Global Sensitivity Analysis of Stochastic Computer Models with Generalized Additive Models

Bertrand Iooss, Mathieu Ribatet, Amandine Marrel

Global sensitivity analysis of stochastic computer models with generalized additive models

Bertrand IOOSS*, Mathieu RIBATET† and Amandine MARREL‡

* CEA Cadarache, DEN/DER/SESI/LCFR, 13108 Saint Paul lez Durance, Cedex, France
† CEMAGREF Lyon, Unité de Recherche Hydrologie-Hydraulique, 3 bis quai Chauveau, CP220, 69336 Lyon cedex 09, France
‡ CEA Cadarache, DEN/DTN/SMTM/LMTE, 13108 Saint Paul lez Durance, Cedex, France

Corresponding author: B. Iooss ; Email: bertrand.iooss@cea.fr
Phone: +33 (0)4 42 25 72 73 ; Fax: +33 (0)4 42 25 24 08

Abstract

The global sensitivity analysis, used to quantify the influence of uncertain input parameters on the response variability of a numerical model, is applicable to deterministic computer codes (for which the same set of input parameters gives always the same output value). This paper proposes a global sensitivity analysis method for stochastic computer codes (having a variability induced by some uncontrolable parameters). The mean and dispersion of the code outputs are modeled by two interlinked Generalized Additive Models (GAM). The "mean" model allows to obtain the controlable parameters sensitivity indices, while the "dispersion" model allows to obtain the uncontrolable parameters ones. The relevance of the proposed model is analyzed with two case studies. Results show that the joint modeling approach leads to more accurate sensitivity index estimations, especially for the joint GAM model.

Keywords: joint modeling, mean and dispersion, generalized linear model, metamodel, uncertainty

1 INTRODUCTION

Many phenomena are modeled by mathematical equations which are implemented and solved by complex computer codes. These computer models often take as inputs a high number of numerical parameters and physical variables, and give several outputs (scalars or functions).

For the development of such computer models, its analysis, or its use the global Sensitivity Analysis (SA) method is an invaluable tool (Saltelli et al. [26], Kleijnen [12], Helton et al. [6]). It takes into account all the variation ranges of the inputs, and tries to apportion the output uncertainty to the uncertainty in the input factors. These techniques, often based on the probabilistic framework and Monte-Carlo methods, require a lot of simulations. The uncertain input parameters are modeled by random variables and characterized by their probabilistic density functions. The SA methods are used for model calibration, model validation, decision making process, i.e. all the processes where it is useful to know which variables mostly contribute to output variability.

The current SA methods are applicable to the deterministic computer codes, codes for which the same set of input parameters always gives the same output values. The randomness is limited to the model inputs, whereas the model itself is deterministic. Most computer codes belong to this kind of model. For example in the nuclear engineering domain, global sensitivity analysis tools have been applied to waste storage safety studies (Helton et al. [6]), environmental models of dose calculations (Iooss et al. [10]), pollutant transport models in the groundwater (Volkova et al. [31]). In such industrial studies, numerical models are often too time consuming for applying directly the global SA methods. To avoid this problem, one solution consists in replacing the time consuming computer code by an approximate mathematical model, called response surface or surrogate model or also metamodel (Sacks et al. [24], Fang et al. [3]). This function must be as representative as possible of the computer code, with good prediction capabilities and must require a negligible calculation time. Several metamodels are classically used: polynomials, splines, neural networks, Gaussian processes (Chen et al. [2], Fang et al. [3]).

In this paper, we are not interested by deterministic computer models but by stochastic numerical models - i.e. the same input parameters set leads to different output values. The model is therefore intrinsically stochastic. For the uncertainty analysis, Kleijnen [12] has raised this question, giving an example concerning a queueing model. In the nuclear engineering domain, examples are given by Monte-Carlo neutronic models used to calculate elementary particles trajectories, Lagrangian stochastic models for simulating a large number

of particles inside turbulent media (in atmospheric or hydraulic environment). In our study, "uncontrolable" parameters correspond to parameters that are known to exist, but unobservable, inaccessible or non describable for some reasons. It includes the important case in which observable vectorial parameters are too complex to be described by a reasonable number of scalar parameters. This last situation concerns the codes in which some simulations of random processes are used: the output values of the computer code depend on the realizations of these random functions. For example, one can quote some partial differential equation resolutions in heterogeneous random media simulated by geostatistical techniques (fluid flows in oil reservoirs, Zabalza-Mezghani et al. [36], acoustical wave propagation in turbulent fluids, Iooss et al. [8]), where the uncontrolable parameter is the simulated spatial field involving several thousand scalar values for each realization.

For an environmental assessment problem, Tarantola et al. [29] propose a first solution by introducing a binomial input parameter ξ governing the simulation of the random field. Therefore, the sensitivity index of ξ quantifies the influence of the random field on the model output variable. However, this method does not give any idea about the influence of the possible interactions between the uncontrolable parameter and the other uncertain input parameters. Moreover, to perform a sensitivity analysis, such approach requires a large number of computer model calculations (several hundreds per input parameter). For most applications, it is impossible due to intractable CPU times, computer codes have to be substituted for metamodels.

For stochastic computer models, classical metamodels (devoted to approximate deterministic computer models) are not pertinent. To overcome this problem, the commonly used Gaussian Process (GP) model is interesting. Kleijnen & van Beers [13] have demonstrated the usefulness of GP for stochastic computer model. Moreover, GP can include an additive error component (called the "nugget effect") by adding a constant term into its covariance function (Rasmussen & Williams [22]). However, it supposes that the error term is independent of the input parameters (homoscedasticity hypothesis), which means that the uncontrolable parameter does not interact with controlable parameters. This hypothesis limits the usefulness of the GP model to particular cases. To construct heteroscedastic metamodels for stochastic

computer codes, Zabalza-Mezghani et al. [35] model the mean and the dispersion of computer code outputs by two interlinked Generalized Linear Models (GLMs). This approach, called the joint model, has been previously studied in the context of experimental data modeling (McCullagh & Nelder [16]). Compared to the GP model, this approach theoretically suits the study of heteroscedastic situations and allows the obtention of a model for the dispersion.

Following the work of Zabalza et al., Iooss & Ribatet [9] have recently introduced the joint model to perform a global sensitivity analysis of a stochastic model. Results show that a total sensitivity index of all the uncontrolable parameters can be computed using the dispersion component of the joint model. However, the parametric form of the GLM framework provides some limitations when modeling complex computer code outputs. To resolve this problem, this paper suggests the use of non parametric models to allow more flexibility and complexity while fitting to the data. Due to its similarity with GLMs, Generalized Additive Models (GAM) are considered (Hastie & Tibshirani [4], Wood & Augustin [34]). GAMs allow variable and model selections *via* a quasi-likelihood function, classical statistical tests on coefficients, and graphical displays.

This paper starts by describing the joint model construction, firstly with the GLM, secondly with the GAM. The third section describes the global sensitivity analysis for deterministic models, and its extension to stochastic models using joint models. Particular attention is devoted to the calculation of variance-based sensitivity indices (the so-called Sobol indices). Considering a simple analytic function, the performance of the proposed approach is compared to other commonly used models. Next, an application on an actual industrial case (groundwater radionucleide migration modeling) is given. Finally, some conclusions synthesize the contributions of this work.

2 JOINT MODELING OF MEAN AND DISPERSION

2.1 Using the Generalized Linear Models

The class of GLM allows to extend the class of the traditional linear models by the use of: (a) a distribution which belongs to the exponential family; (b) and a link function which

connects the explanatory variables to the explained variable (Nelder & Wedderburn [19]). Let us describe the first component of the model concerning the mean:

$$\begin{cases} \mathrm{E}(Y_i) & = \mu_i, \qquad \eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j \ , \\ \mathrm{Var}(Y_i) & = \phi_i v(\mu_i) \ , \end{cases} \tag{1}$$

where $(Y_i)_{i=1\ldots n}$ are independent random variables with mean μ_i; x_{ij} are the observations of the parameter X_j; β_j are the regression parameters which have to be estimated; η_i is the mean linear predictor; $g(\cdot)$ is a differentiable monotonous function (called the link function); ϕ_i is the dispersion parameter and $v(\cdot)$ is the variance function. To estimate the mean component, the functions $g(\cdot)$ and $v(\cdot)$ have to be specified. Some examples of link functions are given by the identity (traditional linear model), root square, logarithm, and inverse functions. Some examples of variance functions are given by the constant (traditional linear model), identity and square functions.

