A randomized Orthogonal Array-based procedure for the estimation of first- and second-order Sobol' indices

Jean-Yves Tissot, Clémentine Prieur

To appear in the *open archive HAL*
Vol. 00, No. 00, Month 20XX, 1–24

A randomized Orthogonal Array-based procedure for the estimation of first- and second-order Sobol' indices

J.- Y. Tissot[a]*, and C. Prieur[a]

[a] *Université de Grenoble, LJK/MOISE BP53 38041 Grenoble cedex, France*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

In variance-based sensitivity analysis, the method of Sobol' [1] allows one to compute Sobol' indices using Monte Carlo integration. One of the main drawbacks of this approach is that estimating Sobol' indices requires a number of simulations which is dependent on the dimension of the model of interest. For example, estimating all the first- or second-order Sobol' indices of a d-dimensional function basically requires $d+1$ or $(d+1)d/2$ independent input vectors, respectively. Some interesting combinatorial results have been introduced to weaken this defect, in particular by Saltelli [2] and more recently by Owen [3], but the quantities they estimate still depend linearly on the dimension d. In this paper, we introduce a new approach to estimate all the first- and second-order Sobol' indices by using only 2 input vectors. We establish theoretical properties of such a method for the estimation of first-order Sobol' indices and discuss the generalization to higher-order indices. In particular, we prove on numerical examples that this procedure is tractable and competitive for the estimation of all first- and second-order Sobol' indices. As an illustration, we propose to apply this new approach to a marine ecosystem model of the Ligurian sea (northwestern Mediterranean) in order to study the relative importance of its several parameters. The calibration process of this kind of chemical simulators is well-known to be quite intricate, and a rigorous and robust — i.e. valid without strong regularity assumptions — sensitivity analysis, as the method of Sobol' provides, could be of great help. This article has supplementary material online.

Keywords: global sensitivity analysis; variance-based sensitivity indices; numerical integration; orthogonal arrays

1. Introduction and notation

Sobol' indices (SI) [1] are quantities defined by normalizing variance components in an ANOVA decomposition [1, 4, 5]. They are an important tool to study the sensitivity of a model output subject to the input parameters since they allow to quantify the relative importance of input factors of a function over their entire range of values. As Sobol' indices essentially consist of integrals, their computation can become rapidly expensive when the number of factors increases. In addition to the method of Sobol', many techniques have been proposed to estimate these indices. They include Fast Amplitude Sensitivity Test (FAST) [6], Random Balance Design (RBD) [7], Bayesian techniques [8], spectral methods based on polynomial chaos expansion [9], or other metamodel-based techniques [10]. A recent review of these methods can be found in Saltelli *et al.* [11], and more specifically a new introduction to FAST and RBD has been recently provided by Tissot *et al.* [12].

Spectral methods — such as FAST, RBD or polynomial chaos expansion-based methods — which exploit the spectral decomposition of the model with respect to a particular

*Corresponding author. Email: jeanyvestissot@free.fr

multivariate basis, can improve the rate of convergence for the estimation of SIs. However, it is only true under strong assumptions on the spectral decomposition of the model itself such as a decay of the spectrum sufficiently fast, the negligibility of high-order spectral coefficients, etc. As a result, these methods are not robust to complex phenomena such as high-frequency variations or discontinuities, and so the method of Sobol' appears as the main method one can trust when no strong a priori knowledge on the model of interest is available.

Let f be a real square integrable function defined on $\mathbb{R}^d$ and $\mathbf{X} = (X_1, \ldots, X_d)$ a random vector with independent components arbitrarily distributed on $\mathbb{R}$. Then let us consider the real random variable $Y = f(\mathbf{X})$ and for any $\mathfrak{u} \subseteq \{1, \ldots, d\}$, denote by $\mathbf{X}_\mathfrak{u}$ the random vector with components X_i, $i \in \mathfrak{u}$. The ANOVA decomposition states that Y can be uniquely decomposed into summands of increasing dimensions

$$Y = \sum_{\mathfrak{u} \subseteq \{1,\ldots,d\}} f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) \tag{1}$$

where the summands in Equation (1) are orthogonal that is, $\mathbb{E}\left(f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) f_\mathfrak{v}(\mathbf{X}_\mathfrak{v})\right) = 0$, for any $\mathfrak{u} \neq \mathfrak{v} \subseteq \{1, \ldots, d\}$. In particular, the sum of functions

$$f_\emptyset + f_1(X_1) + f_2(X_2) + \cdots + f_d(X_d) \tag{2}$$

is the so-called additive part of f, where the constant $f_\emptyset = \mathbb{E}[Y]$ and the random variables $f_i(X_i) = \mathbb{E}[Y|X_i] - \mathbb{E}[Y]$ are the components of lower complexity. The components $f_\mathfrak{u}(\mathbf{X}_\mathfrak{u})$ are explicitely known in terms of conditional expectation. As already mentioned, we have $f_\emptyset = \mathbb{E}[Y]$ and the other terms are recursively defined by

$$f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) = \mathbb{E}[Y|\mathbf{X}_\mathfrak{u}] - \sum_{\mathfrak{v} \subset \mathfrak{u}} f_\mathfrak{v}(\mathbf{X}_\mathfrak{v}).$$

Then, setting $\mathrm{Var}[Y] = \sigma^2$ and $\mathrm{Var}[f_\mathfrak{u}(\mathbf{X}_\mathfrak{u})] = \sigma^2_\mathfrak{u}$, Eq. (1) gives $\sigma^2 = \sum_{\mathfrak{u} \subseteq \{1,\ldots,d\}} \sigma^2_\mathfrak{u}$ and the SIs — also known as global sensitivity indices — are defined as

$$S_\mathfrak{u} = \frac{\sigma^2_\mathfrak{u}}{\sigma^2}, \quad \mathfrak{u} \subseteq \{1, \ldots, d\}.$$

If $\mathfrak{u} = \{i\}$, $S_\mathfrak{u}$ quantifies the main effect due to the factor X_i, and if $|\mathfrak{u}| = \mathrm{Card}(\mathfrak{u}) > 1$, $S_\mathfrak{u}$ quantifies the interaction effect between the factors X_i, $i \in \mathfrak{u}$. Note that $S_\emptyset$ is trivially equal to zero. Similarly one can define $\underline{\tau}^2_\mathfrak{u} = \mathrm{Var}\big[\mathbb{E}[Y|\mathbf{X}_\mathfrak{u}]\big]$ and consider the quantities

$$\underline{S}_\mathfrak{u} = \frac{\underline{\tau}^2_\mathfrak{u}}{\sigma^2}, \quad \mathfrak{u} \subseteq \{1, \ldots, d\}$$

known as lower SIs or closed sensitivity indices. If $\mathfrak{u} = \{i\}$ then $\underline{S}_\mathfrak{u} = S_\mathfrak{u}$, and if $\mathrm{Card}(\mathfrak{u}) > 1$, $\underline{S}_\mathfrak{u}$ quantifies the sum of all the main and interaction effects due to any group of factors $\mathfrak{v} \subseteq \mathfrak{u}$. Note that for any $1 \leq r \leq d$, the knowledge of $\{\underline{S}_\mathfrak{u},\ \mathrm{Card}(\mathfrak{u}) \leq r\}$ or $\{S_\mathfrak{u},\ \mathrm{Card}(\mathfrak{u}) \leq r\}$ are strictly equivalent since, on the one hand, we have $\underline{S}_\mathfrak{u} = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} S_\mathfrak{v}$ and on the other hand, the Möbius inversion formula [see, e.g., 13] gives $S_\mathfrak{u} = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \underline{S}_\mathfrak{v}$.

In practice global sensitivity analysis focuses on the first- or second-order — i.e. $\mathrm{Card}(\mathfrak{u}) = 1$ or 2, respectively — terms. In the present paper, we only focus on the indices $\underline{S}_\mathfrak{u}$ which are the quantities estimated by the method of Sobol'.

The paper proceeds as follows. Section 2 provides a short review of the method of Sobol', gives some notation and explains the main idea of the new method we propose in the present paper. In Section 3, theory is presented including asymptotic properties related to both Latin hypercube sampling (LHS) and replicated Latin hypercube sampling (RLHS). In this section, we also discuss potential generalizations to using randomized orthogonal arrays [see 14, 15]. Numerical illustrations are provided in Section 4, and Section 5 draws conclusions. Detailed proofs of the technical results are presented in an online supplementary document.

2. Background

Like any numerical integration technique, the method of Sobol' can be viewed as the particular combination of a design of experiments (DOE) — i.e. the location the information is collected — and an estimator — i.e. the way the collected information is processed. In this section, we first describe the DOEs of interest and we then come to the definitions of some currently used estimators in the method of Sobol'. Finally, we introduce the concept of RLHS and we explain why this kind of DOE is of great importance in the issue of estimating first-order SIs.

From now on, we assume that $X_1,\dots, X_d$ are independent random variables uniformly distributed on $[0,1]$. However thanks to the *inversion method* [see, e.g., 16] a model $Y = f(X_1,\dots X_d)$ where the X_is are arbitrarily distributed can always be transformed into a model $Y = g(U_1,\dots,U_d)$ where the U_is are random variables uniformly distributed on $[0,1]$, defined by $U_i = F_i(X_i)$ with F_i the cumulative distribution function of X_i. Thus the assumption on the X_is is not restrictive.

2.1 *Designs of experiments*

In the present paper a design of experiments refers to a finite subset of $[0,1]^d$. We do not consider deterministic constructions of DOEs but only random ones, i.e. as in a Monte Carlo method, DOEs consist of n realizations of a set of random d-dimensional vectors.

Estimating a lower SI using the method of Sobol' typically requires $2n$ evaluations of the model. However there exist more sophisticated estimators that may require $3n$ or more evaluations, see e.g. [17]. In the present paper, we focus on the most basic estimators that only require $2n$ simulations. In this case, the first of any double evaluation is a realization of the random variable $Y = f(X_1,\dots,X_d)$, and the complementary evaluation is obtained from the first one by resampling its components indexed by the elements of $\mathfrak{u}^c$. In other words, the complementary evaluation is a realization of the random variable denoted by $Y_{\mathfrak{u}}$ and defined by

$$Y_{\mathfrak{u}} = f(\mathbf{X}_{\mathfrak{u}}, \mathbf{Z}_{\mathfrak{u}^c}) \tag{3}$$

where $\mathbf{Z}$ is a d-dimensional vector uniformly distributed on $[0,1]^d$, and for all i in $\{1,\dots,d\}$

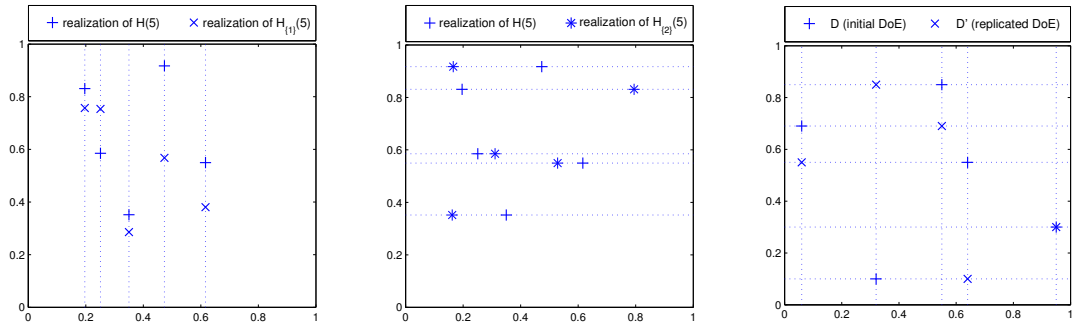
$$(\mathbf{X}_{\mathfrak{u}}, \mathbf{Z}_{\mathfrak{u}^c})_i = \begin{cases} X_i & \text{if } i \in \mathfrak{u} \\ Z_i & \text{otherwise.} \end{cases} \tag{4}$$

Hence the definition of the design of experiments of size $N = 2n$ for estimating $\underline{S}_{\mathfrak{u}}$, denoted by $D_{\mathfrak{u}}(N)$, proceeds as follows. First let $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{Z}^j)_{j=1..n}$ be independent replications of the random vectors $\mathbf{X}$ and $\mathbf{Z}$, respectively, i.e. Monte Carlo sampling. Then

denote the two halves of the main DOE by

$$H(n) = \left\{\mathbf{X}^j,\ 1 \leq j \leq n\right\}$$
$$H_{\mathrm{u}}(n) = \left\{(\mathbf{X}^j_{\mathrm{u}}, \mathbf{Z}^j_{\mathrm{u}^c}),\ 1 \leq j \leq n\right\} \quad \text{(see the definition in Eq.(4))},$$

and define $D_{\mathrm{u}}(N)$ to be the resulting union of both sets. The sampling plans obtained from this pick-freeze procedure are known as *plans based on substituted columns* (see e.g. Morris *et al.* [18]), and they can be used e.g. with the estimators presented in Section 2.2 (see also Janon *et al.* [19]). Figure 1 (a) and (b) show illustrations of such DOEs for estimating first-order SIs in a 2-dimensional space. Note that in this figure $D_{\{1\}}(10)$ and $D_{\{2\}}(10)$ contain two points per level of the first and second axis, respectively; this consists of the main property — or constraint — of the DOE in the issue of estimating SIs using the method of Sobol'.



