

A sensitivity analysis of the PAWN sensitivity index

Arnald Puy^{a,b,*}, Samuele Lo Piano^{c,d}, Andrea Saltelli^{b,d}

^a Department of Ecology and Evolutionary Biology, M31 Guyot Hall, Princeton University, NJ 08544, USA

^b Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Parkveien 9, PB 7805, 5020 Bergen, Norway

^c School of the Built Environment, University of Reading, JJ Thomson Building, Whiteknights Campus, Reading, RG6 6AF, UK

^d Open Evidence, Universitat Oberta de Catalunya, Edifici 22@, Universitat Oberta de Catalunya, 08018 Barcelona, Spain

ARTICLE INFO

Keywords:

Uncertainty
Environmental modelling
Statistics
Risk

ABSTRACT

The PAWN index is gaining traction among the modelling community as a sensitivity measure. However, the robustness to its design parameters has not yet been scrutinized: the size (N) and sampling (ϵ) of the model output, the number of conditioning intervals (n) or the summary statistic (θ). Here we fill this gap by running a sensitivity analysis of a PAWN-based sensitivity analysis. We compare the results with the design uncertainties of the Sobol' total-order index (S_{Ti}^*). Unlike in S_{Ti}^* , the design uncertainties in PAWN create non-negligible chances of producing biased results when ranking or screening inputs. The dependence of PAWN upon (N, n, ϵ, θ) is difficult to tame, as these parameters interact with one another. Even in an ideal setting in which the optimum choice for (N, n, ϵ, θ) is known in advance, PAWN might not allow to distinguish an influential, non-additive model input from a truly non-influential model input.

1. Introduction

Pianosi and Wagener (2015, 2018) have recently published in *Environmental Modelling & Software* a new measure for sensitivity analysis, the PAWN index. Like other moment-independent approaches (i.e. entropy-based (Liu et al., 2006), density-based (Borgonovo et al., 2014; Borgonovo, 2007)), PAWN does not resort to statistical second-order moments such as variance to apportion output uncertainty to the model parameters. Instead, it relies on Cumulative Distribution Functions (CDFs) to characterize the maximum distance between the unconditional output distribution Y_U , i.e. obtained by moving all parameters simultaneously, and the conditional output distribution $Y_{C_{ij}}$, i.e. obtained by fixing the i th parameter to $j = 1, 2, \dots, n$ values or intervals within its uncertainty range. The difference between Y_U and $Y_{C_{ij}}$ is assessed via the Kolmogorov–Smirnov test, although other distance-based tests, such as the Anderson–Darling's, may also be used (Khorashadi Zadeh et al., 2017). The final PAWN index for a given parameter is obtained by calculating the mean, the median, the maximum or any other summary statistic over all the KS values computed between Y_U and $Y_{C_{ij}}$.

The most up-to-date approximation to the PAWN index, named “the generic approach” (Pianosi and Wagener, 2018), is as follows: let there be an \mathbf{A} matrix with $v = 1, 2, \dots, N$ rows and $i = 1, 2, \dots, k$ parameters. After computing the model output Y , the range of variation of the i th parameter is split into $j = 1, 2, \dots, n$ intervals of size N_c (where $N_c \approx N/n$). The model output linked to the j th interval is used as the

conditional model output $Y_{C_{ij}}$. The unconditional model output Y_U can concur with the whole model output or can be a random sub-sample of the same size as N_c . With this approach, the total number of model runs to compute PAWN is fully determined by N (Pianosi and Wagener, 2018).

Based on trials with the Liu et al. (2006) function, the Ishigami and Homma (1990) function, the SWAT model (Khorashadi Zadeh et al., 2017) or a wind-energy converter model (Holl et al., 2016), it has been observed that PAWN might reach convergence much faster than Sobol' indices. A key question, however, is to know how the selection of N, n , the sampling of Y_U or the summary statistic affects the accuracy of PAWN. This is timely given the widespread adoption of the index: since its inception in 2015, PAWN has been cited 94 times, with the number of citations stably increasing from 5 in 2015 to 31 in 2019. Most of the works quoting PAWN are from the environmental sciences (53), followed by engineering (24) and computer science (23) (Scopus search on September 25 2019). Gaining a systematic insight into the internal functioning of PAWN shall thus allow the modelling community to better appraise its robustness, thus increasing its transparency as well as our awareness of its advantages and limitations.

Here we assess the sensitivity of PAWN to the main structural uncertainties involving its calculation, an exercise that might be termed “a sensitivity analysis of a sensitivity analysis” (SA of SA). This expression was used by Paleari and Confalonieri (2016) to study how sensitivity indices are affected by uncertainties in the probability distributions used

* Corresponding author at: Department of Ecology and Evolutionary Biology, M31 Guyot Hall, Princeton University, NJ 08544, USA.
E-mail address: apuy@princeton.edu (A. Puy).

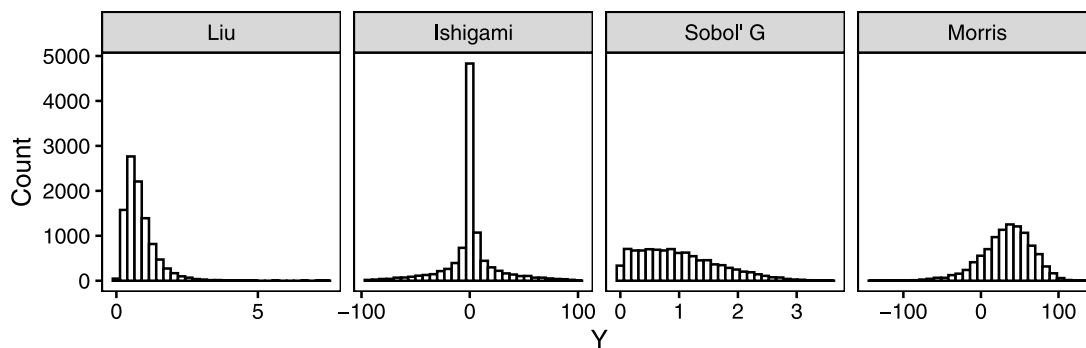


Fig. 1. Model output of the Liu et al. (2006), Ishigami and Homma (1990), Sobol' (1998) and Morris (1991) functions. $N = 4000$.

