A global sensitivity analysis approach for marine biogeochemical modeling

Clémentine Prieur, Laurence Viry, Eric Blayo, Jean-Michel Brankart

A global sensitivity analysis approach for marine biogeochemical modeling

C. Prieur[a], L. Viry[a], E. Blayo[a], J.-M Brankart[b]

[a] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, 38000 Grenoble, France*
[b] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, IGE, 38000 Grenoble, France*
* *Institute of Engineering Univ. Grenoble Alpes*

Abstract

This paper introduces the Sobol' indices approach for global sensitivity analysis, in the context of marine biogeochemistry. Such an approach is particularly well suited for ocean biogeochemical models, which make use of numerous parameters within large sets of differential equations with complex dependencies. This sensitivity analysis allows for a detailed study of the relative influence of a large number of input parameters on output quantities of interest to be chosen. It is able to distinguish between direct effects of these parameters and effects due to interaction between two or more parameters. Although demanding in terms of computation, such a tool is now becoming affordable, thanks to the development of distributed computing environments. An applicative example is presented with the Modecogel biogeochemical model.

Keywords: sensitivity analysis, Sobol' indices, marine biogeochemistry

1. Introduction

Marine biogeochemical models are now commonly integrated as modules within complex ocean circulation modeling systems. They are thus more and more widely used for numerous applications. However these models raise some particular questions, especially regarding their tuning. They generally are systems of nonlinear ordinary differential equations, each equation expressing the time evolution of a given state variable due to hydrodynamical effects (transport and diffusion) and to fluxes between the various components of the ecosystem. A first particularity is that the evaluation of these fluxes involve numerous parameters, typically several times more than the

number of variables. A second particularity is that the values of these parameters are often quite poorly known: as a matter of fact, those parameters depend on the physical and biogeochemical context, while their available reference values were generally estimated in a particular situation (during a field experiment) or in laboratory experiments (i.e. not in a real context). The uncertainty on these values is thus generally quite large (see for example Schartau et al. (2017) for a review on the identification of such parameters).

In this context, *sensitivity analysis* (SA), i.e. methods that aim at quantifying the relative influence of the inputs on some given outputs in a complex system like a numerical model, may be a valuable tool. They indeed can help better understanding the model itself, and identifying which parameters are the most influential and must then be carefully calibrated. A common approach consists in addressing these questions in a somewhat empirical way, by conducting a few experiments in which the values of parameters vary, either one-at-a-time, or simultaneously (e.g. Druon and Le Fèvre (1999); Baklouti et al. (2006); Kriest et al. (2012)). Some additional techniques are also sometimes used, like linear error propagation (Omlin et al., 2001) or a Gaussian emulation machine approach (Scott et al., 2011).

Another approach defines the sensitivity of the output w.r.t. the input as the corresponding gradient (both input and output variables must of course be continuous real-valued quantities). This so-called *gradient based SA* thus consists in computing gradients. This can be quite straightforward for simple cases. For instance, the gradient of a single output quantity Q w.r.t. a constant parameter P can be approximated by $(Q(p+\alpha) - Q(p))/\alpha$, where p is the current value of P and using a small value of α. The computation of this approximate gradient thus just requires running twice the model, using successively p and $p+\alpha$ for parameter P. However getting the gradient can be much more difficult if P is not a constant (e.g. if P is a space and/or time dependent coefficient). In such a case, the preceding approach consisting in computing growth rates requires $N+1$ evaluations of the model, where N is the dimension of P (e.g. the number of space-time grid points). If N is large, an alternative to computing growth rates is to use the so-called *adjoint method*, which provides the exact gradient. This approach was used for instance in Fennel et al. (2001), Faugeras et al. (2003) and Tjiputra et al. (2007).

However, whatever the way to compute the gradient, it is important to understand that the gradient based SA is a local method, in the sense that the gradient is a local notion, computed in the vicinity of the current value

of P. Therefore this gradient, which is a way to quantify the influence of P on Q, can be quite different depending on the value of P. This flaw can be avoided by using a global approach to SA, i.e. an approach that quantifies the influence of P on Q taking into account the possible variations of P. Such a quantification is provided by *Sobol' indices*, that will be presented in the next section. Given their global character, computing such indices may of course require a huge number of model evaluations. However the continuous increase of computer ressources makes it more and more affordable. That is why we think that it is timely for the scientific community to start making a quite systematic use of such tools, in particular when dealing with highly parameterized models. If this becomes to be the case in neighboring disciplines like environmental modeling or marine ecosystem modeling, it is still not to our knowledge in marine biogeochemistry.

In this context, the aim of this paper is to introduce the Sobol' indices approach for global sensitivity analysis, and to illustrate its feasibility and its scientific relevance in the context of marine biogeochemistry. Note however that our goal is much more to present the methodological tools within an applicative example than to conduct an in-depth physical analysis, which would exceed our skills. The outline of this paper is the following. Global SA approach and the Sobol' indices are introduced in Section 2. Then, in Section 3, we present the biogeochemical model, its numerous uncertain parameters and some output quantities we chose. Section 4 is devoted to implementation aspects of the SA, and Section 5 presents some examples of results brought by this method.

2. Global sensitivity analysis, the framework

As indicated above, the main aim of sensitivity analysis is to determine which model inputs are the most influential on some given model outputs. In the following, a model output y will be considered as a deterministic scalar function of some model inputs $\mathbf{x} = (x_1, \ldots, x_d)$, these inputs belonging to a domain Δ:

$$y = y(\mathbf{x}) = y(x_1, \ldots, x_d) \qquad \text{with } \mathbf{x} \in \Delta$$

In Section 3, we will consider various scalar outputs, corresponding to different quantities of interest. In this paper, we adopt the stochastic framework of global sensitivity analysis. Unlike local sensitivity analysis, which analyses how a small perturbation near an input space value $x^0 = (x_1^0, \ldots, x_d^0)$

influences the value of the scalar output, global sensitivity analysis considers the whole variation range in the input space. More precisely, each input parameter is considered as a random variable X_j ($j = 1, \ldots, d$), the uncertainty on its value being modeled by some one-dimensional probability distribution. The output Y can then be considered as a scalar random variable $Y = f(\mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_d)$ with the X_i being assumed independent. Various sensitivity measures have been proposed in the global framework. We focus in this paper on variance-based sensitivity measures, introduced in Sobol' (1993).

2.1. Variance-based sensitivity measures

In order to quantify the influence of the variations of X_j on the variations of Y, let us consider the conditional expectation $\mathrm{E}(Y|X_j = x_j)$. It corresponds to the mean value of Y over the probability distributions of the X_k ($k \neq j$), when X_j is fixed to x_j. The corresponding random variable, when considering the variations of X_j, is $\mathrm{E}(Y|X_j)$, and its variance quantifies the influence of X_j on the dispersion of Y. The so called Sobol' sensitivity indices are obtained by normalizing this variance by the total variance of the output Y, which is assumed to be finite and non null. Thus the first-order Sobol' sensitivity index of input parameter X_j is defined as

$$\mathrm{S}_{\{j\}} = \frac{\mathrm{Var}\big(\mathrm{E}(Y|X_j)\big)}{\mathrm{Var}(Y)}\,. \tag{1}$$

It belongs to the interval $[0, 1]$.

2.2. The Sobol' decomposition

More generally, starting from the functional Analysis of Variance (ANOVA) decomposition (Hoeffding, 1948; Efron and Stein, 1981; Owen, 1992; Sobol', 1993), one can define sensitivity indices of any order $r \in \{1, \ldots, d\}$. Let us first introduce some notation. We assume that f is a real square integrable function, $\mathfrak{u}$ is a subset of $\{1, \ldots, d\}$, $\mathfrak{u}^c$ stands for its complement, its cardinality is denoted by $r = |\mathfrak{u}|$, and $\mathbf{X}_\mathfrak{u}$ represents the random vector with components X_j, $j \in \mathfrak{u}$. The functional ANOVA decomposition then states that $Y = f(\mathbf{X})$ can be uniquely decomposed into summands of increasing size

$$f(\mathbf{X}) = \sum_{\mathfrak{u} \subseteq \{1,\ldots,d\}} f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) \tag{2}$$

where $f_\emptyset = \mathbb{E}[Y]$ and the other components have zero mean value and are mutually uncorrelated. For any $\mathfrak{u} \subseteq \{1, \ldots, d\}$, the Sobol' index (Sobol', 1993) of order $r = |\mathfrak{u}|$ associated to the vector $\mathbf{X}_\mathfrak{u}$ is then defined as

$$\mathrm{S}_\mathfrak{u} = \frac{\sigma^2_\mathfrak{u}}{\sigma^2} = \frac{\mathrm{Var}\big[f_\mathfrak{u}(\mathbf{X}_\mathfrak{u})\big]}{\mathrm{Var}[Y]}\,. \tag{3}$$

The corresponding closed Sobol' index is defined as

$$\mathrm{S}^{\mathrm{closed}}_\mathfrak{u} = \sum_{\mathfrak{v} \subset \mathfrak{u}} \mathrm{S}_\mathfrak{v}\,. \tag{4}$$

The main effect of the j^{th} factor is thus measured by $\mathrm{S}_{\{j\}}$. Then the effect due to the specific interaction between the j^{th} and k^{th} factors ($k \neq j$) is measured by $\mathrm{S}_{\{j,k\}}$. And so on for higher order indices (Saltelli et al., 2000).

