



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>

Peer reviewed version

License (if available):  
CC BY-NC-ND

Link to published version (if available):  
[10.1016/j.envsoft.2018.07.019](https://doi.org/10.1016/j.envsoft.2018.07.019)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Elsevier at <https://www.sciencedirect.com/science/article/pii/S1364815218303220> . Please refer to any applicable terms of use of the publisher.

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

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

Francesca Pianosi <sup>a,b</sup>, Thorsten Wagener <sup>a,b</sup>  
(corresponding author: francesca.pianosi@bristol.ac.uk)

<sup>a</sup>*Department of Civil Engineering, University of Bristol, University Walk,  
BS81TR, Bristol, UK*

<sup>b</sup>*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

---

## 1 Highlights

- 2 • We introduce a new approximation strategy for PAWN indices
- 3 • The strategy is applicable to a generic input-output sample and uses one
- 4 tuning parameter only
- 5 • We demonstrate that the strategy provides robust PAWN sensitivity es-
- 6 timates
- 7 • This approximation strategy facilitates the integration of PAWN into
- 8 multi-method GSA

## 9 Software availability

10 The PAWN algorithm, including the new approximation strategy presented  
11 in this paper, are implemented in Matlab/Octave as part of the SAFE Tool-  
12 box, which is freely available for non-commercial use through the website:  
13 [www.safetoolbox.info](http://www.safetoolbox.info)

## 14 1 Introduction

15 Global Sensitivity Analysis (GSA) is a set of techniques aimed at investigating  
16 the propagation of uncertainty through mathematical models in a structured  
17 way. More specifically, according to the widely used definition by Saltelli et al.  
18 (2008), the aim of GSA is to quantify the relative contribution of the uncer-  
19 tain input factors of a mathematical model to the variability of its outputs.  
20 For model developers, such quantification can aid the process of identifying  
21 a minimum complexity model by eliminating non-influential components. For  
22 model users, it can make the calibration process more efficient by determining  
23 the subset of model parameters whose reduction in uncertainty would mostly  
24 reduce output variability, or it can be used to assess the robustness of the  
25 model predictions against various sources of uncertainty such as errors in the  
26 forcing data or even in uncertain modelling assumptions. GSA is therefore  
27 widely applied in the environmental modelling field to support the construc-  
28 tion, improvement and use of mathematical models (e.g. Beven and Binley  
29 (1992); Spear et al. (1994); Freer et al. (1996); Bastidas et al. (1999); Wagener  
30 and Kollat (2007); Norton (2015); Razavi and Gupta (2015); Xiaomeng et al.  
31 (2015); Ferretti et al. (2016); Pianosi et al. (2016); Petropoulos and Srivastava  
32 (2017)).

33 Many different GSA methods are available depending on the specific purposes  
34 of the analysis as well as the characteristics of the mathematical model be-  
35 ing analysed and its sources of uncertainty (Saltelli et al., 2008; Norton, 2015;  
36 Pianosi et al., 2016). Among them, some of the most widely used are Variance-  
37 Based Sensitivity Analysis (VBSA) methods, which measure output sensitiv-  
38 ity as the proportion of output variance that is attributable to variations of  
39 each uncertain input factor. For an overview of variance-based methods and  
40 their advantages see for example Saltelli et al. (2008) or Pianosi et al. (2016).  
41 Recently, density-based approaches have also gained increasing attention Cas-  
42 taings et al. (2012); Anderson et al. (2014); Peeters et al. (2014); Dell’Oca  
43 et al. (2017); Borgonovo et al. (2017). In these approaches, uncertainty and  
44 sensitivity is characterised by investigating the entire distribution of the model  
45 outputs, instead of its variance only. For this reason, such methods are also  
46 referred to as *distribution-based* or *moment-independent*. Distribution-based

47 strategies are particularly suitable when variance is not an adequate proxy  
48 of uncertainty, for example when the output distribution is highly-skewed or  
49 multi-modal (e.g. Liu et al. (2006)).

50 In a previous paper (Pianosi and Wagener, 2015) we introduced a distribution-  
51 based method, called PAWN, and implemented it as part of our open-source  
52 GSA Toolbox called SAFE (Pianosi et al., 2015). The advantage of PAWN  
53 over other moment-independent methods is that it characterises output dis-  
54 tributions by their cumulative distribution functions, instead of their proba-  
55 bility density functions, which makes the numerical approximation of PAWN  
56 sensitivity indices easy and robust. In Pianosi and Wagener (2015) we demon-  
57 strated the PAWN method by applying it to a standard benchmark function  
58 and a simple rainfall-runoff model (Hymod). In Zadeh et al. (2017) we carried  
59 out a systematic comparison between PAWN and variance-based method on  
60 a medium complexity (26 parameters) hydrological model (SWAT) and found  
61 that PAWN and VBSA had similar convergence rate and screening results,  
62 while PAWN was more effective for parameter ranking as it could better sep-  
63 arate out the relative importance of the influential parameters. Since its pub-  
64 lication, PAWN has been used to investigate the role of uncertain parameters  
65 across a range of environmental modelling fields, including: a transport model  
66 of indoor air pollutant (Sedighian et al., 2015), a computational model of bio-  
67 logical processes (Gillies et al., 2016), rainfall-runoff and land-surface models  
68 in Pianosi and Wagener (2016) and Pianosi et al. (2017), a fluid flow and heat  
69 transport model in geothermal reservoirs (Fox et al., 2016), a groundwater  
70 model for karst systems (Hosseini et al., 2017), and a numerical algorithm for  
71 hillslope-based landscape discretisation (Pilz et al., 2017).

72 Despite this relatively quick uptake of PAWN across different fields of applica-  
73 tion, from our own experience and the feedbacks we received from other users,  
74 we think two main issues remain critical. First, the numerical procedure we  
75 proposed in our original paper to implement PAWN uses a *tailored* sampling  
76 strategy, i.e. a strategy that selects input samples in specific regions of the  
77 input variability space, according to the approximation procedure set out for  
78 the PAWN indices. This is in contrast to *generic* sampling strategies, such as  
79 sampling over a uniform grid, quasi-random sampling (Press et al., 1992) or  
80 (stratified or not) random sampling, e.g. Latin Hypercube (Forrester et al.,  
81 2008), which aim at spreading input samples as uniformly as possible across  
82 the variability space, and can be used across a range of uncertainty and sensi-  
83 tivity analysis methods. The requirement of a tailored sampling strategy thus  
84 makes it more difficult to integrate PAWN into a multi-method GSA study,  
85 such as Pappenberger et al. (2008) or Tang et al. (2007), since its inclusion  
86 would require additional dedicated model evaluations. We believe that this  
87 is a strong limitation given the value of applying multiple GSA methods to  
88 the same problem as a way to validate and complement the results of indi-  
89 vidual methods (Pianosi et al., 2015; Borgonovo et al., 2017). Additionally,

90 the requirement of a tailored sampling strategy prevents the application to  
91 an existing input-output dataset in cases where such a dataset is available  
92 from previous studies. These limitations have motivated researchers to seek  
93 for generic approximation strategies for other GSA methods too, including  
94 variance-based methods. For example, Strong et al. (2014) and Stanfill et al.  
95 (2015) have proposed new approximation strategies to derive first-order and  
96 total-order indices from a generic input-output dataset, as an alternative to  
97 the ‘traditional’ approximators (e.g. Saltelli et al. (2010)) based on ‘re-sample’  
98 matrices, which require a tailored sampling strategy. A general discussion of  
99 the value of approximation procedures that can be applied to given data is  
100 given in Plischke et al. (2013).

