Making best use of model evaluations to compute sensitivity indices

Andrea Saltelli

Making best use of model evaluations to compute sensitivity indices

Andrea Saltelli

Applied Statistics, Institute for the Protection and the Safety of the Citizen (IPSC), The European Commission, Joint Research Centre, TP 361, 21020 Ispra (VA), Italy

Abstract

This paper deals with computations of sensitivity indices in sensitivity analysis. Given a mathematical or computational model $y = f(x_1, x_2, \ldots, x_k)$, where the input factors x_i's are uncorrelated with one another, one can see y as the realization of a stochastic process obtained by sampling each of the x_i from its marginal distribution. The sensitivity indices are related to the decomposition of the variance of y into terms either due to each x_i taken singularly (first order indices), as well as into terms due to the cooperative effects of more than one x_i. In this paper we assume that one has computed the full set of first order sensitivity indices as well as the full set of total-order sensitivity indices (a fairly common strategy in sensitivity analysis), and show that in this case the same set of model evaluations can be used to compute double estimates of:

- the total effect of two factors taken together, for all such $\binom{k}{2}$ couples, where k is the dimensionality of the model;
- the total effect of $k-2$ factors taken together, for all $\binom{k}{2}$ such $(k-2)$ ples.

We further introduce a new strategy for the computation of the full sets of first plus total order sensitivity indices that is about 50% cheaper in terms of model evaluations with respect to previously published works.

We discuss separately the case where the input factors x_i's are not independent from each other.

Keywords: Sensitivity analysis; Sensitivity measures; Sensitivity indices; Importance measures

1. Introduction

Global sensitivity analysis aims to quantify the relative importance of input variables or factors in determining the value of an assigned output variable y. A recent review of applications of this discipline can be found in [14,17]. If the input to the model $y = f(x_1, x_2, \ldots, x_k)$ is composed of independent random variables, the joint probability density function of the input is:

E-mail address: andrea.saltelli@jrc.it (A. Saltelli).
URL address: http://www.jrc.cec.eu.int/uasa.

$$P(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} p_i(x_i). \tag{1}$$

Mean and variance of y can be computed as:

$$E(y) = \int\int \cdots \int f(x_1, x_2, \ldots, x_k) \prod_{i=1}^{k} p_i(x_i)\, \mathrm{d}x_i, \tag{2}$$

$$\begin{aligned} V(y) &= \int\int \cdots \int \big(f(x_1, x_2, \ldots, x_k) - E(y)\big)^2 \prod_{i=1}^{k} p_i(x_i)\, \mathrm{d}x_i \\ &= \int\int \cdots \int f^2(x_1, x_2, \ldots, x_k) \prod_{i=1}^{k} p_i(x_i)\, \mathrm{d}x_i - E^2(y). \end{aligned} \tag{3}$$

If one of the input factors x_j is fixed to a generic value $\tilde{x}_j$, the resulting variance of y will be equal to:

$$\begin{aligned} V\big(y \mid x_j = \tilde{x}_j\big) &= \int\int \cdots \int \big(f(x_1, x_2, \ldots, \tilde{x}_j \ldots, x_k) - E(y \mid x_j = \tilde{x}_j)\big)^2 \prod_{\substack{i=1 \\ i \neq j}}^{k} p_i(x_i)\, \mathrm{d}x_i \\ &= \int\int \cdots \int \big(f^2(x_1, x_2, \ldots, \tilde{x}_j \ldots, x_k)\big) \prod_{\substack{i=1 \\ i \neq j}}^{k} p_i(x_i)\, \mathrm{d}x_i - E^2\big(y \mid x_j = \tilde{x}_j\big). \end{aligned} \tag{4}$$

For the purpose of sensitivity analysis one is interested in eliminating the dependence upon the value $\tilde{x}_j$ by integrating $V(y \mid x_j = \tilde{x}_j)$ over the probability density function of $\tilde{x}_j$, obtaining

$$\begin{aligned} E\big(V(y \mid x_j)\big) &= \int\int \cdots \int f^2(x_i, x_2, \ldots, x_j \ldots, x_k) \prod_{i=1}^{k} p_i(x_i)\, \mathrm{d}x_i \\ &\quad - \int E^2\big(y \mid x_j = \tilde{x}_j\big) p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j. \end{aligned} \tag{5}$$

Note that we have dropped the dependence $\tilde{x}_j$ from the left-hand side, as it disappears due to the integration. Subtracting Eq. (5) from Eq. (3) one obtains:

$$V(y) - E\big(V(y \mid x_j)\big) = \int E^2\big(y \mid x_j = \tilde{x}_j\big) p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j - E^2(y). \tag{6}$$

The left-hand side of Eq. (6) is also equal to $V(E(y \mid x_j))$, and is a good measure of the sensitivity of y with respect to factor x_j. If one divides it by the unconditional variance, one obtains the so-called first order sensitivity index $S_j = V(E(y \mid x_j))/V(y)$. The S_i's are nicely scaled in $[0, 1]$. Eq. (6) is computationally impractical. In a Monte Carlo frame, it implies a double loop: the inner one to compute $E^2(y \mid x_j = \tilde{x}_j)$, and the outer to compute the integral over $\mathrm{d}\tilde{x}_j$. For this reason the integral in (6) has been rewritten by Ishigami and Homma [7] as:

$$\begin{aligned} &\int E^2\big(y \mid x_j = \tilde{x}_j\big) p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_i \\ &\quad = \int \left\{ \int\int \cdots \int f\big(x_1, x_2, \ldots, \tilde{x}_j, \ldots, x_k\big) \prod_{\substack{i=1 \\ i \neq j}}^{k} p_i(x_i)\, \mathrm{d}x_i \right\}^2 p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j \end{aligned}$$

$$= \iint \cdots \int f\left(x_1, x_2, \ldots, \tilde{x}_j ..., x_k\right) f\left(x'_1, x'_2, \ldots, \tilde{x}_j ..., x'_k\right) \prod_{\substack{i=1 \\ i \neq j}}^{k} \left(p_i(x_i)\, \mathrm{d}x_i\right) \prod_{\substack{i=1 \\ i \neq j}}^{k} \left(p_i(x'_i)\, \mathrm{d}x'_i\right) p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j$$

$$= \iint \cdots \int f(x_1, x_2, \ldots, x_j ..., x_k) f\left(x'_1, x'_2, \ldots, x_j ..., x'_k\right) \prod_{i=1}^{k} \left(p_i(x_i)\, \mathrm{d}x_i\right) \prod_{\substack{i=1 \\ i \neq j}}^{k} \left(p_i(x'_i)\, \mathrm{d}x'_i\right). \tag{7}$$

The expedient of using the additional integration variable primed, allows us to realize that the integral in (7) is the expectation value of the function F of a set of $(2k-1)$ factors:

$$\begin{aligned} &F\left(x_1, x_2, \ldots, x_j, \ldots x_k, x'_1, x'_2, \ldots, x'_{j-1}, x'_{j+1}, \ldots x'_k\right) \\ &\quad = f(x_1, x_2, \ldots x_k) f\left(x'_1, x'_2, \ldots, x'_{j-1}, x_j, x'_{j+1}, \ldots x'_k\right). \end{aligned} \tag{8}$$

The integral (7) can be hence computed using a single Monte Carlo loop. The Monte Carlo procedure that follows was proposed by Saltelli et al. [13].

Two input sample matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are generated:

$$\mathbf{M}_1 = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & & & \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}, \qquad \mathbf{M}_2 = \begin{pmatrix} x'_{11} & x'_{12} & \ldots & x'_{1k} \\ x'_{21} & x'_{22} & \ldots & x'_{2k} \\ \ldots & & & \\ x'_{n1} & x'_{n2} & \ldots & x'_{nk} \end{pmatrix}, \tag{9}$$

where n is the sample size used for the Monte Carlo estimate. In order to estimate the sensitivity measure for a generic factor x_j, i.e.

$$\begin{aligned} S_j &= \frac{V(E(y \mid x_j))}{V(y)} = \frac{(U_j - E^2(y))}{V(y)}, \\ U_j &= \int E^2\left(y \mid x_j = \tilde{x}_j\right) p_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j \end{aligned} \tag{10}$$

we need an estimate for both $E(y)$ and U_j. The former can be either obtained from values of y computed on the sample in $\mathbf{M}_1$ or $\mathbf{M}_2$. U_j can be obtained from values of y computed on matrices $\mathbf{M}_1$ and $\mathbf{N}_j$, the latter being defined as:

$$\mathbf{N}_j = \begin{pmatrix} x'_{11} & x'_{12} & \ldots & x_{1j} & \ldots & x'_{1k} \\ x'_{21} & x'_{22} & \ldots & x_{2j} & \ldots & x'_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x'_{n1} & x'_{n2} & \ldots & x_{nj} & \ldots & x'_{nk} \end{pmatrix}, \tag{11}$$

i.e. by:

$$\widehat{U}_j = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f\left(x'_{r1}, x'_{r2}, \ldots, x'_{r(j-1)}, x_{rj}, x'_{r(j+1)}, \ldots, x'_{rk}\right). \tag{12}$$

If one thinks of matrix $\mathbf{M}_1$ as the "sample" matrix, and of $\mathbf{M}_2$ as the "re-sample" matrix, then $\widehat{U}_j$ is obtained from products of values of f computed from the sample matrix times values of f computed from $\mathbf{N}_j$, i.e. a matrix where all factors except x_j are re-sampled. In this way the computational cost associated with a full set of first order indices S_i is $n(k+1)$. One set of n evaluations of f is needed to compute $E(y)$, and k sets of n evaluations of f are needed for the second term in the product (12).

