this, the required weight w_{req} is calculated from Eq. (35). A value of $c_{k,\text{acc}}$ is selected, and the weight factors for each benchmark using Eq. (34) are computed. The criterion in Eq. (33) is checked to within some tolerance. If the criterion is not met, a new value of $c_{k,\text{acc}}$ is chosen until it is. Once the criterion is met, the sample weights are used to compute the calculational margin.

Using c_k as a similarity coefficient to compute a weighting factor representing the degree of relevance is not the only approach possible. A possible alliterative that warrants investigation is a coefficient derived from the mutual information, which, unlike c_k, has no restrictions on the how the bias of each individual benchmark is distributed [25]. Others are certainly possible is well, and a comparison of different similarity coefficients and methods for generating weighting factors is an open area of research for Whisper and similar methods.

III.A.3 Handling small sample sizes

It is conceivable that the criterion in Eq. (33) cannot be met because the number of similar benchmarks in the set is too few. To address this difficulty, an interpolation to a conservative calculational margin may be performed, i.e., a non-coverage penalty is applied. The non-coverage penalty may be based on some user-defined value or, lacking that input from the user, one that Whisper bases on the most biased benchmark in the set.

Let m_0 be the calculational margin that would be computed had all benchmarks been assigned unit weight, i.e., no weighting performed. This sets an upper bound for a calculational margin based upon a benchmark suite that is assumed to cover a wide application space, but perhaps not for the specific properties of the one being analyzed. An intermediate calculational margin $\tilde{m}$ is computed using Eq. (24) based upon the weight factors that were found, which have a sum w_{sum}. The calculational margin m is then determined from the interpolation

$$m = \tilde{m}\frac{w_{\text{sum}}}{w_{\text{req}}} + m_0\left(1 - \frac{w_{\text{sum}}}{w_{\text{req}}}\right). \tag{36}$$

Note that when the weight requirement criterion is met, the m found by Eq. (36) reduces to the value that would have been obtained had no interpolation been performed.

The extreme case is when all of the benchmarks in the set are too dissimilar from the application, which does indeed arise in some circumstances. In this case, there are really no good quantitative measures to establish the calculational margin and any result from Whisper (or any such computational method) should be viewed with skepticism. Whisper will default to using the user-defined value or the most biased benchmark as before to establish the non-coverage penalty. To establish the baseline USL, the MOS for nuclear data variability, as discussed in Sec. III.B.2, is applied in addition to the calculational margin as usual; however, in this case this applied MOS is effectively the unadjusted uncertainty (times some multiple to reflect the desired degree of confidence) in k from nuclear data covariances, which usually (but not always) at the 1-σ level bounds the computational bias. The baseline

USL is further lowered by the fixed MOS for software errors discussed in Sec. III.B.1. These factors taken together leads to the Whisper-reported value for the baseline USL. It is then left to the analyst to decide whether to use the value, further lower it as appropriate, reject it completely in favor of other approaches for determining a USL, or apply non-computational techniques to the criticality safety analysis.

Before proceeding to the discussion of how Whisper assigns the baseline MOS, a few comments on m_0 of Eq. (36) are in order. The value of m_0 is typically a large conservative number dominated by the worst-case benchmark in the entire suite. If the suite is very large, spanning numerous areas of applicability, the value of m_0 can help address the question of how poorly software would predict an arbitrary criticality safety application; that is the basis for interpolating to m_0. The benchmark suite presented in this paper (see Sec. IV) does have a wide range of benchmarks and m_0 is 0.049, which for most applications is reasonably conservative. This approach should be reasonable when there are some similar benchmarks, but not enough in the entire set to meet the criterion in Eq. (33), In the extreme case of no similar benchmarks, however, this may not be sufficient and other actions may need to be taken on the part of the analyst.

III.B Margin of Subcriticality

The margin of subcriticality (MOS) is an additional margin prescribed by the ANSI/ANS-8.24 standard "that is sufficiently large to ensure that the calculated conditions will actually be subcritical." The standard itself does not prescribe how the MOS should be defined, other than to specify that it "should take into account the sensitivity of the system or process to variations" of relevant physical parameters. These statements, taken together, suggest that the criticality safety analyst should study how both the calculated k and the predictive capability of the method and data used for calculating k vary with those physical parameters for the model being analyzed. The MOS should therefore be set in such a way to be bounding and ensure subcriticality for all credible variations in physical parameters

representative of process conditions being analyzed.

Beyond this, additional considerations about the fidelity of the computational software and nuclear data libraries should be considered. These two factors are addressed in this paper because they pertain to the computational techniques and nuclear data, whereas everything else about the application is necessarily left to the criticality safety analyst. Mathematically, the calculational margin can be broken up into three terms,

$$\text{MOS} = \text{MOS}_{\text{software}} + \text{MOS}_{\text{data}} + \text{MOS}_{\text{application}}. \tag{37}$$

The first two of these terms are discussed in turn.

III.B.1 Margin for software errors

Setting a margin to bound the impact of software errors is a qualitative judgment about the reliability and maturity of the software and computational method. As far as computational methods go, continuous-energy Monte Carlo calculations use a high-fidelity representation of the underlying nuclear interactions (no multigroup approximation) with no required approximations from meshing and homogenization of geometric zones. Therefore the intrinsic uncertainty (as opposed to the statistical uncertainty, which is estimated by the software and factored into the bias uncertainty) in k arising from approximations and limitations of the Monte Carlo method itself should be comparatively very small. In practice, analysts often introduce geometric approximations to simplify the creation of the computational models and this would need to be factored into the MOS on a case-by-case basis.

The software for performing a continuous-energy Monte Carlo simulation, however, is complicated and coding errors (i.e., bugs) are inevitable. The frequency and severity of the coding errors is a function of the age of the software, current level of support, current and historical number of users, the degree to which the software has had relevant verification and validation performed, etc.

Based on these considerations, a "detection limit" in k for errors in software can be estimated. The detection limit quantifies the degree, in terms of k, that errors in the coding would produce an incorrect answer that would not be noticed or would not have been noticed already and therefore fixed by the software developers. It is incredibly difficult, and arguably impossible, to quantify this using mathematical and statistical techniques. Rather, the best one can probably hope for in developing a number for this effect is to rely on the experience of experts familiar with both the development of the criticality software and performing calculations with that software. For MCNP6.1, it is the expert opinion of the software developers, considering that the results of k have intentionally not changed substantively in years, that a value of $\mathrm{MOS}_{\mathrm{software}}$ of 0.005 is a reasonably conservative detection limit for the effect of such errors, and this number is used within the development of a baseline USL for this paper.

The exact number can be debated, and different software packages would have different values based upon the considerations mentioned. In determining this number, the criticality safety analyst should consult with the software developers and, if possible, consider the magnitude of such errors that may have historically occurred when using the particular transport software within his or her organization. In either case, it is also the expert judgment of the MCNP developers, considering the physical approximations in state-of-the-art Monte Carlo methods, that a detection limit significantly lower than 0.005 for any Monte Carlo software would be, at present, very difficult to defend.

In principle, a margin could also be devised for deterministic methods. Doing this would be more complicated because the effects of the errors resulting from spatial discretization, energy group collapse (e.g., self shielding of groupwise cross sections), and angular approximations (e.g., quadrature set, spherical harmonics order, etc.) need to be quantified. This applies to both the calculations of k and the sensitivity coefficients. With regard to the sensitivity coefficients, implicit sensitivity coefficients that correct for the effect of group collapse need to be calculated, and the robustness of the techniques for doing this also needs to

be considered. This paper does not attempt to quantify a margin for deterministic software, which would need to be done to apply the Whisper methodology with deterministic methods.

III.B.2 Nuclear data uncertainty/variability

The nuclear data covariance matrix $\mathbf{C}_{xx}$ represents the prior uncertainties of the nuclear data from differential measurements and theoretical models. When preparing a nuclear data evaluation, integral measurements, e.g., critical experiment benchmarks, are used to constrain the choices for cross sections as well, i.e., evaluators do not allow changes to nuclear data libraries that have too adverse an effect on the agreement between calculations and critical experiment benchmarks. Because of this fact, there exists a dependency between the nuclear data and the critical experiment benchmarks, and the actual nuclear data uncertainties are therefore lower than the differential measurements alone would indicate.

A generalized linear least squares (GLLS) method can adjust the nuclear data considering both the nuclear data covariances and the benchmark experiment results. Using the adjusted nuclear data libraries can provide a more realistic estimate of the uncertainty from the nuclear data than simply using the nuclear data covariances to propagate uncertainties directly. This adjusted nuclear data uncertainty may then be used to set a portion of the baseline MOS.

A summary of the GLLS method for nuclear data adjustment is provided here. A more comprehensive overview of the GLLS technique and its development can be obtained in Ref. [6]. These techniques are currently implemented in the TSURFER module of SCALE [13].

Given the sensitivity vector (first derivatives) and the covariance matrix for the benchmark experiments, it is possible to perform an adjustment of the nuclear data that minimizes the bias within the constraints of the covariance data. Specifically, the GLLS technique minimizes the χ^2 statistic, which is a quadratic function that sums the adjusted relative differences of the calculated k values from their respective reference benchmark values and

the proposed change in the nuclear data from their means:

$$\chi^2 = [\Delta\mathbf{k}]^\top \mathbf{C}_{kk} [\Delta\mathbf{k}] + [\Delta\mathbf{x}]^\top \mathbf{C}_{xx} [\Delta\mathbf{x}] . \tag{38}$$

$\Delta\mathbf{k}$ is a column vector representing the relative difference in k for each benchmark after the adjustment predicted by the sensitivity coefficients, $\mathbf{C}_{kk}$ is the relative covariance matrix of the benchmark experiments including benchmark correlations where they are known, $\Delta\mathbf{x}$ is a column vector with elements that are the relative difference of the nuclear data from their means, and $\mathbf{C}_{xx}$ is the relative covariance matrix for the nuclear data. With no adjustment, the $\Delta\mathbf{x}$ are all zero so the χ^2 is determined solely from the differences in the k of the benchmark experiments from their reference values. As the data are adjusted, the differences in the benchmark k may decrease, therefore decreasing the first term in Eq. (38); however, this also increases the differences in the nuclear data and therefore increases the second term in Eq. (38).

The goal of the GLLS adjustment is to find the nuclear data adjustment that minimizes χ^2 by balancing these two competing effects. The minimum χ^2 is

$$\chi^2_{\text{min}} = \mathbf{d}^\top \mathbf{C}_{dd}^{-1} \mathbf{d}, \tag{39}$$

where $\mathbf{d}$ is a column vector of relative differences in the calculations from benchmark experiments, and $\mathbf{C}_{dd}^{-1}$ is the inverse of the covariance matrix of the relative difference vector. The covariance matrix of the difference vector for a set of benchmark experiments can be found using

$$\mathbf{C}_{dd} = \mathbf{B}\mathbf{C}_{kk}\mathbf{B} + \mathbf{S}_{B,kx}\mathbf{C}_{xx}\mathbf{S}_{B,kx}^\top. \tag{40}$$

$\mathbf{B}$ is a diagonal matrix containing the ratio of the benchmark k to calculated k, and $\mathbf{S}_{B,kx}$ is a matrix where each row is the sensitivity vector for each benchmark experiment.

The GLLS method, through its nuclear data adjustment, also allows for an adjustment

of the nuclear data covariances. The adjusted or residual covariance matrix is found by

$$\mathbf{C}_{x'x'} = \mathbf{C}_{xx} - \left[\mathbf{C}_{xx}\mathbf{S}_{B,kx}^{\top}\mathbf{C}_{dd}^{-1}\mathbf{S}_{B,kx}\mathbf{C}_{xx}\right]. \tag{41}$$

This residual covariance matrix $\mathbf{C}_{x'x'}$ may be used to determine an adjusted uncertainty in k because of the uncertainties in nuclear data. The adjusted uncertainties in k for a set of applications may be found with the sandwich rule:

$$\mathbf{C}_{k'k'} = \mathbf{S}_{A,kx}\mathbf{C}_{x'x'}\mathbf{S}_{A,kx}^{\top}. \tag{42}$$

Here $\mathbf{S}_{A,kx}$ is a matrix with rows as the sensitivity vectors for the applications, and $\mathbf{C}_{k'k'}$ is the covariance matrix for k of the applications.

The square root of the diagonal elements of $\mathbf{C}_{k'k'}$ are the adjusted 1σ relative uncertainties of k from the adjusted nuclear data covariances. The portion of the MOS for nuclear data uncertainty/variability for an application i is set using

$$\mathrm{MOS}_{\mathrm{data}} = n_\sigma \mathbf{C}_{k'k',ii}^{1/2}, \tag{43}$$

where $\mathbf{C}_{k'k',ii}^{1/2}$ is the square root of the ith diagonal element of the adjusted application covariance matrix and n_σ is a factor for the desired confidence level q, which for this paper is 0.99. If the underlying data are assumed to be normally distributed, this leads to a value of $n_\sigma = 2.6$, i.e., roughly the number of standard deviations that would enclose 99% of the samples from a normal distribution. The assumption of normality may or may not be justified here, but it is consistent with the assumptions made when applying the sandwich rule in Eq. (42) to compute the adjusted uncertainties. Furthermore, this factor n_σ and the required assumptions behind it are not used for the determination of the calculational margin, but rather to set what should be a reasonable $\mathrm{MOS}_{\mathrm{data}}$ (see Sec. III.B.3 for a demonstration of its reasonableness), a factor applied in addition to the calculational margin.

III.B.3 Comments on the adjusted covariance library

Even though the nuclear data evaluations were influenced by the critical experiments, the information available to the GLLS method does not consider exactly which benchmarks may have influenced particular evaluations and how. Therefore, the resulting covariance libraries will be similarly limited in accuracy, and observations show correlations between nuclear data sets that are likely not real.

Because of the inherent limitations involved in generating the adjusted nuclear covariance data library and the resulting spurious correlations within, using it to calculate c_k similarity coefficients is undesirable, as it introduces what are also most likely spurious dependencies between different benchmarks and applications. Testing performed, but not presented in this paper, demonstrates this.

It is a bit of an inconsistency to use two different nuclear covariance libraries for different purposes, i.e., the prior covariances to estimate similarity coefficients and the adjusted ones to compute the nuclear data uncertainties. Recall, however, that the purpose of performing a GLLS nuclear data adjustment is to attempt, in determining a margin of subcriticality and not the calculational margin, to capture the correlations between the nuclear data evaluations and the critical experiment benchmarks, which are not reflected in the prior nuclear covariance data.

Even with the issue of what are potentially spurious correlations, the adjusted libraries provide more realistic estimates of the uncertainty and variability of k than is obtained when using the prior, unadjusted nuclear covariance data. This is evidenced by the fact that the 1σ uncertainties in k from the prior nuclear covariance data appear to bound most of the discrepancies of the critical experiment benchmark results; whereas, if there is no dependency between the nuclear data evaluations and the benchmark experiments, then the typical rules of standard deviations of normal distributions should apply (after considering experimental correlations of the benchmarks).

For example, the calculated 1σ uncertainty in k of the Jezebel benchmark, a bare sphere

of Pu metal, from using the prior covariances provided with Whisper is about 0.0139. The adjusted nuclear data covariances predict a 1σ uncertainty in k of about 0.00075, which is significantly smaller. Comparisons of different nuclear data libraries (ENDF/B-VII.1, JENDL-4.0, and JEFF-3.1.1) in Ref. [26] show a range of k of about 0.0019 (within 1σ of the benchmark uncertainty of 0.002), which is just within the 99% confidence interval ($n_\sigma = 2.6$) predicted by the adjusted nuclear covariance data libraries. The 1σ nuclear data uncertainty in k from the prior covariances is over a factor of seven larger than the empirically observed variation for this benchmark. Using the prior nuclear covariance data to set MOS_{data} would therefore be overly conservative.

Ideally, there would be a single covariance library that would consider the dependency between the nuclear data evaluations and benchmark experiments. This library would have reduced overall uncertainties (as a result of having additional information in the form of integral benchmark experiments available to the evaluator) and additional correlations between different isotopes and reactions over different energy ranges as a result. Correlations between the nuclear data and the critical benchmark experiments would have to be quantified as part of this hypothetical library as well. Unfortunately, no such library exists, and attempting to produce one considering these difficult to quantify effects is an open area of research.

III.C Estimating Unknown Benchmark Uncertainties

Sometimes it may be desirable or necessary to incorporate benchmarks that do not have rigorously quantified uncertainties. Ideally a benchmark uncertainty would be quantified through detailed analysis of the experiment and the model approximations. Unfortunately, this is not always feasible. An alternative, but ad hoc, approach is to assign a benchmark uncertainty based upon the uncertainties of benchmarks that are similar to the one question.

In these cases, Whisper uses weighted averages of the variances (of those benchmarks with a quantified uncertainty) to determine a representative benchmark uncertainty when one is not provided. The weighting factors in the averaging are the c_k parameter with all

the other benchmarks. For example, if benchmark B does not have a known uncertainty, the other N benchmarks with uncertainties are used to estimate it via a weighted average:

$$\sigma_B^2 = \frac{\sum_{i=1}^{N} c_{k,i}\sigma_i^2}{\sum_{i=1}^{N} c_{k,i}}. \tag{44}$$

Of course, this value is not a true benchmark uncertainty, which takes into account uncertainties in the actual experiment and approximations in creating the benchmark model, but it does serve as a surrogate representing a typical uncertainty for experiments of its type. It is also a more realistic and conservative assumption to apply such an uncertainty as opposed to simply using zero as the uncertainty.

This approach leads to additional correlations of the benchmark uncertainties (not the benchmark measurements of k). Whisper does not currently account for these or any other correlations between or uncertainties of the benchmark uncertainties.

III.D Benchmark Rejection

The benchmark experiments and the input files describing them are going to be of variable quality, and, for a suite that is sufficiently large, a small percentage are likely to be of poor quality. Consequently, the calculational margin can be biased (either conservatively or non-conservatively) because of these errant benchmark descriptions. Furthermore, the nuclear data adjustment used to determine the residual nuclear data uncertainties is less effective in the presence of such benchmarks, and the resulting MOS is larger than it would otherwise be.

The presence of poor quality benchmarks may be detected if the nuclear data cannot be adjusted consistently, considering the uncertainties of the benchmarks and nuclear data, to eliminate the computational biases in the benchmarks. The benchmarks causing this inability to perform a consistent nuclear data adjustment can then be identified and rejected,

i.e., removed from the validation suite.

