Environmental Modelling & Software 79 (2016) 135–152

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Global Sensitivity Analysis of environmental models: Convergence and validation

Fanny Sarrazin [a, *], Francesca Pianosi [a], Thorsten Wagener [a, b]

[a] Department of Civil Engineering University of Bristol, University Walk, BS81TR, Bristol, UK
[b] Cabot Institute, Royal Fort House, University of Bristol, BS8 1UJ, Bristol, UK

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form
11 January 2016
Accepted 1 February 2016
Available online xxx

Keywords:
Sensitivity Analysis
Screening
Ranking
Convergence
Validation
SWAT

ABSTRACT

We address two critical choices in Global Sensitivity Analysis (GSA): the choice of the sample size and of the threshold for the identification of insensitive input factors. Guidance to assist users with those two choices is still insufficient. We aim at filling this gap. Firstly, we define criteria to quantify the convergence of sensitivity indices, of ranking and of screening, based on a bootstrap approach. Secondly, we investigate the screening threshold with a quantitative validation procedure for screening results. We apply the proposed methodologies to three hydrological models with varying complexity utilizing three widely-used GSA methods (RSA, Morris, Sobol'). We demonstrate that convergence of screening and ranking can be reached before sensitivity estimates stabilize. Convergence dynamics appear to be case-dependent, which suggests that "fit-for-all" rules for sample sizes should not be used. Other modellers can easily adopt our criteria and procedures for a wide range of GSA methods and cases.

1. Introduction

Sensitivity Analysis (SA) aims to characterize the impact that changes in the model input factors (e.g. parameters, initial states, input data, time/spatial resolution grid etc.) have on the model output (e.g. a statistic of the simulated time series, such as the average simulated streamflow, or an objective function, like the Root Mean Squared Error). SA is a diagnostic tool that can guide model calibration and verification, support the prioritization of efforts for uncertainty reduction, or help with model-based decision-making (Norton, 2015; Pianosi et al., 2016; Song et al., 2015). Such purposes are generally implemented as four different objectives of GSA: screening (or Factor Fixing), ranking (or Factor Prioritization), Variance Cutting, and Factor Mapping (Saltelli et al., 2008). Screening refers to the identification of those input factors, if any, which have no influence on the model output and therefore can be fixed to any value within their feasible range with negligible implications on the output. For instance, in Kannan et al. (2007) and in Vanuytrecht et al. (2014), screening of model parameters is applied as a preliminary step to inform a subsequent calibration, which is tailored to the subset of influential parameters. Ranking describes the ordering of the input factors according to their relative influence on the model output. It is typically used to enhance our understanding of the model and to identify dominant controls of the model's behaviour (e.g. Van Werkhoven et al., 2008), as well as to prioritize efforts for uncertainty reduction (e.g. Sin et al., 2011), or to support model development (Hartmann et al., 2013). The Variance Cutting setting is used for the reduction of the output variance to a value below a user chosen tolerance. It aims at obtaining specific sensitivities for the different input factors and is, for example, applied in reliability and risk assessment (e.g. Saltelli and Tarantola, 2002). Finally, Factor Mapping aims at identifying those conditions (e.g. sub-ranges of input factors like parameters or forcing inputs) that produce critical values of the output. It can be used to enhance model understanding (e.g. Spear and Hornberger, 1980) or to support robust decision-making (Singh et al., 2014).

Unlike Local Sensitivity Analysis (LSA), where the variability of the model output is explored around some reference input factor setting (e.g. Ljung, 1999 for a general link between LSA and model calibration; Hill and Tiedeman, 2007, for an example application to groundwater models), Global Sensitivity Analysis (GSA) rather attempts to explore the entire space of the input factors. It therefore typically requires larger computational resources than LSA. Generally, the implementation of GSA methods is sampling-based and the value of the sensitivity indices are approximated using

* Corresponding author.
E-mail address: fanny.sarrazin@bristol.ac.uk (F. Sarrazin).

http://dx.doi.org/10.1016/j.envsoft.2016.02.005

Monte Carlo (MC) simulations. A critical step of sampling-based GSA is therefore the choice of the sample size to run the MC experiment. If the sample is too small to adequately cover the input space, the analysis may not provide robust results. On the other hand, for very large sample sizes the computational cost may become very high while not improving the precision of the results significantly. In environmental applications, where models are often complex and simulations expensive, an acceptable trade-off has to be found between the need to obtain robust results and the need to limit computational cost.

The total number of model evaluations (N) used in GSA typically increases with the number of model input factors (M). For some GSA methods, depending on the methodology used to derive the estimates of the sensitivity indices, N is expressed as a function of M and of a base sample size (n) that must be specified by the user (i.e. $N = f(n,M)$). Thus, choosing the value of the total number of model evaluations (N) comes down to choosing the value of the base sample size (n). For other methods, no explicit expression relates N to M and therefore N is directly chosen by the GSA user ($N = n$). Suggestions for the choice of n can be found in the literature for several GSA methods. For instance, Saltelli et al. (2008, Table 6.9) report typical values of n for the Elementary Effect Test (EET, or method of Morris (Morris, 1991)), for Regional Sensitivity Analysis (RSA; Young et al., 1978; Spear and Hornberger, 1980) and for Variance-Based Sensitivity Analysis (VBSA; Sobol', 1990; Saltelli, 2002). However, a relatively limited number of studies actually focus on a rigorous assessment of the convergence of GSA results. Fig. 1 reports several examples taken from the literature regarding the relationship between N and M for EET, RSA and VBSA. From these studies, we make three observations:

1. Previous convergence studies assessed different types of convergence, namely convergence of the sensitivity indices, of the screening results (identification of the non-influential input factors), and of the ranking (ordering of input factors according to their relative importance). This lack of uniformity in the definition of 'convergence' makes it difficult to consistently compare the results obtained for models of different complexities when using different GSA methods. However, a preliminary conclusion that seems to emerge from these studies is that different sample sizes are required for different types of convergence. For instance, in the case of EET, Vanuytrecht et al. (2014) highlight that while a low sample size ($n = 25$) can be suitable for screening, it can be insufficient for factor ranking. Nossent et al. (2011) find that a base sample size of 12,000 is needed to ensure the convergence of Variance-Based sensitivity indices in their specific case study, however, a much smaller sample size ($n < 2000$) is sufficient if one is only interested in ranking the most important input factors.
2. Within a given type of convergence, different values of the base sample size are found for the same method when applied to different models. For instance to ensure convergence of the value of Variance-Based sensitivity indices (Fig. 1 bottom left panel), Tang et al. (2007) use a base sample size n of 8192 (for a case study with 18 input factors), while Yang (2011) uses n equal to 3000 (for a case study with 5 input factors). This suggests that the base sample size may also be a function of the number of input factors or of other characteristics of the model or of the case study. It is also worth noting that these studies show that convergence is often reached using a base sample size significantly larger than the values suggested in Saltelli et al. (2008).
3. Convergence is generally assessed based on a visual analysis of the stability of the results for increasing sample size. Some authors use the confidence intervals of the sensitivity indices for a more quantitative assessment of their convergence (e.g. Campolongo and Saltelli, 1997; Nossent et al., 2011). However, they do not explicitly define a convergence criterion. Herman et al. (2013) and Vanrolleghem et al. (2015) both introduce a quantitative criterion to measure the convergence of the sensitivity indices values (that will be discussed in Section 2.1), but they do not consider the convergence of ranking or screening.

Another issue for GSA is the choice of the screening threshold i.e. a threshold value for the sensitivity indices below which factors are classified as insensitive (more details in Section 2.1). In this respect, the following can be learned from existing studies:

1. For Variance-Based SA, the input factors that have a sensitivity index below 0.01 are often considered non-influential (Tang et al., 2007; Sin et al., 2011; Cosenza et al., 2013; Vanrolleghem et al., 2015). The adequacy of this screening threshold is tested in Tang et al. (2007), however the validation strategy used in that work (based on a visual approach introduced by Andres (1997)) has some limitations that we discuss and overcome here (more details in Section 2.2). Nossent et al. (2011) consider a screening threshold value of zero. They identify as statistically significant any input factor for which the lower bound of the confidence interval on the sensitivity index is positive. This method is quite conservative since, in our experience, a sensitivity index could have positive confidence bounds, and therefore a non-zero value, even if the input factor has negligible effect on the output.
2. EET, which is widely used for screening purpose, provides a relative measure of sensitivity that has a different meaning and range of variation depending on the model output definition in the particular case under study. Therefore, case-specific threshold values are usually taken (Vanuytrecht et al., 2014) and little guidance exists in the literature on this topic. Cosenza et al. (2013) and Vanrolleghem et al. (2015) present an attempt at defining an absolute value for the screening threshold for EET. However, they do not validate the adequacy of their proposed threshold values.

Based on this literature review, we believe that there is a lack of guidance to support GSA users in the choice of an adequate sample size and in the definition of a screening threshold, while there is an opportunity for improving current approaches to the validation of GSA results. Thus, the objectives of the present study are:

1. To define quantitative criteria to assess different types of convergence of GSA results, i.e. convergence of sensitivity indices, ranking and screening.
2. Based on these quantitative convergence measures, to investigate the convergence of three widely used GSA methods and to assess whether it is possible to give general guidelines for an adequate choice of the base sample size.
3. To develop a methodology to quantitatively validate screening results and therefore to formally investigate the adequacy of different choices for the screening threshold.

Here, we consider three widely used GSA methods, the Elementary Effect Test (EET), Regional Sensitivity Analysis (RSA) and Variance-Based Sensitivity Analysis (VBSA), implemented in the Sensitivity Analysis For Everybody (SAFE) toolbox (Pianosi et al., 2015). We apply GSA to three hydrological models of increasing complexity (HyMod, HBV and SWAT). The input factors are the model parameters and the output is the model accuracy. However, our approach could equally be applied to other GSA methods or models, and with different experimental set-ups, i.e. different

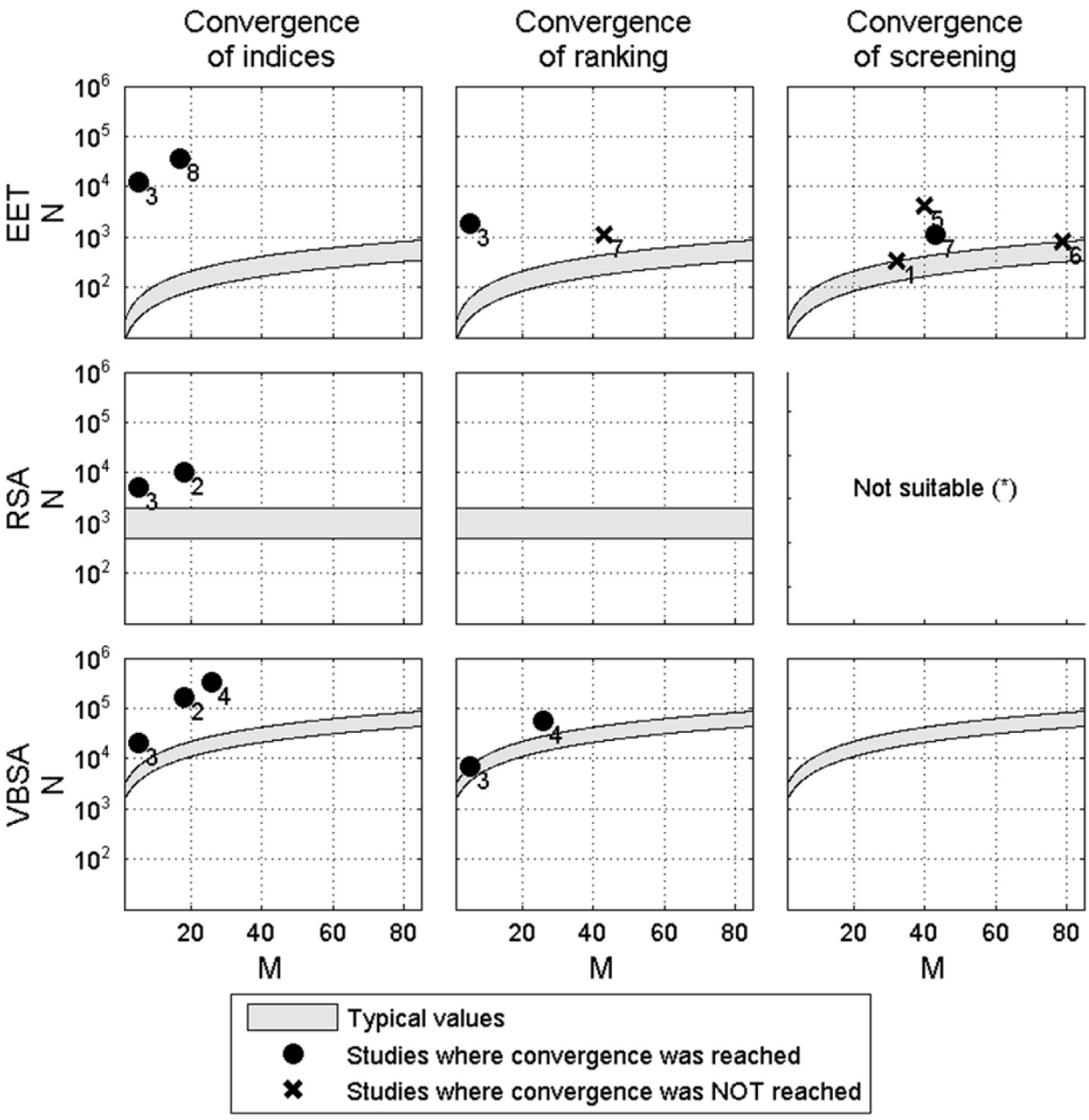

Fig. 1. Number of model evaluations (N) used in GSA against number of input factors (M) from several examples reported in the literature. Circles indicate that convergence is reached (respectively convergence of sensitivity indices, ranking and screening) and crosses indicate that convergence is not reached. The filled areas correspond to typical values of the sample size used in the literature (values from Saltelli et al., 2008, Table 6.9). N is computed as follows: $N = n \cdot (M + 1)$ for applications of Elementary Effect Test (EET), $N = n$ for applications of Regional Sensitivity Analysis (RSA), $N = n \cdot (M + 2)$ for applications of Variance- Based method (VBSA), where n is a base sample size chosen by the GSA user. The results are taken from [1]Campolongo and Saltelli (1997), [2]Tang et al. (2007), [3]Yang (2011), [4]Nossent et al. (2011), [5]Nossent and Bauwens (2012), [6]Cosenza et al. (2013), [7]Vanuytrecht et al. (2014), [8]Vanrolleghem et al. (2015). (*) RSA is not used for screening since it neglects parameter interactions.

definition of the model output and of the input factors subject to GSA (e.g. boundary conditions, errors in input forcing data, model resolution, etc.). Following this introduction, in Section 2 we define the convergence criteria and the validation procedure for the screening results and we describe the workflow adopted for the experiments. Section 3 presents briefly the three GSA methods and the three case studies analysed. We then report the results obtained for convergence and for screening validation in Section 4. We discuss meaning, implications and limitations of these results in Section 5.