The very same procedure for the computation of sensitivity indices was proposed by Sobol'[19]. The problem setting of Sobol' was that of identifying a subset of the k factors that could account for most of the variance of y. Imagine that the factors have been partitioned into a trial set $\mathbf{u} = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ and the remaining set

$\mathbf{v} = (x_{l_1}, x_{l_2}, \ldots, x_{l_{k-m}})$. Then according to Sobol' an idea of the overall effect of the subset $\mathbf{u}$ on the variance of the output can be estimated from:

$$U_v = \int\int \cdots \int f(\mathbf{u}, \mathbf{v}) f(\mathbf{u}', \mathbf{v}) \, \mathrm{d}\mathbf{u} \, \mathrm{d}\mathbf{u}' \, \mathrm{d}\mathbf{v}, \tag{13}$$

$$V\big(E(y \mid \mathbf{v})\big) = U_{\mathbf{v}} - E^2(y), \tag{14}$$

$$V\big(E(y \mid \mathbf{u})\big) + V\big(E(y \mid \mathbf{u}, \mathbf{v})\big) = V(y) - V\big(E(y \mid \mathbf{v})\big). \tag{15}$$

In Eq. (15), $V(E(y \mid \mathbf{u}))$ is the first order effect of the set $\mathbf{u}$, while $V(E(y \mid \mathbf{u}, \mathbf{v}))$ is the interaction term between the sets $\mathbf{u}$ and $\mathbf{v}$. If $V(y) \cong V(E(y \mid \mathbf{v}))$, then $\mathbf{u}$ is non-influent, and all factors in $\mathbf{u}$ can be fixed in a subsequent analysis of the model. Formula (13) shows the same expedient of the additional integration variables already described. The Monte Carlo estimate of $U_{\mathbf{v}}$ is:

$$\widehat{U}_{\mathbf{v}} = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{ri_1}, x_{ri_2}, \ldots, x_{ri_m}, x_{rl_1}, x_{rl_2}, \ldots, x_{rl_{k-m}}) f\big(x'_{ri_1}, x'_{ri_2}, \ldots, x'_{ri_m}, x_{rl_1}, x_{rl_2}, \ldots, x_{rl_{k-m}}\big), \tag{16}$$

i.e. to estimate the total effect of set $\mathbf{u}$ one must now re-sample the variables in the set $\mathbf{u}$. One can easily see that (12) is a particular case of (16). Error estimates for $\widehat{U}_j$'s are discussed in the original reference of Sobol'. A bootstrap based alternative is discussed in [1].

Eq. (15) is a particular case of a general variance decomposition scheme proposed by Sobol', whereby the total unconditional variance can be decomposed as:

$$V(y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{12\ldots k}, \tag{17}$$

where

$$V_i = V\big(E(Y \mid x_i)\big),$$
$$V_{ij} = V\big(E(Y \mid x_i, x_j)\big) - V_i - V_j$$

and so on. The development in (17) contains k terms of the first order V_i, $k(k-1)/2$ terms of the second order V_{ij} and so on, till the last term of order k, for a total of $2^k - 1$ terms. The V_{ij} terms are the second order (or two-way) terms, analogous to the second order effects described in experimental design textbooks (see, e.g., [2]). The V_{ij} terms capture that part of the effect of x_i and x_j that is not described by the first order terms. Formula (17) has a long history, and various authors have proposed different versions of it. A discussion can be found in [1], as well as in [10]. Sobol's version of formula (17) is based on a decomposition of the function f itself into terms of increasing dimensionality, i.e.:

$$f(x_1, x_2, \ldots, x_k) = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + \cdots + f_{12\ldots k}, \tag{18}$$

where each term is function only of the factors in its index, i.e. $f_i = f_i(x_i)$, $f_{ij} = f_{ij}(x_i, x_j)$ and so on. Decompositions (17), (18) are unique provided that the input factors are independent and that the individual terms $f_{i_1 i_2 \ldots i_s}$ in (18) are square integrable and have zero mean over the domain of existence.

One important aspect of Sobol' development is that similar decompositions can be written by taking the factors into subsets, as shown by Eq. (15). This prompted Homma and Saltelli [5] to introduce the new estimate U_{-j}:

$$\widehat{U}_{-j} = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rj}, \ldots, x_{rk}) f\big(x_{r1}, x_{r2}, \ldots, x_{r(j-1)}, x'_{rj}, x_{r(j+1)}, \ldots, x_{rk}\big).$$

As before:

$$V\big(E(y \mid \mathbf{x}_{-j})\big) = \widehat{U}_{-j} - \widehat{E}^2(y), \tag{19}$$

where $V(E(y \mid \mathbf{x}_{-j}))$ is the total contribution to the variance of y due to non-x_j. This implies that the difference $V(y) - V(E(y \mid \mathbf{x}_{-j}))$ is equal to the sum of all terms in the variance decomposition (15) that include x_j. We illustrate this for the case $k = 3$:

$$S_1^T = \frac{V(y) - V(E(y \mid \mathbf{x}_{-1}))}{V(y)} = \frac{E(V(y \mid \mathbf{x}_{-1}))}{V(y)} = S_1 + S_{12} + S_{13} + S_{123}, \tag{20}$$

where, e.g., $S_1 = V(E(y \mid x_1))/V(y)$. Analogous expressions can be written for S_2^T, S_3^T. We have called the S_j^T's "total effect" terms. The total effects are useful for the purpose of sensitivity analysis, as discussed in [18], as they give information on the non additive part of the model. It may be useful to observe here that for a purely additive model, $\sum_{i=1}^{k} S_i = 1$, while for a given factor x_j an important difference between S_j^T and S_j flags an important role of interactions for that factor in y. The same information could be obtained by computing all terms in (17), but these are as many as $2^k - 1$. This problem has been referred to by Rabitz et al. [11] as "the curse of dimensionality". The computational cost of estimating all effects in (17) is in fact as high as $n2^k$, where again n is the sample size used to estimate one individual effect.[1] For these reasons we customarily tend to compute the set of all S_i plus the set of all S_i^T, that gives a fairly good description of the model sensitivities.

This would normally entail a computational cost of $n(2k + 1)$ model evaluations, i.e. nk for the first order terms, again nk for the total effect terms, plus n for $\widehat{E}(y)$. In fact we have found in Homma and Saltelli [5], that better estimates for the first order terms are obtained if the $E^2(y)$ term in (10) is estimated as

$$\widehat{E}^2 = \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f\big(x'_{r1}, x'_{r2}, \ldots, x'_{rk}\big) \tag{21}$$

rather than from

$$\widehat{E}^2 = \left\{\frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk})\right\}^2, \tag{22}$$

i.e. using sample estimates from both $\mathbf{M}_1$ and $\mathbf{M}_2$ matrices rather than from $\mathbf{M}_1$ alone. Eq. (21) is a legitimate estimate of the squared sample mean given the independence of the two sample vectors. It is clear from (10) and (12) that the estimate of S_j goes more naturally to zero for a non-influential factor x_j when (21) is used, as can be seen from:

$$\begin{aligned}\widehat{U}_j - \widehat{E}^2(y) = {} & \frac{1}{n-1}\sum_{r=1}^{n}\big(f(x_{r1}, x_{r2}, \ldots, x_{rk}) f\big(x'_{r1}, x'_{r2}, \ldots, x'_{r(j-1)}, x_{rj}, x'_{r(j+1)}, \ldots, x'_{rk}\big) \\ & - \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f\big(x'_{r1}, x'_{r2}, \ldots, x'_{rk}\big).\end{aligned} \tag{23}$$

On the same ground one can see that the computation of the total effect sensitivity indices is better achieved using (22). $V(y)$ is computed from $\mathbf{M}_1$ for all indices. In conclusion, the standard computational strategy so far employed to compute the full set of total and first order indices entailed a total of $n(2k + 2)$ model evaluations, two samples being used to estimate $\widehat{E}^2$.

Many applications of this strategy to different models can be found in various chapters of Saltelli et al. [14].

An important economy in model evaluation, that is described in [18], is that the S_j^T and S_j terms can also be estimated using an extended version of the Fourier Amplitude Sensitivity Test (FAST). When using extended FAST, the same set of n model evaluations that was used to estimate a given S_j^T can also be used to compute S_j,

[1] $n(2^k - 1)$ would be needed to compute all effects, and n more to compute $\widehat{E}(y)$, $V(y)$.

that makes the entire analysis feasible at the cost of nk model evaluations. For this reason the extended FAST method has been considered so far as the most efficient way to compute the full set of S_j^T and S_j indices. In the present work we introduce an extended version of Sobol' method that out-performs FAST.

2. Extension of the method

We imagine that a model has been characterised using the Sobol' method, computing all S_j^T's and S_j's estimates at the cost of $n(2k+2)$ model evaluations. Can the same coefficient be estimated with a smaller cost? Can additional estimates be obtained with the same sets used to compute the $\widehat{S}_j^T$'s and $\widehat{S}_j$'s? Before we proceed we need to introduce some new notation.

