M&C 2017 - International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering, Jeju, Korea, April 16-20, 2017, on USB (2017)

Continued Investigation of Metrics for Predicting Undersampling Biases in Monte Carlo Simulations[1]

Christopher Perfetti* and Bradley Rearden

Oak Ridge National Laboratory, P.O. Box 2008, Bldg. 5700, Oak Ridge, TN 37831-6170, USA
** perfetticm@ornl.gov*

Abstract – *This paper investigates the strength of statistical metrics for predicting the onset and magnitude of bias in Monte Carlo tally estimates due to fission source undersampling in eigenvalue simulations. Previous studies found that metrics which had showed potential for predicting undersampling biases in flux and eigenvalue estimates in multigroup simulations had difficulty predicting biases for reaction rate estimates in continuous-energy simulations, but the significant degree of stochastic uncertainty present in the tally bias estimates made it difficult draw definitive conclusions. This study utilized approximately 20 times as many active neutron histories to reduce the stochastic uncertainty in tally bias estimates and draw conclusions with a higher degree of certainty. These more highly converged results produced similar trends to what was previously observed – the undersampling metrics were marginally effective at predicting the magnitude of undersampling biases, and stochastic uncertainty once again made it difficult to fully evaluate the strength of the metrics.*

I. INTRODUCTION

This study continued investigations on the viability of several statistical metrics for predicting biases due to undersampling in continuous-energy (CE) Monte Carlo transport simulation tally estimates [1,2]. Undersampling is a phenomenon in which a Monte Carlo simulation does not sample enough particle histories per generation to adequately sample fission sites and interact with tally regions in a problem, resulting in biases in tally estimates that are much larger than the statistical variance. These biases can potentially lead to erroneous conclusions regarding system performance and safety. Previous studies found that biases due to undersampling could be as large as several hundred percent mille (pcm) for eigenvalue estimates [3,4] and up to tens or hundreds of percent for flux tally estimates [4].

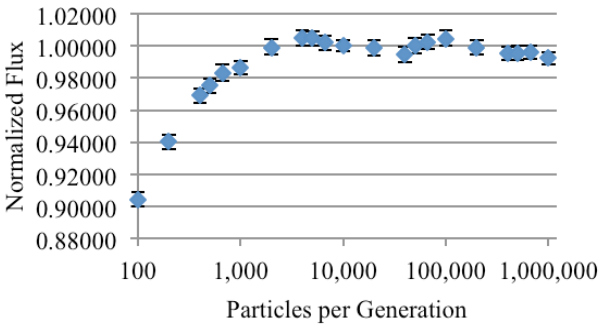

Fig. 1. Flux tallies in the top axial section of a radially reflected PWR assembly [4] show significant bias when fewer than 1,000 histories per generation are simulated.

For example, Perfetti and Rearden considered neutron flux tallied over an axial segment of a radially reflected PWR assembly [4]. Significant undersampling of this region was not expected because the assembly is reflected and the axial power profile is relatively flat. However, the parametric study plotted in Fig. 1 shows undersampling bias far exceeding the statistical error estimates, especially when fewer than 1,000 histories per generation were simulated. This type of bias is not apparent unless analysts perform similar parametric studies, which are computationally expensive.

Several experts have suggested exploring variance reduction schemes to mitigate undersampling bias – for example, the FW-CADIS methodology could be used to distribute histories more uniformly throughout a simulation [5]. Because a rigorous level of convergence is necessary to adequately quantify the magnitude of undersampling biases, it is ideal to assess the performance of potential variance reduction schemes using a metric to infer the magnitude of undersampling biases.

Developers have proposed several statistical metrics to identify undersampling biases without resorting to parametric studies. These metrics can be applied and evaluated while simulations are still running [1,2]. As an example, the Tally Entropy metric shown in Fig. 2 showed promise in predicting the magnitude of undersampling biases for eigenvalue and flux tally estimates in multigroup Monte Carlo simulations [4]. In this study, the effectiveness of each undersampling metric was evaluated by plotting the value of the metric scores against the *Fraction of Undersampling*, which is the relative difference between the biased and reference tally scores [1].

$$Fraction\ of\ Undersampling = \frac{|Biased\ Tally - Reference\ Tally|}{Reference\ Tally} \quad (1)$$

This is an equivalent expression for relative error, which is ideally zero.

