Reliable Confidence Intervals for Monte Carlo-Based Resource Adequacy Studies

Sharifnia, Ensieh; Tindemans, Simon H.

DOI
10.1109/SEST61601.2024.10694674

Publication date
2024

Document Version
Accepted author manuscript

Published in
Proceedings of the 2024 International Conference on Smart Energy Systems and Technologies (SEST)

Reliable Confidence Intervals for Monte Carlo-Based Resource Adequacy Studies

Ensieh Sharifnia, Simon H. Tindemans
Dept. of Electrical Sustainable Energy
Delft University of Technology
Delft, The Netherlands
Email: e.sharifnia@tudelft.nl, s.h.tindemans@tudelft.nl

***Abstract*—Quantitative risk analysis is essential for power system planning and operation. Monte Carlo methods are frequently employed for this purpose, but their inherent sampling uncertainty means that accurate estimation of this uncertainty is essential. Basic Monte Carlo procedures are unbiased and, in the limit of large sample counts, have a well-characterised error distribution. However, for small time budgets and ill-behaved distributions (such as those for rare event risks), we may not always operate in this limit. Moreover, multilevel Monte Carlo was recently proposed as a computationally efficient alternative to regular Monte Carlo. In this approach, great asymptotic speedups are achieved by reducing the number of full model evaluations. This further challenges the assumption that normally distributed errors can be used. This paper investigates the sampling error distributions for a practical resource adequacy case study, in combination with the Multilevel Monte Carlo method. It further proposes a practical test for validating error estimates, based on a bootstrap approach.**

***Index Terms*—central limit theorem, Monte Carlo methods, Multilevel Monte Carlo, resource adequacy, statistical testing**

I. Introduction

The Monte Carlo (MC) method plays a pivotal role in risk analysis for resource adequacy assessment, offering robust support for large and intricate models while effectively capturing the probabilistic nature of power systems [1], [2]. Accurate modeling is crucial for reliable evaluations, necessitating consideration of various aspects of the power system [3]. Despite MC's versatility in handling complexity, it may face computational challenges in convergence, particularly in highly reliable systems such as the power system [1].

The computational burden can be reduced through variance reduction methods, which can be classified into two main categories: intelligent sampling (e.g., [4]) and output interpolation (e.g., [5]). The former focuses on selecting samples, while the latter incorporates simplified models to reduce the variance of estimation [6]. Multilevel Monte Carlo (MLMC) represents a generalized form of output interpolation and has recently been applied in the power system domain [7], [8].

Previous studies have explored the benefits of MLMC in terms of computational speed in the power system domain [7], [8]. However, these papers primarily focused on asymptotic speedup and did not address the technical difficulties of applying MLMC with (relatively) small sample counts. Most critically, the estimation of errors relies on the Gaussian approximation that follows from the *Central Limit Theorem* (CLT), which is inaccurate with small samples from highly irregular distributions. In power system reliability assessment studies, researchers often assume normality in reporting sampling errors [9], [10], which is not always accurate [11].

Practical recommendations for the minimum number of samples required for practical Monte Carlo studies abound, with numbers ranging from tens to hundreds, depending on the properties of the distribution being sampled [11]. In some cases, analytical bounds are available, but in others, repeated numerical simulation is required to determine the number of samples required for convergence to the asymptotic normal distribution [11]. Of course, in practical cases, repeated simulation is not a realistic solution, so rules of thumb or convergence tests are required.

In this paper, we highlight the importance of uncertainty quantification by examining the sampling errors obtained through the MLMC method, specifically analyzing their dependence on overall sample counts. The explorative analysis is done for an ensemble of equivalent simulation runs, but in practical studies, the validity of confidence intervals must be ascertained for a single run. To address this issue, we propose a bootstrap-based test for this purpose, and verify its validity empirically. Results are illustrated using a resource adequacy case study based on the Great Britain system with renewable energy sources and storage units. The proposed validity test is general and applicable to all MC methods.

The remainder of this paper consists of four sections. In Section II, MC methods for reliability analysis are summarized. Section III presents methods for estimating errors, both for ensembles of runs and for individual runs. Section IV demonstrates a practical case study for the validity of error estimates, based on the suggested approach. Conclusions are provided in the last section.

