Comparison of variance-based and moment-independent global sensitivity analysis approaches by application to the SWAT model

FARKHONDEH KHORASHADI ZADEH [(1)*], JIRI NOSSENT [(1,2)], FANNY SARRAZIN [(3)], FRANCESCA PIANOSI [(3)], ANN VAN GRIENSVEN [(1,4)], THORSTEN WAGENER [(3,5)] & WILLY BAUWENS [(1)]

[(1)]*Vrije Universiteit Brussel (VUB), Department of Hydrology and Hydraulic Engineering, Pleinlaan 2, 1050 Brussel, Belgium,*

[(2)]Flanders Hydraulics Research, Department of Mobility and Public Works, Flemish Government, Antwerp, Belgium

[(3)] Department of Civil Engineering, University of Bristol, University Walk, BS81TR, Bristol, UK

[(4)]UNESCO-IHE Institute for Water Education, Core of Hydrology and Water Resources, The Netherlands

[(5)]Cabot Institute, Royal Fort House, University of Bristol, Bristol, BS8 1UJ, UK

* Corresponding author: F. Khorashadi Zadeh, e-mail: Farkhondeh.Khorashadi.Zadeh@vub.ac.be

Abstract

Global Sensitivity Analysis (GSA) is an essential technique to support the calibration of environmental models by identifying the influential parameters (screening) and ranking them.

In this paper, the widely-used variance-based method (Sobol') and the recently proposed moment-independent PAWN method for GSA are applied to the Soil and Water Assessment Tool (SWAT), and compared in terms of ranking and screening results of 26 SWAT parameters. In order to set a threshold for parameter screening, we propose the use of a "dummy parameter", which has no influence on the model output. The sensitivity index of the dummy parameter is calculated from sampled data, without changing the model equations. We find that Sobol' and PAWN identify the same 12 influential parameters but rank them differently, and discuss how this result may be related to the limitations of the Sobol' method when the output distribution is asymmetric.

Keywords: Global sensitivity analysis, Moment-independent method, Variance-based method, PAWN, Sobol', SWAT

Software /data availability

The PAWN method is implemented in the SAFE Matlab/Octave Toolbox for GSA (Pianosi et al., 2015). SAFE is freely available for non commercial purposes at www.bristol.ac.uk/cabot/-resources/safe-toolbox/.

The Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998) is a public domain environmental simulator. The SWAT model as developed by Leta (Leta, 2013; Leta et al., 2015) for the River Zenne (Belgium) is used in this study.

1. Introduction

Due to advancements in the understanding of natural processes and their interactions, and due to the advancements in software engineering and the increased computational power, environmental modelling tools have become more complex over the past decades (e.g. Arnold et al., 1998; Rossman, 2009; DHI, 2011). In general, such complex simulators contain many parameters, most of which cannot be measured directly and can only be inferred by calibration to observed system responses (Yapo et al., 1998; Vrugt et al., 2002). Consequently, parameter estimation has become a major issue, which may limit the applicability of complex simulators (van Griensven et al., 2006). A manual calibration of a model with a large number of parameters is very tedious and time consuming (Vrugt et al., 2003). On the other hand, the efficiency of automatic calibration algorithms is reduced when the number of parameters is large (Duan et al., 1992). In fact, it is not feasible to include all the model parameters in the calibration process (Bekele and Nicklow, 2007; Nossent et al., 2011). In order to support the choice of which model parameters should be the focus of calibration, and which ones could be instead excluded from calibration (and set to 'default' values), Global Sensitivity Analysis (GSA) is becoming popular in environmental modeling practices (e.g. Muleta and Nicklow, 2005; Van Werkhoven et al., 2009; Norton, 2015; Pianosi et al., 2016). GSA indeed allows for the identification of the parameters that have the largest influence on a set of model performance metrics (so called 'factor prioritization') and the identification of non-influential parameters ('factor fixing') (Saltelli et al., 2008; Nossent et al., 2011). Other uses of GSA include the understanding and the interpretation of the model behavior, the prioritization of efforts for uncertainty reduction and the model simplification (Nossent et al., 2011; Pianosi et al., 2016).

The Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998) is a particular example of a relatively complex environmental simulator, which has been widely applied all over the world for watershed management purposes (e.g. Gassman et al., 2010; van Griensven et al., 2012; Bressiani et al., 2015). In SWAT, different watershed processes, including surface runoff, groundwater flow, plant growth, and pesticide and nutrient conversion and transport, are controlled by a large number of parameters (more than 100). Even when some of these parameters can be fixed a priori, calibration of SWAT remains quite challenging given the relatively large number of parameters (26 in our case) that are typically left to be varied simultaneously. Therefore, GSA is often applied prior to the calibration process to identify the most influential parameters and the non-influential ones (Cibin et al., 2010; Nossent et al., 2011; Leta et al., 2015).

Many different GSA methods have been developed (Sobol', 1990; Saltelli et al., 2000; van Griensven et al., 2006; Borgonovo, 2007; Pianosi and Wagener, 2015). Among them, the most well-established and widely-applied one is

probably the variance-based method of Sobol' (Sobol', 1990; applications to environmental models include Pappenberger al., 2008; van Werkhoven et al., 2008; Nossent et al., 2011; Rosolem et al., 2012;; Gan et al., 2014). In general, variance-based methods seek to measure sensitivity to an uncertain input (parameter) using the contribution of that input to the total variance of the model output (a metric of model performance, in the context of model calibration). A well-known merit of variance-based methods is their ability to quantify the individual parameter contribution and the contribution resulting from parameter interactions, independently from assumptions on the form of the input-output relation (e.g. linearity and additivity). Moreover, variance-based sensitivity indices are easy to interpret, as they represent the fraction of the output variance caused by the variation of an input (Saltelli, 2002b).

Variance-based GSA methods use the variance - i.c. the second moment- as a measure of the output uncertainty, and as Saltelli (2002b) underlined, "implicitly assume that this moment is sufficient to describe the output variability". However, it has been recognized that the variance does not adequately represent output uncertainty when the model output is highly-skewed or multi-modal (Liu et al., 2006; Borgonovo et al., 2011; Pianosi and Wagener, 2015). To overcome this limitation, moment-independent GSA measures have been developed (Liu et al., 2006; Borgonovo 2007; Pianosi and Wagener, 2015). These methods -also known as density-based methods- use the entire output distribution to fully characterize the output uncertainty and to quantify the relative influence of the uncertain parameters. The main advantage of these methods, as compared to variance-based ones, is that they do not use a specific moment of the output distribution to measure the output variability and, therefore, are applicable regardless of its shape (e.g. symmetric or highly-skewed).

Pianosi and Wagener (2015) have proposed a moment-independent GSA method, called PAWN. It measures sensitivity based on the difference between the unconditional output distribution, obtained when all the parameters are free to vary, and the conditional output distribution, obtained when one of the parameters is fixed. Hereby, a Cumulative Distribution Function (CDF) is used to characterize the output distribution, whereas other density-based methods (e.g. the entropy-based method (Liu et al., 2006) and the δ-sensitivity measure (Borgonovo, 2007)) used the Probability Density Function (PDF). The main advantage of the PAWN method is that approximating CDFs by using empirical distributions of the data sample is much easier than approximating PDFs, because, it does not require any parameter tuning. This facilitates the analysis of the robustness and the convergence of the estimated sensitivity indices (Pianosi and Wagener, 2015).

In Pianosi and Wagener (2015), the PAWN method was tested on a simple conceptual hydrological model with only 5 parameters. To further investigate its effectiveness and efficiency, it is necessary to apply it to a more complex simulator with a higher number of parameters, such as SWAT, and to compare its results with those of another GSA

method. The main objective of this paper is therefore to evaluate and compare the application of the Sobol' and PAWN methods to a SWAT model. In particular, the two methods will be compared in terms of the rate of convergence of the respective sensitivity indices, and their results for parameter ranking and screening. To this end, 26 parameters of a SWAT model of the upstream sub-catchment of the River Zenne (Belgium) are analysed. As model outputs for sensitivity analysis, we consider two performance metrics for simulating daily river flows at the catchment outlet: the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and the mean error (ME). In performing parameter screening, we propose to calculate the sensitivity index of a "dummy parameter", which has no influence on the model output. The sensitivity index of this dummy parameter is used as a threshold to identify non-influential parameters. It is calculated numerically using sample data, without adding the dummy parameter explicitly to the model. The "dummy parameter approach" provides a practical way to sensibly define a threshold for screening, which is an unresolved issue increasingly discussed in recent GSA literature for both Sobol' and PAWN (e.g. Fanny et al., 2016). However, for the PAWN method, in particular, its effectiveness can be demonstrated by validating the screening results using the two-sample Kolmogorov-Smirnov statistical test (Smirnov, 1948).

2. Materials and methods

2.1. The variance-based Sobol' method

Sobol' (Sobol', 1990) is a "global, quantitative and model free" GSA method (Saltelli, 2002b), which also works properly for non-linear and non-monotonic models. In this method, the contributions of each parameter to the total model output variance, either by variation of the parameter itself or by interactions with other parameters, are quantified and expressed as Sobol' sensitivity indices. These indices provide a quantitative measure of the importance of the parameters and can be used for both factor fixing and factor prioritization (Saltelli et al., 2008).

To further describe the Sobol' method, the following generic model description is used:

$$Y = f(X) = f(X_1, \dots, X_p) \quad (1)$$

where $X = (X_1, \dots, X_p)$ is the set of p model parameters and Y is a scalar model output. For dynamic models, like the SWAT simulator used in this paper, the term "model output" does not refer to the entire simulated time series, but rather to a scalar variable summarizing those time series. In our application, for example, model outputs are two performance metrics measuring the distance between the simulated variable (river flow) and the observations.

The Sobol' method is based on the total variance decomposition (Sobol', 2001), i.e.

$$V(Y) = \sum_{i=1}^{p} V_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} V_{ij} + \cdots + V_{1,\ldots,p} \quad (2)$$

where V_i is the variance contribution of individual parameter X_i to the total variance, V_{ij} is a part of the total variance caused by the interaction between X_i and X_j and $V_{1,\ldots,p}$ is the variance due to the interaction between all parameters.

