Cluster-based GSA: Global sensitivity analysis of models with temporal or spatial outputs using clustering

Sébastien Roux, Samuel Buis, François Lafolie, Matieyendou Lamboni

To cite this version:

Sébastien Roux, Samuel Buis, François Lafolie, Matieyendou Lamboni. Cluster-based GSA: Global sensitivity analysis of models with temporal or spatial outputs using clustering. Environmental Modelling and Software, 2021, 140, pp.105046. 10.1016/j.envsoft.2021.105046 . hal-03203049

HAL Id: hal-03203049
https://hal.inrae.fr/hal-03203049

Submitted on 10 Jun 2022

Cluster-based GSA: global sensitivity analysis of models with temporal or spatial outputs using clustering

Sébastien Roux [12a], Samuel Buis[1b], François Lafolie[b], Matieyendou Lamboni[c,d]

[a]*MISTEA, Univ Montpellier, INRAE, Montpellier SupAgro, Montpellier, France*
[b]*EMMAH, INRAE, Université d'Avignon et des Pays de Vaucluse, Avignon, France*
[c]*Department DFRST, University of Guyane, Cayenne, French Guiana, France*
[d]*228-UMR Espace-Dev, University of Guyane, University of Réunion, Univ Montpellier, IRD, France*

Abstract

A new method named cluster-based GSA is proposed to enhance the sensitivity analysis of models with temporal or spatial outputs. It is based on a tight coupling between Global Sensitivity Analysis (GSA) and clustering procedures. Clustering is introduced to characterize the different behaviors of the model outputs by grouping them into clusters. The cluster-based GSA produces variance-based indices that quantifies how the model inputs drive the model outputs toward a given cluster or how they influence variation along a direction defined by two clusters. Aggregated indices are proposed to summarize the overall influence of model inputs on changes of clusters . The method is applied on two models having temporal outputs: a toy example and a environmental model simulating the decomposition of soil organic matter (CANTIS). In both cases, the influence of the model inputs on the different behaviors of model outputs was efficiently reported by this approach.

Keywords: Regional sensitivity analysis, Multivariate sensitivity analysis, Target sensitivity analysis, Fuzzy clustering, Sobol' indices, Cluster-based GSA

[1] Joint first authors

[2] corresponding author email : sebastien.roux@inrae.fr

Software availability

R codes for the toy model and the cluster-based GSA method are available from a github repository located at https://github.com/sbuis/ClusterBased_GSA. The Cantis model and Vsoil platform are available at http://www6.inrae.fr/vsoil.

1. Introduction

Sensitivity Analysis (SA) is recognized as a central tool in the modelling process, from model development to applications, up to decision making [1, 2]. Environmental modelling applications often involve models with multivariate outputs: typical examples are variables simulated at different instants (temporal outputs) or at different spatial locations (spatial outputs) or both. Standard SA performed on a scalar output often focus on characterizing the influence of model inputs on the variance of the output distribution obtained when these model inputs vary. Due to the increase of dimension, many other characterizations of the output distributions can be of interest for the modeler when performing SA on a temporal or spatial output. We focus specifically in this work on questions relating to the influence of variation in model inputs on the change in shape of model outputs. For instance, when samples of the output distribution are smooth curves differing only in some of their geometric properties (gradients, horizontal or vertical shift ...), a question of interest for the model user can be: which model inputs preferentially lead to particular shapes of these curves (e.g. increasing ones, upper curves ...)?

To study the effect of model inputs on the shape of simulated curves, Campbell et al. [3] proposed to project model outputs onto polynomial or adaptive bases. The outputs are expanded in a new coordinate system defined by a set of basis functions and the SA is then performed on the coefficients of the expansion using any SA method. The geometrical analysis of the outputs variability along each basis functions along with the value of the associated sensitivity indices revealed in this case the impact of model inputs on up–down shift, left–right shift, symmetric kurtosis and tail-fattening components of the simulated curves. This approach was further extended by Lamboni et al. [4, 5] and Savall et al. [6] applied it to a spatio-temporal model describing nitrogen transfers, transformations and losses at the landscape scale. A generic implementation using different bases such as eigenvectors, orthogonal polynomials, b-spline or principal component analysis can be found in Bidot et al. [7]. Recently, Xiao et al. [8] used a wavelet basis to analyze the impact of model inputs on signal features in the time-frequency domain. This approach brings information about the influence of model inputs on some specific variation of shape of the simulated outputs. But the results obtained are constrained by the choice of the basis and may be difficult to interpret.

An alternative approach is to consider that the temporal or spatial output distribution can be decomposed into a small number of homogeneous groups that characterize the diversity of the output shapes. The set of smooth curves taken as an example above can for instance be split into subsets of similar curves which shape differs for some of their geometric properties (e.g. increasing vs decreasing curves or upper vs lower ones . . .). Partitioning the model output space was originally introduced in SA for scalar outputs in the so-called Regional Sensitivity Analysis (RSA) [9]. It consists in splitting the set of simulated outputs into two classes, often called ‘behavioral’ and ’non-behavioral’, and computing sensitivity indices based on a Kolmogorov-Smirnov test. RSA was recently extended to spatio-temporal outputs by applying a clustering algorithm on the simulated outputs and applied to reservoir modeling in [10, 11]. The interest of using a clustering algorithm in this context is to automatically group the simulated temporal or spatial outputs that have similar shapes into a small number of clusters supposed to be representative of the different shapes of interest, and then to identify the model inputs that lead to these different clusters. In reference to the original version of RSA, these groups are denoted as ’model behaviors’ in the following. However, one known drawback of RSA approaches is the difficulty to estimate sensitivity indices with the same precision and flexibility than the widely used Sobol’ indices, which can characterize in a robust way the influence of parameters and their interaction based on the variance decomposition of the considered output. In a recent paper, Raguet et al. [12], citing the work from Lemaitre [13], introduced the concept of Target SA to revisit RSA in the context of reliability engineering where a given model output is split into two groups: a critical domain (associated to rare events) and the rest. They proposed different sensitivity indices and noticeably mentioned the application of Sobol’ indices, i.e. of variance-based Global Sensitivity Analysis (GSA), to the binary function encoding the membership of the simulated output values to the critical domain. This approach seems very promising to bridge the gap between SA including cluster analysis and variance-based SA for temporal or spatial model outputs. However it raises the questions of how to integrate a detection step of behaviors in a GSA procedure and what indices can be derived from it, in particular when more than two behaviors are detected.

In this work we propose to extend the target sensitivity analysis approach to study the impact of model inputs on behaviors characterized by a cluster analysis on temporal or spatial outputs. Although a clear distinction between different model behaviors may not always be possible to achieve, we consider this framework as flexible enough to be of interest for a lot of models and applications. We thus suppose in the following that behaviors can be defined and will come back to this hypothesis in the discussion section. Our approach, named cluster-based GSA, is based on the idea of using the cluster membership functions to compute several variance-based indices allowing a thorough analysis of the influence of model inputs on the different identified model behaviors. The article is organized as follows. We present in Section 2 a motivating example introducing the basic idea of the proposed procedure and in Section 3, the coupling approach that makes use of a clustering of temporal or spatial outputs in order to define several cluster-based sensitivity indices. Section 4 is devoted to the validation of these indices on a dedicated toy model and Section 5 to its application on a realistic environmental model simulating the decomposition of soil organic matter.