Within the joint model framework, the dispersion parameter ϕ_i is not supposed to be constant as in a traditional GLM, but is supposed to vary according to the model:

$$\begin{cases} \mathbb{E}(d_i) & = \phi_i, \qquad \zeta_i = h(\phi_i) = \sum_j u_{ij}\gamma_j \ , \\ \mathrm{Var}(d_i) & = \tau v_d(\phi_i) \ , \end{cases} \tag{2}$$

where d_i is a statistic representative of the dispersion, γ_j are the regression parameters which have to be estimated, $h(\cdot)$ is the dispersion link function, ζ_i is the dispersion linear predictor, τ is a constant and $v_d(\cdot)$ is the dispersion variance function. u_{ij} are the observations of the explanatory variable U_j. The variables (U_j) are generally taken among the explanatory variables of the mean (X_j), but can also be different. To ensure positivity, $h(\phi) = \log\phi$ is often chosen for the dispersion link function. For the statistic representing the dispersion d, the deviance contribution (which is close to the distribution of a χ^2) is considered. Therefore, as the χ^2 is a particular case of the Gamma distribution, $v_d(\phi) = \phi^2$ and $\tau \sim 2$. In particular, for the Gaussian case, these relations are exact: d is χ^2 distributed and $\tau = 2$.

The joint model is fitted using Extended Quasi-Loglikelihood (EQL) (Nelder & Pregibon

[18]) maximization. The EQL behaves as a log-likelihood for both mean and dispersion parameters. This justifies an iterative procedure to fit the joint model. First, a GLM is fitted on the mean; then from the estimate of d, another GLM is fitted on the dispersion. From the estimate of ϕ, weights for the next estimate of the GLM on the mean are obtained. This process can be reiterated as many times it is necessary, and allows to entirely fit our joint model (McCullagh & Nelder [16]).

Statistical tools available in the GLM fitting are also available for each component of the joint model: deviance analysis, Student and Fisher tests, residuals graphical analysis. It allows to make some variable selection in order to simplify model expressions.

Remark: *Let us note that it is possible to build polynomial models for the mean and the variance separately (Vining & Myers [30], Bursztyn & Steinberg [1]). This approach, called the dual modeling, consists in repeating calculations with the same sets of controlable parameters (which is not necessary in the joint modeling approach). The dual modeling approach has been successfully applied in many situations, especially for robust conception problems: optimizing a mean response function while minimizing the variance. However for our purpose (accurate fitting of the mean and dispersion components), it has been shown that this dual model is less performant than the joint model (Zabalza et al. [35], Lee & Nelder [14]): the dual modeling approach fits the dispersion model given the mean model and this approach does not always lead to optimal fits.*

2.2 Extension to the Generalized Additive Models

Generalized Additive models (GAM) were introduced by Hastie & Tibshirani [4, 5] and allow a linear term in the linear predictor $\eta = \sum_j \beta_j X_j$ of equation (1) to be replaced by a sum of smooth functions $\eta = \sum_j s_j(X_j)$. The $s_j(.)$'s are unspecified functions that are obtained by fitting a smoother to the data, in an iterative procedure. GAMs provide a flexible method for identifying nonlinear covariate effects in exponential family models and other likelihood-based regression models. The fitting of GAM introduces an extra level of iteration in which each spline is fitted in turn assuming the others known. GAM terms can be mixed quite generally with GLM terms in deriving a model.

One common choice for s_j is the smoothing spline (Wahba [32]) - i.e. splines with knots at each distinct value of the variables. In regression problems, smoothing splines have to be penalized in order to avoid data overfitting. Wood & Augustin [34] have described in details how GAMs can be constructed using penalized regression splines. This approach is particularly well-suited because it allows the integrated model selection via Generalized Cross Validation (GCV) and related criteria, the incorporation of multi-dimensional smooths and relatively well founded inference using the resulting models. Because numerical models often exhibit strong interactions between input parameters, the incorporation of multi-dimensional smooth (for example the bi-dimensional spline term $s_{ij}(X_i, X_j)$) is particularly important in our context.

GAMs are generally fitted using penalized likelihood maximization. For this purpose, the likelihood is modified by the addition of a penalty for each smooth function, penalizing its "wiggliness". Namely, the penalized loglikelihood is defined as:

$$PL = L + \sum_{j=1}^{p} \lambda_j \int \left(\frac{\partial^2 s_j}{\partial x_j^2} \right)^2 dx_j \tag{3}$$

where L is the loglikelihood function, p is the total number of smooth terms and λ_j are "tuning" constants which compromise between goodness of fit and smoothness.

Estimation of these "tuning" constants is generally achieved using the GCV score minimization. The GCV score is defined as:

$$S_{GCV} = \frac{nd}{(n - DoF)^2} \tag{4}$$

where n is the number of data, d is the deviance and DoF is the effective degrees of freedom, i.e. the trace of the so-called "hat" matrix. Extension to (E)QL models is straightforward by substituting the likelihood function L and the deviance d for their (extended) quasi counterparts.

We have seen that GAMs extend in a natural way GLMs. Therefore, it would be interesting to extend the joint GLM model to a joint GAM one. Such ideas have been proposed in Rigby

& Stasinopoulos [23] where both the mean and variance were modeled using semi-parametric additive models (Hastie & Tibshirani [5]). This model is restricted to observations following a Gaussian distribution and is called Mean and Dispersion Additive Model (MADAM). As our model is based on GAMs and by analogy with the denomination "joint GLM", we call it "joint GAM" in the following. Rigby & Stasinopoulos [23] proposed an algorithm to fit the MADAM model. This fitting procedure is exactly the same than the one for the two interlinked GLMs, apart from the stopping rule. Indeed, the two interlinked GLMs (resp. GAMs) model is fitted when the EQL (resp. PEQL) remains stable within the iterative procedure.

3 GLOBAL SENSITIVITY ANALYSIS

3.1 Deterministic models

The global SA methods are applicable to deterministic computer codes, codes for which the same set of input parameters always leads to the same response value. This is considered by the following model:

$$\begin{aligned} f: \quad \mathbb{R}^p &\to \mathbb{R} \\ \mathbf{X} &\mapsto Y = f(\mathbf{X}) \end{aligned} \tag{5}$$

where Y is the output, $\mathbf{X} = (X_1, \ldots, X_p)$ are p independent inputs, and f is the model function, which is analytically not known. In this section, let us recall some basic ideas on Sobol sensitivity indices applied on this model.

Among quantitative methods, variance-based methods are the most often used (Saltelli et al. [26]). The main idea of these methods is to evaluate how the variance of an input or a group of inputs contributes into the variance of output. We start from the following variance decomposition:

$$\mathrm{Var}\left[Y\right] = \mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(Y|X_i\right)\right] \ , \tag{6}$$

which is known as the total variance theorem. The first term of this equality, named variance of the conditional expectation, is a natural indicator of the importance of X_i into the variance

of Y: the greater the importance of X_i, the greater is $\mathrm{Var}[\mathbb{E}(Y|X_i)]$. Most often, this term is divided by $\mathrm{Var}[Y]$ to obtain a sensitivity index in $[0,1]$.

