Journal of Environmental Informatics 28(1) 11-24 (2016)

Journal of Environmental Informatics

www.iseis.org/jei

An Intercomparison of Sampling Methods for Uncertainty Quantification of Environmental Dynamic Models

W. Gong*, Q. Y. Duan, J. D. Li, C. Wang, Z. H. Di, A. Z. Ye, C. Y. Miao, and Y. J. Dai

College of Global Change and Earth System Science (GCESS) and Joint Center for Global Change Studies, Beijing Normal University, Beijing, 100875, China

Received 26 September 2014; revised 8 April 2015; accepted 10 April 2015; published online 13 July 2015

ABSTRACT. Uncertainty quantification (UQ) of environmental dynamic models requires an efficient way to extract the information about the relationship between input parameter and model output. A uniformly scattered sample set is generally preferred over crude Monte Carlo sampling for its ability to explore the parameter space more effectively and efficiently. This paper compares eight commonly used uniform sampling methods along with the crude Monte Carlo sampling. The efficiency is measured by six uniformity metrics, while the effectiveness is measured by the goodness-of-fit of the surrogate models, and the sensitivity analysis and optimization results. We used two test problems: the Sobol' g-function and the SAC-SMA hydrological model. The results show that among the sampling methods evaluated, the Good Lattice Points (GLP) and Symmetric Latin hypercube (SLH) have the highest uniformity scores, and the Ranked Gram-Schmidt (RGS) de-correlation algorithm can further improve the uniformity of the lattice sample sets. On the other hand, the Quasi-Monte-Carlo (QMC) methods, such as Halton and Sobol' sequences, are not as uniform as their theoretical potential suggests when the number of sample points is low. Further, we found no clear relationship between the sampling methods used and their effectiveness, as the latter is affected by many factors other than the sampling methods, such as the choice of the surrogate modeling methods, sensitivity analysis and optimization methods, and the intrinsic properties of the dynamic models.

Keywords: design of experiment, quasi Monte-Carlo, sampling methods, symmetric Latin hypercube, uniform design

1. Introduction

Computer based environmental dynamic models are important tools for understanding and predicting the impacts of global and environmental changes due to natural or anthropogenic factors. There is a tendency that those models are becoming increasing more complex as they consider more and more physical, chemical and biological processes. As a result, today's models usually contain many model parameters and a large number of model outputs, and, in many cases, require many CPU hours to run. This makes proper parameter specification or model calibration a very difficult task. Furthermore, multi-physics dynamic models also demand multi-objective approach in model calibration (Vrugt et al., 2003; Liu et al., 2005).

There are numerous ways to deal with the unique challenges encountered in uncertainty quantification (UQ) of environmental dynamic models. Particularly, following techniques have been used: (1) sensitivity based parameter screening to reduce the number of parameters to be considered in model calibration; (2) a cheap surrogate model to mimic the response of a dynamic model to different parameter values; and (3) an adaptive resample strategy that wisely use the power of the surrogate model in parameter optimization. All of these techniques need initial sampling of the parameters, which is done by perturbing the adjustable parameters in a specified range and executing the dynamic model to obtain the simulation outputs. Initial sampling is an important step for extracting the information about the relationship between adjustable parameters and simulation outputs. In previous research, we have evaluated various sampling methods for its impact on parameter screening effectiveness (Li et al., 2013; Gan et al., 2014; Di et al., 2014). We also compared different sampling methods on how they impact on adaptive surrogate modeling based optimization (Wang et al., 2014; Gong et al., 2014). These studies emphasize the robustness of the sensitivity analysis and optimization results, while the efficiency aspect of the sampling methods, which is an important consideration for environmental dynamic models, has not been examined in depth.

A good sampling method should be able to explore parameter space more effectively and efficiently. The efficiency can be measured by the number of sample points nee-

[*] Corresponding author. Tel.: +86 10 58804191; fax: +86 10 58804191. *E-mail address:* gongwei2012@bnu.edu.cn (W. Gong).

ISSN: 1726-2135 print/1684-8799 online

doi: 10.3808/jei.201500310

ded to explore the parameter space thoroughly. In general, a uniform sampling is regarded as more efficient and effective than a crude Monte Carlo sampling. Numerous previous studies have examined the efficiency and effectiveness of different uniform sampling methods (Fang et al., 2002; Ye et al., 2000; Morokoff et al., 1995). Fang et al. (2002) derived the theoretical value of Centered L_2-discrepancy (a uniformity metric) of Latin Hypercube (LH) and compared it with crude-Monte-Carlo (MC), and also validated the theoretical results with numerical experiments. Ye et al. (2000) added symmetric property to the classical LH sampling and compared the uniformity of the proposed SLH with LH and MC samples. Fang et al. (1994) compared many numeric-theoretic methods for sampling, pointing out that the GLP method has the lowest discrepancy compared to the Halton sequence, Hammersley sequence, Haber sequence, Hua-Wang cyclotomic field method, among others. But those comparison studies were limited to 2-dimensional because computing the value of discrepancy was very difficult (Hickernell, 1998a and 1998b). Morokoff et al. (1995) made an inter-comparison of three QMC methods: Halton, Sobol' and Faure sequences, and clarified the advantages and weaknesses of these methods and made some suggestions for applications to particular problems. Halton sequence is best for low dimensional problems (i.e., with a dimension of less than 6), while Sobol' sequence is superior for higher dimensions, and Faure sequence falls behind them. Furthermore, Morokoff et al. (1995) also suggested that the QMC methods are suitable for smooth functions, but for less smooth functions the QMCs might be not better than crude-Monte-Carlo.

In this paper, various sampling methods are compared for their effectiveness and efficiency. Six uniformity metrics are used to measure sampling efficiency, and the effectiveness is evaluated by the goodness of fit of the surrogate models as well as surrogate-modeling based sensitivity analysis and optimization results. Two test problems: the Sobol' g-function and the SAC-SMA hydrological model are used for the evaluation. Following sampling methods are included in the inter-comparison: crude Monte Carlo (MC), Latin Hypercube (LH), Symmetric Latin Hypercube (SLH), Good Lattice Point (GLP), Halton sequence, and Sobol' sequence. A simple de-correlation method called the Ranked Gram-Schmidt (RGS) algorithm (Owen, 1994) is applied to LH, SLH and GLP sampling. The RGS algorithm is a post-processor that can remove internal correlation in the sample set. It is easy to use and fast to run, and, can significantly improve the uniformity of LH, SLH and GLP sampling.

This paper is organized as follows: Section 2 gives a brief introduction of the uniformity metrics and sampling methods involved in this paper; Section 3 introduces the background of modeling case studies; Section 4 presents the result and discussions; and Section 5 provides conclusions.

2. Methodology

2.1. Uniformity Metrics

There are many kinds of uniformity metrics, such as discrepancy (Weyl, 1916; Hickernell, 1998), integrated mean squared error (IMSE) (Sacks et al., 1989), entropy (Shewry et al., 1987) and maxmin or minimax distance (Johnson et al., 1990). These metrics describe different aspects of the representation ability of a sample set. In this paper, we calculated 6 different uniformity metrics, including the four discrepancy metrics proposed by Hickernell (1998a; 1998b), and the metric for maximum distance between the sample points and the metric for measuring correlation among the sample points.

The concept of discrepancy comes from the Monte-Carlo integration. For the point set $P_n = \{\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{ks}); k = 1, 2, \ldots, n\}$ in a unit hypercube C^S,the multidimensional integral $I(f) = \int_{C^s} f(\mathbf{x})d\mathbf{x}$ can be estimated using the average value of the uniformly distributed sample points P_n : $Q(f) = n^{-1}\sum_{\mathbf{x}\epsilon P_n} f(\mathbf{x})$. Obviously, the integral $I(f)$ can be estimated more accurately if the point set P_n is uniformly scattered. The error bound of the integral estimation can be expressed as a Koksma-Hlawka inequality (Kuipers et al., 1974):

$$\left|I(f) - Q(f)\right| \le D(P_n)V(f) \tag{1}$$

where $V(f)$ is the total variance of the model output, $f(\mathbf{x})$ and $D(P_n)$ is the discrepancy of point set P_n in the domain C^S. The Koksma-Hlawka inequality suggests that the upper bound of the integral error is controlled by two factors: the fluctuation of $f(\mathbf{x})$ and the model independent point uniformity, $D(P_n)$. Less discrepancy means better uniformity and lower integral error. The discrepancy is defined as the maximum deviation between the volume of a hypercube and the density of sample points falling in the cube (Hua and Wang, 1981; Niederreiter, 1992):

$$D(P_n) = \sup_{\mathbf{x}\in C^s}\left|\frac{\left|P_n \cap [0,\mathbf{x})\right|}{n} - Vol([0,\mathbf{x}))\right| \tag{2}$$

where $[0,\mathbf{x})$ represents the hypercube defined by the two diagonal points, $\left|P_n \cap [0,\mathbf{x})\right|$ represents the number of points falling in the domain $[0,\mathbf{x})$, $Vol([0,\mathbf{x}))$ represents the volume of the hypercube $[0,\mathbf{x})$. The concept of discrepancy was first suggested by Weyl (1916). It is also called star discrepancy. Similarly, we can define L_p-discrepancy as follows:

$$D_p(P_n) = \left[\int_{C^s}\left|\frac{\left|P_n \cap [0,\mathbf{x})\right|}{n} - Vol([0,\mathbf{x}))\right|^p d\mathbf{x}\right]^{1/p} \tag{3}$$

The star discrepancy is a special case when p → ∞. The concept of star discrepancy has played an important role in the theoretical analysis of developing quasi-Monte-Carlo methods. The theoretical orders of many quasi-Monte-Carlo methods have been derived. However, the star discrepancy has many drawbacks that prohibit its application (Fang et al., 2001). First, calculating discrepancy is very time consuming, especially when n and s are large. It is an NP hard problem.

Second, L_p-discrepancy is not sensitive when some points overlap. Third, because of domain [0, **x**), the origin is special in L_p-discrepancy. Last, L_p-discrepancy represents the overall uniformity in the s-dimensional space but omits the uniformity of the projection of P_n to low-dimensional space, e.g., the uniformity of 1D marginal and 2D joint distribution.

To solve the problem, Hickernell (1998a; b) developed a unified definition of discrepancy based on the concept of reproducing kernel of Hilbert space. With the unified framework we can easily develop new discrepancies, and find easier way to compute them. In this paper we consider four discrepancy metrics developed under this framework.

