A comprehensive comparison of total-order estimators for global sensitivity analysis

Arnald Puy*[1,2], William Becker[3], Samuele Lo Piano[4], and Andrea Saltelli[5]

[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] *School of the Built Environment, JJ Thompson Building, University of Reading, Whiteknights Campus, Reading, RG6 6AF, United Kingdom*

[5] *Open Evidence Research, Universitat Oberta de Catalunya (UOC), Barcelona, Spain.*

Abstract

Sensitivity analysis helps identify which model inputs convey the most uncertainty to the model output. One of the most authoritative measures in global sensitivity analysis is the Sobol' total-order index, 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 the benchmark setting defined by the analyst (the sampling method, the distribution of the model inputs, the number of model runs, the test function or model and its dimensionality, the weight of higher order effects or the performance measure selected). Here we compare several total-order estimators in an eight-dimension hypercube where these benchmark parameters are treated as random parameters. This arrangement significantly relaxes the dependency of the results on the benchmark design. We observe that the most accurate estimators are Razavi and Gupta's, Jansen's or Janon/Monod's for factor prioritization, and Jansen's, Janon/Monod's or Azzini and Rosati's for approaching the "true" total-order indices. The rest lag considerably behind. Our work helps analysts navigate the myriad of total-order formulae by reducing the uncertainty in the selection of the most appropriate estimator.

Keywords: Uncertainty analysis; sensitivity analysis; modeling; Sobol' indices; variance decomposition, benchmarking analysis

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]. Among the

*Corresponding author

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 [4–6]. Given a model of the form $y = f(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2, ..., x_i, ..., x_k) \in \mathbb{R}^k$, 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

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

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, a setting known as *factor prioritization* [1].

Homma and Saltelli [7] also proposed the calculation of the total-order index T_i, which measures the first-order effect of a model input jointly with 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 the dimensionality of the uncertain space, a setting 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 metamodeling approaches [8, 9]. The Fourier Amplitude Sensitivity Test (FAST) may also be used to calculate T_i, which involves transforming input variables into periodic functions of a single frequency variable, sampling the model and analysing the sensitivity of input variables using Fourier analysis in the frequency domain [10, 11]. While an innovative approach, FAST is sensitive to the characteristic frequencies assigned to input variables, and is not a very intuitive method - for these reasons it has mostly been superseded by Monte Carlo approaches, or by metamodels when computational expense is a serious issue. In this work we focus on the former.

MC methods require generating a $(N, 2k)$ base sample matrix with either random or quasi-random numbers (e.g. Latin Hypercube Sampling, Sobol' quasi-random numbers [12, 13]), where each row is a sampling point and each column a model input. The first k columns are allocated to an $\boldsymbol{A}$ matrix and the remaining k columns to a $\boldsymbol{B}$ matrix, which are known as the "base sample matrices". Any point in either $\boldsymbol{A}$ or $\boldsymbol{B}$ can be indicated as x_{vi}, where v and i respectively index the row (from 1 to N) and the column (from 1 to k). Then, k additional $\boldsymbol{A}_B^{(i)}$ ($\boldsymbol{B}_A^{(i)}$) matrices are created, where all columns come from $\boldsymbol{A}$ ($\boldsymbol{B}$)

except the i-th column, which comes from $\boldsymbol{B}$ ($\boldsymbol{A}$). The numerator in Equation 3 is finally estimated using the model evaluations obtained from the $\boldsymbol{A}$ ($\boldsymbol{B}$) and $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ ($\boldsymbol{B}_{\boldsymbol{A}}^{(i)}$) matrices. Some estimators may also use a third or $\boldsymbol{X}$ base sample matrices (i.e. $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots, \boldsymbol{X}$), although the use of more than three matrices has been recently proven inefficient by Lo Piano et al. [14].

1.1 Total-order estimators and uncertainties in the benchmark settings

The search for efficient and robust total-order estimators is an active field of research [1, 7, 15–20]. Although some works have compared their asymptotic properties (i.e. [16]), most studies have promoted empirical comparisons where different estimators are benchmarked against known test functions and specific sample sizes. However valuable these empirical studies may be, Becker [21] observed that their results are very much conditional on the choice of model, its dimensionality and the selected number of model runs. It is hard to say from previous studies whether an estimator outperforming another truly reflects its higher accuracy or simply its better performance under the narrow statistical design of the study. Below we extend the list of factors which Becker [21] regards as influential in a given benchmarking exercise and discuss how they affect the relative performance of sensitive estimators.

- *The sampling method:* The creation of the base sample matrices can be done using Monte-Carlo (MC) or quasi Monte-Carlo (QMC) methods [12, 13]. Compared to MC, QMC allows to more effectively map the input space as it leaves smaller unexplored volumes (Fig. S1). However, Kucherenko et al. [22] 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 when benchmarking sensitivity indices [15, 23].

- *The form of the test function:* some of the most commonly used functions in SA are the Ishigami and Homma [24]'s, the Sobol' G and its variants [23, 25], the Bratley and Fox [26]'s or the set of functions presented in Kucherenko et al. [22] [14, 16, 18, 23]. Despite being analytically tractable, these functions capture only one possible interval 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 [27]), and to the higher demand for model transparency [3, 28, 29]. This renders the functional form of the model similar to a random variable [21], something not accounted for by previous works [14, 16, 18, 23].

- *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' [25]'s G or Bratley and Fox [26] 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 [30], including the Soil and Water Assessment Tool (SWAT) model, where $k = 50$ [31], or the Modélisation Environmentale-Surface et Hydrologie (MESH) model, where $k = 111$ [32].

- *The distribution of the model inputs*: the large majority of benchmarking exercises assume uniformly-distributed inputs $p(\boldsymbol{x}) \in U(0,1)^k$ [14, 16, 23, 33]. However, there is evidence 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 [34, 35]. Furthermore, in uncertainty analysis – e.g. in decision theory, the analysts may use distributions with peaks for the most likely values derived, for instance, from an experts elicitation stage.

- *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 (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 [21].

- *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

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

where p is the number of replicas of the sample matrix, and T_i and $\hat{T}_i$ the analytical and the estimated total-order index of the i-th input. 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 $\hat{\boldsymbol{T}} = \{\hat{T}_1, \hat{T}_2, ..., \hat{T}_i, ..., \hat{T}_k\}$ 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 the vectors $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$, which reflect the ranks in $\boldsymbol{T}$ and $\hat{\boldsymbol{T}}$ respectively, such as the Spearman's ρ or the Kendall's W coefficient [21, 36, 37]. 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 [30]. Savage [38] scores or other measures that emphasize this top-down correlation are then a more suitable choice.

Here we benchmark the performance of eight different MC-based formulae available to estimate T_i (Table 1). While the list is not exhaustive, they reflect the research conducted on T_i over the last 20 years: from the classic estimators of Saltelli et al. [1], Homma and Saltelli [7], and Jansen [15] up to the new contributions by Janon et al. [16], Glen and Isaacs [17], Azzini and Rosati [33] and Razavi

and Gupta [20, 39]. In order to reduce the influence of the benchmarking design in the assessment of the estimators' accuracy, we treat 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. 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 setting where all formulae can prove their accuracy. We therefore extend Becker [21]'s approach by testing a wider set of Monte Carlo estimators, by exploring a wider range of benchmarking assumptions and by performing a formal SA on these assumptions. The aim is therefore to provide a much more global comparison of available MC estimators than is available in the existing literature, and investigate how the benchmarking parameters may affect the relative performance of estimators. Such information can help point to estimators that are not only efficient on a particular case study, but efficient and robust to a wide range of practical situations.

2 Assessment of the uncertainties in the benchmarking parameters

In this section we formulate the benchmarking parameters as random variables and assess how the performance of estimators is dependent on them by performing a sensitivity analysis. In essence this is a *sensitivity analysis of sensitivity analyses* [42], and a natural extension of a similar uncertainty analysis in a recent work by Becker [21]. The use of global sensitivity analysis tools to better understand the properties of estimators can give insights into how estimators behave in different scenarios that are not available through analytical approaches.

2.1 The setting

The variability in the benchmark settings $(\tau, N_t, k, k_2, k_3, \phi, \epsilon, \delta)$ is described by probability distributions (Table 2). We assign uniform distributions (discrete or continuous) to each parameter. In particular, we choose $\tau \sim \mathcal{DU}(1, 2)$ 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 base sample matrices. For $\tau = 2$ we use the Sobol' sequence scrambled according to Owen [43] to avoid repeated coordinates at the beginning of the sequence. The total number of model runs and inputs is 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 $\boldsymbol{B}$, $\boldsymbol{B}_A^{(i)}$, $\boldsymbol{A}_B^{(i)}$ or $\boldsymbol{C}_B^{(i)}$ matrices (Table 1), we modify the space defined by (N_t, k) to a non-rectangular domain (we provide more information on this adjustment in Section 2.2).

For ϕ we set $\phi \sim \mathcal{DU}(1, 8)$ to ensure an adequate representation of the most common shapes in the $(0, 1)^k$ domain. Besides the normal distribution truncated at $(0, 1)$ and the uniform distribution, we also take

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

Table 1: Formulae to compute T_i. f_0 and $V(y)$ are estimated according to the original papers. For estimators 2 and 5, $f_0 = \frac{1}{N}\sum_{v=1}^{N} f(\boldsymbol{A})_v$. For estimators 1, 2 and 5, $V(y) = \frac{1}{N}\sum_{v=1}^{N}[f(\boldsymbol{A})_v - f_0]^2$ [1, Eq. 4.16, 7, Eqs. 15, 20]. For estimator 3, $f_0 = \frac{1}{N}\sum_{v=1}^{N}\frac{f(\boldsymbol{A})_v + f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v}{2}$ and $V(y) = \frac{1}{N}\sum_{v=1}^{N}\frac{f(\boldsymbol{A})_v^2 + f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v^2}{2} - f_0^2$ [16, Eq. 15]. In estimator 4, $\langle f(\boldsymbol{A})_v \rangle$ is the mean of $f(\boldsymbol{A})_v$. We use a simplified version of the Glen and Isaacs estimator because spurious correlations are zero by design. As for estimator 7, we refer to it as pseudo-Owen given its use of a $\boldsymbol{C}$ matrix and its identification with Owen [40] in Iooss et al. [41], where we retrieve the formula from. $V(y)$ in Estimator 7 is computed as in Estimator 3 following Iooss et al. [41], whereas $V(y)$ in Estimator 8 is computed as in Estimator 1.

Nž	Estimator	Author
1	$\frac{\frac{1}{2N}\sum_{v=1}^{N}\left[f(\boldsymbol{A})_v - f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v\right]^2}{V(y)}$	Jansen [15]
2	$\frac{V(y) - \frac{1}{N}\sum_{v=1}^{N} f(\boldsymbol{A})_v f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v + f_0^2}{V(y)}$	Homma and Saltelli [7]
3	$1 - \frac{\frac{1}{N}\sum_{v=1}^{N} f(\boldsymbol{A})_v f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v - f_0^2}{V(y)}$	Janon et al. [16] Monod et al. [19]
4	$1 - \left[\frac{1}{N-1}\sum_{v=1}^{N}\frac{[f(\boldsymbol{A})_v - \langle f(\boldsymbol{A})_v \rangle]\left[f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v - \left\langle f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v \right\rangle\right]}{\sqrt{V[f(\boldsymbol{A})_v]V\left[f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v\right]}}\right]$	Glen and Isaacs [17]
5	$1 - \frac{\frac{1}{N}\sum_{v=1}^{N} f(\boldsymbol{B})_v f(\boldsymbol{B}_{\boldsymbol{A}}^{(i)})_v - f_0^2}{V(y)}$	Saltelli et al. [1]
6	$\frac{\sum_{v=1}^{N}[f(\boldsymbol{B})_v - f(\boldsymbol{B}_{\boldsymbol{A}}^{(i)})_v]^2 + [f(\boldsymbol{A})_v - f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v]^2}{\sum_{v=1}^{N}[f(\boldsymbol{A})_v - f(\boldsymbol{B})_v]^2 + [f(\boldsymbol{B}_{\boldsymbol{A}}^{(i)})_v - f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_v]^2}$	Azzini et al. [18] and Azzini and Rosati [33]
7	$\frac{V(y) - \left[\frac{1}{N}\sum_{v=1}^{N}\left\{\left[f(\boldsymbol{B})_v - f(\boldsymbol{C}_{\boldsymbol{B}}^{(i)})_v\right]\left[f(\boldsymbol{B}_{\boldsymbol{A}}^{(i)})_v - f(\boldsymbol{A})_v\right]\right\}\right]}{V(y)}$	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 and Gupta [20, 39] (see SM).

into account four beta distributions parametrized with distinct α and β values and a logitnormal distribution (Fig. 1a). The aim is to check the response of the estimators under a wide range of probability distributions, including U-shaped distributions and distributions with different degrees of skewness.

We link each distribution in Fig. 1a to an integer value from 1 to 7. For instance, if $\phi = 1$, the joint probability distribution of the model inputs is described as $p(x_1, \ldots, x_k) = \mathcal{U}(0,1)^k$. If $\phi = 8$, we create a vector $\boldsymbol{\phi} = \{\phi_1, \phi_2, ..., \phi_i, ..., \phi_k\}$ by randomly sampling the seven distributions in Fig. 1a, and use the i-th distribution in the vector to describe the uncertainty of the i-th input. This last case examines the behavior of the estimators when several distributions are used to characterize the uncertainty in the model input space.

Table 2: Summary of the parameters and their distributions. $\mathcal{DU}$ stands for 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	$\mathcal{DU}(1,2)$

2.1.1 The test function

The parameter ε operationalizes the randomness in the form and execution of the test function. Our test function is an extended version of Becker [21]'s metafunction, which randomly combines p univariate functions in a multivariate function of dimension k. Here we consider the 10 univariate functions listed in Fig. 1b, which represent common responses observed in physical systems and in classic SA test functions (see Becker [21] for a discussion on this point). We note that an alternative approach would be to construct orthogonal basis functions which could allow analytical evaluation of true sensitivity indices for each generated function; however, this extension is left for future work.

We construct the test function as follows:

1. Let us consider a sample matrix such as

$$\boldsymbol{M} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1i} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{v1} & x_{v2} & \cdots & x_{vi} & \cdots & x_{vk} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Ni} & \cdots & x_{Nk} \end{bmatrix} \tag{5}$$

   where every point $\boldsymbol{x}_v = x_{v1}, x_{v2}, \ldots, x_{vk}$ represents a given combination of values for the k inputs and x_i is a model input whose distribution is defined by ϕ.

2. Let $\boldsymbol{u} = \{u_1, u_2, ..., u_k\}$ be a k-length vector formed by randomly sampling with replacement the ten functions in Fig. 1b. The i-th function in $\boldsymbol{u}$ is then applied to the i-th model input: for instance, if $k = 4$ and $\boldsymbol{u} = \{u_3, u_4, u_8, u_1\}$, then $f_3(x_1) = \frac{e^{x_1}-1}{e-1}$, $f_4(x_2) = (10 - \frac{1}{1.1})^{-1}(x_2 + 0.1)^{-1}$, $f_8(x_3) = \frac{\sin(2\pi x_3)}{2}$, and $f_1(x_4) = x_4^3$. The elements in $\boldsymbol{u}$ thus represent the first-order effects of each model input.