101 The second issue with our tailored sampling strategy is that it requires users  
102 to specify three tuning parameters, i.e. the number of unconditional output  
103 samples ( $N_u$ ), the number of conditional output samples ( $N_c$ ), and the number  
104 of conditioning points ( $n$ ). As discussed in Pianosi and Wagener (2015), the  
105 choice of these tuning parameters should be based on a compromise between  
106 approximation accuracy and computational burden, which both increase with  
107 any increase of  $N_u$  or  $N_c$  or  $n$ . In fact, the total number of model evaluations  
108 to approximate all PAWN indices is  $N=N_u+n\times N_c\times M$ , where  $M$  is the num-  
109 ber of uncertain input factors. If each model evaluation is computationally  
110 demanding, either in terms of running time or data storage requirement, one  
111 would like to find the ‘optimal’ combination of  $(N_u, N_c, n)$  to reach sufficient  
112 approximation accuracy at minimum  $N$ . However, such optimal values are  
113 difficult to predict a priori and extrapolating from previous applications may  
114 be risky because the optimal values may change with the problem at hand,  
115 i.e. with the mathematical model, the number of input factors, and possibly  
116 even with the output definition or application domain (Sarrazin et al., 2016).  
117 We indeed know that the approximation accuracy associated with sensitivity  
118 indices at a given sample size can dramatically change with any element of  
119 the experimental set-up, as shown for example in Figure 5 in Pianosi et al.  
120 (2016) or Figures 2 and 3 in Zadeh et al. (2017).

121 In this paper we simultaneously address these two issues by introducing a  
122 new approximation procedure of the PAWN indices that (1) is applicable to  
123 a generic dataset; (2) requires fewer tuning parameters (essentially only the  
124 number of conditioning points  $n$ ) whose choice is easier to make and to evalu-  
125 ate. The approximation procedure was already sketched out in the conclusions  
126 of Pianosi and Wagener (2015) and a similar idea was tested in Pianosi et al.  
127 (2017). Here we further develop those ideas into a new approximation proce-  
128 dure. We test it comprehensively on a benchmark function and on a complex  
129 hydrological model (the Soil Water Assessment Tool, in a set-up that includes  
130 50 uncertain parameters). And finally, we propose a number of simple tools to  
131 assess the accuracy of the resulting PAWN indices as well as their robustness  
132 to the chosen tuning parameter, at negligible additional computing costs.

## 133 2 Methods

134 In this paper, we consider an input-output relationship

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

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

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

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

### 160 2.1 The PAWN method

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

167 A review of several distribution-based methods is given in Pianosi and Wa-  
 168 genger (2015). The distinctive feature of PAWN is that, in contrast to other  
 169 methods, it uses (conditional and unconditional) cumulative distribution func-  
 170 tions (CDFs) of the output instead of probability density functions. The ad-  
 171 vantage of using CDFs is that their approximation from an output sample of  
 172 finite size is easy and robust. Several other advantages of PAWN are discussed  
 173 in Pianosi and Wagener (2015).

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

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

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

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

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

## 188 2.2 Approximating PAWN indices using a tailored sampling strategy

189 In general, given the complexity of the input-output relationship  $f$ , the sensi-  
 190 tivity indices of Eq. (2) cannot be computed analytically and they need to be  
 191 approximated numerically. Pianosi and Wagener (2015) proposed an approx-  
 192 imation procedure based on two simplifications. First, using a finite number  
 193 of conditioning points  $\bar{x}_i^{(1)}, \bar{x}_i^{(2)}, \dots, \bar{x}_i^{(n)}$  for each input factor, instead of all its  
 194 possible values. Second, replacing the distributions  $F_y$  and  $F_{y|x_i}$  by the em-  
 195 pirical distributions  $\hat{F}_y$  and  $\hat{F}_{y|x_i}$  of output samples of finite size. Specifically,  
 196  $\hat{F}_y$  is the empirical distribution of an unconditional sample (YU) obtained by  
 197 varying all input factors simultaneously, and  $\hat{F}_{y|x_i}$  is the empirical distribution  
 198 of a conditional sample (YC <sub>$ik$</sub> ) obtained by varying all factors but the  $i$ -th,  
 199 which is set to the  $k$ -th conditioning value  $\bar{x}_i^{(k)}$ . The PAWN sensitivity index

200 is then approximated by

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

201 The left hand side of Figure 1 provides a visual illustration of this approx-  
 202 imation strategy for the simple case of  $M=3$  input factors. For the sake of  
 203 illustration, the Figure focuses on approximating the PAWN sensitivity index  
 204 of the first input factor ( $x_1$ ). The top left panels (Fig. 1(a) and (b)) show the  
 205 combinations of input factors ( $x_1, x_2, x_3$ ) that need to be evaluated in order to  
 206 obtain the unconditional sample and three conditional samples correspond-  
 207 ing to  $n=3$  fixed values of  $x_1$ . The corresponding output samples (YU, YC<sub>11</sub>,  
 208 YC<sub>12</sub>, YC<sub>13</sub>) are visualised via a scatter plot in Fig. 1(c). The lower panels  
 209 show the further steps for computing the approximate PAWN index  $\hat{S}_1$ : com-  
 210 puting the empirical distributions of YU (red line in (g)) and of YC<sub>11</sub>, YC<sub>12</sub>  
 211 and YC<sub>13</sub> (grey lines), computing the KS at each conditioning point (h), and  
 212 finally taking a statistic, e.g. the median, of those KS values. A similar pro-  
 213 cedure would be applied for approximating the sensitivity indices of  $x_2$  and  
 214  $x_3$ .

215 We call the approach underpinning Eq. (4) a *tailored sampling strategy* because  
 216 a large part of the input samples generated to compute the sensitivity indices,  
 217 namely all those in the conditional samples YC<sub>*ik*</sub> for  $k = 1, \dots, n$ , are concen-  
 218 trated in specific subregions of the input variability space (e.g. the planes in  
 219 Fig.1(b) where the grey circles lie). This is in contrast to *generic sampling*  
 220 *strategies* that would spread input samples as evenly as possible across the  
 221 input space (e.g. the samples in Fig. 1(d)-(e)). Examples of generic sampling  
 222 strategies include latin hypercube sampling (e.g. Sec. 1.4 in Forrester et al.  
 223 (2008)) or quasi-random sampling (e.g. Sec. 7.7 in Press et al. (1992)). Notice  
 224 that while the input samples in YC<sub>*ik*</sub> may be generated by applying a generic  
 225 sampling strategy in the  $(M - 1)$ -dimensional space of all-inputs-but-the-*i*-th  
 226 (for instance, we will use latin hypercube sampling in the following case study  
 227 applications), collectively the ensemble of conditional samples YC<sub>*ik*</sub> does not  
 228 constitute a *generic* dataset in the  $M$ -dimensional input variability space, as  
 229 clearly shown in Fig. 1(b).