To express the sensitivity indices, we use the unique decomposition of any integrable function on $[0,1]^p$ into a sum of elementary functions (see for example Sobol [28]):

$$f(X_1,\cdots,X_p) = f_0 + \sum_{i}^{p} f_i(X_i) + \sum_{i<j}^{p} f_{ij}(X_i,X_j) + \ldots + f_{12..p}(X_1,\cdots,X_p)\ , \tag{7}$$

where f_0 is a constant and the other functions verify the following conditions:

$$\int_0^1 f_{i_1,\ldots,i_s}(x_{i_1},\ldots,x_{i_s})\,dx_{i_k} = 0 \ \ \forall k = 1,\ldots,s\ ,\ \ \forall\{i_1,\ldots,i_s\} \subseteq \{1,\ldots,p\}\ . \tag{8}$$

Therefore, if the X_is are mutually independent, the following decomposition of the model output variance is possible (Sobol [28]):

$$\mathrm{Var}\left[Y\right] = \sum_{i}^{p} V_i(Y) + \sum_{i<j}^{p} V_{ij}(Y) + \sum_{i<j<k}^{p} V_{ijk}(Y) + \ldots + V_{12..p}(Y)\ , \tag{9}$$

where $V_i(Y) = \mathrm{Var}[\mathbb{E}(Y|X_i)]$, $V_{ij}(Y) = \mathrm{Var}[\mathbb{E}(Y|X_iX_j)] - V_i(Y) - V_j(Y), \ldots$ One can thus defines the sensitivity indices by:

$$S_i = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)} = \frac{V_i(Y)}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{V_{ij}(Y)}{\mathrm{Var}(Y)}, \quad S_{ijk} = \frac{V_{ijk}(Y)}{\mathrm{Var}(Y)}, \quad \ldots \tag{10}$$

These coefficients are called the Sobol indices, and can be used for any complex model functions f. The second order index S_{ij} expresses sensitivity of the model to the interaction between the variables X_i and X_j (without the first order effects of X_i and X_j), and so on for higher orders effects. The interpretation of these indices is natural as their sum is equal to one (thanks to equation (9)): the larger and close to one an index value, the greater is the importance of the variable or the group of variables linked to this index.

For a model with p inputs, the number of Sobol indices is $2^p - 1$; leading to an intractable number of indices as p increases. Thus, to express the overall sensitivity of the output to an

input X_i, Homma & Saltelli [7] introduce the total sensitivity index:

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \sum_{j \neq i, k \neq i, j<k} S_{ijk} + \ldots = \sum_{l \in \#i} S_l \ , \tag{11}$$

where $\#i$ represents all the "non-ordered" subsets of indices containing index i. Thus, $\sum_{l \in \#i} S_l$ is the sum of all the sensitivity indices containing i in their index. For example, for a model with three input parameters, $S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}$.

The estimation of these indices can be done by Monte-Carlo simulations (Sobol [28], Saltelli [25]) or by FAST method (Saltelli et al. [27]). Recent algorithms have also been introduced to reduce the number of required model evaluations significantly. As explained in the introduction, an alternative method consists in replacing complex computer models by metamodels which have negligible calculation time. Estimation of Sobol indices by Monte-Carlo techniques with their confidence intervals (requiring thousand of simulations) can then be done using these response surfaces. In practice, when the model has a great number of input parameters, only the first order and total Sobol indices are estimated.

3.2 Stochastic models

In this work, models containing some intrinsic alea, which is described as an uncontrolable random input parameter ε, are called "stochastic computer models". Similarly from equation (5), consider the following (stochastic) model:

$$\begin{array}{rccl} g: & \mathbb{R}^p & \to & \mathbb{R} \\ & \mathbf{X} & \mapsto & Y = f(\mathbf{X}) + \nu(\varepsilon, \mathbf{X} : \varepsilon) \end{array} \tag{12}$$

where $\mathbf{X}$ are the p controlable input parameters (independent random variables), Y is the output, f is the deterministic part of the model function and ν is the stochastic part of the model function. ν is considered to be centered: $\mathbb{E}(\nu) = 0$. The notation $\nu(\varepsilon, \mathbf{X} : \varepsilon)$ means that ν depends only on ε and on the interactions between ε and $\mathbf{X}$. The additive form of equation (12) is deduced directly from the decomposition of the function g into a sum of elementary functions depending on $(\mathbf{X}, \varepsilon)$ (like the decomposition in Eq. (7)).

The joint model introduced in section 2 enables us to recover two GLMs or two GAMs:

$$Y_m = \mathrm{E}(Y|\mathbf{X}) \;=\; \mu \tag{13}$$

by the mean component (Eq. (1)), and

$$Y_d = \mathrm{Var}(Y|\mathbf{X}) = \phi v(\mu) \tag{14}$$

by the dispersion component (Eq. (2)). If there is no uncontrolable parameter ε, it leads to a deterministic model case with $Y_d = \mathrm{Var}(Y|\mathbf{X}) = 0$. By using the total variance theorem (Eq. (6)), the variance of the output variable Y can be decomposed by:

$$\mathrm{Var}\left[Y(\mathbf{X},\varepsilon)\right] = \mathrm{Var}\left[\mathbb{E}\left(Y|\mathbf{X}\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(Y|\mathbf{X}\right)\right] = \mathrm{Var}\left(Y_m\right) + \mathbb{E}\left(Y_d\right) \;. \tag{15}$$

According to model (12), Y_m is the deterministic model part, and Y_d is the variance of the stochastic model part:

$$\begin{aligned} Y_m &= f(\mathbf{X}) \,, \\ Y_d &= \mathrm{Var}\left[\nu(\varepsilon,\mathbf{X}:\varepsilon)|\mathbf{X}\right] \end{aligned} \tag{16}$$

The variances of Y and Y_m are now decomposed according to the contributions of their input parameters $\mathbf{X}$. For Y, the same decomposition than for deterministic models holds (Eq. (9)). However, it includes the additional term $\mathbb{E}\left(Y_d\right)$ (the mean of the dispersion component) deduced from equation (15). Consequently,

$$\mathrm{Var}(Y) = \sum_i^p V_i(Y) + \sum_{i<j}^p V_{ij}(Y) + \sum_{i<j<k}^p V_{ijk}(Y) + \ldots + V_{12..p}(Y) + \mathbb{E}\left(Y_d\right) \;. \tag{17}$$

For the mean component Y_m, we have

$$\mathrm{Var}(Y_m) = \sum_i^p V_i(Y_m) + \sum_{i<j}^p V_{ij}(Y_m) + \sum_{i<j<k}^p V_{ijk}(Y_m) + \ldots + V_{12..p}(Y_m) \;. \tag{18}$$

By noticing that

$$V_i(Y_m) = \mathrm{Var}[\mathbb{E}(Y_m|X_i)] = \mathrm{Var}\{\mathbb{E}[\mathbb{E}(Y|\mathbf{X})|X_i]\} = \mathrm{Var}[\mathbb{E}(Y|X_i)] = V_i(Y) \ , \tag{19}$$

and from equation (10), the sensitivity indices for the variable $Y(\mathbf{X}, \varepsilon)$ according to the controlable parameters $\mathbf{X} = (X_i)_{i=1\ldots p}$ can be computed using:

$$S_i = \frac{V_i(Y_m)}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{V_{ij}(Y_m)}{\mathrm{Var}(Y)}, \quad \ldots \tag{20}$$

These Sobol indices can be computed by classical Monte-Carlo techniques, the same ones used in the deterministic model case. These algorithms are applied on the metamodel defined by the mean component Y_m of the joint GLM or the joint GAM.

Thus, all terms contained in $\mathrm{Var}(Y_m)$ of the equation (15) have been considered. It remains to estimate $\mathbb{E}(Y_d)$ by a simple numerical integration of Y_d following the law of $\mathbf{X}$. Y_d is evaluated with a metamodel, for example the dispersion component of the joint GLM or joint GAM. $\mathbb{E}(Y_d)$ includes all the decomposition terms of $\mathrm{Var}(Y)$ (according to $\mathbf{X}$ and ε) not taken into account in $\mathrm{Var}(Y_m)$ i.e. all terms involving ε. Therefore, the total sensitivity index of ε is

$$S_{T_\varepsilon} = \frac{\mathbb{E}(Y_d)}{\mathrm{Var}(Y)} \ . \tag{21}$$

As Y_d is a positive random variable, positivity of S_{T_ε} is guaranteed. In practice, $\mathrm{Var}(Y)$ can be estimated from the data or from simulations of the fitted joint model:

$$\mathrm{Var}(Y) = \mathrm{Var}(Y_m) + \mathbb{E}(Y_d) \ . \tag{22}$$

If $\mathrm{Var}(Y)$ is computed from the data, it seems preferable to estimate $\mathbb{E}(Y_d)$ with $\mathrm{Var}(Y) - \mathrm{Var}(Y_m)$ to satisfy equation (15). In our applications, the total variance will be estimated using the fitted joint model (Eq. (22)).