For any $j \in \{1, \ldots, d\}$, we also define a total sensitivity index $\mathrm{S}^{\mathrm{tot}}_{\{j\}}$ (Homma and Saltelli, 1996) to express the overall sensitivity to an input X_j as

$$\mathrm{S}^{\mathrm{tot}}_{\{j\}} = \sum_{\mathfrak{v} \subset \{1,\ldots d\} \text{ with } j \in \mathfrak{v}} \mathrm{S}_\mathfrak{v}\,. \tag{5}$$

Example 2.1. *Let $d = 3$, $j = 1$ and $k = 2$. Then one has $\mathrm{S}^{\mathrm{closed}}_{\{1,2\}} = \mathrm{S}_{\{1\}} + \mathrm{S}_{\{2\}} + \mathrm{S}_{\{1,2\}}$ and $\mathrm{S}^{\mathrm{tot}}_{\{1\}} = \mathrm{S}_{\{1\}} + \mathrm{S}_{\{1,2\}} + \mathrm{S}_{\{1,3\}} + \mathrm{S}_{\{1,2,3\}}$.*

2.3. Estimating the Sobol' indices

For simple models, sensitivity analysis sometimes can be performed analytically, by direct examination of their mathematical expression. However this is of course generally not the case for complex models. In that case, one evaluates the model for selected values of the input parameters, and the resulting output values are used to estimate sensitivity indices of interest. For interested readers, Monte Carlo based procedures for the estimation of Sobol' indices are described into details in Appendix A.

Let us summarize the different strategies we apply in Section 5:

- In Subsection 5.1, we apply the replication procedure introduced in Mara and Rakoto Joseph (2008) (and further studied in Tissot and Prieur (2015)) to estimate all first-order Sobol' indices with only two replicated d-dimensional Latin Hypercube Sampling of size n, that is with only $2n$ model evaluations.

- In Subsection 5.2, we apply the replication procedure introduced in Tissot and Prieur (2015) to estimate all closed second-order Sobol' indices with only two replicated d-dimensional randomized orthogonal arrays of strength 2 and size n, that is with only $2n$ model evaluations. Due to constraints in the construction of orthogonal arrays of strength 2, n must be chosen as q^2, with q a prime number greater than, or equal to, $d-1$.

- In Subsection 5.3, we apply the procedure introduced in (Saltelli, 2002, Theorem 1) to estimate all first-order and total Sobol' indices with a cost of $(d+2)n$ model evaluations. This procedure is based on combinatorial arguments for a tricky use of model evaluations.

All these strategies are detailed in Appendix A. A crucial point in the above summary is that the cost (in terms of number of model evaluations) required to estimate total Sobol' indices is only linear in the input space dimension d. In the following (see Section 4), we present an implementation using a grid computing environment, with computing resources distributed on 3 sites (in our case, 10 clusters, 6600 cores and some GPUs). Even in that framework, the estimation of total Sobol' indices remain heavy if d is large. Thus, depending on the budget of the study, one could focus on the two first items, i.e. on the estimation of all first-order and closed second-order Sobol' indices as a first step in the sensitivity analysis.

3. Description of the ocean biogeochemical model

The purpose of this section is to describe the applicative context of this study. The 1D biogeochemical model of the ocean mixed layer is described in subsection 3.1. The uncertain input parameters of the model (vector $\mathbf{x}$ in Section 2) are described in subsection 3.2 and several output quantities (y in Section 2) are listed in subsection 3.3, corresponding to key model results which sensitivity to the input parameters is not obvious.

3.1. The MODECOGeL model

The model used in this paper is MODECOGeL[1]. It was developed for investigating the biogeochemical activity in the Ligurian sea by Lacroix (1998)

[1]MODECOGeL: MODèle d'ECOsystème du GHER (GeoHydrodynamics and Environment Research) et du LOV (Laboratoire d'Océanographie de Villefranche-sur-Mer)

by coupling a 1D hydrodynamic model of the mixed layer to a 12-component ecosystem model.

The hydrodynamic model is a 1D version of the GHER primitive equations model (Nihoul and Djenidi, 1987). The state variables are the horizontal velocity, the potential temperature, the salinity, and the turbulent kinetic energy. A full description of the model can be found in Lacroix and Nival (1998) or Lacroix and Grégoire (2002), where it is applied to simulate the behavior of the system during the FRONTAL oceanographic campaigns from 1984 to 1988. In the present paper, the model is applied to years 2006–2007, and the atmospheric dataset is extracted mainly from the Côte d'Azur meteorological buoy located at the DYFAMED station (BOUSSOLE project) at hourly frequency.

Variable	Acronym	Name
C_1	**NO3**	Nitrate
C_2	**NH4**	Ammonium
C_3	**PicP**	Picophytoplankton
C_4	**NanP**	Nanophytoplankton
C_5	**MicP**	Microphytoplankton
C_6	**NanZ**	Nanozooplankton
C_7	**MicZ**	Microzooplankton
C_8	**MesZ**	Mesozooplankton
C_9	**BAC**	Bacteria
C_{10}	**DON**	Dissolved organic nitrogen
C_{11}	**POM1**	Particulate organic matter (size 1)
C_{12}	**POM2**	Particulate organic matter (size 2)

Table 1: Model state variables.

The ecosystem model provides a 12-component description of the ecosystem of the Ligurian Sea (see state variables in Table 1). The time evolution of each state variable is governed by the equation:

$$\frac{\partial C_i}{\partial t} = \mathtt{ADV}_i + \mathtt{DIFF}_i + \mathtt{SMS}_i \quad \text{with} \quad \mathtt{SMS}_i = \sum_{j \neq i} \mathtt{FLUX}(C_j \rightarrow C_i) \qquad (6)$$

where $\mathtt{ADV}_i$ and $\mathtt{DIFF}_i$ are advection and diffusion terms (governed by the hydrodynamic model), and $\mathtt{SMS}_i$ is the "source minus sink" term summing up

the fluxes ($\texttt{FLUX}(C_j \rightarrow C_i)$) between the various components of the ecosystem (conservation of course imposes that $\texttt{FLUX}(C_j \rightarrow C_i) = -\texttt{FLUX}(C_i \rightarrow C_j)$). Equation (6) is solved numerically (between sea surface and 405 m depth) using a constant vertical discretization (1 m) and a constant time step (6 minutes). Outputs are saved daily at all depths.

Without getting into details (a detailed description of the ecosystem model can be found in Lacroix and Grégoire (2002)), let just emphasize the fact that the mathematical expression of these flux terms does contain numerous parameters, which values are not precisely known. In the following, in line with the objective of this study, we only provide a synthetic description of these model parameters, with a specific focus on the assumptions that we have introduced regarding their respective uncertainty.

3.2. Model parameters

The biogeochemical fluxes ($C_j \rightarrow C_i$) parameterized in MODECOGeL are summarized in Table B.6 (Appendix B). Each flux depends on several parameters, which are indicated by referring to the parameter list in Table C.7 (Appendix C). To give an idea of the role of each parameter in MODECOGeL, the biogeochemical fluxes are organized into several categories: primary production, secondary production, mortality, exudation, excretion, growth of bacteria, decomposition of particulate organic matter, and nitrification. We refer to Appendix B for a detailed description of these different categories. This variety of processes, many of them depending on several factors, explains why there are so many parameters in biogeochemical models. Moreover the parameterization of each process as a joint function of the model state and parameters is often complex and nonlinear, with the consequence that it is usually impossible for the user to appreciate the sensitivity of the whole system to the parameters. A rational and automatic approach is thus needed. To apply the method described in Section 2, a probability distribution must be specified for each input parameter.

This has been done here using the following guidelines:

- In absence of any reliable information about possible correlations, uncertainties on the various parameters are assumed independent.

- A majority of parameters are constrained to be either positive or negative. They are assumed to follow a Gamma distribution.

- Parameters constrained between 0 and 1 are assumed to follow a Beta distribution.
- Some parameters are constrained to be larger than 1. Their logarithm is assumed to follow a Gamma distribution.
- Some parameters are not constrained, and are assumed Gaussian.
- Three different values for standard deviations are used (5%, 20%, 50% of the expected value) according to the confidence we have in the parameters.

The resulting probability distributions are given in Table C.7 (Appendix C).

3.3. Quantities of interest

The quantities of interest (i.e. the output values y) must be defined according to the main scientific objectives of the sensitivity study. In the present case, for this illustrative example, we have chosen to focus on characterizing the simulation of phytoplankton (concentrations C_3, C_4, C_5 in Table 1), which is at the basis of the marine food web. As an additional quantity, we also introduce chlorophyll concentration (noted C_0), which is what is observed by ocean color data, and which can be approximately computed from phytoplankton concentrations using a constant chlorophyll to nitrogen ratio (α):

$$C_0 = \alpha(C_3 + C_4 + C_5) \tag{7}$$

To apply the method described in Section 2, since sensitivity analysis applies to scalar output quantities, we need to reduce the time and space variations of C_0, C_3, C_4, C_5 to some scalar indicators. Let us insist on the important fact that the cost of the method is almost independent of the number of these indicators, but they have to be defined before running the sensitivity study. To illustrate the method, we thus decided to introduce many different quantities of interest characterizing C_0, C_3, C_4, C_5, without limiting our choice to simple linear diagnostics.

Table 2 summarizes the quantities of interest Y_{ij} that we will use in our application. The second index j corresponds to the kind of diagnostic that is computed, and the first index i to the concentration ($i = 0, 3, 4, 5$) to which it is applied. This set of 5 diagnostics is meant to characterize (i) the maximum intensity of the phytoplankton spring bloom, (at the surface and as a vertical average), (ii) the time at which it occurs, and (iii) the overall average over the whole simulation.