2. Motivating example

The basic idea and motivation of the overall approach can be illustrated by looking at a sample of the output distribution obtained when varying some inputs of the CANTIS model that will be further described and analyzed in Section 5. CANTIS output of interest are curves representing the dynamics of a quantity of interest (the biomass of an organic matter pool when studying crop residue over time). By looking at the sample of curves in Fig. 1, it is clear that they have only several typical shapes defined by their monotony, abscissa and value of maximum, and end level. Such parsimony of possible shapes motivates the idea to identify the main groups of shapes that describes the output distribution and to study the influence of input factors on the occurrence of these shapes. For example, one can wonder which parameter drives the output toward monotonic increasing behavior (increase of biomass) or monotonic decreasing behavior (full consumption of initial organic matter stock). Based on the initial idea of target sensitivity analysis, a simple procedure to tackle this issue can be the following: i) Perform simulations on a numerical design-of-experiment associated to a selected SA method (e.g. the Sobol' method), ii) Detect or select groups in the simulated outputs, iii) Build scalar functions representing the membership of each curve to each group, iv) Perform a scalar sensitivity analysis on this new functions. The result would allow discussing the desired model properties. This is the general idea of the Cluster-based GSA that will be formalized in the following section. The result of its complete application on the CANTIS model will be presented in Section 5.

3. Method

In this section, we recall some useful results in the fields of sensitivity analysis and clustering before presenting the proposed cluster-based approach which integrates both aspects.

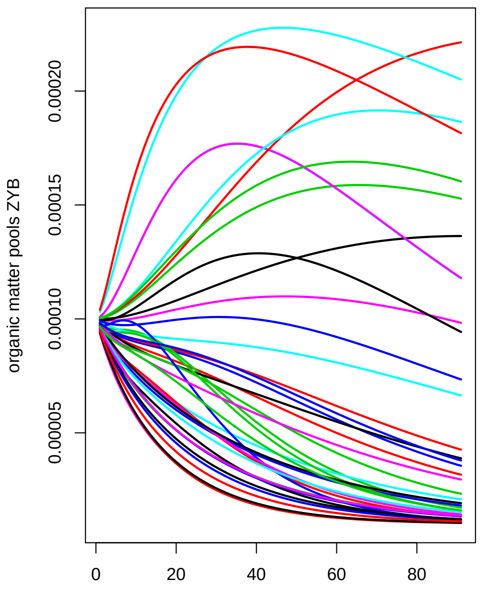

Figure 1: Sample of the output distribution of the CANTIS model

3.1. Variance-based sensitivity indices

Many numerical indices have been proposed to characterize the influence of model inputs on model outputs (see for example [14]). Among these different options, using Sobol' sensitivity indices [15, 16] is a widely accepted choice for global sensitivity analyses of scalar model outputs obtained from model inputs varying independently. Sobol' indices are based on the orthogonal decomposition of the output variance. The first order index (noted here SI_j^Y) corresponds to the part of variance of the scalar output Y due to the variation of an input X_j:

$$SI_j^Y = \frac{\mathbb{V}\left[\mathbb{E}[Y|X_j]\right]}{\mathbb{V}[Y]} \, ,$$

with $\mathbb{E}(Y|X_j)$ the expectation of Y given X_j and $\mathbb{V}(Y)$ the variance of Y. The total order index (noted here TSI_j^Y) corresponds to the part of the variance of Y due to X_j and to all interaction effects of X_j with the other varying inputs:

$$TSI_j^Y = 1 - \frac{\mathbb{V}\left[\mathbb{E}[Y|\boldsymbol{X}_{\sim j}]\right]}{\mathbb{V}[Y]},$$

with $\boldsymbol{X}_{\sim j}$ the vector of all model inputs except X_j.
In the case of a MV output $\boldsymbol{Y}$ made of several scalar components $(Y_k)_k$, Generalized Sensitivity Indices (GSIs) [4, 5, 17] allow the impact of model inputs on the whole MV output to be studied. The first-order GSI of X_j can be defined by:

$$GSI_j = \sum_{k=1}^{N} \frac{\mathbb{V}[Y_k]}{\sum_{m=1}^{N} \mathbb{V}[Y_m]} SI_j^{Y_k}$$

Likewise, the total GSI of X_j can be defined by:

$$GSI_{T_j} = \sum_{k=1}^{N} \frac{\mathbb{V}[Y_k]}{\sum_{m=1}^{N} \mathbb{V}[Y_m]} TSI_j^{Y_k}$$

They are easily computed from the $SI_j^{Y_k}$ and $TSI_j^{Y_k}$. Many software are now available for the estimation of Sobol' indices and GSI (see for example the recent review [18]).

3.2. Clustering

Clustering methods are powerful tools that can reveal hidden structures in multi-dimensional data set by grouping objects using similarity criteria [19]. A successfully applied clustering method would typically provide a small number of groups of objects where the similarity within groups is maximized while the similarity between groups is minimized. In this study we considered fuzzy clustering methods (see for example [20]). As opposed to classical clustering, fuzzy clustering does not produce a binary answer to the question of membership of an object to a given cluster but instead computes the so-called *membership functions* which evaluate the degree of membership of an object to any given cluster. A main advantage of fuzzy clustering over classical clustering is its natural handling of partial or uncertain affectation to clusters using low but non-zero membership functions. In the following, we consider membership functions normalized in $[0, 1]$. Such normalized membership functions correspond to empirical probabilities or weights in clustering approaches based on statistical modeling.

The use of a basic fuzzy clustering algorithm, the fuzzy c-means (or fcm) algorithm [21] proved sufficient to derive a satisfying model analysis in the numerical examples. We briefly recall the principle of this algorithm in the following. The fcm algorithm is the fuzzy extension of the classical K-means algorithm. Its principle is to perform an alternate optimisation procedure in order to minimize the weighted sum of squared distances to cluster centers. More precisely, let u_{ij} be the membership function of object $\boldsymbol{x_j}$ with respect to cluster i, d_{ij} be the distance (e.g. euclidean) between $\boldsymbol{x_j}$ and the i^{th} cluster center $\boldsymbol{c_i}$ and $m > 1$ a parameter controlling the level of 'fuzzyness' of the method. Then, for a given cluster number K, the fcm algorithm seeks for the membership functions $(u^*_{ij})_{ij}$ and centers $(\boldsymbol{c^*_j})_j$ minimizing $\left(\sum_{i=1}^K \sum_{j=1}^n u_{ij}^m d_{ij}^2\right)$ over all membership functions and centers. It proceeds by iteration, updating alternatively the membership functions from previous center estimates (see [21]) :

$$u_{ij} = \frac{d_{ij}^{\frac{-2}{m-1}}}{\sum_{i=1}^K d_{ij}^{\frac{-2}{m-1}}}$$

and the centers from previous membership functions estimates:

$$\boldsymbol{c_i} = \frac{\sum_{j=1}^n u_{ij}^m \boldsymbol{x_j}}{\sum_{j=1}^n u_{ij}^m}$$

3.3. The cluster-based GSA method

3.3.1. Integration of clustering into a sensitivity analysis procedure

A clustering procedure can be easily integrated in a SA workflow: we propose to apply it after the simulations have been performed on the numerical experimental design specific of the SA method used (See Fig. 2). In this way a vector of membership functions is made available on each element of the design and can be used to feed a classical sensitivity analysis approach. The whole set of simulations should be preferably used for clustering if the size and number of objects to be clustered as well as the complexity of the chosen fuzzy clustering algorithm allow it. However, one can also use a subset of model simulations to define the clusters before computing the membership functions for each element of the experimental design in order to reduce the problem size and computational cost of the clustering step.

3.3.2. Cluster-based sensitivity indices

Several sensitivity indices can be easily computed using the membership functions produced by the clustering procedure.