Finally, let us note that it is not possible to quantitatively distinguish the various contributions in S_{T_ε} (S_ε, $S_{i\varepsilon}$, $S_{ij\varepsilon}$, $\ldots$). However, the analysis of the terms in the regression model Y_d and their t-value study give qualitative contributions. For example, if an input parameter

X_i is not present in Y_d, we can deduce the following correct information: $S_{i\varepsilon} = 0$. Moreover, if the t-values analysis and the deviance analysis show that an input parameter X_i has a smaller influence than another input parameter X_j , we can suppose that the interaction between X_i and ε is less influential than the interaction between X_j and ε. Therefore, giving this kind of information is an improvement compared to the Tarantola's method (Tarantola et al. [29], see introduction).

In conclusion, this new approach, based on joint models to compute Sobol sensitivity indices, is useful if the following conditions hold:

- if the computer model contains some uncontrolable parameters (the model is no more deterministic but stochastic);
- if a metamodel is needed due to large CPU times of the computer model;
- if some of the uncontrolable parameters interact with some controlable input ones;
- if some information about the influence of the interactions between the uncontrolable parameters and the other input parameters is of interest.

4 APPLICATIONS

4.1 An analytic test case: the Ishigami function

The proposed method is first illustrated on an artificial analytical model with 3 input variables, called the Ishigami function (Homma & Saltelli [7], Saltelli et al. [26]):

$$Y = f(X_1, X_2, X_3) = \sin(X_1) + 7\sin(X_2)^2 + 0.1X_3^4 \sin(X_1) \ , \tag{23}$$

where $X_i \sim \mathcal{U}[-\pi; \pi]$ for $i = 1, 2, 3$. For this function, all the Sobol sensitivity indices (S_1, S_2, S_3, S_{12}, S_{13}, S_{23}, S_{123}, S_{T_1}, S_{T_2}, S_{T_3}) are known. This function is used in most intercomparison studies of global sensitivity analysis algorithms. In our study, the classical problem is altered by considering X_1 and X_2 as the controlable input random variables, and X_3 as an uncontrolable input random variable. It means that the X_3 random values are not

used in the modeling procedure; this parameter is considered to be inaccessible. However, sensitivity indices have the same theoretical values as in the standard case.

For the model fitting, 1000 samples of (X_1, X_2, X_3) were simulated leading to 1000 observations for Y. The GLM and GAM (with their relative joint extensions) are compared to the Gaussian Process (GP) model including or not the additive error component (the nugget effect). The fitting methodology is the one proposed by Marrel et al. [15] (based on the Welch et al. [33] sequential algorithm) which contains a linear regression component and a GP defined by a generalized exponential covariance. To compare the predictivity of different metamodels, we use the predictivity coefficient Q_2, which is the determination coefficient R^2 computed from a test sample (composed here by 10000 randomly chosen points). For the joint model, Q_2 is computed on the mean component.

4.1.1 Metamodeling

Simple GLM

First, a fourth order polynomial for the GLM is considered. Only the explanatory terms are selected in our regression model using analysis of deviance and the Fisher statistics. The Student test on the regression coefficients and residuals graphical analysis make it possible to judge the model quality. For a simple GLM fitting, one obtains

$$Y = 1.92 + 2.69X_1 + 2.17X_2^2 - 0.29X_1^3 - 0.29X_2^4 \, . \tag{24}$$

The explained deviance of this model is $D_{\text{expl}} = 61.3\%$. The predictivity coefficient is of the same order: $Q_2 = 60.8\%$. We see that it remains 39% of non explained deviance due to the model inadequacy and/or to the uncontrolable parameter.

Joint GLM

One tries to model the data by a joint GLM. The mean component gives the same model (24) as the simple GLM. For the dispersion component, using analysis of deviance techniques, no significant explanatory variable was found. Thus, the dispersion component is supposed to be constant; and the joint GLM is equivalent to the simple GLM approach - but with a

different fitting process. In addition, as one obtains the same explained deviance value as the simple GLM one, it corroborates the joint GLM approach relevance - even for a homoscedastic parameterization.

Simple GAM

We will be now studying the non parametric modeling. A simple GAM gives the following result:

$$Y = 3.76 - 2.67X_1 + s(X_1) + s(X_2) \ , \tag{25}$$

where $s(\cdot)$ is a spline term and where we have kept some parametric terms by applying a term selection procedure. The explained deviance of this model is $D_{\text{expl}} = 76.8\%$: the simple GAM approach clearly outperforms the simple GLM one. Even if this is obviously related to an increasing number of parameters, it is also explained by the fact that GAMs are more adjustable than GLMs: the number of parameters remains very small compared to the data size (1000). This is confirmed by the value of the predictivity coefficient $Q_2 = 75.1\%$ which is very close to the explained deviance (76.8%).

GP model

Let's now compare this GAM with the popular GP metamodel. Without introducing any nugget effect, the obtained GP gives $Q_2 = 72.8\%$. By introducing of the nugget effect (additional error with constant variance), the obtained GP gives $Q_2 = 74.3\%$. Consequently, the GP model including a nugget effect is similar to the simple GAM one. The variance of the nugget effect is estimated to 10% of the total variance, when one expects to obtain the residual variance: $1 - Q_2 = 25.7\%$. We will be discussing in the following section the consequence of this wrong estimation.

Joint GAM

One models now the data by a joint GAM. The resulting model is described by the following features:

$$\begin{aligned} Y_m &= 3.75 - 3.06X_1 + s(X_1) + s(X_2) \ , \\ Y_d &= 0.59 + s(X_1) \ . \end{aligned} \tag{26}$$

The explained deviances are $D_{\text{expl}} = 92.8\%$ for the mean component and $D_{\text{expl}} = 36.7\%$ for the dispersion component. The predictivity coefficient of the mean component is $Q_2 = 75.5\%$,

which is slightly better than the simple GAM and GP results.

Discussion

The explained deviance given by the joint GAM mean component is clearly larger than the one given by the simple GAM approach. This last point demonstrates the efficiency of the joint modeling of the mean and dispersion approach when heteroscedasticity is involved. Indeed, the joint procedure leads to suited prior weights for the mean component. The joint GAM improves both the joint GLM, simple GAM and GP approaches:

(a) due to the GAMs flexibility, the explanatory variable X_1 is identified to model the dispersion component (the interaction between X_1 and the uncontrolable parameter X_3 is therefore retrieved);

(b) the joint GAM explained deviance (93%) for the mean component is clearly larger than the simple GAM and joint GLM ones (Joint GLM: 61%, simple GAM: 77%).

Figure 1 shows the observed response versus the predicted values for the three models Joint GLM, Simple GAM and Joint GAM. In the following graphical analyses, we restrict our attention to these three models, and not to the GP model. Indeed, comparisons are not possible with the GP model because it interpolates the observed responses and the observed residuals are worth zero. Even if the nugget effect introduction allows to obtain non zero residuals, it is not appropriate to perform a statistical analysis of these residuals and a comparison with another model residuals.

On one hand, the advantage of the GAM approaches is visible in the Figure 1 as the dispersion around the $y = x$ line is clearly reduced. On the other hand, Figure 2 shows that the deviance residuals for the mean component of the joint GAM seem to be more homogeneously dispersed around the x-axis; leading to a better prediction on the whole range of the observations. Thus, the joint GAM approach is the most competitive model.

