Optimal Run Strategies in Monte Carlo Iterated Fission Source Simulations

Paul K. Romano*[a], Amanda L. Lund[a], and Andrew R. Siegel[a]
[a]*Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue, Lemont, Illinois 60439*

Abstract — *The method of successive generations used in Monte Carlo simulations of nuclear reactor models is known to suffer from intergenerational correlation between the spatial locations of fission sites. One consequence of the spatial correlation is that the convergence rate of the variance of the mean for a tally becomes worse than $O(N^{-1})$. In this work, we consider how the true variance can be minimized given a total amount of work available as a function of the number of source particles per generation, the number of active/discarded generations, and the number of independent simulations. We demonstrate through both analysis and simulation that under certain conditions the solution time for highly correlated reactor problems may be significantly reduced either by running an ensemble of multiple independent simulations or simply by increasing the generation size to the extent that it is practical. However, if too many simulations or too large a generation size is used, the large fraction of source particles discarded can result in an increase in variance. We also show that there is a strong incentive to reduce the number of generations discarded through some source convergence acceleration technique. Furthermore, we discuss the efficient execution of large simulations on a parallel computer; we argue that several practical considerations favor using an ensemble of independent simulations over a single simulation with very large generation size.*

Keywords — *Monte Carlo, correlation, ensemble*

I. INTRODUCTION

Monte Carlo (MC) methods are widely used for solving neutral particle transport equations as applied to nuclear reactor problems. Because of the long run times typically required, however, in many cases the MC solution is used solely to validate faster, approximate deterministic methods that are used for design calculations. While MC methods have found use as the primary design tool for smaller problems (e.g., research and test reactors), the computational resources required for the analysis of large reactor problems remain out of reach for most analysts. Therefore, any successful effort to reduce the time to solution via MC will help expand the range of problems that can reasonably be handled given a computational resource.

In a MC simulation, each physical quantity that is tallied can be thought of as a random variable. The tally score that results from each source particle is then a realization of the random variable. Assuming that each of these realizations is independent and identically distributed, the variance of the sample mean is inversely proportional to the number of realizations. This is often thought of in the following simplistic manner: To reduce your standard deviation by a factor of 2, you need to simulate four times as many source particles. However, in an iterated fission source simulation using the method of successive generations, the source sites in a given generation may exhibit a strong spatial correlation with the source sites from the previous generation. Consequencely, the individual realizations of tally random variables are not independent of one another; the degree to which realizations are positively correlated with one another depends on both the mean free path of particles in the system and the spatial extent of the tallies themselves.

The spatial correlation of the fission source distribution from generation to generation and its impact on tallies were previously studied by Herman et al.,[1] specifically in the context of large light-water reactors. Miao et al.[2] extended

*Email: promano@anl.gov

that work by looking at the impact of tally size, generation size, and appropriate choice of confidence intervals. Miao et al. also posed two optimization problems[a]:

1. Given a fixed execution time, find the number of generations N and source particles per generation H that minimizes the expected error.

2. Given a fixed level of error, find the combination of N and H that minimizes the total execution time.

These optimizations problems can be viewed as determining parameters that maximize the efficiency of the Monte Carlo process, where the efficiency is inversely proportional to the product of the sample variance and the cost, in this case the execution time, as proposed by Hammersley and Handscomb[3] and later formalized as the "canonical" case in the decision-theoretic framework of Glynn and Whitt.[4] In the nuclear engineering community, this efficiency metric is often called *the* figure of merit.

The conclusion in both cases of the analysis by Miao et al. was that a single generation with H as large as possible would minimize either the error or the total execution time. However, the analysis assumed that the true fission source distribution was known a priori and could be sampled from. For practical problems, the source distribution is not known; instead, an arbitrary guess of the source distribution is made, and fission source iterations are carried out until the source converges. These *inactive* iterations may constitute a non-negligible amount of work, especially if H is large (typically H is kept constant from one source iteration to the next). Li[5] showed how accounting for the inactive generations can change the optimal run strategy for a given problem; however, her work did not develop a general framework for solving an optimization problem as did Miao et al.[2]

In the present work, we take a fresh look at the optimization problems posed by Miao et al. First, we explicitly account for the time spent during inactive generations. We then consider whether any benefit is to be gained from running an ensemble of simulations as suggested by several authors in the literature;[6,7] for example, instead of running a single simulation with 10^9 source particles, one could run 100 simulations each with 10^7 source particles and average the results. A similar approach recently was used to reduce the solution time for high-fidelity CFD simulations of turbulent flow.[8]

II. ANALYSIS

We begin by considering the error minimization problem: Given a fixed execution time, find the set of run parameters that minimizes the expected stochastic error in a tally result. Let B be the total number of source particles simulated. Since the total time is directly proportional to the number of source particles, we can use B as a proxy for the total execution time. Now, given this budget of B source particles, a code user typically has to decide how many source particles per generation, H, will be in a single fission source iteration (generation); how many generations are discarded, N_d, before collecting realizations of our tally random variables; and how many active generations, N, are to be simulated. Before we introduce the notion of an ensemble of independent simulations, let us develop an expression for the variance of the mean in a single simulation.

II.A. Variance

Consider a random variable Y being estimated using a tally in a MC simulation for which each realization is the score resulting from a single source particle. Let σ^2 be the variance of the random variable,

$$\begin{aligned} \sigma^2 &= \mathrm{Var}(Y) \\ &= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] \\ &= \mathrm{E}\left[(Y - \mu)^2\right], \end{aligned} \tag{1}$$

where $\mu = \mathrm{E}[Y]$. Since we are interested in the correlation between entire generations and not individual source particles, we use the average of the realizations in a single generation. Let $\{Y_j \mid 1 \leq j \leq B\}$ be the set of all realizations

[a]Note that the statement of these problems has been reframed in the context of the foregoing analysis.

and $\{Y_j \mid (i-1)H < j \leq iH\}$ be the realizations from source particles in generation i. Then the average in generation i is

$$X_i = \sum_{j=(i-1)H+1}^{iH} \frac{Y_j}{H}. \tag{2}$$

Assuming the Y_j are independent and identically distributed[b] and H is sufficiently large, it holds that X converges in distribution to a normal distribution with mean μ and variance σ^2/H by the central limit theorem.

The variance of the sample mean $\bar{X}$ over N generations is

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(X_i, X_j) \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathrm{Var}(X_i) + 2\sum_{i<j}\mathrm{Cov}(X_i, X_j)\right).
\end{aligned} \tag{3}$$

Since it is a stationary process, the covariance between two generations depends only on the *lag*, that is, the number of generations separating them.[9] We can recast the sum of the covariance terms using the autocovariance of lag k as follows:

$$\mathrm{Var}(\bar{X}) = \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathrm{Var}(X_i) + 2\sum_{k=1}^{N-1}\sum_{i=1}^{N-k}\mathrm{Cov}(X_i, X_{i+k})\right). \tag{4}$$

Using the autocorrelation coefficient (ACC), defined as

$$\rho_k = \frac{\mathrm{Cov}(X_i, X_{i+k})}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_{i+k})}}, \tag{5}$$

and noting that the variance of each realization is the same, we can express the covariance as

$$\mathrm{Cov}(X_i, X_{i+k}) = \mathrm{Var}(X)\rho_k = \frac{\sigma^2\rho_k}{H}. \tag{6}$$

Substituting Eq. (6) in Eq. (4), we find that

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \frac{1}{N^2}\left(\frac{N\sigma^2}{H} + 2\sum_{k=1}^{N-1}\frac{(N-k)\sigma^2\rho_k}{H}\right) \\
&= \frac{\sigma^2}{HN}\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho_k\right) \\
&= \frac{\sigma^2}{HN}r(N),
\end{aligned} \tag{7}$$

where we have defined $r(N)$ as

$$r(N) = 1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho_k. \tag{8}$$

We see from Eq. (7) that $r(N) > 1$ if $\rho_k > 0$; thus, in the presence of correlation, the convergence of the variance is worse than the ideal convergence rate of $1/N$. Eq. (7) also shows that the variance is inversely proportional to the

[b] As pointed out by Miao et al.,[2] Y_j estimates may not be completely independent because multiple source neutrons in a generation may have been produced by the same fission event. However, the variance of the mean of Y_j still scales as the inverse of H.

generation size—in line with the original conclusion of Miao et al. that the best way to reduce variance is to use as large a generation size as possible.

