

Global sensitivity analysis for models with spatially dependent outputs

Amandine Marrel, Bertrand Iooss, Michel Jullien, Béatrice Laurent, Elena Volkova

▶ **To cite this version:**

Amandine Marrel, Bertrand Iooss, Michel Jullien, Béatrice Laurent, Elena Volkova. Global sensitivity analysis for models with spatially dependent outputs. Environmetrics, 2011, 22, pp.383-397. hal-00430171v4

HAL Id: hal-00430171
https://hal.science/hal-00430171v4

Submitted on 22 Sep 2010

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.

Global sensitivity analysis for models with spatially dependent outputs

Amandine MARREL[1], Bertrand IOOSS[2], Michel JULLIEN[3], Béatrice LAURENT[4] and Elena VOLKOVA[5]

[1] IFP, 92852 Rueil-Malmaison cedex, France
[2] EDF R&D, 6 Quai Watier, 78401 Chatou, France
[3] CEA, DEN, F-13108 Saint Paul lez Durance, France
[4] Institut de Mathématiques de Toulouse (UMR 5219), INSA de Toulouse, Université de Toulouse, France
[5] RRC "Kurchatov Institute", Institute of Nuclear Reactors, Russia

Corresponding author: B. Iooss ; Email: biooss@yahoo.fr
Phone: +33 (0)1 30 87 79 69

Abstract

The global sensitivity analysis of a complex numerical model often calls for the estimation of variance-based importance measures, named Sobol' indices. Metamodel-based techniques have been developed in order to replace the cpu time-expensive computer code with an inexpensive mathematical function, which predicts the computer code output. The common metamodel-based sensitivity analysis methods are well-suited for computer codes with scalar outputs. However, in the environmental domain, as in many areas of application, the numerical model outputs are often spatial maps, which may also vary with time. In this paper, we introduce an innovative method to obtain a spatial map of Sobol' indices with a minimal number of numerical model computations. It is based upon the functional decomposition of the spatial output onto a wavelet basis and the metamodeling of the wavelet coefficients by the Gaussian process. An analytical example is presented to clarify the various

steps of our methodology. This technique is then applied to a real hydrogeological case: for each model input variable, a spatial map of Sobol' indices is thus obtained.

Keywords: Computer experiment, Gaussian process, metamodel, functional data, radionuclide migration.

Short title: Spatial global sensitivity analysis

1 INTRODUCTION

Today, in different environments, there are sites with groundwater contaminated because of an inappropriate handling or disposal of hazardous materials or waste. Such environmental or sanitary issues require the development of treatment or remediation strategies and, in all cases, a robust long-term prediction of behaviour. The indispensable simulation of global fluxes, such as water or pollutants, through the different environmental compartments involves many parameters. Numerical modeling is an efficient tool for an accurate prediction of the spreading of the contamination plume and an assessment of environmental risks associated to the site. However, it is well known that many input variables, such as hydrogeological parameters (permeabilities, porosities, etc.) or boundary and initial conditions (contaminant concentrations, aquifer level, etc.), are highly uncertain in the complex numerical models. A systematic and exhaustive 3D characterization of sites is still impossible.

To deal with all these uncertainties, computer experiment methodologies based upon statistical techniques are useful. For instance, we assume that $Y = f(\mathbf{X})$ is the real-valued output of a computer code f. Its input variables are random and modeled by the random vector $\mathbf{X} = (X_1, \ldots, X_d) \in \mathcal{X}$, $\mathcal{X}$ being a bounded domain of $\mathbb{R}^d$, of known distribution. The uncertainty analysis step is used to evaluate statistical parameters, confidence intervals or the density probability distribution of the model response (De Rocquigny et al.,

2008), while the global sensitivity analysis step is used to quantify the influence of the uncertainties of the model input variables (in their whole range of variations) on model responses (Saltelli et al., 2000). Recent studies have applied different statistical methods of uncertainty and sensitivity analysis to environmental models (Helton, 1993; Nychka et al., 1998; Fassò et al., 2003; Volkova et al., 2008; Lilburne and Tarantola, 2009). All these methods have shown their efficiency in providing guidance to a better understanding of the modeling.

However, for the purpose of sensitivity analysis, four main difficulties can arise due to practical problems, especially when focusing on environmental risks:

P1) physical models involve rather complex phenomena (they are non linear and subject to threshold effects) sometimes with strong interactions between physical variables

P2) computer codes are often too cpu time expensive to evaluate a model response, from several minutes to weeks

P3) numerical models take as inputs a large number of uncertain variables (typically $d > 10$)

P4) the outputs of these numerical encompass many variables of interest, that can vary in space and time.

The first problem P1 is solved using variance-based measures. These ones can handle non-linear and non-monotonic relationships between inputs and output (Saltelli et al., 2000). These measures are based upon the functional ANOVA decomposition of any integrable function f (Efron and Stein, 1981) and determine how to share of the variance of the output resulting from a variable X_i or an interaction between variables (Sobol, 1993):

$$S_i = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)} \ , \quad S_{ij} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j\right)\right]}{\mathrm{Var}(Y)} - S_i - S_j \ , \quad S_{ijk} = \ldots \tag{1}$$

The interpretation of these coefficients, namely the Sobol' indices, is natural as all indices lie in $[0, 1]$ and their sum is one in the case of independant input variables. The larger the index value, the greater the importance of the variable related to this index. To express the overall output sensitivity to an input X_i, Homma and Saltelli (1996) introduce the total sensitivity index:

$$S_{T_i} = S_i + \sum_{i<j} S_{ij} + \sum_{i<j<k} S_{ijk} + \ldots = \sum_{l \in \#i} S_l = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_{\sim i}\right)\right]}{\mathrm{Var}(Y)} \tag{2}$$

where $\#i$ represents all the "non-ordered" subsets of indices containing index i and $X_{\sim i}$ is the vector of all inputs except X_i. Thus, $\sum_{l \in \#i} S_l$ is the sum of all the sensitivity indices with index containing i.

Unfortunately, the traditional or advanced Monte Carlo methods, which are used to estimate first order and total Sobol' indices, require a large number of model evaluations (Saltelli et al., 2010). To overcome the problem P2, of too long a calculation time, and make uncertainty and sensitivity analysis tractable, various approaches based upon meta-modeling were recently proposed (Koehler and Owen, 1996; Kleijnen and Sargent, 2000; Oakley and O'Hagan, 2002). The key point consists of replacing the complex computer code by a mathematical approximation, called a metamodel, which is fitted from only a few experiments. The metamodel reproduces the behavior of the computer code in the domain of its influential parameters (Sacks et al., 1989; Fang et al., 2006). Among all the metamodel-based solutions (polynomials, splines, neural networks, etc.), we focus our attention on the Gaussian process (Gp) model. It can be viewed as an extension of the kriging method, which is used for interpolating data in space (Chilès and Delfiner, 1999), to computer code data (Sacks et al., 1989; Oakley and O'Hagan, 2002). Many authors (e.g. Welch et al., 1992; Marrel et al., 2008) have shown how the Gp model can be used as an efficient emulator of code responses, even in high dimensional cases (problem P3).

In this paper, we consider models subject to the four problems together (P1, P2, P3 and P4), which is an usual cas in model-based environmental studies. We mainly pay attention to problem P4, that is the possible high dimension of model outputs. In the application case studied in this paper, the costly numerical model yields spatial concentration maps. These spatial outputs encompass several thousands of grid blocks, each with a concentration value. This kind of problem cannot be tuned to a vectorial output problem because of its dimensionality: the metamodeling of this vectorial output cannot be solved referring to kriging or cokriging techniques (Fang et al., 2006). Therefore, we consider the model output as a functional output synthesized by its projection on an appropriate basis. This problem of building a metamodel (based upon functional decomposition and Gp modeling) for a functional output has recently been addressed for one-dimensional outputs by Shi et al. (2007) and Bayarri et al. (2007) and for two-dimensional outputs by Higdon et al. (2008).

In the case of sensitivity analysis, a functional output is usually considered as a vectorial output and sensitivity indices relative to each input are computed for each discretized value of the output (De Rocquigny et al., 2008). To avoid the large amount of sensitivity index computations when applying such an approach, a few authors referred to various basis decompositions on the functional output, such as the principal component analysis (Campbell et al., 2006; Lamboni et al., 2009). Then, sensitivity indices are obtained for the coefficients of the expansion basis.

However, the full functional restitution of Sobol' indices remains an unexplored challenge. In this paper, we propose an original and complete methodology to compute Sobol' indices at each location of the spatial output map. Our approach consists of building a metamodel based upon wavelet decomposition as in Bayarri et al. (2007) (restricted to the case of a temporal output). This metamodel is then used to compute spatial Sobol' index maps (one map per input variable). A map of the Sobol' index of an input X_i shows the

local and global influences of this input on the output. It can help to better understand the computer code results and can used to reduce more efficiently the uncertainties in the responses. Thus, to reduce the output variability at a given point of the map, we analyze all Sobol' maps and determine the most influential inputs. Then, we can try to reduce the uncertainty of these inputs by accounting for additional measures. In addition, the global influence of each input over the whole space can be investigated to identify areas of influence and non-influence for this input.

Details about the Gp metamodel are given in the following section. Then, a step by step description of our methodology is given in Section 3. A synthetic test function is used to evidence the relevance of our choices and estimate the convergence of the algorithms. Section 4 presents how our methodology is applied to a real environmental problem, which calls for the modeling of radionuclide groundwater migration (MARTHE code). Then, a few points are discussed at the end of this paper.

2 GAUSSIAN PROCESS METAMODELING

This section introduces the Gp metamodel for the case of a single scalar output. We consider n realizations of a computer code. Each realization $y = f(\mathbf{x}) \in \mathbb{R}$ is an output of the computer code and corresponds to a d-dimensional input vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}$. The n points corresponding to the code runs are called the experimental design and are denoted as $X_s = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$. The outputs are denoted as $Y_s = (y^{(1)}, \ldots, y^{(n)})$ with $y^{(i)} = y(\mathbf{x}^{(i)}) \ \forall \ i = 1..n$. Gp modeling treats the deterministic response $y(\mathbf{x})$ as a realization of a random function $Y_{\text{Gp}}(\mathbf{x})$. This includes a regression part and a centered stochastic process (Sacks et al., 1989). It can be written as:

$$Y_{\text{Gp}}(\mathbf{x}) = f_0(\mathbf{x}) + Z(\mathbf{x}) \ . \tag{3}$$

The deterministic function $f_0(\mathbf{x})$ provides the mean approximation of the computer code. In our study, we use a one-degree polynomial model with $f_0(\mathbf{x})$ written as:

$$f_0(\mathbf{x}) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j \ ,$$

where $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_k]^t$ is the regression parameter vector. It has been shown, for example in Martin and Simpson (2005) and Marrel et al. (2008), that such a function is sufficient, and sometimes necessary, to capture the global trend of the computer code.

