Environmental Modelling & Software 26 (2011) 444–457

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis

Jing Yang*,[1]

Department of Geography, University of Guelph, Guelph, ON, N1G 2W1, Canada

ARTICLE INFO

Article history:
Received 13 February 2009
Received in revised form
8 October 2010
Accepted 9 October 2010
Available online 1 November 2010

Keywords:
Sensitivity analysis
Uncertainty analysis
Sobol' method
Morris method
Linear regression method
Regionalized sensitivity analysis
Non-parametric smoothing

ABSTRACT

Sensitivity analysis plays an important role in model development, calibration, uncertainty analysis, scenario analysis, and, hence, decision making. With the availability of different sensitivity analysis techniques, selecting an appropriate technique, monitoring the convergence and estimating the uncertainty of the sensitivity indices are very crucial for environmental modelling, especially for distributed models due to their high non-linearity, non-monotonicity, highly correlated parameters, and intensive computational requirements. It would be useful to identify whether some techniques outperform others with respect to computational requirements, reliability, and other criteria. This paper proposes two methods to monitor the convergence and estimate the uncertainty of sensitivity analysis techniques. One is based on the central limit theorem and the other on the bootstrap technique. These two methods are implemented to assess five different sensitivity analysis techniques applied to an environmental model. These techniques are: the Sobol' method, the Morris method, Linear Regression (LR), Regionalized Sensitivity Analysis (RSA), and non-parametric smoothing. The results show that: (i) the Sobol' method is very robust in quantifying sensitivities and ranking parameters despite a large number of model evaluations; (ii) the Morris method is efficient to rank out unimportant parameters at a medium cost; (iii) the non-parametric smoothing is reliable and robust in quantifying the main effects and low-order interactions while requiring a small number of model evaluations; finally (iv) the other two techniques, that is, LR and RSA, should be used with care.

1. Introduction

Environmental models are becoming increasingly important in the decision making process as they can provide systematic and consistent information on water availability, impacts of climate and land use changes, and analyses of non-point source pollution (Yang et al., 2007). Meanwhile, most environmental models need a so-called calibration process (or uncertainty analysis) because of measurement errors in input data and observations needed for model calibration, errors in model structure, and the large number of non-identifiable parameters (Pappenberger et al., 2006; Yang et al., 2007). Recently applied uncertainty analysis techniques include GLUE (Generalized Likelihood Uncertainty Estimation; Beven and Binley, 1992), SUFI-2 (Sequential Uncertainly Fitting 2; Abbaspour et al., 2007), and various Bayesian frameworks (Yang et al., 2007; Vrugt et al., 2005; Kavetski et al., 2006 etc.). On the other hand, sensitivity analysis (SA) contributes to the assessment of how variations in the output of a model can be apportioned, qualitatively or quantitatively, to different sources of variations, and how the given model depends upon the information fed into it (Saltelli et al., 2000). SA has been playing a very important role in model development, calibration, uncertainty analysis, scenario analysis (e.g., impact of climate change on water resources), and, hence, decision making. As such, SA has been issued in regulatory prescriptions as guidelines for modelling (e.g., European Commission, 2005; The U.S. EPA (Environmental Protection Agency) Council for Regulatory Environmental Modeling, 2003).

In the literature, SA techniques have been categorized multiple ways (e.g., Saltelli et al., 2000, 2004, 2008; Cacuci, 2003; Frey and Patil, 2002; Iman and Helton, 1988). Based on the factor space of interest, these techniques can be divided into two categories: local sensitivity analysis and global sensitivity analysis. Local sensitivity analysis concentrates on examining the local impact of the factors on the model output and is usually derivative based. Local analysis belongs to the class of one-factor-at-a-time (OAT) methods, where each single factor is perturbed in turn, while others are kept constant. Although the computational cost of local sensitivity analysis is low, it can be highly biased for non-linear systems. Therefore, the reliability of its application is not guaranteed. Global sensitivity analysis, on the other hand, aims at quantifying the

* Tel.: +1 519 824 4120x52684; fax: +1 519 837 2940.
E-mail addresses: jingdotyang@hotmail.com, jing.yang@smart.mit.edu.
[1] Now at: Center for Environmental Sensing and Modelling, Singapore-MIT Alliance for Research & Technology, S16-05-08, 3 Science Drive 2, Singapore 117543, Singapore. Tel.: +65 6516 5702; fax: +65 6778 5654.

1364-8152/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.10.007

impact over the entire factor space, and, hence, it has attracted a lot of modelers. Most global sensitivity techniques are Monte-Carlo based. Widely used global sensitivity techniques include: (i) variance-based technique, such as the Sobol' method (Sobol', 1990) and Fourier Amplitude Sensitivity Test (FAST) (Cukier et al., 1973, 1978; McRae et al., 1982; Saltelli et al., 1999); (ii) global screening method, such as the Morris method (Morris, 1991) and Latin Hypercube-OAT (LH-OAT) (van Griensven et al., 2006); (iii) regression/correlation-based techniques, such as regression analysis, correlation measure, stepwise regression analysis (Helton, 1993), response surface methodology (Myers and Montgomery, 1995); (iv) Regionalized Sensitivity Analysis (RSA) (Spear and Hornberger, 1980); to name a few.

More recently, approaches based on the emulation concept have been used for sensitivity analysis purposes: (i) kriging or Gaussian Emulators (Sacks et al., 1989; Oakley and O'Hagan, 2004), (ii) the Random Sampled High Dimensional Model Representation (RS-HDMR) (Li et al., 2002, 2006); and (iii) Smoothing spline ANOVA models (Wahba, 1990; Gu, 2002; Ratto et al., 2007a). These methods utilize the smoothness properties of the functions under analysis and usually dramatically reduce the computational requirements for sensitivity analysis.

In spite of OAT's limitations, most published SA papers are based on OAT and modelers seem reluctant to abandon this practice (Saltelli and Annoni, 2010). Fortunately, there are more and more global SAs (e.g., Varella et al., 2010; Confalonieri et al., 2010; Foscarini et al., 2010; Ravalico et al., 2010) being applied in environmental modelling. Also, there are some interesting studies comparing properties of sensitivity analysis techniques. For example, Tang et al. (2007) tested four sensitivity analysis methods: (i) local analysis using parameter estimation software (PEST), (ii) regional sensitivity analysis (RSA), (iii) analysis of variance (ANOVA), and (iv) the Sobol' method. Pappenberger et al. (2008) performed a comparative study combining the Sobol' method, the Morris method, an entropy-based measure, and RSA. Ratto et al. (2007b) applied the Morris method in conjunction to a multiple predictor smoothing approach (the SDP approach) for the estimation of the main effect of sensitivity indices. Helton et al. (2005) and Storlie and Helton (2008) compared linear measures, linear measures based on rank transformation, the regular linear regression, ranked linear regression, and several smoothing techniques.

Despite reviews and comparison studies, except some additional studies on the effect and adequacy of the sample size on sensitivity analysis (Iman, 1982; Iman and Helton, 1991; Helton and Sallaberry, 2009), there are few studies that have further examined the convergence and uncertainty analysis of the sensitivity indices, particularly for Monte-Carlo based techniques. As far as the convergence is concerned, for a given problem (i.e., the model with parameters) sample sizes (model runs) are given based on expert experiences. For example, for the Sobol' method, Saltelli (1999) suggested the base sample size should be in the range of 100 or higher. In terms of the uncertainty of the sensitivity indices, Archer et al. (1997) proposed the bootstrap technique to estimate the confidence intervals of sensitivity indices for the Sobol' method. However, this method is seldom used in environmental applications.

The objective of this paper is to propose a method to examine the convergence for different Monte-Carlo based SA techniques, to estimate the uncertainty of the sensitivity index, and to assess the performances of five different global sensitivity analysis techniques: (1) the Sobol' method; (2) the Morris method; (3) Linear Regression method; (4) Regionalized Sensitivity Analysis (RSA); and (5) the SDP non-parametric regression/smoothing approach. These tasks are carried out by comparing their applications to an environmental model (HYMOD). In this paper, the term "uncertainty" refers to the "aleatory uncertainty" (or uncertainty due to variability) which represents randomness of samples (see more discussion on uncertainty in Helton and Burmaster, 1996; Paté-Cornell, 1996; Hoffman and Hammonds, 1994; etc.). The remainder of the paper is organized as follows: Section 2 introduces the different sensitivity analysis techniques and briefly discusses their advantages and disadvantages; Section 3 presents the two methods to monitor the convergence and to estimate the uncertainty of sensitivity indices; Section 4 gives a brief description of HYMOD and study area, and the comparison setup of different SA techniques; Section 5 analyzes the results based on different SA techniques and uncertainty methods. Finally, conclusions are drawn in Section 6.

2. Sensitivity analysis techniques

In order to use the same terminology to present each sensitivity technique, a generalized model is defined as:

$$Y = f(X) \tag{1}$$

where X is the array of input factors[2] under consideration, and Y is the corresponding model output which could be a time series (such as streamflow series) or an objective function.

2.1. Sobol' method

The Sobol' method is based on the decomposition of the variance, V, of model output Y:

$$V = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + V_{1,2,\ldots,n} \tag{2}$$

where

$$\begin{aligned} V_i &= V\left(E\left(Y|X_i = x_i^*\right)\right) \\ V_{ij} &= V\left(E\left(Y\middle|X_i = x_i^*, X_j = x_j^*\right)\right) - V_i - V_j \end{aligned} \tag{3}$$

In the above formula, n is the number of factors, and $V(.)$ and $E(.)$ represent variance and expectation operators, respectively.

Normalizing by the unconditional variance V, the corresponding sensitivity indices (S_i, S_{ij}) are defined as:

$$\begin{aligned} S_i &= \tfrac{V_i}{V}, 1 \le i \le n \\ S_{ij} &= \tfrac{V_{ij}}{V}, 1 \le i < j \le n \end{aligned} \tag{4}$$

In the application of variance-based method, modelers are often interested in two sensitivity indices, the first order index or main effect (S_i) and the total sensitivity index or total effect (S_{Ti})

$$S_{Ti} = S_i + \sum_j S_{ij} + \sum_j \sum_k S_{ijk} + \ldots \tag{5}$$

where S_i represents the average output variance reduction that can be achieved when X_i is fixed, while S_{Ti} stands for the average output variance that would remain as long as X_i stays unknown (Tarantola et al., 2002), that is, the total contribution of X_i to the output variation (Archer et al., 1997). The difference between S_{Ti} and S_i denotes the degree of interaction between this factor and other factors. In this paper, S_{Ti} is used as a sensitivity index for ranking parameter and comparing with other techniques.