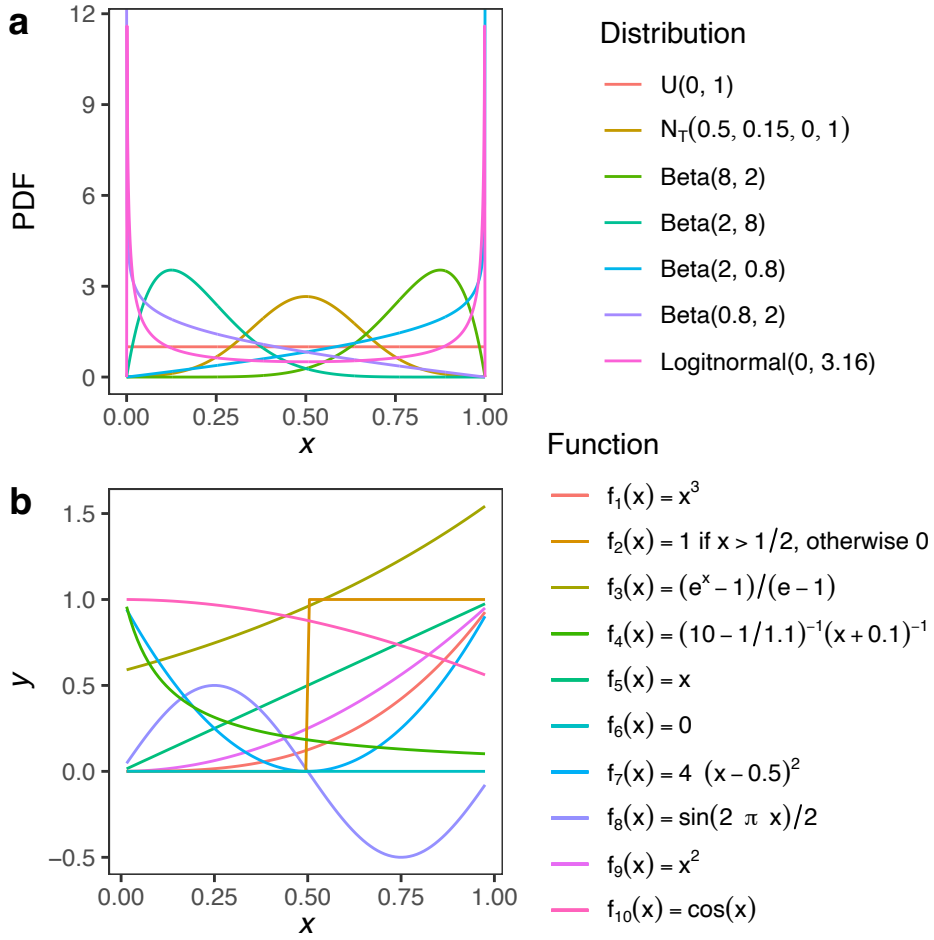

Figure 1: The metafunction approach. a) Probability distributions incuded in ϕ. N_T stands for truncated normal distribution. b) Univariate functions included in the metafunction ($f_1(x)$ = cubic, $f_2(x)$ = discontinuous, $f_3(x)$ = exponential, $f_4(x)$ = inverse, $f_5(x)$ = linear, $f_6(x)$ = no effect, $f_7(x)$ = non-monotonic, $f_8(x)$ = periodic, $f_9(x)$ = quadratic, $f_{10}(x)$ = trigonometric).

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 of the model. Each row in $\boldsymbol{V}$ thus specifies an interaction between two columns in $\boldsymbol{M}$. In the case of $k = 4$ and the same elements in $\boldsymbol{u}$ as defined in the previous example,

$$\boldsymbol{V} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 2 & 3 \\ 2 & 4 \\ 3 & 4 \end{bmatrix} \tag{6}$$

e.g., the first row promotes $f_3(x_1) \cdot f_4(x_2)$, the second row $f_3(x_1) \cdot f_8(x_3)$, and so on until the n-th row. In order to follow the *sparsity of effects principle* (most variations in a given model output should be explained by low-order interactions [44]), the metafunction activates only a fraction of these effects: it randomly samples $\lceil k_2 n \rceil$ rows from $\boldsymbol{V}$, and computes the corresponding interactions

in $\boldsymbol{M}$. $\lceil k_2 n \rceil$ is thus the number of pairwise interactions present in the function. We make k_2 an uncertain parameter described as $k_2 \sim \mathcal{U}(0.3, 0.5)$ in order to randomly activate only between 30% and 50% of the available second-order effects in $\boldsymbol{M}$.

4. Same as before, 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 $\boldsymbol{M}$. For $k = 4$ and $\boldsymbol{u}$ as before,

$$\boldsymbol{W} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 4 \\ 2 & 3 & 4 \end{bmatrix} \tag{7}$$

e.g. the first row leads to $f_3(x_1) \cdot f_4(x_2) \cdot f_8(x_3)$, and so on until the m-th row. The metafunction then randomly samples $\lceil k_3 m \rceil$ rows from $\boldsymbol{W}$ and computes the corresponding interactions in $\boldsymbol{M}$. $\lceil k_3 m \rceil$ is therefore the number of three-wise interaction terms in the function. We also make k_3 an uncertain parameter described as $k_3 \sim \mathcal{U}(0.1, 0.3)$ to activate only between 10% and 30% of all third-order effects 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. Three vectors of coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ of length k, n and m are defined to represent the weights of the first, second and third-order effects respectively. 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 [45] principle (around 80% of the effects are due to 20% of the parameters), found to widely apply in SA [1, 46].

6. The metafunction can thus be formalized as

$$\begin{aligned} y = & \sum_{i=1}^{k} \alpha_i f^{u_i} \phi_i(x_i) \\ & + \sum_{i=1}^{\lceil k_2 n \rceil} \beta_i f^{u_{V_{i,1}}} \phi_i(x_{V_{i,1}}) f^{u_{V_{i,2}}} \phi_i(x_{V_{i,2}}) \\ & + \sum_{i=1}^{\lceil k_3 m \rceil} \gamma_i f^{u_{W_{i,1}}} \phi_i(x_{W_{i,1}}) f^{u_{W_{i,2}}} \phi_i(x_{W_{i,2}}) f^{u_{W_{i,3}}} \phi_i(x_{W_{i,3}}) \,. \end{aligned} \tag{8}$$

Note that there is randomness in the sampling of $\boldsymbol{\phi}$, the univariate functions in $\boldsymbol{u}$ and the coefficients in $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The parameter ε assesses the influence of this randomness by fixing the starting point of the pseudo-random number sequence used for sampling the parameters just mentioned. We use $\varepsilon \sim \mathcal{U}(1, 200)$ to ensure that the same seed does not overlap with the same value of N_t, k or any other parameter, an issue that might introduce determinism in a process that should be stochastic. In Figs. S2–S3 we show the type of T_i indices generated by this metafunction.

Finally, we describe the parameter δ as $\delta \sim \mathcal{DU}(1, 2)$. If $\delta = 1$, we compute the Kendall τ-b correlation coefficient between $\hat{\boldsymbol{r}}$ and $\boldsymbol{r}$, the estimated and the "true" ranks calculated from $\hat{\boldsymbol{T}}$ and $\boldsymbol{T}$ respectively.

This aims at evaluating how well the estimators in Table 1 rank all model inputs. If $\delta = 2$, we compute the Pearson correlation between $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$ after transforming the ranks to Savage scores [38]. This setting examines the performance of the estimators when the analyst is interested in ranking only the most important model inputs. Savage scores are given as

$$Sa_i = \sum_{j=i}^{k} \frac{1}{j}, \tag{9}$$

where j is the rank assigned to the jth element of a vector of length k. If $x_1 > x_2 > x_3$, the Savage scores would then be $Sa_1 = 1 + \frac{1}{2} + \frac{1}{3}$, $Sa_2 = \frac{1}{2} + \frac{1}{3}$, and $Sa_3 = \frac{1}{3}$. The parameter δ thus assesses the accuracy of the estimators in properly ranking the model inputs; in other words, when they are used in a factor prioritization setting [1].

In order to examine also how accurate the estimators are in approaching the "true" indices, we run an extra round of simulations with the MAE as the only performance measure, which we compute as

$$\text{MAE} = \frac{\sum_{i=1}^{k} |T_i - \hat{T}_i|}{k}. \tag{10}$$

Note that, unlike Equation 4, Equation 10 does not make use of replicas. This is because the effect of the sampling is averaged out in our design by simultaneously varying all parameters in many different simulations.

2.2 The execution of the algorithm

We examine how sensitive the performance of total-order estimators is to the uncertainty in the benchmark parameters $\tau, N_t, k, k_2, k_3, \phi, \epsilon, \delta$ by means of a global SA. We create an $\boldsymbol{A}$, $\boldsymbol{B}$ and $k-1$ $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices, each of dimension $(2^{11}, k)$, using Sobol' quasi-random numbers. In these matrices each column is a benchmark parameter described with the probability distributions of Table 2 and each row is a simulation with a specific combination of $\tau, N_t, k, \ldots$ values. Note that we use $k-1$ $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices because we group N_t and k and treat them like a single benchmark parameter given their correlation (see below).

Our algorithm runs rowwise over the $\boldsymbol{A}$, $\boldsymbol{B}$ and $k-1$ $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices, for $v = 1, 2, \ldots, 18,432$ rows. In the v-th row it does the following:

1. It creates five (N_{t_v}, k_v) matrices using the sampling method defined by τ_v. The need for these five sub-matrices responds to the five specific sampling designs requested by the estimators of our study (Table 1). We use these matrices to compute the vector of estimated indices $\hat{\boldsymbol{T}}_i$ for each estimator:
   (a) An $\boldsymbol{A}$ matrix and k_v $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices, each of size (N_v, k_v), $N_v = \lceil\!\lceil \frac{N_{t_v}}{k_v+1} \rceil\!\rceil$ (Estimators 1–4 in Table 1).
   (b) An $\boldsymbol{A}$, $\boldsymbol{B}$ and k_v $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices, each of size (N_v, k_v), $N_v = \lceil\!\lceil \frac{N_{t_v}}{k_v+2} \rceil\!\rceil$ (Estimator 5 in Table 1).
   (c) An $\boldsymbol{A}$, $\boldsymbol{B}$ and k_v $\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ and $\boldsymbol{B}_{\boldsymbol{A}}^{(i)}$ matrices, each of size (N_v, k_v), $N_v = \lceil\!\lceil \frac{N_{t_v}}{2k_v+2} \rceil\!\rceil$ (Estimator 6 in Table 1).

(d) An $\boldsymbol{A}$, $\boldsymbol{B}$ and k_v $\boldsymbol{B}_A^{(i)}$ and $\boldsymbol{C}_B^{(i)}$ matrices, each of size (N_v, k_v), $N_v = \lceil\!\lceil \frac{N_{t_v}}{2k_v+2} \rceil\!\rceil$ (Estimator 7 in Table 1).

(e) A matrix formed by N_v stars, each of size $k_v(\frac{1}{\Delta h} - 1) + 1$. Given that we set Δh at 0.2 (see Supplementary Materials), $N_v = \lfloor\!\lfloor \frac{N_{t_v}}{4k+1} \rfloor\!\rfloor$ (Estimator 8 in Table 1).

The different sampling designs and the value for k_v constrains the total number of runs N_{t_v} that can be allocated to each estimator. Furthermore, given the probability distributions selected for N_t and k (Table 2), specific combinations of (N_{t_v}, k_v) lead to $N_v \leq 1$, which is computationally unfeasible. To minimize these issues we force the comparison between estimators to approximate the same N_{t_v} value. Since the sampling design structure of Razavi and Gupta is the most constraining, we use $N_v = \frac{2(4k+1)}{k+1}$ (for estimators 1–4), $N_v = \frac{2(4k+1)}{k+2}$ (for estimator 5) and $N_v = \frac{2(4k+1)}{2k+2}$ (for estimators 6–7) when $N_v \leq 1$ in the case of Razavi and Gupta. This compels all estimators to explore a very similar portion of the (N_t, k) space, but N_t and k become correlated, which contradicts the requirement of independent inputs characterizing variance-based sensitivity indices [1]. This is why we treat (N_t, k) as a single benchmark parameter in the SA.

2. It creates a sixth matrix, formed by an $\boldsymbol{A}$ and k_v $\boldsymbol{A}_B^{(i)}$ matrices, each of size $(2^{11}, k_v)$. We use this sub-matrix to compute the vector of "true" indices $\boldsymbol{T}$, which could not be calculated analytically due to the wide range of possible functional forms created by the metafunction. Following Becker [21], we assume that a fairly accurate approximation to $\boldsymbol{T}$ could be achieved with a large Monte Carlo estimation.

3. The distribution of the model inputs in these six sample matrices is defined by ϕ_v.

4. The metafunction runs over these six matrices simultaneously, with its functional form, degree of active second and third-order effects as set by ε_v, k_{2_v} and k_{3_v} respectively.

5. It computes the estimated sensitivity indices $\hat{\boldsymbol{T}}_v$ for each estimator and the "true" sensitivity indices $\boldsymbol{T}_v$ using the Jansen [15] estimator, which is currently best practice in SA.

6. It checks the performance of the estimators. This is done in two ways:

   (a) If $\delta = 1$, we compute the correlation between $\hat{\boldsymbol{r}}_v$ and $\boldsymbol{r}_v$ (obtained respectively from $\hat{\boldsymbol{T}}_v$ and $\boldsymbol{T}_v$) with Kendall tau, and if $\delta = 2$ we compute the correlation between $\hat{\boldsymbol{r}}_v$ and $\boldsymbol{r}_v$ on Savage scores. The model output in both cases is the correlation coefficient r, with higher r values indicating a better performance in properly ranking the model inputs.

   (b) We compute the MAE between $\hat{\boldsymbol{T}}_v$ and $\boldsymbol{T}_v$. In this case the model output is the MAE, with lower values indicating a better performance in approaching the "true" total-order indices.

3 Results

3.1 Uncertainty analysis

Under a factor prioritization setting (e.g. when the aim is to rank the model inputs in terms of their contribution to the model output variance), the most accurate estimators are Jansen, Razavi and Gupta, Janon/Monod and Azzini and Rosati. The distribution of r values (the correlation between estimated and "true" ranks) when these estimators are used is highly negatively skewed, with median values of ≈ 0.9. Glen and Isaacs, Homma and Saltelli, Saltelli and pseudo-Owen lag behind and display median r values of ≈ 0.35, with pseudo-Owen ranking last ($r \approx 0.2$). The range of values obtained with these formulae is much more spread out and include a significant number of negative r values, suggesting that they overturned the true ranks in several simulations (Figs. 2a, S4).

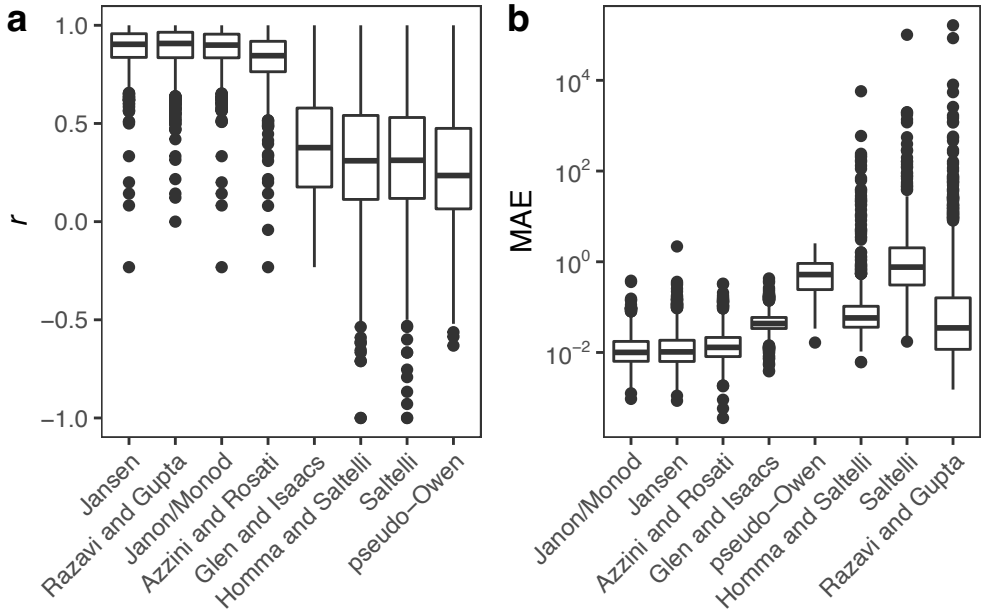

Figure 2: Boxplots summarizing the results of the simulations. a) Correlation coefficient between $\hat{\boldsymbol{r}}$ and $\boldsymbol{r}$, the vector of estimated and "true" ranks. b) Mean Absolute Error (MAE).

When the goal is to approximate the "true" indices, Janon/Monod, Jansen and Azzini and Rosati also offer the best performance. The median MAE obtained with these estimators is generally smaller than Glen and Isaacs' and pseudo-Owen's, and the distribution of MAE values is much more narrower than that obtained with Homma and Saltelli, Saltelli or Razavi and Gupta. These three estimators are the least accurate and produce several MAE values larger than 10^2 in several simulations (Fig. 2b). The volatility of Razavi and Gupta under the MAE is reflected in the numerous outliers produced and sharply contrasts with its very good performance in a factor prioritization setting (Fig. 2a).

