The battle of total-order sensitivity estimators

Arnald Puy*[1,2], William Becker[3], Samuele Lo Piano[4], and Andrea Saltelli[2,3]

[1] *Department of Ecology and Evolutionary Biology, M31 Guyot Hall, Princeton University, New Jersey 08544, USA. E-Mail: apuy@princeton.edu*
[2] *Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Parkveien 9, PB 7805, 5020 Bergen, Norway.*
[3] *European Commission, Joint Research Centre, Via Enrico Fermi, 2749, 21027 Ispra VA, Italy*
[4] *University of Reading, School of the Built Environment, JJ Thompson Building, Whiteknights Campus, Reading, RG6 6AF, United Kingdom*

Summary

Sensitivity analysis helps decision-makers to understand how a given model output responds when there is variation in the model inputs. One of the most authoritative measures in global sensitivity analysis is the Sobol' total-order index (T_i), which can be computed with several different estimators. Although previous comparisons exist, it is hard to know which estimator performs best since the results are contingent on several benchmark settings: the sampling method (τ), the distribution of the model inputs (ϕ), the number of model runs (N_t), the test function or model (ε) and its dimensionality (k), the weight of higher order effects (e.g. second, third, k_2, k_3), or the performance measure selected (δ). Here we break these limitations and simultaneously assess all total-order estimators in an eight-dimension hypercube where $(\tau, \phi, N_t, \varepsilon, k, k_2, k_3, \delta)$ are treated as random parameters. This design allows to create an unprecedentedly large range of benchmark scenarios. Our results indicate that, in general, the preferred estimator should be Razavi and Gupta's, followed by that of Jansen, or Janon/Monod. The remainder lag significantly behind in performance. Our work helps analysts navigate the myriad of total-order formulae by effectively eliminating the uncertainty in the selection of the best estimator.

Keywords: Uncertainty modelling; Sensitivity analysis; Modelling; Sobol' indices; Variance decomposition

1 Introduction

Sensitivity analysis, i.e. the assessment of how much uncertainty in a given model output is conveyed by each model input, is a fundamental step to judge the quality of model-based inferences [1–3]. In operations research, sensitivity analysis helps practitioners grasp the model's behaviour, structure and response, thus increasing the awareness of the managerial

*Corresponding author

problem at hand [4]. Among the many sensitivity indices available, variance-based indices are widely regarded as the gold standard because they are model-free (no assumptions are made about the model), global (they account for interactions between the model inputs) and easy to interpret, as they are anchored in the statistical theory of variance decomposition (ANOVA) [5–8].

Given a model of the form $y = f(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2, ..., x_k) \in \mathbb{R}^k$, where y is a scalar output and $x_1, ..., x_k$ are the k independent model inputs, the variance of y is decomposed into conditional terms as

$$V(y) = \sum_{i=1}^{k} V_i + \sum_{i} \sum_{i<j} V_{ij} + ... + V_{1,2,...,k}, \tag{1}$$

where

$$\begin{aligned} V_i &= V_{x_i}\left[E_{\boldsymbol{x}_{\sim i}}(y|x_i)\right] \\ V_{ij} &= V_{x_i,x_j}\left[E_{\boldsymbol{x}_{\sim i,j}}(y|x_i,x_j)\right] \\ &\quad - V_{x_i}\left[E_{\boldsymbol{x}_{\sim i}}(y|x_i)\right] \\ &\quad - V_{x_j}\left[E_{\boldsymbol{x}_{\sim j}}(y|x_j)\right] \end{aligned} \tag{2}$$

and so on up to the k-th order. The notation $\boldsymbol{x}_{\sim i}$ means *all-but-*x_i. By dividing each term in Equation 1 by the unconditional model output variance $V(y)$, we obtain the first-order indices for single inputs (S_i), pairs of inputs (S_{ij}), and for all higher-order terms. First-order indices thus provide the proportion of $V(y)$ caused by each term, and are widely used to rank model inputs according to their contribution to the model output uncertainty. They are particularly effective when the objective of the analysis is to find the input which, if known, would on average lead to the highest reduction in the variance of y (a goal known as "factor prioritization" since it is helpful in prioritising data collection or research to reduce model uncertainty [1]).

Computing all $2^k - 1$ terms in Equation 1 might be impractical when k is large. In such cases, Homma & Saltelli [9] proposed the calculation of the total-order index T_i, which measures the first-order effect of a model input as well as all its interactions up to the k-th order:

$$T_i = 1 - \frac{V_{\boldsymbol{x}_{\sim i}}\left[E_{x_i}(y|\boldsymbol{x}_{\sim i})\right]}{V(y)} = \frac{E_{\boldsymbol{x}_{\sim i}}\left[V_{x_i}(y|\boldsymbol{x}_{\sim i})\right]}{V(y)} \tag{3}$$

When $T_i \approx 0$, it can be concluded that x_i has a negligible contribution to $V(y)$. For this reason, total-order indices have been applied to distinguish influential from non-influential model inputs and reduce model dimensionality, a setting also known as "factor-fixing" [1].

The most direct computation of T_i is via Monte Carlo (MC) estimation because it does not impose any assumption on the functional form of the response function (unlike meta-modeling approaches), and is simpler to implement than Fourier Amplitude Sensitivity Test (FAST) [10, 11].

In Table 1 we show a number of different MC-based formulae available to estimate T_i [9, 12–20]. All require the creation of a sample matrix using a random or quasi-random

design (e.g. Latin Hypercube Sampling [21], the Sobol' sequence [22, 23]). This matrix is structured such that each row represents a sample point and each column represents a model input. The estimators differ however in the particularities of the sampling design. These differences are summarised as follows, with numbers referring to Table 1:

Table 1: Formulae to compute T_i. For estimators 1–3, $V(y) = \frac{1}{N}\sum_{j=1}^{N}[f(\mathbf{A})_j - f_0]^2$, and $f_0 = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j$. For estimator 4, $f_0 = \left(\frac{1}{N}\sum_{j=1}^{N}\frac{f(\mathbf{A})_j + f(\mathbf{A}_B^{(i)})_j}{2}\right)^2$. In estimator 5, $\langle f(\mathbf{A}_j)\rangle$ is the mean of $f(\mathbf{A}_j)$. We used a simplified version of the Glen and Isaacs estimator because the presence of spurious correlations between $\mathbf{A}$ and $\mathbf{A}_B^{(i)}$) is likely to be close to zero. As for estimator 7, we refer to it as pseudo-Owen given its use of a $\mathbf{C}$ matrix and its identification with Owen [19] in Iooss *et al.* [24], where we retrieved the formula from.

N	Estimator	Author
1	$\frac{\frac{1}{2N}\sum_{j=1}^{N}\left[f(\mathbf{A})_j - f(\mathbf{A}_B^{(i)})_j\right]^2}{V(y)}$	Jansen [15]
2	$\frac{\frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j\left[f(\mathbf{A})_j - f(\mathbf{A}_B^{(i)})_j\right]}{V(y)}$	Sobol' [16]
3	$\frac{V(y) - \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A}_j) f(\mathbf{A}_B^{(i)})_j + f_0^2}{V(y)}$	Homma & Saltelli [9]
4	$1 - \frac{\frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j f(\mathbf{A}_B^{(i)})_j - f_0^2}{\frac{1}{N}\sum_{j=1}^{N}\frac{f(\mathbf{A}_j)^2 + f(\mathbf{A}_B^{(i)})_j^2}{2} - f_0^2}$	Janon *et al.* [25] Monod *et al.* [17]
5	$1 - \left[\frac{1}{N-1}\sum_{j=1}^{N}\frac{[f(\mathbf{A}_j) - \langle f(\mathbf{A}_j)\rangle]\left[f(\mathbf{A}_B^{(i)})_j - \left\langle f(\mathbf{A}_B^{(i)})_j\right\rangle\right]}{\sqrt{V[f(\mathbf{A}_j)]V\left[f(\mathbf{A}_B^{(i)})_j\right]}}\right]$	Glen & Isaacs [18]
6	$1 - \frac{\sum_{j=1}^{N}\left[(f(\mathbf{A})_j - f(\mathbf{B}_A^{(i)})_j)\right]\left[(f(\mathbf{B})_j - f(\mathbf{A}_B^{(i)})_j)\right]}{\frac{1}{2}\sum_{j=1}^{N}\left[(f(\mathbf{A})_j - f(\mathbf{B})_j)^2 + (f(\mathbf{A}_B^{(i)})_j - f(\mathbf{B}_A^{(i)})_j)^2\right]}$	Azzini & Rosati [13]
7	$\frac{V(y) - \left[\frac{1}{N}\sum_{j=1}^{N}\left\{\left[f(\mathbf{B})_j - f(\mathbf{C}_B^{(i)})_j\right]\left[f(\mathbf{B}_A^{(i)})_j - f(\mathbf{A})_j\right]\right\}\right]}{V(y)}$	pseudo-Owen
8	$\frac{E_{x^*_{\sim i}}\left[\gamma_{x^*_{\sim i}}(h_i)\right] + E_{x^*_{\sim i}}\left[C_{x^*_{\sim i}}(h_i)\right]}{V(y)}$	Razavi & Gupta [20, 26] (see SM).

- Estimators 1–5 require a $(N, 2k)$ sample matrix in which the first k columns are allocated to an $\mathbf{A}$ matrix and the remaining k columns to a $\mathbf{B}$ matrix. Then, k additional $\mathbf{A}_B^{(i)}$ matrices are created, where all columns come from $\mathbf{A}$ except the i-th column, which comes from $\mathbf{B}$. The total computational cost of this design is $N_t = N(k+1)$ [9, 15–17, 25].
- Estimator 6 differs from Estimators 1–5 in that it also demands the creation of k additional $\mathbf{B}_A^{(i)}$ matrices (defined analogously to $\mathbf{A}_B^{(i)}$), leading to $N_t = N(2k+2)$ [13, 27].
- Estimator 7 requires a $(N, 3k)$ sample matrix, with the last k columns being allocated

to a $\mathbf{C}$ matrix. The need for k $\mathbf{C}_B^{(i)}$ and $\mathbf{B}_A^{(i)}$ matrices sets the total computational cost at $N_t = N(2k+2)$ [24].

- Estimator 8 relies on a different sampling design based on star centers and cross-sections along k dimensions [20, 26]. Its cost is $N_t = N\left[k((1/\Delta h)-1)+1\right]$, where Δh is a tuning parameter specifying the distance between two points $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ in the input space. We provide further details about estimator 8, known as Variogram Analysis of Response Surfaces (VARS-TO), in the Supplementary Materials.

1.1 Uncertainties in the benchmark settings

The performance of the estimators in Table 1 has been assessed in many studies [12–14, 20, 26, 28]. However, Becker [29] observed that their results are very much conditional on the choice of model or test function, its dimensionality, and the number of available model runs. Moreover, there are other factors besides these that are likely to influence a benchmarking study. Hereafter we provide a more complete list:

- *The sampling method:* The creation of the base sample matrix can be done using Monte-Carlo (MC) or quasi Monte-Carlo (QMC) methods. Compared to MC, QMC allows to more effectively map the input space as it leaves smaller unexplored volumes (Fig. S1). However, Kucherenko *et al.* [30] observed that MC methods might help obtain more accurate sensitivity indices when the model under examination has important high-order terms. Both MC and QMC have been used in the base sample matrix to benchmark sensitivity indices (i.e. Saltelli *et al.* [14] and Jansen [15]).

- *The form of the test function:* some of the most commonly used in sensitivity analysis are the Ishigami function [31], the Sobol' G-function and its variants [14, 32], the Bratley & Fox [33] function, or the set of functions presented in Kucherenko *et al.* [30] [12–14, 28]. Despite being analytically tractable, they represent a fairly narrow range of model behaviour, and the effects of nonlinearities and nonadditivities is typically unknown in real models. This *black-box* nature of models has become more of a concern in the last decades due to the increase in computational power and code complexity (which prevents the analyst from intuitively grasping the model's behaviour Borgonovo & Plischke [4]), and to the higher demand for model transparency [3, 34]. This renders the functional form of the model similar to a random variable [29], something not accounted for by previous works [12–14, 28].

- *The function dimensionality:* many studies focus on low-dimensional problems, either by using test functions that only require a few model inputs (e.g. the Ishigami function, where $k = 3$), or by using test functions with a flexible dimensionality, but setting k at a small value of e.g. $k \leq 8$ (Sobol' [32]'s G or Bratley & Fox [33] functions). This approach trades computational manageability for comprehensiveness: by neglecting higher dimensions, it is difficult to tell which estimator might work best in models with tens or hundreds of parameters. Examples of such models can be readily found in the Earth and Environmental Sciences domain [35], including the Soil and Water Assessment Tool (SWAT) model, where $k = 50$ [36], or the Modélisation Environmentale-Surface et Hydrologie (MESH) model, where $k = 111$ [37].

- *The distribution of the model inputs*: the large majority of benchmarking exercises assume uniformly-distributed inputs $p(\boldsymbol{x}) \in U(0,1)^k$ [12–14, 28]. However, there is evidence suggesting that the accuracy of T_i estimators might be sensitive to the underlying model input distributions, to the point of overturning the model input ranks [38, 39].

- *The number of model runs:* sensitivity test functions are generally not computationally expensive and can be run without much concern for computational time. This is frequently not the case for real models, whose high dimensionality and complexity might set a constraint on the total number of model runs available. Under such restrictions, the performance of the estimators of the total-order index depends on their efficiency i.e. how accurate they are given the budget of runs that can be allocated to each model input. There are no specific guidelines as to which total-order estimator might work best under these circumstances [29].

- *The performance measure selected:* typically, a sensitivity estimator has been considered to outperform the rest if, on average, it displays a smaller mean absolute error (MAE), computed as

$$MAE = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\sum_{i=1}^{k}|T_i - \hat{T}_i|}{k}\right) \tag{4}$$

  where p is the number of replicas of the sample matrix, and $\boldsymbol{T_i}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{i}}$ vectors with the analytical and the estimated T_i values respectively. The use of the MAE is appropriate when the aim is to assess which estimator better approaches the true total-order indices, because it averages the error for both influential and non-influential indices. However, the analyst might be more interested in using the estimated indices $\hat{\boldsymbol{T}}_i$ to accurately rank parameters or screen influential from non-influential model inputs [1]. In such context, the MAE may be best substituted or complemented with a measure of rank concordance between $\boldsymbol{T}_i$ and $\hat{\boldsymbol{T}}_i$, such as the Spearman's ρ or the Kendall's W coefficient [29, 40, 41]. It can also be the case that disagreements on the exact ranking of low-ranked parameters may have no practical importance because the interest lies in the correct identification of top ranks only [35]. Savage Scores or other measures that emphasize this top-down correlation are then a more suitable choice [42].