The partial variance V_i is called the first-order or main effect of X_i on Y. In the Sobol' method, the first-order sensitivity index S_i is obtained by normalizing the main effect V_i by the total variance $V(Y)$:

$$S_i = \frac{V_i}{V(Y)} \quad (3)$$

The first-order sensitivity index S_i can be described as the reduction of the total model output variance that would be obtained on average when the uncertainty about X_i would be removed by setting X_i to a fixed value (Tarantola et al., 2002).

Similarly, the higher order sensitivity indices, which characterize the interactions between the parameters, are calculated using the higher order partial variances (Sobol', 2001). Homma and Saltelli (1996) explicitly introduced the concept of total effect of the parameter X_i on Y, which accounts for the total contribution of parameter X_i to the output variance. Therefore, the total sensitivity index S_{Ti} is the sum of the main effect of X_i and all its interactions with the other parameters up to the p^{th} order. To calculate the total sensitivity index S_{Ti}, the variance $V_{\sim i}$, which is the total contribution of all parameters, except X_i, is used (Homma and Saltelli, 1996):

$$S_{Ti} = 1 - \frac{V_{\sim i}}{V(Y)} \quad (4)$$

The total sensitivity index S_{Ti} represents the fraction of the total output variance that would remain on average as long as X_i stays unknown (Tarantola et al., 2002).

For an additive model and under the assumption of independent model parameters, S_{Ti} and S_i are equal and the sum of all S_i (and all S_{Ti}) is 1. For a non-additive model, where parameter interaction exists, S_{Ti} is greater than S_i and the sum of all S_i is less than 1, while the sum of all S_{Ti} is greater than 1. Therefore, the difference between S_{Ti} and S_i represents the interaction between parameter X_i and the other parameters (Saltelli, 2002b). Obviously, the same information could be obtained by calculating all partial variances in Equation (2). However, for a large number of parameters, this leads to a high computational cost (Rabits and Alis, 2000). For this reason, in the applications of variance-based methods, it is very common to only compute the set of all S_i and S_{Ti}, which provides a quite good representation of the model sensitivities at a more reasonable cost (Saltelli, 2002b).

In practice, for complex and non-linear models, calculating the variances using analytical integrals is usually impossible. The main breakthrough in the Sobol' method was the computation algorithm that allows the direct estimation of the variance-based sensitivity indices from a set of values of $f(X)$ only, rather than the analytical solution (Sobol', 2001). The algorithm was further extended by Homma and Saltelli (1996) and Saltelli (2002a). It uses Monte Carlo integration, which is a numerical integration based on repeated random samples of the model output. Evidently, Monte Carlo integrals are closer to their converged value when more samples are used (Gan et al., 2014).

To estimate the sensitivity indices using Monte Carlo integrals, two independent parameter sample matrices are generated. These matrices are denoted as M_1, the "sample" matrix, and M_2, the "re-sample" matrix (Saltelli, 2002a):

$$M_1 = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & & & \\ X_{N1} & X_{N2} & \dots & X_{Np} \end{pmatrix}, \qquad M_2 = \begin{pmatrix} X'_{11} & X'_{12} & \dots & X'_{1p} \\ X'_{21} & X'_{22} & \dots & X'_{2p} \\ \dots & & & \\ X'_{N1} & X'_{N2} & \dots & X'_{Np} \end{pmatrix} \tag{5}$$

where N is the sample size and p is the number of parameters.

The total model output variance $V(Y)$ is estimated using $M_1$and M_2, as:

$$\hat{V}(Y) = \frac{1}{2N-1}\sum_{r=1}^{N}\{f^2(X_{r1}, X_{r2}, \dots, X_{rp}) + f^2(X'_{r1}, X'_{r2}, \dots, X'_{rp})\} - \hat{f}_0^2 \tag{6}$$

where $\hat{\dots}$ stands for the estimate, $f(X_{r1}, X_{r2}, \dots, X_{rp})$ and $f(X'_{r1}, X'_{r2}, \dots, X'_{rp})$ are the model output evaluated against the parameter combinations in the sample matrix M_1 and the re-sample matrix M_2, respectively, and $\hat{f}_0$ is the expected value of the model output, estimated using the following equation (Homma and Saltelli, 1996).

$$\hat{f}_0^2 = \frac{1}{N}\sum_{r=1}^{N} f(X_{r1}, X_{r2}, \dots, X_{rp}) \times f(X'_{r1}, X'_{r2}, \dots, X'_{rp}) \tag{7}$$

The partial variance V_i, representing that part of the total variance $V(Y)$ that is caused by X_i individually, is estimated by:

$$\hat{V}_i = \frac{1}{N-1}\sum_{r=1}^{N}\{f(X_{r1}, X_{r2}, \dots, X_{rp}) \times f(X'_{r1}, X'_{r2}, \dots, X'_{r(i-1)}, X_{ri}, X'_{r(i+1)}, \dots, X'_{rp})\} - \hat{f}_0^2 \tag{8}$$

where $f(X'_{r1}, X'_{r2}, \dots, X'_{r(i-1)}, X_{ri}, X'_{r(i+1)}, \dots, X'_{rp})$ is the model output computed from a matrix where all parameters are from M_2, except X_i, which comes from M_1. Therefore, to calculate $\hat{V}_i$ for all p parameters, p sets of new N model evaluations are needed.

In order to calculate the total sensitivity index, the variance $V_{\sim i}$ (see Equation (4)), is estimated by:

$$\hat{V}_{\sim i} = \frac{1}{N-1}\sum_{r=1}^{N}\{f(X'_{r1},\, X'_{r2}, \ldots, X'_{rp}) \times f(X'_{r1},\, X'_{r2}, \ldots,\, X'_{r(i-1)},\, X_{ri},\, X'_{r(i+1)}, \ldots, X'_{rp})\} - \hat{f}_0^2 \quad (9)$$

As shown by Equation (9), no further model evaluations are required to calculate the total sensitivity index, once all the model evaluations needed for the first-order sensitivity index are available (see Equation (8)).

Finally, the first order indices S_i and the total sensitivity indices S_{Ti} are estimated using Equations (3) and (4). According to the method explained above, the computational cost for obtaining the full sets of first-order and total sensitivity indices is $N(p+2)$ (Saltelli, 2002a). In fact, 2 sets of N evaluations are needed to compute the model output against the sample matrix M_1 and the re-sample matrix M_2, and p sets of N model evaluations are needed for implementing Equations (8) and (9).

2.2. The density-based PAWN method

In contrast to the Sobol' method, PAWN (Pianosi and Wagener, 2015) is a density-based GSA method, where the entire model output distribution, rather than only its variance, is used to quantify the relative influence of the parameters on the model output. Therefore, by definition, the PAWN method is a moment-independent GSA approach. In general, density-based sensitivity indices measure the sensitivity to parameter X_i by the distance between the unconditional PDF of Y, which is obtained by varying all parameters simultaneously, and the conditional PDFs of Y, which are obtained by varying all parameters but X_i (i.e. X_i is fixed at a nominal value $\overline{X_i}$) (Liu et al., 2006; Borgonovo, 2007). In practice, PDFs are generally unknown and must be approximated using a data sample. However, Pianosi and Wagener (2015) pointed out the difficulties and limitations of deriving empirical PDFs, and suggested using CDFs, instead of PDFs, as the computation of the empirical CDF from a data sample does not require any parameter tuning and is much easier than the approximation of the PDF. Consequently, PAWN is very easy to implement and the analysis of the robustness and convergence of PAWN sensitivity indices is computationally very efficient. Other advantages and limitations of PAWN are discussed in Pianosi and Wagener (2015).

In introducing the PAWN method, Pianosi and Wagener (2015) propose to measure the distance between the conditional and unconditional CDFs by the Kolmogorov-Smirnov statistic (KS) (Kolmogorov, 1933; Smirnov, 1933), as below.

$$KS(X_i) = \max_{Y}\left|F_Y(Y) - F_{Y|X_i}(Y)\right| \quad (10)$$

where $F_Y(Y)$ is the unconditional CDF of the output Y and $F_{Y|X_i}(Y)$ is the conditional CDF when X_i is fixed. As $F_{Y|X_i}(Y)$ characterizes the output distribution when the variability due to X_i is removed, its distance from $F_Y(Y)$ indicates the

effect of X_i on Y. When $F_{Y|X_i}(Y)$ overlaps with $F_Y(Y)$ completely, $KS(X_i)$ is equal to zero, which means that removing the uncertainty about X_i does not affect the output distribution, i.e. X_i has no influence on Y. A large distance, instead, indicates a high influence of the parameter.

Since KS depends on the conditioning value of X_i, the PAWN sensitivity index T_i considers a statistic (e.g. maximum or median) over all possible value of X_i:

$$T_i = \underset{X_i}{\text{stat}}[KS(X_i)] \tag{11}$$

The PAWN index T_i is a global, quantitative and model-independent sensitivity index, which varies between 0 and 1 (the higher the value, the more influential X_i). It is worth nothing that both variance-based and moment-independent global sensitivity measures are part of a common rationale in which a global sensitivity measure can be seen as written in two pieces: an external statistic over the values of X_i and an inner statistic that measures the distance between the conditional and unconditional distributions (Borgonovo et al, 2016). Similar to the Sobol' indices, for complex and non-linear models, the analytical computation of the PAWN index T_i is usually impossible. Pianosi and Wagener (2015), therefore, suggested the following approximate numerical procedure. First, the KS statistic in Equation (10) is approximated by using empirical unconditional and conditional distributions. The empirical unconditional distribution is computed using N_u model evaluations from sampling the entire parameter space. The empirical conditional distributions are computed using N_c model evaluations from sampling all parameters except X_i, which is kept to a fixed value. Second, in Equation (11), the statistic with respect to the conditioning value of X_i is approximated using n randomly sampled values for the fixed parameter X_i. Therefore, the total number of model evaluations required to calculate the PAWN index T_i for all the p parameters is $N_u + n \times N_c \times p$.

