Pianosi, F., & Wagener, T. (2018). Distribution-based sensitivity analysis from a generic input-output sample. *Environmental Modelling and Software*, *108*, 197-207. https://doi.org/10.1016/j.envsoft.2018.07.019

Distribution-based sensitivity analysis from a generic input-output sample

Francesca Pianosi [a,b], Thorsten Wagener [a,b]
(corresponding author: francesca.pianosi@bristol.ac.uk)

[a] *Department of Civil Engineering, University of Bristol, University Walk, BS81TR, Bristol, UK*

[b] *Cabot Institute, University of Bristol, Royal Fort House, BS81UJ, Bristol, UK*

Abstract

In a previous paper we introduced a distribution-based method for Global Sensitivity Analysis (GSA), called PAWN, which uses cumulative distribution functions of model outputs to assess their sensitivity to the model's uncertain input factors. Over the last three years, PAWN has been employed in the environmental modelling field as a useful alternative or complement to more established variance-based methods. However, a major limitation of PAWN up to now was the need for a tailored sampling strategy to approximate the sensitivity indices. Furthermore, this strategy required three tuning parameters whose optimal choice was rather unclear. In this paper, we present an alternative approximation procedure that tackles both issues and makes PAWN applicable to a generic sample of inputs and outputs while requiring only one tuning parameter. The new implementation therefore allows the user to estimate PAWN indices as complementary metrics in multi-method GSA applications without additional computational cost.

Key words: global sensitivity analysis; distribution-based methods; moment-independent methods; multi-method GSA

Highlights

- We introduce a new approximation strategy for PAWN indices
- The strategy is applicable to a generic input-output sample and uses one tuning parameter only
- We demonstrate that the strategy provides robust PAWN sensitivity estimates
- This approximation strategy facilitates the integration of PAWN into multi-method GSA

Software availability

The PAWN algorithm, including the new approximation strategy presented in this paper, are implemented in Matlab/Octave as part of the SAFE Toolbox, which is freely available for non-commercial use through the website: www.safetoolbox.info

1 Introduction

Global Sensitivity Analysis (GSA) is a set of techniques aimed at investigating the propagation of uncertainty through mathematical models in a structured way. More specifically, according to the widely used definition by Saltelli et al. (2008), the aim of GSA is to quantify the relative contribution of the uncertain input factors of a mathematical model to the variability of its outputs. For model developers, such quantification can aid the process of identifying a minimum complexity model by eliminating non-influential components. For model users, it can make the calibration process more efficient by determining the subset of model parameters whose reduction in uncertainty would mostly reduce output variability, or it can be used to assess the robustness of the model predictions against various sources of uncertainty such as errors in the forcing data or even in uncertain modelling assumptions. GSA is therefore widely applied in the environmental modelling field to support the construction, improvement and use of mathematical models (e.g. Beven and Binley (1992); Spear et al. (1994); Freer et al. (1996); Bastidas et al. (1999); Wagener and Kollat (2007); Norton (2015); Razavi and Gupta (2015); Xiaomeng et al. (2015); Ferretti et al. (2016); Pianosi et al. (2016); Petropoulos and Srivastava (2017)).

Many different GSA methods are available depending on the specific purposes of the analysis as well as the characteristics of the mathematical model being analysed and its sources of uncertainty (Saltelli et al., 2008; Norton, 2015; Pianosi et al., 2016). Among them, some of the most widely used are Variance-Based Sensitivity Analysis (VBSA) methods, which measure output sensitivity as the proportion of output variance that is attributable to variations of each uncertain input factor. For an overview of variance-based methods and their advantages see for example Saltelli et al. (2008) or Pianosi et al. (2016). Recently, density-based approaches have also gained increasing attention Castaings et al. (2012); Anderson et al. (2014); Peeters et al. (2014); Dell'Oca et al. (2017); Borgonovo et al. (2017). In these approaches, uncertainty and sensitivity is characterised by investigating the entire distribution of the model outputs, instead of its variance only. For this reason, such methods are also referred to as *distribution-based* or *moment-independent*. Distribution-based strategies are particularly suitable when variance is not an adequate proxy of uncertainty, for example when the output distribution is highly-skewed or multi-modal (e.g. Liu et al. (2006)).

In a previous paper (Pianosi and Wagener, 2015) we introduced a distribution-based method, called PAWN, and implemented it as part of our open-source GSA Toolbox called SAFE (Pianosi et al., 2015). The advantage of PAWN over other moment-independent methods is that it characterises output distributions by their cumulative distribution functions, instead of their probability density functions, which makes the numerical approximation of PAWN sensitivity indices easy and robust. In Pianosi and Wagener (2015) we demonstrated the PAWN method by applying it to a standard benchmark function and a simple rainfall-runoff model (Hymod). In Zadeh et al. (2017) we carried out a systematic comparison between PAWN and variance-based method on a medium complexity (26 parameters) hydrological model (SWAT) and found that PAWN and VBSA had similar convergence rate and screening results, while PAWN was more effective for parameter ranking as it could better separate out the relative importance of the influential parameters. Since its publication, PAWN has been used to investigate the role of uncertain parameters across a range of environmental modelling fields, including: a transport model of indoor air pollutant (Sedighian et al., 2015), a computational model of biological processes (Gillies et al., 2016), rainfall-runoff and land-surface models in Pianosi and Wagener (2016) and Pianosi et al. (2017), a fluid flow and heat transport model in geothermal reservoirs (Fox et al., 2016), a groundwater model for karst systems (Hosseini et al., 2017), and a numerical algorithm for hillslope-based landscape discretisation (Pilz et al., 2017).

Despite this relatively quick uptake of PAWN across different fields of application, from our own experience and the feedbacks we received from other users, we think two main issues remain critical. First, the numerical procedure we proposed in our original paper to implement PAWN uses a *tailored* sampling strategy, i.e. a strategy that selects input samples in specific regions of the input variability space, according to the approximation procedure set out for the PAWN indices. This is in contrast to *generic* sampling strategies, such as sampling over a uniform grid, quasi-random sampling (Press et al., 1992) or (stratified or not) random sampling, e.g. Latin Hypercube (Forrester et al., 2008), which aim at spreading input samples as uniformly as possible across the variability space, and can be used across a range of uncertainty and sensitivity analysis methods. The requirement of a tailored sampling strategy thus makes it more difficult to integrate PAWN into a multi-method GSA study, such as Pappenberger et al. (2008) or Tang et al. (2007), since its inclusion would require additional dedicated model evaluations. We believe that this is a strong limitation given the value of applying multiple GSA methods to the same problem as a way to validate and complement the results of individual methods (Pianosi et al., 2015; Borgonovo et al., 2017). Additionally, the requirement of a tailored sampling strategy prevents the application to an existing input-output dataset in cases where such a dataset is available from previous studies. These limitations have motivated researchers to seek for generic approximation strategies for other GSA methods too, including variance-based methods. For example, Strong et al. (2014) and Stanfill et al. (2015) have proposed new approximation strategies to derive first-order and total-order indices from a generic input-output dataset, as an alternative to the 'traditional' approximators (e.g. Saltelli et al. (2010)) based on 're-sample' matrices, which require a tailored sampling strategy. A general discussion of the value of approximation procedures that can be applied to given data is given in Plischke et al. (2013).

The second issue with our tailored sampling strategy is that it requires users to specify three tuning parameters, i.e. the number of unconditional output samples (N_u), the number of conditional output samples (N_c), and the number of conditioning points (n). As discussed in Pianosi and Wagener (2015), the choice of these tuning parameters should be based on a compromise between approximation accuracy and computational burden, which both increase with any increase of N_u or N_c or n. In fact, the total number of model evaluations to approximate all PAWN indices is $N = N_u + n \times N_c \times M$, where M is the number of uncertain input factors. If each model evaluation is computationally demanding, either in terms of running time or data storage requirement, one would like to find the 'optimal' combination of (N_u,N_c,n) to reach sufficient approximation accuracy at minimum N. However, such optimal values are difficult to predict a priori and extrapolating from previous applications may be risky because the optimal values may change with the problem at hand, i.e. with the mathematical model, the number of input factors, and possibly even with the output definition or application domain (Sarrazin et al., 2016). We indeed know that the approximation accuracy associated with sensitivity indices at a given sample size can dramatically change with any element of the experimental set-up, as shown for example in Figure 5 in Pianosi et al. (2016) or Figures 2 and 3 in Zadeh et al. (2017).

In this paper we simultaneously address these two issues by introducing a new approximation procedure of the PAWN indices that (1) is applicable to a generic dataset; (2) requires fewer tuning parameters (essentially only the number of conditioning points n) whose choice is easier to make and to evaluate. The approximation procedure was already sketched out in the conclusions of Pianosi and Wagener (2015) and a similar idea was tested in Pianosi et al. (2017). Here we further develop those ideas into a new approximation procedure. We test it comprehensively on a benchmark function and on a complex hydrological model (the Soil Water Assessment Tool, in a set-up that includes 50 uncertain parameters). And finally, we propose a number of simple tools to assess the accuracy of the resulting PAWN indices as well as their robustness to the chosen tuning parameter, at negligible additional computing costs.