These undersampling metrics were first applied to eigenvalue and flux estimates in multigroup simulations of single infinitely reflected fuel assemblies [1], and they have been extended to eigenvalue, flux, and reaction-rate tallies for the same systems using CE physics [2].

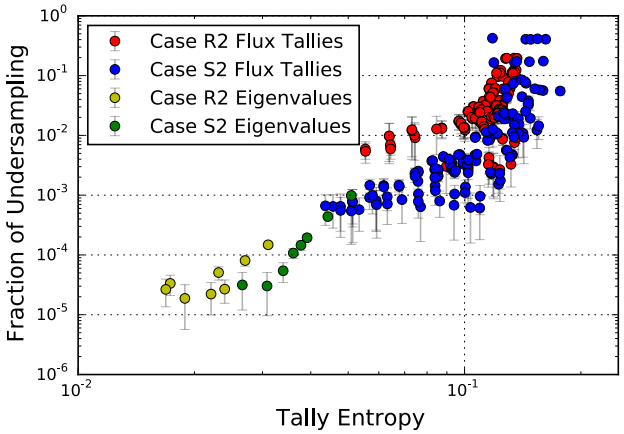

Fig. 2. Magnitude of multigroup undersampling biases (relative difference) vs. the Tally Entropy metric score [1].

II. CHALLENGES TO THE CE TSUNAMI-3D UNDERSAMPLING METRICS

Ideally, these undersampling metrics would be universally applicable, accurately predicting the magnitude of undersampling biases in flux, eigenvalue, reaction rate, and potentially sensitivity tally estimates in systems with substantially different neutron spectra. Unfortunately, the undersampling tallies poorly predict undersampling biases for reaction rates in simulations using CE physics [2]. As shown in Fig. 3 and discussed in Ref. 2, the undersampling metric scores were somewhat correlated to the magnitude of the undersampling biases. Unfortunately, the correlation is less strong than correlations observed for multigroup simulations in Ref. 1. It is difficult to draw conclusions on the effectiveness of these undersampling metrics based on this data because many of these tally estimates were insufficiently converged due to long simulation runtimes and limited computational resources.

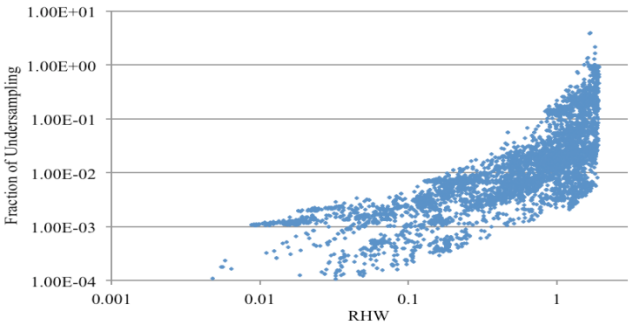

Fig. 3. Magnitude of CE undersampling biases vs. the Heidelberger-Welch RHW metric score [2].

This study uses more highly converged tally and bias estimates to reexamine the predictive capability of the undersampling metrics in CE TSUNAMI-3D, a CE Monte Carlo sensitivity code in the SCALE 6.2 Code System [6]. Eigenvalue, flux, and energy-integrated reaction rate tally estimates are calculated for axial regions of a radially reflected PWR fuel assembly [4]. Undersampling metrics are then calculated and plotted against the *Fraction of Undersampling* calculated from a reference simulation to determine their potential for predicting the magnitude of undersampling biases. Similar to Refs. 1 and 2, the undersampling metrics calculated for this study include:

1. the average number of nonzero tally scores per generation for each tally,
2. the Heidelberger-Welch Relative Half-Width (RHW) for each tally [7],
3. the Tally Entropy for each tally [1], and
4. the true statistical uncertainty in each tally.

References 1 and 2 also included calculation of the Geweke Z-Score; this metric has since been removed from CE TSUNAMI-3D because of its poor correlation with undersampling bias and its relatively large memory footprint.

The simulations presented here calculated undersampling biases and metrics using 30 independent simulations with 100 million active histories, which resulted in the simulation of 20× more active histories than in Perfetti and Rearden's 2016 study [2]. These 30 independent simulations were repeated using different numbers of neutrons simulated per generation (NPG) and active generations, but the same total number of active histories were used. This demonstrated how the impact of undersampling diminishes for simulations with high NPG values. Reference values were calculated for the eigenvalues, fluxes, and reaction rates using 10 independent simulations with 1 billion active histories and 5 million NPG.