All these sources of uncertainty limit our capacity to extrapolate the results of the comparisons between total-order estimators beyond the settings of the existing benchmarking exercises. This raises the question: are some estimators *generally* better than others? And under which circumstances should we choose one estimator over another? It is hard to say from previous works whether one estimator outperforming another truly reflects its higher accuracy or simply its better performance under the narrow statistical design of the study. This ambiguity hampers the selection of the most appropriate estimator given the wide range of simulation models that require sensitivity analysis.

Here we reduce these limitations and compare the performance of all estimators in Table 1 by treating the sampling method τ, the underlying model input distribution ϕ, the number of

model runs N_t, the test function ε, its dimensionality and degree of non-additivity (k, k_2, k_3) and the performance measure δ as random parameters [29]. This better reflects the diversity of models and sensitivity settings available to the analyst. By relaxing the dependency of the results on these benchmark parameters[1], we define an unprecedentedly large battle ground where all formulae can prove their accuracy. We therefore follow the approach of Becker [29] but extend it in a number of ways; notably by testing a wider set of Monte Carlo estimators, by exploring a wider range of benchmarking assumptions and by performing a formal sensitivity analysis on these assumptions.

2 Materials and methods

2.1 Probability distributions and the metafunction approach

To assess how uncertainties in the benchmark settings condition the performance of T_i estimators, we characterized $(\tau, N_t, k, k_2, k_3, \phi, \epsilon, \delta)$ as uncertain parameters and described them using the probability distributions presented in Table 2.

Table 2: Summary of the parameters and their distributions. $\mathcal{DU}$ is discrete uniform.

Parameter	Description	Distribution
τ	Sampling method	$\mathcal{DU}(1,2)$
N_t	Total number of model runs	$\mathcal{DU}(10,1000)$
k	Number of model inputs	$\mathcal{DU}(3,100)$
ϕ	Probability distribution of the model inputs	$\mathcal{DU}(1,8)$
ε	Randomness in the test function	$\mathcal{DU}(1,200)$
k_2	Fraction of pairwise interactions	$\mathcal{U}(0.3,0.5)$
k_3	Fraction of three-wise interactions	$\mathcal{U}(0.1,0.3)$
δ	Selection of the performance measure	$\mathcal{DU}(1,3)$

We chose $\tau \sim \mathcal{DU}(1,2)$ to check how the performance of T_i estimators is conditioned by the use of Monte-Carlo ($\tau = 1$) or Quasi Monte-Carlo ($\tau = 2$) methods in the creation of the sampling matrix [22, 23]. The total number of model runs and inputs was respectively described as $N_t \sim \mathcal{DU}(10,1000)$ and $k \sim \mathcal{DU}(3,100)$ to explore the performance of the estimators in a wide range of N_t, k combinations. Given the sampling constraints set by the estimators' reliance on either a $\mathbf{B}$, $\mathbf{B}_A^{(i)}$, $\mathbf{A}_B^{(i)}$ or $\mathbf{C}_B^{(i)}$ matrices (Table 1), we had to modify the space defined by (N_t, k) to a non-rectangular domain. We provide more information on this adjustment in section 2.2.

As for the model input distributions, we used $\phi \sim \mathcal{DU}(1,8)$ to ensure an adequate representation of the most common shapes in the $(0,1)^k$ domain (Fig. 1a). If $\phi \in (1,7)$, we used the i-th distribution to describe all parameters. If $\phi = 8$, we randomly sampled all distributions and used the i-th to describe x_i.

[1] We refer to the set of benchmarking assumptions as "benchmarking parameters" or "parameters". This is intended to distinguish it from the inputs of each test function generated by the metafunction, which we refer to as inputs.

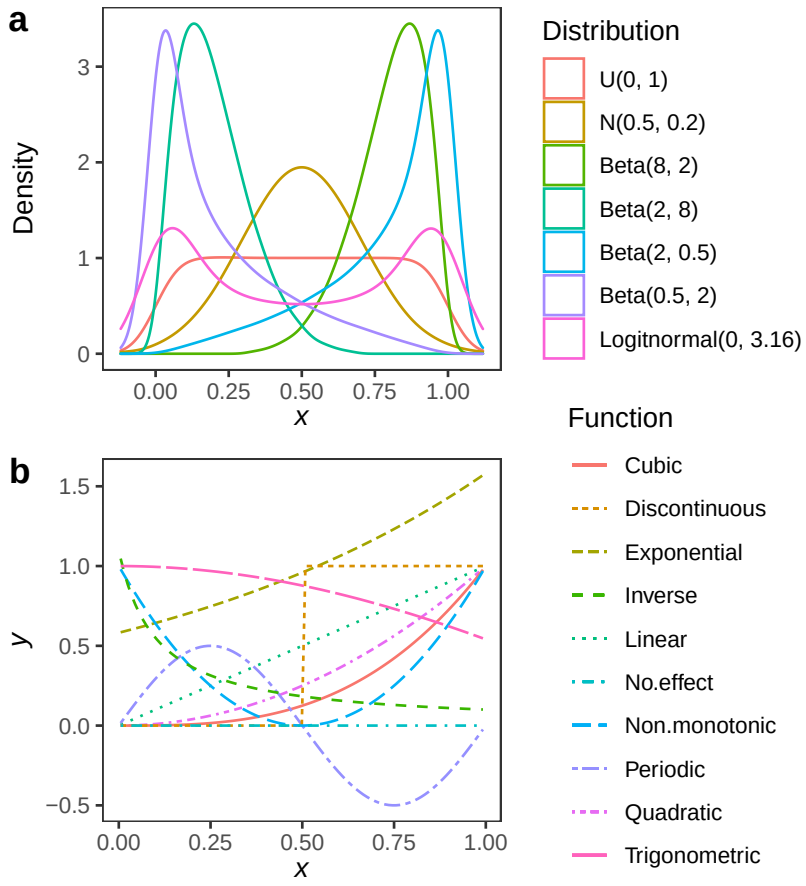

Figure 1: The metafunction approach. a) Probability distributions incuded in ϕ. b) Univariate functions included in the metafunction.

The parameter ε aimed at operationalizing the randomness in the form and execution of the test function, which we describe below. This test function is very similar to Becker [29]'s metafunction: it allows test functions to be generated by randomly combining p univariate functions in a multivariate function of dimension k (Fig. 1b). The inner functioning of our metafunction is the following:

1. *Step 1*: Let $\boldsymbol{M}$ be the (N, k) sample matrix of interest, formed by $i = x_1, x_2, ..., x_k$ model inputs distributed as $\phi_i(x_i)$. Each row and column is a sample point and input respectively.

2. *Step 2*: Let $\boldsymbol{u} = u_1, u_2, ..., u_k$ be a k-length vector formed by randomly sampling with replacement all p functions in Fig. 1b. The elements of this vector represent the first-order effects of each input variable. The i-th function in $\boldsymbol{u}$ is then applied to x_i,

yielding a new sample matrix $\boldsymbol{M}^*$ where the i-th column is the model output of the i-th function.

3. *Step 3*: Let $\boldsymbol{V}$ be a $(n, 2)$ matrix, for $n = \frac{k!}{2!(k-2)!}$, the number of pairwise combinations between the k inputs in $\boldsymbol{M}^*$. Each row in $\boldsymbol{V}$ thus specifies an interaction between two columns in $\boldsymbol{M}^*$. In order to follow the "sparsity of effects principle" (most variations in a given model output should be explained by low-order interactions [43]), the metafunction activates only a fraction of these effects: it randomly samples $k_2 n$ rows from $\boldsymbol{V}$, and computes the corresponding interactions in $\boldsymbol{M}^*$.

4. *Step 4*: Same as in 3, but for third order effects: let $\boldsymbol{W}$ be a $(m, 3)$ matrix, for $m = \frac{k!}{3!(k-3)!}$, the number of three-wise combinations between the k inputs in $\boldsymbol{M}^*$. The metafunction then randomly samples $k_3 n$ rows from $\boldsymbol{W}$, and computes the corresponding interactions in $\boldsymbol{M}^*$. Note that $k_2 > k_3$ because third-order effects tend to be less dominant than two-order effects (Table 2).

5. *Step 5*: Three vectors of coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ are defined to represent the weights of the first, second and third-order effects. These coefficients are generated by sampling from a mixture of two normal distributions $\Psi = 0.3\mathcal{N}(0, 5) + 0.7\mathcal{N}(0, 0.5)$. This coerces the metafunction into replicating the Pareto principle [44] (i.e. around 80% of the effects are due to 20% of the parameters), found to widely apply in sensitivity analysis [45].

The formula of the metafunction is therefore the following:

$$\begin{aligned} Y = & \sum_{i=1}^{k} \alpha_i f^{u_i} \phi_i(x_i) \\ & + \sum_{i=1}^{k_2} \beta_i f^{u_{V_{i,1}}} \phi_i(x_{V_{i,1}}) f^{u_{V_{i,2}}} \phi_i(x_{V_{i,2}}) \\ & + \sum_{i=1}^{k_3} \gamma_i f^{u_{W_{i,1}}} \phi_i(x_{W_{i,1}}) f^{u_{W_{i,2}}} \phi_i(x_{W_{i,2}}) f^{u_{W_{i,3}}} \phi_i(x_{W_{i,3}}) \end{aligned} \tag{5}$$

Note that the metafunction involves randomness in the sampling of the probability distributions in Fig. 1a (*Step 2*), the univariate functions in Fig. 1b (*Step 3*), the fraction of active pairwise and three-wise interactions (k_2, k_3) (*Steps 3–4*) and the coefficients (*Step 5*). The parameter ε assessed the influence of these sources of variability: it fixed the starting point of the pseudo-random number sequence used to sample the parameters just mentioned. We used $\varepsilon \sim \mathcal{U}(1, 200)$ to ensure that the same seed does not overlap with the same value of N_t or k, an issue that might introduce determinism in a process that should be stochastic. In Figs. S2–S3 we show an example of the behaviour of the metafunction and the type of Sobol' indices it generates.

Finally, we described the parameter δ as $\delta \sim \mathcal{DU}(1, 3)$. If $\delta = 1$, we computed the Pearson correlation between $\hat{\boldsymbol{T}}_i$ and $\boldsymbol{T}_i$. This setting assessed how well each estimator approximates

$\boldsymbol{T}_i$. If $\delta = 2$, we computed the Kendall τ-b correlation coefficient between $\hat{\boldsymbol{T}}_i$ and $\boldsymbol{T}_i$, defined as

$$\tau(\boldsymbol{T}_i, \hat{\boldsymbol{T}}_i) = \frac{n_c - n_d}{\sqrt{\left[\frac{n(n-1)}{2} - \sum_i \frac{t_i(t_i-1)}{2}\right)\left(\frac{n(n-1)}{2} - \sum_i \frac{u_j(u_j-1)}{2}\right]}} \tag{6}$$

where n_c and n_d are the number of concordant and discordant pairs, and t_i (u_j) are the number of tied values in the i-th (j-th) group of ties for the first (second) quantity. The use of Kendall τ-b evaluated how well the estimators in Table 1 rank the model inputs. If $\delta = 3$, we computed the Pearson correlation between $\boldsymbol{T}_i$ and $\hat{\boldsymbol{T}}_i$ after transforming the vectors to Savage scores [42], which are given as

$$S_{s_i} = \sum_{i=j}^{k} \frac{1}{j} \tag{7}$$

where j is the rank assigned to the jth element of a vector of length k. This setting examined the performance of the T_i estimators when the analyst is interested in properly ranking only the most important model inputs.

2.2 Sensitivity analysis

To examine how sensitive the performance of the estimators is to $(\tau, N_t, k, k_2, k_3, \phi, \epsilon, \delta)$, we conducted a global sensitivity analysis. We created an $\mathbf{A}$, $\mathbf{B}$ and $k-1$ $\mathbf{A}_B^{(i)}$ matrices, each of dimension $(2^{11}, k)$, using Sobol' quasi-random number sequences [22, 23]. We used $k-1$ $\mathbf{A}_B^{(i)}$ matrices because we grouped N_t and k and treat them like a single benchmark parameter given their correlation (see below). We allocated each element in $(\tau, N_t, k, k_2, k_3, \phi, \epsilon, \delta)$ to a column and described them using the probability distributions listed in Table 2.

Our model ran rowwise over the $\mathbf{A}$, $\mathbf{B}$ and $k-1$ $\mathbf{A}_B^{(i)}$ matrices, as follows: for $j = 1, 2, ..., 2^{11}$ rows, it created four (N_{t_j}, k_j) sub-matrices using the sampling method defined by τ_j. The need for four sub-matrices responds to the four specific sampling designs required to check the performance of all total-order estimators and obtain the vector of estimated indices $\hat{\boldsymbol{T}}_i$:

1. An $\mathbf{A}$, $\mathbf{B}$ and k_j $\mathbf{A}_B^{(i)}$ matrices, each of size (N_j, k_j), $N_j = \lceil\!\lceil \frac{N_{t_j}}{k_j+1} \rceil\!\rceil$ (for the Jansen, Sobol', Glen and Isaacs, Janon/Monod and Homma and Saltelli estimators).

2. An $\mathbf{A}$, $\mathbf{B}$ and k_j $\mathbf{A}_B^{(i)}$ and $\mathbf{B}_A^{(i)}$ matrices, each of size (N_j, k_j), $N_j = \lceil\!\lceil \frac{N_{t_j}}{2k_j+2} \rceil\!\rceil$ (for the Azzini and Rosati estimator).

3. An $\mathbf{A}$, $\mathbf{B}$ and k_j $\mathbf{B}_A^{(i)}$ and $\mathbf{C}_B^{(i)}$ matrices, each of size (N_j, k_j), $N_j = \lceil\!\lceil \frac{N_{t_j}}{2k_j+2} \rceil\!\rceil$ (for the pseudo-Owen estimator).

4. A matrix formed by N_j stars, each of size $k_j((1/\Delta h) - 1) + 1$. Given that we set Δh at 0.2 (see Supplementary Materials), $N_j = \lfloor\!\lfloor \frac{N_{t_j}}{4k+1} \rfloor\!\rfloor$ (for the Razavi and Gupta estimator).