Figure 3 shows the proportion Δ of observations that lie within the $\alpha\%$ theoretical confidence interval in function of the confidence level α. By definition, if a model is suited for both mean and dispersion modelings, the points should be located around the $y = x$ line. As a consequence, this plot is useful to quantify the goodness of fitting accuracy of the models. Figure 3 shows that joint GLM approach is the most accurate model. The joint GAM is less

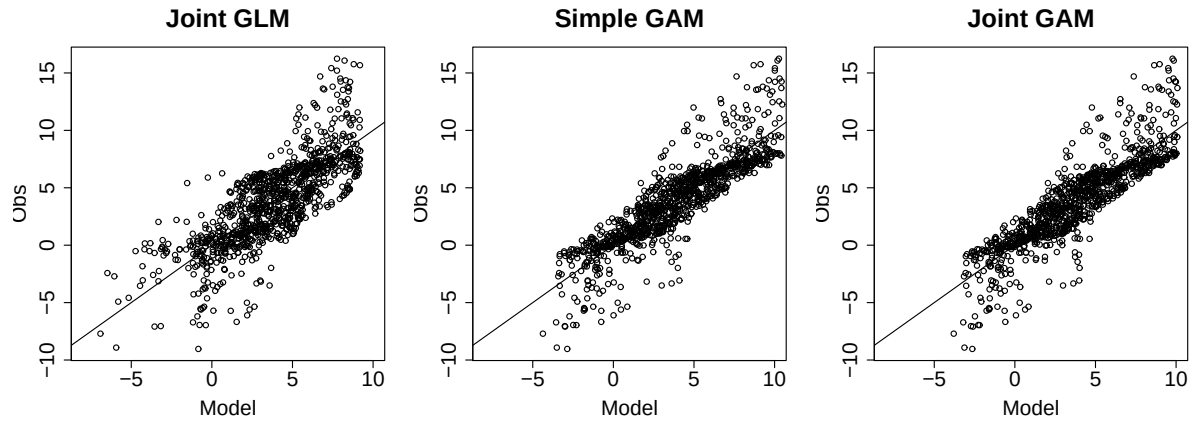

Figure 1: Observed response variable versus the predicted values for the three models: Joint GLM, Simple GAM, Joint GAM (Ishigami application).

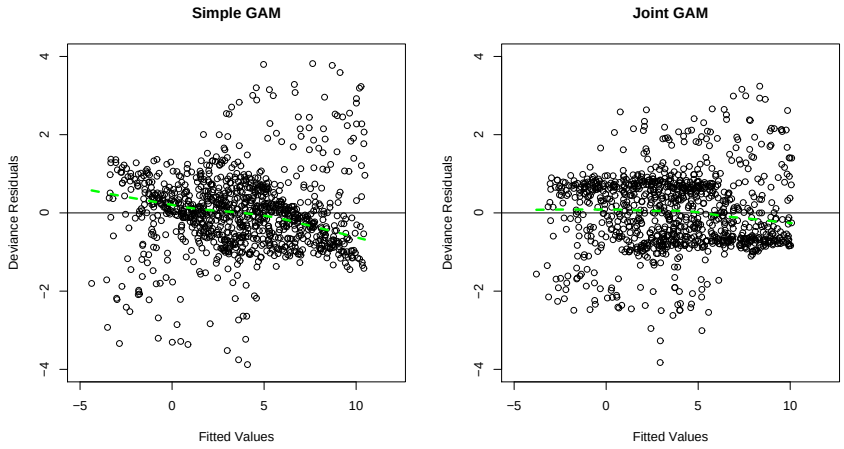

Figure 2: Deviance residuals for the Simple and Joint GAMs versus the fitted values (Ishigami application). Dashed lines correspond to local polynomial smoothers.

relevant but has a homogeneous dispersion around the $y = x$ line. The simple GAM approach systematically lead to overestimations. In particular, it means that the variance, supposed to be a constant, is overestimated and that the dispersion is poorly predicted.

Lastly, from Figures 1–3, the joint GAM seems to be the most competitive one. Indeed, the GAM flexibility allows to model accurately the mean component while the dispersion seems to be correctly modeled.

4.1.2 Sobol indices

Table 1 depicts the Sobol sensitivity indices for the joint GLM and the joint GAM using equations (20) and (21). The standard deviation estimates (sd) are obtained from 100 repe-

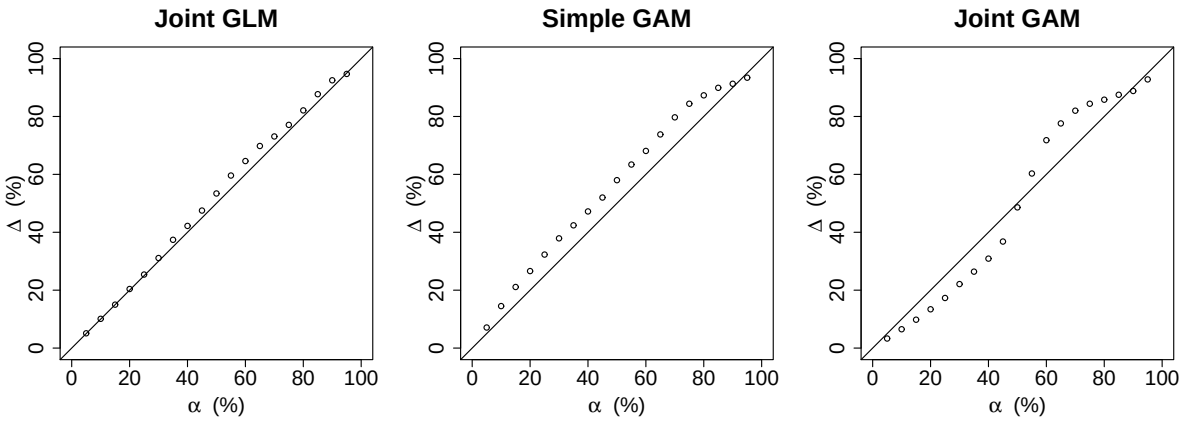

Figure 3: Proportion Δ (in percent) of observations that lie within the $\alpha\%$ theoretical confidence interval in function of the confidence level α (Ishigami application).

titions of the Monte-Carlo estimation procedure (which uses 10^4 model computations for one index estimation). When this Monte-Carlo procedure is used to estimate the Sobol index, we report "MC" in the "Method" column; while "Eq" indicates that the sensitivity indices have been deduced from the joint model regressive equations. Therefore, no estimation errors (sd) are associated to these indices (except for total indices S_{T_i} which can be deduced from S_i). When no quantitative deduction on the sensitivity index can be made with this process, the three column values are marked with the symbol "—".

The joint GLM gives only a good estimation of S_1, while S_2 and S_{T_3} are badly estimated (errors greater than 30%). S_{12} is correctly put to zero by looking directly at the joint GLM mean component formula (the same as the equation (24)). However, some conclusions drawn from the GLM dispersion component formula (which is a constant) are wrong. As no explanatory variable is involved in this formula, the deduced interaction indices are equal to zero: $S_{13} = S_{23} = S_{123} = 0$. Thus, $S_3 = S_{T_3} = 0.366$ while the correct values of S_3 and S_{T_3} are respectively zero and 0.243.

Contrary to the joint GLM, the joint GAM gives good approximations of all the Sobol indices (errors smaller than 7%), including S_{T_3}. Moreover, the deductions drawn from the model formulas (25) are correct ($S_{T_2} = S_2$, $S_{12} = S_{23} = S_{123} = 0$). The only drawback of this method is that some indices remain unknown due to the non separability of the dispersion component effects. However, it can be deduced that S_{13} is non null due to the explicative

effect of X_1 in the dispersion component.

Table 1 gives the Sobol indices computed by the same Monte-Carlo procedure using two classical metamodels as the simple GAM and the GP models. To estimate the first order Sobol indices $S_i = V_i(Y_m)/\mathrm{Var}(Y)$ (for $i = 1, 2$), the metamodel is used to compute $V_i(Y_m)$ and the fitted data (the 1000 observations of Y) to compute $\mathrm{Var}(Y)$. To estimate the total sensitivity index S_{T_3} of the uncontrolable parameter, the metamodel predictivity coefficient $Q_2 = 0.751$ is used.

In fact, by supposing that the metamodels fit correctly the computer code, one deduces that all the unexplained part of these metamodels is due to the uncontrolable parameter: $S_{T_3} = 1 - Q_2$. This is a strong hypothesis, which is verified here due to the simplicity of the analytical function. However, it will not be satisfied for all application cases. Moreover in practical and complex situations, the Q_2 estimation (usually done by a cross-validation method) can be difficult and subject to caution. For the Ishigami function, S_1, S_2, S_{T_3} are correctly estimated. S_{12} can be deduced from the formula (25) for the simple GAM and estimated by Monte-Carlo method for the GP model. However, any other sensitivity indices can be proposed as no dispersion modeling is involved.

