INRAE

Cluster-Based GSA: a clustering-based approach to global sensitivity analysis of models with dynamic and/or spatial outputs

Samuel Buis [1], Sébastien Roux[2], Matieyendou Lamboni[3], François Lafolie[1], Françoise Ruget[1]

[1] INRAE, UMR EMMAH, Avignon, France, samuel.buis@inra.fr

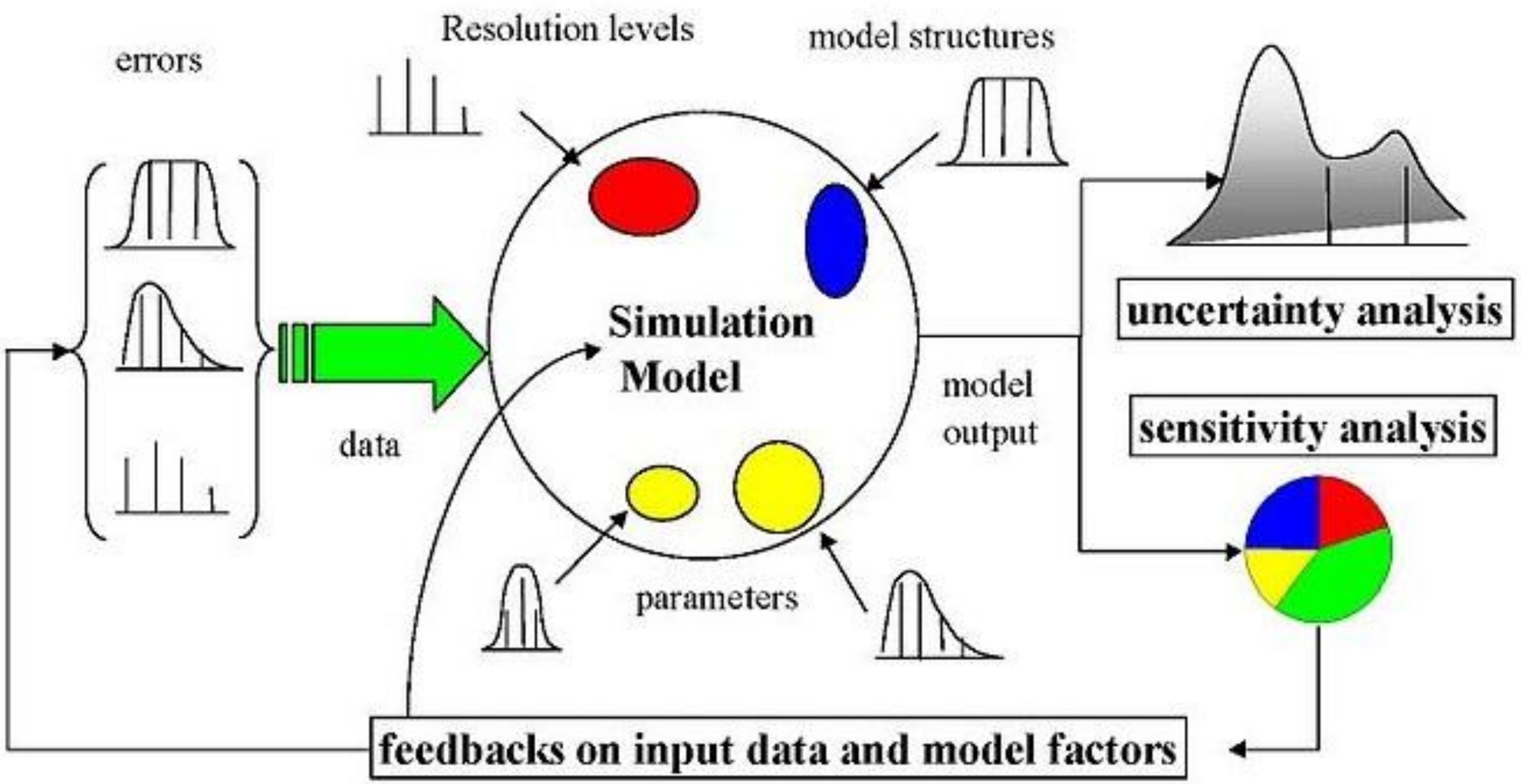
[2] INRAE, UMR MISTEA, Montpellier, France

[3] University of Guyane, 228-UMR-Espace dev, Cayenne, France

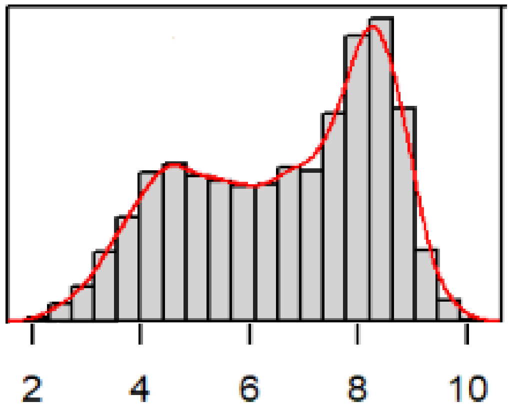
VSoil meeting, March 26-27th, 2026, Avignon

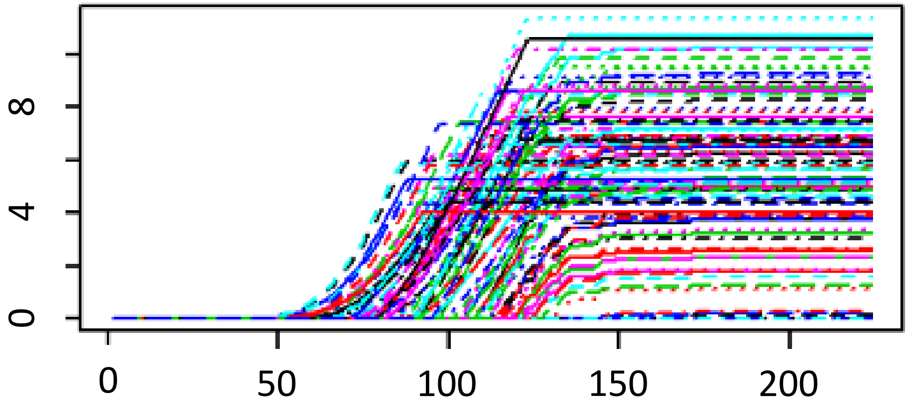
Context

SENSITIVITY ANALYSIS (Saltelli et al. 2008)



Motivation: Regional Sensitivity Analysis





⇒ **Which parameters drive the model ouputs toward different behaviors / regions in the model output space?**

Cluster-based GSA

Principle

Combines clustering methods

=> reveal and characterize multiple distinct behaviors of the model outputs

and variance-based methods

=> identify in a robust way parameters and interactions that drive these different behaviors