To obtain a finer insight into the structure of these results, we plot the total number of model runs N_t against the function dimensionality k (Fig. 3). This maps 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. Under a factor prioritization setting, almost 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. However, some differences unfold at higher dimensions: Saltelli, 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 very good performance overall except in the upper N_t, k boundary, where most of the orange dots concentrate. The estimators of Jansen, Janon/Monod and Razavi and Gupta rank the model inputs almost flawlessly regardless of the region explored in the N_t, k domain (Fig. 3a).

With regards to the MAE, Janon/Monod, Jansen and Azzini and Rosati maintain their high performance regardless of the N_t, k region explored. The accuracy of Razavi and Gupta, however, drops at the upper-leftmost part of the N_t, k boundary, where most of the largest MAE scores are located (MAE > 10). In the case of Saltelli and Homma and Saltelli, the largest MAE values concentrate in the region of small k regardless of the total number of model runs, a domain in which they achieved a high performance when the focus was on properly ranking the model inputs.

The presence of a non-negligible proportion of model runs with $r < 0$ suggests that some estimators significantly overturned the true ranks (Figs 3a, S4). To better examine this phenomenon, we re-plot Fig 3b with just the simulations yielding $r < 0$ (Fig. S5). We observe that $r < 0$ values not only appear 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 emerge at a relatively large N_t and low k in the case of pseudo-Owen, Saltelli and Homma and Saltelli. The Saltelli estimator actually concentrates in the $k < 10$ zone most of the simulations with the lowest negative r values (Fig. S5). This suggests that rank reversing is not an artifact of our study design as much as a by-product of the volatility of these estimators when stressed by the sources of computational uncertainty listed in Table 2. Such strain may lead these estimators to produce a significant fraction of negative indices or indices beyond 1, thus effectively promoting $r < 0$.

We calculate the proportion of $T_i < 0$ and $T_i > 1$ in each simulation that yielded $r < 0$. In the case of Glen and Isaacs and Homma and Saltelli, $r < 0$ values are caused by the production of a large proportion of $T_i < 0$ (25%–75%, the x axis in Fig. 4). Pseudo-Owen and Saltelli suffer this bias too and in several simulations they also generate a large proportion of $T_i > 1$ (up to 100% of the model inputs, the y axis in Fig. 4). The production of $T_i < 0$ and $T_i > 1$ is caused by numerical errors and fostered by the values generated at the numerator of Equation 3: $T_i < 0$ may either derive from $E_{\boldsymbol{x}_{\sim i}}\left[V_{x_i}(y|\boldsymbol{x}_{\sim i})\right] < 0$ (e.g. Homma and Saltelli and pseudo-Owen) or $V_{\boldsymbol{x}_{\sim i}}\left[E_{x_i}(y|\boldsymbol{x}_{\sim i})\right] > V(y)$ (e.g. Saltelli), whereas $T_i > 1$ from $E_{\boldsymbol{x}_{\sim i}}\left[V_{x_i}(y|\boldsymbol{x}_{\sim i})\right] > V(y)$ (e.g. Homma and Saltelli and pseudo-Owen) or $V_{\boldsymbol{x}_{\sim i}}\left[E_{x_i}(y|\boldsymbol{x}_{\sim i})\right] < 0$ (e.g. Saltelli).

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 [21] (Fig. 5, S6). 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, Jansen and Janon/Monod are very good at properly ranking model inputs ($r \approx 0.9$), and are followed very

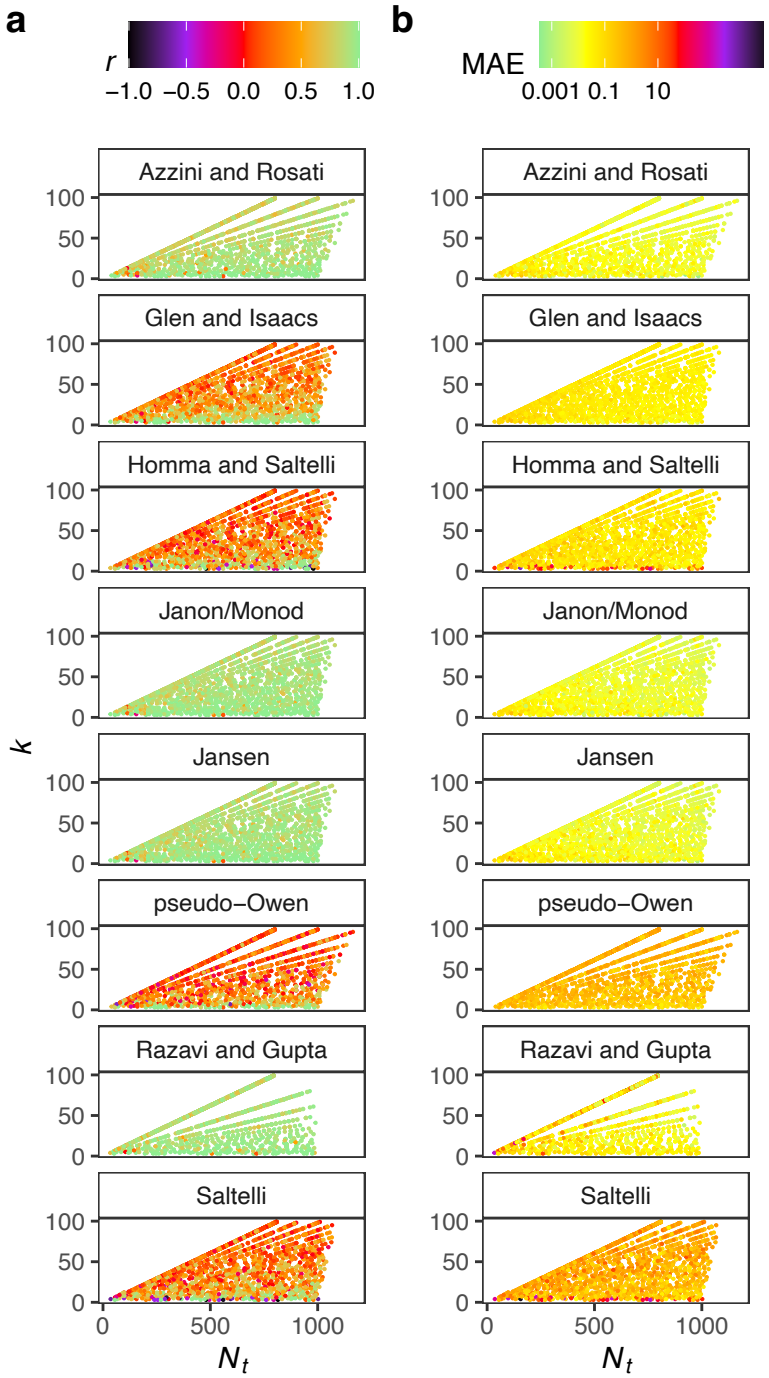

Figure 3: Number of runs N_t against the function dimensionality k. Each dot is a simulation with a specific combination of the benchmark parameters in Table 2. The greener (blacker) the color, the better (worse) the performance of the estimator. a) Accuracy of the estimators when the goal is to properly rank the model inputs, e.g. a factor prioritization setting. b) Accuracy of the estimators when the goal is to approach the "true" total-order indices.

close by Azzini and Rosati ($r \approx 0.8$). Saltelli, Homma and Saltelli and Glen and Isaacs come after ($r \approx 0.3$), with pseudo-Owen scoring last ($r \approx 0.2$). When the N_t/k ratio is increased, all estimators improve their ranking accuracy and some quickly reach the asymptote: this is the case of Razavi and Gupta, Janon/Monod and Jansen, whose performance becomes almost flawless from $(N_t/k) \in [40, 60]$ onwards, and of Azzini

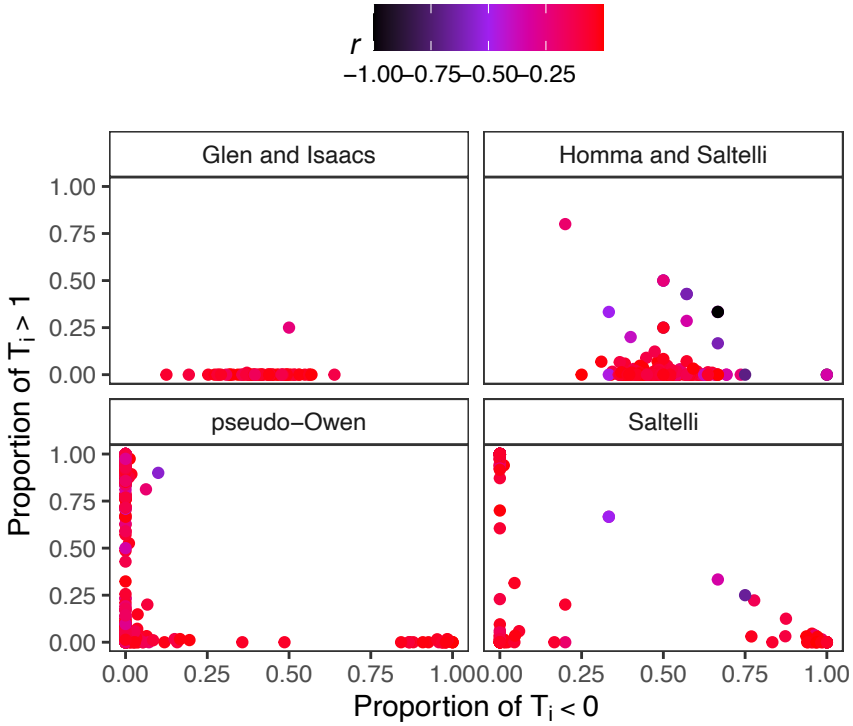

Figure 4: Scatterplot of the proportion of $T_i < 0$ against the proportion of $T_i > 1$ mapped against the model output r. Each dot is a simulation. Only simulations with $r < 0$ are displayed.

and Rosati, which reaches its optimum at $(N_t/k) \in [60, 80]$. The accuracy of the other estimators does not seem to fully stabilize within the range of ratios examined. In the case of Homma and Saltelli and Saltelli, their performance oscillates before plummeting at $(N_t/k) \in [200, 210]$, $(N_t/k) \in [240, 260]$ and $(N_t/k) \in [260, 280]$ due to several simulations yielding large $r < 0$ values (Fig. 5a).

Janon/Monod and Jansen are also the most efficient estimators when the MAE is the measure of choice, followed closely by Azzini and Rosati, Razavi and Gupta and Glen and Isaacs. Saltelli and Homma and Saltelli gain accuracy at higher N_t/k ratios yet their precision diminishes all the same from $(N_t/k) \in [200, 210]$ onwards (Fig. 5b).

3.2 Sensitivity analysis

When the aim is to rank the model inputs, the selection of the performance measure (δ) has the highest first-order effect in the accuracy of the estimators (Fig. 6a). The parameter δ is responsible for between 20% (Azzini and Rosati) and 30% (Glen and Isaacs) of the variance in the final r value. On average, all estimators perform better when the rank is conducted on Savage scores ($\delta = 2$), i.e. when the focus is on ranking the most important model inputs only (Figs. S8–S15). As for the distribution of the model inputs (ϕ), it has a first-order effect in the accuracy of Azzini and Rosati ($\approx$ 10%), Jansen and Janon / Monod ($\approx$ 15%) and Razavi and Gupta ($\approx$ 20%) regardless of whether the aim is a factor prioritization (r) or approaching the "true" indices (MAE). The performance of these estimators drops perceptibly when the model inputs are distributed as $Beta(8, 2)$ or $Beta(2, 8)$ ($\phi = 3$ and $\phi = 4$, Figs. S8-S23), suggesting that

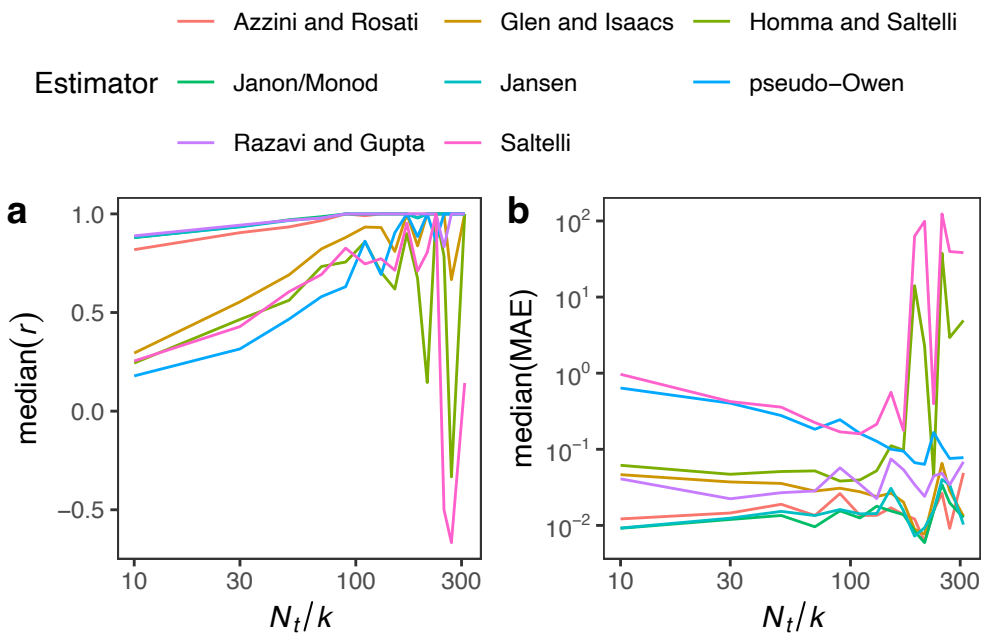

Figure 5: Scatterplot of the model output r against the number of model runs allocated per model input (Nt/k). See Fig. S6 for a visual display of all simulations and Fig. S7 for an assessment of the number of model runs that each estimator has in each N_t/k compartment.

they may be especially stressed by skewed distributions. The selection of random or quasi-random numbers during the construction of the sample matrix (τ) also directly conditions the accuracy of several estimators. If the aim is to approach the "true" indices (MAE), τ conveys from 17% (Azzini and Rosati) to $\approx$ 30% (Glen and Isaacs) of the model output variance, with all estimators except Razavi and Gupta performing better on quasi-random numbers ($\tau = 2$, Figs. S16–S23). In a factor prioritization setting, τ is mostly influential through interactions. Interestingly, the proportion of active second and third-order interactions (k_2, k_3) does not alter the performance of any estimator in any of the settings examined.

To better understand the structure of the sensitivities, we compute Sobol' indices after grouping individual parameters in three clusters, which we define based on their commonalities: the first group includes (δ, τ) and reflects the influence of those parameters that can be defined by the sensitivity analyst during the setting of the benchmark exercise. The second combines ($\varepsilon, k_2, k_3, \phi$) and examines 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 includes (N_t, k) only and assesses the influence of the sampling design in the accuracy of the estimators (we assume 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 $\boldsymbol{B}$, $\boldsymbol{A}_B^{(i)}$), $\boldsymbol{B}_A^{(i)}$ or $\boldsymbol{C}_B^{(i)}$ matrices) (Fig 6b).

