Efficient Sampling Techniques for Uncertainties in Risk Analysis

Renyou Wang,[a] Urmila Diwekar,[a] and Catherine E. Grégoire Padró[b]

[a]Center for Uncertain Systems, Tools for Optimization and Management, (CUSTOM), Department of Chemical Engineering, and Institute for Environmental Science and Policy, University of Illinois at Chicago, Chicago, IL 60607; urmila@uic.edu (for correspondence)

[b]Hydrogen Systems, MST-11, MS D429, P.O. Box 1663, Los Alamos National Laboratory, Los Alamos, NM 87545

Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ep.10020

Risk and policy analysis involves consideration of uncertainties. A quantitative probabilistic method for uncertainty analysis includes: (1) quantifying and assigning probabilistic distributions to the input uncertainties, (2) sampling the distributions of these uncertain parameters in an iterative fashion using Monte Carlo methods, (3) propagating the effects of uncertainties through the model, and (4) predicting the outcomes in terms of probabilistic measures like mean, variance, and fractiles. However, the results of the probabilistic analysis depend on the number of samples chosen. The sample size required for a particular analysis depends on various factors such as type of model, the random number generator used, type of distributions, and the output probabilistic measure and cannot be universally defined. The general tendency is to reduce the samples as much as possible without realizing the effect on decisions. For example, the mean of the output requires a number of samples that is an order of magnitude less compared to the variance. Therefore, it is desirable to use a sampling technique that can predict the output probabilistic measure accurately with the minimum number of samples. In this work, we present new sampling techniques based on Quasi-Monte Carlo sequences and Latin hypercube sampling.© 2004 American Institute of Chemical Engineers Environ Prog, 23: 141–157, 2004

Keywords: *Latin hypercube Hammersley sequence, Hammersley sampling techique, efficiency*

1. INTRODUCTION

Risk assessment involves assessment of the nature and magnitude of risk or harm arising from some articulable hazard [1]. This risk could be in terms of health or it could be ecological or economical. The health risk assessment process can be distilled into the following simple model:

$$\text{Risk} = \text{Dose} \times \text{Toxicity}.$$

It involves the identification of potential health effects associated with exposure of individual or populations to hazardous materials and situations. In the above model, if either exposure or toxicity is lacking, there is no risk. Equally true, however, is the certainty that the determination of the nature and extent of the dose-associated toxicity is not always easy. Ecological risk assessment is fundamentally quite similar to the human health risk assessment, but the terms used to define the process are slightly different. Obviously, there are numerous, inherent uncertainties in the human health and ecological risk assessment processes. Economic risks, on the other hand, are associated with the probability of failure to meet the target and are closely related to the effect of uncertainties.

Simulations with computer models are used for the risk assessment of systems that are too complex to analyze directly. As stated earlier, the risk assessment is often complicated by the uncertainties involved in the system. For many such systems in economics, energy, environment, and other technology-related fields, the science of the problem is reasonably well understood, but uncertainties exist in some important model inputs or parameters. In conducting stochastic risk analysis in these applications, the uncertainty or variability is often expressed in terms of probabilistic distributions. Simulations for stochastic risk analysis require a large number of Monte Carlo simulations. The problem rises as in real-world applications for the uncertainty analysis of well-established models, where the number of sample

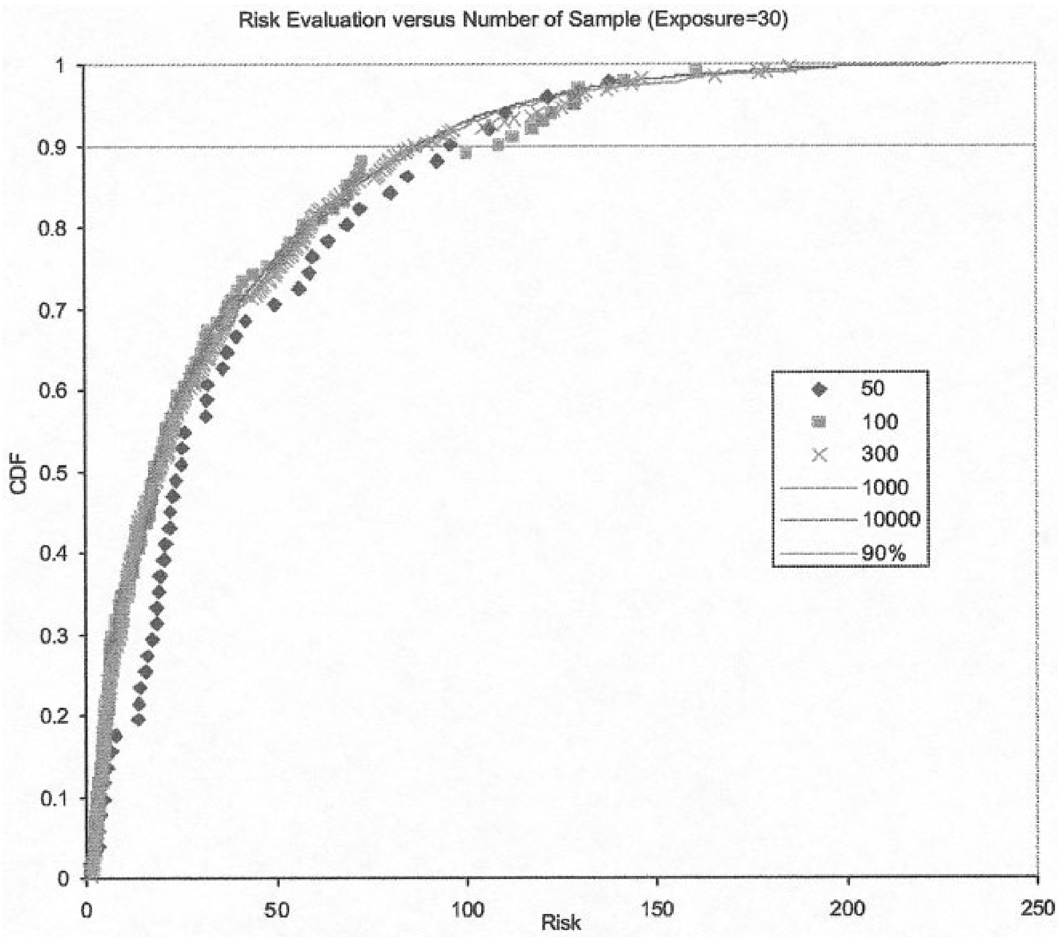


Figure 1. Evaluation of health risk under exposure 30. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

Table 1. Estimate error of risk under exposure 30 at CDF 90%.

Number of sample	50	100	200	300	500	1000	5000	10,000
Risk	95.9	108	82.8	86.6	88.2	88.8	84.9	86.7
Risk error (%)	9.10	23.13	5.82	1.52	0.29	0.98	3.41	1.39

size is often high for using Monte Carlo sampling in order to achieve a high accuracy of the resulting distribution of outcomes. Here the problem exists when the combination of many variables and the complex relationships among the variables result in time-consuming computing processes that require several hours or even days to complete a simulation run for one set of input variables. This set of input variables is only one of the sampling points from uncertain input domain and also the subset of the *k*-dimensional space defined by the range of *k* group sets of uncertain parameters. Because of the expense and time involved, it is unrealistic or sometimes even impossible to have a large number of runs to achieve high accuracy. Unfortunately, a complete analysis of the model is still desired just based on these few runs, sometimes numbering only between 50 and 100. In this case, simulation with the Monte Carlo sampling technique may just offer muddling results and sometimes cause serious troubles or even lead to wrong decisions. For demonstration, let us have a look at a simple example from risk analysis.

1.1. A Simple Example

The following example is taken from Small and Fischbeck [2]. A hypothesized health effect risk(R)-exposure (X) model is assumed as given below for calculation of the risk,

$$R = a + bX^c; \quad X \geq X_t = 0; \quad X < X_t, \tag{1}$$

where a, b, and c are model constants and the threshold X_t, b, and c are uncertain and uniformly distributed U(lower bound, upper bound). The uncertainties in $X_t(U(0, 20))$, $b(U(0.5, 1.5))$, and $c(U(0.5, 1.5))$ are assumed independent. Such risks could exist because of