to describe the model inputs, a work that actually falls into the tradition of “probability of probability” or “probability of frequency” described in Kaplan and Garrick (1981) (see also Aven (2020)). A similar analysis of the sensitivity of results to changes in the range of input factors can be found in Shin et al. (2013). Other approaches include exploring several sensitivity measures Saltelli and Homma (1992), Saltelli et al. (1993), Pappenberger et al. (2008). More recently, Noacco et al. (2019) included considerations of form of the output, sample size, choice of method, measuring interactions, range of distributions, inclusion or exclusion of non-behavioural runs (simulations), and use of dummy variables.

Our experiment departs from these studies in that it assesses the uncertainties embedded in the structural design of sensitivity indices, PAWN in that case. We thus explore the implications of the last ring in the chain of uncertainties characterizing any sensitivity analysis, the last stretches of the garden of forking paths that need to be crossed to arrive at a result in any modelling exercise (Borges, 1941; Gelman and Loken, 2013). Exploring all sources of uncertainty is indeed an important pursuit, including the notions of context, purpose and motivations as suggested in sensitivity auditing (Saltelli et al., 2013). However, we believe that the papers just reviewed (Paleari and Confalonieri, 2016; Pappenberger et al., 2008; Shin et al., 2013; Noacco et al., 2019) are not a SA of a SA but more general instances of uncertainty exploration. We take the meaning of SA literally (i.e. the analysis of the sensitivity of a sensitivity analysis to its own design parameters), and thus exclude in this paper the exploration of sensitivities to changing methods, assumptions or model designs, however worthy these analyses might be.

2. Materials and methods

In order to cross-check our approach, we match the design uncertainties of PAWN against the design uncertainties of the Sobol' total-order index (S_{Ti}^*), a well-established measure to determine how much a given model parameter interacts with the rest.¹ Although different in scope and nature, the fact that unconditional CDFs are also affected by interactions paves the way for the uncertainties in PAWN and S_{Ti}^* to be explored in parallel.

For PAWN, we focused on four uncertain parameters: the total number of model runs (N), the number of conditioning intervals (n), the randomness derived from the sampling of the unconditional model output Y_U (ϵ), and the summary statistic (θ). For S_{Ti}^* , we focused on two: N and θ , the latter reflecting the different estimators existing to compute the total-order effect (Jansen, 1999; Homma and Saltelli, 1996; Sobol', 2007) (Table 1). We assessed how different combinations of values for these uncertain parameters condition the PAWN/ S_{Ti}^* index

¹ The asterisk in S_{Ti}^* is used to differentiate between Sobol' total-order index (S_{Ti}^*), i.e. the index against which PAWN is matched, and the index used to assess the extent to which S_{Ti}^* interacts with its design parameters (N, θ, S_{Ti}).

Table 1

Summary of the parameters and their distribution for both PAWN (N, n, ϵ, θ) and S_{Ti}^* (N, θ). DU stands for discrete uniform.

Parameter	Description	Distribution
N	Total number of runs	$\mathcal{U}(200, 2000)$
n	Number of conditioning intervals	$\mathcal{U}(5, 20)$
ϵ	Randomness in the sampling of Y_U	$\mathcal{U}(1, 10^3)$
θ	Summary statistic/estimator	$\mathcal{D}\mathcal{U}(1, 3)$

using four different test functions that yield a skewed model output (Fig. 1): the Liu et al. (2006)'s, which reads as

$$Y = X_1/X_2 \quad (1)$$

where $X_1 \sim \chi^2(10)$ and $X_2 \sim \chi^2(13.978)$; the Ishigami and Homma (1990)'s, which reads as

$$Y = \sin(X_1) + a \sin(X_2)^2 + bX_3^4 \sin(X_1) \quad (2)$$

where $a = 2, b = 1$ and $(X_1, X_2, X_3) \sim \mathcal{U}(-\pi, +\pi)$; the Sobol' (1998)'s, which reads as

$$Y = \prod_{i=1}^k \frac{|4X_i - 2| + a_i}{1 + a_i} \quad (3)$$

where $k = 8, X_i \sim \mathcal{U}(0, 1)$ and $a = (0, 1, 4.5, 9, 99, 99, 99, 99)$; and the Morris (1991) function, which reads as

$$Y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{i,j} w_i w_j + \sum_{i<j<l}^{20} \beta_{i,j,l} w_i w_j w_l + \sum_{i<j<l<s}^{20} \beta_{i,j,l,s} w_i w_j w_l w_s \quad (4)$$

where $w_i = 2(X_i - 0.5)$ for all i except for $i = 4, 5, 7$, where $w_i = 2(1.1X_i/(X_i + 0.1) - 0.5)$, $X_i \sim \mathcal{U}(0, 1)$, and

$$\begin{aligned} \beta_i &= 20, i = 1, 2, \dots, 10, \\ \beta_{i,j} &= -15, i = 1, 2, \dots, 6, \\ \beta_{i,j,l} &= -10, i = 1, 2, \dots, 5, \\ \beta_{i,j,l,s} &= 5, i = 1, 2, \dots, 4 \text{ (Campolongo et al., 2011)}. \end{aligned}$$

We chose these four skewed benchmark functions in order to provide PAWN with a favourable test ground: density-based measures might better characterize the uncertainty in skewed model outputs than variance-based measures as the former do not rely on any statistical moment (Borgonovo et al., 2011).