Index j	Name	Definition
1	surface maximum	$\max_t C_i(0,t)$
2	time of surface maximum	$\mathrm{argmax}_t C_i(0,t)$
3	maximum of vertical average	$\max_t \frac{1}{Z}\int_0^Z C_i(z,t)\,dz$
4	time of maximum of vertical average	$\mathrm{argmax}_t \frac{1}{Z}\int_0^Z C_i(z,t)\,dz$
5	time and vertical average	$\frac{1}{ZT}\int_0^T\int_0^Z C_i(z,t)\,dz\,dt$

Table 2: Quantities of interest Y_{ij}. The maximum depth for averaging is $Z = 40\,\mathrm{m}$, and T is the total duration of the experiment.

4. Practical aspects of the sensitivity analysis

In this section, we go further into details on the numerical implementation of the sensitivity analysis, i.e. the computation of estimates of all first, closed second-order and total Sobol' indices.

4.1. General issues on the implementation

The different steps required to estimate the Sobol' indices are the following:

Step 1 building of a design of experiments (DoE),

Step 2 evaluation of the model on each set of parameters of this DoE,

Step 3 estimation of Sobol' indices based on these model outputs.

Figure 1 shows these three steps, in the particular case of the estimation of all first and closed second-order indices with the replication procedure introduced in Tissot and Prieur (2015) and implemented in the function `sobolroalhs` of the R package `sensitivity` (Pujol et al., 2017).

In most cases, the cost of the implementation is mostly due to the numerous evaluations of the model on the DoE. In our study, each model evaluation is short (approximately 50s), but $2n$ (resp. $(d+2)n$) evaluations are required to estimate all first-order or all closed second-order (resp. all total) Sobol' indices, where n is typically of the order of $10^3 - 10^6$. Furthermore we must manage model's input and output files. Such a problem thus requires specific

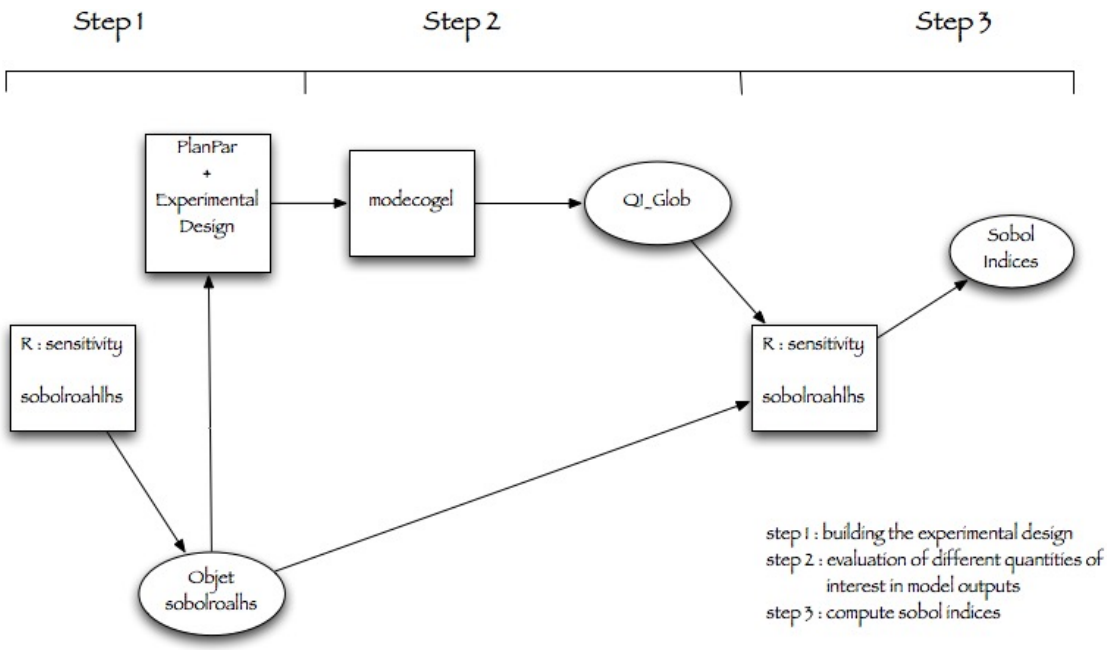


Figure 1: General steps of deployment for the computation of Sobol' indices: example of the estimation of all first-order or all closed second-order Sobol' indices with the `sobolroalhs` function of the R `sensitivity` package (Pujol et al., 2017)

tools to automatically organize data management and access to computational ressources. A grid computing environment, with a distributed storage system and a grid manager, is appropriate for this kind of application since most of the processing is obviously parallel. Note also that, since we have to perform a very large number of model runs, fault tolerance is a necessary property of our computing environment.

In our implementation, we used computing resources available at the University of Grenoble[2], distributed on 3 sites (10 clusters, 6600 cores and some GPUs). These resources are integrated in a local grid[3], associated with several storage nodes distributed on the 3 sites, as close as possible to each supercomputer, and managed by the middleware iRods[4]. Each cluster uses

[2]see https://ciment.ujf-grenoble.fr/wiki-pub/index.php/Welcome_to_the_CIMENT_site!
[3]https://ciment.ujf-grenoble.fr/wiki-pub/index.php/Grid_computing
[4]see irods.org

the local resource manager OAR[5] and the overall access to the 6600 cores of all clusters is achieved through the middleware CIGRI[6] which launches embarrassingly parallel jobs on idle processors. CIGRI acts as a metascheduler of OAR: it retrieves the clusters states through OAR and submits the jobs on free ressources. Furthermore OAR and CIGRI allow the "best effort" mode, i.e. exploiting idle resources of production clusters with a zero priority and resubmitting later the jobs that have been killed because of a local demand for resources.

CIGRI is able to manage job failures in a smart way, allowing the user to submit a large amount of small jobs, and to forget about it until all the jobs are terminated or until CIGRI notifies a serious problem.

As a final step, we need to transfer a lot of input and result files distributed on different iRods resources on a laptop or a computer server of a laboratory. A python script was written to efficiently download the data directly from the outside world.

4.2. Implementation of our study

For the MODECOGeL model presented in Section 3, 74 uncertain input parameters are considered simultaneously, each one associated with a probability distribution provided by physicists using a priori knowledge (see Table C.7).

To build the experimental design and to estimate the Sobol' indices and associated confidence intervals, we used the functions `sobolroalhs` and `sobolSalt` from the R package `sensitivity` (Pujol et al., 2017). Note that R[7] is a free software environment for statistical computing and graphics.

Function `sobolroalhs`, which estimates all first-order and all closed second-order indices, is based on the replication procedure briefly described in Subsection 2.3 and detailed in Appendix A (see Tissot and Prieur, 2015, for more details), making use of four (two for first-order indices and two for closed second-order indices) replicated designs of size n. Let us detail the corresponding grid deployment for the estimation of all first-order Sobol' indices. The different steps are similar if one is interested in the estimation of all closed seond-order indices or in the estimation of all total Sobol' indices, except that the DoEs will differ, as well as the function of the R package

[5]see oar.imag.fr

[6]http://ciment.ujf-grenoble.fr/cigri/dokuwiki/doku.php

[7]http://cran.r-project.org

`sensitivity` we will use. The steps of the grid deployment are the following (see also Figure 2):

- construction of two replicated LHS, each one being of size n;
- evaluation of the model on the DoE: in order to minimize the overhead corresponding to the submission of each evaluation and to optimize the use of the "best effort" mode of the batch manager, the model runs are performed one hundred at a time. Each group of simulations corresponds to 100 different sets of the 74 input parameters, and requires an average cpu time of 84 minutes and a small file in input and output of approximatively 200Ko. For $n = 10^6$, we thus submitted 20 000 jobs on the computing grid with CIGRI. Inputs and outputs are distributed on the storage grid with iRods.
- Merging of the $2n$ files containing the evaluations of the quantities of interest on the DoE.
- Treatment of outliers: a few percents of the model evaluations fail or lead to fully unrealistic results, due to unrealistic combinations of the input parameters. The different quantities of interest are then set to NaN and will be treated as missing values in the computation of Sobol' indices.
- Computation of Sobol' indices with the function `sobolroalhs` of the R package `sensitivity`.

The items which differ for the estimation of closed second-order indices and total Sobol' indices are the first and the fifth ones (see Section 5 for more details).

5. Analysis of Sobol' indices

As indicated in Section 3, we focus in this work on a few quantities of interest (QoI) which are summarized in Table 3. Once again, our aim is not to conduct an extensive in-depth sensitivity analysis of the Modecogel model in its present configuration (we are not specialists of ocean biogeochemistry), but to illustrate the interest of these statistical tools to better understand such complex systems involving many parameters.