The stochastic part $Z(\mathbf{x})$ is a Gaussian centered process fully characterized by its covariance function: $\mathrm{Cov}(Z(\mathbf{x}), Z(\mathbf{u})) = \sigma^2 R(\mathbf{x}, \mathbf{u})$, where σ^2 is the variance of Z and R the correlation function. For simplicity, we consider a stationary process $Z(\mathbf{x})$, which means that correlation between $Z(\mathbf{x})$ and $Z(\mathbf{u})$ is a function of the distance between $\mathbf{x}$ and $\mathbf{u}$. Our study focuses on a particular family of correlation functions that can be written as a product of one-dimensional correlation functions R_l:

$$\mathrm{Cov}(Z(\mathbf{x}), Z(\mathbf{u})) = \sigma^2 R(\mathbf{x} - \mathbf{u}) = \sigma^2 \prod_{l=1}^{d} R_l(x_l - u_l).$$

This form of correlation function is particularly well-suited to simplify mathematical developments in analytical uncertainty and sensitivity analyses (Marrel et al., 2009). More precisely, we use the generalized exponential correlation function:

$$R_{\boldsymbol{\theta},\boldsymbol{p}}(\mathbf{x} - \mathbf{u}) = \prod_{l=1}^{d} \exp(-\theta_l |x_l - u_l|^{p_l}),$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^t$ and $\boldsymbol{p} = [p_1, \ldots, p_d]^t$ are the correlation parameters (also called hyperparameters) with $\theta_l \geq 0$ and $0 < p_l \leq 2 \ \ \forall\, l = 1..d$. This choice is motivated by the wide spectrum of shapes that such a function offers.

If a new point $\mathbf{x}^* = (x_1^*, \ldots, x_d^*) \in \mathcal{X}$ is considered, we obtain the following predictor

and variance formulas:

$$\mathbb{E}[Y_{\text{Gp}}(\mathbf{x}^*)] = f_0(\mathbf{x}^*) + \boldsymbol{k}(\mathbf{x}^*)^t \boldsymbol{\Sigma}_s^{-1}(Y_s - f(X_s)) \ , \tag{4}$$

$$\text{Var}[Y_{\text{Gp}}(\mathbf{x}^*)] = \sigma^2 - \boldsymbol{k}(\mathbf{x}^*)^t \boldsymbol{\Sigma}_s^{-1} \boldsymbol{k}(\mathbf{x}^*) \ , \tag{5}$$

with Y_{Gp} denoting $(Y|Y_s, X_s, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{p})$,

$$\begin{aligned} \boldsymbol{k}(\mathbf{x}^*) &= [\text{Cov}(y^{(1)}, Y(\mathbf{x}^*)), \ldots, \text{Cov}(y^{(n)}, Y(\mathbf{x}^*))]^t \\ &= \sigma^2 [R_{\boldsymbol{\theta},\boldsymbol{p}}(\mathbf{x}^{(1)} - \mathbf{x}^*), \ldots, R_{\boldsymbol{\theta},\boldsymbol{p}}(\mathbf{x}^{(n)} - \mathbf{x}^*))]^t \end{aligned}$$

and the covariance matrix

$$\boldsymbol{\Sigma}_s = \sigma^2 \left(R_{\boldsymbol{\theta},\boldsymbol{p}} \left(\mathbf{x}^{(i)} - \mathbf{x}^{(j)} \right)_{i=1..n, j=1..n} \right) \ .$$

Regression and correlation parameters $\boldsymbol{\beta}$, σ, $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are usually estimated by maximizing likelihood functions (Fang et al., 2006). This optimization problem can be badly conditioned and difficult to solve in high dimensional cases ($d > 5$). Welch et al. (1992) and Marrel et al. (2008) developed algorithms to build Gp metamodels on outputs that have a non-linearity depending on quite a large number of input variables.

The conditional mean (Eq. (4)) is used as a predictor. The variance formula (Eq. (5)) corresponds to the mean squared error (MSE) of this predictor and is also known as the kriging variance. This analytical formula for MSE gives a local indicator of the prediction accuracy. More generally, the Gp model provides an analytical formula for the distribution of the output variable at any arbitrary new point. This distribution formula can be used to develop analytical formula for uncertainty and sensitivity analyses (Oakley and O'Hagan, 2002, 2004). Studying several test functions and one industrial application, Marrel et al. (2009) showed that this analytical approach is efficient to compute the first

order Sobol' indices S_i (Eq. (1)). In addition, it provides confidence intervals for the estimates. However, the analytical approach does not yield any direct estimation of the total Sobol' indices S_{T_i} (Eq. (2)) and deals only with uncorrelated inputs.

3 METHODOLOGY FOR A SPATIAL OUPUT

In this section, we describe the methodology that we use to compute spatial Sobol' index maps (first proposed in Marrel, 2008). We also apply this methodology to an analytical function in order to study the convergence of the algorithms.

3.1 General principles

For a given $\mathbf{x}^*$ value of vector $\mathbf{X} = (X_1, \ldots, X_d)$, the code output is now a deterministic function $y(\mathbf{x}^*, \mathbf{z})$ where $\mathbf{z}$ denotes a vector of dimension p of spatial coordinates. In this paper, we focus on two-dimensional cases. Thus, the target outputs are two-dimensional maps. Thus, $\mathbf{z}$ varies in a grid on a compact set D_z of $\mathbb{R}^2$ and corresponds to an index for the outputs. Variables $\mathbf{X}$ and $\mathbf{z}$ are of very distinct natures: variables $X_1, \ldots, X_d$ which correspond to the inputs of the computer code, are random. They are different for each simulation of the code. Our objective is to perform a sensitivity analysis with respect to these variables. Variables $\mathbf{z}$ are deterministic and vary on a grid of size n_z which corresponds to a discretization of D_z. The grid is the same for each simulation of the code and the output corresponds to the n_z values $y(\mathbf{x}^*, \mathbf{z})$ for $\mathbf{z}$ describing the grid. For example, the MARTHE model described in Section 4 has $d = 20$ input variables and yields at each simulation a map with $n_z = 64 \times 64 = 4096$ points.

Because of the different natures of variables $\mathbf{X}$ and $\mathbf{z}$, the dependency of the output with respect to these two variables is represented from two different ways. For a fixed value of $\mathbf{X}$, we use a projection of map $\mathbf{z} \mapsto Y(\mathbf{X}, \mathbf{z})$, onto an orthonormal wavelet basis.

The coefficients of the projection depend on $\mathbf{X}$. We select the coefficients with the largest variance and model these coefficients with respect to the d-dimensional input variable $\mathbf{X}$. In most applications, the dimension of $\mathbf{X}$ is quite large and each simulation of the code is time-expensive. Therefore, we need a method able to deal with a limited number of simulations and imput vectors of large dimension. In addition, the relationship between the input variables and the coefficients is expected to be highly non-linear. We therefore use the Gp metamodel, described in the previous section, to model the dependency of each selected coefficient with respect to $\mathbf{X}$.

Therefore, for a given input design $X_s = \left(\mathbf{x}^{(i)}\right)_{i=1..n}$ and the n corresponding simulations of the map $(y(\mathbf{x}^{(i)}, \mathbf{z}_j), j = 1, \ldots, n_z)$, $i = 1, \ldots, n$, the three main steps of the method are:

1. Decomposition of the maps $\mathbf{z} \mapsto y(\mathbf{x}^{(i)}, \mathbf{z})$ onto a two-dimensional wavelet basis

2. Selection of the coefficients with the largest variance

3. Modeling of the coefficients with respect to the input variables using a Gp.

At each step, we use various criteria to evaluate the performance of our procedures. We are then able to predict a map $(y(\mathbf{x}^*, \mathbf{z}_j), j = 1 \ldots, n_z)$ for a new value of the input vector $\mathbf{x}^*$. Of course, this method of map prediction (which we call a functional metamodel or also, in our case, a spatial metamodel) has the advantage, compared to the simulation of the code, to be much less time-expensive.

Finally, our functional metamodel allows us to produce maps of sensitivity analysis based upon Sobol' indices by using Monte Carlo methods (see introduction, Eqs. (1) and (2)). As mentioned previously, the direct use of the computer code is impossible because of the required number of function evaluations. This study is restricted to the estimation of the first order indices $S_i(\mathbf{z})$ and total Sobol' indices $S_{T_i}(\mathbf{z})$ for $\mathbf{z} \in D_z$. These two

indices allow us to quantify the individual and total influence for each input. Then, the degree of interaction with other inputs can then be deduced.

3.2 An analytical test case: The Campbell2D function

The analytical function used in this section to perform various tests is inspired by Campbell et al. (2006) who considered a function with four inputs and a one-dimensional output. It was converted to a function with eight inputs ($d = 8$) and a two-dimensional output ($\mathbf{z} = (z_1, z_2)$):

$$Y = g(\mathbf{X}, z_1, z_2) = X_1 \exp\left[-\frac{(0.8z_1 + 0.2z_2 - 10X_2)^2}{60X_1^2}\right] + (X_2 + X_4)\exp\left[\frac{(0.5z_1 + 0.5z_2)X_1}{500}\right] + X_5(X_3 - 2)\exp\left[-\frac{(0.4z_1 + 0.6z_2 - 20X_6)^2}{40X_5^2}\right] + (X_6 + X_8)\exp\left[\frac{(0.3z_1 + 0.7z_2)X_7}{250}\right], \quad (6)$$

where $(z_1, z_2) \in [-90, 90]^2$ represent azimuthal and polar spatial coordinates and $X_i \sim \mathcal{U}[-1, 5]$ for $i = 1 \dots 8$. This function, called the Campbell2D function, gives a spatial map as output (Figure 1). The Campbell2D function has been calibrated in order to give strong spatial heterogeneities, sometimes with sharp boundaries, and very different spatial distributions of the output values according to the $\mathbf{X}$ values.