230 With the tailored sampling strategy, the total number of model evaluations  
 231 to approximate all PAWN sensitivity indices is  $N_u + n \times N_c \times M$ , where  $N_u$  is  
 232 the size of the unconditional sample YU,  $N_c$  is the size of each conditional  
 233 sample YC<sub>*ik*</sub>, and  $M$  is the number of input factors (and hence sensitivity in-  
 234 dices). As discussed in the Introduction, the issue of how to choose the triple  
 235  $(N_u, n, N_c)$  has not been formally investigated and it remains an open question  
 236 in the application of PAWN. This choice is critical given that it affects both  
 237 the accuracy of the PAWN indices and the computational effort (total num-  
 238 ber of model evaluations) to generate them. Another issue with the tailored  
 239 strategy is that much of the computational effort is invested in generating



240 the conditional samples  $YC_{ik}$ , which cannot be re-used in other uncertainty or  
 241 sensitivity analysis methods that would require a generic sample. To overcome  
 242 these two issues we present a novel approach to approximate PAWN indices  
 243 from a generic dataset in the next section.

### 244 2.3 Approximating PAWN indices from a generic dataset

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

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

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

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

273 In either case, the main point here is that both  $N_c$  and  $N_u$  do not need to be  
 274 chosen by the user but they are determined as a consequence of the chosen  
 275 value of  $n$  and  $N$ . This is an advantage with respect to the tailored sampling

276 approach because the number of tuning parameters is reduced to two (instead  
277 of three) but most importantly because selecting their values is much easier. In  
278 fact, when using a generic dataset the computational effort for approximating  
279 the PAWN sensitivity indices is fully controlled by the chosen value for  $N$ .  
280 Hence, an obvious choice is to take the largest value possible for  $N$  given  
281 available computing resources. As for  $n$ , it only has an effect on the splitting  
282 of the input-output dataset  $\langle X, Y \rangle$  but not on its generation. Consequently,  
283 one can attempt different values of  $n$  and evaluate the robustness of GSA  
284 results to this choice without significantly adding to the overall computational  
285 effort. Further ways to assess the robustness of PAWN sensitivity indices to  
286 the chosen sample size  $N$  are discussed in the next subsection and will be  
287 illustrated in the Results section.

#### 288 2.4 Estimating the robustness of PAWN indices

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

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

314 In the case of PAWN sensitivity indices, a dummy parameter should in prin-

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

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

$$\hat{S}_{\text{dummy}} = \text{mean}_{k=1, \dots, n} \max_y |\hat{F}_y^{(k)}(y) - \hat{F}_y^{(n+1)}(y)| \quad (6)$$

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

### 338 3 Results

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

351 to illustrate a simple approach to evaluate the effects of the tuning parameter  
352  $n$  on PAWN sensitivity indices.

### 353 3.1 Application to Ishigami-Homma function

354 We first consider the Ishigami-Homma function

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

355 where  $x_i \sim \mathcal{U}[-\pi, +\pi]$  for  $i=1,2,3$  and  $a=2$  and  $b=1$ . This function is often used  
356 in GSA studies because the variance-based sensitivities of  $y$  can be calculated  
357 analytically (see for instance Chapter 4 in Saltelli et al. (2008)), which makes  
358 it an ideal testing ground of approximate sensitivity estimators. In particular,  
359 the first-order ( $S_i^F$ ) and total-order ( $S_i^T$ ) variance-based sensitivity indices  
360 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  
361 to these indices,  $x_1$  is by far the most influential input,  $x_2$  has very limited  
362 influence and no interactions,  $x_3$  is only influential through interactions with  
363  $x_1$ .

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

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

385 We further explore the impact of the sample size ( $N$ ) and of the chosen num-  
386 ber of conditioning intervals ( $n$ ) in Figure 3. Each panel refers to a different

387 input factor, and it shows the median KS (again, average over 50 bootstrap  
 388 resamples) for different choices of  $n$  (horizontal axis) and for different sample  
 389 size (color). For each combination  $(N, n)$ , the Figure also shows the estimated  
 390 PAWN sensitivity of the dummy parameter, computed by Eq. (6) (dashed  
 391 line). The Figure shows that:

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

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

### 423 3.2 Application to the SWAT model

424 The Soil and Water Assessment Tool (SWAT) is a semi-distributed hydro-  
 425 logical model developed by the USDA Agricultural Research Service (Arnold

426 et al., 1998) and used worldwide to study the impact of land use and man-  
427 agement practices on water quantity and quality at the catchment scale (e.g.  
428 Gassman et al. (2007)). Here, we use a model set-up for the upper Senne River  
429 basin in Belgium, which is described in Leta et al. (2015) and was used in a  
430 previous GSA study by Sarrazin et al. (2016) and comprises 50 uncertain pa-  
431 rameters. The model output  $y$  considered in the GSA is a performance metric,  
432 the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), measuring the distance  
433 between daily flow predictions of the model and available observations. More  
434 information about the model, the application river basin and the definition of  
435 inputs and output for GSA can be found in Sarrazin et al. (2016). A list of the  
436 50 model parameters and their variability ranges is given in the Supplementary  
437 Material of this paper. Here we re-use the input-output dataset generated for  
438 the Regional Sensitivity Analysis in Sarrazin et al. (2016), obtained by Latin  
439 Hypercube sampling and including  $N = 10,000$  samples.

440 First, we approximate the PAWN sensitivity indices by Eq. (5) using  $n=10$   
441 conditioning intervals. We repeat our calculations with 100 bootstrap resam-  
442 ples and compute the averages and confidence intervals of each index, shown  
443 in Figure 4. According to this figure, the most influential parameter is the  
444 11th, followed by parameters 9,32,10. Given that the confidence intervals for  
445 these three parameters are mostly overlapping, we cannot make further dis-  
446 tinctions between them and thus we would put all of them in the 2nd position  
447 of the parameter ranking. Following the same line of reasoning, we would put  
448 parameters 8,17,2,43 in the 3rd position and parameters 14,25,42,34,12,28,35  
449 in the 4th. The remaining parameters have an average sensitivity value close  
450 to that of the dummy parameter (red line in Figure 4), which means we can-  
451 not distinguish whether they actually have an influence on the output or  
452 whether their estimated sensitivity is a pure product of approximation errors.  
453 Hence, we classify them as potentially uninfluential. These ranking results  
454 are consistent with those obtained by Sarrazin et al. (2016) using other GSA  
455 methods. Figure 5 provides a short comparison with those results, focusing in  
456 particular on the top positions of the parameter ranking. The fact that only  
457 a limited number of parameters (4 to 8 in our case) control the output per-  
458 formance metric (Nash-Sutcliffe efficiency) is consistent with several studies  
459 on calibration of hydrological models (e.g. Jakeman and Hornberger (1993)  
460 and Van Werkhoven et al. (2009)). Furthermore, from the parameter list in  
461 the Supplementary Material, one can see that the 4 top-ranking parameters  
462 are: the SCS runoff curve number for moisture condition in the agricultural  
463 areas (11), the hydraulic conductivity in the river channel (9), the average  
464 slope steepness in the agricultural areas (32), and the Manning coefficient for  
465 the channel (10). This is reasonable given the predominance of agricultural  
466 land use in the catchment (62% of the catchment area as reported in Leta  
467 et al. (2015)) and the fact that the chosen output metric (Nash-Sutcliffe ef-  
468 ficiency) emphasises errors in peak flow predictions, which we expect to be  
469 mainly controlled by the parameters that characterise river routing (see for

470 example Van Werkhoven et al. (2008)).