The uncertainty in the functional form of the model [$f(x)$] is responsible for approximately 20% of the variance in the performance of Azzini and Rosati, Janon/Monod or Jansen in a factor prioritization setting. For Glen and Isaacs, Homma and Saltelli, pseudo-Owen or Saltelli, $f(x)$ is influential only through interactions with the other clusters. When the MAE is the performance measure of interest, $f(x)$ has a

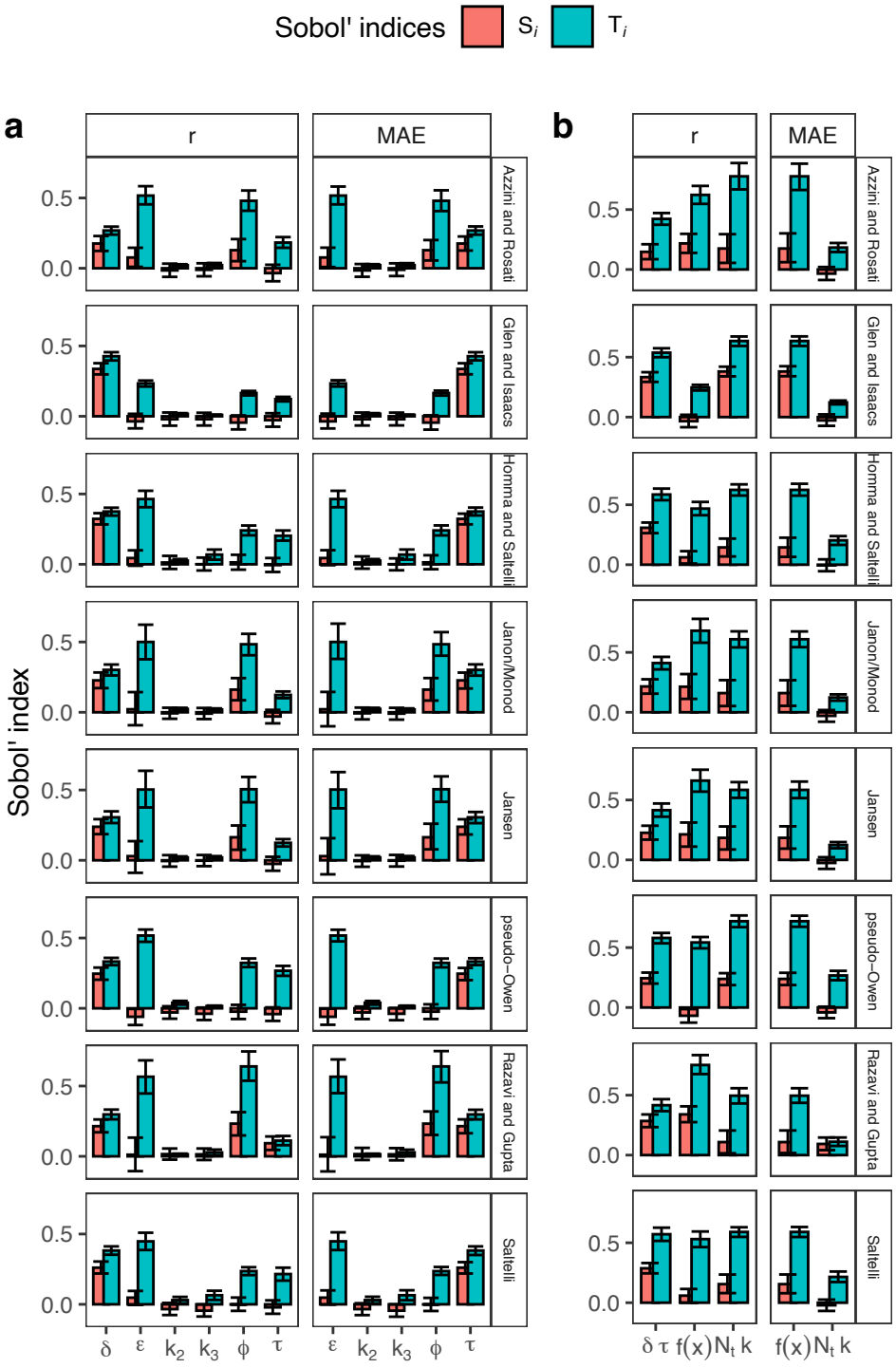

Figure 6: Sobol' indices. a) Individual parameters. b) Clusters of parameters. The cluster $f(x)$ includes all parameters that describe the uncertainty in the functional form of the model (ϵ, k_2, k_3, ϕ). N_t and k are assessed simultaneously due to their correlation. Note that the MAE facet does not include the group ($\delta\tau$) because δ (the performance measure used) is no longer an uncertain parameter in this setting.

much stronger influence in the accuracy of the estimators than the couple (N_t, k), especially in the case of Glen and Isaacs ($\approx$ 40%). In any case, the accuracy of the estimators is significantly conditioned by interactions between the benchmark parameters. The sum of all individual S_i indices plus the S_i index

of the (N_t, k) cluster only explains from $\approx$ 45% (Saltelli) to $\approx$ 70% (Glen and Isaacs) of the estimators' variance in ranking the model inputs, and from $\approx$ 24% (pseudo-Owen) to $\approx$ 60% (Razavi and Gupta) of the variance in approaching the "true" indices.

4 Discussion and conclusions

Here we design an eight-dimension background for variance-based total-order estimators to confront and prove their value in an unparalleled range of SA scenarios. By randomizing the parameters that condition their performance, we obtain a comprehensive picture of the advantages and disadvantages of each estimator and identify which particular benchmark factors make them more prone to error. Our work thus provides a thorough empirical assessment of state-of-the-art total-order estimators and contributes to define best practices in variance-based SA. 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 a sensitivity analysis* (SA of SA) [42].

Our results provide support to the assumption that the scope of previous benchmark studies is limited by the plethora of non-unique choices taken during the setting of the analysis [21]. We have observed that almost all decisions have a non-negligible effect: from the selection of the sampling method to the choice of the performance measure, the design prioritized by the analyst can influence the performance of the estimator in a non-obvious way, namely 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 [14, 16, 18, 20, 23, 33, 39, 40, 42]). Such setting reduces the uncertain space to a minimum and misses the effects that the interactions between the benchmark parameters have in the final accuracy of the estimator. If global SA is the recommended practice to fully explore the uncertainty space of models, sensitivity estimators, being algorithms themselves, should be likewise validated [42].

Our approach also compensates the lack of studies on the theoretical properties of estimators in the sensitivity analysis literature (see for instance [15, 47]), and allows a more detailed examination of their performance than theoretical comparisons. Empirical studies like ours mirror the numerical character of sensitivity analysis when the indices can not be analytically calculated, which is most of the time in "real-world" mathematical modeling.

Two recommendations emerge from our work: the estimators by Razavi and Gupta, Jansen, Janon/Monod or Azzini and Rosati should be preferred when the aim is to rank the model inputs. Jansen, Janon/Monod or Azzini and Rosati should also be prioritized if the goal is to estimate the "true" total-order indices. The drop in performance of Razavi and Gupta in the second setting may be explained by a bias at a lower sample sizes, i.e. a consistent over-estimation of all total-order indices. This is because their estimator relies on a constant mean assumption whose validity degrades with larger values of Δh [20, 39]. In order to remove this bias, Δh should take very small values (e.g., $\Delta h = 0.01$), which may not be computationally feasible.

Since the direction of this bias is the same for all parameters it only affects the calculation of the "true" total-order indices, not the capacity of the estimator to properly rank the model inputs.

It is also worth stating that Razavi and Gupta is the only estimator studied here that require the analyst to define a 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. $\Delta h = 0.002$, $\Delta h = 0.1$, $\Delta h = 0.3$; [20, 21, 39]). Selecting the most appropriate value for a given tuning parameter is not an obvious choice and this uncertainty can make an estimator volatile, as shown by Puy et al. [42] in the case of the PAWN index.

The fact that Glen and Isaacs, Homma and Saltelli, Saltelli and pseudo-Owen do not perform as well in properly ranking the model inputs and approaching the "true" total-order indices may be partially explained by their less efficient computation of elementary effects: by allowing the production of negative terms in the numerator these estimators also permit the production of negative total-order indices, thus leading to biased rankings or sensitivity indices. In the case of Saltelli, the use of a $\boldsymbol{B}$ matrix at the numerator and an $\boldsymbol{A}$ matrix at the denominator exacerbates its volatility (Table 1, Nž 5). Such inconsistency was corrected in Saltelli et al. [23].

The consistent robustness of Jansen, Janon/Monod and Azzini and Rosati makes their sensitivity to the uncertain parameters studied here almost negligible. They are already highly optimized estimators with not much room for improvement. 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. 6b), as well as by their sampling design (N_t, k), which are in any case beyond the analyst's control. 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, Saltelli 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; [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 [21]'s, who observed that Jansen outperformed Janon/Monod under a factor prioritization setting. We did not find any significant difference between these estimators. Although our metafunction approach is based on Becker [21]'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 may account for the slightly different results obtained between the two papers.

Our analysis can be extended to other sensitivity estimators (i.e. moment-independent like entropy-based [48]; the δ-measure [49]; or the PAWN index, [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 [21]'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. [22]. 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 Code availability

The `R` code to replicate our results is available in Puy [52] and in GitHub (https://github.com/arnaldpuy/battle_estimators). The uncertainty and sensitivity analysis have been carried out with the `R` package `sensobol` [53], which also includes the test function used in this study.

6 Acknowledgements

We thank Saman Razavi for his insights on the behavior of the Razavi and Gupta estimator. This work has been funded by the European Commission (Marie Skłodowska-Curie Global Fellowship, grant number 792178 to A.P.).

References

[1] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis. The Primer*. Chichester, UK: John Wiley & Sons, Ltd, Dec. 2008. DOI: 10.1002/9780470725184.

[2] A. Jakeman, R. Letcher, and J. Norton. "Ten iterative steps in development and evaluation of environmental models". *Environmental Modelling & Software* 21.5 (May 2006), 602–614. DOI: 10.1016/j.envsoft.2006.01.004.

[3] S. Eker, E. Rovenskaya, M. Obersteiner, and S. Langan. "Practice and perspectives in the validation of resource management models". *Nature Communications* 9.1 (2018), 1–10. DOI: 10.1038/s41467-018-07811-9.

[4] A. Saltelli. "Sensitivity analysis for importance assessment". *Risk Analysis* 22.3 (June 2002), 579–590. DOI: 10.1111/0272-4332.00040.

[5] B. Iooss and P. Lemaître. "A Review on Global Sensitivity Analysis Methods". *Uncertainty Management in Simulation-Optimization of Complex Systems. Operations Research/Computer Science Interfaces Series, vol 59*. Ed. by G. Dellino and C. Meloni. Boston: Springer, 2015, 101–122. DOI: 10.1007/978-1-4899-7547-8_5. arXiv: 1404.2405.

[6] W. Becker and A. Saltelli. "Design for sensitivity analysis". *Handbook of Design and Analysis of Experiments*. Ed. by A. Dean, M. Morris, J. Stufken, and D. Bingham. Boca Ratón: CRC Press, Taylor & Francis, 2015, 627–674. DOI: 10.1201/b18619.

[7] 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.

[8] L. Le Gratiet, S. Marelli, and B. Sudret. “Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes”. *Handbook of Uncertainty Quantification*. Cham: Springer International Publishing, 2017, 1289–1325. DOI: 10.1007/978-3-319-12385-1_38.

[9] A. Saltelli, S. Tarantola, and K. P.-S. Chan. “A quantitative model-independent method for global sensitivity analysis of model output”. *Technometrics* 41.1 (Feb. 1999), 39. DOI: 10.2307/1270993.

[10] R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. “Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory”. *The Journal of chemical physics* 59.8 (1973), 3873–3878.

[11] R. I. Cukier, H. B. Levine, and K. E. Shuler. “Nonlinear sensitivity analysis of multiparameter model systems”. *Journal of computational physics* 26.1 (1978), 1–42.

[12] I. M. Sobol’. “On the distribution of points in a cube and the approximate evaluation of integrals”. *USSR Computational Mathematics and Mathematical Physics* 7.4 (Jan. 1967), 86–112. DOI: 10.1016/0041-5553(67)90144-9.

[13] I. M. Sobol’. “Uniformly distributed sequences with an additional uniform property”. *USSR Computational Mathematics and Mathematical Physics* 16.5 (Jan. 1976), 236–242. DOI: 10.1016/0041-5553(76)90154-3.

[14] S. Lo Piano, F. Ferretti, A. Puy, D. Albrecht, and A. Saltelli. “Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy”. *Reliability Engineering & System Safety* 206.October 2020 (Feb. 2021), 107300. DOI: 10.1016/j.ress.2020.107300.

[15] 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.

[16] 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.

[17] 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.

[18] I. Azzini, T. Mara, and R. Rosati. “Monte Carlo estimators of first-and total-orders Sobol’ indices” (2020). arXiv: 2006.08232.

[19] H. Monod, C. Naud, and D. Makowski. *Uncertainty and sensitivity analysis for crop models*. Ed. by D. Wallach, D. Makowski, and J. W. Jones. Elsevier, 2006, 35–100.

[20] S. Razavi and H. V. Gupta. “A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application”. *Water Resources Research* 52.1 (Jan. 2016), 440–455. DOI: 10.1002/2015WR017558. arXiv: 2014WR016527 [10.1002].

[21] W. Becker. "Metafunctions for benchmarking in sensitivity analysis". *Reliability Engineering and System Safety* 204 (2020), 107189. DOI: 10.1016/j.ress.2020.107189.

[22] S. Kucherenko, B. Feil, N. Shah, and W. Mauntz. "The identification of model effective dimensions using global sensitivity analysis". *Reliability Engineering & System Safety* 96.4 (Apr. 2011), 440–449. DOI: 10.1016/j.ress.2010.11.003.

[23] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". *Computer Physics Communications* 181.2 (Feb. 2010), 259–270. DOI: 10.1016/j.cpc.2009.09.018.

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

[25] I. M. Sobol'. "On quasi-Monte Carlo integrations". *Mathematics and Computers in Simulation* 47.2-5 (Aug. 1998), 103–112. DOI: 10.1016/S0378-4754(98)00096-2.

[26] P. Bratley and B. L. Fox. "ALGORITHM 659: implementing Sobol's quasirandom sequence generator". *ACM Transactions on Mathematical Software (TOMS)* 14.1 (1988), 88–100.

[27] E. Borgonovo and E. Plischke. "Sensitivity analysis: A review of recent advances". *European Journal of Operational Research* 248.3 (2016), 869–887. DOI: 10.1016/j.ejor.2015.06.032.

[28] A. Saltelli. "A short comment on statistical versus mathematical modelling". *Nature Communications* 10.1 (2019), 8–10. DOI: 10.1038/s41467-019-11865-8.

[29] A. Saltelli, G. Bammer, I. Bruno, E. Charters, M. Di Fiore, E. Didier, W. Nelson Espeland, J. Kay, S. Lo Piano, D. Mayo, R. Pielke Jr, T. Portaluri, T. M. Porter, A. Puy, I. Rafols, J. R. Ravetz, E. Reinert, D. Sarewitz, P. B. Stark, A. Stirling, J. van der Sluijs, and P. Vineis. "Five ways to ensure that models serve society: a manifesto". *Nature* 582.7813 (June 2020), 482–484. DOI: 10.1038/d41586-020-01812-9.

[30] R. Sheikholeslami, S. Razavi, H. V. Gupta, W. Becker, and A. Haghnegahdar. "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.June 2018 (2019), 282–299. DOI: 10.1016/j.envsoft.2018.09.002.

[31] F. Sarrazin, F. Pianosi, and T. Wagener. "Global Sensitivity Analysis of environmental models: Convergence and validation". *Environmental Modelling and Software* 79 (2016), 135–152. DOI: 10.1016/j.envsoft.2016.02.005.

[32] A. Haghnegahdar, S. Razavi, F. Yassin, and H. Wheater. "Multicriteria sensitivity analysis as a diagnostic tool for understanding model behaviour and characterizing model uncertainty". *Hydrological Processes* 31.25 (2017), 4462–4476. DOI: 10.1002/hyp.11358.

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

[34] M. J. Shin, J. H. A. Guillaume, B. F. W. Croke, and A. J. Jakeman. "Addressing ten questions about conceptual rainfall-runoff models with global sensitivity analyses in R". *Journal of Hydrology* 503 (2013), 135–152. DOI: 10.1016/j.jhydrol.2013.08.047.

[35] L. Paleari and R. Confalonieri. "Sensitivity analysis of a sensitivity analysis: We are likely overlooking the impact of distributional assumptions". *Ecological Modelling* 340 (2016), 57–63. DOI: 10.1016/j.ecolmodel.2016.09.008.

[36] C. Spearman. "The proof and measurement of association between two things". *The American Journal of Psychology* 15.1 (Jan. 1904), 72. DOI: 10.2307/1412159.

[37] M. G. Kendall and B. B. Smith. "The problem of m rankings". *The Annals of Mathematical Statistics* 10.3 (Sept. 1939), 275–287. DOI: 10.1214/aoms/1177732186.

[38] I. R. Savage. "Contributions to the theory of rank order statistics - the two sample case". *Annals of Mathematical Statistics* 27 (1956), 590–615.

[39] 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), 423–439. DOI: 10.1002/2015WR017559.

[40] A. B. Owen. "Better estimation of small sobol' sensitivity indices". *ACM Transactions on Modeling and Computer Simulation* 23.2 (2013), 1–17. DOI: 10.1145/2457459.2457460.

[41] 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*. R package version 1.22.1, 2020.

[42] A. Puy, S. Lo Piano, and A. Saltelli. "A sensitivity analysis of the PAWN sensitivity index". *Environmental Modelling and Software* 127 (2020), 104679. DOI: 10.1016/j.envsoft.2020.104679. arXiv: 1904.04488.