We selected the distributions of (N, n, ϵ, θ) based on previous work on PAWN and the S_{Ti}^* index (Pianosi and Wagener, 2015, 2018; Saltelli et al., 2010) and some preliminary tests. Firstly, we observed that $N \approx 2000$ was sufficient for the PAWN index of many model inputs to be very close to convergence. Sobol' S_{Ti}^* , on the other hand, required

a larger number of runs (Fig. S1). This corroborated previous observations on the faster convergence rate of PAWN compared to Sobol' indices (Pianosi and Wagener, 2015; Khorashadi Zadeh et al., 2017). By defining $N \sim \mathcal{U}(200, 2000)$ we set our study in a scenario where the uncertainty with regards to the required sample size needed to compute robust PAWN indices is moderate. Indeed, this is usually the case: the analyst might have a cap on the total number of model runs available with the computing resources at hand, but no prior information on the minimum sample size required to ensure stable sensitivity indices for all model inputs.

We defined the distribution of n based on Pianosi and Wagener (2018), who suggested to start with $n = 10$ and vary n some units up and down to check its effect (provided that $n > 5$).

For ϵ , we set 10^3 different starting points (seeds) for the pseudo-random number sequence used to generate the indices (from 1 to N) to sample Y_U . This ensured (1) negligible chances of the same seed overlapping with the same value for N , thus introducing determinism into a process that should be mainly stochastic, and (2) that the randomness in the sampling of Y_U is assessed in terms of its relative influence in the computation of PAWN.

With regards to θ , for PAWN we used the mean, the median and the maximum as a summary statistic for the KS values when $\theta = 1$, $\theta = 2$ and $\theta = 3$. For S_{Ti}^* , we used the estimators by Jansen (1999), Homma and Saltelli (1996) and Sobol' (2007) when $\theta = 1$, $\theta = 2$ and $\theta = 3$ respectively.

In order to estimate the uncertainty propagated by (N, n, ϵ, θ) (resp. N, θ) to PAWN (resp. S_{Ti}^*), we created a $(2^{13}, 2k)$ sample matrix for each function using Sobol' quasi-random number sequences, where $k = 4$ (resp. $k = 2$), and transformed the columns into their appropriate distributions (Table 1). The first k matrix was labelled **A** and the second k matrix, **B**. Our model ran row-wise in both the **A** and **B** matrices, as follows: based on the information contained in the v th row, it created a Sobol' matrix of size $N^{(v)}$ for PAWN [of size $\text{int}(N^{(v)})/(k+1)$] for S_{Ti}^* , and computed either the Liu et al. (2006), the Ishigami and Homma (1990), the Sobol' (1998) or the Morris (1991) function. Then, for each model input i , it either calculated the PAWN $^{(v)}$ index following the conditions set by $n^{(v)}$, $\epsilon^{(v)}$ and $\theta^{(v)}$, or the $S_{Ti}^{*(v)}$ index according to $\theta^{(v)}$.

We estimated how sensitive PAWN (S_{Ti}^*) indices are to uncertainty in (N, n, ϵ, θ) (resp. N, θ) by means of Sobol' indices (Sobol', 1993). For a model of the form $Y = f(X_1, X_2, \dots, X_k)$, where Y is a scalar and X_1, X_2, \dots, X_k are independent parameters described by known probability distributions, we can measure how sensible Y is to a given parameter X_i with

$$V_i = V_{X_i} [E_{X_{\sim i}}(Y|X_i)] \quad (5)$$

where $E_{X_{\sim i}}(Y|X_i)$ is the expected value of Y calculated over all possible values of all parameters except the i th, which is kept fixed. By dividing Eq. (5) by the unconditional model output variance, we obtain the first order sensitivity index for X_i , which describes the proportion of variance in the model output caused by X_i :

$$S_i = \frac{V_i}{V_Y} \quad (6)$$

We can then decompose the unconditional model output variance V as the sum of conditional variances up to the k th order:

$$V_Y = \sum_{i=1}^k V_i + \sum_i \sum_{i < j} V_{ij} + \dots + V_{1,2,\dots,k} \quad (7)$$

where

$$\begin{aligned} V_{ij} = & V_{X_i, X_j} [E_{X_{\sim i,j}}(Y|X_i, X_j)] \\ & - V_{X_i} [E_{X_{\sim i}}(Y|X_i)] \\ & - V_{X_j} [E_{X_{\sim j}}(Y|X_j)] \end{aligned} \quad (8)$$

From this, we can derive the second-order index S_{ij} , which explains the proportion of variance due to the interaction between X_i and X_j :

$$S_{ij} = \frac{V_{ij}}{V_Y} \quad (9)$$

and so on until order k . However, estimating all terms in Eq. (7) is impractical when k is large, as they result in $2^k - 1$. In this case, we can compute the total order index or S_{Ti} , which measures the proportion of variance due to the first-order effect of X_i jointly with its interactions with the other parameters (Homma and Saltelli, 1996):

$$S_{Ti} = \frac{E_{X_{\sim i}} [V_{X_i}(Y|X_{\sim i})]}{V_Y} \quad (10)$$

For PAWN, since $k = 4$, we computed first (S_i), second (S_{ij}), third (S_{ijk}) and total-order (S_{Ti}) Sobol' indices of (N, n, ϵ, θ) . For S_{Ti}^* , since $k = 2$, we just computed S_i and S_{Ti} . In both settings we used the Saltelli et al. (2010) and the Jansen (1999) estimators to compute S_i and S_{Ti} respectively, as per the established best practices. All the workflow is summarized in Fig. S2 and the *R* code to replicate our results is available in GitHub.