A technical question that was left unaddressed in Pianosi and Wagener (2015) is whether the choice of the KS statistic for measuring the distance between the unconditional and conditional CDFs would affect the PAWN sensitivity results. In this study we thus investigate the use of the Anderson-Darling (AD) statistic (Anderson and Darling, 1952) instead of the KS statistic. Interestingly, the results of comparison (reported in Section A of the Supplementary Materials) show that these two statistics provide very similar parameter rankings for our SWAT model. This result increases the reliability of the conclusions drawn from the application of the PAWN method.

Another advantage of using CDFs when defining the PAWN sensitivity indices is that the two-sample Kolmogorov-Smirnov (KS) (Smirnov, 1948) can be applied to statistically determine non-influential parameters (Pianosi and Wagener, 2015; Sarrazin et al., 2016). Here, the null hypothesis is that the conditional and unconditional CDFs are the same, i.e. the considered parameter is non-influential. The null hypothesis is rejected (i.e. the parameter is

influential) if the *p*-value is equal to or smaller than the selected significance level α (typically set to 5%). The details of calculating *p*-values are described in Massey (1951) and Marsaglia et al. (2003). The significance level is the probability of rejecting the null hypothesis while it is true, i.e. Type I error rate. Therefore, when applying the test, we reject the null hypotheses (and consider parameters influential) with the guarantee that the Type I error rate is no greater than α. In our study, we will use this statistical test for screening parameters when using the PAWN method. We also compare such approach to another screening approach, which applies to both PAWN and Sobol', and is described in the next section.

2.3. Identifying non-influential parameters by using a dummy parameter

In theory, the sensitivity index of a non-influential parameter has a value of zero. The value of zero for the PAWN sensitivity index means that the unconditional CDF coincides with the conditional one, i.e. fixing parameter X_i, has no influence on the model output distribution. The value of zero for the Sobol' total sensitivity index indicates a zero contribution of X_i to the total variance. However, since numerical approximations, rather than analytical solutions, are utilized to calculate the sensitivity indices, small but non-zero indices may be obtained also for the non-influential parameters. For example, in the PAWN method, different samples are used to estimate the unconditional and the conditional CDFs. Since the sample size is limited, there can be small differences between these two estimated distributions, which lead to non-zero sensitivity indices for non-influential parameters. To set a threshold to identify non-influential parameters (i.e. parameter screening), in this paper, we propose to calculate the sensitivity index of a "dummy parameter", which has no influence on the model output. The sensitivity index of this dummy parameter provides an indication of the approximation error of the sensitivity analysis.

The operational way to use the sensitivity index of the dummy parameter for parameter screening is as follows: parameters whose index is above the dummy sensitivity index can therefore reliably be classified as influential; parameters whose index is below the dummy index are non-influential, because the detected contribution to the variance or the difference in conditional and unconditional CDFs is less than the approximation error.

It should be noted that no change in the model equations is needed to account for the dummy parameter, in other words, the dummy parameter is not added to the model. The sensitivity index of the dummy parameter is calculated by using the sampled data. In the following, the procedure and the algebraic equations to calculate the sensitivity index of the dummy parameter in the Sobol' and PAWN methods are explained.

Computation of the Sobol' indices of the dummy parameter:

The first-order and total order Sobol' sensitivity indices of the dummy parameter are calculated according to Equations (8) and (9). In these equations, the only difference between parameter sets of $f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{r(i-1)}^{'}, X_{ri}, X_{r(i+1)}^{'}, \dots, X_{rp}^{'})$ and $f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})$ is in the i^{th} component. When i corresponds to the dummy parameter, the model parameters of the vectors $(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{r(i-1)}^{'}, X_{ri}, X_{r(i+1)}^{'}, \dots, X_{rp}^{'})$ and $(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})$ are identical, and consequently, the model results evaluated against these two vectors are the same. Therefore, for the dummy parameter, $f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})$ replaces with $f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{r(i-1)}^{'}, X_{ri}, X_{r(i+1)}^{'}, \dots, X_{rp}^{'})$ in Equations (8) and (9), as below.

$$\hat{V}_{dummy} = \frac{1}{N-1}\sum_{r=1}^{N}\{f(X_{r1}, X_{r2}, \dots, X_{rp}) \times f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})\} - \hat{f}_0^2 \quad (12)$$

$$\hat{V}_{\sim dummy} = \frac{1}{N-1}\sum_{r=1}^{N}\{f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'}) \times f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})\} - \hat{f}_0^2 \quad (13)$$

where $\hat{V}_{dummy}$ is the variance contribution of individual dummy parameter and $\hat{V}_{\sim dummy}$ is the total contribution of all parameters, except the dummy one.

The total model output variance $V(Y)$ is estimated using Equation (6), just as for any other parameter. Finally, the first-order index S_i and the total sensitivity index S_{Ti} for the dummy parameter are calculated using Equations (3) and (4).

From Equations (12) and (13), it can be noticed that the computation of the first-order and the total order sensitivity indices of the dummy parameter does not require any additional model evaluations beyond those against the sample matrix M_1 ($f(X_{r1}, X_{r2}, \dots, X_{rp})$) and the re-sample matrix M_2 ($f(X_{r1}^{'}, X_{r2}^{'}, \dots, X_{rp}^{'})$), which were already obtained to estimate the sensitivity indices of the other parameters. The dummy sensitivity values can be interpreted as measuring the accuracy with which the (unknown) partial variances are approximated by the sample variances computed on matrices M_1 and M_2. As such, they provide an estimate of the approximation accuracy for the case under study. In theory, the Sobol' sensitivity indices of the dummy parameter are zero. However, in practice, their values depend on the (finite size) samples M_1 and M_2. Consequently, the estimates of the sensitivity indices of the dummy parameter are random and will change from one Sobol' application to another.

Computation of the PAWN index of the Dummy parameter:

Similarly to Sobol', the PAWN sensitivity index of the dummy parameter is calculated using the output samples. In this case, if the i^{th} parameter is the dummy one, the conditional output distribution $F_{Y|X_i}(Y)$ coincides by definition with the unconditional one, which is obtained by the simultaneous variation of all the model parameters. It should be noted

that for the model parameters, the value of the considered parameter remains unchanged for the conditional distribution. Therefore, for the dummy parameter, the unconditional and conditional random samples are from the same distribution. In theory, for infinite sample size, the distance between the CDFs of these two random samples is zero (i.e. KS=0). However, since in the numerical approximation of PAWN indices, all CDFs are empirically approximated using a limited sample size, the KS statistic for the dummy parameter is not zero. It represents the distance between the empirical distributions of two different samples generated from the same distribution. Therefore, it can be interpreted as a measure of the accuracy in approximating CDFs by the limited sample size and hence of the accuracy of the estimated PAWN indices. In operational terms, the PAWN index for the dummy parameter can be computed from at least two independent samples of unconditional output values, i.e. model evaluations against two independent samples where all model parameters are varied simultaneously. In this study however we decided to use 10 independent samples - and take a statistic of the KS value across those samples - so to obtain an estimate of the PAWN sensitivity for the dummy parameter completely consistent with the estimates obtained for the other model parameters. Just as for the Sobol' method, the estimated PAWN index for the dummy parameter is random and changes from one application to another.

2.4. Assessing robustness of sensitivity indices using bootstrapping

In order to assess the robustness of all the sensitivity indices estimated in this study, we computed 95% confidence intervals of the Sobol' and PAWN indices using the bootstrap technique (Efron and Tibshirani, 1994). Bootstrapping has been widely applied to assess the uncertainty of sensitivity indices and to derive their confidence bounds (e.g. Archer et al., 1997; Pappenberger et al., 2008; Yang, 2011; Nossent et al., 2011; Pianosi and Wagener, 2015).

For the Sobol' method, output samples are resampled B times, and for each bootstrap resample, the Sobol' indices are calculated. The obtained distributions of the Sobol' indices are used to derive the upper and lower bounds of the 95% confidence intervals. Using a high number of resamples (i.e. high value for B) leads to a symmetric and median centered sampling distribution, providing an accurate estimation of the confidence intervals (Nossent et al., 2011). Archer et al. (1997) suggested bootstrapping using 1000 or 2000 resamples ($B =$1000 or 2000) for the Sobol' method. Similarly, for the PAWN method, the 95% confidence intervals are estimated by repeating the calculation of the PAWN indices for B' bootstrap resamples of the unconditional and conditional output samples. Pianosi and Wagener (2015) estimated the 95% confidence intervals for the PAWN indices using 1000 bootstrap resamples ($B' =$ 1000). However, as opposed to previous PAWN applications by Pianosi and Wagener (2015), we use resampling without replacement, since the latter approach proved to provide more reliable confidence bounds. Further details are given in Section B of the Supplementary Materials.

2.5. The Soil and Water Assessment Tool (SWAT)

SWAT (Arnold et al., 1998) is a physically-based, semi-distributed environmental simulator that operates on a daily or sub-daily time step. The tool was originally developed to assess the impact of different watershed management practices on water quantity and quality in large river basins. To build up a SWAT model, the main input data includes weather data (e.g. precipitation, temperature, solar radiation and potential evapotranspiration), topographic features, a land use map and a soil type map. SWAT spatially divides a basin into sub-basins based on topographic conditions. Sub-basins are further divided into hydrological response units (HRUs), characterized by a given combination of land use, soil type and slope. The hydrological processes taken into account by SWAT include surface runoff, interception, evapotranspiration, infiltration, lateral flow, groundwater flow and percolation. Furthermore, SWAT can simulate the plant growth and the fate and transport of sediment, nutrients and pesticides. The main outputs are the water flow and the crop, sediment, nutrients and pesticide yields at sub-basin level. The computation of these processes and their outputs is governed by hundreds of parameters, defined at HRU, sub-basin or catchment level.

