EUROPEAN COMMISSION

Community research

PAMINA
Performance Assessment Methodologies in Application to Guide the Development of the Safety Case

(Contract Number: FP6-036404)

TREATMENT OF SPATIALLY DEPENDENT INPUT VARIABLES IN SENSITIVITY ANALYSIS OF MODEL OUTPUT METHODS MILESTONE (N°:M2.2.D.4)

Author: **Bertrand Iooss**
(Commissariat à l'Énergie Atomique)

Date of issue of this report : **17/03/08**

Start date of project : **01/10/2006** Duration : 36 Months

Project co-funded by the European Commission under the Euratom Research and Training Programme on Nuclear Energy within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the partners of the [PAMINA] project	
CO	Confidential, only for partners of the [PAMINA] project	

Foreword

The work presented in this report was developed within the Integrated Project PAMINA: **P**erformance **A**ssessment **M**ethodologies **IN A**pplication to Guide the Development of the Safety Case. This project is part of the Sixth Framework Programme of the European Commission. It brings together 25 organisations from ten European countries and one EC Joint Research Centre in order to improve and harmonise methodologies and tools for demonstrating the safety of deep geological disposal of long-lived radioactive waste for different waste types, repository designs and geological environments. The results will be of interest to national waste management organisations, regulators and lay stakeholders.

The work is organised in four Research and Technology Development Components (RTDCs) and one additional component dealing with knowledge management and dissemination of knowledge:

- In RTDC 1 the aim is to evaluate the state of the art of methodologies and approaches needed for assessing the safety of deep geological disposal, on the basis of comprehensive review of international practice. This work includes the identification of any deficiencies in methods and tools.

- In RTDC 2 the aim is to establish a framework and methodology for the treatment of uncertainty during PA and safety case development. Guidance on, and examples of, good practice will be provided on the communication and treatment of different types of uncertainty, spatial variability, the development of probabilistic safety assessment tools, and techniques for sensitivity and uncertainty analysis.

- In RTDC 3 the aim is to develop methodologies and tools for integrated PA for various geological disposal concepts. This work includes the development of PA scenarios, of the PA approach to gas migration processes, of the PA approach to radionuclide source term modelling, and of safety and performance indicators.

- In RTDC 4 the aim is to conduct several benchmark exercises on specific processes, in which quantitative comparisons are made between approaches that rely on simplifying assumptions and models, and those that rely on complex models that take into account a more complete process conceptualization in space and time.

The work presented in this report was performed in the scope of RTDC 2.

All PAMINA reports can be downloaded from http://www.ip-pamina.eu.

Direction de l'Énergie Nucléaire
Direction du CEA/Cadarache
Département d'Études des Réacteurs
Service d'Études des Systèmes Innovants
Laboratoire de Conduite et Fiabilité des Réacteurs

CEA/DEN/CAD/DER/SESI/LCFR
DO 18 06/03/08
08PPMO000030

BORDEREAU D'ENVOI

Date : 17 MARS 2008 Réf. : CEA/DEN/CAD/DER/SESI/LCFR/NT DO 4 06/03/08

Désignation : Note Technique « Treatment of spatially dependent input variables in sensitivity analysis of model output methods »

Auteurs : B. IOOSS

Destinataires	Nombre	Observations
GALSON SCIENCES LTD		
A. KHURSHEED, D. GALSON	2	Transmission par courriel du document complet
UPV		
J.R. ILARI	1	
TUC		
K. ROEHLIG	1	
Institut Kurchatov		
E. VOLKOVA	1	
CEA/DEN/SACLAY		
DSOE/SIMULATION : B. BRUN	1	
DM2S/SFME/LGLS : V. BERGEAU	1	
DM2S/SFME/MTMS : A. GENTY	1	
CEA/DEN/CADARACHE		
DTN/SMTM/LMTE :		
C. TIFFREAU, F. JOURDAIN, A. MARREL	3	
DER/SESI : J-C. GARNIER	1	
DER/SESI/LCFR :		
F. BERTRAND, B. IOOSS	2	
M. MARQUES, N. PEROT	2	
DER/SESI/LCFR	1	1 exemplaire papier en circulation
DER	1	Transmission par courriel du résumé
DER/SPEX – SPRC – SRES – SSTH	4	
SESI/LCSI - LESA	2	

Le Chef du DER/SESI : Jean-Claude GARNIER

Commissariat à l'énergie atomique
Centre de Cadarache – Bât. 212 – 13108 Saint Paul Lez Durance cedex
Tél. : 33 – (0)4 42 25 30 05 – Fax : 33 – (0)4 42 25 24 08 – mail : nicolas.devictor@cea.fr

Etablissement public à caractère industriel et commercial
R.C.S. PARIS B 775 685 019

Proposal/Contract no.: **FP6-036404**

Project acronym: PAMINA

Project title: **PERFORMANCE ASSESSMENT METHODOLOGIES IN APPLICATION TO GUIDE THE DEVELOPMENT OF THE SAFETY CASE**

Instrument: Integrated Project

Thematic Priority: Management of Radioactive Waste and Radiation Protection and other activities in the field of Nuclear Technologies and Safety

RTDC2 - WP2.2.D – Spatial variability - CEA contribution

Due date of deliverable: 31.03.08
Actual submission date:

Start date of project: 01.10.2006
Duration: 36 months

Revision: 1

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination level

PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

TITLE: TREATMENT OF SPATIALLY DEPENDENT INPUT VARIABLES IN SENSITIVITY ANALYSIS OF MODEL OUTPUT METHODS

AUTHOR(S) : Bertrand IOOSS

ABSTRACT:

This report constitutes deliverable CEA/DEN/DER for the component RTDC 2 of European project PAMINA (Performance Assessment Methodologies in Application to Guide the Development of the Safety Case) 6th PCRD. This task concerns the treatment of the spatial variability of the geological media in the uncertainty and sensitivity analyses of computer codes used for the safety analyses of the deep waste storage facilities. In this report, one restricts to methodological aspects. We describe various techniques to perform global sensitivity analyses of numerical models with spatially variable input parameters. We are specially interested by the models which depend on geostatistical simulations, such as a heterogeneous field of permeability modelled by a random field. This problem is also seen within a more general mathematical framework : analysis of numerical models with functional inputs (random fields, stochastic processes, random sets, ...).

KEY WORDS: IMPACT CALCULATION, ENVIRONMENT, HYDROGEOLOGY, UNCERTAINTY, SENSITIVITY, GEOSTATISTICAL SIMULATION

CEA/DEN/CAD/DER/SESI/LCFR/NT
DO 4 06/03/08
08PPBT000007

DIRECTION DE L'ENERGIE NUCLÉAIRE
DIRECTION DU CENTRE DE CADARACHE
DÉPARTEMENT D'ÉTUDES DES RÉACTEURS
SERVICE D'ÉTUDES DES SYSTÈMES INNOVANTS
LABORATOIRE DE CONDUITE ET FIABILITE DES REACTEURS

NT	CEA/DEN/CAD/DER/SESI/LCFR/NT DO 4 06/03/08	0	5H62	A-SUPAR-02-03	DSOE/SIMUL/SUPAR	1/50
NATURE	CHRONO UNITÉ	INDICE	UNITÉ	ÉLÉMENT D'OTP	CLASSEMENT UNITÉ	PAGE

Note Technique

TITRE : CONSIDERATION DE VARIABLES D'ENTREE SPATIALEMENT HETEROGENES DANS LES METHODES D'ANALYSE DE SENSIBILITE DE SORTIES DE MODELES

TITLE : TREATMENT OF SPATIALLY DEPENDENT INPUT VARIABLES IN SENSITIVITY ANALYSIS OF MODEL OUTPUT METHODS

AUTEUR(S) : B. IOOSS

RESUME : Ce rapport constitue le livrable du CEA/DEN/DER pour le composant RTDC 2 du projet européen PAMINA (Performance Assessment Methodologies IN Application to guide the development of the safety case) du 6ème PCRD. Cette tâche concerne le traitement de la variabilité spatiale des milieux géologiques dans les analyses d'incertitudes et de sensibilité de codes de calcul utilisés pour les analyses de sûreté des installations de stockage profond. Dans ce rapport, on décrit uniquement les aspects méthodologiques, dont notamment différentes techniques pour effectuer des analyses de sensibilité globales sur des modèles numériques ayant des paramètres d'entrée spatialement variables. On s'intéresse plus particulièrement aux modèles qui dépendent de simulations géostatistiques, comme par exemple un champ de perméabilité hétérogène modélisé par un champ aléatoire. Ce problème est également vu dans un cadre mathématique plus général, celui des modèles numériques dépendant d'entrées fonctionnelles (qui peuvent être des champs aléatoires, mais aussi des processus stochastiques, des ensembles aléatoires, ...).

ABSTRACT : This report consists in deliverable CEA/DEN/DER for the component RTDC 2 of European project PAMINA (Performance Assessment Methodologies IN Application to guide the development of the safety case) 6th PCRD. This task concerns the treatment of the spatial variability of the geological media in the uncertainty and sensitivity analyses of computer codes used for the safety analyses of the deep waste storage facilities. In this report, one restricts to methodological aspects. We describe various techniques to perform global sensitivity analyses of numerical models with spatially variable input parameters. We are especially interested by the models which depend on geostatistical simulations, such as a heterogeneous field of permeability modelled by a random field. This problem is also seen within a more general mathematical framework : analysis of numerical models with functional inputs (random fields, stochastic processes, random sets, ...).

KEYWORDS : UNCERTAINTY, SENSITIVITY, METAMODEL, JOINT MODEL, FUNCTIONAL DATA, GEOSTATISTICAL SIMULATION, HYDROGEOLOGY, IMPACT CALCULATION

MOTS CLES : INCERTITUDE, SENSIBILITE, MODELE JOINT, DONNEES FONCTIONNELLES, SIMULATION GEOSTATISTIQUE, HYDROGEOLOGIE, CALCUL D'IMPACT

(04/02 - JA-Document1)

	Rédacteur	Vérificateur	Vérificateur Qualité	Approbateur
Fonction			IQ du SESI	Le Chef du SESI/LCFR
visa				
NOM	Bertrand IOOSS	Nadia PEROT	Florence JOYER	Frédéric BERTRAND
Date	10/03/08	10/03/08	13/03/08	11/03/08

Propriété Industrielle	Cadre de réalisation	Classification					Qualité
		DR	CC	CD	SD	Ss	
En l'absence d'accord ou de contrat, la diffusion des informations contenues dans ce document auprès d'un organisme tiers extérieur au CEA est soumise à l'accord du Chef de Département.	DOB.....DSOE Domaine .....Simulation Projet.....SUPAR Contrat.....					x	Document réalisé sous SMQ certifié AFAQ — ISO 9001

Indice	OBJET DES REVISIONS	DATE	RÉFÉRENCE GCAO
0	Emission initiale	11/03/08	CEA/DEN/CAD/DER/SESI/LCFR/NT DO 4 06/03/08

1. INTRODUCTION

This report consists in deliverable CEA/DEN/DER for the component RTDC 2 of European project PAMINA (Performance Assessment Methodologies IN Application to guide the development of the safety case) 6th PCRD. This task concerns the treatment of the spatial variability of the geological media in the uncertainty and sensitivity analyses of computer codes used for the safety analyses of the deep waste storage facilities. In this report, one restricts to methodological aspects.

The chapter 2 describes various techniques to perform global sensitivity analyses of numerical models with spatially variable input parameters. We are specially interested by the models which depend on geostatistical simulations, such as a heterogeneous field of permeability modelled by a random field. The application (called MARTHE) concerns the modelling of the radionuclide migration of a radioactive waste disposal site.

Chapters 3 and 4 deal with a more general mathematical framework : the uncertainty and sensitivity analyses of numerical models with functional inputs (as for example random fields, stochastic processes and random sets). We explain a new methodology, based on metamodel (i.e. response surface) construction, to deal with CPU time expensive computer code, containing functional inputs. In chapter 3, we apply it on the MARTHE application (a spatial random field input), while in chapter 4, we apply it on another computer model which contains a temporal stochastic process in its input variables.