III. EFFECTIVENESS OF UNDERSAMPLING METRICS

1. Scores per Generation Metric

The scores per generation metric tracks the average number of nonzero tally scores per generation in a tally region. In principle, tallies that receive more scores per generation should be less prone to the effects of undersampling.

Figure 4 plots the scores per generation against the undersampling biases for the eigenvalue, flux, and reaction rate tallies from the 18 different axial levels within the R2 fuel assembly. This figure compiles the results of all seven NPG realizations using a different color for each realization. The reaction rates in this figure include the total, fission,

fission neutron production, capture, (n,γ), elastic scatter, inelastic scatter, (n,2n), (n,α), (n,p), and (n,d) interactions. This figure and all subsequent figures only plot biases with a relative uncertainty of less than 50%.

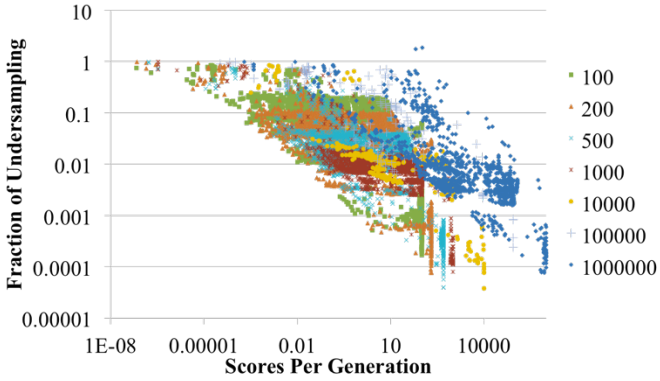

Fig. 4. Relative undersampling biases vs. the scores per generation.

As shown in Fig. 4, tallies that receive more nonzero scores within a generation generally produce smaller undersampling biases. However, as with the results from the previous study, the trend is only weakly predictive. For example, tallies that receive 1 nonzero score on average per generation may produce biases ranging from 0.1% to greater than 100%. Bands of data points appear at several locations on the graph, indicating that a given *Fraction of Undersampling* can occur for a broad range of scores per generation (and vice versa, especially around 100 scores per generation). These bands of data points tend to be clustered for tallies that use the same NPG, which suggests that this metric is not especially effective at determining which tallies are being undersampled within a single simulation. Overall, this metric was effective for generally predicting the presence of undersampling biases, but it was not correlated tightly enough with the magnitude of the undersampling biases to accurately predict the impact of undersampling.

2. Heidelberger-Welch RHW Metric

The Heidelberger-Welch RHW metric examines whether the length of a Markov chain is sufficient to provide accurate estimates for the mean value and uncertainty of a parameter. The metric does this by testing whether tally scores within the Markov chain vary significantly outside the $\alpha = 95\%$ confidence interval of the chain. The statistic for the Heidelberger-Welch RHW test is

$$RHW = \frac{z_{(1-\alpha/2)}\sqrt{\hat{s}_n/n}}{\theta_n}, \qquad (2)$$

where $z_{(1-\alpha/2)}$ represents 100(1- $\alpha/2$)th percentile of a standard normal distribution, n is the length of the Markov chain, θ_n is the estimated mean of the members in the chain, and $\hat{s}_n$ is the estimated variance of the members in the chain [7].

Figure 5 plots the Heidelberger-Welch RHW scores against the undersampling biases for the eigenvalue, flux, and reaction rate tallies in the R2 Case. The data plotted in this figure (and the data later plotted for the Tally Entropy metric) have been filtered to include only tallies that received at least one score per generation (and biases with less than 50% relative uncertainty, as mentioned previously). The RHW metric scores can be used very generally to predict the magnitude of undersampling biases, but, like the scores-per-generation metric, the RHW metric produced wide bands of data points for each NPG realization, indicating that this metric has limited power to assess undersampling bias.