II. Monte Carlo Methods

A. Mathematical Problem Statement

Power system reliability metrics usually take the form of an expectation, i.e. $q = \mathrm{E}[X]$. For example, Expected Energy Not Served (EENS) and Loss of Load Expectation (LOLE/LOLH) are two well-known risk indices in resource adequacy assessment studies. EENS is the expected amount of load curtailment in a year and LOLE/LOLH is the expected number of hours in a year with load curtailment. They can be

defined based on annual traces of load curtailment $c_t(S)$ in hour t and random scenario S as:

$$\text{EENS} = \text{E}\left[\sum_{t=1}^{8760} c_t(S) \times 1h\right], \tag{1}$$

$$\text{LOLE} = \text{E}\left[\sum_{t=1}^{8760} \mathbb{1}_{c_t(S)>0} \times 1h\right], \tag{2}$$

where $E[\cdot]$ denotes the expectation over S and $\mathbb{1}_x$ equals 1 if x is true, and 0 otherwise.

In the more general formulation $q = \text{E}[X]$, the random variable $X = X(s)$ can be seen as the output of any performance function (e.g. c_t) that associates a numerical value to a random state or scenario s. The probabilistic performance of the system is therefore generated by the probabilities of theses states. The probabilistic model in resource adequacy is typically specified using a bottom up model that defines demand levels, component status, generator output levels, etc.

For a large and complex system, it is not possible to apply analytic methods or use state enumeration to compute the quantity of interest $q = \text{E}[X]$. In such cases, Monte Carlo simulation is commonly applied, in which states s are randomly generated using the probabilistic bottom-up model and their performance $X(s)$ is analysed to estimate the value of q using the sample average approximation.

B. Conventional Monte Carlo

A concise overview of conventional Monte Carlo (MC) is presented in this section which provides a point of reference for subsequent sections. MC approximates $q = \text{E}[X]$ by

$$\hat{Q}_{MC} \equiv \frac{1}{n}\sum_{i=1}^{n} X^{(i)}, \tag{3}$$

where variables $X^{(i)}$ are independent and identically distributed to X. $\hat{Q}_{MC}$ is unbiased estimator of q and according to the CLT, for sufficiently large n, ΔQ_{MC} is normally distributed, so that,

$$\Delta Q_{MC} = Q_{MC} - q \sim \mathcal{N}\left(0, \text{Var}(\hat{Q}_{MC})\right). \tag{4}$$

The variance of $\hat{Q}_{MC}$ follows the MC estimator (3):

$$\text{Var}(\hat{Q}_{MC}) = \frac{\text{Var}(X)}{n}. \tag{5}$$

C. Multilevel Monte Carlo

Multilevel Monte Carlo (MLMC) employs a hierarchical structure of models, each with different levels of accuracy and computational cost, to achieve more efficient and accurate estimations [5]. Consider a set of L models $M_1, M_2, ..., M_L$ of the *same* system, which increase in complexity and generate output variables $X_1, ..., X_L$, respectively. The expectation of the top level model $q = E[X_L]$ is the quantity of interest, but the assumption is that evaluating the model M_L is computationally demanding. MLMC uses approximate models to better estimate $E[X_L]$ for a given time budget, using lower level models $M_1, \ldots, M_{L-1}$, which generate output variables $X_1, ..., X_{L-1}$ that are increasingly accurate approximations of X_L, and require increasing computational resources. We have

$$\begin{aligned} q &= E[X_L] \\ &= E[X_1] + E[X_2 - X_1] + ... + E[X_L - X_{L-1}] \\ &= r_1 + ... + r_L, \end{aligned} \tag{6}$$

where r_l is the contribution for level l. r_1 can be considered a crude estimation of q and $r_2, ..., r_L$ are successive refinements. In MLMC, each level contribution r_l is estimated independently by means of MC simulation:

$$\hat{R}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} Y_l^{(i)}, \tag{7}$$

with

$$Y_l^{(i)} = X_{l\circ}^{(i)} - X_{l-1\bullet}^{(i)}, \tag{8}$$

$$X_{0\bullet} \equiv 0. \tag{9}$$

Here, we formally distinguish M_l's outputs in the successive levels by $X_{l\circ}$ and $X_{l\bullet}$. In special cases, these may be drawn from different distributions with the same mean [7], but in the following we shall assume that both models are identical and therefore identically distributed, i.e., $X_{l\circ} \overset{d}{=} X_{l\bullet} \overset{d}{=} X_l$. Combining (6) and (7), the MLMC estimator $\hat{Q}_{ML}$ is defined as

$$\hat{Q}_{ML} \equiv \sum_{l=1}^{L} \hat{R}_l = \sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l} Y_l^{(i)}. \tag{10}$$