(a) realization of $D_{\{1\}}(10)$ (b) realization of $D_{\{2\}}(10)$ (c) D: Latin hypercube

Figure 1. Designs of experiments of size $N = 10$: (a) and (b) are *plans based on substituted columns* $D_{\{1\}}(10)$ and $D_{\{2\}}(10)$ for the respective estimation of $\underline{S}_1$ and $\underline{S}_2$, (c) is the composite DOE based on replicated Latin hypercubes which allows the estimation of both $\underline{S}_1$ and $\underline{S}_2$.

2.2 *Estimators*

Using the previous notation, we now consider, for any j in $\{1, \ldots, n\}$, the output observations

$$Y^j = f(\mathbf{X}^j) \quad \textit{and} \quad Y^j_{\mathrm{u}} = f(\mathbf{X}^j_{\mathrm{u}}, \mathbf{Z}^j_{\mathrm{u}^c}). \tag{5}$$

In the method of Sobol', the estimation of the index $\underline{S}_{\mathrm{u}}$ consists in applying a Monte Carlo (MC) method to both the numerator and denominator of

$$\underline{S}_{\mathrm{u}} = \frac{\mathrm{Var}\left[\mathbb{E}[Y|\mathbf{X}_{\mathrm{u}}]\right]}{\mathrm{Var}[Y]}$$

that can be rewritten (see e.g. [19]), using the notation in Eq. (3), as

$$\underline{S}_{\mathrm{u}} = \frac{\mathrm{Cov}(Y, Y_{\mathrm{u}})}{\mathrm{Var}[Y]} = \frac{\mathbb{E}[YY_{\mathrm{u}}] - \mathbb{E}[Y]\mathbb{E}[Y_{\mathrm{u}}]}{\mathrm{Var}[Y]}. \tag{6}$$

As already mentioned in this section, several estimators have been introduced to perform this numerical integration. In the present paper, we only consider the natural estimator

coming from (6), and an other one due to Monod *et al.* [20]. Using the notation in (5), they are defined by $\widetilde{\underline{S}}_{\mathrm{u},n} = \widetilde{\underline{\tau}}^2_{\mathrm{u},n}/\widetilde{\sigma}^2_n$ where

$$\widetilde{\underline{\tau}}^2_{\mathrm{u},n} = \frac{1}{n}\sum_{j=1}^{n} Y^j Y^j_{\mathrm{u}} - \left(\frac{1}{n}\sum_{j=1}^{n} Y^j\right)\left(\frac{1}{n}\sum_{j=1}^{n} Y^j_{\mathrm{u}}\right) \text{ and } \widetilde{\sigma}^2_n = \frac{1}{n}\sum_{j=1}^{n}\left(Y^j\right)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} Y^j\right)^2 \quad (7)$$

and $\widehat{\underline{S}}_{\mathrm{u},n} = \widehat{\underline{\tau}}^2_{\mathrm{u},n}/\widehat{\sigma}^2_n$ where

$$\widehat{\underline{\tau}}^2_{\mathrm{u},n} = \frac{1}{n}\sum_{j=1}^{n} Y^j Y^j_{\mathrm{u}} - \left(\frac{1}{2n}\sum_{j=1}^{n}\left(Y^j + Y^j_{\mathrm{u}}\right)\right)^2 \quad (8)$$

$$\text{and} \quad \widehat{\sigma}^2_n = \frac{1}{2n}\sum_{j=1}^{n}\left(\left(Y^j\right)^2 + \left(Y^j_{\mathrm{u}}\right)^2\right) - \left(\frac{1}{2n}\sum_{j=1}^{n}\left(Y^j + Y^j_{\mathrm{u}}\right)\right)^2, \quad (9)$$

respectively.

2.3 *Motivation to use replicated Latin hypercube sampling*

Following the previous description of both the DOEs (Section 2.1) and the estimators (Section 2.2), it is easy to understand that estimating all the first-order SIs using this technique requires $n(d+1)$ evaluations of the model, i.e. $d+1$ DOEs each containing n points. More generally, for any k in $\{1,\ldots,d\}$, estimating all the k-th order SIs basically requires $n\left(\binom{d}{k}+1\right)$ evaluations.

Some interesting combinatorial results have been introduced to weaken this defect. For instance, Saltelli [2] shows how to estimate all the second-order SIs by using only $n(2d+2)$ evaluations. In a more recent work, Owen [3] deeply studies these combinatorial aspects. However the quantities of interest estimated in both these papers still require $\mathcal{O}(d)$ evaluations of the model, and this dependency with respect to the dimension d appears as a major drawback in the sensitivity analysis of high-dimensional models potentially depending on more than one hundred input parameters.

We propose in this paper a new approach based on replicated Latin hypercube sampling, which overcomes this issue of dimensionality for the estimation of first-order Sobol' indices. Our approach generalizes to the estimation of higher-order Sobol' indices. However, the generalization to the estimation of higher order interactions requires a constraint linking the dimension d, the strength t of the interactions under study and the number of model evaluations required. This constraint is stated in Section 3.3 (see Inequality (15)). We discuss this constraint in the case of second-order interactions in Remark 3.4. In Section 4, our approach is numerically compared with the estimation procedure presented in Saltelli [2] and with quasi-Monte Carlo approaches.

In the following, if $H(n) = \left\{\mathbf{X}^j = (X^j_1,\ldots,X^j_d),\ 1 \le j \le n\right\}$ is any DOE, we say that a DOE $H'(n)$ is replicated from $H(n)$ if $H'(n)$ is obtained through a column-wise random permutation of $H(n)$. More precisely it means that $H'(n) = \left\{\mathbf{X}'^j = (\pi_1(X^j_1),\ldots,\pi_d(X^j_d)),\ 1 \le j \le n\right\}$, where the π_is are d independent random permutations uniformly distributed on the set of the values $\{X^j_i,\ 1 \le j \le n\}$ for any $i = 1,\ldots,d$. Note that this notion is clearly symmetric, so we can also say that $H(n)$ is replicated from $H'(n)$ or even that $H(n)$ and $H'(n)$ are replicated from each other. As a result, for any axis k in $\{1,\ldots,d\}$, the union of $H(n)$ and one of its replicates $H'(n)$ contains two points per level of the k-th coordinate, see Figure 1 (c). So for any axis k, this composite DOE allows one to estimate the index $\underline{S}_k$. In other words, we obtain a DOE that allows

one to estimate all the first-order SIs using only $2n$ evaluations of the model. These sampling plans are known as *plans based on permuted columns* (see e.g. McKay [21] or Morris *et al.* [18], PC plans). In MKay [21], or Morris *et al.* [18, 22], an arbitrary number r of replications of the initial design $H(n)$ is used to define an estimator of first-order Sobol' indices (see e.g. the definition of this estimator in Section 2.2 of the paper by Morris *et al.* [22]). This estimator is different in nature from ours. The idea to use the estimator defined by (7) in Section 2.2 of the present paper with two half-designs $H(n)$ and $H'(n)$ which are obtained by PC sampling was first introduced by Mara *et al.* [23] for a numerical study on the estimation of first-order Sobol' indices.

3. Theory

Both estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$ introduced in the previous section have strong statistical properties. It is easy to prove that they are *strongly consistent* — i.e. they converge almost surely to the theoretical value $\underline{S}_{\mathfrak{u}}$ — and that the *biases* of both their numerator and denominator are $O(n)$:

$$\mathbb{E}\big[\widetilde{\underline{\tau}}^2_{\mathfrak{u},n}\big] = \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{n}\underline{\tau}^2_{\mathfrak{u}} \quad \textit{and} \quad \mathbb{E}\big[\widetilde{\sigma}^2_n\big] = \sigma^2 - \frac{1}{n}\sigma^2$$

and, as shown by Owen [3],

$$\mathbb{E}\big[\widehat{\underline{\tau}}^2_{\mathfrak{u},n}\big] = \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{2n}(\sigma^2 + \underline{\tau}^2_{\mathfrak{u}}) \quad \textit{and} \quad \mathbb{E}\big[\widehat{\sigma}^2_n\big] = \sigma^2 - \frac{1}{2n}(\sigma^2 + \underline{\tau}^2_{\mathfrak{u}}).$$

But the most important property, proved by Janon *et al.* [19], is that they are *asymptotically normal*. In particular this allows one to derive asymptotic confidence intervals which provide probabilistic bounds on the estimation error.

Now the question that naturally arises is

"Does the new approach based on RLHS we proposed for estimating all the first-order Sobol' indices still have these strong statistical properties?"

The answer is affirmative to the three previous points of interest and we prove it in this section. First we begin by stating important results of convergence on LHS — see Proposition 3.1 in Section 3.1 — as well as on RLHS — see Proposition 3.2 in Section 3.2 — and we then formulate the main result on the estimation of all the first-order SIs using only two RLHS in Theorem 3.1 in Section 3.2. In a final subsection, we address the issue of potential generalizations to randomized OA.

3.1 *On Latin hypercube sampling*

We now assume that the $\mathbf{X}^j$s and the $\mathbf{Z}^j$s are no longer independent replications of $\mathbf{X}$ and $\mathbf{Z}$, but we consider that $\{\mathbf{X}^j,\ 1 \leq j \leq n\}$ and $\{\mathbf{Z}^j,\ 1 \leq j \leq n\}$ are two independent Latin hypercubes of size n. We then prove that both estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$ introduced in the previous section still have the statistical properties presented above. We first introduce the definition of a Latin hypercube:

Definition 3.1 Let d and n in $\mathbb{N}^*$, and consider Π_n the set of all the permutations of $\{1, \ldots, n\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ is a Latin hypercube of size n in $[0,1]^d$ — and we

denote $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ — if for all $j \in \{1,\dots,n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \dots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n} \right) \tag{10}$$

where the π_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n and $[0,1]$, respectively.

Let now $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ and $(\mathbf{Z}^j)_j \sim \mathcal{LH}(n,d)$ be two independent Latin hypercubes of size n in $[0,1]^d$, and for any $\mathfrak{u} \subseteq \{1,\dots,d\}$ let $Y^{j,LHS} = f(\mathbf{X}^j)$ and $Y_{\mathfrak{u}}^{j,LHS} = f(\mathbf{X}_{\mathfrak{u}}^j, \mathbf{Z}_{\mathfrak{u}^c}^j)$.