For the Campbell2D function, it is possible to calculate the first order Sobol' indices $S_i(z_1, z_2)$. Appendix A gives the results of these integrations. The resulting analytical expressions (Eqs. (16) to (23)) provide the exact solutions of the first order Sobol' indices. However, analytical calculations of the total Sobol' indices $S_{T_i}(z_1, z_2)$ (Eq. (2)) are not possible. We estimate $S_{T_i}(z_1, z_2)$, $i = 1, \dots, 8$, by using Saltelli's Monte Carlo algorithm (Saltelli, 2002) with $N = 10^5$. Thus, the Campbell2D function was computed $N(d+2) = 10^6$ times. The estimated errors with such large sample sizes are of the order of 5×10^{-3} (standard deviation estimated via bootstrap). These estimates $S_{T_i}(z_1, z_2)$ are henceforth called exact total Sobol' indices.

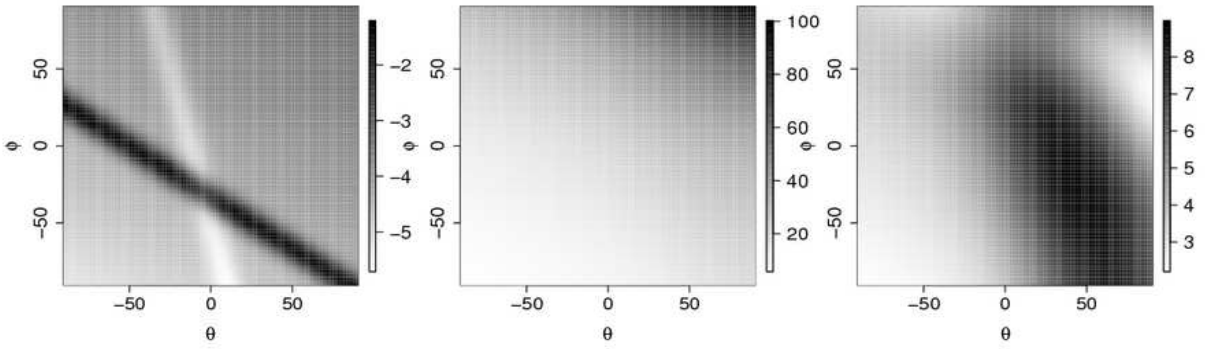


Figure 1: Three different output maps from the Campbell2D function: $\mathbf{x} = (-1,-1,-1,-1,-1,-1,-1,-1)$ (left); $\mathbf{x} = (5,5,5,5,5,5,5,5)$ (center); $\mathbf{x} = (5,3,1,-1,5,3,1,-1)$ (right).

Figure 2 gives the maps of the total Sobol' index estimations. Input X_5 has no influence on the output of the Campbell2D function. Input X_1 has a small influence on the output of the Campbell2D function. Input X_3 has a mild influence in a diagonal axis of the spatial domain. Inputs X_4 and X_8 have mild influences in a large part of the spatial domain. Inputs X_2, X_6 and X_7 have strong influences in different parts of the spatial domain (located in corners for X_2 and X_7). Moreover, the first order Sobol' indices (maps not shown here) for X_3, X_5, X_6 and X_7 are far from the total Sobol' indices. As shown by formula (6), these four variables have some strong interactions (interactions between X_3, X_5 and X_6 and between X_6 and X_7).

3.3 Spatial metamodeling

The spatial metamodeling process is composed of 5 internal steps.

Step 0 - Preparation of the learning sample

When dealing with a large input dimension d, the choice of the input design $X_s = \left(\mathbf{x}^{(i)}\right)_{i=1..n}$ is very important, especially when n is small. For scalar computer model

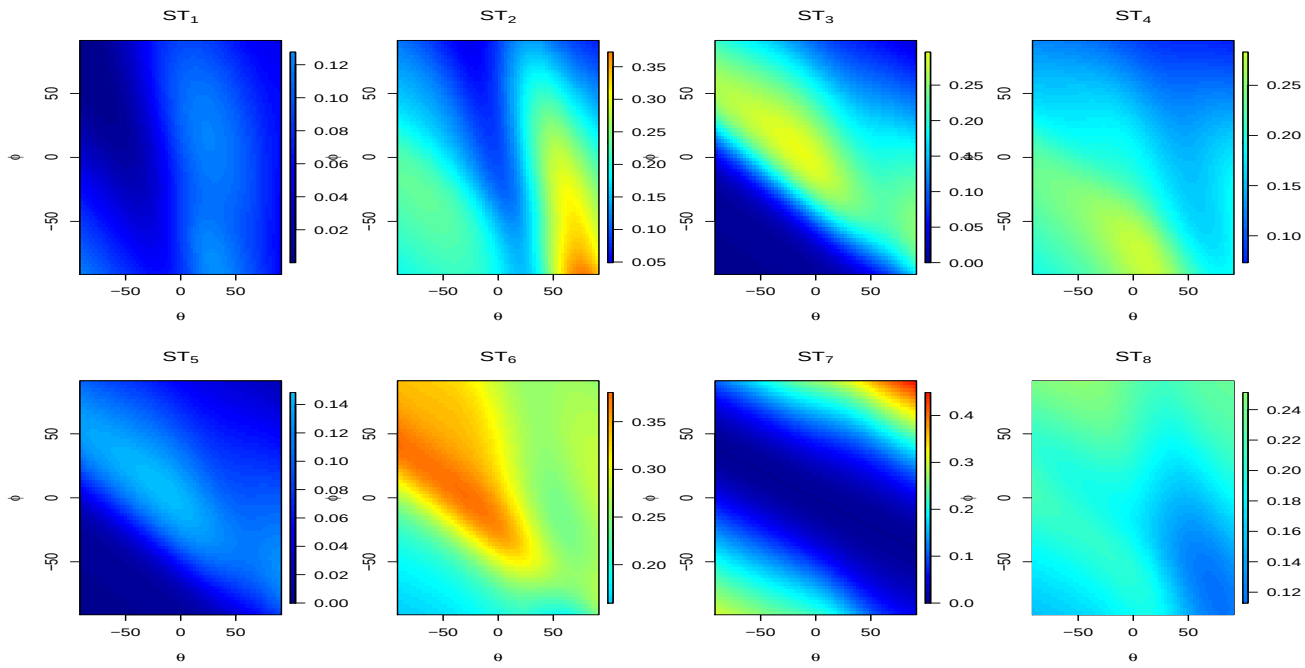


Figure 2: Total Sobol' indices of the 8 input variables of the Campbell2D function, estimated by Monte Carlo algorithm. The color scales are the same for all the plots.

output, numerous authors stressed the strong influence of the input design on the quality of the Gp modeling (Koehler and Owen, 1996; Fang et al., 2006). For instance, maximin Latin hypercube samples and low-discrepancy Latin hypercube samples were shown to provide good results (Marrel, 2008; Iooss et al., 2010). However, building good input designs for functional output still remains an open question which could be the subject of future work.

For our tests with the Campbell2D function, we use maximin Latin hypercube samples. Once the input design is defined, we obtain n simulations of the map $\left(y(\mathbf{x}^{(i)}, \mathbf{z})\right)_{i=1..n}$ by running the numerical model.

Step 1 - Spatial decomposition and selection of coefficients

The spatial decomposition of the output map is made on a basis of orthogonal functions

$\{\phi_j\}_{j\in\mathbb{N}^*}$:

$$Y(\mathbf{X},\mathbf{z}) = \mu(\mathbf{z}) + \sum_{j=1}^{\infty} \alpha_j(\mathbf{X})\phi_j(\mathbf{z}) \text{ with } \alpha_j(\mathbf{X}) = \int_{D_z} [Y(\mathbf{X},\mathbf{z}) - \mu(\mathbf{z})]\phi_j(\mathbf{z})d\mathbf{z} \ , \quad (7)$$

where $\mu(\mathbf{z}) = \mathbb{E}_{\mathbf{X}}[Y(\mathbf{X},\mathbf{z})]$. We define $Y_K(\mathbf{X},\mathbf{z})$ as the truncated decomposition at order K:

$$Y_K(\mathbf{X},\mathbf{z}) = \mu(\mathbf{z}) + \sum_{j=1}^{K} \alpha_j(\mathbf{X})\phi_j(\mathbf{z}) \ . \quad (8)$$

For the function basis, various wavelet bases can be considered (Haar, Daubechies, Symmlet and Coiflet, see Misiti et al., 2007) in order to optimize the compression of the local and global information. In the following tests, we use the Daubechies basis which offered the best results.

The selection of a small number k of coefficients $\alpha_j(\mathbf{X})$ to be modeled with Gp is essential. For instance, MARTHE maps (see Section 4) and Campbell2D maps contain $n_z = 64 \times 64 = 4096$ pixels, which leads to $K = 4096$ wavelet coefficients. Modeling such a number of Gp seems intractable because the building process of one Gp is CPU time consuming (Marrel et al., 2008). It is therefore necessary to model with Gp only the most informative coefficients. The criterion considered for selecting the coefficients involves their variance with respect to $\mathbf{X}$: priority is given to coefficients which explain at most the output map variability. Mathematically, the new order of the coefficients $\{\alpha_1,\ldots,\alpha_K\}$ is written $\left\{\alpha_{(1)},\ldots,\alpha_{(K)}\right\}$ following the inequalities

$$\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_{(1)}(\mathbf{x}^{(i)}) - \overline{\alpha_{(1)}}\right)^2 \geq \ldots \geq \frac{1}{n}\sum_{i=1}^{n}\left(\alpha_{(K)}(\mathbf{x}^{(i)}) - \overline{\alpha_{(K)}}\right)^2 \text{ with } \overline{\alpha_j} = \frac{1}{n}\sum_{i=1}^{n}\alpha_j(\mathbf{x}^{(i)}) \ . \quad (9)$$

The number k of Gp-modeled coefficients will be discussed in Step 3.