Before we proceed further, it is important to note that we have implicitly assumed that ρ_k, and hence $r(N)$, does not depend on H itself. This assumption was studied by Miao et al.[2] where it was shown that the ACCs exhibit no dependence on the generation size (see Fig. 2 in that work). This assumption enables us to evaluate $r(N)$ for a particular choice of H and use it for any choice of H in Eq. (7).

Let us now turn our attention to the issue of running multiple independent simulations. In the absence of correlation between generations, there would naturally be no reason to average the results from multiple independent simulations (an ensemble). One would want to reach a converged source distribution and then simulate as many generations as necessary to reach an acceptable level of error. In the presence of positive correlation, however, each additional active generation does not give us completely new information. On the contrary, if we were to run a new simulation from scratch (starting with an initial source guess and converging to the true source distribution), the first active generation of that simulation would give us a realization of X that is completely independent of all others. This begs the question then of whether it is more beneficial to simulate additional generations in an existing simulation (where we already have a converged source) or to start over from scratch, pay the price of having to converge a new distribution of source sites, and then be rewarded with realizations that are independent of the original simulation.

To generalize the question just posed, we consider the case where we have S independent simulations, each consisting of N_d discard generations, N active generations, and H source particles per generation. Let $\bar{\bar{X}}$ be the average of the results over the S simulations,

$$\bar{\bar{X}} = \frac{1}{S}\sum_{i=1}^{S} \bar{X}_i, \tag{9}$$

where $\bar{X}_i$ is the ith realization of $\bar{X}$, namely, the result of the ith simulation. If a different random seed is used, the $\bar{X}_i$ are independent of one another, and thus

$$\begin{aligned} \mathrm{Var}(\bar{\bar{X}}) &= \mathrm{Var}\left(\frac{1}{S}\sum_{i=1}^{S} \bar{X}_i\right) = \frac{1}{S^2}\sum_{i=1}^{S} \mathrm{Var}(\bar{X}_i) \\ &= \frac{\sigma^2}{SHN} r(N). \end{aligned} \tag{10}$$

Eq. (10) matches our expectations in that using either a larger generation size or multiple independent simulations gives us ideal convergence.

We now have enough information to return to our original question of how to arrange the B source particles in order to minimize our variance. In general, S and N are inversely related; if we have more independent simulations, then each simulation must have fewer active generations to maintain the same number of total source particles. The same is true of H and N—if we have a larger generation size, there must be fewer generations if B is constant. For plausibility, we must have at least one active generation in each simulation. We assume that the number of inactive generations remains the same in each simulation. Since

$$B = SH(N_d + N), \tag{11}$$

this implies that

$$N = \frac{B}{SH} - N_d. \tag{12}$$

The fact that we need at least one active generation implies that

$$\frac{B}{SH} - N_d \geq 1 \tag{13}$$

$$S \leq \frac{B}{H(N_d + 1)}. \tag{14}$$

Eq. (14) gives an upper bound on the number of independent simulations we can use while still having at least one active generation in each. Substituting Eq. (12) into Eq. (10), we obtain

$$\mathrm{Var}(\bar{\bar{X}}) = \frac{\sigma^2}{B - SHN_d} r \left(\frac{B}{SH} - N_d \right). \tag{15}$$

We now have an expression for the variance as a function of S, H, N_d, B, and ρ_k subject to the constraint of Eq. (14). As we will demonstrate in Section III.B, r is a monotonically increasing function of N. This allows us to qualitatively reason how changes in S or H affect the variance. On the one hand, increases in S or H lower the denominator of the fraction in Eq. (15) because more source particles are discarded overall, causing an increase in the variance. On the other hand, increasing S or H causes the argument to the r function to decrease, thereby reducing the overall correlation and the variance itself. Before we consider how to minimize $\mathrm{Var}(\bar{\bar{X}})$ as a function of S and H, let us scrutinize some of the implicit assumptions that have been made.

II.B. Relative Cost of Inactive and Active Generations

During the active generations, a Monte Carlo code has to perform extra work in order to tally the physical quantities that represent our random variables. In a reactor simulation, often millions of physical quantities need to be tallied corresponding to the flux and reaction rates over many non-overlapping physical regions. Thus, in general, the cost of simulating a source particle during active generations is greater than the cost of simulating a source particle during inactive generations. The relative cost depends on many factors, including the particular code being used and the efficiency of the tally implementation, but it is not uncommon to see a factor of 2 increase in the cost during active generations. For our analysis, we will simply introduce another variable f that characterizes the cost of simulating a source particle during active generations relative to inactive generations. If t is the average time to simulate a particle during inactive generations, then the total time for S independent runs is

$$T = tSH(N_d + fN). \tag{16}$$

Solving for N, we find

$$N = \frac{1}{f} \left(\frac{T}{tSH} - N_d \right). \tag{17}$$

Substituting Eq. (17) into Eq. (10), we have

$$\mathrm{Var}(\bar{\bar{X}}) = \frac{\sigma^2}{\frac{T}{ft} - \frac{SHN_d}{f}} r \left(\frac{T}{ftSH} - \frac{N_d}{f} \right). \tag{18}$$

If we define $B' = \frac{T}{ft}$ and $N'_d = \frac{N_d}{f}$, Eq. (18) becomes

$$\mathrm{Var}(\bar{\bar{X}}) = \frac{\sigma^2}{B' - SHN'_d} r \left(\frac{B'}{SH} - N'_d \right). \tag{19}$$

Eq. (19) is identical to Eq. (15) except that B and N_d have been replaced by B' and N'_d, which are a factor of f smaller. Thus, having $f > 1$ can be seen as decreasing the total budget of source particles that we have available, as well as the effective number of discarded generations.

II.C. Efficient Execution on Parallel Computers

The types of large, tightly coupled reactor problems that experience the strongest intergenerational correlation are likely to be simulated in parallel on many processors[c] because of the computational resources required. Thus, we must consider the interplay between parallel execution and the use of multiple independent simulations. When running a simulation in parallel, two strategies exist for scaling a problem to make use of more processors:

[c] In the context of this discussion, it is to be understood that multiple processors can refer to having multiple cores on a single CPU, multiple CPUs in a node, multiple nodes in a cluster, or some combination thereof.

- *Strong scaling*, wherein the problem size is fixed and more processors are used to reduce the total solution time.
- *Weak scaling*, wherein the problem size per processor is fixed and more processors are used to solve a larger problem.

In the context of particle transport, the problem size is the total number of source particles that need to be simulated rather than the spatial extent of the problem geometry. For our error minimization problem, we are interested in a strong-scaling regime because the premise is that we have a fixed budget of time (or equivalently source particles) that we want to optimally subdivide in order to minimize the expected error.

Several issues can arise in a strong-scaling problem. As more processors are added, in an ideal case the solution time would be reduced proportional to the number of processors being used. However, parallel communication (transmitting data between processors) and load imbalances can lead to nonideal scaling. Often, the communication requirements increase with the number of processors, and thus the parallel efficiency may tend to decrease as more processors are added. In addition, for a given problem typically each processor needs a minimum amount of work in order to execute efficiently. If we try to divide the available work over too many processors, each processor may not have enough work. We refer to this situation as *starvation* and note that it implies a hard limit on the number of processors that can be used for a given problem.

