



Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators

Matthias de Lozzo, Amandine Marrel

► To cite this version:

Matthias de Lozzo, Amandine Marrel. Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators. Stochastic Environmental Research and Risk Assessment, 2017, 31, pp.1437-1453. hal-01253686

HAL Id: hal-01253686

<https://hal.science/hal-01253686>

Submitted on 11 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators

Matthias De Lozzo¹ and Amandine Marrel¹

¹CEA,DEN,DER, F-13108 Saint Paul Lez Durance, France

January 11, 2016

Abstract

In a case of radioactive release in the environment, modeling the radionuclide atmospheric dispersion is particularly useful for emergency response procedures and risk assessment. For this, the CEA has developed a numerical simulator, called Ceres-Mithra, to predict spatial maps of radionuclide concentrations at different instants. This computer code depends on many uncertain scalar and temporal parameters, describing the radionuclide, release or weather characteristics. The purpose is to detect the input parameters the uncertainties of which highly affect the predicted concentrations and to quantify their influences.

To this end, we present various measures for the sensitivity analysis of a spatial model. Some of them lead to as many analyses as spatial locations (site sensitivity indices) while others consider a single one, with respect to the whole spatial domain (block sensitivity indices). For both categories, variance-based and dependence measures are considered, based on recent literature. All of these sensitivity measures are applied to the C-M computer code and compared to each other, showing the complementarity of block and site sensitivity analyses. Finally, a sensitivity analysis summarizing the input uncertainty contribution over the entirety of the spatio-temporal domain is proposed.

Keywords: Spatio-temporal models, Global sensitivity analysis, Sobol' indices, Dependence measures, Site & block sensitivity measures, Gaussian process model.

1 Introduction

The Ceres-Mithra (C-M) application is a numerical simulator developed by the CEA for the modeling of radionuclide atmospheric dispersion, after a radioactive material release [26]. This phenomenon can be a consequence of a nuclear power plant accident or accidental releases due to other events. Such a simulator is very useful for emergency response procedures and risk assessment; it allows scientists to make an efficient characterization of the released radioactive substance transport to long distance and an accurate tracking of the more impacted areas. This model takes into account many physical parameters describing in particular the weather conditions and radionuclide release characteristics. The C-M simulator provides, as output, several radionuclide concentration maps. The input variables are highly uncertain due to instrumental limits, lack of knowledge, intrinsic variability and modeling hypotheses. An important objective is to evaluate how these uncertainties can affect the C-M code predictions and to identify the most influential uncertain parameters.

Global sensitivity analysis (GSA) aims at explaining the model output variability according to the input parameter uncertainties [32]. Quantitative GSA methods are helpful in the understanding of physical phenomenon; they can guide its modeling through a model variance reduction by dropping the non-significant variables and increasing the characterization of the influential ones. The input contributions are usually quantified using the Sobol' indices [33], a set of sensitivity measures representing the contribution of any single input parameter or interacting input group to the model output variance. Other sensitivity indices are based on probability density functions [4], model output derivatives [34] or, coarsely, finite differences [27].

Most of the GSA tools deal only with scalar variables while computer codes often consider more complex input or output parameters. In particular, the C-M simulator returns a spatial variable discretized over a grid for different instants and this specificity has to be taken into account. Moreover, the C-M code is CPU time expensive, which limits the number of possible simulations; consequently, quantitative GSA methods requiring

thousands of Monte-Carlo runs cannot be directly applied. This paper proposes solutions to deal with these two problematics.

In order to deal with the CPU time limitation, surrogate modeling is a widely used technique replacing an accurate but costly computer code with a fast mathematical function, called surrogate model or metamodel, copying as much as possible the numerical simulator output behavior [30, 12]. Due to its low calculation time and its satisfying approximation level, a surrogate model can be used in place of the computer code for a GSA based on intensive Monte-Carlo methods. For example, Crestaux et al. [6] and Marrel et al. [22] compute the Sobol' indices using polynomial chaos expansion [16] and Gaussian process model [29] respectively. In the context of the C-M application, Marrel et al. [24] approximate the spatial model output using a proper orthogonal decomposition the coefficients of which are modeled by Gaussian process models.

For the problem of sensitivity analysis in presence of a spatial and more generally multivariate output, some recent works try to provide a solution. From a semantic point of view, Saint-Geours [31] opposes the notions of site and block sensitivity indices, the first ones running a GSA for each local site of the discretized output and the second proposing global sensitivity indices with respect to all the sites included in an area of interest. Among site methods, Marrel et al. [21] map the Sobol' indices over the grid associated to the model output. With respect to block sensitivity indices, Campbell et al. [5] use an empirical orthogonal decomposition of the multidimensional model output (Legendre polynomial, Principal Component Analysis or Partial Least Square decomposition) and apply one dimensional variance-based GSA to the more informative components. To go further than this component by component GSA, Lamboni et al. [19] propose a generalized sensitivity index for each input variable; this is a weighted sum of the first-order Sobol' indices associated to this input and computed for the different components of the decomposition. This method has been applied to an industrial nuclear reactor application [2]. More recently, Gamboa et al. [15] propose an explicit generalization of Sobol' indices for vectorial outputs and study it from a theoretical point of view. The authors mention the link between their sensitivity measures and those proposed by Lamboni et al. [19] when all the components are used in the weighted summation.

All the previous approaches are based on the decomposition of the output variance. Recently, Da Veiga [7] introduced the use of dependence measures in a sensitivity analysis framework. He considers the covariance between some non-linear transformations of the input and output parameters in reproducing kernel Hilbert spaces [17]], or the distance between the output probability density function and the one conditioned by the input variable of interest [35]. De Lozzo and Marrel [8] extend this use to a screening purpose, the objective of which is to separate the input parameters into two groups: the significant and the non-significant. Besides the fact that these dependence measures require less observation than the Sobol' indices for the estimation step, Da Veiga [7] opens the way for their use with multidimensional variables, providing a first application on a problem of pollutant migration in a waste storage site [36]. Lastly, Marrel et al. [23] apply these dependence measures to a 3-length vectorial output and compare them to generalized Sobol' indices on a nuclear accident assessment.

In this paper, we shall focus our attention on the comparison of sensitivity methods based on variance decomposition and dependence measures for the GSA of the spatio-temporal C-M model. These methods are based on generalizations of sensitivity measures for multivariate outputs. This work is guided by many constraints: the spatial-temporal characteristic of the C-M model output, the time-dependence of some inputs and the important CPU cost of each C-M simulation.

Firstly, we describe the Ceres-Mithra atmospheric model in Section 2. Then, sensitivity measures for a spatial output are then presented in Section 3, considering variance decomposition methods and dependence measures, from both site and block points of view. Last, we apply these tools in Section 4 to the C-M problem and compare their results and interpretations.

2 Ceres-Mithra case study

2.1 Atmospheric model description

Following an accidental release of radionuclides in the atmosphere, the Ceres-Mithra (C-M) simulator evaluates instantaneous and time integrated activity concentrations for different points and moments. To do this, several phenomena are simulated: transport, diffusion, impaction and sedimentation. Atmospheric transport modeling is carried out with the Gaussian puff model [26]. This model assumes that a sequence of individual puffs of pollutant is released from the source. Different standard deviation equations can be used; Doury's formulas [10]

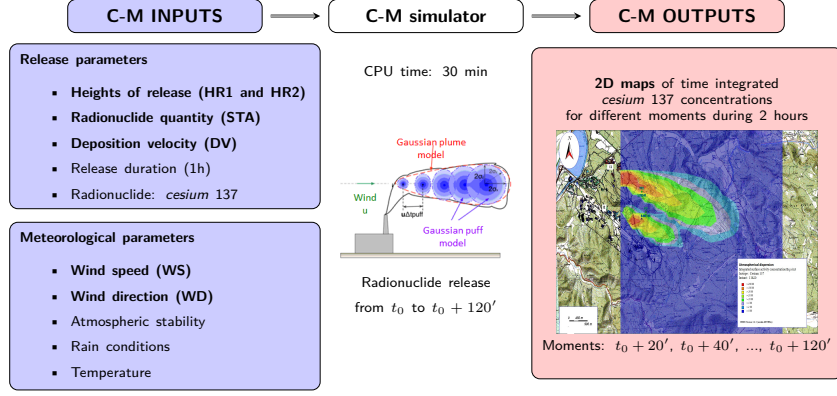


Figure 1: Flowchart of the studied C-M scenario, with the uncertain input parameters in bold.

which are function of travel time are the default option used in this study. Instantaneous and time integrated volume activity concentrations are predicted. Deposits on the ground result from mechanisms of diffusion, impaction and sedimentation from washing out of puffs during rainy situations. The dry deposition velocity is considered independent of distance from the emissary. Quite important rain leads to more significant deposits. The calculation of wet deposition velocity with distance from emissary is possible by taking into account the washing out rate. For aerosols or vapors, depletion due to dry and wet deposition is calculated. Instantaneous and time integrated surface activity concentrations are predicted.