(1) Modified discrepancy. The modified L_p-discrepancy considers not only the uniformity in the s-dimensional space but also any lower dimensional spaces. The simplified equation to compute the modified discrepancy when p = 2 is as follows:

$$MD_2\left(P_n\right)=\left[\begin{array}{l}\left(\frac{4}{3}\right)^s-\frac{2^{1-s}}{n}\sum_{k=1}^{n}\prod_{i=1}^{s}\left(3-x_{ki}^2\right)+\\ \frac{1}{n^2}\sum_{k,l=1}^{n}\prod_{i=1}^{s}\left[2-\max\left(x_{ki},x_{li}\right)\right]\end{array}\right]^{1/2} \qquad (4)$$

(2) Centered discrepancy. In modified discrepancy, the origin is special compared to other points. The Centered L_2-discrepancy replaced the origin with the nearest vertex of the [0, 1] hypercube. The simplified equation for computing the Centered L_2-discrepancy is listed below:

$$CD_2\left(P_n\right)=$$
$$\left[\left(\frac{13}{12}\right)^s-\frac{2^{1-s}}{n}\sum_{k=1}^{n}\prod_{i=1}^{s}\left(2+\left|x_{ki}-\frac{1}{2}\right|-\left|x_{ki}-\frac{1}{2}\right|^2\right)\right. \qquad (5)$$
$$\left.+\frac{1}{n^2}\sum_{k,l=1}^{n}\prod_{i=1}^{s}\left[1+\frac{1}{2}\left|x_{ki}-\frac{1}{2}\right|+\frac{1}{2}\left|x_{li}-\frac{1}{2}\right|-\frac{1}{2}\left|x_{ki}-x_{li}\right|\right]\right]^{1/2}$$

(3) Symmetric discrepancy. This considers the symmetric property of even and odd vertexes:

$$SD_2\left(P_n\right)=\left[\begin{array}{l}\left(\frac{4}{3}\right)^s-\frac{2}{n}\sum_{k=1}^{n}\prod_{i=1}^{s}\left(1+2x_{ki}-2x_{ki}^2\right)\\ +\frac{2^s}{n^2}\sum_{k,l=1}^{n}\prod_{i=1}^{s}\left[1-\left|x_{ki}-x_{li}\right|\right]\end{array}\right]^{1/2} \qquad (6)$$

(4) Wrap-around discrepancy. This connects the 0 and 1 margins end to end:

$$WD_2\left(P_n\right)=\left[-\left(\frac{4}{3}\right)^s+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left[\frac{3}{2}-\right.\right. \qquad (7)$$
$$\left.\left.\left|x_{ki}-x_{ji}\right|\left(1-\left|x_{ki}-x_{ji}\right|\right)\right]\right]^{1/2}$$

In addition to the four discrepancy metrics, another two uniformity measures are also considered. One is the minimum distance between sample points. The points are less uniformly distributed if some points are clustered together, or even overlap. A larger minimum distance means better uniformity. The other metric is the sum of correlation coefficients between each dimension of sample points. Linear correlated sample points have less uniformity, and the correlation coefficient is very sensitive in this case.

2.2. Sampling Methods Evaluated

In this paper, we compared nine different sampling methods: the crude Monte Carlo (MC), Latin Hypercube (LH), Latin Hypercube with de-correlation (LH-dc), Symmetric Latin Hypercube (SLH), Symmetric Latin Hypercube with de-correlation (SLH-dc), Good Lattice Points (GLP), Good Lattice Points with de-correlation (GLP-dc), Halton low discrepancy sequence (Halton), and Sobol’ low discrepancy sequence (Sobol’).

The Latin hypercube (LH) design is a type of stratified lattice design proposed by McKay et al. (1979). Suppose that there are n sample points, the number of factors is s, and the number of levels of each factor is q, then the sampling matrix can be expressed as a U-array: $U_n(q^s) = [u_{ij}]_{n\times s}$, in which each column is a permutation of {1, …, q} and u_{ij} is the level of the j-th factor in the i-th combination. Because of its simplicity, randomness and uniformity, LH sampling has been wildly used in UQ of environmental models. Latin hypercube design is one kind of U-type design (Fang et al., 2006), namely balanced design (Li et al., 1997), or lattice design (Bates et al., 1996), that each factor has n possible value to take: {1, …, q} or $(2i - 1)/2q$ {i = 1, …, q}. An U-array can be transferred to uniform distribution U(0, 1) with the following equation:

$$x_{ij}=\frac{2u_{ij}-1}{2q}, \qquad i=1,\ \ldots,\ n;\ j=1,\ \ldots,\ s \qquad (8)$$

where x_{ij} is the value of the i-th sample with j-th dimension. The set $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{is})$ is also called induced design (Fang et al., 2006). Fang et al. (2002) derived the theoretic expectation and variance of Centered L_2-discrepancy for crude-Monte-Carlo and Latin Hypercube sampling. The average square of centered L_2-discrepancy of MC sampling $R_{n,s}$ is:

$$E\left(CD_2\left(R_{n,s}\right)^2\right)=\left[\left(\frac{5}{4}\right)^s-\left(\frac{13}{12}\right)^s\right]/n \qquad (9)$$

while that of LH sampling L_{n,q^s} is:

$$E\left(CD_2\left(L_{n,q^s}\right)^2\right) = \begin{cases} \left(\frac{13}{12}\right)^s - 2\left(\frac{13}{12} - \frac{1}{12q^2}\right)^s + \frac{1}{n}\left(\frac{5}{4} - \frac{1}{4q^2}\right)^s \\ + \left(1 - \frac{1}{n}\right)\left(\frac{13}{12} + \frac{n-q^2}{6q^2(n-1)} - \frac{1}{4q^2}\right)^s, \quad q \text{ odd} \\ \left(\frac{13}{12}\right)^s - 2\left(\frac{13}{12} + \frac{1}{24q^2}\right)^s + \frac{1}{n}\left(\frac{5}{4}\right)^s \\ + \left(1 - \frac{1}{n}\right)\left(\frac{13}{12} + \frac{n-q^2}{6q^2(n-1)}\right)^s, \quad q \text{ even} \end{cases} \tag{10}$$

Consequently, although the orders of discrepancy of both MC and LH are $O(n^{-1/2})$, the LH sampling is generally more uniform since the average squared CL_2-discrepancy of LH sampling is significantly lower than that of MC.

Based on the framework of Latin Hypercube, Ye et al. (2000) proposed the Symmetric Latin Hypercube (SLH) design that for every point in the design, the reflection of it through the center is also in the design. In other words, for an n-point, n-level, s-dimension SLH, if $(a_1, a_2, \ldots, a_s)$ is one row of the $n \times s$ design matrix, $(n + 1 - a_1, n + 1 - a_2, \ldots, n + 1 - a_s)$ must be another row of the matrix. Although the theoretical expression of SLH design's discrepancy has not been derived, according to the symmetric property and application experiences, SLH is more uniform than the classical LH design.

Another class of sampling methods is quasi-Monte-Carlo method (QMC), or the so-called number-theoretic method (NTM), which is named after the theoretical foundation of these methods. In this paper we involved three of them: GLP method, Halton sequence and Sobol' sequence.

The Good Lattice Point method was originally proposed by Korobov (1959a;b) in USSR and discussed by Fang (1980), Wang et al. (1981), Hua et al. (1981), Sloan (1985), Shaw (1988), Fang et al. (1994) and Fang et al. (2006). The GLP design is generated by the following equations:

$$\begin{cases} q_{ki} = kh_i \left(\bmod\ n\right) \\ x_{ki} = \left(2q_{ki} - 1\right)/n \end{cases}, \quad k = 1,\ \ldots,\ n; \quad i = 1,\ \ldots,\ s \tag{11}$$

where $h_i < n$ and the greatest common divisor of h_i and n is 1. The vector $(n$: $h_1, \ldots, h_s)$ is called the Generating Vector. If the point set $P_n = \{\mathbf{x}_k = (x_{k1}, \ldots, x_{ks}),\ k = 1, \ldots, n\}$ has the lowest discrepancy among all possible generating vectors, the point set P_n is called GLP set. If the number of sample points/levels n is large, the number of possible combinations of generating vectors might be very large and consume a lot of computational resources. To mitigate this problem, Korobov (1959b) suggested to use the Powered Generating Vector: $(n$: $h_1, \ldots, h_s) = (a^0, a^1, \ldots, a^{s-1})$ (mod n), where a satisfies: (1) $1 < a < n$; (2) the greatest common divisor of a and n is 1; (3) $h_1, \ldots, h_s$ are different to each other; (4) $a^{t+1} = 1$(mod n), where $t \geq s - 1$ The powered generating vector is preferred if n is very large. For a given prime number p, the order of discrepancy of GLP set generated from the prime generating vector is:

$$D(p) < c(s)\, p^{-1}\left(\log p\right)^s \tag{12}$$

while that of GLP set generated from the powered generating vector is:

$$D(p) \leq c(s)\, p^{-1}\left(\log p\right)^s \log\log p \tag{13}$$

Consequently, the discrepancy of GLP set generated by the prime generating vector is lower than that generated by the powered generating vector, but the difference is negligible because the term $\log\log p$ is relatively not large. The powered generating vector dramatically reduce the amount of computational resources and is suitable for large problems. The readers can refer to the section 1.3.1 of (Fang et al., 1994) for more information about GLP method.

Another QMC method is Halton sequence (Halton, 1964). The Halton sequence is a generalization of the Van der Corput sequence in high dimensional cases (Niederreiter, 1992; Caflisch, 1998). For the one-dimension case ($s = 1$), the n-th element of the van der Corput sequence is generated as follows:

$$\begin{aligned} n &= a_m a_{m-1} \ldots a_1 a_0 \left(base\ 2\right) \\ x_n &= 0.a_0 a_1 \ldots a_{m-1} a_m \left(base\ 2\right) \end{aligned} \tag{14}$$

where the number n is written in binary (base 2) and the n-th point x_n is the revision of that around a decimal point. This manipulation is called "radical inverse". Generally, the n-th element $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{ns})$ of the s-dimensional Halton sequence can be generated like this: for the i-th dimension, n is expanded in base p_i (the i-th prime number), and x_{p_i} equals to its radical inverse. Halton proved that the discrepancy of the first n points of Halton sequence is:

$$D(n) = O\left(n^{-1}\left(\log n\right)^s\right) \tag{15}$$

The Sobol' QMC sequence proposed by Russian mathematician Sobol' (1967), is also based on radical inverse. In short, each dimension of an s-dimensional Sobol' sequence is a permutation of van der Corput sequence with base 2. If proper permutations are adopted, the Sobol' sequence can be more uniform than the Halton sequence. A comprehensive introduction of Sobol' sequence in English can be found in

Table 1. Parameters of the SAC-SMA Model and their Feasible Ranges and Assigned Values (Wang et al., 2014)

No.	Parameter	Lower bound	Upper bound	Assigned value
1	UZTWM	10.00	300.00	242.868
2	UZFWM	5.00	150.00	49.5779
3	UZK	0.10	0.75	0.4373
4	PCTIM	0.00	0.10	0.011
5	ADIMP	0.00	0.20	0.063
6	ZPERC	5.00	350.00	97.7848
7	REXP	1.00	5.00	1.8564
8	LZTWM	10.00	500.00	325.192
9	LZFSM	5.00	400.00	353.817
10	LZFPM	10.00	1000.00	61.679
11	LZSK	0.01	0.35	0.1092
12	LZPK	0.001	0.05	0.0131
13	PFREE	0.00	0.80	0.262

* The three fixed parameter values are: RSERV = 0.3; RIVA = 0.0; SIDE = 0.0.