The resulting estimators are now denoted $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS} = \widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS} / \widetilde{\sigma}_n^{2,LHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{LHS} = \widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS} / \widehat{\sigma}_n^{2,LHS}$, respectively. Their statistical properties are gathered in the following result:

PROPOSITION 3.1
(i) *If f^4 is integrable then $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{LHS}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS} - \underline{S}_{\mathfrak{u}})$ and $\sqrt{n}(\widehat{\underline{S}}_{\mathfrak{u},n}^{LHS} - \underline{S}_{\mathfrak{u}})$ converge in law to a zero-mean normal distribution with lower variance than the respective variance given in the central limit theorem (CLT) for the basic estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$.*

(iii) *We have*

$$\mathbb{E}\big[\widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS}\big] = \underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} \quad \textit{and} \quad \mathbb{E}\big[\widetilde{\sigma}_n^{2,LHS}\big] = \sigma^2 + B_{n,2}$$
$$\mathbb{E}\big[\widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS}\big] = \underline{\tau}_{\mathfrak{u}}^2 + B_{n,3} \quad \textit{and} \quad \mathbb{E}\big[\widehat{\sigma}_n^{2,LHS}\big] = \sigma^2 + B_{n,3}$$

where

$$-\frac{1}{n-1}\underline{\tau}_{\mathfrak{u}}^2 \le B_{n,1} \le 0\,,\ -\frac{1}{n-1}\sigma^2 \le B_{n,2} \le 0\,,\ -\frac{1}{2(n-1)}(\sigma^2 + \underline{\tau}_{\mathfrak{u}}^2) \le B_{n,3} \le 0\ .$$

Remark 3.1 Due to their intricate structure, the biases of the estimators $\widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS}$, $\widetilde{\sigma}_n^{2,LHS}$, $\widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,LHS}$ and $\widehat{\sigma}_n^{2,LHS}$ can't be easily reduced. Nevertheless we can note that these biases are asymptotically negligible, with a rate of convergence in $O(n^{-1})$ larger than the rate of convergence of the estimators — to their theoretical values — themselves, which is in $O(n^{-1/2})$.

To conclude, we have proven that the estimation of SIs combining MC estimators and LHS has strong statistical properties. In particular, estimators in such a method have similar — and potentially smaller — bias than the classic method based on simple random sampling and asymptotically smaller variance.

3.2 *On replicated Latin hypercube sampling*

We now come to alternative estimators based on RLHS. First in Proposition 3.2, we present a technical result showing that such estimators still have the strong statistical properties of the former estimators. We then prove that only two RLHS are necessary to estimate all the first-order Sobol' indices and that the estimators defined in this efficient strategy have the same statistical properties as in Proposition 3.2.

We begin with the definition of a replicated Latin hypercube:

Definition 3.2 Let d and n in $\mathbb{N}^*$, and consider Π_n the set of all the permutations of $\{1,\dots,n\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{X}'^j)_{j=1..n}$ are two replicated Latin hypercubes of size n in $[0,1]^d$ — and we denote $(\mathbf{X}^j,\mathbf{X}'^j)_j \sim \mathcal{RLH}(n,d)$ — if for all $j\in\{1,\dots,n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(j)-U_{1,\pi_1(j)}}{n},\dots,\frac{\pi_d(j)-U_{d,\pi_d(j)}}{n}\right)$$

and

$$\mathbf{X}'^j = \left(\frac{\pi'_1(j)-U_{1,\pi'_1(j)}}{n},\dots,\frac{\pi'_d(j)-U_{d,\pi'_d(j)}}{n}\right)$$

where the π_i's, the π'_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n, Π_n and $[0,1]$, respectively.

Let now $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ be a Latin hypercube of size n in $[0,1]^d$ and $(\mathbf{Z}^j,\mathbf{Z}'^j)_j \sim \mathcal{RLH}(n,d)$ be two replicated Latin hypercubes of size n in $[0,1]^d$, and for any $\mathfrak{u}\subseteq\{1,\dots,d\}$ define

$$Y^{j,RLHS} = f(\mathbf{X}^j_{\mathfrak{u}},\mathbf{Z}^j_{\mathfrak{u}^c}) \quad \textit{and} \quad Y^{j,RLHS}_{\mathfrak{u}} = f(\mathbf{X}^j_{\mathfrak{u}},\mathbf{Z}'^j_{\mathfrak{u}^c}). \tag{11}$$

Note that $Y^{j,RLHS}$ actually depends on $\mathfrak{u}$, but the essential constraint stating that the random vectors $(\mathbf{X}^j_{\mathfrak{u}},\mathbf{Z}^j_{\mathfrak{u}^c})$ and $(\mathbf{X}^j_{\mathfrak{u}},\mathbf{Z}'^j_{\mathfrak{u}^c})$ have the same components in their i-th coordinate, for all $i\in\mathfrak{u}$ is still checked. However for convenience, we omit $\mathfrak{u}$ in the notation.

The resulting estimators are now denoted by $\widetilde{\underline{S}}^{RLHS}_{\mathfrak{u},n} = \widetilde{\underline{\tau}}^{2,RLHS}_{\mathfrak{u},n}/\widetilde{\sigma}^{2,RLHS}_n$ and $\widehat{\underline{S}}^{RLHS}_{\mathfrak{u},n} = \widehat{\underline{\tau}}^{2,RLHS}_{\mathfrak{u},n}/\widehat{\sigma}^{2,RLHS}_n$, respectively. Note that for the estimators of SIs based on r replicated Latin hypercubes introduced by McKay [21] — see also the summarized presentation by Saltelli *et al.* [24] — no rigorous theoretical study has been proposed, except for the issue of the bias. Morris *et al.* [22] have indeed proposed a sampling procedure which allows to construct an unbiased estimator for each first-order Sobol' index. We imagine that an asymptotic study for the estimator in McKay [21] would be based on r and n growing to infinity. However our procedure is different and just involves $r=2$ replicated LHS. Thus the asymptotic is only formulated as n tends to infinity, where n is the size of the initial half-design.

The statistical properties of $\widetilde{\underline{S}}^{RLHS}_{\mathfrak{u}}$ and $\widehat{\underline{S}}^{RLHS}_{\mathfrak{u}}$ are gathered in the following result:

Proposition 3.2
(i) *If f^4 is integrable then $\widetilde{\underline{S}}^{RLHS}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}^{RLHS}_{\mathfrak{u},n}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}^{RLHS}_{\mathfrak{u},n}-\underline{S}_{\mathfrak{u}})$ and $\sqrt{n}(\widehat{\underline{S}}^{RLHS}_{\mathfrak{u},n}-\underline{S}_{\mathfrak{u}})$ converge in law to a zero-mean normal distribution with the same respective variance given in CLT for the estimators $\widetilde{\underline{S}}^{LHS}_{\mathfrak{u},n}$ and $\widetilde{\underline{S}}^{LHS}_{\mathfrak{u},n}$.*

(iii) *We have*

$$\mathbb{E}\big[\widetilde{\underline{\tau}}^{2,RLHS}_{\mathfrak{u},n}\big] = \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{n}\underline{\tau}^2_{\mathfrak{u}} + B_{n,1} + B_{|\mathfrak{u}|,n} \quad \textit{and} \quad \mathbb{E}\big[\widetilde{\sigma}^{2,RLHS}_n\big] = \sigma^2 + B_{n,3}$$

$$\mathbb{E}\big[\widehat{\underline{\tau}}^{2,RLHS}_{\mathfrak{u},n}\big] = \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{2n}\underline{\tau}^2_{\mathfrak{u}} + B_{n,1} + B_{n,2} + B_{|\mathfrak{u}|,n} \quad \textit{and} \quad \mathbb{E}\big[\widehat{\sigma}^{2,RLHS}_n\big] = \sigma^2 - \frac{1}{2n}\underline{\tau}^2_{\mathfrak{u}} + B_{n,1} + B_{n,2}$$

where

$$
\begin{aligned}
|B_{n,1}| &\leq \left(\frac{d+1}{n}+2\right)\left(\frac{d+1}{n}\right)\mathbb{E}[Y^2]\,,\\
|B_{n,2}| &\leq \frac{\sigma^2}{2n}\,, \quad -\frac{1}{n-1}\sigma^2 \leq B_{n,3} \leq 0\,,\\
|B_{|\mathfrak{u}|,n}| &\leq \left(\frac{d-|\mathfrak{u}|+1}{n}+2\right)\left(\frac{d-|\mathfrak{u}|+1}{n-1}\right)\mathbb{E}[Y^2]\ .
\end{aligned}
$$

Using this technical result, it is possible to estimate all the first-order Sobol' indices with only two replicated Latin hypercubes (RLH) and in addition the new estimators inherit the strong statistical properties stated in Proposition 3.2. The main idea consists in defining, for any $i \in \{1,\dots,d\}$, the two samples Y^j and $Y^j_{\{i\}}$ — necessary to estimate $\underline{S}_{\{i\}}$ — by only considering two RLH. To this end, we first consider $(\mathbf{X}^j, \mathbf{X}'^j)_j$ two replicated Latin hypercubes of size n in $[0,1]^d$ defined by

$$
\mathbf{X}^j = \left(\frac{\pi_1(j)-U_{1,\pi_1(j)}}{n},\dots,\frac{\pi_d(j)-U_{d,\pi_d(j)}}{n}\right),
$$

$$
\mathbf{X}'^j = \left(\frac{\pi'_1(j)-U_{1,\pi'_1(j)}}{n},\dots,\frac{\pi'_d(j)-U_{d,\pi'_d(j)}}{n}\right).
$$

Secondly, let $\pi \in \Pi_n$ be a random permutation independent from the π_is and the π'_is, and for any $i \in \{1,\dots,d\}$ and $j \in \{1,\dots,n\}$ consider

$$
Y^j = f(\mathbf{X}^{\pi_i^{-1}\circ\pi(j)}) \quad \textit{and} \quad Y^j_{\{i\}} = f(\mathbf{X}'^{\pi_i'^{-1}\circ\pi(j)}), \tag{12}
$$

where the symbol $\circ$ denotes the composition of two functions. In this method, the DOE used to estimate $\underline{S}_{\{i\}}$ is

$$
D_{\{i\}}(2n) = \{\mathbf{X}^{\pi_i^{-1}\circ\pi(j)},\ 1\leq j\leq n\} \cup \{\mathbf{X}'^{\pi_i'^{-1}\circ\pi(j)},\ 1\leq j\leq n\} \tag{13}
$$

and viewed as non-ordered set, it does not depend on i and is equal to

$$
D(2n) = \{\mathbf{X}^j,\ 1\leq j\leq n\} \cup \{\mathbf{X}'^j,\ 1\leq j\leq n\}. \tag{14}
$$

As a result, we can construct all the d estimators $\left(\widetilde{\underline{S}}_i\right)_{i=1..n}$ as defined in (7) — resp. $\left(\widehat{\underline{S}}_i\right)_{i=1..n}$ as defined in (9) — by only requiring the $2n$ evaluations of the model f on the DOE $D(2n)$. Then we have the following result.

THEOREM 3.1 *Consider $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{RLH}(n,d)$ and denote $D(2n)$ the union of both these replicated Latin hypercubes as defined in* (14)*. For any $i \in \{1,\dots,d\}$, consider the evaluations of the function f on the reordered set $D_{\{i\}}(2n)$ as defined in* (13)*, and denote them by $(Y^j)_{j=1..n}$ and $(Y^j_{\{i\}})_{j=1..n}$ as presented in* (12)*. Then the estimator $\widetilde{\underline{S}}_{\{i\}}$ as defined in* (7) *— resp. $\widehat{\underline{S}}_{\{i\}}$ as defined in* (9) *— built with Y^j and $Y^j_{\{i\}}$ is strongly consistent, asymptotically normal and has numerator and denominator with bias as stated in (iii) of Proposition 3.2.*

For the proof see the Appendix A.