Step 2 - Modeling the coefficients

For $j = 1, \ldots, K$, the model $A_j(\mathbf{X})$ used for approximating the coefficient $\alpha_j(\mathbf{X})$ is one of the following models listed below:

- Model 1: the empirical mean: $A_j(\mathbf{X}) = \dfrac{1}{n}\sum_{i=1}^{n} \alpha_j\left(\mathbf{x}^{(i)}\right)$;
- Model 2: the linear regression model:

$$A_j(\mathbf{X}) = \beta_{0,j} + \sum_{l=1}^{d} \beta_{l,j}\, X_l \tag{10}$$

  fitted on the learning sample $\left(\mathbf{x}^{(i)}, \alpha_j(\mathbf{x}^{(i)})\right)_{i=1..n}$. We use an AIC selection process to keep only the significant terms in (10);

- Model 3: the Gp model of form (3) as described in Marrel et al. (2008). The deterministic part $f_0(\mathbf{X})$ is a linear regression model as in (10) with a selection process based on AICC (a modified AIC in order to take spatial correlations into account, see Hoeting et al. (2006)). The generalized exponential function is used for the correlation function $R(\cdot)$ of the stochastic part $Z(\mathbf{X})$. The building of this model is rather costly, especially in a high dimensional context ($d > 10$) because of the specific variable selection process proposed by Marrel et al. (2008).

In the following two steps, we compare three different methodologies in order to stress the benefit of an appropriate metamodel choice:

- Method 1: Model 3 for the k selected coefficients and model 1 for the other coefficients
- Method 2: Model 2 for the k selected coefficients and model 1 for the other coefficients
- Method 3: Model 3 for the k selected coefficients, model 2 for the k' following coefficients ($k' \gg k$) and model 1 for the $K - k - k'$ other coefficients. For the

Campbell2D function, setting k' to 500 is a heuristic choice based upon the observation that, in the case studied, the information in terms of variability is explained by 10% of coefficients. More generally, a convergence study can be made in order to find a suitable value for k'.

We now define $\widehat{Y}_{K,k}(\mathbf{X}, \mathbf{z})$ the approximation of $Y_K(\mathbf{X}, \mathbf{z})$ (Eq. (8)) using one of the three previous methods.

Several adequacy criteria can be used to measure the discrepancy between the function $Y(\mathbf{X}, \mathbf{z})$ and its approximation $\widehat{Y}_{K,k}(\mathbf{X}, \mathbf{z})$. We use the mean absolute error, the maximal error and the mean squared error but restrict our presentation to mean squared error results for the sake of consistency. The mean squared error $\mathrm{MSE}(\mathbf{X})$ is written

$$\mathrm{MSE}(\mathbf{X}) = \int_{D_z} \left[Y(\mathbf{X}, \mathbf{z}) - \widehat{Y}_{K,k}(\mathbf{X}, \mathbf{z}) \right]^2 d\mathbf{z} \; . \tag{11}$$

$\mathrm{MSE}(\mathbf{X})$ is estimated by integrating over the n_z grid. For a fixed value of $\mathbf{X}$, this criterion measures the restitution quality in the mean of the overall map. We denote by MSE the expectation (with respect to the variable $\mathbf{X}$) of $\mathrm{MSE}(\mathbf{X})$. When it is possible, we provide new simulations of the map $Y(\mathbf{X}, \mathbf{z})$ for randomized values of $\mathbf{X}$, and we use this test sample to estimate the MSE. For some applications, this is not possible and cross-validation methods can be used to estimate the MSE (see Section 4).

The MSE can also be obtained by first integrating $\left[Y(\mathbf{X}, \mathbf{z}) - \widehat{Y}_{K,k}(\mathbf{X}, \mathbf{z}) \right]^2$ over $\mathbf{X}$ and then by taking the expectation with respect to $\mathbf{z}$. From the MSE, we also define the predictivity coefficient Q_2 which gives us the percentage of the mean explained variance of the output map:

$$Q_2 = 1 - \frac{\mathrm{MSE}}{\mathbb{E}_{\mathbf{z}} \left\{ \mathrm{Var}_{\mathbf{X}} \left[Y(\mathbf{X}, \mathbf{z}) \right] \right\}} \; . \tag{12}$$

The variance is taken with respect to $\mathbf{X}$ because we are interested in the variability induced by the model input vector $\mathbf{X}$. Q_2 corresponds to the coefficient of determination

R^2 computed in prediction (on a test sample or by cross-validation).

Step 3 - Choosing k^*, an optimal value for k

We perform simulations using the Campbell2D function and study convergence of MSE (Eq. (11)) as function of k. Our goal is to compare the three methods proposed in step 2, then to heuristically find an optimal value k^* for k. Indeed, there is a trade-off between keeping k small and minimizing the MSE. The MSE is computed using a test sample of 1000 independent Monte Carlo simulations, giving 1000 output maps.

Figure 3 gives the MSE results as function of k for different values of the learning sample size n. For each method, the MSE curves regularly turn downward as n increases. As expected, method 3, which is the richest in terms of model complexity, gives the best results, especially for small values of k. The usefulness of Gp is proved as we see that method 2 performs badly. It is certainly caused by the behavior of the first selected coefficients, which offer strong and non-linear variations: linear models are irrelevant for modeling these coefficients. For each method, the convergence is reached for k around 20 - 25. We decided to fix the optimal value at $k^* = 30$, which is a reasonable number of Gp models to be built.

In real applications, this methodology for choosing k^* can be applied even if the learning sample size n is limited. For a fixed n, we look for a stabilization of the MSE. If this convergence is not reached, we use a predefined maximal value for k.

It should be noted that if new model runs are available, the analyst has to repeat the process to choose k^*. However, in order to gain some analysis time, we can leave unchanged the ordering of coefficients, which has been obtained with the first set of simulations. In addition, we can just update the predictor (Eq. (4)) by keeping the initial estimation of the correlation parameters (which is the most cpu time consuming step). Such choices have to be made with care.

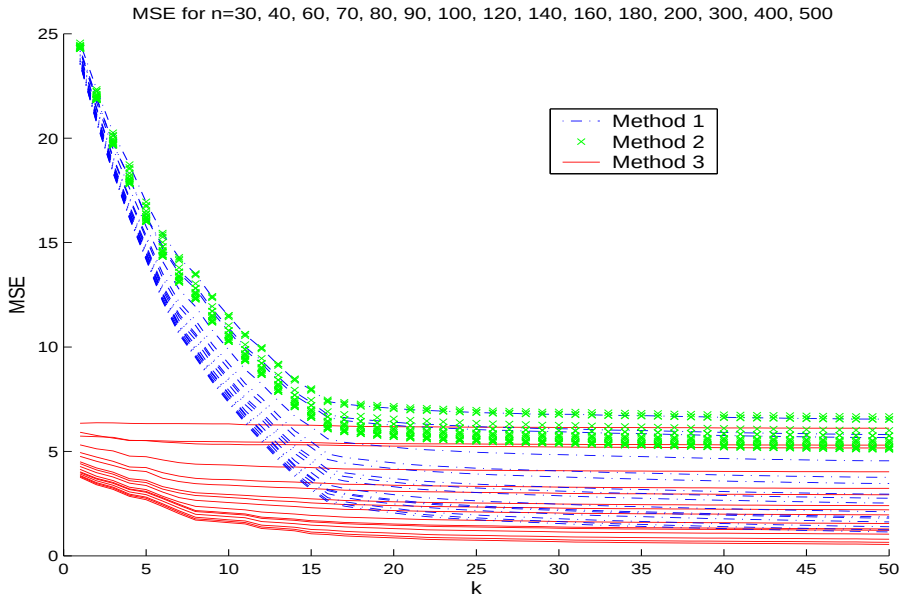


Figure 3: For the Campbell2D function, MSE convergence (as function of k) for the three methods and for various learning sample sizes ($n = 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 180, 200, 300, 400, 500$).

Step 4 - Convergence as function of the learning sample size n

Finally, it is important to study the convergence of the adequacy criteria as function of the learning sample size n. It would allow us to eventually prescribe the need to make new simulations with the code. For the Campbell2D function, Figure 4 gives the MSE results as function of n for different values of k. For each method, the MSE curves regularly turn downward as k increases. In real applications, one can restrict this to the visualization of the k^* curves.

Method 2 performs badly and the stabilization of its curves is obtained earlier. Indeed, adding simulations does not improve the linear models fitted on the k coefficients. For methods 1 and 3, the curve stabilization is not reached at $n = 500$. MSE would decrease for larger values of n, but this decrease becomes slower from $n = 200$ and MSE results are rather satisfactory for this value $n = 200$. In terms of predictivity coefficient (Eq. (12)), we obtain $Q_2 = 96.6\%$ for $n = 500$ and $Q_2 = 92.9\%$ for $n = 200$. For methods 1 and 3,

increasing k and n leads to a systematic decrease of the MSE. It can therefore be argued that MSE tends to zero and that our methodology converges.

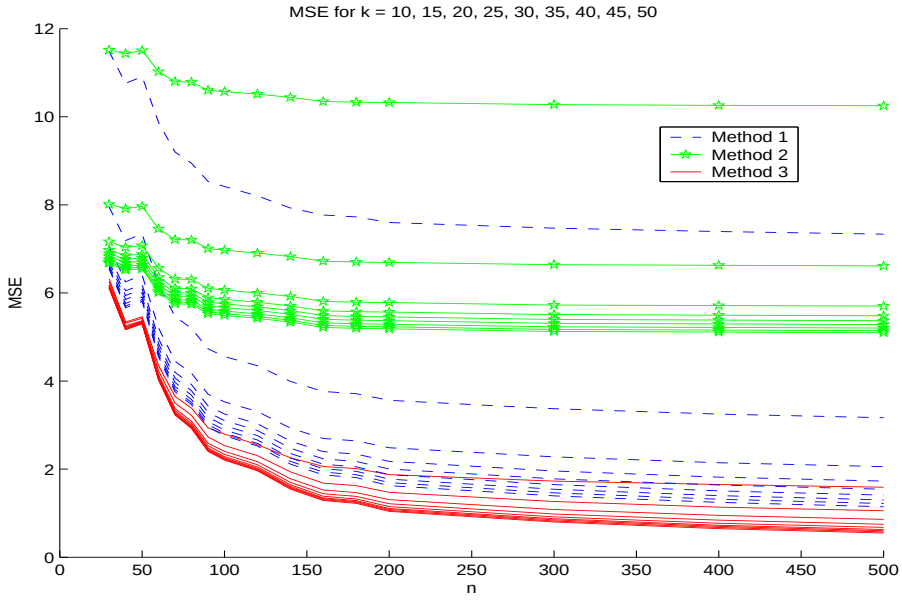


Figure 4: For the Campbell2D function, MSE convergence (as function of n) for the three methods and for various numbers k of Gp-modeled coefficients.