471 Next, we analyse the impact of the sample size  $N$ . To do this, we randomly  
472 extract a subsample of smaller size from our original dataset and repeat the  
473 approximation procedure of the PAWN sensitivity indices. We test  $N=1000$ ,  
474 5000 and 7500 (Figure 6). We find that using  $N=1000$  (top panel) produces  
475 rather imprecise sensitivity indices, in fact almost all confidence intervals over-  
476 lap each other and with the dummy parameter threshold, which prevents us  
477 from inferring a robust parameter ranking. However, already at the next sam-  
478 ple size ( $N=5000$ ) the confidence intervals start to separate out and the rank-  
479 ing of the influential parameters is similar to the one obtained at the highest  
480 sample size ( $N=10000$ ).

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

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

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

532 Based on the analyses presented in this paper, we can give the following prac-  
 533 tical recommendations to future PAWN users:

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



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

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

## 562 **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.

## 563 **References**

- 564 Anderson, B., Borgonovo, E., Galeotti, M., Roson, R., 2014. Uncertainty in  
565 climate change modeling: can global sensitivity analysis be of help? *Risk*  
566 *Analysis* 34 (2), 271 – 293.
- 567 Arnold, J., Srinivasan, R., Muttiah, R., Williams, J., 1998. Large area hydro-  
568 logic modeling and assessment part 1: model development. *J. Am. Water*  
569 *Resour Assoc* 34 (1), 73–89.
- 570 Bastidas, L. A., Gupta, H. V., Sorooshian, S., Shuttleworth, W. J., Yang,  
571 Z. L., 1999. Sensitivity analysis of a land surface scheme using multicriteria  
572 methods. *Journal of Geophysical Research: Atmospheres* 104 (D16), 19481–  
573 19490.
- 574 Beven, K., Binley, A., 1992. The future of distributed models: Model calibra-  
575 tion and uncertainty prediction. *Hydrological Processes* 6 (3), 279–298.

- 576 Borgonovo, E., Hazen, G., Plischke, E., 2016. A common rationale for global  
577 sensitivity measures and their estimation. *Risk Analysis* 36, 1871–1895.
- 578 Borgonovo, E., Lu, X., Plischke, E., Rakovec, O., Hill, M. C., 2017. Making  
579 the most out of a hydrological model data set: Sensitivity analyses to open  
580 the model black-box. *Water Resources Research* 53, 7933–7950.
- 581 Castaings, W., Borgonovo, E., Morris, M., Tarantola, S., 2012. Sampling  
582 strategies in density-based sensitivity analysis. *Environmental Modelling  
583 & Software* 38 (0), 13–26.
- 584 Dell’Oca, A., Riva, M., Guadagnini, A., 2017. Moment-based metrics for global  
585 sensitivity analysis of hydrological systems. *Hydrology and Earth System  
586 Sciences* 21 (12), 6219–6234.
- 587 Efron, B., Tibshirani, R., 1993. *An Introduction to the Bootstrap*. Chapman  
588 & Hall/CRC.
- 589 Ferretti, F., Saltelli, A., Tarantola, S., 2016. Trends in sensitivity analysis  
590 practice in the last decade. *Science of The Total Environment* 568, 666–  
591 670.
- 592 Forrester, A., Sobester, A., Keane, A., 2008. *Engineering Design via Surrogate  
593 Modelling: a Practical Guide*. Wiley.
- 594 Fox, D., Koch, D., Tester, J., 2016. An analytical thermohydraulic model for  
595 discretely fractured geothermal reservoirs. *Water Resources Research* 52 (9),  
596 6792–6817.
- 597 Freer, J., Benev, K., Ambroise, B., 1996. Bayesian estimation of uncertainty  
598 in runoff prediction and the value of data: An application of the GLUE  
599 approach. *Water Resources Research* 32 (7), 2161–2173.
- 600 Gassman, P., Reyes, M. R., Green, C. H., Arnold, J. G., 2007. The Soil and  
601 Water Assessment Tool: Historical development, applications, and future  
602 research directions. *Transactions of the ASABE* 50 (4), 1211–1250.
- 603 Gillies, K., Krone, S., Nagler, J., Schultz, R., 2016. A computational model  
604 of the rainbow trout hypothalamus-pituitary-ovary-liver axis. *PLOS com-  
605 putational biology* 12 (4).
- 606 Hosseini, S., Ataie-Ashtiani, B., Simmons, C., 2017. Spring hydrograph sim-  
607 ulation of karstic aquifers: Impacts of variable recharge area, intermediate  
608 storage and memory effects. *Journal of Hydrology* 552, 225–240.
- 609 Jakeman, A., Hornberger, G., 1993. How much complexity is warranted in a  
610 rainfall-runoff model? *Water Resources Research* 29, 2637–2649.
- 611 Kolmogorov, A., 1933. Sulla determinazione empirica di una legge di dis-  
612 tribuzione. *Giornale dell’Istituto Italiano degli Attuari* 4, 83–91.
- 613 Leta, O., Nossent, J., Velez, C., Shrestha, N., van Griensven, A., Bauwens,  
614 W., 2015. Assessment of the different sources of uncertainty in a SWAT  
615 model of the river senne (belgium). *Environmental Modelling & Software*  
616 68, 129–146.
- 617 Liu, H., Sudjianto, A., Chen, W., 2006. Relative entropy based method for  
618 probabilistic sensitivity analysis in engineering design. *Journal of Mechan-  
619 ical Design* 128, 326–336.
- 620 Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models,

621 Part I - A discussion of principles. *Journal of Hydrology* 10, 282–290.

622 Norton, J., 2015. An introduction to sensitivity assessment of simulation mod-  
623 els. *Environmental Modelling & Software* 69, 166–174.

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

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

631 Petropoulos, G., Srivastava, P., 2017. *Sensitivity Analysis in Earth Observa-*  
632 *tion Modelling*. Elsevier.

633 Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B.,  
634 Wagener, T., 2016. Sensitivity analysis of environmental models: A system-  
635 atic review with practical workflow. *Environmental Modelling & Software*  
636 79, 214 – 232.

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

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

642 Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sen-  
643 sitivity analysis based on cumulative distribution functions. *Environmental*  
644 *Modelling & Software* 67, 1 – 11.