[43] A. B. Owen. "Randomly permuted (t, m, s)-nets and (t, s)-sequences". *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statistics, vol. 106*. 1995, 299–317.

[44] G. E. P. Box, J. S. Hunter, and W. G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, 2005.

[45] V. Pareto. *Manuale di Economia Politica. Vol. 13*. Societa Editrice, 1906.

[46] G. E. P. Box and R. D. Meyer. "An analysis for unreplicated fractional factorials". *Technometrics* 28.1 (1986), 11–18. DOI: 10.1080/00401706.1986.10488093.

[47] I. Azzini, G. Listorti, T. A. Mara, and R. Rosati. “Uncertainty and Sensitivity Analysis for policy decision making. An introductory guide”. Luxembourg, 2020.

[48] H. Liu, W. Chen, and A. Sudjianto. “Relative entropy based method for probabilistic sensitivity analysis in engineering design”. *Journal of Mechanical Design, Transactions of the ASME* 128.2 (2006), 326–336. DOI: 10.1115/1.2159025.

[49] E. Borgonovo. “A new uncertainty importance measure”. *Reliability Engineering and System Safety* 92.6 (2007), 771–784. DOI: 10.1016/j.ress.2006.04.015.

[50] F. Pianosi and T. Wagener. “A simple and efficient method for global sensitivity analysis based on cumulative distribution functions”. *Environmental Modelling and Software* 67 (2015), 1–11. DOI: 10.1016/j.envsoft.2015.01.004.

[51] F. Pianosi and T. Wagener. “Distribution-based sensitivity analysis from a generic input-output sample”. *Environmental Modelling and Software* 108 (2018), 197–207. DOI: 10.1016/j.envsoft.2018.07.019.

[52] A. Puy. *R code of "A comprehensive comparison of total-order estimators for global sensitivity analysis"*. 2020. DOI: 10.5281/zenodo.4946559.

[53] A. Puy, S. L. Piano, A. Saltelli, and S. A. Levin. “sensobol: an R package to compute variance-based sensitivity indices” (Jan. 2021). arXiv: 2101.10103.

A comprehensive comparison of total-order estimators for global sensitivity analysis

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

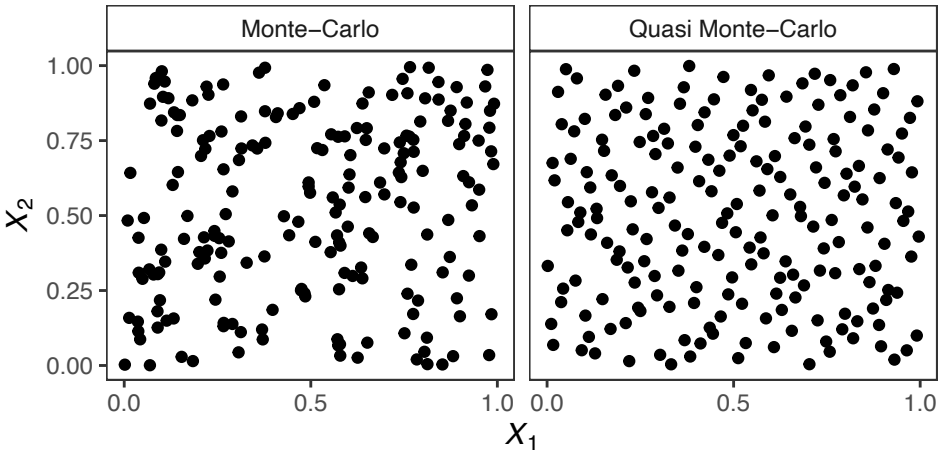

Figure S1: Examples of Monte-Carlo and Quasi Monte-Carlo sampling in two dimensions. $N = 200$.

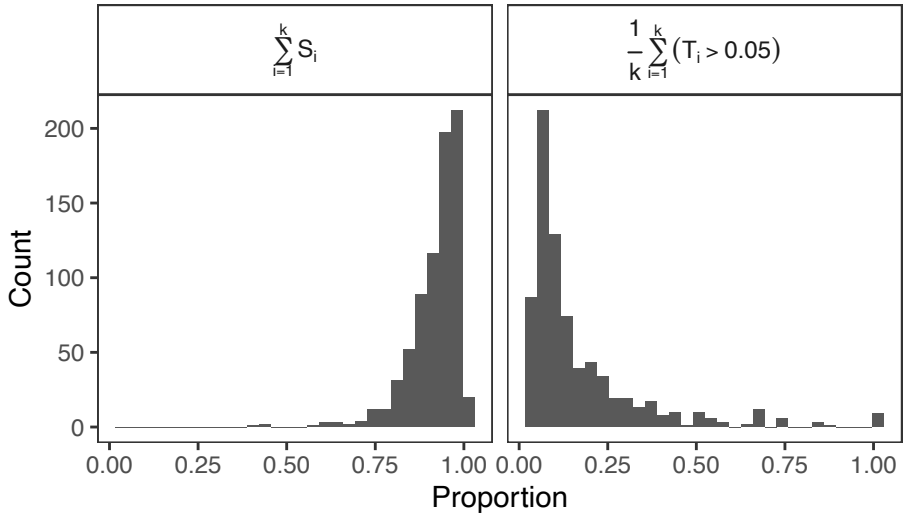

Figure S2: Proportion of the total sum of first-order effects and of the active model inputs (defined as $T_i > 0.05$) after 1000 random metafunction calls with $k \in (3, 100)$. Note how the sum of first-order effects clusters around 0.8 (thus evidencing the production of non-additivities) and how,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

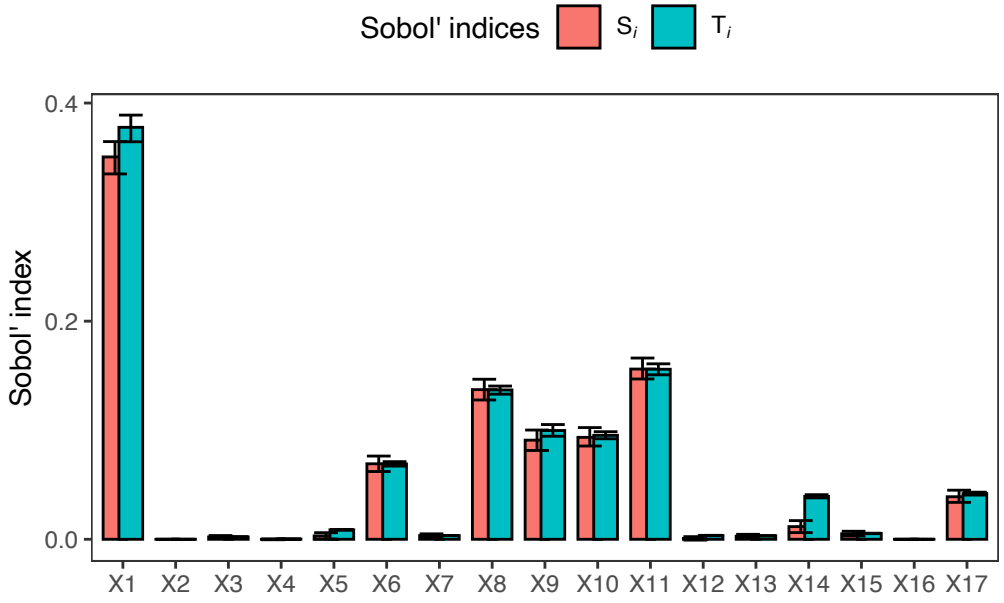

Figure S3: Sobol' T_i indices obtained after a run of the metafunction with the following parameter settings: $N = 10^4$, $k = 17$, $k_2 = 0.5$, $k_3 = 0.2$, $\varepsilon = 666$. The error bars reflect the 95% confidence intervals after bootstrapping ($R = 10^2$). The indices have been computed with the Jansen [3] estimator.

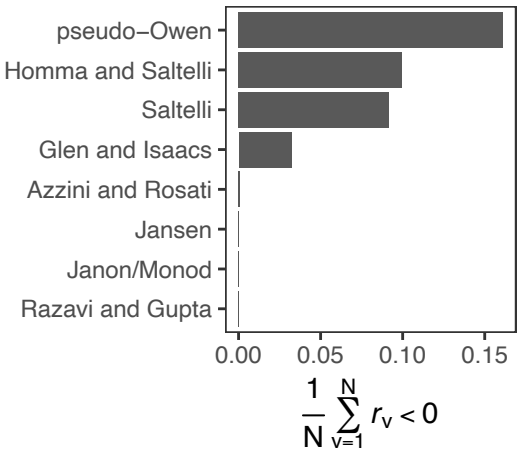

Figure S4: Proportion of model runs yielding $r < 0$.

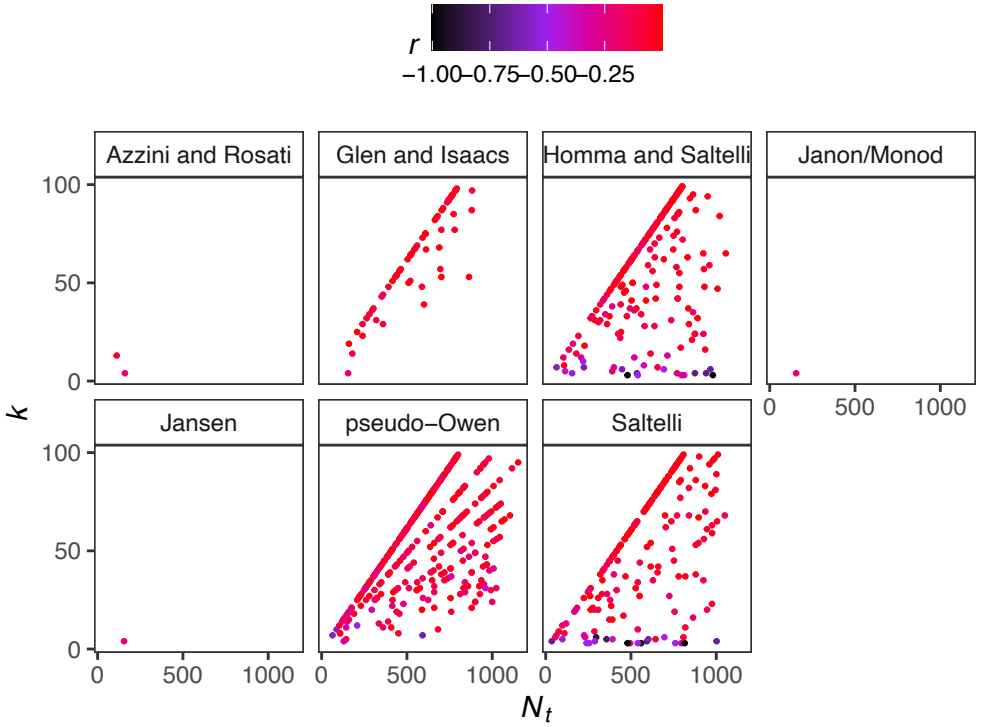

Figure S5: Scatter of the total number of model runs N_t against the function dimensionality k only for $r < 0$.

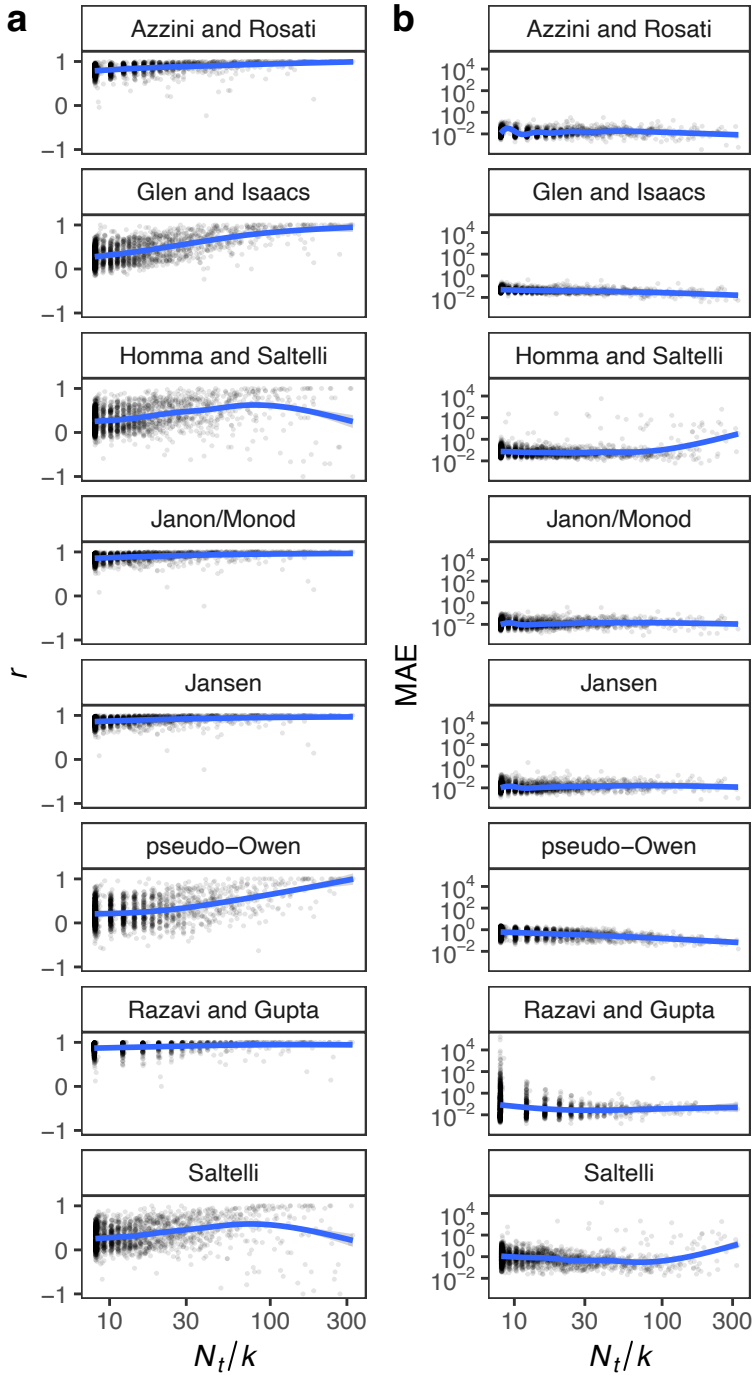

Figure S6: Scatterplot of the correlation between $\boldsymbol{T}_i$ and $\hat{\boldsymbol{T}}_i$ (r) against the number of model runs allocated per model input (Nt/k).

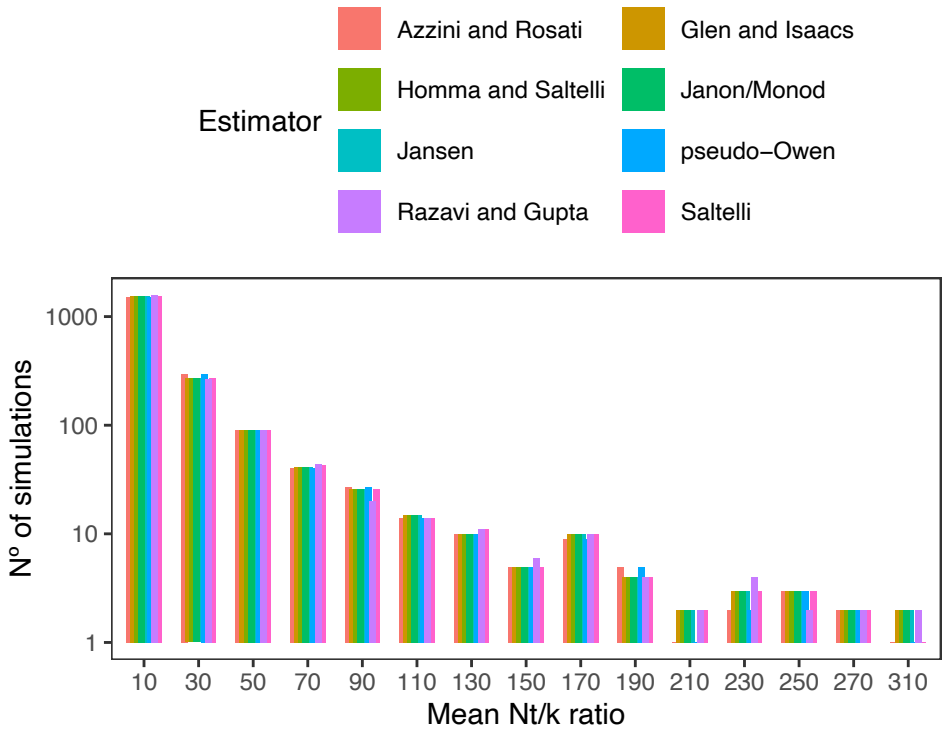

Figure S7: Bar plot with the number of simulations conducted in each of the N_t/k compartments assessed. All estimators have approximately the same number of simulations in each compartment.

