Environmental Modelling and Software 137 (2021) 104960

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

Is VARS more intuitive and efficient than Sobol' indices?

Arnald Puy [a,b,*], Samuele Lo Piano [c], Andrea Saltelli [d]

[a] Department of Ecology and Evolutionary Biology, Princeton University, M31 Guyot Hall, New Jersey, 08544, USA
[b] Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Parkveien 9, PB 7805, 5020, Bergen, Norway
[c] School of the Built Environment, JJ Thomson Building, WhiteKnights Campus, University of Reading, Reading, RG6 6ED, United Kingdom
[d] Open Evidence, Universitat Oberta de Catalunya, Universitat Oberta de Catalunya, Edifici 22@, Barcelona, 08018, Spain

ARTICLE INFO

Keywords:
Uncertainty
Sensitivity analysis
Modeling
Statistics
Design of experiment

ABSTRACT

The Variogram Analysis of Response Surfaces (VARS) has been proposed by Razavi and Gupta as a new comprehensive framework in sensitivity analysis. According to these authors, VARS provides a more intuitive notion of sensitivity and is much more computationally efficient than Sobol' indices. Here we review these arguments and critically compare the performance of VARS-TO, for total-order index, against the total-order Jansen estimator. We argue that, unlike classic variance-based methods, VARS lacks a clear definition of what an "important" factor is, and we show that the alleged computational superiority of VARS does not withstand scrutiny. We conclude that while VARS enriches the spectrum of existing methods for sensitivity analysis, especially for a diagnostic use of mathematical models, it complements rather than replaces classic estimators used in variance-based sensitivity analysis.

1. Introduction

Sensitivity analysis (SA) explores how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input space (Saltelli, 2002).[1] SA is especially needed when complex models, which often formalize partially known processes and include non-linear relations, are used to guide policies in the real world. This is generally the case of models in the Environmental Sciences domain, e.g. on crop water requirements, water availability under climate change, weather forecasting, surface runoff or precipitation and evaporation processes (Döll and Siebert, 2002; Pappenberger et al., 2011; Vieux and Vieux, 2016; Wang et al., 2020). The uncertainties in these models might be either parametric (i.e. exact values for parameters might be unknown, there might be errors in the measurement) or structural (i.e. lack of knowledge on the underlying processes, multiple ways of modeling the same phenomenon), and their combined effect on the model output should be understood to guarantee a robust inference for policy-making. In this context, SA jointly with uncertainty analysis is regarded as an unavoidable step to ensure the quality of the modeling process (Borgonovo and Plischke, 2016; Eker et al., 2018; Jakeman et al., 2006; Saltelli, 2019; Saltelli et al., 2020; Tarantola et al., 2002).

In SA, as in all fields of computational research, different strategies and methods compete to establish themselves as "good", "recommended" or "best" practices. While variance-based methods and Sobol' indices are deemed to belong to the class of recommended methods (Saltelli et al., 2008), other approaches have been proposed to complement or overcome their limitations, i.e. entropy-based methods (Liu et al., 2006), the δ measure (Borgonovo, 2007), the Kuiper' metric (Baucells and Borgonovo, 2013), or the PAWN index (Pianosi and Wagener, 2015, 2018). One of the most recent competitors is the Variogram Analysis of Response Surfaces (VARS), proposed by Razavi and Gupta (2016a, 2016b). According to *Google Scholar* and as of November 2020, the two foundational VARS papers have been cited 86 times, and seem to have been especially embraced by Hydrologists and Water Scientists (Jayathilake and Smith, 2020a, 2020b; Lilhare et al., 2020; Krogh et al., 2017).

Razavi and Gupta (2016a, 2016b) report that VARS outperforms Sobol' indices in two main aspects:

1. It provides a more intuitive assessment of sensitivities and the importance of model inputs in determining the model output.
2. It computes the total-order effect with a much higher computational efficiency (up to two orders of magnitude more efficient).

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

[1] This article is part of a SI on "Sensitivity analysis for environmental modeling."

https://doi.org/10.1016/j.envsoft.2021.104960
Accepted 30 December 2020
Available online 18 January 2021

In the present work we explore these results and benchmark VARS against one of the best Sobol' indices estimator, that of Jansen (1999). Before engaging in the discussion, we briefly recall hereafter some useful formulae needed to understand the two approaches.

1.1. Sobol' indices

The apparatus of variance-based sensitivity indices, described by Sobol', 1993 and extended by Homma and Saltelli (1996) is currently considered as the recommended practice in SA (Saltelli et al., 2008). For a model of k factors $f(\boldsymbol{x}) = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$, the first-order sensitivity index S_i can be written as

$$S_i = \frac{V_{x_i}[E_{\boldsymbol{x}_{\sim i}}(y|x_i)]}{V(y)} \,. \tag{1}$$

The inner mean in Equation (1) is taken over all-factors-but x_i ($\boldsymbol{x}_{\sim i}$), while the outer variance is taken over x_i. $V(y)$ is the unconditional variance of the output variable y. When the factors are independent, S_i can be defined as a first order term in the variance decomposition of y:

$$1 = \sum_{i=1}^{k} S_i + \sum_i \sum_{i<j} S_{ij} + \ldots + S_{1,2,\ldots,k} \,, \tag{2}$$

S_i lends itself to be expressed in plain English as *the fractional reduction in the variance of y which would be obtained on average if x_i could be fixed*. This is because

$$V(y) = V_{x_i}[E_{\boldsymbol{x}_{\sim i}}(y|x_i)] + E_{x_i}[V_{\boldsymbol{x}_{\sim i}}(y|x_i)] \,. \tag{3}$$

$E_{x_i}[V_{\boldsymbol{x}_{\sim i}}(y|x_i)]$ is the average variance that would be left after fixing x_i to a given value in its uncertainty range. For this reason, $V_{x_i}[E_{\boldsymbol{x}_{\sim i}}(y|x_i)]$ must be the average reduction in variance as discussed above. While $V_{\boldsymbol{x}_{\sim i}}(y|x_i)$ can be greater than $V(y)$, $E_{x_i}[V_{\boldsymbol{x}_{\sim i}}(y|x_i)]$ is always smaller than $V(y)$ as per Equation (3).

Another useful variance-based measure is the total-order index T_i (Homma and Saltelli, 1996), which measures the first-order effect of a model input jointly with its interactions up to the k-th order:

$$T_i = \frac{E_{\boldsymbol{x}_{\sim i}}[V_{x_i}(y|\boldsymbol{x}_{\sim i})]}{V(y)} \,. \tag{4}$$

The index is called "total" because it includes all factors in the variance decomposition [see Equation (2)] that include the index i. For instance, for a model with three factors, $T_1 = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3}$, and likewise for T_2 or T_3. The meaning of T_i is *the fraction of variance that would remain on average if x_i is left to vary over its uncertainty range while all other factors are fixed*. Note that the theory of variance-based measures is as flexible as to accommodate "group" or "set" sensitivities. These are simply the first-order effect of a set of factors: if u is the set of factors (x_1, x_2), then $S_u = S_1 + S_2 + S_{1,2}$.

1.2. VARS

VARS is based on variogram analysis to characterise the spatial structure and variability of a given model output across the input space (Razavi and Gupta, 2016a, 2016b). Let us again consider a function of factors $f(\boldsymbol{x}) = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$. If $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are two generic points separated by a distance $\boldsymbol{h}$, then the variogram $\gamma(.)$ is calculated as

$$\gamma(\boldsymbol{x}_A - \boldsymbol{x}_B) = \frac{1}{2} V\left[y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B)\right] \,, \tag{5}$$

and the covariogram $C(.)$ as

$$C(\boldsymbol{x}_A - \boldsymbol{x}_B) = COV\left[y(\boldsymbol{x}_A), y(\boldsymbol{x}_B)\right] \,. \tag{6}$$

Note that

$$V\left[y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B)\right] = V\left[y(\boldsymbol{x}_A)\right] + V\left[y(\boldsymbol{x}_B)\right] - 2COV\left[y(\boldsymbol{x}_A), y(\boldsymbol{x}_B)\right] \,. \tag{7}$$

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

$$\gamma(\boldsymbol{x}_A - \boldsymbol{x}_B) = V\left[y(\boldsymbol{x})\right] - C(\boldsymbol{x}_A, \boldsymbol{x}_B) \,. \tag{8}$$

As mentioned, the points $\boldsymbol{x}_A, \boldsymbol{x}_B$ are spaced by a fixed distance, and V, COV are the variance and covariance, respectively. Note that $\gamma(.)$ is defined by the interval separating $\boldsymbol{x}_A, \boldsymbol{x}_B$. To make this clearer one can write $\boldsymbol{h} = \boldsymbol{x}_A - \boldsymbol{x}_B$, with $\boldsymbol{h} = h_1, h_2, \ldots, h_n$, so that

$$\gamma(\boldsymbol{h}) = \frac{1}{2} E\left[y(\boldsymbol{x} + \boldsymbol{h}) - y(\boldsymbol{x})\right]^2 \,, \tag{9}$$

where the term E^2 in the expression of the variance as the expectation of the square minus the square of the expectation, $V(.) = E(.)^2 - E^2(.)$, is assumed to be zero. The practical formula for computing a multidimensional variogram is