(Bratley et al., 1988). In Sobol' (1967), both of Halton and Sobol' sequence was unified as LP_τ-sequence (see definition 3.8 in Niederreiter, 1978). A more general QMC framework called (*t*, *s*)-nets was proposed by (Niederreiter, 1992) that the properties and theories of these sequence can be summarized in a common framework. The discrepancy of a (*t*, *s*)-net satisfies:

$$D(n) \le C_s \frac{(\log n)^s}{n} + O\left(n^{-1}(\log n)^{s-1}\right) \tag{16}$$

where C_s is a constant depending on the kind of sequence.

We also evaluated the effect of de-correlation sampling post-processing methods. Intuitively speaking, each dimension of a uniformly scattered *s*-dimensional sample set should be independent and the correlation between every two dimensions should be zero. Iman et al. (1982) proposed a ‘ranked Cholesky’ (RC) method that can generate a Latin hypercube sample with user-defined correlation. If an identity matrix is assigned, it can generate a de-correlated Latin hypercube sample set, in which the correlations between each dimension are zero. The Ranked Gram-Schmidt (RGS) algorithm proposed by Owen (1994) is a method that can orthogonalize a vector set in an inner product space and minimize the correlation between each dimension. As shown in (Owen, 1994), a numerical experiment has shown that RGS is more successful than RC at reducing correlations. Because Owen’s RGS method use ranked correlation, it is only applicable for the lattice designs, which are, or can be transferred to an $n \times s$ matrix in which each column is a permutation of 1, 2, …, *n*. RGS decorrelation is not applicable for Halton and Sobol’ sequence because they are not lattice designs and RGS may destroy their space structure.

There are numerous alternative uniform sampling methods (and experimental designs) not considered in this paper. Among them, full factorial design is not considered because it is not suitable for high-dimensional and high-level problems. Similarly Orthogonal Arrays (OA) (Owen, 1992) and the Orthogonal Array Latin Hypercube design (OALH) (Tang, 1993) are not evaluated because it is hard to construct orthogonal arrays for high-dimensional and high-level problems. On the other hand, LH, SLH and GLP methods can provide uniform sample sets that sacrifice the orthogonal property but do not have restrictions on parameter dimensions and levels. There are some techniques which allows the search for the most uniform sample set using optimization methods, such as the Threshold-Accepting method (Fang et al., 2000; Fang et al., 2002), simulated annealing (Morris et al., 1995), and a columnwise-pairwise exchange algorithm (Ye et al., 2000). However, those optimization-based sampling methods may require too many CPU hours to run. Therefore, we do not consider them in this study.

3. Test Problems and Numerical Experimental Setup

The objective of this research is to evaluate the effecttiveness and efficiency of different sampling methods. Two test problems are used for this purpose: the Sobol’ g-function and the Sacramento Soil Moisture Accounting (SAC-SMA) model. They are described as follows.

3.1. Sobol’ g-function

As described in (Sobol', 1993), the Sobol’ g-function is a benchmark test function for sensitivity analysis. It has the following advantages over a real dynamic model: (1) It is fast to run. The Sobol’ g-function has a very simple expression and runs much faster than a dynamic model. (2) It is flexible. The Sobol’ g-function can be extended to any dimensions, and the shape of it can be adjusted by tuning the shape parameter a_i. (3) For sensitivity analysis, the true value of the main (first order) effect and total effect of the Sobol’ g-function can be analytically computed in a very easy manner. Thus, Sobol’ g-function can be used as a standard for evaluating the effecttiveness of sensitivity analysis. The Sobol’ g-function is defined as follows:

$$Y = \prod_{i=1}^{s} g_i\left(X_i\right); \;\; g_i\left(X_i\right) = \frac{\left|4X_i - 2\right| + a_i}{1 + a_i} \tag{17}$$

where X_i are the input factors and the shape parameters $a_i \geq 0$. The conditional variance of factor X_i can be computed using the following equation (Saltelli et al., 2008):

$$V_i = V\left[E\left(Y \mid X_i\right)\right] = \frac{1}{3\left(1 + a_i\right)^2} \tag{18}$$

The main effect is defined as the normalized conditional variance $S_i = V_i/V(Y)$, where $V(Y)$ is the total variance of Y. The higher-order partial variances can be computed by multiplying the lower ones, i.e., $V_{12} = V_1V_2$, so the total variance can be computed as follows:

$$V_{\sim i} = V_i + \sum_{j \neq i}^{n} V_i V_j + \sum_{j,k \neq i}^{n} V_i V_j V_k + \cdots \tag{19}$$

The total effect is defined as $S_{\sim i} = V_{\sim i}/V(Y)$. In this paper, we only show the results of the main effect. The value of shape parameter a_i are as follows: for the 13-parameter case (mimic SAC-SMA), a_i ={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 99, 99, 99}; for the 23-parameter case (mimic WRF), a_i ={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 99, 99, 99}; and for the 40-parameter case (mimic CoLM), a_i ={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99}.

3.2. SAC-SMA Hydrological Model

In previous research by Wang et al. (2014), we have tested the influence of the number of initial sampling points, and compared 2 kinds of quasi-Monte-Carlo sampling method: Halton and Sobol' sequences. In this paper, we use the same experiment setup to test other sampling methods. The observed streamflow was recreated by running the SAC-SMA model with the true observed forcing (precipitation and potential evapotranspiration), and streamflow simulation using the assigned parameters as observations. So that if the optimization algorithm can effectively find the true optimal parameters, the RMSE of streamflow will be close to zero. The data used are from the Leaf River basin near Collins, Mississippi, USA. We use 10 years (Oct, 1948 to Sep, 1958) daily data (mean area precipitation (mm/day), potential evapotranspiration (mm/ day), streamflow (m^3/s)) provided by the U.S. National Weather Service for analysis, with the first 365 days used as the warm-up period. The information about parameters (i.e., names, feasible ranges and assigned values) is presented in Table 1.

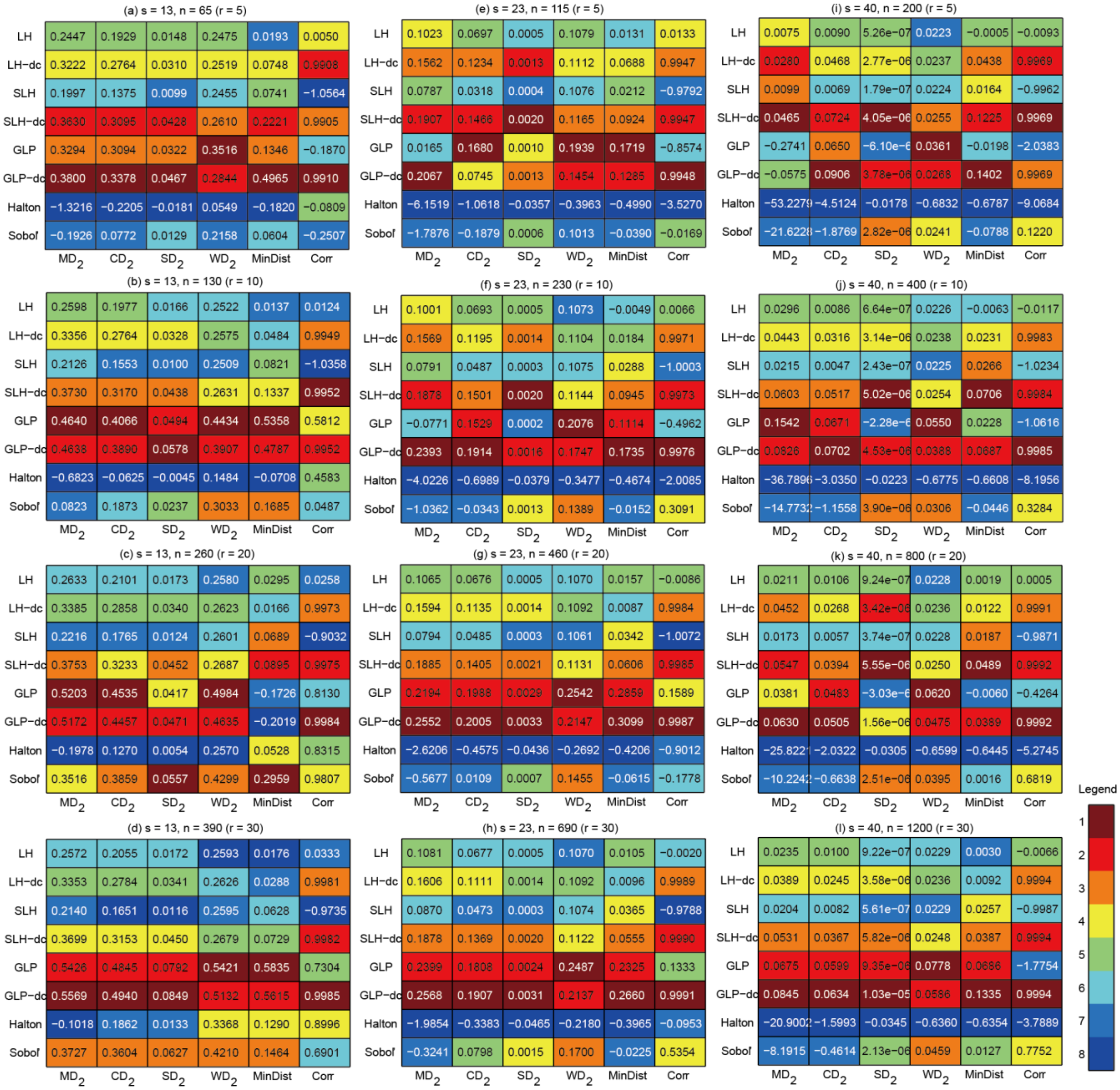
A number of numerical experiments are conducted to evaluate efficiency and effectiveness. Specifically, four sets of evaluations are carried out and these are described below:

(1) The calculation of the uniformity metrics. We first generate parameter samples using the nine sampling methods. For each sampling method, 6 uniformity metrics are computed: MD_2, CD_2, SD_2, WD_2, MinDist, and Corr. For dimension s = 13, 23, 40, we use the number r to set the sample size so $n = s \times r$. For example, for s = 13, the number r = 5, 10, 20, 30 and the sample size n = 65, 130, 260, 390, respectively. To compare the uniformity of each sampling method against the crude Monte Carlo, we define the normalized uniformity metrics as shown in Table 2, where P_n is the evaluated sample set and MC is a crude Monte Carlo sample set with the same dimension and sample size. A larger NM means the sample set is more uniform, whereas NM < 0 means the sample set is not as uniform as MC.