The techniques for this purpose that are employed by Whisper are the same as those developed and implemented within the TSURFER package in SCALE [13, 27]. A brief discussion of the concepts is given here.

The χ^2_{min} calculated by the GLLS nuclear data adjustment measures the degree to which the linear regression can fit the benchmark experimental data within the nuclear covariance data. A χ^2_{min} per degree of freedom of unity (assuming the set of benchmarks is large, otherwise it is slightly less) indicates a perfect regression. Values greater than one indicate that the regression model could not perfectly fit the data, and that there are inconsistencies in the adjustment, i.e., there are likely other sources of bias, the quoted benchmark uncertainties are too small, and/or the nuclear data covariances indicate less uncertainty than there is in actuality. Values less than one are possible, indicating that the quoted uncertainties are larger than they should be. In practice, values of $\chi^2_{\text{min}} > 1$ are observed.

The threshold value of χ^2_{min} that is appropriate is application specific. Empirically determined rules of thumb typically give an acceptable χ^2_{min} between 1.2 and 1.6. For this paper, a threshold χ^2_{min} value of 1.2 is used, which is the default in Whisper.

If the computed value of χ^2_{min} is greater than the threshold, then benchmarks should be rejected until the threshold is met. There are various techniques, with varying degrees of rigor, for selecting which benchmarks to reject. The method used by Whisper and in this paper is the iterative diagonal χ^2 method. The diagonal χ^2 for the ith benchmark is found by

$$\chi^2_{\text{diag},i} = \mathbf{d}_i \mathbf{C}^{-1}_{dd,ii} \mathbf{d}_i, \tag{45}$$

where $\mathbf{d}_i$ is the discrepancy (difference of the calculated k from its reference value) of the ith benchmark and $\mathbf{C}^{-1}_{dd,ii}$ is the ith diagonal element of the inverse of the covariance matrix of the discrepancy vector. The benchmark with the largest $\chi^2_{\text{diag},i}$ is rejected and χ^2_{min} is recomputed. If the new χ^2_{min} is within the threshold χ^2_{min} value, the rejection stops and the remaining benchmarks are accepted. If not, then $\mathbf{C}^{-1}_{dd}$ is recomputed and the process repeats

until the threshold χ^2_{min} is met.

The iterative diagonal χ^2 method is not the most rigorous method available. A more rigorous approach is the $\Delta\chi^2$ method, which recomputes χ^2_{min} without benchmark i for all benchmarks. The benchmark that results in the largest change in χ^2_{min} is rejected (hence the naming of the method). As with the iterative diagonal χ^2 method, the process repeats until the threshold χ^2_{min} is attained. The disadvantage of using the iterative diagonal χ^2 method versus the $\Delta\chi^2$ method is that it tends to reject a greater number of benchmarks that may in actuality be of acceptable quality.

To illustrate the tradeoffs, performing the more rigorous $\Delta\chi^2$ rejection method on the benchmark suite used for the results in this paper containing 1,086 benchmarks using the available computing platform (a single node of the Moonlight cluster at LANL consisting of two Eight-Core Intel Xeon model E5-2670 at 2.6 GHz using 16 OpenMP threads) would take an estimated 3-4 months, which is not practical for a suite of this size. Conversely, the iterative diagonal χ^2 method can complete the entire rejection within about four hours. For reasons of practicality, the iterative diagonal χ^2 method was implemented into Whisper and used in this paper.

III.D.1 Comments on the parametric nature of the benchmark rejection

Regarding the non-parametric nature of the computation of the calculational margin in Whisper, an inconsistency arises when excluding the benchmarks identified by the GLLS rejection. Strictly speaking, the use of a linear regression model based on a χ^2 minimization for the rejection of benchmarks introduces a parametric element to the method. Furthermore, having an insufficient number of similar benchmarks in the set and using the approximate iterative diagonal χ^2 method for the rejection criterion may lead to the method erroneously rejecting benchmarks in some cases or failing to identify outlier benchmarks that ought be rejected in others.

This question and its impact on the analysis are not resolved in this paper and remain

open for further study. The results for Whisper in Sec. V exclude the benchmarks and are not significantly impacted whether they are retained or excluded. Should this be a concern, the Whisper software itself makes it straightforward to include all the benchmarks by simply omitting the exclusion file on the command line (see Sec. III.E and Ref. [14] for further details).

III.E Workflow for the Whisper Software

After installation of the Whisper software, the workflow for using Whisper has two phases: setup and running applications. Complete instructions may be obtained in Ref. [14], but a summary is given here to illustrate how a user interacts with the Whisper software.

III.E.1 Setup for an an area of applicability

The setup phase is typically done once per AOA. First, the suite of benchmarks provided with Whisper (see Sec. IV.B) should be analyzed to determine if it contains an adequate set for the particular AOA. If it does, then the excluded benchmark list from a benchmark rejection and the adjusted covariance libraries are available with Whisper, and the analyst may simply use the available information and skip further setup and begin running applications.

If the current suite in not adequate, then the analyst must find appropriate benchmarks and create MCNP input files, inserting benchmark k and uncertainty information at the bottom of those files as discussed in Ref. [14]. Once this task is done, a script called `whimcnp` is available with Whisper that may be used to run these MCNP input files via batch submission on a computing cluster; the `whimcnp` script will automatically insert appropriate lines of input to generate the applicable sensitivity profiles.

After the MCNP runs have completed, a script called `ww`, the "Whisper Wrapper", will extract the sensitivity profiles from the output files and place them in a personal library directory along with a table of contents file. The Whisper Wrapper will also run Whisper with these files as applications, but this calculation may be aborted for now as it serves no

purpose yet. Next, a shell utility script called `AppendBenchmarks` may be used to merge the new benchmarks with the current set in Whisper.

Following this, a GLLS nuclear data adjustment should be performed to reject inconsistent benchmarks and produce an adjusted covariance data library for future use. Before doing this, however, the analyst should also investigate any available experimental benchmark correlations and modify the provided correlations file accordingly, as these will be considered by the GLLS nuclear data adjustment. Some correlation data is available in the International Criticality Safety Benchmark Experiment Project (ICSBEP) Handbook [28] via DICE [29], the Database for ICSBEP, which is distributed with the Handbook and available online. If enough data about commonalities in the experiments are known, the correlations may also be determined using various techniques [30, 31, 32, 33] that are outside the scope of this paper.

Whisper is run in this mode and the file of recommended benchmarks to exclude and a new adjusted covariance library are produced. This exclusion file should replace the old one available with Whisper and may be done by simply overwriting the file. Replacing the adjusted covariance library is a bit more complicated, and a utility script called `UpdateCovarianceData` is provided that consistently overwrites the available data with new ones.

Once this is finished, all the information is available for analysis of applications within the AOA.

III.E.2 Analysis of applications

Assuming the benchmark suite, benchmark exclusion files, and adjusted nuclear covariance data is appropriate for the AOA being considered, the analyst may begin studying applications.

The first step is to prepare MCNP input files that model the particular application or set of applications being studied. Often this will involve a parametric study to determine

a bounding case. As with the setup phase, the `whimcnp` script may be used to run the MCNP inputs of the applications to obtain the appropriate sensitivity profiles, with the only difference being that there is no available benchmark k and uncertainty. Once these MCNP runs complete, the Whisper Wrapper, `ww`, may be used to extract the sensitivity profiles into an application library and to run the Whisper program; if excluding benchmarks is desired, the user specifies the benchmark exclusion file on the command line.

The Whisper program will read in the benchmark and application sensitivity profile information and the unadjusted and adjusted covariance libraries. The nuclear data uncertainties of the applications will then be computed using the adjusted covariance library, which will be used as part of computing the MOS. Then, for each application, Whisper will compute the c_k similarity parameter for every benchmark, choose an appropriate set, assign weighting factors, and compute a calculational margin from the extreme value distribution. The following information is printed to an output file for each application: the required sample weight, which benchmarks were used to validate each application and their corresponding c_k and weight factors, bias and bias uncertainty, MOS from software and nuclear data uncertainty, and any non-coverage penalty that had to be applied to the calculational margin. After all applications have finished, a summary table of calculational margins, MOS values, baseline USLs, and the amount that k exceeds the baseline USL [the δ_A parameter in Eq. (3)] is written to the output file and the screen.

Given the output, the NCS analyst can then determine the regions of the parameter space with $\delta_A < 0$, which can be assured to be subcritical. More precisely, this value is indicative of a baseline USL or an upper limit on it. It does not relieve the criticality safety analyst from considering any additional factors that are not captured as part of the computational analysis. Before subcriticality of particular configurations is determined, whether these additional factors exist must be considered, and if they do, the impact they would have on the determination of subcriticality must be assessed.

IV BENCHMARK SUITES

Before validation may be performed by Whisper or with any other statistical technique, a suitable benchmark suite must be provided. This section first discusses the general considerations that arise when selecting benchmarks for the suite to be used with Whisper. Next, the benchmark suite that is distributed with Whisper is described.

IV.A Considerations for Creating a Benchmark Suite

For the technique outlined in this paper to be effective, a suitably-sized suite of critical experiment benchmarks should be constructed that, at a minimum, adequately addresses the desired AOA to the extent possible. The method will determine appropriate weighting factors to be applied to each benchmark using the S/U techniques. Furthermore, the same GLLS techniques can be used to reject benchmarks that are inconsistent with the nuclear data adjustment; this indicates that there may be other significant sources of bias other than the nuclear data, e.g., the benchmark may be poorly described, the input file may have errors, etc.

The GLLS nuclear data adjustment is more effective and accurate when it is given more information, i.e., the suite of benchmarks is the largest available even if there are significant portions outside the AOA for the application. An AOA for validation is decided typically on fissionable material and its form (e.g., metal versus oxide), spectral characteristics, reflector materials, etc., and when considering one AOA, there may be shared materials used in similar neutron spectra in other AOAs. The GLLS nuclear data adjustment can take advantage of this fact and provide a more accurate set of adjusted nuclear data and covariances.

If at all possible, benchmarks should be selected so as not to have common sources of potential systematic bias, e.g., the benchmarks experiments should have been performed at a variety of facilities and by different experimenters spanning a large amount of time, the benchmark descriptions should have been evaluated and reviewed by various individuals, etc. Doing this helps minimize systematic bias that may arise from a particular facility,

experimenter, or approach. For benchmarks that have such commonalities, experimental correlations of the benchmark measurements of k should be obtained if they are available or developed when it is feasible.

All of these considerations should be made for any benchmark suite used with Whisper. Next, the specific benchmark suite distributed with Whisper (and used to generate the results in this paper) is discussed along with the results of the benchmark rejection that was performed.

IV.B Whisper-Provided Benchmark Suite

The benchmark suite distributed with Whisper contains 1,086 critical experiment benchmarks from the ICSBEP Handbook [28]. The benchmarks were selected to cover a wide range of fissile materials (uranium at various enrichments, ^{233}U, and Pu), fissionable material form (metal, compound, and solution), and spectral characteristics. The benchmark experiments used in this suite are from numerous different sources, minimizing any potential systematic biases resulting from commonalities in the experiments or evaluations. Table II gives a summary of the benchmarks by ICSBEP identifier. A full listing of the benchmarks is available in Ref. [15].

MCNP models of the benchmark experiments were obtained from the previous NCS validation suite [34, 35], the Mosteller Expanded Criticality Suite for MCNP validation [36], the Kahler validation suite for ENDF/B-VII.1 nuclear data testing [37], and, when needed and unavailable elsewhere, MCNP models were prepared and independently reviewed. All calculations of the critical experiment benchmarks were run with MCNP6.1 using ENDF/B-VII.1 nuclear data.

In five benchmark cases (HEU-MET-FAST-004[2] and the four cases of IEU-MET-FAST-

[2]The HEU-MET-FAST-004 benchmark is a single experiment that was a sphere of highly-enriched uranium metal in a cylindrical water tank. The experiment was performed at Los Alamos National Laboratory in 1976.

001[3]), the benchmark uncertainties are not provided by the ICSBEP Handbook. Estimates of their uncertainties were precomputed by Whisper using Eq. (44) with the c_k values computed for the other 1,081 benchmarks. These approximated uncertainties are used by default in Whisper. The impact of using them should be minor, as it only is applied to five out of 1,086, fewer than 1%, of the benchmarks. None of these are particularly important for the results of the case studies presented in Sec. V.

In cases where two one-sided (i.e., asymmetric) benchmark uncertainties are given, the larger of the two is assumed for conservatism.

The experimental benchmark correlation data from the ICSBEP Handbook (via DICE, the Database for ICSBEP, which is distributed with the Handbook and available through the internet [29]) are used where numerical values are provided.

The covariance data that were used come from the 44-group covariance library [38] that is distributed with SCALE6.1.

With this information, the 1,086 benchmarks are run through a GLLS nuclear data adjustment and a rejection using the iterative diagonal χ^2 method is performed, leaving 972 benchmarks. The benchmarks remaining are summarized in Table II by ICSBEP identifier, and a complete listing is given in Ref. [15]. The inclusion of the experimental benchmark correlations available in DICE at the time of access results in only minor changes to the benchmarks rejected, and impact on the case study results presented in Sec. V is negligible.

V EXAMPLE ANALYSES

Now that the Whisper methodology has been developed (Sec. III) and an underlying benchmark suite created (Sec. IV), illustrative analyses of hypothetical case studies can be performed. Four such cases are considered. The first case is a set of varied masses of spherical Pu metal-water mixtures with neutronically infinite water reflection. The second case is infinite

[3]The IEU-MET-FAST-001 benchmark consists of four experiments that were bare cylindrical configurations of enriched and natural uranium. The experiments were performed at Los Alamos Scientific Laboratory between 1952 and 1954.

square-pitch arrays of light-water moderated LEU-oxide fuel lattices with varied enrichments and pitches. The third case is metal-water mixtures of LEU and Pu at varying concentrations. The fourth case is infinite triangular-pitch lattices containing graphite-moderated MSR fuel with varied ^{233}U/^{232}Th concentrations and pitches.

All material compositions were taken from Ref. [39] except where noted otherwise. As with the benchmarks, the calculations were run with MCNP6.1 using ENDF/B-VII.1 nuclear data. Also, the 44-group covariance data that is distributed with SCALE6.1 was used. The unweighted calculational margin m_0 for the benchmark suite is about 0.049.

All statistically derived quantities from Whisper in this section are taken at the 99% confidence level, which is also the default in Whisper. Comparisons with the standard, parametric approach in Sec. II.A.1 use a 99/99 single-sided tolerance factor. The reason that the 99% confidence level is used as the Whisper default and in this paper, as opposed to the more typical 95% level (or a 95/95 single-sided tolerance factor with the standard, parametric approach), is because the Nuclear Criticality Safety division at LANL has elected to use these confidence levels in its validation of MCNP. There is nothing inherent in the Whisper methodology that prevents the use of a 95% confidence level (or any other), and the Whisper software allows the user to specify the desired confidence level with a "user options" file.

The four cases are now discussed followed by a summary of the results.

V.A Critical Mass of ^{239}Pu

The first application creates a critical mass curve for spherical ^{239}Pu metal-water mixtures with neutronically infinite (thickness of 100 cm) water reflection. More accurately, the curve is for a mass that can be assured to be subcritical for a given hydrogen-to-metal atomic ratio (H/X). A parametric study is performed by varying ^{239}Pu mass (the plutonium is assumed to be pure ^{239}Pu) and H/X in the sphere. The density of plutonium metal is taken as 19.85 g/cm^3 and the density of water is 1.0 g/cm^3. The density and radius of the sphere consisting

of the metal-water mixture are derived from the ^{239}Pu mass and H/X.

The calculated k as a function of H/X and mass is given on the surface plot in Fig. 2. Contour lines (drawn with a basis spline or B-spline) for various values of k are shown as well. The $k = 1$ contour (where criticality is predicted) follows known behavior. The predicted critical mass for dry plutonium metal with neutronically-infinite water reflection is about 5,500 g (not shown in Fig. 2 because H/X is on a log scale). As the plutonium is diluted with water, the critical mass increases to over 10,000 g (moderation is not yet sufficient to thermalize a significant fraction of the neutron population), then decreases to around 500-600 g as the neutron energy spectrum thermalizes (optimal moderation is predicted for an H/X of around 700), and finally increases again sharply as the plutonium becomes too dilute to sustain a nuclear chain reaction.

For criticality safety analysis, a mass below the $k = 1$ contour cannot be taken to be subcritical because of the computational biases and uncertainties discussed in Sec. III. Rather, a curve that considers these effects, i.e., the USL, is desired.

With the method outlined in this paper, the first step toward developing this curve is to determine the calculational margin for each case considered as part of this parametric study. The computed calculational margin from Eq. (27) is displayed in Fig. 3. The calculational margin for $H/X < 1$ is relatively flat around 0.0128, as a consequence of the relative abundance of fast critical benchmarks. For H/X between 1 and 10, however, the calculational margin increases sharply to ranging from 0.028 to 0.031; this is because the system has an intermediate energy spectrum where there are few quality benchmarks available. When H/X exceeds 10, the spectrum begins to thermalize; the calculational margin decreases again as the system becomes neutronically similar to the plutonium solution benchmarks. Note that the calculational margin is lowest in the region around where the $k = 1$ curve predicts optimal moderation, 0.0125, which is slightly lower than the calculational margin in the fast regime. The reason is that there are a large number of solution benchmarks that consistently predict values of k that are higher than the benchmark values.

The bias β computed by Whisper [using Eq. (25)] is negative for all cases and therefore no additional non-conservative bias adjustment is needed [i.e., $\Delta_m = 0$ from Eq. (7)] to the calculational margin in order to prevent the use of non-conservative bias.

The next effect to consider toward determining the USL is the MOS for the nuclear data variability because of their uncertainties [MOS_{data} from Eq. (37)]. The nuclear data adjustment is quite effective at reducing the nuclear data uncertainties; the application uncertainties are reduced by factors of 10-15 in the fast ($H/X < 1$) regime, 7-9 in the intermediate ($1 < H/X < 10$) regime, and 15-30 in the thermal ($H/X > 10$) regime.