Let us call $V^c_{i_1 i_2 \dots i_s}$ a sensitivity measure that is closed within a subset of factors, i.e. $V^c_{i_1 i_2 \dots i_s}$ is the sum of all $V_{i_1 i_2 \dots i_s}$ terms in (17) that is closed in the indices $i_1, i_2, \dots, i_s : V^c_1 = V_1$, $V^c_{ij} = V_i + V_j + V_{ij}$, and so on. Likewise $V^c_{-i_1 i_2 \dots i_s}$ will indicate the sum of all indices that are closed within the complementary set of $i_1, i_2, \dots, i_s$, i.e. $V^c_{-i_1 i_2 \dots i_s} = V^c_{l_1 l_2 \dots l_{k-s}}$, where $i_p \neq l_q$ for all $p \in [1, 2, \dots, s], q \in [1, 2, \dots, k-s]$.

Let $\mathbf{a}_{i_1 i_2 \dots i_s}$ denote the vector of length n containing model evaluations corresponding to the rows of the input factor matrix $\mathbf{N}_{i_1 i_2 \dots i_s}$. As in Eq. (11) above, the matrix $\mathbf{N}_{i_1 i_2 \dots i_s}$ is obtained from matrix $\mathbf{M}_1$ by substituting all columns except $i_1, i_2, \dots, i_s$ by the corresponding columns of matrix $\mathbf{M}_2$. $\mathbf{a}_0$ will hence denote a set of model evaluations corresponding entirely to matrix $\mathbf{M}_2$, while $\mathbf{a}_{i_1 i_2 \dots i_k}$ will indicate the vector of model evaluations corresponding entirely to matrix $\mathbf{M}_1$.

A few equalities are repeated below for reader's convenience:

$$V^c_{i_1 i_2 \dots i_s} = V\left(E(Y \mid x_{i_1} x_{i_2} \dots x_{i_s})\right) = U_{i_1 i_2 \dots i_s} - E^2(y), \tag{24}$$

$$\widehat{U}_{i_1 i_2 \dots i_s} = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \dots, x_{rk}) f\left(x_{ri_1}, x_{ri_2}, \dots, x_{ri_s}, x'_{rl_1}, x'_{rl_2}, \dots, x'_{rl_{k-s}}\right), \tag{25}$$

$$\widehat{U}_{-i_1 i_2 \dots i_s} = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \dots, x_{rk}) f\left(x'_{ri_1}, x'_{ri_2}, \dots, x'_{ri_s}, x_{rl_1}, x_{rl_2}, \dots, x_{rl_{k-s}}\right) \tag{26}$$

with the special cases

$$\widehat{S}_j = \frac{(\widehat{U}_j - \widehat{E}^2(y))}{\widehat{V}(y)}, \tag{27}$$

$$\widehat{S}_j^T = 1 - \frac{(\widehat{U}_{-j} - \widehat{E}^2(y))}{\widehat{V}(y)}. \tag{28}$$

We are now ready to submit the following theorem:

Theorem 1. *Given a model $y = f(x_1, x_2, \dots, x_k)$, it is possible to compute at the cost of $n(k+2)$ model evaluations*:

(1) *One estimate for each of the k indices of the first order $\widehat{S}_j$.*
(2) *One estimate for each of the k total effect indices $\widehat{S}_j^T$.*
(3) *One estimate for each of the $\binom{k}{2} V^c_{-ij}$ closed effect indices.*

Table 1
Terms that can be estimated given the corresponding vectors of model evaluations, $k = 5$

	$\mathbf{a}_0$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$	$\mathbf{a}_{2345}$	$\mathbf{a}_{1345}$	$\mathbf{a}_{1245}$	$\mathbf{a}_{1235}$	$\mathbf{a}_{1234}$	$\mathbf{a}_{12345}$
$\mathbf{a}_0$	$\widehat{V}(y)$											
$\mathbf{a}_1$	$\widehat{S}_1^T$	$\widehat{V}(y)$										
$\mathbf{a}_2$	$\widehat{S}_2^T$	$\widehat{V}_{-12}^c$	$\widehat{V}(y)$									
$\mathbf{a}_3$	$\widehat{S}_3^T$	$\widehat{V}_{-13}^c$	$\widehat{V}_{-23}^c$	$\widehat{V}(y)$								
$\mathbf{a}_4$	$\widehat{S}_4^T$	$\widehat{V}_{-14}^c$	$\widehat{V}_{-24}^c$	$\widehat{V}_{-34}^c$	$\widehat{V}(y)$							
$\mathbf{a}_5$	$\widehat{S}_5^T$	$\widehat{V}_{-15}^c$	$\widehat{V}_{-25}^c$	$\widehat{V}_{-35}^c$	$\widehat{V}_{-45}^c$	$\widehat{V}(y)$						
$\mathbf{a}_{2345}$	$\widehat{S}_1$	$\widehat{E}^2(y)$	$\widehat{V}_{12}^c$	$\widehat{V}_{13}^c$	$\widehat{V}_{14}^c$	$\widehat{V}_{15}^c$	$\widehat{V}(y)$					
$\mathbf{a}_{1345}$	$\widehat{S}_2$	$\widehat{V}_{12}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{23}^c$	$\widehat{V}_{24}^c$	$\widehat{V}_{25}^c$	$\widehat{V}_{-12}^c$	$\widehat{V}(y)$				
$\mathbf{a}_{1245}$	$\widehat{S}_3$	$\widehat{V}_{13}^c$	$\widehat{V}_{23}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{34}^c$	$\widehat{V}_{35}^c$	$\widehat{V}_{-13}^c$	$\widehat{V}_{-23}^c$	$\widehat{V}(y)$			
$\mathbf{a}_{1235}$	$\widehat{S}_4$	$\widehat{V}_{14}^c$	$\widehat{V}_{24}^c$	$\widehat{V}_{34}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{45}^c$	$\widehat{V}_{-14}^c$	$\widehat{V}_{-24}^c$	$\widehat{V}_{-34}^c$	$\widehat{V}(y)$		
$\mathbf{a}_{1234}$	$\widehat{S}_5$	$\widehat{V}_{15}^c$	$\widehat{V}_{25}^c$	$\widehat{V}_{35}^c$	$\widehat{V}_{45}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{-15}^c$	$\widehat{V}_{-25}^c$	$\widehat{V}_{-35}^c$	$\widehat{V}_{-45}^c$	$\widehat{V}(y)$	
$\mathbf{a}_{12345}$	$\widehat{E}^2(y)$	$\widehat{S}_1$	$\widehat{S}_2$	$\widehat{S}_3$	$\widehat{S}_4$	$\widehat{S}_5$	$\widehat{S}_1^T$	$\widehat{S}_2^T$	$\widehat{S}_3^T$	$\widehat{S}_4^T$	$\widehat{S}_5^T$	$\widehat{V}(y)$

An additional theorem is the following:

Theorem 2. *If we modify the setting of Theorem* 1 *by allowing for* $n(2k+2)$ *model evaluations* (*i.e. as many as in the procedure of Section* 1)*, we can obtain*:

(1) *Double rather than single estimates for each of the* $\widehat{S}_j^T$*'s and* $\widehat{S}_j$*'s indices.*
(2) *Double estimates for each of the* $\binom{k}{2} V_{ij}^c$ *terms.*
(3) *Double rather than single estimates for each of the* $\binom{k}{2} V_{-ij}^c$ *terms.*

Theorems 1, 2 constitute the promised extension of Sobol' method. Let us illustrate the new procedures for the case $k = 5$. We have to use this value as lower values of k are special cases and will be treated afterwards. Table 1 gives for each cell what term can be computed by the corresponding $\mathbf{a}_{i_1 i_2 \ldots i_s}$ vectors.

Note that:

(1) Table 1 can be interpreted by referring to Eqs. (24)–(28) above. E.g., we have labelled the entry corresponding to the intersection $\mathbf{a}_0$ and $\mathbf{a}_1$ as $\widehat{S}_1^T$, as $\mathbf{a}_0 \cdot \mathbf{a}_1$ yields $\widehat{U}_{-1}$ that in turn can be used to compute $\widehat{S}_1^T$ (Eq. (28)) and so on for the other terms.
(2) The diagonal has been labelled as providing an estimate of $\widehat{V}(y)$, as this is what can be obtained by the scalar product $\mathbf{a}_{i_1 i_2 \ldots i_s}^2$. In fact each of the $2k+2$ vectors $\mathbf{a}_{i_1 i_2 \ldots i_s}$ can yield an estimate of $\widehat{E}(y)$. Known $\widehat{E}(y)$ each $\mathbf{a}_{i_1 i_2 \ldots i_s}$ can again be used to estimate $\widehat{V}(y)$.
(3) The row labelled $\mathbf{a}_{12345}$ illustrates the same procedure as in Section 1 for the computation of the first order indices and total order indices, i.e. $\widehat{S}_4^T$ is obtained from $\widehat{V}(y)$ and $\widehat{V}_4^T$, this latter being computed from $\mathbf{a}_{12345}, \mathbf{a}_{1235}$.
(4) Looking at the column $\mathbf{a}_0$, one sees that the same set of indices (first order plus total) can be computed from $\mathbf{a}_0$, $\mathbf{a}_{12345}$, and either of the sets $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5\}$ or $\{\mathbf{a}_{-1}, \mathbf{a}_{-2}, \mathbf{a}_{-3}, \mathbf{a}_{-4}, \mathbf{a}_{-5}\} \equiv \{\mathbf{a}_{2345}, \mathbf{a}_{1345}, \mathbf{a}_{1245}, \mathbf{a}_{1235}, \mathbf{a}_{1234}\}$, bringing the computational cost from $n(2k+2)$ down to $n(k+2)$, with a reduction in computational cost that tends to 50% at increasing k values.