2 Methods

In this paper, we consider an input-output relationship

$$y = f(\mathbf{x}) \tag{1}$$

where $\mathbf{x} = |x_1, x_2, ..., x_M| \in \mathcal{X} \subseteq \mathbb{R}^M$ is a vector of M input factors and $y \in \mathbb{R}$ is a (scalar) output variable. The goal of GSA is to quantify the relative contribution of variations in each input factor x_i to the variability of the output y. A quantitative measure of such relative contribution is expressed by the value of a sensitivity index S_i, typically ranging from 0 to 1.

The function f can be available either in closed form or as a numerical procedure to compute y given $\mathbf{x}$. For example, in typical environmental modelling applications the function f typically refers to the numerical procedure for simulating a dynamical system over a given space-time domain. In this case, the output y is a scalar variable that summarises the wide range of variables (often time series, possibly spatially-distributed) provided by the simulation procedure. For example y may be the value of a simulated variable at a time and location of interest, or an aggregate measure of the mismatch between some of the simulated variables and their observations, i.e. an objective or loss function.

When the input-output relationship f is available in closed form (as in the example of Sec. 4.1), it is often referred to as a *model*. When instead it refers to the simulation procedure to compute y from $\mathbf{x}$ (as in Sec. 4.2), it is often referred to as a *response surface*, to avoid confusion with the underlying set of differential equations, which is also called a (simulation) model. Notice that in the latter case, the underlying simulation model might have more inputs than those included in $\mathbf{x}$ and the output y may be defined in different ways. The choice of which variables to include in $\mathbf{x}$ and of how to define one (or multiple) y depends on the underlying motivation for carrying out GSA, and as such it is a subjective choice of the GSA user and will not be discussed here.

2.1 The PAWN method

The key idea of distribution-based methods is that the influence of an input factor is proportional to the amount of change in the output distribution produced by fixing that input. More precisely, the sensitivity of y to x_i is measured by the difference between the unconditional distribution of y, which is induced by varying all input factors simultaneously, and the conditional distribution that is obtained by varying all inputs but x_i.

A review of several distribution-based methods is given in Pianosi and Wagener (2015). The distinctive feature of PAWN is that, in contrast to other methods, it uses (conditional and unconditional) cumulative distribution functions (CDFs) of the output instead of probability density functions. The advantage of using CDFs is that their approximation from an output sample of finite size is easy and robust. Several other advantages of PAWN are discussed in Pianosi and Wagener (2015).

The PAWN sensitivity index for the i-th input factor is defined as

$$S_i = \underset{x_i}{\text{stat}} \max_{y} |F_y(y) - F_{y|x_i}(y|x_i)| \tag{2}$$

where $F_y(y)$ and $F_{y|x_i}(y|x_i)$ are the unconditional and conditional CDFs of the output y, and *stat* is a statistic (e.g. maximum, median or mean) defined by the user. Notice that the inner maximum in Eq. (2), i.e. the maximum absolute difference between CDFs, is no other than the Kolmogorov-Smirnov (KS) statistic, which is widely used as a measure of distance between CDFs (Kolmogorov, 1933; Smirnov, 1939). The PAWN index can thus be reformulated as

$$S_i = \underset{x_i}{\text{stat}}\ \text{KS}(x_i) \quad \text{where} \quad \text{KS}(x_i) = \max_{y} |F_y(y) - F_{y|x_i}(y|x_i)| \tag{3}$$

Other statistics could possibly be used instead of KS. For example, Zadeh et al. (2017) tested the Anderson-Darling statistic and found that, in their application, it provides almost identical sensitivity results as the KS (these results are shown in their Supplementary material). Throughout this paper we will use KS, however our newly proposed approximation strategy could be equally applied to PAWN indices defined using other statistics.

2.2 Approximating PAWN indices using a tailored sampling strategy

In general, given the complexity of the input-output relationship f, the sensitivity indices of Eq. (2) cannot be computed analytically and they need to be approximated numerically. Pianosi and Wagener (2015) proposed an approximation procedure based on two simplifications. First, using a finite number of conditioning points $\bar{x}_i^{(1)}$, $\bar{x}_i^{(2)}$, ..., $\bar{x}_i^{(n)}$ for each input factor, instead of all its possible values. Second, replacing the distributions F_y and $F_{y|x_i}$ by the empirical distributions $\hat{F}_y$ and $\hat{F}_{y|x_i}$ of output samples of finite size. Specifically, $\hat{F}_y$ is the empirical distribution of an unconditional sample (YU) obtained by varying all input factors simultaneously, and $\hat{F}_{y|x_i}$ is the empirical distribution of a conditional sample (YC_{ik}) obtained by varying all factors but the i-th, which is set to the k-th conditioning value $\bar{x}_i^{(k)}$. The PAWN sensitivity index is then approximated by

$$\hat{S}_i = \underset{k=1,...,n}{\text{stat}} \text{KS}(\bar{x}_i^{(k)}) \quad \text{where} \quad \text{KS}(\bar{x}_i^{(k)}) = \max_y |\hat{F}_y(y) - \hat{F}_{y|x_i}(y|x_i = \bar{x}_i^{(k)})| \quad (4)$$

The left hand side of Figure 1 provides a visual illustration of this approximation strategy for the simple case of M=3 input factors. For the sake of illustration, the Figure focuses on approximating the PAWN sensitivity index of the first input factor (x_1). The top left panels (Fig. 1(a) and (b)) show the combinations of input factors (x_1,x_2,x_3) that need to be evaluated in order to obtain the unconditional sample and three conditional samples corresponding to n=3 fixed values of x_1. The corresponding output samples (YU, YC_{11}, YC_{12}, YC_{13}) are visualised via a scatter plot in Fig. 1(c). The lower panels show the further steps for computing the approximate PAWN index $\hat{S}_1$: computing the empirical distributions of YU (red line in (g)) and of YC_{11}, YC_{12} and YC_{13} (grey lines), computing the KS at each conditioning point (h), and finally taking a statistic, e.g. the median, of those KS values. A similar procedure would be applied for approximating the sensitivity indices of x_2 and x_3.

We call the approach underpinning Eq. (4) a *tailored sampling strategy* because a large part of the input samples generated to compute the sensitivity indices, namely all those in the conditional samples YC_{ik} for $k = 1, ..., n$, are concentrated in specific subregions of the input variability space (e.g. the planes in Fig.1(b) where the grey circles lie). This is in contrast to *generic sampling strategies* that would spread input samples as evenly as possible across the input space (e.g. the samples in Fig. 1(d)-(e)). Examples of generic sampling strategies include latin hypercube sampling (e.g. Sec. 1.4 in Forrester et al. (2008)) or quasi-random sampling (e.g. Sec. 7.7 in Press et al. (1992)). Notice that while the input samples in YC_{ik} may be generated by applying a generic sampling strategy in the $(M-1)$-dimensional space of all-inputs-but-the-i-th (for instance, we will use latin hypercube sampling in the following case study applications), collectively the ensemble of conditional samples YC_{ik} does not constitute a *generic* dataset in the M-dimensional input variability space, as clearly shown in Fig. 1(b).

With the tailored sampling strategy, the total number of model evaluations to approximate all PAWN sensitivity indices is $N_u + n \times N_c \times M$, where N_u is the size of the unconditional sample YU, N_c is the size of each conditional sample YC_{ik}, and M is the number of input factors (and hence sensitivity indices). As discussed in the Introduction, the issue of how to choose the triple (N_u,n,N_c) has not been formally investigated and it remains an open question in the application of PAWN. This choice is critical given that it affects both the accuracy of the PAWN indices and the computational effort (total number of model evaluations) to generate them. Another issue with the tailored strategy is that much of the computational effort is invested in generating the conditional samples YC_{ik}, which cannot be re-used in other uncertainty or sensitivity analysis methods that would require a generic sample. To overcome these two issues we present a novel approach to approximate PAWN indices from a generic dataset in the next section.

2.3 Approximating PAWN indices from a generic dataset

So how can we approximate the sensitivity index in Eq. (2) using a generic input-output dataset <X,Y>, for example a dataset generated by Latin hypercube sampling? A possible way to do this is to split the range of variation of each input factor x_i into n equally spaced intervals $\mathcal{I}_k$ and define the conditional samples YC_{ik} accordingly. The unconditional sample YU could instead coincide with the entire sample Y or with a subsample of it. Such a strategy corresponds to approximate PAWN sensitivity indices as:

$$\hat{S}_i = \underset{k=1,\ldots,n}{\mathrm{stat}} \mathrm{KS}(\mathcal{I}_k) \quad \text{where} \quad \mathrm{KS}(\mathcal{I}_k) = \max_{y} |\hat{F}_y(y) - \hat{F}_{y|x_i}(y|x_i \in \mathcal{I}_k)| \qquad (5)$$

A visual illustration of the splitting strategy for creating unconditional and conditional samples from a generic dataset is given on the right hand side of Figure 1 ((d) and (e)). Once the output samples have been built (Fig 1(f)), the subsequent steps for approximating PAWN sensitivity indices are the same than when the tailored sampling strategy is used. A summary comparison of the workflows underpinning Eq. (4) and (5) is given in Figure 2.