3. Results

3.1. Uncertainty analysis

Fig. 2 presents the uncertainty distribution of PAWN and S_{Ti}^* for each model input and function. The results can be matched against Fig. S3, where we display how PAWN and Sobol' indices look like once their design parameters are fixed and the total number of model runs is set at $N = 4000$.

For PAWN, we ran three simulations (Fig. 2a):

1. With θ including the mean, the median and the maximum as possible summary statistics ($\text{max} \in \theta$ setting).
2. With θ including the mean and the median only ($\text{max} \notin \theta$ setting). This aimed at isolating the effect that extreme KS values, which might be obtained for specific conditioning intervals, have in the final PAWN index.
3. With $N \sim \mathcal{U}(2500, 4000)$, $n \sim \mathcal{U}(15, 20)$ and ($\text{max} \notin \theta$) ("Optimum" setting). This latter run reflects an ideal scenario, one in which the number of model runs needed to achieve convergence is known in advance and the uncertainty in n and θ is reduced to the minimum expression.

For S_{Ti}^* we run one simulation to assess its sensitivity to the uncertainty in the total number of model runs and the estimator (N, θ) [with $N \sim (200, 2000)$] (Fig. 2b). Sobol' indices might eventually take on values outside the range $[0, 1]$ due to numerical artefacts created during the computation. This is widely known among sensitivity analysts and managed by considering $(S_{Ti}^* < 0) \approx 0$ and $(S_{Ti}^* > 1) \approx 1$. Fig. 2b only presents $S_{Ti}^* \in [0, 1]$ to allow for a better comparison with the values produced by the PAWN index. A plot showing the distribution of values with $S_{Ti}^* \notin [0, 1]$ is presented in Fig. S4.

3.1.1. Factor prioritization

Fig. 2 shows that, in a factor prioritization context, i.e. when the aim is to sort the parameters according to their contribution to the model output variance (Saltelli et al., 2007), the uncertainty in the value of the design parameters might cause model inputs to overlap, thus raising the likelihood of producing a biased ranking. In order to get precise figures for this overlap, we computed the coefficient of overlapping, i.e. the area lying under the density curves of two different model inputs, following Pastore (2018) (see Fig. S5 for a presentation of Fig. 2 with density curves instead).

In the case of the Ishigami and Homma (1990) function, the overlap between X_2 and X_3 is of 10% if PAWN is used under $\text{max} \in \theta$ (4% under $\text{max} \notin \theta$), despite X_2 being non-influential. The overlap with S_{Ti}^* is 0. With the Sobol' (1998) G function, PAWN mistakes X_1 for X_2 and X_2 for X_3 11% and 20% of the time respectively if $\text{max} \in \theta$ (4% and 10% if $\text{max} \notin \theta$), and might even bias the ranking of X_2 and X_4 (14% and 9% overlap in $\text{max} \in \theta$ and $\text{max} \notin \theta$ respectively). On the contrary, if