Table 1: Sobol sensitivity indices (with standard deviations) for the Ishigami function: exact and estimated values from joint GLM, joint GAM, simple GAM and GP model. "Method" indicates the estimation method: MC for the Monte-Carlo procedure, Eq for a deduction from the model equations and Q_2 for the deduction of the predictivity coefficient Q_2. "—" indicates that the value is not available.

Indices	Exact	Joint GLM			Joint GAM			Simple GAM			GP		
	Values	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	Method
S_1	0.314	0.314	4e-3	MC	0.325	5e-3	MC	0.333	6e-3	MC	0.328	7e-3	MC
S_2	0.442	0.318	5e-3	MC	0.414	5e-3	MC	0.441	6e-3	MC	0.442	7e-3	MC
S_{T_3}	0.244	0.366	2e-3	MC	0.261	2e-3	MC	0.249	—	Q_2	0.257	—	Q_2
S_{12}	0	0	—	Eq	0	—	Eq	0	—	Eq	0.004	8e-3	MC
S_{13}	0.244	0	—	Eq	> 0	—	Eq	—	—	—	—	—	—
S_{23}	0	0	—	Eq	0	—	Eq	—	—	—	—	—	—
S_{123}	0	0	—	Eq	0	—	Eq	—	—	—	—	—	—
S_{T_1}	0.557	0.314	4e-3	Eq	—	—	—	—	—	—	—	—	—
S_{T_2}	0.443	0.318	5e-3	Eq	0.414	5e-3	Eq	—	—	—	—	—	—
S_3	0	0.366	2e-3	Eq	—	—	—	—	—	—	—	—	—

Remark: *By using the GP model including a nugget effect, one can think that estimating the nugget effect would give an estimation of the total sensitivity index. In this example, the variance of the nugget effect has been estimated to* 10% *of the total variance, which is far*

from the exact value (24%). Other tests (not presented here) have shown that the estimation of this nugget effect is not robust. The same problems arise during the estimation of the hyperparameters of the GP covariance function (Fang et al. [3], Marrel et al. [15]). This is caused by a difficult optimization step while fitting the GP model.

In conclusion, this example shows that the joint models, and specially the joint GAM, can adjust complex heteroscedastic situations for which classical metamodels are inadequate. Moreover, the joint models offer a theoretical basis to compute global sensitivity indices of stochastic models.

4.2 Application to an hydrogeologic transport code

This methodology is now applied to a complex industrial model of radioactive pollutants transport in saturated porous media using the MARTHE computer code (developed by BRGM, France). In the context of an environmental impact study, MARTHE has been applied to a model of strontium 90 (^{90}Sr) transport in saturated media for a radwaste temporary storage in Russia (Volkova et al. [31]). Only a partial characterization of the site has been made and, consequently, values of the model input parameters are not known precisely: 20 scalar input parameters have been considered as random variables, each of them associated to a specified probability density function. The model output variables of interest concern the ^{90}Sr concentration values in different spatial locations. One of the main goals of this study is to identify the most influent parameters of the computer code in order to improve the characterization of the site in a judicious way. Because of large computing times of the MARTHE code, the Sobol sensitivity indices are computed using metamodels (boosting regression trees model for Volkova et al. [31] and Gaussian Process model for Marrel et al. [15]).

As a perspective of their work, Volkova et al. [31] propose to study more precisely the influence of the spatial form of an hydrogeologic layer. It consists in performing a geostatistical simulation of this layer (which is a two-dimensional spatial random field), before each calculation of the computer model. This geostatistical simulation is rather complex and the resulting spatial field cannot be summarized by a few scalar values. Therefore, as explained in our introduction, the hydrogeologic layer has to be considered as an uncontrolable param-

eter of the computer model. Additionally to the uncontrolable parameter, 16 scalar input parameters remain uncertain and are treated as random variables. It concerns the permeability of different geological layers, the longitudinal and transversal dispersivity coefficients, the sorption coefficients, the porosity and meteoric water infiltration intensities.

The Latin Hypercube Sampling method is used to obtained a sample of 298 random vectors (each of dimension 16). In addition, 298 independent realizations of the spatial random field (noticed by ε) are obtained by a specific geostatistical simulation algorithm. This leads to obtain 298 observations (after 8 days of calculations) of the output variable of the MARTHE model (^{90}Sr concentration at the domain center). For the GLMs and GAMs construction phase, the large data dispersion suggests the use of logarithmic link functions for g and h (see Eqs (1) and (2)). Due to the large number of inputs, a manual term selection process has been applied. No interaction term has been found to be explicative in the GLMs. However, a bi-dimensional spline term has been added in the GAMs because of convincing deviance contribution and negligible p-value. One synthesizes the results by giving the explained deviance and the explanatory terms involved in the formulas:

- Simple GLM: $D_{\text{expl}} = 60\%$ with the terms $kd1$, $kd2$, $per1$, $per2$.
- Joint GLM: $D_{\text{expl}}(\text{mean}) = 66.4\%$, with the same terms than the simple GLM, $D_{\text{expl}}(\text{dispersion}) = 8.7\%$ with the terms $kd1$ and $per3$.
- Simple GAM: $D_{\text{expl}} = 81.8\%$ with $s(kd1)$, $s(kd2)$, $s(per3)$, $s(per2, kd2)$.
- Joint GAM: $D_{\text{expl}}(\text{mean}) = 98.1\%$ with the same terms than the simple GAM, $D_{\text{expl}}(\text{dispersion}) = 29.7\%$ with $kd1$, $kd2$.
- GP model: the regression and covariance parts include the terms $kd1$, $kd2$, $per1$, $per2$, $per3$. The nugget effect is estimated to 21.1% of the total variance, which shows that the GP model explains 79.9% of the total variance.

$kd1$, $kd2$ and $per1$, $per2$, $per3$ are respectively the sorption coefficients and the permeabilities of the different hydrogeologic layers. One observes that the GAM models outperform the GLM ones. The predictivity coefficient (computed by the leave-one-out method) of the simple GAM

gives $Q_2 = 76.4\%$, while for the simple GLM $Q_2 = 58.8\%$. The GP model is slightly more efficient than the simple GAM ($Q_2 = 80.4\%$). This small improvement may be due to a larger flexibility of GPs and to the specific fitting procedure (Marrel et al. [15]), which is suited to large dimensional problems (16 input parameters here).

Figure 4 shows the deviance residuals versus the fitted values for the joint GLM, simple GAM and joint GAM models. As for the Ishigami application, GP model residuals cannot be compared to these three models residuals. For the joint GLM approach, some outliers are not visible to keep the figure readable. As a consequence, the GAMs clearly lead to smaller residuals. Moreover, the joint GAM outperforms the simple GAM due to the right explanation of the dispersion component. It can be seen that the joint GAM allows to suppress the bias involved by the heteroscedasticity, while simple GAM residuals are affected by this bias. Figure 5 shows the observed values versus the predicted ones. This figure confirms the conclusions drawn from the Figure 4. Indeed, the GAM's flexibility allows to suppress the bias for the smallest data values.

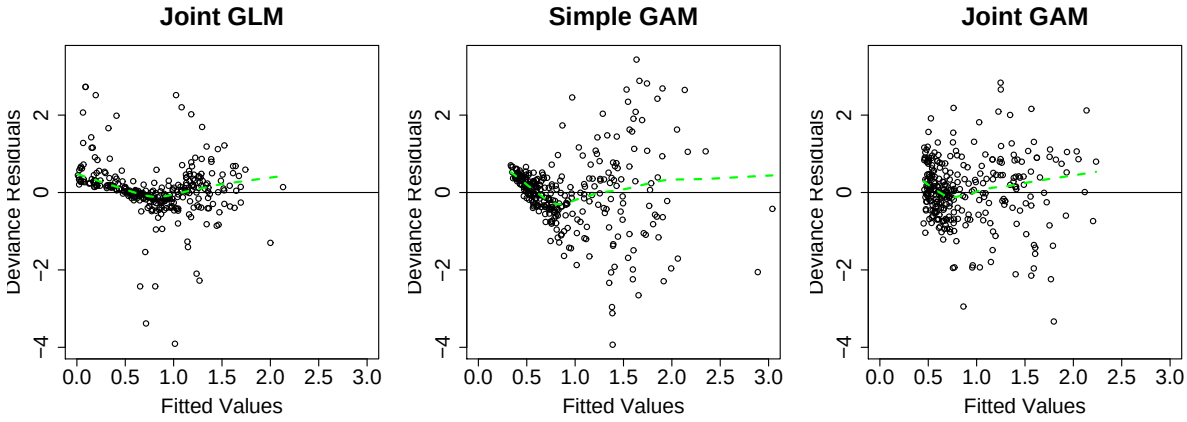

Figure 4: Deviance residuals (mean component) for the Simple GAM, Joint GAM and Joint GLM versus the fitted values (MARTHE application). Dashed lines correspond to local polynomial smoothers.