Besides S_i and S_{Ti}, two other indices, the second-order-closure sensitivity measure (S_{ij}^c) and complementary-closure sensitivity

[2] Input factors can be model parameters and/or input driving force such as rainfall in hydrology. In this study, factors are equivalent to model parameters.

measure (S^c_{-ij}), are also calculated in this study. S^c_{ij} ($=S_i + S_j + S_{ij}$) measures the main effect of the factor group (X_i, X_j) while S^c_{-ij} measures the main effect of the factor group which contains all the factors except (X_i, X_j). These two factors will illustrate the importance of the given group of factors.

As the Sobol' method involves estimating expectation ($E(.)$) and variance ($V(.)$), it requires a large number of model runs: for a base sample size m, it takes at least $m^*(n+2)$ model runs to calculate the (S_i, S_{Ti}) for n factors (Saltelli, 2002). The larger the base sample size, the more accurate the sensitivity indices will be. The number m should be in the range of 100 or higher (Saltelli, 1999), for example, Saltelli (2002) uses $m = 1024$ for the Sobol' g function (see Saltelli, 2002).

The variance-based method displays a number of attractive features for SAs (see e.g., Saltelli et al., 2000): (1) Model independence, that is, it works for non-linear models, non-monotonic models, and model with interaction among factors; (2) the method captures the influence of the full range of variation of each factor; (3) the method captures interaction effects; (4) the method can treat sets of factors as single factors. The major disadvantage is that it requires a large number of model evaluations.

2.2. Morris method

The Morris method is based on replicated and randomized "one-factor-at-a-time" (OAT) design (Morris, 1991). The basic idea is: for a random realization X taken on a grid, the local sensitivity measure (also called elementary effect) is computed for each factor based on OAT as follows:

$$d_i(X) = \frac{f(x_1,\ldots,x_{i-1},x_i+\Delta,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i,\ldots,x_n)}{\Delta} \quad (6)$$

where Δ is the predefined increment, and $X = (x_1,\ldots,x_{i-1}, x_i,\ldots,x_n)$ is a random sample in the parameter space so that the transformed point $(x_1,\ldots,x_{i-1},x_i+\Delta,\ldots,x_n)$ is still within the parameter space. Several elementary effects are computed for each input factor by randomly sampling the input factors on a grid and then a finite distribution of the elementary effects is obtained.

In the original Morris method, two sensitivity measures are used: the sample mean μ and standard deviation σ from the distribution of the elementary effects. The higher μ is, the more important the factor is to the output; the higher σ is, the more non-linear the factor is to the output or the more interactions between the factor and the other factors there are. When the model is non-monotonic, some elementary effects with opposite signs may cancel out when the μ is calculated, and Campolongo et al. (2007) recommended the use of μ^*, the sample mean of the distribution of the absolute values of the elementary effects.

The Morris method takes $m^*(n+1)$ model evaluations to compute the sensitivity indices (μ^*, σ), where m is the number of random realization and n is the number of factors. Compared to the variance-based method, the computational cost of the Morris method is lower.

In the Morris method, one needs to define the number of levels p which defines the optimal increment to be used in (6), $\Delta = p/[2(p-1)]$, with values of p normally within the range of [4,10] (Saltelli, 1999). Usually p equals 4 which means that $\Delta = 2/3$.

The advantages of the Morris method include lack of reliance on assumptions of relative sparsity of important inputs, monotonicity of outputs with respect to factors, or adequacy of a low-order polynomial as an approximation to the computational model (Morris, 1991). The main limitations are: it cannot provide a quantitative estimation of how much a factor contributes to the output variability, and it cannot distinguish the non-linearity of a factor from the interaction with other factors.

2.3. Linear regression analysis

By constructing a linear relationship between factors and model output, regression analysis can also provide a sensitivity measure of how sensitive the model output y is to a factor x_i:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i \quad (7)$$

As soon as the regression coefficients b_i ($i = 1, \ldots, n$) are estimated (e.g., via least squares method), the absolute standardized regression coefficient, SRC, can be taken as the sensitivity measure:

$$\mathrm{SRC}_i = \left| b_i \frac{\widehat{s_i}}{s} \right| \quad (8)$$

here $\widehat{s_i}$ and s are estimated standards deviation for x_i and y.

The advantage of this method is that it is straightforward and simple to apply. However, it is not applicable when the relationship between the factors and model output is non-linear or non-monotonic and when there is a high level of interactions among factors. As such, regression analysis often performs poorly. For non-linear models, the rank transformation can be helpful since the rank transformation can cope with non-linear models and mitigate the detrimental effect of long tailed output distribution (Saltelli and Sobol', 1995). The rank transformation also has two drawbacks: firstly it fails with non-monotonic models, and secondly its main effect is a forced linearization of the system by an artificial increase in the relative weight of the first order terms (Saltelli and Sobol', 1995) so the result cannot be transformed back to the original model. In this study the rank transformation is also applied for reference purposes.

2.4. Regionalized sensitivity analysis (RSA)

RSA was developed in the context of environmental models by Spear and Hornberger (Spear and Hornberger, 1980). The basic idea of RSA is to partition the samples of factor x_i under consideration into two sub-samples (i.e., behavioral and non-behavioral) according to the given criterion (e.g., quantiles of the output distribution, or thresholds or imposed ceilings derived from observed behavior, or legislation prescriptions). If the distributions of an input factor x_i in the two sub-samples are dissimilar then factor x_i is considered influential. The comparison of the two distributions is done by a Smirnov test, in which the maximum distance (d) between the two empirical cumulative distributions is taken. The larger the distance d the more sensitive the factor is.

The method has the advantage of being conceptually simple and easy to implement. Results are easy to understand and the methodology is model-independent. Although under certain circumstances the Smirnov test can highlight some interaction effects (see e.g., Saltelli et al. (2008), pp. 183–211), the RSA cannot quantify, in general, higher order effects or search for interacting structures, except for stylized structures detectable through correlation analysis within each of the two sub-samples. This means that the insignificance of the d statistics does not imply irrelevance of the input factor, due to possible missed interaction effects.

2.5. Recursive, non-parametric smoothing (SDP)

An important approach to sensitivity analysis, which has developed in the last few years, falls within the emulation context.

Table 1
A general comparison of applied sensitivity techniques.

	Sobol' method	Morris method	Linear Regression	RSA	SDP
Sampling strategy	Sobol' quasi-random	Monte-Carlo	Monte-Carlo	Monte-Carlo	Sobol' quasi-random
Computational requirements[a]	$m * (2 * n + 2)$	$m * (n + 1)$	m	m	m
	Computationally demanding	Cheap	Cheap	Depends on the filtering criterion	Cheap
Characteristic of sensitivity measure	Quantitative	Screening	Quantitative/Screening	Screening	Quantitative/Emulation
Parameter interaction	Yes (quantitative)	Yes (qualitative)	Dependent on the regression form	No	Yes (quantitative, up to 2nd−3rd order)
Applicability	Model-independent	Model-independent	Linear model (for monotonic model, rank transformed Y is feasible)	Model-independent	Model-independent
Reliability	High	High	Dependent on R^2	Weak	High (with dependence on R^2)

[a] The number of model runs required to complete the technique based on base sample size m and n factors.

The basic idea is to represent in a direct way the relationship between the model factors and model output, whose form is usually unknown to the analyst. If the emulation exercise is successful, one can obtain a simple relationship between the model factors and model output that fits well the original model and is less computationally demanding. Given the emulator, this can be used to compute any measure of interest, including sensitivities.

The methodology considered here is based on the ANOVA functional decomposition of $Y = f(X_1, \ldots, X_n)$:

$$Y = f_0 + \sum_i f_i(X_i) + \sum_i \sum_{j>i} f_{ij}(X_i, X_j) + \ldots \tag{9}$$

where

$$\begin{aligned} f_0 &= E(Y) \\ f_i(X_i) &= E(Y|X_i) - f_0 \\ f_{ij}(X_i, X_j) &= E(Y|X_i, X_j) - f_i(X_i) - f_j(X_j) - f_0 \end{aligned} \tag{10}$$

There are numerous methods for non-parametric smoothing that allow to build an emulator based on ANOVA functional decompositions, like ANOVA-spline smoothing (Gu, 2002; Wahba, 1990) or the State-Dependent Parameter (SDP) method (Ratto et al., 2007a).

SDP is a non-parametric methodology based on an approach first suggested by Young (1993) and applied for sensitivity analysis and emulation by Ratto et al. (2007a). In its basic form, SDP can be seen as an ANOVA-spline smoothing method that applies recursive algorithms in place of the standard *en-bloc* procedures, which imply large matrix inversions. The recursive form of the SDP approach avoids the tight sample size limitations typical of other emulation approaches (see details in Ratto et al., 2007a).

A brief summary of the fundamental idea of the SDP approach is provided here. The ANOVA expansion, truncated up to the third order, can be viewed as a state-dependent regression model:

$$\begin{aligned} Y_t - f_0 = &\sum_i p_{i,t}(s_{i,t}) + \sum_i \sum_{j>i} p_{ij,t}(s_{i,t}, s_{j,t}) \\ &+ \sum_i \sum_{j>i} \sum_{l>j>i} p_{ijl,t}(s_{i,t}, s_{j,t}, s_{l,t}) + e_t \end{aligned} \tag{11}$$

where each state-dependent parameter $p_{I,t}(s_{I,t})$, $I = i_1, \ldots, i_l$ depends on a state variable $s_{I,t}$ that moves according to a generalized sorting strategy along the co-ordinates of the single factors of the group of factors indexed by I. According to the generalized sorting strategy, the group of input factors of interest I is characterized by a low frequency spectrum (e.g., some quasi-periodic pattern) while the remaining ones present a white spectrum. In this way, the estimation of the various ANOVA terms reduces to the extraction of the low frequency component of the sorted output Y. To do so, the SDPs are modeled by one member of a generalized random walk (GRW) class of non-stationary processes. For instance, the integrated random walk (IRW) process provides good results, since it ensures that the estimated SDP relationship has the smooth properties of a cubic spline.