When using the approximation strategy of Eq. (5), the size of the conditional sample (N_c) does not need to be specified by the user: it simply coincides with the number of points in each interval $\mathcal{I}_k$. However, if input samples are uniformly spread in the given dataset we may expect N_c to be approximately equal to N/n, where N is the size of the generic dataset. Therefore, the user can indirectly control the value of N_c by choosing n: a reduction in n would increase N_c and vice versa. As for the unconditional sample, one option is to let it coincide with the sample Y. This choice would correspond to setting $N_u = N$. Another option is to use a subsample of Y, for example by randomly extracting a subsample of the same size as the conditional ones (i.e. setting $N_u = N_c$). The latter option has the advantage that the compared unconditional and conditional distributions are estimated from the same number of samples. Furthermore, the random extraction from Y can be repeated several times using bootstrapping without replacement as a way to test the robustness of the PAWN sensitivity indices, as will be further described in the next subsection.

In either case, the main point here is that both N_c and N_u do not need to be chosen by the user but they are determined as a consequence of the chosen value of n and N. This is an advantage with respect to the tailored sampling approach because the number of tuning parameters is reduced to two (instead of three) but most importantly because selecting their values is much easier. In fact, when using a generic dataset the computational effort for approximating the PAWN sensitivity indices is fully controlled by the chosen value for N. Hence, an obvious choice is to take the largest value possible for N given available computing resources. As for n, it only has an effect on the splitting of the input-output dataset <X,Y> but not on its generation. Consequently, one can attempt different values of n and evaluate the robustness of GSA results to this choice without significantly adding to the overall computational effort. Further ways to assess the robustness of PAWN sensitivity indices to the chosen sample size N are discussed in the next subsection and will be illustrated in the Results section.

2.4 Estimating the robustness of PAWN indices

When sensitivity indices are computed by an approximate formula such as Eq. (4) or (5), it is very important to evaluate the robustness of the sensitivity values to the chosen sample, particularly if the sample size is small. A computationally efficient way to do this is by repeating sensitivity calculations using different bootstrap resamples (Efron and Tibshirani, 1993) of the available input-output dataset to obtain a distribution of sensitivity indices. The mean of such distributions can be taken as a more robust estimate of the sensitivity indices (at least more robust than the point estimates obtained without bootstrapping) and quantiles can be computed to derive confidence intervals around those estimates (Yang, 2011; Sarrazin et al., 2016).

Additionally, the impact of approximation errors on sensitivity indices can be directly inferred by using a so called *dummy parameter* (Zadeh et al., 2017). A dummy parameter is an input factor that is artificially introduced in the analysis and that, by definition, cannot affect the output variability. However, the sensitivity index of the dummy parameter may still be larger than zero, because of errors in the employed approximation procedure. The value of the dummy sensitivity index thus provides an indication of the extent of approximation errors and can be used to put all other sensitivity results into context. In fact, if an input factor is associated with a sensitivity index significantly larger than the dummy sensitivity, then one can sensibly conclude that this input factor is indeed influential. If instead the approximate sensitivity index is equal or even smaller than the dummy sensitivity, then nothing can be concluded about this input factor because its non-zero sensitivity may be due to an actual effect of the input on the output or be purely a consequence of approximation errors.

In the case of PAWN sensitivity indices, a dummy parameter should in principle have zero sensitivity because if a parameter has no effect on the output, then fixing its value has no effect on the output distribution, hence F_y=$F_{y|x_{\text{dummy}}}$ at all conditioning values of x_{dummy} in Eq. (2) and S_{dummy}=0. However, the dummy parameter might have a positive approximate sensitivity index ($\hat{S}_{\text{dummy}} > 0$) when using Eq. (4) or (5) because the empirical distributions $\hat{F}_y$ of two samples can differ from each other even if the samples are drawn from the same distribution F_y. The approximate sensitivity $\hat{S}_{\text{dummy}}$ can thus be interpreted as a measure of the accuracy in approximating CDFs by empirical distributions and hence of the accuracy of the estimated PAWN indices given the chosen sample size (Zadeh et al., 2017).

In this paper, we will use a very simple and straightforward approach to implement these ideas. As suggested in the previous subsection, we will derive the unconditional sample YU by randomly extracting from Y a subsample of size N_c, and repeat the subsampling for a prescribed number of times. Given that by construction $N_c < N$, we can bootstrap *without replacement* (Efron and Tibshirani, 1993) from YU (the reasons for preferring resampling without replacement when applying PAWN is discussed in the Supplementary Materials of Zadeh et al. (2017)). We will then apply Eq. (5) for each bootstrap resample of YU to derive a distribution of approximate PAWN sensitivity indices, and hence confidence intervals. Finally, we will estimate the PAWN sensitivity of the dummy parameter as:

$$\hat{S}_{\text{dummy}} = \underset{k=1,\ldots,n}{\text{mean}} \max_{y} |\hat{F}_y^{(k)}(y) - \hat{F}_y^{(n+1)}(y)| \tag{6}$$

where $\hat{F}_y^{(k)}$ is the empirical distribution of the k-the boostrap resample of the unconditional output sample YU.

3 Results

In this section we demonstrate the proposed approximation strategy from a generic dataset using two case studies. The former is a standard benchmark function widely used in the GSA literature and also used in Pianosi and Wagener (2015) to demonstrate PAWN with a tailored sampling strategy. The objective of this application is to show whether the two approximation strategies provide similar results and to assess the impact of the tuning parameter n on a simple case study. Then, we apply our new strategy to a much more complex and more realistic case study, the Soil Water Assessment Tool (SWAT) model, in a set-up comprising 50 parameters. The objective of the latter application is to evaluate the scalability of the proposed PAWN approximation strategy to problems with many uncertain input factors. We also explore the impact of sample size on the sensitivity estimates, on input ranking and screening, and to illustrate a simple approach to evaluate the effects of the tuning parameter n on PAWN sensitivity indices.

3.1 Application to Ishigami-Homma function

We first consider the Ishigami-Homma function

$$y = \sin(x_1) + a\sin(x_2)^2 + b\ x_3^4 \sin(x_1) \tag{7}$$

where $x_i \sim \mathrm{U}[-\pi, +\pi]$ for i=1,2,3 and a=2 and b=1. This function is often used in GSA studies because the variance-based sensitivities of y can be calculated analytically (see for instance Chapter 4 in Saltelli et al. (2008)), which makes it an ideal testing ground of approximate sensitivity estimators. In particular, the first-order (S_i^F) and total-order (S_i^T) variance-based sensitivity indices are S_1^F=0.3830, S_1^T=0.9991, S_2^F=S_2^T=0.0009, S_3^F=0, S_3^T=0.6161. According to these indices, x_1 is by far the most influential input, x_2 has very limited influence and no interactions, x_3 is only influential through interactions with x_1.

The Ishigami-Homma function was used in Pianosi and Wagener (2015) as a testing ground for PAWN. In that work, the tailored sampling strategy was used and the KS values were aggregated across conditioning points by taking their median, i.e. stat=median in Eq. (4). With these choices, PAWN sensitivity indices were found to be equal to $\hat{S}_1$=0.48, $\hat{S}_2$=0.14, $\hat{S}_3$=0.30, which provides input rankings of (x_1 as most influential, then x_3, and finally x_2) consistent with the results of a variance-based analysis.

Here we re-compute the PAWN sensitivity indices using the proposed approximation strategy from a generic sample, i.e. according to Eq. (5) instead of Eq. (4). As a generic sampling strategy we use Latin Hypercube and we start by setting the tuning parameters to N=500 samples and n=10 conditioning intervals for each input factor. We repeat the calculations by bootstrapping without replacement, as described in Sec. 3.4. With this set-up, PAWN indices (median KS) are found equal to $\hat{S}_1$=0.50, $\hat{S}_2$=0.16, $\hat{S}_3$=0.29 (averages over 50 bootstrap resamples). These numbers are very consistent with those obtained using the tailored sampling strategy, which means that the two approximation approaches are essentially equivalent in this case. It is worth noticing that here we used a generic dataset of N=500 model evaluations, while in Pianosi and Wagener (2015) we used N_u+n×N_c×M=100+15×50×3=2350 model evaluations (although in Pianosi and Wagener (2015) we did not explore whether using less samples would have produced different results).