In MLMC, samples $\hat{Y}_l^{(i)}$ are randomly and independently selected; only the samples within a level pair ($X_{l\circ}$, $X_{l-1\bullet}$) are jointly selected from a common distribution. Invoking the CLT for each level pair, we know that the MLMC estimator (10) is unbiased and asymptotically normally distributed with variance

$$\text{Var}(\hat{Q}_{ML}) = \sum_{l=1}^{L}\frac{\text{Var}(Y_l)}{n_l}, \quad \text{where} \tag{11}$$

$$\begin{aligned} \text{Var}(Y_l) =& \text{Var}(X_l) + \text{Var}(X_{l-1}) \\ &- 2\text{Cov}(X_{l\circ}, X_{l-1\bullet}). \end{aligned} \tag{12}$$

Like for the regular MC method, for sufficiently large sample counts we invoke the CLT for the sample error as

$$\Delta Q_{ML} = \hat{Q}_{ML} - q \sim \mathcal{N}\left(0, \text{Var}(\hat{Q}_{ML})\right). \tag{13}$$

Clearly, according to (12), the variance (and the error) decreases when the sample correlation in level pairs increases.

The minimum variance for a given computation budget T is achieved when the time spent on each level pair $t_l^{(T)}$ is distributed according to [7].

$$t_l^{(T)} = \alpha_l T, \tag{14a}$$

with

$$\alpha_l = \frac{\sqrt{\text{Var}(Y_l)\tau_l}}{\sum_{l'=1}^{L}\sqrt{\text{Var}(Y_{l'})\tau_{l'}}} \tag{14b}$$

and τ_l is the average time for generating a sample realization y_l (according to its distribution Y_l). Consequently, the optimal number of samples for each level is

$$n_l^{(T)} = \text{round}(t_l^{(T)}/\tau_l). \tag{14c}$$

D. MLMC Parameter Estimation

When implementing an MLMC risk estimation procedure, two of the parameters are generally not known: the variance $\text{Var}(Y_l)$ of each MLMC level and the sample generation time τ_l. Accurately estimating their values is essential for running the MLMC frame work: Eq. (14) to determine the optimal allocation of samples across levels and Eq. (13) to estimate sampling errors.

In practice, both quantities are estimated based on sample realizations. Sample generation times τ_l are estimated as the average time required to generate joint samples $(x_{l\circ}, x_{l-1\bullet})$[1]. Estimation of $\text{Var}(\hat{Q}_{ML})$ is done using (12), where $\text{Var}(X_l)$ is estimated using both sets of realizations $x_{l\circ}, x_{l\bullet}$ (the variances of which were assumed to be identical). However, only the jointly sampled pairs $(x_{l\circ}, x_{l-1\bullet})$ are used to estimate the covariance $\text{Cov}(X_{l\circ}, X_{l-1\bullet})$.

The determination of optimal sample counts relies on sample duration times and level set variances, but those are typically unknown at the start of the simulations. Therefore, a generic MLMC method requires a warm-up phase to generate an initial batch of samples. This should be large enough to get a realistic variance estimation. After the warm up phase, the next step is to allocate sample counts among levels using (14). Estimating parameters based on realizations always has some error, which may end up in suboptimal allocations.

This error gets smaller with more realizations but a long warm up phase is not a solution as it reduces MLMC efficiency when consumes a considerable amount of time. Therefore, it is beneficial to run the simulation in several batches. After each batch, parameters are refined based on the new realizations [12] and t_l (or n_l) is computed for the next batch.

III. Accurate Error Estimation

Results obtained using Monte Carlo methods have inherent sampling uncertainty. Quantifying that uncertainty, e.g. using confidence intervals, is essential for their use. Such confidence intervals are often based on the normal approximation (4) or (13) of the error, making use of the central limit theorem and empirical estimates of the variances (5) or (11). Although this is consistent for very large sample counts, it is not obvious which quantity of samples is required for this to hold sufficiently well. This information is especially critical for MLMC, which often reduces sample counts for detailed models (i.e., the upper levels in the stack) as part of its approach to improve computational efficiency.