The different sampling design structures and the value for k_j constrains the total number of runs N_t that can be effectively allocated to each estimator. Furthermore, given the probability distributions selected for N_t and k (Table 1), specific combinations of (N_{t_j}, k_j) lead to $N_j \leq 1$, which is computationally unfeasible. To minimize these issues, we forced the comparison between estimators to approximate the same N_{t_j} value, regardless of their sampling design specificities: since the sampling design structure of Razavi and Gupta is the most constraining, we used $N_j = \frac{2(4k+1)}{k+1}$ (for estimators 1–5), and $N_j = \frac{2(4k+1)}{2k+2}$ (for estimators 6–7) when $N_j \leq 1$ in the case of Razavi and Gupta. This ensured a robust comparison between the estimators' performance as it compelled them to explore a very similar portion of the (N_t, k) space. However, the trade-off was that N_t and k became correlated, an issue that contradicts the requirement of independent inputs characterizing variance-based sensitivity indices [1]. This is why we decided to treat (N_t, k) as a single benchmark parameter in the sensitivity analysis.

We also created a fifth sub-matrix, this one formed by an $\mathbf{A}$ and $k^{(j)}$ $\mathbf{A}_B^{(i)}$ matrices, each of size $(2^{11}, k^{(j)})$. We used this sub-matrix to compute the vector of true indices $\boldsymbol{T}_i$, which could not be calculated analytically due to the wide range of possible functional forms created by the metafunction. Following Becker [29], we assumed that a fairly accurate approximation to $\boldsymbol{T}_i$ could be achieved with a large Monte Carlo estimation, which we attempted via the Jansen [15] estimator following current established best practices in sensitivity analysis.

We ran the metafunction in these five sub-matrices at once, following the conditions set by $\phi_j, \varepsilon_j, k_{2_j}, k_{3_j}$, and computed the vectors $\hat{\boldsymbol{T}}_{i_j}$ and $\boldsymbol{T}_{i_j}$ for each estimator. Finally, we assessed how well each $\boldsymbol{T}_{i_j}$ approached $\hat{\boldsymbol{T}}_{i_j}$ using the performance measure defined by δ_j. The final model output –once all these uncertain factors were accounted for– is the correlation between $\boldsymbol{T}_{i_j}$ and $\hat{\boldsymbol{T}}_{i_j}$, which we refer to as r.

We ran all simulations in the *R* environment [46]. The code to replicate our results is fully available in Zenodo [47].

3 Results

3.1 Uncertainty analysis

Fig. 2 presents the results of the simulations, with higher r values indicating a better performance, i.e. a better correlation between $\boldsymbol{T}_i$ and $\hat{\boldsymbol{T}}_i$. Five groups of estimators can be identified based on the shape of the empirical distribution for r and its median value, with higher r corresponding to better performance (Fig. 2a):

1. A group formed by pseudo-Owen only, characterized by a positively skewed distribution for $r > 0$, an evident left tail for $r < 0$ and median r values of ≈ 0.2.

2. A group formed by Homma and Saltelli, Sobol', and Glen and Isaacs, showing a positively skewed distribution and a thin tail for $r < 0$, as well as median r values of ≈ 0.3.

3. A group formed by Azzini and Rosati only, which display a negatively skewed distribution and median r values of ≈ 0.6.

4. A group with Janon/Monod and Jansen, characterized by a highly negatively skewed distribution and median r values of ≈ 0.8.

5. A group formed by Razavi and Gupta only, characterized by a highly negatively skewed distribution and median r values of ≈ 0.9.

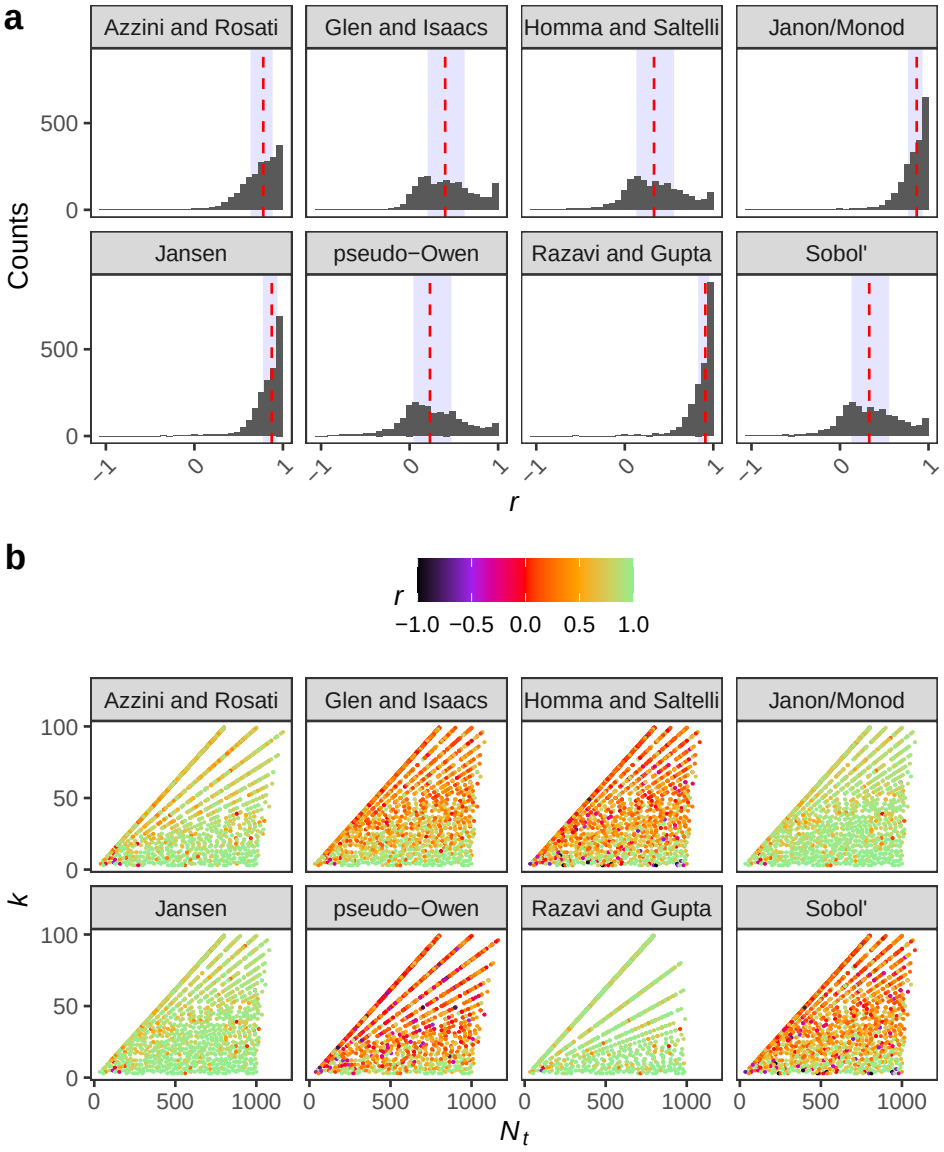

Figure 2: Results of the simulations. a) Histograms of the model output r. The vertical, red dashed line shows the median value, whereas the blue, semi-transparent rectangle highlights the interquartile range. b) Scatterplots of the total number of model runs N_t against the function dimensionality k. The greener (purpler) the color, the better (worse) the performance of the estimator.

To obtain a finer insight into the structure of these results, we plotted the total number of model runs N_t against the function dimensionality k (Fig. 2b). This allowed to map the

performance of the estimators in the input space formed by all possible combinations of N_t and k given the specific design constraints of each formulae. The results suggest that all estimators perform reasonably well at a very small dimensionality ($k \leq 10, r > 0.7$), regardless of the total number of model runs available (Fig. S4). However, some differences unfold at higher dimensions: Sobol', Homma and Saltelli, Glen and Isaacs and especially pseudo-Owen swiftly become inaccurate for $k > 10$, even with large values for N_t. Azzini and Rosati display a good performance overall except in the upper N_t, k boundary, where most of the orange dots concentrate. As for Jansen, Janon/Monod and Razavi and Gupta, they present an almost flawless performance regardless of the region explored in the N_t, k domain.

The presence of a non-negligible proportion of model runs with $r < 0$ (Figs. 2b–S5) suggests that some estimators significantly overturned the true ranks during the computation of $\hat{\boldsymbol{T}}_i$. To better examine this phenomenon, we re-plotted Fig. 2b with just the simulations yielding $r < 0$ (Fig. S6). We observed that $r < 0$ values not only appeared in the region of small N_t, a foreseeable miscalculation derived from allocating an insufficient number of model runs to each model input: they also emerged at a relatively large N_t and low k in the case of pseudo-Owen, Sobol' and Homma and Saltelli. This suggests that rank reversing and/or $\hat{\boldsymbol{T}}_{\boldsymbol{i}}$ inaccuracy is not an artifact of our study design as much as a by-product of the volatility of these estimators when stressed by several sources of computational uncertainty, such as those listed in Table 2.

This is further highlighted in Fig. 3, which shows the Sobol' indices produced in one of the simulations that led Sobol', Homma and Saltelli or pseudo-Owen to comparatively underperform, despite being conducted on a large N_t and small k. In this specific case, the small r values yielded by these estimators are due to their production of negative total-order indices, i.e. $T_i < 0$. This might be partially related to the way they compute T_i, as their formulae allow the production of negative terms in the summation at the numerator and are therefore more vulnerable to error (Table 1).

In order to better examine the efficiency of the estimators, we summarized their performance as a function of the number of runs available per model input N_t/k [29] (Fig. 4–S8). This information is especially relevant to take an educated decision on which estimator to use in a context of a high-dimension, computationally expensive model. Even when the budget of runs per input is low [$(N_t/k) \in [2, 20]$], Razavi and Gupta manages to achieve an almost perfect performance ($r \sim 0.9$). Jansen and Janon/Monod come very close ($r \sim 0.85$), followed by Azzini and Rosati ($r \sim 0.75$), Sobol', Glen and Isaacs, and Homma and Saltelli ($r \sim 0.3$), and finally pseudo-Owen, which ranks last ($r \sim 0.2$). When the N_t/k ratio is increased, all estimators improve their accuracy and some quickly reach the asymptote: this is the case of Razavi and Gupta, Janon/Monod and Jansen, whose performance becomes excellent from $(N_t/k) \in [20, 40]$ onwards. The performance of Azzini and Rosati reaches its optimum at $(N_t/k) \in [40, 60]$, whereas that of Glen and Isaacs at $(N_t/k) \in [80, 100]$. The accuracy of the other estimators fluctuates and does not seem to stabilize within the range of ratios examined. This is especially the case of Homma and Saltelli and Sobol', whose performance oscillates before plummeting at $(N_t/k) \in [280, 300]$ due to several simulations yielding large $r < 0$ values.

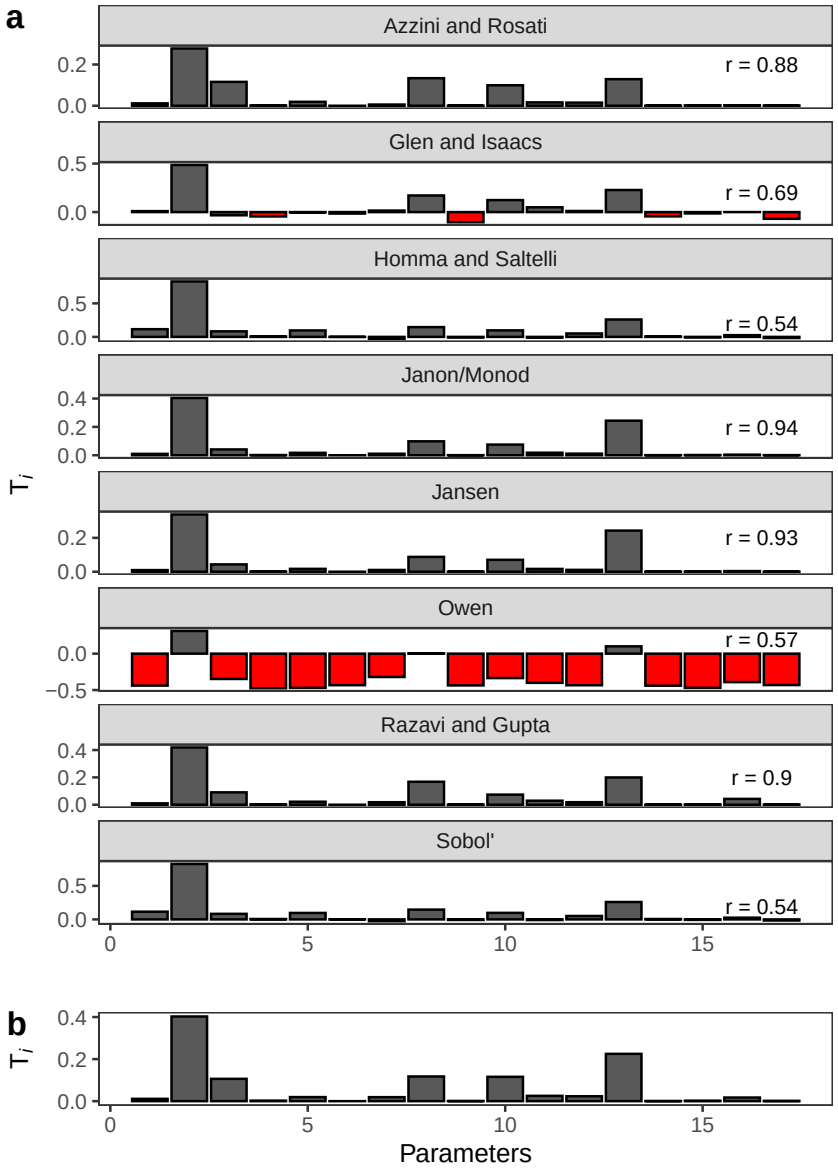

Figure 3: Total-order indices (T_i) obtained in the simulation 782, with the parameters of the metafunction set at $\tau = 1$, $k = 17$, $k_2 = 0.41$, $k_3 = 0.16$, $\varepsilon = 158$, $\phi = 6$, $\delta = 2$. The red bars highlight parameters with $T_i < 0$. a) Estimated T_i indices ($\hat{\boldsymbol{T}}_i$), with $N_t =\sim 691$. b) True T_i indices ($\boldsymbol{T}_i$), computed with the Jansen [15] estimator and $N_t = 2^{11}(17 + 1) = 36,864$. The settings of the other uncertain parameters are the same as in a).