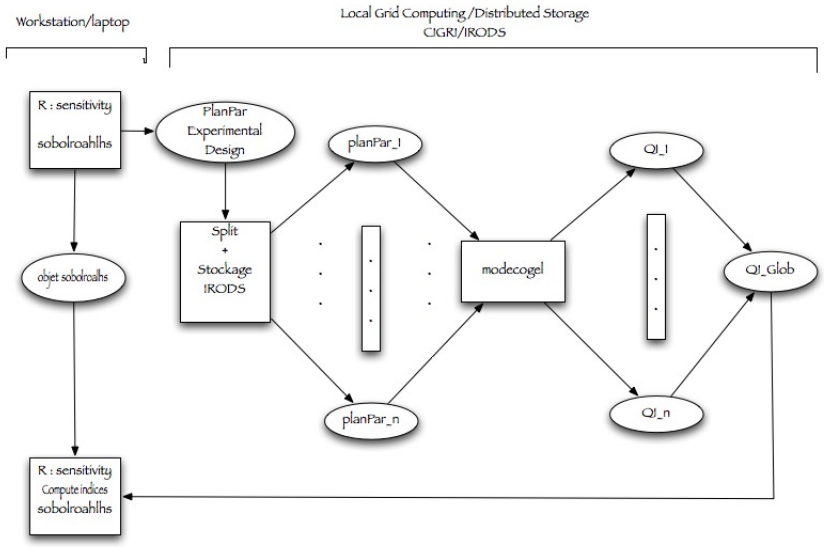


Figure 2: Different steps for the deployment: example of the estimation of all first-order or all closed second-order Sobol' indices with the `sobolroalhs` function of the R `sensitivity` package

5.1. First-order indices

The 74 first-order Sobol' indices were estimated for each QoI, for different values of the sample size: $n = 10^3, 10^4, 10^5, 10^6$. For each index and each value of n, both its estimate and a 95% confidence interval are provided. These results are plotted on Figure 3 for output Y_{01}, namely the maximum surface chlorophyll concentration. As can be seen, the estimation of each index converges in the sense that the length of the 95% confidence interval decreases to zero as n increases, and $n = 10^6$ seems to be a sufficiently large sample size to get accurate estimations.

Only about ten parameters have a first-order index greater than 0.01 (their values are reported in Table 4). Quite similar results are actually obtained for the other QoIs, and it appears that only 15 (resp. 10) model parameters have a first-order Sobol' index greater than 0.01 for at least one (resp. 6) QoIs. This is summarized in Figure 4, where these most influential model parameters are clearly visible. This mostly highlights the important sensitivity of our QoIs to the parameterization of excretion for bacteria, of grazing and ingestion for mesozooplankton, and of the variation of light limitation

Quantity of interest	Y_{ij}	Description
maxc	Y_{01}	annual maximum of surface chlorophyll concentration
timechl	Y_{02}	time of maximum of surface chlorophyll concentration
moyc	Y_{05}	time and vertical average of chlorophyll concentration
maxnp	Y_{41}	surface maximum of nanophytoplankton concentration
timenp	Y_{42}	time of maximum of surface nanophytoplankton concentration
maxmoynp	Y_{43}	maximum of vertical average of nanophytoplankton concentration
timemoynp	Y_{44}	time of maximum of vertical average of nanophytoplankton concentration
moynp	Y_{45}	time and vertical average of nanophytoplankton concentration
maxpp	Y_{31}	maximum of surface picophytoplankton concentration
timepp	Y_{32}	time of maximum of surface picophytoplankton concentration
timemoypp	Y_{34}	time of maximum of vertical average of picophytoplankton concentration
maxmoypp	Y_{33}	maximum of vertical average of picophytoplankton concentration

Table 3: Quantities of interest analyzed in the present work

for phytoplankton.
This low number of influential parameters may indicate that an efficient reduction of this model could be performed, in order to end with a simplified model involving much less parameters. Note however that the sum of first-order indices is equal to 0.423. This is far less than 1, which clearly indicates that the considered outputs cannot be seen as simple additive models of the form $Y = f_1(X_1) + \ldots + f_d(X_d)$. There exists some interaction effects of the parameters on the QoIs, which motivates the investigation of second-order interaction effects.

parameter number (see Table C.7)	67	63	36	30	35	57	18	15	24	73
estimated index	0.10	0.058	0.048	0.035	0.035	0.034	0.017	0.015	0.011	0.011
estimated error	0.0018	0.0019	0.0019	0.0017	0.0018	0.0020	0.0020	0.0016	0.0016	0.0019

Table 4: Estimation of first-order Sobol' indices for the output Y_{01} (annual maximum of chlorophyll concentration). The estimated error is the radius (half of the length) of the 95% confidence interval.

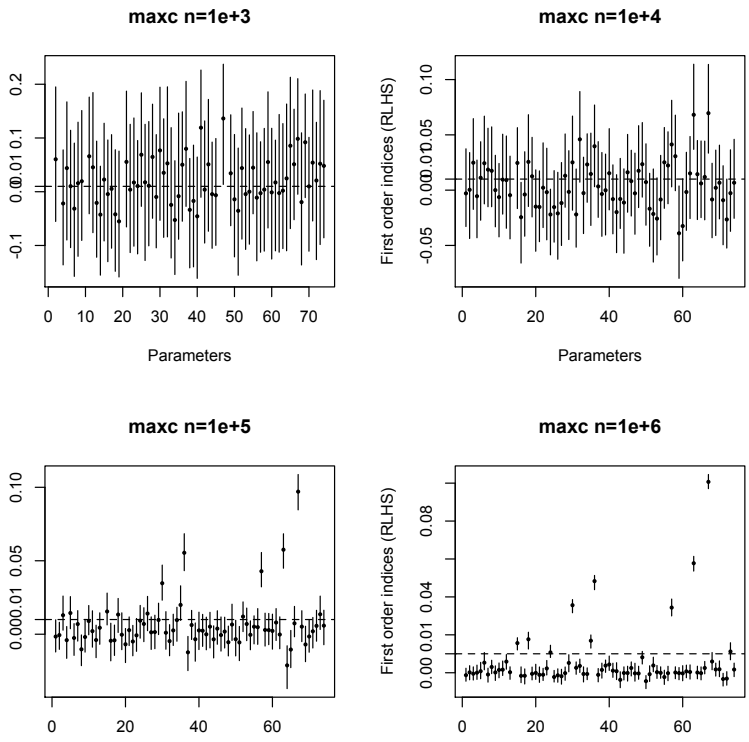


Figure 3: Estimated first-order indices (y-axis) with their 95% confidence interval for the 74 model parameters (x-axis), for $n = 10^3, 10^4, 10^5$ and 10^6, in the case of the output Y_{01}. The dashed horizontal line corresponds to a threshold arbitrarily chosen to 0.01. Confidence intervals were obtained with a bootstrap procedure (e.g. Archer et al. (1997)) and a bootstrap sample size of 100.

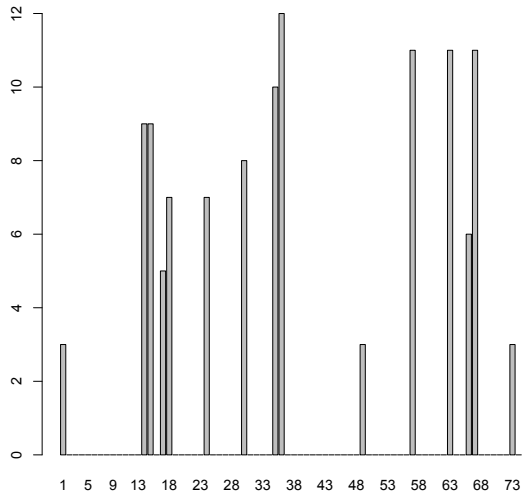


Figure 4: Number of quantities of interest (y-axis) for which the estimate of the first-order index (x-axis) is greater than 0.01

5.2. Second order indices

As indicated before, the `sobolroalhs` function allows the estimation of closed second-order indices at a cost of $2n$ model evaluations. Let us recall that the closed second-order index corresponding to the j^{th} and k^{th} parameters is defined as:

$$\mathrm{S}^{\text{closed}}_{\{j,k\}} = \mathrm{S}_{\{j\}} + \mathrm{S}_{\{k\}} + \mathrm{S}_{\{j,k\}}$$

and corresponds to the sum of the main and interaction effects due to these two parameters.

A specificity for the estimation of closed second-order indices with the replication procedure is that n has to be chosen equal to q^2, where $q \geq d-1$ is a prime number denoting the number of levels of the orthogonal array (see Appendix A for more details). A value of $q = 227$, i.e. $2n = 103\,058$, was necessary to achieve sufficiently accurate estimations of the indices.

Estimates of the $d(d-1)/2 = 2701$ closed second-order indices, along with their corresponding 95% bootstrap confidence intervals, were then computed for each QoI. They are displayed in Figure 5 for the output Y_{01} (annual maximum of chlorophyll concentration, which was already chosen in Figure 3 and Table 4). It clearly appears that very few of these 2701 indices are significant. Given the definition of the $\mathrm{S}^{\text{closed}}_{\{j,k\}}$, most of them obviously correspond to at least one influential parameter listed in Table 4, as clearly displayed in Figure 6a.