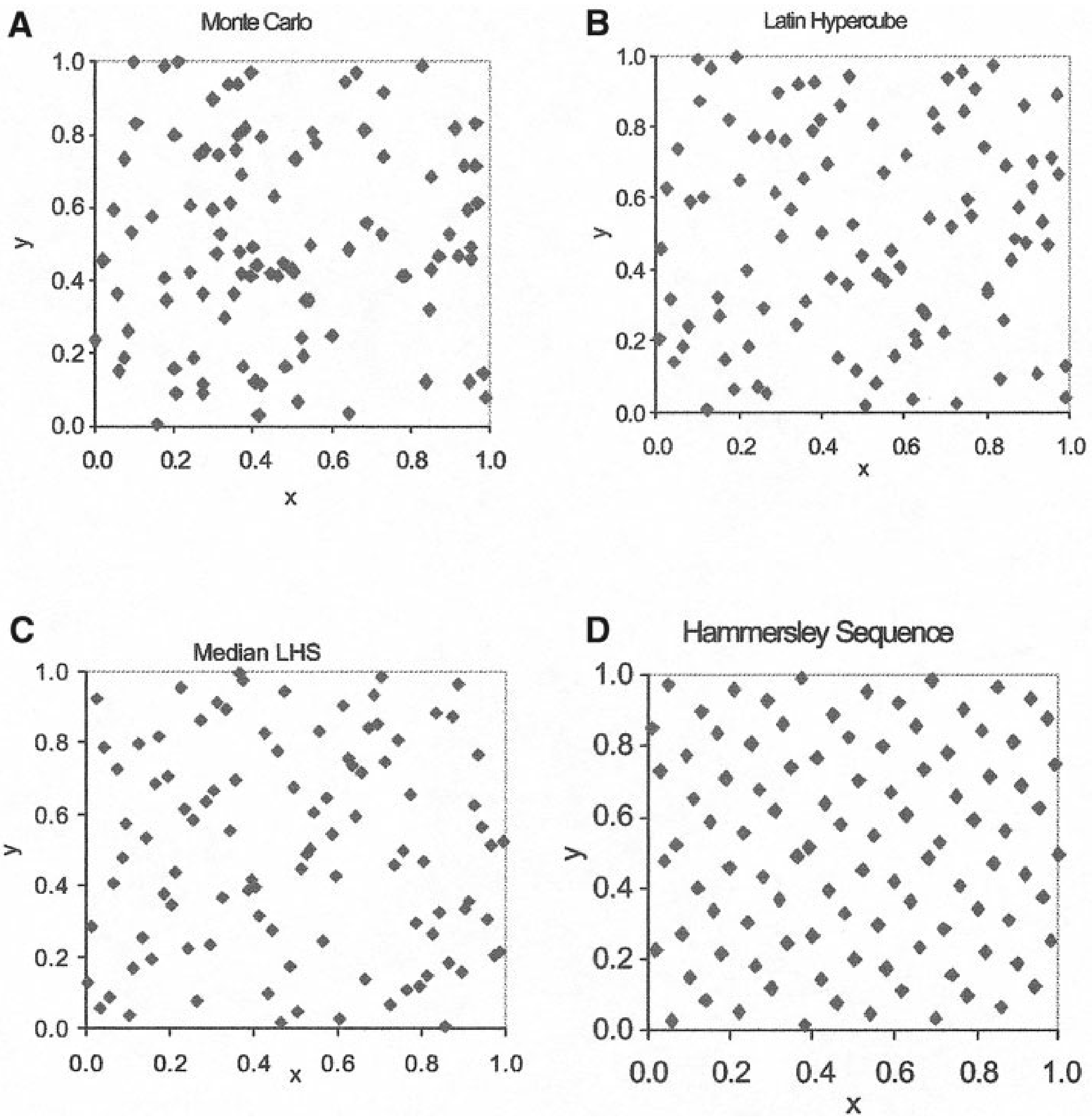


Figure 2. Sample points (100) on a unit square using (A) Monte Carlo sampling, (B) Latin hypercube sampling, (C) median Latin hypercube sampling, and (D) Hammersley sequence sampling techniques ($x - U(0, 1)$, $y - U(0, 1)$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

human exposures to radon, electric and magnetic fields, lead, asbestos, etc. Our objective is to evaluate the risk of such exposures and make decisions. The model shows that health effect exposure risk is affected by all of the above uncertainties and this effect can be calculated using Monte Carlo sampling.

For simplification of analysis without losing generalization, let's examine the risk in case of exposure 30 ($X = 30$). Figure 1 is the CDF curve of risk calculated using Monte Carlo sampling technique. With the same probability of 90%, different risk values could be predicted with different sample size for each Monte Carlo simulation. Assume that the mean of the data with the sample size of 10,000 of 10 runs is the actual data and if we think it is safe when the risk is below 95 with probability of 90%, from Table 1 the calculation with sample size 50 or 100 will show that it is not safe with exposure 30 as the risks are 95.9 and 108.0, respectively. Table 1 also shows that the calculation error with a small sample size of 50 or 100 could approach 23.12%. But the actual risk is around 87.9 and hence exposure 30 is safe. Therefore, we may reach the wrong decision from the small sample size Monte Carlo simulations. This example shows that the sample size can be very important for the analysis of uncertainties in decision making and also in risk analysis.

Because the sample size is closely related to the sampling techniques, a good sampling technique can greatly reduce the sample size so that we could get the correct results with a small number of samples.

1.2. Scope of Work

The purpose of this study is to develop a new efficient sampling technique for uncertainty analysis, es-

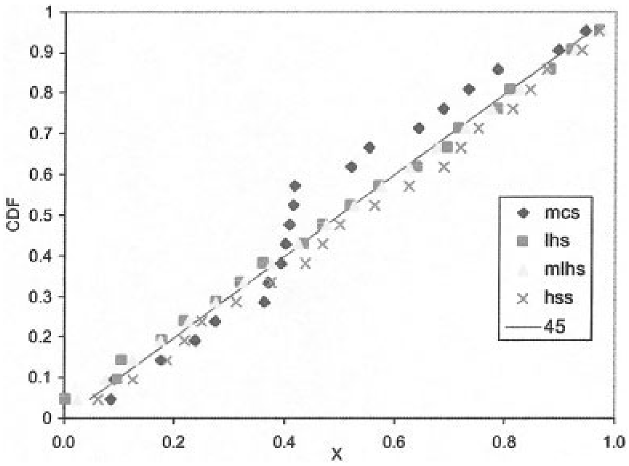


Figure 3. One-dimensional uniformity analysis with sample points (20) using (A) Monte Carlo sampling, (B) Latin hypercube sampling, (C) median Latin hypercube sampling, and (D) Hammersley sequence sampling techniques ($x - U(0, 1)$) [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com].

pecially for large and complicated models with more than a few uncertain input variables. The proposed new sampling technique, the Latin hypercube-Hammersley sequence sampling (LHSS) technique, is based on the analysis of current sampling techniques: Monte Carlo sampling (MCS), Latin hypercube sampling (LHS), median Latin hypercube sampling (MLHS), and the recently developed sampling technique called Hammersley sequence sampling (HSS) [3, 4]. The new sampling technique developed here is by coupling the LHS and HSS techniques. The purpose of such a coupling is to ensure that the new sampling technique inherits all of the advantages of LHS and HSS while the disadvantages of them could be overcome. Various experimental efficiency tests were carried out during this research and the results are compared with those of the other four sampling techniques. The number of input variables used varies between 2 and 10, while seven different kinds of functions, linear additive function, multiplicative function, quadratic function, exponential function, logarithmic function, and health effect risk-exposure function with a threshold, are used in these tests. Different types of distributions and three types of correlation structures for the uncertain input variables have been used in this paper. In addition, one real-world application of this new sampling technique (LHSS) related to economic risk analysis of a renewable energy power system is examined. This case study is about the technical feasibility of a solar energy power system design capacity in a rural village in Central America. In this system, the driving forces—solar insolation and the electric demands of this village—are all uncertain variables and change with time. By simulation study, we can test whether the proposed design capacity of this energy power system is technically feasible, thereby assessing the economic risk of not meeting the power demand for this village.

2. CURRENT SAMPLING TECHNIQUES

The simple Monte Carlo sampling technique was proven to have poor performance. Variance reduction techniques (VRT) are statistical procedures designed to reduce the estimate variance without the requirement of increasing the number of simulation runs n, which is correspondent to the sample size.

James [5] presented four categories of variance reduction techniques. Category 1 includes techniques like the control variate technique that calls for model modification (reduction) and category 3 is specific to correlated uncertain variables. The most generalized and commonly used methods in these four categories are stratified sampling techniques from category 2 and importance sampling techniques from category 3.

In most applications, the actual relationship between successive points in a sample has no physical significance; hence, the independence/randomness of a sample for approximating a uniform distribution is not critical [6]. Once it is apparent that the uniformity properties are central to the design of sampling techniques, constrained or stratified sampling becomes appealing [7].

Stratified sampling techniques ensure that more samples are generated from high-probability regions. On the other hand, importance sampling techniques guarantee full coverage of high-consequence regions in the sample space, even if these regions are associated with low probabilities. This makes importance sampling techniques problem-dependent. Therefore, importance sampling techniques are not further considered here.

Latin hypercube sampling is one form of stratified sampling that can reduce the variance in the Monte Carlo estimate of the integrand significantly [8]. The range of each input uncertain variable is divided into nonoverlapping intervals of equal probability. One value from each interval is selected at random with respect to the probability density in the interval. In MLHS this value is chosen as the midpoint of the interval. MLHS is similar to the descriptive sampling described by Saliby [9]. The n values thus obtained for variable 1 are paired in a random manner with the n values of variable 2 and these n pairs are combined with n values of variable 3 and so on to form n k-tuplets. The random pairing is based on a pseudorandom number generator. Then, the input sample is generated based on the inverse transform method and given by

$$x_{j,i} = F_j^{-1}\left(\frac{(i - 1 + r_i)}{N}\right), \quad i = 1, 2, \ldots, N,$$
$$j = 1, 2, \ldots, k, \tag{2}$$

where the CDF of $\tilde{x}_j$ is equally divided into N partitions, r_i stands for an independent random number on [0, 1], and F_j^{-1} is the inverse transform for the distribution of input variable j. In total, there are N observations generated for each of the multidimensional input variables.