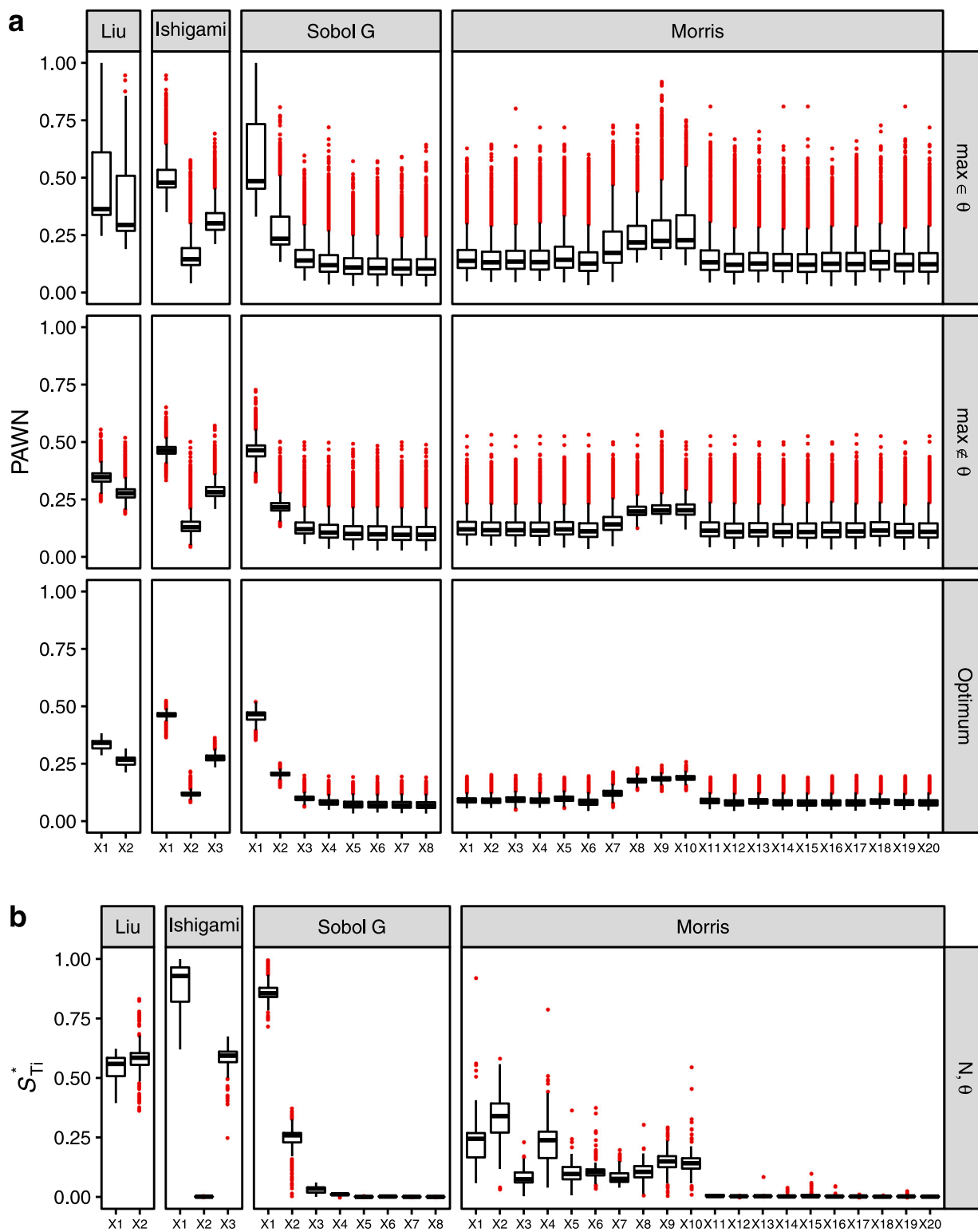


Fig. 2. Uncertainty in the computation of the sensitivity indices. Outliers are shown in red. (a) PAWN. (b) S_{Ti}^* .

S_{Ti}^* is used, the overlap between $X_1 - X_2$, $X_2 - X_3$ and $X_2 - X_4$ is ~0%, 2% and 1% respectively.

The case of the Liu et al. (2006) function deserves a specific comment. Fig. 2 shows that 29% of X_1 and X_2 values overlap if using PAWN under $max \in \theta$ (13% if $max \notin \theta$), whereas for S_{Ti}^* the degree of overlap is 55%. Liu et al. (2006), however, stated that X_1 was more influential than X_2 based on relative-entropy sensitivity methods. Pianosi and Wagener (2015) used the Liu et al. (2006) function to back up their

claim of PAWN outperforming Sobol' indices due to the former being able to discriminate the higher influence of X_1 much better than S_{Ti}^* . However, the analytical values of S_{Ti}^* in the Liu et al. (2006) function are identical at $X_1 = X_2 = 0.546$ (see Table 2 in Liu et al. (2006)). This means that the overlap between X_1 and X_2 shown in Fig. 2b is to be expected and does not result from the sensitivity of S_{Ti}^* to the uncertainties in its own structural design.

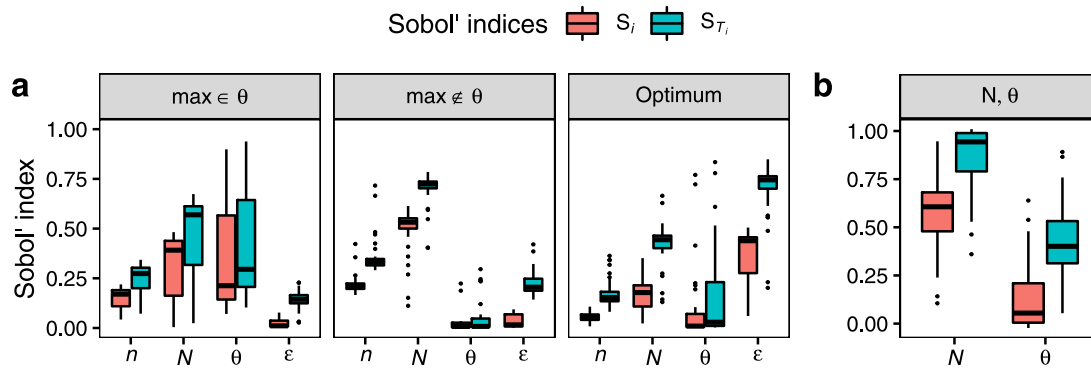


Fig. 3. Sobol' indices. (a) PAWN. (b) $S_{T_i}^*$.

3.1.2. Factor screening

In a factor screening context, i.e. when the aim is to distinguish influential from non-influential parameters (Saltelli et al., 2007), the uncertainty in the PAWN design parameters might also lead to erroneous results, regardless of whether $max \in \theta$ or $max \notin \theta$. In the case of the Sobol' (1998) G function, 10%–14% (7%–8%) of the probability density of X_2 , which has a non-nihil effect, overlaps with X_4, \dots, X_8 , which have no effect at all. The overlap between influential and non-influential model inputs if $S_{T_i}^*$ is used, on the other side, ranges between ~0% and 2%.

The poor screening power of PAWN stands out in the case of the Morris (1991) function: 18%–25% (1%–8% in $max \notin \theta$) of the probability density of X_8, \dots, X_{10} , which have a moderate first-order effect, overlap with that of X_{11}, \dots, X_{20} , which are non-influential (Fig. S2A). In the case of $S_{T_i}^*$, the overlap ranges in 0%–5%. The chances of PAWN mistaking relevant for non-relevant parameters is even higher in the case of parameters whose influence in the model output is through interactions only: this is the case of X_1, \dots, X_6 in the Morris (1991) function, whose degree of overlap with X_{11}, \dots, X_{20} , which are non-influential, range between 75%–95% (Figs. 2-S3). In all these cases, the uncertainty in the design parameters of $S_{T_i}^*$ leads to a 0%–4% overlap between influential and non-influential parameters.

Fig. 2 shows that in the “Optimum” setting the overlap between the model inputs is considerably reduced for PAWN. In the case of the Liu et al. (2006) and the Ishigami and Homma (1990) functions, the percentage of overlap goes down to zero. However, the chances of wrongly ranking/screening the model inputs remain non-negligible for both the Sobol' (1998) G and the Morris (1991) functions. In the former, there is 17%–28% overlap between the slightly influential model input X_3 and X_4, \dots, X_8 , whose effect cannot be differentiated from the approximation error. With regards to the Morris (1991) function, the chances of characterizing as non-influential parameters that have a significant non-additive effect in the model output remain very high: the overlap of X_1, \dots, X_6 with X_{11}, \dots, X_{20} , for instance, range between 35%–90% (Fig. 2). As shown in Fig. S3, the volatility in the computation of PAWN does not allow to distinguish X_{11}, \dots, X_{20} from a dummy, non-influential model input.

