Environmental Modelling & Software 51 (2014) 269–285

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model☆

Yanjun Gan [a], Qingyun Duan [a,*], Wei Gong [a], Charles Tong [b], Yunwei Sun [c], Wei Chu [d], Aizhong Ye [a], Chiyuan Miao [a], Zhenhua Di [a]

[a] State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China
[b] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, USA
[c] Atmosphere, Earth and Energy Division, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, USA
[d] Department of Civil and Environmental Engineering, University of California, Irvine, CA 92617, USA

ARTICLE INFO

Article history:
Received 17 May 2013
Received in revised form
30 September 2013
Accepted 30 September 2013
Available online

Keywords:
Uncertainty quantification
Sensitivity analysis
Parameter screening
Space-filling sampling
PSUADE

ABSTRACT

Sensitivity analysis (SA) is a commonly used approach for identifying important parameters that dominate model behaviors. We use a newly developed software package, a Problem Solving environment for Uncertainty Analysis and Design Exploration (PSUADE), to evaluate the effectiveness and efficiency of ten widely used SA methods, including seven qualitative and three quantitative ones. All SA methods are tested using a variety of sampling techniques to screen out the most sensitive (i.e., important) parameters from the insensitive ones. The Sacramento Soil Moisture Accounting (SAC-SMA) model, which has thirteen tunable parameters, is used for illustration. The South Branch Potomac River basin near Springfield, West Virginia in the U.S. is chosen as the study area. The key findings from this study are: (1) For qualitative SA methods, Correlation Analysis (CA), Regression Analysis (RA), and Gaussian Process (GP) screening methods are shown to be not effective in this example. Morris One-At-a-Time (MOAT) screening is the most efficient, needing only 280 samples to identify the most important parameters, but it is the least robust method. Multivariate Adaptive Regression Splines (MARS), Delta Test (DT) and Sum-Of-Trees (SOT) screening methods need about 400–600 samples for the same purpose. Monte Carlo (MC), Orthogonal Array (OA) and Orthogonal Array based Latin Hypercube (OALH) are appropriate sampling techniques for them; (2) For quantitative SA methods, at least 2777 samples are needed for Fourier Amplitude Sensitivity Test (FAST) to identity parameter main effect. McKay method needs about 360 samples to evaluate the main effect, more than 1000 samples to assess the two-way interaction effect. OALH and LP_τ (LPTAU) sampling techniques are more appropriate for McKay method. For the Sobol' method, the minimum samples needed are 1050 to compute the first-order and total sensitivity indices correctly. These comparisons show that qualitative SA methods are more efficient but less accurate and robust than quantitative ones.

Software availability

Name of software: PSUADE
Developer: Charles Tong
Programming language: C++
Availability: https://computation.llnl.gov/casc/uncertainty_quantification/
Cost: Free for non-commercial academic research

1. Introduction

Computer-based system models have become indispensable in many fields of science and engineering, from finance to life sciences, from quantum physics to earth sciences and environmental engineering. Parameters of these models exert great influence on models' performance. Some of the parameters may be observed or measured, e.g., the physical dimensions of an object or the geomorphological features of a watershed such as slope, area size and elevation. But there are many parameters that are not directly observable, at least not at the scale of modeling units. For example, parameters commonly used in hydrologic models, such as saturated soil hydraulic conductivity or saturated soil matric potential, may be observable at a point scale, but not over a large area. In

* Corresponding author. Tel.: +86 10 5880 4191; fax: +86 10 5880 2165.
E-mail address: qyduan@bnu.edu.cn (Q. Duan).

1364-8152/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.09.031

this case, "effective" values must be estimated so mathematical equations established at a point scale can be extended to an areal scale (Blöschl and Sivapalan, 1995). There is a class of models known as conceptual models whose parameters are generally non-observable and are only related to physical properties indirectly. For example, the parameters in many conceptual rainfall-runoff (CRR) models are not observable and must be calibrated so model simulations closely match observations (Duan et al., 1992).

How to specify system model parameters properly is not a trivial issue (Sorooshian and Gupta, 1983; Duan et al., 1992, 2006; Kavetski et al., 2003). The combined effect of several factors, including errors in observational data, choices of calibration methods and criterias, and model formulation errors, makes parameter estimation being a difficult task. This difficulty is further compounded by over-parameterization problems as today's models are getting increasingly complex in a trend to include more and more sub-physics, but the calibration of these models is still done with rather limited data (Jakeman and Hornberger, 1993; Renard et al., 2010; Clark et al., 2011). Over-parameterization, along with parameter interactions (due to high nonlinearity of model equations), causes model parameters to be not uniquely identifiable. Beven (2006) termed this phenomenon as equifinality, i.e., different parameter sets would result in the same or similar model performance measures. Another potential cause for equifinality may be due to a phenomenon known as "numerical daemon" by Kavetski and Clark (2010). One possible way to mitigate over-parameterization/non-identifiability is reducing the number of parameters to a small number that can be sufficiently calibrated with limited data.

To discern which parameters have the most influence over model performance and to identify what are the most appropriate parameter values, we need to find a way to screen out sensitive parameters and quantitatively evaluate the influence of each parameter on model performance. Sensitivity analysis (SA) has been used by many people for this purpose (Liu et al., 2004; van Griensven et al., 2006; Campolongo et al., 2007; Borgonovo et al., 2012). SA can identify parameters of which a reduction in uncertainty specification will have the most significant impact on improving model performance measures. Thus, if some non-influential parameters can be identified and fixed reasonably at given values over their ranges, the computational cost may decrease without reducing model performance.

There are many different SA approaches. Overall, they can be categorized into two groups: local SA and global SA. The local SA explores the changes of model response by varying one parameter while keeping other parameters constant. The simplest and most common approach is differential SA (DSA), which uses partial derivatives or finite differences of parameters at a fixed parameter location as the measure of parametric sensitivity. Though simple and intuitive, DSA measures only local sensitivity whose value is obviously location dependent. On the other hand, the global SA examines the changes of model response by varying all parameters at the same time. Generalized SA (GSA) method is one of the global SA methods that are designed to overcome the limitations of local SA methods. A version of GSA method, as implemented in Hornberger and Spear (1981), first creates a large number of random parameter sets using the Monte Carlo (MC) (Meteopolis and Ulam, 1949) sampling technique. It then breaks the random parameter sets into behavioral and non-behavioral sets based on a pre-specified threshold for acceptance of model behavior. The frequency density distributions of model performance measures along each parameter axis in the behavioral sub-set are used as indicators of parametric sensitivities. GSA forms the basis for the Generalized Likelihood Uncertainty Estimation (GLUE) method developed by Beven and Binley (1992). GSA is simple to implement and can work with different pseudo-likelihood (i.e., goodness of fit) measures (Beven, 2004), but it is computationally inefficient.

Global SA approaches based on design of experiment (DOE) have gained popularity recently because they offer global sensitivity measures while maintaining computational efficiency. A typical DOE-based SA method involves two steps: first, generating a sample set of parameters within the feasible parameter spaces using a chosen design; and then, obtaining a quantitative attribution of model output variation due to the variation of different parameters. There are many sampling techniques, such as MC, Latin Hypercube (LH) (McKay et al., 1979), Orthogonal Array (OA) (Owen, 1992) and Orthogonal Array based Latin Hypercube (OALH) (Tang, 1993), which are commonly used for DOE-based SA. Some DOE-based SA methods, such as Morris One-At-a-Time (MOAT) (Morris, 1991), Fourier Amplitude Sensitivity Test (FAST) (Cukier et al., 1973), and extended Sobol' method (Saltelli, 2002), require special sampling techniques. More recently, along with the development of response surface methods (RSM), SA based on RSM makes it cheaper for estimating parameter effects (Ratto et al., 2007; Shahsavani and Grimvall, 2011).

Saltelli et al. (2008) provided a comprehensive exposition of contemporarily available SA methods. Tong (2005) developed a software package, called a Problem Solving environment for Uncertainty Analysis and Design Exploration (PSUADE) and containing a wide array of different uncertainty quantification (UQ) methods, including many SA methods. PSUADE has been used successfully for many applications. Hsieh (2006) demonstrated the process of using PSUADE for UQ of the Steven Impact Test problem. Wemhoff and Hsieh (2007) used PSUADE to calibrate the Prout–Tompkins chemical kinetic model. Tong (2008) applied a variety of UQ techniques to the study of a two-dimensional soil-foundation structure-interaction system subjected to earthquake excitation using PSUADE. Tong and Graziani (2008) described a global SA methodology implemented in PSUADE that is specifically designed for general multi-physics application of large complex system models. Snow and Bajaj (2010) adopted the PSUADE for uncertainty analysis of a comprehensive electrostatic Micro-Electro Mechanical Systems (MEMS) switch model.

The aforementioned works have been focused on applying a subset of the UQ methods available within PSUADE. The purpose of this paper is to explore the effectiveness and efficiency of various SA methods in PSUDAE in identifying sensitive parameters of system models, and provide useful guidance on selecting appropriate SA procedures for other applications. We test all available SA methods with a very simple conceptual hydrologic model – Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973). The generality of the findings in this paper would need further works on more complex models and more catchments with different characteristics. This paper is organized as follows. Section 2 offers a brief description of the PSUADE software. Section 3 describes the model, data and experimental methods used in the study. Section 4 presents the results and discussion. And finally, we make some concluding remarks in Section 5.

2. The PSUADE software

PSUADE is a C++ based open-source software package developed to provide an integrated design and analysis environment for performing UQ for large complex system models. This software is available via https://computation.llnl.gov/casc/uncertainty_quantification/. The flow chart for implementing PSUADE for UQ is shown in Fig. 1. The three parts in bold italic are basic elements of PSUADE:

- The experimental design techniques (***Sample generator***)
- The simulator execution environment (***Driver***)
- The analysis toolset (***Analysis tool***)

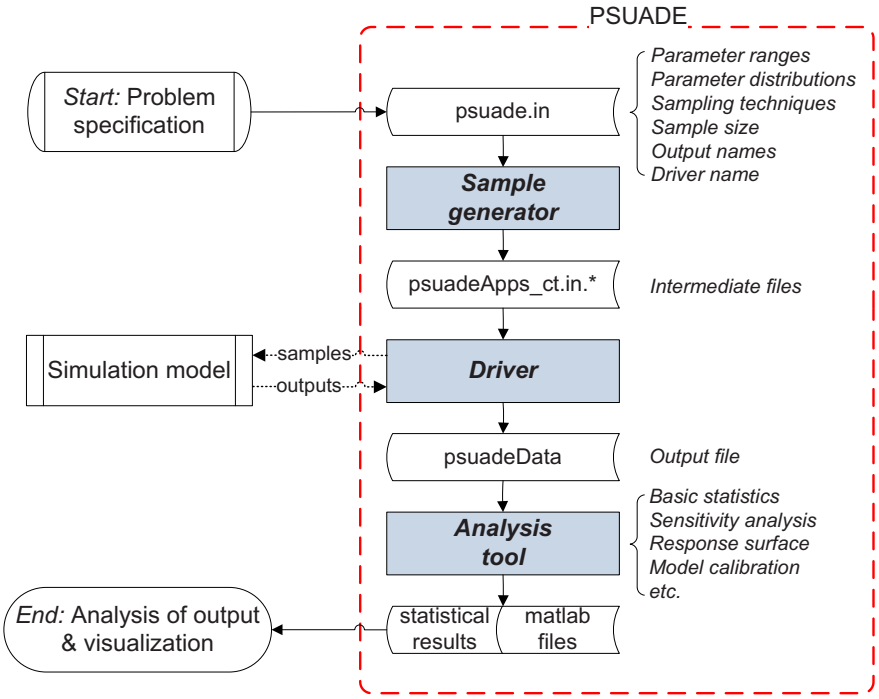

Fig. 1. Flow chart of using PSUADE for uncertainty quantification.

The first part, as defined in the input file "psuade.in", is a sample generator specifying sampling technique, sample size, and parameter ranges and distributions. Sampling techniques available in PSUADE are listed in Table 1. PSUADE supports several probability density functions, such as uniform, normal, lognormal, and triangular distribution. Furthermore, it provides adaptive sampling techniques for global and local refinement.

The second part provides a "non-intrusive" and user-friendly interface for linking simulation executable code and PSUADE. "Non-intrusive" means that users don't need to modify the code of the model. Users can write their own script in any programming language (PSUADE provides default template in Python or C format), which could be used as the "driver" for running a given computer model and collecting the model outputs to the output file "psuadeData".

The third part provides a variety of mathematical/statistical methods for analyzing the input–output relationships. It has a rich set of tools for basic statistical analysis, sensitivity analysis, response surface analysis, and model calibration, etc. Sensitivity analysis methods available in PSUADE are given in Table 2.