Table 2. Normalization of Uniformity Metrics

Original uniformity metrics	Normalized uniformity metrics
MD_2	$\mathrm{NM}(P_n) = 1 - \frac{\mathrm{MD}_2(P_n)}{\mathrm{MD}_2(\mathrm{MC})}$
CD_2	$\mathrm{NM}(P_n) = 1 - \frac{\mathrm{CD}_2(P_n)}{\mathrm{CD}_2(\mathrm{MC})}$
SD_2	$\mathrm{NM}(P_n) = 1 - \frac{\mathrm{SD}_2(P_n)}{\mathrm{SD}_2(\mathrm{MC})}$
WD_2	$\mathrm{NM}(P_n) = 1 - \frac{\mathrm{WD}_2(P_n)}{\mathrm{WD}_2(\mathrm{MC})}$
MinDist	$\mathrm{NM}(P_n) = \frac{\mathrm{MinDist}(P_n)}{\mathrm{MinDist}(\mathrm{MC})} - 1$
Corr	$\mathrm{NM}(P_n) = 1 - \frac{\mathrm{Corr}(P_n)}{\mathrm{Corr}(\mathrm{MC})}$

(2) The evaluation of the effect of sampling methods on surrogate modeling. To evaluate the effectiveness of the sampling methods for surrogate modeling, we considered two different surrogate models (i.e., MARS and GPR). MARS (Multivariate Adaptive Regression Spline) proposed by Friedman (1991) is a regression model for nonlinear, high-dimensional data. It can also be used as a surrogate model (Crino et al., 2007) as well as a sensitivity analysis method (Shahsavani et al., 2010). GPR (Gaussian Processes Regression) is also a flexible nonlinear regression method. An intuitive introduction to GPR was presented by Rasmussen et al. (2006). Gong et al. (2014) and Wang et al. (2014) have shown that GPR has the best goodness-of-fit compared to other surrogate models including MARS.



(a) s = 13, n = 65 (r = 5)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.2447	0.1929	0.0148	0.2475	0.0193	0.0050
LH−dc	0.3222	0.2764	0.0310	0.2519	0.0748	0.9908
SLH	0.1997	0.1375	0.0099	0.2455	0.0741	−1.0564
SLH−dc	0.3630	0.3095	0.0428	0.2610	0.2221	0.9905
GLP	0.3294	0.3094	0.0322	0.3516	0.1346	−0.1870
GLP−dc	0.3800	0.3378	0.0467	0.2844	0.4965	0.9910
Halton	−1.3216	−0.2205	−0.0181	0.0549	−0.1820	−0.0809
Sobol'	−0.1926	0.0772	0.0129	0.2158	0.0604	−0.2507

(b) s = 13, n = 130 (r = 10)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.2598	0.1977	0.0166	0.2522	0.0137	0.0124
LH−dc	0.3356	0.2764	0.0328	0.2575	0.0484	0.9949
SLH	0.2126	0.1553	0.0100	0.2509	0.0821	−1.0358
SLH−dc	0.3730	0.3170	0.0438	0.2631	0.1337	0.9952
GLP	0.4640	0.4066	0.0494	0.4434	0.5358	0.5812
GLP−dc	0.4638	0.3890	0.0578	0.3907	0.4787	0.9952
Halton	−0.6823	−0.0625	−0.0045	0.1484	−0.0708	0.4583
Sobol'	0.0823	0.1873	0.0237	0.3033	0.1685	0.0487

(c) s = 13, n = 260 (r = 20)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.2633	0.2101	0.0173	0.2580	0.0295	0.0258
LH−dc	0.3385	0.2858	0.0340	0.2623	0.0166	0.9973
SLH	0.2216	0.1765	0.0124	0.2601	0.0689	−0.9032
SLH−dc	0.3753	0.3233	0.0452	0.2687	0.0895	0.9975
GLP	0.5203	0.4535	0.0417	0.4984	−0.1726	0.8130
GLP−dc	0.5172	0.4457	0.0471	0.4635	−0.2019	0.9984
Halton	−0.1978	0.1270	0.0054	0.2570	0.0528	0.8315
Sobol'	0.3516	0.3859	0.0557	0.4299	0.2959	0.9807

(d) s = 13, n = 390 (r = 30)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.2572	0.2055	0.0172	0.2593	0.0176	0.0333
LH−dc	0.3353	0.2784	0.0341	0.2626	0.0288	0.9981
SLH	0.2140	0.1651	0.0116	0.2595	0.0628	−0.9735
SLH−dc	0.3699	0.3153	0.0450	0.2679	0.0729	0.9982
GLP	0.5426	0.4845	0.0792	0.5421	0.5835	0.7304
GLP−dc	0.5569	0.4940	0.0849	0.5132	0.5615	0.9985
Halton	−0.1018	0.1862	0.0133	0.3368	0.1290	0.8996
Sobol'	0.3727	0.3604	0.0627	0.4210	0.1464	0.6901

(e) s = 23, n = 115 (r = 5)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.1023	0.0697	0.0005	0.1079	0.0131	0.0133
LH−dc	0.1562	0.1234	0.0013	0.1112	0.0688	0.9947
SLH	0.0787	0.0318	0.0004	0.1076	0.0212	−0.9792
SLH−dc	0.1907	0.1466	0.0020	0.1165	0.0924	0.9947
GLP	0.0165	0.1680	0.0010	0.1939	0.1719	−0.8574
GLP−dc	0.2067	0.0745	0.0013	0.1454	0.1285	0.9948
Halton	−6.1519	−1.0618	−0.0357	−0.3963	−0.4990	−3.5270
Sobol'	−1.7876	−0.1879	0.0006	0.1013	−0.0390	−0.0169

(f) s = 23, n = 230 (r = 10)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.1001	0.0693	0.0005	0.1073	−0.0049	0.0066
LH−dc	0.1569	0.1195	0.0014	0.1104	0.0184	0.9971
SLH	0.0791	0.0487	0.0003	0.1075	0.0288	−1.0003
SLH−dc	0.1878	0.1501	0.0020	0.1144	0.0945	0.9973
GLP	−0.0771	0.1529	0.0002	0.2076	0.1114	−0.4962
GLP−dc	0.2393	0.1914	0.0016	0.1747	0.1735	0.9976
Halton	−4.0226	−0.6989	−0.0379	−0.3477	−0.4674	−2.0085
Sobol'	−1.0362	−0.0343	0.0013	0.1389	−0.0152	0.3091

(g) s = 23, n = 460 (r = 20)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.1065	0.0676	0.0005	0.1070	0.0157	−0.0086
LH−dc	0.1594	0.1135	0.0014	0.1092	0.0087	0.9984
SLH	0.0794	0.0485	0.0003	0.1061	0.0342	−1.0072
SLH−dc	0.1885	0.1405	0.0021	0.1131	0.0606	0.9985
GLP	0.2194	0.1988	0.0029	0.2542	0.2859	0.1589
GLP−dc	0.2552	0.2005	0.0033	0.2147	0.3099	0.9987
Halton	−2.6206	−0.4575	−0.0436	−0.2692	−0.4206	−0.9012
Sobol'	−0.5677	0.0109	0.0007	0.1455	−0.0615	−0.1778

(h) s = 23, n = 690 (r = 30)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.1081	0.0677	0.0005	0.1070	0.0105	−0.0020
LH−dc	0.1606	0.1111	0.0014	0.1092	0.0096	0.9989
SLH	0.0870	0.0473	0.0003	0.1074	0.0365	−0.9788
SLH−dc	0.1878	0.1369	0.0020	0.1122	0.0555	0.9990
GLP	0.2399	0.1808	0.0024	0.2487	0.2325	0.1333
GLP−dc	0.2568	0.1907	0.0031	0.2137	0.2660	0.9991
Halton	−1.9854	−0.3383	−0.0465	−0.2180	−0.3965	−0.0953
Sobol'	−0.3241	0.0798	0.0015	0.1700	−0.0225	0.5354

(i) s = 40, n = 200 (r = 5)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.0075	0.0090	5.26e−07	0.0223	−0.0005	−0.0093
LH−dc	0.0280	0.0468	2.77e−06	0.0237	0.0438	0.9969
SLH	0.0099	0.0069	1.79e−07	0.0224	0.0164	−0.9962
SLH−dc	0.0465	0.0724	4.05e−06	0.0255	0.1225	0.9969
GLP	−0.2741	0.0650	−6.10e−6	0.0361	−0.0198	−2.0383
GLP−dc	−0.0575	0.0906	3.78e−06	0.0268	0.1402	0.9969
Halton	−53.2279	−4.5124	−0.0178	−0.6832	−0.6787	−9.0684
Sobol'	−21.6228	−1.8769	2.82e−06	0.0241	−0.0788	0.1220

(j) s = 40, n = 400 (r = 10)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.0296	0.0086	6.64e−07	0.0226	−0.0063	−0.0117
LH−dc	0.0443	0.0316	3.14e−06	0.0238	0.0231	0.9983
SLH	0.0215	0.0047	2.43e−07	0.0225	0.0266	−1.0234
SLH−dc	0.0603	0.0517	5.02e−06	0.0254	0.0706	0.9984
GLP	0.1542	0.0671	−2.28e−6	0.0550	0.0228	−1.0616
GLP−dc	0.0826	0.0702	4.53e−06	0.0388	0.0687	0.9985
Halton	−36.7896	−3.0350	−0.0223	−0.6775	−0.6608	−8.1956
Sobol'	−14.7732	−1.1558	3.90e−06	0.0306	−0.0446	0.3284