3.3 *Potential generalization to randomized orthogonal arrays*

The main question that arises given the result stated in Theorem 3.1 is: can we estimate all the k-th order Sobol' indices with only two RLHS? On the one hand, the most straightforward answer is clearly negative since RLHS do not have the required strcuture to handle these higher-order Sobol' indices. On the other hand, we have to observe that such a well-suited structure can be built by using orthogonal arrays. So we first begin with the definition of an orthogonal array (OA):

Definition 3.3 An OA in dimension d, with q levels, strength $t \leq d$ and index λ is a matrix with $n = \lambda q^t$ rows and d columns such that in every n-by-t submatrix each of the q^t possible rows — i.e. the distinct t-tuples $(l_1, \ldots, l_t)$ where the l_i's take their values in the set of the q levels — occurs exactly the same number λ of times.

We now recall the definition of a randomized OA [see 14, 15] and introduce the general notion of replicated randomized OA.

Definition 3.4 Let $(A_i^j)_{i=1..d, j=1..n}$ be an OA in dimension d, with n points and q levels in $\{1, \ldots, q\}$, and consider Π_q the set of all the permutations of $\{1, \ldots, q\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ is a randomized OA $(\mathbf{A}^j)_{j=1..n}$ — and we denote $(\mathbf{X}^j)_j \sim \mathcal{LH}\big((\mathbf{A}^j)_j\big)$ — if for all $j \in \{1, \ldots, n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \ldots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q} \right)$$

where the π_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_q and $[0, 1]$, respectively.

Definition 3.5 Let $(A_i^j)_{i=1..d, j=1..n}$ an OA in dimension d, with n points and q levels in $\{1, \ldots, q\}$, and consider Π_q the set of all the permutations of $\{1, \ldots, q\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{X}'^j)_{j=1..n}$ are two replicated randomized orthogonal array $(\mathbf{A}^j)_{j=1..n}$ — and we denote $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{ROA}\big((\mathbf{A}^j)_j\big)$ — if for all $j \in \{1, \ldots, n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \ldots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q} \right),$$

$$\mathbf{X}'^j = \left(\frac{\pi'_1(A_1^j) - U_{1,\pi'_1(A_1^j)}}{q}, \ldots, \frac{\pi'_d(A_d^j) - U_{d,\pi'_d(A_d^j)}}{q} \right),$$

where the π_i's, the π'_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_q, Π_q and $[0, 1]$, respectively.

It is interesting to note that in the particular case of the OA $(\mathbf{A}^j)_{j=1}$ with strength 1 and index unity defined by $\forall i \in \{1, \ldots d\},\ \forall j \in \{1, \ldots n\},\ \ A_i^j = j$, these definitions are exactly Definitions 3.1 and 3.2.

Remark 3.2 Let $(A_i^j)_{1\leq i\leq d,\, 1\leq j\leq n}$ be an orthogonal array in dimension d, with q levels, strength $t \leq d$ and index λ on a q-set of elements, then $(\pi_i(A_i^j))_{1\leq i\leq d,\, 1\leq j\leq n}$ is an OA in dimension d, with q levels, strength $t \leq d$ and index λ on the same q-set of elements.

In this section we consider $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{ROA}((\mathbf{A}^j)_j)$ where $(A_i^j)_{i=1..d, j=1..n}$ is an OA in dimension d, with n points, of strength t and index unity, and denote $D(2n)$ the DOE defined as the union of both these replicated randomized OA: $D(2n) = \{\mathbf{X}^j,\ 1 \leq j \leq n\} \cup \{\mathbf{X}'^j,\ 1 \leq j \leq n\}$.

Then thanks to Definition 3.5 and to Remark 3.2, for any t-tuple of indices $(i_1, \dots i_t)$, there exists a unique permutation $\pi_{i_1,\dots,i_t}$ such that columns $i_1, \dots, i_t$ in $\{\mathbf{X}^j,\ 1 \leq j \leq n\}$ and $\{\mathbf{X}'^{\pi_{i_1,\dots,i_t}(j)},\ 1 \leq j \leq n\}$ are identical. We can thus estimate the t-th order Sobol' index $\underline{S}_{\mathfrak{u}}$, where $\mathfrak{u} = (i_1, \dots, i_t)$, with the DOE $\{\mathbf{X}^j,\ 1 \leq j \leq n\} \cup \{\mathbf{X}'^{\pi_{i_1,\dots,i_t}(j)},\ 1 \leq j \leq n\}$. Once more we can remark that this DOE, as a non-ordered set, does not depend on $(i_1, \dots, i_t)$ and is equal to $D(2n)$. We deduce then that all the t-th order Sobol' indices can be estimated by using only the DOE $D(2n)$. However, due to the constraints on the construction of OAs, this method can not be applied to any values of n. As we note in Remark 3.4 below, the choice of the number of evaluations n depends on t and d.

Remark 3.3 Theoretical properties of these estimators remain open issues and will consist of a further work. The first step for strong consistency will be to state a strong law of large numbers for randomized OA with strength $t > 1$ since, as far as we know, such a result does not exist. Asymptotic normality has already been proved for randomized OA with strength $t = 2$ under smoothness conditions for any dimension $d \geq 3$ or without smoothness conditions but only in dimension $d = 3$ *[see 25, 26]*.

Remark 3.4 Orthogonal arrays with strength larger than 2 don't exist for any number of levels. Following the most famous construction based on Galois fields — Bush's construction *[see 27]* — one can construct any orthogonal array with index unity and strength t in dimension d provided the number of levels is a prime power larger than $t-1$ and $d-1$. More precisely, the number of points in the orthogonal array has to satisfy:

$$n \geq \big(\max(t-1, d-1)\big)^t. \tag{15}$$

Note that n is directly related to the precision in the estimation. Thus it has to be chosen large enough to attain a reasonable precision. In the following (see Section 4.1.1), we prove that even with this construction's constraint, our approach outperforms the one of Saltelli [2] for the estimation of first- and second-order sensitivity indices. For higher order sensitivity indices, Bush's construction can become really restrictive, and new constructions should be investigated.

4. Numerical illustrations

In this section, we propose a thorough comparison of our approach with the approach in Saltelli [2] based on Monte Carlo sampling and with approaches based on quasi-Monte Carlo sampling (see Sections 4.1 and 4.2 below). We also propose in Section 4.3 a comparison with the approach proposed in Morris *et al.* [22] for the estimation of first-order Sobol' indices. At last, we provide in Section 4.4 an application of our estimation strategy to a marine ecosystem simulator.

4.1 *Comparison with the approach in Saltelli [2]*

We first give arguments of the competitivity of our new approach with respect to the one presented in [2].

4.1.1 Efficiency arguments

Let n be the basic sample size. If one considers the result in Theorem 2 of [2], the number of model evaluations needed to estimate all first- and second-order sensitivity indices is equal to $(2d+2)\times n$. Let now p be defined as the smallest prime number such that $p^2 \geq n$. Our approach allows to estimate all first- and second-order indices using only $2n + 2p^2$ model evaluations, as soon as $d \leq \sqrt{n} + 1$ (see Inequality (15) in Remark 3.4). We know moreover from Proposition 3.2 in Section 3 that for first order indices, these estimates are at least as accurate as the ones obtained with Saltelli's approach. To illustrate the strength of our approach, we define the efficiency of our approach as $e = \frac{(2d+2)n}{2n+2p^2}$. We evaluate in Table 1 below this efficiency for different values of the sample size n.

n	1024	2048	4096	8192	16384	32768	65536	131072
$d_{\max}(n)$	33	46	65	91	129	182	257	363
$\frac{e}{d+1} = \frac{2\times n}{2\times n+2\times p^2}$	0.4280	0.4810	0.4772	0.4654	0.4884	0.4732	0.4980	0.4932

Table 1. Efficiency of our approach as $d \leq d_{\max}(n)$, with d the dimension and n the sample size.

Results in Table 1 show that in many practical cases, our approach is competitive (efficiency larger than one).

As in practice n is chosen large enough to attain a reasonable precision, the constraint due to Bush's construction is not a limitation. Indeed, $(2d+2) \times n \leq 2 \times n + 2 \times (d-1)^2$ implies $n \leq d - 1$, and this choice for n is never done in practice as it would provide a very bad precision.

4.1.2 Comparison on the Sobol' g function

The analytical test-case considered in this section is a multiplicative function known as the Sobol' g function — see Saltelli and Sobol' [28]. We consider $Y = f_1(X_1)\times\cdots\times f_d(X_d)$ where the X_is are independent random variables uniformly distributed on $[0,1]$ and for any $i \in \{1,\ldots,d\}$

$$f_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0\,. \tag{16}$$

To allow comparison, we consider the same test-case as in Section 5 in [2], that is $d = 6$ with $\mathbf{a} = (0, 0.5, 3, 9, 99, 99)$, and a sample size $n = 1024$. We compare our approach with the one in Theorem 2 of [2], which requires $(2d + 2) \times n = 14 \times 1024 = 14\,336$ model evaluations. With our approach, we expect to obtain a similar precision with only $2 \times n + 2 \times p^2$ model evaluations, where p is the smallest prime number such that $p^2 \geq \max(n, (d-1)^2) = \max(1024, 25) = 1024$ (see Criterion (15) in Remark 3.4). We thus take $p = 37$, and we apply our approach with $2 \times 1024 + 2 \times 37^2 = 4\,786$ model evaluations. The empirical mean and the empirical variance of the estimators are computed on 10 000 simulations. Concerning Saltelli's approach, each estimate is the average of the two estimates provided in Theorem 2 of [2].

The results in Table 2 go beyond the theoretical result of Proposition 3.2 (see Section 3): our estimation procedure is indeed clearly competitive for the estimation of first-order sensitivity indices, but also for the estimation of second-order sensitivity indices. And probably because of symmetry in the Sobol' g function, the estimation error of $\underline{S}_{\{12\}}$ is

divides by 100 using our approach. Using the restricted procedure of Theorem 1 in [2] allows estimating all the first-order Sobol' indices with a cost of $n \times (d+2) = 1024 \times 8 = 8\,192$ model evaluations, but it provides less accurate results, probably because it does not give a double estimate for each first-order index, contrarily to the procedure of Theorem 2 in [2] (see numerical experiments in Saltelli [2, Section 5]. The sample matrices ($\mathbf{M}_1$ and $\mathbf{M}_2$ in [2]) were obtained from a classical Monte Carlo sampling procedure.