2. Methodology

2.1. Definition of convergence criteria

In this section we provide three definitions of 'convergence' of GSA results and we propose criteria to quantitatively assess the different types of convergence. By 'convergence' we mean here the fact that GSA results do not change (or change to a limited degree) when using a different sample of model evaluations (of equal or larger size). We suggest distinguishing three different types of convergence (Fig. 2):

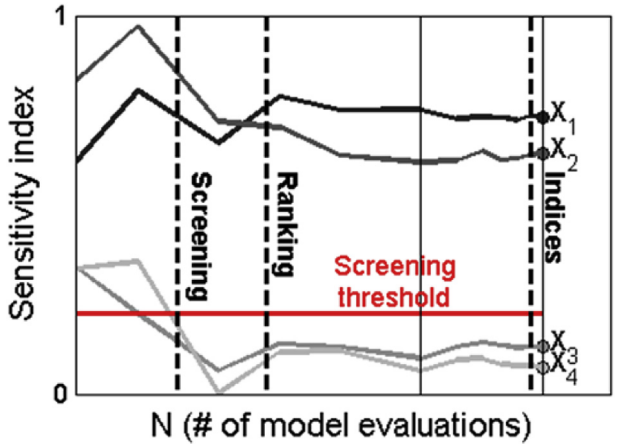

Fig. 2. Definition of convergence for the three GSA objectives. The figure reports the value of the sensitivity index against the number of model evaluations (N) in a hypothetical example with four parameters. Vertical lines indicate the convergence of the screening, ranking and indices. The screening converges when the partitioning between influential parameters and non-influential parameters (indices below the screening threshold) stabilizes. The ranking converges when the ordering among the parameters stabilizes. The sensitivity indices converge when their value stabilizes.

1) Convergence of the sensitivity indices, which is reached if the values of the indices remain stable;
2) Convergence of ranking, which is achieved if the ordering between the parameters remains stable;
3) Convergence of screening, which is reached if the partitioning between sensitive and insensitive parameters remains stable.

We propose three indicators that can be used to assess the three types of convergence defined above. All three indicators satisfy the following properties:

(a) They are quantitative indicators, i.e. they are computed through a numerical, reproducible procedure;
(b) They are efficient, i.e. the numerical procedure for their computation does not require additional model evaluations;
(c) They are easy to interpret and they allow for comparison across case studies and GSA methods.

The convergence indicators are described in the following paragraphs. They all measure the degree of uncertainty in GSA results, which is estimated via the bootstrap technique (Efron and Tibshirani, 1993; Archer et al., 1997). In bootstrapping, many different resamples are constructed by drawing randomly with replacement from the original sample of the model input/output so that no additional model runs are required and property (b) above is respected. The drawback is that resamples are obviously not independent from each other. A discussion regarding the quality of the bootstrap can be found in other studies, e.g. for estimating the mean of a distribution (Yang, 2011) and for its quantiles (Romano and Shaikh, 2012). Under certain conditions, the reliability of the bootstrap technique may be questioned, for instance when the sample size is small (Isaksson et al., 2008). We elaborate on this issue in Section 5.3.

2.1.1. Convergence of the sensitivity indices value

To assess the convergence of the sensitivity indices, we compute the width of the 95% confidence intervals (5% significance level) of the index distribution obtained by bootstrapping. We use the maximum width of the confidence intervals across all the model input factors as a summary statistic:

$$Stat_{indices} = \max_{i=1\ldots M}\left(S_i^{ub} - S_i^{lb}\right) \quad (1)$$

where S_i^{ub} and S_i^{lb} are the upper and lower bounds of the sensitivity index of the i-th input factor while M is the number of input factors. A value of the width of the confidence interval close to zero indicates that the sensitivity index has converged. Since in our study we use normalized sensitivity indices that vary between 0 and 1 (see Section 3.1), we could define an absolute threshold value for $Stat_{indices}$, below which convergence is considered to be reached. In our experience, we found that a reasonable choice for this threshold is 0.05. Other threshold values could be considered, for instance a percentage of the sensitivity index value of the most influential input factor as in Herman et al. (2013).

Our convergence criterion is quite different from the one by Vanrolleghem et al. (2015). In that study, the authors measure the variability of the sum of the sensitivity indices between two random samples of different size. Convergence is reached when this variability is low. We believe that their criterion is a necessary, but not sufficient condition for convergence for two reasons. First, it does not ensure that the sensitivity indices for all input factors have converged individually. Second, it assesses the variability between two random samples only and therefore it could happen that this variability is low even if convergence is not actually reached (the two samples can give similar results 'by chance'). Our criterion instead is based on the statistic of Eq. (1), which measures the variability of the sensitivity estimates of all individual input factors and across multiple resamples.

2.1.2. Convergence of input factor ranking

To assess the convergence of the ranking process, we use a rank correlation coefficient that quantifies the agreement between pairs of rankings generated using different bootstrap resamples. We initially considered and compared different rank correlation coefficients, starting with Spearman's rank correlation coefficient (Spearman, 1904). Its main limitation, when used in our context, is that it gives the same importance to rank differences occurring in the higher positions of the ranking (highly sensitive input factors) as in the lowest (insensitive input factors). However, rank reversals between insensitive input factors are typically of scarce interest since the main aim of ranking is to separate out and sort the most sensitive input factors. This limitation of the Spearman rank correlation coefficient is described by Iman and Conover (1987) and shown in Section C of our Supplementary Material using a theoretical example. Therefore, we decided that this indicator is unsuitable for our purposes.

Other studies attempt to deal with the limitation of the Spearman rank correlation coefficient. Iman and Conover (1987) propose to compute a correlation coefficient based on Savage scores (Savage, 1956) instead of ranks. Dancelli et al. (2013) introduce weights in the Spearman rank correlation coefficient formula (these weights are decreasing functions of the ranks). These two studies partially overcome the limitation of the unweighted Spearman coefficient by giving more weight to rank reversals occurring at the top of the ranking (most sensitive input factors). However, rank reversals of low-sensitivity input factors can still have a significant impact when a model has a large number of low-sensitivity input factors and therefore some of them are in relatively high ranking positions despite having a small sensitivity index (see again Section C of our Supplementary Material). We tackle this situation by proposing the following adjusted and weighted rank correlation coefficient:

$$\rho_{s,j,k} = \sum_{i=1}^{M} \left| R_i^j - R_i^k \right| \frac{\max_{j,k} \left(S_i^j, S_i^k \right)^2}{\sum_{i=1}^{M} \max_{j,k} \left(S_i^j, S_i^k \right)^2} \quad (2)$$

Where S_i^j and S_i^k are the values of the sensitivity index and R_i^j and R_i^k are the ranks of the i-th input factor, estimated using the j-th and the k-th bootstrap resample respectively.

This indicator emphasizes the disagreements in the ranking for the most sensitive input factors while neglecting the disagreements for the least sensitive input factors by directly using the sensitivity values to weight rank reversals. As a weight function we choose the squared maximum sensitivity index value between the two samples. This indicator has a clearly interpretable meaning: it represents the (weighted) average distance in the input factor ranks, obtained over two different bootstrap samples. The choice of the convergence statistic for ranking is further analysed in Section C of our Supplementary Material. In order to aggregate the rank coefficients obtained over all possible pairs of bootstrap resamples, we use the 95% quantile value (5% significance level):

$$Stat_{ranking} = Q_{\substack{0.95 \\ j,k}} \left(\rho_{s,j,k} \right) \quad (3)$$

We consider that the convergence of the ranking is reached when the statistic of Eq. (3) falls below 1. This choice is motivated by the fact that a value of the weighted and adjusted rank correlation coefficient (Eq. (2)) equal to 1 means that, on average, the differences in the ranking for the most sensitive input factors are less than one position.

2.1.3. Convergence of input factor screening

Screening the model input factors (in our application the model parameters) consists in separating them in two groups, the influential and the non-influential (insensitive) ones.

Theoretically, input factors are completely insensitive when the corresponding sensitivity index is equal to zero. In practice, when sensitivity indices are approximated via MC simulation, they are likely to take a very small positive value even when their (unknown) exact value is zero. Moreover, the objective of screening is often to identify not only the parameters that are completely insensitive, but also the ones that have a small but negligible effect. For these reasons, it is common practice to assume a threshold value T for the sensitivity index below which the input factors are considered as insensitive (e.g. Tang et al., 2007; Sin et al., 2011; Cosenza et al., 2013; Vanrolleghem et al., 2015). For a given value of the screening threshold T, the corresponding subset of insensitive input factors X_0 is defined as follows:

$$X_0 = \{x_i \ \ when \ S_i < T\} \quad (4)$$

where x_i is the i-th input factor and S_i is the sensitivity index (bootstrap mean) for the i-th input factor. In principle, the screening convergence might be assessed by measuring the stability in the partitioning as defined by Eq. (4). However, results would be highly dependent on the choice of the screening threshold T, whose exact value is not known a priori. Here, we therefore use a proxy measure for the screening convergence. We set the threshold in Eq. (4) to a relatively high value ($T = 0.05$) so that X_0 takes the meaning of set of "lower-sensitivity" input factors rather than set of "insensitive" input factors. Then, we use as a summary statistic the maximum width of the 95% confidence intervals across the lower-sensitivity input factors in X_0:

$$Stat_{screening} = \max_{x_i \in X_0} \left(S_i^{ub} - S_i^{lb} \right) \quad (5)$$

Similar to the convergence of the sensitivity indices (Eq. (1)), we consider that screening convergence is reached when $Stat_{screening}$ is below a value of 0.05. In other words, we assume that screening convergence has been reached when the sensitivity indices for the lower-sensitivity input factors have converged. We can then investigate whether all input factors in X_0 are actually insensitive using the validation test presented in the next section. This test is also a tool to determine a posteriori the value of the screening threshold T that would identify insensitive input factors in the case under study.

2.2. Validation procedure for screening

In this section we discuss and review two methods that can be used to validate the screening results obtained by GSA. Both methods aim to detect possible effects of the input factors classified as insensitive to avoid classifying influential input factors as insensitive. We denote the model output as y and the vector of input factors subject to GSA as X.

Andres (1997) proposes a method to evaluate the accuracy of the set X_0 defined in Eq. (4) and obtained by a generic GSA approach. In Andres' test, three sets of samples are generated. Set 1 is obtained by sampling the entire input factor space. In set 2 only the non-influential input factors (X_0) are allowed to vary while the influential input factors (denoted hereon by $\overline{X_0}$)are fixed to a prescribed value (for instance, the default parameter value from literature, or the mean of the assumed distribution). Finally in set 3, the influential input factors are sampled within their feasible range while the non-influential input factors are kept fixed. The actual value we fix the input factors at should not matter if the input factors are indeed insensitive. Three sets of model output samples are then obtained through MC simulations: the set of unconditional outputs $\{y\}$ (obtained from input factor set 1) and the two sets of conditional outputs $\{y|\overline{X_0}\}$ (from set 2) and $\{y|X_0\}$ (from set 3). The original test consists of a visual analysis of the two scatter plots that are obtained by plotting the unconditional output samples $\{y\}$ against the two conditional sets $\{y|\overline{X_0}\}$ and $\{y|X_0\}$. The input factors in X_0 are confirmed to be non-influential when in the first plot the conditional samples $y|\overline{X_0}$ align along a horizontal line (i.e. the output does not vary if varying the input factors in X_0 only) and when in the second plot the conditional samples $\{y|X_0\}$ align along a 45° line (i.e. the output variability when varying all input factors but those in X_0 is the same as when varying them all). Tang et al. (2007) and Nossent et al. (2011) use this approach to validate their screening results. These authors also propose to quantify the satisfaction of this screening test by computing the correlation coefficient of the scatter plots. However, the correlation coefficient also takes a value close to one even when the points align along a straight line that does not coincide with a 45° line. Therefore, a high value of the correlation coefficient does not necessarily indicate that the input factors in X_0 are insensitive.