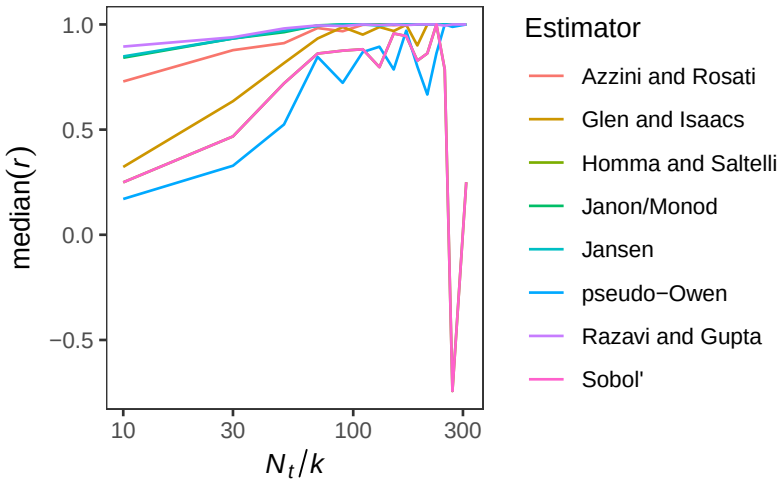

Figure 4: Scatterplot of the model output r against the number of model runs allocated per model input (Nt/k). Note that the line for the Homma & Saltelli [9] estimator does not show because it mimics the trend of Sobol' [16]. See Fig. S7 for a visual display of all simulations and Fig. S8 for an assessment of the number of model runs that each estimator has in each N_t/k compartment.

3.2 Sensitivity analysis

Fig. 5a presents the first (S_i) and total (T_i) - order Sobol' indices of the parameters that condition the performance of the estimators (Table 2). In the Supplementary Materials we also include scatterplots, which allow to map r values back into the uncertain space of each model input (Fig. S9–S16), as well as the unnormalized Sobol' indices, which inform on the overall uncertainty conveyed by each model input (Fig. S17).

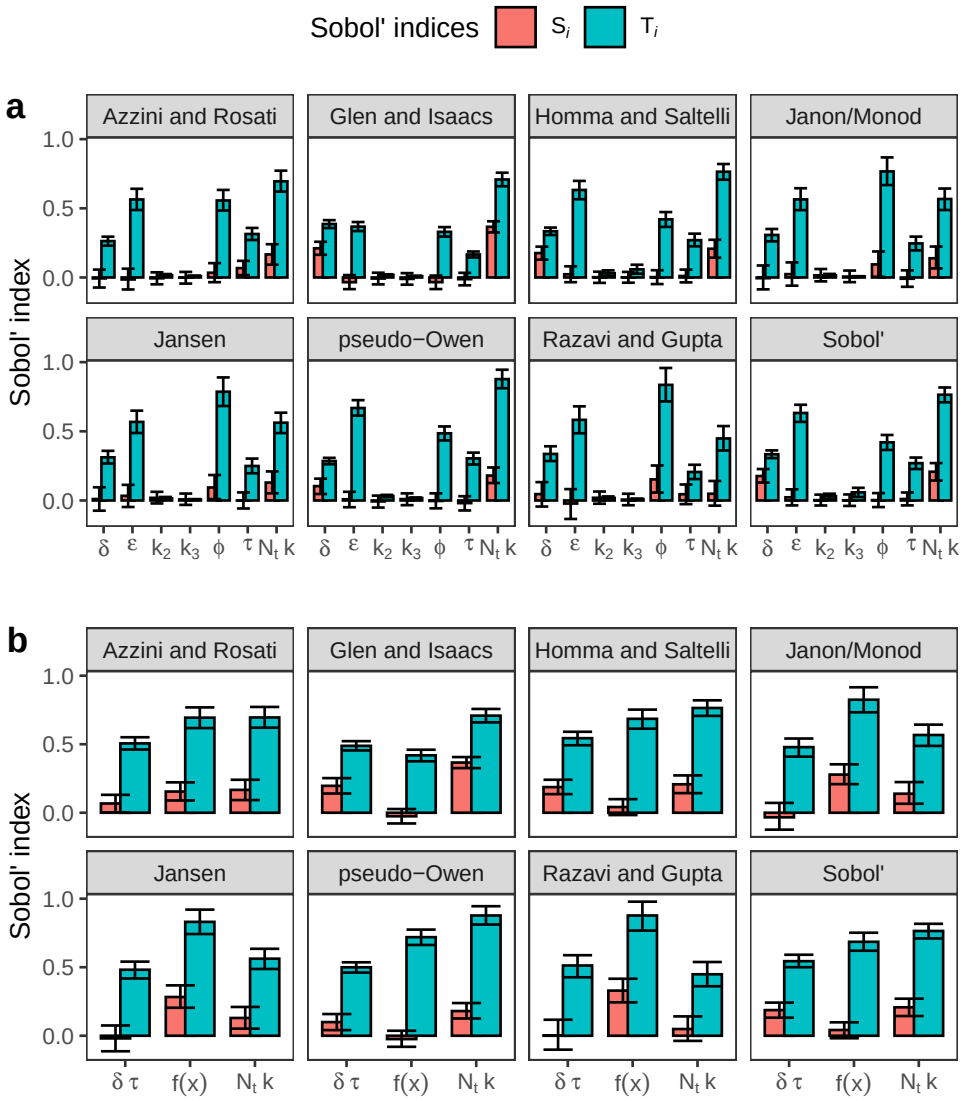

Figure 5: Sobol' indices. a) Parameters. Note that the effect of N_t and k was assessed simultaneously due to their correlation, thus (N_t, k) is treated as a group. b) Clusters. $f(x)$ groups all parameters that describe the model uncertainty $(\varepsilon, k_2, k_3, \phi)$.

The results indicate that the proportion of active second and third-order interactions (k_2, k_3) do not alter the performance of any estimator. The other uncertain benchmark parameters have a non-zero effect and are important especially through interactions, although some also have a first order effect. This is the case of the total number of model

runs and function dimensionality (N_t, k), the underlying model input distributions (ϕ) or the performance measure selected (δ), whose influence varies depending on the estimator: while Glen and Isaacs' performance is significantly sensitive to (N_t, k) ($\sim$ 35%), as well as Sobol's, Homma and Saltelli's and pseudo-Owen's ($\sim$ 20%), Razavi and Gupta's appears to be the less affected by it ($\sim$7%). Razavi and Gupta is actually most conditioned by ϕ ($\sim$15%), which also influences the accuracy of Jansen and Janon/Monod ($\sim$10%). Finally, δ seems to have a considerable first-order effect in the performance of Sobol', Homma and Saltelli, and Glen and Isaacs ($\sim$18%).

To better understand the structure of these sensitivities, we computed Sobol' indices after grouping these parameters in three clusters, which we defined based on their commonalities: the first group included (δ, τ) and reflected the influence of those parameters that can be defined by the sensitivity analyst during the setting of the benchmark exercise. The second combined $(\varepsilon, k_2, k_3, \phi)$ and was defined to examine the overall impact of the model functional form, referred to as $f(x)$, which is often beyond the analyst's grasp. Finally, the third group included (N_t, k) only and was used to assess the influence of the sampling design in the accuracy of the estimators (we assumed that the total number of model runs, besides being conditioned by the computing resources at hand, is also partially determined by the joint effect of the model dimensionality and the use of either a $\mathbf{B}$, $\mathbf{A}_B^{(i)}$), $\mathbf{B}_A^{(i)}$ or $\mathbf{C}_B^{(i)}$ matrices).

The results are presented in Fig. 5b. We observed that the accuracy of Razavi and Gupta, Janon/Monod and Jansen is mostly influenced by $f(x)$ ($\sim$25%), which does not seem to have any significant effect in the performance of Glen and Isaacs, pseudo-Owen, Homma and Saltelli or Sobol'. These are largely influenced by the first-order effect of (N_t, k) and (δ, τ) (10%–20%). Interestingly, the sum of all first-order effects across clusters is much less than one ($\sum_{i=1}^{k} S_i \ll 0.6$) in all estimators, indicating that their performance is significantly influenced by the interactions between at least two clusters simultaneously.

We also detected that some estimators consistently underperformed under some particular probability distributions and performance measures. For instance, Glen and Isaacs, Homma and Saltelli, pseudo-Owen and Sobol' showed a lower accuracy under $\delta = 2$, i.e. when the aim is to rank all model parameters (Figs. S10, S11, S14, S16). In contrast, Jansen, Janon/Monod and Razavi and Gupta underperformed under $x_i \sim \mathcal{U}(0.5, 2)$ and $\delta = 1$, which measured the accuracy of estimators when the measure of interest is the direct correlation between the estimated and true indices $\hat{\boldsymbol{T}}_i$ and $\boldsymbol{T}_i$ (Figs. S12, S13, S15).

Since ϕ and δ were significantly involved in interactions, we computed second-order Sobol' indices to examine which pairs of inputs most affected the estimators' performance. We observed significant interactions between ϕ, δ and ϕ, ϵ, which respectively conveyed as much as 15% and 20% of the uncertainty in the performance of Jansen, Janon/Monod and Razavi and Gupta (Fig. S18).

4 Discussion and conclusions

In this article, we have designed an eight-dimension battleground for total-order estimators to confront and prove their value in an unparalleled range of sensitivity analysis scenarios. By randomising the parameters that condition their performance, we have obtained a comprehensive picture of the advantages and disadvantages of each estimator, and identified

which particular benchmark factors make them more prone to error. Our work thus provides a thorough assessment of state-of-the-art total-order estimators and contributes to define best practices in variance-based sensitivity analysis. The study also aligns with previous works focused on testing the robustness of the tools available to sensitivity analysts, a line of inquiry that can be described as a *sensitivity analysis of sensitivity analysis* (SA of SA) [48].

Our results provide empirical support to the assumption that previous benchmark studies are biased by the plethora of non-unique choices taken during the setting of the analysis [29]. We have observed that almost all decisions have a non-zero effect: from the selection of the sampling method to the choice of the performance measure, the design prioritized by the analyst can potentially bias the results in a non-obvious way, i.e. through interactions. The importance of non-additivities in conditioning performance suggests that the benchmark of sensitivity estimators should no longer rely on statistical designs that change one parameter at a time (usually the number of model runs and, more rarely, the test function; Janon *et al.* [12], Azzini & Rosati [13], Saltelli *et al.* [14], Owen [19], Razavi & Gupta [20, 26], and Lo Piano *et al.* [28]). This setting reduces the uncertain space to a minimum and misses the effects of interactions. If global sensitivity analysis is the recommended practice to fully explore the uncertainty space of models, sensitivity estimators, being algorithms themselves, should be validated alike [48].

Under the wide range of situations created in this study, Razavi and Gupta, Jansen and Janon/Monod have shown an excellent accuracy overall. In terms of pure efficiency, the former appears to be the measure of choice if there are serious constraints on the number of model runs that can be allocated to each model input (i.e. 2–10). To understand why Razavi and Gupta shows a higher accuracy at very low sample sizes, we might need to briefly review the concepts of *economy* and *exploravity*, recently formalized by Lo Piano *et al.* [49] in the case of total-order indices.

Economy refers to the number of elementary effects used to compute T_i given the number of runs, whereas *exploravity* describes the capacity of the estimator in exploring the input space. Both properties can not be maximized at the same time: a higher exploravity leads to a lower economy, and viceversa. In order to display how these features characterize the estimators analysed in this paper, we computed their economy and exploravity following Lo Piano *et al.* [49] (Fig. 6, Table S1). Razavi and Gupta is the estimator with the highest economy because any function value difference in the formula for the estimator is used twice, once in the variogram term and once in the co-variogram term (Table 1). This is particularly effective and compensates for the loss of exploravity derived from its reliance on "enriched" star designs (Supplementary Materials).

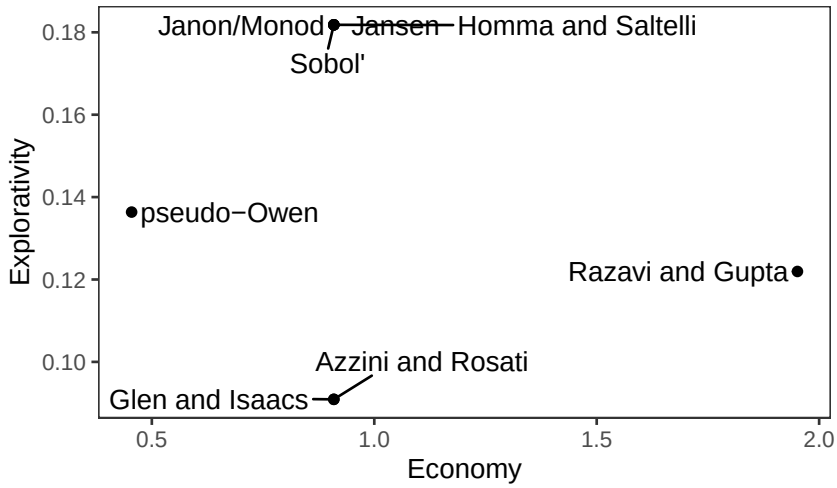

Figure 6: Exploravity *versus* economy plot for $k = 10$. See Table **??** for the formulae used and Fig. S19 for an exploravity and economy plot using other values for k.

The lower exploravity of Razavi and Gupta compared to Jansen or Janon/Monod should not be a critical disadvantage if the model under study is dominated by main effects, as is the case with many physical systems (and with the metafunction used in this paper): Razavi and Gupta's design endows the estimator with a good capacity to appraise nonlinearites in main effects around a few nominal points. However, it remains to be seen how Razavi and Gupta performs when the weight of non-additivities in a given model is much higher than the values explored in this study.

It is also worth highlighting that the use of Razavi and Gupta requires the analyst to define an extra tuning parameter, Δh. In this paper, we have set $\Delta h = 0.2$ after some preliminary trials with the estimator; other works have used different values (e.g. $\Delta h = 0.002$, $\Delta h = 0.1$, $\Delta h = 0.3$; Razavi & Gupta [20, 26] and Becker [29]). Finding the most appropriate value for a given tuning parameter may not be an obvious choice, as shown by Puy *et al.* [48] in the case of the PAWN sensitivity index [50, 51]. To our knowledge, no work has assessed how robust Razavi and Gupta's estimator is when the number of stars N and the tuning parameter Δh are moved simultaneously within reasonable bounds.