2.2 Studied scenario of accidental release

In this paper, we consider the scenario studied by Marrel et al. [24]. It is made up of a *cesium 137* (noted ^{137}Cs) release occurring simultaneously in two installations of a nuclear site, with distinct locations and heights of release, during one hour with a constant ^{137}Cs activity. The radionuclide atmospheric dispersion is studied during two hours from the beginning of the accidental releases, with time steps of 20 minutes. This dispersion process is function of the radionuclide deposition velocity and of the released radionuclide quantity. Concerning the weather conditions, there is no rain and the temperature, the hydrometry and the atmospheric stability are assumed to be constant. Moreover, we supposed that the wind is blowing from the West, with an origin direction lying in $[249^\circ; 333^\circ]$; this situation represents a dry weather class which is the most likely one. For the considered nuclear site, because of its role in the radioactive puff spread, the temporal variation of the wind speed and direction period has to be taken into account in the phenomenon study.

Then, the model output predicted by the C-M simulator is a discretized spatial map of time integrated surface activity concentration of ^{137}Cs (Bq.s.m^{-3}), at moments $t_0 + 20'$, $t_0 + 40'$, ..., $t_0 + 120'$ where t_0 is the beginning of releases. The associated two-dimensional spatial grid is made up of $60 \times 45 = 2700$ points and is the same for any C-M simulations. Figure 1 represents the flowchart of this scenario.

2.3 Input uncertainty modeling

In this problem, six input parameters are considered as uncertain and Table 1 describes the probability distributions modeling these uncertainties. More details about these choices and the law hyperparameters are given in the article [24]. The release heights of both installations (HR1 and HR2), used by the C-M simulator, depend on the atmospheric stability conditions which involve effective heights different from the installation ones. From expert judgment, their uncertainties are modeled by uniform distributions. The released radionuclide quantity, represented by the source term activity (STA), is assumed to follow a log-normal distribution parametrized by the sensor measure error. The same kind of probability law governs the deposition velocity (DV); this assumption is based on expert opinion. Lastly, the wind speed and direction (WS and WD) are random temporal processes approached by first-order autoregressive models (AR1)

$$(X_t - \mu) = a(X_{t-1} - \mu) + \varepsilon_t, \text{ where } \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

where μ, a, σ^2 are the AR1-hyperparameters. For both WD and WS, these parameters have been estimated in a previous study by maximizing the likelihood associated to a database of ten year meteorological records obtained from a sensor placed on the nuclear site, with a 10-minute time step [24]. To simulate a realization of WS (or WD), a trajectory is sampled by 10-minute time step using the AR1 model from a fixed initial value.

Uncertain inputs	Notation	Nature	Probability distribution
First height of release (m)	HR1	real scalar	Uniform
Second height of release (m)	HR2	real scalar	Uniform
Deposition velocity (m.s ⁻¹)	DV	real scalar	Log-uniform
Source term activity (Bq)	STA	real scalar	Log-uniform
Wind direction (°)	WD	real vector	AR(1) truncated process
Wind speed (m.s ⁻¹)	WS	real vector	AR(1) truncated process
Uncertain outputs			
Time integrated ¹³⁷ Cs concentrations (Bq.s.m ⁻³)	TIC	real matrices (1 per time step)	

Table 1: Uncertainty characterization of the inputs and outputs, for the C-M application.

As a result, to obtain a 2-hour trajectory, 12 values of white noise ε_t are sampled. The C-M simulator dealing with a 20-minute time step, the simulated WS and WD trajectories are then averaged by 20-minute interval, leading to two vectors of length 6. Consequently, there is a total of $4 + 12 \times 2 = 28$ random parameters for each simulation: 1 for HR1, 1 for HR2, 1 for STA, 1 for DV, 12 for WD and 12 for WS. At the end, we are interested in the impact of 6 uncertain variables on the radionuclide concentration map: 4 scalar ones (HR1, HR2, DV and STA) and 2 vectorial ones (WD and WS).

3 Sensitivity measures for spatial outputs

We look at the spatial numerical simulator

$$\mathbf{Y} = f(\mathbf{X}) = f(X_1, \dots, X_{d_1}, \mathbf{X}_{d_1+1}, \dots, \mathbf{X}_{d_1+d_2}) \quad (1)$$

where $X_1, \dots, X_{d_1}$ are d_1 real scalar input parameters, $\mathbf{X}_{d_1+1}, \dots, \mathbf{X}_{d_1+d_2}$ are d_2 vectorial input variables, $\mathbf{Y} \in \mathbb{R}^p$ is a vectorial output and $d = d_1 + d_2$ is the total number of parameters. More precisely, for any $i \in \{1, \dots, p\}$, the component Y_i (also noted $Y(\mathbf{X}; z_i)$ or $Y(z_i)$) represents the simulator output value at the spatial point z_i of the grid $\mathcal{Z} = \{z_1, \dots, z_p\}$. The d input parameters are considered as random variables whose laws are perfectly known. About the d_2 vectorial input variables, we have in mind discretized versions of functional data, such as temporal curves or spatial fields. Note that for the C-M application, $d_2 = 2$ and the two corresponding vectorial inputs for the C-M prediction at the time $t = t_0 + i \times 20'$, $i \in \{1, \dots, 6\}$, are $(\mathbf{X}_{d_1+1}, \mathbf{X}_{d_1+2}) = (\mathbf{WD}_{[t]}, \mathbf{WS}_{[t]})$ where

$$\mathbf{WD}_{[t]} = (WD_{[t_0+i \times 20']})_{1 \leq i \leq (t-t_0)/20'}$$

and

$$\mathbf{WS}_{[t]} = (WS_{[t_0+i \times 20']})_{1 \leq i \leq (t-t_0)/20'}$$

with $(WD_{[\bullet]}, WS_{[\bullet]})$ being the value of the wind components at time $\bullet$. The initial values $(WD_{[t_0]}, WS_{[t_0]})$ are fixed and assumed to be known.

In the following, we present two families of sensitivity measures for quantifying the influence of an input parameter on the computer code output $\mathbf{Y}$: a first one based on output variance decomposition and a second one based on dependence measures. Using the terminology introduced by Saint-Geours [31], site and block versions are proposed for both types, what provides sensitivity measure maps corresponding to the grid $\mathcal{Z}$ and scalar sensitivity measures summarizing these informations.

In the framework of the C-M application, the sensitivity measure estimators will be built using $\mathcal{S} = (\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})_{1 \leq i \leq N}$, a N -sample of the pair $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y}^{(i)} = f(\mathbf{X}^{(i)})$.

3.1 Use of Sobol' indices

Since the original paper of Sobol [33], the variance-based sensitivity measures called Sobol' indices are widely used in GSA industrial applications and the literature about their enhancement is abundant.

3.1.1 Classical Sobol' indices for a real scalar output

We consider a numerical simulator $Y = f(X_1, \dots, X_d) \in \mathbb{R}$ function of the independent random input variables $X_1, \dots, X_d$. Assuming that $\mathbb{E}[f^2(\mathbf{X})] < \infty$, the following decomposition is derived [11]:

$$\begin{aligned} f(\mathbf{X}) &= f_\emptyset + \sum_{i=1}^d f_i(X_i) + \sum_{i=1}^d \sum_{i < j}^d f_{ij}(X_i, X_j) \\ &\quad + \dots + f_{1\dots d}(X_1, \dots, X_d) \\ &= \sum_{u \subset \{1, \dots, d\}} f_u(\mathbf{X}_u) \end{aligned} \quad (2)$$

where $f_\emptyset = \mathbb{E}[f(\mathbf{X})]$, $f_i(X_i) = \mathbb{E}[f(\mathbf{X})|X_i] - f_\emptyset$ and $f_u(\mathbf{X}_u) = \mathbb{E}[f(\mathbf{X})|\mathbf{X}_u] - \sum_{v \subset u} f_v(\mathbf{X}_v)$, with the notation $\mathbf{X}_u = (X_i)_{i \in u}$, for any $u \subset \{1, \dots, d\}$. All the 2^d summands in (2) have zero mean and are mutually uncorrelated with each other. This decomposition is unique and, for any $u \subset \{1, \dots, d\}$, leads to the Sobol' indices:

$$S_u = \frac{\mathbb{V}[f_u(\mathbf{X}_u)]}{\mathbb{V}[f(\mathbf{X})]} \text{ and } S_u^T = \sum_{\substack{v \subset \{1, \dots, d\} \\ v \supset u}} S_v,$$

where S_u and S_u^T are the first-order and total Sobol' indices for the input set $\mathbf{X}_u$. $(S_u)_{u \subset \{1, \dots, d\}}$ are the elements of the $f(\mathbf{X})$ variance decomposition according to Equation (2), normalized by the output variance; these elements sum to one.