Figure 4 shows MOS_{data} [Eq. (43)] from this variability corresponding to the 99% confidence level. Again, the trend is much the same as in Fig. 3 for the calculational margin, illustrating the ability or inability of the nuclear data adjustment to consistently reduce the nuclear data uncertainties. The fast range has an MOS_{data} of around 0.0020 for $H/X = 0$ and increases with H/X to 0.0025 at $H/X = 1$. MOS_{data} peaks around 0.0042 for the intermediate spectrum at $H/X = 4$, where the nuclear data adjustment cannot reduce the uncertainties as much because of the scant number of benchmarks in that regime. The smallest MOS_{data} values are about 0.0016 at optimal moderation, where there is a large number of consistent benchmarks, and therefore the nuclear data adjustment is quite successful and can significantly reduce the uncertainty in k from nuclear data.

An additional margin $\text{MOS}_{\text{software}}$ of 0.005 is applied globally to account for the effect of undetected errors in transport software (MCNP6) as discussed in Sec. III.B.1.

Figure 5 displays the resulting USL from Eq. (2) as a function of plutonium mass and H/X. As would be inferred from Figs. 3 and 4 for the calculational margin and MOS, the USL is does not vary much for the fast systems ($H/X < 1$), being around 0.98 at $H/X = 0$ and falling slowly as H/X increases. The USL falls rapidly to 0.955 to 0.960 in the intermediate spectrum regime (H/X between 1 and 10). The USL then increases for $H/X > 10$, peaking to around 0.981 for optimal moderation. The USL is weakly dependent upon Pu mass, having no sharp variations near the predicted critical masses.

The subcritical mass curve (i.e., the value of the plutonium mass that can be taken as subcritical) as a function of H/X corresponds to the $\delta_A = 0$ contour, where δ_A again is the amount, in terms of k, that the USL is exceeded [Eq. (3)]; negative values of δ_A can be taken to be assuredly subcritical. The δ_A parameter as a function of plutonium mass and H/X is shown in Fig. 6. Subcriticality for a dry metal ($H/X = 0$) with "infinite" water reflection can be assured for a mass of about 5,000 g (not shown in Fig. 6 because it is on log scale), which is around 500 g lower than the predicted $k = 1$ value. In the intermediate range, the subcritical mass curve peaks at just under 8,500 g; this is significantly lower than the slightly over 10,000 g critical mass predicted by $k = 1$ (Fig. 2), accounting for the largest difference between the two curves, which is a consequence of having few benchmark critical experiments that have an intermediate spectrum. The subcritical mass decreases to about 400-450 g at optimal moderation, which is only 100-200 g lower than what the $k = 1$ curve would predict; again, this is because of the large number of benchmark critical experiments at optimal moderation that consistently predict a high value of k.

Estimation of the calculational margin, and hence the USL, depends upon the c_k parameter from Eq. (32) for the selection of relevant benchmark critical experiments. The parameter space for this study spans the entire neutron energy spectrum from fast to thermal, and which benchmarks are relevant depends upon the H/X of the application system being considered. For the fast systems ($H/X < 1$) the most relevant benchmark is PU-MET-FAST-011.[4] In the intermediate regime, $1 < H/X < 10$, the most relevant benchmarks are various cases of the PU-COMP-MIXED-001 or PU-COMP-MIXED-002 benchmarks.[5] In this regime, the benchmark data is sparse, and, furthermore, the results of calculations and quoted benchmark experiment values do not agree well. When $H/X > 10$, the PU-

[4]The PU-MET-FAST-011 benchmark is a single experiment that was a sphere of alpha-phase plutonium metal in a cylindrical water tank. The experiment was performed at Los Alamos Scientific Laboratory in 1968.

[5]These PU-COMP-MIXED benchmarks had active fuel that were compacts of polystyrene and PuO_2. The PU-COMP-MIXED-001 cases were unreflected and the PU-COMP-MIXED-002 cases were reflected by plexiglass. The experiments were performed at the Hanford Plutonium Critical Mass Laboratory between 1963 and 1970.

SOL-THERM benchmarks become most relevant; the specific benchmark case that is most relevant varies and correlates with H/X.

To distill the data about which benchmarks are most relevant, suppose the benchmark experiments are grouped by spectra: fast, intermediate/mixed, and thermal. The application cases along the subcritical mass curve (i.e., the case of plutonium mass where $|\delta_A|$ is at a minimum for each H/X) are chosen as the most relevant ones to analyze as they determine the various mass limits. The maximum c_k used to determine the calculational margin for each spectral classification is obtained for each application case along the subcritical mass curve; if no benchmarks of that spectrum are used, then the value is set to zero.

These maximal c_k grouped by benchmark spectrum as a function of application H/X are shown in Fig. 7. The various application regimes as a function of application H/X are illustrated here.

For H/X near zero, the maximum c_k for the fast benchmarks exceeds 0.99 and steadily decreases. The fast benchmarks are most relevant until an H/X of about 2 where the intermediate/mixed become more neutronically similar, and they are completely supplanted for the purposes of determining the calculational margin at an H/X of around 4. The intermediate/mixed benchmarks are relevant for the H/X ranging from about 0.9 to around 40. Between H/X of 10 and 20, the thermal solution benchmarks become more important than the intermediate/mixed ones, and the thermal benchmarks completely overtake the intermediate/mixed (there are sufficiently many that the intermediate/mixed benchmarks are no longer necessary for the validation) at around H/X of around 40. As H/X increases, the maximum c_k for the thermal benchmarks also increases and peaks at a value over 0.999 around optimal moderation, which is at an H/X of around 700.

For comparing the calculational margin obtained with the standard approaches outlined in Sec. II.A, the validation for the ^{239}Pu mass is broken up into the fast, intermediate/mixed, and thermal regimes. The benchmarks that correspond respectively are categorized as Pu or metal, compound (oxide), and solution. The PU-MET-FAST, PU-COMP-MIXED, and

PU-SOL-THERM benchmarks in the validation suite (after rejection by the GLLS nuclear data adjustment) are used to determine the calculational margin for the fast, intermediate, and thermal regimes respectively. The metal and solution (fast and thermal) cases pass the Shapiro-Wilk normality test, and therefore the parametric approach discussed in Sec. II.A.1 is appropriate. The oxides do not pass the Shapiro-Wilk normality test, and the non-parametric, rank-order method in Sec. II.A.2 must be used for that case.

For the fast range, using the PU-MET-FAST benchmarks, the mean multiplication factor $\bar{k}$ based on inverse variance weighting [Eq. (5)] is 1.0004. On average, MCNP6.1 with ENDF/B-VII.1 nuclear data calculates k slightly high; as per standard practice, the bias β [Eq. (6)] is set to zero to not non-conservatively take credit for positive bias. The bias computed by Whisper, which is defined differently as a mean of the extreme value distribution [Eq. (25)], for the fast regime is about -0.0067. The weighted standard deviation in k about the mean s_k [Eq. (9)] is 0.0028 and the average standard deviation in k $\bar{\sigma}_k$ [Eq. (10)] is 0.0020. The pooled standard deviation, and hence the bias uncertainty, σ_β [Eq. (8)] is 0.0035. Because the bias is set to zero, the calculational margin from Eq. (15) with single-sided tolerance factor $\kappa = 3.12$ [Eq. (11)] is 0.0109. The Whisper result is 0.0128, which is certainly conservatively bounding of the result from the standard parametric approach. Note that including all the MIX-MET-FAST benchmarks (after rejection), many of which have ^{239}Pu as the dominant fissile isotope, decreases the calculational margin from the standard parametric approach to 0.0103, mainly because the sample size increased, decreasing κ to 2.91.

For the intermediate range, using PU-COMP-MIXED benchmarks, the rank-order, non-parametric approach is used. In this analysis, the rank-order, non parametric approach is applied using all 34 PU-COMP-MIXED benchmarks as well as (separately) only the 17 PU-COMP-MIXED benchmarks left after the rejection. Either way, the minimum $\tilde{k}$ is 0.98138 with a combined uncertainty of 0.00720. This corresponds to case 5 of PU-COMP-MIXED-001. The value of C_{NP} [Eq. (17)] depends on whether the rejected benchmarks are

considered or not. If the rejected benchmarks are not used, $C_{NP} = 0.582$, and if they are, $C_{NP} = 0.825$. From Table I, the corresponding non-parametric margins m_{NP} are 0.04 and 0.01. This leads to calculational margins that are 0.077 and 0.047 respectively, compared with the calculational margin of about 0.03 computed by Whisper in the intermediate range.

The calculational margins for the non-parametric, rank-order approach are more conservative than the Whisper result, especially if the rejected benchmarks are not included. This difference arises from both the existence of the non-parametric margin and the weighting of benchmarks. With regard to the non-parametric margin, Whisper does have a similar concept, the non-coverage penalty, but it is not applied here as in all cases Whisper was able to find enough relevant benchmarks from other categories (PU-MET-FAST, PU-SOL-THERM, MIX-MET-FAST, MIX-COMP-THERM, and MIX-SOL-THERM) to meet its sample weight requirement. In other words, the S/U methods are able to search the entire benchmark suite beyond the PU-COMP-MIXED benchmarks to find benchmarks that are neutronically similar.

Also, case 5 of PU-COMP-MIXED-001 is never given a particularly high weight for this application; the weight factor is never much greater than 0.21. The c_k value steadily increases as the spectrum softens, from 0.7 when it first appears for the cases with harder spectra to 0.9 when it is overtaken by the large quantity of solution benchmarks for the cases with softer spectra. In the intermediate range, the benchmark that, according to Whisper, is most relevant is case 7 of PU-COMP-MIXED-002, which has a c_k that ranges from 0.88 to 0.96 and a weight of 1.0.

Note that there is seemingly an inconsistency by using $p = 0.95$ as opposed to $p = 0.99$. This is not the case, because p is only used to calculate C_{NP}, which is not used directly except to obtain a non-parametric margin m_{NP} from Table I. Ref. [16], the source of the data in Table I, only gives values for $p = 0.95$. Consistent values could be generated for $p = 0.99$, changing the ranges of C_{NP}, but the results would be identical.

Had the benchmark data for the PU-COMP-MIXED cases been normally distributed,

the parametric approach could have been applied. For illustration, the calculational margin is computed as if this had been the case. The mean multiplication factor $\bar{k}$ based on inverse variance weighting is 1.0099 and therefore the bias β is set to zero; for comparison, the bias as defined by the Whisper methodology is -0.0114. The weighted standard deviation in k about the mean s_k is 0.0089 and the average standard deviation in k $\bar{\sigma}_k$ is 0.0064. The pooled standard deviation, and hence the bias uncertainty, σ_β is 0.0117. Because the bias is set to zero, the calculational margin from Eq. (15) with $\kappa = 3.89$ [Eq. (11)] is 0.0455. This is more conservative than the about 0.03 value that Whisper calculates. The reason for this is the small sample size of PU-COMP-MIXED benchmarks that is used in the standard analysis, i.e., the single-sided tolerance factor κ has a large value that magnifies the bias uncertainty. Including all the PU-COMP-MIXED benchmarks (not rejecting any) decreases the calculational margin to 0.0419 primarily because κ falls to 3.35. If κ had its limiting value of 2.7 for a very large sample size and everything else was fixed with no benchmarks rejected, the calculational margin would be about 0.033, which is slightly more conservative than the Whisper result.

For the thermal range, using PU-SOL-THERM benchmarks, the mean multiplication factor $\bar{k}$ based on inverse variance weighting is 1.0034 and therefore the bias β is set to zero as it is for the Pu metals; for comparison, the bias as defined by the Whisper methodology is -0.0036. The weighted standard deviation in k about the mean s_k is 0.0032 and the average standard deviation in k $\bar{\sigma}_k$ is 0.0030. The pooled standard deviation, and hence the bias uncertainty, σ_β is 0.0044, and the single-sided tolerance factor κ is 2.75. This leads to a calculational margin from the standard parametric approach of 0.0120. The Whisper methodology predicts the lowest calculational margin at optimal moderation as 0.0125, which is slightly higher as the standard, parametric approach. Including all the MIX-SOL-THERM benchmarks (after rejection), many of which are driven by fission of ^{239}Pu, in the analysis using the standard parametric approach increases the calculational margin slightly to 0.0126.

V.B Low-Enriched Uranium Lattice

The second application represents the storage of fresh reactor fuel assemblies immersed in pure light water (density of 1.0 g/cm^3). The hypothetical fuel assembly is a 17 $\times$ 17 fuel pin array in a square-pitch lattice in an infinite array. The elements in the fuel assembly are zircaloy-IV clad UO_2 fuel (mass density 10.97 g/cm^3 with varied enrichments), zircaloy-IV clad UO_2 with 6 weight percent Gd_2O_3, and Al-clad water tubes. All pins and tubes have a height of 500 cm and are reflected by 100 cm of water on top and bottom. Figure 8 gives the layout of the lattice. The fuel pin radius is 0.63245 cm, and the inner and outer cladding radii are 0.64895 and 0.69090 cm respectively (the cladding for both types of fuel and the water tubes have the same dimensions). The assembly is surrounded by a 0.5-cm buffer layer of water on the sides, and there is a 0.1-cm thick boral (with 10% boron by weight) sheath between the assemblies to absorb neutrons; the sheath does not extend into the axial water reflectors. The assembly has reflecting boundaries on the sides (not on the top and bottom) to simulate the effect of these assemblies in a very large array that would be found in a fuel storage pool. For simplicity, grid spacers are not included in the model.

Two parameters are varied as part of the study: the uranium enrichment and the pitch-to-diameter ratio (P/D). The enrichment varies from 1% to 4% in 0.2% increments; isotopes other than ^{235}U and ^{238}U are neglected for simplicity. The P/D varies from 1.0 to 2.0 in increments of 0.05.

Figure 9 gives the calculated k for the configurations with the two parameters varied. The behavior is expected: k increases monotonically with enrichment and has a maximum value at some P/D corresponding to optimal moderation, which varies with enrichment. Contour lines are also included for different values of k. The USL, i.e., the value that can be asserted to be subcritical, lies somewhere below the $k = 1$ contour line.

To determine the contour for the USL, first the calculational margin is estimated (as discussed in Sec. III.A.1). Figure 10 displays the calculational margin for this parametric study. The calculational margin varies from about 0.014 to 0.017 and decreases as P/D

increases (for this range of enrichments). The calculational margin increases sharply to over 0.03 for very low enrichment ($< 2\%$) and P/D (< 1.2). This is because there are no critical experiments in that regime, as it is not possible to make any quantity of a few percent enriched uranium critical without significant moderation and therefore this situation is not a concern for criticality safety.

The next contribution to the USL is from the MOS because of variability in k from nuclear data uncertainties, i.e., MOS_{data} from Eq. (37). These, at the 99% confidence level [Eq. (43)], are shown in Fig. 11. As before, MOS_{data} increases with decreasing P/D, varying from about 0.0016 to 0.0030, and then increases sharply for very low enrichment and P/D, where, again, it is not of practical concern. The nuclear data adjustment is less effective at reducing the uncertainties for the LEU lattice application compared to the the ^{239}Pu mass study in Sec. V.A; the uncertainties in k from nuclear data are reduced by factors of 5-7, compared with the factors of 15-30 for the thermal Pu metal-water mixtures.

Adding an additional margin of $\text{MOS}_{\text{software}} = 0.005$ everywhere to account for errors in transport software (Sec. III.B.1) allows for the calculation of the USL [Eq. (2)] as a function of the parameter space, which is shown in Fig. 12. As expected from the calculational margin and MOS (Figs. 10 and 11), the USL decreases with decreasing uranium enrichment. For the region of practical concern to criticality safety, the USL ranges from about 0.97 to 0.98.

The contour line identifying the region where criticality can be ensured is given in the plot of δ_A [Eq, (3)], the degree to which the calculated k exceeds the USL, in Fig. 13. For $P/D < 1.1$, all uranium enrichments $< 4\%$ can be taken to be subcritical. At about $P/D = 1.55$, the contour reaches its minimum value where the uranium enrichment must be $< 2.7\%$ to be considered subcritical.

While it is of no concern to criticality safety for this particular application, the behavior of the calculational margin seen in the bottom-left corner of Fig. 10 is worth analyzing. These cases are of an intermediate spectrum, having some amount of moderation, but not enough to thermalize the neutrons so that the dominant amount of fission arises from thermal

neutrons. The maximum c_k [Eq. (32)] for the case with the (lowest) enrichment of 1% and $P/D = 1$ is 0.828; the benchmark is actually a fast-spectrum, uranium-metal experiment, IEU-MET-FAST-007.[6] Such a value of c_k is marginal in terms of its neutronic similarity (0.9 or greater is preferred). Accordingly, the total sample weight required increases to compensate for the lack of a similar benchmark experiment.

The reason the calculational margin spikes near the corner is because there is not a sufficient sample weight, and therefore the non-coverage penalty (Sec. III.A.3) is required. For six different cases in the low-enrichment and -P/D regime, Fig. 14 shows the cumulative fraction of the required weight as a function of benchmark c_k (the benchmarks are sorted by c_k and added to the validation by increasing c_k). The value for $c_k = 1$ represents the fraction of the required weight that is available for the sample size in the validation. Ideally, this value would reach unity, indicating that the benchmark suite is sufficient for this particular application. In these cases, however, it is not, and a non-coverage penalty is applied as an additional margin of safety.

Again, the reason this effect arises in this LEU lattice study is that it is impossible to achieve criticality with this low an enrichment and moderation and, therefore, there are no critical experiments possible of this kind. In this case, k is low enough that it would not exceed the reduced USL. As a reminder, criticality safety evaluations do not and should not rely on computational methods alone. Even in cases where the k in an analogous physical regime could hypothetically exceed the lower USL of a calculation, it would behoove the criticality safety analyst to consider the physical possibility of achieving criticality, which is independent of calculational or administrative margins. In other words, for systems where achieving criticality has been shown experimentally to be impossible (even if k can reach a value close to unity and may technically exceed a derived USL), the analyst should rely on that fact and not calculations to set appropriate limits.

[6]The IEU-MET-FAST-007 benchmark is Big Ten, which was an experiment consisting of a large, uranium metal cylindrical core containing disks of varied ^{235}U enrichments with 10% average enrichment surrounded by a depleted uranium reflector. The experiments were performed at Los Alamos National Laboratory with first criticality achieved in 1971.