We further explore the impact of the sample size (N) and of the chosen number of conditioning intervals (n) in Figure 3. Each panel refers to a different input factor, and it shows the median KS (again, average over 50 bootstrap resamples) for different choices of n (horizontal axis) and for different sample size (color). For each combination (N,n), the Figure also shows the estimated PAWN sensitivity of the dummy parameter, computed by Eq. (6) (dashed line). The Figure shows that:

- As the sample size (N) increases, both the bootstrap confidence intervals and the value of $\hat{S}_{\text{dummy}}$ reduce. Furthermore, the median KS values stabilise, at least for larger n values (more on the impact of n in the next point). These patterns are expected and simply prove that the approximation strategy behaves sensibly. The more interesting fact is that using $N = 500$ samples provides very similar results as using N=2000, which means that the proposed approximation strategy provides robust results already at relatively small sample size in this particular case.
- The choice of n seems to have a limited effect on the sensitivity estimates as long as its value is sufficiently high (above 5 in our case). In fact, KS medians are essentially stable for any choice of n >5 and close to the (presumed) correct values (we are now focusing on results for N=500 and N=2000, those obtained with N=100 being too imprecise). Notice that Figure 3 also reports results for n=1, i.e. the limit case where input-output samples are not split into intervals and hence, by definition, the PAWN sensitivity index is equal to 0. This set-up would never be used in practice, however it is shown here to prove that the method behaves consistent with expectations. Finally, sensitivity indices that are lower in values, i.e. those of inputs x_2 and x_3, are more unstable at low values of n while the higher sensitivity index (that of x_1) is almost insensitive to the tuning parameter, which is again quite consistent with expectations. Basically, we see an effect of n only when (a) we use a very small sample size (N=100) and relatively large n so that the number of samples used for estimating output distributions becomes quite low (for example, for $n = 14$ we get N_c=100/14∼7); (b) we use a very small value of n, for example 3 or 4, and then the sensitivity indices of less influential inputs (x_2 and x_3) are badly estimated.

To conclude, application of the proposed approximation strategy to a synthetic test function delivers reliable estimates of PAWN sensitivity indices (median KS) at relatively low sample size (N >=500) and quite irrespectively of the chosen value of the tuning parameter n (provided that n >5).

3.2 Application to the SWAT model

The Soil and Water Assessment Tool (SWAT) is a semi-distributed hydrological model developed by the USDA Agricultural Research Service (Arnold et al., 1998) and used worldwide to study the impact of land use and management practices on water quantity and quality at the catchment scale (e.g. Gassman et al. (2007)). Here, we use a model set-up for the upper Senne River basin in Belgium, which is described in Leta et al. (2015) and was used in a previous GSA study by Sarrazin et al. (2016) and comprises 50 uncertain parameters. The model output y considered in the GSA is a performance metric, the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), measuring the distance between daily flow predictions of the model and available observations. More information about the model, the application river basin and the definition of inputs and output for GSA can be found in Sarrazin et al. (2016). A list of the 50 model parameters and their variability ranges is given in the Supplementary Material of this paper. Here we re-use the input-output dataset generated for the Regional Sensitivity Analysis in Sarrazin et al. (2016), obtained by Latin Hypercube sampling and including $N = 10,000$ samples.

First, we approximate the PAWN sensitivity indices by Eq. (5) using n=10 conditioning intervals. We repeat our calculations with 100 bootstrap resamples and compute the averages and confidence intervals of each index, shown in Figure 4. According to this figure, the most influential parameter is the 11th, followed by parameters 9,32,10. Given that the confidence intervals for these three parameters are mostly overlapping, we cannot make further distinctions between them and thus we would put all of them in the 2nd position of the parameter ranking. Following the same line of reasoning, we would put parameters 8,17,2,43 in the 3rd position and parameters 14,25,42,34,12,28,35 in the 4th. The remaining parameters have an average sensitivity value close to that of the dummy parameter (red line in Figure 4), which means we cannot distinguish whether they actually have an influence on the output or whether their estimated sensitivity is a pure product of approximation errors. Hence, we classify them as potentially uninfluential. These ranking results are consistent with those obtained by Sarrazin et al. (2016) using other GSA methods. Figure 5 provides a short comparison with those results, focusing in particular on the top positions of the parameter ranking. The fact that only a limited number of parameters (4 to 8 in our case) control the output performance metric (Nash-Sutcliffe efficiency) is consistent with several studies on calibration of hydrological models (e.g. Jakeman and Hornberger (1993) and Van Werkhoven et al. (2009)). Furthermore, from the parameter list in the Supplementary Material, one can see that the 4 top-ranking parameters are: the SCS runoff curve number for moisture condition in the agricultural areas (11), the hydraulic conductivity in the river channel (9), the average slope steepness in the agricultural areas (32), and the Manning coefficient for the channel (10). This is reasonable given the predominance of agricultural land use in the catchment (62% of the catchment area as reported in Leta et al. (2015)) and the fact that the chosen output metric (Nash-Sutcliffe efficiency) emphasises errors in peak flow predictions, which we expect to be mainly controlled by the parameters that characterise river routing (see for example Van Werkhoven et al. (2008)).

Next, we analyse the impact of the sample size N. To do this, we randomly extract a subsample of smaller size from our original dataset and repeat the approximation procedure of the PAWN sensitivity indices. We test N=1000, 5000 and 7500 (Figure 6). We find that using N=1000 (top panel) produces rather imprecise sensitivity indices, in fact almost all confidence intervals overlap each other and with the dummy parameter threshold, which prevents us from inferring a robust parameter ranking. However, already at the next sample size (N=5000) the confidence intervals start to separate out and the ranking of the influential parameters is similar to the one obtained at the highest sample size (N=10000).

The effect of the tuning parameter n is then analysed in Figure 7. The Figure depicts the approximate PAWN sensitivity indices obtained using different values of n from 6 to 20. Overall, the changes in value with varying n seem to be minor. We observe a trend of increasing sensitivity values as n increases, i.e. the grey shading gets darker from left to right. However this trend mainly involves parameters with very low sensitivity and does not affect the key conclusion that these parameters are probably uninfluential, as their approximate index remains below that of the dummy parameter (cases flagged by red crosses).

Finally, we investigate the effect of the aggregation statistic. This is shown in Figure 8 and 9, which are the analogues of Figure 6 using stat=mean and stat=max in Eq. (5) instead of the median. Figure 8 shows that using the mean KS provides very similar ranking and screening results as using the median. Figure 9 instead reveals that using the max KS significantly increases the relative importance of some input factors (e.g. parameters 43, 35 and 46). We further investigate this behaviour in Figure 10, which shows the scatter plots of the output samples for parameters 43, 35 and 46 (top panels) and the KS values for different conditioning intervals (bottom panels). We also include the results for the most influential input 11, as a reference. This figure shows that the output is rather insensitive to variations in parameters 43, 35 and 46 for most of their variability ranges with the exception of the very low end, where the KS value is above the dummy parameter threshold. Further analysis (not shown) reveals that in those sub-range the conditional output distributions are shifted to the left of the unconditional ones i.e. lower output values are more frequent. While further investigating the implications of this localised effect is beyond the scope of this paper, we have shown this analysis as an example of how PAWN can also be used to gain insights into the input-output mapping, and reinforce the visual inspection of scatter plots with quantitative evidence of local effects.

4 Conclusions

In this paper we have introduced and discussed a new strategy to approximate PAWN sensitivity indices from a generic sample of model inputs and outputs using only one algorithm tuning parameter (the number n of conditioning intervals). Via application to a benchmark function (with 3 uncertain inputs) and to a complex hydrological model (50 uncertain parameters) we have demonstrated that the new approximation strategy provides results consistent with those of the original approximation strategy and of other GSA methods. Furthermore, the screening and ranking of uncertain inputs based on the new approximation strategy is reliable at reasonably low sample sizes (around N=500 samples in the 3 inputs case and 5000 in the 50 inputs case) and is robust against the choice of n. Obviously we cannot extrapolate from two case studies that these conclusions will hold true for any other application, however in this paper we have also provided a number of visualisation tools, such as those shown in Figure 3, 6 and 7, that can be used to evaluate the impact of N and n in any given application, at no additional computing cost. While we have followed a heuristic approach to the convergence of sensitivity estimates, the same issue is approached from a theoretical perspective in Borgonovo et al. (2016), who investigated the properties of a *partition selection strategy* (the splitting strategy, in our terminology) that ensure converge to true sensitivity values for the case when the aggregation statistic of KS values is the mean. Expanding those theoretical results to other aggregation statistics may be an interesting avenue for future research.