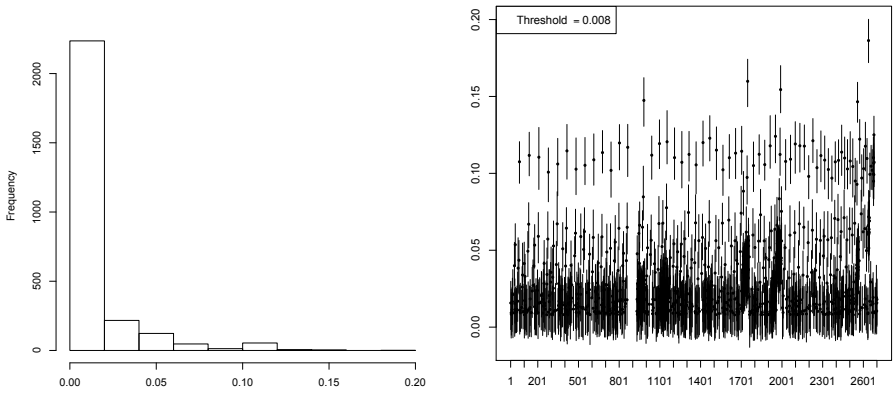
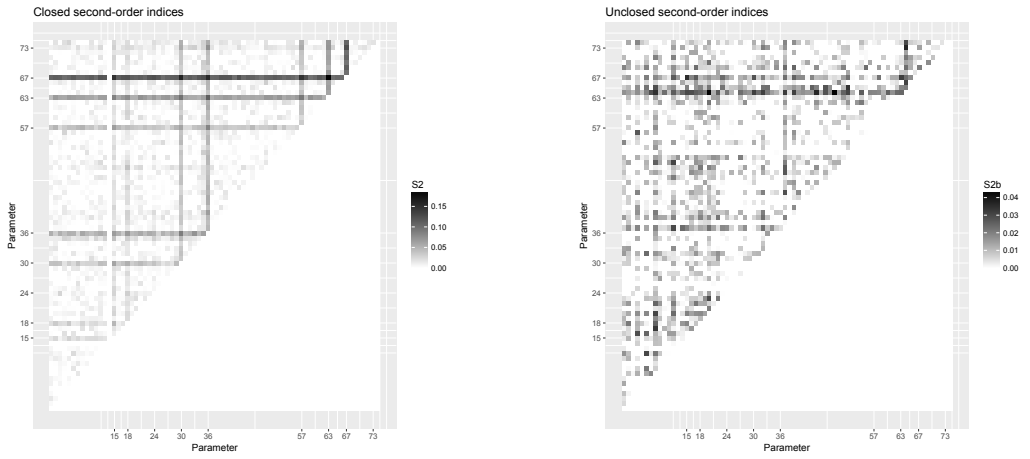


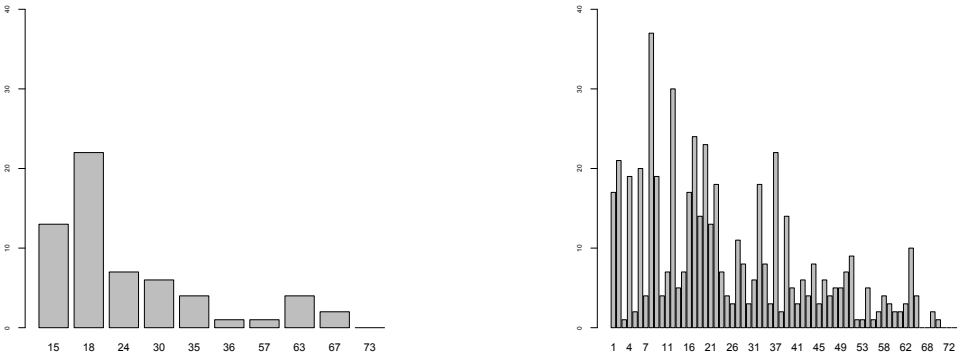
Figure 5: Estimation of second-order closed indices for QoI Y_{01}. Left panel: histogram of the values of the 2701 indices. Right panel: indices greater than 0.008 with their associated 95% bootstrap confidence intervals (with a bootstrap sample size equal to 100).

Then, using jointly these estimates with the previous estimates of first order Sobol' indices, consistent estimates of unclosed second-order Sobol' indices $S_{\{j,k\}}$ were also computed with a bootstrap algorithm. These indices quantify the fraction of variance of the QoI due to these two parameters that cannot be explained by the sum of the main effect of X_i and X_j but only by their interaction. They are displayed in Figure 6b in the particular case of output Y_{01}. New interesting information arise from this plot. First of all, it appears that input parameters with strong direct influence (i.e. strong first order indices) may or may not have significant interactions with other parameters. For instance, parameters 36, 57 and 73 (see Table C.7), related to semisaturation for ingestion by MesZ, fraction of grazing used for growth of MesZ and light attenuation coefficient in sea water, have weak second order indices. Conversely, parameters 15 and 18, related to variation of light limitation for NanP and to optimal temperature for NanP, which already correspond to large first order indices, have also strong interactions with many other parameters (see Figure 7a). Finally, some other parameters that have not been noticed yet, since they have weak first order indices, exhibit strong interactions with numerous other parameters. Let us mention in particular parameters 8 and 12, related to NH4 semisaturation for PicP and to optimal PAR for NanP (see Figure 7b). A summary of these remarks, and additional information, is presented in Figure 7. The threshold in this Figure 7 can be chosen arbitrarily. However the choice 0.01 guarantees a clear separation between inputs with first-order index smaller or larger than the threshold.



(a) Second-order closed indices $S^{closed}_{\{j,k\}}$ (b) Second-order unclosed indices $S_{\{j,k\}}$

Figure 6: Maps (74×74) of the second-order closed and unclosed Sobol indices for QoI Y_{01}. The x and y axes correspond to the number of the parameters, and the grey scale to the value of the index. Note that the numbers indicated on the axes correspond to parameters with high first-order indices.



(a) $S_{\{1\}} \geq 0.01$ (b) $S_{\{1\}} \leq 0.01$

Figure 7: Number of second-order unclosed indices greater than 0.008 (y-axis), for input parameters with high first-order indices (left panel) and for input parameters with weak first-order indices (right panel). x-axis : number of the input parameter.

At this stage, the information brought by first- and second-order indices remains incomplete. Higher order interactions can be expected, especially for highly parameterized models like in the present context of marine biogeochemistry. A natural step forward in the study is then to compute total indices.

5.3. Total order indices

As already defined by Eq. (5), the total Sobol' index $S^{tot}_{\{j\}}$ ($j \in \{1, ..., d\}$) expresses the overall sensitivity of some QoI to the input X_j. As mentioned in Section 2 and Appendix A, a competitive procedure to estimate simultaneously all first and total Sobol' indices is introduced in Saltelli (2002) and implemented in the function `sobolSalt` of the R package `sensitivity`. Its cost is however still quite high, since it requires $(d + 2)n$ model evaluations, but this linear dependency w.r.t. the input space dimension d cannot be avoided. Therefore, depending on the budget available for the study, computing such indices is not always affordable, although they bring an important information, since input parameters with a very low value for their total Sobol' index can be fixed to a nominal value in a calibration procedure.

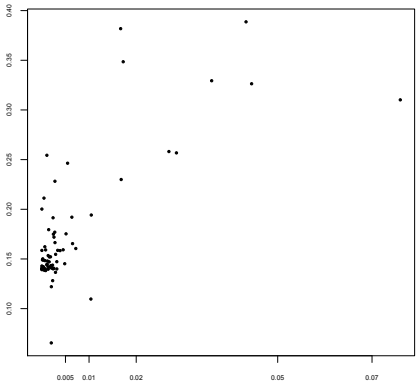
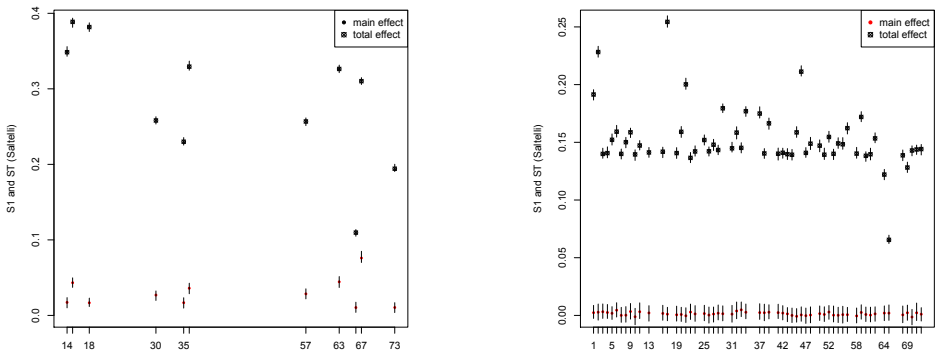


Figure 8: Scatterplot of first-order (x-axis) and total (y-axis) Sobol' indices for output Y_{01}. Note that, for sake of readability, the x-axis is stretched w.r.t. the y-axis.

A scatter plot of first-order and total indices is displayed in Figure 8. We observe globally that the parameters contribute to the total variance mostly through their interactions with other parameters (the dots are far above the line $y = x$). However the relation between first-order and total indices is not linear. More precisely, parameters having a major main effect generally

also contribute to the total variance through their interactions with other parameters (Figure 9a). However the relative importance of these interactions may vary significantly. For instance, parameter 67, which corresponds to the largest first-order index, has a total index smaller than several other parameters with much smaller first-order index. On the other hand, some parameters having a very small main effect contribute to the total variance in a non-negligible manner by their interactions with the other parameters (Figure 9b). This heterogeneous distribution of the relative importance of the first-order effect w.r.t. the total effect is synthesized in Figure 10, where the ratio $\mathrm{S}_{\{j\}}/\mathrm{S}^{\mathrm{tot}}_{\{j\}}$ is displayed.



(a) $S_{\{1\}} \geq 0.01$ (b) $S_{\{1\}} \leq 0.005$

Figure 9: First-order and total Sobol' indices and corresponding confidence intervals, for output Y_{01}. Left panel: parameters with first-order index larger than 0.01 only. Right panel: parameters with first-order index smaller than 0.005 only.