A. Quantifying Convergence of Ensembles

We first focus on methods to verify whether the CLT is applicable for a given case (including a specific number of samples) when we have access to a large set of runs. For regular MC, a single run provides one sample mean $\hat{q}_{MC} = (1/n)\sum_{i=1}^{n} x^{(i)}$. By performing k independent MC runs (e.g. 500) we can build sample mean distribution for n samples. Various quantitative or visual normality tests can be used to judge whether the distribution is considered sufficiently normal, empirically validating the applicability of the CLT for a given number of samples.

Moreover, if the distribution is approximately normal and the MC variance (5) is estimated accurately, we can test the error distribution (4), if a reference value q^{ref} is available. Then we can calculate the z-score, which should be distributed according to a standard normal distribution:

$$z(\hat{q}_{MC}) = \frac{\Delta\hat{q}_{MC}}{\text{Var}(\hat{q}_{MC})} = \frac{\hat{q}_{MC} - q^{ref}}{\text{Var}(\hat{q}_{MC})} \sim \mathcal{N}(0,1), \tag{15}$$

where $\text{Var}(\hat{q}_{MC})$ is the empirically estimated variance. For a number of independent runs, normality (with known mean and variance) can again be tested for using quantitative or visual means. A weaker test that can be performed is to compute the coverage fraction of confidence intervals of different percentages.

For MLMC, identical tests for normality can be performed at two aggregation levels: for individual level pairs (7) or for the overall MLMC estimator (10). Specifically, for the latter we can use a reference result q^{ref} to test for

$$z(\hat{q}_{ML}) = \frac{\Delta\hat{q}_{ML}}{\sqrt{\text{Var}(\hat{q}_{ML})}} = \frac{\hat{q}_{ML} - q^{ref}}{\sqrt{\text{Var}(\hat{q}_{ML})}} \sim \mathcal{N}(0,1). \tag{16}$$

B. An Online Test for Convergence

In resource adequacy assessment and many other real world applications, sampling from the true distribution multiple times to validate the CLT is impractical, and would defeat the purpose of error estimation of a single run. When only a single run (with a single set of samples) is available, the previous tests cannot be used. Hence we propose a test that *can* be used in this setting and that acts as a proxy for whether multiple runs would be normally distributed. That in turn suggests that the resulting confidence intervals can be relied on.

Our proposed test is based on the nonparametric bootstrap procedure, where a single set of realisations is resampled many times and the sample mean is computed for each bootstrap set. According to Singh and Kesar [13], as the number of realizations (n) increases, the difference between the sample mean distribution and the bootstrapped mean distribution tends towards zero. Therefore, as the distribution of sample means approaches normality, so does the bootstrapped distribution. We therefore hypothesize that sufficient normality of the bootstrapped means implies sufficient normality of the sample mean distribution itself, allowing us to assess CLT conditions based on the bootstrapped distribution.

Within the MLMC framework, the bootstrap is applied to the realizations $\mathcal{Y}_l = \{y_l^{(1)}, \ldots, y_l^{(n_l)}\}$ from each MLMC level

[1]We use lowercase letters to show realizations of random variables

l. Each bootstrap iteration b randomly draws n_l values $y_l^{(i,b)}$ with replacement from $\mathcal{Y}_l$ and calculates the mean

$$\hat{r}_l^{(b)} = \sum_{i=1}^{n_l} y_l^{(i,b)}. \tag{17}$$

With 2000 iterations, which is considered to be adequate for applications to confidence intervals [14], there are enough sampled means $\hat{r}_l^{(b)}$ to represent and analyze their distribution. That is, we test for

$$\{\hat{r}_l^{(1)}, \ldots, \hat{r}_l^{(B)}\} \sim \mathcal{N}(\mu_l, \sigma_l^2), \qquad l \in \{1, \ldots, L\}. \tag{18}$$

When the results indicate normality for *all* levels l, this also holds for the sum of the independently generated samples: we can assume that (16) holds and the confidence bounds resulting from the MLMC method can be relied upon.