In this study, we use a variation of the original Andres' test, first introduced by Pianosi and Wagener (2015). It is based on the computation of the Kolmogorov–Smirnov (KS) statistic (Kolmogorov, 1933; Smirnov, 1939; see Wall, 1996 for a general introduction) to estimate the discrepancy between the sets of unconditional and conditional outputs. Specifically in this test, the first step is to compute the empirical unconditional Cumulative Distribution Function (CDF) $F_y(y)$ and the empirical conditional CDF $F_{y|X_0}(y)$ of the model output. Then, a two-sample KS-test can be applied to test the null hypothesis that the two CDFs ($F_y(y)$ and

$F_{y|X_0}(y)$) are drawn from the same distribution. To this end, the KS-statistic is computed:

$$\widehat{KS}(X_0) = \max_y \left| F_y(y) - F_{y|X_0}(y) \right| \tag{6}$$

The null hypothesis is rejected if the KS-statistic between the two CDFs is above a critical value KS_{crit}. For a given significance level of the test α_c, $KS_{crit}(\alpha_c)$ is computed as:

$$KS_{crit}(\alpha_c) = c(\alpha_c)\sqrt{\frac{N_u + N_c}{N_u N_c}} \tag{7}$$

where N_u and N_c are the number of samples used to build the empirical CDFs, and the critical value $c(\alpha_c)$ can be found in the literature (e.g. Wall (1996)).

Given that the KS statistic of Eq. (6) depends on the conditioning values attributed to the input factors in X_0, the test should be repeated at different conditioning values, thus obtaining a set of KS values, which are then aggregated using a summary statistic. Here, we aggregate them by taking the maximum $\widehat{KS}_{max}$over a number n_c of conditioning values:

$$\widehat{KS}_{max} = \max_{X_{0,1}\ldots X_{0,nc}} \left(\widehat{KS}(X_{0,i}) \right) \tag{8}$$

The validation test we apply in our study compares the statistic $\widehat{KS}_{max}$ of Eq. (8) with the critical value $KS_{crit}(\alpha_c)$ of Eq. (7). It is worth noting that unlike the original two-sample KS-test, which consists in the comparison of two CDFs, our validation test consists in the comparison of n_c CDFs (conditional CDFs) to a reference CDF (unconditional CDF). Therefore, given the value of the significance level α_c used to compute the critical value of the KS-statistic of Eq. (7), the significance level of our validation test, denoted as α_t, is higher than α_c. In particular, if the n_c KS-statistics are considered to be independent, it can be shown that $\alpha_t = 1 - (1 - \alpha_c)^n$.

The choice of the sample sizes N_u, N_c and n_c must ensure a sufficient coverage of the input factor space and the convergence and robustness of the results of the KS-test. To set the value of the sample size for conditional (N_c) and unconditional (N_u) outputs, we assessed the convergence of the results of the KS-test. The results and methodology are presented in detail in Section D of our Supplementary Material. From our analysis, the KS-test appears to be very sensitive in that it can detect small deviations between two CDFs. We choose a significance level α_c equal to 0.001 (minimum value of the significance level given in the tables see for instance Wall, 1996). In this way, input factors with very small but non-zero sensitivity are more likely to be detected as insensitive by the KS-test. For the purpose of screening, we believe it is appropriate to identify not only the input factors that are completely insensitive, but also the input factors that have a very small influence on the output, otherwise the screening would be too strict.

We note that a different summary statistic could be chosen instead of the maximum, for instance the mean or the median. The same level of confidence of the validation test α_t can be obtained using any of these summary statistics, provided that an appropriate value of the significance level α_c is chosen to compute the critical KS-statistic. Given the value of α_t, the value of α_c is lower for the maximum than for the median or the mean.

2.3. Workflow for the experiments

Fig. 3 presents the workflow of the analysis that we conduct to investigate and compare the convergence of several GSA methods. First, for each GSA method, we build a dataset of N input/output samples by MC random sampling and model evaluation. We estimate the sensitivity indices and their bootstrapping confidence intervals by resampling with replacement (Fig. 3a). We can then compute the three statistics of Eqs. (1), (3) and (5) and verify whether convergence of the indices, ranking and screening has been reached according to the criteria introduced in Section 2.1. Computations can be repeated using sub-samples of reduced size. A visual summary of the values of sensitivity indices and their uncertainty at different sample size is given by the convergence plot as Fig. 3b. At the sample size when screening convergence is reached, we also apply the validation procedure defined in Section 2.2. For a given value of the assumed screening threshold, we obtain the set of insensitive input factors, compute the KS-statistic (Fig. 3c) and apply the KS-test. We repeat the test for increasing values of the assumed screening threshold and obtain Fig. 3d.

3. GSA methods and experiments

3.1. GSA methods

Here we briefly introduce the three GSA methods investigated in this study. In the following sections, we denote the number of input factors subject to GSA as M, the number of model evaluations performed during GSA as N, and the base sample size chosen by the GSA user as n.

3.1.1. Variance-Based SA (VBSA)

Variance-Based SA (VBSA) is based on the variance decomposition first proposed by Sobol' (1990). Following common practice in GSA applications (Saltelli et al., 2008), we use two sensitivity indices for each input factor: the Main effect (VBM) index S_i^{VBM} and the Total effect (VBT) index S_i^{VBT}, which includes the main effect and interactions. The two sensitivity indices are expressed as follows (Saltelli et al., 2008):

$$S_i^{VBM} = \frac{V_{Xi}(E_{X_{\sim i}}(y|X_i))}{V(y)} \tag{9}$$

$$S_i^{VBT} = \frac{E_{X_{\sim i}}(V_{Xi}(y|X_{\sim i}))}{V(y)} \tag{10}$$

where X_i is the i-th input factor, $X_{\sim i}$ denotes the vector of all input factors but the i-th one, E is the expected value and V is the variance. Both sensitivity indices can be used for ranking the input factors depending on whether the GSA user is interested in main effect only or in main effect and interactions. However, only the total effect index is suitable for screening because it accounts for input factor interactions as well as individual impact on its own.

Here, the indices S_i^{VBM} and S_i^{VBT} are estimated according to the method proposed by Saltelli (2002). First, two independent input samples X_A and X_B are built (each being a matrix of dimension (n,M)). Then, a matrix X_C of dimension ($n.M$,M) is generated by recombination of the samples in X_A and X_B: X_C is composed of M blocks X_{Ci} ($i = 1,\ldots,M$), each being a (n,M) matrix whose columns are all taken from X_B exception made for the i-th column, which is taken from X_A. We denote the three corresponding sets of model outputs as y_A, y_B and y_C. Then, S_i^{VBM} and S_i^{VBT} are computed as follows:

$$S_i^{VBM} = \frac{\frac{1}{n}\sum_{j=1}^{n} y_A^{(j)} y_{Ci}^{(j)} - \left(\frac{1}{n}\sum_{j=1}^{n} y_A^{(j)}\right)^2}{\frac{1}{n}\sum_{j=1}^{n} \left(y_A^{(j)}\right)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} y_A^{(j)}\right)^2} \tag{11}$$

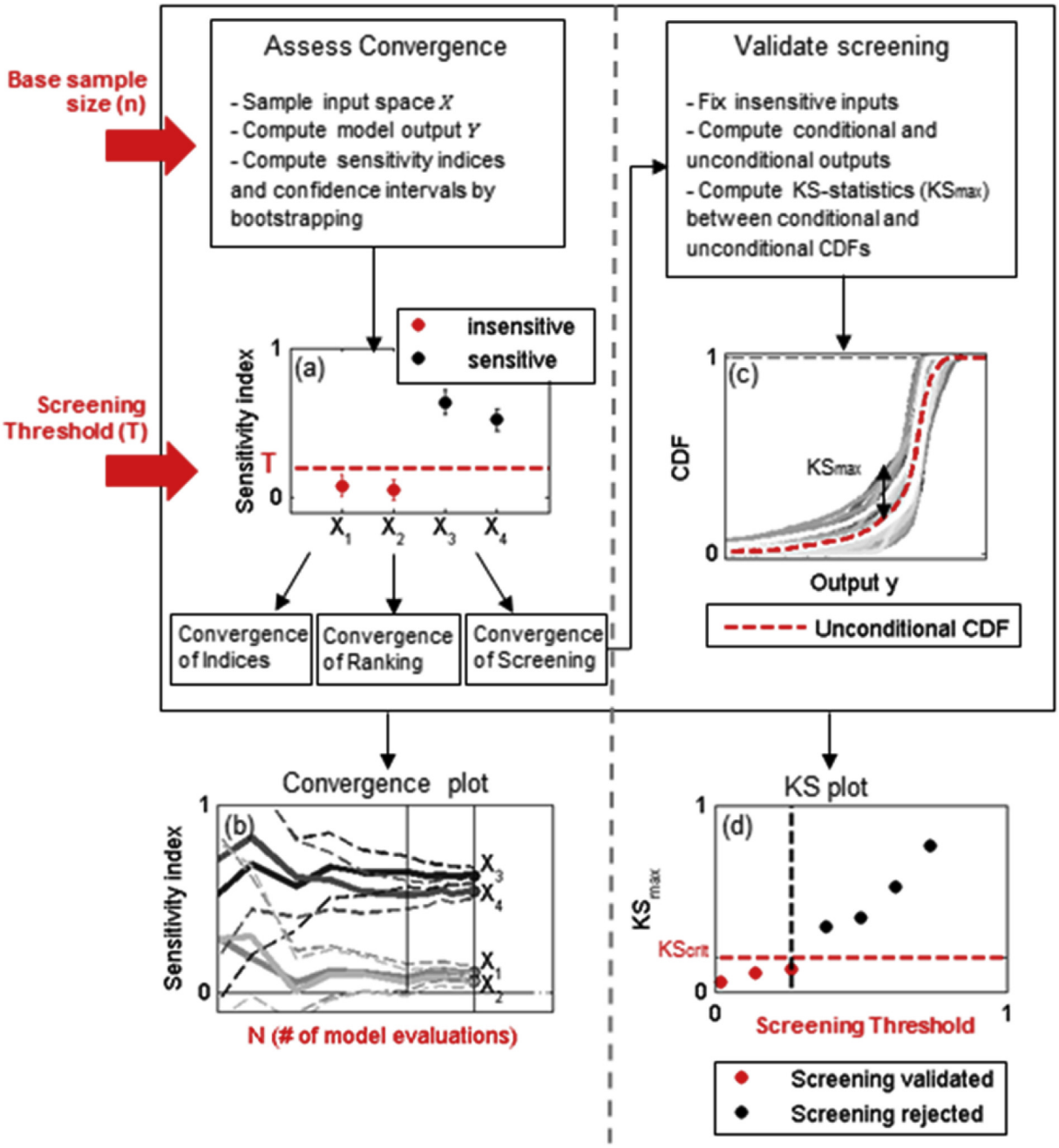

Fig. 3. Workflow for the convergence analysis of GSA and for the screening validation. We investigate the influence of the choice of base sample size n and screening threshold T. We keep the other characteristics of the experimental setup unchanged. We repeat the experiments for different GSA methods. We note that the number of model evaluations N is a function of the base sample size n and the number of input factors M.

$$S_i^{VBT} = \frac{\frac{1}{n}\sum_{j=1}^{n} y_B^{(j)} y_{Ci}^{(j)} - \left(\frac{1}{n}\sum_{j=1}^{n} y_A^{(j)}\right)^2}{\frac{1}{n}\sum_{j=1}^{n} \left(y_A^{(j)}\right)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} y_A^{(j)}\right)^2} \tag{12}$$

The total number of model evaluations for the approximation of both the main and total effects indices is therefore equal to:

$$N = n \cdot (M + 2) \tag{13}$$

In order to maximize the coverage of the input factor space, for each given sample size, we use maximin (maximisation of the minimum inter-point Euclidean distance between any two sampled points) Latin Hypercube Sampling (LHS) to build the matrices X_A and X_B. VBSA is deemed to provide reliable results for screening and ranking purposes (Saltelli et al., 2008). It is often considered as a benchmark to assess the credibility of other GSA methods (see for instance Yang 2011). The sensitivity indices are expressed in terms of percentages of the output variance and always take values between 0 and 1.

3.1.2. Elementary Effect Test (EET)

The Elementary Effect Test (EET, Saltelli et al., 2008) or Method of Morris (Morris, 1991) is a much less computationally expensive method than VBSA and therefore most suitable when dealing with time-consuming models (Saltelli et al., 2008). Campolongo et al. (2007) demonstrated empirically that the sensitivity measure produced by EET could be used as a proxy of the total effect index produced by VBSA, and therefore EET is particularly suitable for screening.