The robustness of Razavi and Gupta, Jansen and Janon/Monod makes their sensitivity to the uncertain parameters studied here almost negligible. Most of their performance is conditioned by the first and total-order effects of the model form jointly with the underlying probability distributions ($f(x)$ in Fig. 5b), which are in any case beyond the analyst's control. This suggests that they are already highly optimized estimators with not much room for improvement. As for the rest, their accuracy might be enhanced by allocating a larger number of model runs per input (if computationally affordable), and especially in the case of Homma and Saltelli, Sobol' and Glen and Isaacs, by restricting their use to low-dimensional models ($k < 10$) and sensitivity settings that only require ranking the most important parameters (a "restricted" *factor prioritisation* setting; Saltelli *et al.* [1]). Nevertheless,

their substantial volatility is considerably driven by non-additivities, a combination that makes them hard to tame and should raise caution about their use in any modeling exercise.

Our results slightly differ from Becker [29]'s, who observed that Jansen slightly outperformed Janon/Monod. We did not find any difference between Jansen and Janon/Monod, which are only marginally surpassed by Razavi and Gupta in efficiency at very small N_t/k ratios. Although our metafunction approach is based on Becker [29]'s, our study tests the accuracy of estimators in a larger uncertain space as we also account for the stress introduced by changes in the sampling method τ, the underlying probability distributions ϕ or the performance measure selected δ. These differences account for the slightly different results obtained between the two papers.

Our study can be extended to other sensitivity estimators (i.e. moment-independent like entropy-based, Liu *et al.* [52]; the δ-measure, Borgonovo [53]; or the PAWN index, Pianosi & Wagener [50, 51]). Moreover, it holds potential to be used overall as a standard crash test every time a new sensitivity estimator is introduced to the modeling community. One of its advantages is its flexibility: Becker [29]'s metafunction can be easily extended with new univariate functions or probability distributions, and the settings modified to check performance under different degrees of non-additivities or in a larger (N_t, k) space. With some slight modifications it should also allow to produce functions with dominant low-order or high-order terms, labeled as Type B and C by Kucherenko *et al.* [30]. This should prompt developers of sensitivity indices to severely stress their estimators so the modeling community and decision-makers fully appraise how they deal with uncertainties.

5 Acknowledgements

This work has been funded by the European Commission (Marie Skłodowska-Curie Global Fellowship, grant number 792178 to A.P.).

References

1. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. *Global Sensitivity Analysis. The Primer* doi:10.1002/9780470725184 (John Wiley & Sons, Ltd, Chichester, UK, Dec. 2008).
2. Jakeman, A., Letcher, R. & Norton, J. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software* **21,** 602–614 (May 2006).
3. Eker, S., Rovenskaya, E., Obersteiner, M. & Langan, S. Practice and perspectives in the validation of resource management models. *Nature Communications* **9,** 1–10 (2018).
4. Borgonovo, E. & Plischke, E. Sensitivity analysis: A review of recent advances. *European Journal of Operational Research* **248,** 869–887 (2016).
5. Saltelli, A. Sensitivity analysis for importance assessment. *Risk Analysis* **22,** 579–590 (June 2002).

6. Iooss, B. & Lematre, P. *Uncertainty Management in Simulation-Optimization of Complex Systems. Operations Research/Computer Science Interfaces Series, vol 59* (eds Dellino, G. & Meloni, C.) 101–122 (Springer, Boston, 2015). doi:10.1007/978-1-4899-7547-8_5. arXiv: 1404.2405.

7. Becker, W. & Saltelli, A. *Handbook of Design of Experiments* (eds Dean, A., Morris, M., Stufken, J. & Bingham, D.) 627–674 (CRC Press, Taylor & Francis, Boca Ratn, 2015). doi:0.1201/b18619.

8. Archer, G. E., Saltelli, A. & Sobol, I. M. Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation* **58,** 99–120 (1997).

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

10. Le Gratiet, L., Marelli, S. & Sudret, B. *Handbook of Uncertainty Quantification* 1289–1325 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-12385-1_38.

11. Saltelli, A., Tarantola, S. & Chan, K. P.-S. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* **41,** 39 (Feb. 1999).

12. Janon, A., Klein, T., Lagnoux, A., Nodet, M. & Prieur, C. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM - Probability and Statistics* **18,** 342–364 (2014).

13. Azzini, I. & Rosati, R. *The IA-Estimator for Sobol' sensitivity indices Ninth International Conference on Sensitivity Analysis of Model Output* (Barcelona, 2019).

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

15. Jansen, M. Analysis of variance designs for model output. *Computer Physics Communications* **117,** 35–43 (Mar. 1999).

16. Sobol', I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* **55,** 271–280 (Feb. 2001).

17. Monod, H., Naud, C. & Makowski, D. *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications* (eds Wallach, D., Makowksi, D. & Jones, J.) 55–99 (Elsevier, 2006). doi:10.1016/j.ress.2007.06.003.

18. Glen, G. & Isaacs, K. Estimating Sobol sensitivity indices using correlations. *Environmental Modelling and Software* **37,** 157–166 (2012).

19. Owen, A. B. Better estimation of small sobol' sensitivity indices. *ACM Transactions on Modeling and Computer Simulation* **23,** 1–17 (2013).

20. Razavi, S. & Gupta, H. V. A new framework for comprehensive, robust and efficient global sensitivity analysis: 2. Application. *Water Resources Research* **52,** 440–455 (Sept. 2016).

21. McKay, M. D., Beckman, R. J. & Conover, W. J. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21,** 239–245 (1979).

22. Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* **7,** 86–112 (Jan. 1967).

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

24. Iooss, B., Janon, A., Pujol, G., with contributions from Baptiste Broto, Boumhaout, K., Veiga, S. D., Delage, T., Amri, R. E., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, P., Marrel, A., Meynaoui, A., Nelson, B. L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T. & Weber, F. *sensitivity: Global Sensitivity Analysis of Model Outputs* 2020.

25. Janon, A., Klein, T., Lagnoux, A., Nodet, M. & Prieur, C. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics* **18,** 342–364 (2014).

26. Razavi, S. & Gupta, H. V. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. *Water Resources Research* **52,** 440–455 (Jan. 2016).

27. Lamboni, M. Global sensitivity analysis: a generalized, unbiased and optimal estimator of total-effect variance. *Statistical Papers* **59,** 361–386 (2018).

28. Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D. & Saltelli, A. Variance-based sensitivity analysis: The quest for better estimators between explorativity and efficiency. *Reliability Engineering & System Safety* (2020).

29. Becker, W. Metafunctions for benchmarking in sensitivity analysis. *Reliability Engineering & System Safety,* 107189 (Aug. 2020).

30. Kucherenko, S., Feil, B., Shah, N. & Mauntz, W. The identification of model effective dimensions using global sensitivity analysis. *Reliability Engineering & System Safety* **96,** 440–449 (Apr. 2011).

31. Ishigami, T. & Homma, T. An importance quantification technique in uncertainty analysis for computer models. *Proceedings. First International Symposium on Uncertainty Modeling and Analysis* **12,** 398–403 (1990).

32. Sobol', I. M. On quasi-Monte Carlo integrations. *Mathematics and Computers in Simulation* **47,** 103–112 (Aug. 1998).

33. Bratley, P. & Fox, B. L. ALGORITHM 659: implementing Sobol's quasirandom sequence generator. *ACM Transactions on Mathematical Software (TOMS)* **14,** 88–100 (1988).

34. Saltelli, A. A short comment on statistical versus mathematical modelling. *Nature Communications* **10,** 8–10 (2019).

35. Sheikholeslami, R., Razavi, S., Gupta, H. V., Becker, W. & Haghnegahdar, A. Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. *Environmental Modelling and Software* **111,** 282–299 (2019).

36. Sarrazin, F., Pianosi, F. & Wagener, T. Global Sensitivity Analysis of environmental models: Convergence and validation. *Environmental Modelling and Software* **79,** 135–152 (2016).

37. Haghnegahdar, A., Razavi, S., Yassin, F. & Wheater, H. Multicriteria sensitivity analysis as a diagnostic tool for understanding model behaviour and characterizing model uncertainty. *Hydrological Processes* **31,** 4462–4476 (2017).

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

39. Paleari, L. & Confalonieri, R. Sensitivity analysis of a sensitivity analysis: We are likely overlooking the impact of distributional assumptions. *Ecological Modelling* **340,** 57–63 (2016).

40. Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* **15,** 72 (Jan. 1904).

41. Kendall, M. G. & Smith, B. B. The problem of m rankings. *The Annals of Mathematical Statistics* **10,** 275–287 (Sept. 1939).

42. Savage, I. R. Contributions to the theory of rank order statistics - the two sample case. *Annals of Mathematical StatisticsStatistics* **27,** 590–615 (1956).

43. Box, G., Hunter, J. & Hunter, W. *Statistics for Experimenters: Design, Innovation, and Discovery* (Wiley, 2005).

44. Pareto, V. *Manuale di Economia Politica. Vol. 13* (Societa Editrice, 1906).

45. Box, G. E. & Meyer, R. D. An analysis for unreplicated fractional factorials. *Technometrics* **28,** 11–18 (1986).

46. R Core Team. *R: A language and environment for statistical computing* Vienna, 2016.

47. Puy, A. *R code of the paper "The Battle of Total-Order Sensitivity Estimators"* 2020. doi:10.5281/ZENODO.3788112.

48. Puy, A., Piano, S. L. & Saltelli, A. A sensitivity analysis of the PAWN sensitivity index. *Environmental Modelling and Software* **127,** 104679 (2020).

49. Lo Piano, S., Puy, A., Borgonovo, E. & Saltelli, A. *Driving trends towards the European Structural Investment Funds payment patters*

50. Pianosi, F. & Wagener, T. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environmental Modelling and Software* **67,** 1–11 (2015).

51. Pianosi, F. & Wagener, T. Distribution-based sensitivity analysis from a generic input-output sample. *Environmental Modelling and Software* **108,** 197–207 (2018).

52. Liu, H., Chen, W. & Sudjianto, A. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design, Transactions of the ASME* **128,** 326–336 (2006).

53. Borgonovo, E. A new uncertainty importance measure. *Reliability Engineering and System Safety* **92,** 771–784 (2007).

The battle of total-order sensitivity estimators

Supplementary Materials

Arnald Puy*[1,2], William Becker[3], Samuele Lo Piano[4], and Andrea Saltelli[2,3]

[1] *Department of Ecology and Evolutionary Biology, M31 Guyot Hall, Princeton University, New Jersey 08544, USA. E-Mail: apuy@princeton.edu*
[2] *Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Parkveien 9, PB 7805, 5020 Bergen, Norway.*
[3] *European Commission, Joint Research Centre, Via Enrico Fermi, 2749, 21027 Ispra VA, Italy*
[4] *University of Reading, School of the Built Environment, JJ Thompson Building, Whiteknights Campus, Reading, RG6 6AF, United Kingdom*

Contents

*Corresponding author

1 Razavi and Gupta's estimator (VARS)

Unlike the other total-order estimators examined in our paper, Razavi and Gupta's VARS (for Variogram Analysis of Response Surfaces [1, 2]) relies on the variogram $\gamma(.)$ and covariogram $C(.)$ functions to compute what they call the VARS-TO, for VARS Total-Order index.

Let us consider a function of factors $\boldsymbol{x} = (x_1, x_2, ..., x_k) \in \mathbb{R}^k$. If $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are two generic points separated by a distance $\boldsymbol{h}$, then the variogram is calculated as

$$\gamma(\boldsymbol{x}_A - \boldsymbol{x}_B) = \frac{1}{2} V\left[y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B)\right] \tag{1}$$

and the covariogram as

$$C(\boldsymbol{x}_A - \boldsymbol{x}_B) = COV\left[y(\boldsymbol{x}_A), y(\boldsymbol{x}_B)\right] \tag{2}$$

Note that

$$V\left[y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B)\right] = V\left[y(\boldsymbol{x}_A)\right] + V\left[y(\boldsymbol{x}_B)\right] - 2COV\left[y(\boldsymbol{x}_A), y(\boldsymbol{x}_B)\right] \tag{3}$$

and since $V\left[y(\boldsymbol{x}_A)\right] = V\left[y(\boldsymbol{x}_B)\right]$, then

$$\gamma(\boldsymbol{x}_A - \boldsymbol{x}_B) = V\left[y(\boldsymbol{x})\right] - C(\boldsymbol{x}_A, \boldsymbol{x}_B) \tag{4}$$

In order to obtain the total-order effect T_i, the variogram and covariogram are computed on all couples of points spaced h_i along the x_i axis, with all other factors being kept fixed. Thus equation 4 becomes

$$\gamma_{x^*_{\sim} i}(h_i) = V(y|x^*_{\sim} i) - C_{x^*_{\sim} i}(h_i) \tag{5}$$

where $x^*_{\sim} i$ is a fixed point in the space of non-x_i. Razavi and Gupta [1, 2] suggest to take the mean value across the factors' space on both sides of equation 5, thus obtaining

$$E_{x^*_{\sim} i}\left[\gamma_{x^*_{\sim} i}(h_i)\right] = E_{x^*_{\sim} i}\left[V(y|x^*_{\sim i})\right] - E_{x^*_{\sim} i}\left[C_{x^*_{\sim} i}(h_i)\right] \tag{6}$$

which can also be written as

$$E_{x^*_{\sim} i}\left[\gamma_{x^*_{\sim} i}(h_i)\right] = V(y)T_i - E_{x^*_{\sim} i}\left[C_{x^*_{\sim} i}(h_i)\right] \tag{7}$$

and therefore

$$T_i = \frac{E_{x^*_{\sim} i}\left[\gamma_{x^*_{\sim} i}(h_i)\right] + E_{x^*_{\sim} i}\left[C_{x^*_{\sim} i}(h_i)\right]}{V(y)} \tag{8}$$

The sampling scheme for VARS does not rely on $\mathbf{A}, \mathbf{B}, \mathbf{A}_B^{(i)}$... matrices, but on star centers and cross sections. Star centers are N random points sampled across the input space. For each of these stars, k cross sections of points spaced Δh apart are generated, including and passing through the star center. Overall, the computational cost of VARS amounts to $N_t = N\left[k((1/\Delta h) - 1) + 1\right]$.