	theoretical value	mean $\widehat{\underline{S}}_{\mathrm{u},n}^{RLHS}$	variance $\widehat{\underline{S}}_{\mathrm{u},n}^{RLHS}$	mean Saltelli 02	variance Saltelli 02
$\underline{S}_{\{1\}}$	0.5868	0.5869	$3.9\,10^{-4}$	0.5867	$4.5\,10^{-4}$
$\underline{S}_{\{2\}}$	0.2608	0.2608	$9.7\,10^{-4}$	0.2605	$8.3\,10^{-4}$
$\underline{S}_{\{3\}}$	0.0367	0.0361	$1.0\,10^{-3}$	0.0364	$9.9\,10^{-4}$
$\underline{S}_{\{4\}}$	0.0059	0.0061	$1.0\,10^{-3}$	0.0058	$9.9\,10^{-4}$
$\underline{S}_{\{5\}}$	0.00005	$< 1e-4$	$1.0\,10^{-3}$	$< 1e-4$	$9.9\,10^{-4}$
$\underline{S}_{\{6\}}$	0.00005	$< 1e-4$	$1.0\,10^{-3}$	$< 1e-4$	$9.9\,10^{-4}$
$\underline{S}_{\{12\}}$	0.9345	0.9345	$1.1\,10^{-5}$	0.9348	$1.0\,10^{-3}$
$\underline{S}_{\{13\}}$	0.6357	0.6357	$2.3\,10^{-4}$	0.6358	$5.4\,10^{-4}$
$\underline{S}_{\{14\}}$	0.5946	0.5946	$2.6\,10^{-4}$	0.5945	$4.7\,10^{-4}$
$\underline{S}_{\{15\}}$	0.5869	0.5867	$2.6\,10^{-4}$	0.5867	$4.5\,10^{-4}$
$\underline{S}_{\{16\}}$	0.5869	0.5870	$2.6\,10^{-4}$	0.5867	$4.5\,10^{-4}$
$\underline{S}_{\{23\}}$	0.3029	0.3029	$6.4\,10^{-4}$	0.3021	$8.8\,10^{-4}$
$\underline{S}_{\{24\}}$	0.2675	0.2673	$6.4\,10^{-4}$	0.2666	$8.4\,10^{-4}$
$\underline{S}_{\{25\}}$	0.2609	0.2610	$6.4\,10^{-4}$	0.2599	$8.4\,10^{-4}$
$\underline{S}_{\{26\}}$	0.2609	0.2608	$6.3\,10^{-4}$	0.2599	$8.4\,10^{-4}$
$\underline{S}_{\{34\}}$	0.0427	0.0428	$7.5\,10^{-4}$	0.0423	$9.8\,10^{-4}$
$\underline{S}_{\{35\}}$	0.0367	0.0368	$7.7\,10^{-4}$	0.0364	$9.8\,10^{-4}$
$\underline{S}_{\{36\}}$	0.0367	0.0376	$7.7\,10^{-4}$	0.0364	$9.8\,10^{-4}$
$\underline{S}_{\{45\}}$	0.0059	0.0059	$7.5\,10^{-4}$	0.0055	$9.8\,10^{-4}$
$\underline{S}_{\{46\}}$	0.0059	0.0059	$7.6\,10^{-4}$	0.0055	$9.8\,10^{-4}$
$\underline{S}_{\{56\}}$	0.0001	$< 10^{-4}$	$7.6\,10^{-4}$	$< 10^{-4}$	$9.6\,10^{-4}$

Table 2. Comparison with Saltelli's [2] approach for the estimation of first and second-order indices of the Sobol' g function in dimension $d = 6$ with $a = (0, 0.5, 3, 9, 99, 99)$, the empirical mean (*resp.* empirical variance) are computed on 10^4 simulations. The total number of model evaluations for Saltelli's [2] approach is equal to $N = 14\,336$ and the one with our approach is equal to $N = 4\,786$.

4.1.3 Asymptotic confidence intervals

In the present subsection and in the following one, we apply the new method proposed in Section 3 to the Ishigami function (see Ishigami and Homma [29]): $f(X_1, X_2, X_3) = \sin(X_1) + 7\sin^2(X_2) + 0.1X_3^4\sin(X_1)$ where the X_is are independent random variable uniformly distributed on $[-\pi, \pi]$. Analytical values of SIs of this model are

$$\underline{S}_1 = 0.3139,\ \underline{S}_2 = 0.4424,\ \underline{S}_3 = 0,\ \underline{S}_{12} = 0.7563,\ \underline{S}_{23} = 0.4424,\ \underline{S}_{13} = 0.5575 \text{ and } \underline{S}_{123} = 1.$$

In the following experiment, we focus on the empirical coverage — i.e. the empirical proportion of confidence intervals (CI) containing the analytical value of the SI — of both estimators at different sample sizes between 10^2 and 10^5, and for 100 000 simulations. We first investigate estimators $\widehat{\underline{S}}_{\{i\},n}$ and $\widehat{\underline{S}}_{\{i\},n}^{RLHS}$, $i \in \{1, \ldots, d\}$ and in both cases, we provide asymptotic CI from the estimation of the asymptotic variance given in Janon *et al.* [19] (see end of the proof of Prop. 2.2). Indeed, as we know that this asymptotic variance is:

$$\sigma_{IID,\mathrm{u}}^2 = \frac{\mathrm{Var}\big[(Y - \mathbb{E}[Y])(Y_{\mathrm{u}} - \mathbb{E}[Y]) - \underline{S}_{\mathrm{u}}/2\big((Y - \mathbb{E}[Y])^2\big)\big(Y_{\mathrm{u}} - \mathbb{E}[Y]\big)\big]}{\mathrm{Var}[Y]^2} \geq \sigma_{RLHS,\mathrm{u}}^2, \tag{17}$$

we can provide an estimator of the asymptotic CI for the classic method

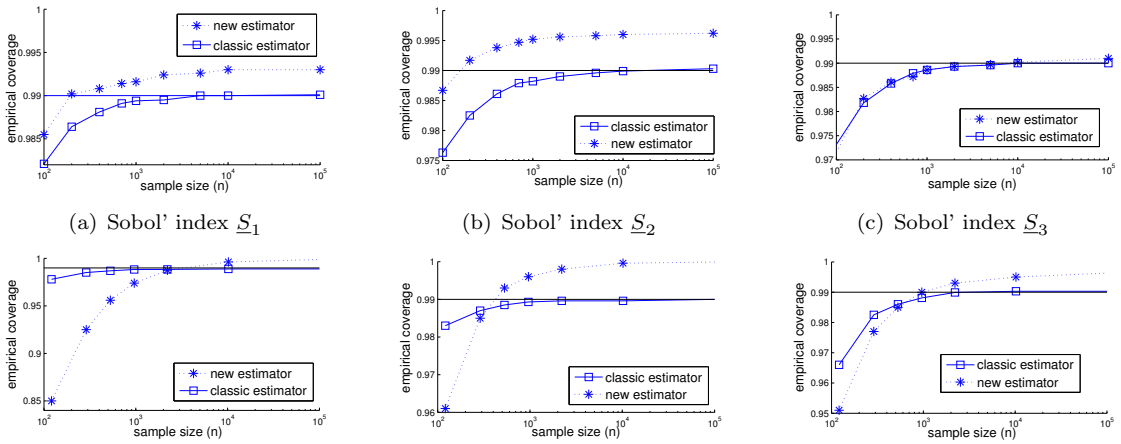
$$\mathcal{I}_{IID,\mathrm{u},\alpha} = \left[\underline{S}_{\mathrm{u}} - \frac{\sigma^2_{IID,\mathrm{u}}\ u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathrm{u}} + \frac{\sigma^2_{IID,\mathrm{u}}\ u_{\alpha/2}}{\sqrt{n}}\right]$$

and another one for the new method

$$\mathcal{I}_{RLHS,\mathrm{u},\alpha} = \left[\underline{S}_{\mathrm{u}} - \frac{\sigma^2_{RLHS,\mathrm{u}}\ u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathrm{u}} + \frac{\sigma^2_{RLHS,\mathrm{u}}\ u_{\alpha/2}}{\sqrt{n}}\right]$$

where $u_{\alpha/2}$ is the normal quantile at the significance level α and n is the sample size. By using the estimator of the asymptotic variance given in (17) in both cases, the CI lengths of the classic and the new estimators are the same. More specifically, the estimated length of the new estimator is greater or equal than its optimal value. We thus deal with a *pessimistic estimated length*. Thus the asymptotic value of the empirical coverage of the new method is greater or equal than the expected one. However at the moment, we do not know how to estimate correctly $\sigma^2_{RLHS,\mathrm{u}}$ because of its singular expression (see Proof of (ii) in Proposition 3.1). We just say a few words about it in the next subsection and more fundamentally, it should consist of a further work.

We also investigate estimators $\widehat{\underline{S}}_{\{i,j\},n}$ and $\widehat{\underline{S}}^{\text{replicated randomized OA2}}_{\{i,j\},n}$, $i \neq j \in \{1,\ldots,d\}$, where the notation replicated randomized OA2 refers to the generalization to replicated randomized OA of strength 2 presented in Section 3.2. In this case, we conjecture that the Central Limit theorem established in (ii) in Proposition 3.2 is also true under some smoothness assumption — note that, here, Ishigami function is $\mathcal{C}^\infty$. Results are gathered in Figures 1–2.



(a) Sobol' index $\underline{S}_1$ (b) Sobol' index $\underline{S}_2$ (c) Sobol' index $\underline{S}_3$

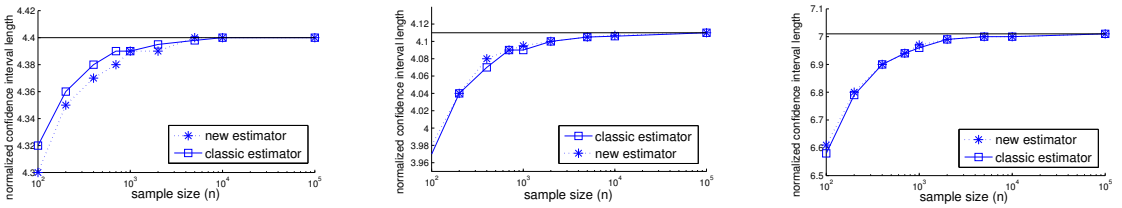
(d) Sobol' index $\underline{S}_{12}$ (e) Sobol' index $\underline{S}_{13}$ (f) Sobol' index $\underline{S}_{23}$

Figure 1. Comparison of the empirical coverage of 99 % CI for the first- and second-order Sobol' indices using the classical estimator based on substituted columns sampling plans and the new estimator based on replicated randomized OA, the empirical coverage is computed on 100 000 simulations of size 10^2 to 10^5.

For the second-order SIs, we can observe that the bivariate stratification has a bad effect on the new estimator at very low sample size, but we can notice its good properties as the number of simulations increases.

4.1.4 Remark on the confidence intervals length of the new estimator

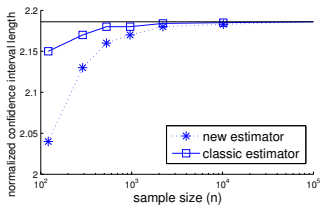
Concerning the estimation of the true CI length for the new estimators based on replicated randomized OA, note that if the asymptotic empirical coverage — estimated using



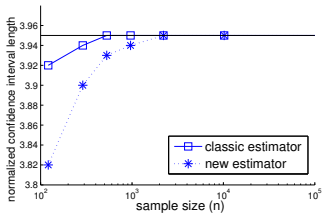
(a) Sobol' index $\underline{S}_1$

(b) Sobol' index $\underline{S}_2$

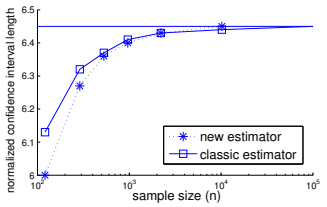
(c) Sobol' index $\underline{S}_3$



(d) Sobol' index $\underline{S}_{12}$



(e) Sobol' index $\underline{S}_{13}$



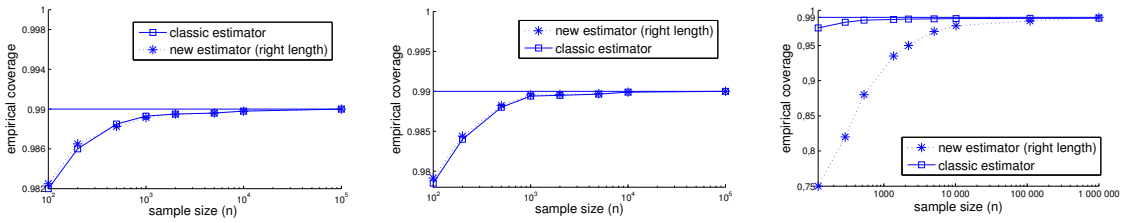
(f) Sobol' index $\underline{S}_{23}$

Figure 2. Comparison of the normalized ($\times\sqrt{n}$) length of the 99 % CI for the first- and second-order Sobol' indices using the classical estimator based on substituted columns sampling plans and the new estimator based on replicated randomized OA, the averaged length is computed on 100 000 simulations of size 10^2 to 10^5.