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum \left[y(\boldsymbol{x}_A) - y(\boldsymbol{x}_B)\right]^2 \,, \tag{10}$$

where the sum is extended to all $N(\boldsymbol{h})$ couples of points $\boldsymbol{x}_A, \boldsymbol{x}_B$ such that their modulo distance $|\boldsymbol{x}_A - \boldsymbol{x}_B|$ is $\boldsymbol{h}$. Razavi and Gupta (2016a, 2016b) suggest some integral measures based on variogram γ, i.e. the integrated variogram $\Gamma(H_i)$:

$$\Gamma(H_i) = \int_0^{H_i} \gamma(h_i) dh_i \,, \tag{11}$$

and recommend the use of IVARS_{10}, IVARS_{30}, and IVARS_{50} (computed for H equal to 10%, 30%, and 50% of the factor range respectively) to explore larger fractions of the variation space of the function, with IVARS_{50} corresponding to the entire interval [in variogram analysis, the maximum meaningful range is one half of the factor range (Cressie, 2015)].

Of important practical use, as we shall see, is the directional variogram along one of the axes of the factors space,

$$\gamma(h_i) = \frac{1}{2} E[y(x_1, \ldots, x_{i+1} + h_i, \ldots, x_n) - y(x_1, \ldots, x_i, \ldots, x_n)]^2 \,, \tag{12}$$

which is evidently computed on all couples of points spaced h_i along the x_i axis, with all other factors being kept fixed. Note that the difference in brackets is what is called in Saltelli et al. (2010) a step along the x_i direction, which is fungible to compute the total sensitivity index T_i.

The equivalent of Equation (8) for the case of the unidirectional variogram $\gamma(h_i)$ is

$$\gamma_{x^*_{\sim i}}(h_i) = V\left(y \middle| x^*_{\sim i}\right) - C_{x^*_{\sim i}}(h_i), \tag{13}$$

where $\boldsymbol{x}^*_{\sim i}$ is a fixed point in the space of non-x_i.

In order for VARS to compute the total-order index T_i [labeled as VARS-TO by Razavi and Gupta (2016a)], the authors suggest taking the mean value across the factors' space on both sides of Equation (13), thus obtaining

$$E_{x^*_{\sim i}}\left[\gamma_{x^*_{\sim i}}(h_i)\right] = E_{x^*_{\sim i}}\left[V\left(y \middle| x^*_{\sim i}\right)\right] - E_{x^*_{\sim i}}\left[C_{x^*_{\sim i}}(h_i)\right], \tag{14}$$

which can also be written as

$$E_{x^*_{\sim i}}\left[\gamma_{x^*_{\sim i}}(h_i)\right] = V(y) T_i - E_{x^*_{\sim i}}\left[C_{x^*_{\sim i}}(h_i)\right], \tag{15}$$

and therefore

$$T_i = \text{VARS-TO} = \frac{E_{x^*_{\sim i}}\left[\gamma_{x^*_{\sim i}}(h_i)\right] + E_{x^*_{\sim i}}\left[C_{x^*_{\sim i}}(h_i)\right]}{V(y)}. \tag{16}$$

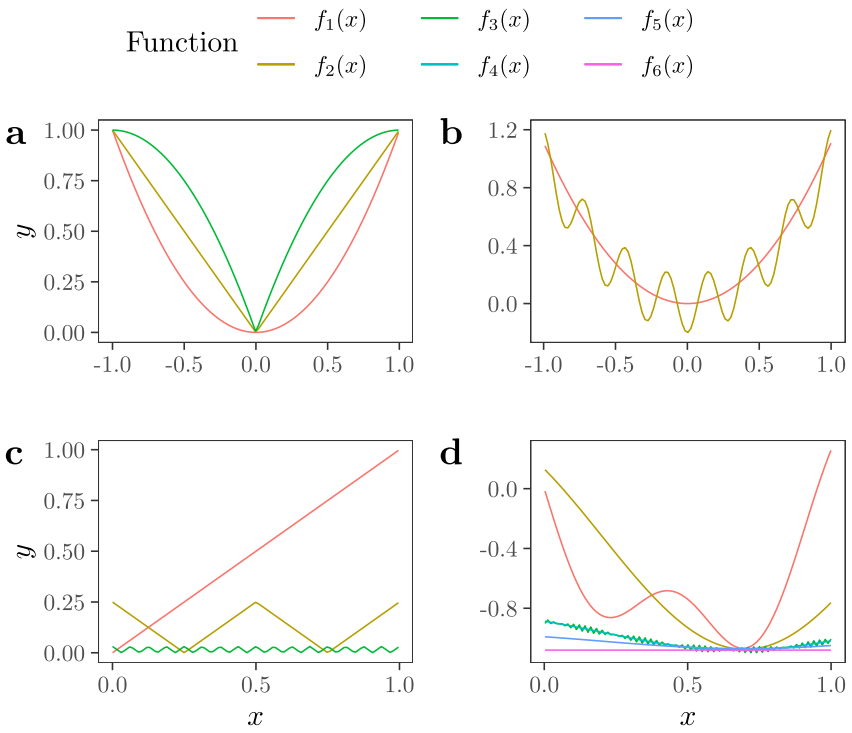

Fig. 1. Examples of functions in Razavi and Gupta (2016a). a) Unimodal functions with different structures. b) Multimodal versus unimodal function with identical variance. c) Functions covering different ranges in the response. d) A six-dimensional response surface. See the Supplementary Materials for a mathematical description of all functions in all sub-plots (see also Equation (17) for the model in d).

2. The issue of intuitiveness and importance

In a paper immediately preceding VARS, Razavi and Gupta (2015) already stressed two main drawbacks of global sensitivity analysis:

1. The incapacity of variance-based Sobol' indices to appraise the spatial distribution of the model response.
2. The dependence of the Morris (1991) approach on the step size defined by the analyst, which can significantly condition the final sensitivity value.

VARS was presented as a comprehensive approach which overcomes these drawbacks by encapsulating, in a single sensitivity framework, a "unified assessment of local and global sensitivity" (Razavi and Gupta 2015, p. 3090). The fact that integrated variogram measures such as $IVARS_{10}$, $IVARS_{30}$ and $IVARS_{50}$ are able to differentiate sensitivities as a function of scale H, whereas Sobol' indices do not, is taken as proof of the limitations of the latter. According to Razavi and Gupta (2016a, pp.427-428, 433-434), this endows VARS with a more intuitive appraisal of sensitivities.

Razavi and Gupta (2016a) construct their case using several functions, which we reproduce hereafter. In Fig. 1a, Sobol' indices do not differentiate f_3 from f_1, whereas VARS points towards f_3 as the most sensitive function. In Fig. 1b, variance-based methods equate f_1 with f_2 because they have identical variance. According to Razavi and Gupta (2016a, p. 428), this "runs counter to our intuitive notion of sensitivity" given the multimodality of f_2. If VARS is used, f_2 is identified as more sensitive than f_1 for $0 \leq h \leq 0.2$.

In Fig. 1c, Sobol' indices do not detect the periodicities of f_2, which Razavi and Gupta suggest might be important in evaluating the impact of a factor from the perspective of model calibration. In Fig. 1d, variance-based methods regard f_2 as more sensitive than f_1. Razavi and Gupta (2016a, p. 433) argue that this is "contrary to intuition" because the effect of f_1 is more complex (bi-modal). $IVARS_{10}$ and $IVARS_{30}$, in contrast, characterise f_1 as more sensitive than f_2.

It is apparent that for Razavi and Gupta (2016a) a sensitivity measure should be able to appraise the function structure. Our impression is that this perception of sensitivity is relevant to specific contexts, e.g. a diagnostic setting in which one is interested in the topology of a given function. However, the key lies in the definition of "importance" pointed to by VARS. In which sense is f_2 more important than f_1 in Fig. 1b, or f_3 more important than f_1 and f_2 in Fig. 1a? If SA is used in an information quality setting (Kenett and Shmueli, 2013), when the aim is to determine which factor has the highest potential to reduce the uncertainty in the inference (i.e. how much is gained by discovering the true value of an uncertain factor), these functions might be regarded as equally sensitive. The same applies to Fig. 1d: given that f_2 changes more decidedly over the interval range than f_1, a larger reduction in uncertainty can be achieved by learning first about f_2 than about f_1.