(5) All indices in rows other than $\mathbf{a}_{12345}$ and columns other than $\mathbf{a}_0$ are novel, in the sense that they were overlooked in the procedure of Section 1. The alternative arranging of the $\mathbf{a}_{i_1 i_2 \ldots i_s}$ terms shows the additional terms that can be computed.
(6) The intersection of vectors $\mathbf{a}_1$ and $\mathbf{a}_{2345}$ has been labelled as an estimate of $\widehat{E}^2(y)$, as all columns in the two sampling matrices are different and the scalar product $\mathbf{a}_{i_1 i_2 \ldots i_s}$, $\mathbf{a}_{j_1 j_2 \ldots j_r}$ provides an estimate of the square of $E(y)$, as in Eq. (21) above.
(7) The two vectors $\mathbf{a}_2$ and $\mathbf{a}_{2345}$ allow the computation of $\widehat{V}^c_{12}$ as columns 1 and 2 are identical in the two sampling matrices.
(8) The two vectors $\mathbf{a}_{2345}$ and $\mathbf{a}_{1345}$ allow the computation of $\widehat{V}^c_{345} = \widehat{V}^c_{-12}$ as columns 3,4,5 are identical in the two sampling matrices.

Looking at this table, for the setting of Theorem 1 (grey cells in Table 1), it is easy to see that we have produced the $\widehat{S}^T_j$'s and $\widehat{S}_j$ indices, with $j \in [1, 2, 3, 4, 5]$ at the reduced cost of $n(k+2) = 7n$ model evaluations, instead of $n(2k+2) = 12n$, with a reduction of 42% in computational cost. Furthermore, we have produced one estimate for each of the $\binom{5}{3} = 10$ indices complementary to the second order ones, that for $k = 5$ happen to be closed indices of the third order. Note that for, e.g., $k = 6$ we would have obtained one estimate for each of the $\binom{6}{4} = 15$ closed indices of the fourth order and so on for larger values of k, and so on based on the known property that $\binom{k}{j} = \binom{k}{k-j}$ for $k \geqslant j$.

For the setting of Theorem 2, it is easy to see that double, rather than single, estimates for each of the $\widehat{S}^T_j$'s, $\widehat{S}_j$ and $\widehat{V}^c_{-ij}$ have been produced. We have additionally obtained double estimates for each of the $\binom{5}{2} = 10$ closed indices of the second order. Additional estimates of $\widehat{E}^2(y)$ are also available as discussed.

The reader might wonder which among the various estimates of $\widehat{E}^2(y)$, $\widehat{V}(y)$ should be used in Eqs. (24)–(28) to obtain, e.g., the $V^c_{i_1 i_2 \ldots i_s}$ from the $\widehat{U}_{i_1 i_2 \ldots i_s}$. In [5] we suggest that the estimate of $\widehat{E}^2(y)$ in (21) obtained from $\mathbf{a}_0$, $\mathbf{a}_{12345}$ should be used for the first order indices and that from (22) based on $\mathbf{a}_0$ alone for the total effect ones.

In the context of the extended procedures (Theorems 1 and 2) presented here, the following approach was taken:

(1) Theorem 1 setting. Eq. (21) is used for the first order indices. This means that for computing, any of the $\widehat{S}_j$, $\mathbf{a}_0$ and $\mathbf{a}_{12345}$ vectors are used to estimate $\widehat{E}^2(y)$ and $\mathbf{a}_{12345}$ to compute $\widehat{V}(y)$. Eq. (22) is used for the total effect, i.e. for any of the $\widehat{S}^T_j$, $\mathbf{a}_0$ alone is used to estimate $\widehat{E}^2(y)$. $\widehat{V}(y)$ is also computed from $\mathbf{a}_0$ for the total effect indices. For the closed effects of order $k-2$ Eq. (22) is used for $\widehat{E}^2(y)$, and the vector to be used in (22) is selected as one of the two that concur in the estimation of the effect. E.g., for $\widehat{V}^c_{-15}$ (in the grey table area), computed from $\mathbf{a}_1$ and $\mathbf{a}_5$, the $\widehat{E}^2(y)$ is computed from $\mathbf{a}_1$ alone (or identically from $\mathbf{a}_5$ alone). $\widehat{V}(y)$ is computed from the same vector used for $\widehat{E}^2(y)$ (either $\mathbf{a}_1$ or $\mathbf{a}_5$).
(2) Theorem 2 setting. Same procedure as Theorem 1 for all double estimates of (i) first order, (ii) total order indices and (iii) order $k-2$ closed indices. For the closed indices of the second order Eq. (21) is used, where $\widehat{E}^2(y)$ is computed using one of the vector that concur in the estimation of the index. E.g., for that estimate of $\widehat{V}^c_{12}$, that is computed from $\mathbf{a}_{1345}$ and $\mathbf{a}_1$, $\widehat{E}^2(y)$ is computed from $\mathbf{a}_{1345}$ and $\mathbf{a}_2$ (or identically from $\mathbf{a}_1$ and $\mathbf{a}_{2345}$). The variance is correspondingly computed from $\mathbf{a}_{1345}$ or $\mathbf{a}_1$. These arrangements can be easily understood by inspecting equations like (23) above.

As we said, $k = 4$ is a special case (Table 2). For this value of k we obtain for the setting of Theorem 2:

(1) Double estimates for each of the $4\widehat{S}_i$ and each of the $4\widehat{S}^T_i$.
(2) Quadruple estimates of the $\binom{4}{2} = 6$ second order terms $\widehat{V}^c_{ij}$.

All 4 estimates of each term V^c_{ij} are independent.

Table 2
Terms that can be estimated given the corresponding vectors of model evaluations, $k = 4$

	$\mathbf{a}_0$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_{234}$	$\mathbf{a}_{134}$	$\mathbf{a}_{124}$	$\mathbf{a}_{123}$	$\mathbf{a}_{1234}$
$\mathbf{a}_0$	$\widehat{V}(y)$									
$\mathbf{a}_1$	$\widehat{S}_1^T$	$\widehat{V}(y)$								
$\mathbf{a}_2$	$\widehat{S}_2^T$	$\widehat{V}_{34}^c$	$\widehat{V}(y)$							
$\mathbf{a}_3$	$\widehat{S}_3^T$	$\widehat{V}_{24}^c$	$\widehat{V}_{14}^c$	$\widehat{V}(y)$						
$\mathbf{a}_4$	$\widehat{S}_4^T$	$\widehat{V}_{23}^c$	$\widehat{V}_{13}^c$	$\widehat{V}_{12}^c$	$\widehat{V}(y)$					
$\mathbf{a}_{234}$	$\widehat{S}_1$	$\widehat{E}^2(y)$	$\widehat{V}_{12}^c$	$\widehat{V}_{13}^c$	$\widehat{V}_{14}^c$	$\widehat{V}(y)$				
$\mathbf{a}_{134}$	$\widehat{S}_2$	$\widehat{V}_{12}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{23}^c$	$\widehat{V}_{24}^c$	$\widehat{V}_{34}^c$	$\widehat{V}(y)$			
$\mathbf{a}_{124}$	$\widehat{S}_3$	$\widehat{V}_{13}^c$	$\widehat{V}_{23}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{34}^c$	$\widehat{V}_{24}^c$	$\widehat{V}_{14}^c$	$\widehat{V}(y)$		
$\mathbf{a}_{123}$	$\widehat{S}_4$	$\widehat{V}_{14}^c$	$\widehat{V}_{24}^c$	$\widehat{V}_{34}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{23}^c$	$\widehat{V}_{13}^c$	$\widehat{V}_{12}^c$	$\widehat{V}(y)$	
$\mathbf{a}_{1234}$	$\widehat{E}^2(y)$	$\widehat{S}_1$	$\widehat{S}_2$	$\widehat{S}_3$	$\widehat{S}_4$	$\widehat{S}_1^T$	$\widehat{S}_2^T$	$\widehat{S}_3^T$	$\widehat{S}_4^T$	$\widehat{V}(y)$

Table 3
Terms that can be estimated given the corresponding vectors of model evaluations, $k = 3$