3. Experimental data, methods, and setup

3.1. Test problem and data

This study intends to fully explore the various SA methods available in PSUADE. The SAC-SMA model is used as a test problem. This CRR model developed by Burnash et al. (1973) is the most widely used hydrological model by the River Forecast Centers (RFCs) of the U.S. National Weather Service for catchment modeling and flood forecasting. Readers who are interested in the details of this model should refer to Burnash (1995).

There are sixteen parameters in the SAC-SMA model (see Table 3). We consider only thirteen of them as tunable parameters, whose feasible ranges are determined based on their physical interpretations and watershed properties, which have been widely referred to in previous literature (e.g., Duan et al., 1994; Burnash,

Table 1
Sampling techniques available in PSUADE.

Sampling technique	Abbreviation	Source
Plackett–Burman	PBD	Plackett and Burman (1946)
Monte Carlo	MC	Meteopolis and Ulam (1949)
Central composite	CCI, CCF, CCC	Box and Wilson (1951)
Box–Behnken	BBD	Box and Behnken (1960)
Full and fractional factorial	FACT, FF	Box and Hunter (1961)
Fourier amplitude sensitivity test	FAST	Cukier et al. (1973)
Latin hypercube	LH	McKay et al. (1979)
LP_τ	LPTAU	Sobol' (1990)
Morris one-at-a-time	MOAT	Morris (1991)
Orthogonal array	OA	Owen (1992)
Orthogonal array based Latin hypercube	OALH	Tang (1993)
Metis	METIS	Karypis and Kumar (1998)
Extended FAST	EFAST	Saltelli et al. (1999)
Extended Sobol'	SOBOL	Saltelli (2002)

Table 2
SA methods available in PSUADE.

SA method	Abbreviation	Source
Correlation analysis	CA	Spearman (1904)
Regression analysis	RA	Galton (1886)
Plackett–Burman screening	PB	Plackett and Burman (1946)
Fractional factorial screening	FF	Box and Hunter (1961)
Morris one-at-a-time screening	MOAT	Morris (1991); Campolongo et al. (2007)
Sum-of-trees screening	SOT	Breiman et al. (1984)
Gaussian process screening	GP	Gibbs and MacKay (1997)
Multivariate adaptive regression splines screening	MARS	Friedman (1991)
Delta test screening	DT	Pi and Peterson (1994)
Fourier amplitude sensitivity test	FAST	Cukier et al. (1973)
McKay main and two-way interaction effect analysis	McKay	McKay (1995); Tong (2005)
Sobol' sensitivity indices	Sobol	Sobol' (1993, 2001)

Table 3
Parameters of SAC-SMA model.

No.	Parameter	Description	Range/value
1	UZTWM	Upper zone tension water maximum storage (mm)	[5.0, 300.0]
2	UZFWM	Upper zone free water maximum storage (mm)	[5.0, 150.0]
3	UZK	Upper zone free water lateral drainage rate (day^{-1})	[0.10, 0.750]
4	PCTIM	Impervious fraction of the watershed area (decimal fraction)	[0.0, 0.10]
5	ADIMP	Additional impervious area (decimal fraction)	[0.0, 0.40]
6	ZPERC	Maximum percolation rate (dimensionless)	[5.0, 350.0]
7	REXP	Exponent of the percolation equation (dimensionless)	[1.0, 5.0]
8	LZTWM	Lower zone tension water maximum storage (mm)	[10.0, 700.0]
9	LZFSM	Lower zone supplemental free water maximum storage (mm)	[5.0, 500.0]
10	LZFPM	Lower zone primary free water maximum storage (mm)	[100.0, 1200.0]
11	LZSK	Lower zone supplemental free water lateral drainage rate (day^{-1})	[0.010, 0.60]
12	LZPK	Lower zone primary free water lateral drainage rate (day^{-1})	[0.0010, 0.050]
13	PFREE	Fraction of water percolating from upper zone directly to lower zone free water (decimal fraction)	[0.0, 0.90]
14	RIVA	Riverside vegetation area (decimal fraction)	0.30
15	SIDE	Ratio of deep recharge to channel base flow (dimensionless)	0.0
16	RSERV	Fraction of lower zone free water not transferrable to lower zone tension water (decimal fraction)	0.0

1995; Gupta et al., 1998; Boyle et al., 2000). Three other parameters RSERV, RIVA, and SIDE are fixed at pre-specified values according to Brazil (1988).

The South Branch Potomac River basin near Springfield, West Virginia in the U.S. was chosen as the study area. The total drainage area upstream of the gauging station (U.S. Geological Survey Station No. 01608500) is about 3800 km^2. Historical precipitation, potential evapotranspiration and streamflow observations from January 1st, 1960 to December 31st, 1979 were obtained from the MOPEX database for this study, where MOPEX stands for Model Parameter Estimation Experiment (Duan et al., 2006). The average annual precipitation over this period is 1021 mm, average annual potential evapotranspiration is 762 mm, and average annual runoff is 39.5 m^3/s. The hydrological simulations were run at a 6-h time step over the entire data period. To evaluate model response to different parameters, we use mean absolute error (MAE) of the simulated and observed daily streamflow discharge (m^3/s) as the objective function, which is a measure of average errors:

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|Q_t^{\mathrm{fcs}} - Q_t^{\mathrm{obs}}\right| \quad (1)$$

where Q_t^{fcs} and Q_t^{obs} are simulated and observed streamflow discharge values at time t, N is the total number of observations. To reduce the influence of incorrect specification of initial conditions, the simulations from the first three months are excluded in the MAE calculation. Since the sensitivities of model parameters are dependent on the choice of objective functions, other objective functions such as root-mean-square error, Nash–Sutcliffe Efficiency may also be used in practical applications. This study focuses on the evaluation of the effectiveness and efficiency of different SA methods, which are not influenced by the choice of objective functions.

3.2. Experimental methods

3.2.1. SA methods

All SA methods as shown in Table 2, except Plackett–Burman (PB) (Plackett and Burman, 1946) and Fractional Factorial (FF) (Box and Hunter, 1961) screening methods, are employed to study the sensitivities of the thirteen SAC-SMA model parameters. PB and FF methods implemented in PSUADE are designed only for two-level design experiments, i.e., the parameters can only be evaluated at two fixed levels. Therefore, it is not suited for continuously varying parameters in SAC-SMA. In addition, those methods are effective only for linear or monotonic parameter–response relationship.

Correlation Analysis (CA) and Regression Analysis (RA) are traditional approaches that are extensively used to assess the strength of the association between two factors (e.g., parameter and response) due to their relatively simple theories (Crawford, 2006). CA measures parameter sensitivity by correlation coefficients, such as Pearson correlation coefficient (PEAR), Spearman rank correlation coefficient (SPEA) and Kendall tau rank correlation coefficient (KEND). These coefficients measure the strength of a linear or monotonic relationship between model parameters and model responses. In this study we take SPEA as the sensitivity measure for CA. RA evaluates parameter sensitivity by standard regression coefficient (SRC) of a regression function relating model parameters and model responses.

MOAT screening is a typical One-At-a-Time (OAT) method for parameter screening (Morris, 1991). Theoretic basis of this method is that the overall effect and interaction effect of each parameter can be approximated by the mean μ and standard deviation σ of the gradients of each parameter sampled from r MOAT paths. Campolongo et al. (2007) proposed a modified mean μ^*, which is an estimate of the mean of absolute gradients, to solve the problem of the effects of opposite signs in gradients. We use the modified mean μ^* (denoted as MOAT-1) and standard deviation σ (denoted as MOAT-2) as the MOAT sensitivity measures.

Multivariate Adaptive Regression Splines (MARS), Delta Test (DT), Sum-Of-Trees (SOT) and Gaussian Process (GP) screening methods can all be regarded as certain types of RSMs, from which one can derive relative scores of parameter overall effects. MARS is an extension of linear models, which makes use of linear regression, the mathematical construction of splines, the binary recursive partitioning and brute search intelligent algorithms (Friedman, 1991; Gutiérrez et al., 2009). It calculates parameter importance scores by refitting the model after dropping all terms involving the parameter in question and calculating the reduction in goodness-of-fit. The least important parameter is the one with the smallest impact on the model quality; similarly, the most important parameter is the one that, when omitted, degrades the model fit the most (Kahng et al., 2010).

DT was originally used for residual noise variance estimation. It is based on the hypotheses of the continuity of the regression function, i.e., if two sample points are close in the parameter space, the responses of these two points will be close enough in the response space. Or else, it can be explained by the influence of noise. DT was devised by Pi and Peterson (1994) for identifying parameter dependencies in continuous functions and was demonstrated and applied by Eirola et al. (2008) for parameter screening. It takes the subset of parameters that minimize the noise variance from all the parameter combinations as sensitive ones. However, this procedure needs an efficient and effective search algorithm to find this subset of parameter combinations. This

search process can be too time-consuming and usually as it is impossible to do an exhaustive search of all combinations. In PSUADE, DT chooses the best 50 subsets and uses them for scoring. It assesses the final choice using forward sweep and uses genetic algorithm to speed up the search.

SOT is the classification (or Bayesian) additive regression tree model based on recursive binary partitioning, which is another useful tool for parameter screening (Breiman et al., 1984; Chipman et al., 2010). Parameter space is recursively split by unbalanced binary tree according to the residual sum of squares of responses until per terminal node has minimum number of sample points. Total number of splittings for each parameter is taken as the ranking criterion.

GP characterizes simulation responses over the parameter space as a multivariate Gaussian distribution. A GP can be expressed as $\boldsymbol{Y} \sim GP(\boldsymbol{\mu}, \boldsymbol{C})$, i.e., random function $\boldsymbol{Y}$ can be specified by its mean function $\mu(\boldsymbol{X})$ and covariance function $C(\boldsymbol{X}, \boldsymbol{X}')$. Different kinds of mean and covariance functions lead to different GPs. Tpros, which is a program written by Gibbs and MacKay (1997) for regression problem using GP, is adopted in PSUADE for parameter screening. Theoretical basis of this approach is that points which are close in parameter space give rise to similar values of response values. A length scale can be found for each parameter that characterizes the distance in a particular parameter direction over which response is expected to vary significantly. A small length scale for a parameter means a more significant influence of this parameter on model response.

FAST, McKay main effect (McKay-1) and two-way interaction effect (McKay-2) analysis, and Sobol' sensitivity indices (Sobol) are all variance-based methods which can be used to quantify the main effect and interaction effect of the parameters. FAST was presented by Cukier et al. (1973) for nonlinear SA of multi-parameter model, in which conditional variances are represented by coefficients from the multiple Fourier series expansion of the response function and the ergodic theorem is applied to transform the multi-dimensional integral into a one-dimensional integral in evaluation of the Fourier coefficients.

McKay-1 makes ANOVA (i.e., analysis of variance)-like decomposition of response variances for calculating correlation ratio, which is a ratio of the variance of expectation conditioned on one parameter and the total variances of response (McKay, 1995). The significance of parameter main effect increases with the parameter correlation ratio. Tong (2005) extended the idea for main effect analysis to two-way interaction effect analysis for uncorrelated parameters (i.e., McKay-2). A high second-order correlation ratio of any two parameters means that they are taken together as important contributors to the response variability.

Another variance-based method in PSUADE is proposed by Sobol' (1993, 2001), which also makes ANOVA-like decomposition of response variances for calculating specific order sensitivity indices. In practice, only the first-order and second-order Sobol' indices are estimated since the number of interaction terms need to be computed for higher order indices will increase exponentially when the number of parameters increases. The statistic "total sensitivity indices", $S_{Ti} = S_i + S_{i,ci} = 1 - S_{ci}$, introduced by Homma and Saltelli (1996) offer a simple way of evaluating the total effects for each parameter. Where S_i (first-order indices) and $S_{i,ci}$ (high-order indices) represent the main effects and interaction effects of parameter i, respectively; and S_{ci} equals the sum of all the other terms except for the terms related to parameter i. PSUADE provides a response surface based Sobol' SA tool for calculating Sobol' first-order (Sobol-1), second-order (Sobol-2) and total (Sobol-t) indices, which makes the computational cost of Sobol' indices cheaper than the direct calculation based on the original model. The default response surface for Sobol' SA is MARS approach. We take the Sobol-1 and Sobol-t as the Sobol' sensitivity measures.

3.2.2. Sampling techniques

All SA methods described above must use sampling techniques to create samples of parameter sets. The most concerned problems in designing a deterministic experiment are whether the sample points are "space-filling" in the design space (Sacks et al., 1989) and how many sample points are sufficient for the experiment (Loeppky et al., 2009).

Among all the sampling techniques available in PSUADE (see Table 1), MC sampling has the longest history and is the most commonly used technique. MC generates sample points randomly from a probability density function over the parameter space. However, large sample size is required to be able to fully explore the parameter space.