Given that SA quantifies the relative influence of each model input in the model output, the concept of sensitivity is ultimately linked to that of "importance". This is why it should be clear what do we mean when we say that a model input is "important", or that a model output is very sensitive to a given model input. Variance-based methods meet this requirement by linking SA to statistical theory via ANOVA (Archer et al., 1997), thus defining SA as "the study of how the variance in the model output is apportioned to different sources of uncertainty in the model input" (Saltelli et al., 2002). The use of variance-based methods such as

Sobol' indices are well defined and associated with clear settings (Saltelli and Tarantola, 2002):

1. Factors prioritization: the aim is to identify the single factor that, if determined (i.e., fixed to its true but unknown value), would lead to the greatest reduction in the variance of the output. This is met by the first-order sensitivity index (S_i).
2. Variance reduction: the aim is to identify the sets of factors (couples, triplets, and so on) leading to the reduction of the output variance below a given threshold, and doing this by fixing the smallest number of factors. This is achieved by using set (group) sensitivity indices.
3. Factors fixing: the objective is to identify factors that can be fixed anywhere in their range of variation without affecting the variance of the output. This is met by the total-order sensitivity index (T_i).

Variance-based methods clearly resolve what is meant by "importance" of a factor. However, this is not as apparent in the case of VARS: if a decision needs to be taken based on the inference provided by a model, which of the variogram-based measures ($IVARS_{10}$, $IVARS_{30}$, $VARS_{50}$, VARS-TO) should be finally used to characterise the factors' importance? and what does "importance" mean for VARS? Razavi and Gupta (2016a)'s statement of VARS being more "intuitive" than Sobol' indices is open to debate: intuition is in the eyes of the beholder, while solid criteria underpin the methodological quality of Sobol' indices.

One way of gaining a factual insight into the alleged "intuitiveness" of VARS is through the analysis of its use by the 86 studies that have cited Razavi and Gupta (2016a, 2016b) up until November 2020. If adopted and used by practitioners other than the VARS authors themselves, and if the VARS framework is applied as recommended by their designers (i.e. by exploring different ranges of the spatial structure of the model response through integrated variograms IVARS and VARS-TO), then the claim by Razavi and Gupta (2016a, 2016b) of VARS being an instinctive, user-friendly framework will find empirical support.

We observed that 53 studies (62%) cite Razavi and Gupta (2016a, 2016b) but do not implement VARS in any specific sensitivity analysis. Of the 33 studies that do apply VARS, 13 (40%) include either Razavi and/or Gupta as lead author/s or co-authors. Hence the number of papers that use VARS and are not contributed by VARS authors amounts to 20, 23% of all VARS citations (Fig. 2a and b).

Out of the 33 studies that do use VARS, there were nine from which we could not retrieve precise information on the VARS metric/s used. As for the remaining 24, 15 studies used just one VARS metric (11 $IVARS_{50}$ and four VARS-TO), two used two ($IVARS_{50}$ and VARS-TO), three used three ([$IVARS_{10}$, $IVARS_{30}$, $IVARS_{50}$] x 2; $IVARS_{10}$, $IVARS_{50}$, VARS-TO) and four studies used all four metrics. The contributions by authors other than Razavi and Gupta have strongly leaned towards the use of a single summary measure: out of the 12 works for which we could retrieve information on the VARS metric used, nine relied merely on one metric (six on $IVARS_{50}$ and three on VARS-TO), with one study using two, three and all four VARS measures.

With regard to the sensitivity settings, VARS has largely been applied to models with up to 20 parameters, with MESH being the model with the highest dimensionality (111). The number of stars has been mostly set between 20 and 50, with a single study raising it to 1000. A large number of works have used $h = 0.1$, with the minimum and maximum h values being 0.01 and 0.3 (Fig. 2c–e).

As yet, such results place the "intuitive nature" of VARS in a disputable position: although cited, its use as a sensitivity measure has been comparatively moderate, and most authors have preferred a single summary VARS metric ($IVARS_{50}$ or VARS-TO, both very similar to the Sobol' total-order index (Razavi and Gupta 2016a, p. 434)) rather than implementing –and interpreting– the whole integrated variogram approach.

The discussion above leads to another aspect listed by Razavi and Gupta 2016a, p. 423) as a motivation for developing VARS: an "ambiguous characterization of sensitivity":

> *(…) different SA methods are based in different philosophies and theoretical definitions of sensitivity. The absence of a unique definition for sensitivity can result in different, even conflicting, assessments of the underlying sensitivities for a given problem.*[2]

We argue that the source of ambiguity in sensitivity analysis is not the lack of a unifying theory, or the fact that many sensitivity measures are available, but in the definition of "importance". Unless the analyst stipulates what she means when she says that a variable is important, different methods can be thrown at the model resulting in different ordering of importance of the input variables, whereby the analyst could be tempted to cherry-pick the method most conforming to one's own bias. By linking the definition of importance to clear settings, Sobol' indices resolve this quandary clearly and transparently, and provide end-users with a plain English description of the results. This comes in handy when the receiver (customer) of the analysis is not another practitioner.

The expedient to produce functions where the validity of Sobol' indices is downplayed is quite common. This approach was also taken by Liu et al. (2006) and Pianosi and Wagener (2015) using Liu's highly-skewed function $y = \frac{x_1}{x_2}$, where $x_1 \sim \chi^2(10)$ and $x_2 \sim \chi^2(13.978)$ (Fig. 3). The reader might wonder why one of the degrees of freedom is expressed with two-digit precision and the other with a five-digits one. The reason is that, with these crisp numbers, T_1 and T_2 are identical and equal to 0.5462, while inspection of Fig. 3b should convince us that x_1 is more important than x_2 by virtue of its longer tail. The Liu function is thus what Lakatos et al. (1976) would have called a *monster example*, designed on purpose to invalidate variance-based methods. However, based on the definition of "importance" of Sobol' indices, the fact that they are equally influential appears totally reasonable.

We conclude by stating that rather than hinting at what should or not should be intuitive, a sensitivity index should pin down its definition of importance in unambiguous terms.

3. The issue of efficiency

Razavi and Gupta (2016a, 2016b) claim that VARS-TO is much more computationally efficient than the total-order estimator of Saltelli et al. (2008, Eq. 4.23) (up to two orders of magnitude), which is taken as a state-of-the-art implementation of the Sobol' approach. They make their case with three different models:

1. The six-dimensional response surface displayed in Fig. 1d, which is a purely additive model. VARS-TO accurately ranks the model inputs with just 60 simulations, beating the Saltelli et al. (2008) estimator of total-order indices at $> 6,000$ simulations (Razavi and Gupta 2016a, pp. 435-436).
2. The five-dimensional conceptual rainfall-runoff model HYMOD (Vrugt et al., 2003). VARS-TO detects the "true" ranking of the model inputs at 500 simulations, while the Saltelli et al. (2008) estimator requires 10,000 simulations (Razavi and Gupta 2016b, pp. 443-444).
3. The 45-dimensional land surface scheme-hydrology model MESH (Pietroniro et al., 2007). The VARS-TO estimate of the total-order effect stabilizes at 5000 simulations, whereas the Saltelli et al.

[2] The extent to which this points to an ambiguity is unclear. In any discipline, including statistics, different methods may naturally exist which become useful in different applications. For instance, the linear relation between two variables x and y might be modelled with Ordinary Least Squares (OLS) if x causes y, or with Standard Major Axis (SMA) if it is unclear which variable is the predictor and which one is the response (Smith, 2009). Does this mean that the characterisation of residuals in regression analysis is an ambiguous branch of statistics?.