Fuzzy Clustering

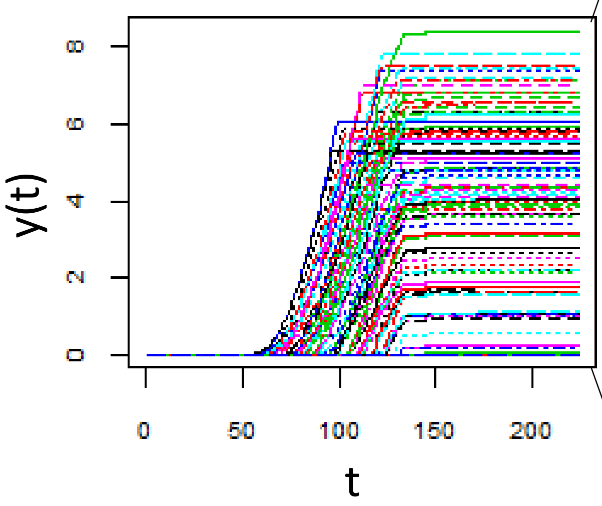
Computes Membership Functions $MF_k(y_i(t))$

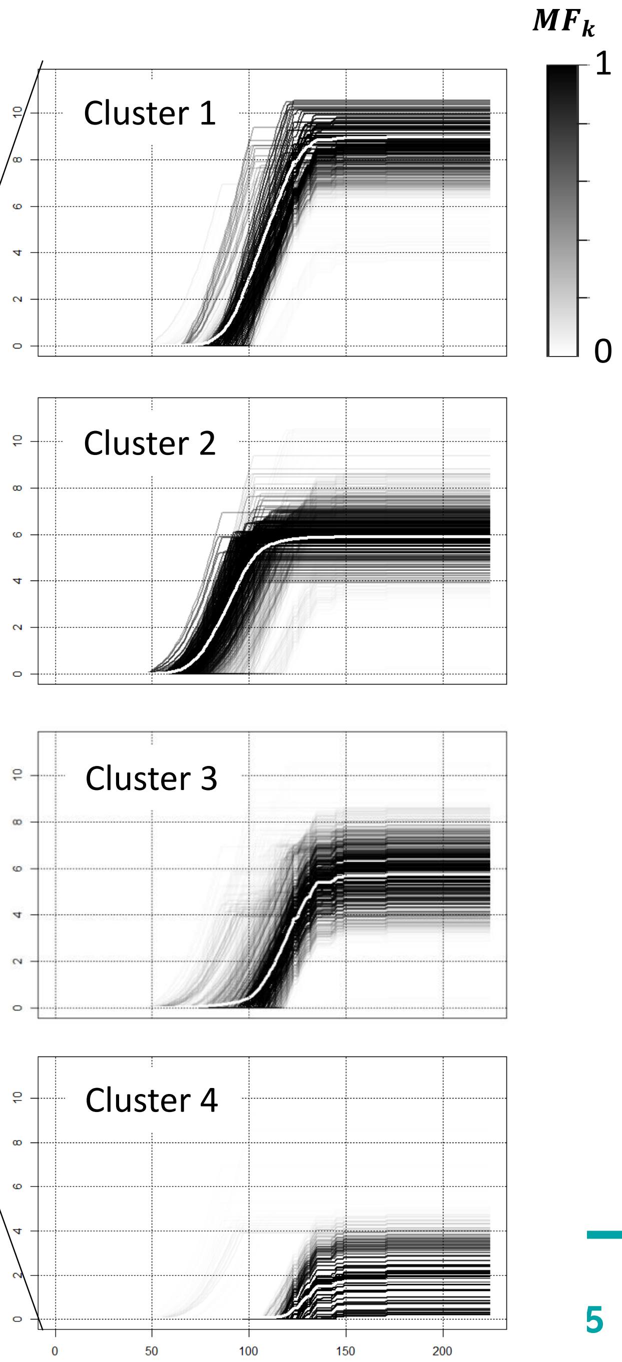
$\Rightarrow$ **Degree of membership of a simulated output to cluster k**

$$MF_k(y_i(t)) \in [0,1]$$

$$\sum_{k=1}^{K} MF_k(y_i(t)) = 1$$

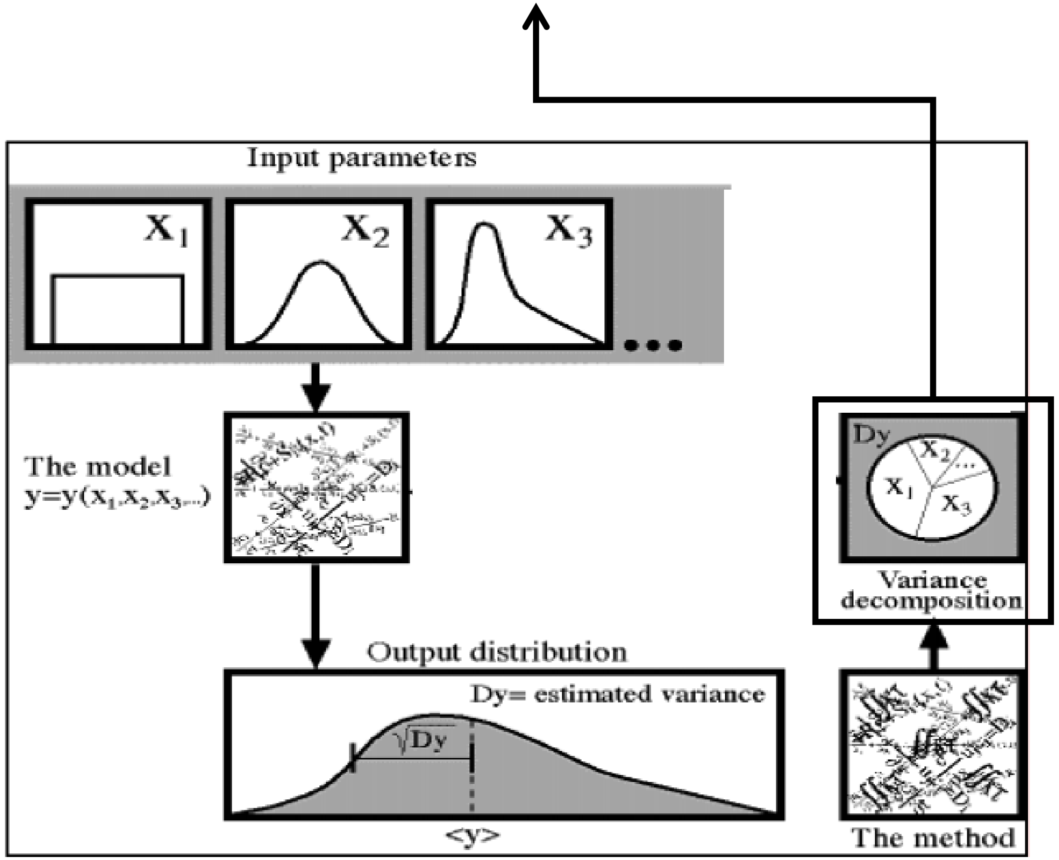
(see e.g. Höppner et al, 1999)