These smoothing techniques normally estimate all the main effects with only a few hundred Monte-Carlo (MC) realizations ($\leq$1000), almost independently of the dimensionality n of the problem. The accurate estimation of interaction effects may require more MC runs, depending on the type of interaction structure. In most cases, sample sizes up to 1000 are usually sufficient for interactions. The main advantages of smoothing methods are that: (i) they provide accurate sensitivity indices at significantly smaller sample sizes with respect to the Sobol' method, and (ii) they provide a full emulator, allowing to 'predict' the model output at new untried points. Moreover, smoothing methods usually provide estimates of the *error* of the sensitivity indices (see Doksum and Samarov, 1995), thus, avoiding the need of the bootstrap technique. The main drawbacks are that: (i) they are less easy to code with respect to the other methods presented here, and (ii) the estimate of total indices may be biased, since it is obtained by summing main effect and low-order interaction terms. This means that, if any high order interaction term exists, this would be missed by any emulation methodology. Fortunately, cross-validation checks evaluating the coverage (R^2) of the ANOVA model allows to highlight this problem, that is, when the R^2 of the ANOVA model is small.

2.6. A general comparison of the applied techniques

Table 1 gives an overview of the applied SA techniques, which includes the sampling scheme, computational requirements, and the characteristics of the sensitivity measure, i.e., interaction, applicability, and reliability.

3. Convergence, uncertainty, and comparison strategy

3.1. Convergence and uncertainty/accuracy

In Monte-Carlo based SAs, a general question is how many samples the given technique needs in order to achieve the convergence or how to monitor the convergence of the given statistic (see Iman, 1982). Obviously, this problem is different from that of the deterministic technique which is often monitored through a comparison of the two consecutive objective functions given the absolute/relative tolerance.

All SA methods are subject to uncertainty as sensitivity indices are estimated on a limited sample (Pappenberger et al., 2006). Though several methodologies under study have their methods to

estimate the uncertainties (e.g., see Saltelli and Sobol' (1995) for Sobol' method), most of the publications (especially in the field of environmental sciences) do not have any uncertainty estimation. A common approach should be proposed for a fair comparison of all the techniques. Here, two methods are presented to monitor the convergence and uncertainty analysis: one is based on the Central Limit Theorem (CLT), and the other on the bootstrap technique.

3.1.1. Convergence and uncertainty analysis based on the Central Limit Theorem

According to the Central Limit Theorem (CLT), given a distribution with mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with mean (μ) and variance σ^2/R as R, the sample size, increases. Based on the CLT, the procedure applied in this paper can be summarized as follows[3]:

(1) Given a base sample of dimension m, independently sample R replicas of such base samples (for the Sobol' and SDP methods, this paper applies scrambled Sobol' quasi-random sequences as in Saltelli et al. (2000)); and, then, the relevant sensitivity indices are computed. This produces R different sensitivity estimates, and the mean and coefficient of variance (standard deviation over the mean) of the calculated sensitivity indices are computed. According to the CLT, R should be larger than 30 (e.g., see p. 359 in Grinstead and Snell, 1997), however, this paper uses 100;
(2) Gradually increase the base sample size m and repeat step (1) until there is not any significant change in the coefficient of variance.

Once the convergence is achieved, the final mean value is used as the sensitivity index while the standard deviation represents its uncertainty, or alternatively, one can use the 95% confidence interval (95% CI) to denote the uncertainty of the sensitivity index. The density function for each sensitivity index, based on Kernel density estimation (Parzen, 1962; Duda and Hart, 1973; Wasserman, 2005), is also computed from 100 CLT replicas.

(3) Determining the most important factor

Since the CLT involves several independent samples, it would be of interest to know if sensitivity results from independent samples agree on the most important factors. In this paper, the top down coefficient of concordance (TDCC) (Iman and Conover, 1987) is used as a measure of agreement among sensitivity results from independent samples. The fundamental idea behind the TDCC is to emphasize the agreement on the most important factors while deemphasizing the disagreement on the less important factors (Helton et al., 2005). The TDCC is defined as follows:

$$\mathrm{TDCC} = \frac{\sum_{i}^{n}\left(\sum_{j}^{R}\mathrm{ss}(O_{ij})\right)^2 - n\cdot R^2}{R^2\cdot\left(n - \sum_{1}^{n}(1/i)\right)} \tag{12}$$

where n is the number of model factors, R the number of independent samples, O_{ij} the computed sensitivity index for the ith factor and the jth independent samples, and $\mathrm{ss}(O_{ij})$ the Savage scores of O_{ij} defined as:

$$\mathrm{ss}(O_{ij}) = \sum_{\mathrm{RANK}(O_{ij})}^{n}(1/k) \tag{13}$$

The higher the TDCC is, the more significant the agreement is on the most important factor.

3.1.2. Convergence and uncertainty analysis based on the bootstrap technique

The procedure described in Section 3.1.1 is only computationally realistic for simple models that require less computational time for model runs and involve few parameters. However, it is not suitable for complicated models which are often characterized by complex response surface, large amounts of parameters and time-consuming computational requirements. In such cases, the CLT approach is not feasible. Therefore, the alternative bootstrap technique is considered, since it provides the cheapest option to estimate confidence intervals for sensitivity indices (Archer et al., 1997):

(1) The estimated statistic (sensitivity index) is plotted against the gradually increasing base sample size m. Once there is not any serious variation for each estimated statistic, convergence is assumed. This method is theoretically derived from the CLT. In the CLT, as the sample size increases, the standard deviation of the estimators decreases, and hence the statistical estimator becomes closer and closer to the average value in the CLT.
(2) The bootstrap technique is applied to estimate the uncertainty of the sensitivity index. This approach relies on re-sampling with replacement. Basically, the m base samples are re-sampled B times with replacement, and at each stage, a new bootstrap estimation of the sensitivity index is computed, leading to a bootstrap estimate of the sampling distribution of the sensitivity indices; and then the 95% CIs of the sensitivity indices are constructed using their sampling distributions. Archer et al. (1997) used this method to estimate the confidence intervals of the sensitivity indices of the Sobol' method, and suggested a value of $B = 1000$ or 2000 to be likely chosen. This study uses $B = 1000$. In the case of the SDP approach, besides the bootstrap technique, this paper uses the estimated standard error of the sensitivity indices, based on asymptotic Gaussian assumptions (Doksum and Samarov, 1995), as a reference.

The advantage of the bootstrap technique is that it is based on a minimum number of assumptions and can be applied when the sample distribution is unknown. And similar to the CLT, their density functions are also computed from 1000 bootstrap replicas.

3.2. Factor ranking

Once the convergence for the SA is reached, factor ranking can be given with certain frequencies based on their rankings in all the CLT replicas or bootstrap replicas. A factor may have more than one rank due to its non-uniqueness ranking in all the CLT replicas or bootstrap replicas.

After the sensitivity index and its uncertainty are computed, the factor rank is very easily computed as follows:

(1) For any two parameters, if there is no overlap/little overlap between the two uncertainty regions (e.g., 95% CI) of sensitivity indices, the parameter with the highest sensitivity index outranks the other one;
(2) For any two parameters, if there is a large area of overlap for the two sensitivity indices, a parameter can only be more sensitive than the other with a certain probability. The overlap in this study is calculated from the density functions which are

[3] In the literature, a similar approach has been applied to calculate the statistical confidence on the complementary cumulative distribution function (e.g., Helton et al., 2000), and 95% prediction uncertainty range (e.g., Beven and Binley, 1992).

estimated based on Kernel density estimation (Parzen, 1962; Duda and Hart, 1973; Wasserman, 2005).

3.3. Reliability

As published in the literature and to be discussed later, different techniques often lead to different sensitivity results (e.g., sensitivity index, uncertainty range, and ranks). Among all the sensitivity techniques, the Sobol' method is the most robust, accurate, reproducible, model-dependent and reliable. This paper takes the result of the Sobol' method as a benchmark and all other results are compared to the benchmark result in terms of the factor rank.

4. Case studies and comparison setup

4.1. Case study

To compare the aforementioned SA techniques, the example chosen is an application of the HYMOD model (Boyle, 2001) to the Leaf River watershed, Mississippi, U.S.. The HYMOD model is a conceptual, five parameter, watershed model. The schematic representation of the HYMOD is shown in Fig. 1, and its five parameters are listed in Table 2. First, HYMOD uses a two parameter (i.e., maximum soil storage capacity, C_{max}, and the degree of spatial variability of soil moisture capacity, b_{exp}) rainfall-excess model (Moore, 1985) to generate the effective rainfall from the precipitation (P) based on the evapotranspiration (ET) and soil storage capacity (C) (left part of the Fig. 1), and then one part (*Alpha*) of the effective rainfall is routed through three identical cascaded reservoirs (with the residence time R_q) to the watershed outlet as quick flow (top middle in Fig. 1), and the other part (1-*Alpha*) of the effective rainfall is routed through a single reservoir (with the residence time R_s) to the watershed outlet as slow flow (bottom middle in Fig. 1). The summation of the quick flow and slow flow is the simulated flow at the watershed outlet (right in Fig. 1). For more details on HYMOD, the readers are referred to Boyle (2001).

The Leaf River watershed drains an area of 1950 km^2. The historical available data (1948–1988), obtained from the Hydrologic Research Laboratory of U.S. National Weather Service, consist of mean area precipitation (mm/day), potential evapotranspiration (mm/day), and streamflow (m^3/s).

In the literature, the HYMOD model and the Leaf River data have been extensively investigated (e.g., Yapo et al., 1998; Boyle et al., 2000; Vrugt et al., 2002, 2003, 2005). This study only uses the three consecutive data sets (28 July 1952–1955) same as in Vrugt et al. (2005).

4.2. Setup of applications

In order to give a fair comparison, all the applications were done with the same model setup, including the same parameter set, parameter prior assumption, and objective function.