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

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

651 Plischke, E., Borgonovo, E., Smith, C. L., 2013. Global sensitivity measures  
652 from given data. *European Journal of Operational Research* 226 (3), 536 –  
653 550.

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

656 Razavi, S., Gupta, H., 2015. What do we mean by sensitivity analysis? the  
657 need for comprehensive characterization of global sensitivity in earth and  
658 environmental systems models. *Water Resources Research* 51 (5), 3070–  
659 3092.

660 Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola,  
661 S., 2010. Variance based sensitivity analysis of model output. Design and  
662 estimator for the total sensitivity index. *Computer Physics Communications*  
663 181 (2), 259–270.

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

- 666 Wiley.
- 667 Sarrazin, F., Pianosi, F., Wagener, T., 2016. Global sensitivity analysis of en-  
668 vironmental models: Convergence and validation. *Environmental Modelling*  
669 & *Software* 79, 135 – 152.
- 670 Sedighian, S., Kim, S. H., Cho, S. Y., Kim, M., Kim, D., Cha, D., 2015.  
671 Parameter ranking system of indoor radon concentration in South Korea,  
672 case studies: Dokdo island, Yang Pyeong and Nae Gi. *International Journal*  
673 *of Environmental Research* 9 (4), 1233–1236.
- 674 Smirnov, N., 1939. On the estimation of the discrepancy between empirical  
675 curves of distribution for two independent samples. *Bulletin Mathématique*  
676 *de l'Université de Moscou* 2 (2).
- 677 Spear, R., Grieb, T., Shang, N., 1994. Parameter uncertainty and interaction  
678 in complex environmental models. *Water Resources Research* 30 (11), 3159–  
679 3169.
- 680 Stanfill, B., Mielenz, H., Clifford, D., Thorburn, P., 2015. Simple approach to  
681 emulating complex computer models for global sensitivity analysis. *Envi-*  
682 *ronmental Modelling & Software* 74, 140 – 155.
- 683 Strong, M., Oakley, J., Brennan, A., 2014. Estimating multiparameter partial  
684 expected value of perfect information from a probabilistic sensitivity analy-  
685 sis sample: a nonparametric regression approach. *Medical Decision Making*  
686 34 (3), 311–326.
- 687 Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007. Comparing sensi-  
688 tivity analysis methods to advance lumped watershed model identification  
689 and evaluation. *Hydrology and Earth System Sciences* 11, 793–817.
- 690 Van Werkhoven, K., Wagener, T., Tang, Y., Reed, P., 2008. Rainfall character-  
691 istics define the value of streamflow observations for distributed watershed  
692 model identification. *Geophysical Research Letters* 35 (L11403).
- 693 Van Werkhoven, K., Wagener, T., Tang, Y., Reed, P., 2009. Complexity re-  
694 duction in multiobjective watershed model calibration. *Advances in Water*  
695 *Resources* 32 (8), 1154–1169.
- 696 Wagener, T., Kollat, J., 2007. Visual and numerical evaluation of hydrologic  
697 and environmental models using the monte carlo analysis toolbox (MCAT).  
698 *Environmental Modelling & Software* 22, 1021–1033.
- 699 Xiaomeng, S., Jianyun, Z., Chesheng, Z., Yunqing, X., Ming, Y., Chonggang,  
700 X., 2015. Global sensitivity analysis in hydrological modeling: Review of  
701 concepts, methods, theoretical framework, and applications. *Journal of Hy-*  
702 *drology* 523, 739–757.
- 703 Yang, J., 2011. Convergence and uncertainty analyses in monte-carlo based  
704 sensitivity analysis. *Environmental Modelling & Software* 26 (4), 444–457.
- 705 Zadeh, F. K., Nossent, J., Sarrazin, F., Pianosi, F., van Griensven, A., Wa-  
706 gener, T., Bauwens, W., 2017. Comparison of variance-based and moment-  
707 independent global sensitivity analysis approaches by application to the  
708 SWAT model. *Environmental Modelling & Software* 91, 210 – 222.

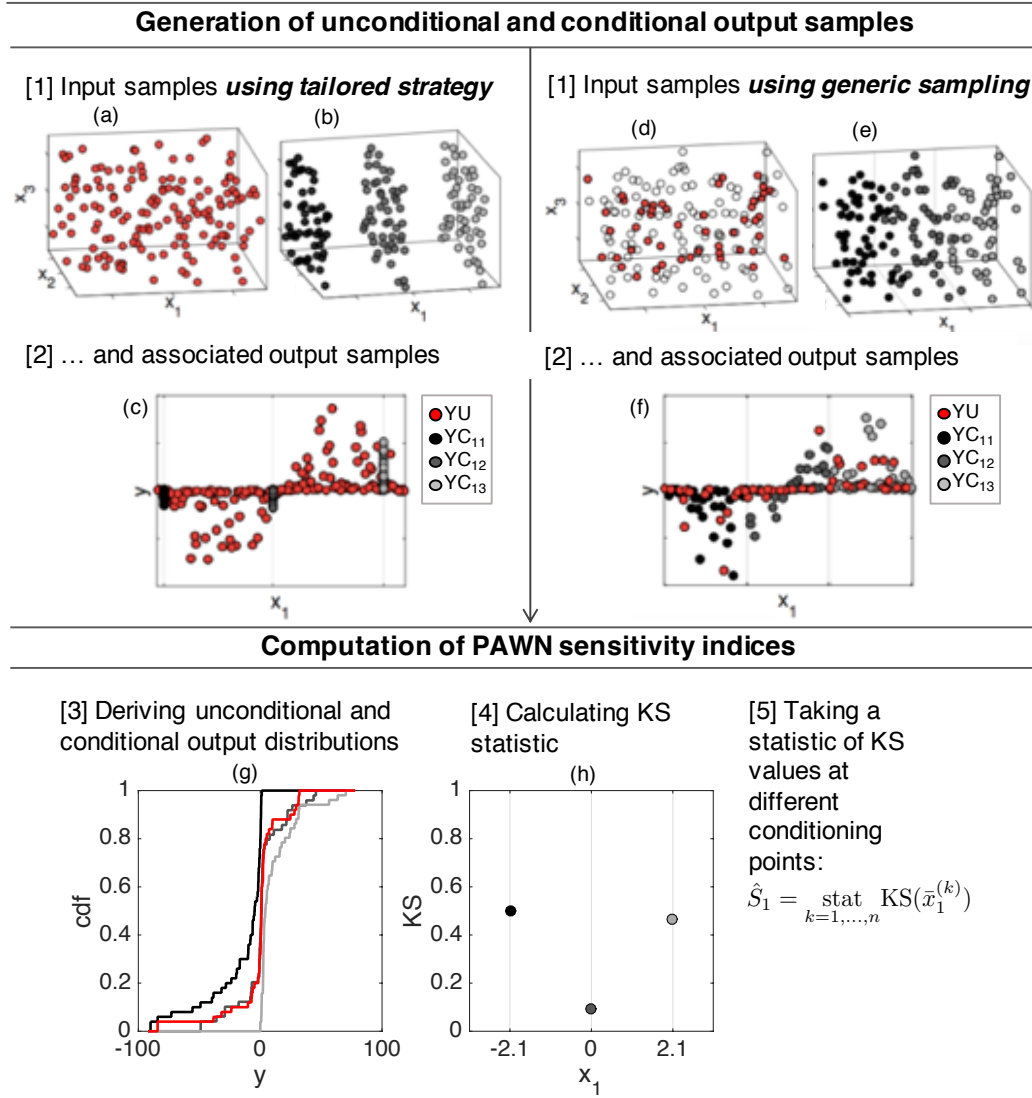


Fig. 1. Example of using a tailored sampling strategy (left) and generic sampling (right) to approximate the PAWN index of input  $x_1$  in a case of  $M=3$  input factors. Left (tailored): (a) Input samples used to derive the unconditional output sample YU. These are generated by randomly sampling the 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).