	$\mathbf{a}_0$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_{23}$	$\mathbf{a}_{13}$	$\mathbf{a}_{12}$	$\mathbf{a}_{123}$
$\mathbf{a}_0$	$\widehat{V}(y)$							
$\mathbf{a}_1$	$\widehat{S}_1^T$	$\widehat{V}(y)$						
$\mathbf{a}_2$	$\widehat{S}_2^T$	$\widehat{S}_3$	$\widehat{V}(y)$					
$\mathbf{a}_3$	$\widehat{S}_3^T$	$\widehat{S}_2$	$\widehat{S}_1$	$\widehat{V}(y)$				
$\mathbf{a}_{23}$	$\widehat{S}_1$	$\widehat{E}^2(y)$	$\widehat{V}_{12}^c$	$\widehat{V}_{13}^c$	$\widehat{V}(y)$			
$\mathbf{a}_{13}$	$\widehat{S}_2$	$\widehat{V}_{12}^c$	$\widehat{E}^2(y)$	$\widehat{V}_{23}^c$	$\widehat{S}_3$	$\widehat{V}(y)$		
$\mathbf{a}_{12}$	$\widehat{S}_3$	$\widehat{V}_{13}^c$	$\widehat{V}_{23}^c$	$\widehat{E}^2(y)$	$\widehat{S}_2$	$\widehat{S}_1$	$\widehat{V}(y)$	
$\mathbf{a}_{123}$	$\widehat{E}^2(y)$	$\widehat{S}_1$	$\widehat{S}_2$	$\widehat{S}_3$	$\widehat{S}_1^T$	$\widehat{S}_2^T$	$\widehat{S}_3^T$	$\widehat{V}(y)$

For $k = 3$ we obtain (Table 3):

(1) Double estimates for each of the $3\widehat{S}_i$ and each of the $3\widehat{S}_i^T$.
(2) Double estimates of the $\binom{3}{2} = 3$ second order terms $\widehat{V}_{ij}^c$.
(3) Two more estimates for each of the $3\widehat{S}_i$.

The case $k = 2$ is non-relevant, as $V_{12}^c = V(y)$.

3. Discussion of the methodological advantages

What benefit does the new computational procedure offer? The main advantage of the new method is that, given the computational effort already made to compute a full set of $\widehat{S}_i$, $\widehat{S}_i^T$ estimates, one can also obtain additional estimates.

Imagine, for the case of $k \geqslant 5$, that the reduced procedure of Theorem 1 has been adopted, and that $\mathbf{a}_0$, $\mathbf{a}_{i_1,i_2,\ldots,i_k}$, and either of the sets $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ or $\{\mathbf{a}_{-1}, \mathbf{a}_{-2}, \ldots, \mathbf{a}_{-k}\}$ has been computed (at the cost of $n(k+2)$ model

evaluations). Beside $\widehat{S}_i$, $\widehat{S}_i^T$, we now also dispose of the $\widehat{S}^c_{-ij}$ indices. These can be very useful. If $\widehat{S}^c_{-ij} \approx 1$, it means that neither x_j nor x_i contribute appreciably to the variance of y, neither by themselves nor in cooperation with other factors. These factors could hence be fixed in a subsequent analysis. Note that the condition $\widehat{S}^c_{-ij} \approx 1$ is equivalent to $\widehat{S}_i^T \approx 0, \widehat{S}_j^T \approx 0$.

If we use instead the extended procedure of Theorem 2, at the cost of $n(2k+2)$ model evaluations, we obtain double estimates of $\widehat{S}_i$, $\widehat{S}_i^T$, double estimates of the $\widehat{S}^c_{-ij}$ indices and finally double estimates of $\widehat{V}^c_{ij}$ for the closed effects of couples of factors, i.e. $V^c_{ij} = V_i + V_j + V_{ij}$. These can be used as such or converted into second order sensitivity coefficients $\widehat{S}_{ij} = (\widehat{V}^c_{ij} - \widehat{V}_i - \widehat{V}_j)/\widehat{V}(y)$. A full set of second order coefficients is likely to tell us much of what we need to know about a model sensitivity, also because interaction of higher orders should in general be less frequent, as discussed in [10].We know from the value of S_j if a factor is influent at the first order, and from S_j^T whether it has important interactions. If this is the case, inspection of the S_{ij} for all $i \neq j$ will allow us to identify which factor x_j interacts with.

4. The case of the correlated input

Sensitivity analysis for correlated input is discussed in [8,16]. For this setting, the important computational simplifications described in Section 1 are not applicable, and Eq. (17) loses its uniqueness. In these cases there is no alternative to the computation of the double loop needed to estimate conditional variances such as $V(E(y \mid x_j))$, Eq. (6). For the purpose of Monte Carlo simulations, correlated input can be generated using Markov Chain Monte Carlo (MCMC), or procedures based on Cholesky decompositions (see, e.g., [6]) or on Latin Hypercube Sampling (LHS, [8]). The problem with correlated sample, in brief, is that the reduction in variance that can be achieved by fixing one factor depends on whether or not other factors have been fixed, and the incremental reduction in variance for each factor depends on the order in which factors are fixed.

We have discussed in [16] two general settings for sensitivity analysis. Each setting is based on a bet posed on the model $y = f(x_1, x_2, \ldots, x_k)$, for the general case where the input can be correlated. In the first bet, one has to make a rational guess on which parameter would induce the largest reduction in variance if it were fixed to its "true" value. Because such true value is in general unknown, the bet can be rationally placed by computing the estimates $\widehat{V}(E(y \mid x_j))$, whether or not the input is correlated.

For the second setting, of relevance in risk analysis and control theory, the bet is on the identification of the smallest subset of $\mathbf{x}$ capable of inducing a target reduction in the unconditional variance $V(y)$, as in the work of Sobol' [19], discussed in Section 1. For the *uncorrelated* case, a rational selection strategy for the subset of interest is based on the computation of the full sets of S_j and S_j^T. This strategy is meant to fight the curse of dimensionality, as attempting all combinations of factors, in a brute-force search for the smallest subset of $\mathbf{x}$ that gives the desired reduction in $V(y)$, would be computationally prohibitive; one would have to compute all $2^k - 1$ terms in Eq. (17) to start with. As described in [16], an iterative procedure can be adopted for the uncorrelated case, that includes as a step the computation of the full set of S_j and S_j^T.

For the correlated case, one might still engage in a brute force search computing all possible closed terms $V^c_{i_1 i_2 \ldots i_s}$. Note that for the correlated case the $V^c_{i_1 i_2 \ldots i_s}$ can no longer be decomposed meaningfully into a sum of lower dimensionality terms, but would still allow a perfectly informed choice, as would the $V_{i_1 i_2 \ldots i_s}$ in the uncorrelated case. Also for the correlated case, we suggest in [16] a cheaper, albeit approximate, alternative that involves the computation of the S_j and S_j^T for the non-correlated problem.

Hence, in the context covered by these problem settings, the procedure proposed in Section 2 can still be usefully applied.

5. Test cases

We illustrate the algorithm on non-correlated test cases. We have selected an analytic function due to Sobol' and known as "Sobol' g function", and a more complex numeric calculus test case originating from modelling of petroleum generation in sedimentary basins. The cost of computing the former can be assumed as zero, while the computation of a single output time series for the latter takes about 0.05 s on a 8-nodes Linux cluster, 16 CPU's with a Pentium 3 processor running under Linux RedHat.

The g function is a strongly non-monotonic, non-additive function of k factors x_i, assumed identically and uniformly distributed in the unit cube $I^k = \{x \mid 0 \leqslant x_i \leqslant 1; i = 1, 2, \ldots, k\}$.

$$g(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} g_i(x_i) \tag{29}$$

with

$$g_i(x_i) = \frac{|4x_i - 2| + a_i}{1 + a_i}. \tag{30}$$

For each of the $g_i(x_i)$ functions $\int_0^1 g_i(x_i)\,\mathrm{d}x_i = 1$, and for $x_i \in [0, 1]$ the function's variation is

$$1 - \frac{1}{1 + a_i} \leqslant g_i(x_i) \leqslant 1 + \frac{1}{1 + a_i}. \tag{31}$$

The importance of each factor x_i is driven by its associated coefficient a_i. For $a_i = 0$, the factor is important ($0 \leqslant g_i(x_i) \leqslant 2$). For, e.g., $a_i = 9$ the factor in non-important ($0.9 \leqslant g_i(x_i) \leqslant 1.1$), while for $a_i = 99$ the factor can be considered as non-influent ($0.99 \leqslant g_i(x_i) \leqslant 1.01$). Analytical expressions are available for the sensitivity indices ([1,15]):

$$\int_{I^k} f(x_1, x_2, \ldots, x_k)\,\mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_k = 1, \tag{32}$$

$$V_{i_1 i_2 \ldots i_s} = V_{i_1} V_{i_2} \ldots V_{i_s}, \tag{33}$$

$$V_i = \int_0^1 \left[g_i(x_i) - 1\right]^2 \mathrm{d}x_i = \tfrac{1}{3}(1 + a_i)^{-2}. \tag{34}$$

In Figs. 1–3 we have computed the sensitivity indices for a 6-dimensional g-function with $\mathbf{a} = \{0, 0.5, 3, 9, 99, 99\}$ using first the standard procedure of Section 1, at the cost of $n(2k + 2)$ model evaluations, then with the restricted procedure of Theorem 1 at the cost of $n(k + 2)$ model evaluations. Finally we have used the setting of Theorem 2, at the cost of $n(2k + 2)$ model evaluations. For all experiments $n = 1024$, and the standard error associated with the computation of the sensitivity indices was computed using bootstrap, as described in [1], with a bootstrap sample dimension of 10,000.