Figure 6 shows the proportion Δ of observations that lie within the $\alpha\%$ theoretical confidence interval versus the confidence interval α. It can be seen that the joint GAM is clearly the most accurate model. Indeed, all its points are close to the theoretical $y = x$ line, while the joint GLM (resp. simple GAM) systematically leads to underestimations (resp. overestimations). Consequently, from the Figures 4–6, one deduces that the joint GAM model is

the most competitive one. On one hand, the mean component is modeled accurately without any bias. On the other hand, the dispersion component is competitively modeled leading to reliable confidence intervals.

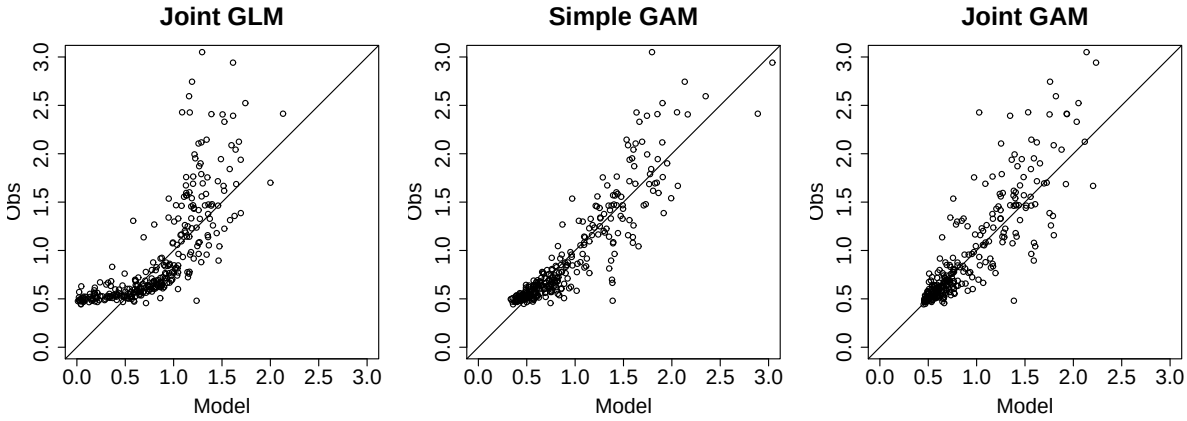

Figure 5: Observed response variable versus the predicted values for the three models: Joint GLM, Simple GAM, Joint GAM (MARTHE application)

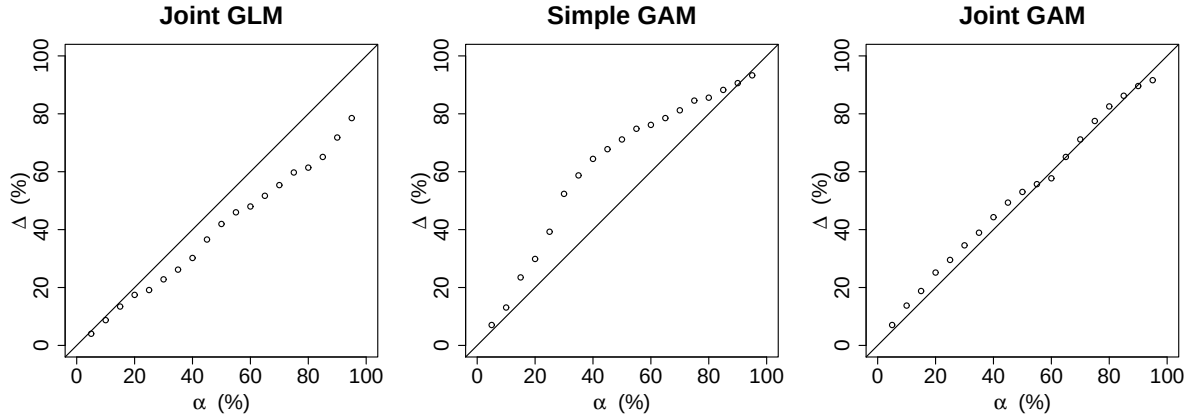

Figure 6: Proportion Δ (in percent) of observation that lie within the α% theoretical confidence interval in function of the confidence level α. MARTHE application.

Table 2 gives the main Sobol sensitivity indices for the joint GLM, joint GAM, simple GAM and GP models (using 10^4 model computations for one index estimation). The Sobol indices of the interactions between controlable parameters are not given (except between $kd2$ and $per2$) because these interactions are not included in the formulas of the two joint models. Therefore, their Sobol indices are zero. The two joint models give similar results for all first order sensitivity indices. The sorption coefficient of the second layer $kd2$ explained more than 52% of the output variance, while the permeability of the second layer $per2$ explained more

than 5%. Some large differences arise in the total influence of the uncontrolable parameter ε: 38.2% for the joint GLM and 27.7% for the joint GAM. Moreover, the joint GLM shows an influence of the interaction between $per3$ and ε, while the joint GAM shows an influence of the interaction between $kd2$ and ε. In this application, we consider the joint GAM results more reliable than the joint GLM ones because the joint GAM captures more efficiently the mean and dispersion components of the data than the joint GLM.

Table 2: Estimated Sobol sensitivity indices (with standard deviations obtained by 100 repetitions) for the MARTHE code. "Method" indicates the estimation method: MC for the Monte-Carlo procedure, Eq for a deduction from the model equations and Q_2 for the deduction of the predictivity coefficient Q_2. "—" indicates that the value is not available.

Indices	Joint GLM			Joint GAM			Simple GAM			GP		
	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	Method
S(kd1)	0.002	0.6e-2	MC	0.037	1.0e-2	MC	0.140	1.0e-2	MC	0.126	1.3e-2	MC
S(kd2)	0.522	0.6e-2	MC	0.524	1.0e-2	MC	0.550	1.1e-2	MC	0.603	0.9e-2	MC
S(per1)	0.018	0.7e-2	MC	0	—	Eq	0	—	Eq	0.012	1.1e-2	MC
S(per2)	0.052	0.6e-2	MC	0.078	1.0e-2	MC	0.044	1.0e-2	MC	0.048	1.2e-2	MC
S(per3)	0	—	Eq	0.005	1.0e-2	MC	0.008	1.0e-2	MC	0.003	1.1e-2	MC
S(kd2,per2)	0	—	Eq	0.063	1.0e-2	MC	0.026	1.0e-2	MC	0.021	1.4e-2	MC
$S_T(\varepsilon)$	0.382	0.2e-2	MC	0.277	0.3e-2	MC	0.235	—	Q_2	0.196	—	Q_2
S(kd1,ε)	> 0	—	Eq	> 0	—	Eq	—	—	—	—	—	—
S(kd2,ε)	0	—	Eq	> 0	—	Eq	—	—	—	—	—	—
S(per1,ε)	0	—	Eq	0	—	Eq	—	—	—	—	—	—
S(per2,ε)	0	—	Eq	0	—	Eq	—	—	—	—	—	—
S(per3,ε)	> 0	—	Eq	0	—	Eq	—	—	—	—	—	—

By comparing the joint GAM results with the simple GAM and GP model results, some significant differences can be printed out:

- The $kd1$ first order sensitivity index is overestimated using the simple GAM and GP model (14.0% and 12.6% instead of 3.7% for the joint GAM). Indeed, the deviance analysis of the joint GAM dispersion component shows a high contribution of $kd1$, which means that the interaction between $kd1$ and the uncontrolable parameter is probably large. For a standard metamodel, like the simple GAM and GP models, this interaction is not found out and leads to a wrong estimation of the first order sensitivity index of $kd1$.