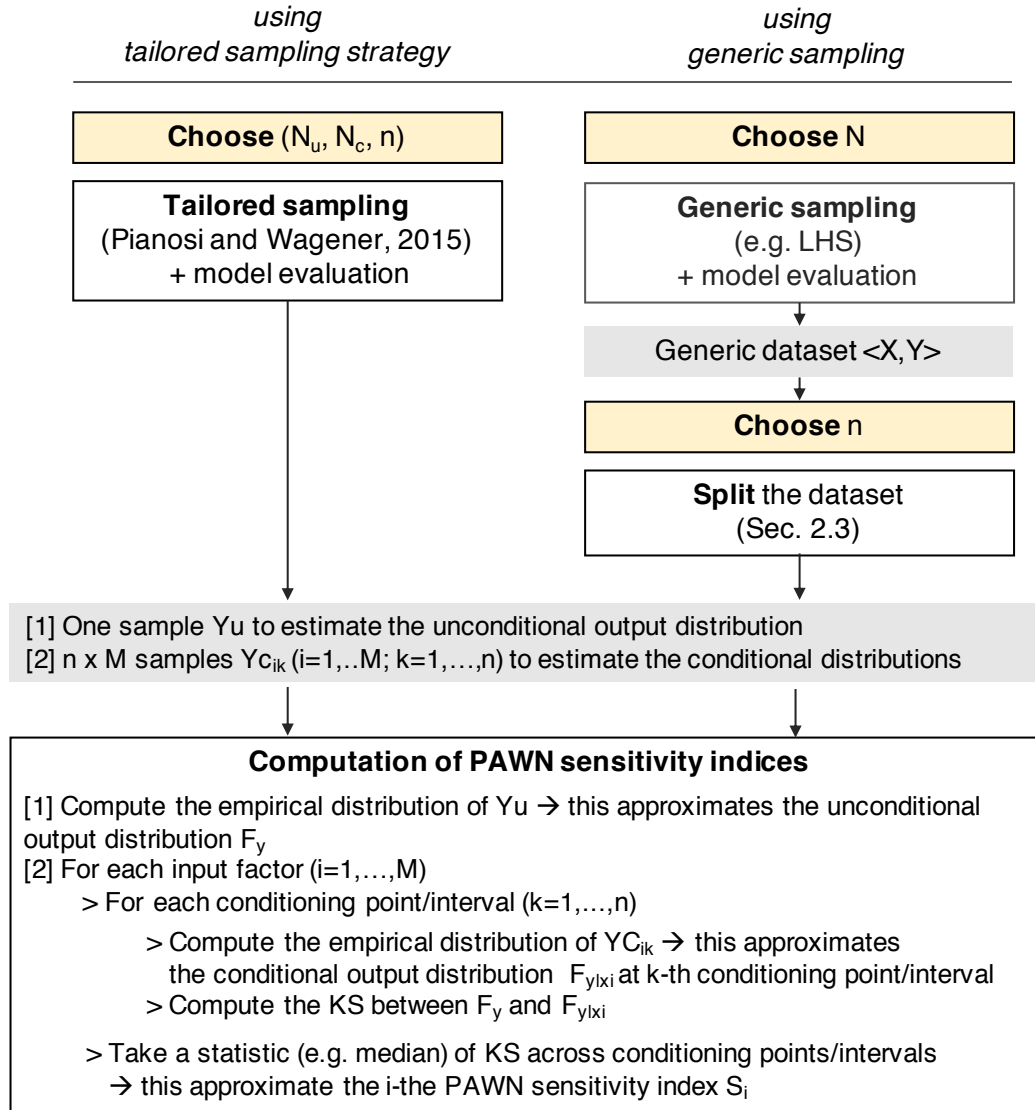


Fig. 2. Schematic of the steps needed to apply PAWN using a tailored sampling strategy (left) and generic sampling (right). In the latter case, if a generic input/output dataset is already available, the very first step of sampling and model evaluation can be skipped and the subsequent steps applied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

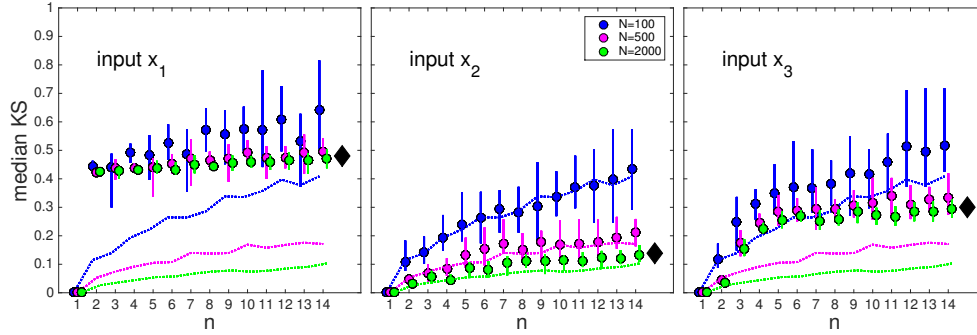


Fig. 3. PAWN indices from generic sample for the three input factors of the Ishigami-Homma function. Each subplot report results for one input factor. The PAWN index is defined as the median KS across conditioning intervals (i.e. Eq. (5) where  $\text{stat}=\text{median}$ ). PAWN indices are approximated 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).

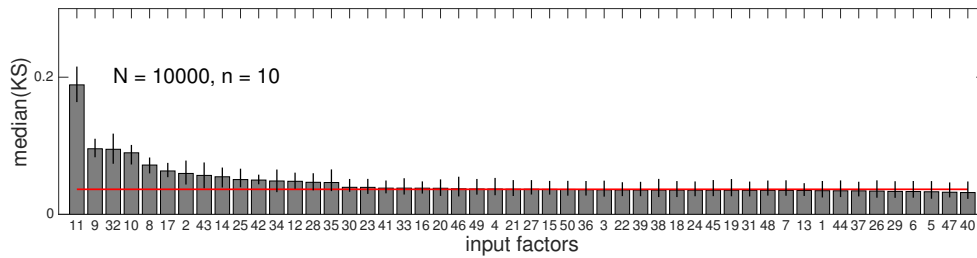


Fig. 4. PAWN sensitivity indices from generic sample for the 50 input parameters of the SWAT simulation model. The PAWN index is defined as the median KS across conditioning intervals (i.e. Eq. (5) where  $\text{stat}=\text{median}$ ). Bootstrapping is used to estimate the 95% confidence interval (vertical 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.