Indices of cluster membership. This first type of indices is simply obtained by applying a sensitivity analysis method on each scalar membership function in order to estimate their Sobol' indices. Depending on the method used,

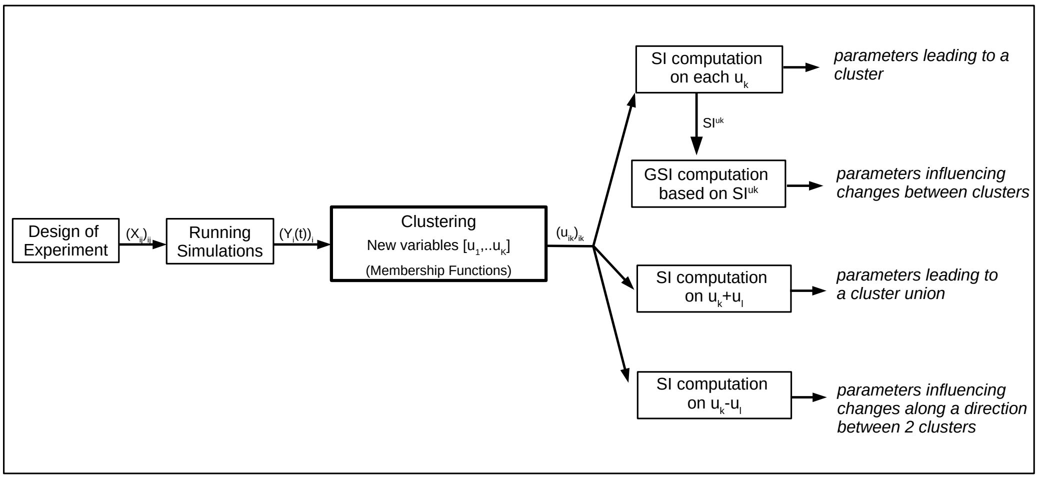
Design of Experiment
$(X_{ij})_{ij}$
Running Simulations
$(Y_i(t))_i$
Clustering
New variables $[u_1,..u_K]$
(Membership Functions)
$(u_{ik})_{ik}$
SI computation on each u_k
parameters leading to a cluster
SI^{uk}
GSI computation based on SI^{uk}
parameters influencing changes between clusters
SI computation on u_k+u_l
parameters leading to a cluster union
SI computation on u_k-u_l
parameters influencing changes along a direction between 2 clusters

Figure 2: Integration of a clustering step in sensitivity analysis workflow. The clustering is applied on the simulations and its outputs, the membership functions, feed scalar sensitivity analysis procedures to produce different cluster-based sensitivity indices.

first order, total or specific interaction indices can be computed. For a model input X_j, first and total order indices related to the k-th cluster are defined using the membership functions u_k as follows:

$$SI_j^{u_k} = \frac{\mathbb{V}\left[\mathbb{E}\left[u_k|X_j\right]\right]}{\mathbb{V}\left[u_k\right]}$$

$$TSI_j^{u_k} = 1 - \frac{\mathbb{V}\left[\mathbb{E}\left[u_k|X_{\sim j}\right]\right]}{\mathbb{V}\left[u_k\right]}$$

These indices, by definition, quantify the effect of model input parameters onto the variability of the membership function of a given cluster. They allow therefore the parameters driving the model outputs to a targeted behavior defined by a cluster to be pointed out.

Aggregated indices of changes between clusters (Cluster-based GSIs). These indices are based on the idea of computing aggregated indices on the vector of membership functions $\boldsymbol{u} = [u_1, .., u_K]$, seen as a new low-dimensional multivariate variable. The GSIs computation scheme presented in Section 3 gives the following expressions for the first and total indices for model input

X_j using indices $SI_j^{u_k}$ and $TSI_j^{u_k}$:

$$GSI_j^{clu} = \sum_{k=1}^{N} \frac{\mathbb{V}[u_k]}{\sum_{m=1}^{N} \mathbb{V}[u_m]} SI_j^{u_k}$$

$$GSI_{T_j}^{clu} = \sum_{k=1}^{N} \frac{\mathbb{V}[u_k]}{\sum_{m=1}^{N} \mathbb{V}[u_m]} TSI_j^{u_k}$$

These cluster-based GSIs summarize the overall influence of parameters on changes of behaviors defined by the different clusters. They are easily computed from indices computed on the membership functions. They aggregate the information contained in these indices while taking into account differences in their variance.

Indices of directions of change between clusters and of membership to an union of clusters. The membership functions can be combined to define other transformation functions that provide complementary information about the impact of model inputs on changes of model behaviors defined by the different clusters.

- Sum of membership functions u_i+u_j. The sensitivity indices associated to the sum of two or more membership functions allow discussing which parameters influence the membership of an union of clusters.

- Difference of membership functions $u_i - u_j$. The sensitivity indices associated to the difference between two membership functions allow discussing which parameters influence variations of the vector of membership functions projected along the direction defined by 2 clusters. It should be noted that i) changes of membership orthogonal to the direction defined by the two cluster centers do not contribute to the variance explained by this index, ii) changes from cluster i to cluster j in situations where all changes of membership from on cluster i is transferred to cluster j contribute, iii) changes from cluster i (or j) to another cluster also contribute, but to a lower extent, to the variance explained by the index. These indices thus mostly bring information about the impact of model inputs on transitions between two given clusters.

4. Application to a toy model

4.1. Model description

A toy model named TC (for ToyCurves) was introduced to perform a qualitative validation of the cluster-based sensitivity indices and illustrate their interpretation capabilities. Validation of the results is made by linking the geometrical effects of parameters deduced from the model definition (e.g. triggering the emergence of a pattern, modifying its shape, etc) with the result of the clustering.

The TC model has six parameters and produces on the $[0,1]$ interval a curve defined as a sum of a vertical offset plus two shifted triangles. As can be seen in Fig. 3, parameter X_1 drives the height of the first triangle while two parameters (X_1, X_2) drive the height of the second in an interacting way. Their abscissas are centered respectively at $t = 0.15$ and $t = 0.75$ with perturbations controlled by parameters (X_4, X_6) for Triangle 1 and (X_5, X_6) for Triangle 2. Parameter X_3 controls the height of the global shift. The mathematical definition of the model is the following: Let $\mathit{Trg}(c,h)(.)$ denote a triangle function producing over [0,1] a triangle of height h and width 0.3 centered at $t = c$. A possible expression of this function is $\mathit{Trg}(c,h)(t) = \frac{h}{0.15} \cdot \max\left(0, (0.15 - |t-c|)\right)$. Then the TC model is defined as follows for an input vector $\boldsymbol{X} = [X_1, .., X_6]$:

$$\begin{aligned} \mathit{TC}(\boldsymbol{X})(t) = {} & \frac{X_3}{5} + \mathit{Trg}(0.25 + 2.(X_4 - 0.5).X_6, X_1)(t) \\ & + \mathbb{1}_{[0.5,1]}(X_1).\mathbb{1}_{[0.5,1]}(X_2).\mathit{Trg}(0.75 + 2.(X_5 - 0.5).X_6, 2.(X_2 - 0.5)(t) \end{aligned} \tag{1}$$

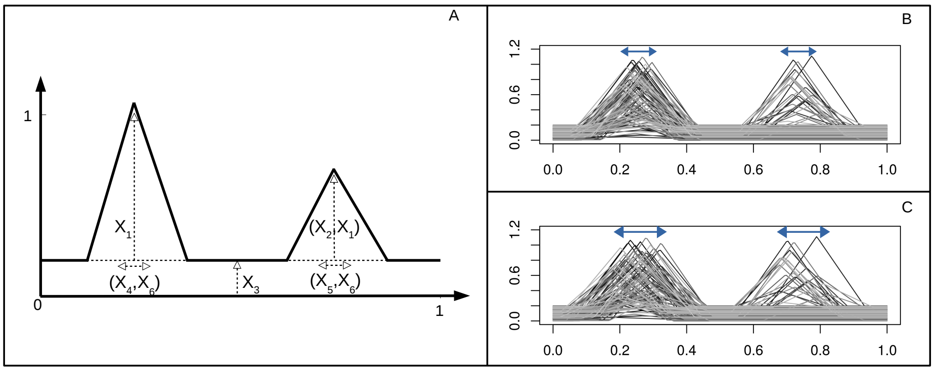

Figure 3: The toy model (A) and some samples for two parameter settings; Setting 1: small t-shift (B), Setting 2: large t-shift (C).