LHS shows a better 1-dimensional uniformity property and better represents its distribution. But the con-

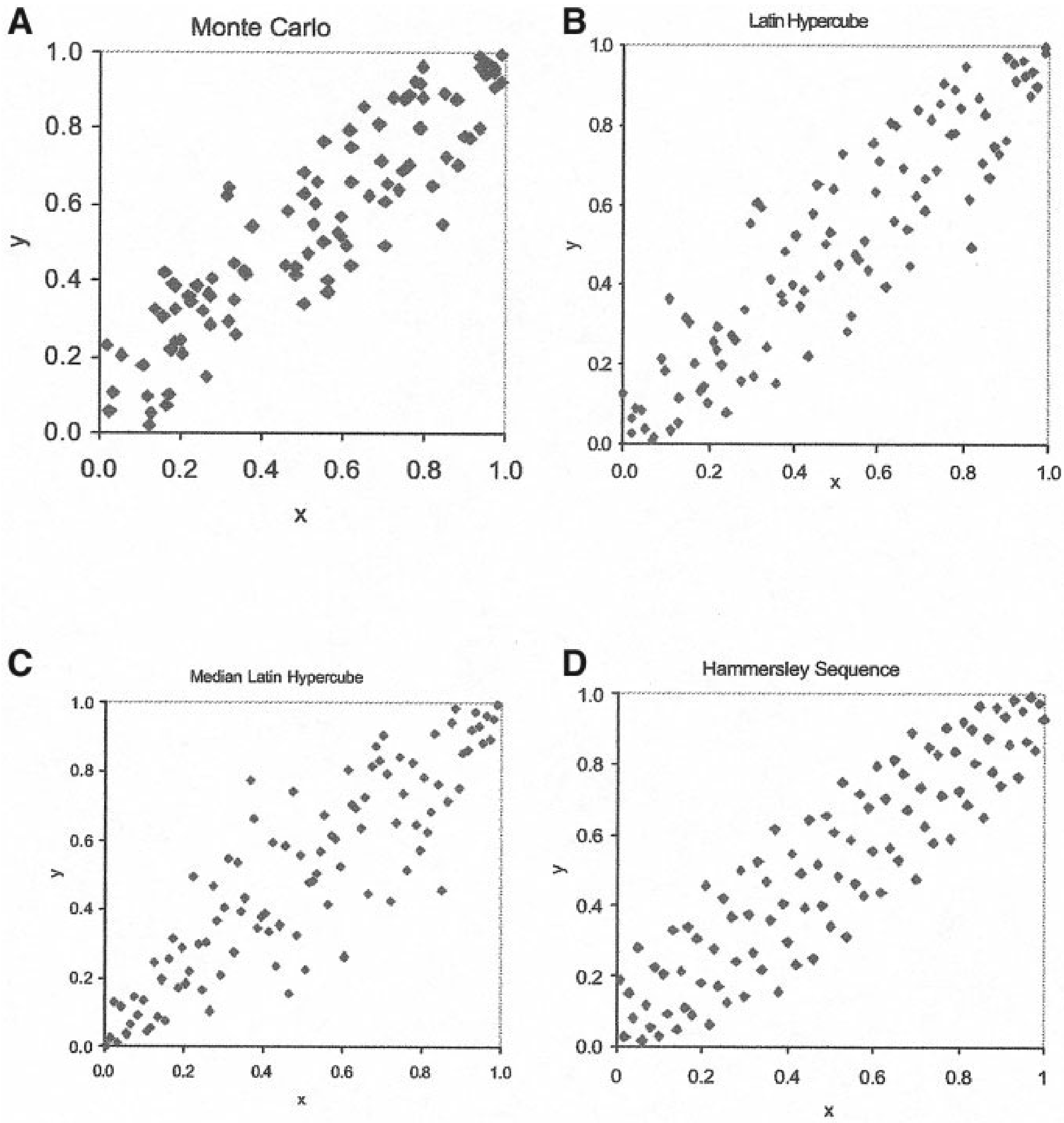


Figure 4. Sample points (100) on a unit square with correlation of 0.9 using (A) Monte Carlo sampling, (B) Latin hypercube sampling, (C) median Latin hypercube sampling, and (D) Hammersley sequence sampling techniques ($x - U(0, 1)$, $y - U(0, 1)$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

vergence rate is low due to its poor multidimensional uniformity.

Until now, one of the most efficient sampling techniques was Hammersley sequence sampling [3, 4] with which the samples are generated according to Hammersley sequence [10]. The Hammersley sequence was first proposed by Hammersley as a quasi-random number generator. In 1991, Wozniakowski [11] proposed the use of the shifted Hammersley sequence to increase the randomness of the sequence. It has long been believed that no multidimensional infinitive sequence can have discrepancy an order of magnitude smaller than $N^{-1}\log^{k}N$. The sequences of Halton [12], Faure [13], and Sobol [14, 15] all have $O(N^{-1}\log^{k-1}N)$. But Micchelli and Wahba [11] conjectured that Hammersley sequence points should lead to an order of magnitude $N^{-1}\log^{k-1}N$ and the discrepancy of optimal shifted Hammersley sequence points can even be up to an order $N^{-1}\log^{(k-1/2)}N$ [10].

Based on the low discrepancy of HSS points, Diwekar and colleagues further developed the HSS sampling technique [3, 4, 16, 17]. Their research work shows that the HSS technique has a better performance than that of MCS or LHS and is at least 3 to 100 times faster for convergence.

Figure 2 shows the 2-dimensional uniformity properties of four main sampling techniques, MCS, LHS, MLHS, and HSS. Figure 2 qualitatively shows that the Hammersley sequence sampling has the best uniformity property. However, the conclusion could be different if we consider 1-dimensional uniformity properties. Figure 3 shows the 1-dimensional uniformity of all

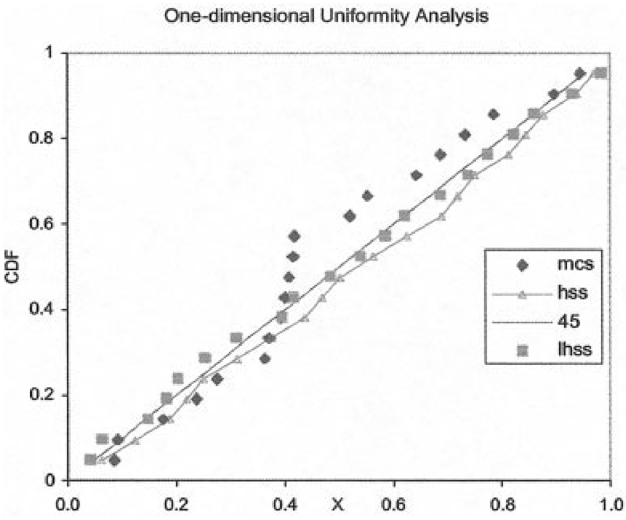


Figure 5. One-dimensional uniformity analysis with sample points (20) using (A) Monte Carlo sampling, (B) Hammersley sequence sampling, and (C) Latin hypercube-Hammersley sequence sampling techniques ($x - U(0, 1)$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

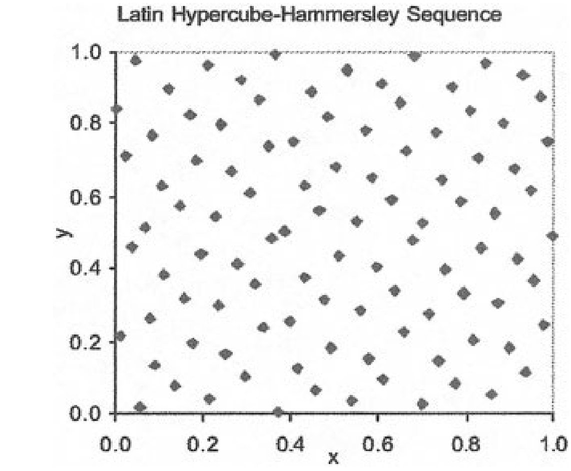


Figure 6. Sample points (100) on a unit square using Latin hypercube Hammersley sequence sampling techniques ($x - U(0, 1)$, $y - U(0, 1)$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

four sampling techniques. The better 1-dimensional uniformity is indicated by closeness to the 45° line with a uniform interval between the adjacent sample points. From Figure 3, LHS and MLHS but not HSS have better 1-dimensional uniformity.

The comparison of the sampling based on HSS with that based on Sobol's sequence [14, 15] also shows the better efficiency of HSS when the number of uncertain input variables is greater than 2.

3. LATIN HYPERCUBE HAMMERSLEY SEQUENCE SAMPLING TECHNIQUE

Based on the above analysis, the new sampling, Latin hypercube Hammersley sequence sampling, is developed by generating the sample values with LHS to utilize its better 1-dimensional uniformity and pairing them with HSS to use its better multidimensional uniformity. The distribution-free rank approach [18, 19] is employed for the pairing implementation processes.

3.1. Distribution-Free Rank Approach

The distribution-free approach to inducing correlation among input variables was proposed by Iman and Conover [18] in 1982. Until now, in risk assessment, most implementations of the correlation have relied upon this method. Haas [20] proposed another approach, but it is powerful only when the correlation is high, above 0.7, and for bivariate analysis. The following section is a simple description of the method, and for a detailed description of this methodology the original paper should be consulted.