	$\underline{S}_1$	$\underline{S}_2$	$\underline{S}_{12}$	$\underline{S}_{13}$	$\underline{S}_{23}$
pessimistic estimated lengths	4.40	4.15	2.19	3.95	6.45
corrected lengths	3.96	3.28	1.53	2.37	5.16

Table 3. Comparison between *pessimistic estimated lengths* and *corrected lengths* of the 99 % CI for $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$, the lengths are averaged on 100 000 simulations of size 10^2 to 10^5

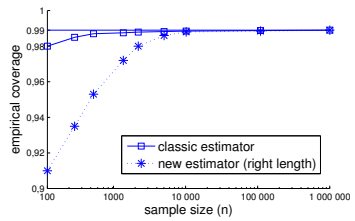
Formula (17) — is $1-\alpha'$ instead of the expected value $1-\alpha$, then it means that the true asymptotic CI should be $u_{\alpha/2}/u_{\alpha'/2}$ time as long, where $u.$ denote the normal quantiles. More specifically in our first application, we obtain in this way the true asymptotic normalized ($\times\sqrt{n}$) CI length (*corrected length*) of $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$; they are gathered in Table 3. Moreover considering these true normalized CI lengths, we can observe on Figure 3 that the empirical coverage of the new estimator converges to the expected level 0.99 as n increases, and so we confirm the reliability of the empirical CI constructed with the true asymptotic length. Unfortunately, evaluating the true asymptotic CI length



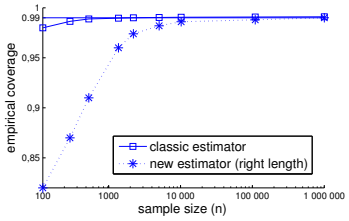
(a) Sobol' index $\underline{S}_1$

(b) Sobol' index $\underline{S}_2$

(c) Sobol' index $\underline{S}_{12}$



(d) Sobol' index $\underline{S}_{13}$



(e) Sobol' index $\underline{S}_{23}$

Figure 3. Comparison of the empirical coverage of *corrected* 99 % CI for the first- and second-order Sobol' indices using the classical estimator based on substituted columns sampling plans and the new estimator based on replicated randomized OA, the empirical coverage is computed on 100 000 simulations of size 10^2 to 10^5.

is infeasible in practice since it requires a lot of replications to estimate the empirical coverage. So the issue related to the construction of optimal CI remains open.

4.2 *Performance comparison between the new method and the quasi-Monte Carlo estimation*

We now come to the comparison between the new method based on RLHS and the method of Sobol' performed with Sobol' sequences, also known as LP_τ sequences and that consist of (t, s)-sequences in base 2 — see, e.g., Niederreiter [30] for (t, s)-sequences and Saltelli *et al.* [31] for the application to the method of Sobol'. The numerical application we propose consists in comparing the estimation error — the mean squared error — in the computation of all the first-order Sobol' indices $\underline{S}_i$ between these two approaches using exactly the same total number of model evaluations. Low discrepancy sequences are known for their efficiency but keep in mind that the total number of model evaluations required to estimate all the first-order Sobol' indices is $(d+1)n$ when using (t, s)-sequences while only $2n$ are necessary when using RLHS. The analytical test-case is the one of Section 4.1.2, that is the g-function defined by (16). The numerical test is divided into three cases:

Case i) in dimension $d = 3$ with g-function parameters $\mathbf{a} = (0, 1, 9)$
Case ii) in dimension $d = 12$ with g-function parameters $\mathbf{a} = (0, 0, 0, 0, 1, 1, 1, 1, 9, 9, 9, 9)$
Case iii) in dimension $d = 24$ with g-function parameters $\mathbf{a} = (\underbrace{0, \ldots, 0}_{8 \text{ times}}, \underbrace{1, \ldots, 1}_{8 \text{ times}}, \underbrace{9, \ldots, 9}_{8 \text{ times}})$.

The theoretical values of the Sobol' indices are the following:

Case i) $\underline{S}_1 = 0.742$, $\underline{S}_2 = 0.185$, $\underline{S}_3 = 0.007$
Case ii) $\underline{S}_1 = \cdots = \underline{S}_4 = 0.098$, $\underline{S}_5 = \cdots = \underline{S}_8 = 0.024$, $\underline{S}_9 = \cdots = \underline{S}_{12} = 0.001$,
Case iii) $\underline{S}_1 = \cdots = \underline{S}_8 = 0.018$, $\underline{S}_9 = \cdots = \underline{S}_{16} = 0.004$, $\underline{S}_{17} = \cdots = \underline{S}_{24} = 10^{-4}$.

The computations are performed at different total numbers of runs:

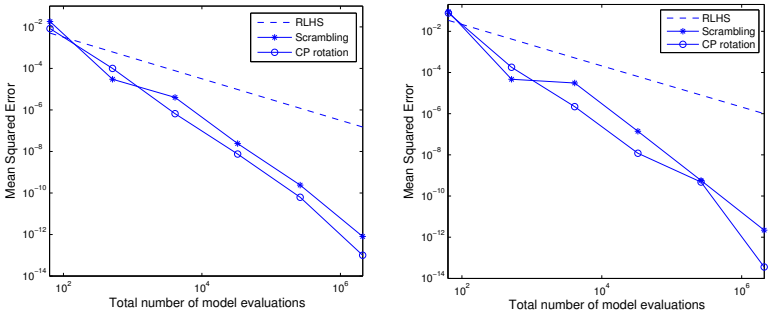
Case i) $N = 64$, 512, 4096, 32768, 262144, 2097152
Case ii) $N = 208$, 1664, 13312, 106496, 851968, 6815744
Case iii) $N = 200$, 1600, 12800, 102400, 819200, 6553600.

In this numerical experiment, as we measure the error in term of mean squared error, we consider randomized Sobol' sequences. More precisely we use both the following well-known methods of randomization:

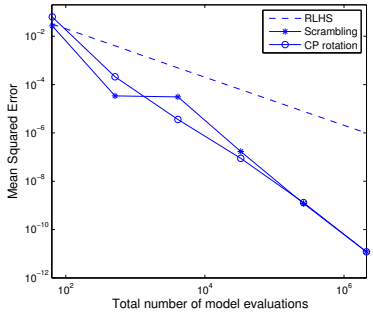
1. Cranley-Patterson rotation, that consists in adding a random vector to all the points of a DOE — where the addition is the componentwise addition modulo 1.
2. Owen's scrambling, that essentially consists in randomly permuting the levels of a (t, s)-sequence keeping the low discrepancy structure unchanged — see Owen [32–34].

The results are gathered in Figures (4–6).

First we can observe that the mean squared error (MSE) of all the estimates computed by using RLHS decreases with a rate of convergence of $O(n^{-1})$ in the three test-cases, while the rate of convergence of the low discrepancy method — with scrambling as well as Cranley-Patterson rotation — is $O(n^{-2})$ in dimension 3 but only $O(n^{-1})$ in dimension 24. Second note that MSEs of the estimates computed by using RLHS at the lowest total number of model evaluations keep in the same range of values, while initial MSEs in low discrepancy methods become worse as the dimension increases. Consequently, the use of Sobol' sequences clearly leads to better results in dimension 3, but appears as a poor choice in the other cases. Indeed in dimension 12, we can observe that Sobol' sequences

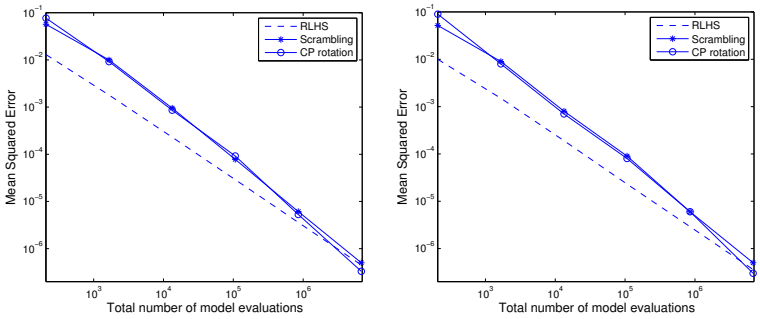


(a) Estimation error of $\underline{S}_1$ (b) Estimation error of $\underline{S}_2$

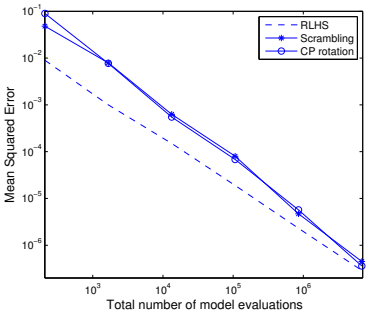


(c) Estimation error of $\underline{S}_3$

Figure 4. Plots of the mean squared error for the estimation of Sobol' indices of a g-function in dimension 3 with parameters $(0, 1, 9)$, the total number of model evaluations growing from 10^2 to 10^6. NB: CP rotation stands for Cranley-Patterson rotation.



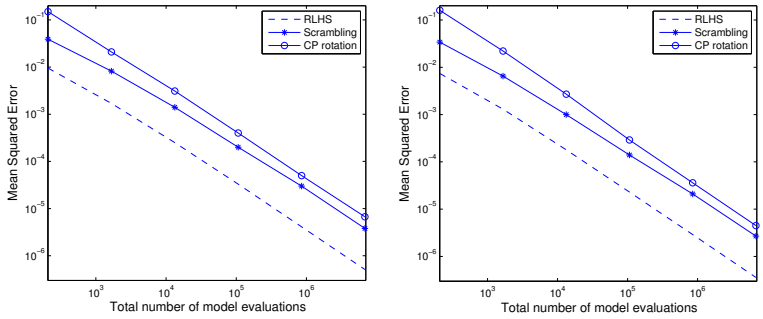
(a) Estim. error of $\underline{S}_1, \ldots, \underline{S}_4$ (b) Estim. error of $\underline{S}_5, \ldots, \underline{S}_8$



(c) Estim. error of $\underline{S}_9, \ldots, \underline{S}_{12}$

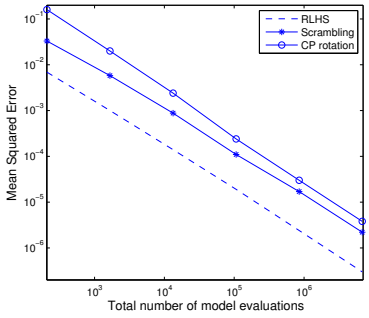
Figure 5. Plots of the mean squared error for the estimation of Sobol' indices of a g-function in dimension 12 with parameters $(0, 0, 0, 0, 1, 1, 1, 1, 9, 9, 9, 9)$, the total number of model evaluations growing from 10^3 to 10^6. NB: CP rotation stands for Cranley-Patterson rotation.

provide better results than RLHS only asymptotically — we can see in Figure 5 that at least 10^7 model evaluations are necessary — and in dimension 24, MSEs computed by



(a) Estim. error of $\underline{S}_1, \ldots, \underline{S}_8$

(b) Estim. error of $\underline{S}_9, \ldots, \underline{S}_{16}$



(c) Estim. error of $\underline{S}_{17}, \ldots, \underline{S}_{24}$

Figure 6. Plots of the mean squared error for the estimation of Sobol' indices of a g-function in dimension 24 with parameters $(0, \ldots, 0, 1, \ldots, 1, 9, \ldots, 9)$, the total number of model evaluations growing from 10^3 to 10^6. NB: CP rotation stands for Cranley-Patterson rotation.

using RLHS are always a factor 10 or 15 better than those of low discrepancy methods.