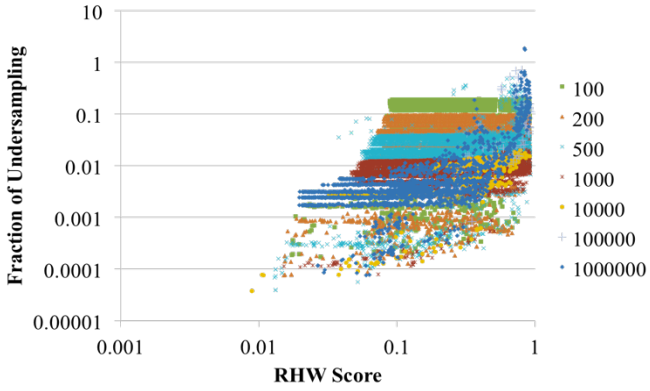

Fig. 5. Effectiveness of the Heidelberger-Welch RHW metric at predicting undersampling biases.

Interestingly, the absolute uncertainty in the RHW metric scores from the repeated simulations (plotted in Fig. 6) shows more promise. This alternative metric produced data points that correlated much more strongly with the *Fraction of Undersampling*, and although the bands of data still exist, they are clustered together more tightly than those produced by any of the other undersampling metrics.

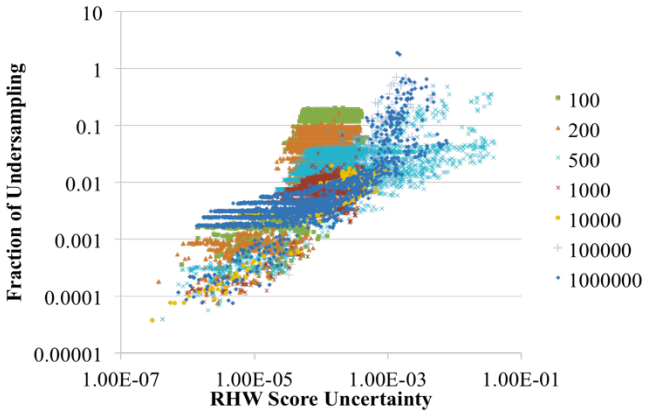

Fig. 6. Effectiveness of the absolute ncertainty in the Heidelberger-Welch RHW metric at predicting undersampling biases.

3. Tally Entropy Metric

The Tally Entropy metric [1] was developed using the information theory concept of Shannon Entropy. The Shannon Entropy, H, of a sampling process with N possible outcomes is

$$H = -\sum_{n}^{N} p_n \log_2(p_n), \tag{3}$$

where p_n is the probability of outcome n [8]. For example, a fair coin toss has $p_{\text{heads}} = p_{\text{tails}} = 0.5$, so tossing a coin once samples $H = 1$ shannon of information. Reactor physicists typically use the natural logarithm instead of $\log_2$.

Brown and Ueki used Shannon Entropy as a metric to detect unconverged fission sources in Monte Carlo simulations [9]. Ueki and Brown calculated the Shannon Entropy of the fission source by imposing a spatial mesh over the model. The probability of sampling a fission site in mesh element n is p_n, which can be estimated from the fraction of fission sites sampled in each mesh element. Shannon Entropy that has not yet converged to a steady value indicates that the fission source is still evolving and that additional inactive generations should be simulated.

Unfortunately, Shannon Entropy cannot be used in this way to assess the convergence of Monte Carlo tallies. This is because undersampled tallies may produce falsely converged Shannon Entropy estimates that are different but indistinguishable (a priori) from the entropy that would be produced by a converged set of tallies. Therefore, an alternative approach was developed for using the concept of Shannon Entropy to diagnose undersampling in Monte Carlo tally estimates.

For this metric, p_n is defined as the probability that history n is the one that contributes a particular increment to the tally. Therefore, p_n is estimated as the fractional contribution of history n to tally i within generation j. It is estimated by dividing the tally score from history n by the sum of the tally scores produced in generation j:

$$p_n = \frac{Tally\ Score\ of\ Particle\ n}{Sum\ of\ all\ Tally\ i\ Scores\ in\ Gen.\ j}. \tag{4}$$

After p_n is estimated for the histories within a generation, the tally- and generation-specific entropy is

$$H_{i,j} = -\sum_{history\ n}^{N_{i,j}} p_n \ln(p_n), \tag{5}$$

where $N_{i,j}$ is the number of histories in generation j that produced nonzero scores for tally i.