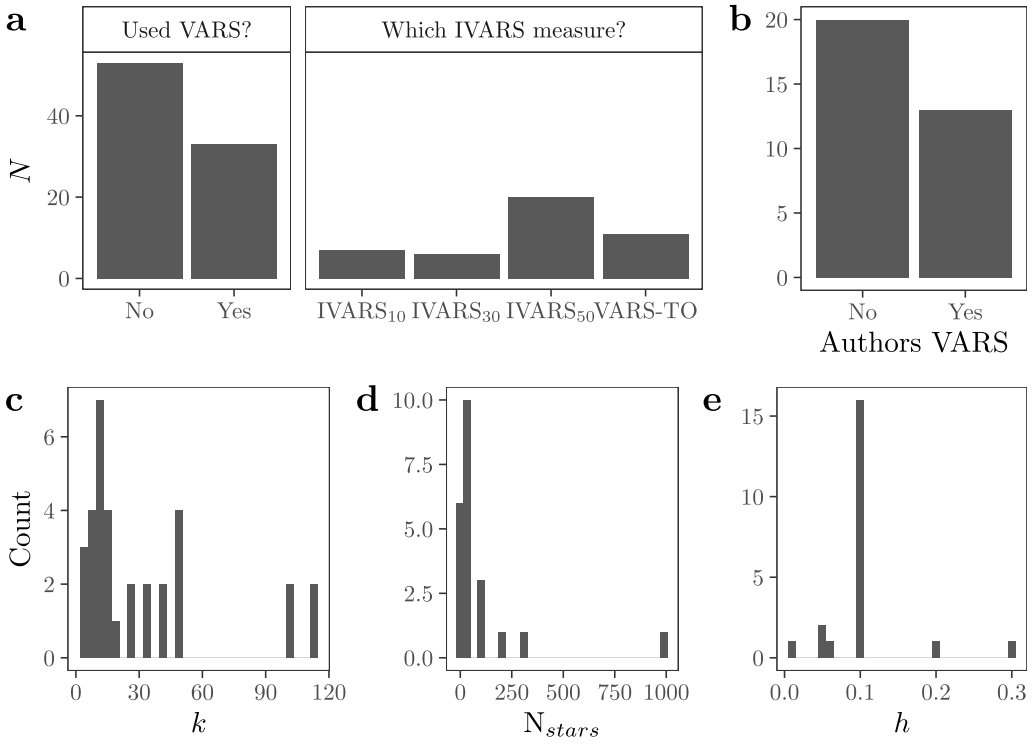

Fig. 2. Results of the survey conducted on all papers that cite Razavi and Gupta (2016a, 2016b) up to November 2020. a) Use of VARS. b) Number of studies that use VARS and include (Yes) or do not include (No) the VARS authors themselves. c) Distribution of the dimensionality of the models for which VARS has been applied. d) Distribution of the number of stars used. e) Distribution of the values set for h.

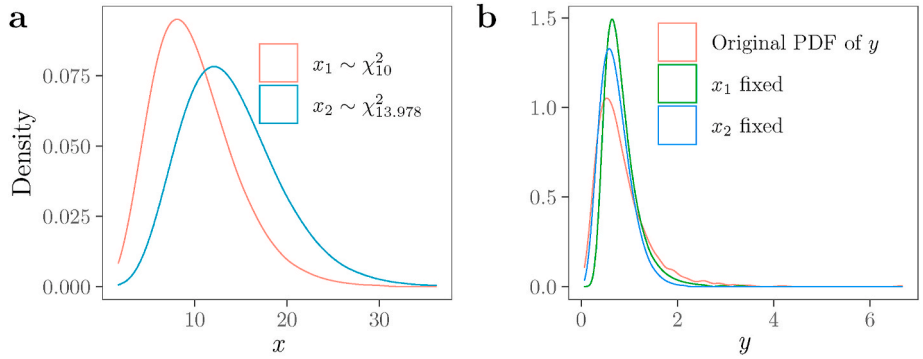

Fig. 3. The highly skewed function of Liu et al. (2006). a) Distribution of x_1 and x_2. b) Comparison of impacts of inputs.

(2008) estimator requires more than 100,000 simulations (Razavi and Gupta, 2016b, pp. 453-454).

Do these examples truly prove that VARS-TO is between 20 and 100 times more efficient than the Sobol'-based approach to total-order indices?

3.1. The case of the six-dimensional response surface model

To properly answer this question in the case of the six-dimensional model (Fig. 1d), whose functional form reads as

$$\begin{aligned} g_1(x_1) &= -\sin(\pi x_1) - 0.3\sin(3.33\pi x_1) \\ g_2(x_2) &= -0.76\,\sin\,[\pi(x_2 - 0.2)] - 0.315 \\ g_3(x_3) &= -0.12\,\sin\,[1.05\pi(x_3 - 0.2)] - 0.02\sin(95.24\pi x_3) - 0.96 \\ g_4(x_4) &= -0.12\,\sin\,[1.05\pi(x_4 - 0.2)] - 0.96 \\ g_5(x_5) &= -0.05\,\sin\,[\pi(x_5 - 0.2)] - 1.02 \\ g_6(x_6) &= -1.08 \\ y &= f\left[g_1(x_1) + g_2(x_2) + \ldots + g_6(x_6)\right], \end{aligned} \tag{17}$$

we should first focus on the sampling design of VARS and Sobol' indices.

The computation of VARS relies on stars and is referred to as STAR-VARS by Razavi and Gupta (2016b): the analyst first randomly selects N_{star} points across the factor space, i.e. via random numbers, Latin Hypercube Sampling (LHS) or Sobol' Quasi Random Numbers (QRN). These are the "star centres" and their location can be denoted as $s_v = s_{v_1}, \ldots, s_{v_i}, \ldots, s_{v_k}$, where $v = 1, 2, \ldots, N_{star}$. Then, for each star centre, a cross section of equally spaced points Δh apart needs to be generated for each of the k factors, including and passing through the star centre (Fig. 4, left side plot). The cross section is produced by fixing $s_{v_{\sim i}}$ and varying s_i. Finally, for each factor all pairs of points with h values of $\Delta h, 2\Delta h, 3\Delta h$ and so on should be extracted. The total computational cost of this design is $N_t = N_{star}\left[k\left(\frac{1}{\Delta h} - 1\right) + 1\right]$.

Sobol' indices also rely on a star-based sampling strategy: they require a $(N, 2k)$ base sample matrix, designed via LHS or QRN, in which the rightmost k columns are allocated to an $\boldsymbol{A}$ matrix and the leftmost k

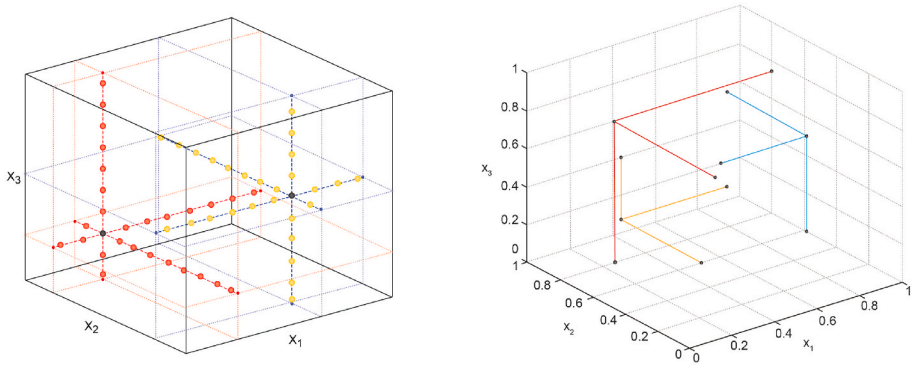

Fig. 4. Sampling design of VARS-TO (STAR-VARS, left) and Sobol' indices (right). For VARS-TO the plot shows a star-based sampling in three dimensions, with $\Delta h = 0.1$ and number of stars $N_{star} = 2$. Black dots are the star centres, and the coloured dots are the additional $\frac{1}{\Delta h}$ points along the three axes. Adapted from Razavi and Gupta (2016b, Fig. 1). For Sobol' based-indices, the plot also displays a three-dimensional model, with the links being steps in the x_i direction and $N = 4$. Adapted from Becker et al. (2015, Fig. 18.7).

columns to a $\boldsymbol{B}$ matrix. Then, k extra $(N,k)\boldsymbol{A}_{\boldsymbol{B}}^{(i)}$ matrices are created, where all columns come from $\boldsymbol{A}$ except the i-th, which comes from $\boldsymbol{B}$. This design creates stars with centres and points a step away in the x_i direction (Fig. 4, right-side plot). The cost of this design for T_i is $N_t = N(k+1)$, where N is the row dimension of the base sample matrix.

When the function or the model under study is fully additive, as in the six-dimensional surface model mentioned above (Fig. 1d), the computation of VARS-TO can be done with a single cross-section in the space of $x^*_{\sim i}$ for each model input. VARS-TO thus becomes a first-order index *de facto*, as one model input remains constant while all the others vary. The natural term of comparison is thus the Sobol' first-order index, and not the total. In that sense, and for any function which behaves non-additively for at least one factor, i.e. $F = f(x_i) + g(\boldsymbol{x}_{\sim i})$, the first-order effect S_i can be computed very easily, since

$$S_i = \frac{E_{x_i}[f(x_i)]^2 - E^2_{x_i}[f(x_i)]}{V(F)}, \tag{18}$$

i.e. S_i is only a function of x_i and hence it can be computed with a single trajectory along x_i, irrespective of its position in $\boldsymbol{x}_{\sim i}$. We provide the proof in Section 2.1 of the Supplementary Materials.

We used Equation (18) to compute S_i for the six-dimensional model, aiming at replicating the results by Razavi and Gupta (2016a, see their Fig. 6). For VARS-TO and Sobol'-based indices they tested their probability of failure, defined as the probability of obtaining erroneous ranks for the model inputs of the six-dimensional model (Fig. 1d and Equation (17)). We observed that, if Equation (18) is used to compute Sobol'--based indices, all model inputs are accurately ranked at $N_t = 896$ (Fig. 5a), contrasting with the $N_t > 6,000$ obtained by Razavi and Gupta (2016a). This example suggests that VARS-TO is indeed more efficient than a Sobol'-based approach when the model is fully additive and the aim is to rank the parameters, but significantly less than what the authors claimed it to be. It is also worth noting that there are other approaches that might permit a more efficient computation of first-order indices (Plischke, 2010; Mara et al., 2017; Strong et al., 2012).