4.2. Numerical configuration

We considered two levels of horizontal shift driven by parameter X_6. Setting 1, referred to as 'small t-shift', is obtained by fixing $X_6 = 0.05$ and corresponds to small horizontal perturbations of the triangle centers. Setting 2, referred to as 'large t-shift' is obtained with $X_6 = 0.08$ and produces larger perturbations. The TC model is analysed for these two settings when the other parameters $(X_1, .., X_5)$ have independent uniform distributions within $[0, 1]$. Samples of the resulting curves are presented in Fig.3. A Sobol' algorithm ("SobolJansen" from the Sensitivity package [22]) was used with $N = 7000$ to define an experimental design and to compute the various sensitivity indices based on the clustering outputs. 95% confidence intervals on the sensitivity indices have been estimated using bootstrap replicates.

4.3. Clustering

A fuzzy clustering algorithm was applied for each setting on the curves produced by the TC model on the experimental design defined by the Sobol' algorithm. We applied a fuzzy c-means algorithm (see Section 3.2) targeting three classes and using an euclidean distance between curves. This choice led to a robust convergence with respect to the starting point and to clear and interpretable clusters. The results are presented in Fig.4 and Table 1. The clusters obtained on Setting 1 are characterized by differences in terms of number (1 or 2) and amplitude of triangles, while the clustering obtained for Setting 2 stressed the importance of the first triangle center location for defining groups in the set of output curves: Cluster 1 and Cluster 3 only differs by the horizontal position of the first triangle center (see Fig.4).

These results highlight that the TC model has different dominant behaviors depending on the parameter setting. The influence of the model inputs on these behaviors will now be studied using the cluster-based sensitivity indices.

	Setting 1: small t-shift	Setting 2: large t-shift
Cluster 1	single small and early maximum	one early-right maximum and small late maximum
Cluster 2	two maxima	single small and early maximum
Cluster 3	single large and early maximum	one early-left maximum and small late maximum

Table 1: Description of the clusters obtained for the two settings of the toy model

4.4. Results of the sensitivity analysis

The objective of this section is to illustrate how the different indices defined in Section 3.3.2 accurately report interpretable and known effects of the toy model parameters.

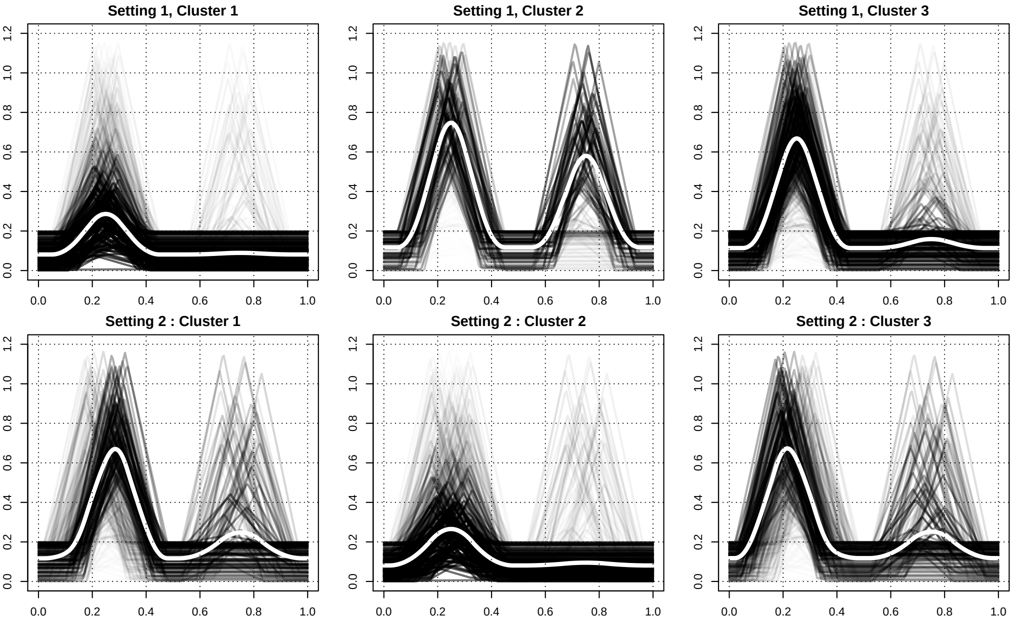

Figure 4: Clustering obtained on two settings of the TC model (first line for Setting 1, second line on Setting 2). Cluster centers are drawn as bold white curves. Simulated curves are represented with a grey level depending on their membership level to a given cluster (black = high membership, light grey = low membership). Qualitative description of the clusters can be found in Table 1.

4.4.1. Indices of cluster membership

Sensitivity indices computed for each cluster membership function for the two settings are presented in Fig.5 and discussed more precisely in the following:

- Setting 1: The clusters obtained for this setting differ by the number and amplitude of maxima, not by their horizontal shift. As X_4 and X_5 precisely drive the horizontal shift of the two triangles, they shall not exhibit strong influences on the membership functions, which is effectively verified ($\max_k TSI_4^{u_k} = 0.034, \max_k TSI_5^{u_k} = 0.004$). As the amplitude of the first triangle is only defined by X_1 and as the appearance of the second triangle occurs only for large values of X_1 and X_2 (and corresponds to high amplitude of the first triangle), X_1 is expected to be the main parameter influencing the membership function variability of cluster 1, whereas X_1 and X_2 should explain the membership to cluster 2 and 3 with a strong interaction. As can be seen in Fig. 5, these properties are verified by the computed indices.

- Setting 2: Unlike Setting 1, Setting 2 produced clusters differing by the shift of the first triangle. Parameter X_2, which drives the height of the second triangle when X_1 is high enough, shall not influence the membership to any of the three clusters obtained in this setting. This is verified in Fig.5: $\max_k TSI_2^{u_k} = 0.055$. The same fact can be observed and verified for parameter X_5 which only appears in the model definition in the computation of the shift of the second triangle and thus exhibits small sensitivity indices (remember that the second cluster is not a key differentiating feature between clusters). On the contrary, parameter X_4 should drive the membership to cluster 1 and 3, as they differ each other (and with cluster 2) by the shift and amplitude of the first triangle. This property is verified in the results shown Fig.5: $SI_4^{u_1} = 0.262, SI_4^{u_3} = 0.255$.

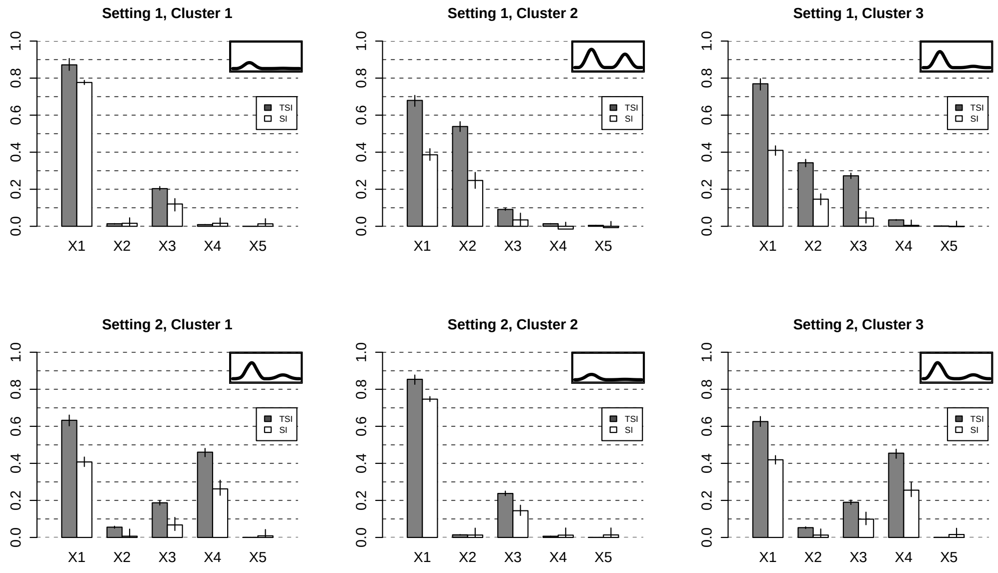

Figure 5: Sensitivity indices (1st order and total) on membership functions obtained on the toy model: first line on Setting 1 ('small t-shift') , second line on Setting 2 ('large t-shift'). Cluster centers are recalled on top of each graph.