EET is a global extension of One-factor-At-the-Time (OAT) Local SA methods. It is based on the computation of several Elementary Effects (EEs). Specifically, the *EE* of the *i*-th input factor x_i at given baseline point X^j and for a predefined perturbation Δ is given by:

$$EE_i^j = \frac{y\left(x_1^j, x_2^j, \ldots, x_{i-1}^j, x_i^j + \Delta, \ldots x_M^j\right) - y\left(x_1^j, x_2^j, \ldots, x_{i-1}^j, x_i^j, \ldots x_M^j\right)}{\Delta} \tag{14}$$

For each input factor, EEs are computed at n randomly chosen baseline points across the input factor space. The estimated mean (μ_i) of the EEs is taken as a measure of the total effects of the *i*-th input factor. The standard deviation (σ_i) of the EEs can be interpreted as the intensity of the interactions of the *i*-th input factor with other input factors. In order to avoid compensations between EEs of opposite sign, we use the mean of the absolute values of the EEs (μ_i^*) in this study, as first suggested by Campolongo et al. (2007), i.e.:

$$\mu_i^* = \frac{1}{n}\sum_{j=1}^{n}\left|EE_i^j\right| \tag{15}$$

The sensitivity index of Eq. (15) provides a semi-quantitative measure of sensitivity, particularly suitable to rank the factors on an interval scale (Saltelli et al., 2008). To define baseline points and the perturbation Δ, we implement the radial design strategy proposed by Campolongo et al. (2011) since it was shown that radial based design is computationally efficient. In this approach, n baseline points are sampled across the input factor space, and associated with other n auxiliary points, also randomly chosen. Then, the perturbation Δ is computed as the difference between the *i*-th coordinate of the auxiliary and baseline point. Here, the baseline and auxiliary points were generated by maximin Latin Hypercube sampling so to maximize the coverage of the input factor space. The total number of model evaluations required to compute the mean EEs for all input factors is equal to:

$$N = n\cdot(M+1) \tag{16}$$

We note that the value of μ_i^* has no specific meaning per se, as it depends on the scale and units of measurements of the model output *y*. Therefore, to allow for comparison between different case studies, we define a normalized mean of the EEs as our sensitivity index, i.e. the ratio between μ_i^* and the maximum value of the mean EEs across all the input factors:

$$S_i^{EET} = \frac{\mu_i^*}{\max_k \mu_k^*} \tag{17}$$

The sensitivity index of Eq. (17) now takes values between 0 and 1 regardless of the units of measurement of *y*, and it expresses input factor sensitivity as a fraction of the sensitivity for the most influential input factor. The index still provides a semi-quantitative measure of sensitivity.

3.1.3. Regional Sensitivity Analysis (RSA)

Regional Sensitivity Analysis (RSA, Young et al., 1978; Spear and Hornberger, 1980) is a GSA method, which is widely used because of its ease of implementation and because it allows for Factor Mapping (see for instance Freer et al., 1996; Wagener et al., 2001; Sieber and Uhlenbrook, 2005). Though it is of limited use for screening, since it does not detect interactions between input factors (for instance, factors combined as products or quotients may compensate, see p.190 in Saltelli et al. (2008)) and therefore a zero-value sensitivity index produced by RSA is a necessary, but not sufficient condition for an input factor to be non-influential.

The method first decomposes the set of input factor samples into two groups, depending on whether their associated output exceeds a prescribed threshold value (e.g. a certain level of performance). The two marginal CDFs $F_i^B(x_i)$ and $F_i^{\overline{B}}(x_i)$ for the two groups (B and $\overline{B}$), i.e. behavioural (acceptable) and non-behavioural (poor) model predictions, are then derived and compared.

In the present study, we quantify the discrepancy between the behavioural and the non-behavioural CDFs by means of the Kolmogorov–Smirnov statistic. The sensitivity index for the *i*-th input factor x_i is then expressed as follows:

$$S_i^{RSA} = \max_{x_i}\left|F_i^B(x_i) - F_i^{\overline{B}}(x_i)\right| \tag{18}$$

The sensitivity index of Eq. (18) varies between 0 and 1 and is semi-quantitative.

3.2. Models and data

Three hydrological conceptual-type models of varying complexity are investigated in this study (HyMod, HBV and SWAT) and are applied to three different catchments. In our application, the input factors are the model parameters. Fig. 4 presents the available datasets for the three case studies and Section A of our Supplementary Material provides a brief description of the model parameters.

3.2.1. HyMod model (5 parameters)

The HyMod model was first introduced by Boyle (2001) and is described in Wagener et al. (2001). It has been widely applied because of its simplicity (5 parameters) (e.g Vrugt et al., 2002; Kollat et al., 2012; Gharari et al., 2013). The HyMod model produces a time series of stream flow predictions and is forced by precipitation and potential evapotranspiration. It is composed of a soil moisture routine (parameters BETA and SM) and a routing module (parameters ALPHA, RS and RF). The latter module consists of two sets of parallel linear reservoirs, namely three linear reservoirs for the fast runoff component and a single linear reservoir for the slow runoff component. In this study, the model is evaluated with daily time step data over a simulation horizon of ten years, starting on 01/11/1948, including a one-year warm-up period. The application study site is the Leaf River catchment, a 1950 km^2 catchment located north of Collins, Mississippi, USA. Sorooshian et al. (1983) provide a detailed description of the Leaf River catchment.

3.2.2. HBV model (13 parameters)

The HBV model was introduced by Bergström (1995) and is described in Seibert (1997). Although developed initially for applications in Scandinavia, the HBV model was used in many studies around the world (e.g. Grillakis et al., 2010; Kollat et al., 2012). The model produces a time series of stream flow predictions and is driven by precipitation, mean temperature, and potential evapotranspiration. We implement a version with 13 parameters. The model includes a snow module (parameters TS, CFMAX, CFR, CWH),

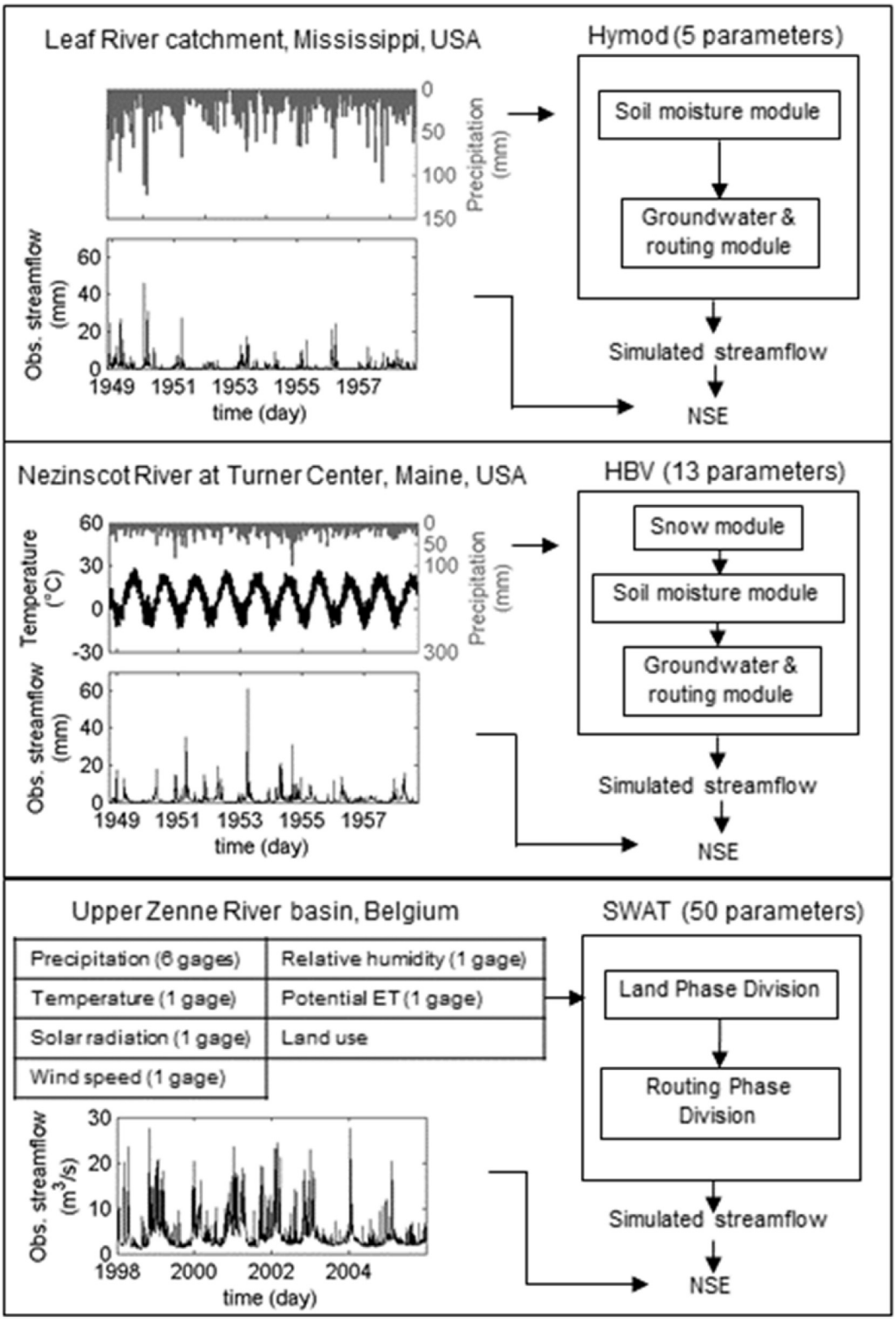

Fig. 4. Data, routine and model output for the three case studies analysed.

a soil moisture module (parameters β, LP, FC) and a groundwater and routing module (parameters PERC, K0, K1, K2, UZL, MAXBAS). The latter is composed of an upper reservoir for the fast response, a lower reservoir for the slow response and a channel routing component with a triangular weighting function. Here, HBV is evaluated with daily time step data over a simulation horizon of ten years, starting on 01/10/1948, including a one-year warm-up period. We use hydrologic years to remove the carry-over impacts of snow storage. The application study site is the Nezinscot River at Turner Center, Maine, USA (USGS 01055500), a catchment of

438 km^2 (Duan et al., 2006).

3.2.3. SWAT model (50 parameters)

The Soil and Water Assessment Tool, SWAT (Arnold et al., 1993, 1998), is a semi-distributed hydrological model developed by the USDA Agricultural Research Service. The model is used worldwide to study the impact of catchment management on water availability (e.g. Tram et al., 2014), sediments (e.g. Ali et al., 2014), nutrients (e.g. Bouraoui and Grizzetti, 2008) and agricultural yields (e.g. Bannwarth et al., 2014), and the impact of land use (e.g. Vaché et al., 2002) and climate change (e.g. Bae et al., 2011). It is a complex model with more than 100 parameters (though not all are typically calibrated) that includes the major catchment processes. The simulation of the hydrology is separated in two divisions in the SWAT model. The first division is the land phase of the hydrologic cycle, which controls the amount of water, sediment, nutrient and pesticide loadings to the main channel in each sub-basin. It includes multiple modules: climate of the watershed (weather generator, soil temperature), snow pack, canopy interception, surface runoff, soil moisture, groundwater, surface storage (ponds), tributary channels, plant growth and erosion, sediment, nutrient and pesticide yield. The second division is the water or routing phase of the hydrologic cycle, which is composed of four components: water, sediments, nutrients and organic chemicals. It includes the routing in the main channel or reach to the outlet and the routing in the reservoirs. A catchment is partitioned into multiple sub-basins, which are then divided into Hydrological Response Units (HRUs). Each HRU has unique land cover, soil characteristics, and management combination and therefore requires specific values for its parameters. The flow at the outlet of the basin is evaluated with daily time-step data over a simulation horizon of eight years, starting on 01/01/1998, including a three-year-warm-up period. The application study site is the upper Zenne River basin, Belgium, a 642 km^2 catchment (Leta et al., 2015). We use a SWAT model version that includes 21 sub-basins and 155 HRUs. We study the sensitivity of 26 flow parameters typically considered for GSA (see for instance Nossent and Bauwens, 2012). In order to add more parameters to GSA, we analyse the sensitivity of 6 of these 26 parameters separately for the five land use types present in the basin - Agriculture (A), Urban (U), Forest (F), Pasture (P) and Range Brush (R). We therefore consider 50 parameters for SA. It is worth noting that these 6 parameters defined at the land use scale are controlling the properties of a part of the catchment only. Therefore, they are likely to be less sensitive than the corresponding parameters defined at the catchment scale. In the present study, we use the 2009 version of the SWAT model (Neitsch et al., 2009).

The ranges of the parameters are taken from Wagener et al. (2001) for HyMod, from Kollat et al. (2012) for HBV. and from personal communication for SWAT. An initial analysis was conducted to refine those ranges for the particular application sites (see Section B of our Supplementary Material) since the chosen ranges influence the results of GSA (Wang et al., 2013; Kelleher et al., 2013). Section A of our Supplementary Material reports the ranges that are used for the analysis.