Variance-Based methods

Parts of variance explained by the different parameters



From Saltelli et al (2001)

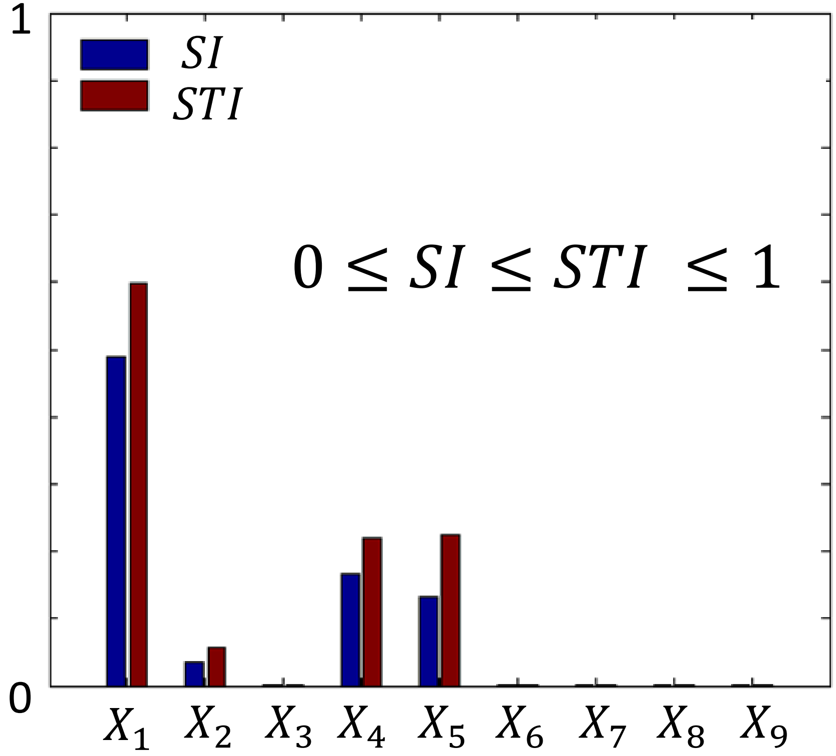
Sobol' Indices:

X_i alone

$$SI(X_i) = \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(Y)}$$

X_i and its interactions

$$STI(X_i) = \frac{V_{X_i}(E_{X_{\sim i}}(Y|X_i))}{V(Y)}$$



Cluster-based GSA indices

- **Clust-SI: sensitivity indices on membership functions (u_k)**

=> Which parameters (and interactions) drive the model outputs toward a targeted cluster k?

$$SI_k(X_j)=\frac{V(E(u_k|X_j))}{V(u_k)} \qquad TSI_k(X_j)=1-\frac{V(E(u_k|X_{-j}))}{V(u_k)}$$

- **dClust-SI: sensitivity indices on membership function differences**

=> Which parameters (and interactions) drive the model outputs from one cluster k to another i?

$$SI_{kl}(X_j)=\frac{V(E((u_k-u_l)|X_j))}{V(u_k-u_l)} \qquad TSI_{kl}(X_j)=1-\frac{V(E((u_k-u_l)|X_{-j}))}{V(u_k-u_l)}$$

- **Clust-GSI: aggregated indices on the vector of membership functions**

=> Which parameters (and interactions) globally impact changes between clusters

$$SI(X_j)=\frac{\sum_{k=1}^{K}V(u_k)SI_k(X_j)}{\sum_{k=1}^{K}V(u_k)} \qquad TSI(X_j)=\frac{\sum_{k=1}^{K}V(u_k)TSI_k(X_j)}{\sum_{k=1}^{K}V(u_k)}$$

Example

The CANTIS model

Carbon And Nitrogen Transformations in Soils (Garnier et al., 2001)

- **within Vsoil platform (https://www6.inrae.fr/vsoil)**
- **Standard batch conguration**
- **Simulation on a 3 months period**

- **Output of interest: ZYB**
  - ZYmogenous microbial Biomass
  - controls CO_2 emissions and soil mineral N content