Most of the time, we only consider the effect of a single variable X_i , $i \in \{1, \dots, d\}$, rather than an input group one $\mathbf{X}_u$. In this case, the Sobol' indices takes the form:

$$S_i = \frac{\mathbb{V}[\mathbb{E}[f(\mathbf{X})|X_i]]}{\mathbb{V}[f(\mathbf{X})]} \text{ and } S_i^T = \sum_{\substack{v \subset \{1, \dots, d\} \\ v \ni i}} S_v.$$

Meaning. The first-order Sobol' index S_i represents the part of the output variance explained by the isolated input parameter X_i . The total one S_i^T corresponds to S_i completed by the part of the output variance explained by X_i in interaction with any other input or group of inputs.

Estimation. Many estimators exist for the approximation of the first-order and total Sobol' indices [32], using evaluations of the numerical simulator or evaluations of a surrogate model, when this model is CPU time expensive. Most of the estimation methods are implemented in the R package `sensitivity` [28].

Limitation. Classical Sobol' indices are dedicated to scalar outputs. For the spatial model (1), they can be used for every element of the grid $\mathcal{Z}$ and thus, provide a classical Sobol' index map for any input parameter (see, e.g., [24]); in this case, these measures are called "site sensitivity indices" and make sense. On the other hand, for many years, extensions and other approaches have been proposed to deal with multidimensional input and output parameters (see, e.g., the review in [31]). In the following, we focus on recent works dedicated to multidimensional outputs.

3.1.2 Generalization of the Sobol' indices to a spatial output

Gamboa et al. [15] introduce generalized Sobol' indices for any input parameter X_i , $i \in \{1, \dots, d\}$, in the explanation of the multidimensional output variance. To achieve this, they consider the multidimensional model $\mathbf{Y} = f(\mathbf{X}) \in \mathbb{R}^p$ satisfying $\mathbb{E}[\|f(\mathbf{X})\|^2] < \infty$ and apply the Hoeffding decomposition to the function $f: \mathbb{R}^d \mapsto \mathbb{R}^p$:

$$f(\mathbf{X}) = f_\emptyset + f_i(X_i) + f_{\sim i}(X_{\sim i}) + f_{i, \sim i}(X_i, \mathbf{X}_{\sim i})$$

where $f_\emptyset = \mathbb{E}[\mathbf{Y}]$, $f_i = \mathbb{E}[\mathbf{Y}|X_i] - f_\emptyset$, $f_{\sim i} = \mathbb{E}[\mathbf{Y}|\mathbf{X}_{\sim i}] - f_\emptyset$ and $f_{i, \sim i} = \mathbf{Y} - f_{\sim i} - f_i - f_\emptyset$, with $\mathbf{X}_{\sim i} = (X_j)_{\substack{1 \leq j \leq d \\ j \neq i}}$. Then, they compute the covariance matrix of both sides of this equality and apply the trace operator in order to turn the matrix format into a scalar one:

$$\text{trace}(\Sigma) = \text{trace}(C_i) + \text{trace}(C_{\sim i}) + \text{trace}(C_{i, \sim i})$$

where $\Sigma = \mathbb{V}[f(\mathbf{X})] \in \mathcal{M}_p(\mathbb{R})$, $C_i = \mathbb{V}[f_i(X_i)]$, and so on. Finally, dividing both sides by $\text{trace}(\Sigma)$ leads notably to the first-order index:

$$S_i^{[G]} = \frac{\text{trace}(C_i)}{\text{trace}(\Sigma)} = \frac{\sum_{l=1}^p \mathbb{V}[Y_l] S_{i,l}}{\sum_{l=1}^p \mathbb{V}[Y_l]} \quad (3)$$

where $\mathbb{V}[Y_l]$ is the variance associated to the l^{th} scalar output and $S_{i,l}$ is the first-order Sobol' index associated to the l^{th} output for the i^{th} input parameter.

In practice, this index is approximated by the pick-and-freeze estimator [15]:

$$S_{i,N}^{[G]} = \frac{\sum_{l=1}^p \left(\sum_{j=1}^N a_{ijl} - \left(\sum_{j=1}^N b_{ijl} \right)^2 \right)}{\sum_{l=1}^p \left(\sum_{j=1}^N c_{ijl} - \left(\sum_{j=1}^N b_{ijl} \right)^2 \right)}$$

with $a_{ijl} = \frac{Y_l^{(j)} Y_l^{[i],(j)}}{N}$, $b_{ijl} = \frac{Y_l^{(j)} + Y_l^{[i],(j)}}{2N}$ and $c_{ijl} = \frac{(Y_l^{(j)})^2 + (Y_l^{[i],(j)})^2}{2N}$. where $\mathcal{S} = (\mathbf{X}^{(j)}, \mathbf{Y}^{(j)})_{1 \leq j \leq N}$ is a sample of N independent realizations of $(\mathbf{X}, \mathbf{Y})$ and where $\mathbf{Y}^{[i],(j)} = f(X_i^{(j)}, \check{\mathbf{X}}_{\sim i}^{(j)})$ with $\check{\mathbf{X}}_{\sim i}^{(j)}$ being an independent copy of $\mathbf{X}_{\sim i}^{(j)}$. Total indices can be computed in the same way thanks to the relation:

$$S_i^{[G],T} = S_i^{[G]} + S_{i,\sim i}^{[G]} = 1 - S_{\sim i}^{[G]}.$$

These generalized Sobol' indices are equal to the sensitivity measures proposed by Lamboni et al. [19], when all the terms of the output proper orthonormal decomposition (POD) used in their work are kept in the global sensitivity index computation.

3.2 Hilbert-Schmidt independence criterion

Da Veiga [7] introduces the use of dependence measures for global SA and De Lozzo and Marrel [8] extend this work to a screening purpose. Among these dependence measures, we focus on the Hilbert-Schmidt Independence Criterion (HSIC) which is based on input and output parameter space transformations. These new indices can be extended to multivariate inputs and outputs. Marrel et al. [23] have recently applied them on a 3-length vectorial output for a nuclear assessment study. These tools have also been used on a spatial output field for an industrial application of pollutant migration in a waste storage site [7]. In this section, we present the HSIC firstly, in a site point of view, and secondly in a block one.

3.2.1 HSIC definition, associated sensitivity measure and estimator

Gretton et al. [17] propose the Hilbert-Schmidt independence criterion (HSIC) for detecting the dependence between the random variables $W_1 \in \mathbb{W}_1$ and $W_2 \in \mathbb{W}_2$, based on the covariance between some transformations of W_1 and W_2 . For any $i \in \{1, 2\}$, an universal reproducing kernel Hilbert space (RKHS) $\mathbb{F}_i$ composed of functions mapping from $\mathbb{W}_i$ to $\mathbb{R}$ and defined by the kernel function k_i is considered. $\langle f, g \rangle_{\mathbb{W}_i}$ denotes the scalar product over $\mathbb{W}_i$ between the elements f and g of $\mathbb{W}_i$. See [1] for details on RKHS theory.

Then, the operator of crossed-covariance $C_{W_1 W_2}$ is the linear operator mapping from $\mathbb{F}_2$ to $\mathbb{F}_1$ and defined for any $f \in \mathbb{F}_1$ and any $g \in \mathbb{F}_2$ by:

$$\langle f, C_{W_1 W_2} g \rangle_{\mathbb{F}_1} = \text{Cov}(f(W_1), g(W_2)).$$

This operator generalizes the notion of covariance between W_1 and W_2 .