C. Quantativve Statistical Tests

Common statistical tests for normality of the generating distribution of samples are the Kolmogorov-Smirnov (KS), Shapiro and Anderson-Darling (AD) tests. Among these, AD is the most suitable one for examining normality of MC/MLMC samples, z-scores (e.g., (16)) and bootstrap sample means (18): AD is more sensitive to the tail distribution, which is the criterion that is most important for the coverage of MC confidence intervals. The AD test is based on the A^2 statistic, which is based on the weighted distance between empirical cumulative distribution function (EDF) and cumulative distribution function (CDF) of null hypothesis (in our case a normal distribution).

Although A^2 will fluctuate depending on the sample drawn, sufficiently large values can be taken as strong evidence against the hypothesis that the samples come from the test distribution (in our case, the normal distribution). Tables with cut-off values are available, depending on the number of samples. For sample counts over 20, a cut-off value of 1.0 roughly corresponds to a p-value of 1%.

When applying the AD test to a bootstrapped distribution, sample duplication generally results in higher values of the A^2 statistic. Hence, instead of relying on tabular information, an appropriate threshold θ is derived empirically in the next section. Having identified an appropriate threshold for A^2, we propose to validate the result of an MLMC risk calculation as follows: after using the bootstrap procedure for each level l, the test $A^2(\{\hat{r}_l^{(1)}, \ldots, \hat{r}_l^{(B)}\}) \leqslant \theta$ is used, for a normal distribution with unknown mean and standard deviation. This test should be satisfied for each level in order for the overall result to be considered dependable. For multi-risk analyses like resource adequacy, this procedure should be applied to all risk indices and if the test fails for one of these, more samples are required at the required level to ensure validity of the results.

IV. Result and Discussion

A. Test System

To demonstrate the challenges and reliability of MLMC, a generation adequacy application with multiple battery storage units is used [7]. This study is based on data from the Great Britain (GB) system; annual demand traces and wind traces are randomly drawn from historical data and a synthetic data set respectively.

Two battery energy dispatching models (M_2, M_1) are used in the MLMC. The reference model (M_2) is the EENS-minimising dispatch policy given in [15] which provides X_2. A simplified model (M_1), which adds the power and energy parameters of all storage units into a single unit that is dispatched greedily to avoid load shedding, which results in an optimistic assessment of EENS. We refer readers to [7] for further details about data and models.

Note that we use a two level MLMC setup to avoid unnecessary complexity caused by several levels and keep the results clear and simple. However, the finding can be extended for MLMC with several levels or regular (single level) MC methods. All the experiments were implemented in Python and the Anderson-Darling test from the Python package `scipy.stats` was used.

B. Ensemble Convergence

First, the convergence of MLMC runs was investigated using an ensemble of independent runs. Prior to starting this calculation, a long test run was initiated to determine the optimal ratio of samples between the upper level pair (Y_2) and the lower level (Y_1) as $n_2 : n_1 = 1 : 55$. This avoided having a warm-up phase and ensured repeatability of the results. The ensemble consisted of 800 runs with $n_2 = 1,100$ and $n_1 = 60,500$, but results were also investigated for intermediate sample counts. In all cases, the average results across all complete runs served as reference value q^{ref} for LOLE, EENS as well as reference values for their partial results r_1 and r_2. The errors of these reference values were considered negligible.

Figure 1 illustrates the EDF of risk values computed by the MLMC with sample sizes of $n_1 = 1100$ and $n_2 = 20$. In each subplot the solid line represents the EDF of 800 runs, alongside a dashed line depicting the analytical CDF of the best fitting normal distribution. The first two columns show the MLMC level contributions $\hat{r}_l$, while the subsequent columns display MLMC LOLE and EENS estimates, and their z-scores, which should conform to a standard normal distribution. The EDFs of $\hat{r}_1$ closely resemble normal distributions compared to $\hat{r}_2$ due to the former's higher sample size, as evidenced by the AD test statistic A^2. However, the LOLE EDF in r_2 deviates further from normal distribution, attributed to its discrete impact function. Discrepancies between the CDF and EDF for the z-score for EENS suggest potentially unreliable risk estimation with this simulation budget. Nevertheless, the empirical coverage probability of 95% confidence intervals remains reasonable at 93%.