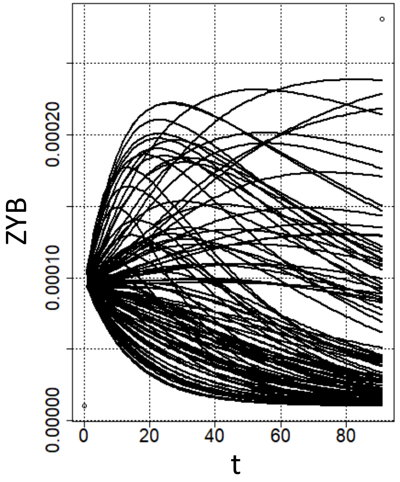
Uncertain parameters considered

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	-	$U[0.3, 0.6]$

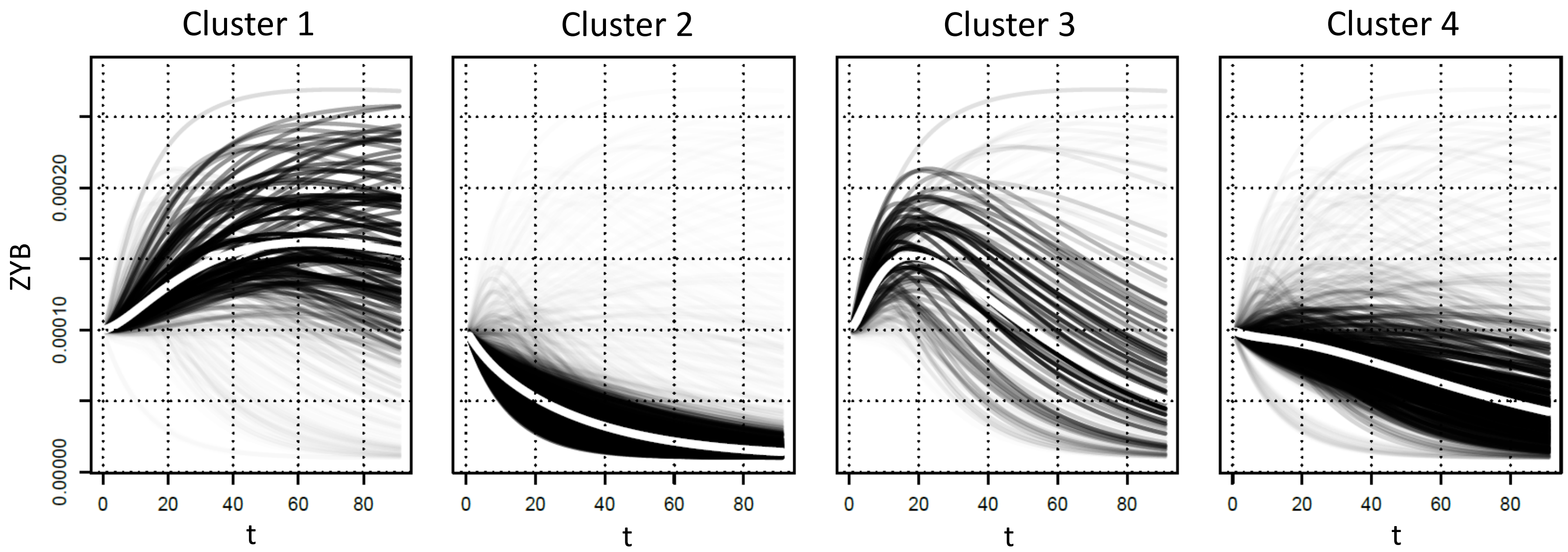
Example

Clustering

- **sobolSalt Design-Of-Experiment** **(R package sensitivity),**
  N=2500 (55000 parameter values) => CANTIS



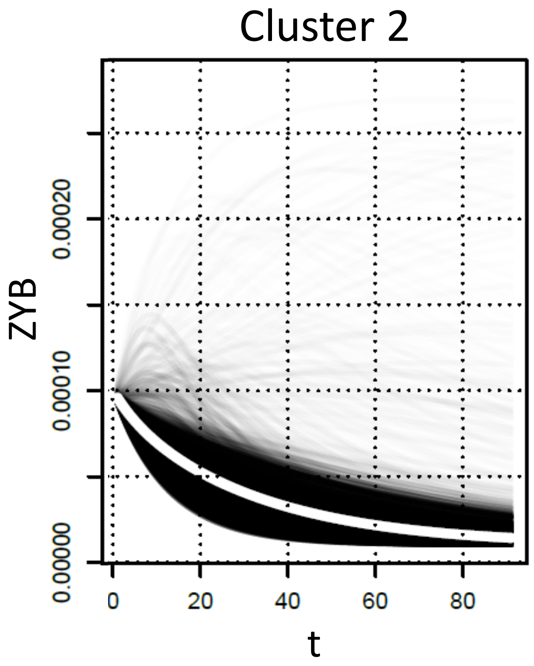
- **Fuzzy c-means clustering** **(R package fclust)**

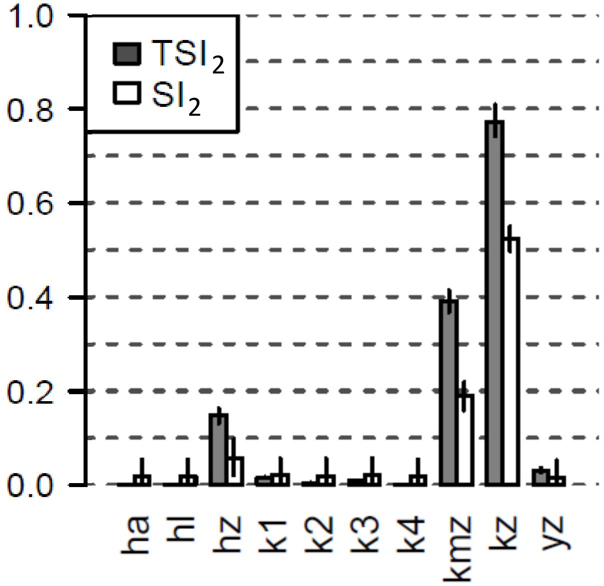


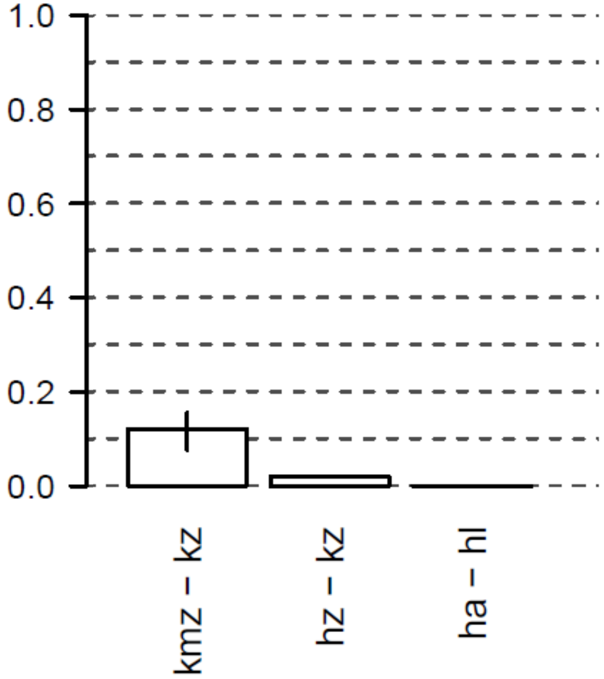
Example

Sobol' Indices

Which parameters lead to continuously decreasing ZYB dynamics?