Based on the analyses presented in this paper, we can give the following practical recommendations to future PAWN users:

- Always use the new approximation strategy instead of the tailored strategy presented in Pianosi and Wagener (2015). The functions to implement the new approximation strategy are now included in our open-source SAFE Toolbox (Pianosi et al., 2015).
- If a generic input-output dataset is available, you can re-use it to apply PAWN, otherwise generate one of size N as large as possible, compatibly with available computing resources. In both cases, compute the PAWN indices using both the complete dataset and subsets of smaller sizes, as done for example in Figure 6, to verify that the key conclusions about the ranking and screening of the input factors are not significantly affected by the value of N. If instead they are, the sample size should probably be increased.
- Use n=10 to start with but check the effects of varying n of some units up and down, as done in Figure 7.
- In all the analyses, use bootstrapping to derive confidence intervals and thus infer whether differences in sensitivity indices are large enough to

discriminate between the relevant inputs, or they should be put in the same ranking position. Use the KS of the dummy parameter to identify inputs whose measured sensitivity is too low to be distinguishable from approximation errors.

Once again we stress that all these analyses (i.e. reducing N, changing n, bootstrapping, and calculation of the dummy KS) can be performed over the available dataset and do not require to re-run the model, hence they come at almost no additional computing cost. We hope this increased efficiency and simplicity of the new approximation strategy will contribute to increase the uptake of the PAWN method and facilitate its use as a complement of variance-based sensitivity analysis and its integration into multi-method approaches to GSA in general.

Acknowledgements

The initial development of the PAWN method and of the SAFE Toolbox was supported by the Natural Environment Research Council [Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1]. F Pianosi is partially funded by a UK EPSRC Living with Environmental Uncertainty Fellowship [grant number EP/R007330/1]. T Wagener is partially supported by a Royal Society Wolfson Research Merit Award. The authors are grateful to the researchers who used PAWN so far and, through their feedbacks and comments, motivated and gave directions for this study.

References

Anderson, B., Borgonovo, E., Galeotti, M., Roson, R., 2014. Uncertainty in climate change modeling: can global sensitivity analysis be of help? Risk Analysis 34 (2), 271 – 293.

Arnold, J., Srinivasan, R., Muttiah, R., Williams, J., 1998. Large area hydrologic modeling and assessment part 1: model development. J. Am. Water Resour Assoc 34 (1), 73–89.

Bastidas, L. A., Gupta, H. V., Sorooshian, S., Shuttleworth, W. J., Yang, Z. L., 1999. Sensitivity analysis of a land surface scheme using multicriteria methods. Journal of Geophysical Research: Atmospheres 104 (D16), 19481–19490.

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

Borgonovo, E., Hazen, G., Plischke, E., 2016. A common rationale for global sensitivity measures and their estimation. Risk Analysis 36, 1871–1895.

Borgonovo, E., Lu, X., Plischke, E., Rakovec, O., Hill, M. C., 2017. Making the most out of a hydrological model data set: Sensitivity analyses to open the model black-box. Water Resources Research 53, 7933–7950.

Castaings, W., Borgonovo, E., Morris, M., Tarantola, S., 2012. Sampling strategies in density-based sensitivity analysis. Environmental Modelling & Software 38 (0), 13–26.

Dell'Oca, A., Riva, M., Guadagnini, A., 2017. Moment-based metrics for global sensitivity analysis of hydrological systems. Hydrology and Earth System Sciences 21 (12), 6219–6234.

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

Ferretti, F., Saltelli, A., Tarantola, S., 2016. Trends in sensitivity analysis practice in the last decade. Science of The Total Environment 568, 666–670.

Forrester, A., Sobester, A., Keane, A., 2008. Engineering Design via Surrogate Modelling: a Practical Guide. Wiley.

Fox, D., Koch, D., Tester, J., 2016. An analytical thermohydraulic model for discretely fractured geothermal reservoirs. Water Resources Research 52 (9), 6792–6817.

Freer, J., Benev, K., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. Water Resources Research 32 (7), 2161–2173.

Gassman, P., 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.

Gillies, K., Krone, S., Nagler, J., Schultz, R., 2016. A computational model of the rainbow trout hypothalamus-pituitary-ovary-liver axis. PLOS computational biology 12 (4).

Hosseini, S., Ataie-Ashtiani, B., Simmons, C., 2017. Spring hydrograph simulation of karstic aquifers: Impacts of variable recharge area, intermediate storage and memory effects. Journal of Hydrology 552, 225–240.

Jakeman, A., Hornberger, G., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resources Research 29, 2637–2649.

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

Leta, O., Nossent, J., Velez, C., Shrestha, N., van Griensven, A., Bauwens, W., 2015. Assessment of the different sources of uncertainty in a SWAT model of the river senne (belgium). Environmental Modelling & Software 68, 129–146.

Liu, H., Sudjianto, A., Chen, W., 2006. Relative entropy based method for probabilistic sensitivity analysis in engineering design. Journal of Mechanical Design 128, 326–336.

Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models, Part I - A discussion of principles. Journal of Hydrology 10, 282–290.

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

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

Peeters, L. J. M., Podger, G. M., Smith, T., Pickett, T., Bark, R. H., Cuddy, S. M., 2014. Robust global sensitivity analysis of a river management model to assess nonlinear and interaction effects. Hydrology and Earth System Sciences 18 (9), 3777–3785.

Petropoulos, G., Srivastava, P., 2017. Sensitivity Analysis in Earth Observation Modelling. Elsevier.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., 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, F., Iwema, J., Rosolem, R., Wagener, T., 2017. A Multimethod Global Sensitivity Analysis Approach to Support the Calibration and Evaluation of Land Surface Models. Amsterdam:Elsevier, pp. 125–144.

Pianosi, F., Sarrazin, F., Wagener, T., 2015. A matlab toolbox for global sensitivity analysis. Environmental Modelling & Software 70, 80 – 85.

Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based oncumulative distribution functions. Environmental Modelling & Software 67, 1 – 11.

Pianosi, F., Wagener, T., 2016. Understanding the time-varying importance of different uncertainty sources in hydrological modelling using global sensitivity analysis. Hydrological Processes 30 (22), 3991–4003.

Pilz, T., Francke, T., Bronstert, A., 2017. lumpR 2.0.0: an R package facilitating landscape discretisation for hillslope-based hydrological models. Geoscientific Model Development 10 (8), 3001–3023.

Plischke, E., Borgonovo, E., Smith, C. L., 2013. Global sensitivity measures from given data. European Journal of Operational Research 226 (3), 536 – 550.

Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical Recipes in C. Cambridge University Press.

Razavi, S., Gupta, H., 2015. What do we mean by sensitivity analysis? the need for comprehensive characterization of global sensitivity in earth and environmental systems models. Water Resources Research 51 (5), 3070–3092.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications 181 (2), 259–270.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis, The Primer.

Wiley.
Sarrazin, F., Pianosi, F., Wagener, T., 2016. Global sensitivity analysis of environmental models: Convergence and validation. Environmental Modelling & Software 79, 135 – 152.
Sedighian, S., Kim, S. H., Cho, S. Y., Kim, M., Kim, D., Cha, D., 2015. Parameter ranking system of indoor radon concentration in South Korea, case studies: Dokdo island, Yang Pyeong and Nae Gi. International Journal of Environmental Research 9 (4), 1233–1236.
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 (2).
Spear, R., Grieb, T., Shang, N., 1994. Parameter uncertainty and interaction in complex environmental models. Water Resources Research 30 (11), 3159–3169.
Stanfill, B., Mielenz, H., Clifford, D., Thorburn, P., 2015. Simple approach to emulating complex computer models for global sensitivity analysis. Environmental Modelling & Software 74, 140 – 155.
Strong, M., Oakley, J., Brennan, A., 2014. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a nonparametric regression approach. Medical Decision Making 34 (3), 311–326.
Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrology and Earth System Sciences 11, 793–817.
Van Werkhoven, K., Wagener, T., Tang, Y., Reed, P., 2008. Rainfall characteristics define the value of streamflow observations for distributed watershed model identification. Geophysical Research Letters 35 (L11403).
Van Werkhoven, K., Wagener, T., Tang, Y., Reed, P., 2009. Complexity reduction in multiobjective watershed model calibration. Advances in Water Resources 32 (8), 1154–1169.
Wagener, T., Kollat, J., 2007. Visual and numerical evaluation of hydrologic and environmental models using the monte carlo analysis toolbox (MCAT). Environmental Modelling & Software 22, 1021–1033.
Xiaomeng, S., Jianyun, Z., Chesheng, Z., Yunqing, X., Ming, Y., Chonggang, X., 2015. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. Journal of Hydrology 523, 739–757.
Yang, J., 2011. Convergence and uncertainty analyses in monte-carlo based sensitivity analysis. Environmental Modelling & Software 26 (4), 444–457.
Zadeh, F. K., Nossent, J., Sarrazin, F., Pianosi, F., van Griensven, A., Wagener, T., Bauwens, W., 2017. Comparison of variance-based and moment-independent global sensitivity analysis approaches by application to the SWAT model. Environmental Modelling & Software 91, 210 – 222.