	PAWN from generic dataset	Method of Morris & VBSA	Regional Sensitivity Analysis
Ranking of influential inputs	11	11	11
	9	9	32
	32	10	
	10		
	8	32	9
	17	8	8
	2	2	2
	43	43	43
<i>(from Sarrazin et al (2016))</i>			

	1st
	2nd
	3rd
	4th

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.



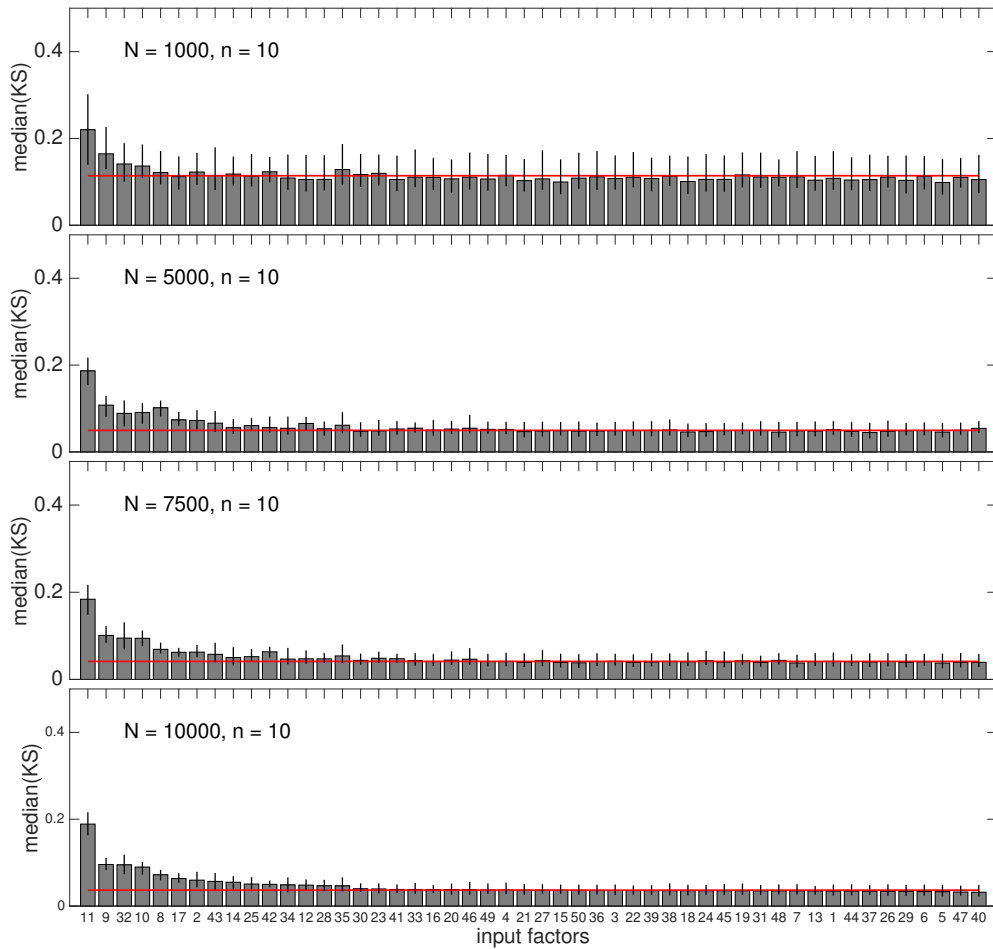


Fig. 6. Effect of the sample size  $N$  on the PAWN sensitivity indices approximated from a generic sample. Notice that the results in the bottom panel are the same as in Fig. 4 and are only reported to facilitate comparison. In all panels the input parameters are presented in the same order: this order 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.

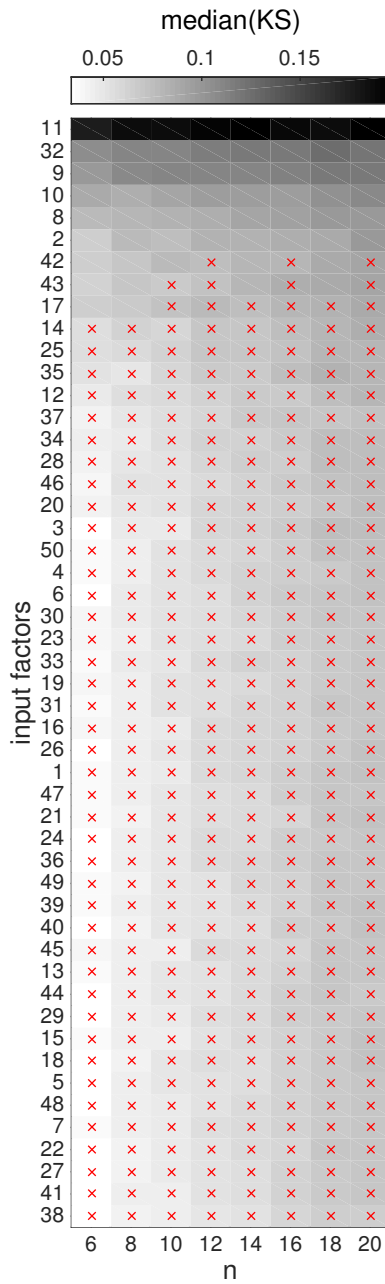


Fig. 7. Effect of the tuning parameter  $n$  (number of conditioning intervals) on the PAWN sensitivity indices approximated from generic sample (sample size  $N=5000$ ). Red crosses are used to mark sensitivity indices whose value is not higher than the KS of the dummy parameter, and hence is within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