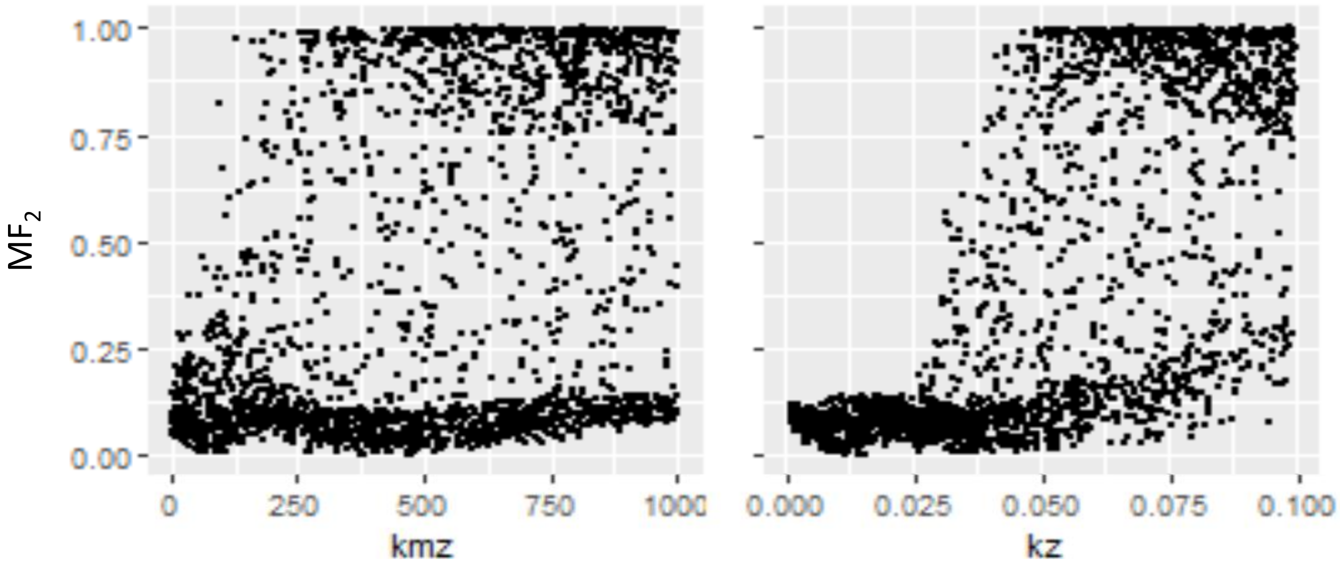


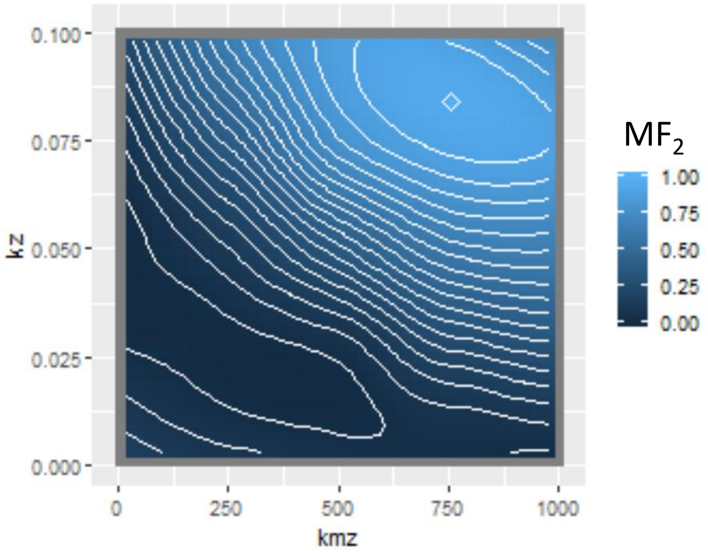
The variance of MF_2 is mostly explained by kz and kmz (1st orders and interaction)
=> they lead to minimal values of ZYB all along the simulation.

Example

Graphical sensitivity analysis

Which values of the parameters lead to continuously decreasing ZYB dynamics?