To improve the representativeness of sample points, McKay et al. (1979) proposed LH sampling. For a n-dimension p-level parameter space, only p sample points are generated by LH sampling, whereby each level exists only once when sample points are projected to any single dimension. Sample size of replicated LH (rLH) sampling is $N = \lambda \times p$, where λ is the replication times. LH sampling became popular in the 1980s, and a lot of improvements have since been made to it. For example, Owen (1992) used OA to define the generalization of LH sampling. For a n-dimension p-level parameter space, an OA sampling of strength t ($t < n$) generates p^t sample points, such that all possible level combinations for every t-dimension p-level subspace occur exactly once. Sample size of replicated OA (rOA) sampling is $N = \lambda \times p^t$, where p should be a prime number or 4, and t is usually set to 2. Clearly, a strength one OA sampling is equivalent to a LH sampling. Meanwhile, Tang (1993) presented OALH sampling, which uses OA to construct LH. A strength t OALH sampling not only preserves the stratification properties of LH sampling in single-parameter space, but also in t-dimensional space. The sample size of OALH sampling is the same as that of OA sampling.

Besides the aforementioned sampling techniques, some other "space-filling" methods also have received much attention over the past few decades. For example, Karypis and Kumar (1998) introduced METIS method for partitioning large irregular graphs and large meshes, and computing fill-reducing orderings of sparse matrices, which is based on multilevel graph partitioning algorithms. Statnikov and Matusov (2002) described LP_τ (LPTAU) method for generating deterministic and uniformly distributed sequence of points in a multidimensional space, which provides a way to add more sample points to the initial sample with the same uniformity characteristics.

Also, there are some sampling techniques that were designed for specific SA methods, e.g., the sampling techniques for FAST analysis (Cukier et al., 1973), MOAT screening (Morris, 1991) (denoted same as corresponding SA methods), and the sampling technique proposed by Saltelli (2002) for Sobol' sensitivity indices (denoted as SOBOL). They are generally based on simple random sampling, and different conditions need to be satisfied for their sample sizes. FAST transforms a multi-dimensional integral into a one-dimensional integral. Different forms of transformation lead to different distributions of sample points. Minimum sample size of the classic FAST is determined by $N = 2 \times M_s \times \omega_{\max} + 1$, where M_s is the maximum harmonic which should be no less than 4 (usually taken to be 4 or 6) and $\omega_{\max}$ is the maximum frequency which is determined by the number of factors.

As for MOAT sampling, range of each parameter is partitioned into $p-1$ equal intervals, thus the parameter space is an n-dimension p-level orthogonal grid, where each parameter can take on values from these p predetermined values. First, r points ($r \times n$

sample matrix M_0, each row is a n-dimension sample point) are randomly generated from the orthogonal grid; and then, for each of the r points, other sample points are generated by perturbing one dimension at a time under a $p/[2 \times (p-1)]$ space step until all the n dimensions have been varied for only one time. Therefore, total sample points will be $(n + 1) \times r$.

Sampling technique of SOBOL is similar to that of MOAT. It starts with two random $r \times n$ sample matrices M_0 and M_{n+1} (each row is an-dimension sample point). For each of the r sample points, the ith (i from 1 to n) sample point is generated from both two matrices, where the ith column is the same as M_{n+1} while other columns are the same as M_0. Thus the total number of sample points will be $(n + 2) \times r$.

3.3. Experimental setup

In evaluating each SA method, we attempt to answer the following questions:

(1) Is the method capable of identifying sensitive and insensitive parameters correctly? (effectiveness)
(2) Given that a method is effective, what is the minimum number of samples for each specific sampling technique? (efficiency)

Table 4 shows the experimental setup for analyzing the effectiveness of different SA methods. Of the ten SA methods (McKay-1 and McKay-2 are taken as one method), the first seven are qualitative methods and the last three are quantitative ones. Qualitative methods provide a heuristic score to intuitively represent the relative sensitivity of parameters, while quantitative methods tell how sensitive the parameter is by computing the impact of the parameter on the total variance of model output. A brief description of different SA measures is given in Appendix A. Sampling techniques used for these SA methods are selected according to the recommendation from PSUADE user's manual. A rough rule of thumb about the sample size is that at least $10 \times n$ sample points are needed to identify key factors (i.e., parameters), where n is the number of experimental factors (Levy and Steinberg, 2010). In order to get reliable SA results, we set the maximum sample sizes to between 2777 and 3000, which are more than twenty times the required minimum sample size 130 ($=10 \times 13$). The detailed settings of sample sizes are illustrated as follows. Sample sizes of MC, METIS and LH sampling are set to 3000 since there are no specific requirements for them. Sample size of OA sampling is set to 2890 ($=10 \times 17^2$). As for FAST, the maximum harmonic $M_s = 4$, and the maximum frequency $\omega_{max} = 347$ when $n = 13$. Therefore, the minimum sample size of FAST is 2777. For MOAT and SOBOL sampling techniques, we use 200 replications, resulting in sample sizes of 2800 and 3000, respectively.

Table 4
Experimental setup for effectiveness analysis of SA methods.

SA method	SA measures	Sampling technique	Sample size
CA	Spearman rank correlation coefficient (SPEA)	MC	3000
RA	Standard regression coefficient (SRC)	MC	3000
MOAT	Modified mean and standard deviation (MOAT-1, MOAT-2)	MOAT	2800
MARS	MARS sensitivity score (MARS)	METIS	3000
SOT	SOT sensitivity score (SOT)	METIS	3000
DT	DT sensitivity score (DT)	METIS	3000
GP	GP sensitivity score (GP)	METIS	3000
FAST	FAST first-order index (FAST)	FAST	2777
McKay	McKay first-order correlation ratio (McKay-1)	rLH	3000
McKay	McKay second-order correlation ratio (McKay-2)	rOA	2890
Sobol	Sobol' first and total indices (Sobol-1, Sobol-t)	SOBOL	3000

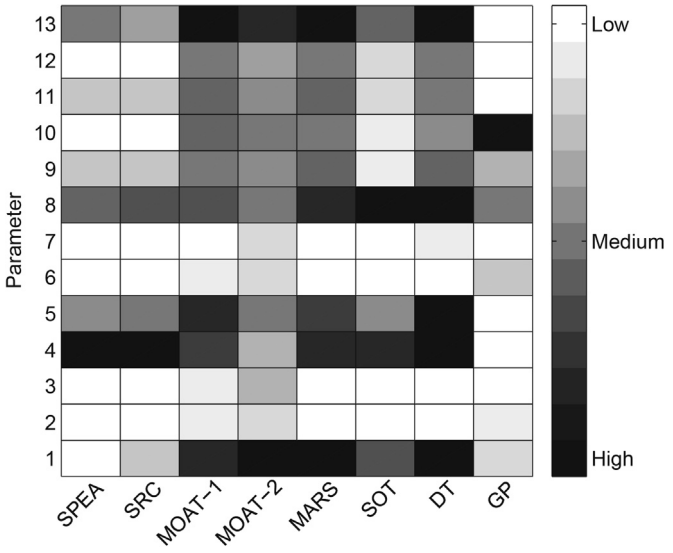

Fig. 2. Parameter sensitivity rankings of different qualitative SA methods.

All suitable sampling techniques as described above are used to explore the efficiency of effective SA methods. Sample size starts from a large one (result of which will be taken as the benchmark) and decreases sequentially to the minimum size in order to know exactly how many sample points are sufficient for specific sampling techniques. Detailed experimental setups for efficiency analysis of different SA methods are illustrated together with their results.

4. Results and discussion

4.1. Evaluation of qualitative SA methods

4.1.1. Effectiveness of qualitative SA methods

The first question we want to explore for qualitative SA methods is if all these methods are effective in screening out sensitive parameters from the insensitive ones. In order to compare all the qualitative SA results in a concordant way, we ranked all thirteen parameters based on respective SA measures for different methods (see Fig. 2), where rankings are represented by brightness of the color. The darker brightness means the higher ranking (i.e., more sensitive parameter). As shown in Fig. 2, parameter sensitivity rankings of different SA measures vary from each other. Overall, parameter sensitivity rankings of SPEA, SRC and GP are inconsistent with those obtained by other methods. In fact, it should be noted that CA's sensitivity index SPEA is a rank-transformed statistic which only works with monotonic relationships between model parameters and model responses, while RA's sensitivity index SRC is based on a linear hypothesis. Most hydrologic models, however, are nonlinear and parameter–response relationship is non-monotonic. For example, Duan et al. (1992) drew the response surface of SIXPAR model. Although the SIXPAR model is very simple, the response surface showed very strong non-monotonic, nonlinear interactions between the parameters. Consequently, CA and RA may not give the correct assessment of model parameter sensitivities. According to parameter sensitivity rankings of all qualitative SA measures except CA, RA and GP, the top five parameters are 1, 4, 5, 8 and 13, which can be taken as highly sensitive parameters; the middle four

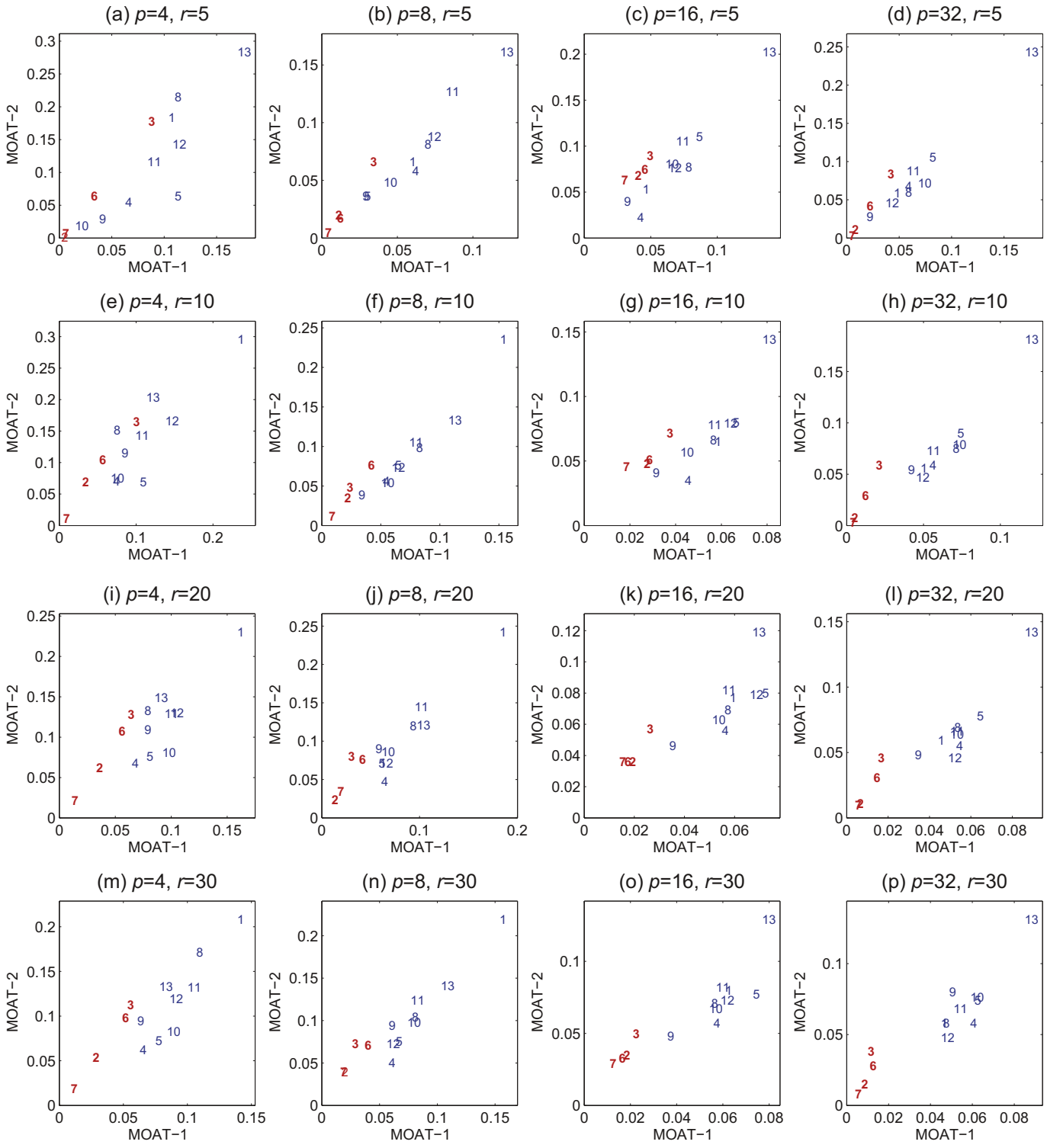

Fig. 3. Sensitivity analyzing results of MOAT screening using different replication times r and different levels p: Numbers in red bold font are supposed to be insensitive parameters. The closer the parameter to upper right, the more sensitive the parameter is. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parameters are 9, 10, 11 and 12, which can be taken as marginally sensitive parameters; and the last four parameters are 2, 3, 6 and 7, which can be taken as insensitive parameters. As for SPEA and SRC indices, more parameters besides 2, 3, 6 and 7 are also identified as insensitive ones. The most striking divergence between them is the classification of parameter 1, which is identified as highly sensitive by most indices, but is seen as insensitive by SPEA and marginally sensitive by SRC. In addition, insensitive parameters identified by GP are 3, 4, 5, 7, 11, 12 and 13, which include not only insensitive parameters but also marginally and highly sensitive parameters identified by other measures. Furthermore, we found that GP takes more time to compute than other methods.