3.3. Output definition

We compute the Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970) as scalar output for sensitivity analysis:

$$NSE = 1 - \frac{\sum_{t=1}^{H}\left(y_{o,t} - y_{s,t}\right)^2}{\sum_{t=1}^{H}\left(y_{o,t} - \overline{y_o}\right)^2} \tag{19}$$

where H is the number of time steps, $y_{o,t}$ is the observed value of the stream flow at time t, $\overline{y_o}$ is the average value of the observations and $y_{s,t}$ is the simulated value of the stream flow at time t. NSE is widely used in hydrology. The better the simulations match the observations, the more *NSE* tends towards a value of 1. Instead, values of *NSE* below 0 indicate that the average of the observations provides a better estimate of the observed stream flow than the model simulations. It is worth noting that the *NSE* has a tendency to focus on fitting high flows due to the use of squared residuals. Different results for the GSA could be obtained with different model outputs, such as a different performance measure (e.g. bias or absolute mean error, which focuses on the water balance), a prediction function (e.g. mean annual stream flow) or an output related to a state variable (e.g. soil moisture). For a rigorous diagnostic of the parameter sensitivity, different model outputs and environmental conditions should be taken into account (Van Werkhoven et al., 2008; Vanuytrecht et al., 2014). This is, however, beyond the scope of the present study, which only aims to provide indications on how to choose the sample size and validate screening in GSA through particular case studies.

3.4. Experimental set-up

The computational experiments were performed with the SAFE toolbox (Pianosi et al., 2015; available at http://www.bris.ac.uk/cabot/resources/safe-toolbox/). SAFE implements the three GSA methods tested (EET, RSA, VBSA), the bootstrap technique, tools for the convergence analysis and the HyMod and the HBV models as test examples. Table 1 summarizes the experimental set-up adopted for the analyses.

First, we generated a Latin Hypercube Sample of maximum size and we derived smaller samples by randomly taking sub-samples of the original LHS. The advantage of this approach is that it limits the number of model evaluations. However, the sub-samples are dependent and therefore, the evolution of the sensitivity indices for increasing sample size is smoother than it would be when using independent samples. The computation of the bootstrap estimates partially overcomes this issue, since the bootstrap technique approximates the sample distribution. For a sub-sample of a given size N, the bootstrap estimates were obtained by resampling with replacement within this sub-sample. Moreover, the stratified structure of the Latin Hypercube is not maintained in the sub-samples. For RSA, we reduced the sample size by dropping parameterizations using the maximin criterion (i.e. we randomly generate ten subsamples and we take the one with the maximal value of the minimum Euclidian inter-point distance) in order to cover the search space as evenly as possible. For EET and VBSA instead, due to the particular structure of the samples, the parameterizations in the initial LHS were dropped randomly without using the maximin criterion. We note that strategies exist to avoid this loss of stratification (see for instance Andres (1997) for further details).

4. Results

In this section we present the results of the convergence study and of the screening threshold investigation. Fig. 5 shows the evolution of the sensitivity indices for increasing numbers of model evaluations for the three GSA methods and the three case studies. It is worth noting that an apparent convergence of the bootstrap mean of the sensitivity index (flattening of the line) is a necessary but not sufficient condition for convergence. It can happen that the bootstrap mean takes similar values in two different samples 'by chance' while the actual statistical convergence is not reached yet (the confidence intervals are still very wide). Therefore, it is important to also include information about the confidence

Table 1
Experimental set-up for the analyses.

Experimental set-up for sampling of the parameter space		
Parameter distribution	Uniform distribution (no a priori knowledge)	
Sampling strategy	Maximin Latin Hypercube (uniform 1D margins and maximisation of the minimum inter-point Euclidean distance)	
Experimental set-up for convergence analysis		
Threshold value for RSA	0.4	
Number of bootstrap replicates (N_{boot})	1000	
Maximum number of model evaluations for HyMod (N_{max})	EET	78,000
	RSA	10,000
	VB	420,000
Maximum number of model evaluations for HBV (N_{max})	EET	560,000
	RSA	20,000
	VB	600,000
Maximum number of model evaluations for SWAT (N_{max})	EET	102,000
	RSA	30,000
	VB	520,000
Experimental set-up for validation of screening		
Sample size for unconditional output (N_u)	2000	
Sample size for conditional output (N_c)	1600	
Number of conditioning values (n_c)	20	

intervals. Fig. 6 shows the evolution of the convergence statistics defined in Section 2.1. In this Figure, vertical lines indicate the number of model evaluations, N, suggested in Saltelli et al. (2008, Table 6.9) for the three GSA methods. These values correspond to a base sample size n equal to 10 for EET, 2000 for RSA and 1000 for VBSA. Table 2 reports the values of the sample sizes that ensure convergence of the indices ($Stat_{indices} < 0.05$), of ranking ($Stat_{ranking} < 1$) and of screening ($Stat_{screening} < 0.05$) when using our suggested convergence statistics. Note that for VBSA, the number of model evaluations necessary to reach convergence refers to the joint estimation of the two indices (VBM and VBT), because these two indices are obtained from the same sample of model evaluations.

4.1. Convergence of sensitivity indices

The top panels in Fig. 6(a–c) show the values of the convergence statistic for the value of the sensitivity indices for increasing sample size. They show that the sensitivity indices converge first for RSA compared to the other methods (see also Table 2). RSA requires at most 15,000 model evaluations for the three case studies considered. EET and VBSA require a much larger number of model evaluations, generally of the order of magnitude of several hundreds of thousands, which is prohibitive when simulations are computationally expensive. In particular for the SWAT model (Fig. 6c), sensitivity indices have not reached convergence even after 102,000 model evaluations for EET and 520,000 model evaluations for VBSA for both Main effect (VBM) and Total effect (VBT). Moreover, we observe that with the typical values suggested in the literature (vertical dashed lines in Fig. 6), the width of the confidence intervals of the sensitivity indices are quite wide, especially for EET. However, for RSA, $Stat_{indices}$ is already quite close to its threshold value after 2000 model evaluations since the width of the confidence intervals is equal to 0.09 for the HyMod and the SWAT models.

4.2. Convergence of parameter ranking

The middle panels in Fig. 6(d–f) show the value of the convergence statistic for ranking for increasing sample size, for the three models. Fig. 5 reports the ordering of the most sensitive parameters.

We first observe from Fig. 5 that the three GSA methods provide different rankings of importance for the model parameters. This is reasonable since the three methods measure sensitivity according to different rationales and assumptions. The rankings given by EET and VBT indices are generally quite consistent with each other. In particular, the two methods identify the same group of most sensitive parameters (ALPHA and RF for HyMod in Fig 5a,d; FC and TS for HBV in Fig. 5e,h; CN2_A, CH–K2 and CH–N2 for SWAT in Fig. 5i,l) and of least sensitive parameters. For HBV and SWAT, the ranking given by RSA (Fig. 5g,k) differs from the one produced by EET and VBSA, which might be explained by the fact that RSA does not detect many types of interactions (see p.190 in Saltelli et al. (2008)).

We find that the ranking generally converges faster than the estimates of the sensitivity indices when comparing the middle panels of Fig. 6(d–f) with the top panels (a–c) and the corresponding values of the sample sizes reported in Table 2. However, the results are different for RSA applied to SWAT. The two more sensitive parameters (CN2_A and SLOPE_A) clearly separate out while the other parameters have very similar values of the sensitivity indices. Since this happens also for parameters that have a relatively high sensitivity index, minor fluctuations in these indices values can lead to large differences in ranking.

When comparing the convergence of the ranking across the three case studies, we observe that the number of model evaluations N required for convergence usually increases with the number of parameters, M, as expected. Interestingly, this does not seem to be the case for EET. We indeed observe that for the HBV model (Fig. 6e), 7000 model evaluations are necessary for the convergence of the ranking provided by EET while for the SWAT model (Fig. 6f) only 4590 model evaluations are necessary. The rate of convergence for the ranking appears to depend on the specific case study and on the relative value of the sensitivity indices among the different parameters. For EET applied to the SWAT model (Fig. 5i), the sensitivity indices of the three most influential parameters are significantly higher than all the others, while for the HBV model (Fig. 5e) they are more evenly spread. As a result, the ranking of the most influential parameters stabilizes faster for SWAT than for HBV.

Analysing the rate of convergence across the three GSA methods, we observe that the convergence of the ranking is reached quickest for the RSA method compared to the other methods for the HyMod and the HBV models (Fig. 6, Table 2). EET appears to require fewer model evaluations than VBSA while providing a ranking consistent with VBT. We also note that the ranking obtained with the number of model evaluations suggested in the literature (vertical dashed lines in Fig. 6) is generally not robust for the two more complex models (HBV and SWAT) since $Stat_{ranking}$ takes values above 1 at these sample sizes. In particular for EET, with a base sample size of $n = 10$ ($N = 140$ for HBV and $N = 520$ for SWAT), $Stat_{ranking}$ is higher than 4 for these two models. However, $Stat_{ranking}$ underestimates the rate of convergence for the VBSA method applied to the SWAT model (Fig. 6f). We indeed note that the curve for VBM has large oscillations and only converges for a very high number of model evaluations. This is due to rank reversals happening between some low-sensitivity parameters, while the ranking of the most sensitive parameters stabilizes already after a much lower number of model evaluations. This shows that rank reversals for low-

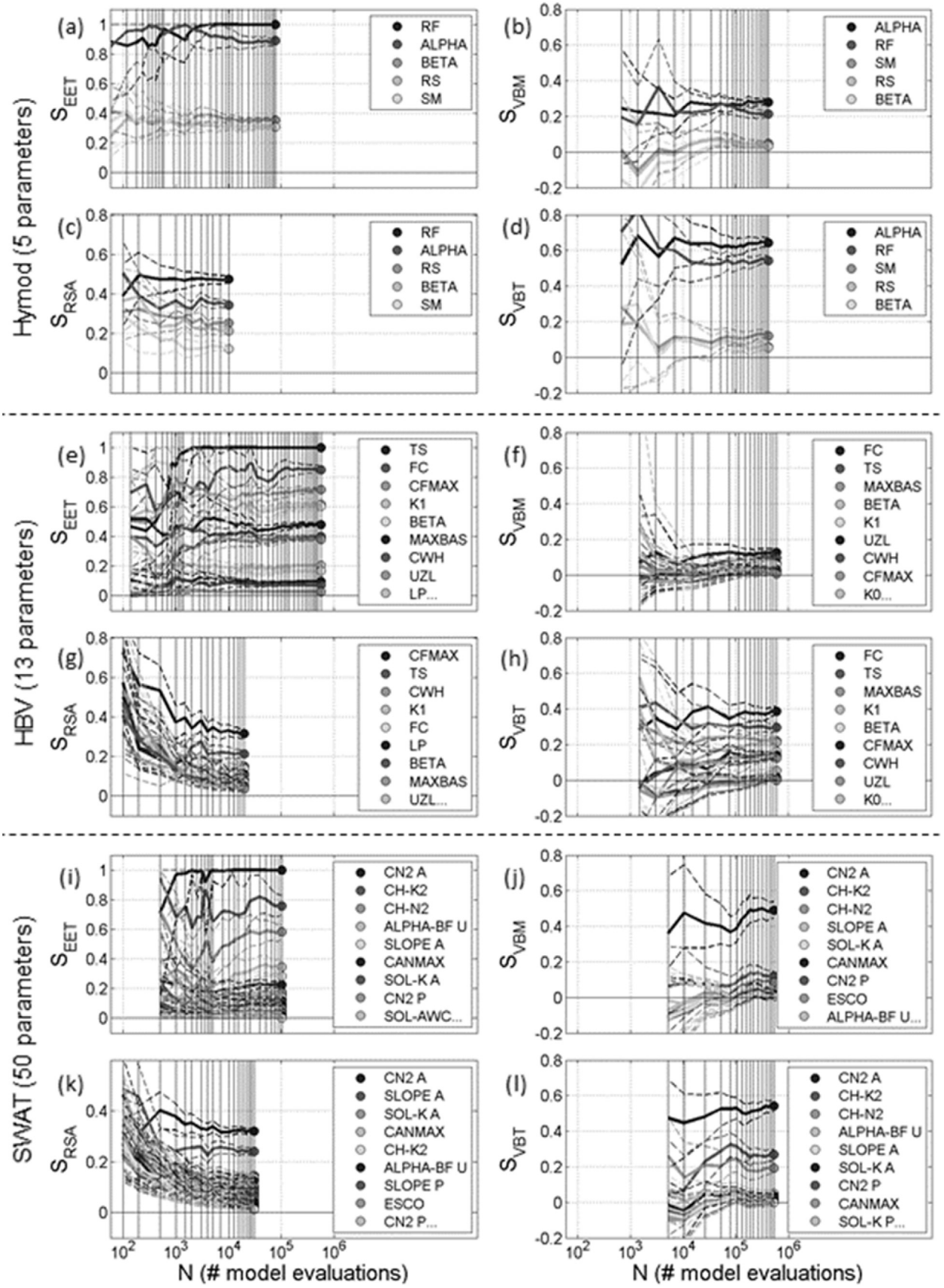

Fig. 5. Convergence plots: the figure represents the sensitivity indices of the model parameters for HyMod, HBV and SWAT, estimated using an increasing number of model evaluations *N*, computed for the different GSA methods, Elementary Effect Test (EET), Regional Sensitivity Analysis (RSA), Variance-based methods Main effect (VBM) and Total effect (VBT). The solid lines are the bootstrap means of the sensitivity indices and the dashed lines are the 95% bootstrap confidence intervals.