2. SENSITIVITY ANALYSIS OF A HYDROGEOLOGICAL MODEL DEPENDENT ON GEOSTATISTICAL SIMULATIONS OF THE PERMEABILITY FIELD

2.1 ABSTRACT

This research focused on developing a transport model for ^{90}Sr in a saturated porous medium for an interim radwaste storage site in Moscow (Russia). A previous global sensitivity analysis of the model led to the conclusion that the calculated concentrations for the piezometers were mainly influenced by a) the distribution coefficient of ^{90}Sr in the different layers of the domain and b) the intensity of infiltration in the pipe leakage zones, and not by the hydrodynamic parameters (dispersivity, porosity, etc.)

The effect of the shape of the coarse sandy layer is investigated in this new study. A technique based on geostatistical simulation is being developed to simulate the shape of this layer. Standard techniques can no longer be used to deal with the sensitivity analysis of such a model, i.e. with random input fields. Different qualitative and quantitative analytical methods are proposed. These methods make it possible to quantify the effect of poorly understanding the shape of the coarse sand layer. The sensitivity analysis also makes it possible to detect other influential parameters for this new modelling technique. The precision on these influential parameters or on the shape of the layer can therefore be improved, thereby considerably decreasing the uncertainty on model prediction.

2.2 INTRODUCTION

This study follows on from the internship by E. Volkova carried out within the collaborative agreement between the Kurchatov Institute (Moscow, Russia) and the French Atomic Energy Commission (France). This aim of this internship was to develop a transport model for water-saturated porous media using the MARTHE software (BRGM[1]) for an interim radwaste storage site (STDR) in Moscow (Russia). The main objective was to model the transport of strontium-90, a mobile radioactive element found in soils – in the upper aquifer of the site. To begin with, modelling was performed during a pre-established measurement period (2002 – 2004) in order to compare the model results with the in-field measurements. For predictive purposes, modelling was then continued up to 2010 to determine the degree of potential contamination in the aquifer. In order to identify the code input parameters considered to have the most effect on the calculation result, statistical methods for uncertainty and sensitivity analyses were used (Law & Kelton [10], Helton *et al.* [6], Devictor & Bolado [4], Iooss [7]). Considering the complexity of the numerical model and the long CPU time, an intermediary phase was required to build the metamodels for a limited number of code simulations (Volkova *et al.* [14]). Additional research was also carried out to improve the metamodel building phase (Marrel *et al.* [11]).

This first study made it possible to reach the conclusion that the concentration in the piezometers, as modelled by MARTHE, was more influenced by the distribution coefficient of ^{90}Sr in the different layers of the domain and by the intensity of infiltration in the pipe leakage zones, than by the other model parameters. Therefore, the improved precision of these parameters would make it possible to considerably decrease uncertainty on the model prediction. Owing to this result, one of the lines of research suggested by this work consists in quantifying the influence of the shapes of certain less-permeable zones of the model. During the first study, these shapes (resulting from interpolation methods based on in-field data) were assumed to be unchanging. In the second study, we suggested varying them using geostatistical simulations.

[1] French public organisation specialised in geosciences

Research therefore aimed to study the effect of geostatistical simulations on variations in the output concentrations calculated by the MARTHE model. Input geostatistical simulations of numerical models were only briefly discussed in the sensitivity analyses, whether this be on a theoretical, methodological or practical level. The different solutions discussed in the literature are examined in looss & Ribatet [8] (see chapter 4) who came to the conclusion that these solutions cannot be applied when the model is non-linear, requires long CPU times, and has many uncertain input parameters and output variables. Furthermore, the solution recommended by looss *et al.* [9] (see chapter 3) based on a double metamodel has proven to be difficult to apply to this application. This is due to the input dimensions, the inconsistency in concentration values (some are very low while others are extremely high), and most probably the very strong influence of geostatistical simulations in relation to other scalar input parameters. Other approaches have to be considered, as the functional decomposition of the random field (Busby et al. [0]).

It was possible to foresee simpler methods in the second study owing to feedback from the first study. The only difference between the two models lies in the replacement of four random input parameters (permeability in different zones) by the simulation of the random field. The same sets of simulation runs were performed on the other random input parameters. This particularly made it possible to calculate this sensitivity of the model to the random field through comparisons with the results of the first model (without the random field).

This chapter will describe the new transport scenario based on the MARTHE code, using geostatistical simulations of the shape of zones. The fourth section will discuss the application of different techniques used to qualitatively and quantitatively analyse the effect of geostatistical simulations on model output variables. The fifth section discusses the use of metamodels and the associated results. A concise summary of all these results is given by way of conclusion.

2.3 SCENARIO WITH GEOSTATISTICAL SIMULATIONS

The construction of the first model for the STDR site (Volkova *et al.* [14]) required using the available data. The modelled hydrogeological aquifer was sub-divided into three layers representing the different geological media: two layers of fine-grained sand and an intermediary layer of coarse-grained sand. The thickness of each layer is variable and was calculated by interpolation of the lithologic data taking into account all the drill holes in the area. The data also revealed that the coarse sand layer was not found everywhere in the domain: there are drill holes that do not contain this layer. The shape of these zones void of coarse sand was interpolated using specialised software and the drill hole data. Figure 2.1 shows the limit according to a horizontal plane between the intermediary layer of coarse sand and the upper layer, making it possible to visualise the absence of this layer in certain places (blue zones).

The numerical model – based on MARTHE – consists in propagating a source term over a period from August 2002 to December 2010. The numerical diagram in question requires a CPU time of about 15 minutes for one simulation. A total of 300 different simulations were performed by randomly varying 20 scalar input parameters using a Monte-Carlo technique. The total CPU time added up to three full days. All these calculations were analysed in the first study (Volkova *et al.* [14]).

Our next objective was to vary the shape of the zones in which there was no coarse sand. A method of geostatistical simulations was established in order to do this. This method is based on a simulation method for categorical variables (called "indicator principle component approach") resulting from the "ipcsim" program of the GSLib software (Deutsch & Journel [3]). A spatial two-category variable was used to simulate the shape of the zones void of coarse sand. The first category corresponds to the presence of the layer of coarse sand, whereas the second indicates the zones void of coarse sand. A Gaussian-type variogram was chosen with a correlation length adapted to the heterogeneities that require simulating. The variogram was built on the basis of the coarse sand layer thickness observed in the 66 drill holes on the site. Furthermore, the simulations

were conditioned: for all simulated fields, the absence or presence of coarse sand corresponded to all the available observations. Two different simulations are given in Figure 2.2.

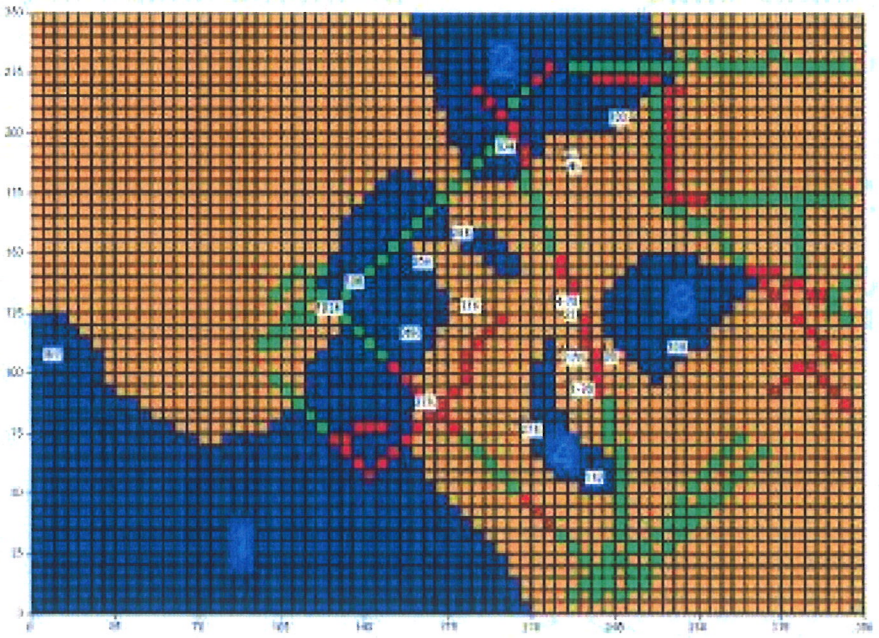

Figure 2.1: Materialisation of the model taking into account different zones in the second layer. Blue zones (numbered 1 to 4) – no coarse sand; Orange zones – presence of coarse sand; Green zones – moderate infiltration near the pipes; Red zones – high infiltration near the pipes; the other number correspond to the 20 piezometers.

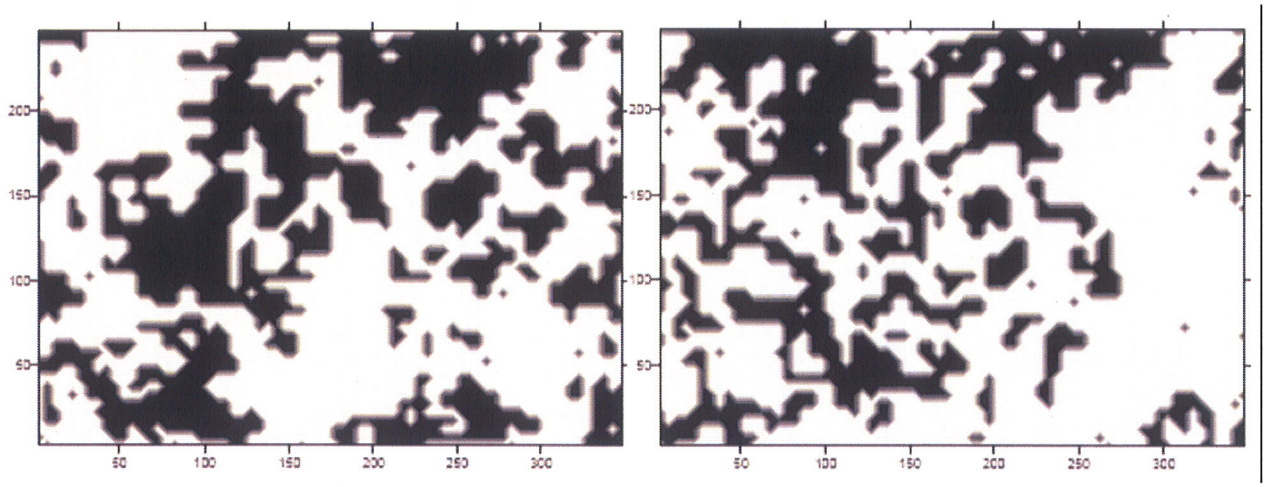

Figure 2.2: Examples of fields indicating the absence of the second layer (black zones).

For each calculation, a generated field was used in the matrix of permeability values for the second layer of the model. When a mesh corresponds to the first category (presence of coarse sand), the permeability of this grid cell is considered equal to the parameter per2 (characteristic value for coarse sand). When a mesh corresponds to the second category (no coarse sand), the permeability for this mesh is considered equal to the permeability of the first layer (upper layer) of the model (parameter per1). The same procedure was used to generate volume distribution coefficients (by using the parameters kd1 and kd2), as well as longitudinal (parameters d1 and d2) and transverse (parameters dt1 and dt2) dispersivity values for the second layer of the model.

The initial study involved 20 scalar input parameters. Seeing that the variable shapes were taken into account for the second layer, the parameterisation of the model must be reviewed. In the first study, 5 parameters were associated with the permeability of the model's second layer: perz1, perz2, perz3, perz4 – for the four blue zones (Figure 2.1) and per2 for the orange zone of the second layer. The value of per2 for zones with coarse sand was applied for the second layer (white zones in Figure 2.2), whereas per1 was applied for zones void of coarse sand (black zones in Figure 2.2). Four input parameters were therefore deleted. These are permeabilities for the different blue zones in Figure 2.1.
Table 2.1 provides a summary of all the scalar parameters considered uncertain in this new model:

- nominal values of parameters,
- distribution types,
- parameters of their distribution (minimum and maximum limits for the uniform distribution and the coefficients of the Weibull law). The parameters of the uniform distribution have been obtained by bibliographical analyses while the parameters of the Weibull distribution have been obtained from a probability distribution fitting procedure on experimental data.