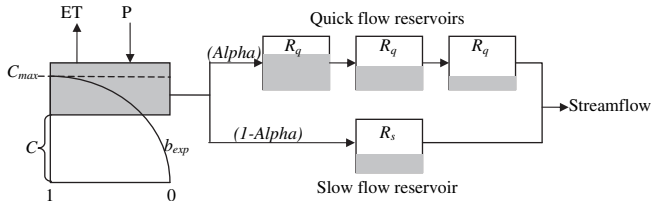

Fig. 1. A schematic representation of the HYMOD model. P, ET and C stand for precipitation, evapotranspiration, and catchment storage capacity, respectively, and C_{max}, b_{exp}, *Alpha*, R_q and R_s are the five model parameters. (Adapted from Vrugt et al., 2003, by permission from American Geophysical Union).

Table 2
The HYMOD parameters and their initial uncertainty ranges.

Parameter	Meaning of the parameter	Initial uncertainty range
C_{max}	Maximum soil storage capacity [mm]	[200.00, 500.00]
b_{exp}	Degree of spatial variability of soil moisture capacity [−]	[0.10, 2.00]
Alpha	Distribution factor between two reservoirs [−]	[0.50, 0.99]
R_s	Residence time of the slow flow reservoir [days]	[0.00, 0.10]
R_q	Residence time of the quick flow reservoir [days]	[0.30, 0.70]

In hydrology, the most frequently used objective function is the Nash-Sutcliffe, NS (Nash and Sutcliffe, 1970), which is also applied in this study:

$$\mathrm{NS} = 1 - \frac{\sum_{t_i=1}^{q}\left(y_{t_i}^{M}(X) - y_{t_i}\right)^2}{\sum_{t_i=1}^{q}\left(y_{t_i} - \bar{y}\right)^2} \tag{14}$$

where q is the number of observed data points, and y_{t_i} and $y_{t_i}^{M}(X)$ represents the observation and model simulation with parameters X at time t_i, respectively, and $\bar{y}$ is the average value of the observations. *NS* takes values from $-\infty$ to 1, and the larger the NS is, the better the simulation is. When NS equals 1, the simulation perfectly matches the observation.

For the prior distributions of parameters in Table 2, since they are subjective, the priors are not very easily assessed (see discussion in Beven and Binley, 1992). In this paper, all parameter priors are assumed to be independently and uniformly distributed.

Besides the objective function, and parameter selection and parameter priors, one needs to setup the p value for the Morris method and filtering criteria (methods to separate behavioral and non-behavioral sets) for RSA. In order to study the behavior of different "p"s on sensitivity results, several "p"s (4, 8, 20, 40, and 60) are tried, although p values greater than 10 are seldom used. For RSA, NS = 0.6 is taken as a threshold to separate parameter sets into behavioral (NS greater than or equal to 0.6) and non-behavioral (NS smaller than 0.6) sets, which is often used in hydrology to assess model behavior. Furthermore, the following filtering criteria are also considered: NS = 0.5 and 0.7, and 70–30% (i.e., 30% of the parameters with highest *NS* values are classified as behavioral and the rest 70% as non-behavioral), 80–20%, and 90–10%.

5. Application and discussion

In this section, the five SA techniques are implemented to the case study, i.e., the application of HYMOD in the Leaf River watershed, firstly based on CLT (Section 3.1.1), and then on the bootstrap technique (Section 3.1.2). Their results are discussed independently and then the comparison is done between them.

5.1. Result based on CLT

Fig. 2 shows the convergence of the applications of different SA techniques based on CLT. In each plot of Fig. 2, each line represents an evolution of the "coefficient of variation" for each parameter with increasing base sample size m (expressed as total model runs in Fig. 2). And Table 3 lists converged means, standard deviations, 95% CIs, and ranks of the sensitivity indices for each SA application.

As can be seen in Fig. 2, the Sobol' method took around 1500 base samples (i.e., 1500 * (2 * 5 + 2) = 18,000 model runs for each independent CLT replica) to converge, the Morris method with p = 40 needed 500 base samples (i.e., 500 * (5 + 1) = 3000 model

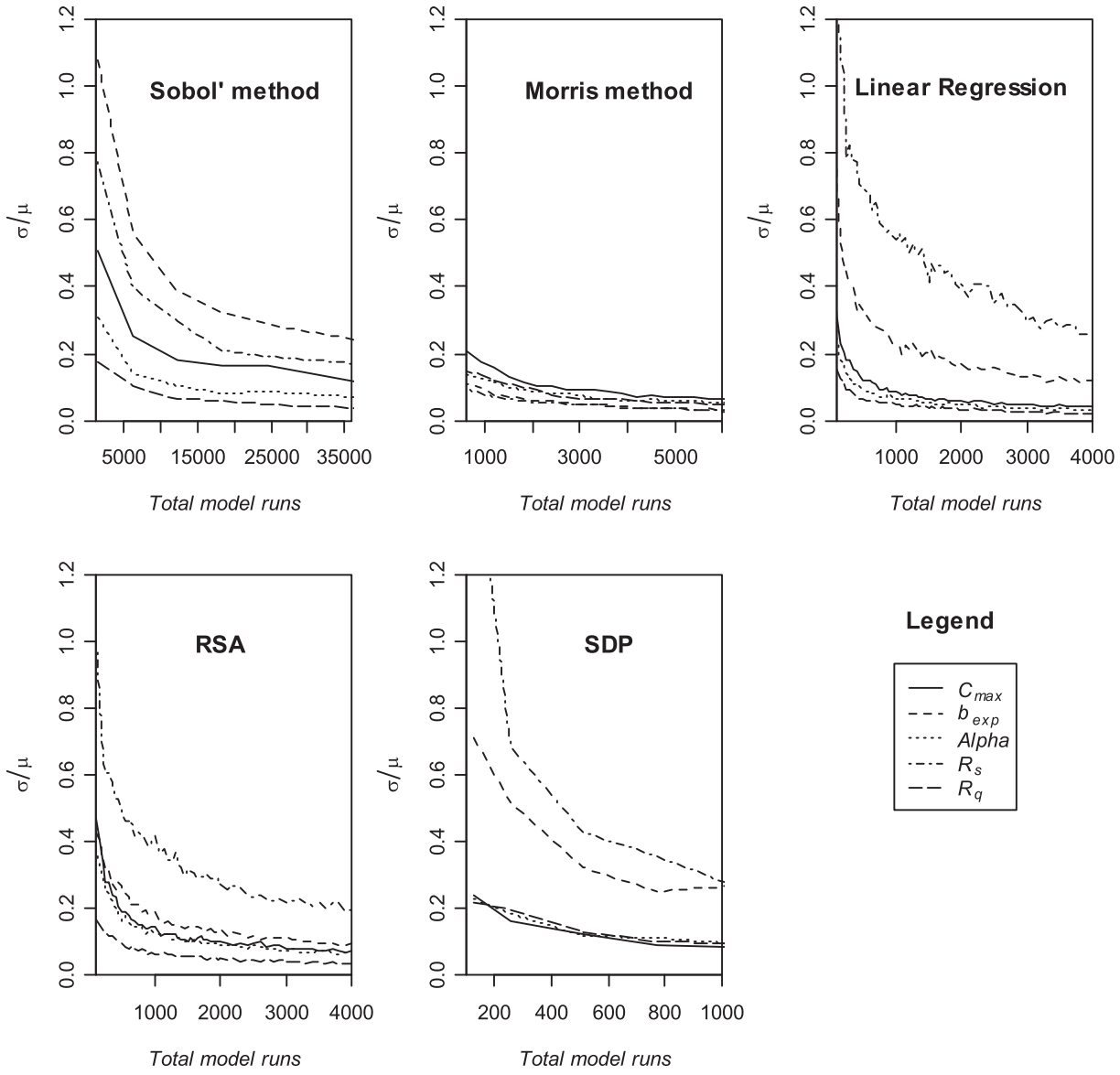

Fig. 2. Convergence of the applications of different SA techniques based on CLT. Each line in each plot denotes an evolution of the "coefficient of variance" (σ/μ) of the sensitivity index with the increasing base sample size (expressed as the total model runs).

Table 3
Applications of different sensitivity techniques to the HYMOD model based on CLT.

	Sobol' method (18,000[a])			Morris method (3000[a])			Linear regression (3000[a])			RSA (3000[a])			SDP (500[a])		
	Mean	Stdev	Rank[b]	Mean	Stdev	Rank	Mean	Stdev	Rank	Mean	Stdev	Rank	Mean	Stdev	Rank
C_{max}	0.277 (0.194, 0.339) [c]	0.044	**3**	0.693 (0.583, 0.823)	0.064	**3**	0.230 (0.207, 0.249)	0.011	**3**	0.177 (0.151, 0.208)	0.014	**4**/5 (0.98)[d]	0.219 (0.169, 0.273)	0.028	**3**
b_{exp}	0.038 (0.016, 0.065)	0.013	**4**	0.325 (0.295, 0.355)	0.016	**4**	0.115 (0.088, 0.141)	0.015	**4**	0.288 (0.226, 0.345)	0.031	**3**/2 (0.94)	0.029 (0.013, 0.049)	0.01	**4**/5 (0.99)
Alpha	0.440 (0.379, 0.518)	0.037	**2**	0.972 (0.847, 1.117)	0.071	**2**	0.283 (0.264, 0.307)	0.011	**2**	0.344 (0.293, 0.391)	0.025	**2**/3 (0.94)	0.377 (0.282, 0.443)	0.044	**2**
R_s	0.009 (0.006, 0.013)	0.002	**5**	0.176 (0.161, 0.192)	0.008	**5**	0.047 (0.021, 0.078)	0.014	**5**	0.119 (0.076, 0.169)	0.025	**5**/4 (0.98)	0.008 (0.003, 0.013)	0.003	**5**/4 (0.99)
R_q	0.712 (0.663, 0.785)	0.041	**1**	1.277 (1.129, 1.410)	0.080	**1**	0.419 (0.399, 0.436)	0.010	**1**	0.419 (0.388, 0.452)	0.017	**1**	0.650 (0.498, 0.800)	0.083	**1**

[a] This is the number of model runs required for each independent CLT replica to convergence.
[b] The bold number is the most possible rank for the corresponding parameter while others are the other possible ranks.
[c] This indicates the 95% CI of the sensitivity measure.
[d] This denotes the frequency of the main rank for the given parameter and the default is 1.

runs for each independent CLT replica), Linear Regression and RSA needed 2500–3000 model runs for each independent CLT replica, and SDP needed only 500 model runs for each independent CLT replica. As far as the convergence speed goes, SDP is the fastest; the Morris method, Linear Regression, and RSA have similar convergence speed; and the Sobol' method is the slowest. For CLT, there is no strict requirement on the base sample size for convergence, and it would depend on the modelling purpose – for example, screening out most insensitive parameters requires small base sample sizes – as a small base sample size leads to a large uncertainty range, and, hence, it is hard to distinguish the parameter's behavior (e.g., ranking), while more precise estimation (a narrow uncertainty range) requires more model simulations (large sample size).