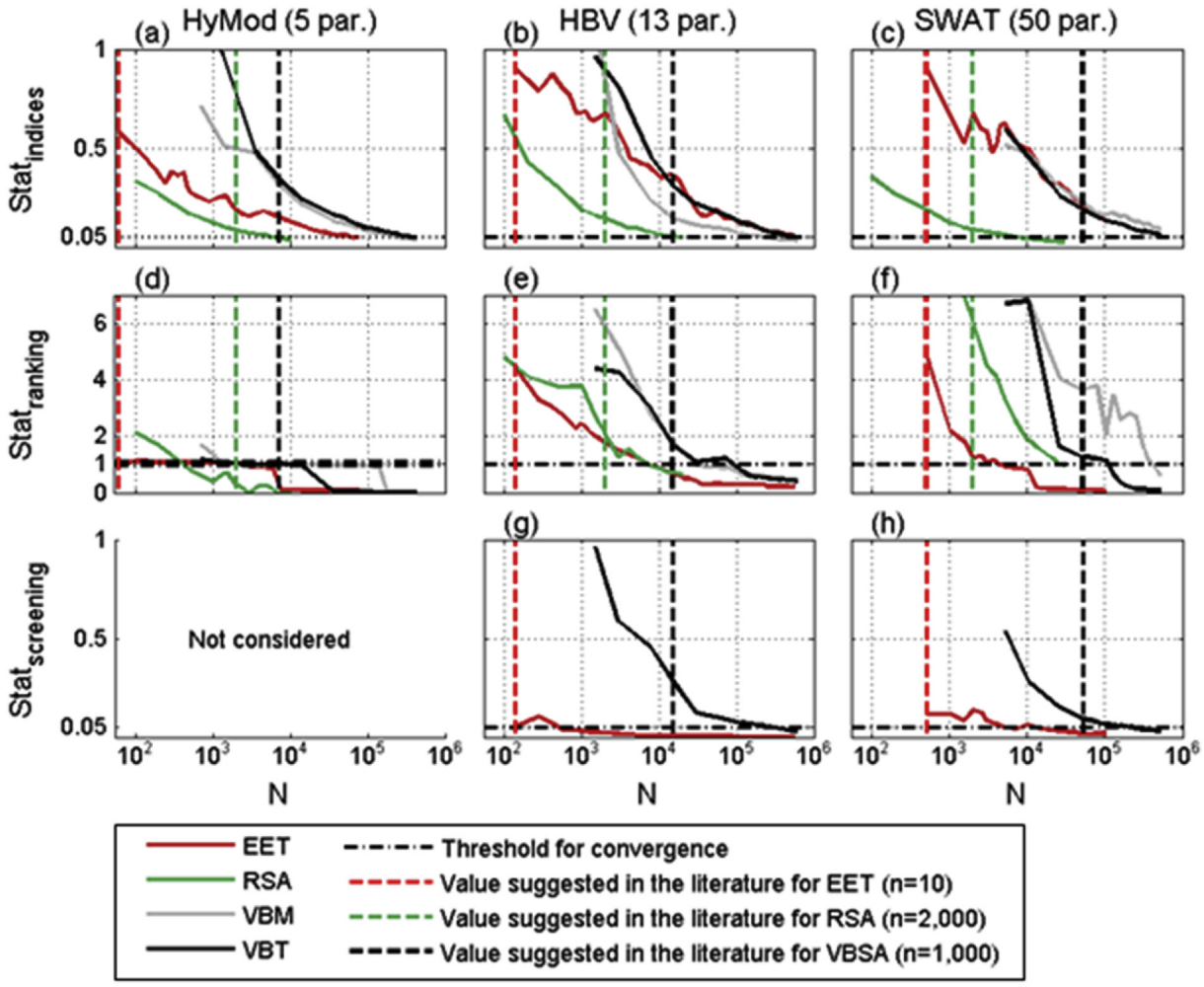

Fig. 6. Value of the convergence statistics against number of model evaluations N. $Stat_{indices}$ (convergence of sensitivity indices) is the maximum width of the bootstrap confidence intervals, $Stat_{ranking}$ (convergence of ranking) is an adjusted rank correlation coefficient, and $Stat_{screening}$ (convergence of screening) is the maximum width of the bootstrap confidence intervals for the lower-sensitivity parameters. The results are reported for the three case studies. The dashed vertical lines report the number of model evaluations suggested in the literature (Saltelli et al., 2008, Table 6.9) and the dashed horizontal black lines represent the threshold values for the convergence statistics, below which we assume that convergence is reached.

Table 2
Number of model evaluations N (and corresponding base sample size n) necessary to reach convergence of sensitivity indices ($Stat_{indices} < 0.05$), ranking ($Stat_{ranking} < 1$) and screening ($Stat_{screening} < 0.05$) based on analysis. M is the number of parameters. RSA and VBM are not used for screening. We do not screen the HyMod parameters (they are all influential).

GSA method	Objective	Number of model evaluations for convergence		
		HyMod ($M = 5$)	HBV ($M = 13$)	SWAT ($M = 50$)
EET	Indices	$N = 60{,}000$ ($n = 10{,}000$)	$N = 560{,}000$ ($n = 40{,}000$)	$N > 102{,}000$ ($n > 2000$)
$N = n \cdot (M + 1)$	Ranking	$N = 3000$ ($n = 500$)	$N = 7000$ ($n = 500$)	$N = 4590$ ($n = 90$)
	Screening		$N = 560$ ($n = 40$)	$N = 5100$ ($n = 100$)
RSA	Indices	$N = 7000$ ($N/M = 1400$)	$N = 15{,}000$ ($N/M = 1154$)	$N = 7000$ ($N/M = 140$)
$N = n$	Ranking	$N = 500$ ($N/M = 100$)	$N = 7000$ ($N/M = 538$)	$N = 25{,}000$ ($N/M = 500$)
	Screening			
VBM	Indices	$N = 210{,}000$ ($n = 30{,}000$)	$N = 225{,}000$ ($n = 15{,}000$)	$N > 520{,}000$ ($n > 10{,}000$)
$N = n \cdot (M + 2)$	Ranking	$N = 1400$ ($n = 200$)	$N = 30{,}000$ ($n = 2000$)	$N = 416{,}000$ ($n = 8000$)
	Screening			
VBT	Indices	$N = 350{,}000$ ($n = 50{,}000$)	$N = 450{,}000$ ($n = 30{,}000$)	$N > 520{,}000$ ($n > 10{,}000$)
$N = n \cdot (M + 2)$	Ranking	$N = 3500$ ($n = 500$)	$N = 112{,}500$ ($n = 7500$)	$N = 130{,}000$ ($n = 2500$)
	Screening		$N = 262{,}500$ ($n = 17{,}500$)	$N = 208{,}000$ ($n = 4000$)

sensitivity parameters can still have some influence on our proposed indicator $Stat_{ranking}$ (see Section C of our Supplementary Material for further analysis).

4.3. Convergence of parameter screening

The bottom panels in Fig. 6(g,h) show the value of the convergence statistic for screening for increasing sample size. Fig. 7

reports the results of the validation test.

4.3.1. Convergence of sensitivity indices for low-sensitivity parameters

For the purpose of screening, we consider only EET and VBT while we exclude RSA and VBM because they do not account for interactions. Furthermore, screening is not applied to the HyMod model since all its five parameters are found to be influential in our experimental set-up. We compute the convergence statistic for screening for the other two case studies (HBV and SWAT) (Fig. 6g,h).

Comparing the bottom panels of Fig. 6(g,h) with the top panels (b,c), we observe that the convergence of the sensitivity indices for the lower-influence parameters ($Stat_{screening}$) is quicker compared to the other parameters ($Stat_{indices}$). In particular, for EET, after 560 model evaluations ($n = 40$) for HBV and 5100 model evaluations ($n = 100$) for SWAT, the lower-sensitivity parameters have already converged while it takes hundreds of thousands of model evaluations for all the sensitivity indices to converge (see also Table 2). The convergence of the indices for the lower-sensitivity parameters requires still more model evaluations than usually suggested in the literature.

Analysing the results obtained with EET across the two case studies, we observe that the convergence of the screening for HBV is reached for a smaller number of model evaluations than for SWAT, which is expected because HBV has a lower number of parameters. However, for VBT, the screening converges slightly earlier for the SWAT model than the HBV model. Therefore, as observed for the convergence of the ranking, the number of model evaluations required to stabilize the sensitivity indices of lower-sensitivity parameters does not necessarily increase with the number of parameters considered.

Moreover, we notice some oscillations in the value of $Stat_{screening}$ for EET when the number of model evaluations is small (Fig. 6g,h). For small sample sizes, the bootstrap technique is not able to assess the 'true' variability of the sensitivity indices because the small samples may not contain enough information. We indeed note that the width of the bootstrap confidence interval for some low-influence parameters increases significantly with the sample size early on, before decreasing and reaching convergence when further increasing the sample size.

4.3.2. Validation of screening

We validated the screening and investigated the value of the screening threshold at the sample sizes for which convergence is reached (reported in Table 2). Fig. 7 shows the estimated KS_{max} for increasing values of the assumed screening threshold for the two models and the two GSA methods. The Figure also shows the critical values of the KS-statistics at different significance levels. As explained in Section 2.2, we used a small value of the significance level (0.001) when applying the KS-test, so that the screening is not too conservative.

The screening results of EET and VBT are consistent. For the HBV model (Fig. 7a,c), one insensitive parameter is identified (K2) by both EET and VBT. For the SWAT model (Fig. 7b,d), 27 insensitive parameters are identified by EET and only 21 of those 27 insensitive parameters are identified with VBT. Therefore, for SWAT, EET identifies a higher number of non-influential parameters for a much smaller number of model evaluations than VBT. The reason is that for VBT, the 95% confidence intervals of the sensitivity index for the lower-sensitivity parameters are strongly overlapping while for EET we observe much less overlap. As a result, for SWAT, EET is able to differentiate the sensitivities among the lower-sensitivity

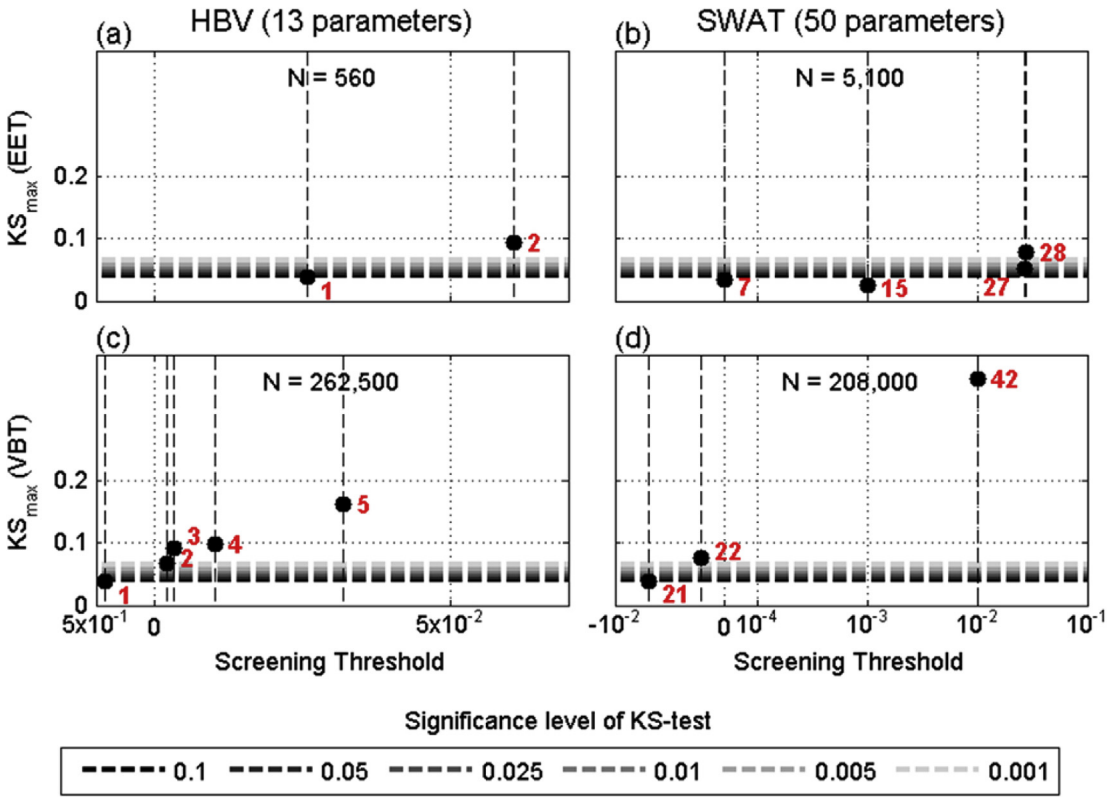

Fig. 7. Validation of screening results: KS_{max} against assumed Screening Threshold on the sensitivity indices, computed by EET and VBT methods for the HBV and SWAT models. N is the number of model evaluations used to compute the sensitivity indices (this was chosen to ensure screening convergence, see also Table 2). The critical values of the KS-statistic are reported for different significance levels. Next to each data point is the number of parameters that are fixed to compute the KS-statistic.

parameters, allowing for a better separation between parameters that have a negligible impact on the output and parameters that have a low but non-negligible influence.