A random process with N outcomes can produce a minimum entropy of zero and a maximum entropy of $\ln(N)$. The signal will produce zero entropy if only one of the outcomes is possible, and it will produce maximum entropy when ($p_1 = p_2 = \cdots = p_n$). This maximum-entropy condition happens to be ideal for scoring unbiased Monte Carlo tally estimates: each tally should receive scores from many histories in each generation, and each history should contribute a similarly sized score to the tally estimate. Therefore, the Tally Entropy convergence metric predicts undersampling biases by comparing the Shannon Entropy of the tally to its theoretical maximum. The Tally Entropy test statistic for tally i is therefore calculated by the following equation:

$$Tally\ Entropy_i \equiv \frac{\langle \ln(N_{i,j}) \rangle - \langle H_{i,j} \rangle}{\langle \ln(N_{i,j}) \rangle}, \tag{6}$$

where the $\langle\ \rangle$ operator denotes the average of a value over all active generations.

Figure 7 plots the Tally Entropy scores against the undersampling biases for the eigenvalue, flux, and reaction rate tallies in the R2 case. The correlation between the Tally Entropy metric and the uncertainty biases looks quite similar to the correlation observed in Fig. 5 for the RHW metric, except that the Tally Entropy metric suffered to lesser degree to the previously observed banding of data points. With the exception of data points from the NPG=1,000,000 case, a boundary line can be drawn showing the maximum expected bias that will occur for a given tally entropy metric score.

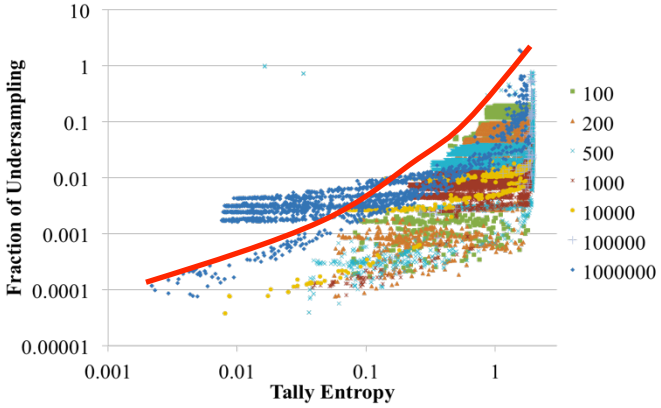

Fig. 7. Effectiveness of the Tally Entropy metric at predicting undersampling biases. The red line represents the maximum undersampling bias expected for a given Tally Entropy score.

Although the Tally Entropy metric seems to be more effective than the scores per generation and RHW metrics at predicting the occurrence of undersampling, the metric still leaves much to be desired. The red bounding line in Fig. 7 may be effective at identifying the minimum Tally Entropy score needed to achieve a given amount of undersampling bias, but a large Tally Entropy does not guarantee a large

bias. For example, observing a Tally Entropy of 0.5 indicates a *Fraction of Undersampling* between 0.02% and 10%. This wide range of bias may not be useful to analysts because it will cause them to falsely reject tallies that do not possess a large degree of bias. Thus, although this metric may be successful at predicting the maximum amount of bias present, its correlation with the magnitude of undersampling biases is not strong enough to reliably predict the size of an undersampling bias.

As shown in Fig. 8, the absolute uncertainty in the Tally Entropy scores was also examined to identify trends in the undersampling biases. The Tally Entropy uncertainty showed promise for predicting the magnitude of smaller undersampling biases, but it exhibited nonlinear behavior for larger biases (particularly for biases above 0.01 with a Tally Entropy uncertainty near 1.0E-04). As with the other metrics, the uncertainty in the Tally Entropy was not especially effective at predicting the magnitude of undersampling biases.

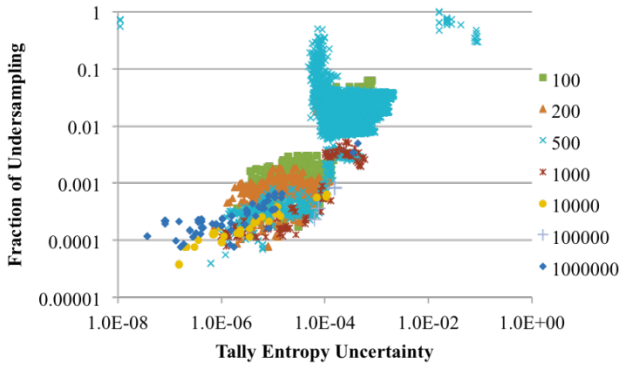

Fig. 8. Effectiveness of the Tally Entropy metric absolute uncertainty at predicting undersampling biases.