4.3 *Comparison with the approach proposed in Morris et al. [22] for the estimation of first-order Sobol' indices*

In this section, we propose to compare our approach with the one in Morris *et al.* [22] on the example they provide in the Section 4 of their article. Their example is the Sobol' g function in dimension $d = 8$ with $\mathbf{a} = (0, 1, 1, 2, 3, 5, 8, 13)$. They compute the mean and standard error for their estimates over 1000 simulations. For each of these 1000 simulations they use a *permuted column* (PC) sampling plan composed with 8 replications of an initial design of size 8. They compare the results for an i.i.d. initial sample with the results obtained with 8 replicated LHS. They also provide the results for the strategy based on the UPCS (*unbiased permuted column samples*) with or without LHS. In Table 4 below we compare our results (random balance strategy, **RB strategy**) with the ones they obtain with *random arrays, Latin hypercube sampling* (**Mor1**), and with *orthogonal arrays, Latin hypercube sampling* (**Mor2**) by computing the empirical mean (μ) and standard deviation (σ) on 1000 simulations, each of which is performed at a total cost of 64 model evaluations.

It is easy to see on these results the bias of their standard approach, based on random arrays with Latin hypercube sampling. The strategy they propose to remove the bias, based on the construction of a strength 2 OA is very efficient as we can see. Due to the construction of the UPCS, they need at least d replications of the initial design. The bias in our approach is not very important, even with 64 model evaluations, except for small indices such as $\underline{S}_{\{7\}}$ and $\underline{S}_{\{8\}}$. For small indices, it is known that there exist better strategies proposed e.g. by Owen [17]. We must also achnowledge on this example that for the estimation procedure proposed in Morris *et al.* [22] the empirical standard deviations are smaller. Thus, if we are only interested by first-order Sobol' indices the

	analytical	μ **RB strategy**	σ **RB strategy**	μ **Mor1**	σ **Mor1**	μ **Mor2**	σ **Mor2**
$\underline{S}_{\{1\}}$	0.4890	0.5032	0.1285	0.5450	0.0838	0.4850	0.0615
$\underline{S}_{\{2\}}$	0.1223	0.1198	0.1413	0.2091	0.0843	0.1183	0.0334
$\underline{S}_{\{3\}}$	0.1223	0.1216	0.1366	0.2061	0.0864	0.1176	0.0346
$\underline{S}_{\{4\}}$	0.0543	0.0556	0.1336	0.1532	0.0780	0.0514	0.0264
$\underline{S}_{\{5\}}$	0.0306	0.0246	0.1322	0.1278	0.0674	0.0293	0.0237
$\underline{S}_{\{6\}}$	0.0136	0.0123	0.1302	0.1150	0.0674	0.0127	0.0190
$\underline{S}_{\{7\}}$	0.0060	0.0005	0.1269	0.1099	0.0614	0.0051	0.0172
$\underline{S}_{\{8\}}$	0.0025	0.0042	0.1300	0.1074	0.0640	0.0022	0.0163

Table 4. Comparison with Morris *et al.* [22] approach for the estimation of first-order indices of the Sobol' g function in dimension $d = 8$ with $\mathbf{a} = (0, 1, 1, 2, 3, 5, 8, 13)$, the empirical mean ($\mu$) and the empirical standard deviation (σ) are computed on 1000 simulations. The total number of model evaluations for each approach is equal to $N = 64$.

approach in Morris *et al.* [22] should be prefered to the approach in Saltelli [2] or to our approach. However, we do not propose in this article a systematic comparison on various examples, and our conclusions remain limited to the example above.

4.4 *Application to a marine ecosystem simulator*

We now illustrate the new method to a one-dimensional coupled hydrodynamical– biological model developed and applied on the Ligurian Sea (northwestern Mediterranean). This ecosystem simulator, MODèle d'ÉCOsystème du GHER et du LOBEPM[1] (MODECOGeL), combines a 1D (vertical) version of the 3D GHER model which takes into account momentum and heat surface fluxes computed from a real meteorological data set, and a biogeochemical model defined by a nitrogen cycle of 12 biological state variables (see Figure 7) controlled by 87 input parameters, see Lacroix *et al.* [35]. Here we focus on the chlorophyll-a concentration which is defined as a function of time and depth $\texttt{chla}(t,z) = 1.59 \times \big(\texttt{pp}(t,z) + \texttt{np}(t,z) + \texttt{mp}(t,z)\big)$ where `pp`, `np` and `mp` are the phyto-, nano- and microphytoplankton biomasses, respectively. The behavior of these three state

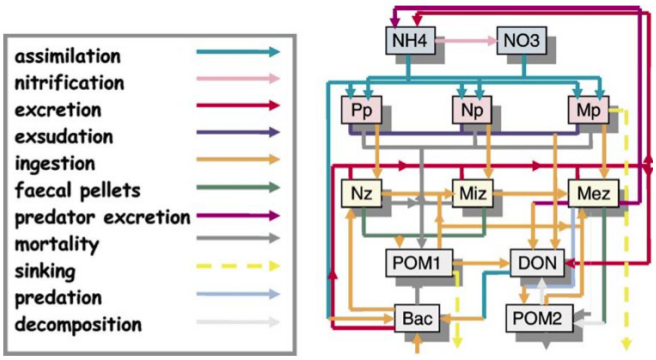


Figure 7. Biogeochemical model (NH4: Ammonium; NH3: nitrate; Pp, Np, Mp: pico-, nano-, microphytoplankton; Nz, Miz, Mez: nano-, micro-, mesozooplankton; PON1, PON2: type 1 and 2 particulate organic nitrogen; Bac: bacteria; DON: dissolved organic nitrogen).

variables are modeled by the following reaction-diffusion and reaction-advection-diffusion

[1]GHER: GeoHydrodynamics and Environment Research, Université de Liège, Belgium. LOBEPM: Laboratoire d'Océanologie Biologique et d'Écologie du Plancton Marin, Université Pierre et Marie Curie, France

equations

$$\frac{\partial\,\mathtt{pp}}{\partial t} = \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{pp}}{\partial z}\right) + \big((1-exud_{\mathtt{pp}})\mu_{\mathtt{pp}} - mort_{\mathtt{pp}}\big)\mathtt{pp} - ing_{\mathtt{pp},\mathtt{nz}}\mathtt{nz}$$

$$\frac{\partial\,\mathtt{np}}{\partial t} = \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{np}}{\partial z}\right) + \big((1-exud_{\mathtt{np}})\mu_{\mathtt{np}} - mort_{\mathtt{np}}\big)\mathtt{np} - ing_{\mathtt{np},\mathtt{miz}}\mathtt{miz}$$

$$\frac{\partial\,\mathtt{mp}}{\partial t} = \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{mp}}{\partial z}\right) + \big((1-exud_{\mathtt{mp}})\mu_{\mathtt{mp}} - mort_{\mathtt{mp}}\big)\mathtt{mp} - ing_{\mathtt{mp},\mathtt{mez}}\mathtt{mez} - sin_{\mathtt{mp}}\frac{\partial\mathtt{mp}}{\partial z}$$

where `nz`, `miz` and `mez` are the nano-, micro- and mesozooplankton biomasses, respectively, and the other notations are

λ	vertical turbulent diffusivity ($\mathrm{m}^2.\mathrm{s}^{-1}$)
$exud_A$	exudation of A (percentage)
μ_A	growth rate of A (day^{-1})
$mort_A$	mortality rate of A (day^{-1})
$ing_{A,B}$	ingestion rate of A by predator B (mgChl)
$sin_{\mathtt{mp}}$	sinking velocity of microphytoplankton ($\mathrm{m.day}^{-1}$)

In our experiment, we focus on two different outputs: the annual maximum of chlorophyll-a concentration in surface water Y_{surf} and the annual maximum of the mean of cholorphyll-a concentration between 20 and 50 meters in depth Y_{depth}. These are practical indicators of biological activity. We are interested in the influence of eight parameters among the 87 input factors. On the one hand, we consider 6 a priori influent parameters $\mu_{max\mathtt{pp}}$, $\mu_{max\mathtt{np}}$, $\mu_{max\mathtt{mp}}$, $I_{opt\mathtt{pp}}$, $I_{opt\mathtt{np}}$ and $I_{opt\mathtt{mp}}$ where μ_{maxA} and I_{optA} denote the maximum growth rate of A and the optimum insolation for A, respectively. These input factors are directly related to the growth rate of A, μ_A (see details in Supplementary materials). On the other hand, we consider the maximum growth rate of bacteria $\mu_{max\mathtt{bac}}$ and the sinking velocity of particulate organic nitrogen (type 1) $sin_{\mathtt{pon1}}$ which have a priori a negligible effect on chlorophyll-a concentration since they do not act directly on `pp`, `np` and `mp` but on the state variables `bac` and `pon1`. We take these eight parameters to be independent gamma distributed random variables with parameters given in Table 5. We estimate all the first- and second-order SIs of both outputs Y_{surf} and Y_{depth} by using the estimators $\widehat{\underline{S}}_{\mathtt{u}}^{RLHS}$ defined in Sections 3.1 and 3.2 with sample sizes $n = 65536$ and $n = 66049$, respectively.

The first-order SIs are estimated by using nested replicated latin hypercubes following Qian's construction Qian [36]. They allow to visualize empirical convergence of the estimated indices as shown in Figure 8. The estimated indices at the biggest sample size ($n = 65536$) are reported in Tables 6 and 7; we can notice that both outputs do not define an additive model since in both cases, the sum of the first-order SIs are less than sixty percents. We also notice that $\mu_{max\mathtt{pp}}$ is important in both outputs, while three other a priori important parameters — $\mu_{max\mathtt{np}}$, $I_{opt\mathtt{np}}$ and $I_{opt\mathtt{mp}}$— have actually no effect. At last, it is surprising to observe that the parameter $\mu_{max\mathtt{bac}}$, which does not act directly on both outputs, has non-zero values.

The second-order SIs are estimated by using a replicated latin hypercube based on an orthogonal array with 257 levels, index 1 and strength 2 — i.e. $n = 66049$ — following Bush's construction, see Bose [27]. The results are reported in Tables 8 and 9; they confirm that $\mu_{max\mathtt{pp}}$ has the main role in both outputs since the non-negligible second-order SIs are all related to the latter. As a conclusion, we can notice that both outputs are extremely complex and contain, without any doubt, interactions of order more than or equal to 3. Such an analysis with the MC estimator of SIs would be less efficient without the new approach we proposed in this paper. More precisely, both order 1 and order 2

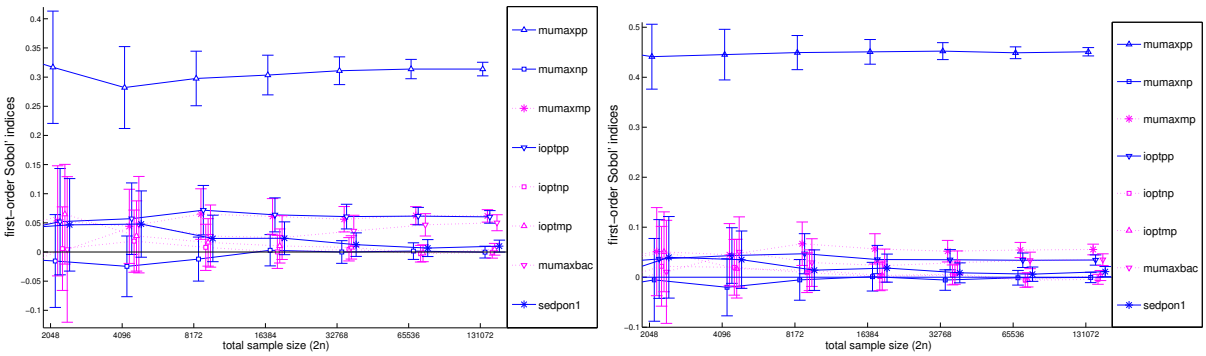


Figure 8. First-order SIs with error bars — 99% CI — for both Y_{surf} (left) and Y_{depth} (right).

analysis using the classical MC estimator — i.e. estimating all the first- or second-order SIs— only could use a sample size equal to $263\,170 = 2 \times 65\,536 + 2 \times 66\,049$ instead of $1\,179\,648 = (2 \times 8 + 2) \times 65\,536$ for a better or equal precision, see Section 4.1.1.

	label	k	θ	mean	standard deviation
$\mu_{max\texttt{pp}}$ (day^{-1})	1	9	0.33	3	1
$\mu_{max\texttt{np}}$ (day^{-1})	2	9	0.28	2.5	0.83
$\mu_{max\texttt{mp}}$ (day^{-1})	3	9	0.22	2	0.67
$I_{opt\texttt{pp}}$ (W.m^{-2})	4	9	1.11	10	3.33
$I_{opt\texttt{np}}$ (W.m^{-2})	5	9	1.67	15	5
$I_{opt\texttt{mp}}$ (W.m^{-2})	6	9	2.22	20	6.67
$\mu_{max\texttt{bac}}$ (day^{-1})	7	9	0.22	2	0.67
$sin_{\texttt{pon1}}$ (m.day^{-1})	8	9	0.17	1.5	0.5

Table 5. Distributions of variables using gamma density $f(x; k, \theta) = x^{k-1} \exp(-x/\theta)/(\Gamma(\theta)\theta^k)$, where $\Gamma(\cdot)$ is the gamma function.