	Parameters	Notation	Model valu	Distributio type	Interval or distribution parameters
1	Permeability of layer 1	per1	8	Uniform	1 - 15
2	Permeability of layer 2	per2	15	Uniform	5 - 20
3	Permeability of layer 3	per3	8	Uniform	1 - 15
8	Longitudinal dispersivity of layer 3	d1	0.8	Uniform	0.05 - 2
9	Longitudinal dispersivity of layer 3	d2	0.8	Uniform	0.05 - 2
10	Longitudinal dispersivity of layer 3	d3	0.8	Uniform	0.05 - 2
11	Transverse dispersivity of layer 1	dt1	0.08	Uniform	0.01 *d1 – 0.1*d1
12	Transverse dispersivity of layer 2	dt2	0.08	Uniform	0.01*d2 – 0.1*d2
13	Transverse dispersivity of layer 3	dt3	0.08	Uniform	0.01*d3 – 0.1*d3
14	Volumetric distribution coefficient c. 1	kd1	5.1	Weibull	1.1597, 19.9875
15	Volumetric distribution coefficient c. 2	kd2	0.34	Weibull	0.891597, 24.4455
16	Volumetric distribution coefficient c. 3	kd3	5.1	Weibull	1.27363, 22.4986
17	Porosity (all layers)	por	0.3	Uniform	0.3 – 0.37
18	Infiltration type 1	i1	0.0001	Uniform	0 – 0.0001
19	Infiltration type 2	i2	0.004	Uniform	i1 – 0.01
20	Infiltration type 3	i3	0.02	Uniform	i2 - 0,1

Table 2.1: Input parameters for the sensitivity analysis with nominal values, distribution type and distribution coefficients.

There are now 16 random scalar input parameters in addition to the spatial parameter for the shape of the second layer, taken into account by the geostatistical simulations. As in the first study, the output variables under investigation correspond to the calculated concentrations in 2010 for the different piezometers on the site, which results in a total of 20 output variables.

Owing to an increased heterogeneity in the medium where the radionuclide spreads, MARTHE calculations based on geostatistical simulations are much more CPU time-consuming than the previous calculations. The numerical scheme results in a CPU time of about 40 minutes for one simulation. For each simulation, the 16 scalar input parameters were varied randomly (Latin hypercube sampling (LHS) plan). In total, 300 simulations of the MARTHE model were performed, resulting in 8 full days of calculations.

In fact, the simulation sets for the 16 scalar input parameters are the same as those used in the previous study (excepting the four input parameters in excess). It will therefore be possible to calculate the correlation coefficients between the outputs of the two numerical models in order to measure the degree of similarity between the two models, and thereby determine the impact of the

geostatistical simulations. However, as the two output variables relative to the two models in the same piezometer are not statistically independent, it will not be possible to perform standard statistical tests to compare their distribution.

2.4 EFFECT OF GEOSTATISTICAL SIMULATIONS

2.4.1 Comparison of distributions

To analyse the effect of geostatistical simulations on the output of the MARTHE model, a preliminary approach consisted in studying the change in the distributions of the model output variables between calculations without geostatistical simulation and calculations with geostatistical simulation. On the basis of 300 simulations performed for each model, the histograms for each output were represented by superposing that of the geostatistical-free model on that of the geostatistical model.

Figure 2.3 shows the superposition of the histograms for concentrations from two different piezometers. Visually speaking, such a representation provides a lot of information. It can be seen in these two examples that the addition of the geostatistical simulation variable has little impact on the concentration distributions.

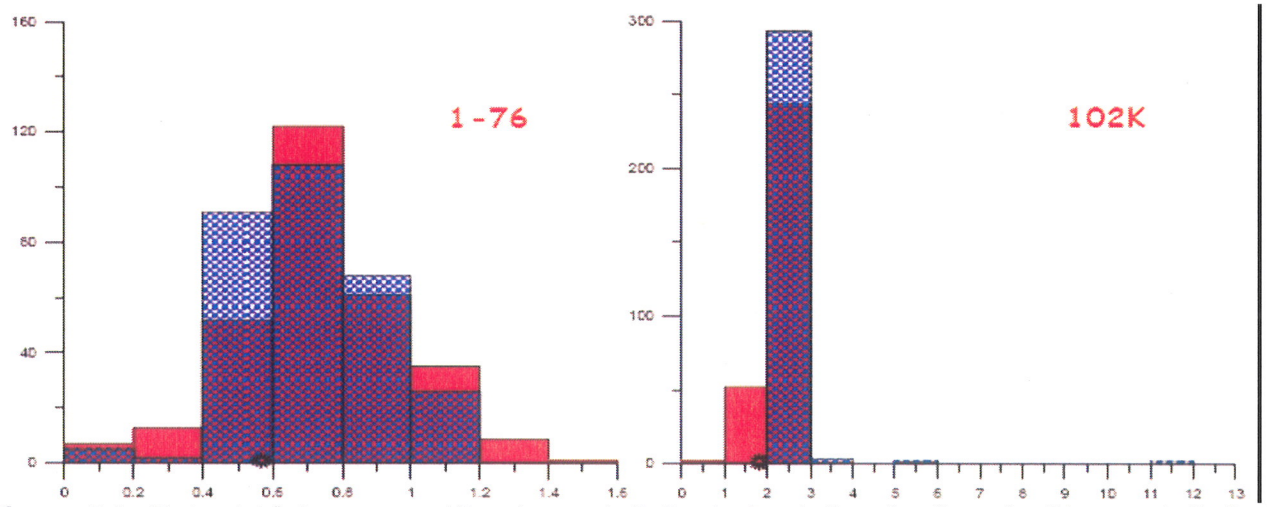

Figure 2.3: Output histograms without geostatistical simulation (red) and with geostatistical simulation (blue) for two MARTHE outputs (concentrations for piezometers p1-76 and p102K).

However, comparison is difficult considering that the histograms have been divided up into different classes of values, which prevents us from observing any subtlety in the distribution of the values. To resolve this problem, density smoothing techniques were used via a non-parametric method based on smoothing kernels (Gaussian in this case). We chose to use the "density" function of the R statistics software (R Development Core Team [12], Venables & Ripley [13]). Figure 2.4 represents the smoothed densities of the calculated concentrations over the 20 piezometers. Each figure makes it possible to visualise the impact of adding the geostatistical simulation variable (red curves in relations to the blue curves).

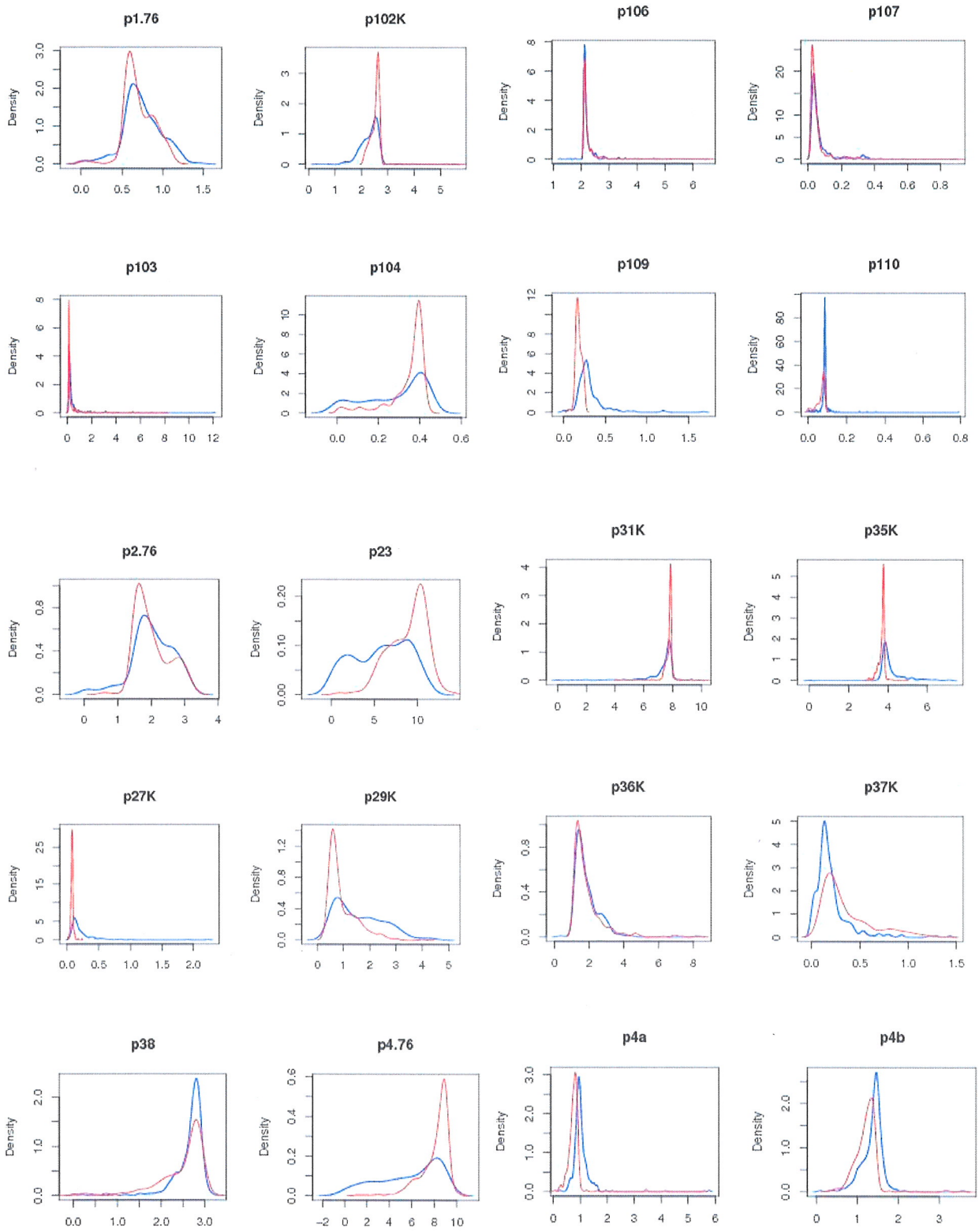

Figure 2.4: MARTHE output densities both without geostatistical simulation (blue curves) and with geostatistical simulation (red curves).

In view of these figures, the MARTHE outputs can be qualitatively classified into three categories:

- outputs where the distributions between calculations without geostatistical simulations and calculations with geostatistical calculations are similar (p103, p106, p107, p36K), which means that the shape of the second layer has little impact on MARTHE;

- outputs where the distributions between calculations without geostatistical simulations and calculations with geostatistical calculations are slightly different (p1-76, p102K, p110, p2-76, p27K, p31K, p38, p4a, p4b);
- outputs where the distributions between calculations without geostatistical simulations and calculations with geostatistical calculations are different (p104, p109, p23, p29K, p35K, p37K, p4-76), which means that the shape of the second layer has a great impact on MARTHE.

There is apparently no link between this classification and the spatial location of the piezometers on the site (see Figure 2.1). As in the previous study, this confirms the complexity and the non-linearity of the simulated numerical models.

Generally speaking, it was observed that the distribution of values did not tend to spread, as we could have expected when integrating geostatistical simulations into the model. The distribution even tends to retract (see piezometers p102K, p104, p109, p27K, p29K, p31K, p35K, p4-76), which does not at all prejudice the effect of the geostatistical simulations. This first purely-visual analysis requires confirmation (or negation) by quantitative analysis.

2.4.2 Correlations

2.4.2.1 Correlation coefficients

The standard correlation coefficient (Pearson) measures the degree of linear correlation between two quantitative numerical variables. Spearman's correlation coefficient measures the degree of monotonic correlation between two quantitative variables. It is therefore worth calculating these coefficients between the outputs of two numerical models: a model with 20 scalar input parameters without geostatistical simulation and another model with 16 scalar input parameters with geostatistical simulation. For each of the 300 calculations for the two models, 16 input parameters are similar. The Pearson and Spearman correlation coefficients therefore measure the linear or monotonic effect of replacing four input parameters (permeabilities in zones without coarse sand) by a geostatistical simulation of the shape of these zones. They are very sensitive to several extreme values of a sample, which means they must be considered qualitatively and with caution.