(k) s = 40, n = 800 (r = 20)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.0211	0.0106	9.24e−07	0.0228	0.0019	0.0005
LH−dc	0.0452	0.0268	3.42e−06	0.0236	0.0122	0.9991
SLH	0.0173	0.0057	3.74e−07	0.0228	0.0187	−0.9871
SLH−dc	0.0547	0.0394	5.55e−06	0.0250	0.0489	0.9992
GLP	0.0381	0.0483	−3.03e−6	0.0620	−0.0060	−0.4264
GLP−dc	0.0630	0.0505	1.56e−06	0.0475	0.0389	0.9992
Halton	−25.8221	−2.0322	−0.0305	−0.6599	−0.6445	−5.2745
Sobol'	−10.2242	−0.6638	2.51e−06	0.0395	0.0016	0.6819

(l) s = 40, n = 1200 (r = 30)

	MD$_2$	CD$_2$	SD$_2$	WD$_2$	MinDist	Corr
LH	0.0235	0.0100	9.22e−07	0.0229	0.0030	−0.0066
LH−dc	0.0389	0.0245	3.58e−06	0.0236	0.0092	0.9994
SLH	0.0204	0.0082	5.61e−07	0.0229	0.0257	−0.9987
SLH−dc	0.0531	0.0367	5.82e−06	0.0248	0.0387	0.9994
GLP	0.0675	0.0599	9.35e−06	0.0778	0.0686	−1.7754
GLP−dc	0.0845	0.0634	1.03e−05	0.0586	0.1335	0.9994
Halton	−20.9002	−1.5993	−0.0345	−0.6360	−0.6354	−3.7889
Sobol'	−8.1915	−0.4614	2.13e−06	0.0459	0.0127	0.7752

Figure 1. Uniform metrics of 8 compared sampling methods against crude Monte-Carlo. '*s*' is the number of dimensions and '*n*' is the sample size. De-correlated samples are labeled with '-dc'.

The effectiveness of a surrogate model built using different sampling methods is evaluated against the one based on an independent Monte Carlo sample set. The Root Mean Squared Error (RMSE) of 2000 Monte Carlo sample points is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \qquad (20)$$

where Y_i is the output of the surrogate model and $\hat{Y}_i$ is the corresponding output of the original model, and n = 2000 is the size of test set. Smaller RMSE means better goodness-of-fit, and the surrogate model is effective if the RMSE of an independent test set is small enough. To get a stable result, for random initial sampling method, such as MC, LH, LH-dc, SLH and SLH-dc, the surrogate modeling experiment was replicated for 10 times, and only the mean RMSEs of 10 replications are shown in the results section.

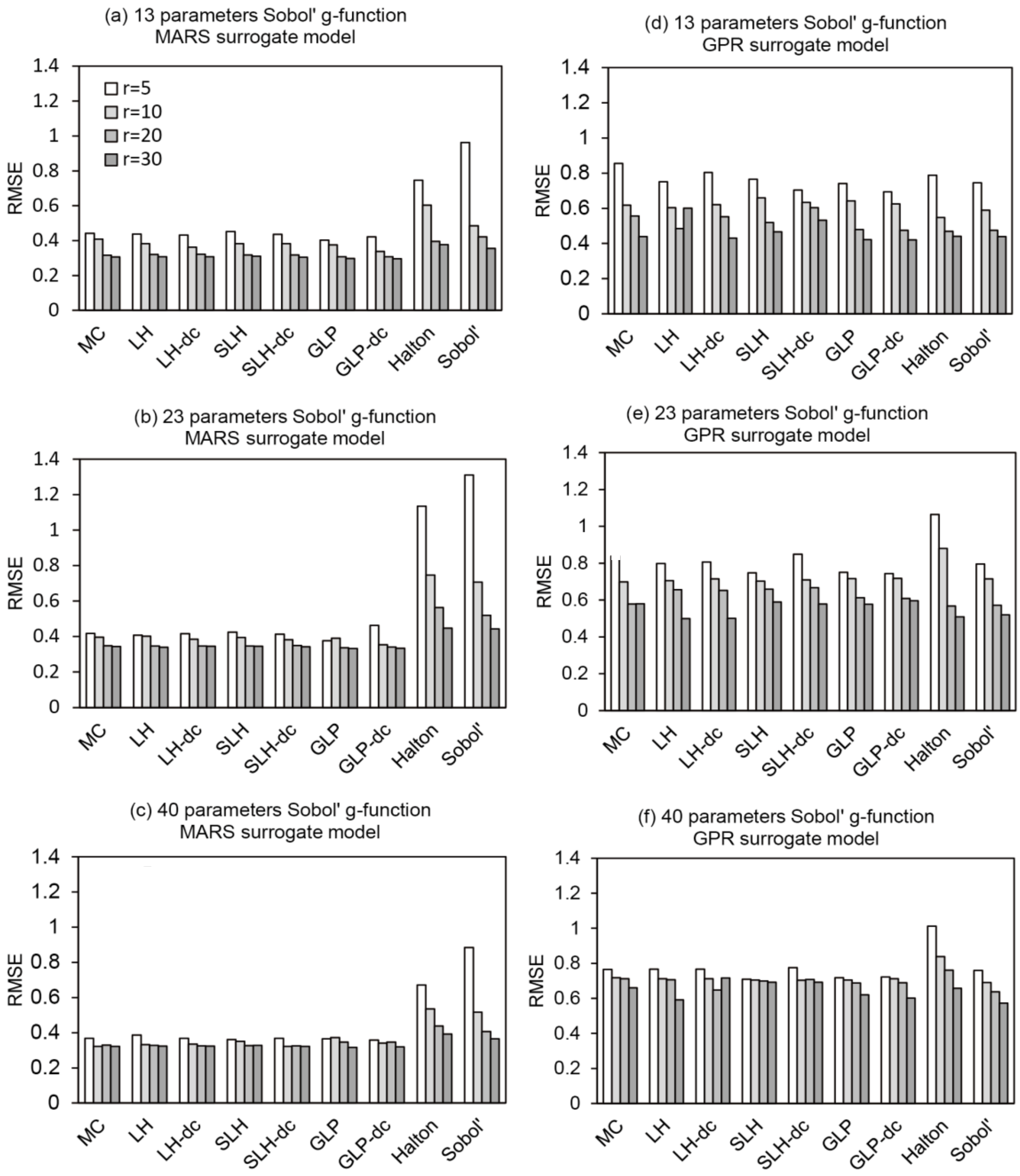


Figure 2. RMSEs of the MARS and GPR surrogate models built with different sampling methods (Sobol' g-function).

(3) The effect of sampling methods on sensitivity analysis. In this set of numerical experiments, we examine the effect of different sampling methods on the sensitivity analysis results. The SA method used for this purpose is the RSMSobol' method, a surrogate modeling based quantitative sensitivity analysis method that calculates both the main and total effect of each parameter (Sobol', 1993; Sobol', 2001; Storlie et al., 2009). The RSMSobol' method is revised from the original Sobol' variance decomposition method by running the Sobol' calculation on the response surface of a surrogate model. It is as effective as the original Sobol' method, but is more efficient computationally. In the case study with the Sobol' g-function, we computed the main effect of the g-function using the RSMSobol method and compared it with the theoretical values using Equation (18). In the case of SAC-SMA model, only the main effect given by RSMSobol was presented. The sensitivity analysis of random sampling methods (MC, LH, LH-dc, SLH, SLH-dc) was repeated for 10 times and only the mean values of RSMSobol results are shown.

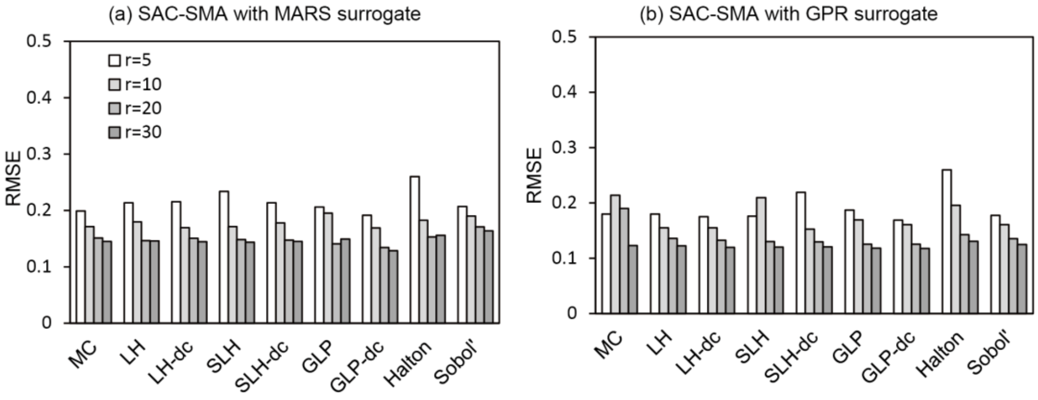


Figure 3. RMSEs of the MARS and GPR surrogate models built with different sampling methods (SAC-SMA hydrological model).

(4) The effect of sampling methods on parameter optimization. The objective of this test is to compare the influence of initial sampling on surrogate modeling based optimization. In this test case, we use the SAC-SMA model with recreated streamflow using assigned parameters listed in Table 1. Because optimization result is significantly influenced by the adaptive sampling strategy, we only compared the optimal point (i.e. the point having minimum RMSE) based on the initial sample sets generated by different sampling methods. To obtain statistically meaningful results for different sampling methods, the optimization was replicated for 10 times using different random realizations. The influence of the sample size was also investigated.

4. Results

4.1. Uniformity Metrics

First we evaluated the efficiency of 9 sampling methods with 6 uniformity metrics. Figure 1 shows the uniformity metrics of each case. The numbers in each grid are the normalized uniformity metrics defined in Table 2. As shown in the legend, different ranks have different colors. Color ‘red’ implies more uniform, and color ‘blue’ means less uniform. To obtain statistically robust results, the MC, LH, LH-dc, SLH and SLH-dc samplings are repeated 100 times, and only the mean value of uniform metrics are shown in this figure. Figure 1 reveals some interesting information: (1) For most cases, GLP (and also GLP-dc) produces the most uniform sample set, and SLH (and also SLH-dc) ranks the second. (2) The de-correlation post-processing method can significantly improve the uniformity of a sample set. In most cases, the de-correlated sample set is more uniform than the original sample set. (3) The quasi-random sampling methods, Halton and Sobol’, are not as uniform, even compared to the crude Monte Carlo method if the sample size is small (r = 5 or 10). The uniformity of Halton and Sobol’ sampling methods improves when the sample size is sufficiently large (r = 20 or 30). (4) The ranks given by different uniform metrics vary slightly, but they are similar to each other.