An examination of mean sensitivity indices with increasing base sample size shows: while the estimations based on RSA remain almost constant only when the sample size m is over 1500; those of other SAs remain constant at a smaller base sample size (500 for the Sobol' method, 100 for the Morris method, 500 for Linear Regression, and 250 for SDP). This indicates RSA needs large base sample size to obtain stable sensitivity indices.

In Table 3, for the Sobol' method, sensitivity indices are distinctly different from each other in terms of means, 95% CIs and ranks; and apparently R_q outranks other parameters, followed by *Alpha*, C_{max}, b_{exp}, and R_s, respectively. There is not any overlap between their 95% CIs, indicating that a simulation with a larger sample may give more precise sensitivity indices but no new information on ranking. The uncertainty in estimated sensitivity indices (both first order S_i and total order S_{Ti}) is also visualized in the top-left box-plot of Fig. 3. In the plot, the upper box (filled in grey) in each column (for each parameter) indicates the box-and-whisker for S_{Ti}, and the lower box (filled in white) is for S_i. The difference between S_{Ti} and S_i represents the interaction of this parameter with all other parameters. For instance, the average of this difference for R_q is about 0.36, indicating a strong interaction with other parameters. As can also be seen, the strongest interacting parameters are R_q, *Alpha*, and C_{max}, while b_{exp} or R_s shows little interactions with other parameters. Table 4 further corroborates this pattern and characterizes their interactions. The group (R_q, *Alpha*, C_{max}) makes up to around 95% (the bold number in the lower triangular matrix) of the *NS*'s variation, while group (b_{exp}, R_s) only contributes around 1.5% (the bold number in the upper triangular matrix), and the interaction of these two groups represents only 3%. This means that the model output (*NS* coefficient in this study) is highly dependent on the maximum soil storage capacity (C_{max}) and quick flow component (*Alpha* and R_q), while the degree of spatial variability of soil moisture capacity (b_{exp}) and slow flow component (R_s) are not as important. Fig. 5 introduces the density functions of total effects (the solid line), and it does not show any overlap between density functions.

Results of the Morris method with different "p"s (i.e., 4, 8, 20, 40 and 60) after convergence show that as p increases, sensitivity ranks between parameters are more and more distinguishable. For example, when $p = 4$, one can only distinguish two groups of parameters, that is, the most sensitive group (R_q, *Alpha*, C_{max}) and the almost insensitive group (b_{exp}, R_s), and within each group difference in their means and standard deviations are not so significant. Using $p = 40$, their sensitivity ranks are more or less separable at a very low base sample size (100), and this will allow us to compare the results with the quantitative method (the Sobol' method). Therefore, in this paper, all the results concerning the Morris method are based on $p = 40$. These results (see in Table 3) are quite similar to those of the Sobol' method: the means and 95% CIs of the sensitivity indices are clearly separated; the sensitivity ranks have the same order as those of the Sobol' method, that is, R_q is the most sensitive parameter followed by *Alpha*, C_{max}, b_{exp} and R_s. This similarity between the Sobol' total effects and the μ^* measure has been highlighted by Campolongo et al. (2007) and Saltelli et al. (2008). Indeed, it shows that μ^* can be used as an excellent proxy of the total effect, when the computational cost of the model under analysis does not allow the use of the Sobol' method. The interactions/non-linearities of parameters are shown in the middle-left plot of Fig. 3. The larger the standard deviation, the stronger is the non-linearity/interaction of the parameter with other parameters. Similar to the Sobol' method, the strongest interacting parameters are R_q, *Alpha*, and C_{max}, while b_{exp} or R_s show less interaction with other parameters. This pattern confirms that: the Morris method provides a good approximation to the Sobol' method in terms of parameter rankings and interaction characterization. The density functions in the top right of Fig. 5 (solid lines) illustrate similar results.

Considering Linear Regression, it gives similar results to the Sobol' method and the Morris method in terms of distinguishable means, non-overlapped 95% CIs, and same parameter ranking, that is, R_q is the most sensitive parameter followed by *Alpha*, C_{max}, b_{exp} and R_s (see in Table 3). However, Linear Regression provides a R^2 below 40% which indicates that it accounts for less than 40% of the model uncertainty; and the reliability of the results is questionable. Furthermore, the result based on the rank transformation does not show any improvement on R^2, which indicates that Linear Regression based on rank transformation is also not applicable to the case study. The middle-left plot of Fig. 5 illustrates the density functions of the sensitivity indices.

As for RSA, there are lots of overlaps between their 95% CIs as indicated by the density functions (solid lines) in the middle right of Fig. 5; indicating the potential non-uniqueness of the ranking. The column "Rank" of RSA in Table 3 lists all possible ranks for each parameter and the major rank number is marked in bold based on its frequency in all the rankings of 100 CLT replicas. For example, C_{max} has the frequency of 98% to rank fourth and 2% to rank fifth, while b_{exp} has the frequency of 94% to rank third and 6% to rank second. Results with different filtering criteria (NS = 0.5 and 0.7, and 70–30%, 80–20%, and 90–10% separations) show that different parameter filtering criteria lead to different sensitivity indices and parameter ranks, and even the most sensitive and insensitive parameters change. As expected, the result of RSA is highly dependent on the choice of the filtering criterion. For NS = 0.7, there is no distinct convergence trend when the sample size m is smaller than 2000. This is because this filtering criterion with a smaller base sample size leads to few behavioral parameter sets (for instance, when the base sample size $m = 1000$, one CLT replica only led to four behavioral parameter sets, 0.4% of whole samples), which is insufficient to reliably estimate the behavioral cumulative distribution, in other words, RSA lacks statistical power in this case.

As for the SDP approach, sensitivity indices are estimated quite reliably at a very small sample size. For example, $m = 500$ (i.e., 500 model runs for each independent CLT replica) is almost sufficient to unambiguously distinguish the ranks of the five parameters except that b_{exp} and R_s have the small frequency of 1% to rank fifth and fourth, respectively. The estimation of first order indices is extremely efficient and reliable, and converges exactly to the values obtained with the Sobol' method. Total indices are estimated simply by summing the main effects and the second order effects involving each parameter. This is an approximation and one has to expect a downward bias, due to missing higher order interactions, in particular a third order term between C_{max}, *Alpha* and R_q (see discussion for the Sobol' method above). Nonetheless, this approximate analysis provides good estimates of the total effects, with the ranking exactly matching those of the Sobol' method. In particular, the shift in the ranks of C_{max} and *Alpha* between first order (where

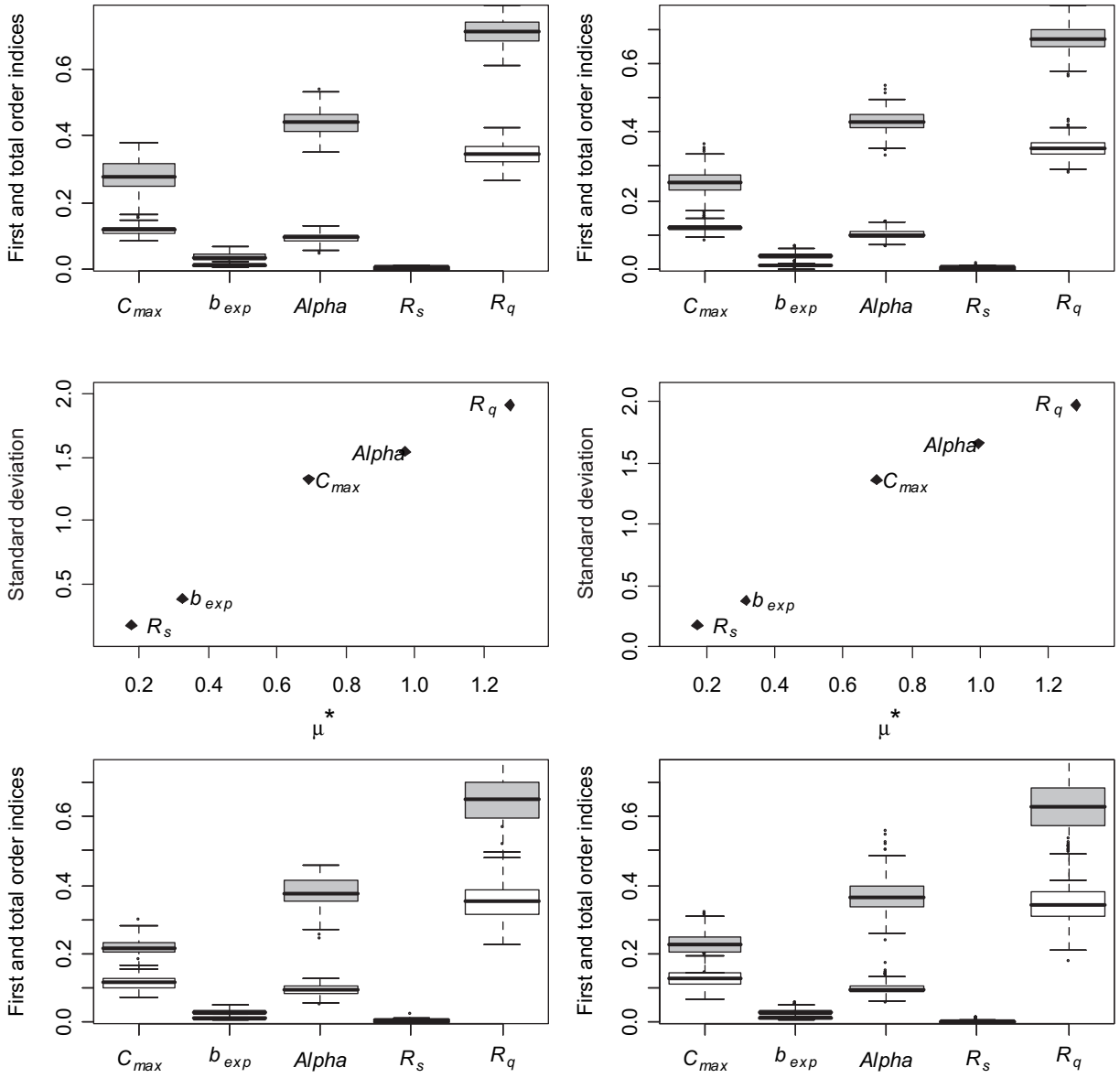

Fig. 3. The interaction of the HYMOD parameters (from top to bottom: Sobol' method, Morris method and SDP). The three left plots are based on CLT, and the three right plots are based on the bootstrap technique.