Model parameter sensitivities are heavily impacted by several factors, including the choice of analysis methods, evaluation metrics, and system physical characteristics (Tang et al., 2007; van Werkhoven et al., 2008). In our experiment, sensitivity categories of most qualitative SA methods are consistent when using the same evaluation metric for the study area. The rationality of SA results thus can be analyzed from the perspective of physical characteristics of the study area. The South Branch Potomac River basin is a humid watershed and is usually wet during most of the year. In this watershed, surface runoff makes up a large proportion of the total runoff. Therefore, parameters related to surface runoff generation and evapotranspiration should be sensitive. In all thirteen parameters, parameters 4 and 5 are permanent and temporary impervious fraction of the watershed area respectively, which contribute to direct runoff. Parameters 1 and 8 are maximum capacities of the upper zone tension water storage and the lower zone tension water storage respectively, which not only have direct relationships with evapotranspiration of the upper zone and lower zone, but also have great influences on the production of surface runoff and base flow. Parameter 13 is the percentage of percolated water which is available to the lower free water storages before all lower zone tension water deficiencies are satisfied. It controls the proportion of lower zone tension water and free water, and indirectly influences the evapotranspiration of the lower zone. From a physical perspective, it is reasonable that these five parameters are identified as highly sensitive. On the other hand, parameters 9, 10, 11, and 12 are those related to the generation of primary and supplemental base flow. Parameters 2 and 3 are those related to the process of interflow drainage, and parameters 6 and 7 are related to percolation from upper zone to lower zone. It is found that parameters 9, 10, 11, and 12 are more sensitive than parameters 2, 3, 6, and 7, which indicates that base flow plays a more important role than the interflow in this area. This makes sense since the source of runoff is mainly come from base flow in dry season.

Furthermore, it can be observed from the results of MOAT-1 and MOAT-2 that parameters with low overall effects also have low interaction effects. Although this conclusion is model-specific, it reminds us about the importance of interaction effect. Comparing results of MARS, DT and SOT, it is obvious that parameter sensitivity scores from MARS and DT are very close, with both measures giving a higher score for marginal sensitive parameters than SOT. Therefore, it could be inferred that MARS and DT screening may be good at identifying insensitive parameters since the differences between marginal and insensitive parameters are significant. On the contrary, SOT screening may be good at identifying highly sensitive parameters due to the big differences between highly and marginal sensitive parameters.

4.1.2. Efficiency of qualitative SA methods

Since CA, RA and GP are regarded as ineffective SA methods in this experiment, they were not analyzed further in this paper. Other methods are not only effective but also have great flexibilities. Therefore, it is necessary to compare their performances with various sampling techniques and different sample sizes for their efficiency. This is the second question we will explore for qualitative SA methods.

(1) MOAT screening

Recall that sample points of MOAT are taken from n-dimension p-level orthogonal grid by the way of perturbing one parameter at a time under a $p/[2 \times (p-1)]$ space step, and sample size of MOAT is determined by $N = (n + 1) \times r$. In order to know exactly the least sample points for effective screening, different replication times r ($r = 5, 10, 20, 30$) are set while using different levels p ($p = 4, 8, 16$ and 32). Evaluation criterion is whether sensitive and insensitive parameters can be identified successfully. SA results of different combinations of levels and replications are shown in Fig. 3. The sensitive (blue numbers) and insensitive (red bold numbers) labels in Fig. 3 were determined according to the consensus results from Fig. 2. Overall the results from MOAT experiments shown in Fig. 3 are not very consistent. Specifically, the results in row 1 ($r = 5$) are quite different from the results in other rows ($r = 10, 20, 30$). This suggests clearly that 5 replications are not enough to produce reliable results for screening SAC-SMA model parameters. We also notice that the SA results for different levels ($p = 4, 8, 16$, and 32) are not stable under the same replication times. Especially, $p = 4$, as used by Morris (1991) and Campolongo et al. (2007) and $p = 8$ seem to be insufficient and produce different SA results. This study suggests that for MOAT screening to work properly, levels should be 16 or 32 and replication times should be at least 20 or more, i.e., requiring no less than 280 model runs in this case.

(2) RSM-based screening

To investigate which sampling techniques are more appropriate and how many sample points are sufficient for MARS, DT and SOT screening, six different sampling techniques, MC, LH, OA, OALH, LPTAU and METIS, are compared. Further, different sample sizes are tested for each sampling technique. SA results of MARS, DT and SOT screening are given in Figs. 4–6, respectively. In these figures, sample sizes of the last column in each subfigure are the maximum sample sizes we set, the results of which are taken as the benchmark for effectiveness. For MC, LH, LPTAU and METIS sampling, sample sizes of other columns start from a small number with an increment of 100 sample points. For OA and OALH sampling, recall that sample size is $N = \lambda \times p^t$. Given $\lambda = 1$ and $t = 2$, sample size N is determined by the prime number p. Therefore, sample sizes 169, 289, 361, 529, 841, and so on, are generated from prime number 13, 17, 19, 23, 29, and so on. When the insensitive parameters identified by three consecutive columns identical with the last column, the first sample size of the three is thought to be sufficient for the corresponding sampling technique. From the figures, in each subfigure, the sample size of the fourth column from the right is the minimum sample size required by corresponding sampling technique. It therefore implies that MC, OA and OALH sampling are the most suitable sampling techniques for MARS screening, and about 400 sample points are sufficient. At the same time, MC, OA and OALH sampling are also the most suitable sampling techniques for DT screening, and about 400 sample points are sufficient. MC and OALH sampling are the most suitable sampling techniques for SOT screening, and about 600 sample points are sufficient. Overall, MC and OALH sampling are more suitable than others for RSM-based screening methods.

More sample points are required by SOT screening than MARS and DT screening to identify insensitive parameters, no matter which sampling technique is used, which supports the inference that MARS and DT are good at identifying insensitive parameters while SOT are good at identifying highly sensitive parameters. Furthermore, sample sizes required by different sampling techniques have bigger differences for SOT screening than for MARS and DT screening.

4.2. Evaluation of quantitative SA methods

All quantitative SA methods tested here are based on variance decomposition of model responses. These methods provide quantitative measures of how much each varying parameter contributes to the overall variance of the model response.

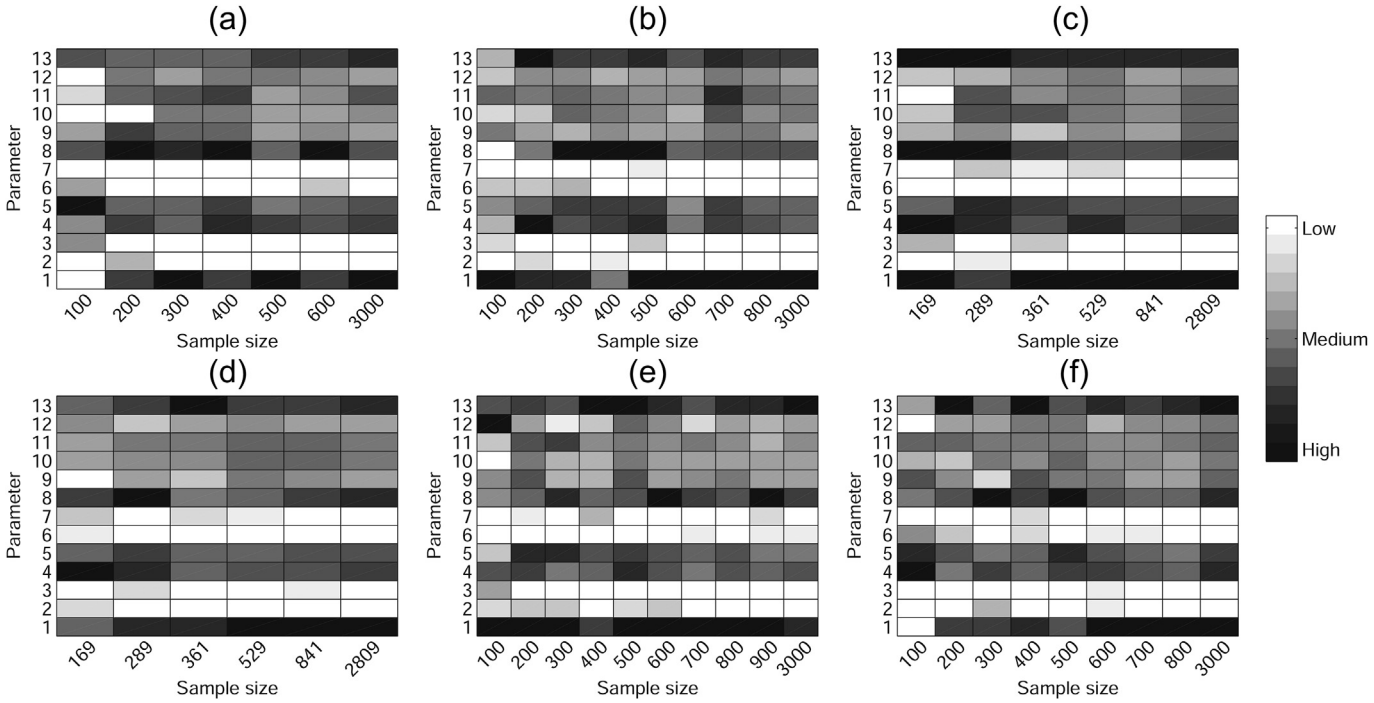

Fig. 4. Parameter sensitivity rankings of MARS screening using (a) MC sampling, (b) LH sampling, (c) OA sampling, (d) OALH sampling, (e) LPTAU sampling, and (f) METIS sampling.

4.2.1. Effectiveness of quantitative SA methods

The first question about the quantitative SA methods is also their effectiveness. We ranked all thirteen parameters for different quantitative SA methods and the results are shown in Fig. 7. Measures of FAST, McKay-1 and Sobol-1 are representations of parameter main effects (first-order effects). It is apparent that insensitive parameters identified by these three measures are consistent, i.e., parameters 2, 3, 6 and 7, which are also consistent with the qualitative SA results. Furthermore, parameters with low main effects may have significant interaction effects with other parameters, which should also be treated as important ones. More attention should be paid to those parameters in order to avoid the so-called type II error, i.e., important parameters are treated as unimportant ones. In all these quantitative SA measures, McKay-2 is capable of presenting two-way interaction effects (first-order effects plus second-order effects),

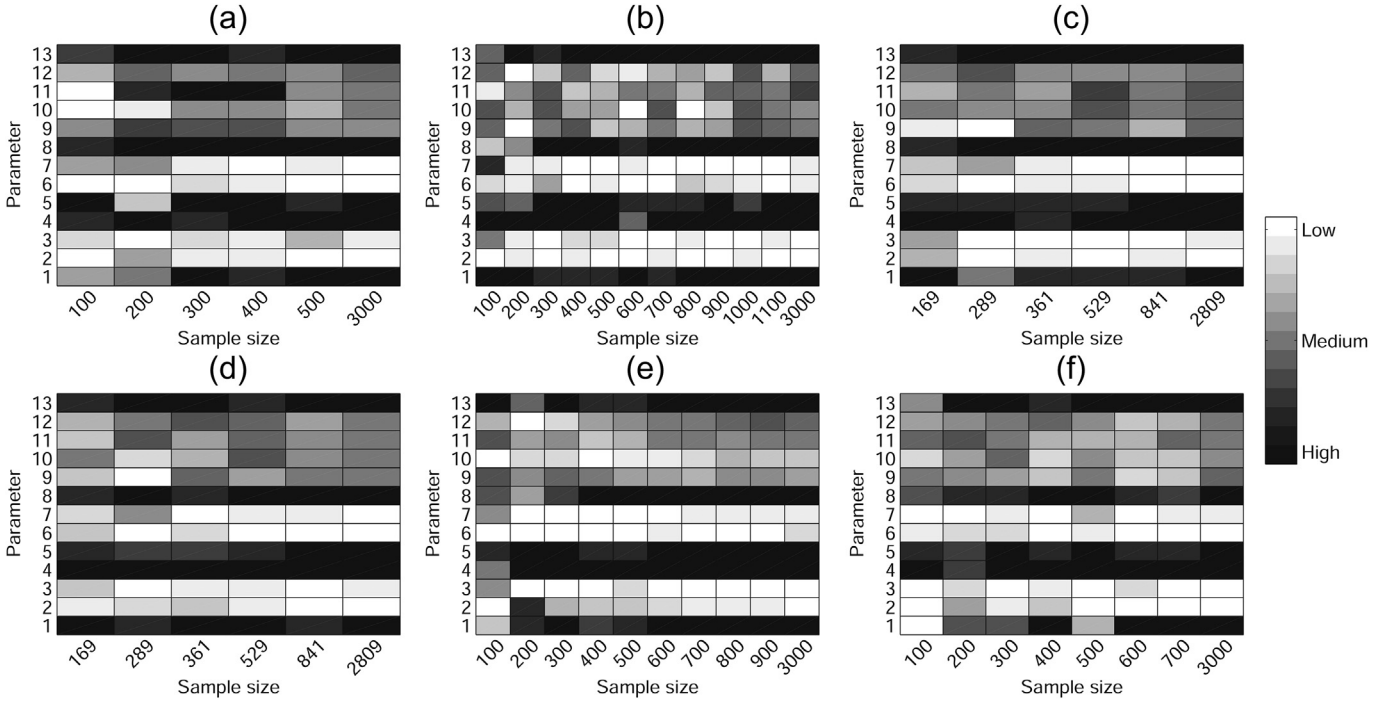

Fig. 5. Parameter sensitivity rankings of DT screening using (a) MC sampling, (b) LH sampling, (c) OA sampling, (d) OALH sampling, (e) LPTAU sampling, and (f) METIS sampling.