4.2. Surrogate Modeling

The previous section has clearly established what sampling methods are the most efficient according to the six uniformity metrics. Here we investigate if the efficient sampling methods lead to better surrogate models. In test case with the Sobol’ g-function, the RMSEs of the MARS and GPR surrogate models built with different sampling methods are presented in Figure 2. From Figure 2 we have the following interesting findings: (1) Compared to the crude Monte-Carlo, the RMSEs of LH, LH-dc, SLH, SLH-dc, GLP, and GLP-dc are very similar, and always lower than the other two QMC methods: Halton and Sobol’. (2) The RMSE values of Halton and Sobol’ sampling methods are also large when r = 5 and 10. But they can be reduced when r = 20 and 30. This finding confirms the finding from Figure 1 that the uniformity of the Halton and Sobol’ method improves with increasing sample size. (3) In comparison of the MARS and GPR surrogate models, the RMSEs given by MARS are generally smaller than that given by GPR, which seems to deviate from the conclusion of Wang et al. (2014) and Gong et al., (2014). As shown in Figure 1 of Sobol' (1993), the shape of Sobol’ g-function seems like a symmetric hinge. So the MARS surrogate is more suitable because it can fit the hinge with its inherent hinge function (Hastie et al., 2009), while the GPR acts like an interpolation approach that may give over-smoothed prediction at the valley bottom of Sobol’ g-function. This observation implies that different problems may prefer different type of surrogate models, and it is essential to prudently evaluate the fitness of candidate surrogate models and select the best one for sensitivity analysis and for optimization, respectively.

The RMSEs of surrogate models of SAC-SMA test case are presented in Figure 3. The RMSEs of MC, LH, LH-dc, SLH, SLH-dc, GLP, and GLP-dc are very similar, while that of Sobol’ sequence becomes quite lower. The RMSEs of Halton sequence is still high. Unlike Figure 2, the RMSEs provided by GPR are low than that of MARS, confirming the finding of Wang et al. (2014) and Gong et al. (2014) that GPR is suitable for constructing the surrogate model for SAC-SMA hydrological model.

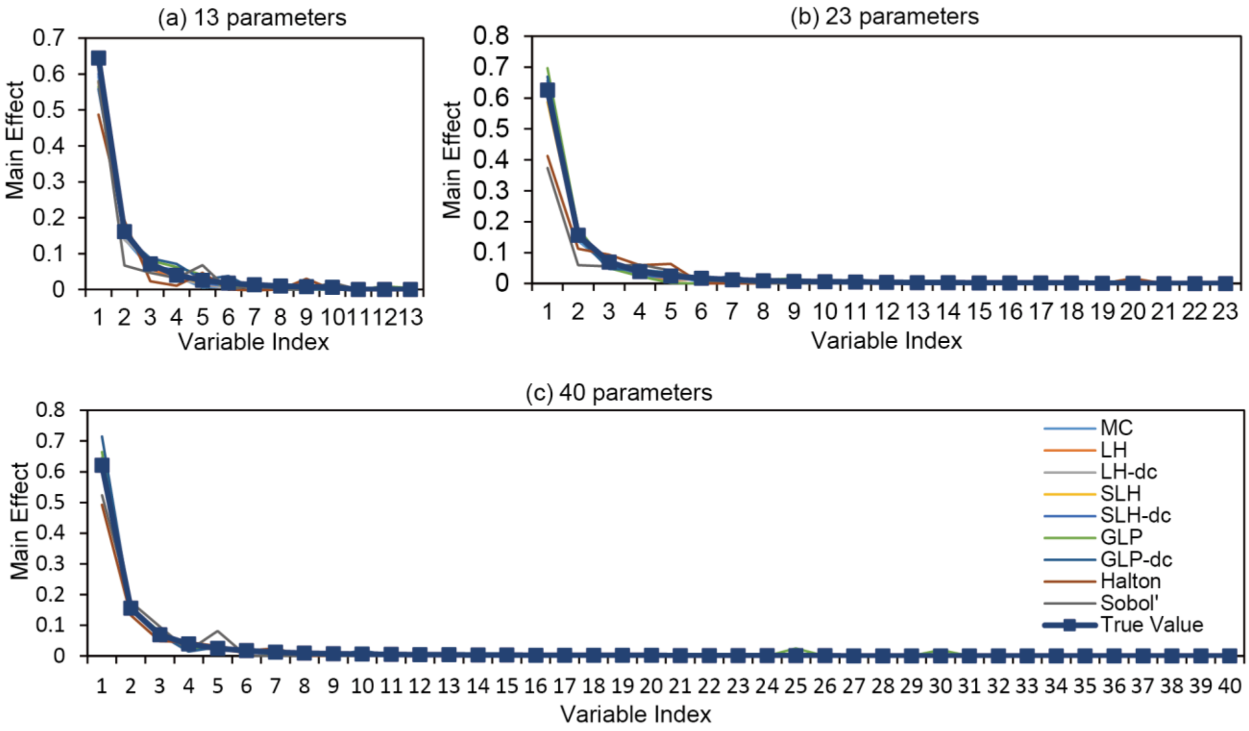


Figure 4. RSMSobol sensitivity analysis results comparing with the true value of main effect (Sobol' g-function with MARS surrogate).

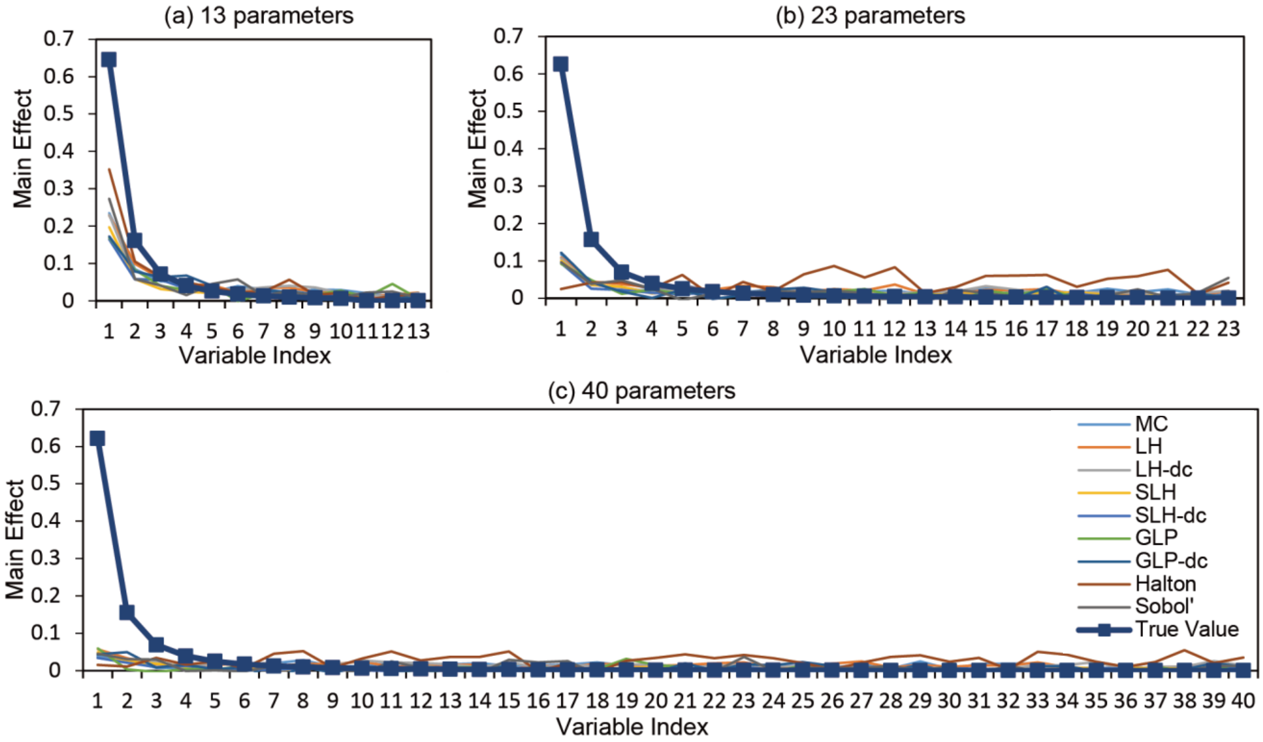


Figure 5. RSMSobol sensitivity analysis results comparing with the true value of main effect (Sobol' g-function with GPR surrogate).

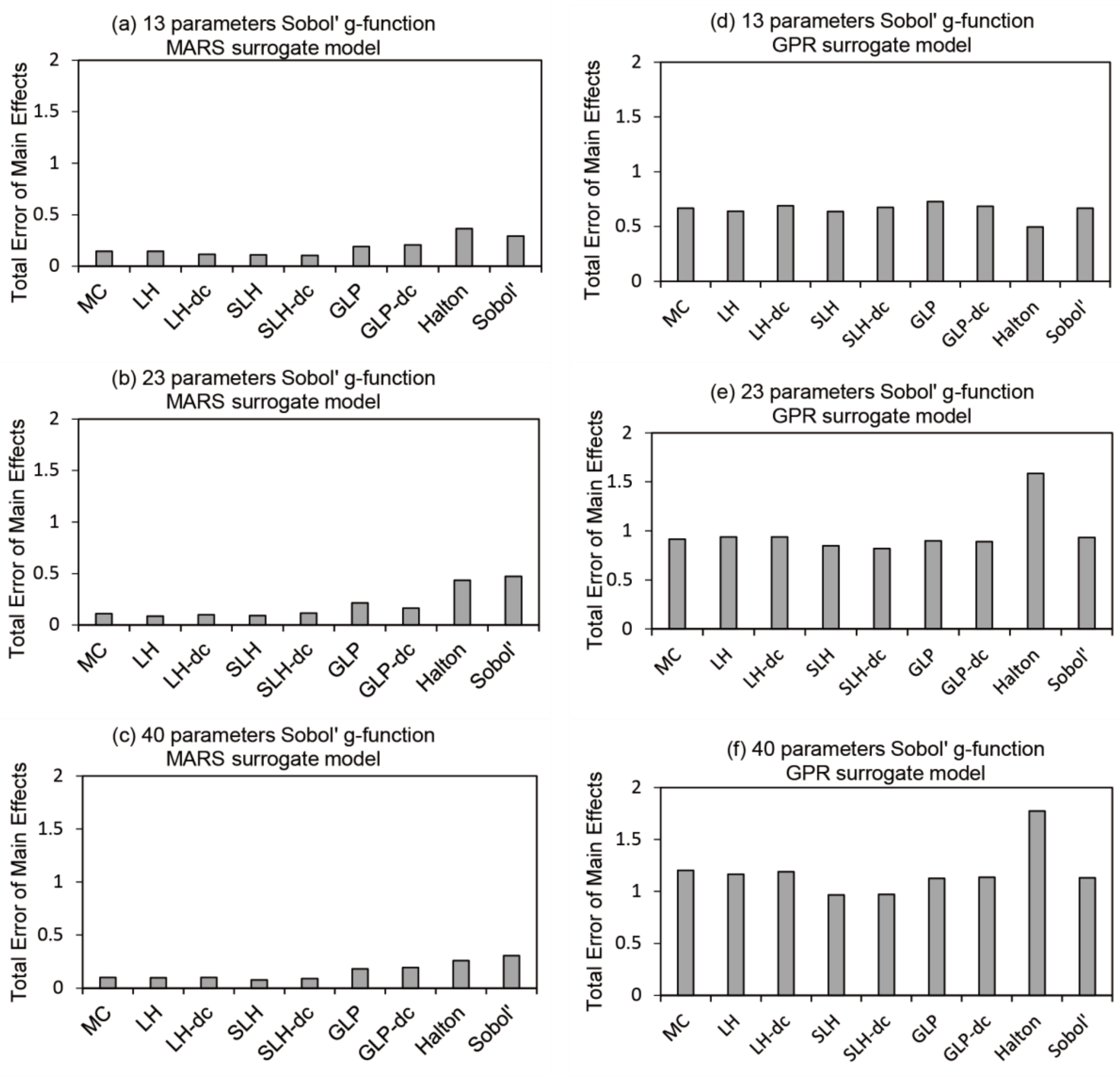


Figure 6. Total error of main effect given by RSMSobol built with different sampling methods (Sobol’ g-function).