Table 2.2 shows the two correlation coefficients between each output of the two models. In the following analysis, the highest coefficient (absolute value) between that of Pearson and that of Spearman was used for each output.

Piezometer	**p1-76**	**p102K**	**p103**	**p104**	**p106**	**p107**	**p109**	**p110**	**p2-76**	**p23**
Pearson	35.16	-1.16	69.48	58.41	18.97	63.64	24.14	-13.01	11.20	63.48
Spearman	43.68	76.52	87652	58.57	40.71	46.04	25.81	21.45	28.52	67.98
Piezometer	**p27K**	**p29K**	**p31K**	**p35K**	**p36K**	**p37K**	**p38**	**p4-76**	**p4a**	**p4b**
Pearson	12.70	70.89	31.80	-13.00	60.63	34.44	58.56	68.66	46.90	47.43
Spearman	12.89	79.33	52.34	-20.45	58.93	36.88	-20.4	73.63	4.09	45.04

Table 2.2: Correlation coefficients (expressed as a percentage) between the outputs of the model without geostatistical simulation and the outputs of the model with geostatistical simulation.

In view of Table 2.2, the MARTHE outputs can be classified into three categories:

- outputs where the correlation coefficients are particularly high (> 66%): p102K, p103, p23, p29K, p4-76, which means that the shape of the second layer has little impact on MARTHE;

- outputs where the correlation coefficients are close to 50% (between 33% and 66%): p1-76, p104, p106, p107, p31K, p36K, p37K, p38, p4a, p4b;
- outputs where the correlation coefficients are low (<33%): p109, p110, p2-76, p27K, p35K, which means that the shape of the second layer has a high impact on MARTHE.

Upon preliminary analysis, the results of this interpretation seem to differ from the results mentioned in paragraph 2.4.1.

Based on these results, it was nevertheless remarked that the piezometers revealing a high impact of the shape of the second layer were located in the lower left zone of the model (see Figure 2.1), i.e. out of the way in relation to the spread of contamination over time. This seems to correspond to piezometers with low concentrations (because contamination diffusion is poor in this direction), which probably means that the data cannot be considered robust enough in relation to a change in the model (like the same of layer).

2.4.2.2 Correlation ratios

When measuring the correlation between a quantitative variable Y and a qualitative variable X (which has a certain number of categories), the following correlation ratio is applied:

$$\eta^2_{Y|X} = \frac{Var[E(Y|X)]}{Var(Y)}$$

which corresponds to the variance of the quantitative variable Y for each category of X. For our application, we can therefore assume that Y is the output variable (^{90}Sr concentration), and that X corresponds to an input parameter with two categories: with and without geostatistical simulation. The ratio of the empirical correlation can be calculated according to the following equation:

$$e^2(Y) = \frac{n_1 n_2}{n^2} \frac{(\overline{Y_1} - \overline{Y_2})^2}{\sigma^2(Y)}$$

where n is the total size of the data sample *(X,Y)*, n_1 and $\overline{Y_1}$ are the frequency and the mean of Y_1 corresponding to the first category of X (no simulations), n_2 and $\overline{Y_2}$ are the frequency and mean of corresponding to the second category of Y (with simulations), and $\sigma(Y)$ is the variance of Y. In this case, n = 600 and $n_1 = n_2 = 300$.

When e^2 is equivalent to zero, there is generally independence between X and Y: the input parameter X does not affect the variation in the output of model Y. This means that the presence (or absence) of geostatistical simulations does not affect the concentrations obtained. When e^2 is close to 1 (or to 100%), it means that there is strong dependence between the categories of the input parameter X and the response Y.

The correlation ratios are given in Table 2.3. In this table, certain results seem incorrect as they are greater than 100%. In fact, the hypotheses associated with the calculation of these correlation ratios are not met because the data is not inter-independent. They are highly correlated for certain outputs between two categories (as seen in the previous paragraph).

Piezometer	**p1-76**	**p102K**	**p103**	**p104**	**p106**	**p107**	**p109**	**p110**	**p2-76**	**p23**
e^2	0.65	49.37	0.02	1.90	**4.98**	0.18	8.81	0.29	0.05	5351
Piezometer	**p27K**	**p29K**	**p31K**	**p35K**	**p36K**	**p37K**	**p38**	**p4-76**	**p4a**	**p4b**
e^2	10.93	240.7	168.8	**1.6$_4$1**	**1.64**	13.30	12.69	2899	26.40	9.77

Table 2.3: Correlation ratios between the output of two numerical models and the "presence/ absence" category of the geostatistical simulation (as a percentage).

In view of this table, it is possible to classify MARTHE outputs into three categories:

- outputs where the correlation ratios are less than 10%: p1-76, p103, p104, p107, p109, p110, p2-76, p36K, p4b, which means that the shape of the second layer has little effect on the MARTHE model;
- outputs where the correlation ratios are greater than 10% and less than 100%: p102K, p27K, p37K, p38, p4a;
- outputs where the correlation ratios are abnormal because they are higher than 100%: p23, p29K, p31K, p35K, p4-76, which means that the shape of the second layer has a high impact on the MARTHE model.

As a first approach, these results seem rather different from the results in paragraph 2.4.2.1, and more similar to those in paragraph 2.4.1. However, the confidence level given to these results is relative seeing that the basic hypothesis of inter-independence of data was not met.

2.4.2.3 Comparison of sensitivity analyses

The next step was to perform sensitivity analyses in relation to the scalar input parameters for the two MARTHE models: the model without geostatistical simulation and the model with geostatistical simulation. Owing to the high non-linearity and even the non-monotony of the outputs of the two MARTHE numerical models, we decided to use sensitivity indices based on the statistical tests that only require the independence of the initial sample, and no hypothesis on the complexity of the model. This was the case here, as the sets of input parameters were established randomly (according to the LHS plan).

Helton *et al.* [6] provide a detailed list of indices based on statistical tests (parametric or non-parametric). The idea is to partition the variation range of an input (and possibly the output) into several parts, before testing to see if the distribution is different between the parts. It is therefore necessary to use tests that enable the comparison of several (especially more than two) samples: F-ANOVA, χ^2, Kruskal-Wallis, etc. The results are given in terms of the p-value, and are therefore not expressed quantitatively as is the case for the standard regression coefficients which take into account contribution to the output variance. However, the hypothesis of linearity or of monotony is no longer required to validate the indices. The only hypothesis given concerns the independence of observations. For the LHS plans, where observations are not independent, it seems that this hypothesis is not restrictive and that the tests are both robust and fully applicable. Certain tests are more robust that others; some do not assume that the sample follows a normal law. We used the SSURFER software (Iooss [7]) in the suite, which is limited to the use of the "common locations" (CL) test based on the Kruskal-Wallis test (Conover [2]).

The Kruskal-Wallis test can be considered as a generalisation of the Wilcoxon-Mann-Whitney test with more than two samples. It is a non-parametric test and therefore does not offer any hypothesis on the shape of the underlying distributions. Like many non-parametric tests, it does not focus on the observation values but on their ranks, once these observations have been pooled into the one mega-sample. The statistics of the Kruskal-Wallis test are built on the mean of the observation ranks in the different samples. This is why the Kruskal-Wallis test is sometimes called: univariate ANOVA on the ranks. As we have 300 data for each model, we chose a breakdown into 10 segments in relation to the values of an input parameter, which results in tests between 10 samples of 30 data each. A test was performed for each input parameter and each output. The result of a statistical test is a p-value. A p-value equal to 0.1 means that the postulated hypothesis cannot be rejected at confidence levels greater than or equal to 90%, i.e. that each sample has the same Kruskal-Wallis statistics. Our final interpretation will therefore be: if the p-value is greater than 0.1, the samples are statistically homogeneous and therefore the input parameter does not affect the output. Below the threshold of 0.1, it can be said that the different samples are no longer statistically homogeneous, and that the lower the p-value, the greater the impact of the input parameter on the output.

Table 2.4 provides the results of sensitivity analyses based on the CL test for each output and each model (with and without geostatistical simulation). For the sake of simplifying our interpretation, we have not given the p-values but only the classification (in order of decreasing influence) of the inputs whose p-value is below 0.1.

	géosta	kd	kd	kd		d3	l1	i2	i3	per	per	per	pz	pz	pz	pz	po		
p1-76	Withou	2	1				3				4								
	With		1							2									
p102K	Withou	2	1						3			4		6					
	With		1								2								
p103	Withou	1	4	5					3		2	6							
	With	2	1			4				3									
p104	Withou	1							2		3								
	With	1	3							2									
p106	Withou	1							3				2						
	With	1		3						2									
p107	Withou	1							3				2						
	With	2							4	1		3							
p109	Withou					6		3	1	2	5	7					4		
	With	2								1		4							
p110	Withou	2	3						1			4				5	6		
	With	1	3							2									
p2-76	Withou	2	1				4			5	3								
	With	4	1					3		2									
p23	Withou	3	1						2		4								
	With		1				4			2		3							
p27K	Withou		2						1			3							
	With	3	1					5	2		4								
p29K	Withou	2	1						3		6	5	4	7					
	With	2	1	5						3	4	6							
p31K	Withou		1						2			3							
	With		1																
p35K	Withou		1						2	5		4	3						
	Avec		1	3						2									
p36K	Withou	1		2		4						5	3						
	With	1		3						2	4								
p37K	Withou	5	3	6					1	2	4								
	With		1							2	3								
p38	Withou	1	2	4					3										
	With	2	1	5						3	4								
p4-76	Withou	3	1						2					5					
	With	2	1							3									
p4a	Withou							2	1						3				
	With	3	1	4					6		5	2							
p4b	Withou		2					4	1	7	3	6		8	5				
	With		1	3					5		4	2							

Table 2.4: Results of the sensitivity analysis between the inputs and the outputs using the CL test – Classification of the parameters whose p-values are below 0.1 (Blue: model without geostatistical simulation. Red: model with geostatistical simulation).

As a preliminary analysis, the results of the model without geostatistical simulation (blue) seem consistent with the results of Volkova *et al.* [14] which use indices based on the Sobol ranks and indices calculated using metamodels. For certain output variables (non monotonic and poorly modelled by a metamodel), slight differences can appear; the classification in Table 2.4 are to take priority. The CL test is robust and not based on any hypothesis (Helton *et al.* [6]). The most influential input parameters for this model are therefore the distribution coefficients of layers 1 and 2 (kd1 and kd2), and the intensity of infiltration in the leakage zones of pipes (i3). To a lesser extent, the permeability of layers 2 and 3 (per2 and per3) are also often influential. It can also be seen that the permeabilities in the different zones without coarse sand are influential for several piezometers.

For the model with geostatistical simulation, the three most influential parameters are present: kd1, kd2 and the permeability in layer 1 (per1). As previously mentioned, per2 and per3 are influential to a lesser extent. Last of all, the distribution coefficient of layer 3 (kd3) also has a rather significant impact. In terms of the sensitivity analysis, the only remarkable difference between the model without geostatistical simulation and that with geostatistical simulation is the influence of i3 and per1:

- The permeability of the first layer plays an important role in the model with geostatistical simulation, whereas it hardly has any influence in the first model. This is logical in terms of the numerical model: in the first model, there were 4 different parameters for the permeability of zones without coarse sand in layer 2. In the new model, the permeability of all these zones is combined with the permeability of layer 1 to shape parameter per1. It can be said that the influence of parameters perz1, perz2, perz3 and per1 can be explained by the influence of parameter per1. From a hydrogeological viewpoint, the transport of radionuclides is partially determined by the direction of flow in the ground water, which depends on the shapes of the different zones of permeability. The permeability of these zones in the second layer of the model (where the concentration of radionuclides is taken into account) is per1 and per2. For the model without geostatistical simulation, the permeabilities of these zones were perz1, perz2, perz3 and perz4.