However, why do Razavi and Gupta rely exclusively on the "probability of failure" in the ranking as a performance metric? Sorting the parameters by their influence in the model output is indeed a common setting in sensitivity analysis. However, other goals may exist: the analyst might be more interested in getting exact values for the sensitivity indices in order to ascertain, for instance, how much the uncertainty would be reduced if the "true" value of an uncertain factor is discovered. In such context, a performance measure such as the mean absolute error (MAE) between the estimated ($\widehat{\boldsymbol{T}}$) and the analytical ($\boldsymbol{T}$) values might be more appropriate. The MAE has been a very widespread performance measure in sensitivity analysis (Saltelli et al., 2010; Lo Piano et al., 2020), and is computed as

$$\text{MAE} = \frac{1}{p}\sum_{v=1}^{p}\left(\frac{\sum_{i=1}^{k}\left|T_i - \widehat{T}_i\right|}{k}\right), \tag{19}$$

where p is the number of replicas of the sample matrix, and T_i and $\widehat{T}_i$ the analytical and the estimated total-order index of the i-th input.

Had Razavi and Gupta relied on the MAE rather than on the "probability of failure" as a performance measure, their assessment of the efficiency of VARS-TO and Sobol'-based indices would have been very different: throughout the range of explored model runs, VARS-TO never approaches Equation (18) in terms of accuracy (Fig. 5b). These results exemplify how sensitive the outcome of a benchmarking exercise can be to the particular settings defined by the analyst: merely the use of a different performance measure can completely tip the balance from one estimator to another. In the section below, we show how to minimize this source of bias to get a more accurate picture of the true performance of VARS-TO compared to Sobol'-based estimators (Puy et al., 2020a).

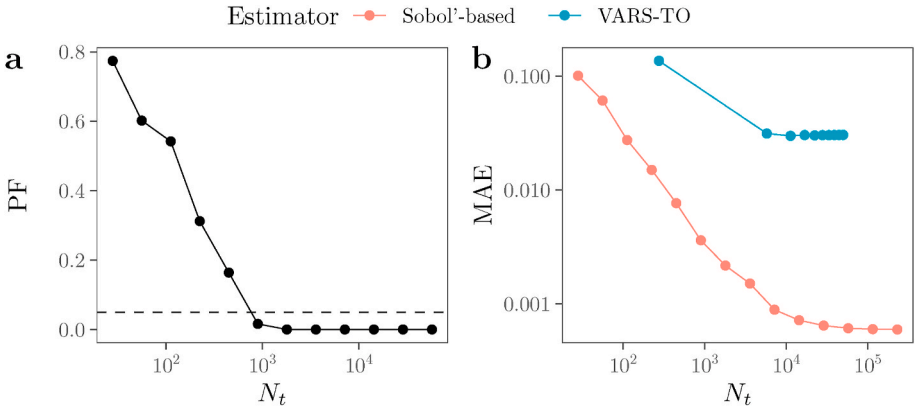

Fig. 5. Assessment of VARS-TO and the Sobol'-based estimator [Equation (18)] with the six-dimensional model as a test function. a) Probability of failure (PF) of Equation (18) in correctly ranking the model inputs. Each dot summarises the PF over 500 quasi-random number matrices with different starting points. The horizontal dashed line is at PF = 0.05. For more details about the computational methodology, see Razavi and Gupta (2016a). b) Mean Absolute Error (MAE) [see Equation (19)], with $p = 50$. See Section 2.1.1 of the Supplementary materials for the computation of the analytical values of the six-dimensional model.

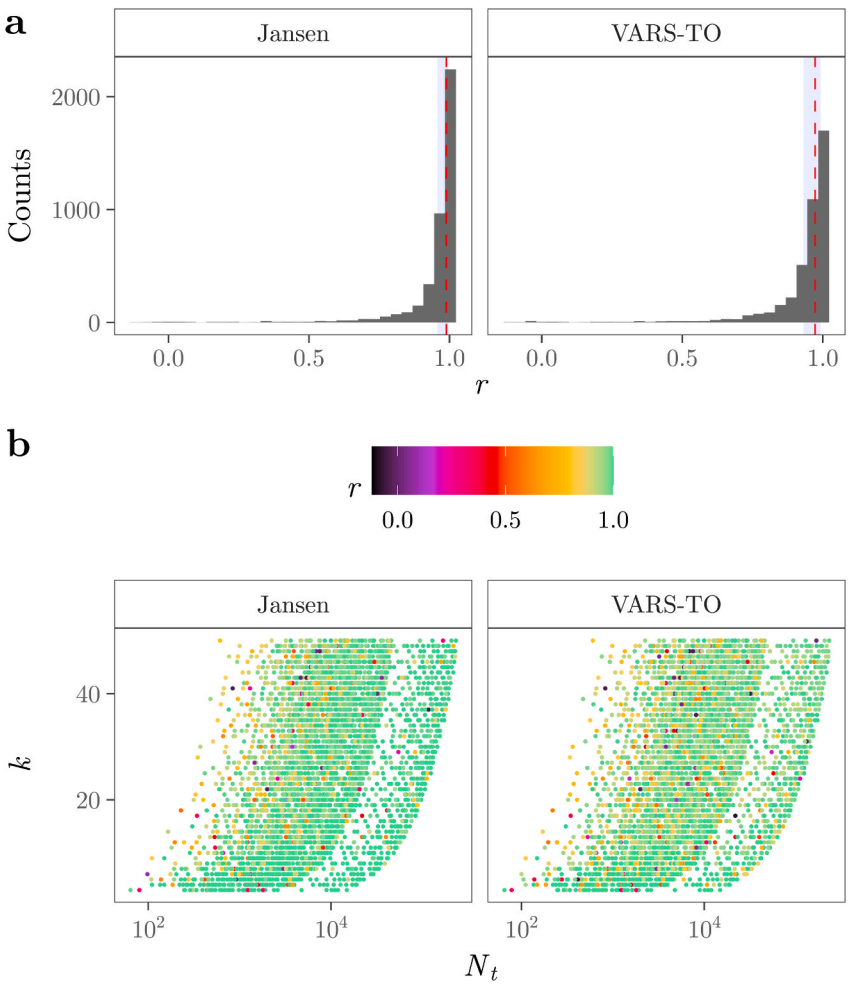

Fig. 6. Uncertainty analysis conducted on 2^{12} simulations. a) Histograms of r. The vertical red, dashed line shows the median value. The transparent, blue rectangle frames the 0.25, 0.75 quantiles. b) Scatterplots showing the performance of Jansen and VARS-TO as a function of the total number of model runs N_t and the model dimensionality k. Each simulation is a dot. The greener (darker) the colour, the better (worse) the performance. The white space between 10^3 and 10^4 in the x axis is caused by the uneven distribution selected for h (see Table 1 and the Supplementary Materials). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.2. The case of the HYMOD and MESH models

Unlike the six-dimensional model, HYMOD and MESH are non-additive models. Hence a single trajectory is not enough and several cross-sections in the space of $x^*_{\sim i}$ should be drawn to fully explore the hypercube. Under such settings, the comparison between VARS-TO and a Sobol'-based estimator of the total-order index is the appropriate methodological choice.

But does the higher accuracy of VARS-TO reported by Razavi and Gupta (2016a, 2016b) for these two models truly evidence its superiority over Sobol'-based total-order indices? We argue that the following issues make Razavi and Gupta (2016a, 2016b)'s claim controversial:

- The use of the Saltelli et al. (2008) total-order estimator as "state--of-the-art". Amongst all estimators available for computing T_i, that of Saltelli et al. (2008) ranks close to last on accuracy and performance and is significantly outperformed by the Jansen or the Janon/Monod estimators (Jansen, 1999; Puy et al., 2020a; Monod et al., 2006; Janon et al., 2014). Furthermore, Saltelli et al. (2010) demonstrated that configurations based on $\boldsymbol{B}$, $\boldsymbol{B}_A^{(i)}$ matrices (as is the case of the Saltelli estimator) were surpassed in performance by those relying on $\boldsymbol{A}$, $\boldsymbol{A}_B^{(i)}$ matrices (i.e. the Jansen estimator) when Quasi-Random numbers were used to create the sample matrix.
- The extrapolation of the results obtained with HYMOD and MESH to mean that VARS-TO is *generally* better than Sobol'-based indices. Puy et al. (2020a) recently showed that, once the benchmark settings are randomised (i.e. the model and its dimensionality, the sampling method, the total number of model runs, the fraction of active second and third-order effects, the distribution of the model inputs and the performance measure), VARS-TO loses much of its purported computational superiority: it only very slightly outperforms the Sobol'-based estimators Jansen (1999) and Janon/Monod (Monod et al., 2006; Janon et al., 2014) when there are serious constraints on the number of model runs that can be allocated to each model input (i.e. 2-10). At larger sample sizes, the performances of VARS and Jansen and Janon/Monod are exactly the same (Puy et al., 2020a). This randomisation is required to reduce the dependence of the results on the choices taken by the analyst: as we have seen in the case of the six-dimensional model (see Section 3.1), even the use of a given performance measure over another might completely swap the outcome of an analysis.