A comparison is performed with the standard approaches in Sec. II.A using all LEU-COMP-THERM benchmarks in the post-rejection validation suite. The benchmarks do not pass the Shapiro-Wilk normality test, and therefore the non-parametric, rank-order method (Sec. II.A.2) is required. The minimum $\tilde{k}$ is 0.98838 with a 1σ uncertainty (benchmark and statistical) of 0.00410. The benchmark with the minimum $\tilde{k}$ is case 1 of the LEU-COMP-THERM-025 benchmark,[7] which has a benchmark k of 1.0000 and therefore the corresponding minimum k is the same as $\tilde{k}$, 0.98838. The non-parametric confidence C_{NP} from Eq. (17) with $p = 0.95$ is about 0.9999 and the corresponding non-parametric margin m_{NP} from Table I is 0.0, so no additional margin beyond that given by Eq. (15) needs to be applied. Using $n_\sigma = 2.6$ (the 99% confidence level), the calculational margin computed with Eq. (15) is 0.0223, which is significantly more conservative than the Whisper calculational margins that range from 0.014 to 0.017.

The reason for this difference is that a single benchmark, case 1 of LEU-COMP-THERM-025, is determining the calculational margin by itself. There is no consideration given to weighting by relevance to the application being studied, as all LEU-COMP-THERM benchmarks are treated equally in the standard non-parametric, rank-order method. Whisper never actually uses the LEU-COMP-THERM-025-001 in the validation (i.e., its weight is zero) because its c_k is too low ($c_{k,\text{acc}}$ of Sec. III.A.2 ranges from 0.4 to 0.6 for $H/D \geq 1.1$) to be considered relevant when considering the availability of other more relevant LEU-COMP-THERM and LEU-SOL-THERM benchmarks in the suite. This demonstrates an advantage of the Whisper methodology, which weights benchmarks based upon their relevance to the particular application being analyzed.

A comparison is also performed with the parametric approach in Sec. II.A.1 that would have been used had the benchmarks been normally distributed. The average scaled k, $\bar{k}$, based on inverse variance weighting [Eq. (5)] is 0.9989. The bias β [Eq. (6)] from the

[7]The LEU-COMP-THERM-025 benchmark is a series of experiments that are water-moderated hexagonal pitch lattices of 7.5% enriched UO_2 rods with stainless steel cladding. These experiments were performed in 1965 at the Kurchatov Institute in Russia.

standard approach is -0.0011; for comparison, the bias computed from the Whisper methodology [Eq. (25)] for a typical case near the subcritical contour (enrichment of 2.8% and $P/D = 1.35$) is about -0.0084. The weighted standard deviation in k about the mean s_k [Eq. (9)] is 0.0017 and the average standard deviation in k $\bar{\sigma}_k$ [Eq. (10)] is 0.0018. The pooled standard deviation σ_β [Eq. (8)] is 0.0025. The resulting calculational margin from Eq. (15) with $\kappa = 2.70$ [Eq. (11)] is then 0.0080. The Whisper methodology predicts a range of calculational margins; its minimum is about 0.0135, which bounds the one obtained with the standard parametric approach.

V.C Uranium-Plutonium Metal-Water Mixtures

The third application is a 4×3 array of cans containing metal-water mixtures of U and Pu, representing a hypothetical analysis of solutions encountered during reprocessing of spent nuclear fuel. The cans are SS-304 with a thickness of 0.1 cm. They have an interior diameter and height of 15 cm and 50 cm respectively. The center-to-center distance between the cans is 20 cm. The cans sit on a Los Alamos concrete floor that is 50 cm thick and extends for 500 cm from the center of the cans on the edge. The air between the cans is simulated as void.

The metal concentrations of U and Pu range from 0 to 200 g/L in increments of 10 g/L, and they are varied independently such that all mixtures within the range are considered. The U has 3% ^{235}U with the remainder being ^{238}U. The Pu consists of 62% ^{239}Pu, 22% ^{240}Pu, 12% ^{241}Pu, and 4% ^{242}Pu. All percentages are by weight.

Figure 15 gives k for the uranium and plutonium concentrations. As expected, the greater the plutonium concentration, the greater the k. Perhaps not as obvious is that increased uranium concentration actually decreases the reactivity. The reason is that the uranium is enriched only enriched to 3%, and therefore increasing the amount of uranium predominantly increases the presence of ^{238}U, which increases the total amount of resonance capture in the solution and therefore decreases k.

Contour lines for various values of k are given in Fig. 15 as well. Note that the calculated critical line occurs only for plutonium concentrations > 180 g/L, and the amount of plutonium needed to achieve criticality (as predicted by MCNP6.1 with ENDF/B-VII.1 nuclear data) only increases with increased uranium concentration. As will be seen, the actual value that can be treated as critical is significantly lower.

The calculational margin computed by Eq. (27) is displayed in Fig. 16 with contour lines to illustrate the behavior, which is relatively flat over this entire range, its value being from about 0.034 to 0.035. The calculational margin peaks at around 0.0353 for plutonium concentrations of 50 to 120 g/L with uranium concentrations ranging from 0 to 90 g/L.

The MOS from the uncertainties in k from nuclear data [i.e., MOS_{data} from Eq. (37) and computed by Eq. (43)], which are displayed in Fig. 17 at the 99% confidence level, are likewise somewhat insensitive to the plutonium and uranium concentrations, ranging from 0.003 to 0.004. Of particular note is the trend that the uncertainty in k from nuclear data increases as the uranium concentration decreases and the plutonium concentration decreases. This is important because, as seen in the behavior in k, the USL will tend to decrease as the calculated system k increases; however, the magnitude of the effect of uncertainties in k from nuclear data is small relative to that of the calculational margin. The effectiveness of the nuclear data adjustment for reducing the nuclear data uncertainties decreases with LEU concentration. For low LEU concentrations, the reduction is about a factor of 10, and then decreases to about a factor of 5 at the 200 g/L of LEU limit.

The resulting USL, including an additional global margin of 0.005 for undetected errors in transport software ($\text{MOS}_{\text{software}}$ described in Sec. III.B.1), is shown in Fig. 18 and ranges from about 0.956 to 0.957, which is, again, a small variation over the range of parameters considered. The δ_A parameter, representing the amount of k over the USL [Eq. (3)], is displayed in Fig. 19. At 20 g/L of uranium, the maximum plutonium concentration in an assurably subcritical configuration is about 140 g/L, which is significantly lower than the 180 g/L that would be obtained if the calculated k (Fig. 15) were perfectly accurate.

The allowable plutonium concentration increases monotonically and slowly with increasing uranium concentration.

A comparison is performed with the parametric approach in Sec. II.A.1. All 21 MIX-SOL-THERM benchmarks (prior to rejection) are used in this case, because the remaining 12 after rejection make the sample size marginal. The benchmarks pass the Shapiro-Wilk normality test, and therefore the parametric approach is valid.

The average scaled k, $\bar{k}$, based on inverse variance weighting [Eq. (5)] is 1.0018. The bias β [Eq. (6)] from the standard approach is set to zero to not non-conservatively take credit for positive bias. The weighted standard deviation in k about the mean s_k [Eq. (9)] is 0.0082 and the average standard deviation in k $\bar{\sigma}_k$ [Eq. (10)] is 0.0019. The pooled standard deviation σ_β [Eq. (8)] is 0.0084. The resulting calculational margin from Eq. (15) with $\kappa = 3.78$ [Eq. (11)] is then 0.0318. The Whisper methodology predicts the calculational margin at about 0.035, which bounds the one obtained with the standard parametric approach.

Both Whisper, as demonstrated here, and other GLLS-based methods for determining the calculational margin are still able to produce meaningful results when the statistical foundations of many of the other approaches become very questionable when the sample size is marginal. S/U-based methods for validation are most useful in these situations.

V.D Molten Salt Reactor Lattice

The final case deliberately attempts to stress the method by applying it to a case where there are no obviously relevant benchmark critical experiments. This case is a hypothetical MSR with channels of ^{233}U/^{232}Th fuel in a molten salt within a graphite moderator. The model is an infinite unit cell in a hexagonal lattice. The unit cell consists of a fuel channel for the molten salt-fuel mixture with a diameter of 1.2 cm. Surrounding the fuel channel is a graphite moderator with reflecting boundaries on the sides. The P/D is varied and ranges from 1.00 to 2.20 in increments of 0.05. The channel height is 200 cm. Above and below is a uniform molten salt fuel layer 20 cm tall with another 10 cm of graphite.

The molten salt fuel mixture consists of (atom percentages) 55% LiF, 24% BeF_2, 20% ZrF_4, and 1% Th/UF_4. The lithium is enriched to 99.95% ^{7}Li with the remainder being ^{6}Li. The uranium is 99% ^{233}U and 1% ^{232}U. In this parametric study, the fraction of uranium (the remainder being thorium) ranges from 5% to 8% and is varied in in increments of 0.2%.

Figure 20 shows k as a function of uranium fraction and P/D. Since the physics is very similar to the LEU lattice of Sec. V.B, the trend in k with respect to the two analogous parameters is quite similar. Because the materials in the MSR lattice are quite different and more exotic (at least with respect to current typical reactor designs) than the LEU lattice, there are few benchmark critical experiments available, and therefore the USL is significantly lower.

The calculational margin [computed using Eqs. (27) and (36)] is displayed in Fig. 21 and is relatively insensitive to the variation of the uranium concentration and lattice P/D; the value is around 0.0475. Note that this is significantly larger than the calculational margins seen in the previous three example studies and is near the bounding, unweighted value of 0.049 [from Eq. (27) without using Eq. (36)]. The maximum c_k [Eq. (32)] for this case ranges from 0.24 to 0.29, indicating very poor coverage in the benchmark suite — the ICSBEP Handbook does not provide any experiments that appear relevant to this application and neither does the International Handbook of Evaluated Reactor Physics Benchmark Experiments [40]. The most relevant benchmarks are, unsurprisingly, the ^{233}U compound (lattice) and solution systems.

The calculational margin is driven by the comparatively large bias of these systems. It turns out that the non-coverage penalty (Sec. III.A.3) does not play a significant role for this application even though the amount of sample weight available for the validation is only 60 to 80% of the required sample weight. This is because the calculational margin $\tilde{m}$, the calculational margin that is calculated assuming the sample weight is sufficient, is already near m_0, the value used as the bounding calculational margin. In other words, the most negatively biased results in the benchmark set provided by Whisper are the ^{233}U

cases, which are the ones being used in Whisper's attempt to establish a baseline USL. The resulting non-coverage penalty is about 0.001 or less.

Had the case of having no relevant benchmark occurred for an application driven by ^{235}U or Pu, the non-coverage penalty would have likely been significantly higher. This is because the benchmarks in the set featuring those fissionable isotopes have a much less negative bias associated with them, and the interpolation to m_0 would have resulted in a larger increase in the calculational margin.

Figure 22 shows the MOS for the uncertainties in k from nuclear data [i.e., MOS_{data} from Eq. (37) and computed by Eq. (43)] at the 99% confidence level. The nuclear data uncertainty MOS_{data} ranges from just above 0.013 to just below 0.019, which is also significantly higher than what is observed in the hypothetical studies in Secs. V.A, V.B, and V.C. The reason these are so much higher is the relative scarcity of ^{233}U benchmarks; the nuclear data adjustment cannot reduce the nuclear data uncertainties to nearly the degree that it can for the plutonium and uranium systems. The uncertainty in k from nuclear data tends to decrease as uranium concentration and P/D increases, which is a conservative effect, i.e., the margin decreases as k would tend to increase. Because of the lack of benchmarks, the nuclear data adjustment is quite ineffective at reducing the nuclear data uncertainties; the uncertainties in k are reduced by only 10-15%, compared to factors of 5-30 for the other studies.

After the application of a global margin of 0.005 for transport software ($\text{MOS}_{\text{software}}$ of Sec. III.B.1), the resulting USL is obtained using Eq. (2); this is presented in Fig. 23. Based on the variation of the nuclear data uncertainty MOS_{data}, the similar trend is observed that the USL increases (conservatively) with increasing uranium concentration and P/D. The USL ranges from 0.930 to 0.935 for much of the studied parameter space. This value is much lower than seen in the studies in Secs. V.A, V.B, and V.C, but this reflects both the lack of neutronically similar benchmark critical experiments and the large biases in the (marginally appropriate) ones that do exist.

A plot of δ_A [Eq. (3)] and the corresponding subcriticality contour is shown in Fig. 24. The contour follows the similar shape of the curves for k in Fig. 20 but is significantly lower than the $k = 1$ curve.

Since there are no clearly applicable benchmarks available, no comparison is made with the standard approaches discussed in Sec. II.A. However, the USL results for a system with no benchmarks are on the same order as analyses of other systems using the sensitvity/uncertainty methods in SCALE. For example, a validation study for mixed-oxide fuel reported USL values of 0.928 and 0.936 depending on the trending parameters used [41]. Similar analyses of the a ^{233}U storage array resulted in USL values on the order of 0.955 [42].

Whether to accept the region of subcriticality identified by Whisper can be debated. The Whisper method did the best it could with the data that was provided (or really lack thereof in terms of its relevance to the application), and produced a quite conservative USL. In such a case as this, the criticality safety analysts should certainly give due consideration to whether this USL is conservative enough and if additional margin is appropriate, more reactive materials may be substituted that are better known (e.g., ^{239}Pu in place of ^{233}U), other non-computational techniques should be used to bound the system k, and/or additional measurements are needed to ensure subcriticality.

V.E Summary & Discussion of Results

Table III gives a few of the results obtained from Whisper. A few sets of parameters are taken for each hypothetical application case; the parameters chosen lead to configurations near where subcriticality can be assured (i.e., the $\delta_A = 0$ contour). Note that the MOS values in Table III contain both the margin for software errors ($\text{MOS}_{\text{software}}$ of Sec. III.B.1) and nuclear data uncertainties (MOS_{data} of Sec. III.B.2) at the 99% confidence level and the USL given is a baseline value before any additional margin about the application is applied, as appropriate, by the analyst.

Table IV compares the calculational margins obtained by the standard approaches dis-

cussed in Sec. II.A (both the parametric and the non-parametric techniques for all cases regardless of whether or not they passed the Shapiro-Wilk normality test) and Whisper. The calculational margins shown in Table IV correspond to the largest calculational margin from the respective cases in Table III.

Whisper typically obtains similar or more conservative calculational margins than the standard parametric approach and less stringent calculational margins than the rank-order approach.

The reason that Whisper calculational margins are typically more conservative than the standard parametric approach is because an extreme value distribution is used. The exception in Table IV is the plutonium-oxide cases where Whisper estimates a significantly lower value than the standard parametric approach. This is because the S/U techniques employed by Whisper are able to identify plutonium metal or solution benchmarks that are neutronically similar, albeit having a different fissionable material form, for the validation, whereas the analysis done with the standard parametric approach, lacking insight into neutronic similarity of benchmarks with other fissionable material forms, used only the very limited set of plutonium-oxide benchmarks leading to a very large single-sided tolerance factor κ; also, the weighting based on similarity in Whisper tended to discount the few benchmarks leading to a higher bias uncertainty, which is information that is unavailable to the standard parametric approach.

Whisper calculational margins tend to be less stringent than the non-parametric, rank-order approach because Whisper uses neutronic similarity weighting in its extreme value calculation, whereas the rank-order approach simply uses the worst case, i.e., lowest normalized k, from the set of benchmarks chosen by the analyst, which, lacking similarity information, may or may not be the most applicable to the current application model being analyzed.

As seen in Table III, the MOS values obtained by Whisper tend to be significantly lower than those typically used in NCS analysis, which are usually 0.02 or higher. Recall the MOS

accounts for all aspects of the process and simulation and should be large enough that the analyst can ensure subcriticality. The MOS values derived from Whisper only account for two aspects of this, undetected errors in software and expected uncertainty/variability from nuclear data libraries. Because these factors are now explicitly quantified, the NCS analyst may use this information to determine what, if any, extra margin is appropriate to ensure subcriticality for the process being analyzed.

VI SUMMARY & FUTURE WORK

This paper introduced Whisper, a method and software package for calculating baseline USLs for criticality safety analysis that can be integrated into the workflow of using continuous-energy Monte Carlo software such as MCNP. The Whisper methodology uses S/U techniques to assist with the selection and weighting of benchmark critical experiments relevant to an application. Once these weights are known, the calculational margin is computed with an extreme value distribution. A baseline MOS is determined from the combination of a term for undetected errors in software and nuclear data processing software (set at 0.005 in this paper) and the uncertainty in k from nuclear data following a GLLS data adjustment.

A benchmark suite, starting with 1,086 benchmark critical experiments and actually using 972 benchmarks after a rejection of outliers identified by a GLLS nuclear data adjustment with an iterative-diagonal χ^2 rejection technique, was discussed and used in the generation of results. Four hypothetical criticality safety case studies were analyzed: the minimum critical mass of ^{239}Pu, an infinite array of LEU lattice fuel assemblies, a series of mixed uranium-plutonium metal-water mixtures, and an infinite MSR lattice with ^{233}U/^{232}Th fuel. For these studies, Whisper generated results that were used to produce curves identifying regions in the parameter space that can be assured to be subcritical. Comparisons with a traditional parametric approach for validation show that Whisper obtains similar or more conservative calculational margins, and comparisons with a non-parametric, rank-order approach show that Whisper obtains less stringent ones.

In the near term, a study should be performed comparing the calculational margins generated by Whisper and the GLLS methods that are implemented in other software such as SCALE. The two methods are both S/U based, but use different approaches to generate the calculational margin, and assessing the similarity or difference between the two results would be insightful.

Going forward, the largest theoretical improvements to Whisper that are needed are further study of similarity coefficients, weighting by benchmark uncertainties, handling of small sample sizes, accounting for correlations of benchmark critical experiments, and a generalization that would allow the incorporation of other responses.

Whisper currently used c_k, the correlation coefficient as a measure of similarity. This choice restricts the measured values to be normally distributed. Alternatives, such as a coefficient derived from the mutual information [25], that allow for non-normally distributed quantities would be advantageous in these cases. In either case, other measures of similarity and methods for generating weighting factors should be investigated.

One drawback of Whisper is that the extreme value distribution tends to bias toward benchmark critical experiments with very large uncertainties. For the analyses performed in Sec. V and in Ref. [15], the benchmark critical experiments were used without any weighting based upon the magnitude of the benchmark uncertainties, and, consequently, a few benchmarks with abnormally large uncertainties may determine the USL. In practice, the ICSBEP process filters out those experiments with too high of an uncertainty to be useful for criticality safety validation. This approach is rather unsatisfying as sometimes it would be useful or even necessary to add experiments with larger uncertainties than would normally qualify for the ICSBEP Handbook, and a more rigorous approach for discounting benchmarks with abnormally large benchmark uncertainties needs to be investigated.