4. True Uncertainty in Tallies

The true uncertainty in the tallied parameters was also examined as a metric to predict the magnitude of undersampling bias [2]. The true uncertainty was calculated by taking the standard deviation of the tally results from the 30 independent repeated simulations for each NPG realization. As shown in Fig. 9, the true uncertainty of a tally seems reasonably correlated to the magnitude of the undersampling bias in that tally. These results seem to show the same banding effect that was observed for the other metrics, but to a much smaller degree. These data were filtered to include only the data points plotted that produced a relative uncertainty in their biases of less than 100%. (Previous plots used a 50% cutoff.) The sharp cutoff along the bottom edge of the data cluster is due to this filtering and should not be considered strength of correlation.

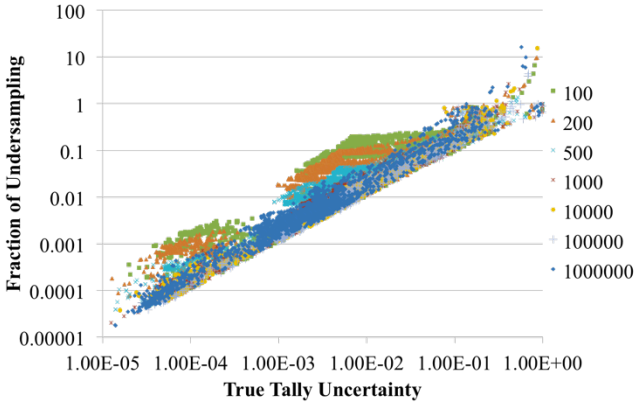

Fig. 9. Effectiveness of the tallies' true uncertainty at predicting undersampling biases.

While examining the true uncertainty in tallies may seem like a reasonable approach for quantifying the magnitude of the undersampling bias, this metric does not hold up to closer examination. Figure 10 plots the same data shown in Fig. 9, but only for the flux and eigenvalue tallies from the simulations. Although the flux tally true uncertainties show good correlation with the magnitude of the undersampling biases within a single simulation, the different NPG realizations produce lines of data, where tallies with a certain uncertainty produce undersampling biases that can differ by more than an order of magnitude. These results suggest that the true tally uncertainty may not effectively predict undersampling across different simulations, but it may be useful for predicting the undersampling bias in difficult-to-tally parameters using the known bias in an easier-to-tally parameter in that simulation.

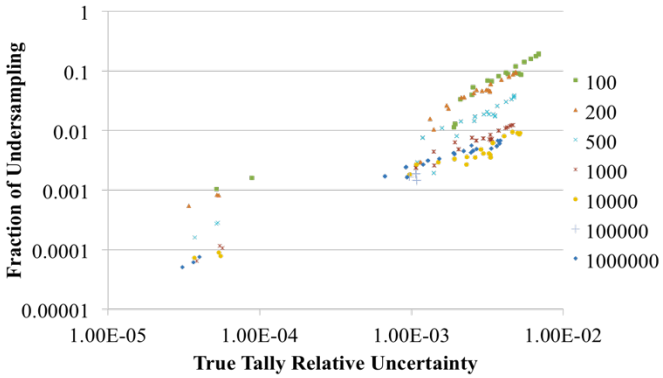

Fig. 10. Effectiveness of the eigenvalue and flux tallies' true uncertainty at predicting undersampling biases.

IV. CONCLUSIONS

This paper details a more thorough continuation of a previous study [2] for using statistical metrics to predict the presence and magnitude of undersampling biases using the CE TSUNAMI-3D within the SCALE code system [5]. All metrics that were examined showed limited correlation with

the magnitude of undersampling biases in eigenvalue estimates, flux tallies, and reaction rates, but no metric was able to consistently predict the magnitude of undersampling biases. Although this study used 20 times as many histories as the previous study [2], it was still difficult to obtain tally estimates that were sufficiently converged to yield low uncertainty bias estimates. Figure 11 plots the cumulative probability distribution of the relative uncertainty in undersampling biases for the roughly 11,500 tally estimates examined for each NPG realization.