3.2. Sensitivity analysis

Fig. 3 presents the Sobol' first (S_i) and total (S_{T_i}) indices for each of the settings of our analysis after pooling the values from all functions and parameters [the Sobol' indices for each function and design parameter are shown in Figs. S6–S7 (PAWN), and Figs. S8–S9 ($S_{T_i}^*$)]. Fig. 3 thus informs on how much uncertainty each design parameter contributes to each sensitivity index. In Fig. S10 of the Supplementary Information file, we prove that the results displayed in Fig. 3 are robust without having to correct the stronger weight that the Morris function might have in defining the trends due to its much larger number of model inputs.

As shown in Fig. 3a, the first-order effect of θ and N is much more variable in the $max \in \theta$ setting, suggesting that their degree of contribution to the PAWN index uncertainty might considerably be function-dependent. This variability is highly reduced in the $max \notin \theta$ setting, which provides a more robust account of the extent to which each design parameter contributes to define the PAWN index. In this setting, the selection of the total number of model runs (N) and the number of conditioning intervals (n) convey up to c. 60% and c. 20% of the PAWN index uncertainty respectively. The stochasticity in the sampling of $Y_U(\epsilon)$ (ϵ) is mostly influential through interactions in both settings, whereas the selection of the summary statistic (θ) has a nearly nihil effect in the $max \notin \theta$ setting. Remarkably, interactions are also significant in the “Optimum” setting, especially those involving ϵ , n and N .

To gain further insights into the structure of the non-additivities in PAWN, we computed second and third-order effects, shown in Fig. 4 (the second and third-order Sobol' indices for each function and parameter are shown in Figs. S11–S14). In the $max \notin \theta$ setting, the interactions that have a significant effect on the model output involve the initial sample size (N) with the number of conditioning intervals (n) or the stochasticity in the sampling of $Y_U(\epsilon)$ (ϵ). Such second-order effects might contribute up to 15% of the PAWN index uncertainty. This also applies to the “Optimum” setting, where the interaction between N and ϵ or n and ϵ has an even higher effect (up to 25% uncertainty). These three design parameters have significant third-order interactions in all three settings (Fig. 4b).

Regarding $S_{T_i}^*$ (Fig. 3b), both N and θ have a non-negligible first-order effect on the index. The wide boxplots suggest that both the extent of their influence and the degree of non-additivities in defining the final $S_{T_i}^*$ index depends on the model under scrutiny (Figs. S8–S9).

4. Discussion and conclusions

Sensitivity analysis is an important tool to check the robustness of a model in the context of its validation. However, also the measurement of the sensitivity of the output variables to input parameters rests on modelling hypotheses, i.e. sensitivity analysis is a modelling process per se, based on the use of an algorithm — like the models being investigated. Here we tested the robustness of this aspect of the modelling process by assessing the dependency of the PAWN index to its design parameters, and matched the results against the Sobol' total-order index ($S_{T_i}^*$). Our work thus sheds some light on the uncertainties concealed in the last stage of any sensitivity analysis: the implementation of the sensitivity algorithm. This practice could help ensuring the adequacy and the range of applicability of the tools which modellers deploy to improve the quality of their modelling exercises.

Two elements emerge from our work: the PAWN index is more sensitive to the design parameters than $S_{T_i}^*$, and this sensitivity has a complex pattern which makes the use of PAWN, or better the tuning of the PAWN design parameters, a delicate task. The chances of incurring

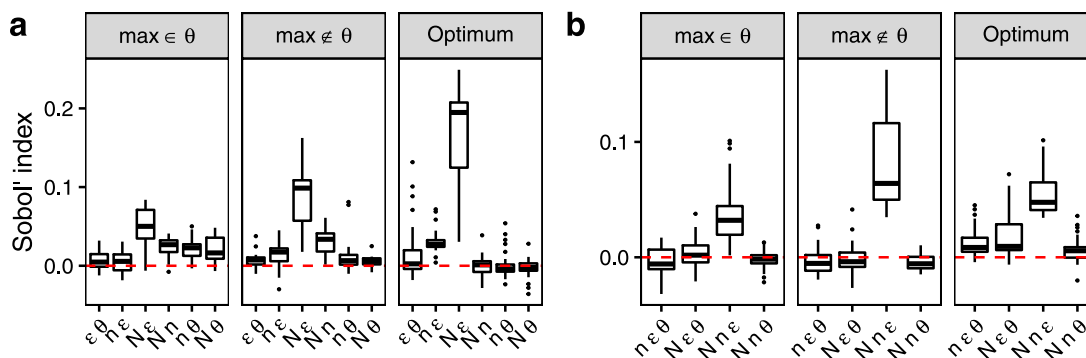


Fig. 4. High-order interactions between the PAWN design parameters. Only those boxplots with all values above the red dashed horizontal line reflect a true effect. (a) Second-order interactions. (b) Third-order interactions.

in false positives (mistaking a non-influential parameter for influential, with a waste of computational resources) and/or false negatives (mistaking an influential parameter for non-influential, with a loss of useful information) are significant even when the design parameter space does not include the maximum as a possible summary statistic, but only measures of central tendency. Even in an ideal setting, where the uncertainty in n and θ is highly reduced and N ensures convergence, the PAWN index might be incapable to differentiate between non-influential model inputs and influential model inputs whose effect in the model output is fully through interactions. Mora et al. (2019) also observed that the high level of noise in PAWN might produce biased rankings when the model inputs have different orders of magnitude in their contribution to the response.