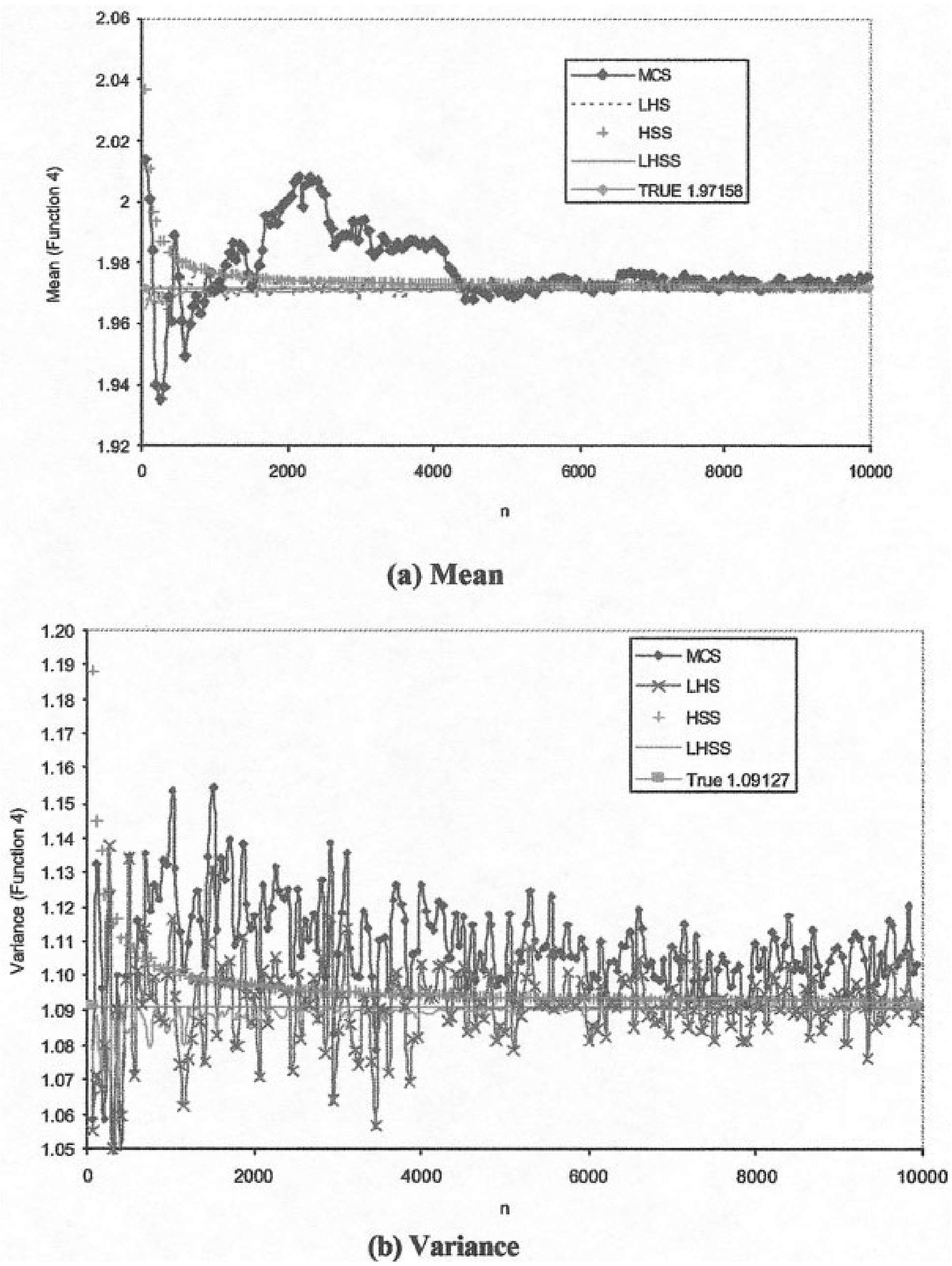
Suppose a random row vector $\vec{X}$, whose elements are represented by an ($N*k$) matrix, has a correlation matrix I. This means that the elements of $\vec{X}$ are uncorrelated. In implementing the correlation relation among input variables in order to obtain a sample in which the sample correlations better match the assumed/intended correlations, the transformed vector $\vec{X}\vec{P}^{\mathrm{T}}$ could be used to achieve the desired correlation matrix $\vec{C}$ if $\vec{P}\vec{P}^{\mathrm{T}} = \vec{C}$ where $\vec{P}$ is a lower triangular matrix [21]. Figure 4 shows the sample points for the correlation matrix given below.

$$\vec{C} = \begin{bmatrix} 0 & 0.9 \\ 0.9 & 0 \end{bmatrix}$$

3.2. LHHS Sampling

In the process of generating samples with LHHS, the sample values of each input variable are first generated using LHS. The next step is to pair them and combine the input vectors. The conventional method is to pair all of them randomly. However, the sample correlation matrix of input variables generated by either LHS or MCS with random pairing processes is not exactly equal to I and it also shows bad uniformity. In our approach, restricted pairing procedure is used in all cases. Even when the input variables are independent, the restricted pairing procedure is still employed for the desired correlation matrix I to make sure there is no actual dependence among the input variables. Diwekar *et al.* [3, 4, 15, 16] already showed that Hammersley sequence points have better multidimensional uniformity. In order to characterize the new sampling technique this property, the HSS matrix $\vec{H}(N*k)$ corresponding to van der Waerden scores matrix in Iman and Conover's approach in LHS, is used in pairing procedures. To avoid the problem associated with $\vec{H}(N*k)$ not having a correlation matrix equal to I, the sample correlation matrix $\vec{R}(k*k)$ associated with $\vec{H}(N*k)$ is used to find a matrix $\vec{S}$ so that

$$\vec{S}\vec{R}\vec{S}^{\mathrm{T}} = \vec{C}, \tag{3}$$



(a) Mean

(b) Variance

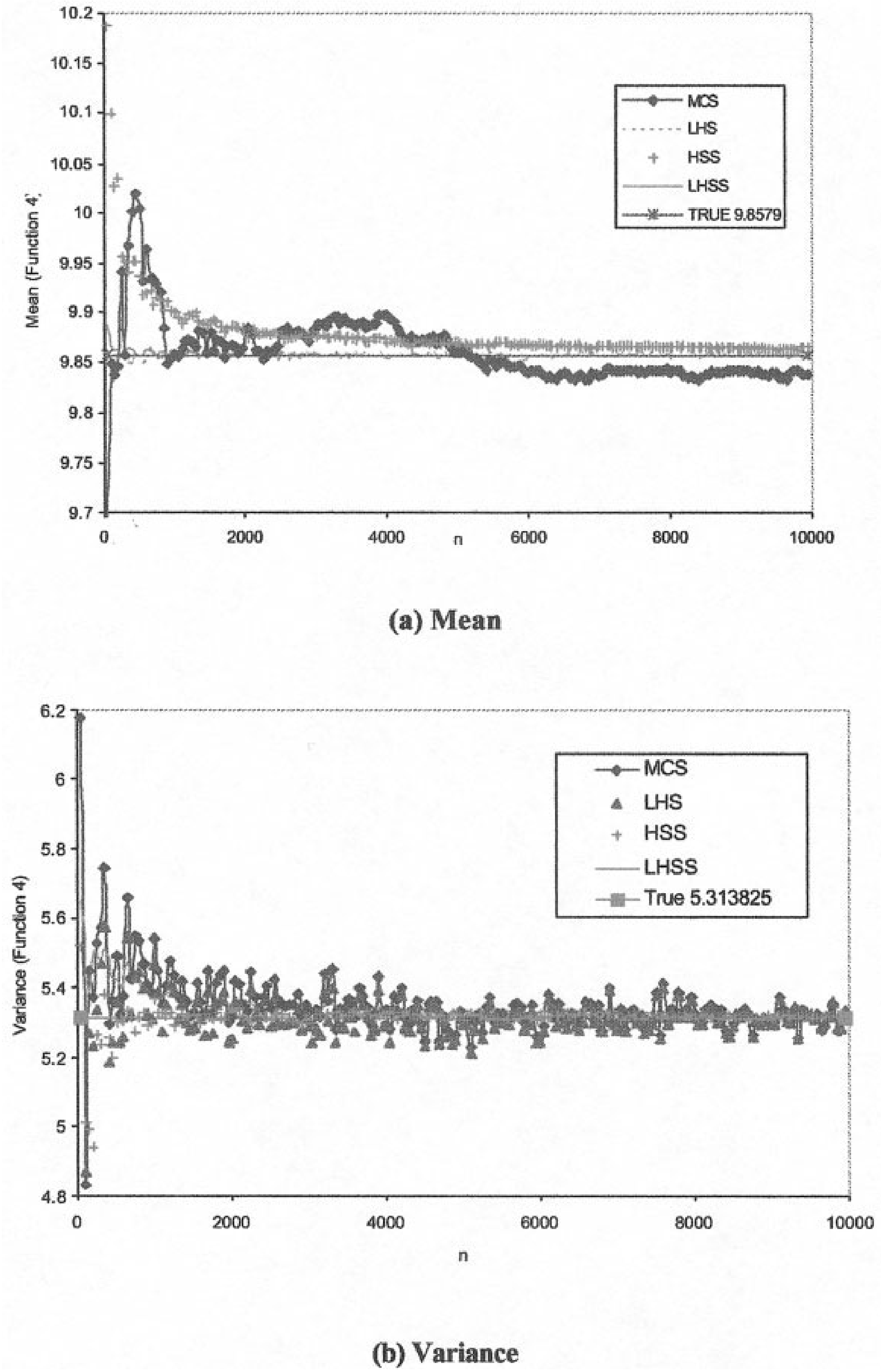
Figure 7. The mean (a) and variance (b) of function 4 as a function of sample size for MCS, LHS, HSS, and LHSS for two input variables without correlations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

where $\vec{C}$ is the desired sample correlation matrix. The same as above, the Cholesky factorization is used to find a lower triangular matrix $\vec{Q}$ such that

$$\vec{Q}\vec{Q}^{\mathrm{T}} = \vec{R}. \tag{4}$$

Therefore, the solution of $\vec{S}$ can be found, which is given by $\vec{S} = \vec{P}\vec{Q}^{-1}$, and correspondingly the transformation factor for the rank matrix is changed to $\vec{S}$ and the rank matrix becomes $\vec{H}^{*} = \vec{H}\vec{S}^{\mathrm{T}}$. The correlation matrix of $\vec{H}^{*}$ is exactly equal to the desired correlation matrix $\vec{C}$. The sample can therefore be paired according to the new rank matrix $\vec{H}^{*}$ rather than $\vec{H}$.