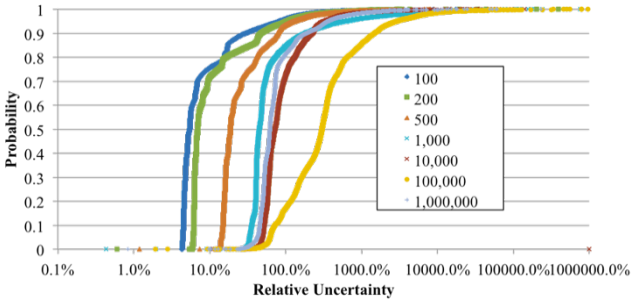

Fig. 11. Uncertainty in undersampling bias estimates.

As shown in Fig. 11, only the NPG=100 and NPG=200 cases produced bias estimates that generally contained less than 10% relative uncertainty. This occurred because the undersampling biases produced by these cases were larger in magnitude, making them easier to resolve. Many other NPG realization bias estimates reported an excess of 50% relative uncertainty, making it difficult to determine if the undersampling metrics performed poorly or if the data were too poorly resolved to evaluate the performance of the metrics. However, even the NPG=100 and NPG=200 cases did not produce undersampling metric data that correlated strongly with the magnitude of the undersampling biases, so the poor predictability observed for the undersampling metrics may not be due to insufficiently resolved bias estimates.

The NPG=1,000,000 data in Fig. 11 produced lower uncertainty estimates than several cases with lower NPG values because these cases were simulated to a finer degree of convergence than the other cases. The NPG=1,000,000 data were more finely resolved because they were originally intended to serve as the reference data. Each of the other NPG realizations used the same number of active histories in each simulation and the same number of repeated simulations.

Accurate, efficient reactor and safety analyses require effective measures to guarantee the reliability of simulation results, and the methods investigated in this study failed to reliably predict the magnitude of undersampling biases. Future work includes using undersampling metrics (without running computationally expensive simulations to resolve potentially small undersampling biases) to evaluate the effectiveness of variance reduction schemes, such as the FW-CADIS method [5], for the potential to mitigate undersampling biases. The results from this study suggest that future studies on undersampling could benefit from developing a more rigorous set of metrics for detecting and predicting undersampling biases.

REFERENCES

1. C. M. PERFETTI, B. T. REARDEN, W. J. MARSHALL, "Diagnosing Undersampling in Monte Carlo Eigenvalue and Flux Tally Estimates," *Nucl. Sci. Eng.*, **185**, 1, (2017).
2. C. M. PERFETTI, B. T. REARDEN, "A New CE TSUNAMI-3D Capability for Calculating Undersampling Metrics and Biases," *Trans. Am. Nucl. Soc.*, 114, 441–444 (2016).
3. F. B. BROWN, "'K-effective of the World' and Other Concerns for Monte Carlo Eigenvalue Calculations," *Progress in Nuclear Science and Technology*, **2**, pp. 738–742 (2011).
4. C. M. PERFETTI and B. T. REARDEN, "Quantifying the Effect of Undersampling in Monte Carlo Simulations Using SCALE," *Proc. PHYSOR 2014*, Kyoto, Japan, September 28–October 3, 2014, American Nuclear Society (2014).
5. B. T. REARDEN and M. A. JESSEE, Eds., *SCALE Code System*, ORNL/TM-2005/39, Version 6.2, Oak Ridge National Laboratory, Oak Ridge, Tennessee (2016). Available from Radiation Safety Information Computational Center as CCC-834.
6. P. HEIDELBERGER and P. D. WELCH, "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations," *Simulation Modeling and Statistical Computing*, **24**(4), pp. 233−245 (1981).
7. C. E. SHANNON, "A Mathematical Theory of Communication," *Bell System Technical Journal*, **27**(3), pp. 379−423 (1948).
8. T. UEKI and F.B. BROWN, "Stationarity Modeling and Informatics-Based Diagnostics in Monte Carlo Criticality Calculations," *Nucl. Sci. and Eng.*, **149**, 38 (2005).
9. J. C. WAGNER, D. E. PEPLOW, and S. W. MOSHER, "FW-CADIS Method for Global and Semi-Global Variance Reduction of Monte Carlo Radiation Transport Calculations" *Nucl. Sci. and Eng.*, 177(**1**), 37−57 (2014).