2.6. The case study

A daily time step SWAT model of the upstream sub-catchment of the River Zenne (Belgium) (Leta, 2013; Leta et al., 2015), is selected as a case study. The River Zenne drains an area of 1162 km^2, located in the central part of Belgium. The upstream sub-catchment, with an area of about 747 km^2, is dominated by agricultural land (56%), followed by pasture and mixed forest. The watershed has a temperate maritime climate and is usually wet during most of the year. The predominant precipitation type is rainfall, ranging from 700 mm/y to 1200 mm/y for the simulation period (1998-2005). Daily precipitation data is obtained from 6 stations. To calibrate and validate the model, daily stream flow data at two stations is used. More details about the data and the model can be found in (Leta, 2013; Leta et al., 2015). In this study, the first three years (1998-2000) are used for warming up the model. GSA is performed using the period 2001-2005, which includes wet, normal and dry years (Leta et al., 2015). The annual precipitation and the mean flow at the outlet of the catchment are given in Table 1.

Table 1. The annual precipitation and the mean flow at the outlet of the study area

	2001	2002	2003	2004	2005
Annual precipitation (mm/yr)	1100	1200	700	900	800
Mean flow (m^3/s)	5.96	5.77	3.5	3.19	3.22

2.7. Setting-up numerical experiments for the Sobol' and PAWN application to the SWAT model

In this study, the results of applying the Sobol' and PAWN methods to the SWAT model are compared, in terms of convergence rate, parameter screening and ranking. For this purpose, 26 parameters ($p = 26$) that affect the hydrological cycle of the SWAT model are selected to be analyzed (see list in Table 2). The flow at the outlet of the

basin is considered as SWAT output. The parameters selection is based on expert knowledge and on the fact that these parameters are commonly considered in sensitivity analysis and calibrations of SWAT models (Cibin et al. 2010; Nossent and Bauwens, 2012; Leta et al., 2015). Moreover, these parameters are used by the sensitivity analysis algorithm (i.e. Latin-hypercube-One-factor-at-a-time (LH-OAT)) incorporated in the SWAT simulator (van Griensven et al., 2006). Therefore, our selection reflects the set of parameters that SWAT model users would typically consider for calibration and/or for applying the simpler LH-OAT sensitivity analysis method.

In this research, a change of a given parameter is applied to the all HRU's in the basin, resulting in one sensitivity index per parameter for the whole basin. Spatial variability for the impact of the model parameters at the HRU level, according to different land uses, soil and slope types, is therefore not considered. However, sensitivity indices can, in principle, be computed at the HRU level, at the price of increasing the total number of parameters incorporated in the GSA (it is multiplied by the number of HRU's) (Nossent et al., 2011). The ranges of variations of our selected 26 parameters (reported in Table 2) are determined based on the SWAT manual (Arnold et al., 2011) and on the results of previous applications of SWAT to this and similar catchments (Leta, 2013; Leta et al., 2015; Nossent and Bauwens, 2012). Since there is no prior information on parameter distributions, parameter values are sampled from a uniform distribution within these ranges.

The Sobol' quasi-random sampling technique (Sobol', 1976) is used to create the parameter samples for both the Sobol' and PAWN methods. According to Sobol' (1976), quasi-random numbers enhance the convergence rate as compared to regular Monte Carlo random numbers.

AS mentioned in Section 2.1, the effects of the uncertain parameters are assessed relative to a scalar variable that summarizes the simulated time series. Typically, in the sensitivity analysis literature, a performance measure is used as a scalar variable when the objective is to inform the calibration procedure (e.g. van Griensven et al., 2006; Pappenberger et al., 2008; Nossent et al., 2011). Obviously, the definition of the performance measure affects the sensitivity analysis results, because different performance measures may have different sensitivity to the model parameters (Pianosi et al., 2016). In this study, the comparison of the Sobol' and PAWN methods is performed twice: once using the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) as performance metric, and once using the mean error (ME), i.e.

$$NSE = 1 - \frac{\sum_{t=1}^{M}(y_t^o - y_t^s)^2}{\sum_{t=1}^{M}(y_t^o - \overline{y^o})^2} \quad (14)$$

$$ME = \frac{\sum_{t=1}^{M}(y_t^o - y_t^s)}{M} \quad (15)$$

where y_t^o is the observed flow on day t, y_t^s is the simulated flow on day t, $\overline{y^o}$ is the average of the observations and M is the total number of days.

In order to analyze and compare the convergence rate of both methods, the sensitivity indices are calculated for increasing sample sizes. For the Sobol' method, the maximum sample size considered is 9000 (N=9000), resulting in 252,000 $(= N(p+2))$ model evaluations. For the PAWN method, sensitivity indices are calculated using 10 conditioning values ($n = 10$) for each parameter and up to 1000 random samples for approximating the unconditional and conditional CDFs ($N_u = N_c = 1000$), requiring 261,000 model runs ($= N_u + n \times N_c \times p$). The maximum KS value, estimated for each parameter, is considered as a PAWN sensitivity index. The maximum sample sizes are selected based on the recommended values in the literature (Sobol', 1967; Tang et al., 2007; Saltelli et al., 2008; Nossent et al., 2011; Pianosi and Wagener, 2015). The 95% confidence intervals for the sensitivity indices are calculated using 1000 bootstrap resamples, as mentioned in Section 2.4.

Table 2. SWAT parameters considered for the Sobol' and PAWN sensitivity analysis, and their ranges of variation

Parameter	Definition	Process	Range
Alpha_Bf	Baseflow recession factor (1/day)	Groundwater	[0,1]
Biomix	Biological mixing efficiency (-)	Evapotranspiration	[0,1]
Blai	Maximum potential leaf area index for crop (mm)	Evapotranspiration	[0.5,10]
Canmax	Maximum canopy index (mm)	Evapotranspiration	[0,10]
Ch_K2	Hydraulic conductivity in main channel (mm/h)	Routing	[0,150]
Ch_N2	Manning coefficient for channel (-)	Routing	[0,1]
Cn2	SCS runoff curve number for moisture condition II (-)	Surface runoff	[35,98]
Epco	Plant uptake compensation factor (-)	Evapotranspiration	[0.1,1]
Esco	Soil evaporation compensation factor (-)	Evapotranspiration	[0,1]
Gw_Delay	Groundwater delay (days)	Groundwater	[1,60]
Gw_Revap	Groundwater 'revap' coefficient (-)	Groundwater	[0.02,0.2]
Gwqmn	Threshold storage in shallow aquifer for return flow (mm)	Groundwater	[10,500]
Rchrg_Dp	Groundwater recharge to deep aquifer (-)	Groundwater	[0,1]
Revapmn	Threshold storage in shallow aquifer for 'revap' (mm)	Groundwater	[1,500]
Sftmp	Snowfall temperature (°C)	Snow	[-5,5]
Slope	Average slope steepness (m/m)	Lateral flow	[0,1]
Slsubbsn	Average slope length (m)	Routing	[10,150]
Smfmn	Minimum melt rate for snow (mm/°C/day)	Snow	[0,10]
Smfmx	Maximum melt rate for snow (mm/°C/day)	Snow	[0,10]
Smtmp	Snow melt base temperature (°C)	Snow	[-5,5]
Sol_Alb	Soil albedo (-)	Evapotranspiration	[0,0.25]
Sol_Awc	Available water capacity of the soil layer (mm)	Soil water	[0,1]
Sol_K	Soil conductivity (mm/h)	Soil water	[0,2000]
Surlag	Surface runoff lag coefficient (-)	Surface runoff	[0.5,10]
Tlaps	Temperature laps rate (°C/km)	Evapotranspiration	[-10,10]
Timp	Snow pack temperature lag factor (-)	Snow	[-10,10]

3. Results

3.1. The distributions of the performance measures

As mentioned above, two different performance measures are considered for the GSA: the NSE and the ME. The 9000 random sample generated for Sobol' application (i.e. model performance against random parameter matrices M_1) are used to obtain the empirical distributions of the NSE and ME. As shown in Figure 1, the distribution of NSE is negatively-skewed, while the distribution of ME is slightly bi-modal with a second (small) peak on the right. The statistical analysis, based on the KS test (Smirnov, 1948), the Jarque-Bera test (Jarque and Bera, 1987) and the Lilliefors test (Lilliefors, 1967), strongly rejects that the NSE and ME have a normal distribution. Using these two performance measures with different empirical PDF shapes allows to compare the results of the Sobol' and PAWN methods in both a situation of strongly non-symmetric distribution (NSE) and a situation of slightly bi-modal distribution (ME).

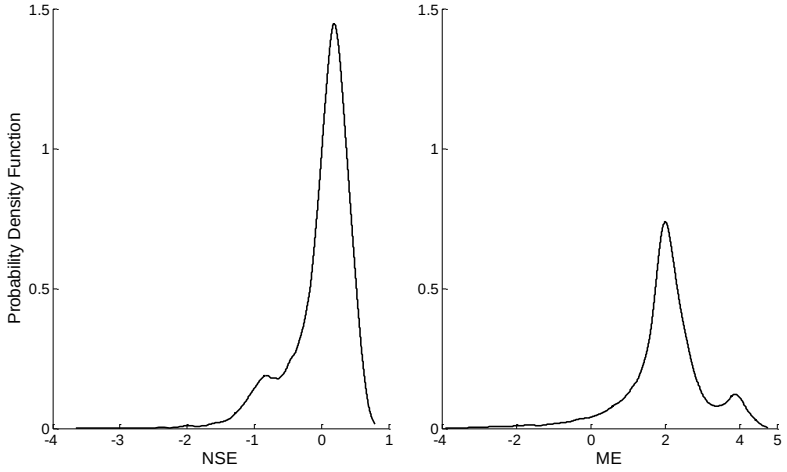

Figure 1. Estimated Probability Density Function of the Nash-Sutcliffe efficiency (NSE) and the mean error (ME), based on a sample of 9000 model evaluations against randomly sampled parameter sets. The distribution of NSE is negatively-skewed, while the distribution of ME is slightly bi-modal.