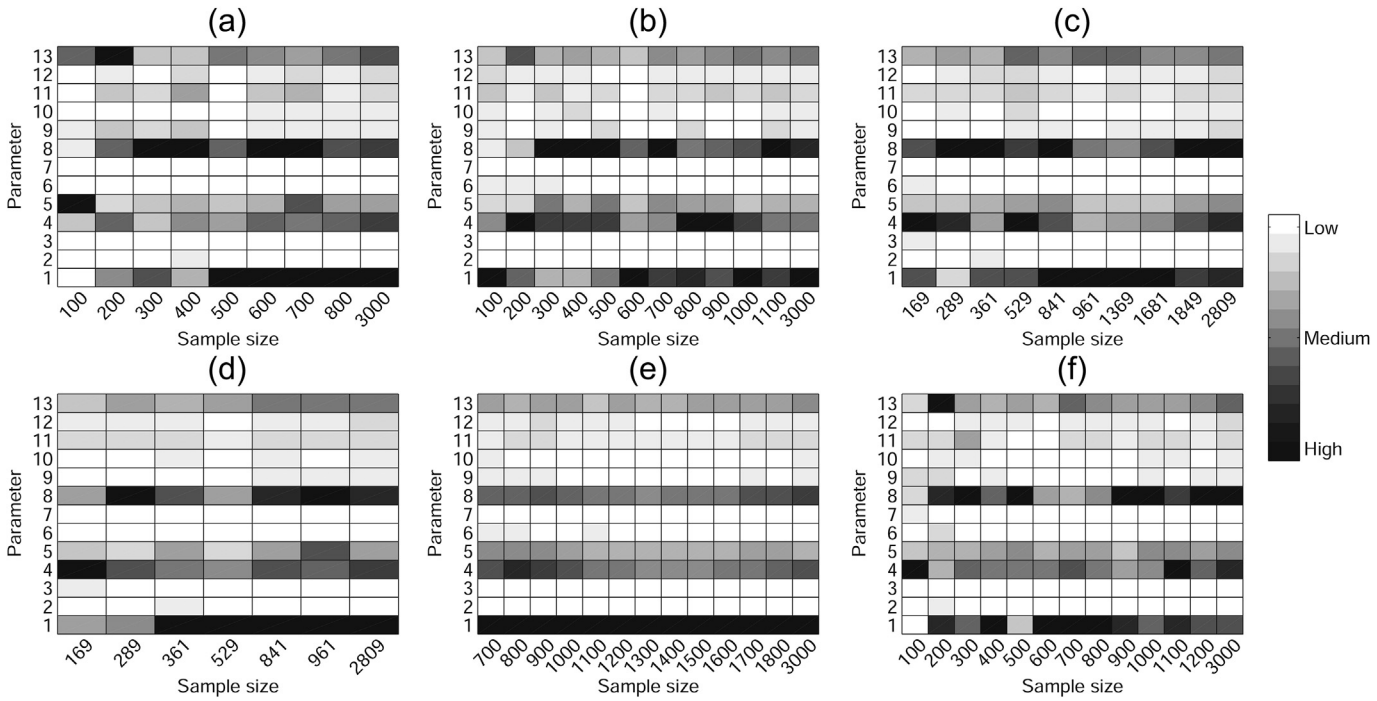

Fig. 6. Parameter sensitivity rankings of SOT screening using (a) MC sampling, (b) LH sampling, (c) OA sampling, (d) OALH sampling, (e) LPTAU sampling, and (f) METIS sampling.

and Sobol-t is able to display total effects (first-order effects plus all interaction effects). It is observed that parameters with lower main effects also have lower interaction effects. Overall, all these quantitative SA measures are effective for identifying parameter sensitivities.

4.2.2. *Efficiency of quantitative SA methods*

Although FAST is effective for SA, sampling technique of this method is fixed and minimum sample size is not changeable when number of parameters is already determined. Therefore, we will not give further consideration about FAST here. What are the

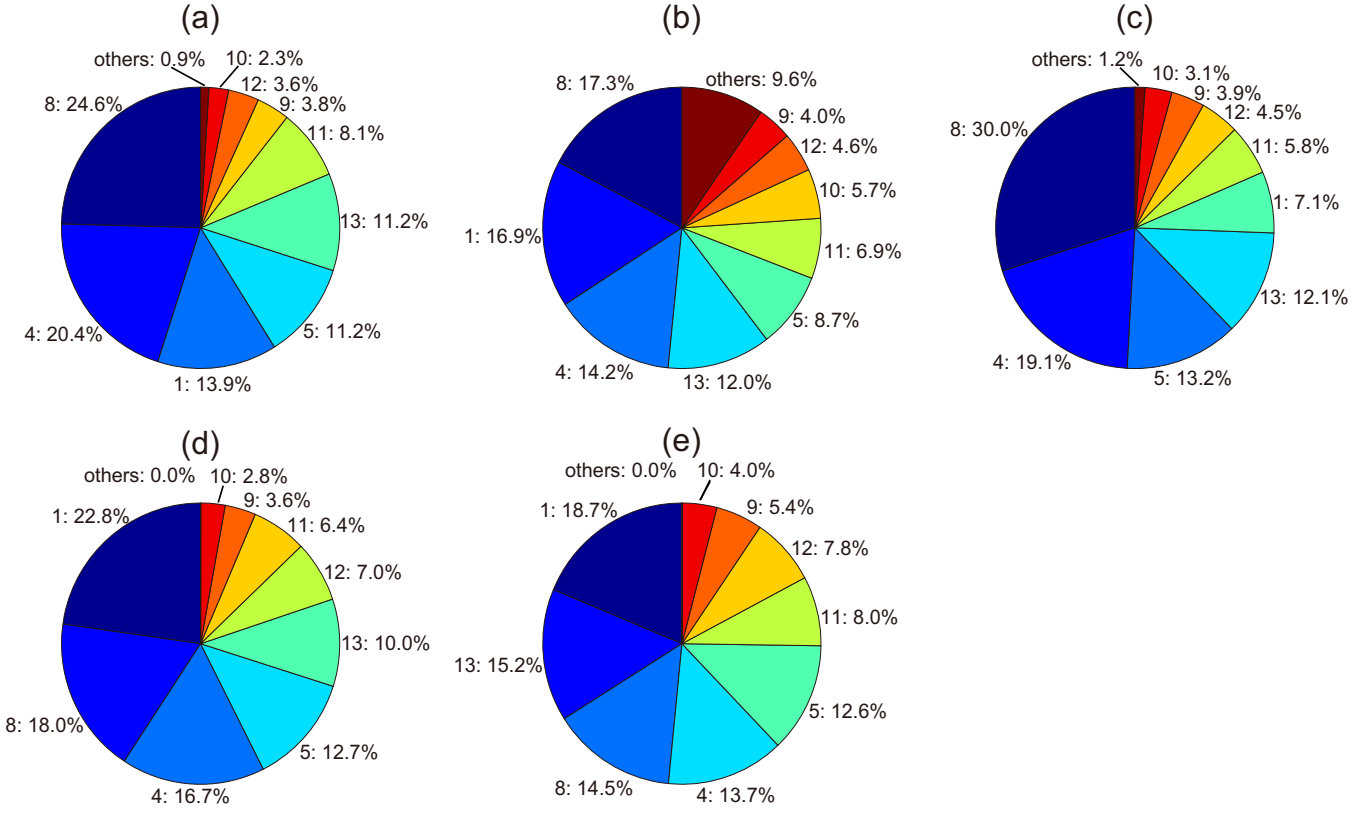

Fig. 7. Parameter sensitivity rankings of different quantitative SA methods. (a) FAST, (b) Mckay-1, (c) Mckay-2, (d) Sobol-1, and (e) Sobol-t.

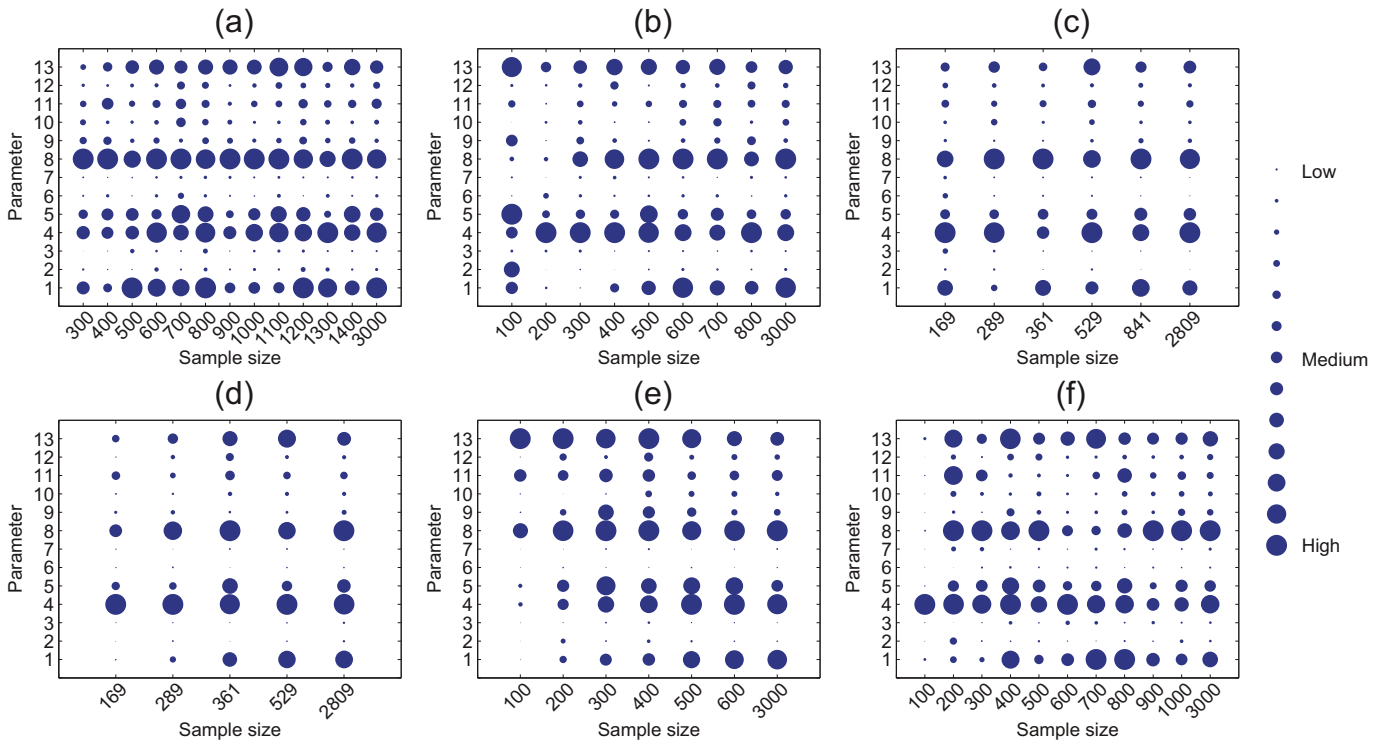

Fig. 8. Parameter sensitivity rankings of McKay main effect analysis using (a) MC sampling, (b) rLH sampling, (c) OA sampling, (d) OALH sampling, (e) LPTAU sampling, and (f) METIS sampling.

efficiencies of McKay main effect and two-way interaction effect analysis, and Sobol' sensitivity indices are addressed here.