Whisper currently handles the lack of critical experiment data with an ad hoc interpolation to an unweighted calculational margin for a very large and comprehensive benchmark suite. In these cases, it is especially incumbent upon the criticality safety analyst to treat

the results from Whisper with caution, as they may be misleading. While it is unlikely that this problem can be solved for all cases—in the end, the lack of data must be handled based upon experience and engineering judgment—some further research in the handling of small sample sizes and exactly what constitutes such is needed.

The Whisper methodology only partially accounts for experimental correlations in the benchmark critical experiments; these are taken into account in the nuclear data adjustment, but they are neglected in the computation of the calculational margin. Testing during the development of this method, which is not presented in this paper, showed that the assumption of independence typically leads to more conservative calculational margins. The effect on the calculations of $\mathrm{MOS}_{\mathrm{data}}$ and the benchmark rejection, which did account for them where available, showed the impact to be minor on the overall results of the cases presented in this paper. Other work using different approaches, however, has shown that experimental correlations can have a significant impact on the bias and bias uncertainty [43, 44], and therefore directly change the calculational margin. Since few benchmark critical experiment correlations have currently been quantified, the impact of neglecting benchmark correlations in determining the calculational margin are currently restricted to a small portion of practical AOAs; however, more correlation information is gradually becoming available, and the Whisper methodology should be adjusted to account for them consistently. Additionally, it may be possible to extend Whisper to account for the additional correlations of benchmark uncertainties that arise as a result of estimating unknown benchmark uncertainties (see Sec. III.C).

Currently, the Whisper methodology is restricted to using critical benchmark experiment values of the effective multiplication k. While this is of primary interest to criticality safety, other measured quantities in benchmark experiments may be useful as well. Recently, techniques have been developed that allow continuous-energy Monte Carlo calculations of sensitivity coefficients of more general responses [45]. Furthermore, the possibility of adapting and applying a quantity called the coverage ratio [46] to connect other measured responses

such as k should also be investigated in the context of Whisper and similar methods.

ACKNOWLEDGMENTS

This work was jointly funded by the DOE/NNSA Nuclear Criticality Safety Program (NCSP) and the Advanced Scientific Computing (ASC) program. The authors would like to thank the generous consultation by staff at Oak Ridge National Laboratory, including and in no particular order: B. Rearden, C. Perfetti, W.J. Marshall, D. Mueller, and D. Bowen. The authors would also like to thank M. Mitchell at LANL who provided useful discussions related to how computational analysis relates to the more extensive process of performing criticality safety evaluations.

References

[1] "Nuclear Criticality Safety in Operations with Fissionable Material Outside Reactors," American Nuclear Society, American National Standard ANSI/ANS-8.1-1998 (1998) (Reaffirmed 2007).

[2] "Validation of Neutron Transport Methods for Nuclear Criticality Safety Calculations," American Nuclear Society, American National Standard ANSI/ANS-8.24-2007 (2007).

[3] J.T. GOORLEY, et al., "Initial MCNP6 Release Overview", *Nucl. Technol.*, **180**, 298-315 (2012).

[4] M.B. CHADWICK, et al., "ENDF/B-VII.1: Nuclear Data for Science and Technology: Cross Sections, Covariances, Fission Product Yields and Decay Data," *Nucl. Data Sheets*, **112**, 2887 (2011).

[5] J.L. CONLIN, et al., "Continuous Energy Neutron Cross Section Data Tables Based upon ENDF/B-VII.1," Los Alamos National Laboratory Report, LA-UR-13-20137, (2013).

[6] B.L. BROADHEAD, et al., "Sensitivity- and Uncertainty-Based Criticality Safety Validation Techniques," *Nucl. Sci. Eng.*, **146**, 340-366 (2004).

[7] B.T. REARDEN, "Perturbation Theory Eigenvalue Sensitivity Analysis with Monte Carlo Techniques," *Nucl. Sci. Eng.*, **146**, 367-382 (2004).

[8] B.C. KIEDROWSKI, F.B. BROWN, "Adjoint-Based k-Eigenvalue Sensitivity Coefficients to Nuclear Data Using Continuous-Energy Monte Carlo," *Nucl. Sci. Eng.*, **174**, 227-244 (2013).

[9] C.M. PERFETTI, W.R. MARTIN, B.T. REARDEN, M.L. WILLIAMS, "Advanced Methods for Eigenvalue Sensitivity Coefficient Calculations," *Trans. Am. Nucl. Soc.*, **107**, 575-578 (2012).

[10] H.J. SHIM, C.H. KIM, "Adjoint Sensitivity and Uncertainty Analyses in Monte Carlo Forward Calculations," *J. Nucl. Sci. Technol.*, **48**, 1453-1461 (2011).

[11] K.F. RASKACH, A.A. BLYSKAVKA, "An Experience of Applying Iterated Fission Probability Method to Calculation of Effective Kinetics Parameters and keff Sensitivities with Monte Carlo," *Proc. PHYSOR 2010 – Advances in Reactor Physics to Power the Nuclear Renaissance*, Pittsburgh, Pennsylvania, USA, May 9-14 (2010).

[12] S.M. BOWMAN, "SCALE 6: Comprehensive Nuclear Safety Analysis Code System," *Nucl. Technol.*, **174**, 126-148 (2011).

[13] B.T. REARDEN, et. al., "Sensitivity and Uncertainty Analysis Capabilities and Data in SCALE," *Nucl. Technol.*, **174**, 236-288 (2011).

[14] B.C. KIEDROWSKI, "User Manual for Whisper (v1.0.0), Software for Sensitivity- and Uncertainty-Based Nuclear Criticality Safety Validation," Los Alamos National Laboratory Report, LA-UR-14-26436 (2014).

[15] B.C. KIEDROWSKI, et al., "Validation of MCNP6.1 for Criticality Safety of Pu-Metal, -Solution, and -Oxide Systems," Los Alamos National Laboratory Report, LA-UR-14-23352 (2014).

[16] K.D. KIMBALL, E.F. TRUMBLE, "Statistical Methods for Accurately Determining Criticality Code Bias," *Proc. Topl. Mtg. Criticality Safety Challenges in the Next Decade*, Chelan, Washington, USA, Sep. 7-11 (1997).

[17] M.G. NATRELLA, "Experimental Statistics," *National Bureau of Standards Handbook 91*, US Department of Commerce (1963).

[18] D. BISWAS, Q. AO, R. REED, "Comparison of a Few Statistical Methods for Validation," *Proc. Nuclear Criticality Safety Division Topical Mtg.*, Richland, Washington, USA, Sep. 13-17 (2009).

[19] J.J. LICHTENWALTER, S.M. BOWMAN, M.D. DEHART, and C.M. HOPPER, "Criticality Benchmark Guide for Light-Water-Reactor Fuel in Transportation and Storage Packages," NUREG0CR-6361 (ORNL/TM-13211), U.S. Nuclear Regulatory Commission/Oak Ridge National Laboratory (1997).

[20] Q. AO, "A Statistical Methodology for Validating Criticality Analysis Code," *Proc. 8th International Conference on Nuclear Criticality*, St. Petersburg, Russia, May 28 - Jun. 1 (2007).

[21] Q. AO, "USLSA – A Statistical Tool for Criticality Analysis Code Validation," *Trans Am. Nucl. Soc.*, **96**, 271-273 (2007).

[22] Q. AO, "PARANAL: An Efficient Tool for Parametric Analysis of Criticality Safety," *Trans. Am. Nucl. Soc.*, **100**, 359-361 (2009).

[23] F. FERNEX, Y. RICHET, E. LETANG, "MACSENS: A New MORET Tool to Assist Code Bias Estimation," *Proc. Nuclear Criticality Safety Division Topical, Integrating Criticality Safety into the Resurgence of Nuclear Power*, Knoxville, TN USA Sep. 19-22 (2005).

[24] B.C. KIEDROWSKI, F.B. BROWN, P.P.H. WILSON, "Adjoint-Weighted Tallies for k-Eigenvalue Calculations with Continuous-Energy Monte Carlo," *Nucl. Sci. Eng.*, **168**, 226-241 (2011).

[25] P. ATHE, H.S. ABDEL-KHALIK, "Mutual Information: A Generalization of Similarity Indices," *Trans. Am. Nucl. Soc.*, **111**, 1299-1302 (2014).

[26] S.C. VAN DER MARCK, "Benchmarking ENDF/B-VII.1, JENDL-4.0 and JEFF-3.1.1 with MCNP6," *Nucl. Data Sheets*, **113**, 2935-3005 (2012).

[27] J.J. WAGSCHAL, C.M. HOPPER, "Determination of Consistent Benchmarks Used for Nuclear Criticality Safety Analysis Applications," *Trans. Am. Nucl. Soc.*, **93**, 257-259 (2005).

[28] "International Handbook of Evaluated Criticality Safety Benchmark Experiments," Organization for Economic Co-operation and Development Nuclear Energy Agency, NEA/NSC/DOC(95)03 (2013).

[29] "Database for the International Criticality Safety Benchmark Evaluation Project (DICE)," Organization for Economic Co-operation and Development Nuclear Energy Agency, Available online at URL: https://www.oecd-nea.org/science/wpncs/icsbep/dice.html (accessed Apr. 18, 2014).

[30] M. BOCK, M. STUKE, "Determination of Correlations Among Benchmark Experiments by Monte Carlo Sampling Techniques," *Proc. ANS Nuclear Criticality Safety Division Topical Meeting (NCSD2013)*, Wilmington, NC, USA Sep. 29 - Oct. 3 (2013).

[31] B.T REARDEN, K.J. DUGGAN, F. HAVLUJ, "Quantification of Uncertainties and Correlations in Criticality Experiments with SCALE," *Proc. ANS Nuclear Criticality Safety Division Topical Meeting (NCSD2013)*, Wilmington, NC, USA Sep. 29 - Oct. 3 (2013).

[32] T. IVANOVA, E. IVANOV, G.E. BIANCHI, "Establishment of Correlations for Some Critical and Reactor Physics Experiments," *Nucl. Sci. Eng.*, **178**, 311-325 (2014).

[33] W.J. MARSHALL, B.T. REARDEN, "Determination of Experimental Correlations Using the Sampler Sequence Within SCALE 6.2," *Trans. Am. Nucl. Soc.*, **111**, 867-870 (2014).

[34] R.J. KAMM, "Validation of MCNP5 on the Ganglion Cyst Computer Cluster with Various Cross Section Libraries," Los Alamos National Laboratory Report, NCS-TECH-007-002 (2007).

[35] F.B. BROWN, B.C. KIEDROWSKI, J.S. BULL, "Verification of MCNP5-1.60 and MCNP6.1 for Criticality Safety Applications," Los Alamos National Laboratory Report, LA-UR-13-22196 (2013).

[36] R.D. MOSTELLER, "An Expanded Criticality Validation Suite for MCNP," Los Alamos National Laboratory Report, LA-UR-10-06230, Rev. 3 (2010).

[37] A.C. KAHLER, et al., "ENDF/B-VII.1 Neutron Cross Section Data Testing with Critical Assembly Benchmark and Reactor Experiments," *Nucl. Data Sheets*, **112**, 2997 (2011).

[38] M.L. WILLIAMS, B.T. REARDEN, "SCALE-6 Sensitivity/Uncertainty Methods and Covariance Data," *Nucl. Data Sheets*, **109**, 2796-2800 (2008),

[39] R.J. MCCONN, et al., "Compendium of Material Composition Data for Radiation Transport Modeling," Pacific Northwest National Laboratory Report, PNNL-15870, Rev. 1 (2011).

[40] "International Handbook of Evaluated Reactor Physics Benchmark Experiments," Organization for Economic Co-operation and Development Nuclear Energy Agency, NEA/NSC/DOC(2006)1 (2013).

[41] M.E. DUNN, B.T. REARDEN, "Application of Sensitivity and Uncertainty Analysis Methods to a Validation Study for Weapons-Grade Mixed-Oxide Fuel," *Proc. 2001 Embedded Topical Mtg. on Practical Implementation of Nuclear Criticality Safety*, Reno, Nevada, USA, Nov. 11-15 (2001).

[42] D.E. MEULLER, B.T. REARDEN, D.F. HOLLENBACH, "Application of the SCALE TSUNAMI Tools for the Validation of Criticality Safety Calculations Involving ^{233}U," Oak Ridge National Laboratory Report, ORNL/TM-2008/196 (2009).

[43] T.T. IVANOVA, et. al., "Influence of the Correlations of Experimental Uncertainties on Criticality Prediction," *Nucl. Sci. Eng.*, **145**, 97-104, (2003).

[44] M. BOCK, M. BEHLER, "Impact of Correlated Benchmark Experiments on the Computational Bias in Criticality Safety Assessment," *Proc. ANS Nuclear Criticality Safety Division Topical Meeting (NCSD2013)*, Wilmington, NC, USA Sep. 29 - Oct. 3 (2013).

[45] C.M. PERFETTI, B.T. REARDEN, "A New Method for Calculating Generalized Response Sensitivities in Continuous-Energy Monte Carlo Applications in SCALE," *Trans. Am. Nucl. Soc.*, **109**, 739-742 (2013).

[46] P. ATHE, U. MERTYUREK, H.S. ABDEL-KHALIK, "Determination of Bias, Bias Uncertainty, and Coverage Using Data Assimilation," *Trans. Am. Nucl. Soc.*, **111**, 1299-1302 (2014).

Table I: Non-Parametric Margins for Rank-Order Method ($p = 0.95$)

C_{NP}	m_{NP}
$C_{NP} > 0.9$	0.00
$0.8 < C_{NP} \leq 0.9$	0.01
$0.7 < C_{NP} \leq 0.8$	0.02
$0.6 < C_{NP} \leq 0.7$	0.03
$0.5 < C_{NP} \leq 0.6$	0.04
$0.4 < C_{NP} \leq 0.5$	0.05
$C_{NP} \leq 0.4$	Additional data needed.

Table II: Benchmark Suite Summary

ICSBEP Identifier	# Included	# After Rejection
HEU-MET-FAST	251	222
HEU-MET-INTER	4	2
HEU-MET-THERM	4	2
HEU-MET-MIXED	8	8
HEU-COMP-INTER	1	1
HEU-COMP-THERM	25	17
HEU-SOL-THERM	93	91
IEU-MET-FAST	12	12
IEU-COMP-THERM	1	1
LEU-COMP-THERM	182	178
LEU-SOL-THERM	27	25
MIX-MET-FAST	33	32
MIX-MET-MIXED	1	1
MIX-COMP-FAST	2	2
MIX-COMP-INTER	1	1
MIX-COMP-THERM	15	15
MIX-SOL-THERM	21	12
PU-MET-FAST	53	49
PU-COMP-FAST	1	1
PU-COMP-INTER	1	1
PU-COMP-MIXED	34	17
PU-SOL-THERM	158	142
U233-MET-FAST	10	8
U233-COMP-THERM	9	9
U233-SOL-INTER	33	24
U233-SOL-THERM	106	94
Total	1086	972

Table III: Summary of Whisper Results

Case	CM	MOS	USL
Pu Mass ($H/X = 0$)	0.0128	0.0070	0.980
Pu Mass ($H/X = 4$)	0.0310	0.0092	0.959
Pu Mass ($H/X = 700$)	0.0125	0.0066	0.981
LEU Lattice (Enirch = 2.8%, $P/D = 1.35$)	0.0162	0.0072	0.976
LEU Lattice (Enirch = 2.8%, $P/D = 1.80$)	0.0137	0.0066	0.976
U/Pu Mix (U = 20 g/L, Pu = 140 g/L)	0.0351	0.0090	0.956
U/Pu Mix (U = 200 g/L, Pu = 180 g/L)	0.0340	0.0089	0.957
MSR Lattice (^{233}U = 6.4%, $P/D = 1.20$)	0.0475	0.0206	0.932
MSR Lattice (^{233}U = 6.0%, $P/D = 1.80$)	0.0471	0.0203	0.932

Table IV: Comparison of Calculational Margins Using Standard Approaches and Whisper

Case	Benchmark Set	#	Normal?	Parametric	Rank-Order	Whisper
Pu Mass (Fast)	PU-MET-FAST	50	yes	0.0109	0.0281	0.0128
Pu Mass (Inter)	PU-COMP-MIXED	19	no	0.0455	0.0773	0.0310
Pu Mass (Thermal)	PU-SOL-THERM	154	yes	0.0120	0.0228	0.0125
LEU Lattice	LEU-SOL-THERM	178	no	0.0080	0.0223	0.0162
U/Pu Mix	MIX-SOL-THERM	21	yes	0.0318	0.0454	0.0351
MSR Lattice	None Available	0	—	—	—	0.0475

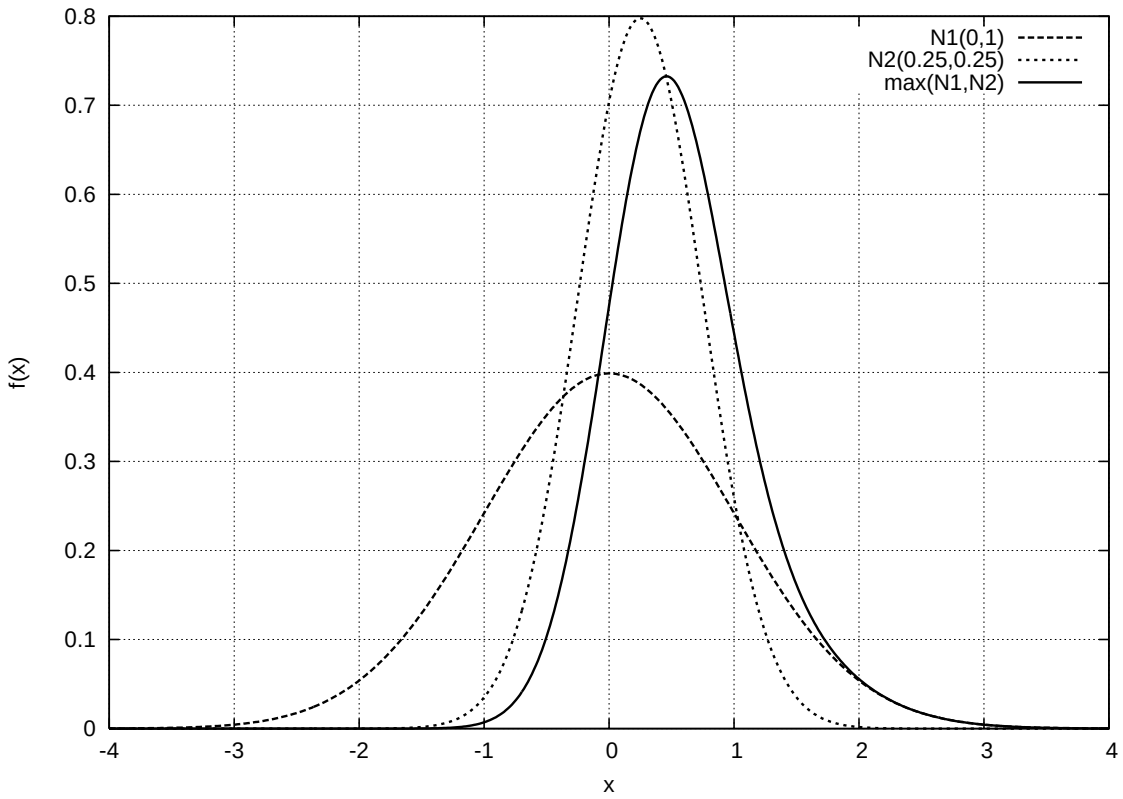

Figure 1: Probability density for the maximum of two normal distributions with different means and variances.