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.314	0	0.061	0.060	0	0.003	0.051	0.010
estimated error	0.010	0.011	0.012	0.011	0.010	0.010	0.013	0.012

Table 6. Estimation of first-order SIs for the output Y_{surf}. The estimated error is the radius of the 99% CI.

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.451	0	0.055	0.034	0	0	0.035	0.011
estimated error	0.009	0.010	0.010	0.010	0.010	0.010	0.012	0.010

Table 7. Estimation of first-order SIs for the output Y_{depth}. The estimated error is the radius of the 99% CI.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.374	0.479	0.424	0.339	0.324	0.400	0.318	0.069	0.066	0.016
estimated error	0.012	0.011	0.013	0.011	0.011	0.010	0.011	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.015	0.069	0.015	0.125	0.074	0.075	0.128	0.072	0.077	0.070
estimated error	0.010	0.015	0.010	0.011	0.011	0.011	0.013	0.011	0.011	0.011

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.121	0.066	0.017	0.055	0.014	0.056	0.009	0.050
estimated error	0.013	0.011	0.010	0.015	0.010	0.014	0.010	0.015

Table 8. Estimation of second-order SIs for the output Y_{surf}. The estimated error is the radius of the 99% CI.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.506	0.593	0.510	0.455	0.450	0.515	0.447	0.056	0.034	0.005
estimated error	0.010	0.009	0.010	0.009	0.009	0.008	0.009	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.008	0.055	0.009	0.087	0.057	0.064	0.109	0.063	0.041	0.043
estimated error	0.010	0.014	0.010	0.011	0.011	0.011	0.013	0.011	0.010	0.010

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.082	0.041	0.009	0.040	0.007	0.046	0.006	0.041
estimated error	0.013	0.010	0.010	0.014	0.010	0.014	0.010	0.014

Table 9. Estimation of second-order SIs for the output Y_{depth}. The estimated error is the radius of the 99% CI.

5. Conclusion

We have introduced a new method to estimate all the first-order SIs by using only 2 samples of size n where n does not depend on the dimension anymore. We also explained how this approach extends to second-order or even k-th order sensitivity indices, with a sample size n which now depends both on the dimension d and on the strength k of the interactions. We also explained and illustrated on numerical examples that, at least when considering the estimation of all the first- and second-order indices, this approach outperforms the one introduced in Saltelli (2002) in many frameworks. We derive theoretical results in the particular case of first-order SIs from the work by Janon *et al.* [19] on asymptotical properties of SIs and from the work by Loh [37] on asymptotical properties of LHS. Further works will consist in deriving these theoretical results to higher-order SIs and in improving the method by studying how we can estimate correctly the asymptotic variance of the new estimator.

Acknowledgments

The authors are grateful to Eric Blayo, Jean-Michel Brankart and Pierre Brasseur for valuable discussions on the simulator MODECOGeL and more generally on marine ecosystem models. They also thank Art Owen for his helpful comments. This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015).

Supplementary materials

These supplementary files can be downloaded from the journal website as a single archive.

Proofs: Detailed proofs of Propositions 1 and 2 (PDF document).
Phytoplankton model: Equations of the phytoplankton growth model (PDF document).

Appendix A. Proof of Theorem 3.1

We have $\mathbf{X}^{\pi_i^{-1}\circ\pi(j)} = (\widetilde{\mathbf{X}}^j_{\{i\}}, \mathbf{Z}^j_{\{i\}^c})$ and $\mathbf{X}'^{\pi_i'^{-1}\circ\pi(j)} = (\widetilde{\mathbf{X}}^j_{\{i\}}, \mathbf{Z}'^j_{\{i\}^c})$ where

$$\widetilde{\mathbf{X}}^j_{\{i\}} = \frac{\pi(j) - U_{i,\pi(j)}}{n}$$

$$\mathbf{Z}^j_{\{i\}^c} = \left(\frac{\breve{\pi}_1(j) - U_{1,\breve{\pi}_1(j)}}{n}, \ldots, \frac{\breve{\pi}_{i-1}(j) - U_{i-1,\breve{\pi}_{i-1}(j)}}{n}, \frac{\breve{\pi}_{i+1}(j) - U_{i+1,\breve{\pi}_{i+1}(j)}}{n}, \ldots, \frac{\breve{\pi}_d(j) - U_{d,\breve{\pi}_d(j)}}{n}\right)$$

$$\mathbf{Z}'^j_{\{i\}^c} = \left(\frac{\breve{\pi}'_1(j) - U_{1,\breve{\pi}'_1(j)}}{n}, \ldots, \frac{\breve{\pi}'_{i-1}(j) - U_{i-1,\breve{\pi}'_{i-1}(j)}}{n}, \frac{\breve{\pi}'_{i+1}(j) - U_{i+1,\breve{\pi}'_{i+1}(j)}}{n}, \ldots, \frac{\breve{\pi}'_d(j) - U_{d,\breve{\pi}'_d(j)}}{n}\right)$$

with $\breve{\pi}_k = \pi_k \circ \pi_i^{-1} \circ \pi$ and $\breve{\pi}'_k = \pi'_k \circ \pi_i'^{-1} \circ \pi$. Then note that π, the $\breve{\pi}_k$s and the $\breve{\pi}'_k$s are independent random permutations, and deduce that Y^j and $Y^j_{\{i\}}$ given in (12) are defined as in (11). Thus Proposition 3.2 applies and the conclusion follows.

References

[1] Sobol' IM. Sensitivity analysis for nonlinear mathematical models. Mathematical Modeling and Computational Experiment. 1993;1:407–414.

[2] Saltelli A. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications. 2002;145:280–297.

[3] Owen AB. Variance components and generalized sobol' indices. SIAM/ASA J Uncertainty Quantification. 2013;1:19–41.

[4] Hoeffding WF. A class of statistics with asymptotically normal distributions. Annals of Mathematical Statistics. 1948;19:293–325.

[5] Efron B, Stein C. The jackknife estimate of variance. The Annals of Statistics. 1981;9(3):586–596.

[6] Cukier RI, Levine HB, Shuler KE. Nonlinear sensitivity analysis of multiparameter model systems. Journal of Computational Physics. 1978;26:1–42.

[7] Tarantola S, Gatelli D, Mara TA. Random balance designs for the estimation of first-order global sensitivity indices. Reliability Engineering and System Safety. 2006;91:717–727.

[8] Oakley JE, O'Hagan A. Probabilistic sensitivity analysis of complex computer models: A bayesian approach. JRSS Series B. 2004;66:751–769.

[9] Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering and System Safety. 2008;93:964–979.

[10] Chen W, Jin R, Sudjianto A. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. ASME Journal of Mechanical Design. 2005; 127:875–886.

[11] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Saisana DGM, Tarantola S. Global sensitivity analysis: The primer. John Wiley & Sons; 2008.

[12] Tissot JY, Prieur C. Variance-based sensitivity analysis using harmonic analysis. 2012. Available from: http://hal.archives-ouvertes.fr/hal-00680725.

[13] Stanley RP. Enumerative combinatorics, volume 1 (2nd edition). Cambridge University Press; 2012.

[14] Owen AB. Orthogonal arrays for computer experiments, integration and visualization. Statistica Sinica. 1992;2:439–452.

[15] Lemieux C. Monte Carlo and quasi-Monte Carlo sampling. Springer Series in Statistics. Springer, New York; 2009.

[16] Devroye L. Non-uniform random variate generation. Springer-Verlag, New York; 1986.

[17] Owen AB. Better estimation of small sobol' sensitivity indices. ACM TOMACS. 2012;23(2).

[18] Morris MD, Moore LM, McKay MD. Sampling plans based on balanced incomplete block

designs for evaluating the importance of computer model inputs. J Statist Plann Inference. 2006;136(9):3203–3220.

[19] Janon A, Klein T, Lagnoux-Renaudie A, Nodet M, Prieur C. Asymptotic normality and efficiency of two Sobol index estimators. ESAIM: Probability and Statistics. 2013;Available from: http://hal.inria.fr/hal-00665048.

[20] Monod H, Naud C, Makowski D. Uncertainty and sensitivity analysis for crop models. In: Wallach D, Makowski D, Jones JW, editors. Working with dynamic crop models: Evaluation, analysis, parameterization, and applications. chap. 4. Elsevier, Amsterdam; 2006. p. 55–99.

[21] McKay MD. Evaluating prediction uncertainty. Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory. 1995;:1–79.

[22] Morris MD, Moore LM, McKay MD. Using orthogonal arrays in the sensitivity analysis of computer models. Technometrics. 2008;50(2):205–215.

[23] Mara TA, Rakoto Joseph O. Comparison of some efficient methods to evaluate the main effect of computer model factors. Journal of Statistical Computation and Simulation. 2008; 78(2):167–178.

[24] Saltelli A, Chan K, Scott M. Sensitivity analysis. John Wiley & Sons; 2000.

[25] Loh WL. A multivariate central limit theorem for randomized orthogonal array sampling designs in computer experiments. The Annals of Statistics. 2008;36:1983–2023.

[26] Loh WL. A combinatorial central limit theorem for randomized orthogonal array sampling designs. Ann Statist. 1996;24(3):1209–1224.

[27] Bose R. On the application of the theory of Galois fields to the problem of construction of hyper-graeco-latin squares. Sankhya. 1938;3:323–338.

[28] Saltelli A, Sobol' IM. About the use of rank transformation in sensitivity analysis of a model. Reliability Engineering and System Safety. 1995;50:225–239.

[29] Ishigami T, Homma T. An importance quantification technique in uncertainty analysis for computers models. First International Symposium on Uncertainty Modeling and Analysis Proceedings. 1990;:398–403.

[30] Niederreiter H. Random number generation and quasi-Monte Carlo methods. Vol. 63 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; 1992.

[31] Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance-based sensitivity analysis of model output. design and estimator for the total index. Computer Physics Communications. 2010;181:259–270.

[32] Owen AB. Randomly permuted (t,m,s)-nets and (t,s)-sequences. In: Niederreiter H, Shiue PJS, editors. Monte carlo and quasi-monte carlo methods in scientific computing. Springer-Verlag, New York; 1995. p. 299–317.

[33] Owen AB. Monte carlo variance of scrambled equidistribution quadrature. SIAM J Numer Anal. 1997;34(5):1884–1910.

[34] Owen AB. Scrambled net variance for integrals of smooth functions. Ann Statist. 1997; 25(4):1541–1562.

[35] Lacroix G, Nival P. Influence of meteorogical variability on primary production dynamics in the ligurian sea (nw mediterranean sea) with 1d hydrodynamic/biological model. Journal of Marine Systems. 1998;37:229–258.

[36] Qian PZG. Nested latin hypercube sampling. Biometrika. 2009;96(4):957–970.

[37] Loh WL. On latin hypercube sampling. The Annals of Statistics. 1996;24(5):2058–2080.