For EET (Fig. 7a,b), the screening is validated for screening threshold values of 2.6×10^{-2} and 2.7×10^{-2} for HBV and SWAT respectively. For the VBT method (Fig. 7c,d), it is validated for a screening threshold value of -8.4×10^{-3} and -4.0×10^{-3} for HBV and SWAT respectively. We note that the value of the screening threshold is slightly negative for VBT because the value of the sensitivity indices for the lower-sensitivity parameters has still not perfectly converged (the width of the confidence interval is equal to 0.05 and not 0). The values of the screening thresholds for VBT are lower than the value of 0.01 generally used in the literature (see for instance Tang et al., 2007; Sin et al., 2011; Cosenza et al., 2013; Vanrolleghem et al., 2015). For the SWAT model (Fig.7d), using a threshold of 0.01 would lead to a KS-statistic well above the critical values. On the other hand, the subset of parameters with a sensitivity index below a value of 0.01 seems to remain stable when the sample size is increased, since it does not change up to the maximum sample size analysed ($N = 520{,}000$) (result not shown). Therefore, a screening threshold of 0.01 as used before appears to be too high to screen the parameters of the SWAT model because it would lead to a larger number of parameters being deemed insensitive.

5. Discussion

5.1. Implications of the results for GSA implementation

The results of the present study can be used to derive guidelines for the implementation of GSA, while of course considering that we used specific case studies for testing, as well as a specific experimental set-up, e.g. the choice of parameter ranges, of the output function, of the simulation period, of the sampling strategy. Shin et al. (2013) for example demonstrate the influence of some of these choices on GSA results.

Firstly, we observe that different base sample size values are needed for different models. Interestingly, from our results, no clear relationship emerges between the number of parameters M and the number of model evaluations N needed to reach convergence (Table 2). The values of the base sample size n for EET and VBSA and of the ratio N/M for RSA vary between models. Moreover, while the number of model evaluations generally increases with model complexity in our case studies, for EET, the ranking of the 50 model parameters of the SWAT model converges before the ranking of the 13 model parameters of the HBV model. Besides model complexity, we believe that the rate of convergence depends on other factors, such as the relative sensitivity of the parameters (i.e. the closer the sensitivity of the influential parameters, the slower the convergence of the ranking). We expect to observe even more variations between convergence rate of GSA when considering a larger range of models and types of parameter variation. In particular, highly non-linear models (e.g. models that exhibit low frequency high consequence effects) are likely to show different convergence dynamics. Therefore, our study demonstrates that it is essential to check the convergence of the GSA results within the case under study and not only rely on numbers reported in previous studies. In particular, we find that sample sizes needed to reach the convergence are larger than suggested in the literature (e.g. Saltelli et al., 2008, Table 6.9) (Fig. 6, Table 2). This is consistent with results reported in the few other studies that specifically focus on GSA convergence (Fig. 1).

Secondly, we note that the convergence of ranking and screening can be reached at significantly lower number of model evaluations than the convergence of the sensitivity indices. This observation emphasizes the importance of distinguishing between the three types of convergence (indices, ranking and screening) to use computational resources efficiently. Full convergence of the value of the sensitivity indices may not be needed if one is interested in ranking (Factor Prioritization) or screening (Factor Fixing). In this case, trying to reach the full convergence of the value of the sensitivity indices constitutes a Type III error or framing error (Saltelli et al., 2008), given that the right answer (sample size for convergence) is sought for the wrong questions (what is the exact value of the sensitivity indices?). If instead the objective would be Variance Cutting, a converged value of VBSA sensitivity indices is needed. For RSA and EET, the values of the sensitivity indices do not provide much information by themselves and determining converged values seems to be of little interest.

Furthermore, we find that it is important to validate the screening results to avoid classifying influential parameters as insensitive. For EET, we obtain similar values of the screening threshold (between 0.025 and 0.03) for both the HBV and the SWAT models. A more exhaustive analysis would be needed in order to study the applicability of these values to other case studies. For VBT, we obtain negative values of the screening threshold very close to zero, even after a high number of model evaluations (more than 200,000). This indicates that the sensitivity indices for the insensitive parameters have not perfectly convergence because their actual value should be zero or very small but positive. Consequently, the exact (converged) value of the screening threshold T cannot be determined in these case studies. Nonetheless, we showed that the typical value of the screening threshold used for the VBT method (0.01) is not suitable for screening the parameters of the SWAT model. Some influential parameters appear to have a sensitivity index below this threshold value and the actual value of the screening threshold is thus below 0.01.

Finally, we can draw a comparison between the different GSA methods. In the case studies analysed here, EET provides ranking and screening of the model parameters consistent with VBT using much fewer model evaluations. For the SWAT model, after applying the validation test, we even identified a higher number of insensitive parameters with EET compared to VBT. Therefore, for SWAT, a complex model, VBT may not be suitable, since it would require a very high number of model evaluations and EET may be a better choice. Moreover, the results provided by RSA converge quickly in the three case studies. More broadly, given that the different GSA methods rely on different assumption, we recommend applying different GSA methods to the same case study within the limits of the available computational resources.

5.2. Consistency of our results with previous studies

Our study confirms the results found in the literature regarding the relative computational cost of the three GSA methods EET, RSA and VBSA (comparing our results in Table 2 and Fig. 6 with the results of other studies summarized in Fig. 1). We also find that for a given number of parameters, the ranking provided by RSA converges before the ranking provided by VBSA. Likewise, ranking and screening provided by EET converge before the ranking and the screening provided by VBSA. In particular, Campolongo et al. (2007) empirically demonstrated that EET allows for screening the model parameters with much fewer model evaluations than VBSA, which is corroborated by our study. However, we show that the result does not hold when the objective is the full convergence of the indices: the order of magnitude of the number of model evaluations necessary to reach convergence of the indices is the same for EET and VBSA, while RSA is much less computationally expensive. Moreover, for HyMod and HBV, convergence of indices and ranking given by Variance-Based Main effect (VBM) requires fewer model

evaluations than Variance-Based Total effect (VBT). Nossent et al. (2011) already noticed that the convergence of VBM tends to be faster than the convergence of VBT. No conclusion can be drawn from the results of the SWAT model on this point, because sensitivity indices did not converge for either VBM or VBT even after 520,000 model evaluations and the convergence statistic for ranking underestimates the rate of convergence in particular for VBM (see Section 4.2).

It is worth comparing our results with the study by Yang (2011) for the HyMod model applied to the Leaf River Catchment. Although the experimental set-up differs between Yang's and our study (simulation period, threshold for RSA, sampling strategy), we observe similarities in the results. The sample sizes for the convergence of the sensitivity indices (EET and RSA) and of ranking (VBSA and EET) have the same order of magnitude. Yet, we note a significant difference for the convergence of the sensitivity indices for VBSA ($n = 30{,}000$ in our study while $n = 3000$ in Yang's study). In Yang's study, the convergence analysis is performed qualitatively and in ours, the quantitative criterion used may be conservative. No sample size is explicitly given for the convergence of the ranking for RSA in Yang's paper, while screening convergence is not considered.

Finally, we observe a general coherence between EET and VBT results since both methods consistently separate out the most sensitive and the least sensitive parameters, in agreement with previous studies (e.g. Campolongo et al., 2007; Confalonieri et al., 2010; Herman et al., 2013). However, also in accordance with these previous studies, differences are observed in parameter ranking. EET may be a suitable alternative to VBT when model simulations are computationally expensive depending on the specific case study.

5.3. Limitations of our study

Our study introduces methods to formally assess the convergence of GSA results, relying on the definition of quantitative convergence statistics (Section 2.1). We set threshold values on these convergence statistics below which we assume that convergence has been reached. We believe that these threshold values ensure a sufficient degree of convergence of GSA results in order to obtain reliable results, although they could be conservative. The adjusted and weighted rank correlation coefficient here proposed, was shown to be more suitable than the Spearman rank correlation coefficient for comparison of parameter rankings. In fact, it emphasizes the differences in ranking for the more influential parameters while it reduces the impact of the low-sensitivity parameters. However, low-sensitivity parameters can still contribute to the value of the statistic when their sensitivity index is not negligible and when their rank is highly variable (see Section C of our Supplementary Material and discussion in Section 4.2). Therefore, the convergence statistic for ranking may underestimate the rate of convergence, which leads again to an estimate on the conservative side. In fact, when the statistic takes low values, GSA users can be quite confident that ranking convergence has been reached.

The methodology introduced to assess convergence may not be suitable when sample sizes are very small. In this case, the sample may not provide sufficient coverage of the parameter space so that bootstrapping may show wrongly low uncertainty of the sensitivity index estimates (see discussion in Section 4.3). This problem has for example been observed by Isaksson et al. (2008). At very small sample sizes, our analysis could be misleading and incorrectly suggest that GSA has converged. Likewise, when low frequency high consequence events can occur in a model (i.e. a small number of input data points can produce a large effect on the output), bootstrapping might fail to assess the uncertainty of the sensitivity index estimates (if these highly influential values are not present in the sample).

The methodology for screening (introduced in Section 2.1) applies for models with a reasonable number of input factors so that it is computationally affordable to estimate the sensitivity for every single input factor. When in contrast the number of input factors is very high compared to available computational resources (like in supersaturated designs), it might not be possible to estimate the effect of every single input factor. Nevertheless, the methodology proposed in our study could be applied for such models if input factors are assigned to a given number of groups, and if GSA is performed by taking these groups of input factors as inputs. In this case, screening consists in identifying the influential groups (i.e. at least one input factor in the group is sensitive) and the non-influential ones (i.e. all input factors in the group are insensitive). We refer to Saltelli et al. (2008) for more details on screening for supersaturated designs and group sampling.

Finally, we propose a quantitative validation method for the screening results based on the computation of the KS-statistic between unconditional output (obtained by varying all parameters) and conditional output (only influential parameters are varied) (Section 2.2). One main drawback of this method is that it requires further model evaluations for the computation of the conditional outputs. Further investigation is needed regarding the robustness of the KS-test. One possibility would be to use the bootstrap technique to compute the KS-statistic. However, in our study we found that bootstrapping tends to overestimate the KS-statistic (more details are given in Section D of our Supplementary Material).

6. Conclusions

We examine three widely used GSA methods, the Elementary Effect Test (EET, or method of Morris), Regional Sensitivity Analysis (RSA) and Variance-Based Sensitivity Analysis (VBSA, or Sobol' method). These methods are based on the computation of sensitivity indices through Monte Carlo simulations to measure the influence of parameters or other model inputs on the model output. We test these methods for the model parameters of three hydrological models with increasing complexity, the HyMod (5 parameters), HBV (13 parameters) and SWAT (50 parameters) models. The methods introduced here can be generalized to other case studies and other types of input factors as long as the associated uncertainty (distribution) can be quantified.

The methodological contribution of this paper is twofold. First, we define quantitative criteria to assess the convergence of sensitivity indices, of ranking (ordering among the influential parameters) and of screening (identification of the insensitive parameters). Second, we propose a quantitative and unconditional method to validate the screening results to avoid classifying influential parameters as non-influential.

The results of our study show that EET can provide a good approximation of the ranking and screening given by VBSA for much fewer model evaluations, as has already been noted in previous studies. As far as RSA is concerned, it appears to converge quickly in the case studies considered, although, as discussed in previous studies, it cannot be used when the objective is to study parameter interactions.

Our study demonstrates that it is indeed important to separately assess the convergence of sensitivity indices, ranking and screening, since these three objectives may require different numbers of model evaluations. It is not always necessary to reach the full convergence of the value of all the sensitivity indices. In fact, the parameter ranking and the sensitivity indices of the low-influential parameters (and therefore the screening) may converge first. We also observed that typical values of the sample

size sometimes suggested in the literature (e.g. Table 6.9 in Saltelli et al. (2008)) can be insufficient to reach convergence of GSA results, as observed for the two more complex models analysed here (HBV and SWAT). Moreover, we found that typical values of the screening threshold can be inadequate and lead to a misclassification of influential parameters as insensitive.

Since no clear relationship emerged between the number of model parameters subject to GSA and the number of model evaluations necessary to reach convergence, we recommend that GSA users always check the convergence of their results within their specific case study. Likewise, the choice of the screening threshold should always be validated in order to avoid classifying influential parameters as non-influential. In this paper, we introduce and test a number of convergence criteria and a validation procedure to formally do this. In particular, the convergence analysis can be done without additional model evaluations when using bootstrapping.

Further investigation is needed for:

- the convergence statistic for ranking in order to make it potentially less conservative,
- the KS-test in order to formalize the assessment of its convergence,
- the bootstrap technique in order to overcome its limitations in particular for the KS-statistic,
- the LHS design in order to develop and test strategies that would avoid a loss of stratification when increasing or decreasing the sample size.

Additionally, future work should include a comparison of the convergence speed between different sampling strategies to help GSA users with this choice.

Acknowledgements

This work is supported by a University of Bristol Alumni Postgraduate Scholarship to F.S. Partial support for F.P. and T.W. was provided by Natural Environment Research Council [Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1]. We thank Olkeba Tolessa Leta who set up the SWAT model for the Zenne River basin. We thank Farkhondeh Khorashadi Zadeh and Ann van Griensven for providing support for the implementation of the SWAT model. We thank Toby Dunne for computer support. We thank three anonymous referees for the many useful suggestions that have contributed to a significant improvement of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2016.02.005.

References

Ali, Y.S.A., Crosato, A., Mohamed, Y.A., Abdalla, S.H., Wright, N.G., 2014. Sediment balances in the Blue Nile river basin. Int. J. Sediment Res. 29 (3), 316–328. http://dx.doi.org/10.1016/S1001-6279(14)60047-0.