In this pairing process, when a correlation structure is not specified, variance of inflation factor (VIF), defined as the largest element on the diagonal of the inverse of the correlation matrix, is computed to detect the large pairing correlations. As the VIF gets much larger than 1, there may be some undesirably large pairing correlations. For VIF 10, there can be serious collinearity [22, 23]. The calculation results for the new sampling show that when the



(a) Mean

(b) Variance

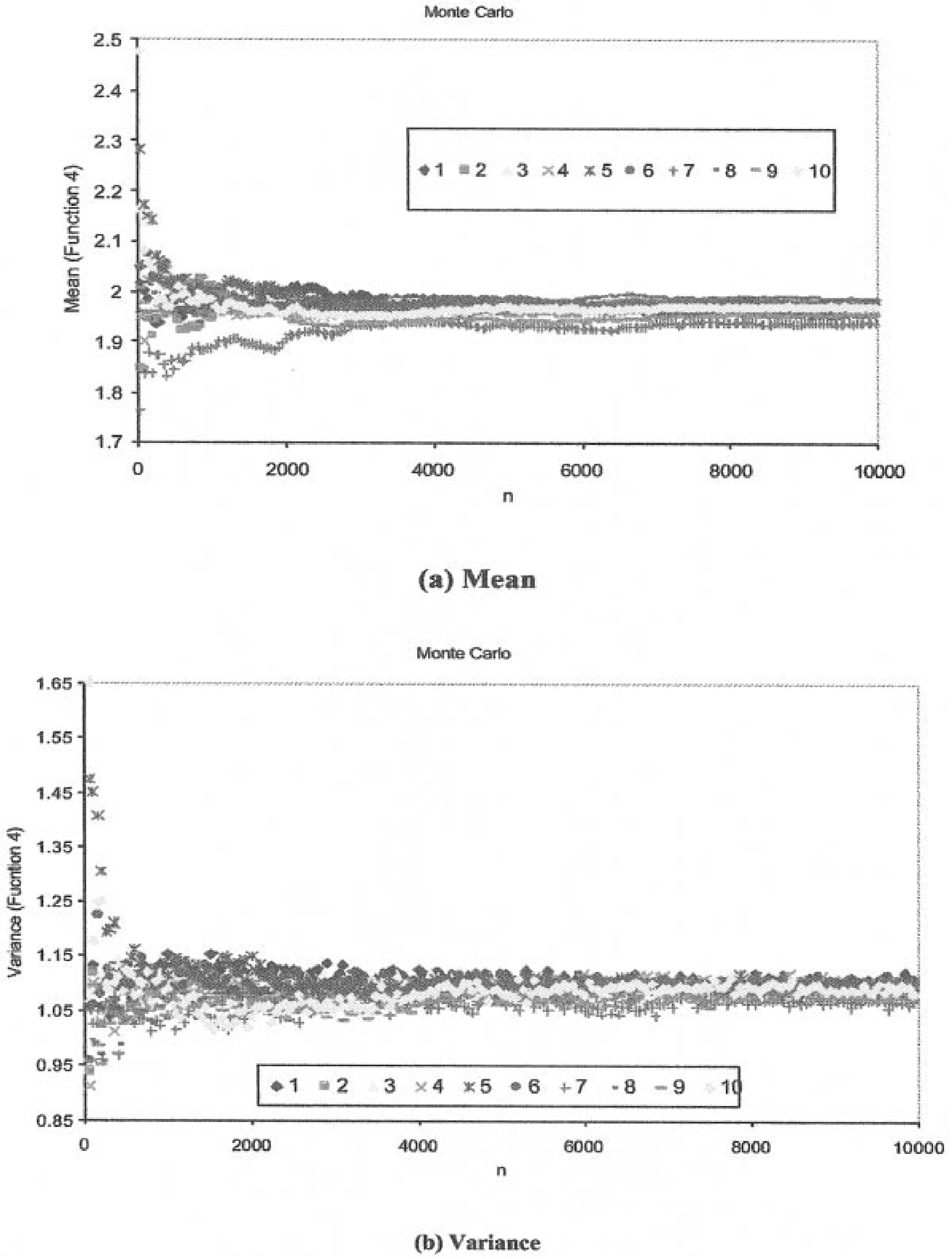
Figure 8. The mean and variance (function 4) as a function of sample size for MCS, LHS HSS, and LHSS for 10 input variables without correlations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

sample size N is much larger than the sampling input variables number, VIF is very close to 1. Therefore, there is no inflation problem for this restricted pairing processes.

4, EFFICIENCY ANALYSIS

The new sampling is expected to have not only a better 1-dimensional uniformity property but also a better multidimensional uniformity property. Figure 5 shows the 1-dimensional uniformity property of LHHS and other sampling techniques. The interesting result shows that LHHS has inherited the better 1-dimensional uniformity property of LHS. Compared with HSS, LHHS has improved greatly.

Figure 6 graphs the sample points with LHHS on a unit square. From this qualitative picture, LHHS seems



(a) Mean

(b) Variance

Figure 9. The mean and variance of output of function 4 as a function of sample size (n) for MCS with two input variables and 10 different initial random seeds. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

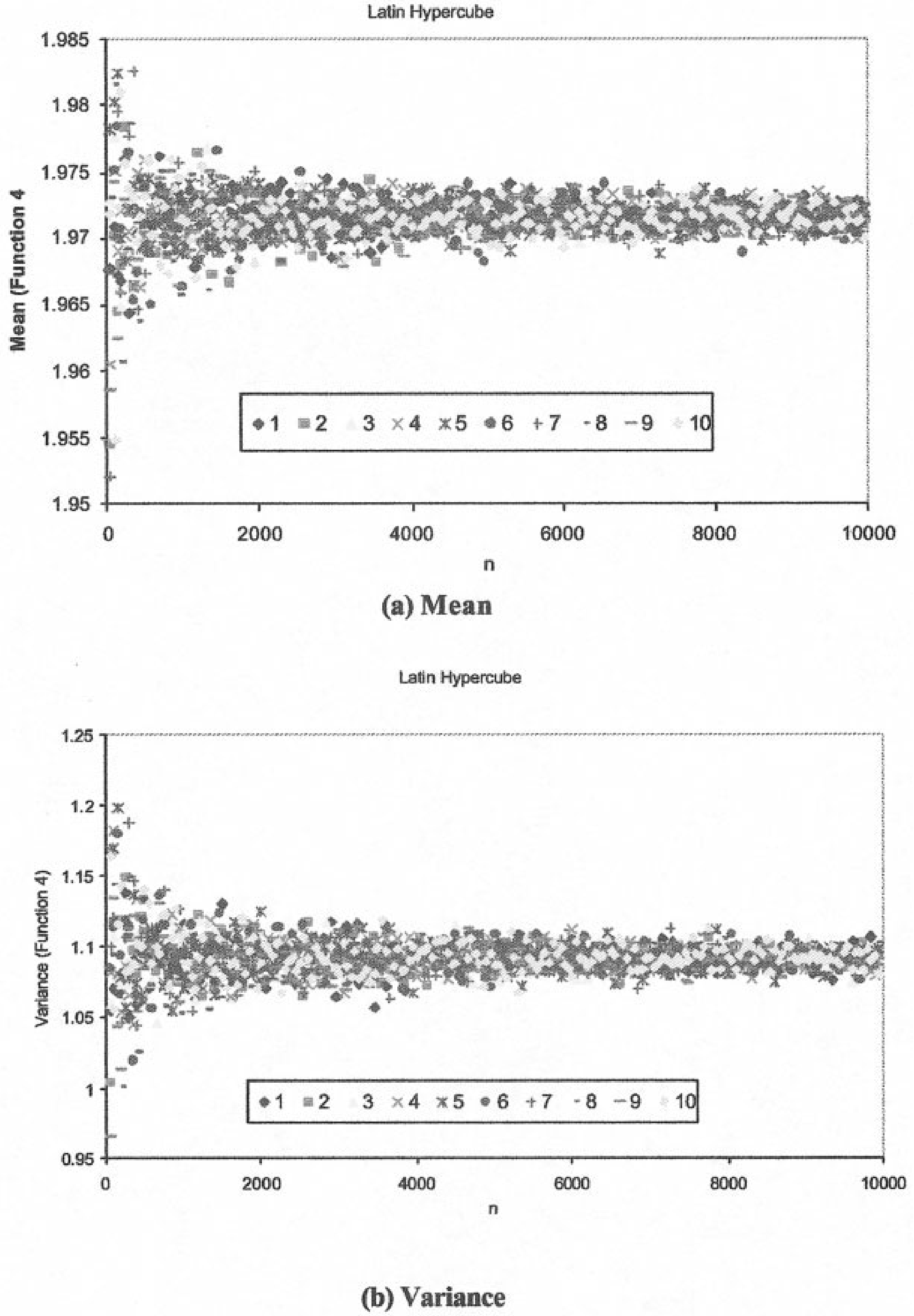
to have also inherited the good multidimensional uniformity property of HSS.

4.1. Experimental Tests

In this section, the performance of the LHHS technique is compared with that of the MCS, LHS, MLHS, and HSS techniques by propagating samples derived from each of the techniques for a set of k-input variables through various functions and measuring the number of samples required for the output error within a fixed error of the "true" mean and variance. By testing different functions and distributions with various correlation structures, the convergence property of LHHS can be tested and the efficiency of LHHS can be proved.