Figure 2 shows the convergence of relative errors to a normal distribution by plotting the A^2 statistic of normalized error (z-statistic) of LOLE and EENS estimates alongside that of their contributing levels. In this case, the convergence of the errors in $\hat{q}_{ML}$ is dominated by those in $\hat{r}_1$, and the two

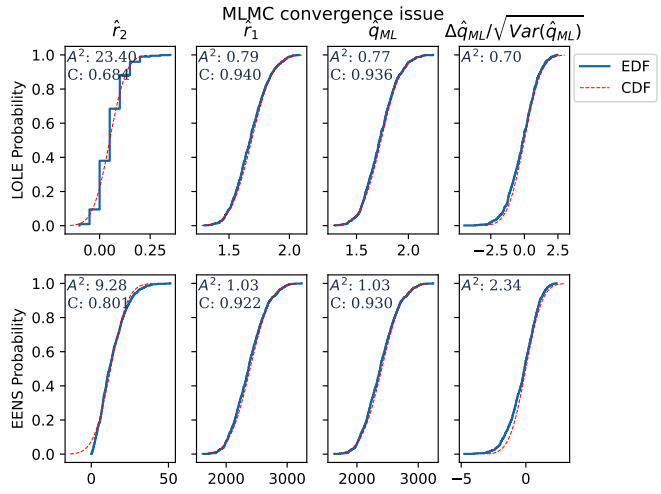

Fig. 1. Comparison between analytical CDF and EDF of MLMC simulation with $n_1 = 1.1k, n_2 = 20$ samples over 800 runs. Values of the A^2 statistic and the empirical coverage C of the 95% confidence interval (probability of having the reference value q^{ref} in the interval) are reported.

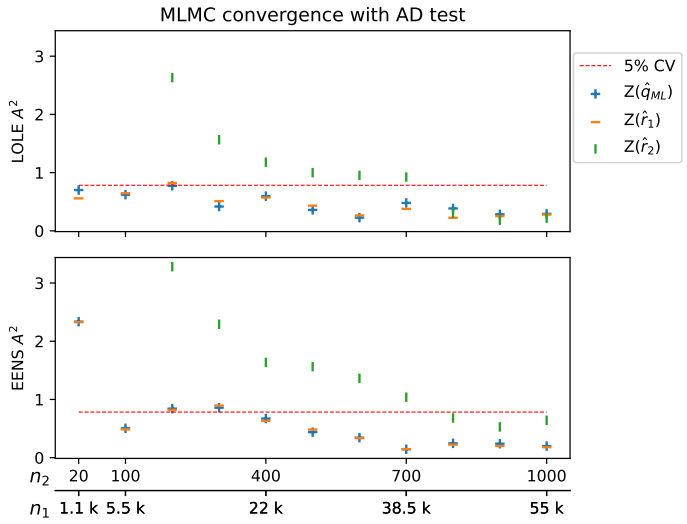

Fig. 2. Values of A^2 statistic measured over 800 runs for LOLE (top) and EENS (bottom). Three sets of symbols indicate the normalized errors $z(\cdot)$ of the MLMC estimate $\hat{q}_{ML}$ and the level contributions $\hat{r}_1$ and $\hat{r}_2$. The dashed line is the cut-off value for the 5% significance level of the AD test.

follow a similar pattern. The estimation of $\hat{r}_2$ requires around 800 samples to sufficiently converge, which takes substantially longer than the convergence of $\hat{r}_1$ when the optimal long-term sample proportion of $1 : 55$ is used. Taking a conservative approach, requiring convergence of both levels also guarantees convergence of the overall result.

C. Single-Run Convergence Tests

To determine the threshold θ for sufficient normality of the bootstrap distribution, we use two criteria. First, according to Fig. 2, for both LOLE and EENS the z-score of MLMC is well-behaved for $n_2 \geqslant 800$, so we expect to see a large probability of acceptance for those samples. Second, if we