The parallel efficiency of a given code or simulation may depend on many factors. For our model here, we make a simple assumption about the parallel efficiency to observe what effect it might have on our choice of run strategy. Let us assume that the average time to simulate a source particle increases with the number of processors in the following manner:

$$t' = t(1 + \epsilon(P)), \tag{20}$$

where $\epsilon(P)$ is the "overhead" that increases monotonically with the number of processors, P. Furthermore, we suppose that the overhead increases linearly as a function of P,

$$\epsilon(P) = \frac{\epsilon_0 P}{P_0} \tag{21}$$

where the overhead is ϵ_0 with P_0 processors. Now let us assume that if we are running S independent simulations on P_0 processors, each independent simulation will use P_0/S processors. This implies that the average time to simulate a source particle is

$$t' = t(1 + \epsilon_0 S^{-1}). \tag{22}$$

If we replace t in Eq. (18) with t' from Eq. (22) and again define $B' = \frac{T}{ft}$ and $N'_d = \frac{N_d}{f}$, the variance is

$$\mathrm{Var}(\bar{\bar{X}}) = \frac{\sigma^2}{\frac{B'}{1+\epsilon_0 S^{-1}} - SHN'_d} r\left(\frac{B'}{SH(1+\epsilon_0 S^{-1})} - N'_d\right). \tag{23}$$

II.D. Optimization: Minimizing Variance

We see that Eq. (23) has a slightly more complicated dependence on S and H than do Eqs. (15) and (19). In Eqs. (15) and (19), which are equivalent to assuming $\epsilon_0 = 0$, we see that S and H appear together as a product. This implies that increasing one of these variables has the same effect on variance as increasing the other, as we noted earlier. By considering their product as a single variable, the optimization problem we have posed can be solved explicitly. Because it's not possible to explicitly solve the optimization problem while accounting for nonideal parallel scaling, we demonstrate its effect later through a numerical experiment in Section III.C.

By treating the product of S and H as a single variable, our optimization problem can be restated as "What combination of S and H results in a minimum variance?" Let us define $\eta = SH$, allowing us to rewrite Eq. (10) as

$$\mathrm{Var}(\bar{\bar{X}}) = \frac{\sigma^2}{\eta N} r(N) \tag{24}$$

and Eq. (12) as

$$N = \frac{B}{\eta} - N_d. \tag{25}$$

Our goal is to find η that minimizes $\mathrm{Var}(\bar{\bar{X}})$. Since Eq. (25) gives a one-to-one relationship between N and η, it suffices to find N that minimizes $\mathrm{Var}(\bar{\bar{X}})$, in other words, find N_0 such that

$$\frac{d}{dN}\,\mathrm{Var}(\bar{\bar{X}})\bigg|_{N=N_0} = 0. \tag{26}$$

Taking the derivative of Eq. (24), we have

$$\begin{aligned}\frac{d}{dN}\,\mathrm{Var}(\bar{\bar{X}}) &= \frac{d}{dN}\left(\frac{\sigma^2(N+N_d)}{BN}r(N)\right)\\ &= \frac{\sigma^2}{B}\left[\left(1+\frac{N_d}{N}\right)\frac{dr}{dN} - \frac{rN_d}{N^2}\right].\end{aligned} \tag{27}$$

Multiplying by NB/σ^2, we arrive at the following condition on N that minimizes the variance:

$$g(N) \equiv (N+N_d)\frac{dr}{dN} - \frac{rN_d}{N} = 0. \tag{28}$$

In general, $r(N)$ is not differentiable because it assumes values only at integral values of N. That being said, it is possible to construct a continuous function that assumes the values of $r(N)$ for $N \in \mathbb{Z}^+$ by performing a least-squares fit of the autocorrelation coefficients as a function of k. Given that the ACCs appear to decay exponentially with increasing lag, we assume the following form:

$$\rho_k = \rho_0 \sum_{j=1}^{J} \lambda_j^k \tag{29}$$

where J is the total number of decay modes. It is shown later that for the problem analyzed in this paper, only five terms are needed. The values of λ_j^k are assumed to be between zero and one so that $\rho_{k+1} < \rho_k$ for all k. Substituting Eq. (29) in Eq. (8), we obtain

$$r(N) = 1 + 2\rho_0 \sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\sum_{j=1}^{J}\lambda_j^k. \tag{30}$$

Interchanging the summations over k and j leads to

$$r(N) = 1 + 2\rho_0 \sum_{j=1}^{J}\left(\sum_{k=1}^{N-1}\lambda_j^k - \frac{1}{N}\sum_{k=1}^{N-1}k\lambda_j^k\right). \tag{31}$$

The partial sums over k can be rewritten by using the following identities:

$$\sum_{k=1}^{N-1} x^k = \frac{x^N - x}{x-1} \tag{32}$$

$$\sum_{k=1}^{N-1} kx^k = \frac{(N-1)x^{N+1} - Nx^N + x}{(x-1)^2}. \tag{33}$$

Using Eqs. (32) and (33) in Eq. (31) and simplifying, we obtain

$$r(N) = 1 + 2\rho_0 \sum_{j=1}^{J}\frac{\lambda_j^{N+1} - N\lambda_j^2 + (N-1)\lambda_j}{N(\lambda_j-1)^2}. \tag{34}$$

Taking the derivative of Eq. (34) gives us

$$\frac{dr}{dN} = 2\rho_0 \sum_{j=1}^{J}\frac{\lambda_j(N\lambda_j^N \log\lambda_j - \lambda_j^N + 1)}{N^2(\lambda_j-1)^2}. \tag{35}$$

We now have a formalism to find the value of N, and hence the value of η, that minimizes $\mathrm{Var}(\bar{\bar{X}})$. We reiterate here that the solution will give us a unique value of N and η but not of S or H. Any combination of S and H that satisfies $\eta = SH$ will result in the minimum variance attainable.

II.E. Optimization: Minimizing Solution Time

In the preceding section, we considered how to find the value of N and η that minimizes the variance given a fixed B. We can also consider the reverse problem: Given a fixed variance, what are the values of N and η that minimize the time to solution, for which B is a proxy? Let us call V the desired variance. Then

$$V = \frac{\sigma^2}{\eta N} r(N). \tag{36}$$

Solving Eq. (36) for η yields

$$\eta = \frac{\sigma^2}{VN} r(N). \tag{37}$$

Substituting Eq. (37) into Eq. (11) then gives us B solely as a function of N:

$$B = \frac{\sigma^2 (N + N_d)}{VN} r(N). \tag{38}$$

To find the value of N that minimizes B, we take the derivative and set it to zero:

$$\frac{dB}{dN} = \frac{\sigma^2}{V} \left[\left(1 + \frac{N_d}{N} \right) \frac{dr}{dN} - \frac{r N_d}{N^2} \right] = 0. \tag{39}$$

Multiplying Eq. (39) by NV/σ^2 gives us the same equation as Eq. (28). We can summarize this finding by stating that for a given N_d and $r(N)$, there is a unique solution N for both optimization problems. If we are minimizing the variance, the corresponding value of η can be found via Eq. (25). If we are minimizing the solution time through B, the corresponding value of η can be found via Eq. (37).

III. RESULTS

Eq. (23) shows the most general form of the variance as a function of various parameters. Many of these parameters are representative of the problem itself and can be considered constant:

- The variance per source particle σ^2 will depend on the definition of the tally and whether any variance reduction schemes are applied in the Monte Carlo code.
- The average time per source particle t and the tally overhead f are properties of the Monte Carlo code itself.
- In the variance minimization problem, the total simulation time T is considered constant.
- The parallel scaling constant ϵ_0 is a property of the code, choice of tallies, and the system it is executing on.
- The autocorrelation coefficients ρ_k indicate the degree of serial correlation overall in the system but also depend on the choice of tallies.