4.1.1. Framework of the Experimental Test

The comparison is performed by propagating samples derived from different techniques for a set of input variables $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k)$ through various functions $\tilde{Y} = f(\tilde{X})$ and measuring the number required to converge to the mean and variance of the derived distribution for Y. The design of the test includes varying type of functions, the number of input variables, type of input distributions, and the correlation structures between them. The number of input variables



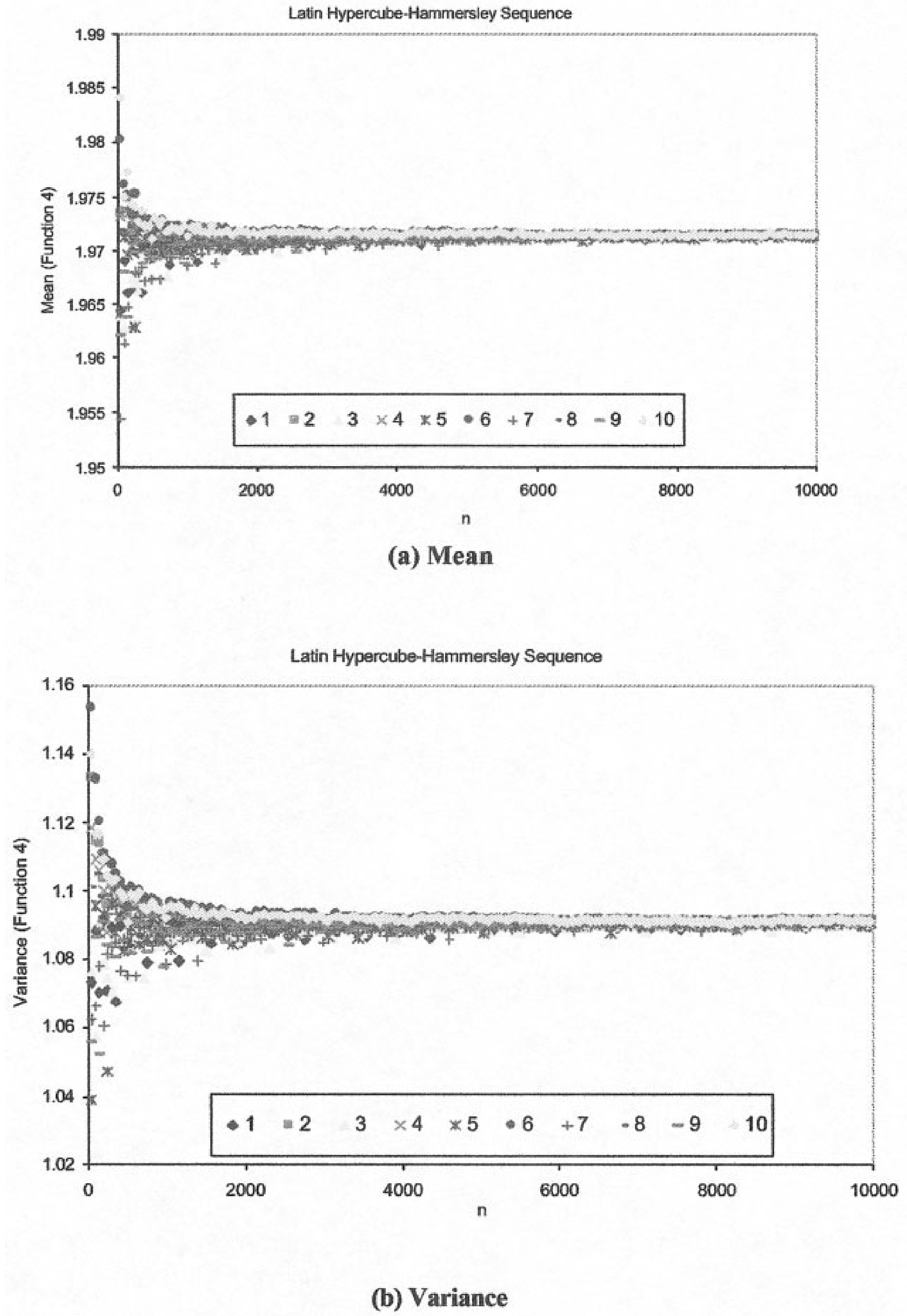
(a) Mean

(b) Variance

Figure 10. The mean and variance of output of function 4 as a function of sample size (n) for LHS with two input variables and 10 different initial random seeds. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

used was varied from 2 to 10. Seven different functions were used: linear additive function, multiplicative function, quadratic function, exponential function, logarithmic function, cosine function, and health effect risk-exposure function, which we discussed in Section 1. In addition, three correlation structures, including correlation coefficients 0.5, 0.9, and 0, and four types of distributions of inputs, uniform, normal, lognormal, and ā distributions, are tested. This testing matrix represents a total of 420 data sets (5 × 4 × 3 × 7) for each set of input variables. As we vary the variable number from 2 to 10, a total of 3780 data sets are generated for testing.

In experimental tests, we adopted the following decision rule: We use the analytical results for true mean and variance. If we cannot get the analytical results or



(a) Mean

(b) Variance

Figure 11. The mean and variance of output of function 4 as a function of sample size (n) for LHSS with two input variables and 10 different initial random seeds. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

the function is too complicated, the true mean and variance are estimated by propagating a very large number of samples using Monte Carlo. Once the true values have been found, performances of different sampling techniques are compared by estimating the number of samples needed to meet the required errors. It is graphically presented by plotting the calculated value of the mean and variance as a function of the number of samples in the calculation. Sampling schemes that add one point at a time (such as MCS and

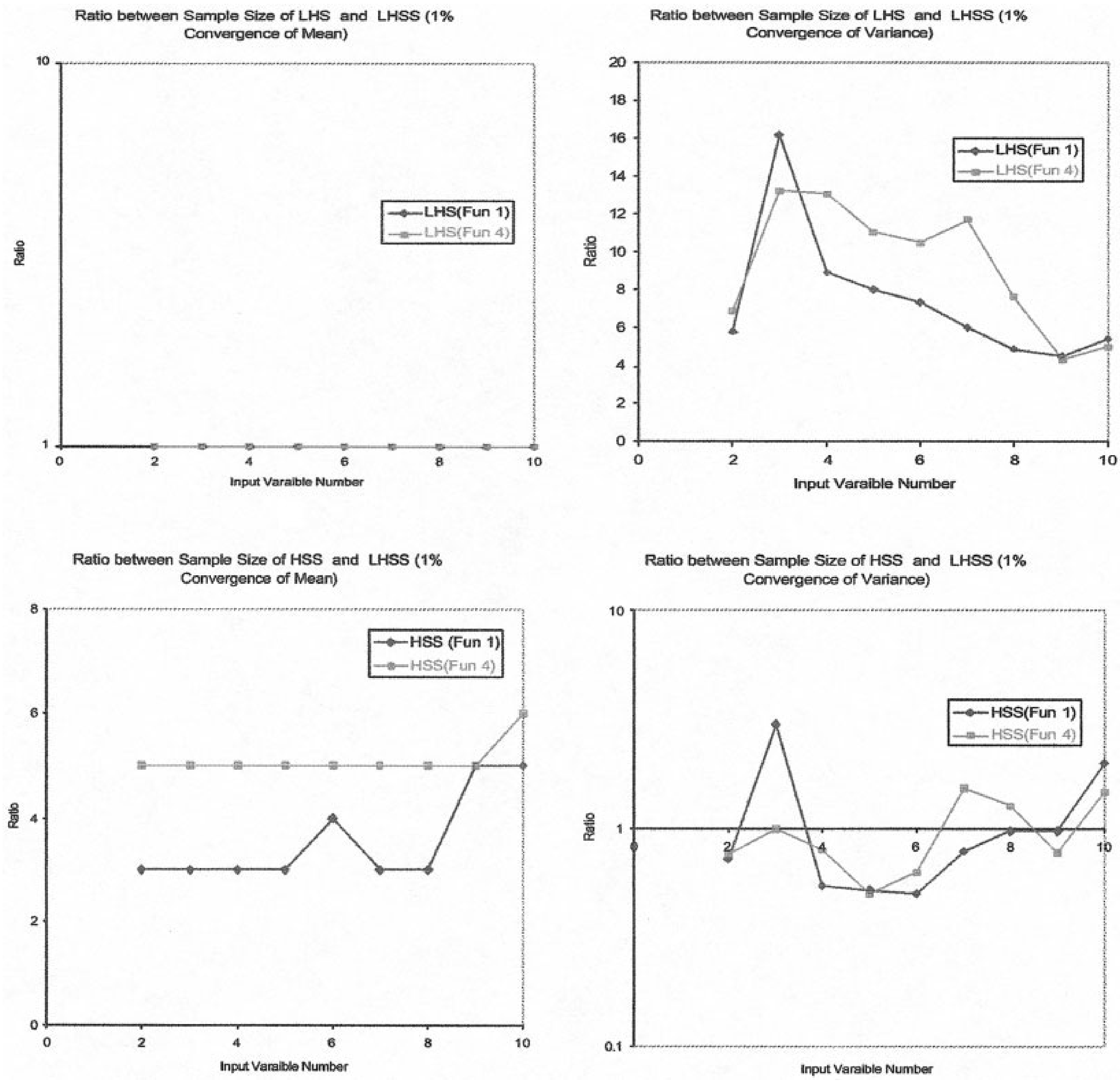


Figure 12. Ratio of sample size of LHS and HSS to LHHS versus different input variable numbers for 1% convergence of mean and variance of functions 1 and 4. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

HSS) typically have less fluctuation compared to schemes not retaining the original samples while generating the next one. In order to avoid the misleading of convergence and fluctuating, we choose the following test functions that are analytically simple for which true values can be calculated exactly.

1. Linear additive function: $Y = \sum_i X_i$
2. Multiplicative function: $Y = \Pi_i X_i$
3. Quadratic function: $Y = \sum X_i^2$
4. Exponential function: $Y = \sum_i [X_i \times \exp(X_{i+1})]$
5. Logarithmic function: $Y = \sum_i [X_i \times \log(X_{i+1})]$
6. Cosine function: $Y = \sum_i \cos X_i$
7. Health effect risk-exposure function:

$R = a + bX^c; \quad X > X_t$

$= 0; \quad X \leq X_t$

The last threshold function is added because it is one kind of typical function in risk analysis and also for practical use.

4.1.2. Experimental Test Results and Discussions

A summary result of experimental tests is presented in this section. Figure 7 shows the mean and variance of the output of function 4 as a function of sample size for MCS, LHS, HSS, and LHHS for two input variables.