In real applications, if no additional simulation can be made, this step can be optional. However, in the opposite case, these curves would help us to decide if our simulation number is sufficient and which method we have to choose. Moreover, knowing that method 3 can be costly, we can decide to choose method 1 if their MSEs are similar. In practical terms, we start from an initial n_0 (random selection of n_0 simulations among the n simulations) and randomly add simulations until n. The choice of a low-discrepancy sequence would also allow the space-filling properties of the design to be kept while increasing n.

In conclusion, by analyzing all these convergence plots, we choose in the next section to use a learning sample size $n = 200$ and to model $k^* = 30$ Gps using method 3 in order to compute Sobol' indices.

Coarse estimation of the computational time of the different steps

Concerning the computational time needed to carry out all our methodology, the most costly steps are the construction of each Gp metamodel for the k^* wavelet coefficients and the validation step (i.e. computation of MSE or Q_2 by cross validation). All the other steps such as the wavelet decomposition, the selection of coefficients or the prediction of the functional metamodel for any new input value are negligible in terms of computational time.

So, the first main difficulty is the hyperparameter estimation of the k^* Gp metamodels. Indeed, each computation of the likelihood requires the inversion of correlation matrix and consequently, the maximum likelihood estimation can be CPU time consuming. In the case of 10 inputs for example and a few hundreds of simulations, a Gp modeling usually requires several minutes on a standard PC (Pentium 4, 1.8GHz). So, for tens of coefficients to be modeled, the step 2 can take one hour.

The second difficulty is the validation step, also because of the time required by the maximum likelihood estimation. To reduce its computational cost, a k-fold cross-validation is preferable in practice and limits the time required for cross validation to just a few hours. Another solution is to leave unchanged the hyperparameters of Gp at each loop of cross validation. Only the Gp predictor is updated. The cross validation is then a little biased but, for a few hundreds of simulations, this bias becomes quickly negligible.

As a conclusion, only the step of the Gp modeling is computationally expensive. For instance, in the Campbell2D function study, with $d = 8$ inputs, $n_z = 4096$ pixels, k ranging from 10 to 50 Gp models and $n = 200$ simulations, the metamodeling process from steps 1 to 3 (without the convergence plot in function of n) required approximately one day. For the MARTHE test case, with $d = 20$, $n_z = 4096$, $n = 300$, $k = 100$ and with a 10-fold cross-validation process, the computation of all the Sobol' indices has required approximately two days. These operational cost may appear to be high but this process is only made once to obtain a full functional metamodel. Afterwards, any evaluation of

the metamodel will require a negligible computational time compared to a simulation of the initial MARTHE simulator.

3.4 Global sensitivity analysis

At this stage, we have a functional metamodel allowing us to predict new output concentration maps for any new set of input variables. This metamodel has been obtained with only $n = 200$ computations with the Campbell2D function. To estimate Sobol' indices of the overall output map of the Campbell2D function, we then perform thousands of simulations on our functional metamodel. This method is called hereafter the functional metamodel-based approach.

Note that there is no direct link between the Sobol' indices for the wavelet coefficients and the Sobol' indices for the output. Indeed from (8), we have

$$Y_K(\mathbf{X}, \mathbf{z}) = \mu(z) + \sum_{j=1}^{K} \alpha_j(\mathbf{X})\phi_j(\mathbf{z}),$$

where $(\alpha_j(\mathbf{X}))_{1\leq j\leq K}$ denote the wavelet coefficients. The sensitivity map with respect to the variable X_i is $S_i(\mathbf{z}) = \dfrac{\mathrm{Var}\{\mathbb{E}[Y_K(\mathbf{X}, \mathbf{z})]|X_i\}}{\mathrm{Var}[Y_K(\mathbf{X}, \mathbf{z})]}$. Hence,

$$S_i(\mathbf{z}) = \frac{\sum_{j,l=1}^{K} \mathrm{Cov}\{\mathbb{E}[\alpha_j(\mathbf{X})|X_i], \mathbb{E}[\alpha_l(\mathbf{X})|X_i]\}\phi_j(\mathbf{z})\phi_l(\mathbf{z})}{\mathrm{Var}[Y_K(\mathbf{X}, \mathbf{z})]}.$$

If the functions $(\phi_j(\mathbf{z}))_{1\leq j\leq K}$ have disjoint supports, all the terms with $l \neq j$ in the above formula equal zero and Sobol' indices for the wavelet coefficients could be used to compute Sobol' indices for the output. In this paper, this is not the case as we use the Daubechies basis for these functions. This basis gave much better results than bases with disjoint supports functions (such as the Haar basis).

Thus, to estimate Sobol' indices of the overall output map of the Campbell2D function,

we perform thousands of simulations on our functional metamodel. Because of constraints of memory allocation (due to the size of the output map and our vectorial programming constraints), it is not possible to use Saltelli's Monte Carlo algorithm (Saltelli, 2002). Therefore, we use the following procedure for each of the n_z nodes of the grid:

- For the variance of the conditional expectation of each input variable X_i ($i = 1, \ldots, 8$), we perform 1000 Monte Carlo computations to estimate $\mathbb{E}(Y|X_i)$ (integration over 7 dimensions) and 200 Monte Carlo computations to estimate $\mathrm{Var}[\mathbb{E}(Y|X_i)]$ (integration over one dimension).

- For the variance of the conditional expectation of each $X_{\sim i}$ ($i = 1, \ldots, 8$), we perform 100 Monte Carlo computations to estimate $\mathbb{E}(Y|X_{\sim i})$ (integration over one dimension) and 1000 Monte Carlo computations to estimate $\mathrm{Var}[\mathbb{E}(Y|X_{\sim i})]$ (integration over 7 dimensions).

- The variance of the output $\mathrm{Var}(Y)$ is obtained using 2×10^4 simulations (integration over 8 dimensions).

- Thus, the first order Sobol' index estimates (noticed S_i^{Gp}) are obtained from Eq. (1) and the total Sobol' index estimates (noticed $S_{T_i}^{\mathrm{Gp}}$) are obtained from Eq. (2).

Finally, we obtain the Sobol' indices $S_i^{\mathrm{Gp}}(\mathbf{z})$ and $S_{T_i}^{\mathrm{Gp}}(\mathbf{z})$ for all the n_z grid points.

Figure 5 shows the Sobol' index maps for X_2 and X_6, which are the most influential input variables in the Campbell2D function (see Fig. 2). Results for X_2 are particularly convincing: first order and total sensitivity values obtained with the functional metamodel-based approach are accurate everywhere in the spatial domain D_z. Results for X_6 are fairly good for the first order Sobol' index and less precise for the total Sobol' index. However, the spatial influence zone of X_6 in the upper left corner is well retrieved by the functional metamodel-based approach. In fact, X_2 corresponds to a solely influential input variable while X_6 has significant interactions with other input variables (mainly

with X_3). Therefore, because of a more difficult Gp fitting process, the Gp models of the wavelet coefficients of X_6 are less precise than the Gp models of the wavelet coefficients of X_2. However, we argue that the important information is present in the spatial Sobol' map of $S_{T_6}^{\mathrm{Gp}}(\mathbf{z})$.

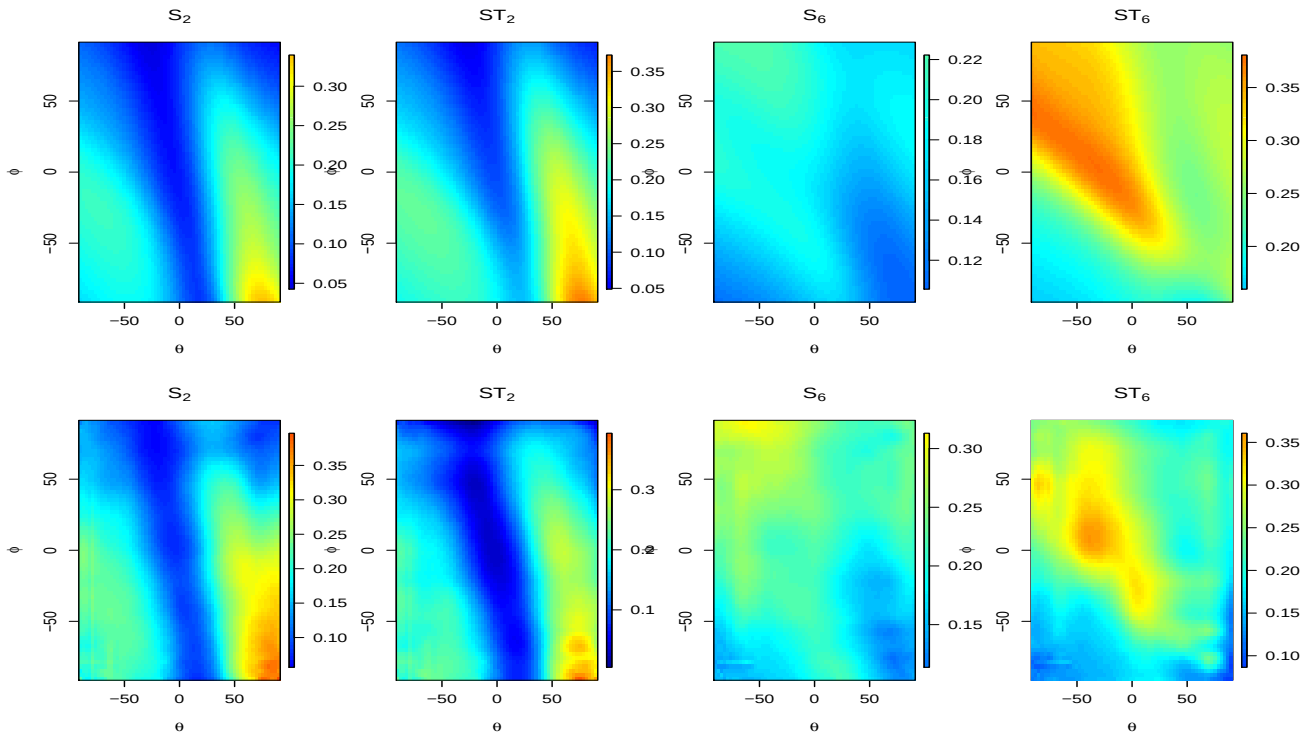


Figure 5: For the Campbell2D function and variables X_2 and X_6, comparison between exact first order and total Sobol' indices (top) and functional metamodel-based Sobol' indices (bottom). The color scales are the same for all the plots.