Comparing Figs. 1(a)–1(b) (S_j and S_j^T by the standard procedure) with 2(a)–2(b) (S_j and S_j^T by the procedure of Theorem 1), we can see that the quality of the estimates is the same. Fig. 2(c) shows the advantage brought by the term of the 4th order, especially to identify couples of non-influent factors $ij = \{45, 46, 56\}$.

Moving to the procedure of Theorem 2, Figs. 3(a)–3(b), we see that the confidence bound of the estimates is lower (each estimate is the average of 2). Similarly for Fig. 3(c), to be compared with 2(c). Additional insight into the structure of the model is offered by the new Fig. 3(d), with the second order indices.

The PMOD model computes the generation and expulsion of hydrocarbons from a host rock, and is part of a suite of computer models used to estimate the oil potential of sedimentary basins. PMOD has been originally developed at Lawrence Livermore Laboratory ([3]) and adapted at ENI-AGIP for its basin modelling environment ([12]).

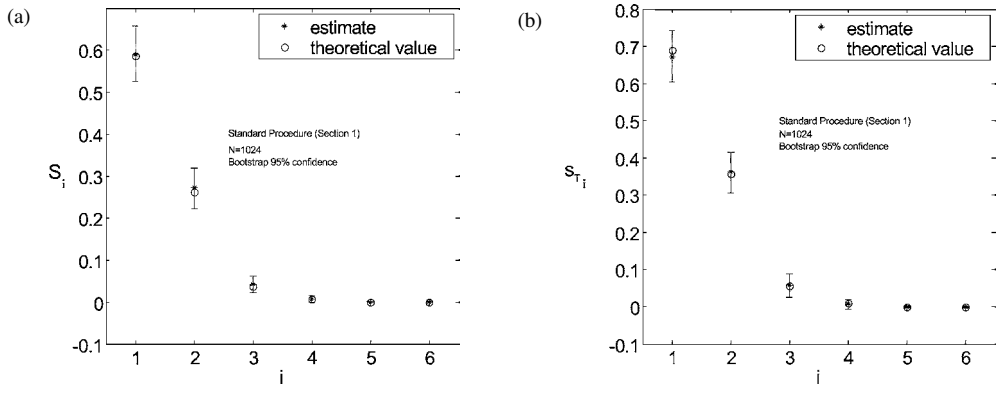

Fig. 1. First order (a) and total effect (b) indices for the g function with bootstrap-estimated 95% confidence bounds using the standard procedure of Section 1.

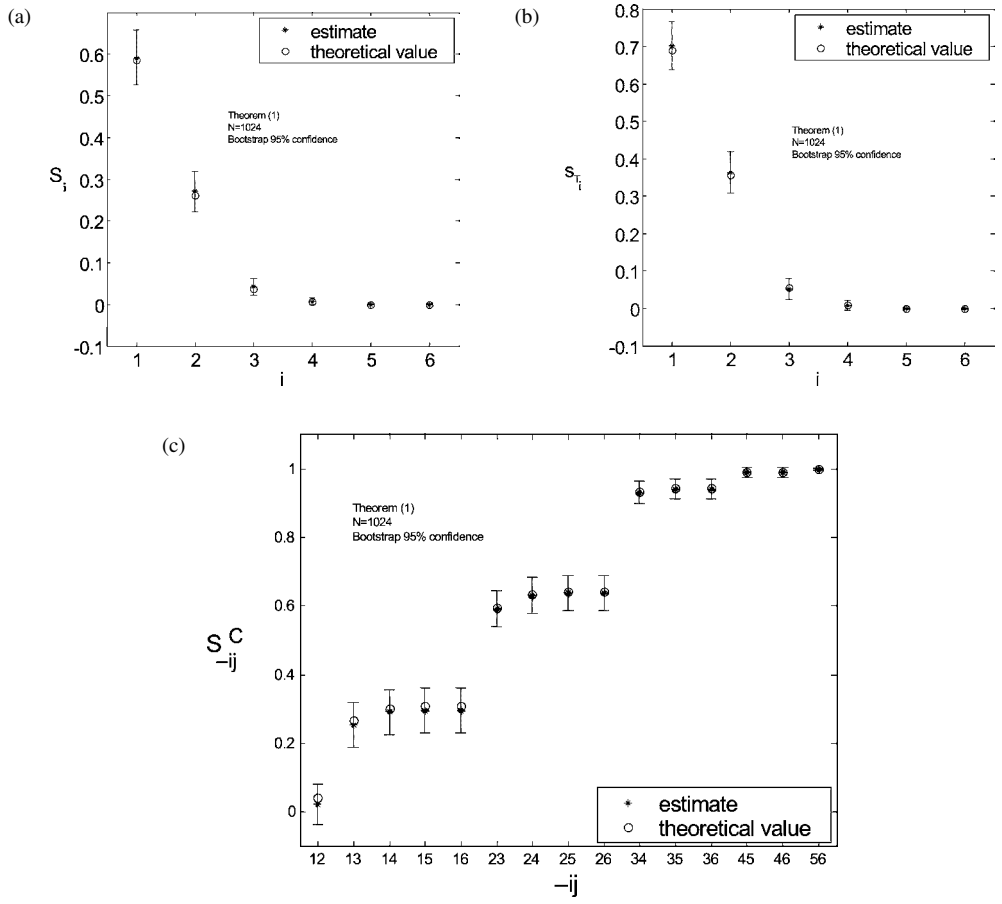

Fig. 2. First order (a), total effect (b), and closed effect of order $(6-2)=4$ (c) for the g function with bootstrap-estimated 95% confidence bounds using the procedure of Theorem 1.

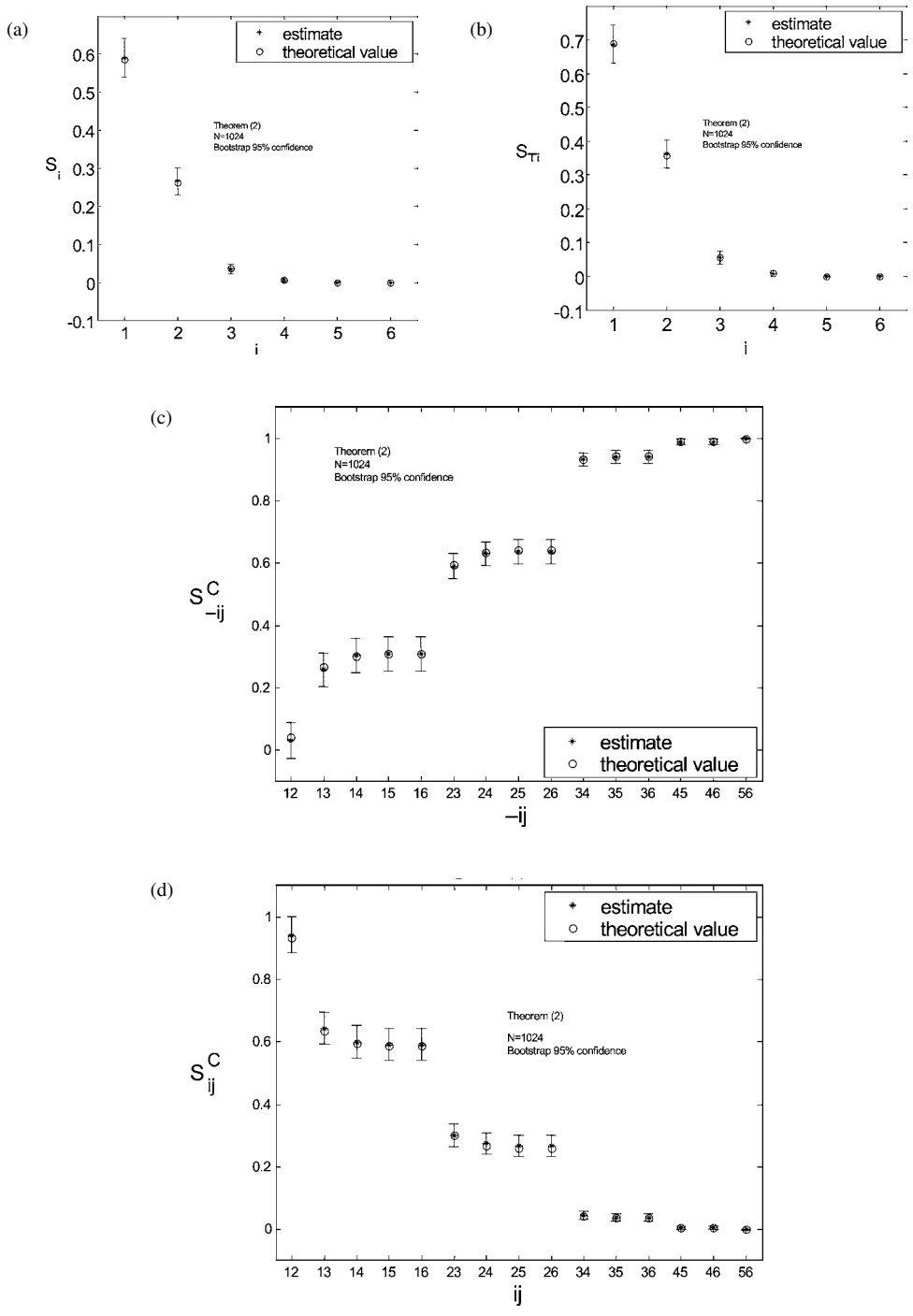

Fig. 3. First order (a), total effect (b), closed effect of order $(6-2)=4$ (c), and closed effect of order 2 (d) for the g function with bootstrap-estimated 95% confidence bounds using the procedure of Theorem 1.