The number of discarded generations N_d can also be considered a property of the system being simulated since it depends on the eigenvalues of the fission operator. However, various methods for accelerating source convergence may significantly reduce N_d. In our analysis, we will compare several choices of N_d to study its impact. This leaves S and H as the user-defined parameters that we wish to optimize. We saw that under the assumption of perfect parallel scaling, Eq. (23) reduces to Eq. (15) wherein S and H become a single variable η. By assuming an exponential fit to the ACCs, we derived a function $g(N)$ whose root minimizes the variance. Selecting values of all the constants will then allow us to solve the variance minimization problem.

TABLE I
Parameters for one-group reflective box problem.

Parameter	Value
Σ_t	$0.300\,\mathrm{cm}^{-1}$
Σ_s	$0.270\,\mathrm{cm}^{-1}$
Σ_f	$0.012\,\mathrm{cm}^{-1}$
$\nu\Sigma_f$	$0.018\,\mathrm{cm}^{-1}$
Σ_a	$0.018\,\mathrm{cm}^{-1}$
Dimensions	400 cm×400 cm×400 cm 25 cm×25 cm×25 cm

III.A. Autocorrelation Coefficients

Evaluating $r(N)$ requires the autocorrelation coefficients, ρ_k, for tallies of interest. Following the work of Miao et al.,[2] rather than using a full-core problem to evaluate ACCs, we use a simple one-group reflective box problem with tallies defined to mimic the observed correlation of assembly-size and pin-size tallies in the BEAVRS benchmark.[10] Parameters for the one-group problem are listed in Table I.

The reflective box problem was modeled in OpenMC.[11] While OpenMC is most often used as a continuous-energy Monte Carlo code, it also has the capability to use multigroup cross sections. Simulations with H=10^6 source particles per generation and N=1000 generations were run for two different configurations. In the first configuration, the dimensions of the box were 400 cm×400 cm×400 cm, and in the second configuration the dimensions were 25 cm×25 cm×25 cm. In both configurations, a fission rate tally over a 16×16×16 mesh covering the entire box was defined. Thus, for the second configuration, the edge length of each mesh element is 16 times smaller than for the first configuration. Miao et al.[2] previously demonstrated that the first configuration has ACCs comparable to tallies over an assembly in the BEAVRS core. We also show below that the second configuration has ACCs comparable to tallies over a pin in the BEAVRS core. We will henceforth refer to the mesh tallies in the first and second configurations as the *assembly mesh tally* and the *pin mesh tally*, respectively.

An estimator for the ACCs can be determined by expanding Eq. (5) in terms of expected values

$$\rho_k = \frac{\mathrm{E}[X_i X_{i+k}] - \mathrm{E}[X_i]\,\mathrm{E}[X_{i+k}]}{\sqrt{\left(\mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2\right)\left(\mathrm{E}[X_{i+k}^2] - \mathrm{E}[X_{i+k}]^2\right)}}. \tag{40}$$

Replacing the expected value of each random variable by a sample mean gives us

$$\hat{\rho}_k = \frac{\nu \sum_{i=1}^{\nu} X_i X_{i+k} - \left(\sum_{i=1}^{\nu} X_i\right)\left(\sum_{i=1}^{\nu} X_{i+k}\right)}{\sqrt{\left(\nu \sum_{i=1}^{\nu} X_i^2 - \left(\sum_{i=1}^{\nu} X_i\right)^2\right)\left(\nu \sum_{i=1}^{\nu} X_{i+k} - \left(\sum_{i=1}^{\nu} X_{i+k}\right)^2\right)}}, \tag{41}$$

where $\nu = N - k$. One can show[12] that Eq. (41) is a biased estimator with a negative bias of order $O(\nu^{-1})$. Since in this case N=1000, the bias is expected to be negligible. Eq. (41) was used to evaluate sample autocorrelation coefficients for lags up to 200 for each mesh cell in the assembly and pin mesh tallies. Rather than analyzing each mesh cell separately, a single spatially averaged correlation coefficient was calculated as

$$\bar{\rho}_k = \frac{\sum_{m=1}^{M} s_m^2 \hat{\rho}_{m,k}}{\sum_{m=1}^{M} s_m^2}, \tag{42}$$

TABLE II
Parameters from least-squares fit in Eq. (29) to spatially averaged sample autocorrelation coefficients for the assembly mesh tally.

Parameter	Value
ρ_0	0.1933
λ_1	0.9949
λ_2	0.9645
λ_3	0.8923
λ_4	0.7455
λ_5	0.5946

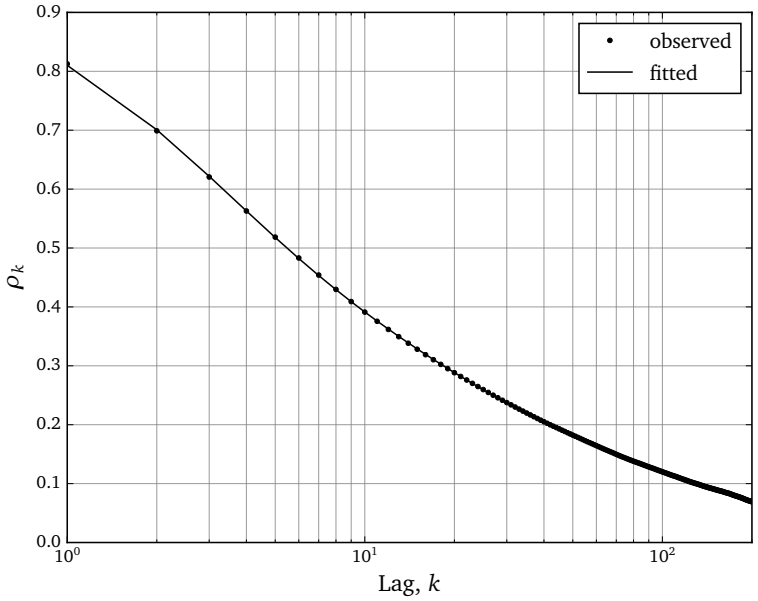

Fig. 1. Spatially averaged autocorrelation coefficients for the assembly mesh tally.

where s_m^2 is the sample variance in the mth mesh cell and $\hat{\rho}_{m,k}$ is the sample autocorrelation coefficient of lag k in the mth mesh cell. The variance weighting is discussed at length by Herman et al.[1]

The simulation results allow us to calculate estimates of ρ_k up to $k = 200$. For $k > 200$, the estimates become less reliable because of the noise of the simulation and a lower value of ν. Instead, we performed a least-squares fitting of the spatially averaged sample ACCs for the assembly mesh tally using Eq. (29) with $J = 5$. This allows us to both evaluate ACCs at lags greater than 200 and take derivatives of $r(N)$ as discussed in Section II.D. Table II shows the resulting least-squares parameters for the assembly mesh. For the pin mesh tally, we assume that $\rho_k = 0$ for $k > 200$. The spatially averaged sample ACCs and the least-squares fit for the assembly mesh tally are shown in Fig. 1. The spatially averaged ACCs for the pin mesh tally up to lag 200 are shown in Fig. 2. Comparing the ACCs in Figs. 1 and 2 with those for the BEAVRS model obtained by Herman,[6] one can conclude that the correlation behavior for this simple one-group problem is equivalent to what would be observed in a complex reactor model.