- The intensity of infiltration i3 in the pipe leakage zones does not play a major role in the model with geostatistical simulation. This behaviour can be explained in the same way as beforehand. Without geostatistical simulation, the flow in the model domain was mainly influenced by the force of the leaks, with the forms of the less-permeable zones having been set. With geostatistical simulation, the direction of flow depends on the shapes of the zones (influenced by per1 and per2).

Through a finer analysis (output by output) of the influences of the inputs between the two models, notable differences can be seen for practically all outputs. Even though the most influential parameter is often the same, the following parameters differ most of the time. This confirms the potential impact of the geostatistical simulations on the dynamics of the MARTHE model.

2.5 METAMODELS

When a numerical model is costly and non-linear, it is worth considering replacing it by a simpler mathematical function, called a "metamodel" whose CPU time is negligible. Different types of metamodels are available (Fang *et al.* [5]). We decided to focus on the use of Gaussian processes, which is a very popular metamodel at the moment, as well as on the development of a joint model more suitable for our problem.

2.5.1 Gaussian processes

In this section, we recommend building a "Gaussian process" type metamodel for each output of each numerical model. We used the methodology developed by Marrel *et al.* [11]. We adjusted a metamodel for each output of each model. The predictive quality of the metamodel is given by the predictivity coefficient Q_2, calculated by a cross-validation procedure. The Q_2 coefficient represents the variance contribution explained by the metamodel in relation to the total variance of the output variable.

The construction of a metamodel is possible as a function of the scalar input parameters. As the random field results from geostatistical simulation (not being a scalar parameter), it is not possible to integrate it into the construction of the metamodel. For the model with geostatistical simulation, the metamodel can therefore not explain the variance contribution of the variable which is due to the effect of the geostatistical simulation. To indicate the effect of the geostatistical simulation, we can

therefore imagine using the Q_2 difference between the model without geostatistical simulation and that with geostatistical simulation

Table 2.5 provides the Q_2 results for the two models and their difference. The few negative values correspond to situations where the outputs of the model with geostatistical simulation were easier to adjust that those without geostatistical simulation. The heterogeneous shape of layer 2 probably tended to cancel out the non-linearities of the physical model. Generally speaking, a metamodel is considered predictive when Q_2 is greater than 80%. This is the case for 8 out of 19 situations for the model without geostatistical simulation, whereas this practically never occurs for the model with geostatistical simulation.

In view of Table 2.5, the MARTHE outputs can be classified into three categories:

- outputs where the prediction differences are low (< 15%): p2-76, p27K, p29K, p36K, p37K, p4a, p4b, which means that the shape of the second layer has little impact on the MARTHE model;
- outputs where the prediction differences are average (between 15% and 30%): p103, p104, p106, p109, p23, p35K, p38;
- outputs where the prediction differences are particularly high (> 30%): p1-76, p102K, p107, p31K, p4-76, which means that the shape of the second layer has a high impact on the MARTHE model.

Piezometer	p1.76	p102K	p103	p104	p106	p107	p109	p110	p2.76	p23
Without geostat.	84	78	50	96	45	86	50		86	94
With geostat.	47	0	33	71	63	22	69		63	77
Difference	37	78	17	25	-18	64	-19		13	17
Piezometer	**p27K**	**p29K**	**p31K**	**p35K**	**p36K**	**p37K**	**p38**	**p4.76**	**p4a**	**p4b**
Without geostat.	43	93	69	56	60	90	52	96	9	37
With geostat.	31	80	0	37	71	80	73	57	13	26
Difference	12	13	69	19	-11	10	-21	39	-4	11

Table 2.5: Predictivity coefficients Q_2 (data given as a percentage and calculated by cross-validation) of adjusted Gaussian process models.

As a first approach, these results appear rather different to the results mentioned in paragraphs 2.4.2.1 and 2.4.2.2. In view of these results and the difficulty in obtaining good metamodels by Gaussian processes for the MARTHE data, it was decided that this Q_2 difference criterion was not suitable for the given situation.

2.5.2 Joint modelling using Generalized Additive Model (GAM)

The joint model consists in adjusting a double metamodel to the data: one component for the mean and another component for dispersion. The input random field is in fact rejected as an uncontrollable (or noisy) parameter of the model. The mean and dispersion components are therefore expressed via other scalar input parameters of the model. Thus, the dispersion component includes everything that the scalar parameters only could not explain, i.e. the effect of the random field (including its interactions with the scalar parameters).

Iooss *et al.* [9] (see chapter 3) illustrated the use of the joint model with a MARTHE output. This result is therefore partial. Processing other outputs is problematic because MARTHE data with geostatistical simulation is particularly complex: the random field effect is often predominant, "outliers" threshold effects, highly-scattered values, etc. The joint model requires a manual, iterative adjustment procedure for each component, which is therefore not very suitable. Developing an automatic yet accurate adjustment procedure for this model therefore remains to be accomplished.

2.6 CONCLUSION

Developed at least twenty years ago, global sensitivity analysis methods for numerical models are limited to scalar input parameters. This research work consisted in investigating and in applying several simple tools to assess the sensitivity of a model output variable in relation to a stochastic input that is non-quantifiable by a small number of scalar parameters. Our application focused on a hydrogeological model dependent on the simulations of a random input field. This random field was built with the aim of correctly representing the heterogeneity of a hydrogeological parameter (permeability in a layer of coarse sand).

All of our analyses were based on the fact that we were no longer able to perform calculations with the numerical model. We only had the relatively limited calculation results obtained using both the model with geostatistical simulation and the model without geostatistical simulation, which correspond to the same sets of scalar input parameters. The different methods used to compare the results of these calculations (with and without geostatistical simulation) were tested:

- comparison of the output distributions of the two models. This analysis made it possible to graphically visualise the impact of the geostatistical simulation, without however quantifying it;
- analysis of the correlation coefficients between the outputs of the two models. These indicators were used to measure the change induced by the geostatistical simulations. They are sensitive to a few extreme values in a sample, and have to be considered qualitatively and with precaution;
- study of the correlation ratios between the output of both models and the "with/without geostatistical simulation" category. These indicators made it possible to quantitatively measure the effect of a qualitative variable (with/without geostatistical simulation) upon a quantitative variable (model output). However, the reliability of these results is relative seeing that the basic hypothesis of independence between data was not met in our study;
- comparison of sensitivity analyses for the two models in relation to scalar inputs. This is a qualitative method only, designed to check if both models have different influential input parameters. It is used to better understand the behaviour of models and the effect of geostatistical simulation upon other input parameters (via their interactions with the random field);
- comparison of the construction of metamodels (based on Gaussian processes) for both models. This is the same type of method as the previously one but this time based on the construction of a metamodel. This method produced very little information in our case;
- construction of a metamodel based on the joint modelling of the mean and of the dispersion. This method is the most accurate and made it possible to obtain a total sensitivity index for the "random field" input variable. However, this method requires the construction of a metamodel for the mean and a metamodel for dispersion. This method could not be implemented in our application owing to the extreme complexity of our data and an excessive number of input parameters. Iooss *et al.* [9] (see chapter 3) nevertheless managed to apply the method to an output of the model with geostatistical simulation.

In conclusion, our results showed the significant impact of geostatistical simulations on the dynamic of the MARTHE model. This demonstration supports the assumption supposing the limitation of our modelling in relation to the modelling of different layers with different hydrogeological behaviours. The piezometers that proved to be the most sensitive to the shape of the second layer are in fact those located in the background in relation to the direction in which contamination spreads over time. This corresponds to piezometers to low concentration values.

2.7 ACKNOWLEGMENTS

This work has been done with Elena Volkova (Kurchatov Institute, Moscow) and Amandine Marrel (CEA Cadarache).

2.8 REFERENCES

[0] D. Busby, T. Romary, M. Feraille and S. Touzani, An integrated approach for uncertainty and sensitivity analysis in reservoir forecasting. *Journal of Computational Geoscience,* submitted, 2008.

[1] T. Hastie and R. Tibshirani. *Generalized additive models.* Chapman & Hall/CRC, 1990.

[2] W.J. Conover. *Practical non parametric statistics, Third edition,* Wiley, New-York, 1999.

[3] C.V. Deutsch and A.V. Journel. *GSLib - Geostatistical software library and user's guide.* Second edition, Oxford University Press, 1997.

[4] N. Devictor and R. Bolado Lavin. *Uncertainty and sensitivity methods in support of PSA level 2,* SARNET report, SARNET-PSA2-P06 Revision 0, European Commission, 2006.

[5] K.T. Fang, R. Li and A. Sudjianto. *Design and modelling of computer experiments.* Chapman & Hall/CRC, 2006.

[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] B. Iooss. *Manuel utilisateur du logiciel SSURFER V1.2: programmes en R d'analyse d'incertitudes, de sensibilités, et de construction de surfaces de réponse.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 6 08/03/06, CEA Cadarache, 2006.

[8] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety,* submitted, 2008. Available at URL: http://fr.arxiv.org/abs/0802.1009v1.

[9] B. Iooss, M. Ribatet and A. Marrel. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics,* submitted, 2008. Available at URL: http://fr.arxiv.org/abs/0802.0443v1.

[10] A.M. Law and W.D. Kelton. *Simulation modelling and analysis,* third edition, New-York: McGraw-Hill, 2000.

[11] A. Marrel, B. Iooss, F. Van Dorpe and E. Volkova. An efficient methodology for modelling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis,* accepted, in revision, 2008. Available at URL: http://fr.arxiv.org/abs/0802.1099v1.

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

[13] W.N. Venables and B.B. Ripley. *Modern applied statistics with S.* Springer, 4th ed., 2002.

[14] E. Volkova, B. Iooss and F. Van Dorpe. Global sensitivity analysis for a numerical model of radionuclide migration from the "RRC Kurchatov Institute" waste disposal site. *Stochastic Environmental Research and Risk Assessment,* 22:17-31, 2008.

3 GLOBAL SENSITIVITY ANALYSIS OF STOCHASTIC COMPUTER MODELS WITH GENERALIZED ADDITIVE MODELS

3.1 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 chapter proposes a global sensitivity analysis method for stochastic computer codes (having a variability induced by some uncontrollable parameters). The mean and the dispersion of the code outputs are modeled by two interlinked Generalized Additive Models (GAM). The "mean" model allows to obtain the controllable parameters sensitivity indices, while the "dispersion" model allows to obtain the uncontrollable 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.

3.2 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 (looss 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 chapter, we are not interested by deterministic computer models but by stochastic numerical models - i.e. when the same input parameters set leads to different output values. The model is therefore intrinsically stochastic because of some "uncontrollable" parameters. 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, "uncontrollable" 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, looss et al. [8]), where the uncontrollable 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 uncontrollable 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 uncontrollable parameter does not interact with controllable 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., looss & 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 uncontrollable 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 chapter 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 chapter starts by describing the joint model construction, firstly with the GLM, secondly with the GAM. The fourth 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.

3.3 JOINT MODELING OF MEAN AND DISPERSION

3.3.1 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 :

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

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 :

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

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 controllable 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 competitive 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.*

3.3.2 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.

3.4 GLOBAL SENSITIVITY ANALYSIS

3.4.1 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 :

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

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 Var$[\mathbb{E}(Y|X_i)]$. Most often, this term is divided by 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.

3.4.2 Stochastic models

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

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

where $\mathbf{X}$ are the p controllable 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)).

For a stochastic model (12), the joint model introduced in section 3.3 enables us to recover two GLMs or two GAMs :

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

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

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

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

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

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

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

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,

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

For the mean component Y_m, we have

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

By noticing that

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

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

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

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 $\text{Var}(Y_m)$ 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 $\text{Var}(Y)$ (according to $\mathbf{X}$ and ε) not taken into account in $\text{Var}(Y_m)$ i.e. all terms involving ε. Therefore, the total sensitivity index of ε is

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

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

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

If $\text{Var}(Y)$ is computed from the data, it seems preferable to estimate $\mathbb{E}(Y_d)$ with $\text{Var}(Y) - \text{Var}(Y_m)$ 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, $S_{i\varepsilon}$, $S_{ij\varepsilon}$, ...). However, the analysis of the terms in the regression model Y_d and their t-values 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 uncontrollable 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 uncontrollable parameters interact with some controllable input ones ;
- if some information about the influence of the interactions between the uncontrollable parameters and the other input parameters is of interest.