4.4.2. Indices of direction of change between clusters

Sensitivity indices obtained on membership function differences are presented in Fig.6 for each of the three directions of change defined by two clusters $((1,2),(2,3),(3,1))$ and discussed below:

- Setting 1: A interesting property is the importance of X_2 for explaining the variability of the model outputs along direction $(2,3)$. Indeed, these two clusters differ from the amplitude of the second triangle in the presence of a high early triangle. From the model definition, such a transition is obtained when X_2 varies while X_1 is kept large. This fact is retrieved in Fig. 6: X_2 is the parameter with the highest sensitivity on direction $(2,3)$, with a high level of interaction with parameter X_1 ($\mathit{SI}_2^{u_{23}} = 0.357, \mathit{TSI}_2^{u_{23}} = 0.826$). The indices obtained for the two other directions show, as expected, the high importance of X_1 since it controls the height of the first triangle.

- Setting 2: The indices computed on membership functions have highlighted the importance of parameter X_4. Its effect is completely revealed in the analysis of indices associated to direction $(3,1)$ corresponding to the two clusters differing only by a small shift of the early triangle center: $\mathit{SI}_4^{u_{13}} = 0.554, \mathit{TSI}_4^{u_{13}} = 1.00$. This corresponds precisely to the geometrical effect of X_4 in the model definition. However, the transition between cluster 1 and 3 only occurs when X_1 takes high values and, thus, there is also a quite high interaction effect between X_1 and X_4.

4.4.3. Aggregated indices on changes between clusters

We computed the cluster-based GSIs for the two settings with a focus on how they stress out different parameter influence as compared to classical GSIs computed on the dynamic outputs. The results are presented in Fig.7. A first remark is that classical GSIs and cluster-based GSIs are different, even if in all cases they rank X_1 as the most influential parameter. Such result is not suprising, as it is an application of GSI on two different functions. A nice property of these results lies in the indices of the other parameters: while for both settings the classical GSIs lead to conclude that (X_1, X_2, X_4) play a role to explain the global variability of the output curves (with a slightly higher effect of X_4 for Setting 2), cluster-based GSIs clearly highlight their differential effect in the two settings. Indeed, in addition to X_1, only (X_2, X_3) have non-negligible cluster-based GSIs for Setting 1 and (X_2, X_4) of Setting 2. Cluster-based GSIs are thus able to report the effect of X_2 to explain the dominant amplitude-based behaviors in Setting 1 and the effect of X_4 to explain the dominant shift-based behaviors in Setting 2.

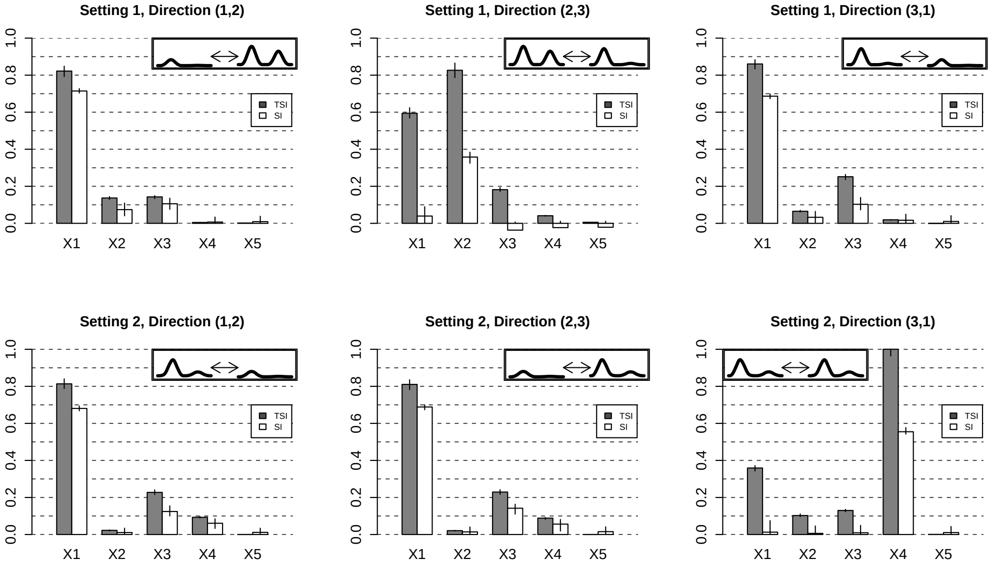

Figure 6: Sensitivity indices (1st order and total) on membership function differences obtained on the toy model: first line on Setting 1 ('small t-shift') , second line on Setting 2 ('large t-shift'). Cluster directions are recalled on top of each graph.

5. Application to a realistic environmental model

5.1. Model description

The CANTIS model simulates Carbon And Nitrogen Transformations in Soils [23]. Soil organic matter is an important natural resource sensitive to direct and indirect human impacts. Simulation models play an important role to integrate and examine the understanding of its dynamics, to evaluate human impacts on ecosystem function, and to manage soil organic matter for greenhouse gas mitigation, improved soil health and sustainable use as a natural resource [24]. CANTIS is made of a set of first-order ordinary differential equations modeling the dynamics of several soil organic matter pools interacting during their evolution. These pools correspond to: soil humified organic matter, crop residues and two microbial pools growing on the humified compartment and the crop residues, respectively. Crop residues are split in four pools to account for the large variety of biochemical composition of the residues. The model includes more than 20 parameters. While some of these are well known, a dozen of them are affected by significant uncertainties. The impact of these uncertainties on the dynamics of simulated outputs

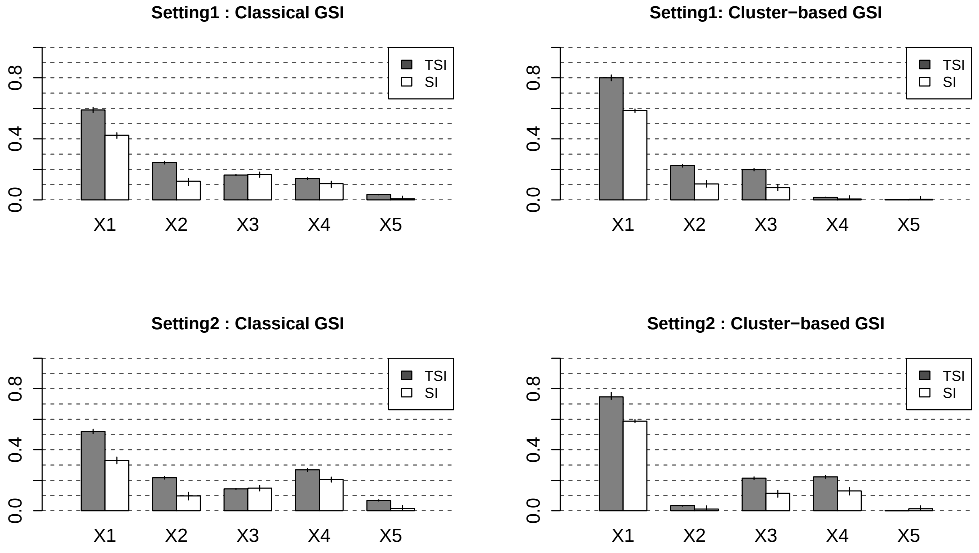

Figure 7: Comparison of classical GSIs with cluster-based GSIs (GSIs applied on the vector of membership functions) on the two settings of the toy model.

is far to be well known due to the strong interactions between the simulated processes and to the high variability of their dynamics. This motivated the application of cluster-based sensitivity analysis on this model.