4.3. Sensitivity Analysis

The results from the previous section suggest that the more efficient sampling methods may not lead to better surrogate models. Here we examine if efficient sampling methods will lead to more accurate sensitivity analysis results. For Sobol’ g-function, the main effects given by the RSMSobol sensitivity analysis are shown in Figure 4 (MARS surrogate) and Figure 5 (GPR surrogate). The total errors, which are the sum of the absolute errors of each factor compared to the analytical true value given in Equation (18), are plotted in Figure 6. In this step, we only present the results of r = 10, as the results of other sample sizes are similar to r = 10. Figures 4, 5, and 6 indicated that for sensitivity analysis, the influence of the surrogate modeling methods is much more important than the sampling methods. For RSMSobol with MARS surrogate, MC, LH, LH-dc, SLH, SLH-dc GLP and GLP-dc samplings provide similarly small errors, and the errors of the Halton and Sobol’ samplings are relatively large. For RSMSobol with GPR surrogate, the errors are much larger than that of MARS surrogate, no matter what kind of sampling approach is used. The RSMSobol with MARS surrogate can correctly identify the sensitive parameters, while with GPR surrogate the sensitivity analysis result is quite misleading and ineffective.

For the SAC-SMA model, the main effects given by RSMSobol approach with MARS and GPR surrogate model are presented in Figure 7. Because for the SAC-SMA model the theoretical values of main effects are unknown, we only show the sensitivity index $S_i = V_i/V(Y)$ provided by different sampling and surrogate models, and evaluate their differences. Similarly with the results of Sobol’ g-function, the influence of the surrogate modeling methods is also more significant than the sampling methods. In Figure 7(a), lztwm, lzfsm, lzfpm, lzsk are identified as sensitive parameters, however in

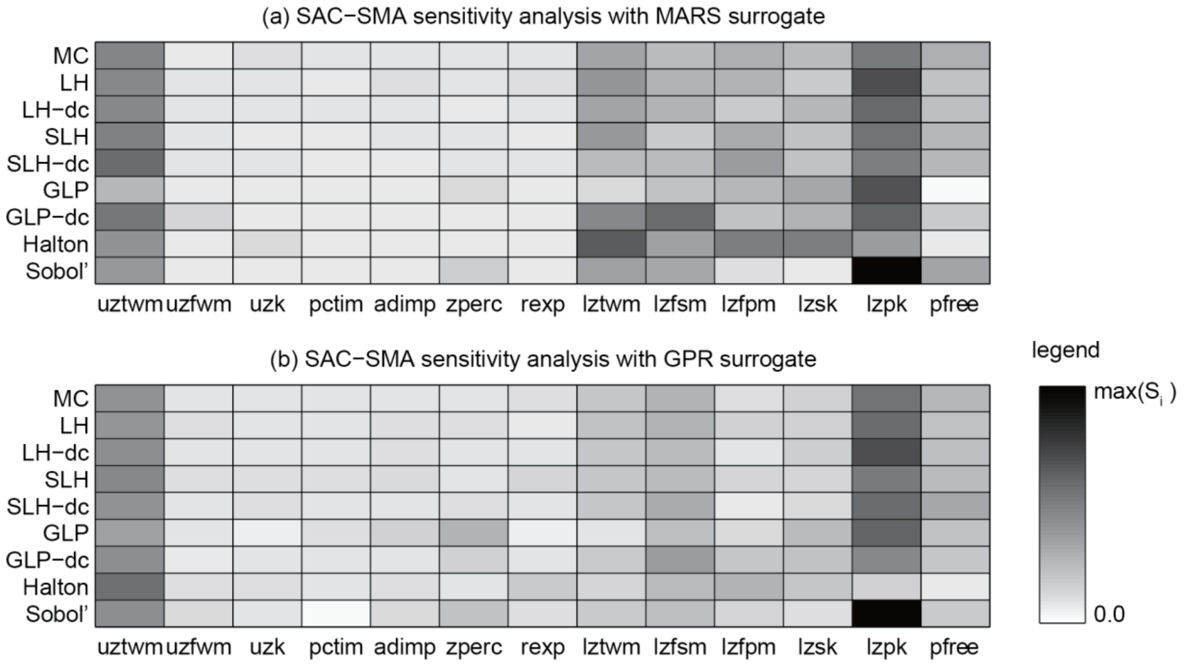


Figure 7. RSMSobol sensitivity analysis results of SAC-SMA hydrological model with MARS and GPR surrogates.

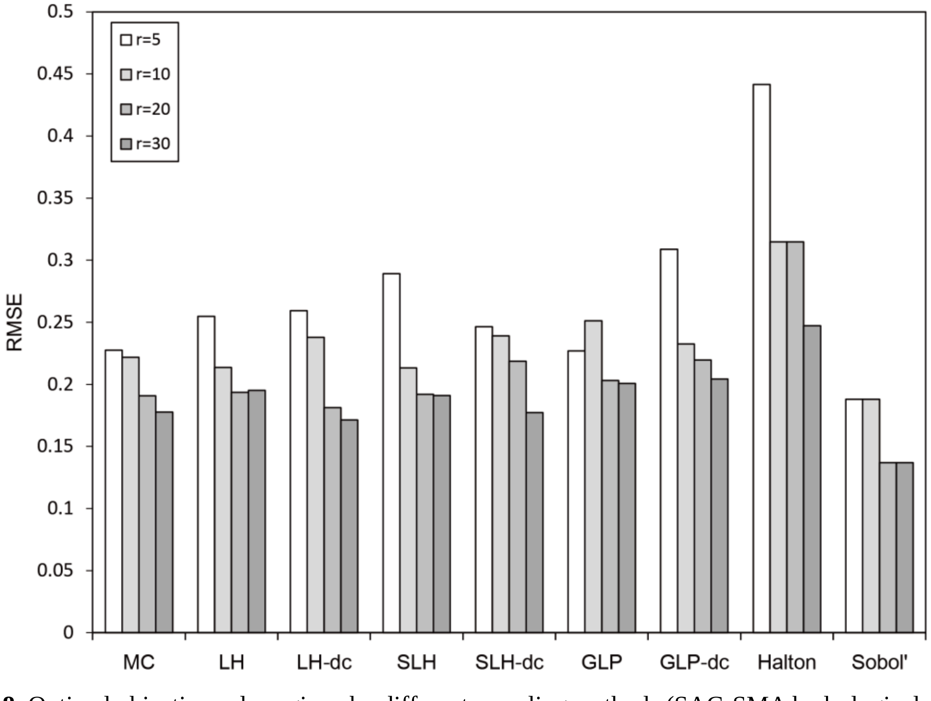


Figure 8. Optimal objective values given by different sampling methods (SAC-SMA hydrological model).

Figure 7(b), they are still sensitive but less significant. On the influence of sampling methods, with MARS surrogate, GLP and Halton sampling failed to screen out pfree as a sensitive parameter, while Sobol' sequence failed to screen out lzfpm and lzsk. With GPR surrogate, Halton sequence failed to screen out lzpk and pfree.

4.4. Optimization

Now we found that the efficiency of sampling methods is less important in surrogate modeling and sensitivity analysis results compared to other factors such as surrogate modeling methods or the type of test problems. This section investigates how the efficiency of the sampling methods is related to the robustness of optimization results. Based on Wang et al. (2104), the sampling methods and sample sizes of initial sampling have significant influence on surrogate-based optimization when Halton and Sobol' sampling methods are used. Here we examined the optimization results of other sampling methods along with Halton and Sobol' methods. The optimal objective values given by different sampling methods and sample sizes are shown in Figure 8. As shown in this figure, the Sobol' sequence provides best optimal value, which confirms the conclusion of Wang et al. (2014). The optimal values given by Halton sequence is not as good as other sampling methods, and the MC, LH, LH-dc, SLH, SLH-dc, GLP and GLP-dc have similar performance.

Interestingly, although the Sobol' sequence did not have outstanding performance in surrogate modeling and sensitivity analysis experiments, it did produce the best optimal parameter set compared to other sampling methods. The possible explanation might be the effectiveness of the sampling methods in terms of optimization results depend on many factors in addition to sampling methods, including the choice of surrogate modeling methods and optimization search methods. To confirm the generality of this observation, more test problems should be conducted.

5. Discussion and Conclusions

In this paper, the effectiveness and efficiency of nine sampling methods for uncertainty quantification are evaluated. First we used six kinds of uniformity metrics to evaluate the uniformity of sample sets, then we compared the results of surrogate modeling, sensitivity analysis and parameter optimization with test problems. The main findings are summarized as below.