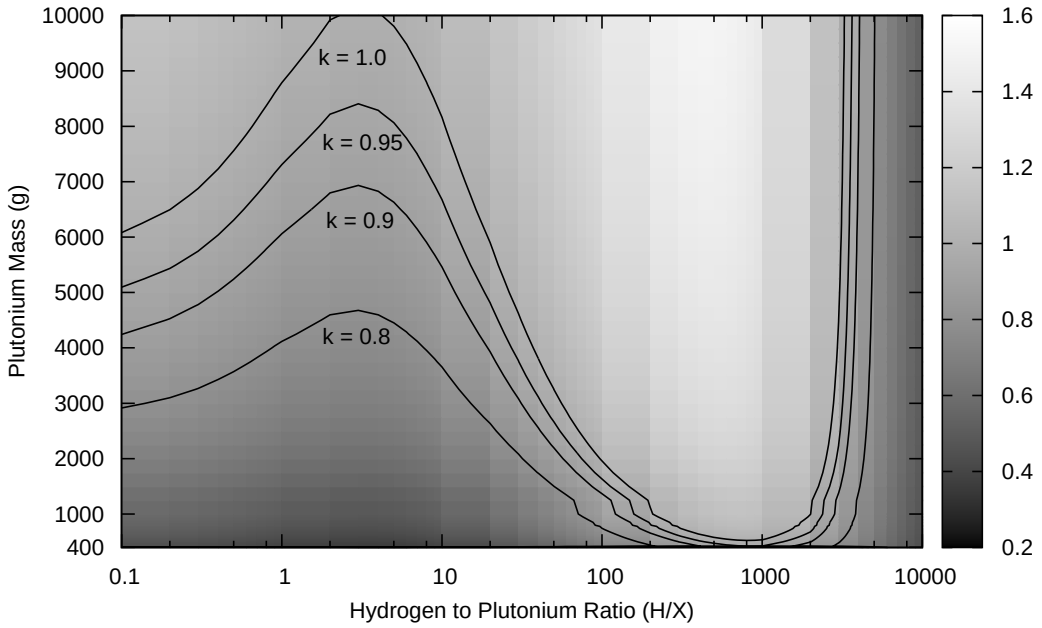

Figure 2: Variation of k as a function of ^{239}Pu mass and H/X for the ^{239}Pu critical mass study.

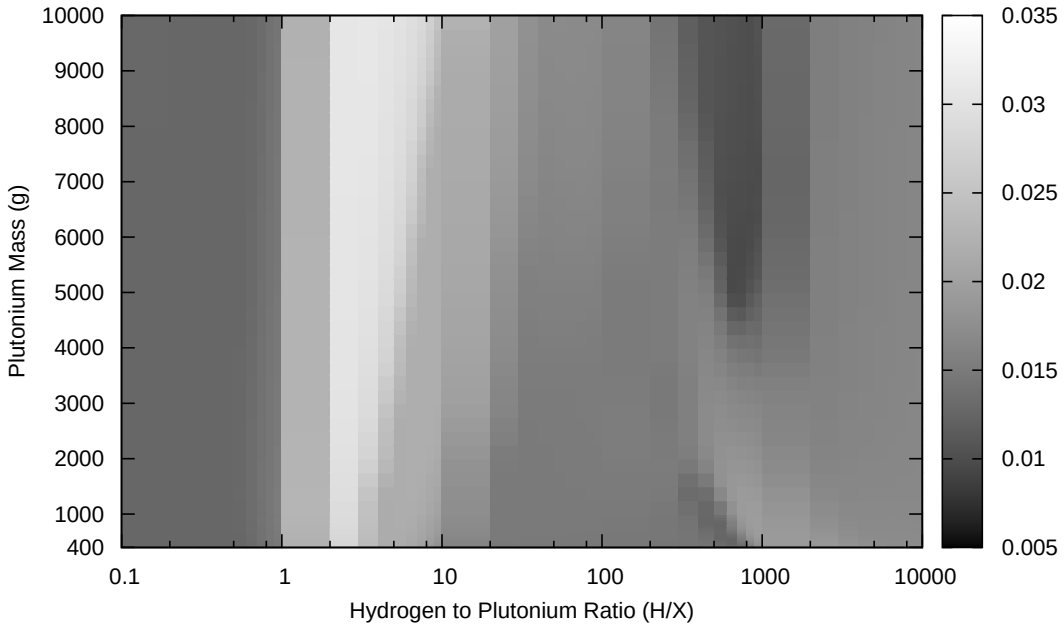

Figure 3: Calculational margin as a function of ^{239}Pu mass and H/X for the ^{239}Pu critical mass study

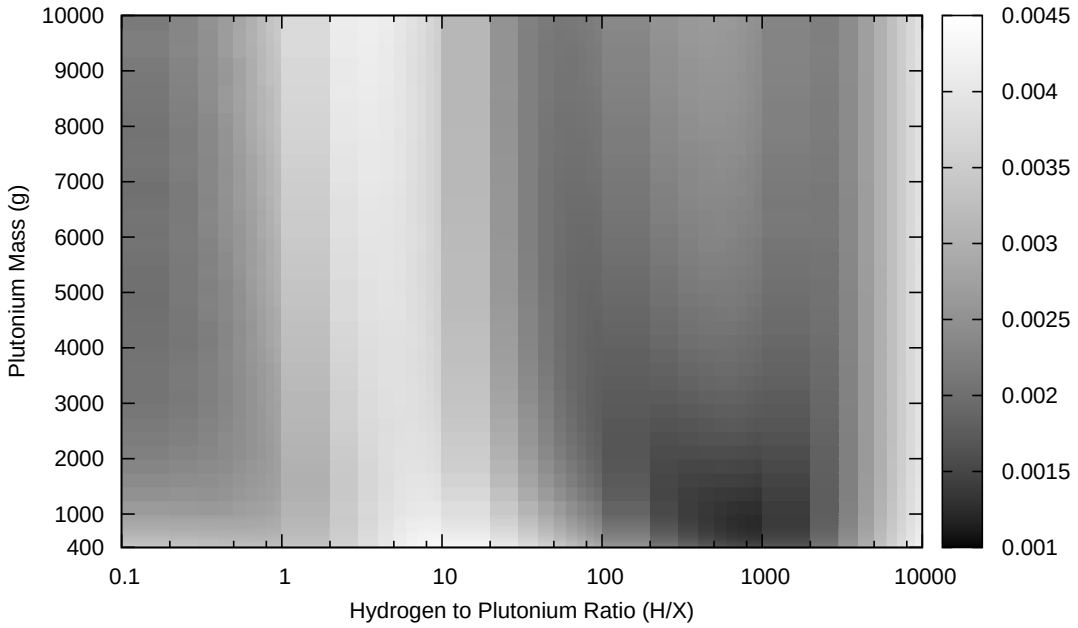

Figure 4: Margin of subcriticality for nuclear data uncertainties (99% confidence level) as a function of ^{239}Pu mass and H/X for the ^{239}Pu critical mass study.

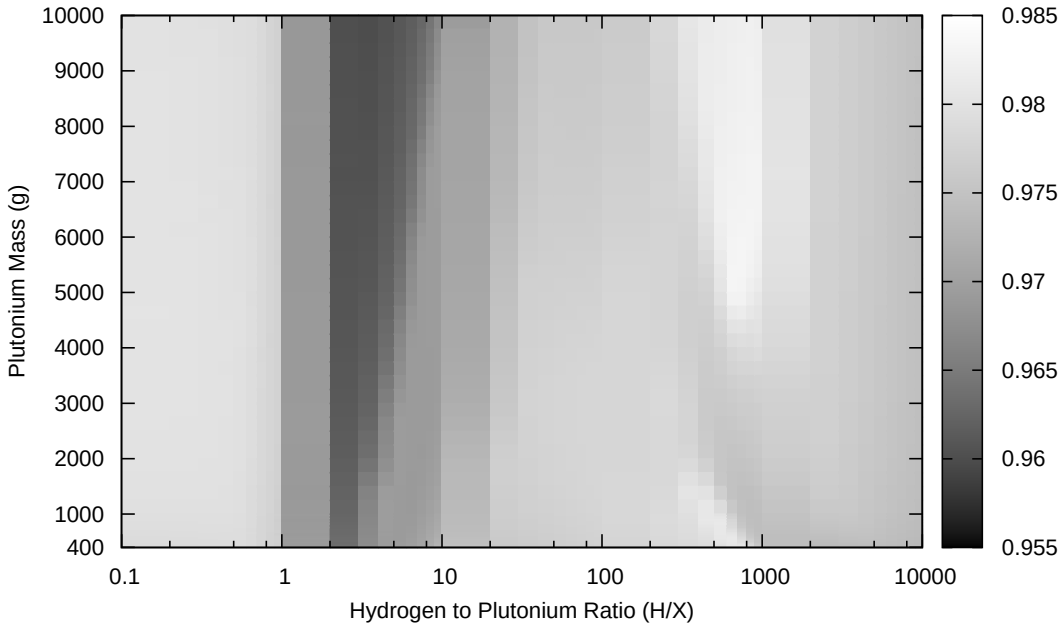

Figure 5: Upper subcritical limit as a function of ^{239}Pu mass and H/X for the ^{239}Pu critical mass study.

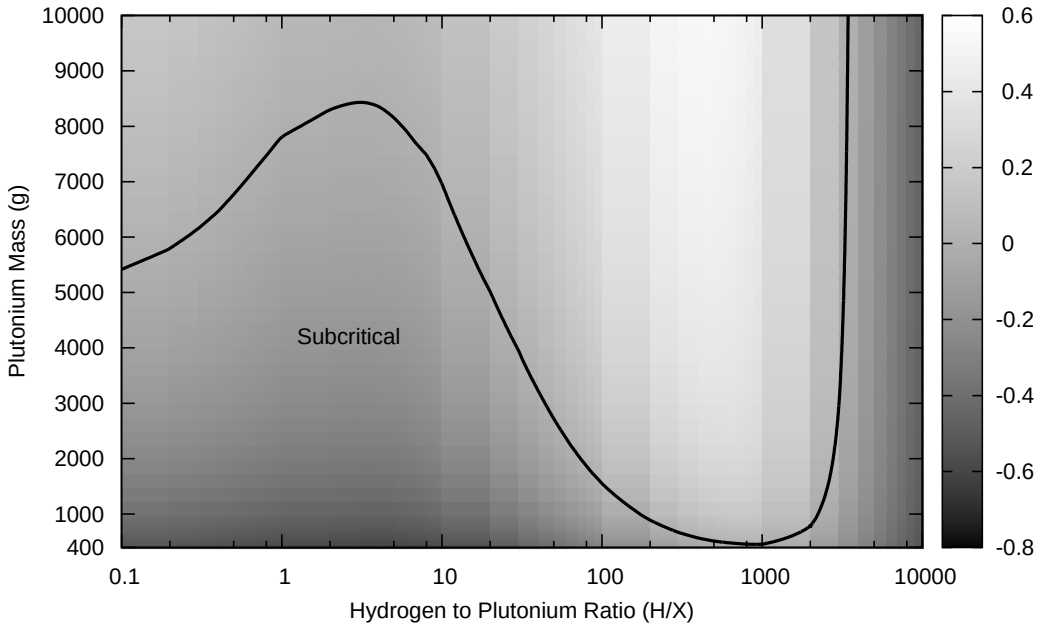

Figure 6: Variation of δ_A, the amount k exceeds the USL, as a function of ^{239}Pu mass and H/X for the ^{239}Pu critical mass study.

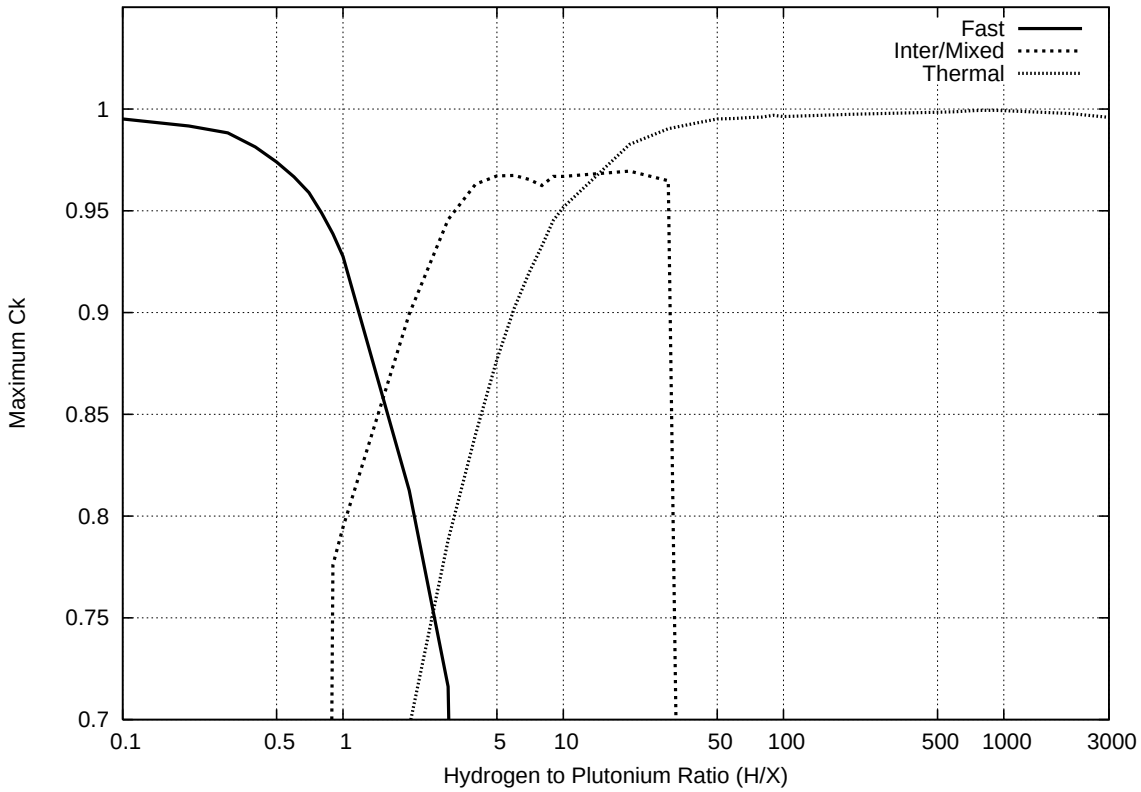

Figure 7: Maximum c_k for benchmarks grouped by neutron energy spectrum as a function of H/X along the subcriticality contour.

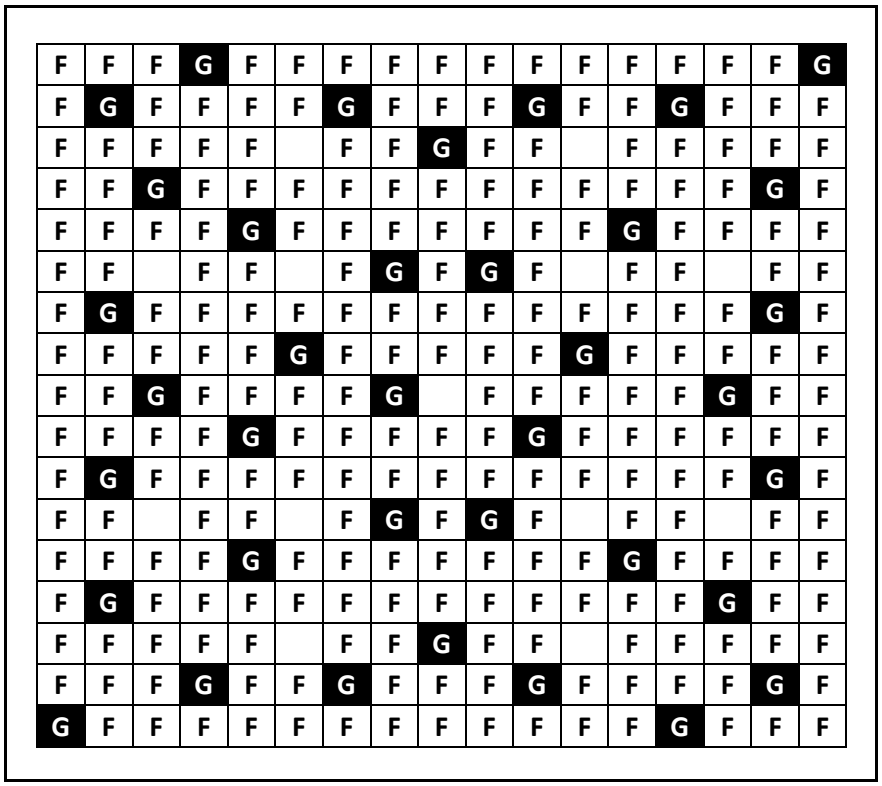

Figure 8: Pin layout for the LEU lattice. **F** represents a fuel pin, **G** represents a fuel pin with Gd_2O_3, and an empty space is a water/instrument tube.