(1) McKay main effect and two-way interaction effect analysis

rLH sampling was recommended for McKay main effect analysis based on PSUADE user's manual, here we set the replication times to 2. Other feasible sampling techniques such as MC, OA, OALH, LPTAU and METIS sampling were also tried. Different sample sizes were set for each sampling technique. Parameter sensitivity rankings are given in Fig. 8. The fourth column from the right of each subfigure is the minimum sample size required by the corresponding sampling technique. It then can be observed that sample points needed by MC, LH, OA, OALH, LPTAU and METIS sampling are 1200, 600, 361,289, 400 and 800, respectively. It is quite clear that less sample points are needed by OA, OALH and LPTAU sampling.

rOA sampling was advocated in PSUADE user's manual for two-way interaction effect analysis. Other sampling techniques feasible for this method including MC, LH, OALH, LPTAU and METIS sampling were also tested. In addition, in order to get reliable results, sample points required by this method should be no less than 1000. Sample sizes of MC, LH, LPTAU and METIS sampling are set to 1000,whereas the sample sizes of OA and OALH sampling are set to 1058 ($=2 \times 23^2$) and 1369 ($=1 \times 37^2$), respectively. SA results are represented in Fig. 9. It can be observed from Fig. 9 that most correlation ratios of every two parameters have high degree of consistency for all results except for the results obtained by rOA sampling. Correlation ratios are almost the same when using replicated OA sampling, which demonstrates that OA sampling is very tricky and more replication times are required for identifying parameter two-way interaction effects.

(2) Sobol' first-order and total effects analysis

Sobol' first-order and total sensitivity indices can be estimated at the cost of $N = (n + 2) \times r$ model evaluations when using SOBOL sampling. Different replication times r were tried to evaluate the efficiency of this method. In Fig. 10, parameter sensitivity analyzing results of Sobol' first-order indices (black bar) and total indices (black bar plus white bar) of different sample sizes are compared. Overall, when sample size is larger than 1050, it is clear that parameters with low first-order (main) effect also have low interaction effect and insensitive parameters 2, 3, 6 and 7 can be correctly identified.

Comparing to McKay's main effect and two-way interaction effect analysis, Sobol' method needs more sample points due to its computational complexity for high-order Sobol' index terms.

5. Summary and conclusions

This study conducted a comprehensive evaluation of the effectiveness and efficiency of various SA methods available from PSUADE by using the SAC-SMA model as a test problem. The strengths and limitations of several qualitative and quantitative SA methods are explored. Based on the highly consistent results of different SA methods, parameters 1 (UZTWM), 4 (PCTIM), 5 (ADIMP), 8 (LZTWM) and 13 (PFREE) can be regarded as highly sensitive parameters; parameters 9 (LZFSM), 10 (LZFPM), 11 (LZSK) and 12 (LZPK) as marginally sensitive parameters; parameters 2 (UZFWM), 3 (UZK), 6 (ZPERC) and 7 (REXP) as insensitive parameters in the thirteen parameters we considered. In model calibration, the insensitive parameters may be fixed at prescribed values in order to improve model parameter identifiability.

Some general conclusions can be drawn when applying different SA methods available in PSUADE. For qualitative SA methods: (1) Traditional methods, such as correlation and regression analysis, are not suitable for nonlinear and non-monotonic problems like the SAC-SMA model. (2) GP screening is ineffective in screening out

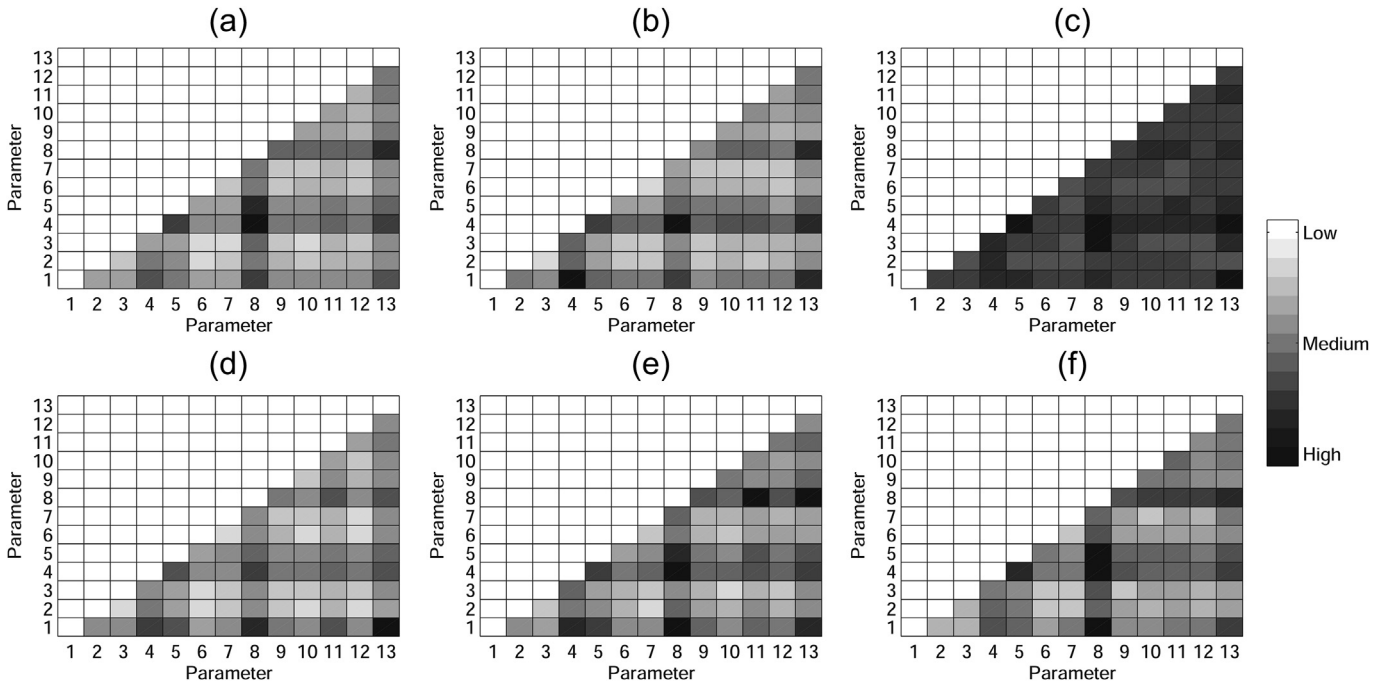

Fig. 9. Sensitivity analyzing results of McKay two-way interaction effect analysis using (a) MC sampling, (b) LH sampling, (c) rOA sampling, (d) OALH sampling, (e) LPTAU sampling, and (f) METIS sampling.

sensitive parameters. (3) MOAT screening can provide a qualitative evaluation of overall and interaction effects at a low cost. Minimum sample points needed by MOAT is 280 when the level is set at $p = 16$ or $p = 32$. But MOAT results are not robust under different combination of levels and replication times; (4) Other sensitivity screening methods based on non-parametric RSM have different characteristics. MARS and DT screening are efficient at identifying insensitive parameters, whereas SOT is good at identifying highly sensitive parameters. MC, OA and OALH sampling are more appropriate for MARS screening, and sample points required by them are about 400; meanwhile, MC, OA and OALH sampling are also the most suitable sampling techniques for DT screening, and

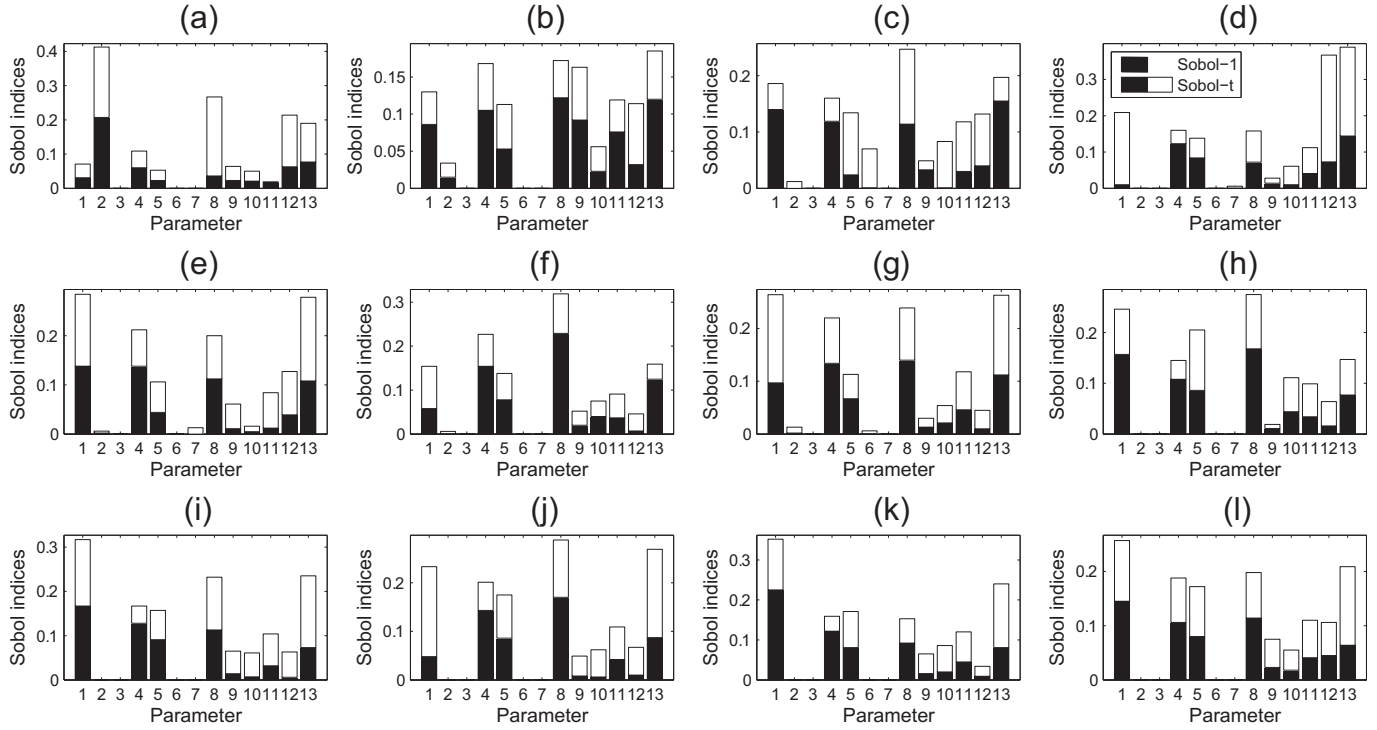

Fig. 10. Sensitivity analyzing results of Sobol' first-order effects (black bar) and total effects (black bar plus white bar). (a) $r = 20$ ($N = 300$), (b) $r = 30$ ($N = 450$), (c) $r = 40$ ($N = 600$), (d) $r = 50$ ($N = 750$), (e) $r = 60$ ($N = 900$), (f) $r = 70$ ($N = 1050$), (g) $r = 80$ ($N = 1200$), (h) $r = 90$ ($N = 1350$), (i) $r = 100$ ($N = 1500$), (j) $r = 110$ ($N = 1650$), (k) $r = 120$ ($N = 1800$), and (l) $r = 200$ ($N = 3000$).

sample points required by them are about 400; and MC and OALH sampling are fit for SOT screening, and sample points required by them are about 600.

For quantitative SA methods: (1) FAST and McKay main effect analysis are capable of computing parameter first-order effects. Minimum sample size of classic FAST method is not changeable when the number of parameters is determined, in this case at 2777. OA, OALH and LPTAU sampling are more appropriate for McKay main effect analysis; sample points required by them are 361, 289 and 400, respectively. (2) McKay two-way interaction analysis can be used for analyzing first-order plus second-order effects. MC, LH, OALH, LPTAU and METIS sampling are more suitable than OA sampling; 1000 sample points are sufficient for identifying two-way interaction effects of this experiment. (3) Theoretically, Sobol' method can compute sensitivity indices of any orders. In practice, the computational cost may be too expensive for indices larger than second-order, PSUADE allows computation for commonly used Sobol' first-order, second-order and total sensitivity indices. To identify parameter sensitivities correctly, minimum sample points needed by SOBOL sampling are 1050.

In general, qualitative SA methods are more efficient for parameter screening than quantitative ones. However, they cannot give a quantitative description for specific order sensitivity effects. On the other hand, quantitative SA methods are more accurate and robust than qualitative ones, but more sample points are required by them. For a complex system model with a lot of parameters, qualitative SA methods can be first used for a rough parameter screening, which will prune the most insensitive parameters with low evaluation costs. Then quantitative SA methods can be adopted for a further SA of this simplified system model.

Acknowledgments

This study is supported by the National Basic Research Program of China (973 Program) (No. 2010CB428402) and National Natural Sciences Foundation of China (41075075). The work by the authors from LLNL was performed under the auspices of the U.S. Department of Energy under Contract No. DE-AC52-07NA27344. We would like to acknowledge the valuable suggestions from Dr. Xuesong Zhang. We also want to thank three anonymous referees for their thorough and constructive reviews.