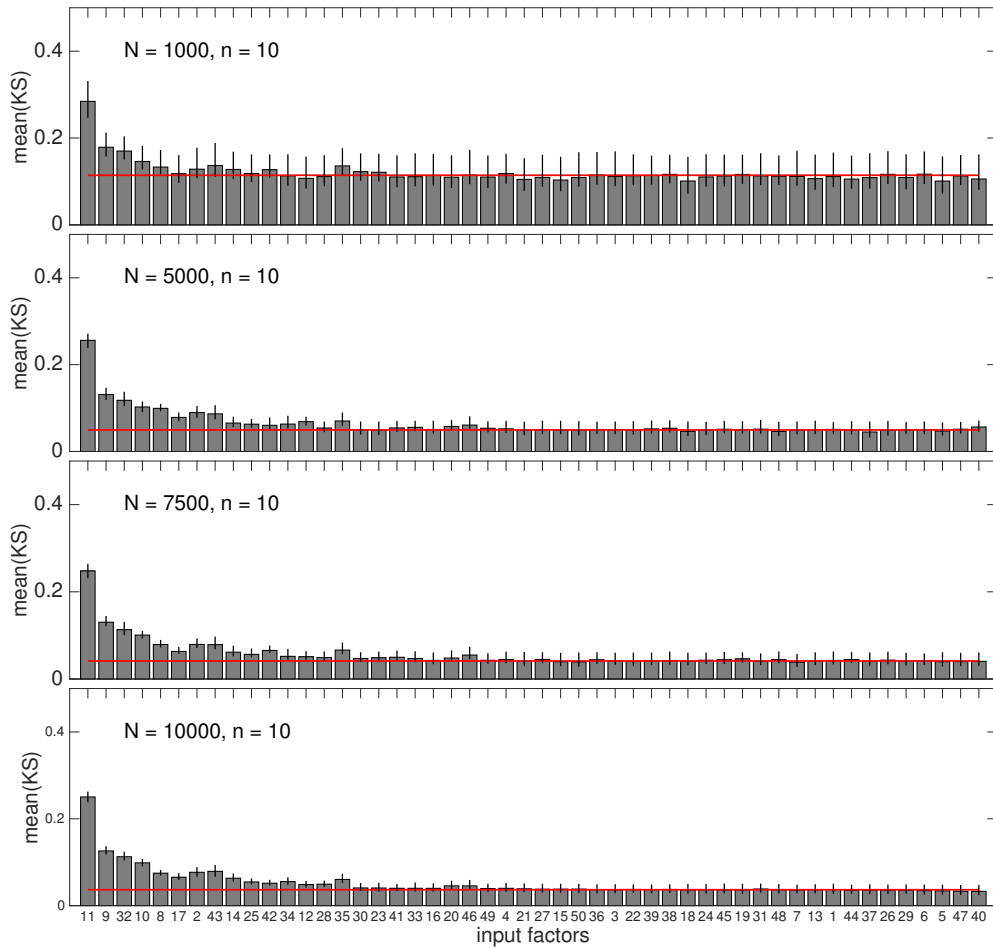


Fig. 8. Same as in Figure 6 but defining the PAWN index as the mean KS across conditioning intervals, i.e.  $\text{stat}=\text{mean}$  in Eq. (5) instead of median.

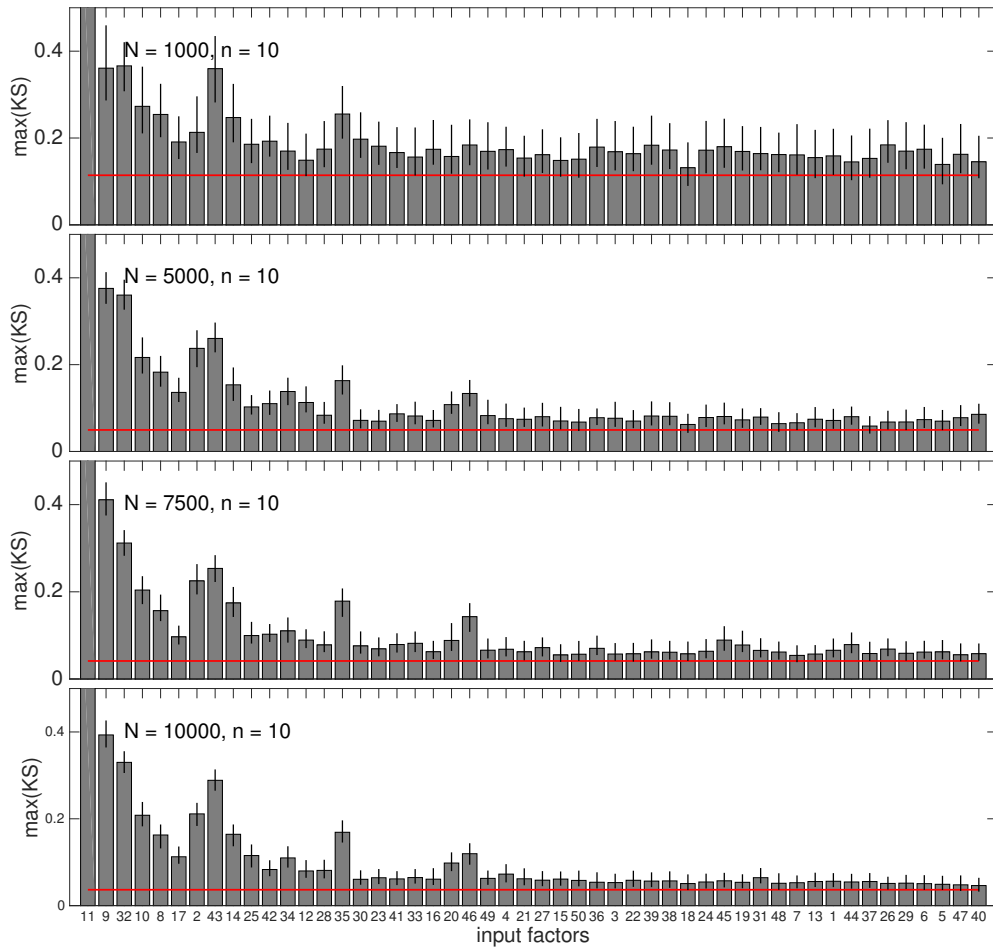


Fig. 9. Same as in Figure 6 but defining the PAWN index as the maximum KS across conditioning intervals, i.e.  $\text{stat}=\text{max}$  in Eq. (5).

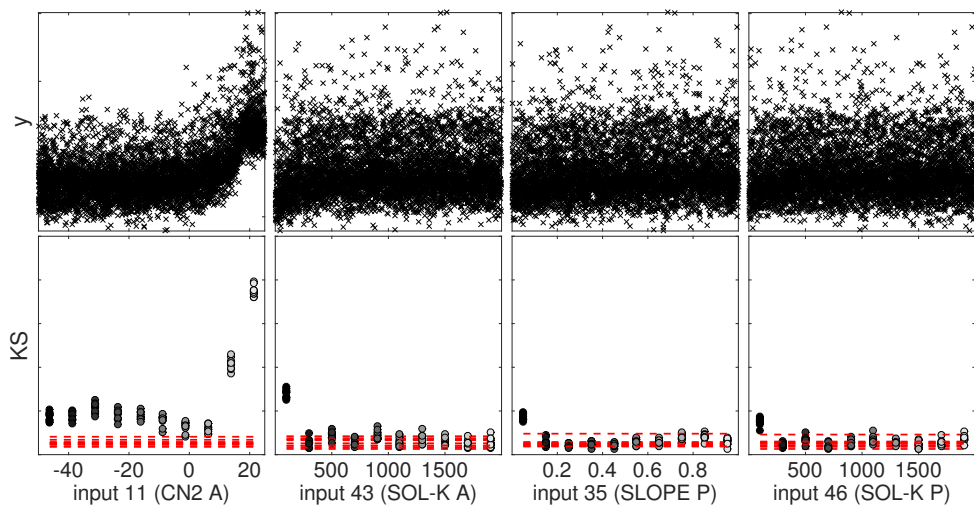


Fig. 10. Scatter plots and KS statistics of four selected input parameters of the SWAT model: number 11 is the one consistently ranked as most influential, number 43, 35 and 46 are classified as influential if using the maximum KS as PAWN sensitivity index, while they are not if using the mean or median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