- For the simple metamodels, using the relation $S_T(\varepsilon) = 1 - Q_2$, the total sensitivity index of the uncontrolable parameter is underestimated: 23.5% (simple GAM) and 19.6% (GP model) instead of 27.7% (joint GAM). The classical metamodels tend to explain some

parts of the data which can be adequately included in the dispersion component of the joint GAM during the iterative fitting algorithm.

- Contrary to the other metamodels, the joint GAM allows to prove that only $kd1$ and $kd2$ interact with the uncontrolable parameter.

As a conclusion, these sensitivity analysis results will be very useful to the physicist or the modeling engineer during the model construction and calibration steps. In this specific application, the sensitivity analysis shows that the geometry of the second hydrogeological layer has a strong influence (up to 28%) on the predicted ^{90}Sr concentration. Therefore, an accurate modeling of this geometry, coupled with a better knowledge of the most influential parameter $kd2$, are the key steps to an important reduction of the model prediction uncertainties.

5 CONCLUSION

This paper proposes a solution to resolve the problem of uncertainty and sensitivity analyses on stochastic computer models (Kleijnen [12]). A natural solution is to model the mean and the dispersion of the code outputs by two explanatory models. The classical way is to separately build these models. In this paper, the use of the joint modeling is preferred. This theory, proposed by Pregibon [20] and extensively developed by Nelder [17], is a powerful tool to fit the mean and dispersion components simultaneously. Zabalza et al. [35] already applied this approach to model stochastic computer code. However, the behavior of some numerical models can be highly complex and non linear. In the present paper, some examples show the limit of this parametric joint model. Being inspired by Rigby & Stasinopoulos [23] who use non parametric joint additive models (restricted to Gaussian cases), we propose to use a more general framework using GAMs. Like GLMs, GAMs are a suited framework because it allows variable and model selections *via* quasi-likelihood function, classical statistical tests on coefficients and graphical displays.

An analytic case on the Ishigami function shows that the joint GAM is adapted to complex heteroscedastic situations where classical response surfaces are inadequate. Moreover, it offers a theoretical basis to compute Sobol sensitivity indices in an efficient way. The performance

of the Joint GAM approach was assessed on an industrial application. Compared to other methods, the modeling of the dispersion component allows to obtain a robust estimation of the total sensitivity index of the uncontrolable parameter, which leads to correct estimations of the first order indices of the controlable parameters. In addition, it reveals the influential interactions between the uncontrolable parameter and the other input parameters.

The joint GAM has proven its flexibility to fit complex data: we have obtained the same performance for its mean component as the powerful Gaussian Process model. Moreover, the analytical formulas available with the joint GAM are very useful to complete the sensitivity analysis results and to improve our model understanding and knowledge. Finally, the joint GAM can also serve in propagation uncertainty and reliability studies of complex models, with unquantifiable random input variables, to obtain some mean predictions with their confidence intervals.

For some applications, joint GAM could be inadequate, and other models can be proposed. For example, for Gaussian observations, Juutilainen & Röning [11] have used a neural network model for mean and dispersion. It is shown to be more efficient than joint GLM and joint additive models in a context of numerous explanatory variables (25) and of a large amount of data (100000). Moreover, in the joint GAM as in the joint GLM, only diagnosis tools to analyze separately the two components of the joint model are available. It would be very convenient in the future to have accurate tools to analyse the two components simultaneously.

In the whole, all statistical analysis were performed using the R software environment [21]. In particular, the following functions and packages were useful: the "glm" function to fit a simple GLM and the "mgcv" (Multiple Smoothing Parameter Estimation by GCV) package to fit a simple GAM. We also developed the "JointModeling" package to fit joint models (including joint GLM and joint GAM).

6 ACKNOWLEGMENTS

We would like to thank I. Zabalza-Mezghani for her help while studying the joint GLM and A. Antoniadis for the discussion about the use of non parametric GLMs. We are also grateful

to E. Volkova who has performed the MARTHE simulations, G. Pujol for some remarks and F. Serre and his daughter for the help with the english. Lastly, we thank the editor and the referees for their comments which significantly improved this paper.

References

[1] D. Bursztyn and D.M. Steinberg. Screening experiments for dispersion effects. In A. Dean and S. Lewis, editors, *Screening - Methods for experimentation in industry, drug discovery and genetics*. Springer, 2006.

[2] V.C.P. Chen, K-L. Tsui, R.R. Barton, and M. Meckesheimer. A review on design, modeling and applications of computer experiments. *IIE Transactions*, 38:273–291, 2006.

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

[4] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

[5] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1990.

[6] J.C. Helton, J.D. Johnson, C.J. Salaberry, and C.B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91:1175–1209, 2006.

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

[8] B. Iooss, C. Lhuillier, and H. Jeanneau. Numerical simulation of transit-time ultrasonic flowmeters due to flow profile and fluid turbulence. *Ultrasonics*, 40:1009–1015, 2002.

[9] B. Iooss and M. Ribatet. Analyse de sensibilité globale de modèles numériques à paramètres incontrôlables. In *Proceedings of 38èmes Journées de Statistique*, Clamart, France, May-June 2006.

[10] B. Iooss, F. Van Dorpe, and N. Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety*, 91:1241–1251, 2006.

[11] I. Juutilainen and J. Röning. A comparaison of methods for joint modelling of mean and dispersion. In *Proceedings of the 11th Symposium on ASMDA*, Brest, France, May 2005.

[12] J. Kleijnen. Sensitivity analysis and related analyses: a review of some statistical techniques. *Journal of Statistical Computation and Simulation*, 57:111–142, 1997.

[13] J. Kleijnen and W.C.M. van Beers. Robustness of kriging when interpolating in random simulation with heterogeneous variances: some experiments. *European Journal of Operational Research*, 165:826–834, 2005.

[14] Y. Lee and J.A. Nelder. Robust design via generalized linear models. *Journal of Quality Technology*, 35(1):2–12, 2003.

[15] A. Marrel, B. Iooss, F. Van Dorpe, and E. Volkova. An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics and Data Analysis*, submitted.

[16] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

[17] J.A. Nelder. A large class of models derived from generalized linear models. *Statistics in Medicine*, 17:2747–2753, 1998.

[18] J.A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74:221–232, 1987.

[19] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society A*, 135:370–384, 1972.

[20] D. Pregibon. Review of "Generalized Linear Models" by McCullagh and Nelder. *Annals of Statistics*, 12:1589–1596, 1984.

[21] R Development Core Team. R: A language and environment for statistical computing. 2006. ISBN 3-900051-07-0.

[22] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[23] R.A. Rigby and D.M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6:57–65, 1996.

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

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

[26] A. Saltelli, K. Chan, and E.M. Scott, editors. *Sensitivity analysis*. Wiley Series in Probability and Statistics. Wiley, 2000.

[27] A. Saltelli, S. Tarantola, and K. Chan. A quantitative, model-independent method for global sensitivity analysis of model output. *Technometrics*, 41:39–56, 1999.

[28] I.M. Sobol. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.

[29] S. Tarantola, N. Giglioli, N. Jesinghaus, and A. Saltelli. Can global sensitivity analysis steer the implementation of models for environmental assesments and decision-making? *Stochastic Environmental Research and Risk Assesment*, 16:63–76, 2002.

[30] G.G. Vining and R.H. Myers. Combining Taguchi and response-surface philosophies - a dual response approach. *Journal of Quality Technology*, 22:38–45, 1990.

[31] E. Volkova, B. Iooss, and F. Van Dorpe. Global sensitivity analysis for a numerical model of radionuclide migration from the rrc "kurchatov institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assesment*, in press.

[32] G. Wahba. *Spline models for observational data*. SIAM, 1990.

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

[34] S.N. Wood and N.H. Augustin. GAMs with inegrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling*, 157:157–177, 2002.

[35] I. Zabalza, J. Dejean, and D. Collombier. Prediction and density estimation of a horizontal well productivity index using generalized linear models. In *ECMOR VI, Peebles*, September 1998.

[36] I. Zabalza-Mezghani, E. Manceau, M. Feraille, and A. Jourdan. Uncertainty management: From geological scenarios to production scheme optimization. *Journal of Petroleum Science and Engineering*, 44:11–25, 2004.