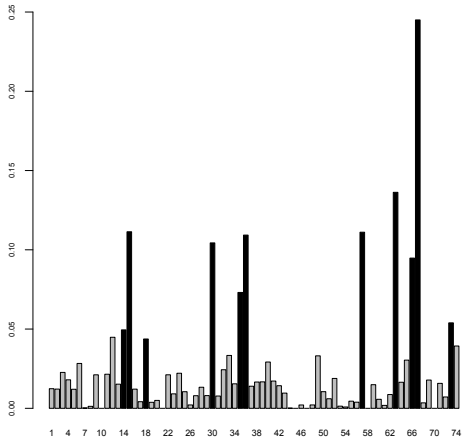


Figure 10: Estimation of the ratio $S_{\{j\}}/S^{tot}_{\{j\}}$ for $j = 1, \ldots, 74$. Black bars correspond to parameters with first-order index larger than 0.01.

5.4. Computation costs

The computation cost of the different procedures is displayed in Table 5, along with the estimated error related to the Sobol' indices (i.e. the half-length of the 95% bootstrap confidence interval). As mentioned earlier, the computation of total indices requires a number of model runs proportional to $d + 2$, which is much more than for the evaluation of only first- and second-order indices. Moreover the estimation error is inversely proportional to $\sqrt{n}$, as it is mainly the case in Monte-Carlo procedures, consistently with the rate of convergence in the central limit theorem. Note that other designs of experiments such as quasi Monte Carlo procedures are not considered in this paper, since we do not assume any regularity on the underlying model. In the present case, given the order of magnitude of the indices, $n = 10^5$ appears as a good compromise between accuracy and computation cost.

Estimation of Sobol' indices	`sobolSalt` n=10^5		`roalhs` n=10^3	`roalhs` n=10^4	`roalhs` n=10^5	`roalhs` n=10^6	`roalhs` q=227
	S1	ST	S1	S1	S1	S1	S2
Estimated Error							
maximum value	0.0076	0.0064	0.077	0.025	0.0077	0.0025	0.024
mean value	0.0065	0.0047	0.062	0.020	0.0064	0.0020	0.018
standard deviation	$2.0\,10^{-7}$	$2.3\,10^{-7}$	$2.7\,10^{-5}$	$3.4\,10^{-6}$	$5.7\,10^{-8}$	$3.2\,10^{-8}$	$2.7\,10^{-7}$
Number of evaluations	$7.6\,10^6$		$2\,10^3$	$2\,10^4$	$2\,10^5$	$2\,10^6$	$q^2 \simeq 10^5$

Table 5: Statistics (maximum and mean values, standard deviation) related to the estimated error over all 74 parameters, and number of model runs required for the estimation of the Sobol' indices.

6. Conclusion

A global sensitivity analysis, based on Sobol' indices, is presented in the context of a realistic ocean biogeochemical model. Without going into complex physical interpretations, for which the authors' skills are very limited, the aim of this study is to demonstrate the potential interest of this systematic and mathematically sound approach for such applications. This method quantifies the influence of uncertain input parameters on an arbitrary number of output quantities of interest, either through their direct effect or through their mutual interactions. In the present study, we started by computing first- and second-order indices. This highlighted on one hand the strong influence of some input parameters and of some second order interactions, and on the other hand that some higher order interactions were probably significant. Total Sobol' indices were thus computed, which brought complementary information. Note also that such total indices can be a very helpful information in the perspective of model reduction, since it indicates parameters which value could be frozen.

Such a global sensitivity analysis, though requiring a large number of model runs, is now fully feasible in a grid computing environment. Large sets of input parameters can be considered, and indices as expensive as total indices can be computed. Note that the full study was performed with a full Monte-Carlo procedure, thus avoiding any restrictive hypothesis on the effective dimension of the model nor on its regularity.

Appendix A. Estimation procedure for Sobol' indices

In this paper, we used a Monte Carlo based procedure for the estimation of Sobol' indices (Sobol', 1993), which we describe in details below. Let us first introduce some notation. Let $\mathfrak{u}$ be a non-empty subset of $\{1, \dots d\}$. Let $\mathbf{X}^1 = (X_1^1, \dots, X_d^1)$ and $\mathbf{X}^2 = (X_1^2, \dots, X_d^2)$ be two independent copies of the input vector $\mathbf{X} = (X_1, \dots, X_d)$. Let $(\mathbf{X}^{1,i})_{i=1,\dots,n}$ be an *independent and identically distributed* (i.i.d.) sample of size n of vector $\mathbf{X}^1$ and $(\mathbf{X}^{2,i})_{i=1,\dots,n}$ be an i.i.d. sample of size n of vector $\mathbf{X}^2$. For any $\mathfrak{u} \subset \{1, \dots, d\}$, we define the d-dimensional vector $\mathbf{X}_\mathfrak{u} = (X_{\mathfrak{u},1}, \dots, X_{\mathfrak{u},d})$ in the following way: for any $j \in \mathfrak{u}$, $X_{\mathfrak{u},j}$ is equal to X_j^1, and for any $j \in \mathfrak{u}^c$, $X_{\mathfrak{u},j} = X_j^2$. The n-sample $(\mathbf{X}_\mathfrak{u}^i)_{i=1,\dots,n}$ is defined in a similar manner.

And we define the corresponding output variables:

$$Y = f(\mathbf{X}^1)\,,\ Y_\mathfrak{u} = f(\mathbf{X}_\mathfrak{u}) \text{ and } Y^i = f(\mathbf{X}^{1,i})\,,\ Y_\mathfrak{u}^i = f(\mathbf{X}_\mathfrak{u}^i). \tag{A.1}$$

It is possible to prove that $\mathrm{S}_\mathfrak{u}^{\text{closed}} = \dfrac{\mathrm{Cov}(Y, Y_\mathfrak{u})}{\mathrm{Var}[Y]}$ (see, e.g., Lemma 1.2 in Janon et al. (2014)). Based on the above formula, we propose the following estimator for $\mathrm{S}_\mathfrak{u}^{\text{closed}}$:

$$\hat{\mathrm{S}}_{\mathfrak{u},n}^{\text{closed}} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^n Y^i Y_\mathfrak{u}^i - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n \frac{Y^i + Y_\mathfrak{u}^i}{2}\right)^2}{\dfrac{1}{n}\displaystyle\sum_{i=1}^n \frac{(Y^i)^2 + (Y_\mathfrak{u}^i)^2}{2} - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n \frac{Y^i + Y_\mathfrak{u}^i}{2}\right)^2}\,. \tag{A.2}$$

This estimator was first introduced in Monod et al. (2006). Its asymptotic properties are stated in Janon et al. (2014, Propositions 2.2 and 2.5). See also Gamboa et al. (2016) for further asymptotic and non-asymptotic results.

Since $\mathrm{S}_{\{j\}} = \mathrm{S}_{\{j\}}^{\text{closed}}$ (see Eq. (4)), we therefore estimate $\mathrm{S}_{\{j\}}$ with $\hat{\mathrm{S}}_{\{j\},n} = \hat{\mathrm{S}}_{\{j\},n}^{\text{closed}}$.

Recall now the law of total variance which can be found, e.g., in (Weiss, 2006, pages 385-386):

$$\mathrm{Var}[Y] = \mathrm{Var}\big(\mathrm{E}(Y|\mathbf{X}_\mathfrak{u})\big) + \mathrm{E}\big(\mathrm{Var}(Y|\mathbf{X}_\mathfrak{u})\big)\,. \tag{A.3}$$

From the definition of total Sobol' indices, we can prove that:

$$\mathrm{S}_{\{j\}}^{\text{tot}} = \frac{\mathrm{E}\big(\mathrm{Var}(Y|X_{\{j\}^c})\big)}{\mathrm{Var}[Y]}\,. \tag{A.4}$$

From (A.3) and (A.4) we deduce $\mathrm{S}^{\mathrm{tot}}_{\{j\}} = 1 - \mathrm{S}^{\mathrm{closed}}_{\{j\}^c}$. We thus define the estimator of $\mathrm{S}^{\mathrm{tot}}_{\{j\}}$ as:

$$\hat{\mathrm{S}}^{\mathrm{tot}}_{\{j\},n} = 1 - \hat{\mathrm{S}}^{\mathrm{closed}}_{\{j\}^c,n}\,.$$

These estimators require a large number of model evaluations. Indeed, the estimation of all first-order indices actually requires $(d+1)n$ evaluations of the model: Y^i, $Y^i_{\{j\}}$, $i = 1, \ldots, n$, $j = 1, \ldots, d$. The estimation of all closed second-order indices requires $(\binom{d}{2} + 1)n$ model evaluations of the model: Y^i, $Y^i_{\{j,k\}}$, $i = 1, \ldots, n$, $j \neq k = 1, \ldots, d$.

The Monte Carlo sample size n is directly related to the accuracy of the estimation via the rate of convergence in the central limit theorem, which is of order $\sqrt{n}$ (see Proposition 2.2 in Janon et al. (2014)).

To circumvent this linear (*resp.* quadratic) dependence of the cost (in terms of number of model evaluations) in the input space dimension d, the authors in Tissot and Prieur (2015) proposed a procedure based on replicated Latin Hypercube Sampling (LHS) (e.g. Lemieux, 2009) (*resp.* replicated randomized orthogonal arrays) to estimate all first-order (*resp.* all closed second-order) Sobol' indices. The ideas in Tissot and Prieur (2015) generalize some pioneer work in Mara and Rakoto Joseph (2008). We refer to Tissot and Prieur (2015) for a detailed description of the procedure, and a rigorous analysis of its asymptotic properties. In the present paper, we applied the replication procedure to estimate all first-order Sobol' indices in Subsection 5.1 and to estimate all closed second-order Sobol' indices in Subsection 5.2. More precisely, the replication procedure in Subsection 5.1 allows to estimate all the first-order Sobol' indices with only two replicated d-dimensional LHS of size n, that is with only $2n$ model evaluations. The replication procedure in Subsection 5.2 allows to estimate all closed second-order Sobol' indices with only two replicated d-dimensional randomized orthogonal arrays of strength two and size n, that is with only $2n$ model evaluations. Due to constraints in the construction of orthogonal arrays of strength two, n must be chosen as q^2, with q a prime number greater or equal to $d-1$. Note that orthogonal arrays were introduced for the first time in Kishen (1942).