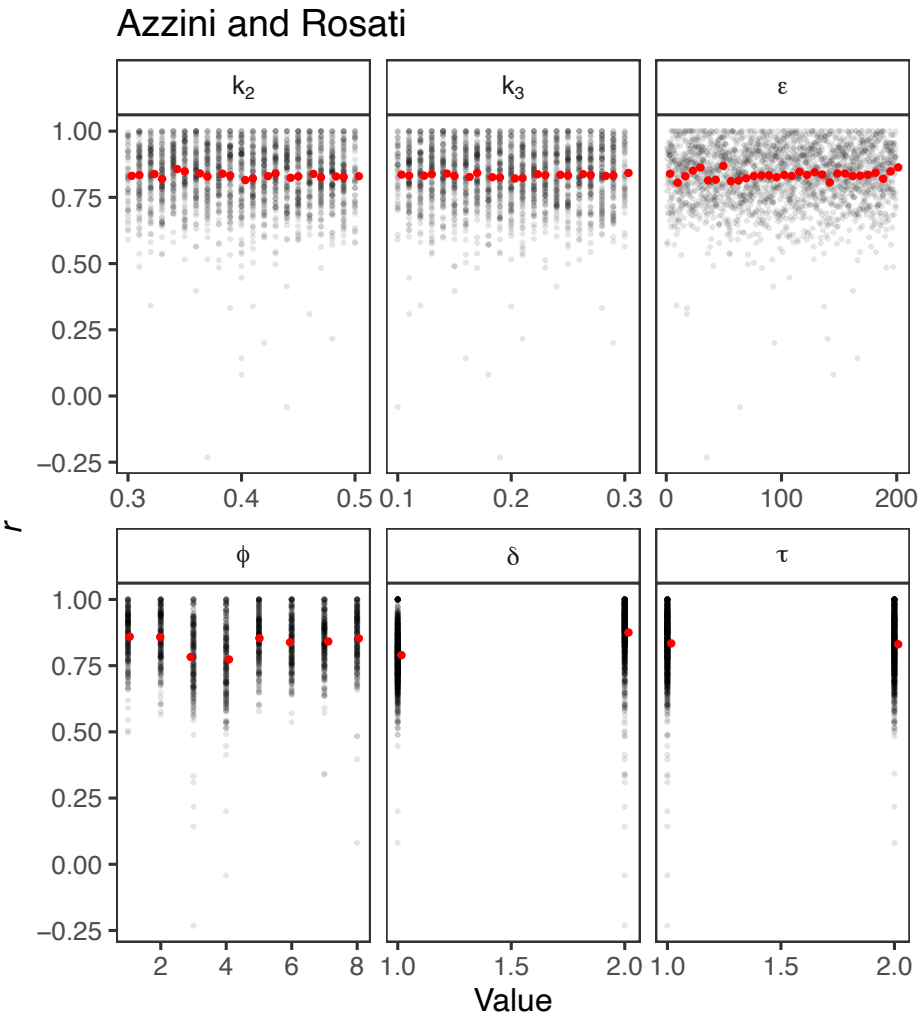

Figure S8: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

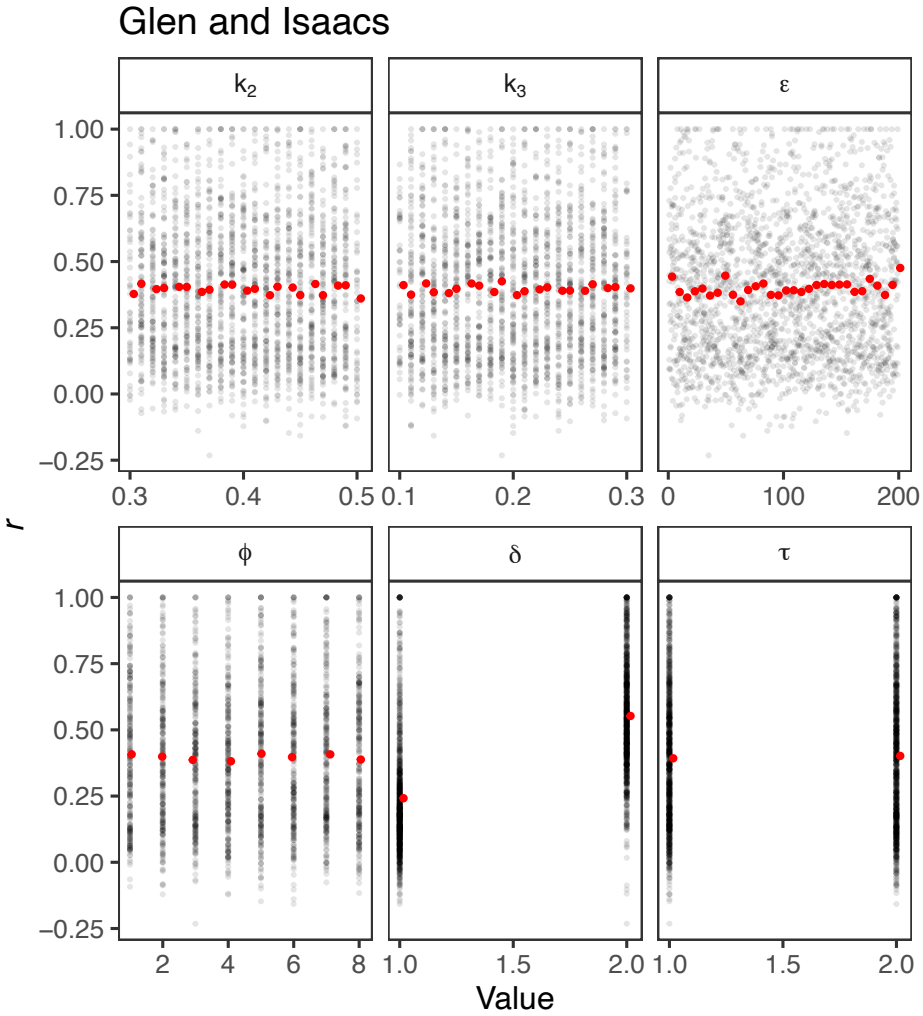

Figure S9: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

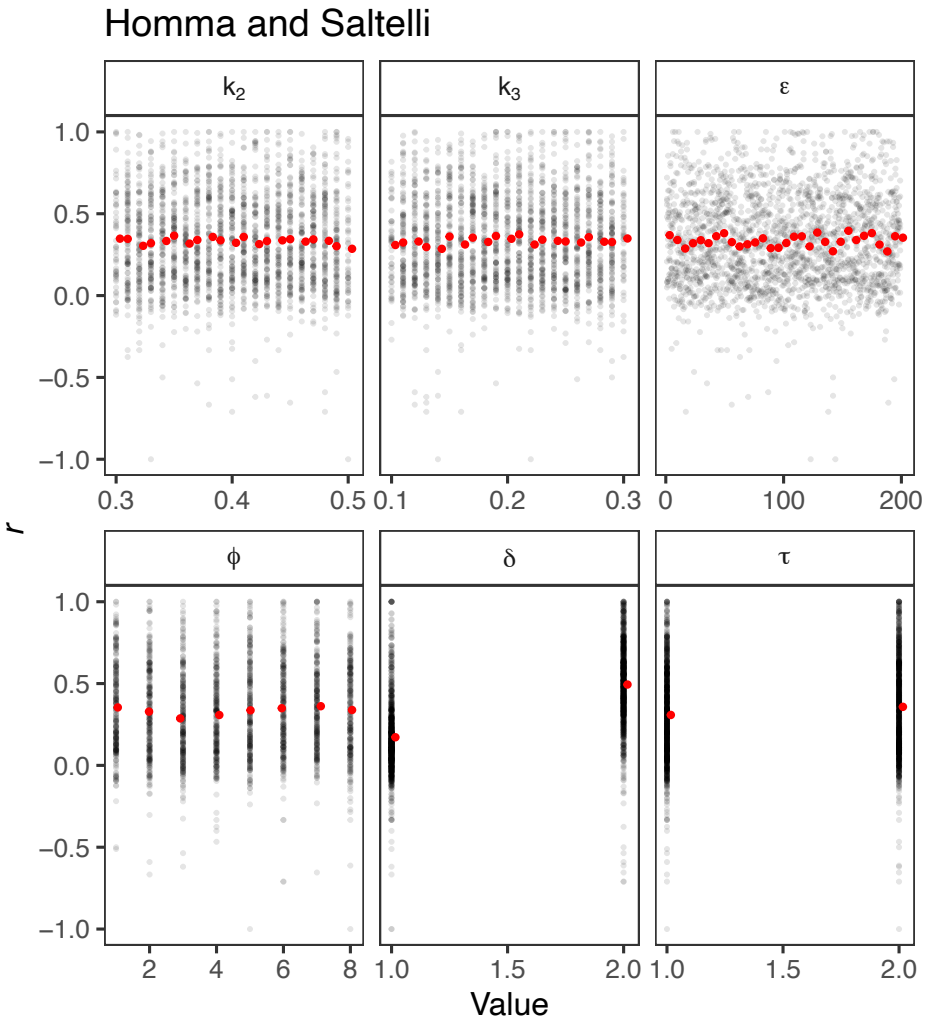

Figure S10: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

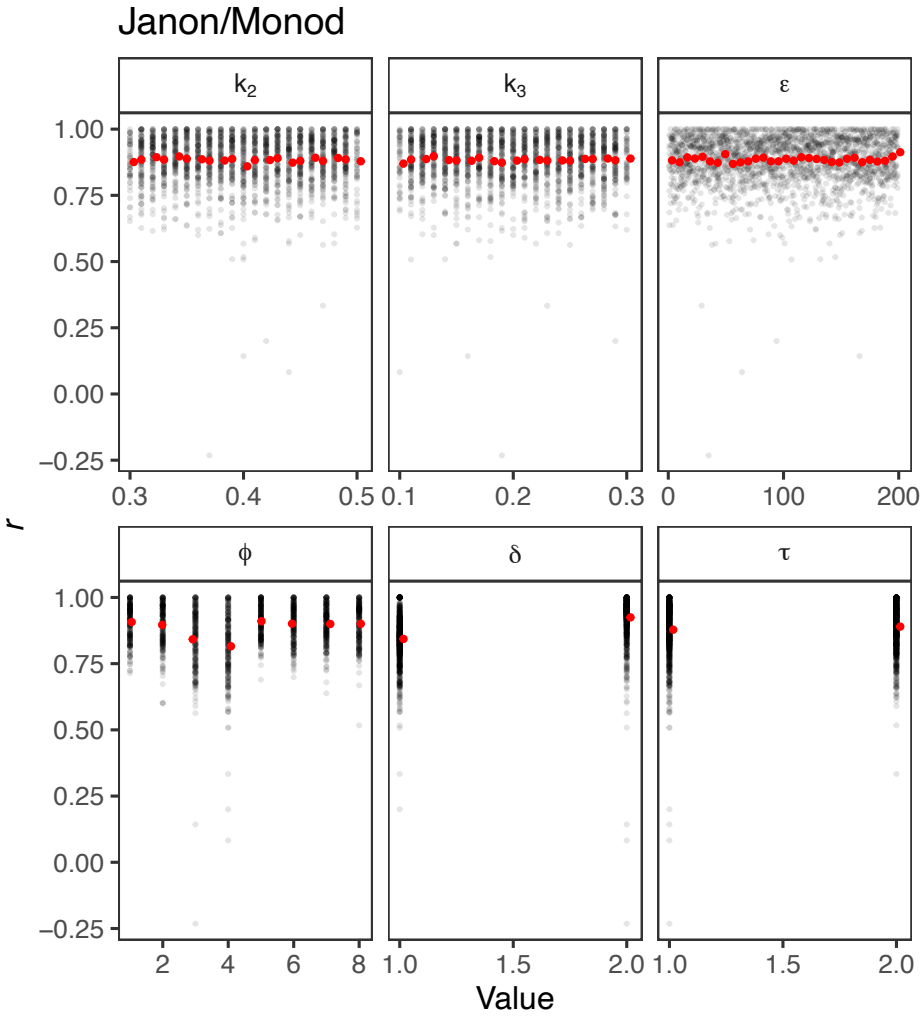

Figure S11: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

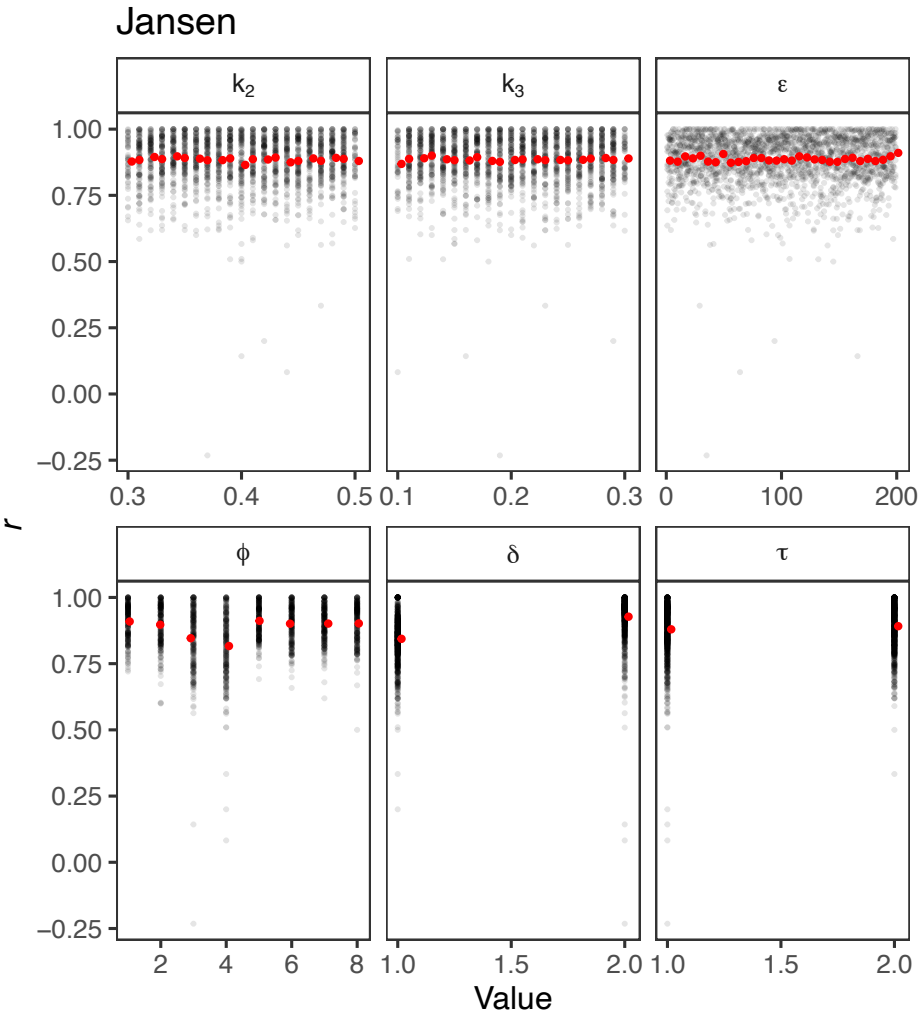

Figure S12: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

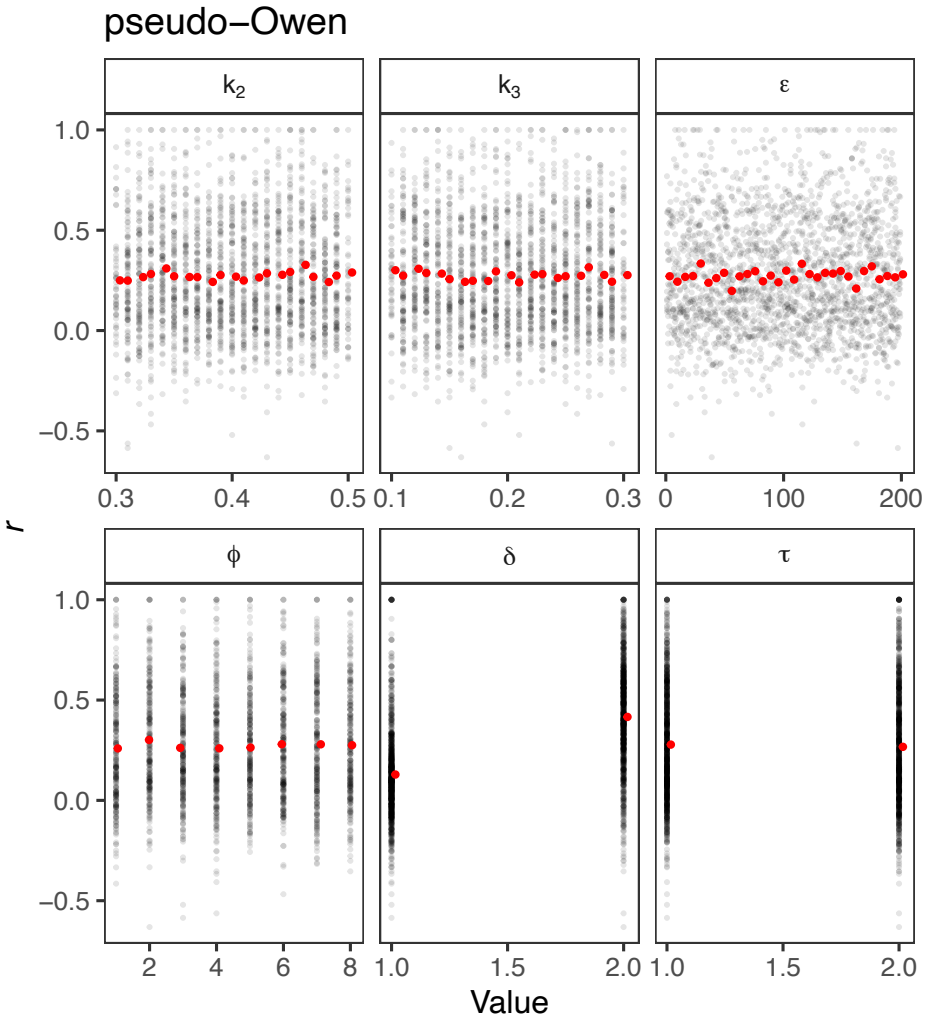

Figure S13: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

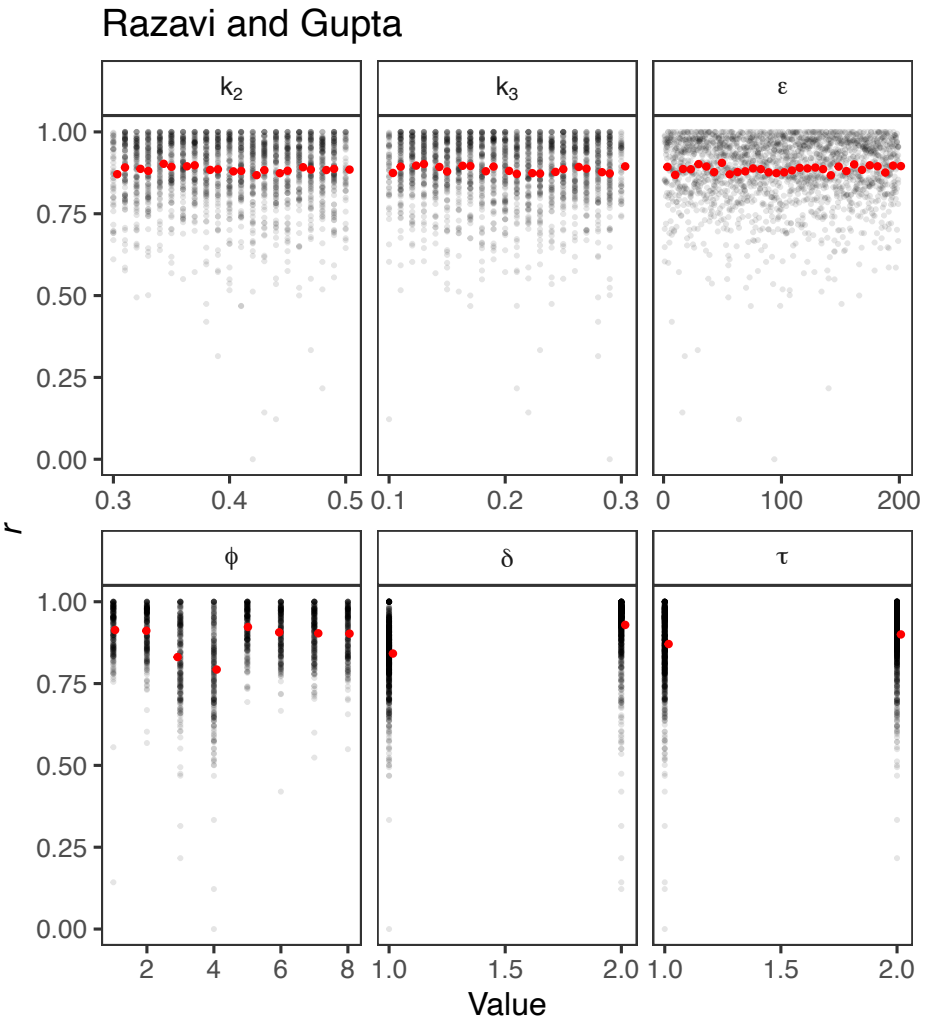

Figure S14: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