In order to obtain a more comprehensive view of the performance of VARS-TO against Sobol' indices, we have reproduced the work by Puy et al. (2020a) with the following changes and/or additions:

1. We have tested VARS-TO against the Jansen (1999) formula, one of the most precise and accurate Sobol' total-order estimators (Puy et al., 2020a).
2. We have increased the range of the proportion of active second and third-order effects in the test functions (i.e. between 50-100% and 30–100% respectively; in Puy et al. (2020a) they ranged between 30-50% and 10–30% respectively). This aimed at checking how VARS-TO performs under serious non-additivities.
3. We have taken into account the algorithmic uncertainties of VARS-TO, i.e. the number of stars N_{star} and the distance between pairs h,

which ultimately condition its computational cost. These design parameters need to be set by the analyst before executing the algorithm, and the specific value that might work best is unclear. Different authors have used different values for h ($\Delta h = 0.002, \Delta h = 0.1, \Delta h = 0.3$; [19, 20, 23, 47]; see Fig. 2e). Razavi et al. (2019) recommend $h = 0.1$ and $h < 0.1$ if more accurate results are needed. As shown by Puy et al. (2020b) for PAWN, the uncertainty in the design parameters of a sensitivity index might contribute appreciably to its volatility.

We compared the performance of VARS-TO and the Jansen estimator by treating the main benchmark settings listed in Table 1 as uncertain parameters described by probability distributions. We created a $(2^{12}, k)$ sample matrix using Sobol' (1967, 1976) quasi-random numbers, in which each row is a sample point and each column an uncertain parameter. For $v = 1, 2, \ldots, 2^{12}$ rows, we computed VARS-TO and the Jansen total-order index according to the specifications set by $N_{star_v}, h_v, \ldots, \delta_v$. The final model output was r_v, the correlation coefficient between the indices estimated by VARS-TO and Jansen ($\widehat{\boldsymbol{T}}_v$) and the "true" indices ($\boldsymbol{T}_v$), computed with a large sample size ($N = 2^{12}$) and the Jansen (1999) estimator. The larger the r_v, the better the estimation of the "true" sensitivity indices by VARS-TO or Jansen. We argue that this approach allows us to examine the accuracy of VARS-TO more comprehensively given the enormous range of sensitivity problems that it is able to explore (Puy et al., 2020a, Becker, 2020) (potentially more than 3 billion scenarios in this case). If VARS-TO outperforms Sobol'-based indices unequivocally, as asserted by Razavi and Gupta (2016a, 2016b), its computational advantage should emerge against Jansen as well. The Supplementary Materials thoroughly detail the rationale and the execution of the experiment.

Fig. 6a shows that both Jansen and VARS-TO are very accurate as the empirical distribution of r is highly right-skewed. If anything, Jansen seems to outperform VARS-TO overall due to its slightly narrower distribution (95% CI 0.93–0.99, median of 0.99 for Jansen; 95% CI 0.87–0.99, median of 0.97 for VARS). This is also apparent in Fig. 6b, with VARS-TO presenting more simulations with redder/orange colours (approx. $r \leq 0.85$). A closer look at the performance of both approaches reveals that Jansen maintains a higher median accuracy at higher dimensions (Fig. 7a), whereas VARS-TO confirms its slightly higher efficiency only when the number of runs that can be allocated per model input (N_t/k) is considerably constrained (< 50 in this case, Fig. 7b) (Puy et al., 2020a). VARS-TO also displays a larger volatility at $100 > (N_t/k)$ (Fig. 7b), suggesting that Jansen might become more stable in a larger number of sensitivity problems if the number of model runs per input is increased. These results rest on solid grounds as the number of simulations for which we have computed the median N_t/k is almost identical for Jansen and VARS-TO (Fig. 7c). Overall, this proves that both estimators have a very similar efficiency and reliability.

Table 1
Summary of the uncertain parameters and their distributions. $\mathscr{D}\mathscr{U}$ is discrete univariate. See the Supplementary Materials for a description of the rationale behind the selection of the uncertain parameters and their distributions.

Parameter	Description	Distribution
N_{star}	Number of star centres	$\mathscr{D}\mathscr{U}(3, 50)$
h	Distance between pairs	$\mathscr{D}\mathscr{U}(0.01, 0.05, 0.1, 0.2)$
k	Number of model inputs	$\mathscr{D}\mathscr{U}(3, 50)$
ε	Randomness in the test function	$\mathscr{D}\mathscr{U}(1, 200)$
τ	Sampling method	$\mathscr{D}\mathscr{U}(1, 2)$
φ	Probability distribution of the model inputs	$\mathscr{D}\mathscr{U}(1, 8)$
k_2	Fraction of pairwise interactions	$\mathscr{D}\mathscr{U}(0.5, 1)$
k_3	Fraction of three-wise interactions	$\mathscr{D}\mathscr{U}(0.3, 1)$
δ	Selection of the performance measure	$\mathscr{D}\mathscr{U}(1, 3)$

We also computed Sobol' indices to assess which uncertain parameter most influences the performance of VARS-TO (Fig. 8). We observed that c. 30% the variance in its performance is driven by the underlying probability distribution of the model inputs φ, which appears as the most influential parameter. The other parameters are important through interactions, especially the functional form of the model (ε), the sampling method (τ), the model dimensionality (k) and the performance measure selected (δ), in that order. The proportion of second and third-order effects (k_2, k_3) has no effect, which means that VARS-TO is very robust against non-additivities.

Compared to Jansen, VARS-TO significantly underperformed when the model inputs were normally distributed (e.g. when $\varphi = 2$, Figs.S5, 9). We observed that this was caused by high-order interactions between the sampling design of VARS-TO (Fig. 4, left side) and at least five different uncertain parameters, $N_{star}, h, k, \varphi, \tau$.

To understand these interactions, let us first assume that we use random numbers ($\tau = 1$) to sample our star centres, which Razavi et al. (2019, Table 1) list as a possible sampling strategy to compute VARS-TO. These star centres are located at $s_v = s_{v_1}, \ldots, s_{v_i}, \ldots, s_{v_k}$, where $v = 1, 2, \ldots, N_{star}$. The higher the N_{star} and k, the higher the chances that a value at the boundary of $(0, 1)$ is included in s_v. Given that VARS-TO requires fixing $s_{v_{\sim i}}$ while varying s_i at a step defined by h, this value at the periphery of $(0, 1)$ will be repeated in the $\left[\left(\frac{1}{h}\right) - 1\right](k-1)$ coordinates, which can be manifold if k is high and h is low. Once the model inputs are transformed into a normal distribution, it will turn into an extreme value and will disrupt both the model output and the computation of VARS-TO for the $\boldsymbol{x}_{\sim i}$ parameters.

Let us now assume that we do not use random numbers to sample the star centres, but Sobol' Quasi-Random (QRN) number sequences ($\tau = 2$). They are also contemplated by Razavi et al. (2019, Table 1) as a sampling strategy to compute VARS-TO. Although the design of QRN makes the sampling of star centres at the very periphery $(0, 1)$ very unlikely, cross-sections can indeed sample the boundary of the domain: for instance, if $h = 0.1$ and $s_{v_i} = 0.5$, the cross-section of the x_i parameter will be the vector $\boldsymbol{x}_i = 0.1, 0.2, \ldots s_{v_i}, \ldots, 1$. This will cause VARS-TO to crash as 1 becomes infinity under a normal distribution. Even if the STAR-VARS algorithm is modified to prevent 1 from being sampled (e.g. by replacing 1 by 0.999, as we did in this study), some cross-sections will still sample values very close to 1 by design, especially if h is set at a small value. These values will be extreme values under a normal distribution, disrupting again the computation of the model output and the VARS-TO index –in this case, for the x_i parameter.