3.5 APPLICATIONS

3.5.1 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 controllable input random variables, and X_3 as an uncontrollable 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.

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 uncontrollable 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 uncontrollable 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 3.1 shows the observed response against 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 3.1 as the dispersion around the $y = x$ line is clearly reduced. On the other hand, Figure 3.2 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.

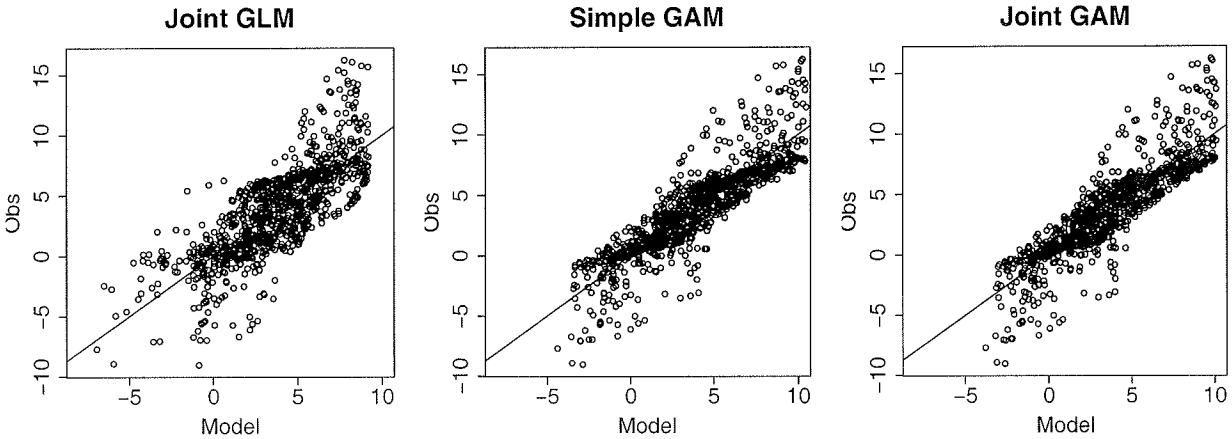

Fig. 3.1 – **Observed response variable versus the predicted values for the three models : Joint GLM, Simple GAM, Joint GAM (Ishigami application).**

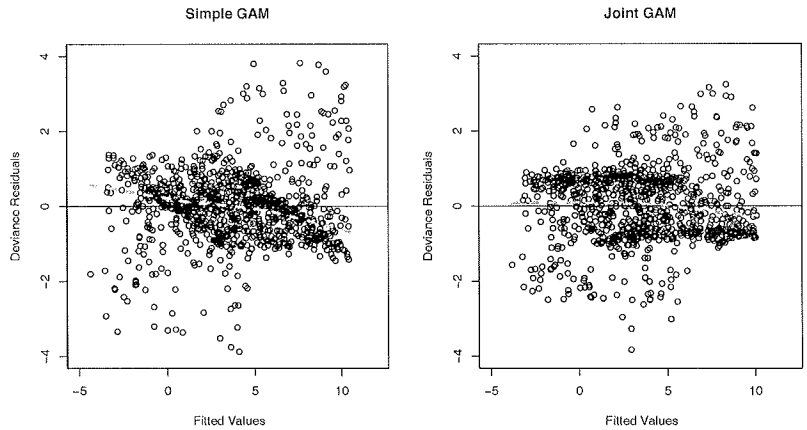

Fig. 3.2 – **Deviance residuals for the Simple and Joint GAMs versus the fitted values (Ishigami application). Dashed lines correspond to local polynomial smoothers.**

Figure 3.3 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 3.3 shows that joint GLM approach is the most accurate model. The joint GAM is less 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.

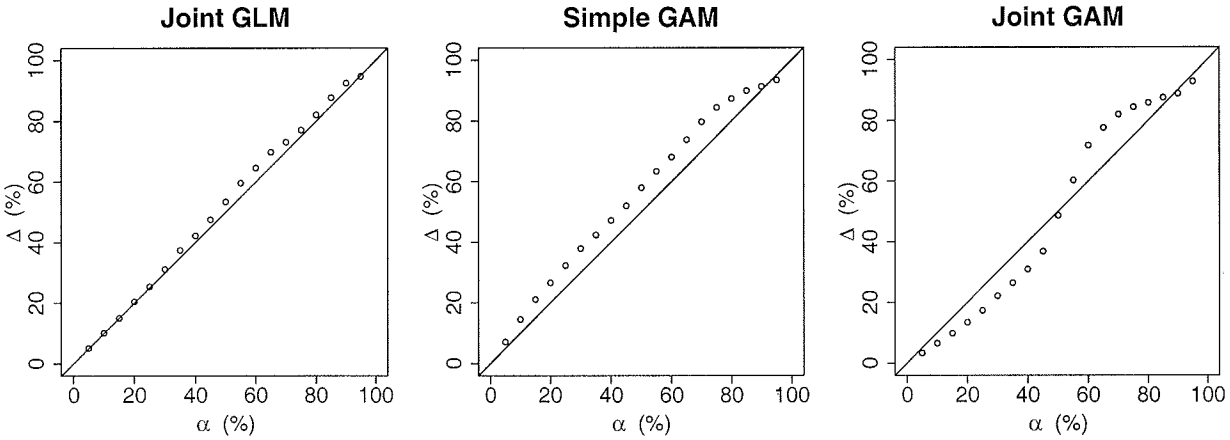

Fig. 3.3 – **Proportion Δ (in percent) of observations that lie within the α% theoretical confidence interval in function of the confidence level α (Ishigami application).**

Lastly, from Figures 3.1–3.3, 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.

Sobol indices

Table 3.5.1 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 repetitions 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 3.5.1 gives also 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 uncontrollable 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

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	—	—	—	—	—	—	—	—	—

TAB. 3.1 – **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.**

the unexplained part of these metamodels is due to the uncontrollable 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.

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 it is difficult to have an optimization algorithm which gives an efficient and robust estimation of the nugget effect.*

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 efficiently global sensitivity indices of stochastic models.

3.5.2 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 influential 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, this hydrogeologic layer form has to be considered as an uncontrollable parameter of the computer model. Additionally to the uncontrollable 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 3.4 shows the deviance residuals against 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 3.5 shows the observed values against the predicted ones. This figure confirms the conclusions drawn from the Figure 3.4. Indeed, the GAM's flexibility allows to suppress the bias for the smallest data values.

Figure 3.6 shows the proportion Δ of observations that lie within the $\alpha\%$ theoretical confidence interval against 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 3.4–3.6, 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.

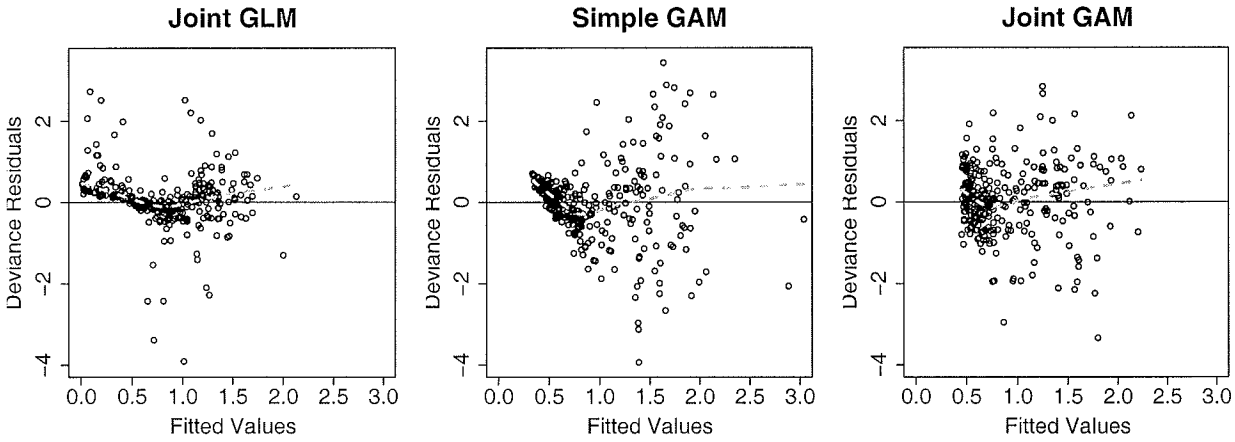

Fig. 3.4 – **Deviance residuals (mean component) for the Simple GAM, Joint GAM and Joint GLM versus the fitted values (MARTHE application). Dashed lines correspond to local polynomial smoothers.**

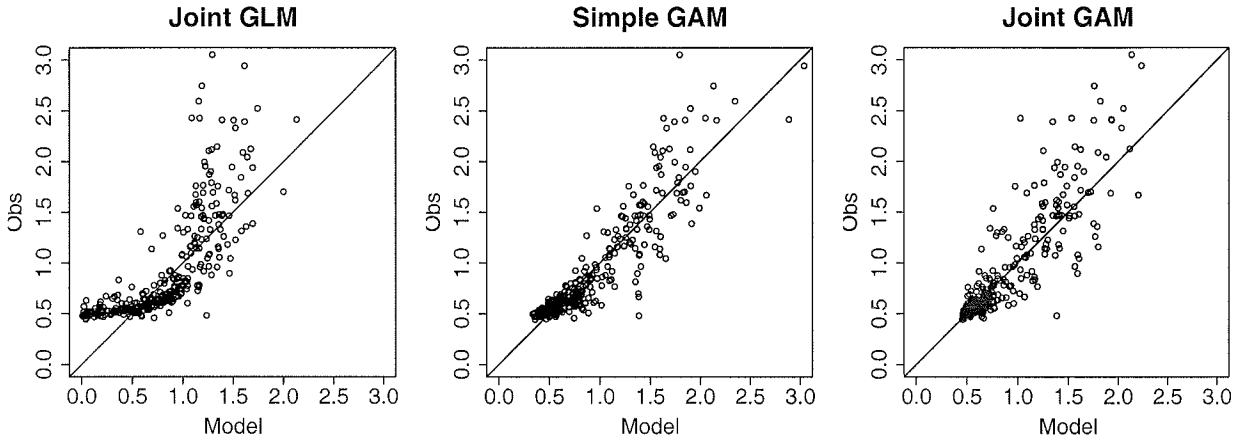

Fig. 3.5 – **Observed response variable versus the predicted values for the three models : Joint GLM, Simple GAM, Joint GAM (MARTHE application)**

Table 3.5.2 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 controllable 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 uncontrollable 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.

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 uncontrollable 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

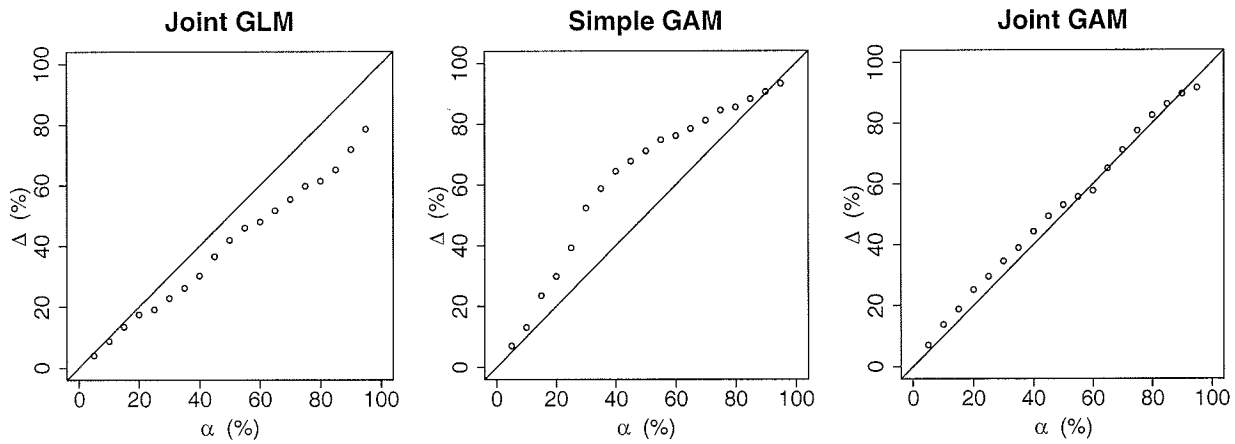

Fig. 3.6 – **Proportion Δ (in percent) of observation that lie within the α% theoretical confidence interval in function of the confidence level α. MARTHE application.**

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 uncontrollable 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 uncontrollable 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.