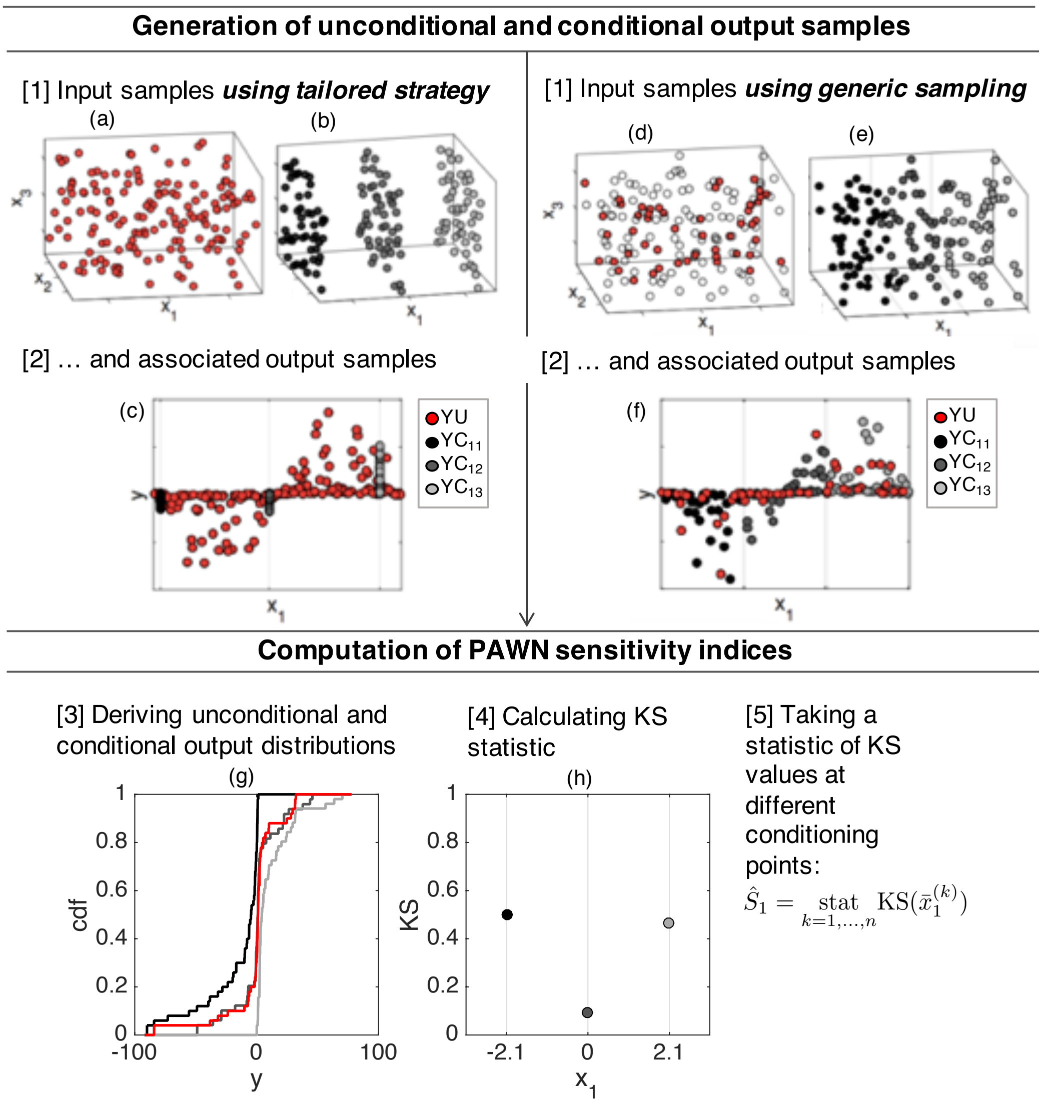

Fig. 1. Example of using a tailored sampling strategy (left) and generic sampling (right) to approximate the PAWN index of input x_1 in a case of M=3 input factors. Left (tailored): (a) Input samples used to derive the unconditional output sample YU. These are generated by randomly sampling the entire space of input variability. (b) Input samples used to derive three conditional samples YC_{11}, YC_{12} and YC_{13}. These are generated by fixing x_1 at selected conditioning values (for the sake of clarity, only n=3 conditioning values are shown here). (c) Scatter plot of the unconditional (red) and conditional (grey) output samples YU, YC_{11}, YC_{12} and YC_{13} against x_1. Right (generic): similar to the left hand side but this time the input samples in (d) and (e) are the same. A random subset (highlighted in red) is used to derive YU, and the three subsets obtained by splitting the variability range of x_1 into 3 intervals (grey) are used to derive YC_{11}, YC_{12} and YC_{13}. After sampling, the approximation of the PAWN sensitivity index follows the same steps: (g) unconditional output distribution (red) and the three conditional distributions (grey) when x_1 is fixed to a given value (interval). (h) KS statistic (maximum absolute difference) between the unconditional distribution and each of the three conditional ones, plotted against the conditioning value (centre of the interval).

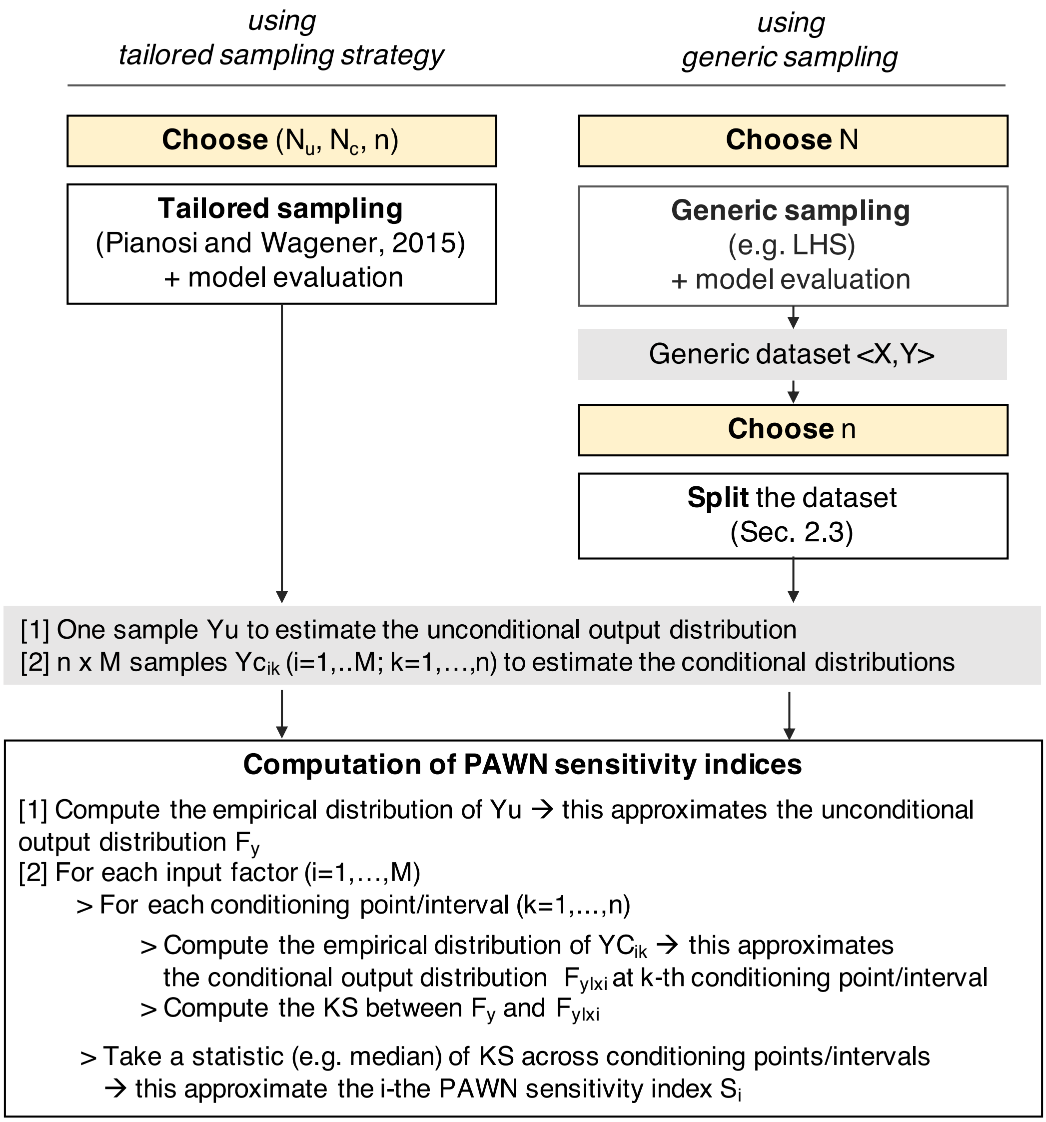

Fig. 2. Schematic of the steps needed to apply PAWN using a tailored sampling strategy (left) and generic sampling (right). In the latter case, if a generic input/output dataset is already available, the very first step of sampling and model evaluation can be skipped and the subsequent steps applied to the available dataset.