Appendix A. Sensitivity measures

A.1 Spearman rank correlation coefficient (SPEA)

SPEA is denoted by ρ and is defined as:

$$\rho_{x_i,y} = \frac{\sum_{k=1}^{N}\left(x_i^k - \overline{x_i}\right)\left(y^k - \overline{y}\right)}{\sqrt{\sum_{k=1}^{N}\left(x_i^k - \overline{x_i}\right)^2 \cdot \sum_{k=1}^{N}\left(y^k - \overline{y}\right)^2}} \tag{A.1}$$

where x_i is the rank of the ith input X_i, y is the rank of the output Y, and k represents the kth sample point. The higher the $\rho_{xi,y}$ value, the more sensitive is input X_i.

A.2 Standard regression coefficient (SRC)

The generalized form of a linear regression model is:

$$\widehat{Y_k} = b_0 + \sum_{i=1}^{n} b_i X_i^k \tag{A.2}$$

While the actual computational model output is expressed as

$$Y_k = b_0 + \sum_{i=1}^{n} b_i X_i^k + \varepsilon_k \tag{A.3}$$

where b_i is the regression coefficient of the ith input X_i, and ε_k is the error term between the computational model output and the regression results of the kth sample. Under the assumption of Gaussian errors, the regression coefficient can be computed using the least squares approach. Utilizing the means and standard deviations of input and output, the regression model is usually normalized to

$$\frac{\widehat{Y_k} - \overline{Y}}{\widehat{s}} = \sum_{i=1}^{n} \frac{b_i \widehat{s}_i}{\widehat{s}} \frac{X_i^k - \overline{X_i}}{\widehat{s}_i} \tag{A.4}$$

where SRC $= b_i\widehat{s}_i/\widehat{s}$ is defined as standard regression coefficient, and

$$\widehat{s} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(Y_k - \overline{Y}\right)^2} \tag{A.5}$$

and

$$\widehat{s}_i = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(X_i^k - \overline{X_i}\right)^2} \tag{A.6}$$

are the standard deviations of Y and X_i, respectively. The higher the SRC value, the more sensitive is input X_i.

A.3 Morris One-At-a-Time(MOAT) screening

Assume that we have a n-dimension p-level orthogonal input space, where each X_i may take on values from $\{0, 1/(p-1), 2/(p-1), \ldots, 1\}$. The elementary effect of the ith input is defined as

$$d_i = \left(f(X_1, \ldots, X_{i-1}, X_i + \Delta, X_{i+1}, \ldots, X_n) - f(X)\right)/\Delta \tag{A.7}$$

where Δ is a predetermined multiple of $1/(p-1)$. When p is even, usually $\Delta = p/[2(p-1)]$. After repeating this procedure r times, we can get the final Morris measures of the ith input

$$\mu_i = \sum_{j=1}^{r} d_i(j)/r \tag{A.8}$$

and

$$\sigma_i = \sqrt{\sum_{j=1}^{r}\left(d_i(j) - \mu_i\right)^2/r} \tag{A.9}$$

where μ_i and σ_i are the mean and standard deviation of d_i, respectively. A revised mean was given by Campolongo et al. (2007) as

$$\mu_i^* = \sum_{j=1}^{r}|d_i(j)|/r \tag{A.10}$$

For MOAT method, the higher the μ_i (or μ_i^*) value, the more sensitive is input X_i. On the other hand, the higher the σ_i value, the more interaction input X_i has with other inputs.

A.4 Multivariate Adaptive Regression Splines (MARS) screening

The MARS model can be represented as:

$$Y = f(X) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+ \quad \text{(A.11)}$$

where a_0 is a constant, a_m are fitting coefficients, M is the number of basis functions, K_m is the number of knots, s_{km} takes on values of either 1 or −1 and indicates the right/left sense of the associated step function, $v(k,m)$ is the label of the independent variable, and t_{km} indicates the knot location.

MARS builds a model in two phases: the forward and the backward pass, which is the same as that used by recursive partitioning of trees. The forward pass usually builds an overfit model using all input variables, while the backward pass prunes the overfit model by removing one input variable from the model at a time. The index called generalized cross-validation (GCV) is then computed for both the overfit model and the pruned model:

$$\mathrm{GCV}(M) = \frac{1}{N} \frac{\sum_{k=1}^{N} \left(Y_k - \hat{Y} \right)^2}{\left[1 - \frac{C(M)}{N} \right]^2} \quad \text{(A.12)}$$

with

$$C(M) = 1 + c(M)d \quad \text{(A.13)}$$

where N is the number of observations in the data set, k is the number of non-constant terms, d is the effective degrees of freedom, and $c(M)$ is a penalty for adding a basis function. The increase in GCV values between the pruned model and the over-fitted model is considered as the importance measure of the removed variable (Steinberg et al., 1999). The score of the ith ($i = 1, 2, \ldots, n$) variable is given by

$$\mathrm{score}(i) = \frac{\Delta g(i)}{\max\{\Delta g(1), \Delta g(2), \ldots, \Delta g(n)\}} \times 100 \quad \text{(A.14)}$$

where $\Delta g(i)$ is the increase in GCV when ith variable is removed. The larger the GCV increase, the more important is the removed variable.

A.5 Delta Test (DT) screening

DT model is a method based on nearest neighbors (NN) for estimating the variance of the residuals. Assume that we have n input variables, and sample points $X_k \in [0,1]^n$ for $1 \le k \le N$. Let $Y_k = f(X_k) + \varepsilon_k$, where f is a continuous function with bounded first and second partial derivatives, and the residuals $\varepsilon_k \sim (0, \sigma^2)$. Then the points $(X_k, Y_k)_{k=1}^N$ comprise imitation data set. Let the DT metric that restricted to the variable subset space S be

$$\delta_S = \frac{1}{N} \sum_{k=1}^{N} \left(Y_k - Y_{N_S(k)} \right)^2 \approx \mathrm{Var}(\varepsilon) \quad \text{(A.15)}$$

where the nearest neighbor of kth sample is

$$N_S(k) = \mathrm{argmin}_{l \neq k} \|X_k - X_l\|_S^2 \quad \text{(A.16)}$$

and the semi-norm

$$\|X_k - X_l\|_S^2 = \sum_{p \in S} \left(X_k^{(p)} - X_l^{(p)} \right)^2 \quad \text{(A.17)}$$

Thus the DT metrics for all $2^n - 1$ non-empty variable subsets can be calculated. PSUADE takes the first 50 subsets which have the lowest value of DT metrics for sensitivity scoring. The score of the ith (i=1,2,…,n) variable is given by

$$\mathrm{score}(i) = \frac{\sum_{m=1}^{50} \delta_S^{(m)} \times I_i^{(m)}}{\sum_{m=1}^{50} \delta_S^{(m)}} \times 100 \quad \text{(A.18)}$$

where $\delta_S^{(m)}$ is the DT metric of the mth subset; $I_i^{(m)} = 1$ if the ith variable is included in the mth subset, or else $I_i^{(m)} = 0$. A higher score means a more sensitive parameter.

A.6 Sum-Of-Trees (SOT) screening

A SOT model is fundamentally an additive model with multivariate components (Chipman et al., 2010). Let T denotes a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with each of the b terminal nodes of T. Thus the SOT model can be represented as

$$Y = \sum_{j=1}^{m} g(X; T_j, M_j) + \varepsilon \quad \text{(A.19)}$$

where for each binary regression tree T_j and its associated terminal node parameters M_j, $g(X; T_j, M_j)$ is the function which assigns $\mu_{ij} \in M_j$ to X, m is the total number of trees, and $\varepsilon \sim N(0, \sigma^2)$.

In PSUADE, residual sum of squares are used as the criteria for node splitting. Variable which has the maximum decrease of residual sum of squares will be chosen to split the node. The splitting process will not be stopped until per terminal node has minimum number of data points. Total number of splittings for each variable is then taken as the scoring criterion of sensitivity. The score for ith input variable is expressed as

$$\mathrm{score}(i) = \frac{p(i)}{\max\{p(1), p(2), \ldots, p(n)\}} \times 100 \quad \text{(A.20)}$$

where $p(i)$ is the number of splittings for ith variable. The more splittings the variable has, the more sensitive is the variable.

A.7 Gaussian Process (GP) screening

The joint distribution of the random variables $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ is a GP:

$$P(\boldsymbol{Y} | \boldsymbol{C}, \boldsymbol{X}) = \frac{1}{Z} \exp\left[-\frac{1}{2} (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}))^T \boldsymbol{C}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X})) \right] \quad \text{(A.21)}$$

with $\boldsymbol{C} = \{C(X_k, X_l; \Theta)\}_{k,l=1}^N$ is a parameterized covariance function with hyperparameters Θ, $\boldsymbol{\mu}$ is the mean function given a-priori, and Z is the normalization factor.

A form of covariance function in Gibbs and MacKay (1997) is:

$$C(X_k, X_l; \Theta) = \theta_1 \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{\left(X_k^{(i)} - X_l^{(i)} \right)^2}{r_i^2} \right\} + \theta_2 + \varepsilon_{kl}(X_k, X_l) \quad \text{(A.22)}$$

where θ_1 is the hyperparameter gives the overall vertical scale, θ_2 is the hyperparameter gives the vertical uncertainty, $\varepsilon_{kl}(X_k, X_l)$ is the noise model, $X_k^{(i)}$ and $X_l^{(i)}$ are the ith components of sample points X_k and X_l respectively, and r_i is the length scale that characterizes

the distance in the direction of ith variable over which Y is expected to vary significantly. PSUADE takes the length scales as the scoring criteria. The score for ith input variable is expressed as

$$\text{score}(i) = \frac{1/r_i}{\max\{1/r_1, 1/r_2, \ldots, 1/r_n\}} \times 100 \tag{A.23}$$

Thus a small length scale of the ith variable means that this variable has significant influence on the response Y, i.e., ith variable is sensitive.

A.8 Fourier Amplitude Sensitivity Test (FAST) analysis

Let us consider the function $Y = f(\boldsymbol{X}) = f(X_1, X_2, \ldots, X_n)$, where $X_i \in [0,1]; i = 1, 2, \ldots, n$. The key idea of FAST is applying the ergodic theorem to transform the n-dimension integral $\int_0^1 \int_0^1 \cdots \int_0^1 f(\boldsymbol{X}) \mathrm{d}X_1 \mathrm{d}X_2 \cdots \mathrm{d}X_n$ to one-dimension integral. Consider a set of transfer functions:

$$X_i = G_i(\sin(\omega_i s)), \quad i = 1, 2, \ldots, n. \tag{A.24}$$

where $\{\omega_i\}$ is a set of integer angular frequencies, $s \in (-\pi, \pi)$. The mean and variance of Y then can be approximated by

$$E(Y) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \mathrm{d}s \tag{A.25}$$

and

$$V(Y) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} f^2(s) \mathrm{d}s - E^2(Y) \tag{A.26}$$

By applying the Parseval's theorem to the formulations of mean and variance, we can get

$$V(Y) \approx 2 \sum_{p=1}^{\infty} \left(A_p^2 + B_p^2\right) \tag{A.27}$$

where $A_p = 1/2 \int_{-\pi}^{\pi} f(s) \cos(ps) \mathrm{d}s$ and $B_p = 1/2 \int_{-\pi}^{\pi} f(s) \sin(ps) \mathrm{d}s$ are the Fourier coefficients. Thus the FAST first-order sensitivity index can be defined as

$$S_i = \frac{V_i}{V(Y)} = \frac{2 \sum_{q=1}^{\infty} \left(A_{q \cdot \omega_i}^2 + B_{q \cdot \omega_i}^2\right)}{2 \sum_{p=1}^{\infty} \left(A_p^2 + B_p^2\right)} \approx \frac{\sum_{q=1}^{M} \left(A_{q \cdot \omega_i}^2 + B_{q \cdot \omega_i}^2\right)}{\sum_{i=1}^{n} \sum_{q=1}^{M} \left(A_{q \cdot \omega_i}^2 + B_{q \cdot \omega_i}^2\right)} \tag{A.28}$$

where M is the maximum harmonic, usually to be 4 or 6. S_i is the fraction of the output variance due to the input variable X_i. A large index means a significant first-order effect.

A.9 McKay main and two-way interaction effects analysis

Let $E(Y)$ and $V(Y)$ be the prediction mean and variance of an output Y, thus $V(Y)$ can be decomposed as

$$V(Y) = V[E(Y|X_i)] + E[V(Y|X_i)] = V\left[E(Y|X_i, X_j)\right] + E\left[V(Y|X_i, X_j)\right] \tag{A.29}$$

where X_i and X_j are the ith and jth input respectively, $V[E(Y|X_i)]$ is the variance of the conditional expectation of Y conditioned on X_i, and $E[V(Y|.X_i)]$ is the residual term measuring the estimated variance of Y by fixing X_i; $V[E(Y|X_i, X_j)]$ is the variance of the conditional expectation of Y conditioned on X_i and X_j, and $E[V(Y|X_i, X_j)]$ is the residual term measuring the estimated variance of Y by fixing X_i and X_j. The indices of McKay main effect and two-way interaction effect analysis are defined as

$$\eta_i^2 = \frac{V[E(Y|X_i)]}{V(Y)} = \frac{V[E(Y|X_i)]}{V[E(Y|X_i)] + E[V(Y|X_i)]} \tag{A.30}$$

and

$$\eta_{ij}^2 = \frac{V\left[E(Y|X_i, X_j)\right]}{V(Y)} = \frac{V\left[E(Y|X_i, X_j)\right]}{V\left[E(Y|X_i, X_j)\right] + E\left[V(Y|X_i, X_j)\right]} \tag{A.31}$$

The above two indices are also called correlation ratios. The former measures the relative contribution of input X_i to the output variance, while the latter measures the relative contributions of input X_i and X_j together to the output variance.