2 Figures

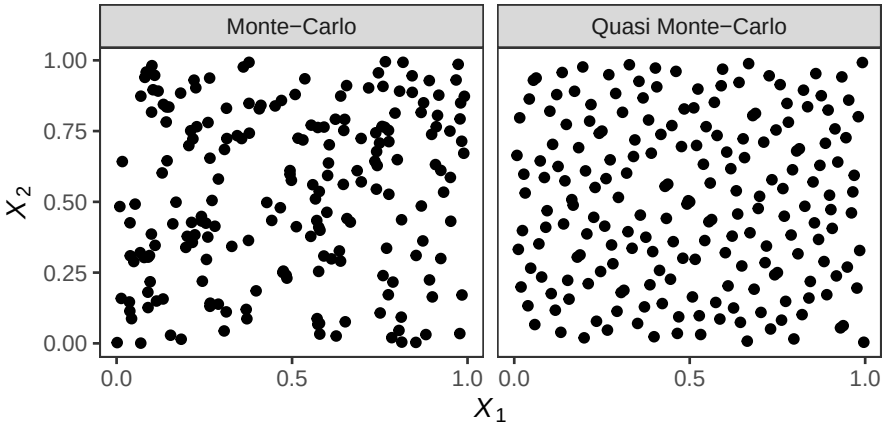

Figure S1: Examples of Monte-Carlo and Quasi Monte-Carlo sampling in two dimensions. $N = 200$.

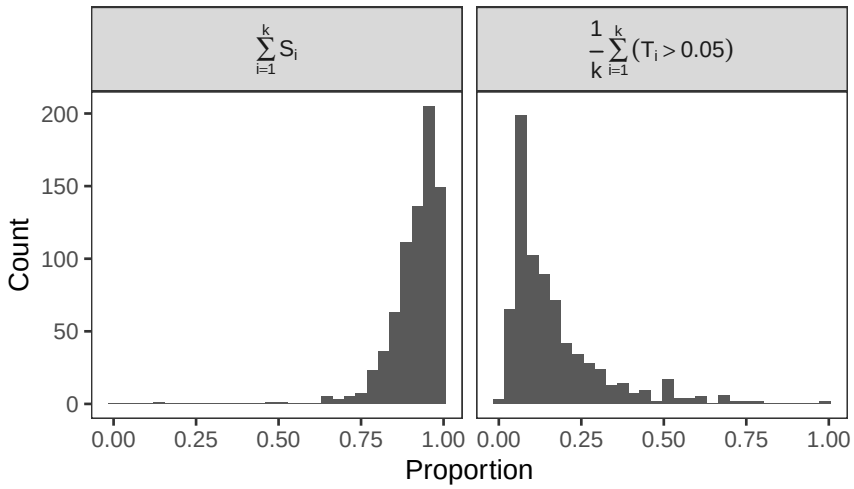

Figure S2: Proportion of the total sum of first-order effects and of the active model inputs (defined as $T_i > 0.05$) after 1000 random metafunction calls with $k \in (3, 100)$. Note how the sum of first-order effects clusters around 0.8 (thus evidencing the production of non-additivities) and how, on average, the number of active model inputs revolves around 10–20%, thus reproducing the Pareto principle.

Figure S3: Sobol' T_i indices obtained after a run of the metafunction with the following parameter settings: $N = 10^3$, $k = 21$, $k_2 = 0.5$, $k_3 = 0.3$, $\varepsilon = 3$. The error bars reflect the 95% confidence intervals after bootstrapping ($R = 10^2$). The indices have been computed with the Jansen [3] estimator.

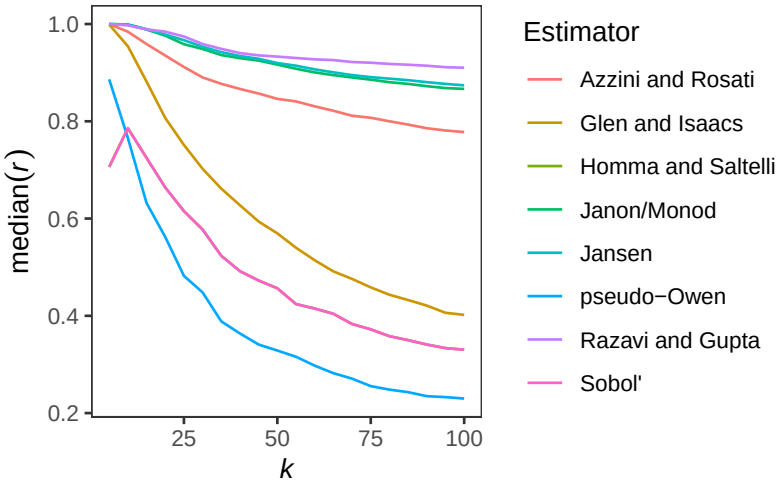

Figure S4: Scatterplot of the model output r against the function dimensionality. Note that the line for the Homma and Saltelli [4] estimator does not show because it mimics the trend of Sobol' [5].

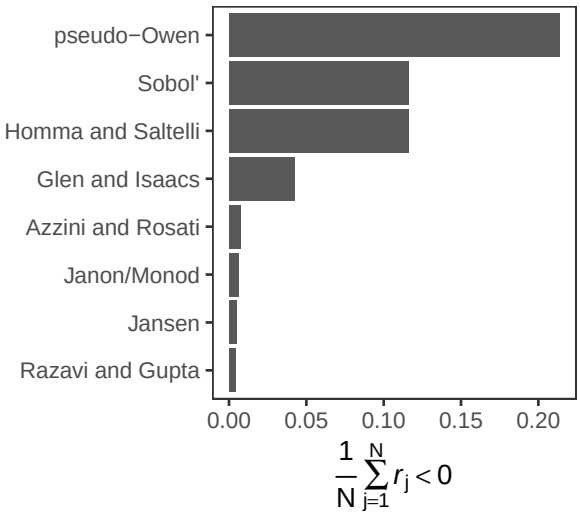

Figure S5: Proportion of model runs yielding $r < 0$.

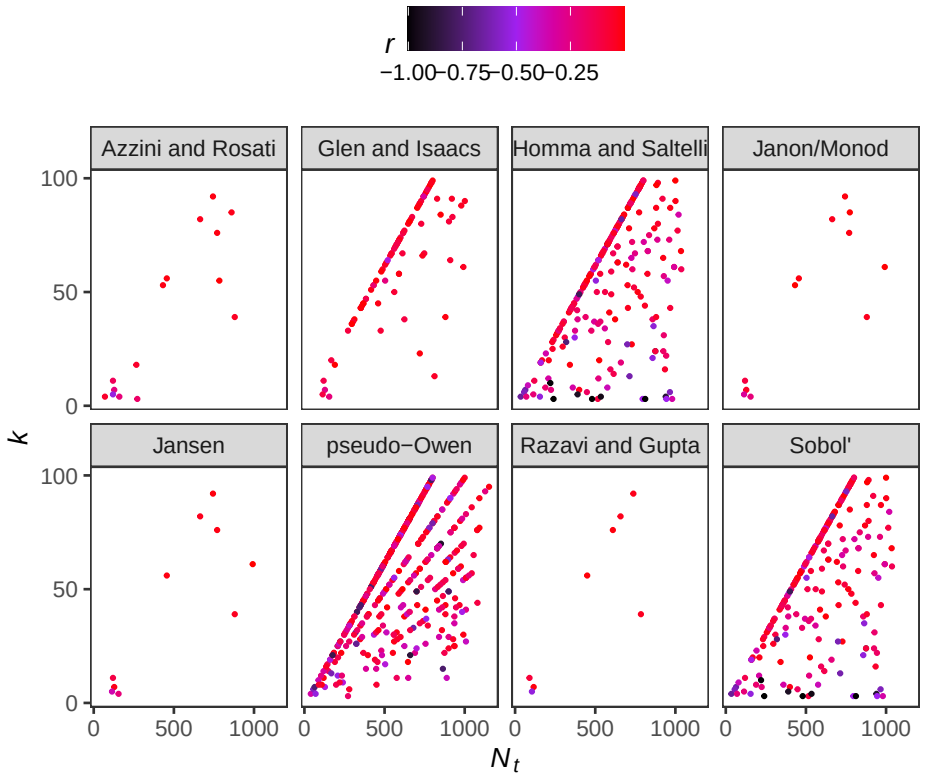

Figure S6: Scatter of the total number of model runs N_t against the function dimensionality k only for $r < 0$.

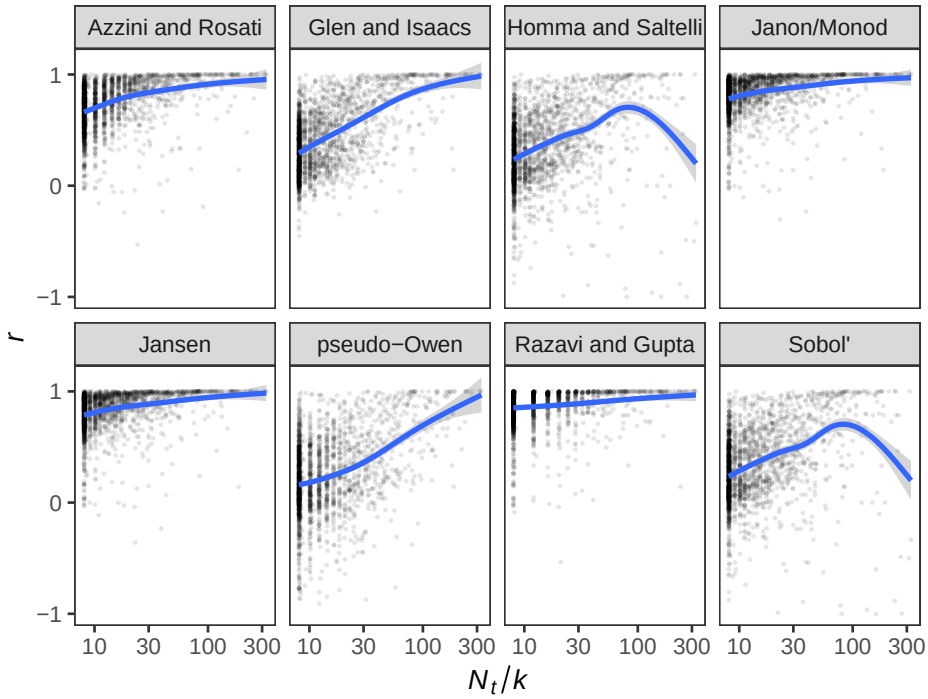

Figure S7: Scatterplot of the correlation between $\boldsymbol{T}_i$ and $\hat{\boldsymbol{T}}_i$ (r) against the number of model runs allocated per model input (Nt/k).

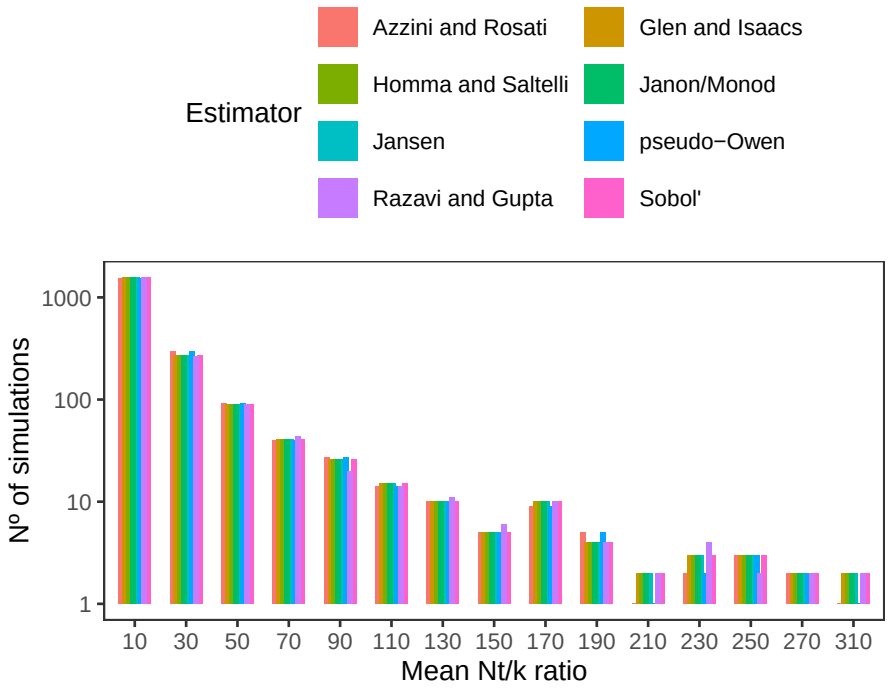

Figure S8: Bar plot with the number of simulations conducted in each of the N_t/k compartments assessed. All estimators have approximately the same number of simulations in each N_t/k compartment

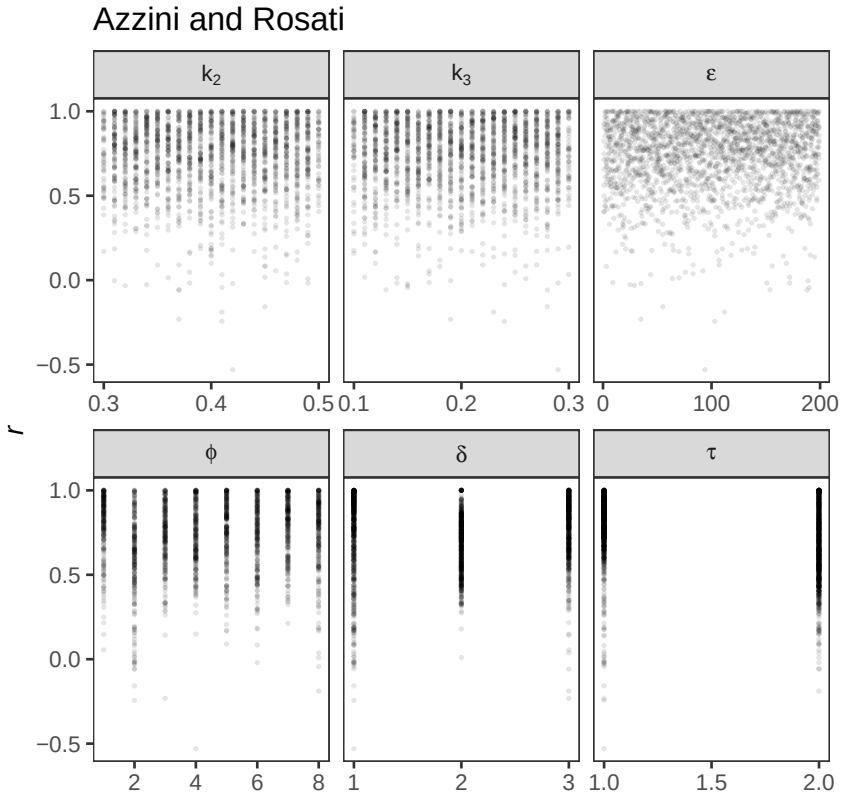

Figure S9: Scatterplots of the model output against the uncertain model parameters for the Azzini and Rosati [6] estimator.