For all the input variables, the relative mean absolute errors of the first order Sobol' indices,

$$\mathrm{rMAE}(S_i) = \frac{\mathbb{E}_{\mathbf{z}}|S_i^{\mathrm{Gp}}(\mathbf{z}) - S_i(\mathbf{z})|}{\mathbb{E}_{\mathbf{z}}[S_i(\mathbf{z})]}\,, \tag{13}$$

were estimated for $i = 1, \ldots, 8$ (see Table 1). The results of Table 1 show that the estimations of the sensitivity maps for X_2 and X_6 correspond to one of the most difficult

Table 1: For the Campbell2D function, relative mean absolute errors (in percent) of the first order sensitivity indices estimated via the functional metamodel-based approach.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
8.75	16.25	16.35	12.8	—	13.17	11.80	9.96

cases. Figure 5 shows that a mean absolute error of a 15%-order is quite satisfactory in terms of sensitivity maps. Therefore, all the results for the other input variables show that our functional metamodel-based approach gives precise results. Note that the rMAE value for X_5 is not given because $S_5(z_1, z_2) = 0 \ \ \forall (z_1, z_2) \in [-90, 90]^2$, and the denominator in Eq (13) is equal to zero.

In conclusion, we have shown the efficiency of this new spatial global sensitivity analysis method for this analytical and relatively complex test function: all sensitivity index spatial maps have been obtained using only $n = 200$ computations of the Campbell2D function.

4 APPLICATION

4.1 The environmental problem

In the period between 1943 and 1974 radioactive waste was buried in eleven temporary repositories built on a specially allocated site at the RRC Kurchatov Institute (KI) in the Moscow area (Russia). The site used for radioactive waste interim storage covers an area of about 2 hectares and is situated near the KI external perimeter in the immediate vicinity of the city's residential area. A radioactive survey of the site and its adjacent area performed in the late 1980s - early 1990s and in 2002 showed that radioactive contamination is not only present on the surface but has a tendency to spread into the groundwater. The porous media of the site is represented principally by sands alternatively with clays that form several horizontal superposed aquifers. To analyze radioactive contamination of

groundwater, about a hundred exploration wells were drilled on the site. As a result of the survey, it was discovered that contamination of groundwater concerns mainly connected to ^{90}Sr. Since the radiation survey results have demonstrated the necessity to clean up the site, rehabilitation activities on radwaste removal and liquidation of old repositories were performed at the site between 2002 and 2006. A network of observation wells is used to control groundwater conditions of the two upper aquifers. This network consists of twenty observation wells for the upper moraine aquifer and nine for the second Jurassic aquifer. It is used for a regular recording of groundwater levels, its chemical and radionuclide composition (see Velikhov et al., 2007).

A numerical model of ^{90}Sr transport in groundwater was developed for the RRC Kurchatov Institute (KI) radwaste disposal site (Volkova et al., 2008). It aimed to provide a correct prediction of further contamination plume spreading since 2002 (using an interpolated concentration map) and up to the end of the year 2010, to show the risks associated with contamination and to serve as a basis for engineering decision-making. The numerical model was constructed using the MARTHE hydrogeological program package (developed by BRGM, the French Geological Survey). It is a three-dimensional combined transient flow and transport convection-dispersion model taking into account sorption and radioactive decay. Three layers were singled out; horizontal, vertical and temporal meshes were chosen in accordance with the migration characteristics of the sand. Initial concentration plume in 2002 and spreading prediction made for the year 2010 are shown in Figure 6. As can be seen, contamination plume predicted for the year 2010 is not uniform and is more diffused than the initial one. This is due, above all, to the influence of intensive infiltration assigned to several zones of the model domain that results in local dispersion of the contamination plume.

It has been shown in Volkova et al. (2008) that the shape of the predicted contamination plume depends on the model input values (hydraulic conductivity, infiltration pa-

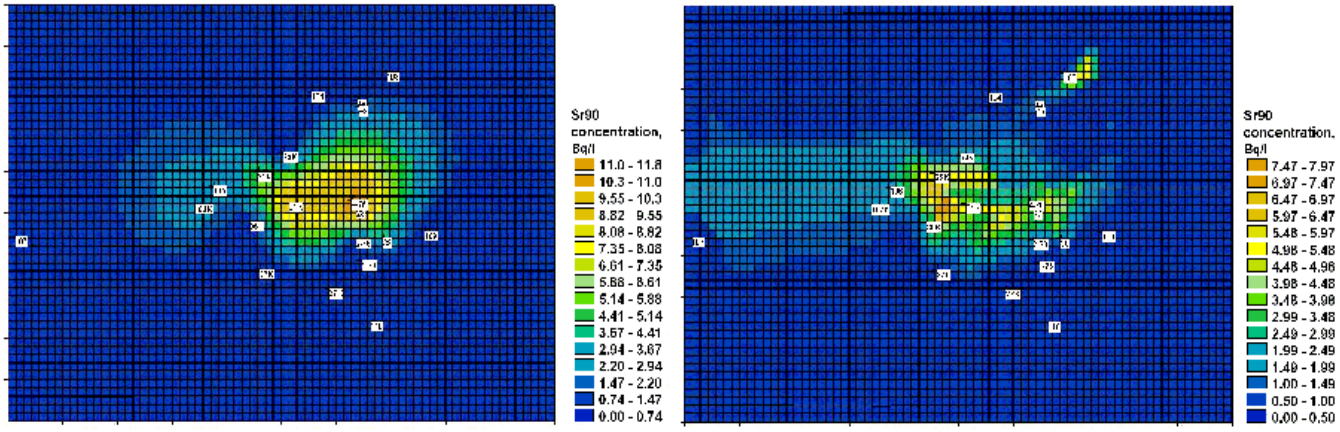


Figure 6: Initial (left, 2002) and predicted (right, 2010) ^{90}Sr concentrations (hot colors represent higher levels of concentration). Initial concentrations range from 0 to 12 Bq/l while final concentrations range from 0 to 8 Bq/l. The small white rectangles represent the location of the observation wells.

rameters, sorption distribution coefficients, etc.). Indeed, a large part of the model input variables are exposed to some uncertainty, since their values have been obtained through expert judgment, model calibration, field experiments and laboratory experiments. These uncertainties lead to uncertainties in model prediction. In order to evaluate the degree of input influence on the resulting contamination plume shape and concentration values predicted in observation wells, it was proposed to perform global sensitivity analysis on this numerical model (called MARTHE in the following sections).

4.2 Global sensitivity analysis on scalar outputs

From expert judgment and laboratory experiments, probability distributions (uniform and Weibull laws) were assigned to 20 random input variables of MARTHE. 300 Monte Carlo simulations, based upon Latin hypercube sampling of the input variables (McKay et al., 1979), were performed (requiring four calculation days). For each simulated set of input variables, MARTHE computes transport equations of ^{90}Sr and predicts the evolution of

^{90}Sr concentration. The 20 uncertain model parameters are the permeability of different geological layers composing the simulated field, longitudinal and transverse dispersivity coefficients, and sorption distribution coefficients. To perform global sensitivity analysis and in particular to compute Sobol' indices, previous studies have concentrated on 20 scalar outputs of ^{90}Sr concentration values, predicted for the year 2010, in 20 piezometers located on the waste repository site.

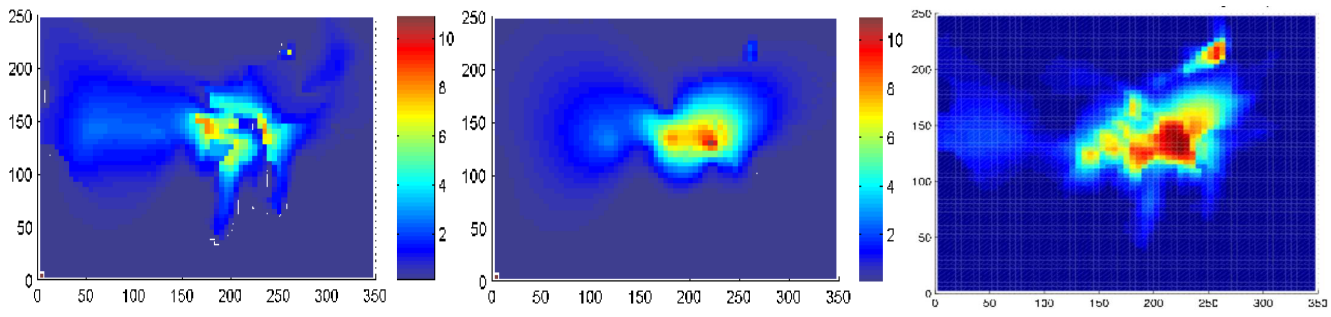
Because of the long computing time of MARTHE and of the non-linearity of the relationships between inputs and outputs, Volkova et al. (2008) proposed to fit a metamodel (based upon the boosting of regression trees) on each output using the learning sample (300 observations). The boosting trees method consists of a sequential construction of weak models (here regression trees with low interaction depth), that are then aggregated. This leads to a relatively efficient metamodel (but difficult to interpret). Then Sobol' indices were computed by intensive Monte Carlo simulations using this metamodel. In Marrel et al. (2008), each output was modeled by a Gp metamodel. The Gp metamodel outperforms the linear regression and the boosting regression trees metamodel in terms of predictivity of the output values.

As a result of these sensitivity analyses, we note that the calculated concentrations at the piezometric locations are mainly influenced by the distribution coefficient of ^{90}Sr in the first and second layers of the domain and by the intensity infiltration in the pipe leakage zones, and to a lesser extent by the hydrodynamic parameters (dispersivity, porosity, etc.). However, we are aware that spatial information has been lost in these analyses, due to the limited amount of output values that we have considered (concentrations located at 20 locations). Our goal was then to compute Sobol' indices in the whole spatial concentration map, predicted by the model for 2010.

4.3 Global sensitivity analysis on the output concentration map

The methodology presented in the previous section was then applied to MARTHE. Remember that this model contains $d = 20$ input random variables and that the $n = 300$ simulations have been performed following a Latin hypercube sample in a previous work. In previous studies, 20 scalar output variables had been considered and we hoped to obtain more information by using all the spatial information contained in the maps. We used the 300 spatial output maps, discretized in $n_z = 4096$ pixels and predicting the ^{90}Sr concentration values in 2010.

Figure 7 (a) and (b) shows two output maps and exemplifies the potential variability between the maps and their contour irregularity. Another output map (Figure 6, right) confirms this observation. The variance of the 300 maps (Figure 7 (c)) allows us to illuminate the strong-variability zones (central spot), the mild-variability zones (on the left and at the top of the central spot) and the zones with no variability where the concentration values are equal to zero (the major part of the maps). All this corroborates the need for a non-trivial functional metamodel, such as our wavelet-Gp based metamodel decribed in Section 3.3.