III.B. Model Predictions

We can now evaluate our model in Eq. (23) to estimate the variance as a function of S and H with the constraint imposed by Eq. (17). Let us first assume that $f = 1$ and $\epsilon_0 = 0$ such that Eq. (23) reduces to Eq. (15). For the other parameters, we adopt values from an MC21 solution[13] of the OECD/NEA Monte Carlo performance benchmark.[14] Specifically, to satisfy the error requirements of this benchmark, we used $B = 200 \cdot 10^9$ source particles in total,

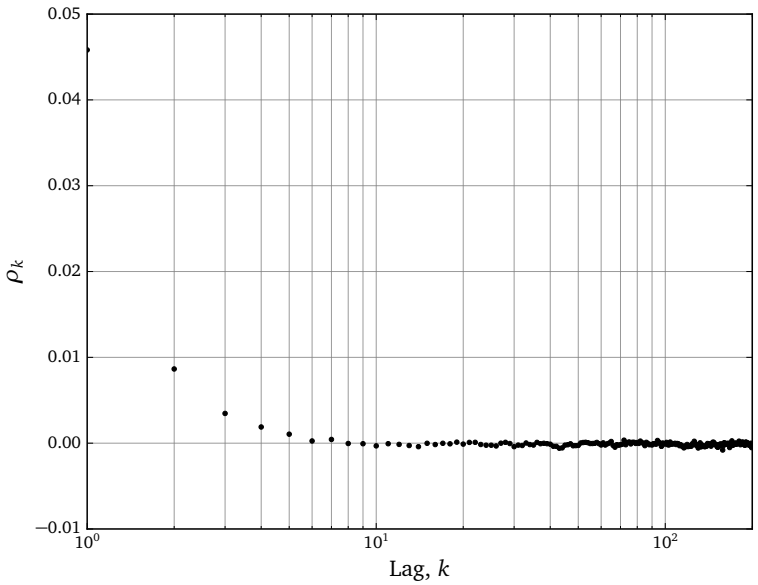

Fig. 2. Spatially averaged autocorrelation coefficients for the pin mesh tally.

divided into generations of size $H = 4 \cdot 10^6$. Holding these values constant and increasing S from 1 to the maximum permitted by Eq. (14), we obtain the solid curves in Fig. 3 for various choices of N_d. The largest value, $N_d = 600$, corresponds to what Kelly et al. used for the MC21 simulation.

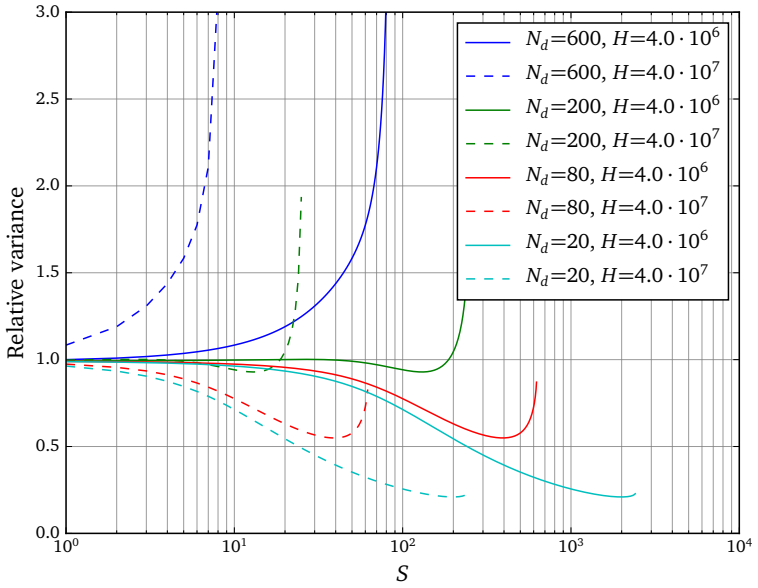

Fig. 3. Relative variance of assembly tally as a function of independent simulations, S, for selected values of source particles per generation, H, and inactive generations, N_d.

Fig. 3 demonstrates two competing effects. Firstly, as the number of independent simulations is increased, we benefit from the fact that there is less correlation overall, that is, the argument to the r function is smaller. Second, each simulation discards N_d generations so that there are fewer source particles being simulated during active generations; that is, the denominator in the first term of Eq. (15) is greater. When $N_d = 600$, the reduction in $r(N)$ is outweighed by the increase in the number of discarded source particles as S is increased. For lower values of N_d, however, there is a clear benefit from the reduction in $r(N)$ up to a certain point. This is clear when considering $r(N)$ itself as shown in Fig. 4. When $N_d = 600$ and $S = 1$, there are about 50,000 active generations; from Fig. 4, N clearly is so large that

we are in the asymptotic region. Thus, increasing the number of simulations (decreasing N) has little effect on the overall correlation. For S to be large enough to reach the non-asymptotic region on the $r(N)$ curve, it would no longer satisfy Eq. (14); all the source particles would be wasted on discard generations. This is evident at the end of the solid blue curve in Fig. 3 where the variance increases sharply. On the other hand, as N_d is decreased, one can make S sufficiently large that overall correlation is reduced and enough source particles remain.

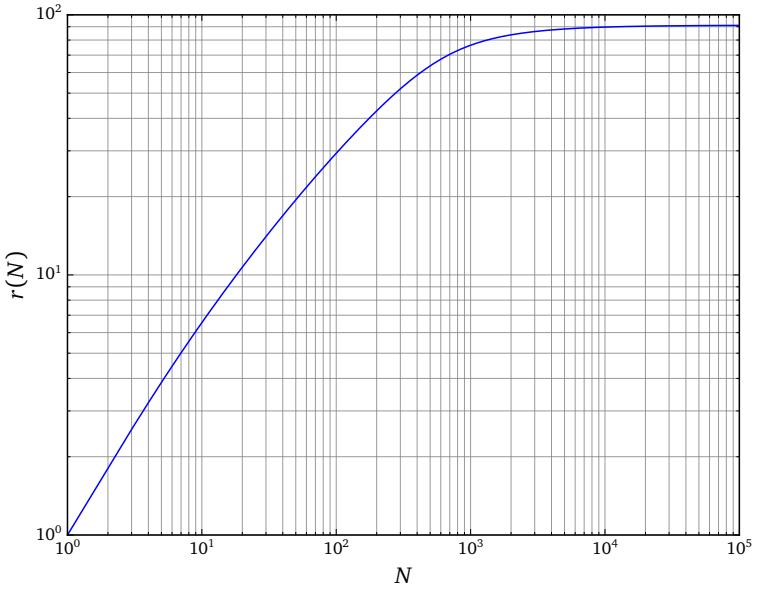

Fig. 4. Cumulative correlation, $r(N)$, as a function of the number of active generations, N.

Also displayed in Fig. 3 is the variance as a function of S for the same values of N_d and a value of H that is an order of magnitude greater. In each case, we see that increasing H simply moves each curve to the left. This is expected—as explained earlier, the product of S and H can be considered a single variable η. Thus, increasing H is equivalent to decreasing S. In fact, if rather than holding H constant and increasing S we hold S constant and increase H, as is done in Fig. 5, we see that the variance behaves exactly as in Fig. 3.

By looking at the variance as a function of S and H, we immediately infer values that minimize the variance. Alternatively, we reach the same result by solving Eq. (28) for N using the expressions for $r(N)$ and its derivative in Eqs. (34) and (35), respectively. Fig. 6 shows $g(N)$ for each value of N_d. As an example, for $N_d = 80$, $g(N) = 0$ at approximately $N = 49$. Subsequently solving Eq. (12) for S given $H = 4 \cdot 10^6$ yields $S = 392$. Examining Fig. 3, one can confirm that this corresponds exactly to the minimum of $\mathrm{Var}(\bar{\bar{X}})$.

All the predictions so far have been based on spatially averaged ACCs from the assembly mesh tally. If instead we use the ACCs from the pin mesh tally, the behavior is much different. Fig. 7 shows the relative variance for the spatially averaged pin tally as a function of the number of simulations. In this case, the correlation is not sufficiently large for there to be any benefit to running multiple independent simulations regardless of how many generations we assume must be discarded.