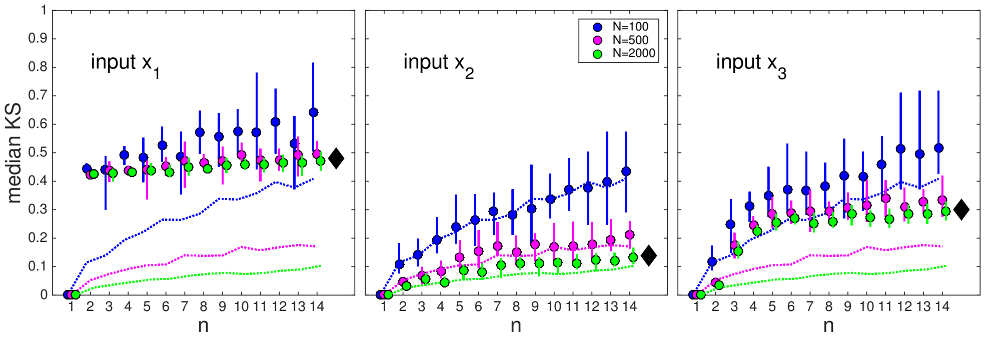

Fig. 3. PAWN indices from generic sample for the three input factors of the Ishigami-Homma function. Each subplot report results for one input factor. The PAWN index is defined as the median KS across conditioning intervals (i.e. Eq. (5) where stat=median). PAWN indices are approximated using an increasing sample size (N) and increasing number of conditioning intervals (n). For each combination of (N,n), bootstrapping is used to estimate the 95% confidence interval (vertical line) and mean value (circle) of each PAWN index. Dashed lines show the KS of the dummy parameter computed according to Eq. (6) at each combination of (N,n). For comparison, the Figure also shows the PAWN indices approximated using the tailored sampling strategy (black diamond).

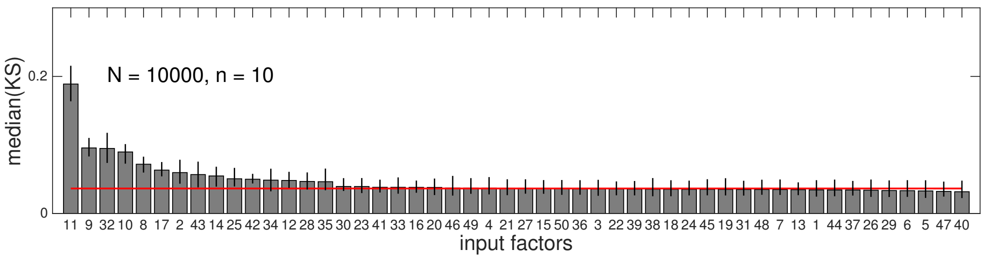

Fig. 4. PAWN sensitivity indices from generic sample for the 50 input parameters of the SWAT simulation model. The PAWN index is defined as the median KS across conditioning intervals (i.e. Eq. (5) where stat=median). Bootstrapping is used to estimate the 95% confidence interval (vertical line) and mean value (bar height) of each PAWN index. The red line shows the KS of the dummy parameter computed by Eq. (6). Input parameters are sorted according to their PAWN index values.

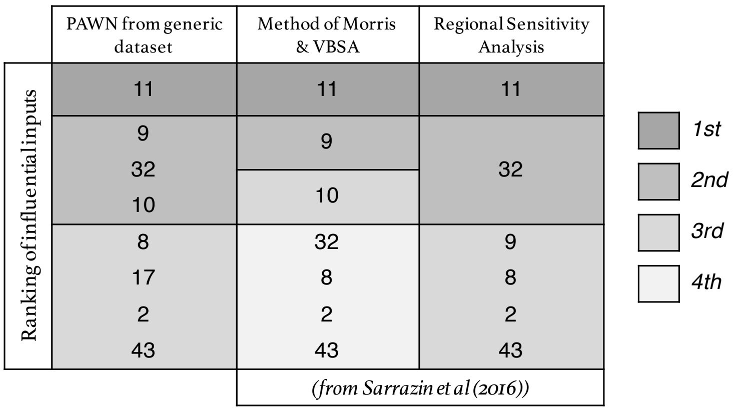

Fig. 5. Comparison between the ranking of influential parameters derived from the PAWN indices and those obtained in Sarrazin et al. (2016) by applying the method of Morris, Variance-Based Sensitivity Analysis (VBSA) and Regional Sensitivity Analysis.

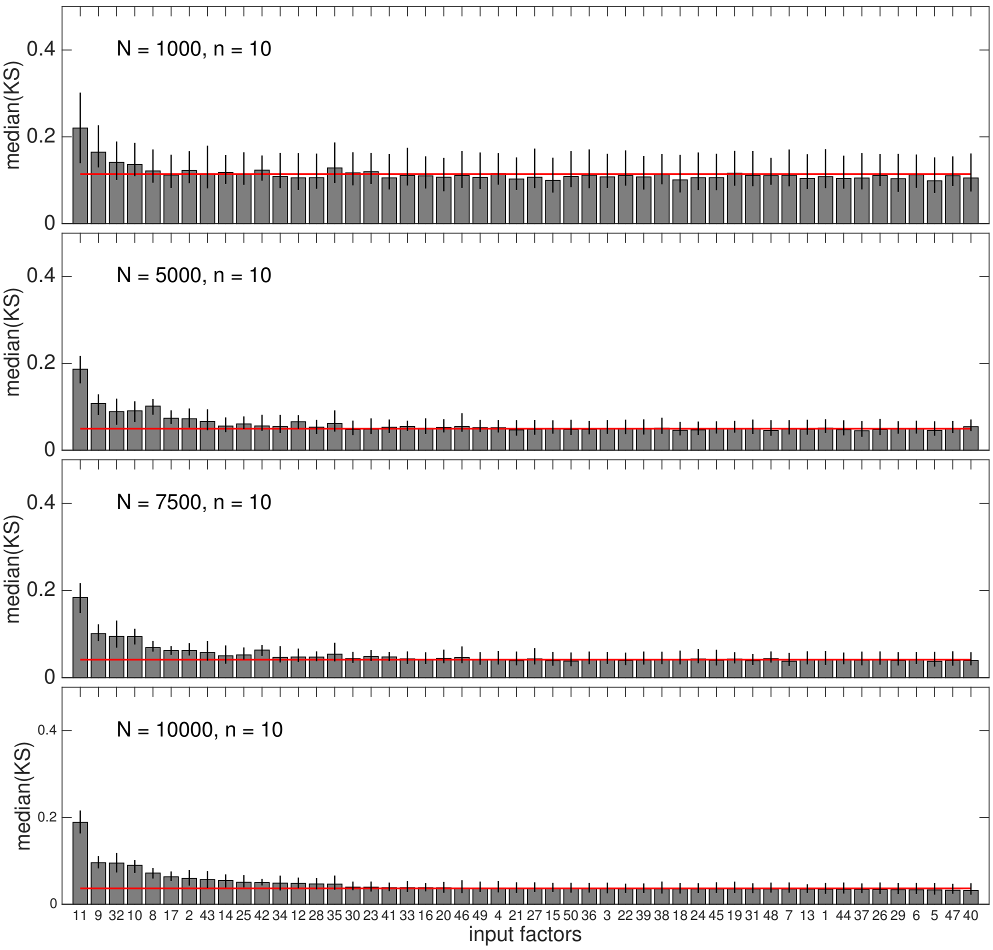

Fig. 6. Effect of the sample size N on the PAWN sensitivity indices approximated from a generic sample. Notice that the results in the bottom panel are the same as in Fig. 4 and are only reported to facilitate comparison. In all panels the input parameters are presented in the same order: this order coincides with their ranking (from most influential to least) in the bottom panel but not necessarily in the others given that the PAWN sensitivity estimates are different. The red line depicts the dummy parameter result.

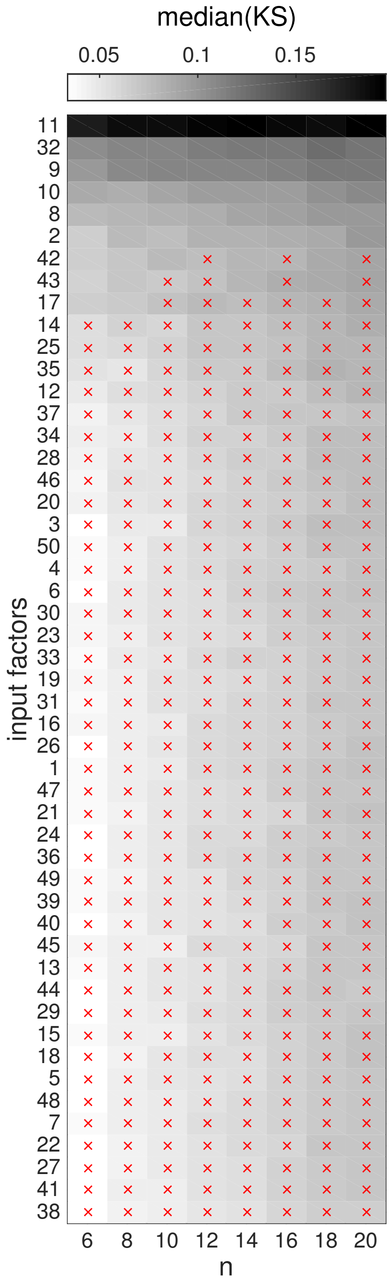

Fig. 7. Effect of the tuning parameter n (number of conditioning intervals) on the PAWN sensitivity indices approximated from generic sample (sample size N=5000). Red crosses are used to mark sensitivity indices whose value is not higher than the KS of the dummy parameter, and hence is within margins of approximation errors.