(a) (b) (c)

Figure 7: (a) and (b): Two final concentration maps of MARTHE (units in Bq/l). (c): Variance of the 300 concentration maps (colors are in logarithmic scales, ranging from 0 to 10).

As step 0 was already done, we applied the remaining steps of the spatial global sensitivity analysis methodology (see Section 3) using our learning sample of size $n = 300$. From steps 2 and 3, we retained method 3 with the choice of $k^* = 100$ modeled coefficients with Gp: the stabilization of MSE was observed for this value of k^*. The number of coefficients modeled with linear models is $k' = 900$. Step 4 was not applied to this application case. Indeed, MARTHE simulations have been performed in a previous study (Volkova et al., 2008) and the computer code is no longer available. Therefore, no additional point could be added and step 4 would be useless.

In the MARTHE application, no test basis was available to compute the MSE in prediction. The MSE estimate was obtained via a 10-fold cross-validation technique. The learning sample was randomly divided into 10 sub-samples. Then, we iterated 10 times the following process: learning the functional metamodel on 9 sub-samples and estimating the MSE on the remaining sub-sample. Our final MSE estimate is the mean of the 10 obtained MSE values: MSE= 0.039. In terms of predictivity coefficient (Eq. (12)), we obtain $Q_2 = 72.1\%$. All the details of this study are given in Marrel (2008).

At present, the functional metamodel can be used to estimate first order and total Sobol' indices. We used Saltelli's Monte Carlo algorithm (as for the total Sobol' indices in Section 3.2) with $N = 10^3$. Indeed, the low computational cost of our metamodel makes it possible to carry thousands of simulations, but not billions because of memory allocation problems (see Section 3.4). The final computation cost of Saltelli's algorithm is $N(d+2)$, which leads to a number of 22000 metamodel-based simulations in our case. As a final result, we obtain 20 maps of first order Sobol' indices and 20 maps of total Sobol' indices (two maps for each input).

Figure 8 (a), (b) and (c) shows three maps of total Sobol' indices S_{T_i} corresponding to the three main influential variables. The 17 remaining input variables have no influence in any zone of the spatial output domain. These results are completely coherent with pre-

vious studies which have detected the predominant influence of these three variables. Our new results have provided some additional spatial information. For example, we locate more precisely the influence zones of the distribution coefficient of the first hydrogeological layer. Such information is precious for model engineers. It could help them to determine according to the spatial location of large variability zones the kind of additional information which is needed. Subsequent decisions could be to place new piezometers in specific geographical zones. The methodological developments highlight not only the direct application to post-treatment processes but also enable us to propose a new characterization strategy.

Figure 8 (d) gives spatial information about the MARTHE model. It clarifies the obvious correlation between the MARTHE hydrogeological scenario and our obtained spatial maps of sensitivity indices: influential $kd1$ zones correspond to the absence of the second hydrogeological layer while influential $kd2$ zones correspond to its presence. In Figure 8 (c), we also retrieve the high infiltration lines of Figure 8 (d) and see their spatial area of influence.

In our radioactive waste problem, the Sobol' maps of each uncertain input parameter clearly provide guidance to a better understanding of the simulator forecast and can be used to reduce the response uncertainties most efficiently. For example, if we want to reduce the predicted concentration uncertainty at a specific point of the map, we analyze all the Sobol' maps and determine the most influential inputs at this point. Then, we can try to reduce the uncertainty of these inputs by additional measures. Moreover, spatial maps for sensitivity indices can reveal gradient of influence of uncertain parameters, linked to the physics of the phenomenon (e.g. influence of a parameter varying in function of the flow direction). The global influence of each input over the whole space can also be used to identify areas of influence and areas of non-influence of this input and can be linked, as for $kd1$ and $kd2$, to a map of a geological parameter. If we now consider the strong

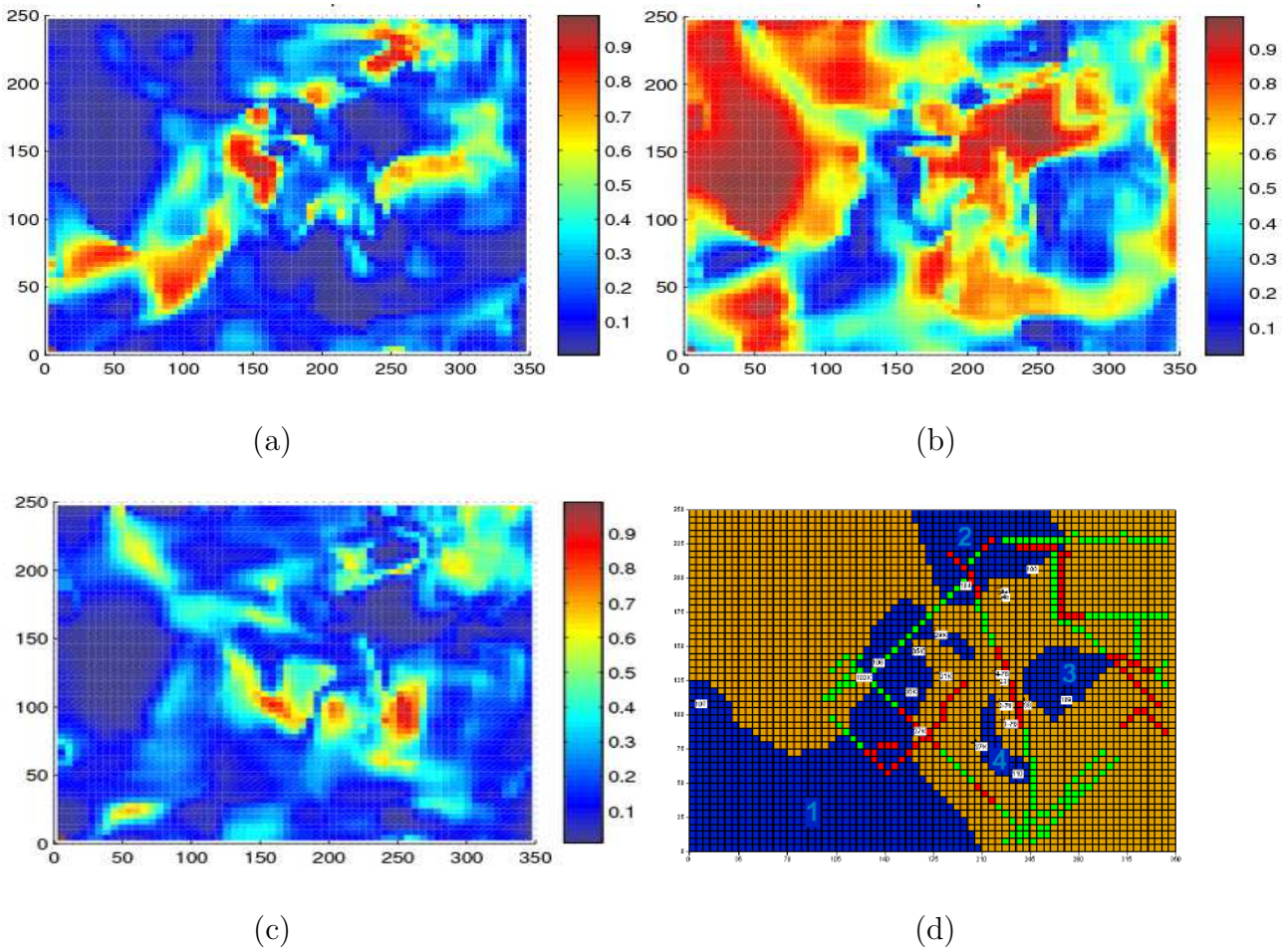


Figure 8: Total Sobol' indices of three input variables of MARTHE: (a) kd1 (distribution coefficient of the first layer), (b) kd2 (distribution coefficient of the second layer) and (c) i3 (high infiltration rate). (d): MARTHE hydrogeological model: blue zones (numbered from 1 to 4) correspond to low conductivity zones (absence of coarse sand in the second layer); lines present zones of high infiltration rates.

infiltration coefficient denoted as $i3$ and its sensitivity map, we can deduce that $i3$ is only influential around the pipe and its influence is very limited outside the pipe area. The lack of knowledge on this parameter does not induce a big uncertainty on the concentration forecast at the site boundary and consequently on the decision relative to the need of a site rehabilitation.

5 CONCLUSION

In this paper, a new methodology was introduced to compute spatial maps of variance-based sensitivity indices (such as the Sobol' indices) for numerical models giving spatial maps as outputs. Such situations often occur in environmental modeling problems. One critical issue with our method is due to the reduced number of model output maps available because of the high cpu time cost of the numerical model. A functional basis decomposition (wavelet basis) linked to a metamodel technique (based upon the Gp model) is proposed and used to solve this problem. Choosing a wavelet basis is well-suited for our application cases (analytical and real models) because strong spatial heterogeneities and sharp boundaries are observed in the model output maps. In addition,the Gp model is appropriate for handling the large differences between the output maps obtained for various inputs. This induces strong non-linear variations in the Gp-modeled wavelet coefficients. The resulting functional metamodel is a fast emulator (i.e. with negligible cpu time) of the computer code. It can be used for uncertainty propagation issues, optimization problems and, as advocated in this paper, for sensitivity index estimation.

An analytical test function was presented to explain the different steps, criteria and modeling choices of our methodology. The convergence of our Gp-based functional metamodel was also investigated. Then, our methodology was applied to a real case to stress its concrete applicability. We particularly emphasized the relevance of the additional information (in addition to the expert and model knowledge) brought by the spatial maps of first order and total sensitivity indices. These sensitivity maps allow us for spatially identifying the most influential inputs, for detecting zones with input interactions and for determining the zone of influence for each input.

Our methodology can be extended to any computer codes with functional outputs: codes with outputs depending on time, codes depending on other physical processes (such as a function of temperature), codes with outputs varying in space and time. In the third

case, the temporal and the spatial scales must be carefully distinguished. It would be interesting in a future work to apply our method to the MARTHE spatio-temporal evolutions of the concentration values (between 2002 and 2010). In addition, improvements could be proposed. For example, the vaguelette-wavelet decomposition (Abramovich and Silverman, 1997; Ruiz-Medina et al., 2007) would be an interesting substitute to the wavelet decomposition. It would allow a simultaneous treatment of all the spatial output maps and a direct standardization of all decomposition coefficients. Last, dealing with the functional input case remains an important and challenging issue to disseminate the global sensitivity analysis into environmental modeling communities. Iooss and Ribatet (2009) and Lilburne and Tarantola (2009) proposed some preliminary methodologies to account for the spatially distributed inputs when computing Sobol' indices.