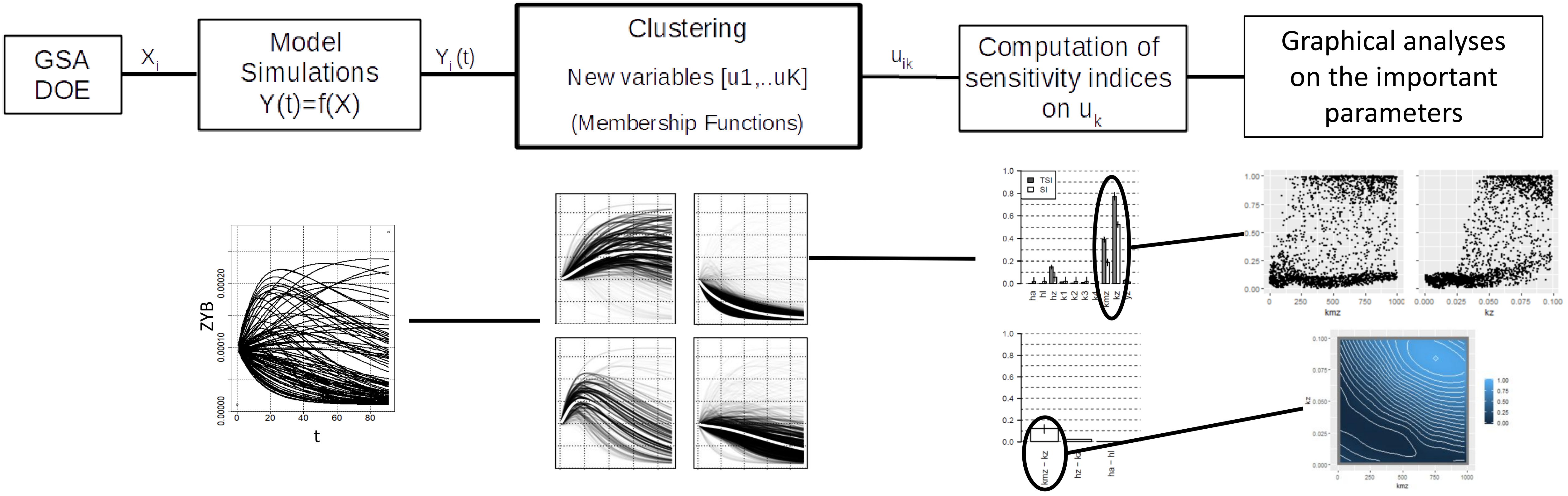




=> only medium to high values

=> only when both take medium to high values

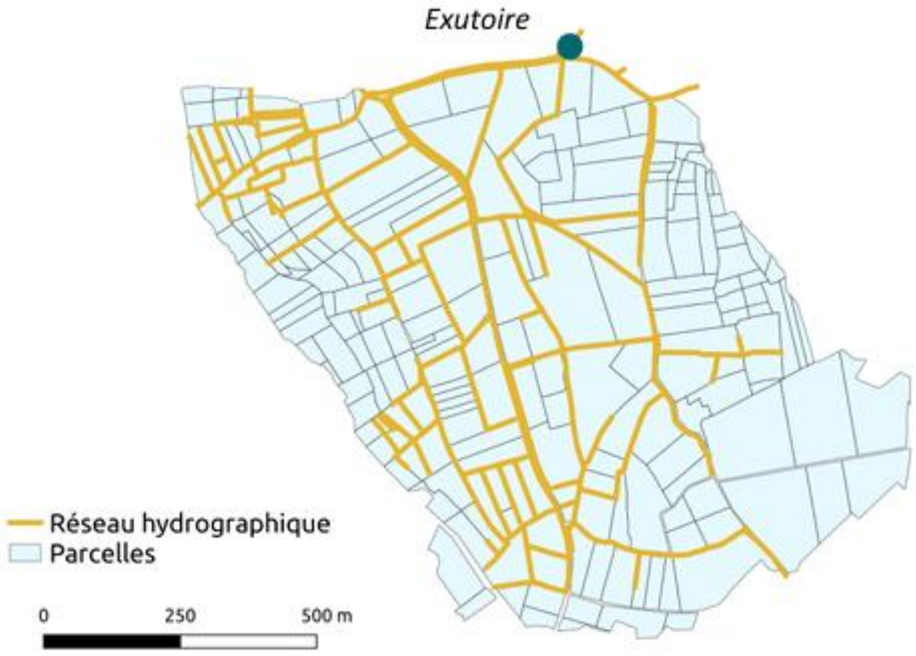
Summary of the approach



Example with spatial outputs: context

EauCharb'Oc - Défi Clé Water Occitanie

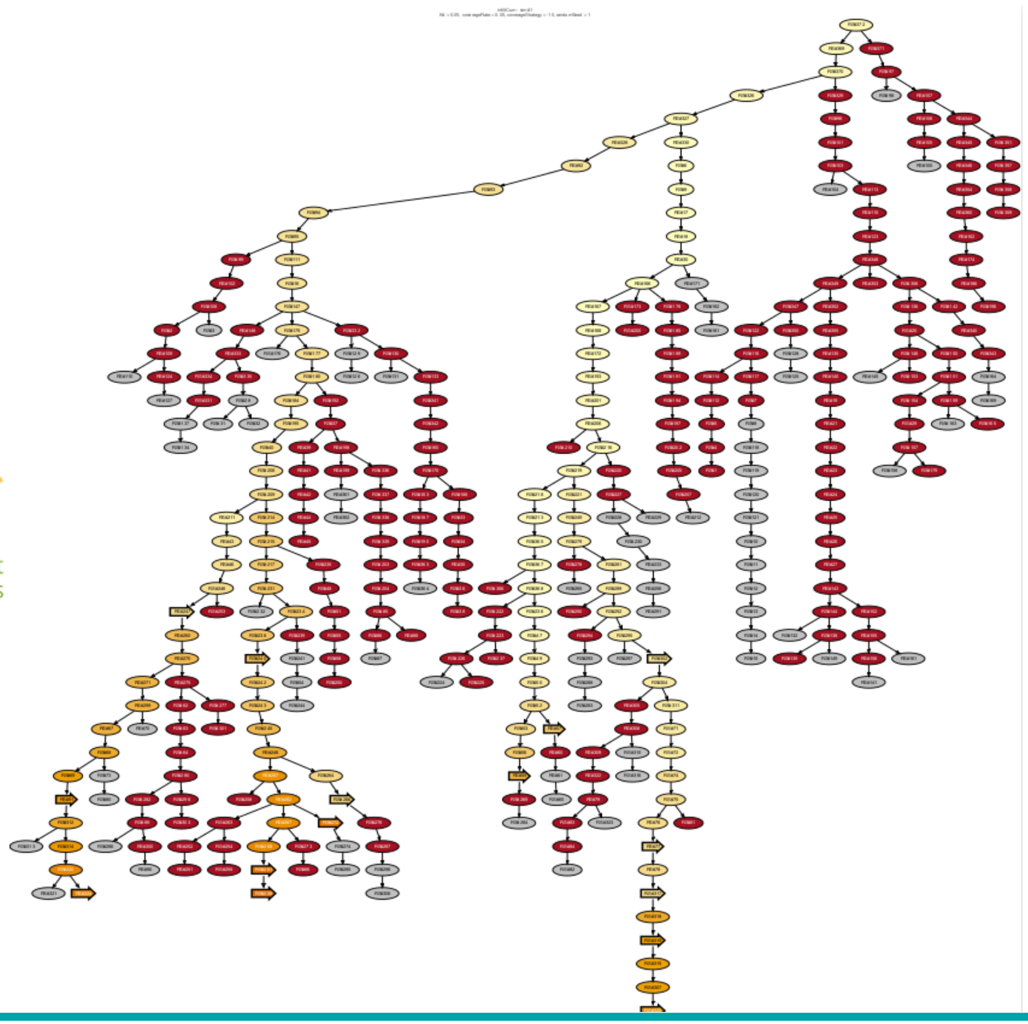
Transfer and retention of pesticides in agricultural ditch networks



Roujan (0.91 km2)

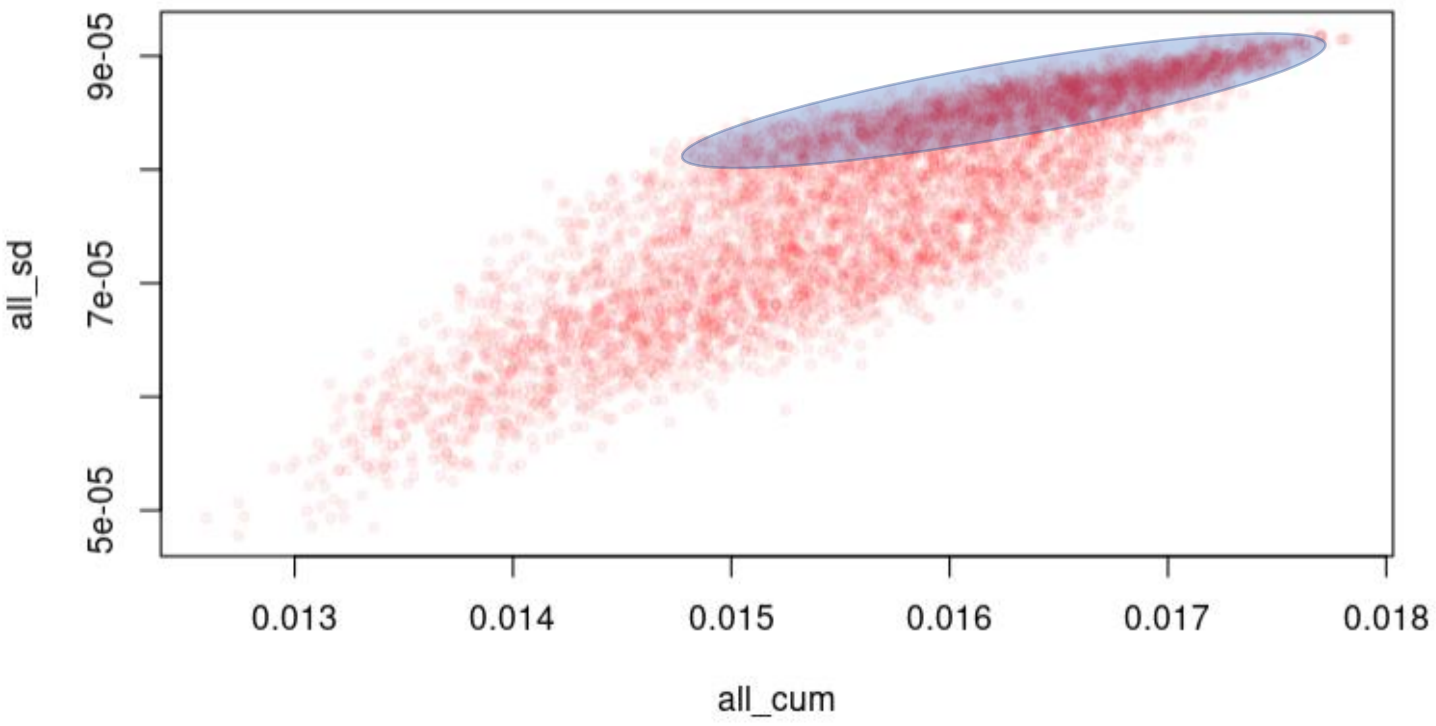
OpenFLUID
Software Environment
for Spatial Modelling in Landscapes