C_{max} ranks second and *Alpha* ranks third) and total order (where C_{max} ranks third and *Alpha* ranks second) is highlighted in Fig. 3, which is exactly the same results as those of the Sobol' method. In Fig. 3 (bottom left), one can appreciate the excellent results of SDP with only 500 model runs, with respect to Sobol' method using several thousands of model runs (18,000). In addition to this, the non-parametric regression approach provides a full emulator to the HYMOD code. The R^2 of the second order ANOVA model is about 94% (in sample), while tests of forecasts at untried points (out of sample cross-validation) provide a $R^2 = 91\%$. The unexplained 6% of sample fit is linked to third order interactions, presumably to the interaction between C_{max}, *Alpha* and R_q. This means that the estimated non-parametric ANOVA model describes reasonably well the behavior of the Nash-Sutcliffe objective function, which is a non-linear and non-monotonic function, as confirmed by the very small R^2 (<40%) of the Linear Regression.

Table 4
The estimated two-dimensional-closure sensitivity measure S^c_{ij} (upper triangular matrix) and two-dimensional-complementary-closure sensitivity measure S^c_{-ij} (lower triangular matrix).

	C_{max}	b_{exp}	*Alpha*	R_s	R_q
C_{max}		0.136	0.260	0.123	0.520
b_{exp}	0.686		0.112	**0.015**	0.367
Alpha	0.368	0.534		0.101	0.678
R_s	0.714	**0.953**	0.555		0.349
R_q	0.123	0.267	0.141	0.279	

As mentioned before, the results of the Sobol' method are taken as the reference for parameter rankings. Nonetheless, the Morris method provides the same ranking as that obtained by the Sobol' method. So does Linear Regression though its result is questionable because of low R^2. Compared to the Sobol' method, RSA mis-ranks C_{max} and b_{exp}, while in the Sobol' method, C_{max} belongs to most sensitive group and b_{exp} does not, and the ranking is highly dependent on the choice of the filtering criteria. Therefore, RSA should be used with care. The SDP method provides almost the same ranking as the Sobol' method except that b_{exp} and R_s have the frequency of 1% to rank fifth and fourth, respectively. The main

effects are extremely precise and reliable at a computational cost that is much cheaper than that for the Sobol' method. The total effects obtained from the SDP method, although slightly downward biased due to the missing third order interactions, provide an excellent approximation of the values obtained by the Sobol' method, with exactly the same ranking.

The TDCC analysis shows that for all the sensitivity techniques, TDCC values are significant at low sample sizes and increase as the sample size increases, and the most important factors identified are the same, that is, R_q, as listed in Table 3. This identifies the systematically consistent results on the most important factor for all the sensitivity techniques compared. The convergence based on CLT for each technique is a sufficient but not necessary condition for the TDCC analysis as TDCC focuses more on the rank instead of the numerical stability of the indices.

5.2. Result based on bootstrap technique

Fig. 4 shows the plot of sensitivity index versus base sample size m (expressed as total model runs in Fig. 4), which helps us to visually judge the convergence of simulations. Table 5 lists the converged sensitivity indices, 95% CIs and the rankings estimated based on the bootstrap technique.

The Sobol' method converged at the base sample size around 3000 (equivalent to 36000 model runs), although for values in the interval of [500, 1000] the total effects already provide a clear distinction and ranking of importance. The Morris method converged at m around 2000 (12,000 model runs). It already provides a clear distinction between the two negligible input factors ($b_{\exp}$ and R_s) and the other three parameters, at very small m values. This pattern confirms the good efficiency of the Morris method for screening purposes, where very few model runs have to be taken to highlight the negligible input factors. Linear Regression converged at m around 1000 (1000 model runs). RSA converged at about 5000 (5000 model runs). SDP estimates are rather stable beyond $m = 512$ (i.e., 512 model runs in total). As far as parameter ranking is concerned (without consideration of uncertainty), the Sobol' method, the Morris method, Linear Regression and SDP converged at the base sample size of 300 or even earlier, while RSA required a larger base sample size. This means that the rank result

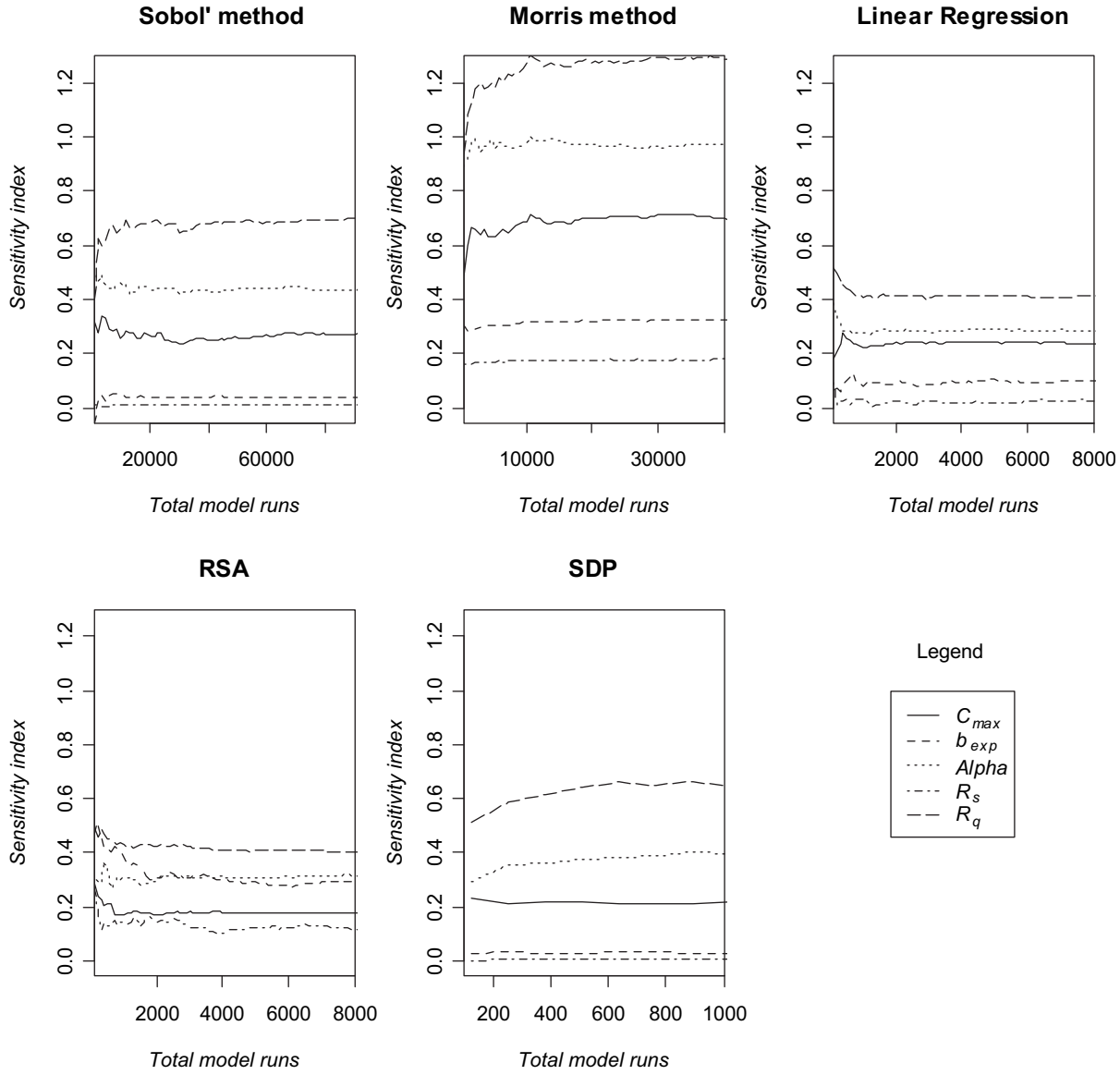

Fig. 4. Sensitivity index versus the increasing base sample size (expressed as total model runs).

Table 5
Applications of different sensitivity techniques to the HYMOD model based on the bootstrap technique.

	Sobol' method (36,000a)			Morris method (12,000[a])			Linear regression (1000[a])			RSA (5000[a])			SDP (512[a])		
	Sen. index	95% CI	Rank	Sen. index	95% CI	Rank	Sen. index	95% CI	Rank	Sen. index	95% CI	Rank	Sen. index	95% CI	Rank
C_{max}	0.258	(0.189, 0.318)	**3**	0.699	(0.643, 0.757)	**3**	0.225	(0.183, 0.264)	**3**/2 (0.947)[b]	0.177	(0.165, 0.203)	**4**/5 (0.976)	0.217	(0.169 0.291)	**3**/2 (0.997)
b_{exp}	0.039	(0.023, 0.056)	**4**	0.316	(0.299, 0.332)	**4**	0.081	(0.038, 0.128)	**4**/5 (0.932)	0.293	(0.266, 0.339)	**3**/2 (0.683)	0.030	(0.018 0.047)	**4**
Alpha	0.434	(0.375, 0.483)	**2**	0.994	(0.922, 1.072)	**2**	0.277	(0.233, 0.316)	**2**/3 (0.947)	0.307	(0.284, 0.354)	**2**/3 (0.683)	0.376	(0.288 0.466)	**2**/3 (0.997)
R_s	0.008	(0.005, 0.012)	**5**	0.175	(0.168, 0.183)	**5**	0.028	(0.002, 0.080)	**5**/4 (0.932)	0.126	(0.090, 0.180)	**5**/4 (0.976)	0.005	(0.002 0.009)	**5**
R_q	0.672	(0.604, 0.739)	**1**	1.279	(1.195, 1.368)	**1**	0.408	(0.378, 0.442)	**1**	0.407	(0.391, 0.434)	**1**	0.643	(0.493 0.806)	**1**

[a] The total model runs required for the application to converge.
[b] The bold number is the main rank (and possibly followed by the secondary rank) of the given parameter, and the number in the brackets denotes the frequency of the main rank (the default is 1).

based on RSA is very sensitive at a small base sample size. Compared to the convergence in Section 5.1, the bootstrap technique takes a longer base sample size but less total model runs to converge.