5.2. Numerical configuration

For this study, a standard batch configuration has been used with initial contents defined as (i) non-limiting for nitrogen, (ii) corresponding to a residue incorporation for organic matter, and (iii) sufficient for ammonium to generate a nitrification at the beginning of the simulation. Simulation duration has been set to 3 months, allowing the different processes of the model to express in our configuration.

Ten parameters were selected to vary in the sensitivity analysis experiments. Their name, definition and uncertainty distributions are detailed in Appendix A.

The simulated variable studied is the zymogenous microbial biomass, designed by ZYB in the following. This is the microbial biomass growing on the crop residues. The size of this pool varies in time depending on the availability of the residues and on environmental conditions. The dynamics of this microbial biomass is also closely related to the quality of the residues and also linked with the humification process which is a transient state towards storage of carbon in soils. In addition, the activity of this biomass largely controls the carbon dioxide emissions and the soil mineral nitrogen concentrations, two variables measured during organic matter recycling experiments. It appears interesting to identify the parameters that have the largest influence on its dynamics.

The SobolSalt function of the R package sensitivity [22] has been used to generate the numerical design and compute both first order, second order and total Sobol' sensitivity indices. It implements the asymptotically efficient formulas given in section 4.2 of [25]. The sample size has been set to $n = 2500$.

Simulations have been realized using the INRAE Virtual Soil Platform. The platform provides an easy way to use and couple numerical modules representing processes occurring in soils. Detail information about the platform and how to use it and contribute can be found in : http://www6.inrae.fr/vsoil.

5.3. Clustering

Results of simulations conducted on the experimental design showed a large diversity of dynamics for ZYB. Applying the clustering on time increments curves with 4 clusters lead to a fairly clear distinction of interpretable behaviors as shown in Fig.8:

- Cluster 1 gathers concave ZYB dynamics increasing continuously during almost all 3 months or starting to decrease in second half of the simulation period,
- Cluster 2 gathers convex ZYB dynamics decreasing continuously from the beginning of the simulation,
- Cluster 3 gathers non-monotonic concave ZYB dynamics with a maximum reached in the first half of the simulated period,
- Cluster 4 gathers mostly continuously decreasing ZYB dynamics, convex or concave, with a large variation of slopes and of values at the end of the simulated period with respect to curves in cluster 2.

The four clusters illustrate that very different behaviors can be simulated. Then, the question that emerges is that of the identification of the parameters and/or of their interactions that drive the distinction between these behaviors.

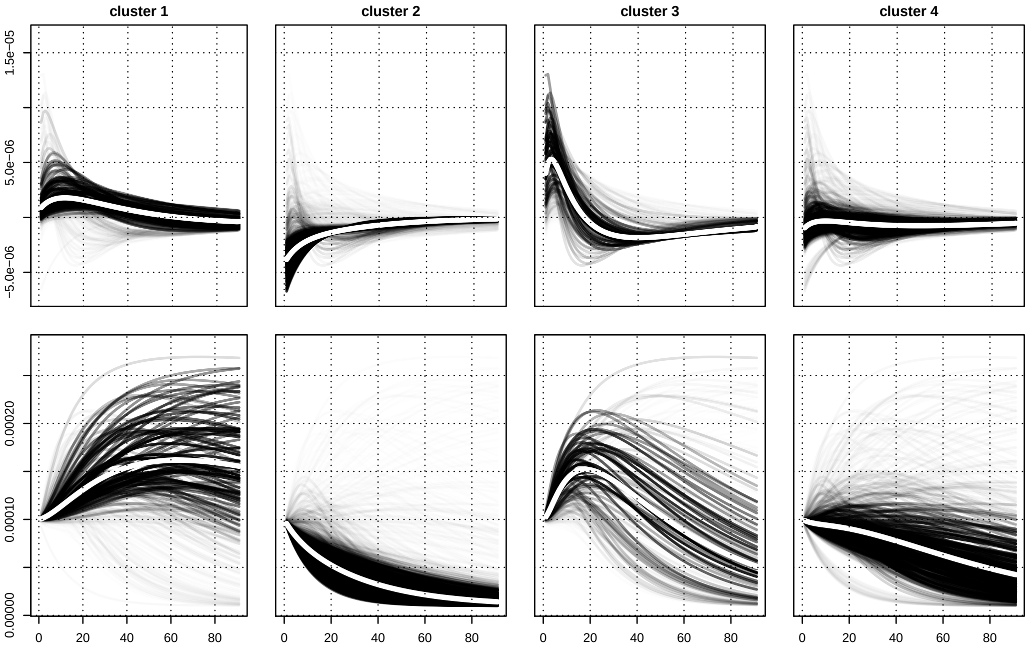

Figure 8: Clustering obtained on Zymogenous Biomass time increments simulated by the Cantis model (on first line the time increments and on second line the total biomass. Cluster centers are drawn as bold white curves. Simulated curves are represented with a grey level representing their membership level to a given cluster (black = high membership, light grey = low membership). X axes represent the time in days.

5.4. Sensitivity Analysis

Results of sensitivity analyses using the cluster-based approach, presented in Fig.9 and 10, showed that only 3 parameters out of the 10 varying in this study explain the variations of *ZYB* dynamics: h_z, the humification coefficient of dead *ZYB*, k_{mz}, the Michaëlis-Menten constant for SOL decomposition and k_z, the *ZYB* mortality rate. While the sensitivity to these parameters was expected, the analyses, however, revealed that the biomass dynamic is relatively insensitive to the parameters k_1, k_2, k_3 and k_4 controlling the decomposition of the four pools composing the crop residues. This is probably due to the quite low level of uncertainty associated to these parameters in this study.

Cluster-based GSIs exhibit a large impact of interactions between these parameters which is not the case of GSIs computed on *ZYB* (Fig.9). Cluster-based GSIs reveal thus much more complexity in the relationships between these parameters and *ZYB* dynamics than GSIs computed on *ZYB*.

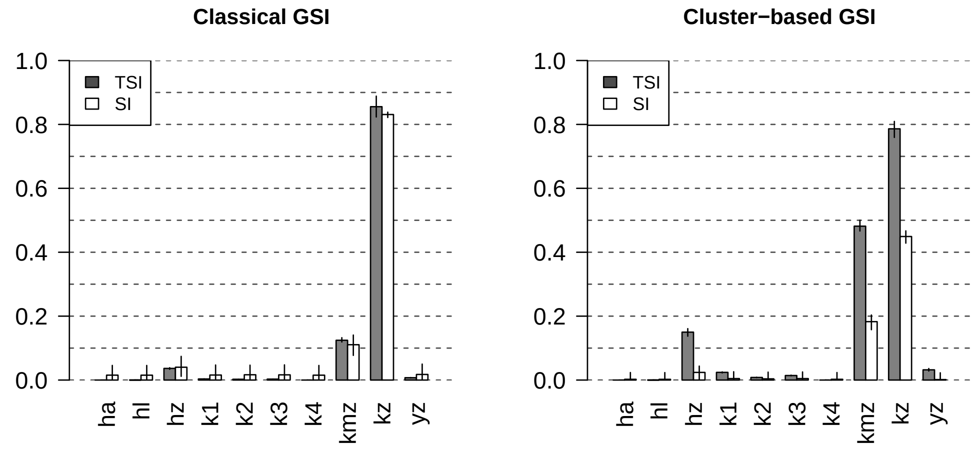

Figure 9: Classical GSI indices obtained on Zymogenous biomass times series (on the left) and on vectors of membership functions (Cluster-Based GSIs, on the right).

Fig.10 shows that k_z explains about 80% of the variance of cluster 1 membership function by itself. The value of this parameter is thus of extreme importance for maintaining the dynamics of *ZYB* at high levels. Not surprisingly, high values of the membership function are obtained for low values of k_z, that slow down the mortality of *ZYB*, as shown in Fig.1 of supplementary materials.