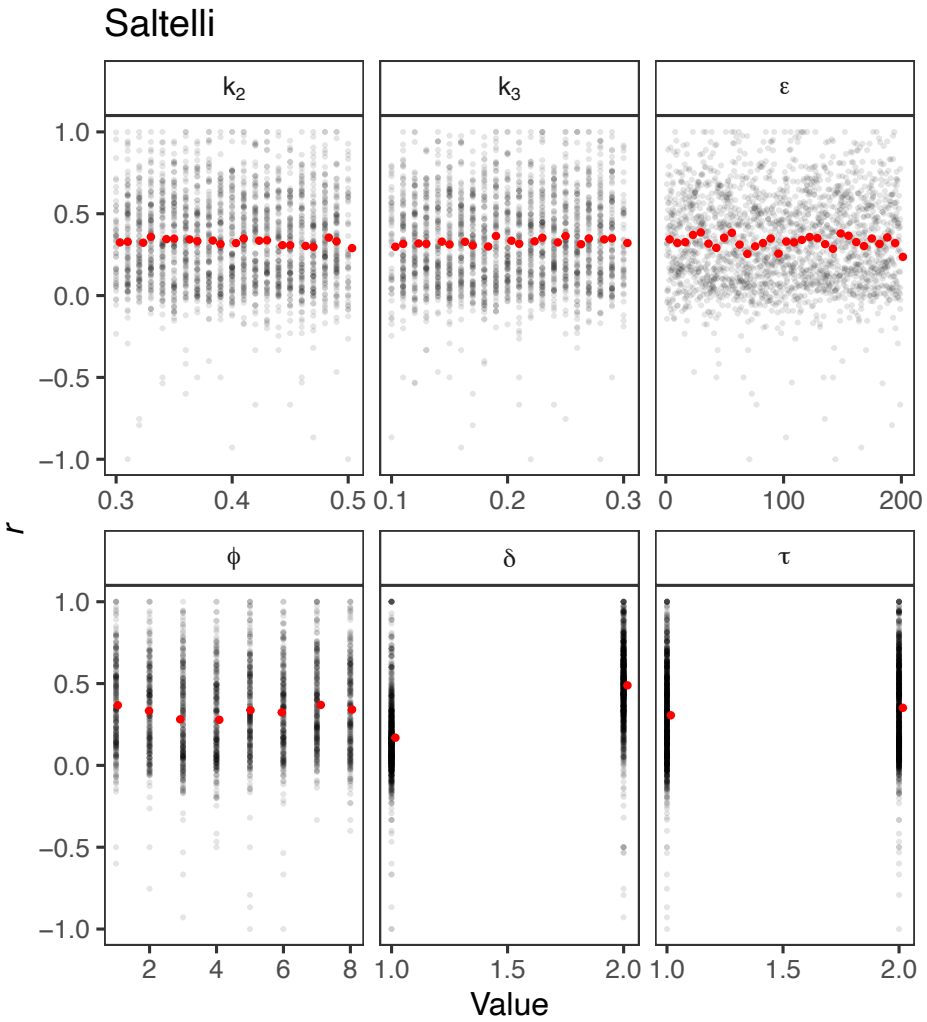

Figure S15: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

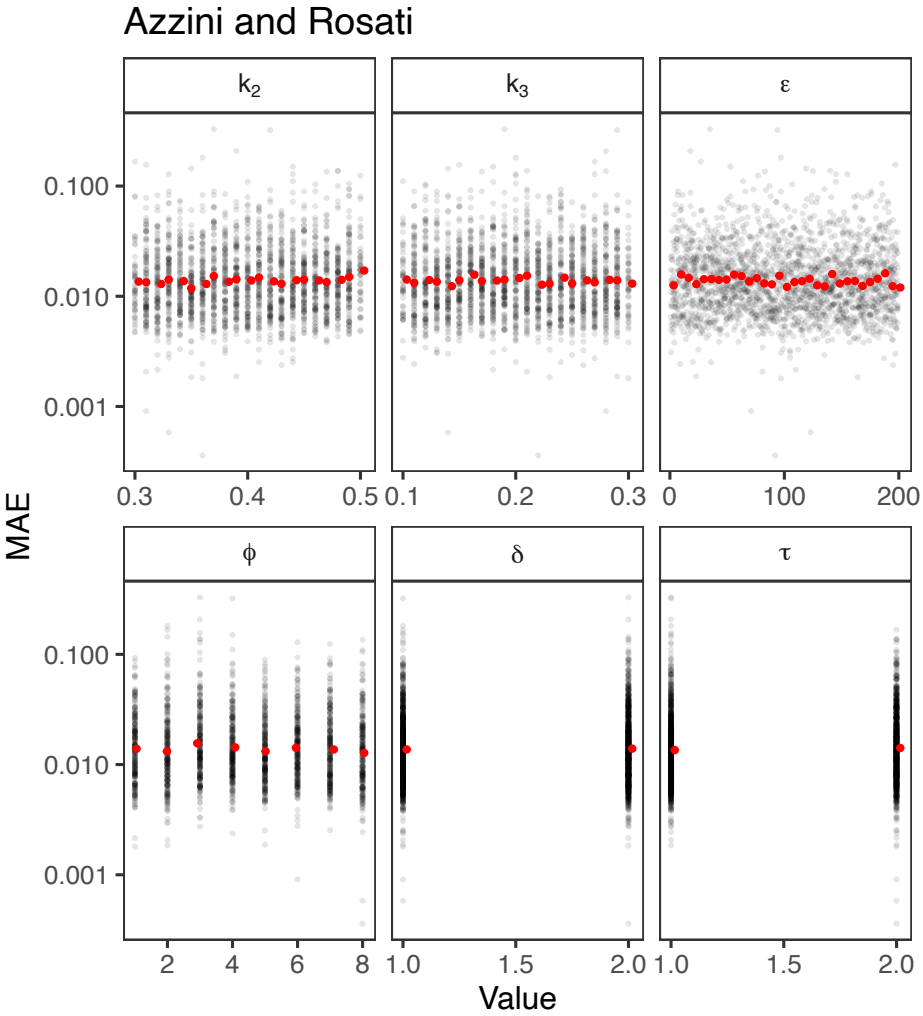

Figure S16: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

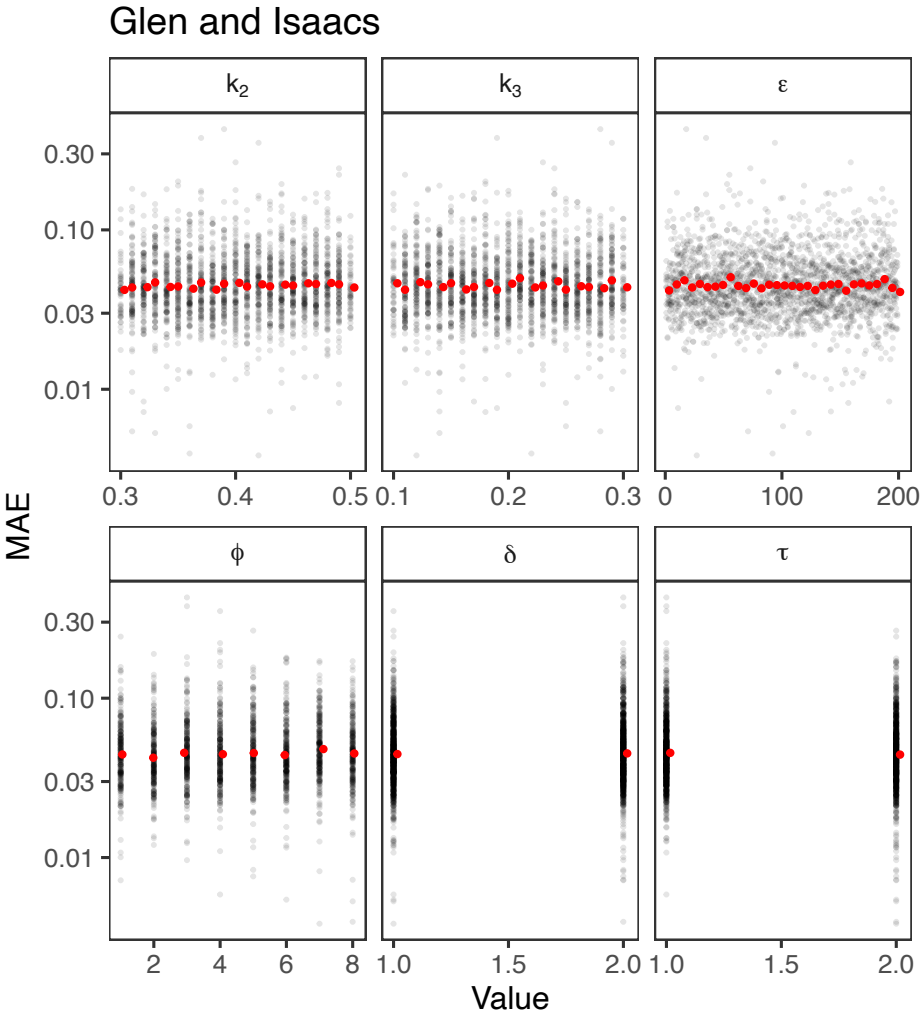

Figure S17: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

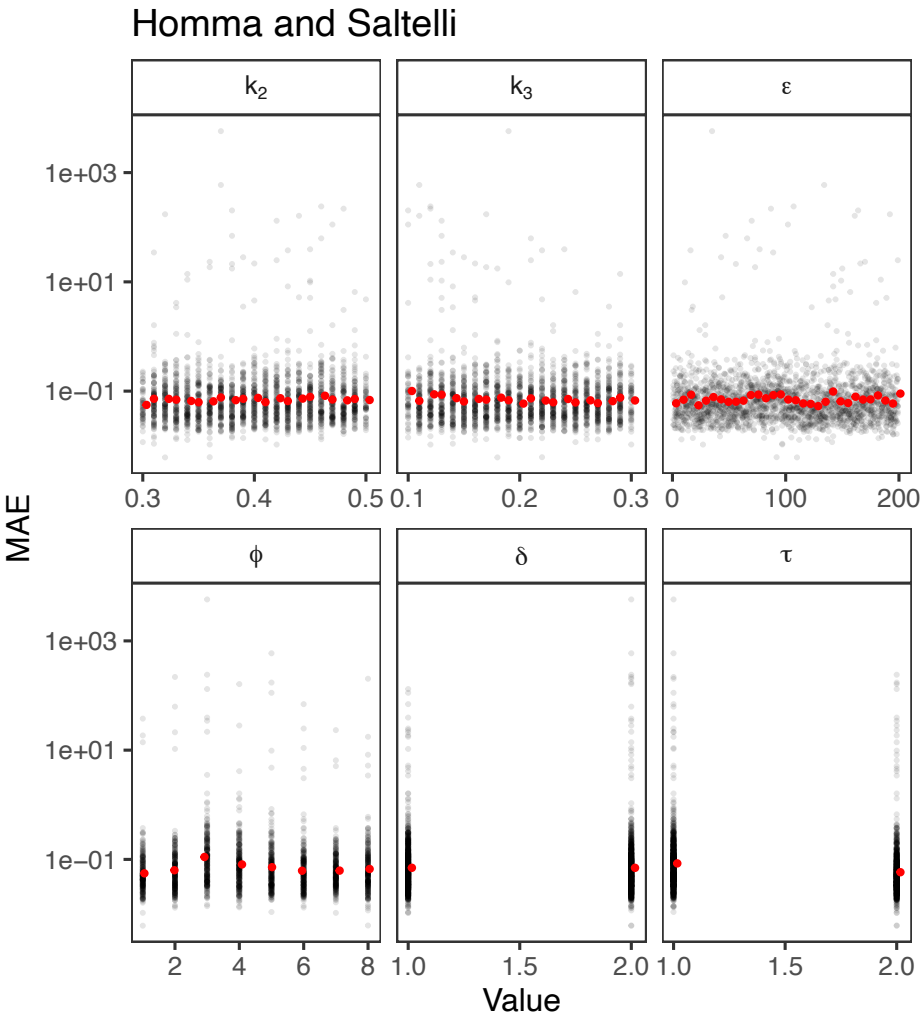

Figure S18: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

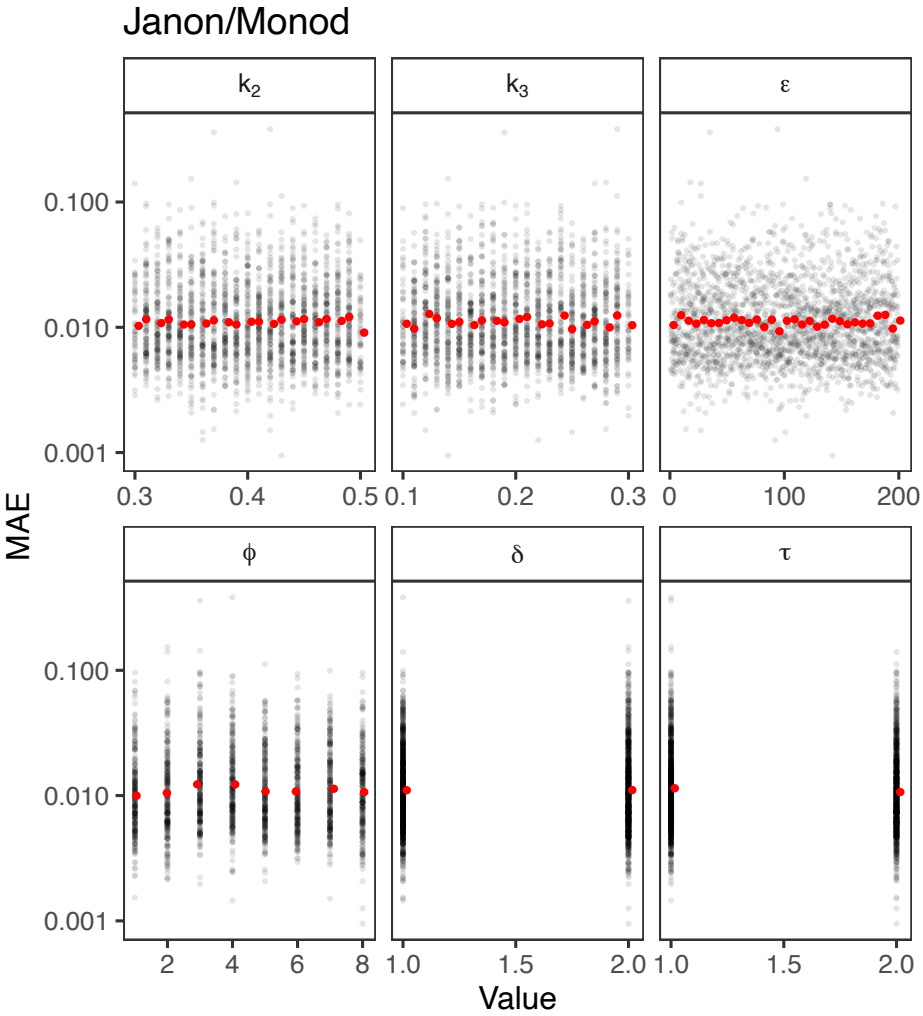

Figure S19: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

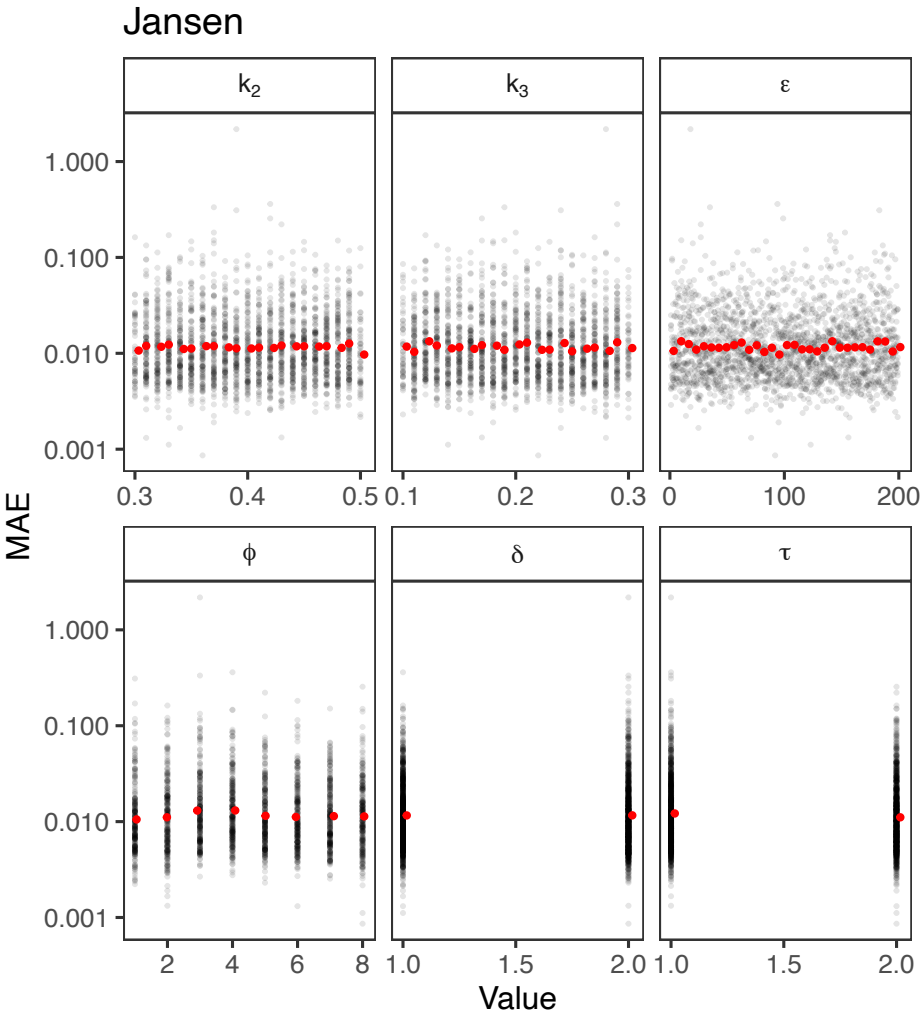

Figure S20: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

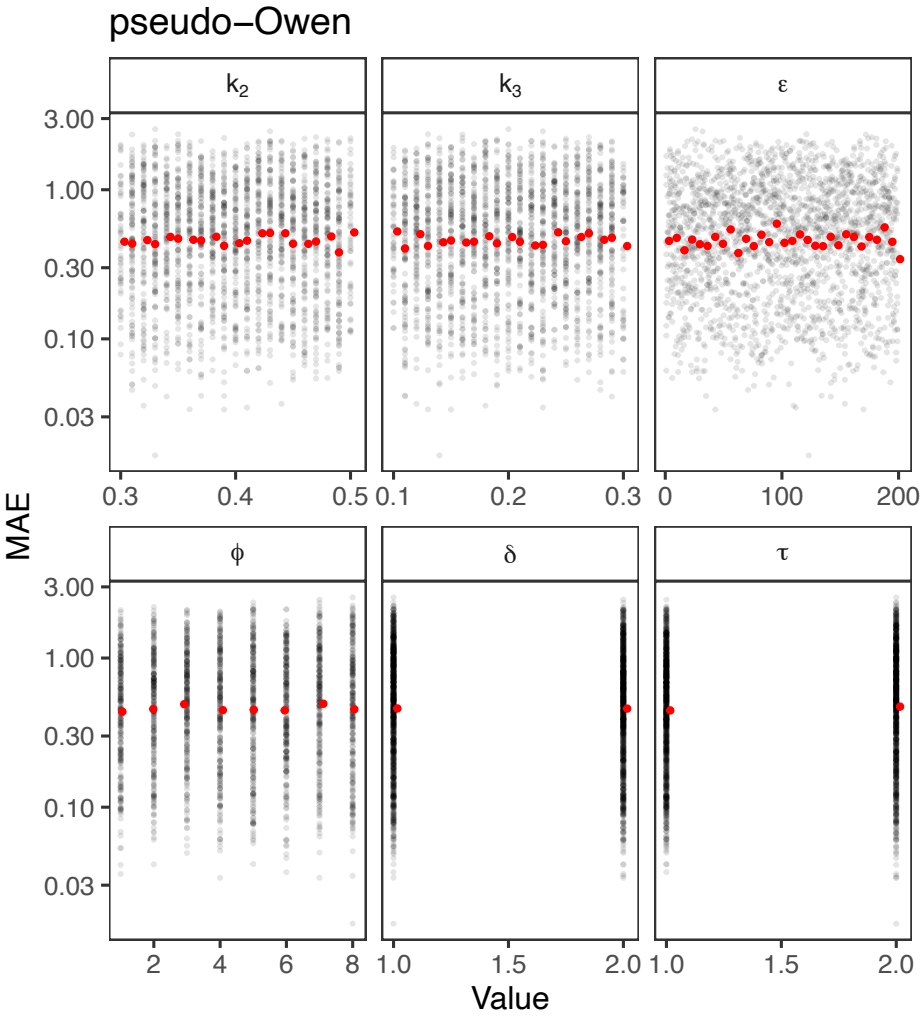