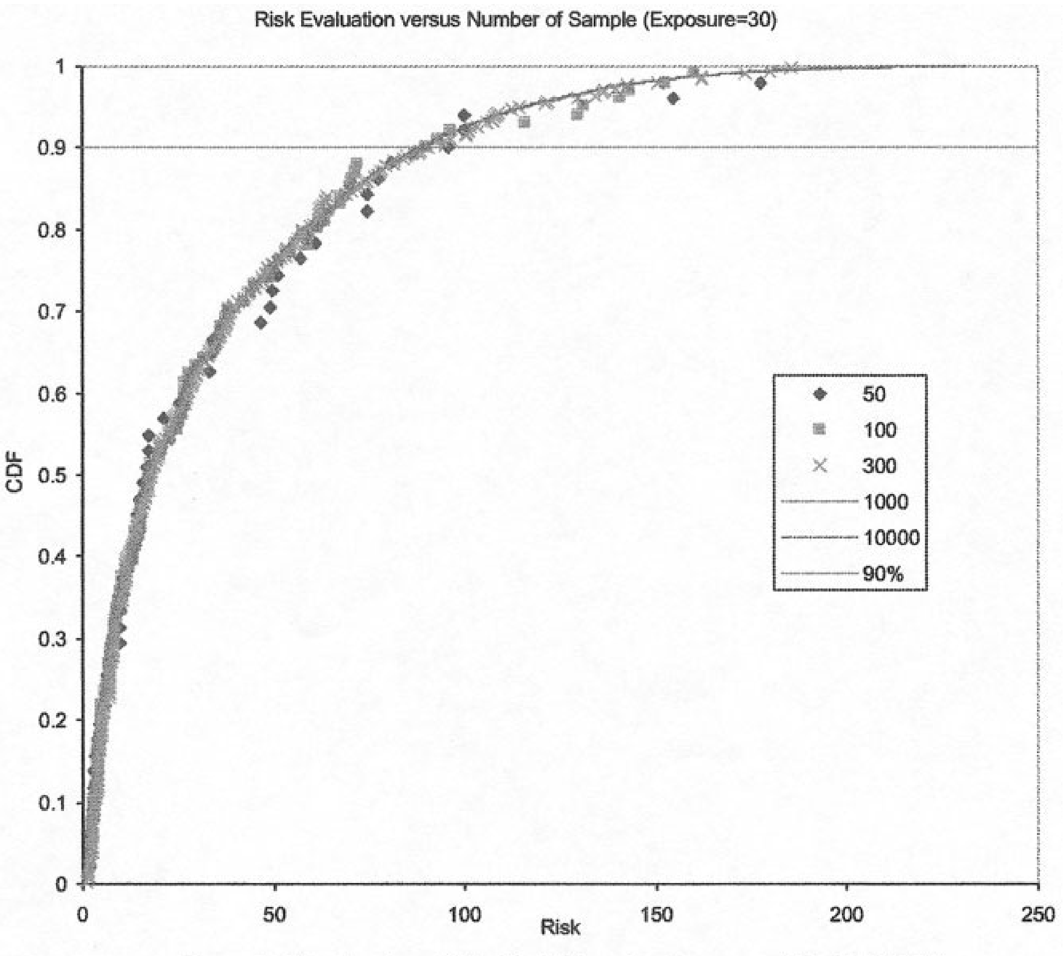


Figure 13 Evaluation of Health Risk under Exposure 30 with LHSS

Figure 13. Evaluation of health risk under exposure 30 with LHSS. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

Table 2. Risk value versus number of sample with probability of 90%.

Number of sample	50	100	200	300	500	1000	5000	10,000
Risk	94.2	89.8	92.0	90.2	90.6	88.6	87.5	87.9
Risk error (%)	7.16	2.11	4.60	2.61	3.10	0.78	0.48	0.01

Figure 8 shows the mean and variance of the output of functions 4 as functions of sample size for MCS, LHS, HSS, and LHHS for 10 input variables. In these graphs, the input variables are uncorrelated and uniformly distributed $U(0.1, 1)$. From these figures, if calculated with LHHS, the output distribution, either mean or variance, shows a better agreement with the true distribution. By running the propagation process many times, Figures 9-11 are graphed for mean and variance of function 1 with MCS, LHS, and LHHS for the random seed number being 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, respectively. The calculation with LHHS shows a lower error bar if compared with that of calculation with MCS and LHS sampling techniques. Therefore, LHHS performed better in estimating the mean and variance of the output. It has been found that LHHS inherits both merits of LHS and HSS while their shortcomings are avoided.

Figure 12 plots the ratio of the LHHS to the HSS, as well as LHS sample sizes as a function of the design parameters of the numerical experiments. The convergence sample sizes for these sampling schemes were obtained by 10 different sample sets. The sample size used for this comparison is the required samples to converge to within 1% of the actual value of mean and variance. Each subgraph plots the ratio of sample size against the number of input variables for functions 1 and 4. The results show that for mean LHHS and LHS perform similarly but LHHS is an order of magnitude better than LHS in calculation of variance. LHHS is better than HSS for calculation of mean but for variance

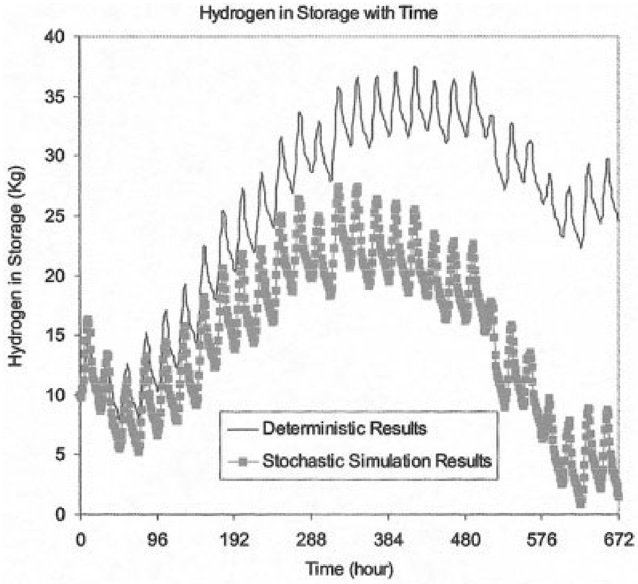


Figure 14. Hydrogen in storage versus time calculated with nominal values of random input variables and stochastic simulation results. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

calculation the performance varies depending on the number of variables and functionality. Results similar to the variance calculations were observed for fractile calculations. Therefore, it can be said that LHHS and HSS are the preferred sampling techniques for uncertainty analysis and choice of the final technique depends upon the problem at hand.

It should also be noted that the results presented here are qualitatively representative of the general trends observed for the data sets that have been analyzed with no exception to periodical functions. Therefore, the conclusions from these results can show the general performance of different sampling techniques.

4.2. Health-Effect Risk Model Revisited

In the Introduction, we discussed the health effect risk model calculation results with MCS. In this section, we present the results calculated with LHHS. Figure 13 shows the CDF curve of risk calculated using LHHS, and Table 2 tabulates the error percentage of risk values at CDF of 90% and calculated with different sample sizes. Figure 13 and Table 2 show that even with a sample size of 100 or 200, the predicting risk value will fall into the range within the convergence error of 5%. Therefore, the predicting efficiency has been improved by using LHHS instead of MCS.

5. RENEWABLE ENERGY POWER SYSTEM CASE STUDY

This case study is based on the renewable energy power system model from National Renewable Energy Laboratory (NREL). After evaluation of the power system for a small coastal village in Central America (approximately 100 homes) with limited access to the national power grid, a hydrogen energy system, photo voltaic (PV)-electrolysis/compressed gas/proton exchange membrane fuel cell (PEMFC) system, was selected as a minigrid because this area has high solar insolation [24, 25]. This comprehensive database of subsystem models can be linked and integrated using commercially available software (ASPEN Plus). In this integrated model, hourly average data for the solar resource based on a typical meteorological year and a projected demand profile for regions similar to the proposed site should be used. For illustrative purposes, the study period in this paper is 1 month, with a 2-week "good resource" period, followed by a 2-week "poor resource" period.

The original simulation model produces deterministic results for a particular set of input assumptions. The system was found adequate to meet the power demand. However, uncertainties and seasonal variabilities were not considered. Therefore, in our study, we should first add in stochastic blocks into the ASPEN simulator give it the capability to handle uncertainties systematically. After the stochastic capability has been built and the uncertainties involved in the model have been identified, it is then possible to do the simulation study.

5.1. Incorporation of Stochastic Models into ASPEN Plus

The approach adopted here involves adding a stochastic modeling capability for uncertainty analysis to ASPEN Plus 10.1 (Aspen Technology, Inc.). To implement the stochastic modeling capability, ASPEN's modular nature (consisting of unit operation modules or blocks) is utilized. This incorporation is based on the work of Diwekar and Rubin [26]. To accommodate the diverse nature of uncertainty, uniform, triangular, beta, normal, lognormal, and user-specified distributions are made available within the stochastic lock. Fifty uncertainties with 1500 cycles can be analyzed according to the current settings of the system. Five sampling techniques are all available for comparing the results [27]. In case we need more cycles and need to analyze more uncertainties, the settings of the library file can be easily changed to meet such requirements.