III.C. Model Predictions with Inefficiencies

In the preceding section, we looked at predictions from our model assuming the code being used had no overhead from tallies ($f = 1$) and perfect parallel scaling ($\epsilon_0 = 0$). As discussed earlier, having $f > 1$ simply amounts to different values of B and N_d, so nothing is to be learned from experimenting with different values of f that we could not already infer from the results given thus far. To better understand the impact of parallel scaling, we again look at the case of $B = 200 \cdot 10^9$ and $H = 4 \cdot 10^6$. This time, we use Eq. (23) to determine the variance with three different values of ϵ_0 corresponding to 100%, 75%, and 50% parallel efficiencies with $S = 1$. Since the parallel efficiency would simply be the inverse of $1 + \epsilon_0 S^{-1}$, an efficiency of 75% corresponds to $\epsilon_0 = 1/3$ and an efficiency of 50% corresponds to $\epsilon_0 = 1$. Fig. 8 shows the relative variance for the case analyzed before but with only two choices of N_d

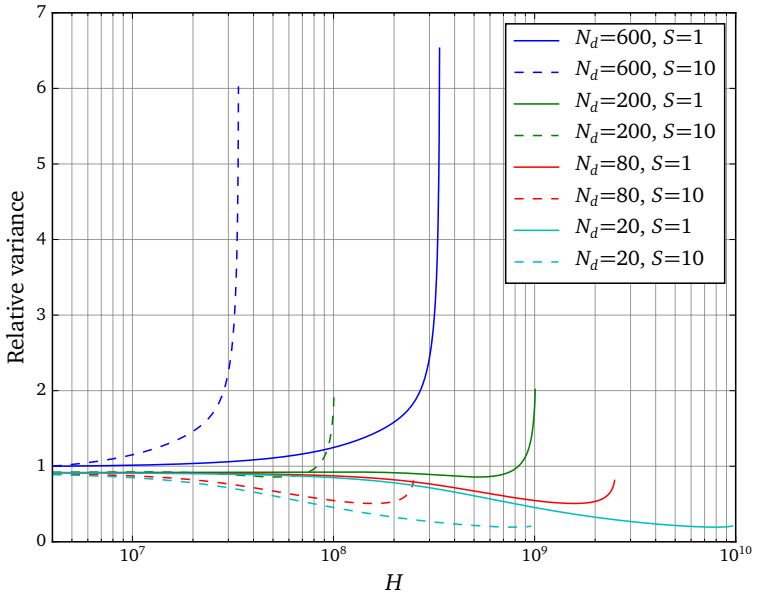

Fig. 5. Relative variance of assembly tally as a function of source particles per generation, H, for selected values of independent simulations, S, and inactive generations, N_d.

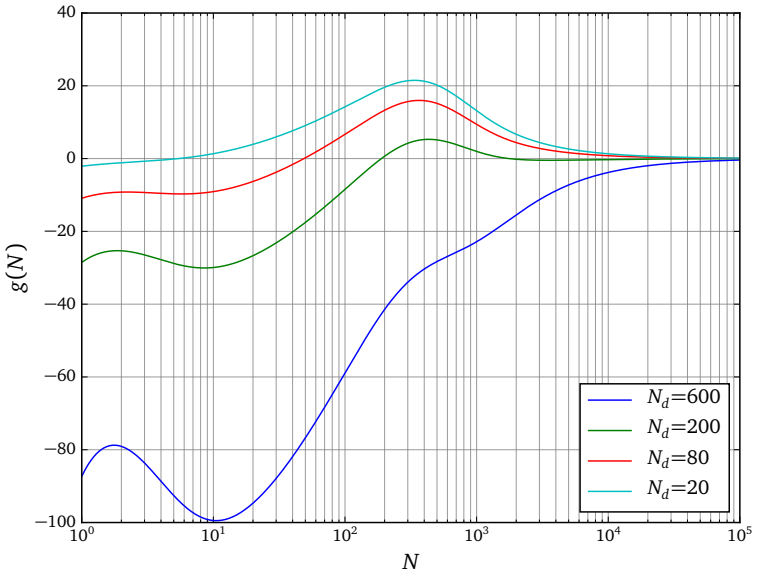

Fig. 6. Function whose root corresponds to the number of active generations, N, that minimizes the variance or solution time.

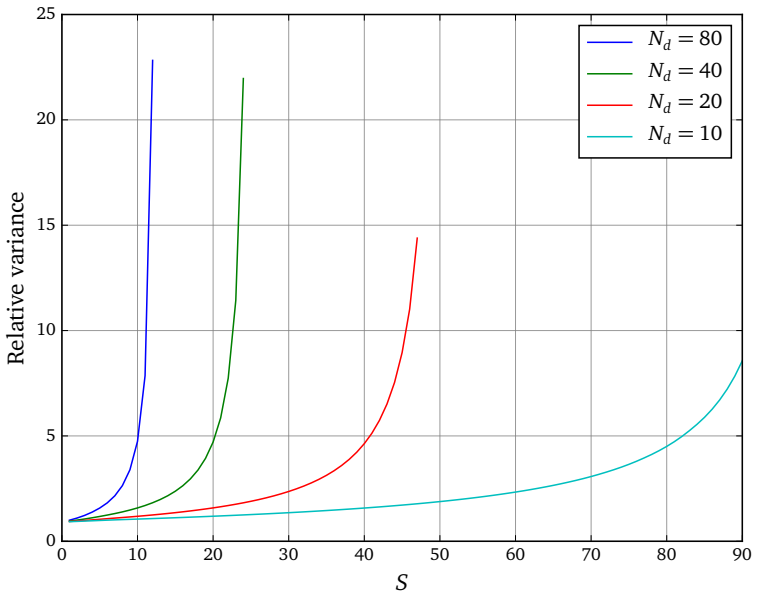

Fig. 7. Relative variance of the pin mesh tally as a function of the number of independent simulations.

given. As we saw before, when $\epsilon_0 = 0$, there is no benefit in running multiple independent simulations if $N_d = 600$. For $\epsilon_0 = 1/3$ and $\epsilon_0 = 1$, however, there is a separate benefit in running multiple simulations in that each simulation uses fewer processors and therefore attains better parallel efficiency.

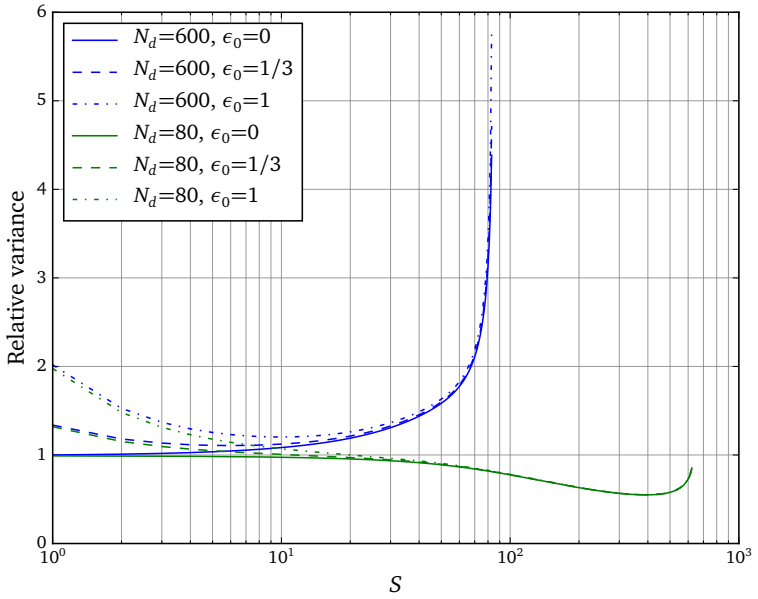

Fig. 8. Relative variance of assembly tally as a function of the number of independent simulations for $B = 200 \cdot 10^9$.