Finally, the Hilbert-Schmidt Independence Criterion (HSIC) [17] is defined as the Hilbert-Schmidt norm of the operator $C_{W_1 W_2}$ [9]:

$$\text{HSIC}(W_1, W_2) := \|C_{W_1 W_2}\|_{\text{HS}}^2 = \sum_{i,j} \langle u_i, C_{W_1 W_2} v_j \rangle_{\mathbb{W}_1}$$

where $(u_i)_{i \geq 0}$ and $(v_j)_{j \geq 0}$ are orthonormal bases of $\mathbb{F}_1$ and $\mathbb{F}_2$, respectively. Considering W'_1 and W'_2 independent copies of W_1 and W_2 , $\text{HSIC}(W_1, W_2)$ can be rewritten:

$$\begin{aligned} & \text{HSIC}(W_1, W_2) \\ &= \mathbb{E}_{W_1, W'_1, W_2, W'_2} [k_1(W_1, W'_1) k_2(W_2, W'_2)] \\ &+ \mathbb{E}_{W_1, W'_1} [k_1(W_1, W'_1)] \mathbb{E}_{W_2, W'_2} [k_2(W_2, W'_2)] \\ &- 2 \mathbb{E}_{W_1, W_2} [\mathbb{E}_{W'_1} [k_1(W_1, W'_1)] \mathbb{E}_{W'_2} [k_2(W_2, W'_2)]] . \end{aligned}$$

Dependence. This dependence measure is equal to zero if and only if the random variables W_1 and W_2 are independent. This assertion is no longer valid if the RKHS are not universal.

GSA. Applying the HSIC formula to the model (1) in a GSA purpose, Da Veiga [7] proposes the following sensitivity measure for the input parameter X_k , $k \in \{1, \dots, d\}$:

$$R_k = \frac{\text{HSIC}(X_k, \mathbf{Y})}{\sqrt{\text{HSIC}(X_k, X_k)\text{HSIC}(\mathbf{Y}, \mathbf{Y})}}. \quad (4)$$

Estimator. Considering the N -sample S , an estimator of the measure $\text{HSIC}(X_k, Y)$ is proposed in [17]:

$$\widehat{\text{HSIC}}(X_k, \mathbf{Y}) = \frac{1}{(N-1)^2} \text{Tr}(K_{X_k} H K_Y H)$$

where $H = I_N - \frac{1}{N}$, $K_{X_k} = (k_{X_k}(X_k^{(i)}, X_k^{(j)}))_{1 \leq i, j \leq N}$ and $K_Y = (k_Y(\mathbf{Y}^{(i)}, \mathbf{Y}^{(j)}))_{1 \leq i, j \leq N}$. Then, R_i can be estimated by plug-in:

$$\hat{R}_k = \frac{\widehat{\text{HSIC}}(X_k, \mathbf{Y})}{\sqrt{\widehat{\text{HSIC}}(X_k, X_k)\widehat{\text{HSIC}}(\mathbf{Y}, \mathbf{Y})}}.$$

The computational cost of these estimators are in $\mathcal{O}(N^2)$.

Kernel choice. The kernel functions involved in the HSIC definition can belong to various classes of kernel functions, such as the Gaussian, Laplacian or Matérn families [14]. These functions often require hyperparameter values which can be deduced from heuristic processes or fixed in order to maximize the HSIC value [3]. The Gaussian kernel function is often considered for a scalar or vectorial variable $\mathbf{w} \in \mathbb{R}^q$:

$$k(\mathbf{w}, \mathbf{w}') = \exp\left(-\sum_{k=1}^q \frac{(w_k - w'_k)^2}{\sigma_k^2}\right) \quad (5)$$

where σ_k is estimated by the empirical standard deviation associated to the sample $(w_k^{(1)}, \dots, w_k^{(N)})$, or by the empirical median associated to $(|w_k^{(i)} - w_k^{(j)}|)_{1 \leq i, j \leq N}$.

Block and site sensitivity indices. Referring back to the spatial model (1) and considering the input parameter X_k , $k \in \{1, \dots, d\}$, the block sensitivity index is given by Equation (4) and the site sensitivity indices are equal to

$$R_k(z_i) = \frac{\text{HSIC}(X_k, Y(z_i))}{\sqrt{\text{HSIC}(X_k, X_k)\text{HSIC}(Y(z_i), Y(z_i))}}. \quad (6)$$

for any $i \in \{1, \dots, p\}$.

For high dimensional input and output variables such as temporal curves or spatial fields, Da Veiga [7] advises the use of kernels depending on semi-metrics dedicated to functional data [13]. In the following, we present this method for a random output variable $Z(t)$ indexed by a real scalar parameter t .

3.2.2 HSIC for functional variables

Given a functional random variable $W(t)$, $t \in [t_1, t_p]$, such that $\mathbb{E}[W^2(t)] < \infty$, the functional principal component analysis (FPCA) gives the following truncated expansion $\tilde{W}^{(q)}(t) = \sum_{k=1}^q \left(\int_{t_1}^{t_p} W(t')v_k(t')dt'\right)v_k(t)$ where $v_1, \dots, v_q$ are the q first orthonormal eigenfunctions of the covariance operator $\Gamma(s, t) = \mathbb{E}[W(s)W(t)]$. Considering the semi-norm

$$\|W\|_q^{\text{PCA}} = \sqrt{\sum_{k=1}^q \left(\int_{t_1}^{t_p} W(t)v_k(t)dt\right)^2},$$

Ferraty and Vieu [13] define the associated $\mathbb{L}^2$ semi-metric, for any $i, j \in \{1, \dots, N\}$, by

$$\begin{aligned} d_q^{\text{PCA}}(W^{(i)}, W^{(j)}) &= \|W^{(i)} - W^{(j)}\|_q^{\text{PCA}} \\ &= \sqrt{\sum_{k=1}^q \left(\int_{t_1}^{t_p} (W^{(i)}(t) - W^{(j)}(t)) v_k(t) dt \right)^2} \end{aligned} \quad (7)$$

they call it the PCA semi-metric. Then, because we do not observe the function W but a discretized version $\mathbf{W} = (W(t_1), \dots, W(t_p))^T$, we use the ℓ^2 semi-metric

$$\begin{aligned} d_q^{\text{PCA}}(W^{(i)}, W^{(j)}) &= \sqrt{\sum_{k=1}^q \left(\sum_{l=1}^p m_l (W^{(i)}(t_l) - W^{(j)}(t_l)) V_{lk} \right)^2} \\ &= \|V' M (\mathbf{W}^{(i)} - \mathbf{W}^{(j)})\|_2 \end{aligned} \quad (8)$$

where V is the $p \times q$ real matrix whose columns are the eigenvectors associated to the q greatest eigenvalues of the covariance matrix $\hat{\Gamma} M$ where $\hat{\Gamma} = \frac{1}{N} \sum_{i=1}^N \mathbf{W}^{(i)} \mathbf{W}^{(i)T}$ and $M = \text{diag}(m_1, \dots, m_p)$ is the diagonal matrix made of quadrature weights, usually $m_i = t_i - t_{i-1}$, for $i \in \{1, \dots, p\}$.

In practice, the number of principal components q is such that some ratio α of the total variance is explained by the basis spanned by these first components:

$$q = \underset{\beta \in \{1, \dots, p\}}{\text{arginf}} \left\{ \frac{\sum_{k=1}^{\beta} \lambda_k}{\sum_{l=1}^p \lambda_l} \geq \alpha \right\}$$

where $\lambda_1 \geq \dots \geq \lambda_J$ are the eigenvalues of $\hat{\Gamma} W$.

Finally, given a kernel function k acting on $\mathbb{R}$, the kernel function applied to the observations $W^{(i)}$ and $W^{(j)}$ is

$$k_W(W^{(i)}, W^{(j)}) = k \left(d_q^{\text{PCA}}(W^{(i)}, W^{(j)})^2 \right).$$

In particular, for the Gaussian case, we write

$$k_W(W^{(i)}, W^{(j)}) = \exp \left(-d_q^{\text{PCA}}(W^{(i)}, W^{(j)})^2 \right). \quad (9)$$

Extension to a spatial variable. The PCA semi-metric (Equation (7)) has been proposed by Ferraty and Vieu [13] for functional variables indexed by a real scalar, such as spectrometric or temporal curves. Its use has been recently extended to spatial variables [7]. However, the approximation of Expression (7) by Expression (8) can be questionable when the discretization $\mathcal{Z}$ is irregular or not sufficiently fine; this criticism is emphasized by the dimension of the functional variable index. In the C-M case study for example, the output variable is indexed by a two-dimensional parameter $z \in \mathcal{Z}$ representing the spatial location, even by a three-dimensional one $(z, t) \in \mathcal{Z} \times \{t_0 + 20', t_0 + 40', \dots, t_0 + 120'\}$ if we consider the spatio-temporal aspect parametrized by a large time step. The mesh density, relatively to the physical phenomenon scale, is often lower in this kind of problems because of the high CPU time associated to a fine mesh in two or three dimensions and the use of Equation (8) is not appropriate.