Table 4
List of the uncertain input factors and their stochastic properties

Factor's name	Factor's number	Type	Range of values	PDF
"KEM/FIZ" files	1	Discrete	$1, \ldots, 8$	Uniform
"phi-stress" curves	2	Discrete	$1, 2, 3$	Uniform
TOC	3	Continuous	0.005–0.05	Uniform
Porosity	4	Continuous	Min = 0.04 Mode = 0.05 Max = 0.09	Triangular
Permeability	5	Continuous	1.e–9–1.e–6	Log-uniform
Source thickness	6	Continuous	Min = 907 Mode = 1814 Max = 2721	Triangular
Time-series	7	Discrete	$1, 2, \ldots, 32$	Uniform

Some of the uncertain input factors in PMOD are time-dependent physical quantities: the rock's total organic carbon content (TOC), the rock's porosity, permeability and thickness. In the analysis, the values at the initial time point, that corresponds to 30 million years before present (mybp), have been considered, neglecting the time dependency.

One of the inputs, the PHI stress variable, describes the mechanic behaviour of the rock. Due to lack of data on the specific site, three different curves relative to similar sedimentary basins for other areas have been used, and the model selects randomly which of the three to use at runtime. Similarly for the so-called "KEM/FIZ" files, that describe the stoichiometrics and kinetics of the chemical system considered. Eight alternative such descriptions were generated by the experts, so that sampling from these might be considered as representative of the system chemistry's variability. Also in this case the model selects at runtime one of the eight files. Finally the model needs as input 4 highly correlated time series (temperature, pressure, effective stress and hydrostatic pressure). Thirty-two such multivariate series have been generated by the experts and the model selects one set at random for each execution of the PMOD model. A summary of the 7 input factors is given in Table 4. All these factors are considered independent from each other.

A sensitivity analysis for this model had already been performed before the algorithms presented in the present paper were developed ([20]), where all coefficients of the first and total order had been computed. We have hence repeated the analysis using the setting of Theorem 2. The model output is composed of cumulative expelled amounts of oil, gas (CH4) and wet gas (CHx) at selected time points (i.e. at 30, 11.5, 8.5, 4.8, 1.9, 0 mybp). Figs. 4(a)–4(c) show the first order sensitivity indices obtained with the two approaches at different time points for the ouput CH4, as almost identical results hold for CHx and oil. Similarly for the total order indices, Figs. 5(a)–5(c).

In Figs. 6(a)–6(c) the coefficients of the second orders have been also computed using the set of Theorem 2, and compared with estimates obtained previously using an independent sample of size n for each index (i.e. $21n$ additional runs, from [20]).

For all these Figs. 5–6 there is a general agreement between the two methods, and the confidence bounds, computed exactly as in the previous case study, are lower for the estimate from Theorem 2, as expected.

Finally the closed indices of order 5 are given in Fig. 7, as computed with Theorem 2. The results for Theorem 1 are very similar, as expected, and not shown here.

With the new procedures, we have been able to compute at no extra (or at a reduced) computational cost the coefficients of order $(k-2)$, that allow us to identify the non influential factors. Given the highly non linear and non additive nature of PMOD, there are time points, such as $t = 8.5$ mybp, where none of the factors is non-influent, apart from source thickness and porosity (Fig. 7(b)). At time $t = 0$ mybp the coefficients of order $(k-2)$ help us to rule out as non-influent all factors except TOC.

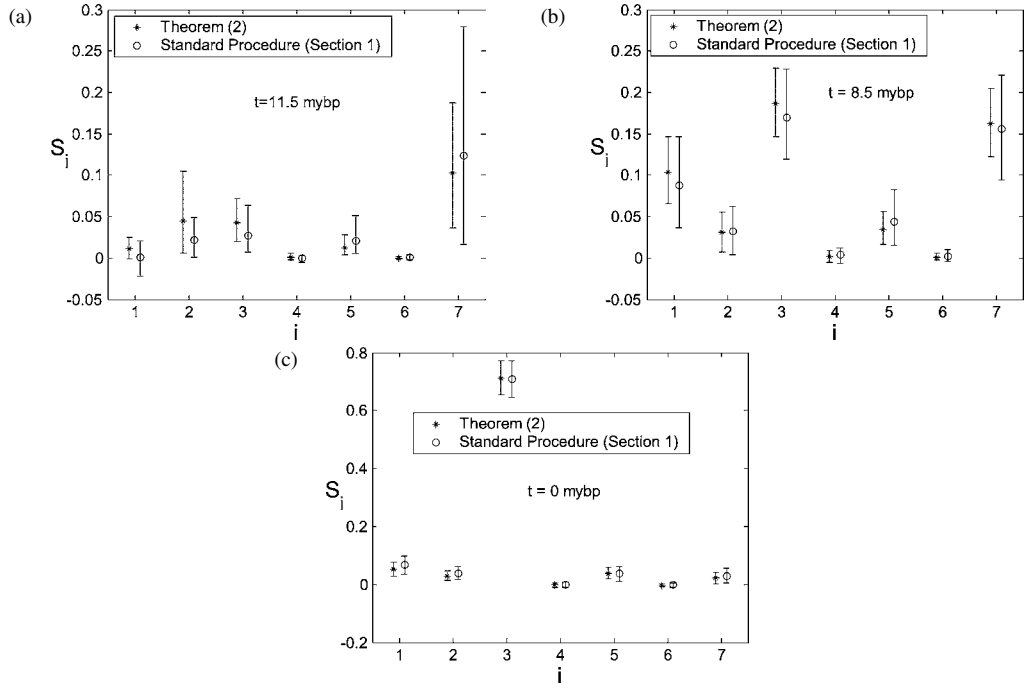

Fig. 4. Main effect sensitivity indices, with bootstrap-estimated 95% confidence bounds, for CH4 at three different time points: 11.5 (a), 8.5 (b) and 0 (c) million years before present.

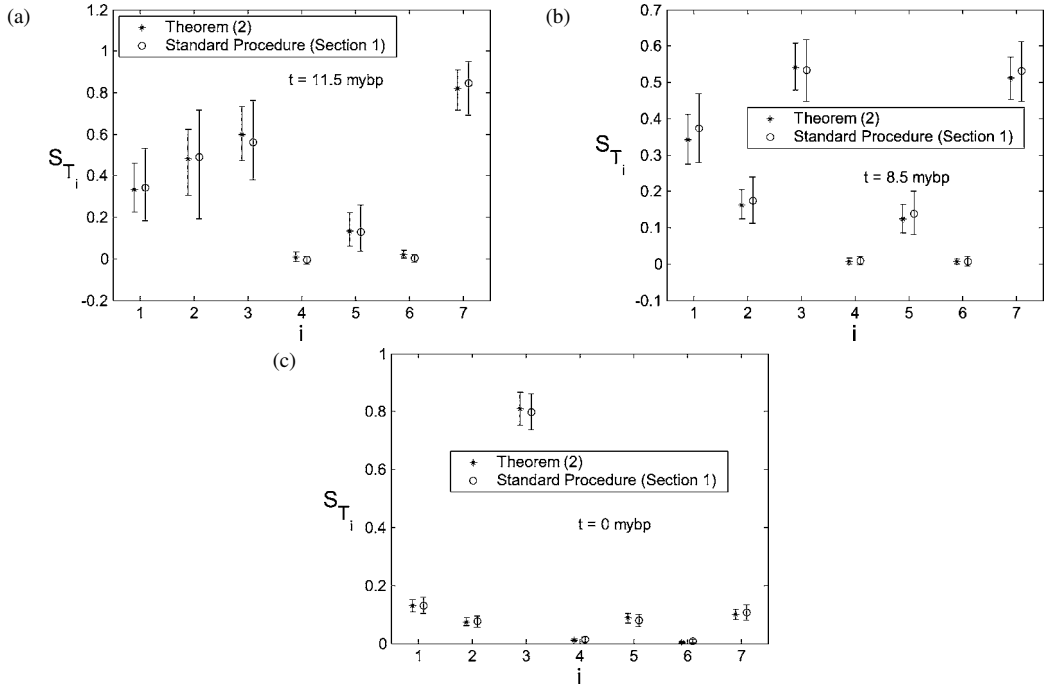

Fig. 5. Total effect sensitivity indices, with bootstrap-estimated 95% confidence bounds, for CH4 at three different time points: 11.5 (a), 8.5 (b) and 0 (c) million years before present.