We believe that this explains the high-order interactions involving $N_{star}, h, k, \varphi, \tau$, which are non-negligible (Fig. 8). The effect of N_{star} and h in VARS-TO was not explored by Puy et al. (2020a) nor by Becker (2020), who documented a slightly higher performance of VARS-TO against Jansen and Janon/Monod. Our results indicate that VARS-TO loses this marginal edge once these internal uncertainties jointly with the uncertainties in the benchmark settings are considered in the computations. Even if the use of VARS-TO is restricted to non-normal distributions ($\varphi \neq 2$), its performance would still be slightly outdone by Jansen (95% CI 0.96–0.99, median of 0.99 for Jansen; 95% CI 0.94–0.99, median of 0.97 for VARS).

4. Conclusions

We have revised the Variogram Analysis of Response Surfaces (VARS), a new framework for sensitivity analysis developed by Razavi and Gupta (2016a, 2016b). We have specifically focused on two aspects that, according to Razavi and Gupta (2016a, 2016b), make VARS outperform Sobol' indices: its more intuitive appraisal of sensitivities and of the importance of model inputs, and its 20–100 times higher computational efficiency.

The claim that VARS is more intuitive than Sobol' indices can hardly be reversed as it ultimately is a matter of personal taste, disciplinary

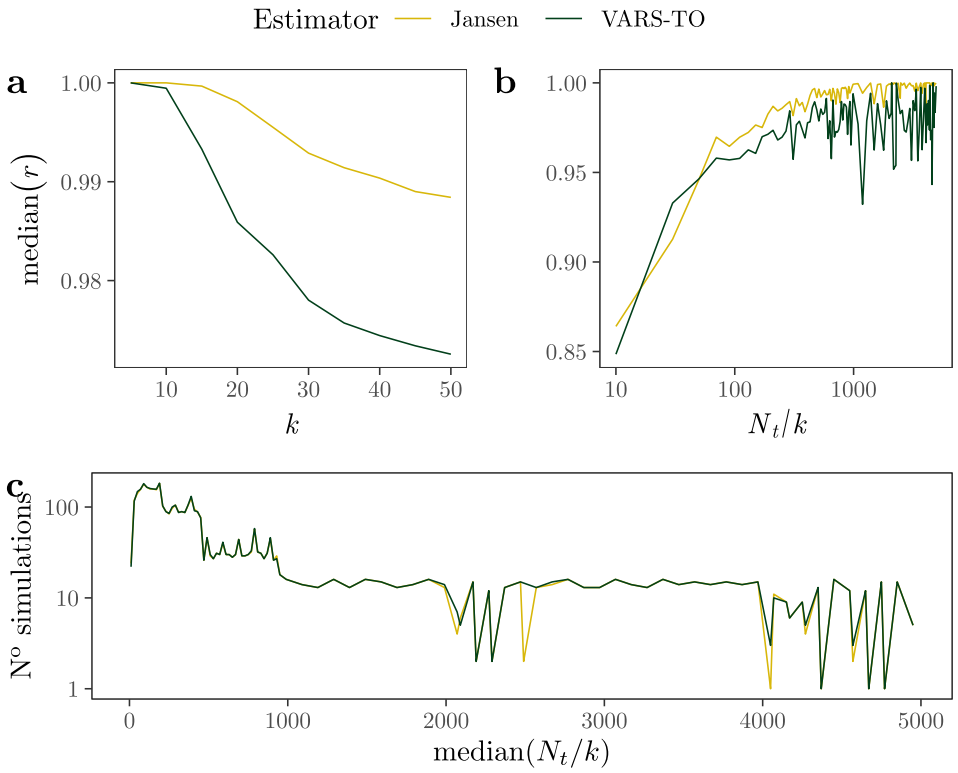

Fig. 7. Comparison between the accuracy and efficiency of VARS-TO and Jansen (1999). a) Evolution of the median r value across different dimensions k. b) Evolution of the median r value across the different number of runs allocated to each model input (N_t/k). c) Number of simulations with the same N_t/k ratio. Both lines almost fully overlap.

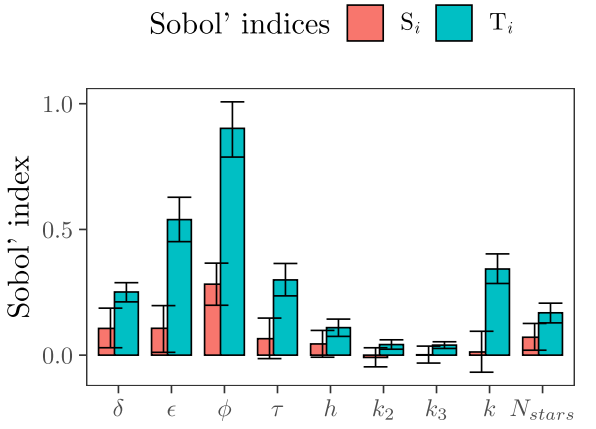

Fig. 8. Sobol' indices for VARS-TO. The error bars show the 95% confidence intervals, computed with the percentile method after bootstrapping ($R = 500$).

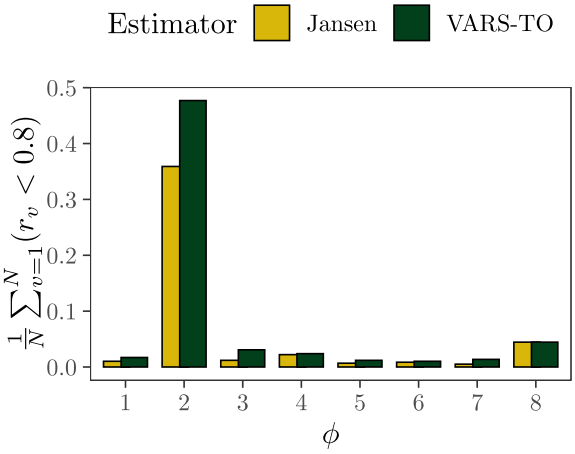

Fig. 9. Proportion of model runs with $r < 0.8$ as a function of φ. The normal distribution $[\mathcal{N} \sim (0.5, 0.2)]$ is triggered by $\varphi = 2$. See Fig. S1 for a recoding of all levels of φ into their probability distributions.

orientation and objective of the modeling activity: a geographer working in a diagnostic model setting might indeed prefer VARS's approach to the model structure due to its capacity to distinguish response surfaces. However, the professed higher intuitive nature of VARS ties in poorly with the available evidence: less than one half of the studies citing VARS actually implement it in a case study, and almost half of those that use it include the VARS authors themselves. Furthermore, most works do not explore the full range of the response surface but rely exclusively on summary metrics such as $IVARS_{50}$ or VARS-TO, which are very similar to the Sobol' total-order index.

We argue that Sobol' indices provide a clearer description of what an "important" model input is given its connection to statistical theory and ANOVA. The use of Sobol' first or total order indices is associated with clear research settings and their meaning can be easily conveyed to stakeholders or non-specialists, which adds to their transparency. VARS, in contrast, allows the analyst to zoom into the structure of the model output and assess its dependency on the model inputs through the integrated variograms $IVARS_{10}$, $IVARS_{30}$ and $IVARS_{50}$, as well as through the variance-based total-order effect VARS-TO. But what is their definition of importance? How useful is it for a stakeholder to know that a parameter is "important" under $IVARS_{10}$ and not as much under $IVARS_{30}$, for instance? Which IVARS measure should she finally rely onto before making a policy decision? If the answer is the summary measure VARS-TO, then it is unclear how VARS advances Sobol' indices given the reliance of VARS-TO on variance and covariance matrices.

The purported much higher efficiency of VARS-TO is very

contentious. The observation that it is more than 100 times more efficient than Sobol'-based total-order indices rests on an exercise with a fully additive model, in which VARS-TO is compared against one of the less accurate total-order estimators [Saltelli et al. (2008), see Puy et al. (2020a)], and the performance measure chosen is the "probability of failure" of properly ranking the model inputs. If the comparison is instead conducted between VARS-TO and a lightweight Sobol'-based first-order estimator [Equation (18)], the advantage of the former shrinks to being "only" 15 times more efficient. VARS-TO completely loses all its edge if the Mean Absolute Error (MAE) replaces the "probability of failure" as a performance measure: in this setting, Equation (18) is the one showing an accuracy up to 100 times higher than VARS-TO.

Nevertheless, the advantage of VARS-TO over Sobol'-based indices is still remarkable if the goal is to rank parameters, and suggests that VARS-TO should be the sensitivity measure of choice if computational efficiency is a priority and the model is additive. However, this condition is unlikely to apply to models of the Earth and Environmental domain, either because they encompass multiplicative terms and exponentials or because their mathematical complexity prevents the analyst from knowing their behavior before running the simulations.

The assertion that VARS-TO is at least 20 times more efficient than Sobol'-based total-order indices is not confirmed by our results. VARS-TO only very slightly outperforms one of the most accurate Sobol' total-order estimators, that of Jansen (1999), when the number of model runs per model input is very small. However, it comes second to Jansen at increasing dimensionalities and in overall performance. Such results have been obtained after randomising the benchmark settings, thus creating a set of sensitivity problems much wider than those represented by the HYMOD and MESH models, and by simultaneously examining the internal uncertainties of VARS-TO (N_{star}, h). Its sampling design makes it especially vulnerable to the high-order interactions between the sampling method (τ), the number of stars (N_{star}), the function dimensionality (k), the distance between pairs (h) and the underlying distribution of the model inputs (φ), especially if they follow a normal distribution.