5.2. Identification of Uncertainties

The integrated system consists of three major subsystems: the PV-electrolysis subsystem to supply the electrical needs and produce hydrogen when excess power is available, the compressed gas subsystem for hydrogen storage, and the proton exchange membrane fuel cell (PEMFC) system for production of the electricity when the PV system cannot meet the power demand during the day, as well as to meet the demand at night. In this system, the two driving forces, solar insolation and electricity demand, are uncertain variables. Therefore, it is necessary to perform the uncertainty analysis. According to our study of the system, five main sources of uncertainties, solar insolation, electricity demand, temperature of the PV system, and temperature of cooling water, were identified. All data and information associated with the identified uncertainties can be obtained from the meteorological data

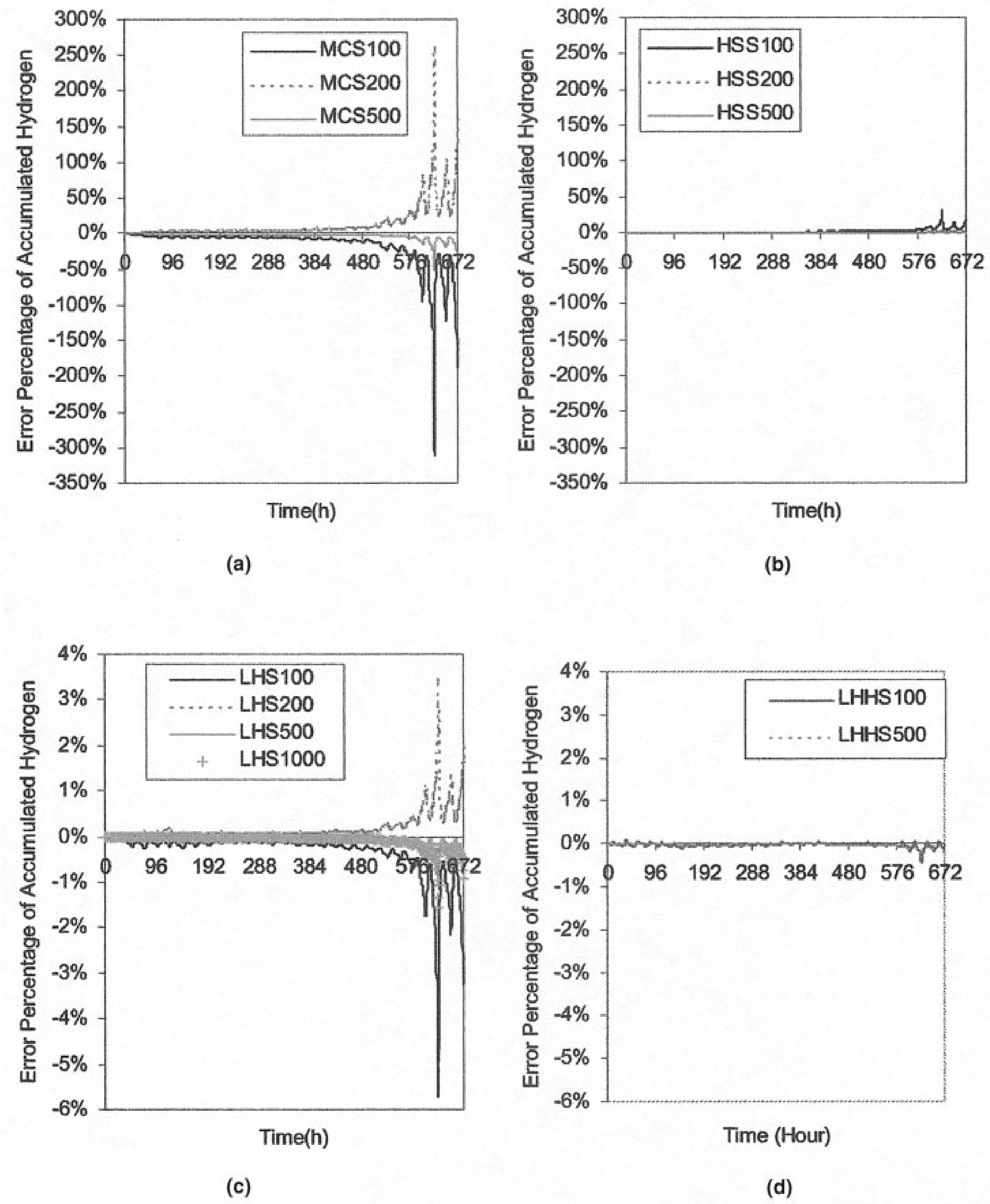


Figure 15. The percentage of predicting error of accumulated hydrogen in storage with MCS, LHS, HSS, and LHHS. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

of this village in past years. In this case study, all uncertainties are assumed to be uniformly distributed.

5.3. Results and Discussion

In the integrated system design, we are interested in whether the PV panels can meet the power demand. If the PV panels can meet the load, then the power demand for the fuel cell is set to zero and the fuel cell system solves for a zero power demand. If the PV panels cannot meet the load, the PV system power output is set to be the maximum that can be provided by it and then its power shortfall is calculated. If this shortfall is less than or equal to the maximum power for the fuel cell, the power demand for the fuel cell is set equal to this power shortfall. In the case when the fuel cell cannot meet the shortfall (which should be avoided in system design), the power demand for the fuel cell is set equal to the maximum power output of the fuel cell.

The net accumulated storage of the hydrogen, which is produced by the system mostly in the daytime and used in the nighttime, is the output we are most interested in. After a cycle, we can determine the technical feasibility by the value of hydrogen in storage. If it is too large, the design capacity is unnecessarily high and it can be too expensive (economically risky). But if it is negative, the design capacity cannot meet the power demand requirement. The peak of this value during the cycle can also determine the size of the storage system. Figure 14 shows the relationship of the storage of hydrogen with the time for the deterministic model calculated with nominal values for all random input variables and stochastic simulation results. The big difference between them shows the value of stochastic simulation (VSS). Figure 15 shows the results calculated with MCS, LHS, HSS, and LHHS. The error of calculation results with MCS will be accumulated up to 310% when the sample size is 100 and no big improvements were achieved with sample size 200 or 500. Compared with the calculation results with MCS, the calculation results with LHS were greatly improved but the error will also be accumulated even when the sample size is 1000. The simulation error can be large with HSS when the sample size is too small (100) but the error is not accumulated when the sample size is 200 or more. The encouraging results are those calculated with LHHS. The simulation error is within 0.5% with LHHS with a sample size of 500 or even 100. Therefore, in this renewable energy power system design application, the new sampling technique could be very important for avoiding muddling results.

6. CONCLUSIONS

In this paper, a new efficient sampling technique, Latin hypercube-Hammersley sequence sampling technique, for uncertainty in risk analysis is presented. This sampling technique is derived by coupling two well-known sampling techniques: LHS and HSS. LHHS has been tested for efficiency by using various functions, distributions, and correlation structures with different numbers of input uncertain variables. For all test cases, LHHS and HSS have been found to be consistently better for predicting overall performance than the existing sampling techniques: MCS, LHS, and MLHS. The renewable energy power system case study further confirmed the efficient performance of LHHS in handling large-scale system problems.

ACKNOWLEDGMENTS

The authors thank Dr. Victoria L. Putsche for her original modeling work for the case study presented in this paper. They are also indebted to the National Renewable Energy Laboratories for financial support for this project.

LITERATURE CITED

1. Voltaggio, T. (1999). Public communication and risk management for controversial projects, in managing environmental risks, Topical Conference Proceedings, New York: American Institute of Chemical Engineers.
2. Small, M.J., & Fischbeck, P.S. (1999). False precision in Bayesian updating with incomplete models, Human and Ecological Risk Assessment, 5, 291–304.
3. Kalagnanam, J., & Diwekar, U. (1997). An efficient sampling technique for off-line quality control, Technometrics, 39, 308–319.
4. Diwekar, U., & Kalagnanam, J. (1997). An efficient sampling technique for optimization under uncertainty, AIChE Journal, 43, 440–459.
5. James, B.A.P., (1985). Variance reduction techniques, Journal of Operational Research Society, 36, 525–530.
6. Knuth, D.E. (1998). The art of computer programming—Volume 1, Fundamental algorithms. 3rd ed., Reading, MA: Addison–Wesley Longman.
7. Morgan, M.G., & Henrion, M. (1990). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis, Cambridge, UK: Cambridge University Press.
8. McKay, M.D., & Beckman, R.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245.
9. Saliby, E. Descriptive sampling: a better approach to Monte Carlo simulations, Journal of Operations Research, 41, 1133–1142.
10. Hammersley, J.M. (1960). Monte Carlo methods for solving multivariable problems, Annals of the New York Academy of Science, 86, 844–874.
11. Wozniakowski, H. (1991). Average case complexity of multivariate integration, Bulletin of the American Mathematical Society, 24 (1), 185–194.
12. Halton, J.H. (1991). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, Numerische Mathematik, 2, 84–90.
13. Fox, B.L. (1986). ACM algorithm 647: implementation and relative efficiency of quasi-random sequence generators. ACM Transactions on Mathematical Software, 12, 362–376.
14. Sobol, I.M. (1979). On the systematic search in a hypercube, SIAM Journal of Numerical Analysis, 16, 790–793.
15. Bratly, P., & Fox, B.L. (1988). Algorithm 659: implementing Sobol's quasi random sequence generator, ACM Transactions on Mathematical Software, 14, 88–100.
16. Diwekar, U., & Kalagnanam, J. (1996). Robust design using an efficient sampling technique, Computers and Chemical Engineering, 20(Suppl.), S389–S394.
17. Diwekar, U. (2003). A novel sampling approach to combinatorial optimization under uncertainty, Computational Optimization and Applications, 24, 335–371.
18. Iman, R.L., & Conover, W.J. (1982). A distribution-free approach to inducing rank correlation among input variables, Communications on Statistics: Simulation and Computing, 11, 311–334.
19. Iman, R.L., & Helton, J.C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models, 8, Risk Analysis, 71–90.
20. Haas, C.N. (1999). On modeling correlated random

variables in risk assessment, Risk Analysis, 19, 1205–1214.

21. Anderson, T.W. (1958). An introduction to multivariate statistical analysis (pp. 19–20), New York: Wiley.
22. Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, Technometrics, 12, 591–612.
23. Marquardt, D.W., & Snee, R.D. (1975). Ridge regression in practice, The American Statistician, 29, 3–20.
24. Gregoire-Padro, C.E., & Putsche, V.L. (1998). Summary report for component model case studies, Filename: P01A. INP, National Renewable Energy Laboratory.
25. Gregoire-Padro, C.E., Putsche, V.L., & Fairlie, M.J. (1998). Modeling of hydrogen energy systems for remote applications, National Renewable Energy Laboratory.
26. Diwekar, U.M., & Rubin, E.S. (1991). Stochastic modeling of chemical processes Computers And Chemical Engineering, 15, 105–114.
27. Diwekar, U. (1999). User's manual for stochastic simulation in Aspen Plus, Prepared for U.S. Department of Energy, Carnegie Mellon University.