III.D. Measurements from OpenMC

The results that have been presented so far were based on analytical predictions in Eqs. (15) and (23). To confirm that the actual variance as a function of the model parameters behaves as we have predicted, we carried out a series of runs with OpenMC. Our goal is to estimate the variance for a given choice of the parameters N_d, S, H, and B. However, we cannot rely on the variance reported by OpenMC because when calculating the sample variance of the

mean, one must assume that each realization is independent. To circumvent the variance underprediction, we can take advantage of the fact that an analytical solution is known to the reflective box problem; namely, the flux is uniform across the box with $\phi = S/\Sigma_a$ where S is the source per unit volume. Since the neutron production cross section is constant, it follows that the source is also uniform across the box. OpenMC reports each tallied quantity per source particle, so the expected fission rate across the entire geometry as reported by OpenMC is simply Σ_f/Σ_a. Given a $\Lambda \times \Lambda \times \Lambda$ mesh over the box, the expected fission rate in each mesh cell is thus

$$\mu = \frac{\Sigma_f}{\Lambda^3 \Sigma_a}. \tag{43}$$

Now let us define the spatially averaged fission rate

$$\langle \bar{\bar{X}} \rangle = \frac{1}{M} \sum_{m=1}^{M} \bar{\bar{X}}_m, \tag{44}$$

where $\bar{\bar{X}}_m$ is the fission rate in the mth mesh cell. Let Z be the average squared difference between the spatially averaged fission rate and the true value,

$$Z = \left(\langle \bar{\bar{X}} \rangle - \mu \right)^2. \tag{45}$$

If we take the expectation of Z, we see that

$$\mathrm{E}[Z] = \mathrm{E}\left[\left(\langle \bar{\bar{X}} \rangle - \mu \right)^2 \right] \tag{46}$$

$$= \mathrm{Var}(\langle \bar{\bar{X}} \rangle) + \left(\mathrm{E}[\langle \bar{\bar{X}} \rangle] - \mu \right)^2. \tag{47}$$

If we assume that $\langle \bar{\bar{X}} \rangle$ is an unbiased estimator, then $\mathrm{E}[\langle \bar{\bar{X}} \rangle] = \mu$ and

$$\mathrm{E}[Z] = \mathrm{Var}(\langle \bar{\bar{X}} \rangle) = \frac{1}{M^2} \sum_{m=1}^{M} \mathrm{Var}(\bar{\bar{X}}_m). \tag{48}$$

Our procedure then is to run OpenMC to generate multiple estimates of $\bar{\bar{X}}_m$, which we denote $\hat{\bar{\bar{X}}}_{m,\ell}$, and then estimate the average squared difference

$$\hat{Z} = \frac{1}{L} \sum_{\ell=1}^{L} \left(\frac{1}{M} \sum_{m=1}^{M} \hat{\bar{\bar{X}}}_{m,\ell} - \mu \right)^2 \tag{49}$$

which gives us an estimate of the sum of the actual variances in each mesh cell per Eq. (48).

To validate the analytical results, we look at the behavior of the variance for simulations with $H = 10^5$, $B = 10^8$, and four different values of N_d: 80, 40, 20, and 10. For each value of N_d, S was varied from one up to the maximum allowed by Eq. (14). For each value of S, N was modified by using Eq. (12) to preserve the total number of generations. Each set of S simulations (with N active generations) was carried out $L = 20$ times[d]. The results of each of these L sets was averaged over all mesh cells and compared with μ via Eq. (49). The observed variance for the spatially averaged assembly mesh tally calculated via Eq. (49) is shown in Fig. 9. Along with the observed variances displayed with solid lines, predictions based on Eq. (15) are shown with dashed lines. The observed variances were normalized so that the first point on the $N_d = 80$ curve matches the prediction. We see that the actual errors observed in the OpenMC runs match the predicted behavior. While the curves show the same general shape as what is predicted, they do not match exactly. The reason for this could be the arbitrary normalization, the fact that the ACCs used to make the predictions could themselves be off slightly, or an artifact of the run strategy. In particular, each run was started with a uniform source distribution (the true source distribution), so the earlier fission generations might have been closer to the true source distribution (thereby reducing the error) than they would have been otherwise.

[d]This gives us a much better estimate of the variance, especially for $S = 1$ where we have only a single estimate of the fission rate for each mesh cell.

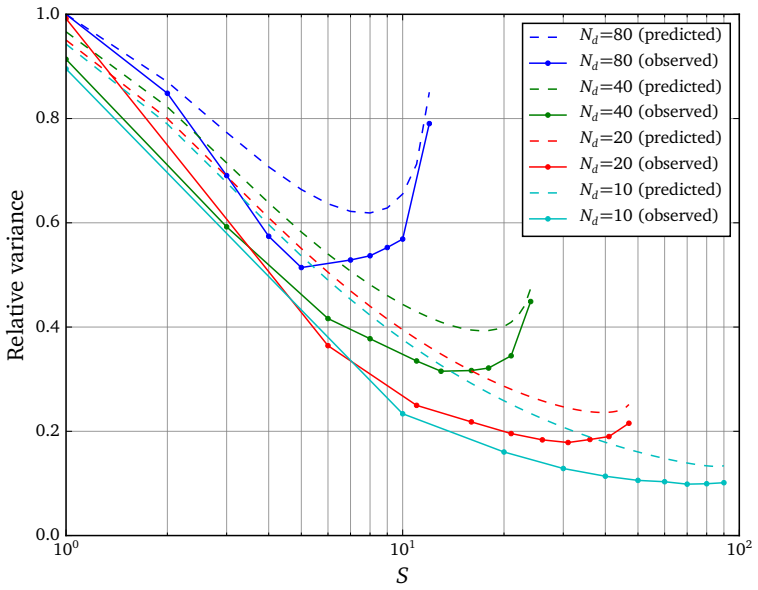

Fig. 9. Observed variance of spatially averaged assembly mesh tally as a function of the number of independent simulations compared to predictions.

The observed variance for the spatially averaged pin mesh tally and the model predictions from Eq. (15) are shown in Fig. 10. Once again, we see that the predicted behavior of the variance shows remarkable agreement with the actual error observed.

IV. CONCLUSIONS

Simulations of large light-water reactors using Monte Carlo methods have been known to suffer suboptimal convergence due to serial correlation that arises when using the method of successive generations. In this paper, we have demonstrated through both analysis and simulation that under certain conditions the solution time for highly correlated reactor problems may be significantly reduced by running either an ensemble of multiple independent simulations or simply using a larger generation size. The expression for the variance, Eq. (23), as a function of the number of independent simulations, S, and the number of particles per generation, H, shows two basic competing effects when considering how to choose S and H. On the one hand, increasing S or H can reduce the cumulative correlation effect as embodied in $r(N)$ and thereby reduce the variance. On the other hand, running more simulations results in more source particles being discarded as each simulation has to converge on the source distribution before tallies can begin; ergo, fewer source particles are available during active generations.

Evaluating Eq. (23) at various values of B, H, N_d, S, and ρ_k, we obtained the following insights:

- When using ACCs representative of an assembly-size tally, the correlation is strong enough that an order of magnitude reduction in the variance can be achieved by increasing S or H. For ACCs representative of a pin-size tally, however, no improvement is to be expected from using larger S or H.

- As the effective number of discarded generations, N_d is decreased, one is able to tolerate a larger S or H because fewer source particles need be discarded per simulation. This implies that the effective use of source convergence acceleration techniques would increase the viability of using an ensemble of independent simulations.

- If the number of active generations is very high, as was the case in Fig. 3, we reach a point on the $r(N)$ curve where it asymptotes to a constant value. This implies an asymptotic $1/N$ convergence in variance as discussed by Miao et al.[2] In the context of the present analysis, it also implies that very large values of S would be required to reduce the overall correlation.