VARS nonetheless represents a relevant addition to the family of sensitivity analysis methods, with the additional merit of having been developed to appraise the response surface of a model. Furthermore, the conceptual framework of VARS comes with a software described as "next-generation" by Razavi (2019). Time will tell whether VARS ends up unseating Sobol'-based indices as the recommended best practice in sensitivity analysis.

Code availability

Fully documented code is freely available in Puy (2020) and in GitHub (https://github.com/arnaldpuy/VARS_paper).

Data availability

A .csv file of the studies citing VARS as of November 2020 can be found in GitHub (https://github.com/arnaldpuy/VARS_paper).

Declaration of competing interest

We do not have any conflict of interest.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2021.104960.

References

Archer, G.E., Saltelli, A., Sobol, I.M., 1997. Sensitivity measures, anova-like techniques and the use of bootstrap. J. Stat. Comput. Simulat. 58, 99–120.
Baucells, M., Borgonovo, E., 2013. Invariant probabilistic sensitivity analysis. Manag. Sci. 59, 2536–2549.
Becker, W., Saltelli, A., 2015. In: Dean, A., Morris, M., Stufken, J., Bingham, D. (Eds.), Handbook of Design and Analysis of Experiments. CRC Press, Taylor & Francis, Boca Ratón, pp. 627–674.
Becker, W., 2020. Metafunctions for benchmarking in sensitivity analysis. Reliab. Eng. Syst. Saf. 204, 107189.
Borgonovo, E., 2007. A new uncertainty importance measure. Reliab. Eng. Syst. Saf. 92, 771–784.
Borgonovo, E., Plischke, E., 2016. Sensitivity analysis: a review of recent advances. Eur. J. Oper. Res. 248, 869–887.
Cressie, N.A.C., 2015. Statistics for Spatial Data, Revised Edition. Wiley, London.
Döll, P., Siebert, S., 2002. Global modeling of irrigation water requirements. Water Resour. Res. 38, 8-1–8–10.
Eker, S., Rovenskaya, E., Obersteiner, M., Langan, S., 2018. Practice and perspectives in the validation of resource management models. Nat. Commun. 9, 1–10.
Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. 52, 1–17.
Jakeman, A., Letcher, R., Norton, J., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Software 21, 602–614.
Janon, A., Klein, T., Lagnoux, A., Nodet, M., Prieur, C., 2014. Asymptotic normality and efficiency of two Sobol index estimators. ESAIM P. S. 18, 342–364.
Jansen, M., 1999. Analysis of variance designs for model output. Comput. Phys. Commun. 117, 35–43.
Jayathilake, D.I., Smith, T., 2020a. Predicting the temporal transferability of model parameters through a hydrological signature analysis. Front. Earth Sci. 14, 110–123.
Jayathilake, D.I., Smith, T., 2020b. Understanding the role of hydrologic model structures on evapotranspiration-driven sensitivity. Hydrol. Sci. J. 65, 1474–1489.
Kenett, R.S., Shmueli, G., 2013. On information quality. J. Roy. Stat. Soc. A 177, 3–38.
Krogh, S.A., Pomeroy, J.W., Marsh, P., 2017. Diagnosis of the hydrology of a small Arctic basin at the tundra-taiga transition using a physically based hydrological model. J. Hydrol. 550, 685–703.
Lakatos, I., 1976. In: Worrall, J., Zahar, E. (Eds.), Proofs and Refutations: The Logic of Mathematical Discovery. Cambridge University Press.
Lilhare, R., Pokorny, S., Déry, S.J., Stadnyk, T.A., Koenig, K.A., 2020. Sensitivity analysis and uncertainty assessment in water budgets simulated by the variable infiltration capacity model for Canadian subarctic watersheds. Hydrol. Process. 34, 2057–2075.
Liu, H., Chen, W., Sudjianto, A., 2006. Relative entropy based method for probabilistic sensitivity analysis in engineering design. Journal of Mechanical Design, Transactions of the ASME 128, 326–336.
Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D., Saltelli, A., 2020. Variance-based sensitivity analysis: the quest for better estimators between explorativity and efficiency. Reliab. Eng. Syst. Saf. 107300.
Mara, T.A., Belfort, B., Fontaine, V., Younes, A., 2017. Addressing factors fixing setting from given data: a comparison of different methods. Environ. Model. Software 87, 29–38.
Monod, H., Naud, C., Makowski, D., 2006. Uncertainty and sensitivity analysis for crop models. In: Wallach, D., Makowski, D., Jones, J. (Eds.), Working with Dynamic Crop Models. Elsevier, pp. 35–100.
Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174.
Pappenberger, F., Thielen, J., Del Medico, M., 2011. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. Hydrol. Process. 25, 1091–1113.
Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environ. Model. Software 67, 1–11.
Pianosi, F., Wagener, T., 2018. Distribution-based sensitivity analysis from a generic input-output sample. Environ. Model. Software 108, 197–207.
Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D., Caldwell, R., Evora, N., Pellerin, P., 2007. Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth Syst. Sci. 11, 1279–1294.
Plischke, E., 2010. An effective algorithm for computing global sensitivity indices (EASI). Reliab. Eng. Syst. Saf. 95, 354–360.
Puy, A., 2020. R Code of the Paper "Is VARS More Intuitive and Efficient than Sobol'. Zenodo.
Puy, A., Becker, W., Lo Piano, S., Saltelli, A., 2020a. The Battle of Total-Order Sensitivity Estimators arXiv: 2009.01147.
Puy, A., Lo Piano, S., Saltelli, A., 2020b. A sensitivity analysis of the PAWN sensitivity index. Environ. Model. Software 127, 104679.
Razavi, S., 2019. VARS-TOOL: A toolbox for comprehensive, efficient, and robust sensitivity and uncertainty analysis. Environ. Model. Software 112, 95–107.
Razavi, S., Gupta, H.V., 2015. What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models. Water Resour. Res. 51, 3070–3092.
Razavi, S., Gupta, H.V., 2016a. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resour. Res. 52, 423–439.
Razavi, S., Gupta, H.V., 2016b. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application. Water Resour. Res. 52, 440–455.
Saltelli, A., 2002. Sensitivity analysis for importance assessment. Risk Anal. 22, 579–590.

Saltelli, A., 2019. A short comment on statistical versus mathematical modelling. Nat. Commun. 10, 8–10.
Saltelli, A., Tarantola, S., 2002. On the relative importance of input factors in mathematical models: Safety a ssessment for nuclear waste disposal. J. Am. Stat. Assoc. 97, 702–709.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer. John Wiley & Sons, Ltd, Chichester, UK.
Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181, 259–270.
Saltelli, A., Benini, L., Funtowicz, S., Giampietro, M., Kaiser, M., Reinert, E., van der Sluijs, J.P., 2020. The technique is never neutral. How methodological choices condition the generation of narratives for sustainability. Environ. Sci. Pol. 106, 87–98.
Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2002. Sensitivity Analysis in Practice 10-11. John Wiley & Sons, Ltd, Chichester, UK, pp. 1109–1125.
Smith, R.J., 2009. Use and misuse of the reduced major axis for line-fitting. Am. J. Phys. Anthropol. 140, 476–486.
Sobol', I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput. Math. Math. Phys. 7 (4), 86–112.
Sobol', I.M., 1976. Uniformly distributed sequences with an additional uniform property. USSR Comput. Math. Math. Phys. 16 (5), 236–242.
Sobol', I.M., 1993. Sensitivity analysis for nonlinear mathematical models. Mathematical Modeling and Computational Experiment 1, 407–414.
Strong, M., Oakley, J.E., Chilcott, J., 2012. Managing structural uncertainty in health economic decision models: a discrepancy approach. J. Roy. Stat. Soc. C Appl. Stat. 61, 25–45.
Tarantola, S., Giglioli, N., Jesinghaus, J., Saltelli, A., 2002. Can global sensitivity analysis steer the implementation of models for environmental assessments and decisionmaking? Stoch. Environ. Res. Risk Assess. 16, 63–76.
Vieux, B.E., 2016. In: Vieux, B.E. (Ed.), Distributed Hydrologic Modeling Using GIS. Springer Netherlands, Dordrecht, pp. 165–187.
Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour. Res. 39.
Wang, Z., Timlin, D., Kouznetsov, M., Fleisher, D., Li, S., Tully, K., Reddy, V., 2020. Coupled model of surface runoff and surface-subsurface water movement. Adv. Water Resour. 137, 103499.