3.2. The convergence analysis of the Sobol' and PAWN methods

In this study, numerical approximation algorithms, based on Monte Carlo simulations, are applied in both the Sobol' and PAWN methods to calculate the sensitivity indices. Therefore, obtaining converged values for the sensitivity indices is a crucial issue to guarantee reliable estimates of the sensitivity indices and a reliable parameter ranking.

In order to investigate the convergence of the sensitivity indices and compare the convergence rate of the Sobol' and the PAWN methods, the Sobol' total sensitivity indices and the PAWN indices are calculated for an increasing sample size using both methods. As equal sample sizes result in different required numbers of model simulations in the

Sobol' and the PAWN methods, the comparison is performed based on the required number of model evaluations for increasing sample size. Since small sample sizes result in large variations in the sensitivity indices, the comparison is started from a sample size of 900 ($N = 900$) and a sample size of 100 ($N_u = N_c = 100$) for the Sobol' and PAWN methods, respectively. It should be noted that the number of conditioning values for each parameter in the PAWN method is unchanged ($n = 10$). The fluctuations and the slope of the graphs for the increasing sample sizes are used as a measure to graphically analyze and compare the convergence. A graph with no significant fluctuation and a horizontal slope indicates almost complete convergence. According to such visual analysis, we found that for both methods, the sensitivity indices have converged- for most of the parameters- after 250,000 model simulations. The latter is equivalent to almost 1000 samples for the PAWN method ($N_u = N_c = 1000,\ n = 10$) and 9000 samples for the Sobol' method ($N = 9000$). As shown in Figure 2, which reports the evolution of the sensitivity indices of a selected number of parameters, the sensitivity index of the most influential parameter, Cn2 (curve number), converges very quickly to its final value. In both the Sobol' and PAWN methods, for the ME performance measure, the parameters with lower sensitivity indices (i.e. less influential parameters) do not have completely horizontal graphs. For example, the PAWN sensitivity index of Smtmp (snow melt base temperature) is still decreasing at large number of model evaluations, even if changes are rather small and possibly not affecting the conclusion about the relatively negligible influence of that parameter. A similar trend is observed for the other less-influential parameters and for the dummy parameter.

In order to complete the convergence analysis, the parameter ranking results for an increasing number of simulations are also evaluated (Figure 3). For the top ranked parameters, such as Cn2 and Slope, both methods provide stable results even with a limited sample size, while the parameter ranking for less-influential parameters shows fluctuations, especially for the PAWN method. The reason is mainly related to the small and nearly equal values for the sensitivity indices of the less-influential parameters, which causes shifts in the parameter ranking for these parameters, even for small changes in the sensitivity indices. A similar observation for the less-influential parameters is also reported by Nossent et al. (2011). It is worth noticing that such variations in indices and ranking positions of the less-influence parameters may be of minor importance when, as often the case, the aim of GSA is to properly rank parameters that do have a significant influence on the model outputs.

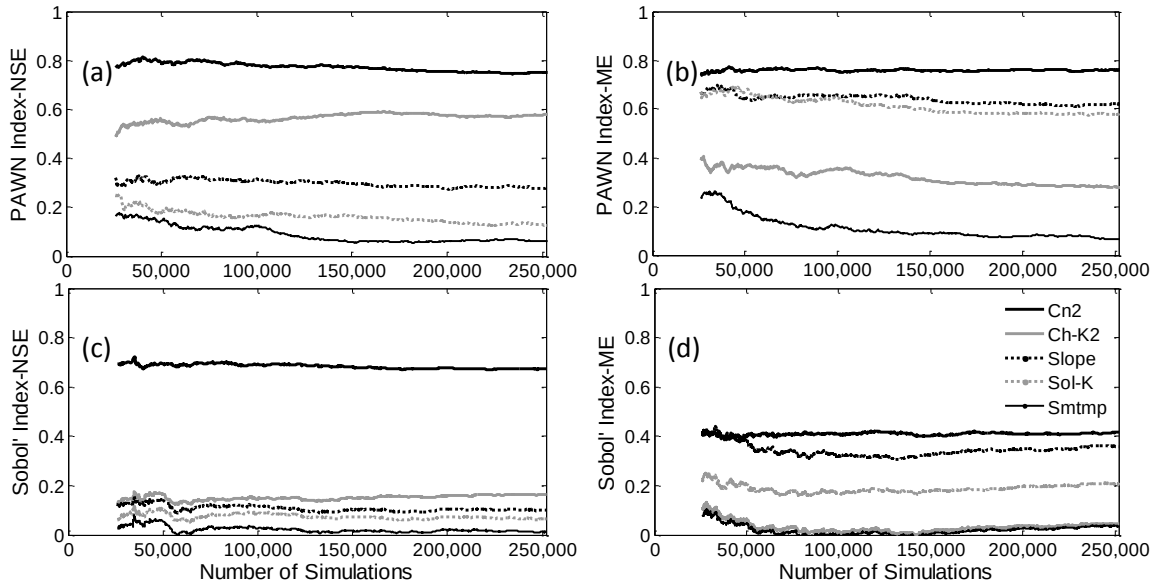

Figure 2. The convergence rates of the Sobol' and the PAWN methods are evaluated using the evolution of sensitivity indices for increasing sample sizes. The sensitivity indices using the NSE (a and c) reach stable values, while the results for ME (b and d), especially for less-influential parameters, do not have completely horizontal graphs.

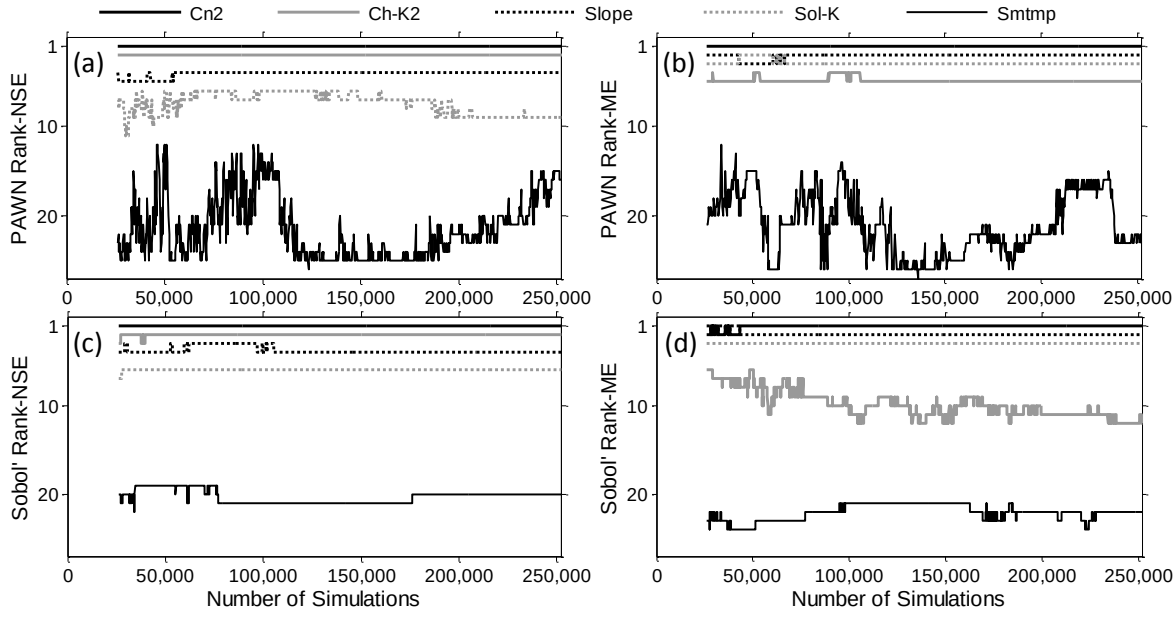

Figure 3. The convergence rate of the Sobol' and PAWN methods is evaluated using the evolution of the parameter ranking. (a) PAWN ranking for the NSE, (b) PAWN ranking for the ME, (c) Sobol' ranking for the NSE and (d) Sobol' ranking for the ME. Both methods provide a stable ranking for the top ranked parameters, while the ranking of the low-ranked parameters fluctuates, especially for the PAWN method.

3.3. Comparison of the Sobol' and PAWN parameter ranking and screening results

In this section, the results of the Sobol' total sensitivity indices and the PAWN indices of the 26 SWAT parameters, together with the dummy parameter, are presented and parameter rankings are compared (Figure 4). The Sobol' and PAWN sensitivity indices are estimated using sample sizes of 9000 (N=9000) and 1000 ($N_u = N_c = 1000$), respectively (maximum sample sizes considered in this study). The red lines in Figure 4 illustrate the 95% confidence intervals, estimated using percentiles of 1000 bootstrap resamples. Parameters are sorted in order of increasing sensitivity index, to allow immediate evaluation of the ranking. Since the Sobol' and PAWN methods have completely different background and rationale, comparing the values of the sensitivity indices does not provide any meaningful insights, however, the parameter ranking and screening results – based on the respective sensitivity indices- can be compared.

Comparison of Sobol' and PAWN ranking and screening results for the NSE

Based on the PAWN and Sobol' sensitivity indices for the NSE (Figures 4(a) and (b)), Cn2 is clearly the most important parameter, followed by Ch_K2 (hydraulic conductivity of the river bed) in both methods. As expected, the sensitivity index of the dummy parameter is small, but not zero. For the PAWN method, the confidence intervals of the sensitivity indices of the parameters with a rank 12 and worse overlap with that of the dummy parameter (see Figure 4(a)), and hardly distinguishable between each other. Therefore, this group of parameters is considered as non-influential. The same applies to parameters ranked 12 or worse by the Sobol' method (see Figure 4(b)). Although there are some differences in the parameter rankings produced by Sobol' and PAWN, the separation between the top 11 ranked parameters and the other "non-influential" ones, i.e. the screening result, is the same for both methods.

When the influential parameters are considered, it is interesting to notice that the PAWN method leads to more distinctive sensitivity indices, as compared to Sobol'. For example, Alpha_Bf (baseflow recession factor) and Slope (average slope steepness) have almost the same Sobol' sensitivity indices for the NSE, with confidence bounds largely overlapping, while PAWN indices are completely distinctive.