If higher order interactions are expected in the model (of any order greater or equal to three), one may be interested in estimating total Sobol' indices. The replication procedure does not adapt to the estimation of total Sobol' indices, mainly because of the constraints in the construction of orthogonal arrays of strength higher or equal to three. In Saltelli (2002), the authors

propose two different procedures, both based on combinatorial tricks: the first one allows the estimation of all first-order and total Sobol' indices with a cost of $(d+2)n$ model evaluations (see Theorem 1 in Saltelli (2002), the second one leads to a double estimate of all first-order, closed second-order and total Sobol' indices at a cost of $(2d+2)n$ model evaluations (see Theorem 2 in Saltelli (2002)). To our knowledge, the procedure in (Saltelli, 2002, Theorem 1) is the most competitive one if the estimation of total Sobol' indices is involved (see also Gilquin et al. (2017)). It is the one we have applied in Subsection 5.3.

Appendix B. Description of biogemical fluxes

The biogeochemical fluxes ($C_j \to C_i$) parameterized in MODECOGeL are summarized in Table B.6. Each flux depends on several parameters, which are indicated by referring to the parameter list in Table C.7 in Appendix C. To give an idea of the role of each parameter in MODECOGeL, the biogeochemical fluxes are organized into several categories using different colors:

- **Primary production (green)** is the growth of phytoplanktons by photosynthesis. In Table B.6, this corresponds to all fluxes from nutrients (C_1, C_2) to phytoplanktons (C_3, C_4, C_5). Parameters govern the maximum growth rate (1–3), and the dependence to nutrient concentrations (4–10), to solar irradiance (11-16), and to temperature (18–20).

- **Secondary production (blue)** is the growth of zooplanktons by grazing of phytoplanktons or by assimilation of bacteria and particulate organic matters. In Table B.6, this corresponds to all fluxes to zooplanktons (C_6, C_7, C_8). Parameters govern the ingestion rate (28–33), the dependence to prey concentration (34–36), the efficiency according to the type of prey (37–43), and the fraction actually used for growth (55-59).

- **Mortality (red)** of living species, including a parameterization of predation by higher trophic levels. In Table B.6, this corresponds to all fluxes from phytoplanktons or zooplanktons or bacteria (C_3 to C_9) to particulate organic matter (C_{11} or C_{12}). Parameters govern mortality rates (44–50) and predation (51–53).

- **Exudation (magenta)** by phytoplanktons. In Table B.6, this corresponds to all fluxes from phytoplanktons (C_3, C_4, C_5) to dissolved organic nitrogen (C_{10}). Parameters are exudation rates (25–27).

- **Excretion (pink)** by zooplanktons and bacteria. In Table B.6, this corresponds to all fluxes from zooplanktons or bacteria (C_6 to C_9) to ammonium and dissolved organic nitrogen (C_2 and C_{10}). Parameters govern excretion rates (60–63), the dependence to temperature (64–67), the tradeoff between ammonium and dissolved organic matter (68), and the excreted fraction of predation (54).

- **Growth of bacteria (yellow)** from ammonium and dissolved organic matter. In Table B.6, this corresponds to all fluxes to bacteria (C_9). Parameters govern the growth rate (23–24).
- **Decomposition of particulate organic matter (orange)**. In Table B.6, this corresponds to all fluxes from particulate organic matter (C_{11} or C_{12}) to dissolved organic nitrogen (C_{10}). Parameters are decomposition rates (69–70).
- **Nitrification (brown)**. In Table B.6, this corresponds to the flux from ammonium (C_2) to nitrate (C_1). The parameter is the nitrification rate (72).

	Nutrients		Phytoplanktons			Zooplanktons			BAC	DON & POM		
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
C_1			1,4,5, 11,14, 17,20	2,4,6, 12,15, 18,21	3,4,7, 13,16, 19,22							
C_2	72	72	1,8, 11,14, 17,20	2,9, 12,15, 18,21	3,10, 13,16, 19,22				23,24			
C_3			1,5,8, 11,14, 17,20, 25,44			28,31, 34,55				25	44	
C_4				2,6,9, 12,15, 18,21, 26,45			29,32, 35,56			26	45	
C_5					3,7,10, 13,16, 19,22, 27,46			30,33, 36,37, 57		27	46	
C_6						28,31, 34,47, 55,60, 64	29,32, 35,39, 56			60,64, 68	47	
C_7							29,32, 35,48, 56,61, 65	30,33, 36,40, 57		61,65, 68	48	
C_8								30,33, 36,49, 52,53, 54,57, 62,66		54,62, 66,68		49,51, 52,53
C_9						28,31, 34,38, 55	29,32, 35,56		23,24, 50,63, 67		50	
C_{10}									23,24			
C_{11}							29,32, 35,41, 58	30,33, 36,42, 59		69	69	
C_{12}								30,33, 36,43, 59		70		70

Table B.6: Biogeochemical fluxes from variable C_i (line i) to variable C_j (column j). Numbers in the boxes refer to parameter indices, given in Table C.7.

Appendix C. Model parameters

This table describes the different probability distributions chosen for input parameters.

Index	Name	Unit	Pdf	Mean	Std	Std/Mean
1	**PicP** max growth rate	t^{-1}	$\Gamma(25, 0.12)$	3.	0.6	20%
2	**NanP** max growth rate	t^{-1}	$\Gamma(25, 0.1)$	2.5	0.5	20%
3	**MicP** max growth rate	t^{-1}	$\Gamma(25, 0.08)$	2.	0.4	20%
4	dependence of **NO3** limitation to **NH4**	C^{-1}	$\Gamma(400, 0.00365)$	1.46	0.073	5%
5	**NO3** semisaturation for **PicP**	C	$\Gamma(4, 0.125)$	0.5	0.25	50%
6	**NO3** semisaturation for **NanP**	C	$\Gamma(4, 0.175)$	0.7	0.35	50%
7	**NO3** semisaturation for **MicP**	C	$\Gamma(4, 0.25)$	1.0	0.5	50%
8	**NH4** semisaturation for **PicP**	C	$\Gamma(4, 0.075)$	0.3	0.15	50%
9	**NH4** semisaturation for **NanP**	C	$\Gamma(4, 0.125)$	0.5	0.25	50%
10	**NH4** semisaturation for **MicP**	C	$\Gamma(4, 0.175)$	0.7	0.35	50%
11	optimal PAR for **PicP**	I	$\Gamma(25, 0.4)$	10.	2.	20%
12	optimal PAR for **NanP**	I	$\Gamma(25, 0.6)$	15.	3.	20%
13	optimal PAR for **MicP**	I	$\Gamma(25, 0.8)$	20.	4.	20%
14	variation of light limitation for **PicP**	–	$-\Gamma(4, 0.2)$	-0.8	0.4	50%
15	variation of light limitation for **NanP**	–	$-\Gamma(4, 0.175)$	-0.7	0.35	50%
16	variation of light limitation for **MicP**	–	$-\Gamma(4, 0.15)$	-0.6	0.3	50%
17	optimal temperature for **PicP**	T	$\mathcal{N}(15, 3^2)$	15.	3.	20%
18	optimal temperature for **NanP**	T	$\mathcal{N}(15, 3^2)$	15.	3.	20%
19	optimal temperature for **MicP**	T	$\mathcal{N}(16, 3.2^2)$	16.	3.2	20%
20	variation of temp. limitation for **PicP**	–	$-\Gamma(4, 0.125)$	-0.5	0.25	50%
21	variation of temp. limitation for **NanP**	–	$-\Gamma(4, 0.125)$	-0.5	0.25	50%
22	variation of temp. limitation for **MicP**	–	$-\Gamma(4, 0.1375)$	-0.55	0.275	50%
23	bacteria growth limitation	–	$\Gamma(4, 0.15)$	0.6	0.3	50%
24	semisaturation for **BAC** growth	C	$\Gamma(4, 0.125)$	0.5	0.25	50%
25	exudation ratio for **PicP**	–	$\Gamma(4, 0.015)$	0.06	0.03	50%
26	exudation ratio for **NanP**	–	$\Gamma(4, 0.0125)$	0.05	0.025	50%
27	exudation ratio for **MicP**	–	$\Gamma(4, 0.01)$	0.04	0.02	50%
28	max ingestion rate for **NanZ**	t^{-1}	$\Gamma(25, 0.12)$	3.	0.6	20%
29	max ingestion rate for **MicZ**	t^{-1}	$\Gamma(25, 0.08)$	2.	0.4	20%
30	max ingestion rate for **MesZ**	t^{-1}	$\Gamma(25, 0.06)$	1.5	0.3	20%
31	threshold ingestion for **NanZ**	C	$\Gamma(4, 0.0125)$	0.05	0.025	50%
32	threshold ingestion for **MicZ**	C	$\Gamma(4, 0.0075)$	0.03	0.015	50%
33	threshold ingestion for **MesZ**	C	$\Gamma(4, 0.0025)$	0.01	0.005	50%
34	semisaturation for ingestion by **NanZ**	C	$\Gamma(4, 0.125)$	0.5	0.25	50%
35	semisaturation for ingestion by **MicZ**	C	$\Gamma(4, 0.1875)$	0.75	0.375	50%
36	semisaturation for ingestion by **MesZ**	C	$\Gamma(4, 0.25)$	1.	0.5	50%
37	efficiency of **MesZ** on **MicP**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
38	efficiency of **NanZ** on **BAC**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
39	efficiency of **MicZ** on **NanZ**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
40	efficiency of **MesZ** on **MicZ**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
41	efficiency of **MicZ** on **MOP1**	–	$\beta(19.8, 79.2)$	0.2	0.04	20%
42	efficiency of **MesZ** on **MOP1**	–	$\beta(19.8, 79.2)$	0.2	0.04	20%
43	efficiency of **MesZ** on **MOP2**	–	$\beta(19.8, 79.2)$	0.2	0.04	20%
44	mortality rate for **PicP**	t^{-1}	$\Gamma(4, 0.015)$	0.06	0.03	50%
45	mortality rate for **NanP**	t^{-1}	$\Gamma(4, 0.0125)$	0.05	0.025	50%
46	mortality rate for **MicP**	t^{-1}	$\Gamma(4, 0.01)$	0.04	0.02	50%
47	mortality rate for **NanZ**	t^{-1}	$\Gamma(4, 0.015)$	0.06	0.03	50%