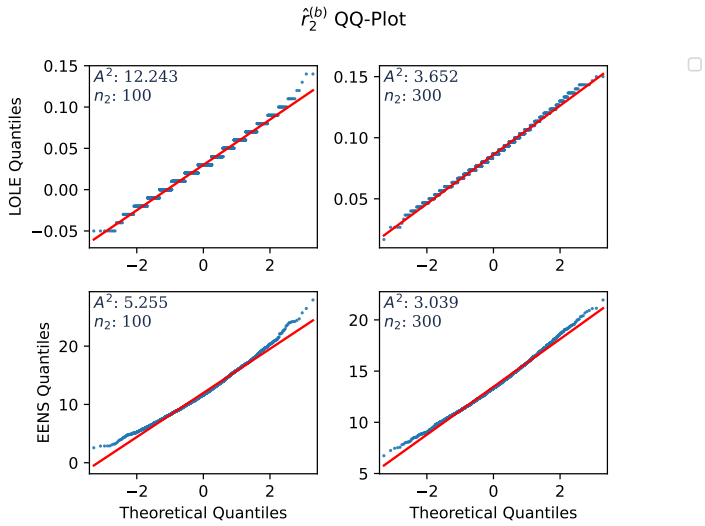

Fig. 3. QQ-plot of bootstrap means $r_2^{(b)}$ (upper level of MLMC) for LOLE (top) and EENS (bottom). The solid line indicates the reference in all subplots. Results for $n_2 = 100$ on the left; for $n_2 = 300$ on the right.

use the threshold to accept individual runs, those runs should *collectively* form a well-behaved distribution. This can again be tested using A^2 statistic.

To propose the appropriate cut-off for the normality test, a new set of 500 MLMC runs was conducted using sample sizes up to $n_2 = 1000$. For each run, the A^2 statistic was computed for the bootstrapped values LOLE and EENS, and accepted if both were below the threshold. Table I displays the fraction of accepted samples for each combination of MC samples and threshold value. In parentheses, the quality of accepted results is quantified by reporting the A^2 values of all accepted runs, for LOLE and EENS.

For cutoff values of 1.0 and 2.0, the acceptance probabilities are very low, even for $n_2 \geqslant 800$. The high A^2 values for LOLE with $n_2 = 300$ and $n_2 = 500$ with a cutoff value of 5.0 indicate that some MLMC estimations approved by the bootstrap procedure may not be reliable. Cut-off values of 3.0 and 4.0 both yield acceptable results, and any values within this range could be considered appropriate depending on policy considerations. Further tests remain to be done to check how generalizable this threshold value is. The computational overhead of the test is on the order of seconds, for relevant sample sizes. This is typically negligible relative to the overall study time.

An alternative approach to determine an appropriate threshold is to visually inspect QQ-plots of the bootstrap means $r_l^{(b)}$ against the quantiles of a normal distribution. Figure 3 shows that the QQ-plot bootstrap means $r_2^{(b)}$ increasingly approximate a normal distribution as the number of initial samples increases from 100 to 300 (the number of bootstrap resamples remains 2000). The results for $n_2 = 300$ may be considered acceptable, as is confirmed by the (bootstrapped) A^2 values around 3.5.

TABLE I
BOOTSTRAP FILTERED RESULTS: ACCEPTANCE FRACTION, A^2 FOR LOLE (L: #) AND EENS (E: #)

n_2 Cut-off	100	200	300	400	500	600	700	800	900	1000
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
2.0	0.00	0.00	0.00	0.02 ($L:0.17$) ($E:0.40$)	0.08 ($L:0.45$) ($E:0.26$)	0.24 ($L:1.13$) ($E:0.87$)	0.44 ($L:0.32$) ($E:0.43$)	0.64 ($L:0.37$) ($E:0.16$)	0.76 ($L:0.51$) ($E:0.25$)	0.85 ($L:0.42$) ($E:0.30$)
3.0	0.00	0.01	0.06 ($L:0.17$) ($E:0.26$)	0.31 ($L:0.71$) ($E:0.26$)	0.65 ($L:0.73$) ($E:0.47$)	0.88 ($L:0.47$) ($E:0.48$)	0.94 ($L:0.27$) ($E:0.17$)	0.97 ($L:0.32$) ($E:0.28$)	0.98 ($L:0.28$) ($E:0.27$)	0.99 ($L:0.35$) ($E:0.36$)
4.0	0.00	0.05 ($L:0.40$) ($E:0.35$)	0.35 ($L:0.44$) ($E:0.28$)	0.77 ($L:0.61$) ($E:0.31$)	0.92 ($L:0.68$) ($E:0.36$)	0.99 ($L:0.43$) ($E:0.49$)	1.00 ($L:0.28$) ($E:0.18$)	1.00 ($L:0.28$) ($E:0.21$)	1.00 ($L:0.27$) ($E:0.27$)	1.00 ($L:0.32$) ($E:0.36$)
5.0	0.01	0.18 ($L:0.51$) ($E:0.51$)	0.68 ($L:0.81$) ($E:0.71$)	0.94 ($L:0.50$) ($E:0.32$)	0.99 ($L:0.83$) ($E:0.35$)	1.00 ($L:0.44$) ($E:0.53$)	1.00 ($L:0.28$) ($E:0.19$)	1.00 ($L:0.27$) ($E:0.21$)	1.00 ($L:0.27$) ($E:0.27$)	1.00 ($L:0.32$) ($E:0.36$)