Comparison of Sobol' and PAWN ranking and screening results for the ME

As expected, using the ME as a model output provides different sensitivity indices (Figures 4(d) and (e)) and consequently different parameter ranking, as compared to NSE. Although Cn2 is still the most important parameter, the sensitivity indices of Slope and Sol_K (soil conductivity) are increased considerably and these two parameters are ranked 2nd and 3rd, respectively, for both the PAWN and Sobol' methods. Just as observed for the NSE case, the parameters ranked 13 and worse have almost the same sensitivity indices in both methods, and within the range of the variability of the dummy parameter. So, these parameters are considered non-influential. Again similarly to NSE,

the separation between the top 12 and the other parameters is similar for Sobol' and PAWN, although not exactly the same: PAWN in fact includes Revapmn in the list of the top 12 influential parameters and excludes Gw_Revap, while Sobol' does the opposite. It is also worth noticing that the separation between influential and non-influential parameters is very similar for NSE and ME (the only difference is the replacement of Gw_Revap by Revapmn according to PAWN). It implies that the choice between these two different performance measures, in this case study, does not significantly affect the parameter screening results. Finally, again similarly to NSE case, the Sobol' method does not clearly discriminate the relative importance of the influential parameters (the parameters ranked 5 and worse have nearly the same sensitivity indices in Figure 4(d)).

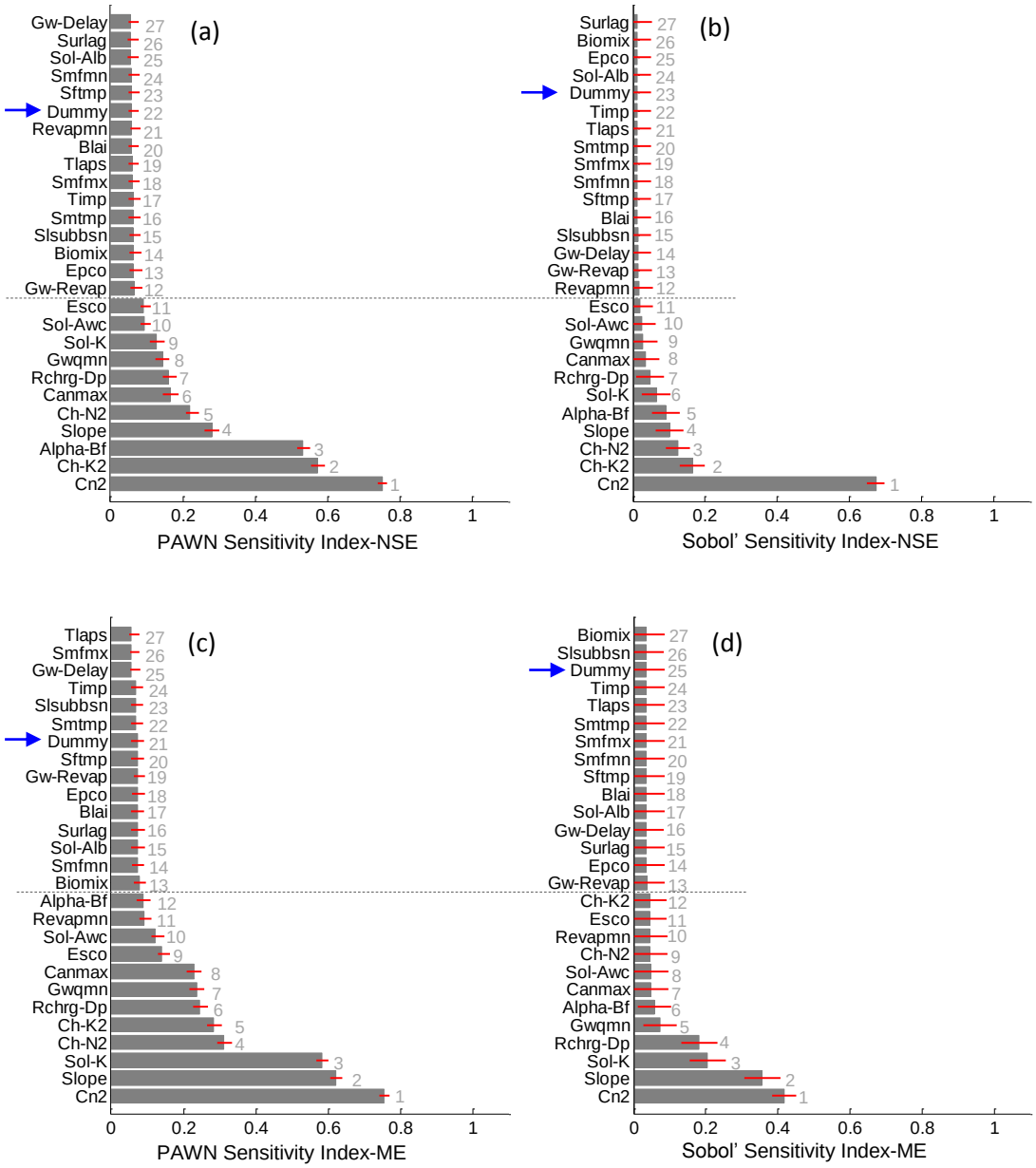

Figure 4. Applying the Sobol' and the PAWN methods results in different sensitivity indices for the SWAT model parameters using the NSE and ME performance measures. The numbers represent the parameter ranking obtained based on the sensitivity indices. The red lines represent 95% confidence intervals obtained by bootstrapping. (a) PAWN indices for NSE, (b) Sobol' indices for NSE, (c) PAWN indices for ME and (d) Sobol' indices for ME.

3.4. The two-sample Kolmogorov-Smirnov test for screening

As explained in section 2.2, when using the PAWN method, the two-sample Kolmogorov-Simonov test can be applied as a more formal approach to separate the influential and non-influential parameters (screening).

Figure 5 shows the test results for the NSE case. The *p*-values of the two-sample KS test for different conditioning values of the parameters are shown as circles. The red dashed line represents the significance level (α) of 5%. For the 9 lowest ranked parameters and for the dummy parameter, the *p*-value is larger than α at all conditioning values. Therefore, the null hypothesis of the test (i.e. the conditional and unconditional output distributions are the same) cannot be strongly rejected for those 9 parameters, indicating that they are non-influential. For the remaining 17 parameters, the *p*-value is smaller than α for at least one conditioning value and, thus, the test indicates that these parameters are influential. However, it should be noted that the test is performed with a significant level of 5%, which means that there is still 5% probability that a parameter is identified influential while it is not (null hypothesis is rejected while it is true). As shown in the figure, for 6 parameters (Timp, Smtmp, Slsubbsn, Biomix, Epco and Gw-revap), the *p*-value is smaller than α only for one (out of 10) conditioning value. Therefore, based on these results, it is difficult to strongly conclude that these 6 parameters are influential. On the other hand, for the top 11 parameters in Figure 5, the *p*-values are lower than 5% for most of the conditioning values, and consequently it is possible to strongly conclude that these top 11 parameters are influential. Interestingly, this group of influential parameters is the same as the one identified in section 3.3 using the “dummy parameter” approach (Figure 4(a)), which can be regarded as an indication of the validity of that approach.

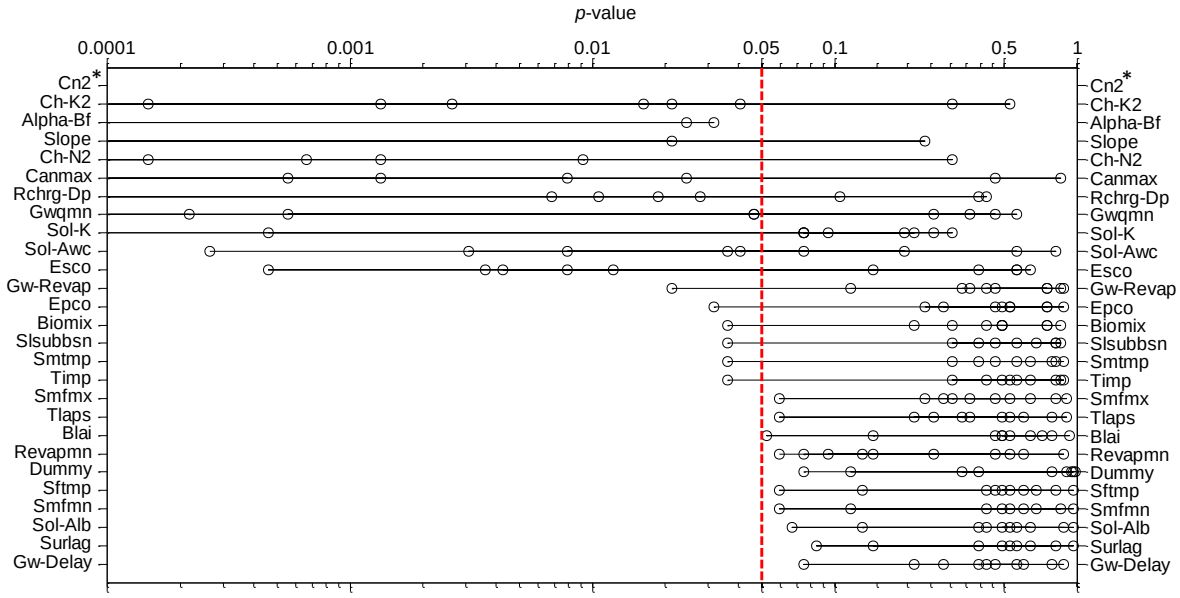

Figure 5. The *p*-values (circles) of the two-sample KS test for different conditioning values of the SWAT parameters, together with the dummy parameter, for the NSE performance measure. The red dashed line represents a significance level of 5% (α). A *p*-value smaller than α implies that the parameter is influential. The *p*-values of parameter Cn2 (marked by an asterisk in the figure) are not shown because they are lower than 0.0001 for all 10 different conditioning values.