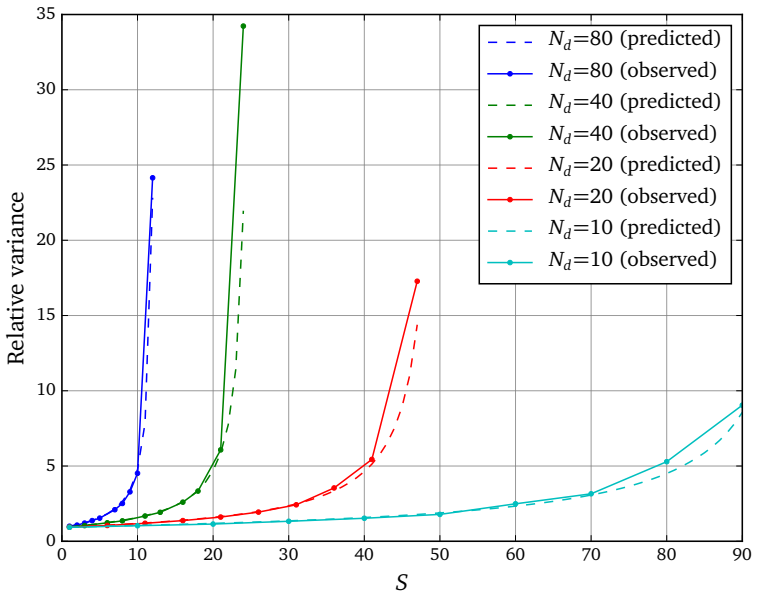

Fig. 10. Observed variance of representative pin tally as a function of the number of independent simulations

While we have seen that increasing S or H generally has the same effect, practical considerations may lead to preferring one or the other. As we have pointed out in Section II.C, if one assumes that the code being used does not scale perfectly ($\epsilon_0 > 0$), there is a secondary benefit to using more simulations in the sense that it can help improve the parallel efficiency of each simulation. Using independent simulations, as opposed to a larger generation size, can also help optimize throughput on machines managed by a job queue. It's typically easier to secure resources for a smaller job; thus, rather than having one simulation with very large H that requires many compute nodes (and may have to queue for a long time), we can divide the simulation into an ensemble of smaller jobs, each of which can execute independently, to achieve the same level of error. This strategy also provides some fault tolerance since one could still accumulate results even if one of the many independent simulations were to fail. Moreover, using multiple independent simulations can provide an unbiased estimate of the true variance.[7] Such estimates can also be obtained in a single simulation with large H, for example, by redefining a tally realization to include multiple fission generations,[7,13,15,16] with slightly greater difficulty.

While our analysis demonstrates how one can determine run parameters that minimize the variance for a given budget of source particles, its practical utility is diminished by the fact that it requires knowledge of the ACCs, which are generally not known a priori. Notwithstanding, the results indicate that for problems experiencing strong intergenerational correlation, one is likely better off using a large generation size (or an ensemble of independent simulations) to the extent that it is practical. Our analysis and results generalize those of Miao et al.,[2] in which it was suggested that very large generations would minimize variance, by explicitly accounting for the time spent in discarded generations.

In our analysis, we have assumed that all tally estimators are unbiased. Researchers have demonstrated, however, that the use of a finite generation size and fission source renormalization results in a bias[17,18] because of undersampling that decreases with increasing H. While most works in the literature have focused on the bias in estimates of k-effective, several recent papers that studied local tallies have demonstrated that biases of up to 20% are possible.[19,20] Thus, for a given tally in a problem, there may be a minimum H needed to ensure that the estimate is not biased. Further investigation in this area is needed to better understand how the bias depends on properties of the tallies, for example, the spatial extent and whether it is divided up into energy bins.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Jilang Miao and Benoit Forget for their constructive comments and suggestions. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S.

Department of Energy Office of Science and the National Nuclear Security Administration.

REFERENCES

1. B. R. HERMAN et al., "Analysis of tally correlation in large light water reactors," *Proc. PHYSOR – The Role of Reactor Physics toward a Sustainable Future*, Kyoto, Japan, Sep. 28–Oct. 3, 2014.

2. J. MIAO, B. FORGET, and K. SMITH, "Analysis of Correlations and their Impact on Convergence Rates in Monte Carlo Eigenvalue Simulations," *Ann. Nucl. Energy*, **92**, 81 (2016).

3. J. M. HAMMERSLEY and D. C. HANDSCOMB, *Monte Carlo Methods*, Springer Netherlands (1964).

4. P. W. GLYNN and W. WHITT, "The Asymptotic Efficiency of Simulation Estimators," *Oper. Res.*, **40**, *3*, 505 (1992).

5. L. LI, *Acceleration Methods for Monte Carlo Particle Transport Simulations*, PhD thesis, Massachusetts Institute of Technology, 2017.

6. B. R. HERMAN, *Monte Carlo and thermal hydraulic coupling using low-order nonlinear diffusion acceleration*, PhD thesis, Massachusetts Institute of Technology, 2014.

7. H. J. SHIM, S. H. CHOI, and C. H. KIM, "Real variance estimation by grouping histories in Monte Carlo eigenvalue calculations," *Nucl. Sci. Eng.*, **176**, 58 (2014).

8. V. MAKARASHVILI, E. MERZARI, A. OBABKO, P. FISCHER, and A. SIEGEL, "Accelerating the high-fidelity simulation of turbulence: ensemble averaging," *Proc. ASME 2016 Fluids Engineering Division Summer Meeting*, Washington, DC, Jul. 10–14, 2016.

9. T. UEKI, F. B. BROWN, D. K. PARSONS, and D. E. KORNREICH, "Autocorrelation and Dominance Ratio in Monte Carlo Criticality Calculations," *Nucl. Sci. Eng.*, **145**, 279 (2003).

10. N. HORELIK, B. HERMAN, B. FORGET, and K. SMITH, "Benchmark for Evaluation and Validation of Reactor Simulations (BEAVRS)," *Proc. Int. Conf. Mathematics and Computational Methods Applied to Nuclear Science and Engineering*, Sun Valley, Idaho, May 5–9, 2013.

11. P. K. ROMANO, N. E. HORELIK, B. R. HERMAN, A. G. NELSON, and B. FORGET, "OpenMC: A state-of-the-art Monte Carlo code for research and development," *Ann. Nucl. Energy*, **82**, 90 (2015).

12. A. STUART, *Kendall's Advanced Theory of Statistics*, Wiley (2010).

13. D. J. KELLY, T. M. SUTTON, and S. C. WILSON, "MC21 Analysis of the Nuclear Energy Agency Monte Carlo performance benchmark problem," *Proc. PHYSOR – Advances in Reactor Physics – Linking Research, Industry, and Education*, Knoxville, Tennessee, Apr. 15–20, 2012.

14. J. E. HOOGENBOOM, W. R. MARTIN, and B. PETROVIC, "The Monte Carlo performance benchmark test—Aims, Specifications, and First Results," *Proc. Int. Conf. on Mathematics and Computational Methods Applied to Nuclear Science and Engineering*, Rio de Janeiro, Brazil, 2011.

15. E. M. GELBARD and R. PRAEL, "Computation of standard deviations in eigenvalue calculations," *Prog. Nucl. Energy*, **24**, 237 (1990).

16. P. K. ROMANO and B. FORGET, "Reducing Parallel Communication in Monte Carlo Simulations via Batch Statistics," *Trans. Am. Nucl. Soc.*, **107**, 519 (2012).

17. R. J. BRISSENDEN and A. R. GARLICK, "Biases in the estimation of K_{eff} and its error by Monte Carlo methods," *Ann. Nucl. Energy*, **13**, *2*, 63 (1986).

18. E. M. GELBARD and A. G. GU, "Biases in Monte Carlo Eigenvalue Calculations," *Nucl. Sci. Eng.*, **117**, 1 (1994).

19. F. B. BROWN, "A Review of Best Practices for Monte Carlo Criticality Calculations," *Proc. Nuclear Criticality Safety*, Richland, Washington, Sep. 13–17, 2009.

20. C. M. PERFETTI and B. T. REARDEN, "Quantifying the effect of undersampling in Monte Carlo simulations using SCALE," *Proc. PHYSOR – The Role of Reactor Physics toward a Sustainable Future*, Kyoto, Japan, Sep. 28–Oct. 3, 2014.