237 field plots
372 ditch reach

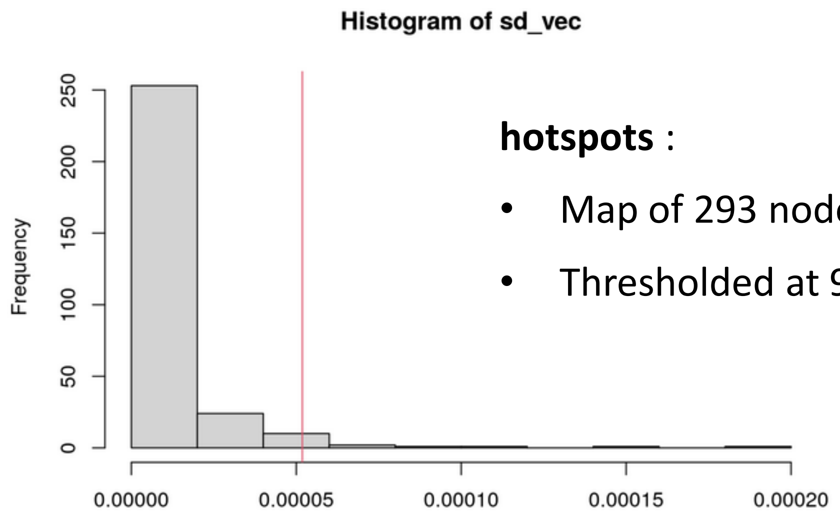


Example with spatial outputs: reduction to 2D

Summarize information in a 2D plot: cumulative VS variability of infiltration

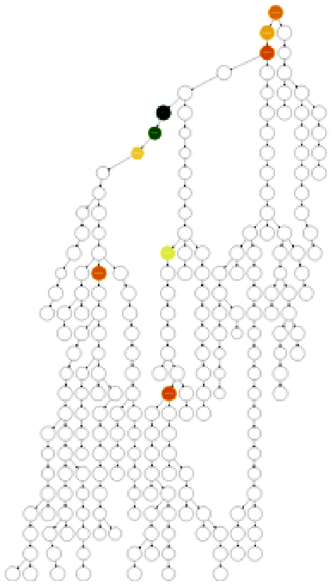


Example with spatial outputs: reduction to hotspots



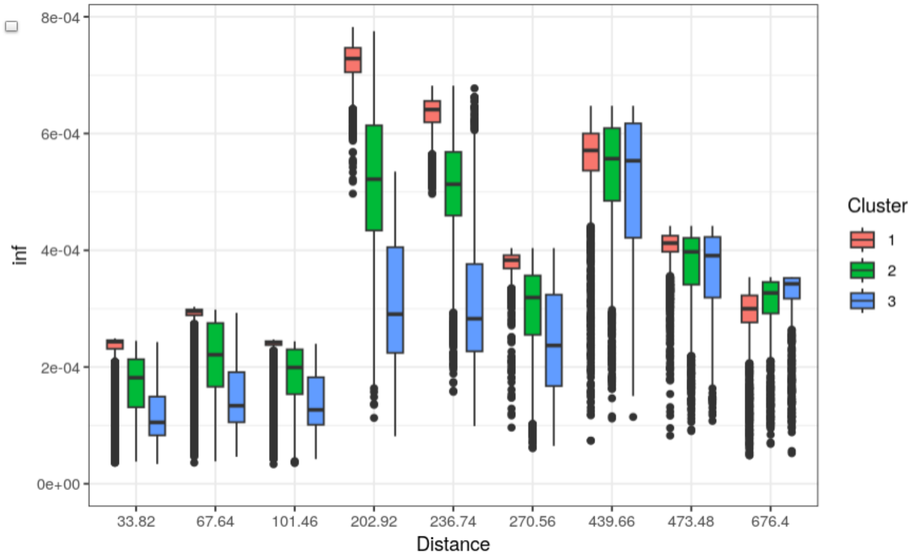
hotspots :

- Map of 293 node-to-node standard deviations (from 6000 simulations)
- Thresholded at 97% (top 9 most variable nodes)



Cluster visualization and interpretation:

Analysis of the 9 hotspots, e.g., in relation to distance from the outlet



Conclusion

Cluster-based GSA (Roux, Buis et al, 2021 EMS)

- a new method to quantify the **influence of input factors on targetted behaviors** of model outputs (**of any dimension**)
- validated on a toy model and successfully applied to complex ones
- **generic** wrt clustering (crisp or fuzzy, automatic or manual, distance used, …) and GSA methods used
- can be efficiently combined with graphical SA to identify regions of model inputs that lead to regions of model outputs (Factor Mapping)
- **relies on the interpretability of the clusters**

Perspective:

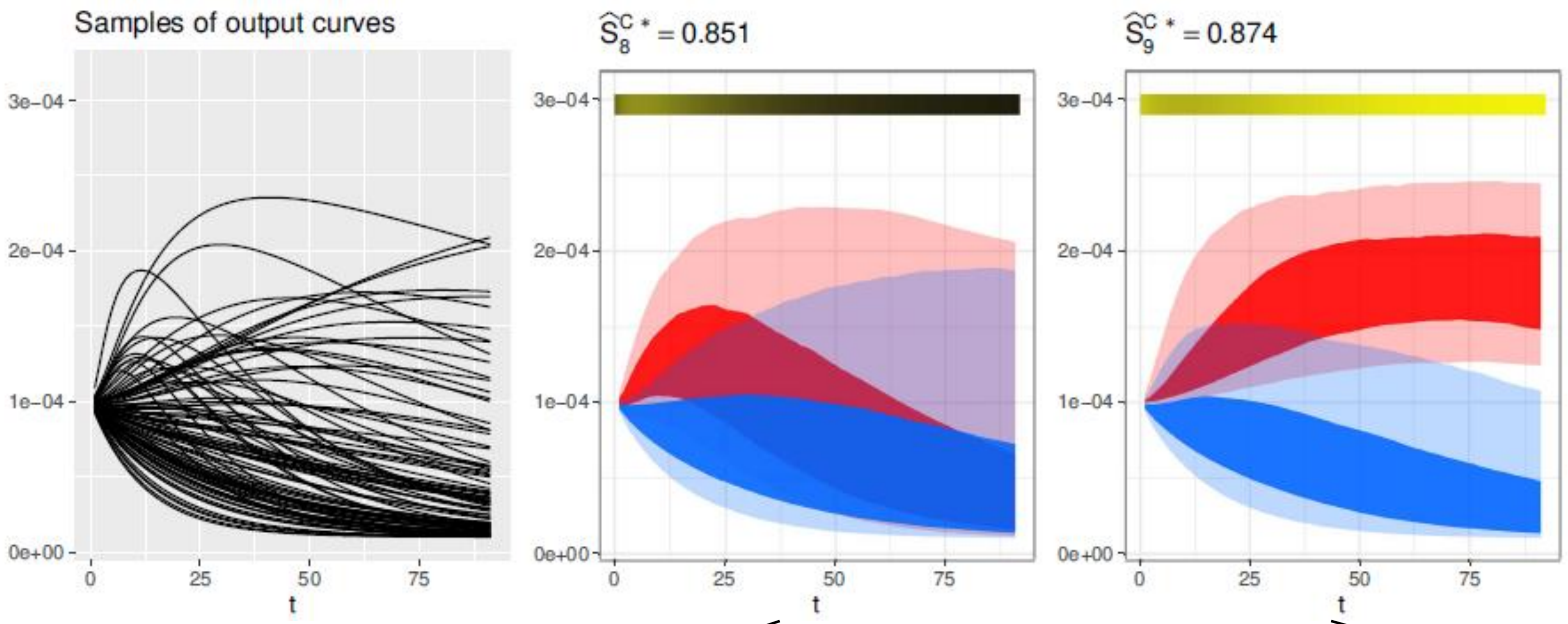
- provide indices to quantify how parameters influence the variation of model outputs within a cluster