In Table 5, there are some overlaps in the 95% CI and in the density functions (dashed lines in Fig. 5) between the parameters for all the sensitivity techniques except the Sobol' method, the Morris method and SDP (for SDP, the only overlapped area between C_{max} and *Alpha* is very small as indicated by their frequencies in the last column of Table 5). For Linear Regression, the overlaps are between R_s and b_{exp}, and C_{max} and *Alpha*. For RSA, the overlaps are between R_s and C_{max}, and b_{exp} and *Alpha*. As a comparison, the results of SDP based on asymptotic Gaussian assumption (not shown) provide wider uncertainty estimation and more overlaps than both the bootstrap technique and the CLT do. From the overlaps of density functions (dashed lines) in Fig. 5, one can easily see the significance of the overlap for Linear Regression and RSA, and the insignificance for the other three techniques.

As far as the parameter ranking is concerned, the Morris method provides the unique and same order as the Sobol' method does, so does SDP (except C_{max} and *Alpha*, having the very small frequency of 0.3% to rank second and third, respectively, which is negligible). Linear Regression gives the same ranking order as the Sobol' method, although C_{max} and *Alpha* have a frequency of 5.3% to rank second and third, and b_{exp} and R_s have a frequency of 6.8% to rank fifth and fourth, respectively. Compared to the Sobol' method, RSA gives a different ranking to C_{max} and b_{exp} even if the converged indices are concerned, and there is a strong overlap between the density functions of b_{exp} and *Alpha* (dashed lines of the middle right plot in Fig. 5) with the frequency of 68.3% to be their major ranks. Fig. 3 shows the box-plots of the sensitivity indices for the Sobol' method (top right) and SDP (bottom right), and interactions in the sensitivity indices for the Morris method (middle right). All plots indicate the high interactions between R_q, *Alpha*, and C_{max}, while little interactions are detected for the other two parameters.

5.3. Comparison of results based on CLT and bootstrap technique

Based on Tables 3 and 5 and Figs. 2–5, a comparison can be made between the results with respect to both the CLT and the bootstrap technique:

1. For each SA technique, all the converged sensitivity indices based on the bootstrap technique (i.e., the column "Sens. index" of each technique in Table 5) are close to their corresponding mean sensitivity indices based on CLT (i.e., the column "Mean" of each technique in Table 3), and within the 95% CI ranges of CLT with the exception of the index for b_{exp} from Linear Regression (which can be explained by the fact that base sample size based on the bootstrap technique is only one-third of the one based on CLT). This applies to the major rankings of sensitivity indices. All above observations imply that the analysis based on the bootstrap technique is a good approximation to that based on CLT as far as sensitivity index and its ranking are concerned;
2. For the same SA technique, the application based on the bootstrap technique normally needs less model runs than that based on CLT. This is a clear and obvious advantage of the bootstrap technique, making it applicable to complex systems which requires longer time for a run;
3. A comparison of the estimated density functions both based on CLT and bootstrap technique (Fig. 5) shows: (i) For the Sobol' and Morris methods, the density functions from CLT (solid lines) are slightly flatter than those based on the bootstrap technique. This indicates that the bootstrap analysis tends to

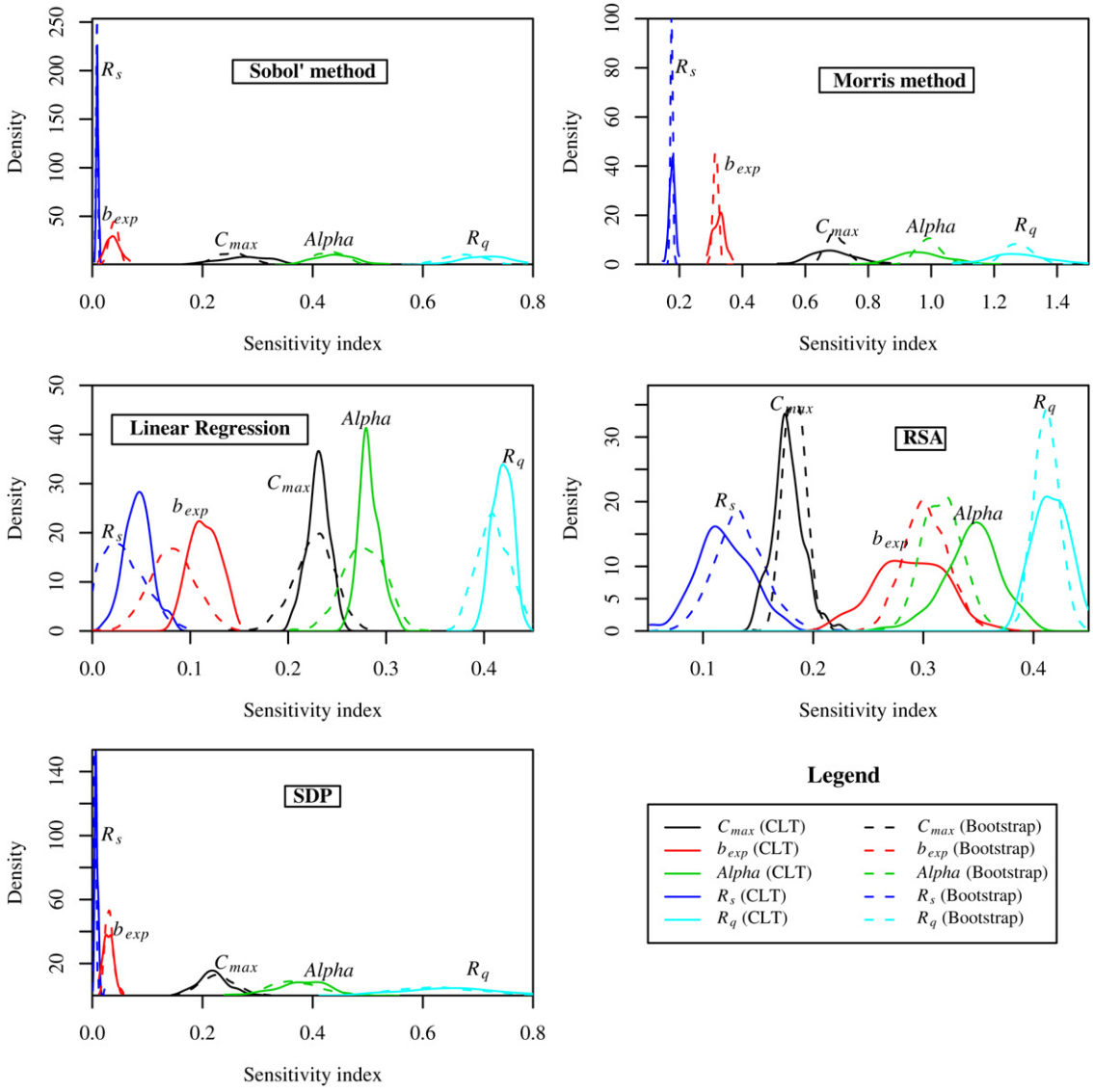

Fig. 5. Estimated density functions of sensitivity indices for each SA technique based on CLT (solid lines) and bootstrap technique (dashed lines).

generate a sharper distinction than sensitivity estimated with respect to 'reality' (represented by CLT analysis); (ii) For Linear Regression and RSA, the estimated density functions based on the bootstrap technique shift a bit and there are strong overlaps in some density functions (e.g., the overlap between b_{exp}'s and *Alpha*'s for RSA). This can be reflected by their overlap of the 95% CIs and rankings in Table 5. For Linear Regression, estimated density functions based on the bootstrap technique are flatter than those based on CLT due to the fact that the base sample size is only one-third of CLT. The converse is true for RSA; (iii) For SDP, density functions based on bootstrap technique are similar to those based on CLT. The ranking is almost unique and unambiguous except for the insignificant frequencies of b_{exp} and R_s to rank fifth and fourth based on CLT, and of C_{max} and *Alpha* to rank second and third based on the bootstrap technique, respectively.

Finally, as shown by the plots of interaction in Fig. 3, there is no significant difference between the analyses based on CLT and bootstrap technique for the Sobol' method, the Morris method, and SDP.

6. Conclusion

In this paper, two methods are proposed to monitor the convergence and uncertainty analyses of sensitivity indices of SAs: the CLT and the bootstrap techniques. These methods are implemented with five different Monte-Carlo based SA techniques to the HYMOD model. The results can be summarized as follows:

- As far as sensitivity index and its ranking are concerned, the bootstrap technique provides a good approximation to that based on CLT. For the Sobol' method, the Morris method and SDP, the uncertainty analysis based on the bootstrap technique provides a good approximation to the analysis based on CLT.
- The Morris method provides a very good proxy to the Sobol' total effect due to its effective screening capability. This is especially useful for computationally expensive models, where it is necessary to screen out unimportant factors.
- Although Linear Regression provides the same rank order as the Sobol' method, the usefulness is questionable as it considers less than 40% of the model uncertainty in the case study.

- The result of RSA primarily depends on the choice of the filtering criterion. RSA should be used with care.
- The SDP method provides excellent results with very few model evaluations. SDP results obtained with 512 model runs compare well with the results of the Sobol' method using several thousands of runs. For the SDP method, the uncertainty analysis based on the asymptotic Gaussian assumptions provides wider estimates with respect to the bootstrap and CLT techniques; the latter two methods provide similar results. In addition to the sensitivity estimation, the SDP approach, as other non-parametric regression approaches, provides a full emulator of the original computational model, and allow inferring the values of the model output at untried points.