A.10 Sobol' sensitivity indices

Let the function $Y = f(\boldsymbol{X}) = f(X_1, X_2, \ldots, X_n)$, where $X_i \in [0,1]$; $i = 1, 2, \ldots, n$. Assume the model output can be decomposed into terms of increasing dimensionality as follows:

$$\begin{aligned} Y = f(\boldsymbol{X}) = & f_0 + \sum_{i=1}^{n} f_i(X_i) + \sum_{i=1}^{n} \sum_{j>i}^{n} f_{i,j}(X_i, X_j) + \ldots \\ & + f_{1,2,\ldots,n}(X_1, X_2 \cdots, X_n) \end{aligned} \tag{A.32}$$

where f_0 is a constant, $f_i(X_i)$ are the functions of one variable, $f_{i,j}(X_i, X_j)$ are the functions of two variables, etc. The total variance of the output can be written as

$$V(Y) = \int_0^1 \cdots \int_0^1 f^2(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} - f_0^2 \tag{A.33}$$

while the contribution of a generic term $f_{i_1, \cdots, i_s} (1 \le i_1 < \cdots < i_s \le n)$ to the total variance can be written as

$$V_{i_1, \cdots, i_s} = \int_0^1 \cdots \int_0^1 f_{i_1, \cdots, i_s}^2 (X_{i_1}, \cdots, X_{i_s}) \mathrm{d}X_{i_1} \ldots \mathrm{d}X_{i_s} \tag{A.34}$$

Thus the ANOVA-like decomposition of total variance can be expressed as

$$V(Y) = \sum_{s=1}^{n} \sum_{i_1 < \cdots < i_s}^{n} V_{i_1, \cdots, i_s} = \sum_{i=1}^{n} V_i + \sum_{i=1}^{n} \sum_{i<j}^{n} V_{i,j} + \cdots + V_{1,2,\cdots,n} \tag{A.35}$$

The Sobol' sensitivity indices are defined as

$$S_{i_1, \cdots, i_s} = \frac{V_{i_1, \cdots, i_s}}{V(Y)}, \quad 1 \le i_1 < \cdots < i_s \le n. \tag{A.36}$$

Theoretically, this global method can compute sensitivity index of any order. However, the computation for high order is impractical when the number of input variables is large. The measure proposed by Homma and Saltelli (1996) provides a simple way for computing the total effect of each input variable. The total effect of the ith input is defined as

$$S_{Ti} = S_i + S_{i,ci} = 1 - S_{ci} \quad \text{(A.37)}$$

where S_i and $S_{i,ci}$ are representations of first-order effect and high-order effect, respectively; S_{ci} is the sum of all the $S_{i1,\ldots,is}$ terms that excludes the index i.

References

Beven, K., 2004. Rainfall-Runoff Modelling: The Primer. John Wiley & Sons, Ltd, Chichester, West Sussex.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1–2), 18–36. http://dx.doi.org/10.1016/j.jhydrol.2005.07.007.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6 (3), 279–298. http://dx.doi.org/10.1002/hyp.3360060305.

Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Process. 9 (3–4), 251–290. http://dx.doi.org/10.1002/hyp.3360090305.

Borgonovo, E., Castaings, W., Tarantola, S., 2012. Model emulation and moment-independent sensitivity analysis: an application to environmental modelling. Environ. Model. Softw. 34, 105–115. http://dx.doi.org/10.1016/j.envsoft.2011.06.006.

Box, G.E.P., Behnken, D.W., 1960. Some new three level designs for the study of quantitative variables. Technometrics 2 (4), 455–475.

Box, G.E.P., Hunter, J.S., 1961. The 2^{k-p} fractional factorial designs: part I. Technometrics 3 (3), 311–351.

Box, G.E.P., Wilson, K.B., 1951. On the experimental attainment of optimum conditions. J. R. Stat. Soc. Ser. B (Methodol.) 13 (1), 1–45.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36 (12), 3663–3674. http://dx.doi.org/10.1029/2000WR900207.

Brazil, L.E., 1988. Multilevel Calibration Strategy for Complex Hydrologic Simulation Models. Ph.D. Thesis. Department of Civil Engineering, Colorado State University, Fort Collins, CO.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Burnash, R.J.C., 1995. The NWS river forecast system-catchment modeling. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO, pp. 311–366.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers. US Department of Commerce, National Weather Service, Sacramento, CA.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Model. Softw. 22 (10), 1509–1518. http://dx.doi.org/10.1016/j.envsoft.2006.10.004.

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298. http://dx.doi.org/10.1214/09-AOAS285.

Clark, M.P., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47, W09301. http://dx.doi.org/10.1029/2010WR009827.

Crawford, S.L., 2006. Correlation and regression. Circulation 114 (19), 2083–2088. http://dx.doi.org/10.1161/CIRCULATIONAHA.105.586495.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I theory. J. Chem. Phys. 59 (8), 3873–3878. http://dx.doi.org/10.1063/1.1680571.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F., 2006. Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320 (1–2), 3–17. http://dx.doi.org/10.1016/j.jhydrol.2005.07.031.

Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J. Hydrol. 158 (3–4), 265–284. http://dx.doi.org/10.1016/0022-1694(94)90057-4.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031. http://dx.doi.org/10.1029/91WR02985.

Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., Verleysen, M., 2008. Using the delta test for variable selection. In: ESANN 2008 Proceedings, European Symposium on Artificial Neural Networks, Bruges, Belgium, pp. 25–30.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67. http://dx.doi.org/10.1214/aos/1176347963.

Galton, F., 1886. Regression towards mediocrity in hereditary stature. J. Anthropol. Inst. G B Irel. 15, 246–263.

Gibbs, M., MacKay, D.J.C., 1997. Efficient Implementation of Gaussian Processes (Unpublished manuscript).

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763. http://dx.doi.org/10.1029/97WR03495.

Gutiérrez, Á.G., Schnabel, S., Contador, J.F.L., 2009. Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. Ecol. Model 220 (24), 3630–3637. http://dx.doi.org/10.1016/j.ecolmodel.2009.06.020.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. 52 (1), 1–17. http://dx.doi.org/10.1016/0951-8320(96)00002-6.

Hornberger, G.M., Spear, R.C., 1981. An approach to the preliminary analysis of environmental systems. J. Environ. Manag. 12 (1), 7–18.

Hsieh, H., 2006. Application of the PSUADE Tool for Sensitivity Analysis of an Engineering Simulation. Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. 29 (8), 2637–2649 http://dx.doi.org/10.1029/93WR00877.

Kahng, A.B., Lin, B., Samadi, K., 2010. Improved on-chip router analytical power and area modeling. In: 2010 Asia and South Pacific Design Automation Conference, Taipei, Taiwan, pp. 241–246.

Karypis, G., Kumar, V., 1998. METIS: a Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices (Version 4.0). University of Minnesota, Minneapolis, MN.

Kavetski, D., Clark, M.P., 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. Water Resour. Res. 46 (10), W10511. http://dx.doi.org/10.1029/2009WR008896.

Kavetski, D., Kuczera, G., Franks, S.W., 2003. Semi-distributed hydrological modeling: a "saturation path" perspective on TOPMODEL and VIC. Water Resour. Res. 39 (9), 1246. http://dx.doi.org/10.1029/2003WR002122.

Levy, S., Steinberg, D.M., 2010. Computer experiments: a review. Adv. Stat. Anal. 94 (4), 311–324. http://dx.doi.org/10.1007/s10182-010-0147-9.

Liu, Y., Gupta, H.V., Sorooshian, S., Bastidas, L.A., Shuttleworth, W.J., 2004. Exploring parameter sensitivities of the land surface using a locally coupled land-atmosphere model. J. Geophys. Res. 109, D21101. http://dx.doi.org/10.1029/2004JD004730.

Loeppky, J.L., Sacks, J., Welch, W.J., 2009. Choosing the sample size of a computer experiment: a practical guide. Technometrics 51 (4), 366–376. http://dx.doi.org/10.1198/TECH.2009.08040.

McKay, M.D., 1995. Evaluating Prediction Uncertainty. Los Alamos National Laboratory (LANL), Los Alamos, NM.

McKay, M.D., Beckman, R.J., 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/10.1080/00401706.2000.10485979.

Meteopolis, N., Ulam, S., 1949. The Monte Carlo method. J. Am. Stat. Assoc. 44 (247), 335–341.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experments. Technometrics 33 (2), 161–174. http://dx.doi.org/10.1080/00401706.1991.10484804.

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

Pi, H., Peterson, C., 1994. Finding the embedding dimension and variable dependencies in time series. Neural Comput. 6 (3), 509–520. http://dx.doi.org/10.1162/neco.1994.6.3.509.

Plackett, R.L., Burman, J.P., 1946. The design of optimum multifactorial experiments. Biometrika 33 (4), 305–325.

Ratto, M., Pagano, A., Young, P., 2007. State dependent parameter metamodelling and sensitivity analysis. Comput. Phys. Commun. 177 (11), 863–876. http://dx.doi.org/10.1016/j.cpc.2007.07.011.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. Water Resour. Res. 46, W05521. http://dx.doi.org/10.1029/2009WR008328.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Stat. Sci. 4 (4), 409–423. http://dx.doi.org/10.1214/ss/1177012413.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145 (2), 280–297. http://dx.doi.org/10.1016/S0010-4655(02)00280-1.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., 2008. Global Sensitivity Analysis: the Primer. John Wiley & Sons, Ltd, Chichester, West Sussex.

Saltelli, A., Tarantola, S., Chan, K.P.S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41 (1), 39–56. http://dx.doi.org/10.2307/1270993.

Shahsavani, D., Grimvall, A., 2011. Variance-based sensitivity analysis of model outputs using surrogate models. Environ. Model. Softw. 26 (6), 723–730. http://dx.doi.org/10.1016/j.envsoft.2011.01.002.

Snow, M.G., Bajaj, A.K., 2010. Uncertainty Quantification Study for a Comprehensive Electrostatic MEMS Switch Model. PRISM. NNSA Center for Prediction of Reliability, Integrity and Survivability of Microsystems, West Lafayette, IN.

Sobol', I.M., 1990. Quasi-Monte Carlo methods. Progr. Nucl. Energy 24 (1–3), 55–61. http://dx.doi.org/10.1016/0149-1970(9090022-W).

Sobol', I.M., 1993. Sensitivity analysis for nonlinear mathematical models. Math. Model. Comput. Exp. 1, 407–414.

Sobol', I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. 55 (1–3), 271–280. http://dx.doi.org/10.1016/S0378-4754(0000270-6).

Sorooshian, S., Gupta, V.K., 1983. Automatic calibration of conceptual rainfall-runoff models: the question of parameter observability and uniqueness. Water Resour. Res. 19 (1), 260–268. http://dx.doi.org/10.1029/WR019i001p00260.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101.

Statnikov, R.B., Matusov, J.B., 2002. Multicriteria Analysis in Engineering: Using the PSI Method with MOVI 1.0. Kluwer Academic Publishers, Dordrecht.

Steinberg, D., Colla, P.L., Martin, K., 1999. MARS User Guide. Salford Systems, San Diego, CA.

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

Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrol. Earth Syst. Sci. 11 (2), 793–817. http://dx.doi.org/10.5194/hess-11-793-2007.

Tong, C., 2005. PSUADE User's Manual. Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

Tong, C., 2008. Quantifying Uncertainties of a Soil-foundation Structure-interaction System under Seismic Excitation. Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

Tong, C., Graziani, F., 2008. A practical global sensitivity analysis methodology for multi-physics applications. In: Graziani, F. (Ed.), Computational Methods in Transport: Verification and Validation. Springer-Verlag, Berlin, pp. 277–299.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. J. Hydrol. 324 (1), 10–23. http://dx.doi.org/10.1016/j.jhydrol.2005.09.008.

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2008. Characterization of watershed model behavior across a hydroclimatic gradient. Water Resour. Res. 44 (1), W01429. http://dx.doi.org/10.1029/2007WR006271.

Wemhoff, A.P., Hsieh, H., 2007. TNT Prout-Tompkins Kinetics Calibration with PSUADE. Lawrence Livermore National Laboratory (LLNL), Livermore, CA.