According to the uniform metrics, Symmetric Latin Hypercube (SLH) and Good Lattice Points (GLP) are the most efficient sampling methods in this comparison. If the sampling procedure needs replication, SLH is preferred because it is a random sampling method. GLP is preferred if the computational resources are rather limited that only a small number of samples is affordable. On the other hand, the Halton and Sobol' quasi-random sampling methods are even not as uniform as crude Monte Carlo when the number of samples is not large enough. Compared to previous studies like (Morokoff et al., 1995) and (Fang et al., 1994), we have extended the inter-comparison to higher dimension (s = 13, 23 and 40) and various sample sizes, and acquired similar results. It might be interesting to extend the comparison to even higher dimensions, and check the extendibility of such conclusion to various kinds of problems. Considering the theoretical order of discrepancy, we have confirmed the conclusion of Morokoff et al. (1995) that some QMC methods' actual discrepancies fall behind their theoretical values. For an instance, theoretically the order of discrepancies of GLP, Halton and Sobol' QMC methods are quite similar, but actually the computed discrepancy of GLP is better than the other two, especially when the number of sample points is very small. The RGS de-correlation can significantly improve the uniformity metrics (efficiency) of lattice designs.

Interestingly, although the efficiency of a sampling method can be objectively measured using problem independent uniformity metrics, the effectiveness largely depends on many other factors. As indicated by the test problems, the type of surrogate model, sensitivity analysis method, and the intrinsic properties of the environmental dynamic model have more significant affects to the final results than sampling methods. For each practical problem, it is necessary to prudently choose appropriate UQ methods.

We hope our work is useful for scientists who are interested in sensitivity analysis, surrogate modeling and parameter optimization for environmental dynamic models. Any discussion and collaborations on the sampling and relative topics are warmly welcome, and the source code used in this paper is available from the first author.

Acknowledgments. This research is supported by the Chinese Ministry of Science and Technology 973 Research Program (No. 2015CB953703), the Special Fund for Meteorological Scientific Research in Public Interest (GYHY201506002, CRA~40: the 40~ year CMA global atmospheric reanalysis), the Natural Science Foundation of China (No.41075075, No.41375139 and No.51309011) and the Fundamental Research Funds for the Central Universities - Beijing Normal University Research Fund (No.2013YB47).

References

Bates, R.A., Buck, R.J., Riccomagno, E., and Wynn, H.P. (1996). Experimental Design and Observation for Large Systems. J. *Roy. Stat. Soc. Ser. B. (Stat. Method.),* 58(1), 77-94. http://dx.doi.org/10.2307/2346166

Bratley, P. and Fox, B.L. (1988). Algorithm 659: Implementing sobol's quasirandom sequence generator. *ACM Trans. Math. Software,* 14(1), 88-100. http://dx.doi.org/10.1145/42288.2143 72

Caflisch, R.E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica,* 7, 1-49. http://dx.doi.org/10.1017/S0962492900002804

Crino, S. and Brown, D.E. (2007). Global optimization with multivariate adaptive regression splines. *IEEE Trans. Syst., Man, Cybern. B, Cybern.,* 37 (2), 333-340. http://dx.doi.org/10.1109/TSMCB.2006.883430

Fang, K. (1980). The uniform design: application of number-theoretic methods in experimental design. *Acta Math. Appl. Sinica,* 3(4), 363-372. http://dx.doi.org/10.1080/00401706.2000.10486045

Fang, K., Li, R., and Sudjianto, A. (2006). *Design and modeling for computer experiments,* Boca Raton, FL, Chapman & Hall / CRC.

Fang, K., Lin, D.K.J., Winker, P., and Zhang, Y. (2000). Uniform design: Theory and application. *Technometrics,* 42(3), 237-248. http://dx.doi.org/10.2307/1271079

Fang, K., Ma, C., and Winker, P. (2002). Centered L2-Discrepancy of random sampling and latin hypercube design, and construction of uniform designs. *Math. Comput.,* 71(237), 275-296. http://dx.doi.org/10.2307/2698872

Fang, K., Wang, Y., and Bentler, P.M. (1994). Some applications of number-theoretic methods in statistics. *Stat. Sci.,* 9(3), 416-428. http://dx.doi.org/10.1214/ss/1177010392

Fang, K. and Ma, C. (2001). *Orthogonal and uniform experiment design,* Beijing, China, Science Press.

Fang, K. and Wang, Y. (1994). *Number-theoretic methods in statistics,* London, UK, Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-3095-8

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Ann. Stat.,* 19(1), 1-14. http://dx.doi.org/10.1214/aos/1176347 963

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., Ye, A., and Miao, C. (2014). Multi-objective parameter optimization of common land model using adaptive surrogate modelling. *Hydrol. Earth Syst. Sci. Discuss.,* 11(6), 6715-6751. http://dx.doi.org/10.5194/hessd-11-6715-2014

Halton, J.H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM,* 7(12), 701-702. http://dx.doi.org/10.1145/355588.365104

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning 2nd,* New York, USA, Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Hickernell, F.J. (1998). A generalized discrepancy and quadrature error bound. *Math. Comput.,* 67(221), 299-322. http://dx.doi.org/10.1090/S0025-5718-98-00894-1

Hickernell, F.J. (1998). Lattice rules: how well do they measure up? *Random and Quasi-Random Point Sets,* Hellekalek, P. and Larcher, G., Springer-Verlag: 106-166.

Hua, L. K. and Wang, Y. (1981). *Application of number theory to numberical analysis,* Berlin and Beijing, Springer and Science Press.

Iman, R.L. and Conover, W.J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Commun. Stat. Simulation Comput.,* 11(3), 311-334. http://dx.doi.org/10.1080/03610918208812265

Johnson, M.E., Moore, L.M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Stat. Plann. Inference,* 26(2), 131-148. http://dx.doi.org/10.1016/0378-3758(90)90122-B

Korobov, N.M. (1959). Computation of multiple integrals by the method of optimal coefficients. Vestnik Moskow *Univ. Sec. Math. Astr. Fiz. Him.,* 4, 19-25.

Korobov, N.M. (1959). The approximate computation of multiple integrals. *Dokl. Akad. Nauk. SSSR,* 124, 1207-1210.

Kuipers, L. and Niederreiter, H. (1974). *Uniform distribution of sequences,* New York, USA, Wiley.

Li, J., Duan, Q.Y., Gong, W., Ye, A., Dai, Y., Miao, C., Di, Z., Tong, C., and Sun, Y. (2013). Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis. *Hydrol. Earth Syst. Sci.,* 17(8), 327 9-3293. http://dx.doi.org/10.5194/hess-17-3279-2013

Li, W.W. and Wu, C.F.J. (1997). Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics,* 39(2), 171-179. http://dx.doi.org/10. 2307/1270905

Liu, Y.Q., Gupta, H.V., Sorooshian, S., Bastidas, L.A., and Shuttleworth, W.J. (2005). Constraining land surface and atmospheric parameters of a locally coupled model using observational data. *J. Hydrometeorol.,* 6(2), 156-172. http://dx.doi.org/10.1175/JHM407.1

McKay, M.D., Beckman, R.J., and Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics,* 21(2), 239-245. http://dx.doi.org/10.2307/1268522

Morokoff, W.J. and Caflisch, R.E. (1995). Quasi-Monte Carlo integration. *J. Comput. Phys.,* 122(2), 218-230. http://dx.doi.org/10.1006/jcph.1995.1209

Morris, M.D. and Mitchell, T.J. (1995). Exploratory designs for computational experiments. *J. Stat. Plann. Inference,* 43(3), 381-402. http://dx.doi.org/10.1016/0378-3758(94)00035-T

Niederreiter, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.,* 84, 957-1041. http://dx.doi.org/10.1090/S0002-9904-1978-14532-7

Niederreiter, H. (1992). *Random number generation and Quasi-Monte Carlo methods,* Philadelphia, USA, SIAM. http://dx.doi.org/10.1137/1.9781611970081

Owen, A.B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Stat Sin.,* 2(2), 439-452.

Owen, A.B. (1994). Controlling correlations in Latin hypercube samples. *J. Am. Stat. Assoc.,* 89(428), 1517-1522. http://dx.doi.org/10.2307/2291014

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian processes for machine learning. massachusetts,* USA, MIT Press.

Sacks, J., Schiller, S.B., and Welch, W.J. (1989). Designs for computer experiments. *Technometrics,* 31(1), 41-47. http://dx.doi.org/10.2307/1270363

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis: the Primer,* Chichester, UK, Wiley-Interscience.

Shahsavani, D., Tarantola, S., and Ratto, M. (2010). Evaluation of MARS modeling technique for sensitivity analysis of model output. *Proc. Soc. Behavior. Sci.,* 2(6), 7737-7738. http://dx. doi.org/10.1016/j.sbspro.2010.05.204

Shaw, J.E.H. (1988). A quasirandom approach to integration in bayesian statistics. *Ann. Stat.,* 16(2), 895-914. http://dx.doi.org/10.2307/2241763.

Shewry, M.C. and Wynn, H.P. (1987). Maximum entropy sampling. *J. Appl. Stat.,* 14(2), 165-170. http://dx.doi.org/10.1080/02664768700000020

Sloan, I.H. (1985). Lattice methods for multiple integration. *J. Comput. Applied Math.,* 12-13, 131-143. http://dx.doi.org/10.1016/0377-0427(85)90012-3

Sobol', I.M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.,* 7(4), 86-112. http://dx.doi.org/10.1016/0041-5553(67)90144-9

Sobol', I.M. (1967). The use of Haar series in estimating the error in the computation of infinite-dimensional integral. *Dokl. Akad. Nauk SSSR,* 175, 34-37.

Sobol', I.M. (1993). Sensitivity analysis for nonlinear mathematical models. *Math. Model. Comput. Exp.,* 1(4), 407-414.

Sobol', I.M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simulation,* 55(1-3), 271-280. http://dx.doi.org/10.1016/S0378-4754(00)00270-6

Storlie, C.B., Swiler, L.P., Helton, J.C., and Sallaberry, C.J. (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliab. Eng. Syst. Saf.,* 94(11), 1735-1763. http://dx.doi.org/10.1016/j.ress.2009.05.007

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *J. Am. Stat. Assoc.,* 88(424), 1392-1397. http://dx.doi.org/10.2307/2291282

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., and Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.,* 39(8), 1214. http://dx.doi.org/10.1029/2002WR0 01746.

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., and Miao, C. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environ. Model. Software,* 60, 167-179. http://dx.doi.org/10.1016/j.envsoft.20 14.05.026

Wang, Y. and Fang, K. (1981). A note on uniform distribution and experimental design. *Ke Xue Tong Bao,* 26(6), 495-489.

Weyl, H. (1916). Über die Gleichverteilung der Zahlem mod Eins. *Math. Ann.,* 77(3), 313-352. http://dx.doi.org/10.1007/BF01475864

Ye, K.Q., Li, W., and Sudjianto, A. (2000). Algorithmic construction of optimal symmetric Latin hypercube designs. *J. Stat. Plann. Inference,* 90(1), 145-159. http://dx.doi.org/10.1016/S0378-3758(00)00105-1