The two-sample KS test is also performed for the ME performance measure and the results of the *p*-values are compared with the significance level of 5% (α) in Figure 6. Similar to the results for the NSE, Cn2 is the most significant parameter with *p*-values lower than 0.0001 for all 10 conditioning values (and, therefore, not visible in Figure 6).

As shown in Figure 6, the null hypothesis of the test is not rejected for the 3 lowest ranked parameters (*p*-values are larger than α for all conditioning values). These 3 parameters can thus be considered non-influential. For the remaining parameters, the *p*-value is smaller than α for at least one conditioning value (the null hypothesis is rejected), which indicates that these parameters are influential. The conclusion applies also to the dummy parameter, for which the *p*-value is lower than 0.05 for one conditioning value. However, we are sure that the dummy parameter has no effect on the model output. The reason for this unexpected result is that, as already explained above, the set-up of the test allows for Type I error (classifying a parameter as influential while it is not) with 5% probability. For the same reason, we cannot strongly conclude that all the other parameters with *p*-value lower than 0.05 at only one conditioning value are influential. On the other hand, the results of the test for the top 12 parameters strongly indicate that these parameters are influential. Similar to the results for the NSE, this group of influential parameters is the same as the one identified in Section 3.3, using the sensitivity index of the dummy parameter as a threshold for screening (Figure 4(c)).

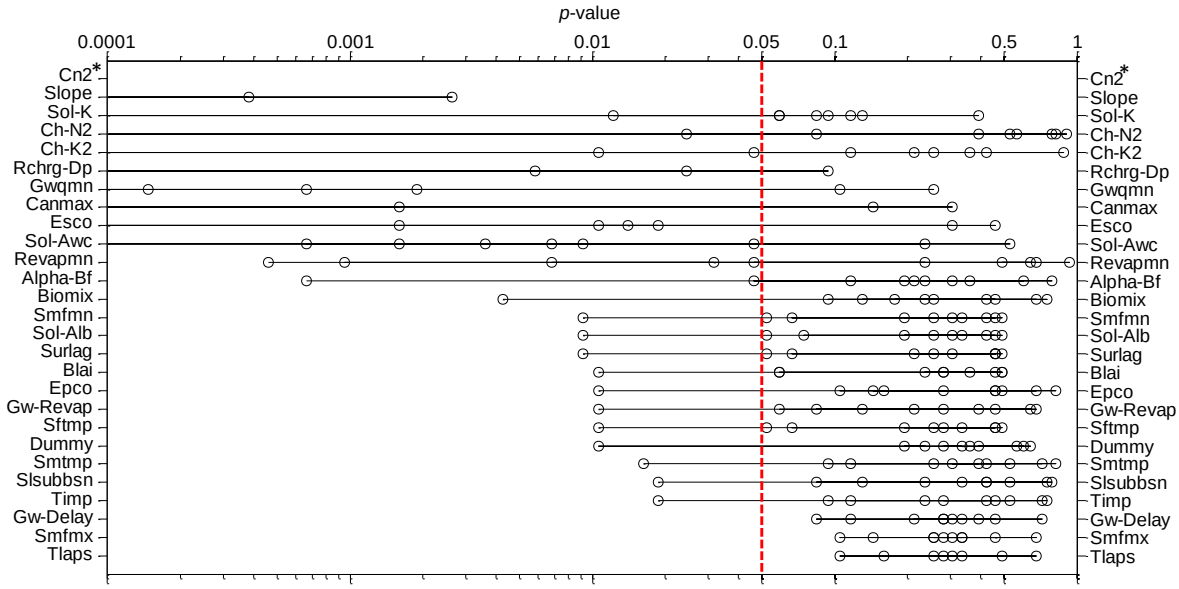

Figure 6. The *p*-values (circles) of the two-sample KS test for different conditioning values of the SWAT parameters, together with the dummy parameter, for the ME performance measure. The red dashed line represents a significance level of 5% (α). A *p*-value smaller than α implies that the parameter is influential. The *p*-values of parameter Cn2 (marked by an asterisk in the figure) are not shown because they are lower than 0.0001 for all 10 different conditioning values.

4. Discussion

The results of this study are used to compare the Sobol' and PAWN methods for the global sensitivity analysis to 26 parameters of the SWAT model. However, it should be noted that the comparison is performed for a specific case study (River Zenne, Belgium). Previous studies, for example Cibin et al., (2010), pointed out the effects of contrasting climate conditions and flow regimes on the Sobol' sensitivity analysis results of SWAT models. Moreover, many other choices made in the experimental set-up of GSA, including the choice of the parameters subject to GSA and their ranges, the selection of the simulation period and of the scalar output, can strongly influence GSA results (Pappenberger et al., 2008; Shin et al., 2013).

A visual analysis of the evolution of sensitivity indices and associated parameter rankings with increasing sample size shows that the parameters with the highest sensitivity indices converge quickly to their final sensitivity indices values for both Sobol' and PAWN. The quick convergence of the Sobol' total sensitivity indices for the most influential SWAT parameters was also reported by Nossent et al. (2011). Similar to the sensitivity indices, the parameter rankings of the top-ranked parameters converged to their final ranks quickly in both methods, even with a limited sample size. As pointed out by Nossent et al. (2011), for the Sobol' method, a sample size of 2000 was enough for the significant parameters to attain their final rank. For the PAWN method, according to results of our analysis, a sample size of 200 was actually sufficient to obtain stable parameter ranking for the top-ranked parameters. These results are also

consistent with the findings reported by Sarrazin et al. (2016), which shows that stable parameter ranking and screening can be obtained at significantly lower sample size (i.e. lower number of model evaluations) than stable estimates of the sensitivity indices. This can be an advantage for both the Sobol' and PAWN methods, as the stable parameter ranking and parameter screening can be obtained with a limited computational cost.

Overall, ranking and screening results of Sobol' and PAWN are very consistent for both the considered performance measures (NSE and ME). However, for both the NSE and ME, the difference in value between the PAWN indices of the influential parameters is more marked, as compared to Sobol', where all influential parameters are associated with an almost the same sensitivity index value. This difference between PAWN and Sobol' may be related to the fact that the distributions of the performance measures are rather skewed (NSE) and slightly bi-modal (ME), which limits the ability of the Sobol' method to properly quantify the relative influence of parameters on the model output. As discussed in the Introduction, the effectiveness of variance-based methods, such as Sobol', depends on the level of symmetry of the output distribution (Borgonovo, 2007), and variance-based sensitivities become less reliable for highly skewed or multi-model distributions (Liu et al., 2006; Pianosi and Wagener, 2015).

According to both the Sobol' and PAWN indices, the curve number (Cn2) is the most influential parameter for both the NSE and ME performance measures. Actually, Cn2 has been reported as an important parameter affecting flow simulation in all of the SWAT applications (Gassman et al., 2007; Cibin et al., 2010; Nossent et al, 2011; Leta et al., 2015). In general, the SWAT parameters identified as influential by both Sobol' and PAWN are almost the same as those of previous GSA applications to this study area and similar catchments (Leta et al., 2015; Nossent at al., 2011).

Comparing GSA results for different performance measures (NSE and ME), we found that the selection of the performance measure as scalar model output affects the parameter ranking but not the parameter screening. Overall, this choice seem to be less crucial than in other applications of GSA, where even smaller differences in the definition of the performance metric (for example, considering the mean of the squared errors or the mean of the absolute errors) significantly affected Sobol' results (Pappenberger et al. (2008)).

Finally, we found that our proposed "dummy parameter" approach is indeed an easy-to-implement and effective way to set a screening threshold. Application of such approach to both Sobol' and PAWN results provided the same separation of influential and non-influential parameters. The statistical two-sample KS test was also applied in PAWN, which confirmed that the top 11 ranked parameters for the NSE performance measure are strongly influential. However, the application of the two-sample KS test also illustrates the possibility of occurrence of Type I errors (coherently with the chosen confidence level), thus highlighting the statistical nature of the test and hence the need for interpreting its results coherently with such statistical nature.

5. Conclusions

In this paper, we compared the application of two GSA techniques, the variance-based Sobol' method and the density-based PAWN method, to the analysis of 26 parameters of the SWAT model, a hydrological model widely-used for water quality and quantity simulations. The comparison was performed in terms of convergence rate and parameter ranking and screening results. Moreover, the use of a "dummy parameter" approach as a viable option to set a threshold value for parameter screening was demonstrated for both Sobol' and PAWN.

Considering the results, there was no difference between the Sobol' and PAWN methods in term of convergence rate and screening. Both methods can identify the set of 14 (or 15, depending on the performance metric) non-influential parameters with a relatively limited number of model evaluations. Therefore, they are equally useful for informing about which parameters could be excluded from computationally expensive automatic calibration. However, in terms of parameter ranking, the difference between the relative importance of the influential parameters was better quantified by the PAWN method, as compared to Sobol'. One possible explanation for this is that the distributions of the model outputs (i.e. NSE and ME) were non-symmetric, undermining Sobol' implicit assumption that variance is a good proxy for output uncertainty. We think these findings are encouraging towards promoting PAWN as an alternative method for GSA of environmental models.

Since all the above results were obtained for a specific model, case study and GSA set-up, further research is needed to investigate how transferable our conclusions are to other models and applications. Moreover, we would like to highlight that in this paper we presented Sobol' and PAWN as alternative methods. In fact, the numerical approximation of the respective indices requires a tailored sampling strategy, and therefore the application of PAWN after Sobol' (or vice versa) require running the model thousands or even hundred thousands more times, which would be unfeasible for many time-consuming simulation models. However, ongoing research (e.g. Strong and Oakley, 2013, Plischke et al, 2013, Pianosi et al., 2016) is aiming at developing new approximation strategies to compute Sobol' and PAWN indices from a single output sample. Once established, these strategies will open up the possibility of applying both methods to the same set of model evaluations, and in general of applying multiple GSA methods at the same computational cost as individual GSA methods (Pianosi et al., 2016). Looking forward, we thus think that variance-based and density-based approaches can be regarded as complimentary approaches that will be applied in the future to investigate model output sensitivities from different and complimentary angles.