3.6 CONCLUSION

This chapter 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 chapter, 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 chapter, 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 more general models 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 uncontrollable parameter, which leads to correct estimations of the first order indices of the controllable parameters. In addition, it reveals the influential interactions between the uncontrollable parameter

Indices	Joint GLM			Joint GAM			Simple GAM			Gp		
	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	Method	Values	*sd*	M
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	]
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	]
S(per1)	0.018	0.7e-2	MC	0	—	Eq	0	—	Eq	0.012	1.1e-2	]
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	]
S(per3)	0	—	Eq	0.005	1.0e-2	MC	0.008	1.0e-2	MC	0.003	1.1e-2	]
S(kd2,per2)	0	—	Eq	0.063	1.0e-2	MC	0.026	1.0e-2	MC	0.021	1.4e-2	]
$S_T(\varepsilon)$	0.382	0.2e-2	MC	0.277	0.3e-2	MC	0.235	—	Q_2	0.196	—	
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	—	—	—	—	—	

TAB. 3.2 – **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.**

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, the "mgcv" (Multiple Smoothing Parameter Estimation by GCV) package to fit a simple GAM, and the "sensitivity" package to compute Sobol indices. We also developed the "JointModeling" package to fit joint models (including joint GLM and joint GAM).

3.7 ACKNOWLEGMENTS

This work has been done with Mathieu Ribatet (CEMAGREF Lyon) and Amandine Marrel (CEA Cadarache).

3.8 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*, accepted, in revision, 2008. Available at URL : http ://fr.arxiv.org/abs/0802.1099v1.

[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. looss, 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*, 22 :17–31, 2008.

[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 integrated 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.

4 GLOBAL SENSITIVITY ANALYSIS OF COMPUTER MODELS WITH FUNCTIONAL INPUTS

4.1 ABSTRACT

Global sensitivity analysis is used to quantify the influence of uncertain input parameters on the response variability of a numerical model. The common quantitative methods are applicable to computer codes with scalar input variables. This chapter aims at illustrating different variance-based sensitivity analysis techniques, based on the so-called Sobol indices, when some input variables are functional, such as stochastic processes or random spatial fields. In this work, we focus on large cpu time computer codes which need a preliminary metamodeling step before performing the sensitivity analysis. We propose the use of the joint modeling approach, i.e., modeling simultaneously the mean and the dispersion of the code outputs using two interlinked Generalized Linear Models (GLM) or Generalized Additive Models (GAM). The "mean" model allows to estimate the sensitivity indices of each scalar input variables, while the "dispersion" model allows to derive the total sensitivity index of the functional input variables. The proposed approach is compared to some classical SA methodologies on an analytical function. Lastly, the proposed methodology is applied to an industrial computer code that simulates the nuclear fuel irradiation.

4.2 INTRODUCTION

Modern computer codes that simulate physical phenomenas often take as inputs a high number of numerical parameters and physical variables, and return several outputs - scalars or functions. For the development and the use of such computer models, Sensitivity Analysis (SA) is an invaluable tool. The original technique, based on the derivative computations of the model outputs with respect to the model inputs, suffers from strong limitations for most of computer models. More recent global SA techniques take into account all the variation ranges of the inputs and aim to apportion the whole output uncertainty to the input factor uncertainties (Saltelli et al. [18]). The global SA methods can also be used for model calibration, model validation, decision making process, i.e., any process where it is useful to know which variables that mostly contribute to the output variability.

The common quantitative methods are applicable to computer codes with scalar input variables. For example, in the nuclear engineering domain, global SA tools have been applied to numerous models where all the uncertain input parameters are modeled by random variables, possibly correlated - such as thermal-hydraulic system codes (Marquès et al. [12]), waste storage safety studies (Helton et al. [6]), environmental model of dose calculations (Iooss et al. [9]), reactor dosimetry processes (Jacques et al. [10]). Recent research papers have tried to consider more complex input variables in the global SA process, especially in petroleum and environmental studies :

- Tarantola et al. [23] work on an environmental assessment on soil models which use spatially distributed maps affected by random errors. For the SA, they propose to replace the spatial input by a "trigger" parameter that governs the random field simulation ;
- Ruffo et al. [15] evaluate an oil reservoir production using a model that depends on different heterogeneous geological media scenarios. These scenarios, which are of limited number, are then substituted for a discrete factor (a scenario number) before performing the SA ;
- Iooss et al. [8] (see chapter 3) study a groundwater radionuclide migration model which is based on geostatistical simulations of the hydrogeological layer heterogeneity. The authors propose to consider the spatial input parameter as an "uncontrollable" parameter.

In this chapter, we tackle the problem of the global SA for numerical models and when some input parameters ε are functional. $\varepsilon(u)$ is a one or multi-dimensional stochastic function where u can be spatial coordinates, time scale or any other physical parameters. Our work focuses in models that depend on scalar parameter vector $\boldsymbol{X}$ and need some stochastic processes simulations or random fields $\varepsilon(u)$ as input parameters. The computer code output value Y depends on the realizations of these random functions. These models are typically non linear with strong interactions between input

parameters. Therefore, we concentrate our methodology on the variance based sensitivity indices estimation ; that is, the so-called Sobol indices (Sobol [22], Saltelli et al. [18]).

To deal with this situation, a first natural approach consists in the discretization of the input functional parameter $\varepsilon(\boldsymbol{u})$ or its decomposition into an appropriate basis of orthogonal functions. Then, for all the new scalar parameters which represent $\varepsilon(\boldsymbol{u})$, sensitivity indices are computed. However, in the case of complex functional parameters, this approach seems to be rapidly intractable as these parameters cannot be represented by a small number of scalar parameters (Tarantola et al. [23]). Moreover, when dealing with non physical parameters (for example coefficients of orthogonal functions used in the decomposition), sensitivity indices interpretation may be labored. Indeed, most often, physicists would prefer to obtain one global sensitivity index related to $\varepsilon(\boldsymbol{u})$.

The following section presents three different strategies to compute the Sobol indices with functional inputs : (a) the macroparameter method, (b) the trigger parameter method and (c) the proposed joint modeling approach. Section 4.4 compares the relevance of these three strategies an analytical function : the WN-Ishigami function. Then, the proposed approach is illustrated on an industrial computer code simulating fuel irradiation in a nuclear reactor.

4.3 COMPUTATIONAL METHODS OF SOBOL INDICES

First, let us recall some basic notions about Sobol indices. Let define the model

$$\begin{aligned} f : \mathbb{R}^p &\to \mathbb{R} \\ \boldsymbol{X} &\mapsto Y = f(\boldsymbol{X}) \end{aligned} \tag{1}$$

where Y is the code output, $\boldsymbol{X} = (X_1, \ldots, X_p)$ are p independent inputs, and f is the model function, which is analytically not known. The main idea of the variance-based SA methods is to evaluate how the variance of an input or a group of input parameters contributes to the output variance. These contributions are described using the following sensitivity indices :

$$S_i = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_iX_j\right)\right]}{\mathrm{Var}(Y)} - S_i - S_j, \quad S_{ijk} = \ldots \tag{2}$$

These coefficients, namely the Sobol indices, can be used for any complex model functions f. The second order index S_{ij} expresses the model sensitivity 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 all indices lie in $[0, 1]$ and their sums are equal to one. The larger an index value is, the greater is the importance of the variable or the group of variables related 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 output sensitivity to an input X_i, Homma & Saltelli [7] introduce the total sensitivity index :

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \sum_{j \neq i, k \neq i, j<k} S_{ijk} + \ldots = \sum_{l \in \#i} S_l \tag{3}$$

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 having i in their index. The estimation of these indices (Eqs. (2) and (3)) can be performed by Monte-Carlo simulations based on independent samples (Sobol [21], Saltelli [17]), or by refined sampling designs introduced to reduce the number of required model evaluations significantly, for instance FAST (Saltelli et al. [20]) and quasi-random designs (Saltelli et al. [19]).

Let us now consider a supplementary input parameter which is a functional input variable $\varepsilon(\boldsymbol{u}) \in \mathbb{R}$ where $\boldsymbol{u} \in \mathbb{R}^d$ is a d-dimensional location vector. $\varepsilon(\boldsymbol{u})$ is defined by all its marginal and joint probability distributions. In this work, it is supposed that random function realizations can be simulated. For

example, these realizations can be produced using geostatistical simulations (Lantuéjoul [11]) or stochastic processes simulations (Gentle [4]). Our model writes now

$$Y = f(\boldsymbol{X}, \varepsilon) \tag{4}$$

and in addition to the Sobol indices related to $\boldsymbol{X}$, our goal is to derive methods to compute the sensitivity indices relative to ε, i.e., S_ϵ (first order index), S_{T_ε} (total sensitivity index), $S_{i\varepsilon}$ (second order indices), $S_{ij\varepsilon}$, ...

4.3.1 The macroparameter method

To solve the problem of correlated input parameters in the Sobol indices calculations, Jacques et al. [10] propose the use of multi-dimensional sensitivity indices (Sobol [22]) : each group of correlated parameters is considered as a multi-dimensional parameter or macroparameter. The different Sobol indices (first order, second order, ..., total) are then computed using independent Monte-Carlo sampling techniques (Sobol [21], Saltelli [17]). These techniques allow correlations between input parameters ; while it is prohibited with other methods - for example FAST.

In our context, this approach seems to be relevant as the input functional parameter $\varepsilon(\boldsymbol{u})$ can be considered as an unique input multi-dimensional parameter (i.e. a macroparameter). For instance, the first order Sobol index related to $\varepsilon(\boldsymbol{u})$ is defined as previously by

$$S_\varepsilon = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|\varepsilon\right)\right]}{\mathrm{Var}(Y)} \tag{5}$$

A simple way to estimate $S_\varepsilon = D_\varepsilon / D$ is based on the Sobol [21] algorithm :

$$\hat{f}_0 = \frac{1}{N}\sum_{k=1}^{N} f(\boldsymbol{X}_k^{(1)}, \varepsilon_k) \tag{6a}$$

$$\hat{D} = \frac{1}{N}\sum_{k=1}^{N} f^2(\boldsymbol{X}_k^{(1)}, \varepsilon_k) - \hat{f}_0^2 \tag{6b}$$

$$\hat{D}_\varepsilon = \frac{1}{N}\sum_{k=1}^{N} f(\boldsymbol{X}_k^{(1)}, \varepsilon_k) f(\boldsymbol{X}_k^{(2)}, \varepsilon_k) - \hat{f}_0^2 \tag{6c}$$

where $(\boldsymbol{X}_k^{(1)})_{k=1...N}$ and $(\boldsymbol{X}_k^{(2)})_{k=1...N}$ are two independent sets of N simulations of the input vector $\boldsymbol{X}$ and $(\varepsilon_k)_{k=1...N}$ is a sample of N realizations of the random function $\varepsilon(\boldsymbol{u})$. To compute the sensitivity indices S_i, the same algorithm is used with two independent samples of $(\varepsilon_k)_{k=1...N}$. In the same way, the total sensitivity index S_{T_ε} is derived from the algorithm of Saltelli [17].

The major drawback of this method is that it may be cpu time consuming. An accurate estimation of Sobol indices by this naive Monte-Carlo method requires more than thousand model evaluations for one input parameter. In complex industrial applications, it is intractable due to the cpu time cost of one model evaluation and the possible large number of input parameters.