To face this limitation, we propose in this paper an alternative kernel function to (9) based on a proper orthogonal decomposition (POD) of the model output vector:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathcal{V} H$$

with $\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}]$, $\mathcal{V}$ is the matrix whose columns $\mathbf{v}_1, \dots, \mathbf{v}_p$ are the eigenvectors of $\Sigma = \mathbb{V}[Y]$, also called components, and $H = \mathcal{V}^T (\mathbf{Y} - \boldsymbol{\mu})$ is a random vector. The eigenvalues associated to $\mathbf{v}_1, \dots, \mathbf{v}_p$ are noted $\lambda_1, \dots, \lambda_p$.

From a N -sample $\mathcal{S} = (\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})_{1 \leq i \leq N}$, the covariance matrix Σ is estimated by

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{Y}^{(i)} - \bar{\mathbf{Y}}) (\mathbf{Y}^{(i)} - \bar{\mathbf{Y}})^T \quad (10)$$

where $\bar{\mathbf{Y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{Y}^{(i)}$. Similarly, we note $\hat{\lambda}_k$ and $\hat{\mathbf{v}}_k$, the k^{th} eigenvalue and eigenvector of $\hat{\Sigma}$ respectively and we consider the score matrix $\hat{H} = \left(\hat{H}_{ik} \right)_{\substack{1 \leq i \leq N \\ 1 \leq k \leq p}}$ where $\hat{H}_{ik} = \langle \mathbf{Y}^{(i)} - \bar{\mathbf{Y}}, \hat{\mathbf{v}}_k \rangle_2$, $\langle \cdot, \cdot \rangle_2$ being the scalar product on $\mathbb{R}^p$. Finally, we propose the following kernel function:

$$k_q(Y, Y') = \exp \left(- \sum_{k=1}^q \frac{\langle \mathbf{Y} - \mathbf{Y}', \hat{\mathbf{v}}_k \rangle_2^2}{\hat{\sigma}^2} \right) \quad (11)$$

where $\hat{\sigma}$ is either the empirical standard deviation associated to the output sample $\left(\mathbf{Y}_j^{(i)} \right)_{\substack{1 \leq i \leq N \\ 1 \leq j \leq p}}$ or the empirical median associated to the output distance sample $\left(\left| \mathbf{Y}_k^{(i)} - \mathbf{Y}_k^{(j)} \right| \right)_{\substack{1 \leq i \neq j \leq N \\ 1 \leq k \leq p}}$.

Equation (11) replaces the classical squared Euclidean distance between $\mathbf{W}$ and $\mathbf{W}'$ by the same distance applied on linear transformations of these variables, which are function of the output POD.

3.3 Randomized dependence coefficient

We also consider in this paper the randomized dependence coefficient (RDC), which is another dependence measure recently proposed by Lopez-Paz et al. [20] and estimating the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (HRG):

$$\text{hgr}(X_k, \mathbf{Y}) = \sup_{f, g} \text{corr}(g(X_k), h(\mathbf{Y})),$$

where g and h are Borel-measurable functions with finite variance. To our knowledge, this dependence measure has never been used in a GSA purpose. This dependence measure has an advantageous computational cost in $\mathcal{O}(N \log(N))$ whereas the HSIC one is in $\mathcal{O}(N^2)$. For any N -sample $\mathcal{S} = (\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})_{1 \leq i \leq N}$ of $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^p$, the RDC associated to X_k , $k \in \{1, \dots, d\}$, is defined for any $(\gamma, \delta) \in \mathbb{N}_+ \times \mathbb{R}_+$ by

$$\text{RDC}_k = \sup_{\alpha, \beta \in \mathbb{R}^\gamma} \text{corr} \left(\alpha^T \Phi_{\mathbf{P}(X_k)}^{\gamma, \delta}, \beta^T \Phi_{\mathbf{P}(\mathbf{Y})}^{\gamma, \delta} \right) \in [0, 1].$$

Vectors $\mathbf{P}(X_k)$ and $\mathbf{P}(\mathbf{Y})$ are the empirical-copula transformations of the samples X_k and $\mathbf{Y}$ respectively and

$$\Phi_{\mathbf{P}(\mathbf{Y})}^{\gamma, \delta} = \left(\sin \left(\mathbf{w}_i^T (\mathbf{P}(\mathbf{Y}))_{j,:}^T + \mathbf{b}_i \right) \right)_{\substack{1 \leq i \leq \gamma \\ 1 \leq j \leq N}} \in \mathcal{M}_{\gamma \times N}(\mathbb{R})$$

with $\mathbf{w}_{\mathbf{Y}, 1}, \dots, \mathbf{w}_{\mathbf{Y}, \gamma} \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(0, \delta I_p)$ and $\mathbf{b}_{\mathbf{Y}, 1}, \dots, \mathbf{b}_{\mathbf{Y}, \gamma} \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(0, \delta)$, the same thing being done for X_k . The RDC between X_k and $\mathbf{Y}$ measures the largest canonical correlation between random non-linear projections of their respective empirical copula-transformations and is used for testing their dependence, based on the property “ $\text{RDC}_k = 0$ if and only if X_k and $\mathbf{Y}$ are independent”. The reader is referred to [20] for more details. In our study, we estimate the RDC using the default values of γ and δ given by the authors.

The RDC can deal with vectorial inputs and outputs. Consequently, a first approach to take into account functional variables consists in considering discretized versions of these quantities in a vectorial way.

4 Application to the Ceres-Mithra case study

Now, we apply the sensitivity indices presented in Section 3 to the Ceres-Mithra (C-M) study case described in Section 2. Both site and block versions are considered and the performances and interpretations of these different measures are compared. In a last part, we extend the notion of block sensitivity index to the temporal evolution of the spatial output, in order to summarize the contribution of an input uncertainty to the output one over the whole spatio-temporal domain.

The estimation of Sobol' indices requiring a high number of evaluations, the surrogate model built by Marrel et al. [24] is necessary because of the important CPU time cost of each C-M run. This metamodel is an orthogonal decomposition of the spatial output, whose coefficients are approximated by Gaussian process models (GPMs). These GPMs have been parametrized using a learning sample of size N , based on a *maximin* Latin hypercube sampling design [25, 18], with $N = 200$, and validated by cross-validation. Note that, from the AR1 model properties, these metamodels use the independent time increments of the wind components rather than the original vectors $\mathbf{WD}_{[t]}$ and $\mathbf{WS}_{[t]}$, in order to consider independent inputs in the surrogate

modeling. For an equitable comparison, this surrogate model is also used for the sensitivity indices based on a HSIC and RDC dependence measure, even if these indices require less simulations and could have been estimated directly from the C-M simulator. The different sensitivity indices are estimated from many thousands simulations of the previous metamodel.

For an easier lecture, site and block sensitivity indices (4) and (6) based on the HSIC and the RDC are normalized in order to sum to one:

$$\tilde{R}_k = \frac{R_k}{\sum_{i=1}^d R_i} \text{ and } \tilde{R}_k(z_i) = \frac{R_k(z_i)}{\sum_{i=1}^d R_i(z_j)},$$

for any $j \in \{1, \dots, p\}$ and any $k \in \{1, \dots, d\}$.

4.1 Site sensitivity indices

Firstly, for any input parameter and any instant, we consider a sensitivity map where each pixel represents the value of a given sensitivity index at this specific point. Figure 2 represents the temporal evolution of the sensitivity maps associated to the significantly influential input variables in the HSIC sense, with Gaussian kernel and σ calibration using the empirical standard deviation. The heights of release and deposition velocity are not represented because the estimated sensitivity indices are negligible what reveals that their uncertainty has no impact on the output variability. This figure is completed by Figure 3 representing these results averaged over the spatial domain (only the spatial points where the empirical output variance is significant are considered). After 20 minutes, the source term activity (STA) has an important impact (around 25%) on the variability of the plume central part, while the wind direction (WD) highly impacts bottom and top sides of the plume and the wind speed (WS) explains almost the totality of the uncertainty localized at the right extremity. Logically, wind conditions play an important role in the plume shape, WD and WS focusing on the plume width and length respectively. When the time passes, STA still has an almost exclusive influence in the central part but also to the right side which is opposed to the wind origin direction, while the wind conditions highly impact top and bottom parts (between 35% and 70%). Finally, these results are qualitatively similar to those obtained with the total Sobol' indices in Marrel et al [24], but the latter give a higher weight to the wind direction. These quantitative differences between dependence measures and variance-based indices have been highlighted by De Lozzo and Marrel [8] on test functions comprising linear, monotonous or non-monotonous effects.