Acknowledgement

The author gratefully thanks Dr. Marco Ratto at Joint Research Centre of European Commission (Italy) for constructive comments especially on the SDP method, Prof. Alain N. Rousseau of INRS-ETE (Canada), Prof. Wanhong Yang at University of Guelph (Canada), and Prof. Lloyd Chua at Nanyang Technological University (Singapore) for reviewing and correcting the paper, and finally the anonymous reviewers for the constructive comments.

References

Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., Srinivasan, R., 2007. Modelling hydrology and water quality in the pre-ailpine/alpine Thur watershed using SWAT. Journal of Hydrology 333 (2–4), 413–430.

Archer, G.E.B., Saltelli, A., Sobol, I.M., 1997. Sensitivity measures, ANOVA-like techniques and the use of bootstrap. Journal of Statistical Computation and Simulation 58 (2), 99–120.

Beven, K., Binley, A., 1992. The future of distributed models – model calibration and uncertainty prediction. Hydrological Processes 6 (3), 279–298.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resources Research 36 (12), 3663–3674.

Boyle, D.P., 2001. Multicriteria Calibration of Hydrological Models, Ph.D. Dissertation, Department of Hydrology and Water Resources, University of Arizona, Tucson.

Cacuci, D., 2003. Sensitivity and Uncertainty Analysis. In: Theory, vol. 1. Chapman and Hall.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling and Software 22, 1509–1518.

Confalonieri, R., Bellocchi, G., Tarantola, S., Acutis, M., Donatelli, M., Genovese, G., 2010. Sensitivity analysis of the rice model WARM in Europe: exploring the effects of different locations, climates and methods of analysis on model sensitivity to crop parameters. Environmental Modelling and Software 25 (4), 479–488.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schailby, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. Journal of Chemical Physics 59 (8), 3873–3878.

Cukier, R.I., Levine, H.B., Shuler, K.E., 1978. Nonlinear sensitivity analysis of multiparameter model systems. Journal of Computational Physical 26 (1), 1–42.

Doksum, K., Samarov, A., 1995. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. The Annals of Statistics 23 (5), 1443–1473.

Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, ISBN 0-471-22361-1.

EPA Council for Regulatory Environmental Modeling, 2003. Draft Guidance on the Development, Evaluation, and Application of Regulatory Environmental Models. http://www.modeling.uga.edu/tauc/otherpapers.html.

European Commission, 2005. Impact Assessment Guidelines, SEC(2005) 791. http://ec.europa.eu/governance/docs/index_en.htm.

Foscarini, F., Bellocchi, G., Confalonieri, R., Savini, C., Van den Eede, G., 2010. Sensitivity analysis in fuzzy systems: integration of SimLab and DANA. Environmental Modelling and Software 25 (10), 1256–1260.

Frey, H.C., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Analysis 22 (3), 553–578.

Grinstead, C.M., Snell, J.L., 1997. Introduction to Probability: Second Revised Edition. American Mathematical Society. ISBN: 10: 0-8218-0749-8, ISBN: 13: 978-0-8218-0749-1.

Gu, C., 2002. Smoothing Spline ANOVA Models. Springer-Verlag, New York.

Helton, J.C., Burmaster, D.E., 1996. Guest editorial: treatment of aleatory and epistemic uncertainty in performance assessments for complex systems. Reliability Engineering and System Safety 54 (2–3), 91–94.

Helton, J.C., Sallaberry, C., 2009. Computational implementation of sampling-based approaches to the calculation of expected dose in performance assessments for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada. Reliability Engineering and System Safety 94, 699–721.

Helton, J.C., Davis, F.J., Johnson, J.D., 2005. A comparison of uncertainty and sensitivity analysis results obtained with random and latin hypercube sampling. Reliability Engineering and System Safety 89 (3), 305–330.

Helton, J.C., 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive-waste disposal. Reliability Engineering and System Safety 42 (2–3), 327–367.

Helton, J.C., Martell, M.-A., Tierney, M.S., 2000. Characterization of subjective uncertainty in the 1996 performance assessment for the waste isolation pilot plant. Reliability Engineering and System Safety 69 (1–3), 191–204.

Hoffman, F.O., Hammonds, J.S., 1994. Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. Risk Analysis 14 (5), 707–712.

Iman, R.L., Conover, W.J., 1987. A measure of top-down correlation. Technometrics 29 (3), 351–357.

Iman, R.L., Helton, J.C., 1988. An investigation of uncertainty and sensitivity analysis techniques for computer-models. Risk Analysis 8 (1), 71–90.

Iman, R.L., Helton, J.C., 1991. The repeatability of uncertainty and sensitivity analyses for complex probabilistic risk assessments. Risk Analysis 11 (4), 591–606.

Iman, R.L., 1982. Statistical methods for including uncertainties associated with the geologic isolation of radioactive waste which allow for a comparison with licensing criteria. In: Kocher, D.C. (Ed.), Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste, Gatlinburg, TN, March 9–13, 1981. US Nuclear Regulatory Commission, Directorate of Technical Information and Document Control, Washington, DC, pp. 145–157.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling. 1 Theory. Water Resources Research 42, W03407. doi:10.1029/2005WR004368.

Li, G., Wang, S.W., Rabitz, H., 2002. Practical approaches to construct RS-HDMR component functions. Journal of Physical Chemistry 106, 8721–8733.

Li, G., Hu, J., Wang, S.W., Georgopoulos, P., Schoendorf, J., Rabitz, H., 2006. Random sampling-high dimensional model representation (RS-HDMR) and orthogonality of its different order component functions. Journal of Physical Chemistry A 110, 2474–2485.

McRae, G.J., Tilden, J.W., Seinfeld, J.H., 1982. Global sensitivity analysis – a computational implementation of the Fourier amplitude sensitivity test (FAST). Computers & Chemical Engineering 6 (1), 15–25.

Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 30 (2), 273–297.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174.

Myers, R.H., Montgomery, D.C., 1995. Response Surface Methodology: Process and Product in Optimization using Designed Experiments. Wiley & Sons, Inc., New York, NY, USA.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10 (3), 282–290.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal Royal Statistical Society, Series B 66, 751–769.

Pappenberger, F., Iorgulescu, I., Beven, K.J., 2006. Sensitivity analysis based on regional splits and regression trees (SARS-RT). Environmental Modelling and Software 21 (7), 976–990.

Pappenberger, F., Beven, K.J., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. Advances in Water Resources 31, 1–14.

Parzen, E., 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Paté-Cornell, M.E., 1996. Uncertainties in risk analysis: six levels of treatment. Reliability Engineering and System Safety 54 (2–3), 95–111.

Ratto, M., Pagano, A., Young, P., 2007a. State dependent parameter meta-modelling and sensitivity analysis. Computer Physics Communications 177, 863–876.

Ratto, M., Young, P.C., Romanowicz, R., Pappenberger, F., Saltelli, A., Pagano, A., 2007b. Uncertainty, sensitivity analysis and the role of data based mechanistic modelling in hydrology. Hydrology and Earth System Sciences 11, 1249–1266.

Ravalico, J., Dandy, G., Maier, H., 2010. Management Option Rank Equivalence (MORE) – A new method of sensitivity analysis for decision-making. Environmental Modelling and Software 25 (2), 171–181.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Statistical Science 4, 409–435.

Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environmental Modelling and Software 25 (12), 1508–1517.

Saltelli, A., Sobol', I.M., 1995. About the use of rank transformation in sensitivity analysis of model output. Reliability Engineering and System Safety 50 (3), 225–239.

Saltelli, A., Tarantola, S., Chan, K., 1999. Quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41 (1), 39–56.

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity Analysis. John Wiley & Sons Publishers. Probability and Statistics Series.

Saltelli, A., Tarantola, A., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: a Guide to Assessing Scientific Models. John Wiley & Sons, Ltd.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer. Wiley & Sons, Chichester, United Kingdom.

Saltelli, A., 1999. Sensitivity analysis: could better methods be used? Journal of Geophysical Research-Atmospheres 104 (D3), 3789–3793.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications 145 (2), 280–297.

Sobol', I.M., 1990. Sensitivity estimates for nonlinear mathematical models. Matematicheskoe Modelirovanie 2, 112–118 (in Russian), translated in Mathematical Modelling and Computationial Experiments 1, 407–414 (1993).

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in Peel Inlet, II, identification of critical uncertainties via generalised sensitivity analysis. Water Resources Research 14, 43–49.

Storlie, C.B., Helton, J.C., 2008. Multiple predictor smoothing methods for sensitivity analysis: example results. Reliability Engineering and System Safety 93, 55–77.

Tang, Y., Reed, R., Wagener, T., van Werkhoven, K., 2007. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrology and Earth System Sciences 11, 793–817.

Tarantola, S., Giglioli, N., Jesinghaus, J., Saltelli, A., 2002. Can global sensitivity analysis steer the implementation of models for environmental assessments and decision-making? Stochastic Environmental Research and Risk Assessment 16 (1), 63–76.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, A., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. Journal of Hydrology 324 (1–4), 10–23.

Varella, H., Guérif, M., Buis, S., 2010. Global sensitivity analysis measures the quality of parameter estimation: the case of soil parameters and a crop model. Environmental Modelling and Software 25 (3), 310–319.

Vrugt, J.A., Bouten, W., Gupta, H.V., Sorooshian, S., 2002. Toward improved identifiability of hydrologic model parameters: the information content of experimental data. Water Resources Research 38 (12), 1312,. doi:10.1029/2001WR001118.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research 39 (8), 1201,. doi:10.1029/2002WR001642.

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. Water Resources Research 41, W01017. doi:10.1029/2004WR003059.

Wahba, G., 1990. Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics.

Wasserman, L., 2005. All of Statistics: a Concise Course in Statistical Inference. Springer Texts in Statistics.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007. Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. Journal of Hydrology 340 (167), 167–182.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. Journal of Hydrology 204 (1–4), 83–97.

Young, P.C., 1993. Time variable and state dependent modelling of nonstationary and nonlinear time series. In: Rao, T.S. (Ed.), Developments in Time Series Analysis. Chapman and Hall, London, pp. 374–413.