6 ACKNOWLEDGMENTS

This work was backed by the "Risk Control" project that is managed by the CEA/Nuclear Energy Division/Nuclear Development and Innovation Division, and by the "Monitoring and Uncertainty" project of IFP. This work has also been backed by French National Research Agency (ANR) through COSINUS program (project COSTA BRAVA n^{o}ANR-09-COSI-015). We are grateful to Mickaele Le Ravalec for her help with the English.

APPENDIX A: SOBOL INDICES FOR THE CAMPBELL2D FUNCTION

The analytical derivations of the first order Sobol' indices S_i (Eq. (1)) of the Campbell2D function (6) consists, first of all, in obtaining analytical expressions of the conditional expectations $\mathbb{E}(Y|X_i)$ (for $i = 1, \ldots, 8$). The multiple integrations are made following the

uniform distribution on $[-1,5]$ (we have $\mathbb{E}(X_i)=2$ and $\mathrm{Var}(X_i)=3\ \forall\, i=1,\ldots,8$). The terms of these integrals which do not depend on X_i can be directly put to zero (because these terms disappear when the variance over X_i is taken). In the next step, we take the variance over X_i of the expressions of the conditional expectations (which leads to simple integrals). In some cases, analytical simplifications can be made but in other cases, these variances cannot be simplified and the integrals are evaluated by Monte Carlo.

We recall that $(z_1,z_2)\in[-90,90]^2$ and we define the following variable changes:

$$\theta_1=0.8z_1+0.2z_2\ ,\ \theta_2=0.5z_1+0.5z_2\ ,\ \phi_1=0.4z_1+0.6z_2\ ,\ \phi_2=0.3z_1+0.7z_2\ . \quad (14)$$

The Campbell2D function is now written

$$\begin{aligned} g(\mathbf{X},z_1,z_2) &= X_1\exp\left[-\frac{(\theta_1-10X_2)^2}{60X_1^2}\right]+(X_2+X_4)\exp\left[\frac{\theta_2X_1}{500}\right] \\ &+X_5(X_3-2)\exp\left[-\frac{(\phi_1-20X_6)^2}{40X_5^2}\right]+(X_6+X_8)\exp\left[\frac{\phi_2X_7}{250}\right]\ , \end{aligned} \quad (15)$$

We also define $\Phi(x)$ as the cumulative distribution function of a standardized Gaussian variable. The first order Sobol' indices for the 8 input variables are written:

$$S_1(z_1,z_2)=\mathrm{Var}\left\{\sqrt{\frac{\pi}{60}}X_1^2\left[\Phi\left(\frac{50-\theta_1}{\sqrt{30}X_1}\right)-\Phi\left(-\frac{10+\theta_1}{\sqrt{30}X_1}\right)\right]+4\exp\left(\frac{\theta_2X_1}{500}\right)\right\}\ , \quad (16)$$

$$S_2(z_1,z_2)=\begin{cases} \mathrm{Var}\left\{\dfrac{250X_2}{3\theta_2}\left[\exp\left(\dfrac{\theta_2}{100}\right)-\exp\left(-\dfrac{\theta_2}{500}\right)\right]+\displaystyle\int_{-1}^{5}\frac{x}{6}\exp\left[-\frac{1}{2}\left(\frac{\theta_1-10X_2}{\sqrt{30}x}\right)^2\right]dx\right\} & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \theta_2\neq 0\ , & \\ \mathrm{Var}\left\{X_2+\displaystyle\int_{-1}^{5}\frac{x}{6}\exp\left[-\frac{1}{2}\left(\frac{\theta_1-10X_2}{\sqrt{30}x}\right)^2\right]dx\right\}\ \text{if } \theta_2=0\ , & \end{cases} \quad (17)$$

$$S_3(z_1, z_2) = \frac{\pi}{120} \left\{ \int_{-1}^{5} \frac{x^2}{6} \left[\Phi\left(\frac{100 - \phi_1}{\sqrt{20}x} \right) - \Phi\left(\frac{-20 - \phi_1}{\sqrt{20}x} \right) \right] dx \right\}^2 , \tag{18}$$

$$S_4(z_1, z_2) = \begin{cases} \frac{1}{3} \left\{ \frac{250}{\theta_2} \left[\exp\left(\frac{\theta_2}{100} \right) - \exp\left(-\frac{\theta_2}{500} \right) \right] \right\}^2 & \text{if } \theta_2 \neq 0 , \\ 3 \text{ if } \theta_2 = 0 , \end{cases} \tag{19}$$

$$S_5(z_1, z_2) = 0 , \tag{20}$$

$$S_6(z_1, z_2) = \begin{cases} \frac{1}{3} \left\{ \frac{125}{\phi_2} \left[\exp\left(\frac{\phi_2}{50} \right) - \exp\left(-\frac{\phi_2}{250} \right) \right] \right\}^2 & \text{if } \phi_2 \neq 0 , \\ 3 \text{ if } \phi_2 = 0 , \end{cases} \tag{21}$$

$$S_7(z_1, z_2) = \begin{cases} \frac{8}{3} \frac{125}{\phi_2} \left[\exp\left(\frac{\phi_2}{25} \right) - \exp\left(-\frac{\phi_2}{125} \right) \right] - \frac{4}{9} \left\{ \frac{250}{\phi_2} \left[\exp\left(\frac{\phi_2}{50} \right) - \exp\left(-\frac{\phi_2}{250} \right) \right] \right\}^2 & \\ \qquad \text{if } \phi_2 \neq 0 , & \\ 0 \text{ if } \phi_2 = 0 , & \end{cases} \tag{22}$$

$$S_8(z_1, z_2) = S_6(z_1, z_2) . \tag{23}$$

References

Abramovich, F. and Silverman, B. (1997). The vaguelette-wavelet decomposition approach to statistical inverse problems. *Biometrika*, 85:115–129.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, 35:1874–1906.

Campbell, K., McKay, M., and Williams, B. (2006). Sensitivity analysis when model ouputs are functions. *Reliability Engineering and System Safety*, 91:1468–1472.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. Wiley, New-York.

De Rocquigny, E., Devictor, N., and Tarantola, S., editors (2008). *Uncertainty in industrial practice*. Wiley.

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

Fang, K.-T., Li, R., and Sudjianto, A. (2006). *Design and modeling for computer experiments*. Chapman & Hall/CRC.

Fassò, A., Esposito, A., Porcu, E., Reverberi, A. P., and Vegliò, F. (2003). Statistical sensitivity analysis of packed column reactors for contaminated wastewater. *Environmetrics*, 14:743–759.

Helton, J. (1993). Uncertainty and sensitivity analysis techniques for use in performance assesment for radioactive waste disposal. *Reliability Engineering and System Safety*, 42:327–367.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103:571–583.

Hoeting, J., Davis, R., Merton, A., and Thompson, S. (2006). Model selection for geostatistical models. *Ecological Applications*, 16:87–98.

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

Iooss, B., Boussouf, L., Feuillard, V., and Marrel, A. (2010). Numerical studies of the metamodel fitting and validation processes. *International Journal of Advances in Systems and Measurements*, 3, in press.

Iooss, B. and Ribatet, M. (2009). Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, 94:1194–1204.

Kleijnen, J. and Sargent, R. (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120:14–29.

Koehler, J. and Owen, A. (1996). Computer experiments. In Ghosh, S. and Rao, C., editors, *Design and analysis of experiments*, volume 13 of *Handbook of statistics*. Elsevier.

Lamboni, M., Makowski, D., Lehuger, S., Gabrielle, B., and Monod, H. (2009). Multivariate global sensitivity analysis for dynamic crop models. *Fields Crop Research*, 113:312–320.

Lilburne, L. and Tarantola, S. (2009). Sensitivity analysis of spatial models. *International Journal of Geographical Information Science*, 23:151–168.

Marrel, A. (2008). *Mise en oeuvre et exploitation du métamodèle processus gaussien pour l'analyse de modèles numériques - Application à un code de transport hydrogéologique*. Thèse de l'INSA Toulouse.

Marrel, A., Iooss, B., Laurent, B., and Roustant, O. (2009). Calculations of the Sobol indices for the Gaussian process metamodel. *Reliability Engineering and System Safety*, 94:742–751.

Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis*, 52:4731–4744.

Martin, J. and Simpson, T. (2005). Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43:853–863.

McKay, M., Beckman, R., and Conover, W. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.-M. (2007). *Matlab - Wavelet toolbox user's guide*. The Mathworks.

Nychka, D., Cox, L., and Piegorsch, W., editors (1998). *Case studies in environmental statistics*. Springer Verlag.

Oakley, J. and O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution. *Biometrika*, 89:769–784.

Oakley, J. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 66:751–769.

Ruiz-Medina, M., Angulo, J., and Fernández-Pascual, R. (2007). Wavelet-vaguelette decomposition of spatiotemporal random fields. *Stochastic Environmental Research and Risk Assessment*, 21:273–281.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4:409–435.

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

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communication*, 181:259–270.

Saltelli, A., Chan, K., and Scott, E., editors (2000). *Sensitivity analysis*. Wiley Series in Probability and Statistics. Wiley.

Shi, J., Wang, B., Murray-Smith, R., and Titterington, D. (2007). Gaussian process functional regression modeling for batch data. *Biometrics*, 63:714–723.

Sobol, I. (1993). Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414.

Velikhov, E. P., Ponomarev-Stepnoi, N. N., Volkov, V. G., Gorodetskii, G. G., Zverkov, Y. A., Ivanov, O. P., Koltyshev, S. M., Muzrukova, V. D., Semenov, S. G., Stepanov, V. E., Chesnokov, A. V., and Shisha, A. D. (2007). Rehabilitation of the radioactively contaminated objects and territory of the Russian Science Center Kurchatov Institute. *Atomic Energy*, 102:375–381.

Volkova, E., Iooss, B., and Van Dorpe, F. (2008). Global sensitivity analysis for a numerical model of radionuclide migration from the RRC "Kurchatov Institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assesment*, 22:17–31.

Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T., and Morris, M. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34(1):15–25.