PAWN especially underperformed when used in the Sobol' (1998) ($k = 8$) and the Morris (1991) function ($k = 20$). This raises a red flag for analysts when using PAWN in high-dimensional models such as those commonly being employed in the Earth and Environmental Sciences domain, which might easily include tens of parameters (Sheikholeslami et al., 2019). In these contexts, the use of PAWN might indeed allow to significantly reduce the number of model runs required to conduct a sensitivity analysis, but at the expense of significant risks of obtaining a biased result due to its structural design.

The fact that the sensitivity of PAWN to its design parameters is complex, including important interactions up to the third order effect, implies that finding the perfect range of design parameters to use PAWN safely and efficiently is not easy. The significant non-additivity of PAWN only unfolds once the values of its main design parameters are moved simultaneously within reasonable uncertainty ranges; this is, when all the forking paths and divergences leading towards its computation are assessed at once. Instead, in the paper where the “generic approach” is presented, Pianosi and Wagener (2018) analysed the influence of (N, n, θ) on PAWN by combining different discrete point-estimates for N and n (see their Fig. 3), or by changing the value for either N or n while keeping the other design parameters fixed (see their Figs. 6 and 7). This approach is very similar to a one-at-a-time (OAT) sensitivity analysis, a method that may fail to detect interactions between model parameters due to its incomplete examination of the uncertainty space (Saltelli and Annoni, 2010). Our work shows that even by increasing N , ensuring a “high” N/n ratio or fixing θ at a central tendency measure, significant interactions between N and n , N and ϵ or between N , n and ϵ make PAWN considerably difficult to tame.

In our study we also observed the existence of potential interactions between the total number of model runs N and the estimator θ for $S_{T_i}^*$. However, the variance in $S_{T_i}^*$ deriving from the uncertainty in its design parameters is comparatively much smaller, thus reducing the chances of obtaining a biased ranking or producing a wrong screening. It should be stated that our research design for $S_{T_i}^*$ included the use of different estimators out of our willingness to fully explore its potential

uncertainty space. However, there is actually no uncertainty in θ when computing Sobol' total-order indices: Saltelli et al. (2010) showed that the best estimate of the degree to which a given model input interacts with the rest is obtained with the Jansen (1999) estimator, which should be the default choice. This means that the uncertainty in the design parameters of $S_{T_i}^*$ actually narrows down to the total number of model runs N only. It is thus to be expected that given two algorithms, the one depending on a higher number of design parameters will be the more delicate to use. A similar discussion led Campolongo et al. (2011) to conclude that the screening method of Morris could be effectively substituted by $S_{T_i}^*$.

We would like to stress that there are many other uncertainties beyond those embedded in sensitivity indices that have the potential to condition the results of a given sensitivity analysis. These do not only involve quantifiable elements such as the selection of the distributions, sample size or the model output form (Kaplan and Garrick, 1981; Aven, 2020; Noacco et al., 2019), but also uncertainties in the entire knowledge and model-generating processes (Saltelli et al., 2013). Algorithmic uncertainties ingrained in the design parameters of sensitivity indices, however, are especially important for they condition the computational path towards the final sensitivity value. Although several alternative routes might be available, these multiple options are eliminated once the computation is triggered, yielding the illusion of a deterministic process. We believe that a good way to know whether a sensitivity index is robust is to check how volatile it is when all paths involved in its effective computation are walked at once.

The present findings do not suggest discarding PAWN as a sensitivity measure. Moment independent measures have a role to play in sensitivity analysis of output with long-tailed distributions. Additionally, they may find an ideal use in settings where the output of interest is itself in the form of a difference between two cumulative distributions. The present analysis aims at encouraging developers of sensitivity indices to fully explore the structural uncertainty of their algorithms in order to deliver transparent and robust sensitivity tools.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Francesca Pianosi, Razi Sheikholeslami, Thorsten Wagener and two anonymous reviewers for their constructive comments on previous versions of this manuscript. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 792178 (A.P.).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.envsoft.2020.104679>.