The variance of cluster 2 membership function is mostly explained by k_z and k_{mz} (first orders and interaction). The values of these parameters may thus lead to minimal values of *ZYB* all along the simulations. Such behaviors are obtained when both k_z and k_{mz} take medium to high values (see Fig. 1 and 2 of Supplementary materials). A high value of k_z implies a high mortality while at the same time high values of k_{mz} implies a slow growth of the biomass. A combination of such parameter values thus impedes the growth of the biomass and leads to a decrease from the initial value. Such a situation probably corresponds to a low decomposition of the organic matter pools RDM, HCEL, CEL and LIG. This result exhibits that in this configuration of parameter values, the mortality is such that it compensates the growth due to residue decomposition.

k_{mz} explains about 70% of the variance of cluster 3 membership function by itself. This parameter thus clearly controls rapid dynamics of *ZYB* with a high increase at the beginning of the simulation period and an early peak

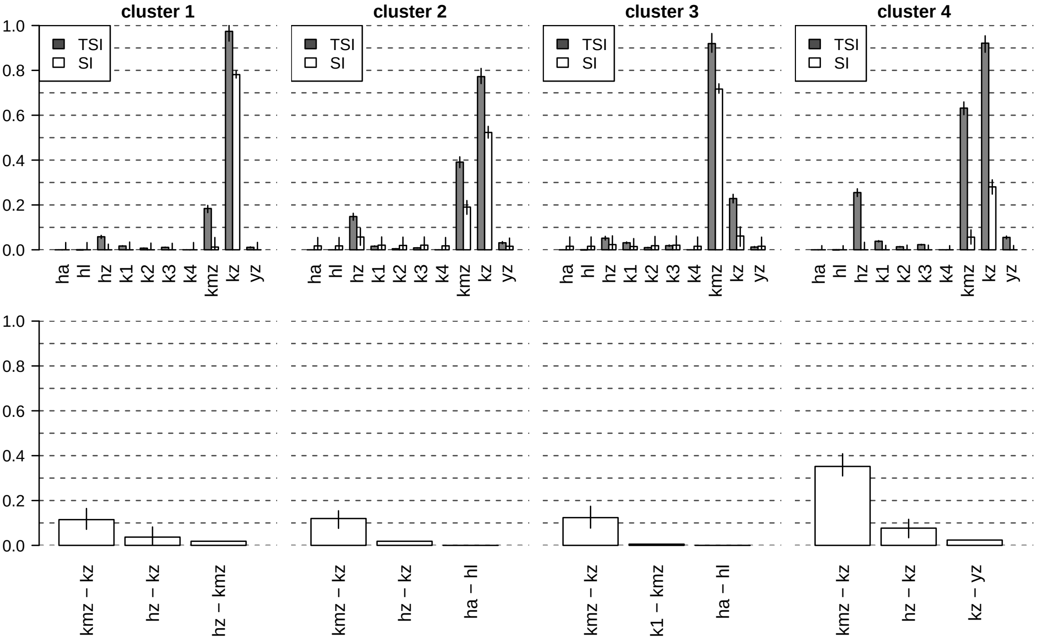

Figure 10: Sensitivity indices of membership functions obtained for Zymogenous biomass clustering. On first line the main and total indices, and on second line the three largest 2nd-order interaction indices.

followed by a rapid decrease. Fig.1 of supplementary materials shows that these curves are obtained for low values of k_{mz}. This means that in this configuration a rapid colonization of the residues by the biomass occurs which induce a lower availability of the substrate and thus a quite early recycling of the biomass.

The probability of ZYB dynamics to belong to cluster 4, i.e. to decrease continuously but not too fast, is mostly dependent on k_z and on its second-order interactions with k_{mz} and h_z. Fig. 1 and 2 of supplementary materials show that very low values of k_z do not allow to conduct to such behavior while medium to high values may, preferentially if associated with high to low values of h_z and k_{mz}. This shows that relatively low values of k_{mz}, which allows a good growth at the beginning, are balanced by the death. A low value of h_z contributes to a recycling of the dead biomass by itself and hence provides nutrients and in consequence prevents from a fast decrease due to exhaustion of substrate. This combination of parameters values thus leads to a death rate of the biomass slightly larger than the growth which produces a relatively slow decrease of ZYB biomass.

5.5. Conclusion

Clustering of *ZYB* dynamics revealed a large diversity of simulated behaviors, some of them rather rarely observed in natural systems. The cluster-based GSA allowed identifying the parameters that govern these behaviors. It has also indicated which density plots to look at in order to fully understand the processes involved and their interactions. The analysis particularly highlighted the importance of the competition between growing and recycling of dead biomass on *ZYB* dynamics.

6. Discussion

6.1. Genericity of cluster-based GSA

The cluster-based GSA was illustrated on temporal outputs and using a simple fuzzy clustering algorithm along with a Monte Carlo scheme to estimate Sobol' indices. However, the fundamental principle of the method is to compute indices based on the cluster membership functions. As a first consequence, the scope of application that can be dealt with cluster based GSA is wide and concerns any output for which considering homogeneous groups may be of interest for a modeler. It may be applied on a single temporal or spatial output but also on a single scalar output or on multiple outputs partitioned simultaneously.

A second consequence concerns the choice of the clustering algorithm. A simple fuzzy clustering algorithm (the fuzzy c-means) used with euclidean distances proved sufficient for the two case studies. However, the cluster-based GSA is in fact generic with respect to the fuzzy algorithm used. The only requirement is to compute membership functions or membership probabilities. Thus, more advanced clustering methods and/or distances might be considered in more challenging situations. It should also be noted that the method can directly be transposed to crisp clustering algorithms in which case membership functions take their values in $\{0, 1\}$. This includes the case of expert-based manual clustering where the user identifies and classifies himself the different behaviors to analyze.

Finally, the same remark holds for the computational scheme used to compute Sobol' indices. We used a standard Monte Carlo approach to estimate the Sobol' indices, but some more advanced methods, such as the ones based on metamodels, may be required for models with high computational cost. Concerning the cluster-based GSI, we proposed here to compute them from the vector of membership functions with the usual approach based on a decomposition of the trace of the model output covariance matrix [4, 17]. Different new aggregated indices have recently been introduced [26, 27] to explicitly take into account linear dependencies between the considered outputs. Cluster-based GSA is generic enough to allow computing these indices. Computing several types of GSI might bring complementary information on the global impact of model inputs on the vector of membership functions, but is beyond the scope of this paper.

6.2. Limits of cluster-based GSA

As indicated in introduction, the ability of the cluster-based indices to provide an analysis useful to model users is dependent on the model properties on the considered experimental design. If model outputs have no structure, then clustering algorithms will not find clearly separated clusters and there will be no additional information brought by cluster-based indices. It may be the case for models whose outputs of interest are strongly linked to highly variable forcing input data considered on a large domain (e.g. weather data, soil characteristics ...). In such case, the preferred option for a sensitivity study remains the 'point-based' approach that consists in computing sensitivity indices at each time step and/or spatial location ([28, 29, 30]). Note however that clearer patterns may appear in highly varying outputs by selecting temporal and/or spatial sub-domains of interest. Reordering spatial locations or time w.r.t. well chosen indicators may also be investigated for an easier interpretation, and thus clustering, of such kind of outputs (see e.g. [31]).

6.3. Within cluster analysis

As mentioned earlier in the paper, this work is related to the concept of target sensitivity analysis [12] which involves binary partitions of the output space corresponding to a critical domain. As these authors mentioned, such partitions raised two types of question on the parameter influence: how model inputs drive the model outputs to the critical domain (target SA) but also how they influence the outputs variability inside the restricted domain (what they called conditional SA). The same complementary analysis could be considered in the case of the cluster-based GSA, namely how parameters influence the variation of model outputs within a cluster. Such questions however raised the same difficulties: parameter distributions when considered conditionally to the membership of a cluster become strongly dependent and require different estimation methods [12].