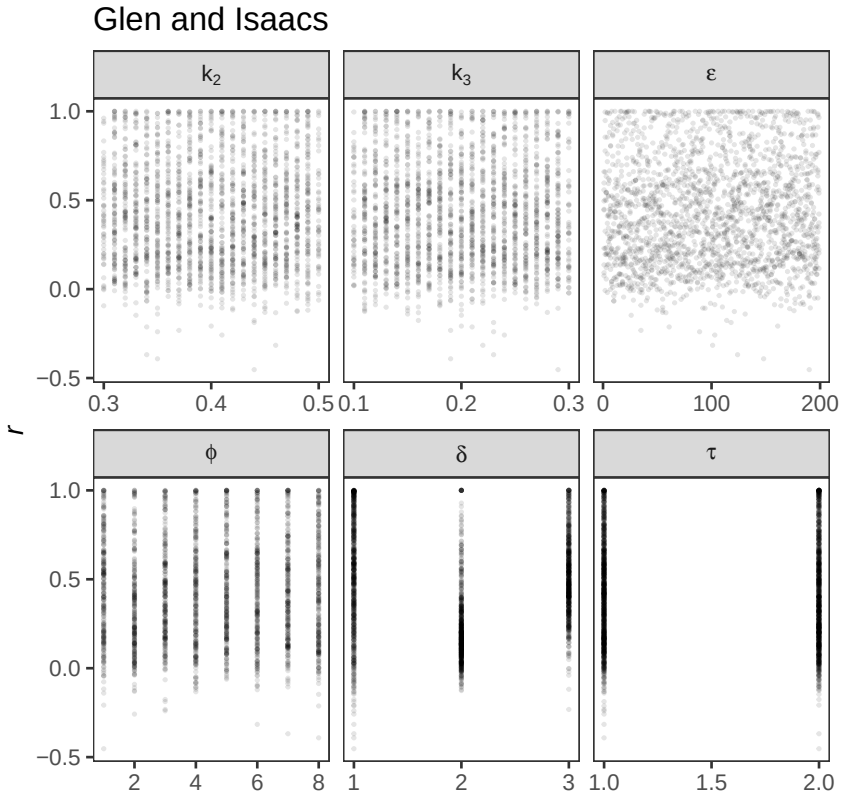

Figure S10: Scatterplots of the model output against the uncertain model parameters for the Glen and Isaacs [7] estimator.

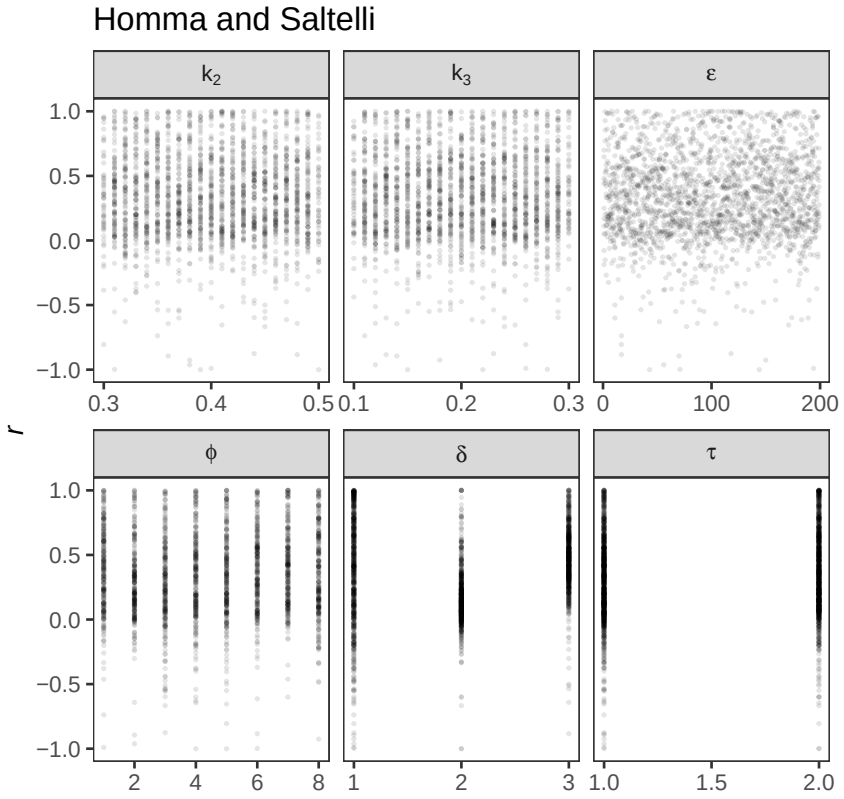

Figure S11: Scatterplots of the model output against the uncertain model parameters for the Homma and Saltelli [4] estimator.

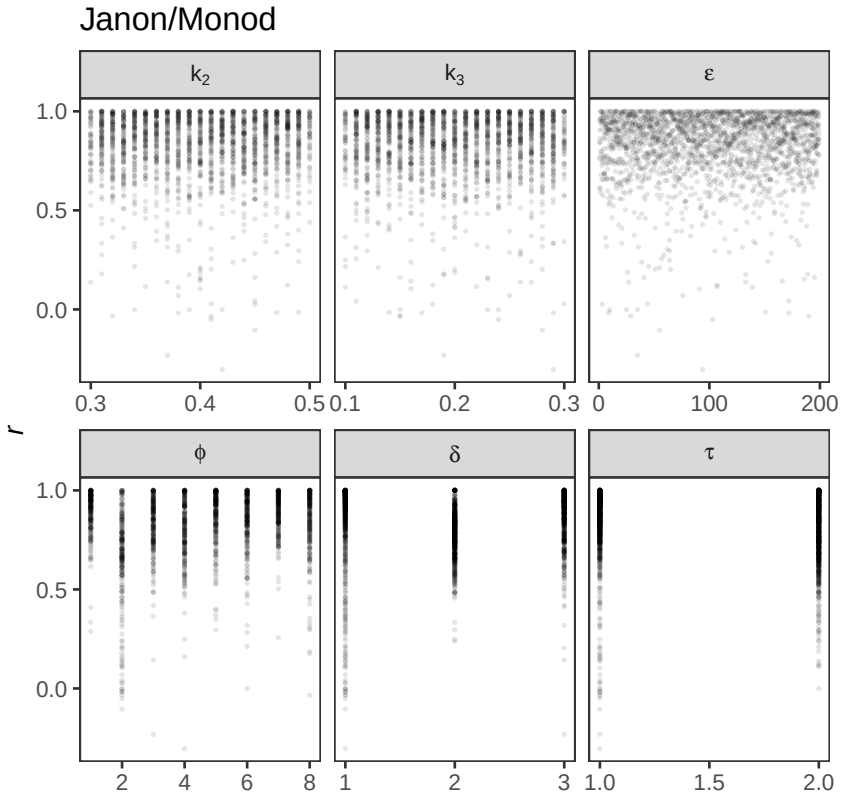

Figure S12: Scatterplots of the model output against the uncertain model parameters for the Janon/Monod [8, 9] estimator.

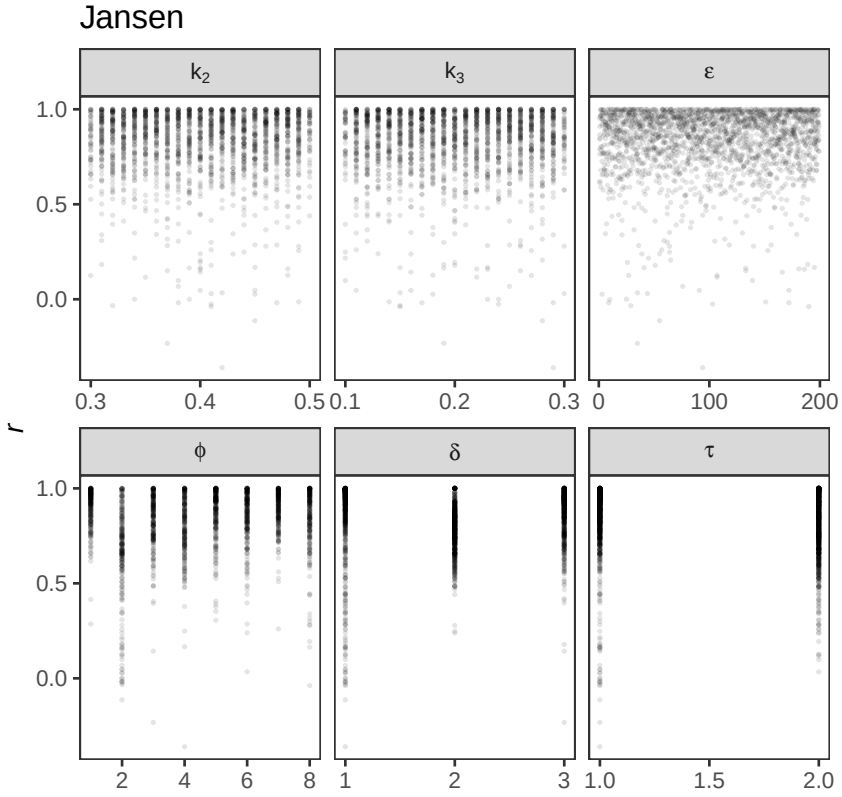

Figure S13: Scatterplots of the model output against the uncertain model parameters for the Jansen [3] estimator.

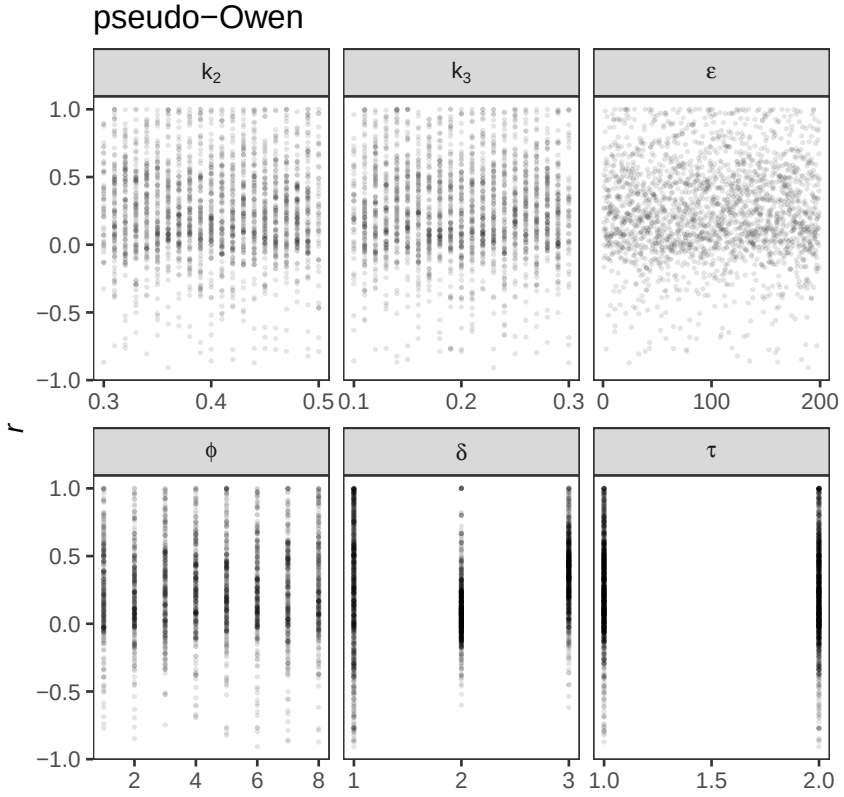

Figure S14: Scatterplots of the model output against the uncertain model parameters for the pseudo-Owen [10] estimator.

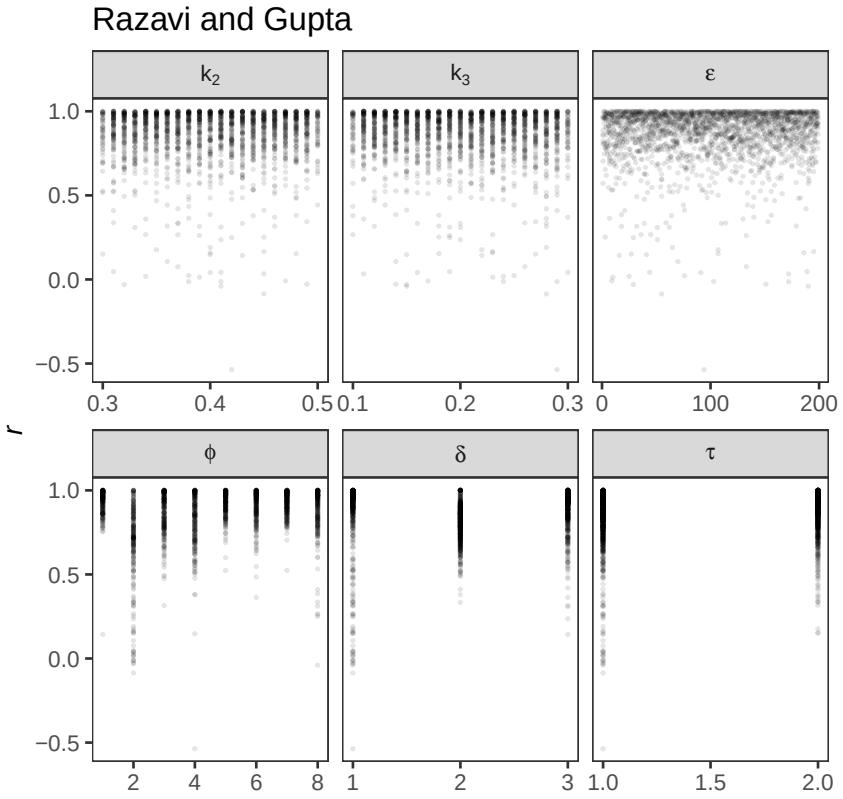

Figure S15: Scatterplots of the model output against the uncertain model parameters for the Razavi and Gupta [1, 2] estimator.