References

- Aven, T., 2020. Three influential risk foundation papers from the 80s and 90s: Are they still state-of-the-art?. *Reliab. Eng. Syst. Saf.* 193 (February 2019), 106680. <http://dx.doi.org/10.1016/j.res.2019.106680>.
- Borges, J., 1941. El Jardín de los Senderos que se Bifurcan, Buenos Aires.
- Borgonovo, E., 2007. A new uncertainty importance measure. *Reliab. Eng. Syst. Saf.* 92 (6), 771–784. <http://dx.doi.org/10.1016/j.res.2006.04.015>.
- Borgonovo, E., Castaings, W., Tarantola, S., 2011. Moment independent importance measures: new results and analytical test cases. *Risk Anal.* 31 (3), 404–428. <http://dx.doi.org/10.1111/j.1539-6924.2010.01519.x>.
- Borgonovo, E., Tarantola, S., Plischke, E., Morris, M.D., 2014. Transformations and invariance in the sensitivity analysis of computer experiments. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (5), 925–947. <http://dx.doi.org/10.1111/rssb.12052>.
- Campolongo, F., Saltelli, A., Cariboni, J., 2011. From screening to quantitative sensitivity analysis. A unified approach. *Comput. Phys. Comm.* 182 (4), 978–988. <http://dx.doi.org/10.1016/j.cpc.2010.12.039>.
- Gelman, A., Loken, E., 2013. The garden of forking paths : Why multiple comparisons can be a problem , even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. <https://osf.io/n3axs/>.
- Holl, M., Janke, T., Pelz, P.F., Platzer, M., 2016. Sensitivity analysis of a techno-economic optimal wind-energy converter. In: *International Conference on Next Generation Wind Energy (ICNGWE)*, pp. 1–24.
- Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* 52, 1–17. [http://dx.doi.org/10.1016/0951-8320\(96\)00002-6](http://dx.doi.org/10.1016/0951-8320(96)00002-6).
- Ishigami, T., Homma, T., 1990. An importance quantification technique in uncertainty analysis for computer models. In: *Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, vol. 12, pp. 398–403.
- Jansen, M., 1999. Analysis of variance designs for model output. *Comput. Phys. Comm.* 117 (1–2), 35–43. [http://dx.doi.org/10.1016/S0010-4655\(98\)00154-4](http://dx.doi.org/10.1016/S0010-4655(98)00154-4).
- Kaplan, S., Garrick, J.B., 1981. On the quantitative definition of risk. *Risk Anal.* 1 (1), 11–27. <http://dx.doi.org/10.1093/annhyg/24.2.245>.
- Khorashadi Zadeh, F., 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. *Environ. Model. Softw.* 91, 210–222. <http://dx.doi.org/10.1016/j.envsoft.2017.02.001>.
- Liu, H., Chen, W., Sudjianto, A., 2006. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Trans. ASME, J. Mech. Des.* 128 (2), 326–336. <http://dx.doi.org/10.1115/1.2159025>.
- Mora, E.B., Spelling, J., van der Weijde, A.H., 2019. Benchmarking the PAWN distribution-based method against the variance-based method in global sensitivity analysis: Empirical results. *Environ. Model. Softw.* 122 (September), <http://dx.doi.org/10.1016/j.envsoft.2019.104556>.
- Morris, M., 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics* 33 (2), 161–174.
- Noacco, V., Sarrazin, F., Pianosi, F., Wagener, T., 2019. Matlab/R workflows to assess critical choices in global sensitivity analysis using the SAFE toolbox. *MethodsX* 6, 2258–2280. <http://dx.doi.org/10.1016/j.mex.2019.09.033>.
- Paleari, L., Confalonieri, R., 2016. Sensitivity analysis of a sensitivity analysis: We are likely overlooking the impact of distributional assumptions. *Ecol. Model.* 340, 57–63. <http://dx.doi.org/10.1016/j.ecolmodel.2016.09.008>.
- Pappenberger, F., Beven, K.J., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. *Adv. Water Resour.* 31 (1), 1–14. <http://dx.doi.org/10.1016/j.advwatres.2007.04.009>.
- Pastore, M., 2018. Overlapping: a R package for estimating overlapping in empirical distributions. *J. Open Source Softw.* 3 (32), 1023. <http://dx.doi.org/10.21105/joss.01023>.
- Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environ. Model. Softw.* 67, 1–11. <http://dx.doi.org/10.1016/j.envsoft.2015.01.004>.
- Pianosi, F., Wagener, T., 2018. Distribution-based sensitivity analysis from a generic input-output sample. *Environ. Model. Softw.* 108, 197–207. <http://dx.doi.org/10.1016/j.envsoft.2018.07.019>.
- Saltelli, A., Andres, T.H., Homma, T., 1993. Sensitivity analysis of model output. An investigation of new techniques. *Comput. Statist. Data Anal.* 15 (2), 211–238. [http://dx.doi.org/10.1016/0167-9473\(93\)90193-W](http://dx.doi.org/10.1016/0167-9473(93)90193-W).
- Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. *Environ. Model. Softw.* 25 (12), 1508–1517. <http://dx.doi.org/10.1016/j.envsoft.2010.04.012>.
- 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. *Comput. Phys. Comm.* 181 (2), 259–270. <http://dx.doi.org/10.1016/j.cpc.2009.09.018>.
- Saltelli, A., Guimaraes Pereira, A., van der Sluijs, J.P., Funtowicz, S.O., 2013. What do I make of your latinorum? sensitivity auditing of mathematical modelling. *Int. J. Innov. Policy* 9 (2/3/4), 213–234. <http://dx.doi.org/10.1504/IJFIP.2013.058610>, [arXiv:1211.2668](https://arxiv.org/abs/1211.2668).
- Saltelli, A., Homma, T., 1992. Sensitivity analysis for model output. performance of black box techniques on three international benchmark exercises. *Comput. Statist. Data Anal.* 13 (1), 73–94. [http://dx.doi.org/10.1016/0167-9473\(92\)90155-9](http://dx.doi.org/10.1016/0167-9473(92)90155-9).
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2007. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, Chichester, UK, <http://dx.doi.org/10.1002/9780470725184>.
- Sheikholeslami, R., Razavi, S., Gupta, H.V., Becker, W., Haghnegahdar, A., 2019. Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. *Environ. Model. Softw.* 111 (June 2018), 282–299. <http://dx.doi.org/10.1016/j.envsoft.2018.09.002>.
- Shin, M.J., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J., 2013. Addressing ten questions about conceptual rainfall-runoff models with global sensitivity analyses in R. *J. Hydrol.* 503, 135–152. <http://dx.doi.org/10.1016/j.jhydrol.2013.08.047>.
- Sobol', I.M., 1993. Sensitivity analysis for nonlinear mathematical models. *Math. Model. Comput. Exp.* 1 (4), 407–414. <http://dx.doi.org/10.18287/0134-2452-2015-39-4-459-461>.
- Sobol', I.M., 1998. On quasi-Monte Carlo integrations. *Math. Comput. Simulation* 47 (2–5), 103–112. [http://dx.doi.org/10.1016/S0378-4754\(98\)00096-2](http://dx.doi.org/10.1016/S0378-4754(98)00096-2).
- Sobol', I., 2007. Global sensitivity analysis indices for the investigation of nonlinear mathematical models. *Mat. Model.* 19 (11), 23–24.