Figure S21: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

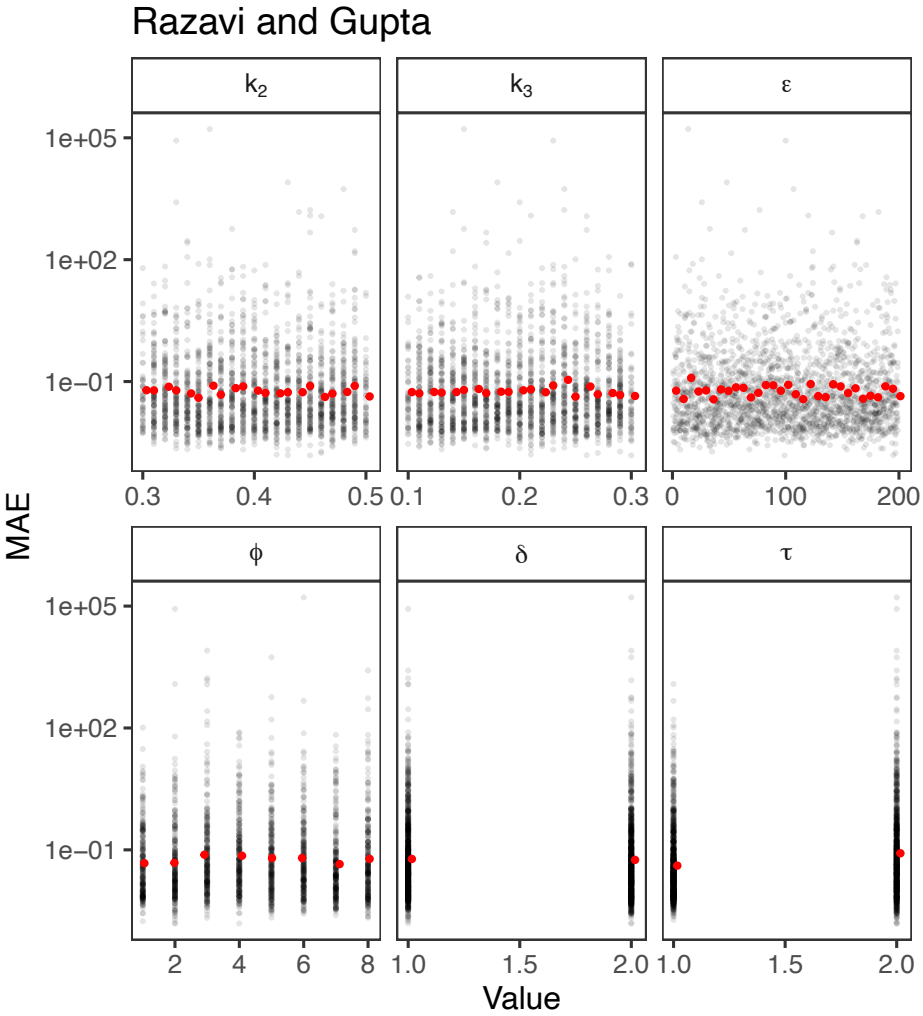

Figure S22: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

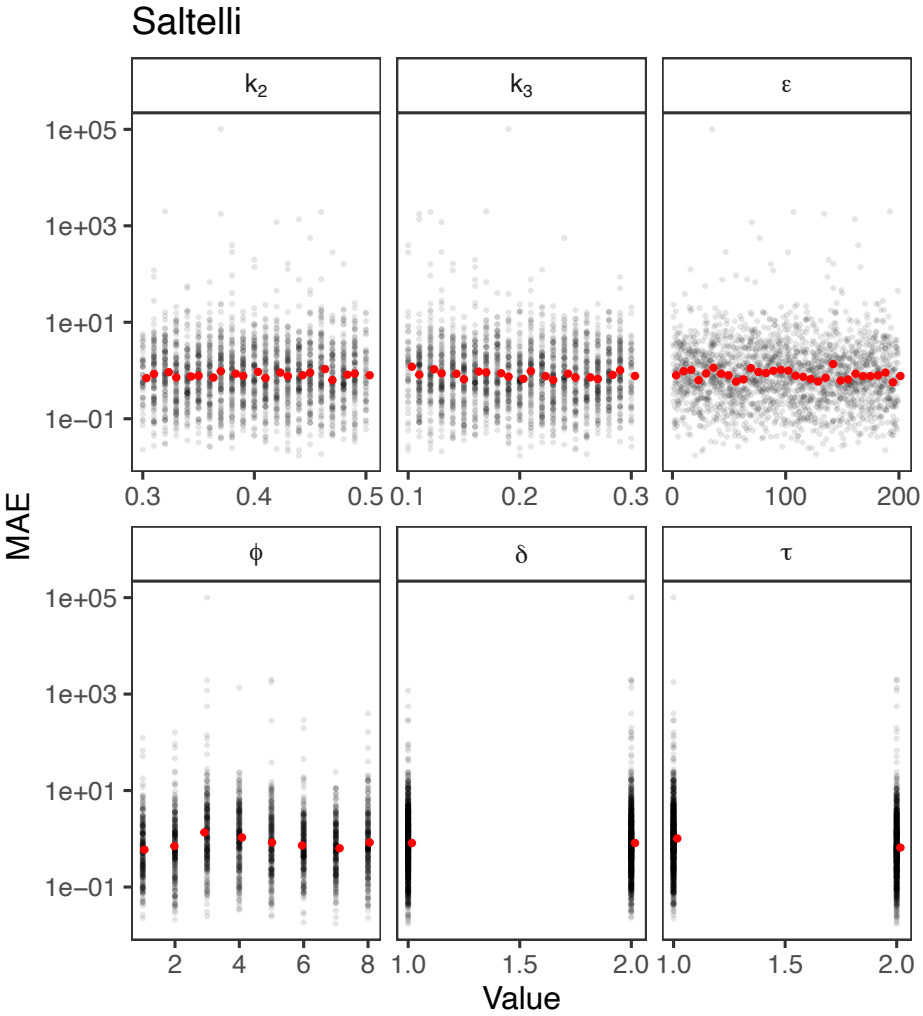

Figure S23: Scatterplots of the model inputs against the model output. The red dots show the mean value in each bin (we have set the number of bins arbitrarily at 30).

References

[1] S. Razavi and H. V. Gupta. "A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application". *Water Resources Research* 52.1 (Jan. 2016), 440–455. DOI: 10.1002/2015WR017558. arXiv: 2014WR016527 [10.1002].

[2] 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), 423–439. DOI: 10.1002/2015WR017559.

[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.