Figures 4 and 5 show that the RDC leads to conclusions qualitatively closer to the HSIC's ones than to the Sobol' ones. Moreover, HR1, HR2 and DV have local contributions to the output variability up to 20% in terms of RDC, which are greater than in terms of HSIC or Sobol' indices. The associated sensitivity maps are not represented in this paper for clarity reasons and because they are significantly less influential than STA, WD and WS. The qualitative similarity can be explained by the fact that both indices are dependence measures based on transformations of the model parameters. Out of the scope of this paper, complementary studies about the RDC should be realized for a better use and comprehension of this dependence measure in GSA. Particularly, it seems appropriate for a screening purpose, due to its definition and to many advantages such as a low computational cost and the intrinsic possibility to deal with vectorial input and output parameters.

4.2 Block sensitivity indices

Now, we turn to a block sensitivity analysis where the interest is to summarize the contribution of an input parameter to the output uncertainty by a single real scalar, rather than by a sensitivity map. Concretely, we compare the following sensitivity indices:

- the generalized Sobol' indices for the whole output (see Expression (3));
- the HSIC-based sensitivity indices for the whole output (see Expressions (4-5)), with a Gaussian kernel function whose hyperparameters σ are calibrated by empirical standard deviations or the medians;
- the distance correlation (a particular HSIC with a non-parametric kernel, rather than a Gaussian one);
- the RDC applied to the first principal components (PCs) of the output proper orthogonal decomposition.

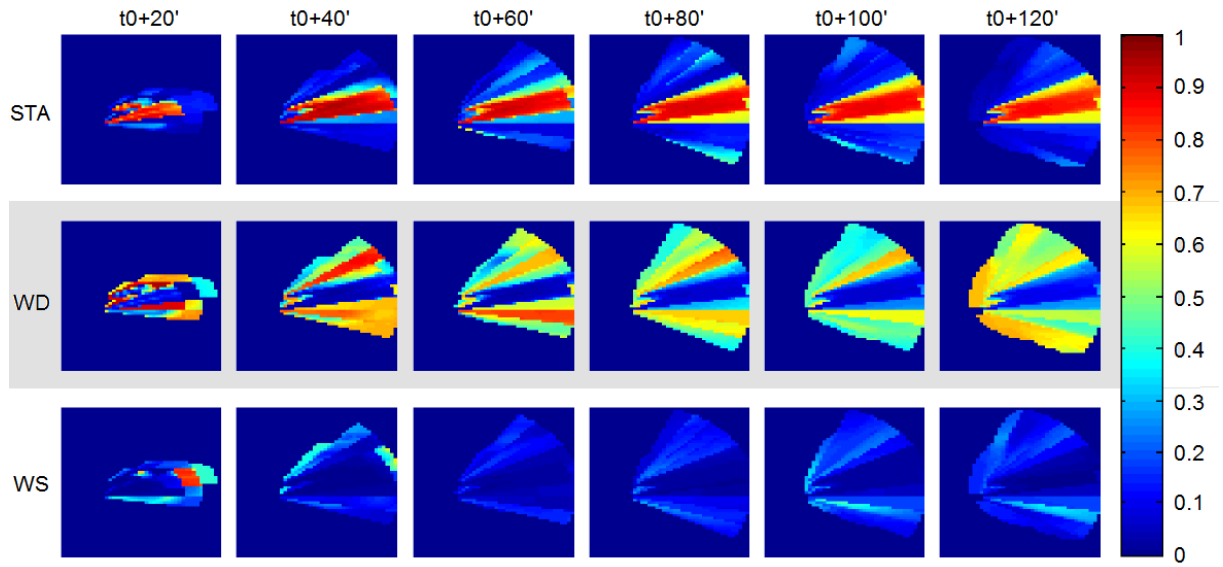


Figure 2: Temporal evolution of normalized site HSIC for the significantly influential inputs.

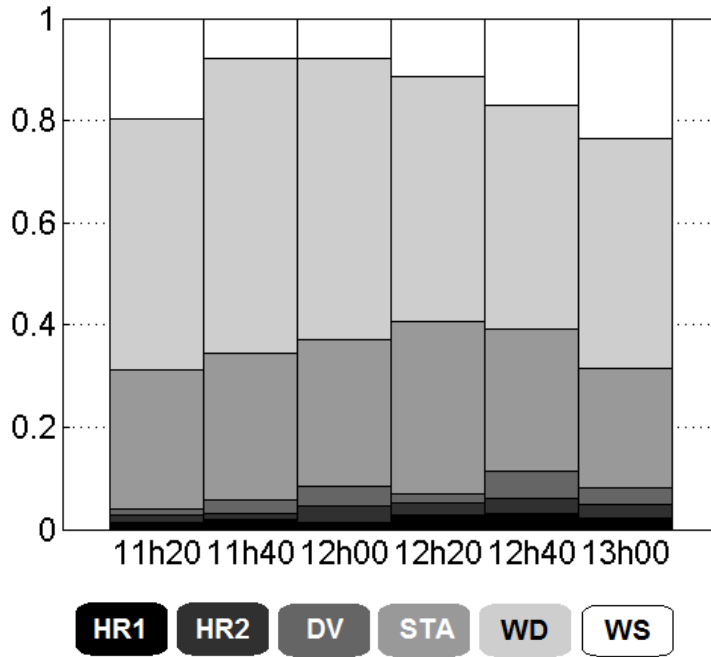


Figure 3: Temporal evolution of normalized site HSIC averaged over the spatial domain.

4.2.1 Spatial consideration

Figure 6 represents the time evolution of the block sensitivity indices based on HSIC, RDC, distance correlation (a particular HSIC without hyperparameter; see [7]) and Sobol' decomposition for the whole model output. Firstly, all measures agree on the dominance of the wind direction and source term activity. Secondly, Sobol' indices and distance correlations lead to similar conclusions which do not vary with time: the WD contribution to the output variability is between 60% and 80%, the STA one between 15% and 20% and the WS one between 5% and 20%. Then, the HSIC gives different results according to the choice of the hyperparameter $\hat{\sigma}$, except that the group STA-WD explains more than 80% of the output behavior (up to 90% for some instants) and the group HR1-HR2-DV has almost no effect on the output uncertainty. When the empirical median is used and following the timeline, the STA/WD contribution goes from 25%/60% after 20 minutes to 40%/50% at the final time. In other words, the wind direction is clearly the more significant input parameter

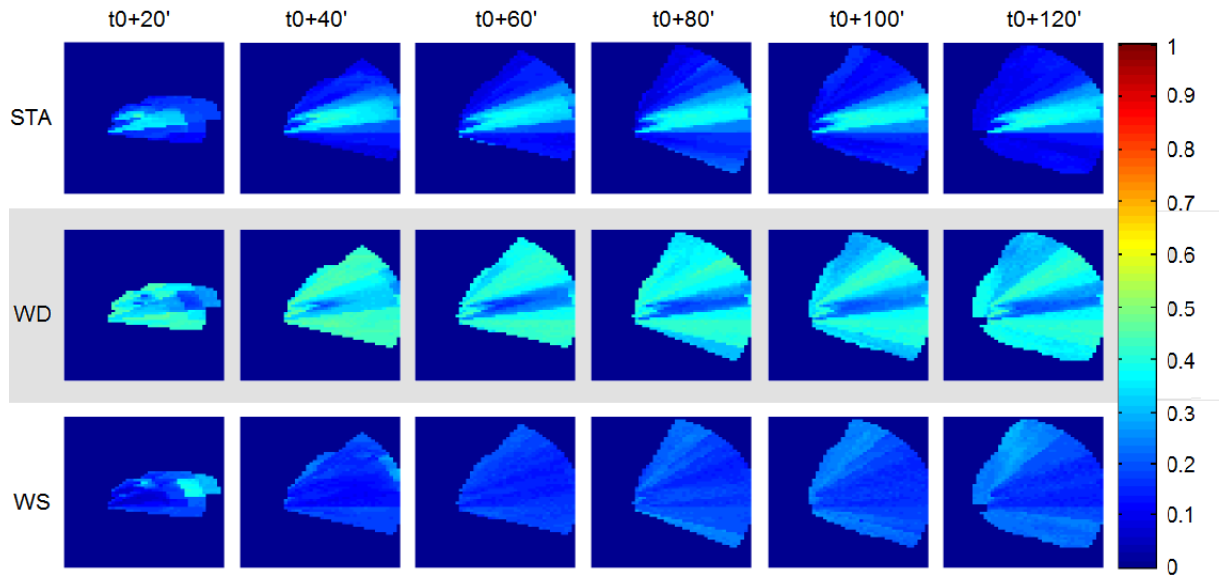


Figure 4: Temporal evolution of normalized site RDC for the significantly influential inputs.