Index	Name	Unit	Pdf	Mean	Std	Std/Mean
48	mortality rate for **MicZ**	t^{-1}	$\Gamma(4, 0.0125)$	0.05	0.025	50%
49	mortality rate for **MesZ**	t^{-1}	$\Gamma(4, 0.0075)$	0.03	0.015	50%
50	mortality rate for **BAC**	t^{-1}	$\Gamma(4, 0.015)$	0.06	0.03	50%
51	threshold for predation	C	$\Gamma(4, 0.005)$	0.02	0.01	50%
52	maximum predation rate on **MesZ**	t^{-1}	$\Gamma(4, 0.25)$	1.	0.5	50%
53	semisaturation for predation on **MesZ**	C	$\Gamma(4, 0.25)$	1.	0.5	50%
54	excreted fraction of predation on **MesZ**	–	$\beta(2.33, 4.67)$	0.333	0.167	50%
55	fraction of grazing used for growth of **NanZ**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
56	fraction of grazing used for growth of **MicZ**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
57	fraction of grazing used for growth of **MesZ**	–	$\beta(4.2, 1.05)$	0.8	0.16	20%
58	fraction of POM used for growth of **MicZ**	–	$\beta(12, 12)$	0.5	0.1	20%
59	fraction of POM used for growth of **MesZ**	–	$\beta(12, 12)$	0.5	0.1	20%
60	excretion rate for **NanZ**	t^{-1}	$\Gamma(4, 0.0375)$	0.15	0.075	50%
61	excretion rate for **MicZ**	t^{-1}	$\Gamma(4, 0.025)$	0.1	0.05	50%
62	excretion rate for **MesZ**	t^{-1}	$\Gamma(4, 0.0125)$	0.05	0.025	50%
63	excretion rate for **BAC**	t^{-1}	$\Gamma(4, 0.0375)$	0.15	0.075	50%
64	temperature variation of excretion for **NanZ**	–	LogGamma	1.05	0.0525	5%
65	temperature variation of excretion for **MicZ**	–	LogGamma	1.05	0.0525	5%
66	temperature variation of excretion for **MesZ**	–	LogGamma	1.02	0.051	5%
67	temperature variation of excretion for **BAC**	–	LogGamma	1.04	0.052	5%
68	fraction of excretion as **DOM**	–	$\beta(2.75, 8.25)$	0.25	0.125	50%
69	POM1 decomposition rate	t^{-1}	$\Gamma(4, 0.01625)$	0.065	0.0325	50%
70	POM2 decomposition rate	t^{-1}	$\Gamma(4, 0.015)$	0.06	0.03	50%
71	sedimentation velocity for **MicP**	V	$\Gamma(4, 0.25)$	1.	0.5	50%
72	nitrification rate	t^{-1}	$\Gamma(4, 0.0075)$	0.03	0.015	50%
73	light attenuation coefficient in sea water	–	$\Gamma(25, 0.0016)$	0.04	0.008	20%
74	fraction of photosynthetically active radiation	–	$\Gamma(25, 0.02)$	0.5	0.1	20%

Table C.7: Model parameters X_i. Units are: time t in days, concentration C in mmolN/m^3, irradiance I in W/m^2, and velocity V in m/day. The notation $-\Gamma(.,.)$ means that the parameter is negative and that its opposite follows a $\Gamma(.,.)$ distribution.

References

G. Archer, A. Saltelli, and I. Sobol'. Sensitivity measures, ANOVA-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58(2):99–120, 1997.

M. Baklouti, V. Faure, L. Pawlowski, and A. Sciandra. Investigation and sensitivity analysis of a mechanistic phytoplankton model implemented in a new modular numerical tool (eco3m) dedicated to biogeochemical modelling. *Progress in Oceanography*, 71(1):34–58, 2006.

J.-N. Druon and J. Le Fèvre. Sensitivity of a pelagic ecosystem model to variations of process parameters within a realistic range. *Journal of Marine Systems*, 19(1-3):1–26, 1999.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.

B. Faugeras, M. Lévy, L. Mémery, J. Verron, J. Blum, and I. Charpentier. Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? a case study assimilating data in the northwestern Mediterranean sea at the JGOFS-DYFAMED station. *Journal of Marine Systems*, 40:99–125, 2003.

K. Fennel, M. Losch, J. Schröter, and M. Wenzel. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. *Journal of Marine Systems*, 28(1-2):45–63, 2001.

F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol pick freeze Monte Carlo method. *Statistics*, 50(4):881–902, 2016.

L. Gilquin, E. Arnaud, C. Prieur, and A. Janon. Making best use of permutations to compute sensitivity indices with replicated designs. working paper or preprint, June 2017. URL `https://hal.inria.fr/hal-01558915`.

W. F. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19:293–325, 1948.

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

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

K. Kishen. On latin and hyper-graeco-latin cubes and hyper-cubes. *Current Science*, 11(3):98–99, 1942.

I. Kriest, A. Oschlies, and S. Khatiwala. Sensitivity analysis of simple global marine biogeochemical models. *Global Biogeochemical Cycles*, 26(2), 2012.

G. Lacroix. *Simulation de l'écosystème pélagique de la mer Ligure à l'aide d'un modèle unidimensionnel. Etude du bilan de matière et de la variabilité saisonnière, interannuelle et spatiale.* PhD thesis, Université Pierre et Marie Curie, Paris, France, 1998.

G. Lacroix and M. Grégoire. Revisited ecosystem model (MODECOGeL) of the Ligurian sea: seasonal and interannual variability due to atmospheric forcing. *J. Marine Syst.*, 37(4):229–258, 2002.

G. Lacroix and P. Nival. Influence of meteorological variability on primary production dynamics in the Ligurian Sea (NW Mediterranean sea) with a 1D hydrodynamic/biological model. *J. Marine Syst.*, 16(1-2):23–50, 1998.

C. Lemieux. *Monte Carlo and quasi-Monte Carlo sampling*. Springer Science & Business Media, 2009.

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

H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J.W. Jones, editors, *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*, chapter 4, pages 55–99. Elsevier, 2006.

J. C. J. Nihoul and S. Djenidi. Perspectives in three-dimensional modelling of the marine system. In J.C.J. Nihoul and B.M. Jamart, editors, *Three-Dimensional Models of Marine and Estuarine Dynamics*, pages 1–34. Elsevier, Amsterdam, 1987.

M. Omlin, R. Brun, and P. Reichert. Biogeochemical model of lake Zürich: sensitivity, identifiability and uncertainty analysis. *Ecological Modelling*, 141(1-3):105–123, 2001.

A. B. Owen. Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, pages 439–452, 1992.

G. Pujol, B. Iooss, A. Janon, P. Lemaitre, L. Gilquin, L. Le Gratiet, T. Touati, B. Ramos, J. Fruth, and S. De Veiga. *Sensitivity: Global Sensitivity Analysis of Model Outputs*, 2017. R package version 1.15.0.

A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297, 2002.

A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

M. Schartau, P. Wallhead, J. Hemmings, U. Löptien, I. Kriest, S. Krishna, B.A. Ward, T. Slawig, and A. Oschlies. Reviews and syntheses: parameter identification in marine planktonic ecosystem modelling. *Biogeosciences (BG)*, 14(6):1647–1701, 2017.

V. Scott, H. Kettle, and C. Merchant. Sensitivity analysis of an ocean carbon cycle model in the north Atlantic: an investigation of parameters affecting the air-sea CO2 flux, primary production and export of detritus. *Ocean Science*, 7(3):405–419, 2011.

I. Sobol'. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414, 1993.

J.-Y. Tissot and C. Prieur. A randomized orthogonal array-based procedure for the estimation of first- and second-order Sobol' indices. *Journal of Statistical Computation and Simulation*, 85(7):1358–1381, 2015.

J.F. Tjiputra, D. Polzin, and A. Winguth. Assimilation of seasonal chlorophyll and nutrient data into an adjoint three-dimensional ocean carbon cycle model: sensitivity analysis and ecosystem parameter optimization. *Global biogeochemical cycles*, 21(1), 2007.

N. A. Weiss. *A course in probability*. Addison-Wesley, Boston, MA, 2006. URL `http://cds.cern.ch/record/735662`.