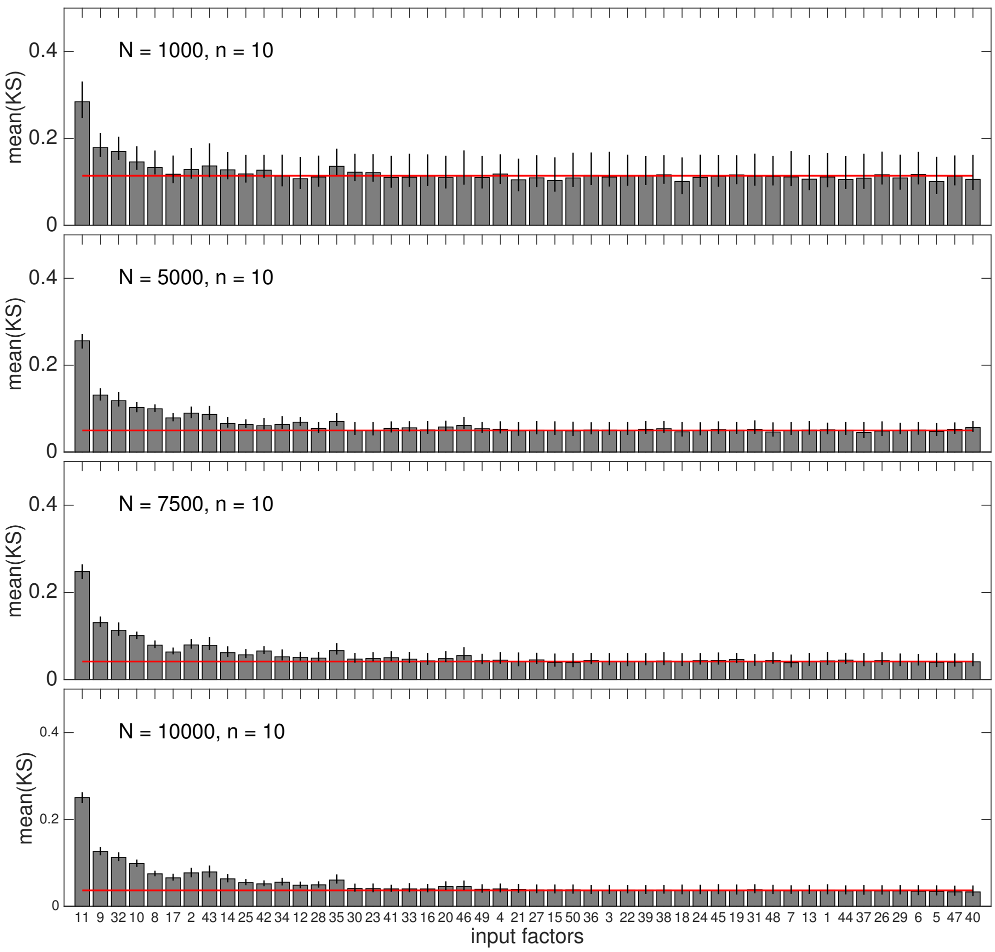

Fig. 8. Same as in Figure 6 but defining the PAWN index as the mean KS across conditioning intervals, i.e. stat=mean in Eq. (5) instead of median.

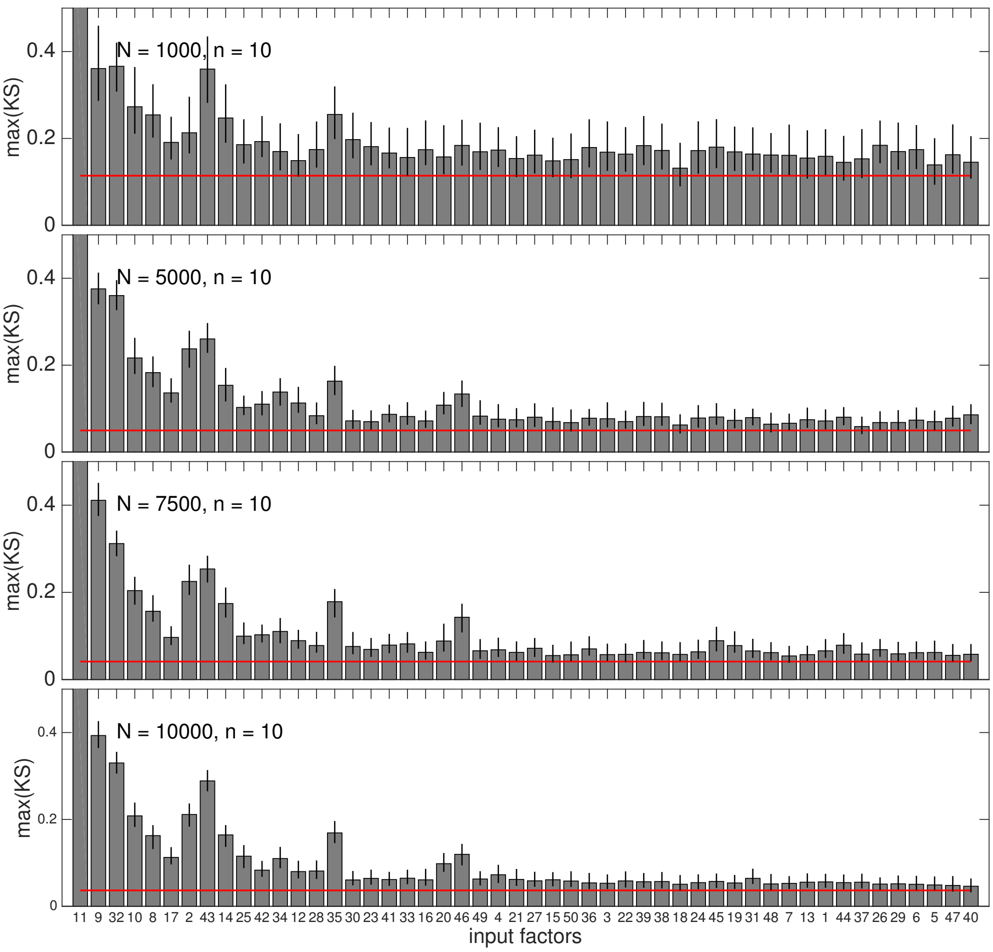

Fig. 9. Same as in Figure 6 but defining the PAWN index as the maximum KS across conditioning intervals, i.e. stat=max in Eq. (5).

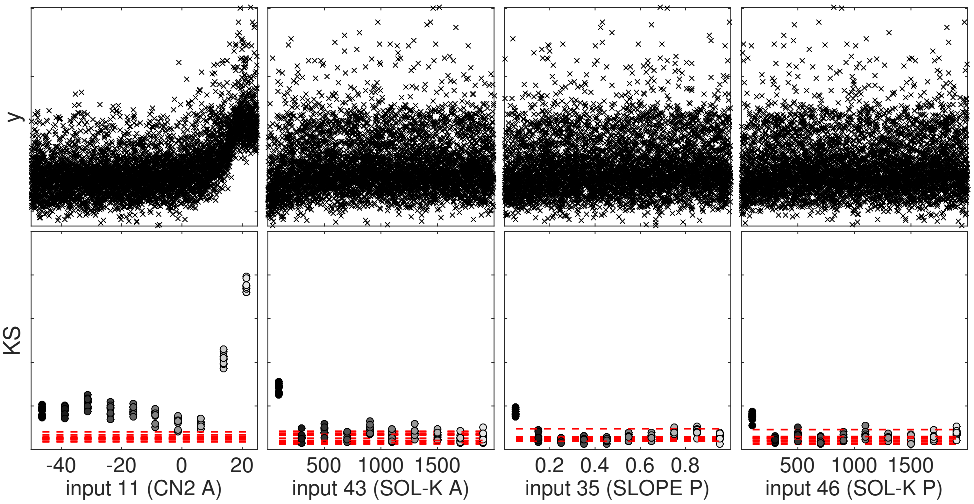

Fig. 10. Scatter plots and KS statistics of four selected input parameters of the SWAT model: number 11 is the one consistently ranked as most influential, number 43, 35 and 46 are classified as influential if using the maximum KS as PAWN sensitivity index, while they are not if using the mean or median KS.