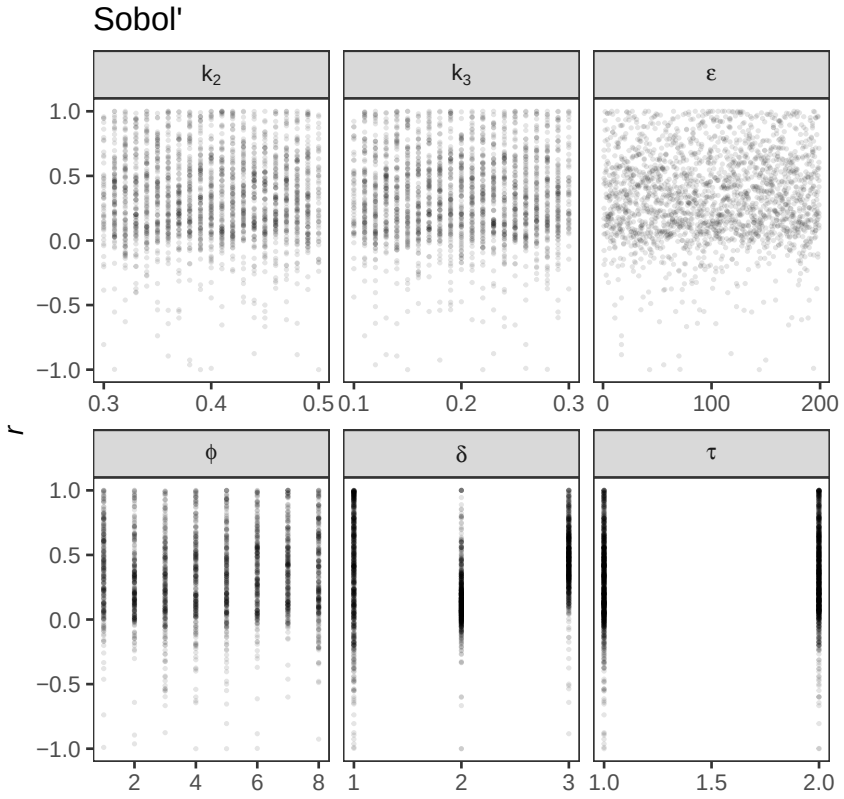

Figure S16: Scatterplots of the model output against the uncertain model parameters for the Sobol' [5] estimator.

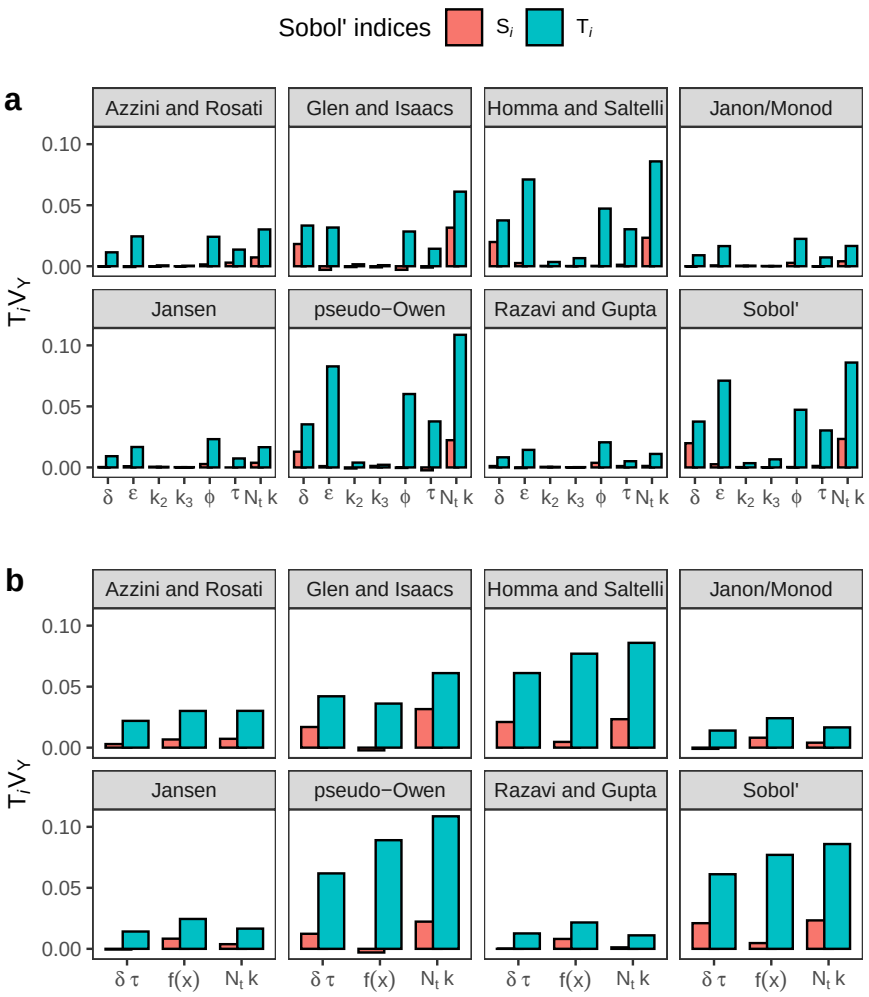

Figure S17: Unnormalized Sobol' indices. a) Parameters. Note that the effect of N_t and k was assessed simultaneously due to their correlation. b) Clusters. $f(x)$ groups all parameters that describe the model uncertainty $(\varepsilon, k_2, k_3, \phi)$.

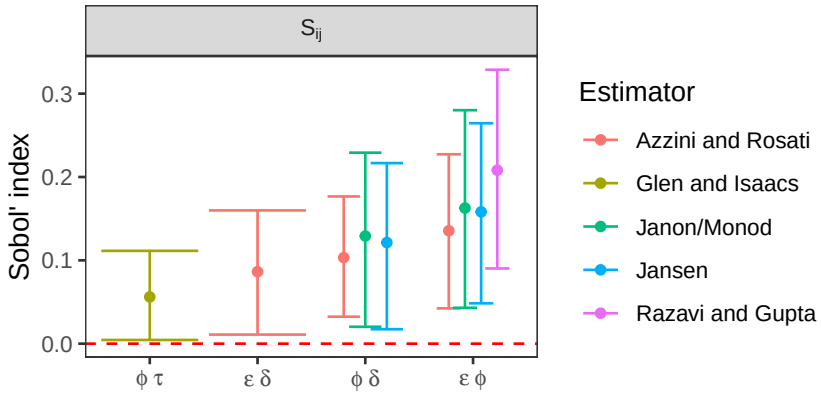

Figure S18: Second-order (S_{ij}) effects. We only show those effects whose lower confidence interval do not overlap with 0.

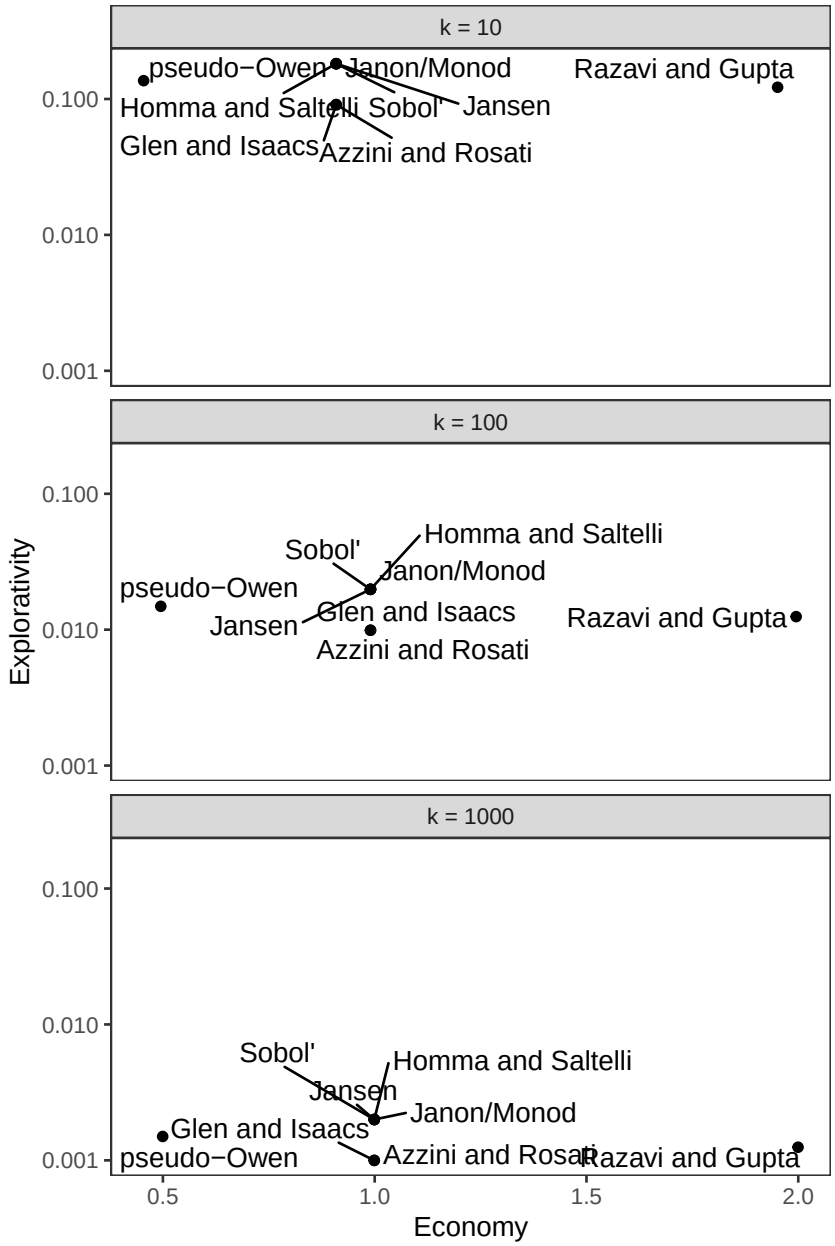

Figure S19: Economy and explorativity of the T_i estimators at $n = 2$ and different dimensionalities ($k = 10$, $k = 100$, $k = 1000$) (see Table S1).

3 Tables

Table S1: Economy and explorativity of the T_i estimators. n equals the number of matrices, k is the dimensionality and Δh is the distance between two points $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$. The economy and explorativity formulae for Razavi and Gupta has been calculated for $\Delta h = 0.2$, the value used in our paper. See Lo Piano et al. [11] for a full description of how the formulae have been derived for each estimator.

Estimator	Economy	Explorativity
Jansen [3] Sobol' [5] Homma and Saltelli [4] Janon/Monod [12, 13]	$\frac{k}{k+1}$	$\frac{2}{k+1}$
Glen and Isaacs [7] Azzini and Rosati [6]	$\frac{k}{k+1}$	$\frac{1}{k+1}$
pseudo-Owen [10]	$\frac{k}{2(k+1)}$	$\frac{3}{2(1+k)}$
Razavi and Gupta [1, 2]	$\frac{\frac{2k}{\Delta h}}{1+\frac{k}{\Delta h}}$	$\frac{1+\frac{1}{\Delta h}}{1+\frac{k}{\Delta h}}$

References

[1] S. Razavi and H. V. Gupta. "A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory". *Water Resources Research* 52.1 (Jan. 2016), 440–455. DOI: 10.1002/2015WR017559. arXiv: 2014WR016527 [10.1002].

[2] S. Razavi and H. V. Gupta. "A new framework for comprehensive, robust and efficient global sensitivity analysis: 2. Application". *Water Resources Research* 52 (Sept. 2016), 440–455. DOI: 110.1002/2015WR017558.

[3] M. Jansen. "Analysis of variance designs for model output". *Computer Physics Communications* 117.1-2 (Mar. 1999), 35–43. DOI: 10.1016/S0010-4655(98)00154-4.

[4] T. Homma and A. Saltelli. "Importance measures in global sensitivity analysis of nonlinear models". *Reliability Engineering & System Safety* 52 (1996), 1–17. DOI: 10.1016/0951-8320(96)00002-6.

[5] I. M. Sobol'. "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates". *Mathematics and Computers in Simulation* 55.1-3 (Feb. 2001), 271–280. DOI: 10.1016/S0378-4754(00)00270-6.

[6] I. Azzini and R. Rosati. "The IA-Estimator for Sobol' sensitivity indices". *Ninth International Conference on Sensitivity Analysis of Model Output*. Barcelona, 2019.

[7] G. Glen and K. Isaacs. "Estimating Sobol sensitivity indices using correlations". *Environmental Modelling and Software* 37 (2012), 157–166. DOI: 10.1016/j.envsoft.2012.03.014.

[8] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. "Asymptotic normality and efficiency of two Sobol index estimators". *ESAIM - Probability and Statistics* 18.Toulouse 3 (2014), 342–364. DOI: 10.1051/ps/2013040. arXiv: 1303.6451.

[9] H. Monod, C. Naud, and D. Makowski. "Uncertainty and sensitivity analysis for crop models". *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*. Ed. by D. Wallach, D. Makowksi, and J. Jones. Elsevier, 2006, 55–99. DOI: 10.1016/j.ress.2007.06.003.

[10] B. Iooss, A. Janon, G. Pujol, with contributions from Baptiste Broto, K. Boumhaout, S. D. Veiga, T. Delage, R. E. Amri, J. Fruth, L. Gilquin, J. Guillaume, L. Le Gratiet, P. Lemaitre, A. Marrel, A. Meynaoui, B. L. Nelson, F. Monari, R. Oomen, O. Rakovec, B. Ramos, O. Roustant, E. Song, J. Staum, R. Sueur, T. Touati, and F. Weber. *sensitivity: Global Sensitivity Analysis of Model Outputs*. 2020.

[11] S. Lo Piano, A. Puy, E. Borgonovo, and A. Saltelli. "Driving trends towards the European Structural Investment Funds payment patters".

[12] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. "Asymptotic normality and efficiency of two Sobol index estimators". *ESAIM: Probability and Statistics* 18.3 (2014), 342–364. DOI: 10.1051/ps/2013040. arXiv: arXiv:1303.6451v1.

[13] H. Monod, C. Naud, and D. Makowski. *Uncertainty and sensitivity analysis for crop models*. 2006, 35–100. DOI: 10.1016/j.ress.2007.06.003.