Andres, T.H., 1997. Sampling methods and sensitivity analysis for large parameter sets. J. Stat. Comput. Simul. 57 (1–4), 77–110. http://dx.doi.org/10.1080/00949659708811804.

Archer, G.E.B., Saltelli, A., Sobol', I.M., 1997. Sensitivity measures, ANOVA-like techniques and the use of bootstrap. J. Stat. Comput. Simul. 58 (2), 99–120. http://dx.doi.org/10.1080/00949659708811825.

Arnold, J.G., Allen, P.M., Bernhardt, G., 1993. A comprehensive surface-groundwater flow model. J. Hydrol. 142 (1–4), 47–69. http://dx.doi.org/10.1016/0022-1694(93)90004-S.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part 1: model development. J. Am. Water Resour. Assoc. 34 (1), 73–89. http://dx.doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Bae, D.H., Jung, I.W., Lettenmaier, D.P., 2011. Hydrologic uncertainties in climate change from IPCC AR4 GCM simulations of the Chungju Basin, Korea. J. Hydrol. 401 (1–2), 90–105. http://dx.doi.org/10.1016/j.jhydrol.2011.02.012.

Bannwarth, M.A., Sangchan, W., Hugenschmidt, C., Lamers, M., Ingwersen, J., Ziegler, A.D., Streck, T., 2014. Pesticide transport simulation in a tropical catchment by SWAT. Environ. Pollut. 191, 70–79. http://dx.doi.org/10.1016/j.envpol.2014.04.011.

Bergström, S., 1995. The HBV model (Chapter 13). In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, Colorado, USA, pp. 443–476.

Bouraoui, F., Grizzetti, B., 2008. An integrated modelling framework to estimate the fate of nutrients: application to the Loire (France). Ecol. Model. 212 (3–4), 450–459. http://dx.doi.org/10.1016/j.ecolmodel.2007.10.037.

Boyle, D., 2001. Multicriteria Calibration of Hydrological Models. Ph.D. thesis. Departement of Hydrology and Water Resources, University of Arizona, Tucson.

Campolongo, F., Saltelli, A., 1997. Sensitivity analysis of an environmental model: an application of different analysis methods. Reliab. Eng. Syst. Saf. 57 (1), 49–69. http://dx.doi.org/10.1016/S0951-8320(97)00021-5.

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.

Campolongo, F., Saltelli, A., Cariboni, J., 2011. From screening to quantitative sensitivity analysis. A unified approach. Comput. Phys. Commun. 182 (4), 978–988. http://dx.doi.org/10.1016/j.cpc.2010.12.039.

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. Environ. Model. Softw. 25 (4), 479–488. http://dx.doi.org/10.1016/j.envsoft.2009.10.005.

Cosenza, A., Mannina, G., Vanrolleghem, P.A., Neumann, M.B., 2013. Global sensitivity analysis in wastewater applications: a comprehensive comparison of different methods,. Environ. Model. Softw. 49, 40–52. http://dx.doi.org/10.1016/j.envsoft.2013.07.009.

Dancelli, L., Manisera, M., Vezzoli, M., 2013. On two classes of Weighted Rank Correlation measures deriving from the Spearman's ρ. In: Giudici, P., Ingrassia, S., Vichi, M. (Eds.), Statistical Models for Data Analysis. The Springer, pp. 107–114.

Duan, et al., 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.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, New York, USA.

Freer, J., Beven, K., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resour. Res. 32 (7), 2161–2173. http://dx.doi.org/10.1029/95WR03723.

Gharari, S., Hrachowitz, M., Fenicia, F., Savenije, H.H.G., 2013. An approach to identify time consistent model parameters: sub-period calibration. Hydrol. Earth Syst. Sci. 17, 149–161. http://dx.doi.org/10.5194/hess-17-149-2013.

Grillakis, M.G., Tsanis, I.K., Koutroulis, A.G., 2010. Application of the HBV hydrological model in a flash flood case in Slovenia. Nat. Hazards Earth Syst. Sci. 10, 2713–2725. http://dx.doi.org/10.5194/nhess-10-2713-2010.

Hill, M.C., Tiedeman, C.R., 2007. Effective Groundwater Model Calibration: with Analysis of Data, Sensitivities, Predictions, and Uncertainty. John Wiley & Sons.

Hartmann, A., Wagener, T., Rimmer, A., Lange, J., Brielmann, H., Weiler, M., 2013. Testing the realism of model structures to identify karst system processes using water quality and quantity signatures. Water Resour. Res. 49 (6), 3345–3358. http://dx.doi.org/10.1002/wrcr.20229.

Herman, J.D., Kollat, J.B., Reed, P.M., Wagener, T., 2013. Technical Note: method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models. Hydrol. Earth Syst. Sci. 17, 2893–2903. http://dx.doi.org/10.5194/hess-17-2893-2013.

Iman, R.L., Conover, W.J., 1987. A measure of top-down correlation. Technometrics 29 (3), 351–357. http://dx.doi.org/10.2307/1269344.

Isaksson, A., Wallman, M., Göransson, H., Gustafsson, M.G., 2008. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognit. Lett. 29 (14), 1960–1965. http://dx.doi.org/10.1016/j.patrec.2008.06.018.

Kannan, N., White, S.M., Worrall, F., Whelan, M.J., 2007. Sensitivity analysis and identification of the best evapotranspiration and runoff options for hydrological modelling in SWAT-2000. J. Hydrol. 332 (3–4), 456–466. http://dx.doi.org/10.1016/j.jhydrol.2006.08.001.

Kelleher, C., Wagener, T., McGlynn, B., Ward, A.S., Gooseff, M.N., Payn, R.A., 2013. Identifiability of transient storage model parameters along a mountain stream. Water Resour. Res. 49 (9), 5290–5306. http://dx.doi.org/10.1002/wrcr.20413.

Kollat, J.B., Reed, P.M., Wagener, T., 2012. When are multiobjective calibration trade-offs in hydrologic models meaningful? Water Resour. Res. 48 (3) http://dx.doi.org/10.1029/2011WR011534.

Kolmogorov, A., 1933. Sulla determinazione empirica di una legge di distribuzione. G. dell'Istituto Ital. degli Attuari 4, 83–91.

Leta, O.T., Nossent, J., Velez, C., Shrestha, N.K., van Griensven, A., Bauwens, W., 2015. Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). Environ. Model. Softw. 68, 129–146. http://dx.doi.org/10.1016/j.envsoft.2015.02.010.

Ljung, L., 1999. System Identification: Theory for the User, second ed. PTR Prentice Hall, Upper Saddle River, NJ.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174. http://dx.doi.org/10.2307/1269043.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - A discussion of principles. J. Hydrol. 10 (3), 282–290. http://dx.doi.org/10.1016/0022-1694(70)90255-6.
Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2009. Soil and Water Assessment Tool Theoretical Documentation - Version 2009. Texas Water Resources Institute, College Station. Texas. Technical Report no. 406.
Norton, J., 2015. An introduction to sensitivity assessment of simulation models. Environ. Model. Softw. 69, 166–174. http://dx.doi.org/10.1016/j.envsoft.2015.03.020.
Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol' sensitivity analysis of a complex environmental model. Environ. Model. Softw. 26 (12), 1515–1525. http://dx.doi.org/10.1016/j.envsoft.2011.08.010.
Nossent, J., Bauwens, W., 2012. Multi-variable sensitivity and identifiability analysis for a complex environmental model in view of integrated water quantity and water quality modeling. Water Sci. Technol. 65 (3), p539–549. http://dx.doi.org/10.2166/wst.2012.884.
Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environ. Model. Softw. 67, 1–11. http://dx.doi.org/10.1016/j.envsoft.2015.01.004.
Pianosi, F., Sarrazin, F., Wagener, T., 2015. A Matlab toolbox for global sensitivity analysis. Environ. Model. Softw. 70, 80–85. http://dx.doi.org/10.1016/j.envsoft.2015.04.009.
Pianosi, F., Wagener, T., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., 2016. Sensitivity Analysis of Environmental Models:a Systematic Review with Practical Workflow, Environmental Modelling & Software. http://dx.doi.org/10.1016/j.envsoft.2016.02.008 in Press.
Romano, J.P., Shaikh, A.M., 2012. On the uniform asymptotic validity of subsampling and the bootstrap. Ann. Stat. 40 (6), 2798–2822. http://dx.doi.org/10.1214/12-AOS1051.
Saltelli, A., 2002. Making best use of model valuations to compute sensitivity indices. Comput. Phys. Commun. 145 (2), 280–297. http://dx.doi.org/10.1016/S0010-4655(02)00280-1.
Saltelli, A., Tarantola, S., 2002. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. J. Am. Stat. Assoc. 97, 702–709.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer. Wiley.
Savage, I.R., 1956. Contributions to the theory of rank order statistics-the two-sample case. Ann. Math. Stat. 27 (3), 590–615. http://dx.doi.org/10.1214/aoms/1177728170.
Seibert, J., 1997. Estimation of parameter uncertainty in the HBV model. Nord. Hydrol. 28 (4–5), 247–262.
Shin, M., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J., 2013. Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R. J. Hydrol. 503 http://dx.doi.org/10.1016/j.jhydrol.2013.08.047, 135–125.
Sieber, A., Uhlenbrook, S., 2005. Sensitivity analyses of a distributed catchment model to verify the model structure. J. Hydrol. 310 (1–4), 216–235. http://dx.doi.org/10.1016/j.jhydrol.2005.01.004.
Sin, G., Gernaey, K.V., Neumann, M.B., van Loosdrecht, M.C., Gujer, W., 2011. Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. Water Res. 45 (2), 639–651. http://dx.doi.org/10.1016/j.watres.2010.08.025.
Singh, R., Wagener, T., Crane, R., Mann, M.E., Ning, L., 2014. A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: application to a watershed in Pennsylvania, USA. Water Resour. Res. 50 (4), 3409–3427. http://dx.doi.org/10.1002/2013WR014988.
Smirnov, N., 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin Mathématique de l'Université de Moscou 2.
Sobol', I.M., 1990. Sensitivity estimates for nonlinear mathematical models, Matematicheskoe Modelirovanie 2, 112-118 (in Russian), translated in English (1993). In: Mathematical Modelling and Computational Experiments, 1, pp. 407–414.
Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., Xu, C., 2015. Global sensitivity analysis in hydrological modeling: review of concepts, methods, theoretical framework, and applications. J. Hydrol. 523, 739–757. http://dx.doi.org/10.1016/j.jhydrol.2015.02.013.
Sorooshian, S., Gupta, V.K., Fulton, J.L., 1983. Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: influence of calibration data variability and length on model credibility. Water Resour. Res. 19 (1), 251–259. http://dx.doi.org/10.1029/WR019i001p00251.
Spear, R.C., Hornberger, G.M., 1980. Eutrophication in peel inlet – II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14 (1), 43–49. http://dx.doi.org/10.1016/0043-1354(80)90040-8.
Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101. http://dx.doi.org/10.2307/1412159.
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, 793–817. http://dx.doi.org/10.5194/hess-11-793-2007.
Tram, V.N.Q., Liem, N.D., Loi, N.K., 2014. Assessing water availability in PoKo catchment using SWAT model. Khon Kaen Agric. J. 42 (Suppl. 2), 73–84.
Vaché, K.B., Eilers, J.M., Santelmann, M.V., 2002. Water quality modeling of alternative agricultural scenarios in the U.S corn belt. J. Am. Water Resour. Assoc. 38 (3), 773–787. http://dx.doi.org/10.1111/j.1752-1688.2002.tb00996.x.
Vanrolleghem, P., Mannina, G., Cosenza, A., Neumann, M., 2015. Global sensitivity analysis for urban water quality modelling: terminology, convergence and comparison of different methods. J. Hydrol. 522, 339–352. http://dx.doi.org/10.1016/j.jhydrol.2014.12.056.
Vanuytrecht, E., Raes, D., Willems, P., 2014. Global sensitivity analysis of yield output from the water productivity. Environ. Model. Softw. 51, 323–332. http://dx.doi.org/10.1016/j.envsoft.2013.10.017.
Van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2008. Characterization of watershed model behavior across a hydroclimatic gradient. Water Resour. Res. 44 (1) http://dx.doi.org/10.1029/2007WR006271.
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 Resour. Res. 38 (12), 48.1–48.13. http://dx.doi.org/10.1029/2001WR001118.
Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5 (1), 13–26. http://dx.doi.org/10.5194/hess-5-13-2001.
Wall, J.V., 1996. Practical statistics for astronomers - II. Correlation, data-modelling and sample comparison. Q. J. R. Astron. Soc. 37, 519–563.
Wang, J., Li, X., Fang, F., 2013. Parameter sensitivity analysis of crop growth models based on the extended Fourier amplitude sensitivity test method. Environ. Model. Softw. 48, 171–182. http://dx.doi.org/10.1016/j.envsoft.2013.06.007.
Yang, J., 2011. Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. Environ. Model. Softw. 26 (4), 444–457. http://dx.doi.org/10.1016/j.envsoft.2010.10.007.
Young, P.C., Spear, R.C., Hornberger, G.M., 1978. Modelling badly defined systems: some further thoughts. In: Proceedings SIMSIG Conference. Canberra, pp. 24–32.