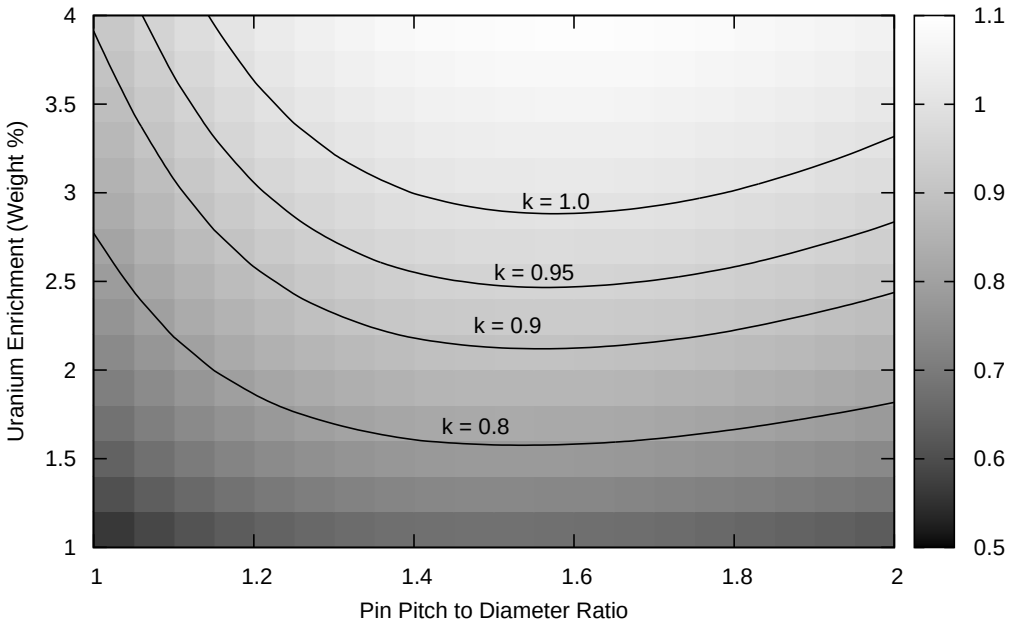

Figure 9: Variation of k as a function of uranium enrichment and P/D for the LEU lattice test case.

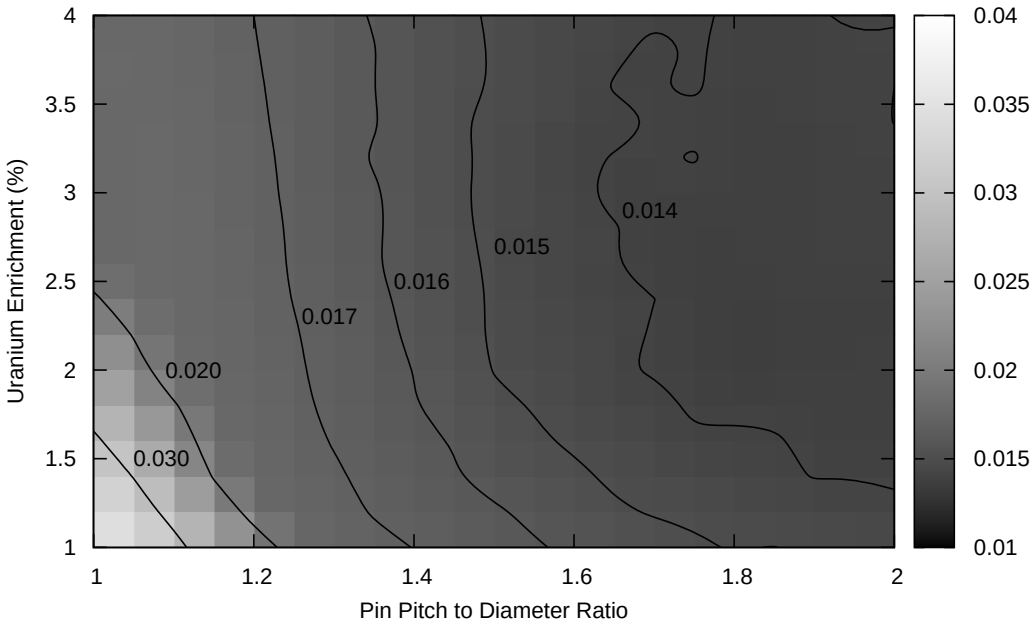

Figure 10: Calculational margin as a function of uranium enrichment and P/D for the LEU lattice test case.

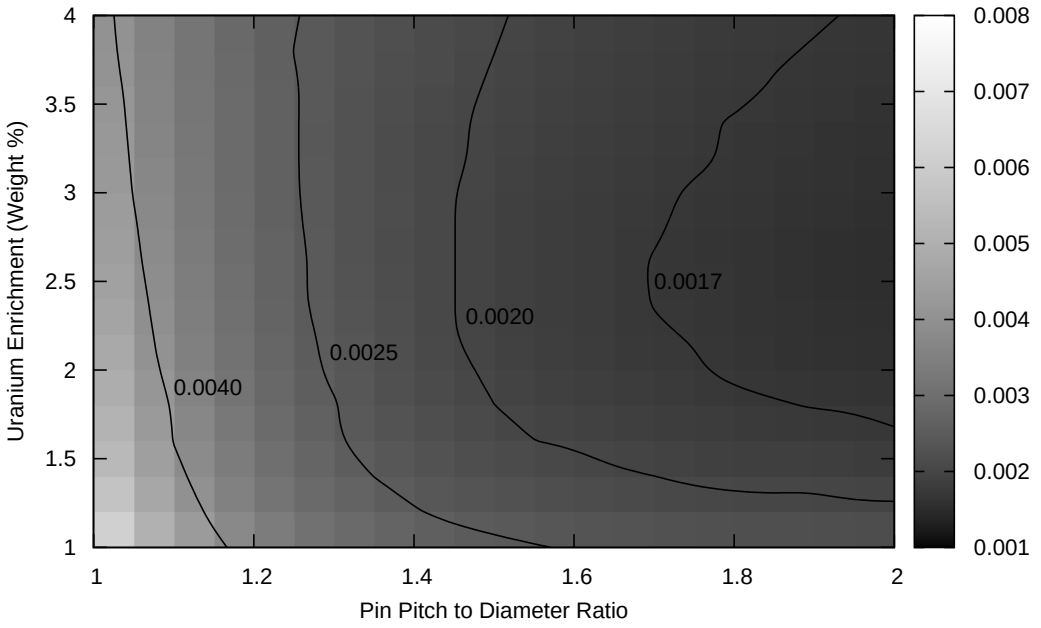

Figure 11: Margin of subcriticality for nuclear data uncertainties (99% confidence level) as a function of uranium enrichment and P/D for the LEU lattice test case.

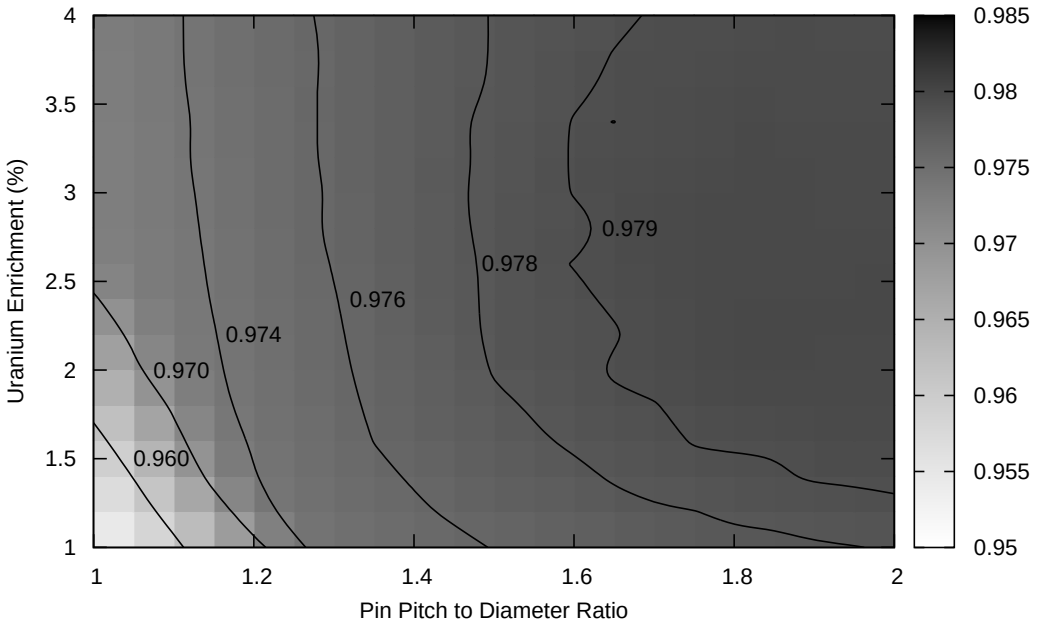

Figure 12: Upper subcritical limit as a function of uranium enrichment and P/D for the LEU lattice test case.

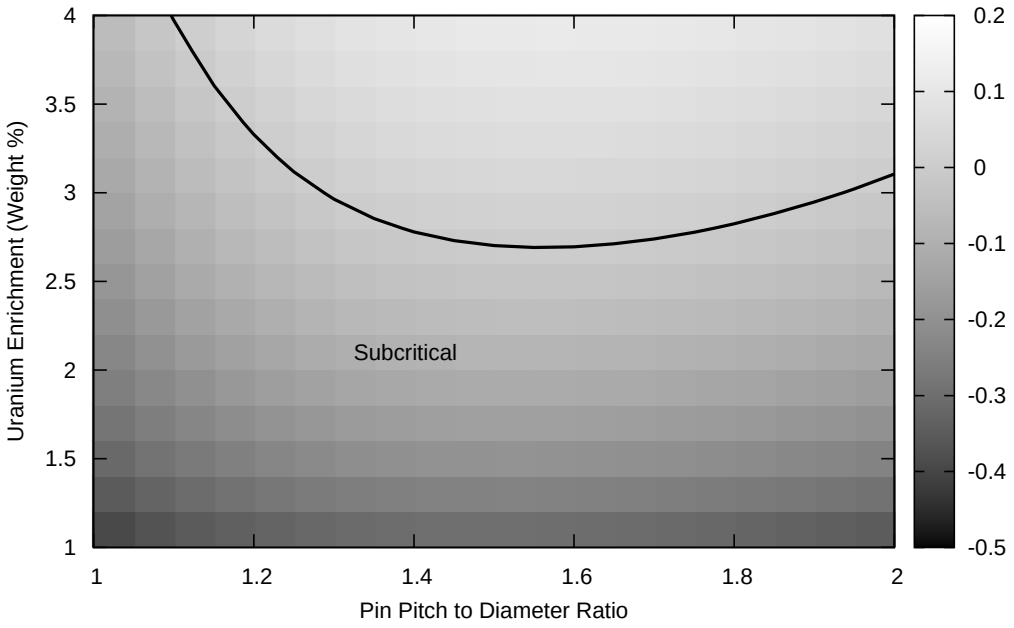

Figure 13: Variation of δ_A, the amount k exceeds the USL, as a function of uranium enrichment and P/D for the LEU lattice test case.

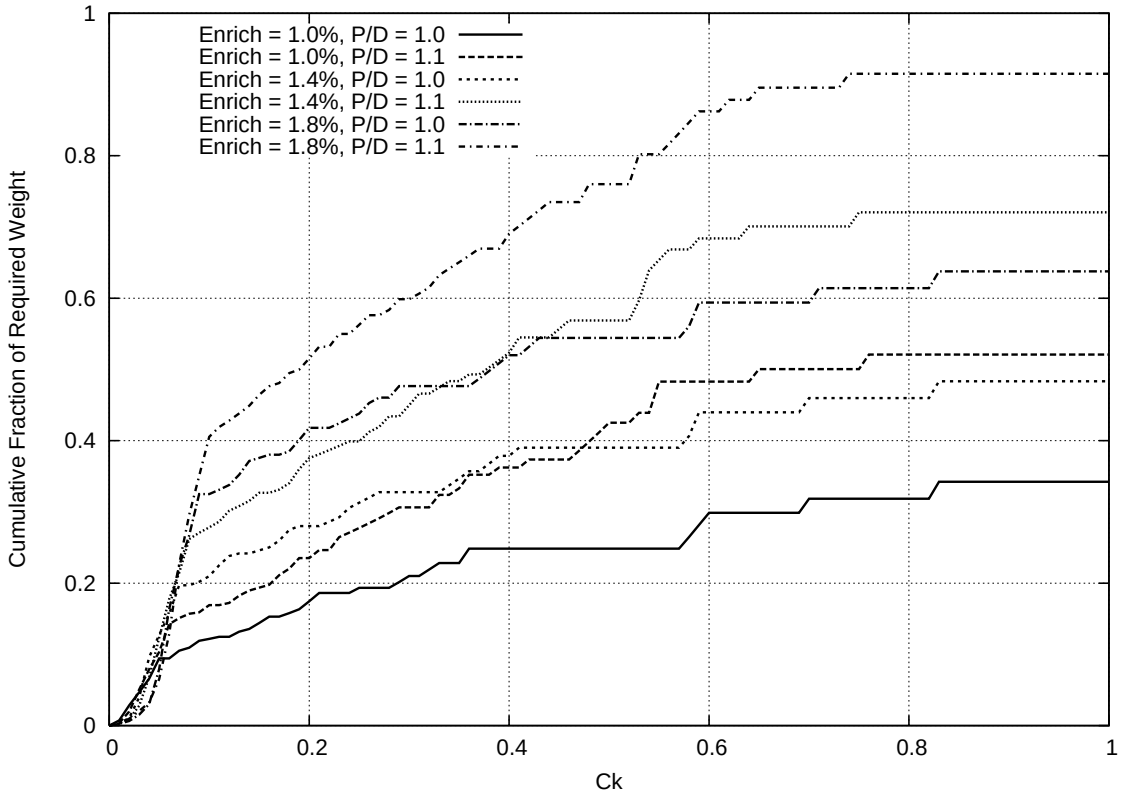

Figure 14: Cumultative fraction of required weight as a function of c_k for select cases of the LEU lattice test case.

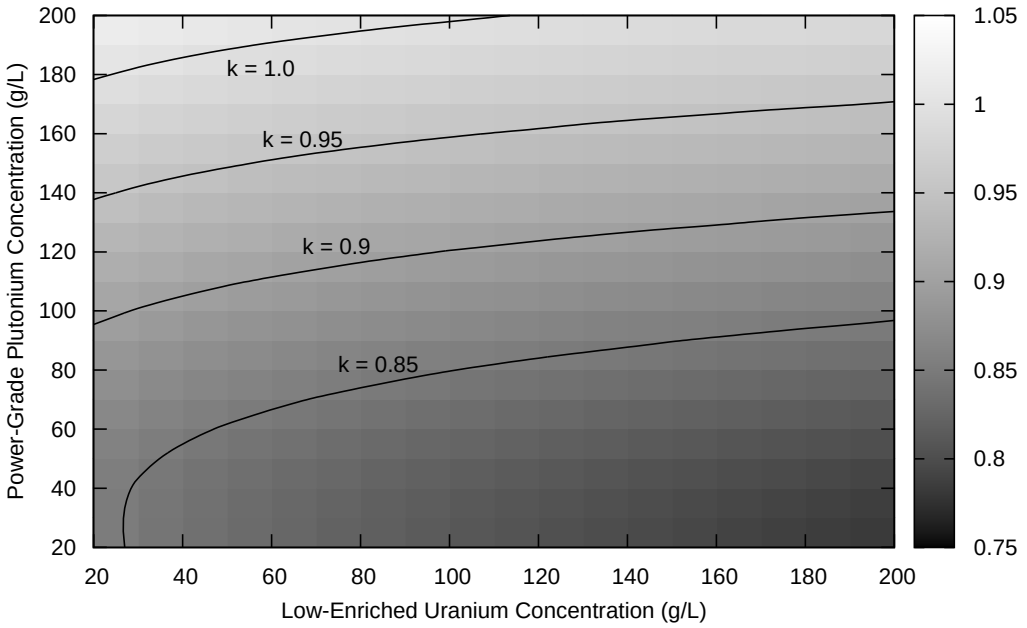

Figure 15: Variation of k as a function of uranium and plutonium concentrations in g/L for the mixed uranium-plutonium metal-water mixture test case.

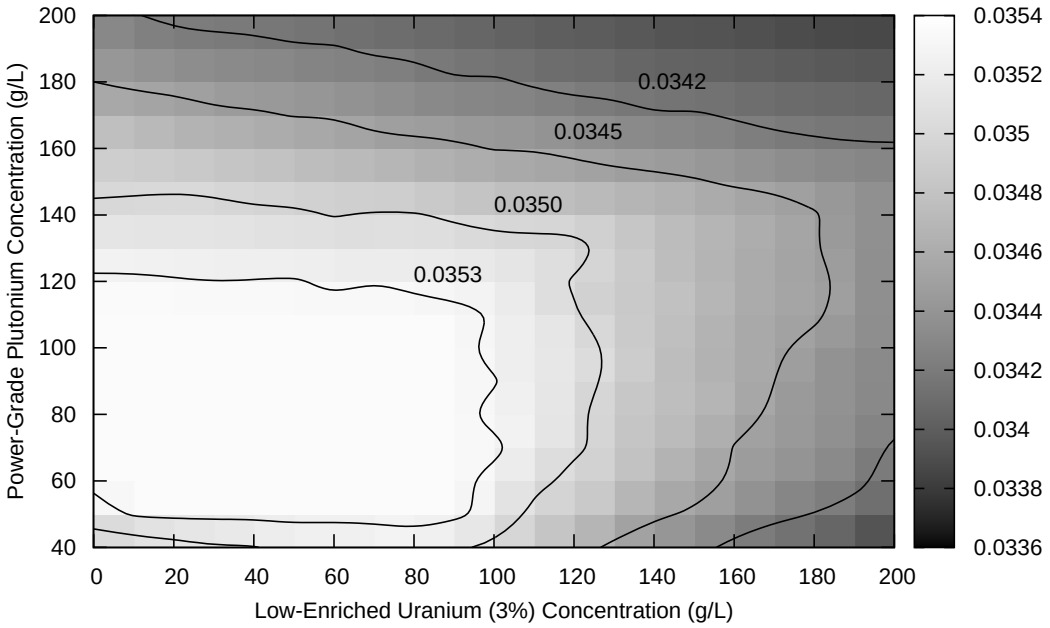

Figure 16: Calculational margin as a function of uranium and plutonium concentrations in g/L for the mixed uranium-plutonium metal-water mixture test case.