Thank you
for your attention

Extension: maximized RSA (mRSA)

Principle

Revealing the behavior most impacted by variations of a parameter (or a group or an interaction,..)



kmz controls the occurrence of rising/falling biomass curves

kz controls the occurrence of curves associated with high / low biomass levels

Extension: maximized RSA (mRSA)

Algorithm

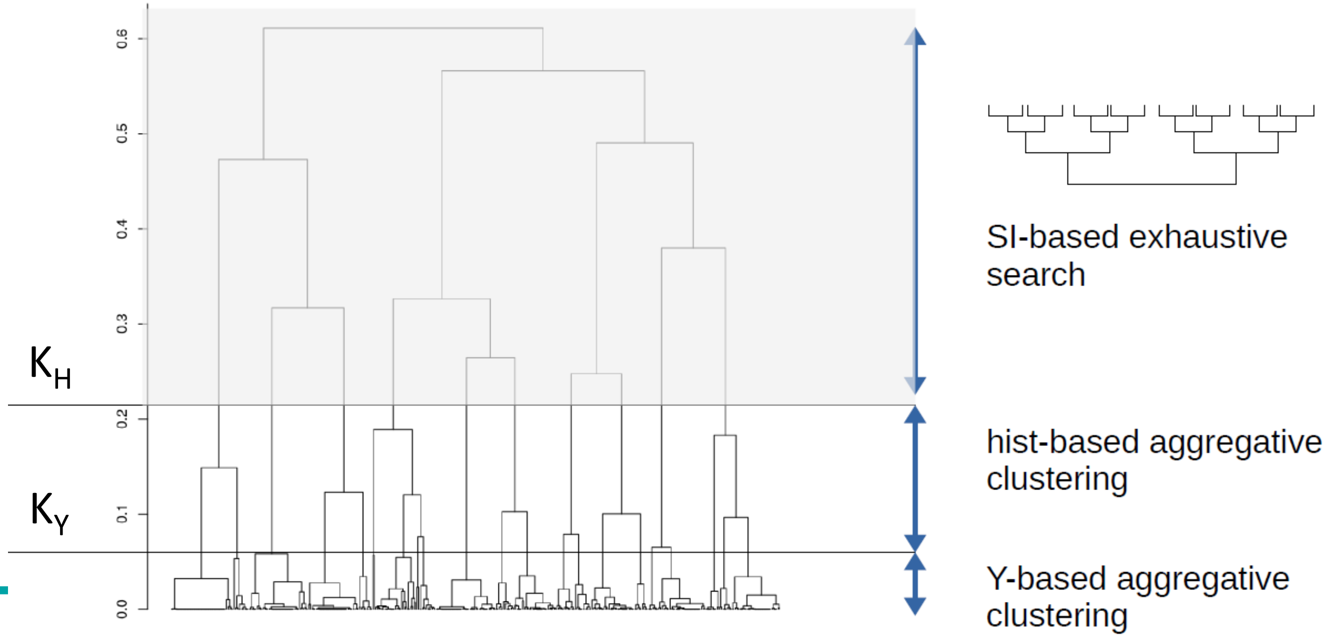
$\Rightarrow$ defining a **sensitivity score** associated to the influence of an input factor in driving the ouput toward a region of the output space

$\Rightarrow$ defining an **optimization algorithm** to find the clustering of the output space that maximizes the criterion

Algorithm 1: mRSA

Input: $(\mathbf{X}, \mathbf{Y})$
Parameters: $(i, ClustFun^Y, K_Y, n_X, K_H, \gamma)$
Output: $(P_{k^*}, \overline{P_{k^*}}), \widehat{S}_i^{P_{k^*}}$

Apply $ClustFun^Y$ on $\mathbf{Y}$ to get K_Y elementary clusters $(C_1, ..., C_{K_Y})$
Compute all histograms $(h^{C_k})_{k=1...K_Y}$ for input X_i discretized with n_X bins
Apply $ClustFun^H$ on $(h^{C_1}, ..., h^{C_{K_Y}})$ to get K_H metaclusters $(\widehat{C}_1, ..., \widehat{C}_{K_H})$
for all 2-partitions $(P_k, \overline{P_k})$ of $(\widehat{C}_1, ..., \widehat{C}_{K_H})$ **do**
  Compute an estimate $\widehat{S}_i^{P_k}$ of $S_i^{P_k}$ from $(\mathbf{X}, \mathbf{Y})$
  Compute $\gamma_k = 1/N \min(\mathrm{Card}(P_k), \mathrm{Card}(\overline{P_k}))$, where N = number of rows in $\mathbf{X}$
end for
$k^* = \arg\max_{k, \gamma_k \geq \gamma} \widehat{S}_i^{P_k}$



Conclusion

Two methods proposed to analyze the relationships between model input factors and behaviors of simulated model outputs (of any dimension)

CB-GSA

Roux, Buis et al. (2021) EMS

https://github.com/sbuis/ClusterBased_GSA

Revealing the parameters that drive the model outputs toward specific behavior(s)

mRSA

Roux, Loisel & Buis (2025) IJUQ

https://forge.inrae.fr/sebastien.roux/mrsa_paper

Revealing the behavior most impacted by variations of specific parameter(s)

Can be used with any GSA method
Make use of different types of clustering

expertise-driven clustering
Clustering based on Y

data-driven clustering
Optimization based on Y
(fuzzy or crisp)

sensitivity-driven clustering
Optimization based on Y and X

Prior partitions

Optimized partitions

… but fully rely on the interpretability of the clusters