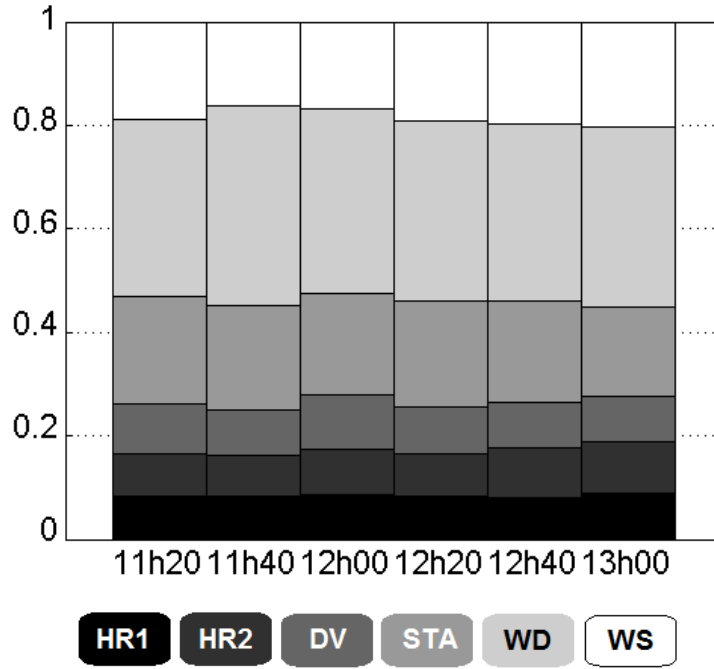


Figure 5: Temporal evolution of normalized site RDC averaged over the spatial domain.

at the beginning of the release while after 2 hours, its contribution is close to the source term activity one. When the empirical standard deviation is considered for σ , the wind component contributions slightly change over time while the other input uncertainties keep the same impacts on the output variability. We can see that different calibrations of σ lead to results slightly different. Particularly, the choice of the median seems to give a higher relative contribution to the less influential input parameters. Finally, according to the RDC and independently from the considered instant, wind components and source term activity explain almost 80% of the output variability (30% for STA, 30% for WD and 20% for WS) and the group HR1-HR2-DV explains a total of 20% of the output variability. These results are rather different from those obtained with the HSIC-based sensitivity measures or with the Sobol' indices and distance correlations. However, we can notice that the latter are closed to the RDC applied to the first component of the output POD, which explains more than 50% of the variability presented in the output of the learning sample used for the surrogate model construction. When the number of components increases, the RDC tends to the previous results for the whole

output. These differences of results can be due to the formulation of RDC which computes empirical copula transformations for each pixel of the output map or for each component of the POD. Further studies could be performed with advanced tools dealing with copula for spatial output.

To conclude, block HSIC and Sobol' sensitivity indices lead to consistent results but some quantitative differences occur. This is essentially due to the fact that these indices are different by definition, with Sobol' indices measuring the input contributions to the output variance and HSIC ones measuring the dependence between the input parameters and the output in a more general way.

4.2.2 Spatio-temporal consideration

Finally, we apply the sensitivity indices based on the HSIC and the RDC to the whole temporal scenario in order to summarize the input uncertainty contribution to the output variability by a single percentage rather than one by time step. We consider that these sensitivity indices are more appropriate for structured outputs, such as spatio-temporal variables, than the generalized Sobol' indices. For the HSIC approach, we simply extend the Gaussian kernel version (Expression (11)) dealing with the principal components of a spatial map to the spatio-temporal cases:

$$k_q(W, W') = \exp \left(- \sum_{t=1}^6 \sum_{k=1}^{q_t} \frac{\langle \mathbf{W}_t - \mathbf{W}'_t, \hat{\mathbf{v}}_{t,k} \rangle^2}{\hat{\sigma}^2} \right),$$

where $\hat{\sigma}$ is the empirical median computed from the sample $\left(\left| Y_{[t_0+t \times 20'], k}^{(i)} - Y_{[t_0+t \times 20'], k}^{(i')} \right| \right)_{1 \leq i, i' \leq n, 1 \leq t \leq 6, 1 \leq k \leq n_y}$ or the empirical standard deviation computed from the sample $\left(Y_{[t_0+t \times 20'], k}^{(i)} \right)_{1 \leq i, i' \leq n, 1 \leq t \leq 6, 1 \leq k \leq n_y}$. In this way, we obtain the results presented in Table 2. From this point of view, it is clear that the uncertainty on the wind direction and the source term activity plays an important role in the model output variability (between 35 and 50% for each of these input parameters). From this, characterization efforts should be oriented on these two variables in order to efficiently reduce the uncertainty on the simulator forecasts. To a lesser extent, the wind speed plays also a significant role. As is the case for site and block points of view, the use of RDC leads to different conclusions, with a greater weight for the variables HR1, HR2 and STA, even if WD, STA and WS remain the most influential parameters.

It is important to point out that the use of this sole all-in-one GSA is not recommended because it can hide very different input parameter contributions occurring at some instants of the plume temporal evolution. On the contrary, we advise this synthetic approach in addition to block GSAs (even also site GSAs) for the different instants.

	HSIC with median $\text{med}(Y - Y')$	HSIC with standard deviation $\text{std}(Y)$	RDC
HR1	3%	1%	9%
HR2	3%	1%	13%
DV	4%	1%	14%
STA	40%	35%	21%
WD	46%	49%	21%
WS	5%	12%	21%

Table 2: Normalized block sensitivity measures over the whole time and space variation domain.

5 Conclusion

This paper presents several sensitivity indices dealing with spatial model outputs. They are based either on dependence measures or on variance decompositions. More precisely, the dependence measures are function of the Hilbert-Schmidt Independence Criterion using a Gaussian kernel while the variance-based measures are the conventional Sobol' indices.

For the latter, a recent extension to spatial output is presented, based on output covariance matrix decomposition. A link is mentioned between both these generalized Sobol' indices and those of Lamboni et al. [19],

which aggregate elementary sensitivity analyses applied to the components of an output proper orthogonal decomposition.

Moreover, we mention that dependence and variance based sensitivity measures can be computed either for any single pixel of the output map, thereby producing spatial maps of site sensitivity indices, or for the whole output, leading to block sensitivity measures summarizing the contribution of the input parameter uncertainties to the entire spatial output variability. Both approaches are compared on an environmental application and appear to be complementary: the first one focuses principally on local perturbations due to the input variable uncertainties while the second one looks for the origin of the whole output uncertainty. In this way, site sensitivity measures advantageously highlight possible local contributions which are hidden by global measures. However, in a qualitative GSA with a screening purpose, spatial maps can be too informative, thereby making their comparison and the ranking of input variables difficult. In this case, block sensitivity measures become a useful tool that can be used to prioritize characterization efforts, for example. Through the industrial application and the comparison of the different indices (dependence measures and Sobol' indices) and their different versions (block and site ones), we give a first methodological guide for the global sensitivity analysis in a high dimensional context. Lastly, results obtained with variance-based and dependence measures are consistent but the second ones have the great advantage of requiring much less observation for the estimation step which can be carried out directly from the computer code, without any surrogate model.

Future research could focus on the randomized dependence coefficient in a GSA framework, this measure having proved to be encouraging from a CPU time point of view and from its ability to deal with vectorial output parameters. We have successfully applied it on the environmental study case and compared it with the other dependence measure based on HSIC, even if its interpretation is less straightforward.