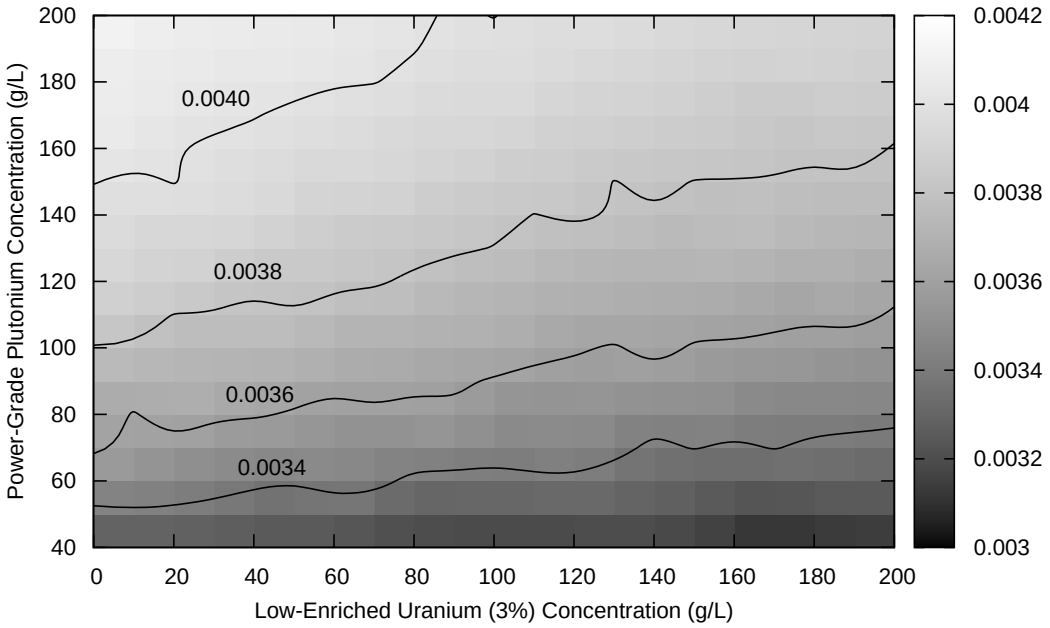

Figure 17: Margin of subcriticality for nuclear data uncertainties (99% confidence level) as a function of uranium and plutonium concentrations in g/L for the mixed uranium-plutonium metal-water mixture test case.

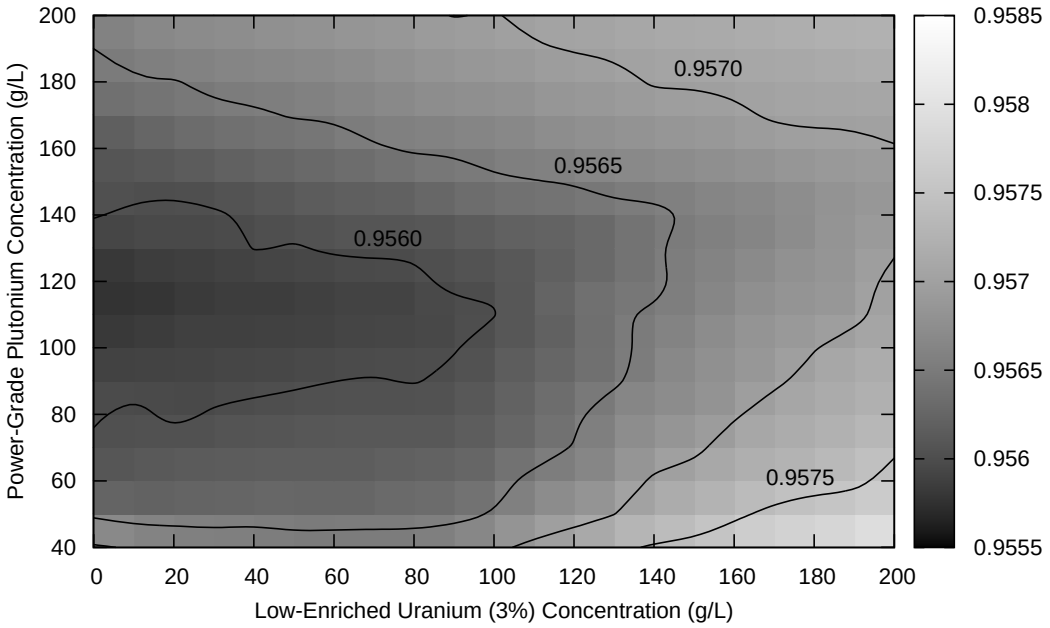

Figure 18: Upper subcritical limit as a function of uranium and plutonium concentrations in g/L for the mixed uranium-plutonium metal-water mixture test case.

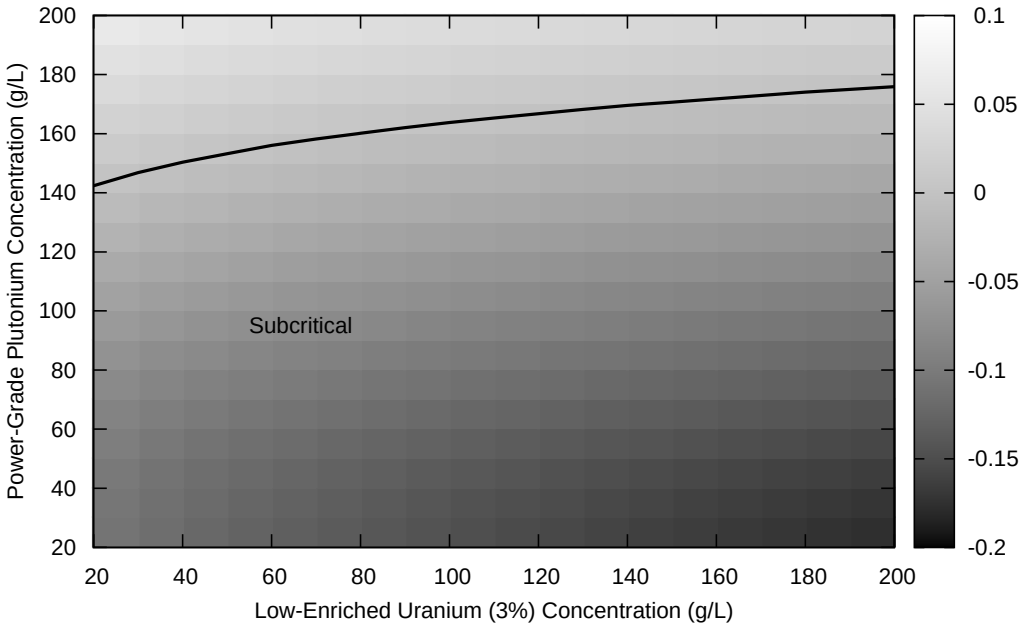

Figure 19: Variation of δ_A, the amount k exceeds the USL, as a function of uranium and plutonium concentrations in g/L for the mixed uranium-plutonium metal-water mixture test case.

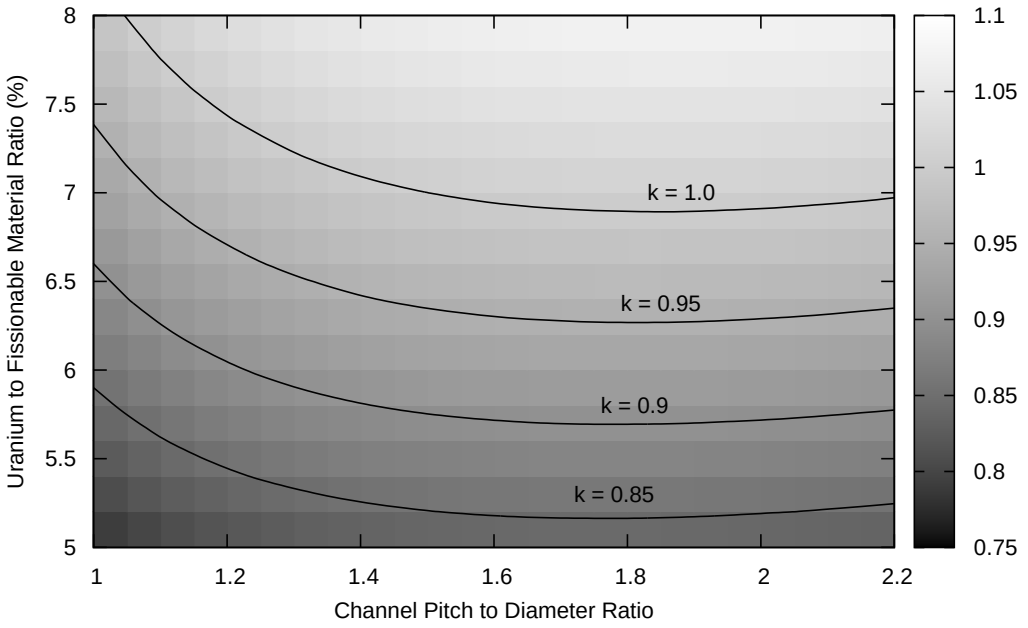

Figure 20: Variation of k as a function of uranium fraction and P/D for the MSR lattice test case.

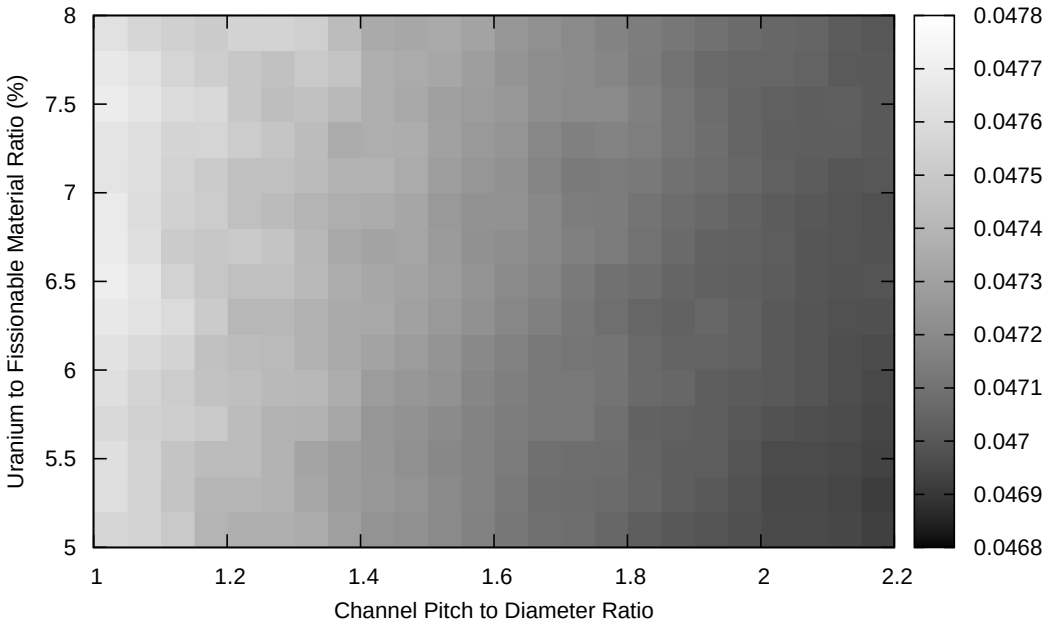

Figure 21: Calculational margin as a function of uranium fraction and P/D for the MSR lattice test case.

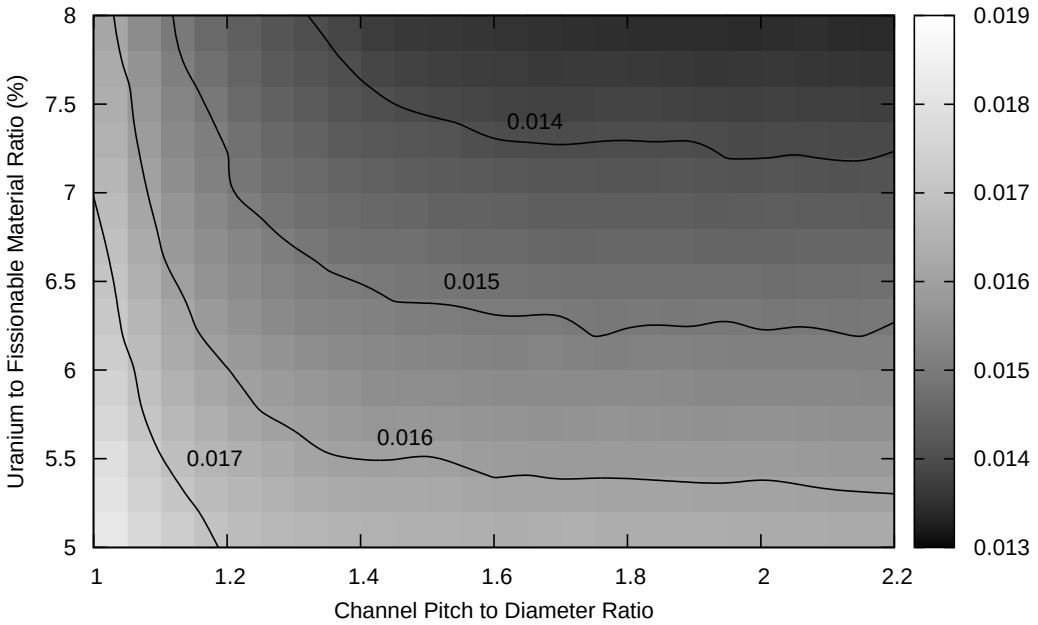

Figure 22: Margin of subcriticality for nuclear data uncertainties (99% confidence level) as a function of uranium fraction and P/D for the MSR lattice test case.

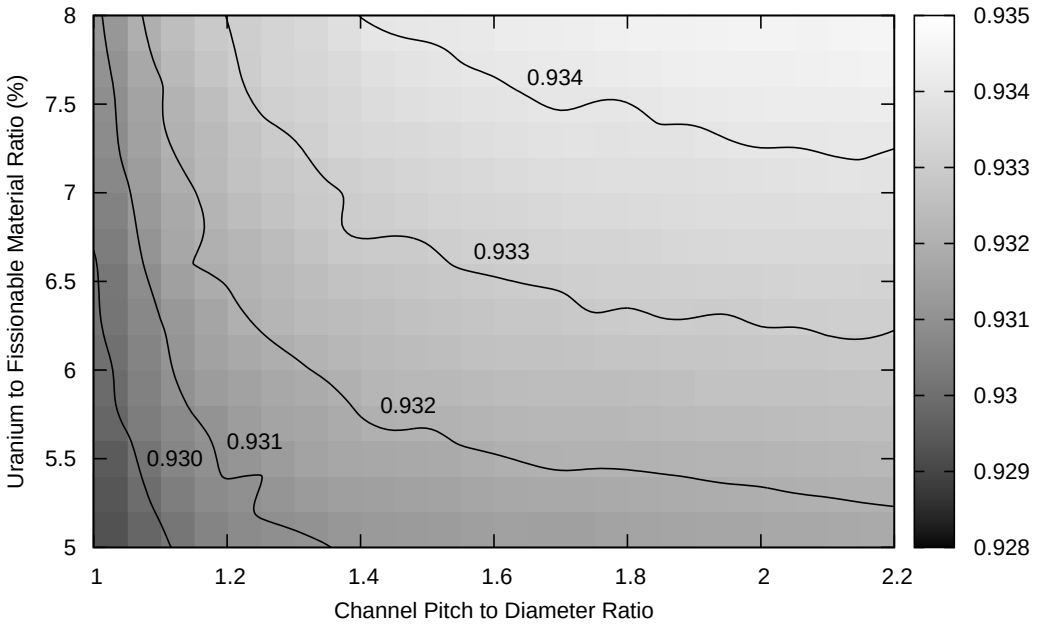

Figure 23: Upper subcritical limit as a function of uranium fraction and P/D for the MSR lattice test case.

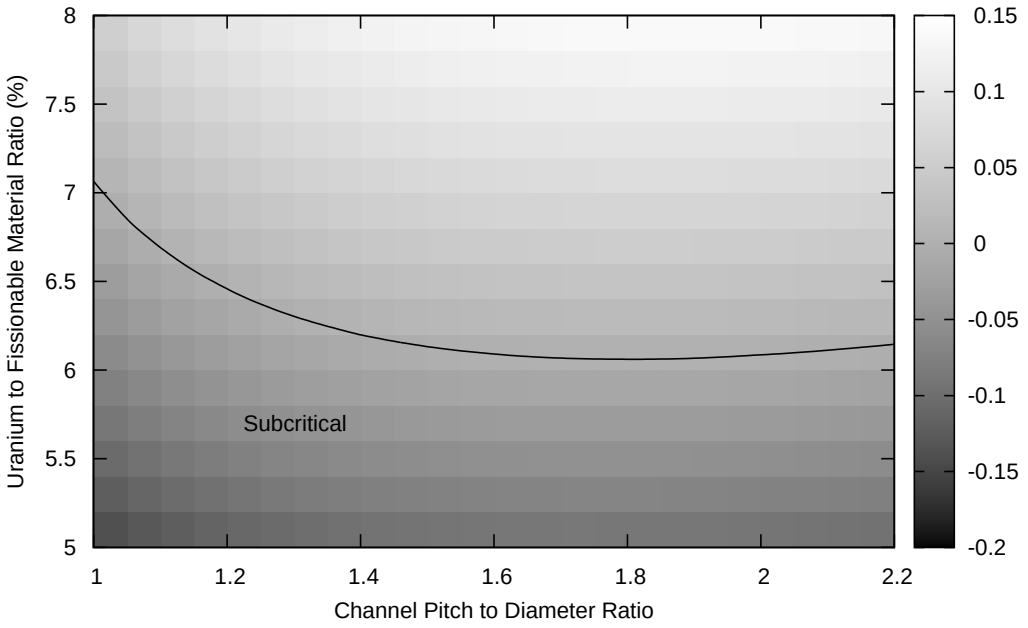

Figure 24: Variation of δ_A, the amount k exceeds the USL, as a function of uranium fraction and P/D for the MSR lattice test case.