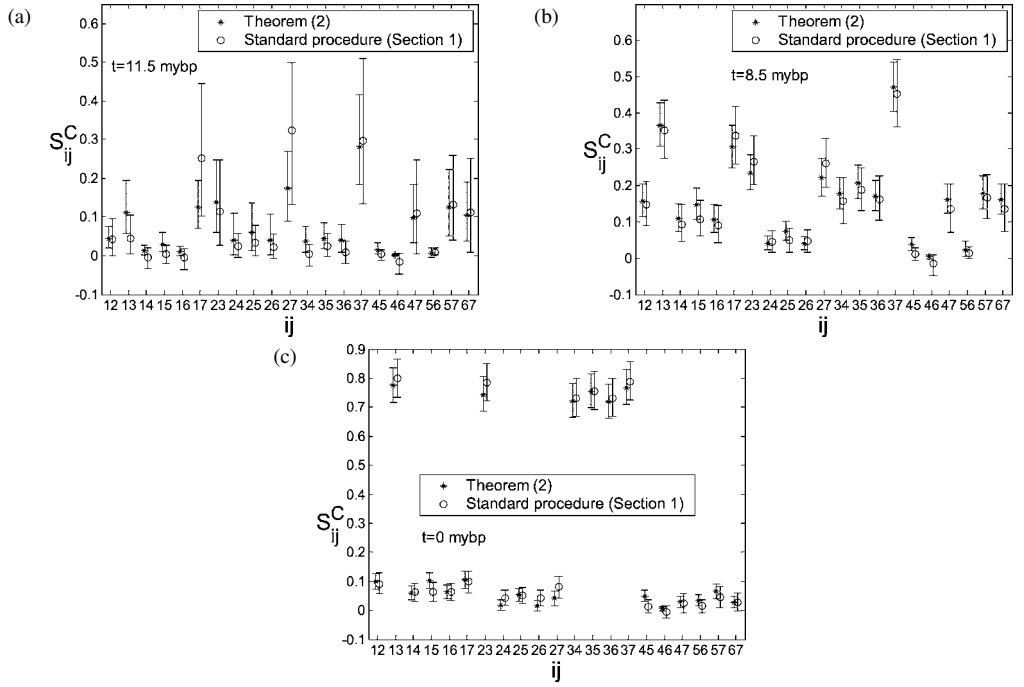

Fig. 6. Second order closed indices, with bootstrap-estimated 95% confidence bounds, for CH4 at three different time points: 11.5 (a), 8.5 (b) and 0 (c) million years before present.

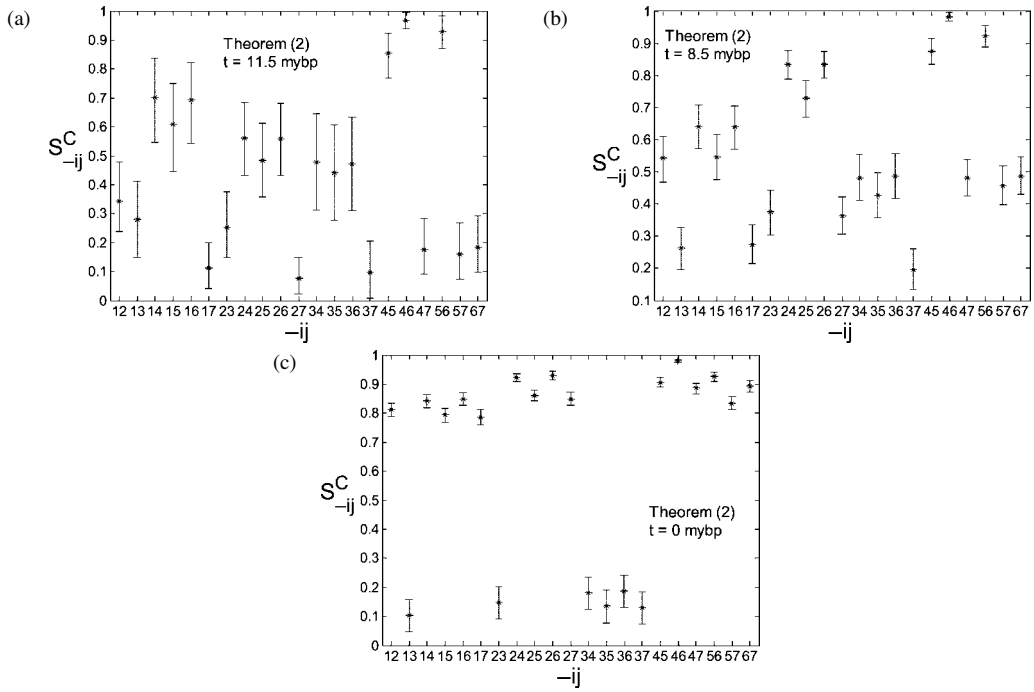

Fig. 7. Fifth order closed sensitivity indices, with bootstrap-estimated 95% confidence bounds, for CH4 at three different time points: 11.5 (a), 8.5 (b) and 0 (c) million years before present.

6. Conclusions

The present paper has suggested some efficient procedures for numerical experiments aimed at sensitivity analysis of model output. We have focused here on the computation of sensitivity indices that are based on decomposing the variance of the target function in a quantitative fashion. The approach presented here opens a road to fight the so-called "curse of dimensionality", that hinders the use of quantitative sensitivity analysis for computationally expensive models. The analyst willing to use such methods disposes now of two approaches to tackle the system. One is a parsimonious procedure (Theorem 1) that gives all effects of the first and total order, plus all those of order $k-2$, at the cost of $n(k+2)$ simulations. We have thus both reduced the computational cost for the standard procedure of some 50% and extended it to compute the indices of order $k-2$.

A second possible approach is the more expensive procedure (Theorem 2, cost $= n(2k+2)$) that gives more robust estimates of the index of the first and total orders, plus estimates of all indices of order 2 and $k-2$.

Even assuming for n the value of 1000, the two procedures appear affordable for models whose cost per run is in the range from milliseconds or lower to some minutes. For models whose execution is in the tenths of minutes to a day range, quantitative methods are not applicable and efficient qualitative methods such as that of Morris [9] should be used (see [4] for a review).

Acknowledgements

The author is grateful to Marco Ratto for MATLAB coding of the algorithm, and Stefano Tarantola, Michaela Saisana, for reviewing this manuscript. Paolo Ruffo and Anna Corradi of ENI (AGIP Division) kindly produced the PMOD evaluations and allowed their use in the present work. ENI-AGIP funded part of the this study.

References

[1] G. Archer, A. Saltelli, I.M. Sobol', Sensitivity measures, ANOVA like techniques and the use of bootstrap, J. Stat. Comput. Simulation 58 (1997) 99–120.

[2] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, John Wiley and Sons, New York, 1978.

[3] R.L. Braun, A.K. Burnham, User's manual for PMOD, A pyrolysis and primary migration model, Lawrence Livermore National Laboratory report, 1993.

[4] F. Campolongo, A. Saltelli, in: A. Saltelli, K. Chan, M. Scott (Eds.), Design of experiment, Sensitivity Analysis, Probability and Statistics Series, John Wiley, 2000, pp. 51–63.

[5] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of model output, Reliability Engrg. System Safety 52 (1) (1996) 1–17.

[6] R.L. Iman, W.J. Conover, A distribution free approach to inducing rank correlation among input variables, Comm. Stat. B 11 (3) (1982) 311–334.

[7] T. Ishigami, T. Homma, An importance quantification technique in uncertainty analysis for computer models, in: Proceedings of the ISUMA'90, First International Symposium on Uncertainty Modelling and Analysis, December 3–6, University of Maryland, 1990.

[8] M.D. McKay, Variance-based methods for assessing uncertainty importance in NUREG-1150 analyses, LA-UR-96-2695, 1996, pp. 1–27.

[9] M.D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics 33 (2) (1991) 161–174.

[10] H. Rabitz, Ö.F. Aliş, in: A. Saltelli et al. (Eds.), Managing the Tyranny of Parameters in Mathematical Modelling of Physical Systems, 2000.

[11] H. Rabitz, Ö.F. Aliş, J. Shorter, K. Shim, Efficient input–output representations, Comput. Phys. Comm. 117 (1,2) 11–20.

[12] P. Ruffo, Personal communication (2001).

[13] Saltelli A., T.H. Andres, T. Homma, Some new techniques in sensitivity analysis of model output, Comput. Statist. Data Anal. 15 (1993) 211–238.

[14] A. Saltelli, K. Chan, M. Scott (Eds.), Sensitivity Analysis, Probability and Statistics Series, John Wiley, 2000.

[15] A. Saltelli, I.M. Sobol', About the use of rank transformation in sensitivity analysis of model output, Reliability Engrg. Syst. Safety 50 (1995) 225–239.

[16] A. Saltelli, S. Tarantola, On the relative importance of input factor in mathematical models, J. Amer. Statist. Assoc. (2002), in press.

[17] A. Saltelli, S. Tarantola, F. Campolongo, Sensitivity analysis as an ingredient of modelling, Statist. Sci. 15 (4) (2000) 377–395.
[18] A. Saltelli, S. Tarantola, K. Chan, A quantitative, model independent method for global sensitivity analysis of model output, Technometrics 41 (1) (1999) 39–56.
[19] I.M. Sobol', Sensitivity estimates for nonlinear mathematical models, Mat. Model. 2 (1990) 112–118;[2] Transl.: I.M. Sobol', Sensitivity analysis for nonlinear mathematical models, Math. Modelling Comput. Exp. 1 (1993) 407–414.
[20] S. Tarantola, A. Corradi, P. Ruffo, A. Saltelli, Global sensitivity analysis techniques for the analysis of the oil potential of sedimentary basins, in: P. Prado, R. Bolado (Eds.), 3rd International Symposium on Sensitivity Analysis of Model Output, Proceedings of SAMO 2001, CIEMAT Publication, Madrid, 2001.

[2] In Russian.