References

- [1] Aronszajn, N.: Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**(3), 337–404 (1950)
- [2] Auder, B., Crecy, A.D., Iooss, B., Marquès, M.: Screening and metamodeling of computer experiments with functional outputs. application to thermal-hydraulic computations. *Reliability Engineering & System Safety* **107**, 122 – 131 (2012)
- [3] Balasubramanian, K., Sriperumbudur, B.K., Lebanon, G.: Ultrahigh dimensional feature screening via rkhs embeddings. In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, vol. 31, pp. 126–134 (2013)
- [4] Borgonovo, E.: A new uncertainty importance measure. *Reliability Engineering & System Safety* **92**(6), 771 – 784 (2007)
- [5] Campbell, K., McKay, M.D., Williams, B.J.: Sensitivity analysis when model outputs are functions. *Reliability Engineering & System Safety* **91**(10-11), 1468–1472 (2006)
- [6] Crestaux, T., Le Maître, O., Martinez, J.M.: Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety* **94**(7), 1161 – 1172 (2009)
- [7] Da Veiga, S.: Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation* **85**(5), 1283–1305 (2015)
- [8] De Lozzo, M., Marrel, A.: New improvements in the use of dependence measures for sensitivity analysis and screening. Submitted, hal-01090475 (2015)
- [9] Deza, M.M., Deza, E.: Encyclopedia of distances. In: *Encyclopedia of Distances*. Springer Berlin Heidelberg (2009)
- [10] Doury, A.: Pratiques françaises en matière de dispersion quantitative de la pollution atmosphérique potentielle liée aux activités nucléaires. In: *Proceedings of the seminar on radioactive releases and their dispersion in the atmosphere following a hypothetical reactor accident*, pp. 403–448. *RISØ*, vol I (1980)
- [11] Efron, B., Stein, C.: The jackknife estimate of variance. *The Annals of Statistics* **9**(3), 586–596 (1981)
- [12] Fang, K., Li, R., Sudjianto, A.: Design and modeling for computer experiments. *Computer science and data analysis series*. Chapman & Hall/CRC (2006)

- [13] Ferraty, F., Vieu, P.: Nonparametric functional data analysis : theory and practice. Springer series in statistics. Springer, New York (2006)
- [14] Fukumizu, K., Gretton, A., Lanckriet, G.R., Schölkopf, B., Sriperumbudur, B.K.: Kernel choice and classifiability for rkhs embeddings of probability distributions. In: Advances in Neural Information Processing Systems 22, pp. 1750–1758 (2009)
- [15] Gamboa, F., Janon, A., Klein, T., Lagnoux, A.: Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics* **8**(8), 575–603 (2014)
- [16] Ghanem, R.G., Spanos, P.D.: Stochastic finite elements: a spectral approach. Springer-Verlag, New York (1991)
- [17] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Proceedings Algorithmic Learning Theory, pp. 63–77. Springer-Verlag (2005)
- [18] Johnson, M.E., Moore, L.M., Ylvisaker, D.: Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**(2), 131–148 (1990)
- [19] Lamboni, M., Monod, H., Makowski, D.: Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety* **96**(4), 450 – 459 (2011)
- [20] Lopez-Paz, D., Hennig, P., Schölkopf, B.: The randomized dependence coefficient. In: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) Advances in Neural Information Processing Systems 26, pp. 1–9. Curran Associates, Inc. (2013)
- [21] Marrel, A., Iooss, B., Jullien, M., Laurent, B., Volkova, E.: Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics* **22**, 383–397 (2011)
- [22] Marrel, A., Iooss, B., Laurent, B., Roustant, O.: Calculations of sobol indices for the gaussian process metamodel. *Rel. Eng. & Sys. Safety* **94**(3), 742–751 (2009)
- [23] Marrel, A., Marie, N., De Lozzo, M.: Advanced surrogate model and sensitivity analysis methods for sodium fast reactor accident assessment. *Reliability Engineering & System Safety* **138**, 232–241 (2015)
- [24] Marrel, A., Perot, N., Mottet, C.: Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators. *Stochastic Environmental Research and Risk Assessment* **29**(3), 959–974 (2015)
- [25] McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**(2), 239–245 (1979)
- [26] Monfort, M., Patryl, L., P, A.: Presentation of the ceres platform used to evaluate the consequences of the emissions of radionuclides in the environment. *HARMO* pp. 1–4 (2010)
- [27] Morris, M.D.: Factorial sampling plans for preliminary computational experiments. *Technometrics* **33**(2), 161–174 (1991)
- [28] Pujol, G., Iooss, B., with contributions from Paul Lemaitre, A.J., Gilquin, L., Gratiot, L.L., Touati, T., Ramos, B., Fruth, J., Veiga, S.D.: sensitivity: Sensitivity Analysis (2014). URL <http://CRAN.R-project.org/package=sensitivity>. R package version 1.10.1
- [29] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005)
- [30] Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators. *Statistical Science* **4**, 409–435 (1989)
- [31] Saint-Geours, N.: Sensitivity analysis of spatial models: application to cost-benefit analysis of flood risk management plans. Ph.D. thesis, Université Montpellier II - Sciences et Techniques du Languedoc (2012)
- [32] Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S.: Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications* **181**(2), 259–270 (2010)

- [33] Sobol, I.: Sensitivity estimates for nonlinear mathematical models. *MMCE* **1**, 407–414 (1993)
- [34] Sobol, I., Gresham, A.: On an alternative global sensitivity estimators. In: *Proceedings of SAMO 1995*, pp. 40–42. Belgirate (1995)
- [35] Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**(6), 2769–2794 (2007)
- [36] Volkova, E., looss, B., Van Dorpe, F.: Global sensitivity analysis for a numerical model of radionuclide migration from the rrc kurchatov institute radwaste disposal site. *Stochastic Environmental Research and Risk Assessment* **22**(1), 17–31 (2008)

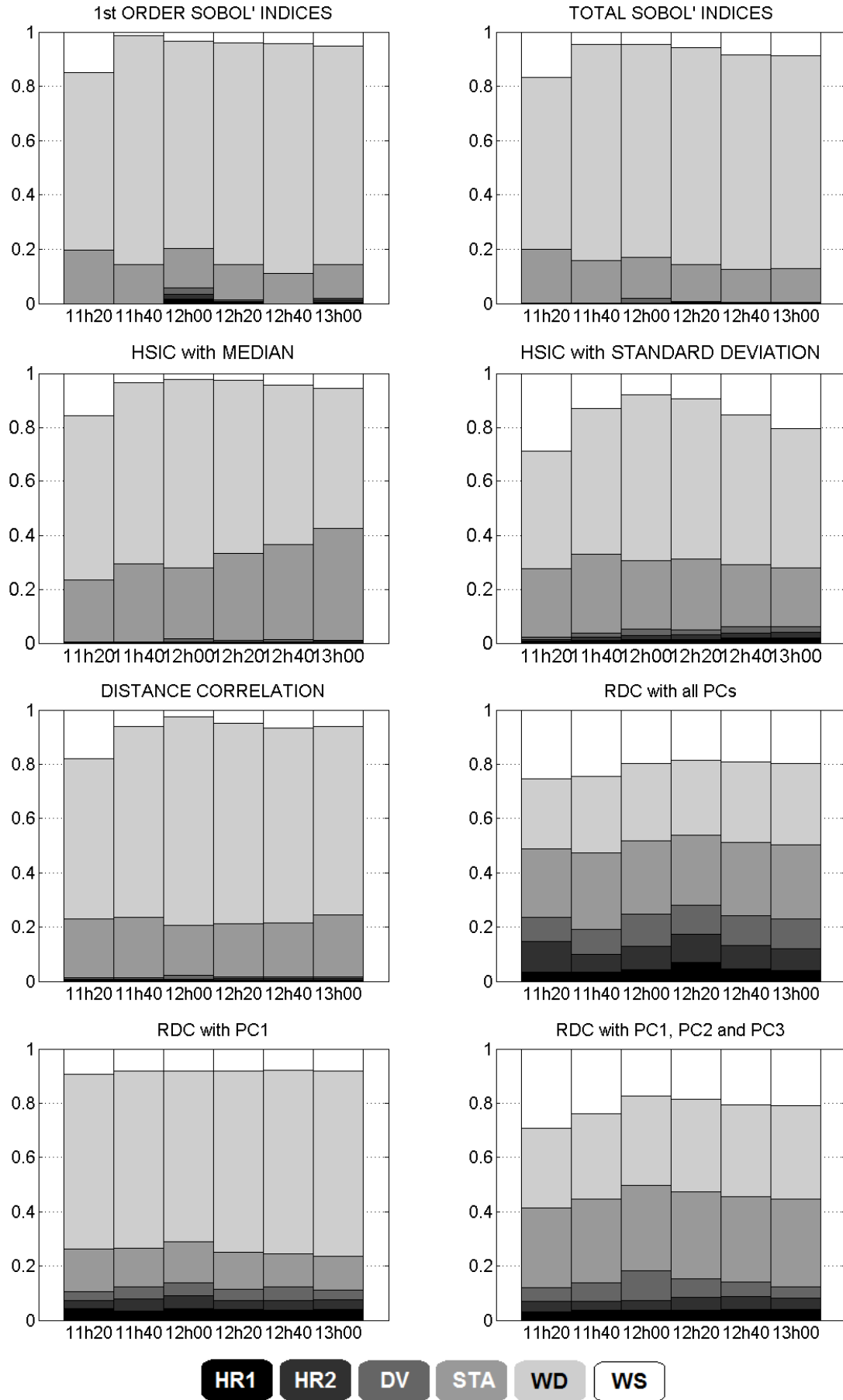


Figure 6: Temporal evolution of normalized block sensitivity measures.