Acknowledgment

The authors would like to thank the Flanders Hydraulics Research for supporting and coordinating the project of "Development of conceptual models for an integrated river basin management". We also thank Dr. Olkeba Tolessa Leta for setting up the SWAT model for the River Zenne (Belgium). This work is partially supported by a University of Bristol Alumni Postgraduate Scholarship to Fanny Sarrazin. Partial support for Francesca Pianosi and Thorsten Wagener was provided by the Natural Environment Research Council (Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1).

Appendix .Supplementary Materials

References

Anderson, T. W. and Darling, D. A. (1952). A symptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics, 193-212.

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

Arnold, J. G., Srinivasan, R., Muttiah, R. S. and Williams, J. R. (1998). Large area hydrologic modeling and assessment part I: Model development1. *Journal of the American Water Resources Association*, 34(1), 73-89.

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B. and Neitsch, S.L. (2011). Soil and Water Assessment Tool. Input/output File Documentation, Version 2009. Texas Water Resources Institute

Baroni, G. and Tarantola, S. (2014). A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environmental Modelling & Software*, 51, 26-34.

Baucells, M., and Borgonovo, E. (2013). Invariant probabilistic sensitivity analysis. *Management Science*, 59(11), 2536-2549.

Bekele, E. G. and Nicklow, J. W. (2007). Multi-objective automatic calibration of SWAT using NSGA-II. *Journal of Hydrology*, 341(3), 165-176.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6), 771-784.

Borgonovo, E., Castaings, W. and Tarantola, S. (2011). Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, 31(3), 404-428.

Borgonovo, E., Tarantola, S., Plischke, E. and Morris, M. D. (2014). Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(5), 925-947.

Borgonovo, E., Hazen, G. B. And Plischke, E. (2016). A common rationale for global sensitivity measures and their estimation. *Risk Analysis,* 1-24.

Bressiani, D. D. A., Gassman, P. W., Fernandes, J. G., Garbossa, L. H. P., Srinivasan, R., Bonumá, N. B. and Mendiondo, E. M. (2015). Review of Soil and Water Assessment Tool (SWAT) applications in Brazil: Challenges and prospects. *International Journal of Agricultural and Biological Engineering*, 8(3), 9-35.

Cibin, R., Sudheer, K. P. and Chaubey, I. (2010). Sensitivity and identifiability of stream flow generation parameters of the SWAT model. *Hydrological processes*, 24(9), 1133-1148.

DHI., 2011. MIKE 11 Water Quality (Ecolab) Reference manual. *DHI Water & Environment.*

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water resources research*, 28(4), 1015-1031.

Efron, B. and Tibshirani, R. J. (1994). An introduction to the bootstrap.CRC press.

Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., ..., and Di, Z. (2014). A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environmental Modelling & Software*, 51, 269-285.

Gassman, P. W., Reyes, M. R., Green, C. H., & Arnold, J. G. (2007). The soil and water assessment tool: historical development, applications, and future research directions. *Transactions of the ASABE*, 50(4), 1211-1250.

Gassman, P. W., Arnold, J. J., Srinivasan, R. and Reyes, M. (2010). The worldwide use of the SWAT Model: Technological drivers, networking impacts, and simulation trends. In *21st Century Watershed Technology: Improving Water Quality and Environment Conference Proceedings, 21-24 February 2010, Universidad EARTH, Costa Rica* (p. 1). American Society of Agricultural and Biological Engineers.

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

Helton, J. C. and Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81(1), 23-69.

Hill, M. C. and Tiedeman, C. R. (2006). *Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty*. John Wiley & Sons.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1-17.

Jarque, C. M., and Bera. A. K.(1987). A Test for Normality of Observations and Regression Residuals.*International Statistical Review*, 55(2), 163–172.

Kepner, W. G., Semmens, D. J., Bassett, S. D., Mouat, D. A. and Goodrich, D. C. (2004). Scenario analysis for the San Pedro River, analyzing hydrological consequences of a future environment. *Environmental Monitoring and Assessment*, *94*(1-3), 115-127.

Kolmogorov, A. (1933). *Sulla determinazione empirica di una legge di distribuzione*. *Giornale Dell'istituto Italiano Degli Attuari,* 4, 83–91.

Leta, O. T. (2013). Catchment Processes Modeling, Including the Assessment of Different Sources of Uncertainty, Using the SWAT Model: the River Zenne Basin (Belgium) Case Study. PhD Thesis, Vrije Universiteit Brussel, Brussels, Belgium, 283 pp.

Leta, O. T., Nossent, J., Velez, C., Shrestha, N. K., van Griensven, A. and Bauwens, W. (2015). Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). *Environmental Modelling and Software*, 68, 129-146.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.

Liu, H., Chen, W. and Sudjianto, A. (2006). Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design*, 128(2), 326-336.

Marsaglia, G., Tsang, W. and Wang, J. (2003). Evaluating Kolmogorov's Distribution. Journal of Statistical Software. 8(18).

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. 46(253) 68–78.

Moore, C., and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resources Research*, *41*(5)

Muleta, M. K. and Nicklow, J. W. (2005). Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. *Journal of Hydrology*, 306(1), 127-145.

Nash, J., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, 10(3), 282-290.

Norton, J. (2015). An introduction to sensitivity assessment of simulation models. *Environmental Modelling & Software*, *69*, 166-174.

Nossent, J., Elsen, P. and Bauwens, W. (2011). Sobol' sensitivity analysis of a complex environmental model. *Environmental Modelling & Software*, 26(12), 1515-1525.

Nossent, J. and 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 Science and Technology, 65(3), 539-549.

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

Pettitt, A. N. (1976). A two-sample Anderson-Darling rank statistic. Biometrika, 63(1), 161-168.

Pianosi, F. and Wagener, T. (2015). PAWN: a simple and efficient method for Global Sensitivity Analysis based on cumulative distribution functions. *Environmental Modelling & Software*, 67, 1-11.

Pianosi, F., Sarrazin, F. and Wagener, T. (2015). A Matlab toolbox for global sensitivity analysis. Environmental Modelling & Software, 70, 80-85.

Pianosi, F., Beven, K., Freer, J.W. Hall, J. Rougier, J. Stephenson, D.B., Wagener, T. (2016), Sensitivity analysis of environmental models: A systematic review with practical workflow, *Environmental Modelling& Software*, 79, 214-232.

Pianosi, Iwema, Rosolem and Wagener (2016), A Multi-method Global Sensitivity Analysis Approach to Support the Calibration and Evaluation of Land Surface Models, In: Sensitivity Analysis in Earth Observation Modelling. Edited by Petropoulos et al., Elsevier

Plischke, E., Borgonovo, E. and Smith, C. L. (2013). Global sensitivity measures from given data, Eur. J. Oper. Res., 226, 536-550,

Rabitz, H. and Alis, O. F. (2000). Managing the tyranny of parameters in mathematical modelling of physical systems. *Sensitivity Analysis, Saltelli A, Chan K, Scott M Editors*, 385-397.

Rossman, L.A., 2009. Storm Water Management Model user's manual, Version 5.0.

Rosolem, R., Gupta, H. V., Shuttleworth, W. J., Zeng, X. and Gonçalves, L. G. G. (2012). A fully multiple-criteria implementation of the Sobol′ method for parameter sensitivity analysis. *Journal of Geophysical Research: Atmospheres* (1984–2012), 117(D7).

Saltelli, A., Chan, K. and Scott, E. M. (Eds.) (2000). *Sensitivity analysis* (Vol. 134).New York, Wiley.

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

Saltelli, A. (2002b). Sensitivity analysis for importance assessment. *Risk Analysis*, 22(3), 579-590.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). Global sensitivity analysis: the primer. John Wiley and Sons.

Sarrazin, F., Pianosi, F. and Wagener, T. (2016). Global sensitivity analysis of environmental models: Convergence and validation. *Environmental Modelling & Software*, 79, 135-152.

Shin, M. J., Guillaume, J. H., Croke, B. F. and Jakeman, A. J. (2013). Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R. *Journal of Hydrology*, 503, 135-152.

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *BulletinMath´ematique de l'Universit´e de Moscou*, 2 (2).

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 279-281.

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational mathematics and mathematical physics*, (7), 86-112.

Sobol', I. M. (1976). Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics*, 16(5), 236-242.

Sobol', I. M. (1990).On sensitivity estimation for nonlinear mathematical models. *MatematicheskoeModelirovanie*, 2(1), 112-118.

Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, 55(1), 271-280.

Shrestha, N. K., Leta, O. T., De Fraine, B., Van Griensven, A. and Bauwens, W. (2013). OpenMI-based integrated sediment transport modelling of the river Zenne, Belgium. *Environmental Modelling & Software*, 47, 193-206.

M. Strong and J. E. Oakley, (2013) "An efficient method for computing partial expected value of perfect information for correlated inputs," Med. Decis., 33, 755-766.

Tang, Y., Reed, P., Wagener, T. and Van Werkhoven, K. (2007). Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrology and Earth System Sciences Discussions*, 11(2), 793-817.

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

van Werkhoven, K., Wagener, T., Reed, P. and Tang, Y. (2008). Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resources Research*, *44*(1).

van Werkhoven, K., Wagener, T., Reed, P. and Tang, Y. (2009). Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Advances in Water Resources*, *32*(8), 1154-1169.

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

van Griensven, A., Ndomba, P., Yalew, S. and Kilonzo, F. (2012). Critical review of SWAT applications in the upper Nile basin countries. *Hydrology and Earth System Sciences*, 16(9), 3371-3381.

Vrugt, J. A., Bouten, W., Gupta, H. V. and Sorooshian, S. (2002). Toward improved identifiability of hydrologic model parameters: The information content of experimental data. *Water Resources Research*, *38*(12), 48-1.

Vrugt, J. A., Gupta, H. V., Bouten, W. and Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*,*39*(8).

Yang, J. (2011). Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. *Environmental Modelling & Software*, *26*(4), 444-457.

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

Zhang, X., Srinivasan, R., Zhao, K. and Liew, M. V. (2009). Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrological Processes*, *23*(3), 430-441.