7. Conclusion

In this work, we showed how to integrate a cluster analysis inside a global sensitivity analysis workflow in order to discuss in detail the influence of model inputs on the shape of model temporal or spatial outputs. The cluster analysis is used to partition the simulated outputs into homogeneous clusters that characterize the diversity of the output shapes, i.e. the different model behaviors. Several dedicated Sobol' indices built from the cluster membership functions have been proposed to quantify (i) how the inputs drive the outputs to a given cluster, (ii) how they influence the transitions between two given clusters and (iii) how they influence overall changes between clusters. The insights gained by this approach were validated on a toy example with respect to expected model properties. The cluster-based GSA approach was then applied to the CANTIS model. The cluster analysis successfully summarized the main dynamics of the simulated output curves. The cluster-based indices revealed i) the two main factors influencing these behaviors and ii) which density plots to look at in order to fully understand the processes involved and their interactions. Based on these results, we think that the cluster-based GSA is a promising method to improve the understanding of spatio-temporal models that exhibit different shapes of simulated outputs.

8. Acknowledgments

This work was part of the EA-ASM project funded by the French National Institute for Agricultural Research (INRAE). We thank Romain Narcy and Hervé Monod for fruitful discussions on multivariate sensitivity analysis, and Patrice Loisel for discussions on clustering algorithms. We also thank the team and the head of "la relation internationale de l'Université de Guyane" for their financial support during SAMO 2019 conference. Nathalie Moitrier, Cédric Nouguier and Nicolas Moitrier are thanked for their help in using the Vsoil platform.

References

[1] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, T. Wagener, Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software 79 (2016) 214–232.

[2] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. H. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, H. R. Maier, The future of sensitivity analysis: An essential discipline for systems modeling and pol-

icy support, Environmental Modelling & Software 137 (2021) 104954. doi:https://doi.org/10.1016/j.envsoft.2020.104954.

[3] K. Campbell, M. D. McKay, B. J. Williams, Sensitivity analysis when model outputs are functions, Reliability Engineering & System Safety 91 (10-11) (2006) 1468–1472.

[4] M. Lamboni, D. Makowski, S. Lehuger, B. Gabrielle, H. Monod, Multivariate global sensitivity analysis for dynamic crop models, Field Crops Research 113 (3) (2009) 312–320.

[5] M. Lamboni, H. Monod, D. Makowski, Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models, Reliability Engineering & System Safety 96 (4) (2011) 450–459.

[6] J. F. Savall, D. Franqueville, P. Barbillon, C. Benhamou, P. Durand, M.-L. Taupin, H. Monod, J.-L. Drouet, Sensitivity analysis of spatio-temporal models describing nitrogen transfers, transformations and losses at the landscape scale, Environmental Modelling & Software 111 (2019) 356–367.

[7] C. Bidot, M. Lamboni, H. Monod, multisensi: Multivariate Sensitivity Analysis, r package version 2.1-1 (2018).
URL https://CRAN.R-project.org/package=multisensi

[8] S. Xiao, Z. Lu, P. Wang, Multivariate global sensitivity analysis for dynamic models based on wavelet analysis, Reliability Engineering & System Safety 170 (2018) 20–30.

[9] R. Spear, G. Hornberger, Eutrophication in peel inlet—ii. identification of critical uncertainties via generalized sensitivity analysis, Water Research 14 (1) (1980) 43–49.

[10] D. Fenwick, C. Scheidt, J. Caers, Quantifying asymmetric parameter interactions in sensitivity analysis: application to reservoir modeling, Mathematical Geosciences 46 (4) (2014) 493–511.

[11] J. Park, G. Yang, A. Satija, C. Scheidt, J. Caers, Dgsa: A matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments, Computers & geosciences 97 (2016) 15–29.

[12] H. Raguet, A. Marrel, Target and conditional sensitivity analysis with emphasis on dependence measures, arXiv preprint arXiv:1801.10047 (2018).

[13] P. Lemaître, Analyse de sensibilité en fiabilité des structures, Ph.D. thesis, Université Bordeaux (2014).

[14] P. Wei, Z. Lu, J. Song, Variable importance analysis: a comprehensive review, Reliability Engineering & System Safety 142 (2015) 399–432.

[15] I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, Mathematical Modelling and Computational Experiments 1 (4) (1993) 407–414.

[16] A. Saltelli, K. Chan, E. M. Scott, et al., Sensitivity analysis, Wiley New York, 2000.

[17] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, Sensitivity analysis for multidimensional and functional outputs, Electron. J. Statist. 8 (1) (2014) 575–603.

[18] D. Douglas-Smith, T. Iwanaga, B. F. Croke, A. J. Jakeman, Certain trends in uncertainty and sensitivity analysis: An overview of software tools and techniques, Environmental Modelling & Software 124 (2020) 104588.

[19] R. Xu, D. C. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks 16 (3) (2005) 645–678.

[20] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy cluster analysis: methods for classification, data analysis and image recognition, John Wiley & Sons, 1999.

[21] J. C. Bezdek, R. Ehrlich, W. Full, Fcm: The fuzzy c-means clustering algorithm, Computers & Geosciences 10 (2-3) (1984) 191–203.

[22] G. Pujol, B. Iooss, A. Janon, Sensitivity: Global sensitivity analysis of model outputs, R package https://cran.r-project.or/package=sensitivity (2017).

[23] P. Garnier, C. Néel, B. Mary, F. Lafolie, Evaluation of a nitrogen transport and transformation model in a bare soil, European Journal of Soil Science 52 (2) (2001) 253–268.

[24] E. E. Campbell, K. Paustian, Current developments in soil organic matter modeling and the expansion of model applications: a review, Environmental Research Letters 10 (12) (2015) 123004.

[25] H. Monod, C. Naud, D. Makowski, Uncertainty and sensitivity analysis for crop models, in: Elsevier (Ed.), Working with dynamic crop models: Evaluation, analysis, parameterization, and applications, 2006, Ch. 4, pp. 55–100.

[26] L. Xu, Z. Lu, S. Xiao, Generalized sensitivity indices based on vector projection for multivariate output, Applied Mathematical Modelling 66 (2019) 592–610.

[27] M. Lamboni, Multivariate sensitivity analysis: Minimum variance unbiased estimators of the first-order and total-effect covariance matrices, Reliability Engineering & System Safety 187 (2019) 67–92.

[28] H. Varella, M. Guérif, S. Buis, Global sensitivity analysis measures the quality of parameter estimation: The case of soil parameters and a crop model, Environmental Modelling & Software 25 (3) (2010) 310–319.

[29] A. Mesgouez, S. Buis, G. Lefeuve-Mesgouez, G. Micolau, Use of global sensitivity analysis to assess the soil poroelastic parameter influence, Wave Motion 72 (2017) 377–394.

[30] C. Massmann, T. Wagener, H. Holzmann, A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales, Environmental modelling & software 51 (2014) 190–194.

[31] J. Herman, P. Reed, T. Wagener, Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior, Water Resources Research 49 (3) (2013) 1400–1414.

Appendix A. List of Cantis model parameters selected for SA

Name	Definition	Unit	Distribution
k_1	RDM decomposition rate	day^{-1}	$U[0.08, 0.24]$
k_2	HCE decomposition rate	day^{-1}	$U[0.032, 0.096]$
k_3	CEL decomposition rate	day^{-1}	$U[0.0485, 0.1455]$
k_4	LIG decomposition rate	day^{-1}	$U[0.0007, 0.0021]$
k_{mz}	Michaëlis - Menten constant for SOL decomposition	-	$U[1, 1000]$
k_z	ZYB decomposition rate	day^{-1}	$U[0.001, 0.1]$
h_a	Humification coefficient for AUB	-	$U[0, 1]$
h_l	Humification coefficient of LIG	-	$U[0, 1]$
h_z	Humification coefficient for ZYB	-	$U[0, 1]$
y_z	C assimilation yield by ZYB	-	$U[0.3, 0.6]$

where RDM is rapidly decomposable material, HCE is hemicelluloses, CEL is cellulose, LIG is lignin, SOL is soluble organic matter, ZYB is zymogenous biomass, AUB is autochtonous biomass and C is carbon.