V. CONCLUSION

This paper underscores the critical role of uncertainty quantification in Monte Carlo methods, particularly in the context of Multilevel Monte Carlo (MLMC), where only a few samples from the reference model are evaluated. It proposes a practical approach to validate error estimates, using a bootstrap technique when only a single set of realizations is accessible.

Our experimental results demonstrate that the bootstrap test provides an effective measure of the reliability of MC/MLMC error estimates. This can provide a valuable addition to 'black box' sampling-based reliability assessment tools. Future research will focus on embedding convergence testing in an automated MLMC framework and quantifying the overall computational efficiency.

REFERENCES

[1] M. Firouzi, A. Samimi, and A. Salami, "Reliability evaluation of a composite power system in the presence of renewable generations," *Reliability Engineering System Safety*, vol. 222, 2022.

[2] S. Li, C. Ye, Y. Ding, Y. Song, and M. Bao, "Reliability assessment of renewable power systems considering thermally-induced incidents of large-scale battery energy storage," *IEEE Transactions on Power Systems*, vol. 38, no. 4, pp. 3924–3938, 2022.

[3] B. Zhang, M. Wang, and W. Su, "Reliability analysis of power systems integrated with high-penetration of power converters," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 1998–2009, 2021.

[4] Y. Zhao, Y. Han, Y. Liu, K. Xie, W. Li, and J. Yu, "Cross-entropy-based composite system reliability evaluation using subset simulation and minimum computational burden criterion," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5198–5209, 2021.

[5] M. B. Giles, "MLMC techniques for discontinuous functions," *arXiv preprint arXiv:2301.02882*, 2023.

[6] A. S. Meliopoulos, M. Papic, S. H. Tindemans, S. Ekisheva, M. Yue, and D. M. Logan, "Composite power system reliability with renewables and customer flexibility," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, IEEE, 2022, pp. 1–8.

[7] S. Tindemans and G. Strbac, "Accelerating system adequacy assessment using the multilevel Monte Carlo approach," *Electric Power Systems Research*, vol. 189, p. 106 740, 2020.

[8] E. Sharifnia and S. H. Tindemans, "Multilevel Monte Carlo with surrogate models for resource adequacy assessment," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, IEEE, 2022, pp. 1–6.

[9] M. Papic and D. Logan, "Bibliography on composite system reliability assessment 2000-2020," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, IEEE, 2022, pp. 1–8.

[10] S. Peyghami, P. Palensky, and F. Blaabjerg, "An overview on the reliability of modern power electronic based power systems," *IEEE Open Journal of Power Electronics*, vol. 1, pp. 34–50, 2020.

[11] C. Canals and A. Canals, "When is n large enough? looking for the right sample size to estimate proportions," *Journal of Statistical Computation and Simulation*, vol. 89, no. 10, pp. 1887–1898, 2019.

[12] B. Welford, "Note on a method for calculating corrected sums of squares and products," *Technometrics*, vol. 4, no. 3, pp. 419–420, 1962.

[13] K. Singh, "On the asymptotic accuracy of Efron's bootstrap," *Ann. Stat.*, pp. 1187–1195, 1981.

[14] B. Efron and T. Hastie, *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Cambridge University Press, 2021.

[15] M. P. Evans, S. H. Tindemans, and D. Angeli, "Minimizing unserved energy using heterogeneous storage units," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3647–3656, 2019.