4.3.2 The "trigger" parameter method

Dealing with spatially distributed input variables, Tarantola et al. [23] propose an alternative that uses an additional scalar input parameter ξ - called "trigger" parameter. $\xi \sim U[0,1]$ governs the random function simulation. For each simulation, if $\xi < 0.5$, the functional parameter $\varepsilon(\boldsymbol{u})$ is fixed to a nominal value $\varepsilon_0(\boldsymbol{u})$ (for example the mean $\mathbb{E}[\varepsilon(\boldsymbol{u})]$), while if $\xi > 0.5$, the functional parameter $\varepsilon(\boldsymbol{u})$ is simulated. Using this methodology, it is possible to estimate how sensitive the model output is to the presence of the random function. Tarantola et al. [23] use the Extended FAST method to compute the first order and total sensitivity indices of 6 scalar input factors and 2 additional "trigger" parameters.

For their study, the sensitivity indices according to the "trigger" parameters are small and the authors conclude that it is unnecessary to model these spatial errors more accurately.

Contrary to the previous method, there is no restriction about the sensitivity indices estimation procedure - i.e. Monte-Carlo, FAST, quasi Monte-Carlo. However, there are two major drawbacks for this approach :

- As the macroparameter method, it also requires the use of the computer model to perform the SA and it may be problematic for large cpu time computer models. This problem can be compensate by the use of an efficient quasi Monte-Carlo algorithm.
- As underlined by Tarantola et al. [23], ξ reflects only the presence or the absence of the stochastic errors on $\varepsilon_0(\boldsymbol{u})$. Therefore, the term $\text{Var}[\mathbb{E}(Y|\xi)]$ does not quantify the contribution of the random function variability to the output variability $\text{Var}(Y)$. We will discuss about the significance of $\text{Var}[\mathbb{E}(Y|\xi)]$ later, during our analytical function application.

4.3.3 The joint modeling approach

To perform a variance-based SA for time consuming computer models, some authors propose to approximate the computer code by a mathematical function (Marseguerra et al. [13], Volkova et al. [24]), often called response surface or metamodel (Fang et al. [2]). For metamodels with sufficient prediction capabilities, the bias due to the use of the metamodel instead of the true model is negligible. Several choices of metamodel can be found in the literature : polynomials, splines, Gaussian processes, neural networks, ... Thus, for the functional input problem, one strategy may be to fit a metamodel with a multi-dimensional scalar parameters representing $\varepsilon(\boldsymbol{u})$ as an input parameter - i.e. its discretization or its decomposition into an appropriate basis. However, this approach seems to be impracticable due to the potential large number of scalar parameters.

A second option is to substitute each random function realization for a discrete number, which can correspond to the scenario parameter of Ruffo et al. [15] (where the number of geostatistical realizations is finite and fixed, and where each different value of the discrete parameter corresponds to a different realization). However, in the general context, this restriction of the possible realizations of the input random function to a few ones is not acceptable.

The last solution considers $\varepsilon(\boldsymbol{u})$ as an uncontrollable parameter and a metamodel is fitted in function of the other scalar parameters $\boldsymbol{X}$:

$$Y_m = \mathbb{E}(Y|\boldsymbol{X}) \tag{7}$$

Therefore, using the relation

$$\text{Var}(Y) = \text{Var}[\mathbb{E}(Y|\boldsymbol{X})] + \mathbb{E}[\text{Var}(Y|\boldsymbol{X})] \tag{8}$$

it can be easily shown that the sensitivity indices of Y according to the scalar parameters $\mathbf{X} = (X_i)_{i=1...p}$ write (looss et al. [8])

$$S_i = \frac{\text{Var}[\mathbb{E}(Y_m|X_i)]}{\text{Var}(Y)}, \quad S_{ij} = \frac{\text{Var}[\mathbb{E}(Y_m|X_iX_j)]}{\text{Var}(Y)} - S_i - S_j, \quad \ldots \tag{9}$$

and can be computed by classical Monte-Carlo techniques applied on the metamodel Y_m. Therefore, using equation (8), the total sensitivity index of Y according to $\varepsilon(\boldsymbol{u})$ corresponds to the expectation of the unexplained part of $\text{Var}(Y)$ by the metamodel Y_m :

$$S_{T_\varepsilon} = \frac{\mathbb{E}[\text{Var}(Y|\boldsymbol{X})]}{\text{Var}(Y)} \tag{10}$$

Using this approach, our objective is altered because we cannot decompose the ε effects into elementary effect (S_ε) and interaction effects between ε and the scalar parameters $(X_i)_{i=1...p}$. However, we see below that our technique allows a qualitative appraisal of the interaction indices.

The sensitivity index estimations from equations (9) and (10) raise two difficulties :

1. It is well known that classical parametric metamodels (based on least squares fitting) are not adapted to estimate $\mathbb{E}(Y|\boldsymbol{X})$ accurately due to the presence of heteroscedasticity (induced by the effect of ε). Such cases are analyzed by looss et al. [8] (see chapter 3). The authors show that heteroscedasticity may lead to sensitivity indices misspecifications.
2. Classical non parametric methods, such as Generalized Additive Model (Hastie and Tibshirani [5]) and Gaussian process (Sacks et al. [16]) which can provide efficient estimation of $\mathbb{E}(Y|\boldsymbol{X})$ (examples are given in looss et al. [8]), even in high dimensional input cases ($p > 5$), are based on homoscedasticity hypothesis and do not propose the estimation of $\text{Var}(Y|\boldsymbol{X})$.

To solve the second problem, Zabalza-Mezghani et al. [26] propose the use of a theory developed for experimental data (McCullagh and Nelder [14]) : the simultaneous fitting of the mean and the dispersion by two interlinked Generalized Linear Models (GLM), which is called the joint modeling. Besides, to resolve the first problem, this approach has been extended by looss et al. [8] (see chapter 3) to non parametric models. This generalization allows more complexity and flexibility while fitting the data. The authors propose the use of Generalized Additive Models (GAMs) based on penalized smoothing splines (Wood [25]). GAMs allow model and variable selections using quasi-likelihood function, statistical tests on coefficients and graphical display. However, compared to other complex metamodels, GAMs impose an additive effects hypothesis. Therefore, two metamodels are obtained : one for the mean component $Y_m = \mathbb{E}(Y|\boldsymbol{X})$; and the other one for the dispersion component $Y_d = \text{Var}(Y|\boldsymbol{X})$. The sensitivity indices of $\boldsymbol{X}$ are computed using Y_m with the standard procedure (Eq. (9)), while the total sensitivity index of $\varepsilon(\boldsymbol{u})$ is computed from $\mathbb{E}(Y_d)$ (Eq. (10)). Using the explicit formula on Y_d and the associated regression diagnostics, qualitative sensitivity indices for the interactions between $\varepsilon(\boldsymbol{u})$ and the scalar parameters of $\boldsymbol{X}$ can also be deduced.

4.4 APPLICATION TO AN ANALYTICAL EXAMPLE

The three previously proposed methods are first illustrated on an artificial analytical model with two scalar input variables and one functional input :

$$Y = f(X_1, X_2, \varepsilon(t)) = \sin(X_1) + 7\sin(X_2)^2 + 0.1[\max_t(\varepsilon(t))]^4 \sin(X_1) \tag{11}$$

where $X_i \sim \mathcal{U}[-\pi;\pi]$ for $i = 1, 2$ and $\varepsilon(t)$ is a white noise, i.e. an i.i.d. stochastic process $\varepsilon(t) \sim \mathcal{N}(0,1)$. In our model simulations, $\varepsilon(t)$ is discretized in one hundred values : $t = 1 \ldots 100$. The function (11) is similar to the well-known Ishigami function (Homma and Saltelli [7]) but substitute the third parameter for the maximum of a stochastic process. Consequently, we call our function the white-noise Ishigami function (WN-Ishigami). Although the WN-Ishigami function is an artificial model, the introduction of the maximum of a stochastic process inside a model is quite realistic. For example, some computer models simulating physical phenomena can use the maximum of time-dependent variable - river height, rainfall quantity, temperature. Such input variable can be modeled by a temporal stochastic process.

As for the Ishigami function, we can immediately deduce from the formula (11) the sensitivity indices which are worse zero :

$$S_\varepsilon = S_{12} = S_{2\varepsilon} = S_{12\varepsilon} = 0 \tag{12}$$

Then, we have

$$S_{T_1} = S_1 + S_{1\varepsilon}, \quad S_{T_2} = S_2, \quad S_{T_\varepsilon} = S_{1\varepsilon} \tag{13}$$

In the following, we focus our attention on the estimation of S_1, S_2 and S_{T_ε}.

Because of a particularly complex probability distribution of the maximum of a white noise, there is no analytical solution for the theoretical Sobol indices S_1, S_2 and $S_{1\varepsilon}$ for the WN-Ishigami function. Even with the asymptotic hypothesis (number of time steps tending to infinity), where the maximum of the white noise follows Generalized Extreme Value distribution, theoretical indices are unreachable. Therefore, our benchmark Sobol indices values are derived from the Monte-Carlo method. However, these benchmark values can be considered as relevant because of the negligible computation time required to evaluate equation (11).

Indices	Macroparameter				"Trigger" parameter			
	Sobol algo		Saltelli algo		Sobol algo		Saltelli algo	
	Values	sd	Values	sd	Values	sd	Values	sd
S_1	0.540	1.3e-2	0.551	1.6e-2	0.304	1.3e-2	0.330	1.8e-2
S_{T_1}	—	—	0.808	2.0e-2	—	—	0.656	1.4e-2
S_2	0.197	1.1e-2	0.207	0.8e-2	0.329	1.4e-2	0.348	1.5e-2
S_{T_2}	—	—	0.212	0.7e-3	—	—	0.532	1.3e-2
$S_{1\varepsilon}$	0.268	2.4e-2	—	—	0.177	2.2e-2	—	—
S_{T_ε}	—	—	0.248	1.3e-2	—	—	0.336	1.4e-2

TAB. 4.1 – **Sobol sensitivity indices (with standard deviations sd) obtained from two Monte-Carlo algorithms (Sobol [21] and Saltelli [17]) and two integration methods of the functional input ε (macroparameter and "trigger" parameter) on the WN-Ishigami function. "—" indicates that the value is not available.**

4.4.1 The macroparameter and "trigger" parameter methods

Table 4.1 contains the Sobol indices estimates using the macroparameter and "trigger" parameter methods. As explained before, we can only use the two algorithms based on independent Monte-Carlo samples : the algorithm of Sobol [21] which computes S_1, S_2, $S_{1\varepsilon}$, and the algorithm of Saltelli [17] which computes the first order indices S_1, S_2 and the total sensitivity indices S_{T_1}, S_{T_2}, S_{T_ε}. For the estimation, the size of the Monte-Carlo samples are limited to $N = 10000$ because of memory computer limit. Indeed, the functional input $\varepsilon(u)$ contains for each simulation set 100 values. Then, the input sample matrix has the dimension $N \times 102$ which becomes extremely large when N increases. To evaluate the effect of this limited Monte-Carlo sample size N, each Sobol index estimate is associated to a standard-deviation estimated by bootstrap - with 100 replicates of the input-output sample. The obtained standard-deviations are relatively small, of the order of 0.01, which is rather sufficient for our exercise.

For the macroparameter method, the theoretical relations between indices given in (13) are verified. We are therefore confident with the estimates obtained with this method (which is in addition theoretically well-founded) and we choose the Sobol indices obtained with Saltelli's algorithm as the indices references :

$$S_1 = 55.1\%, \quad S_2 = 20.7\%, \quad S_{T_\varepsilon} = 24.8\%$$

The S_ε, S_{12}, $S_{\varepsilon 2}$ and $S_{12\varepsilon}$ indices (Eq. (12)) are not reported in table 4.1 as estimates are negligible.

With the "trigger" parameter method, the obtained values in table 4.1 are not close to the reference values. The inadequacies are larger than 30% for all the indices, and can be larger than 60% for a few ones (S_2 and S_{T_2}). Moreover, the relations given in (13) are not satisfied at all. Actually, replacing the input parameter $\varepsilon(u)$ by ξ which governs the presence or the absence of the functional input parameter changes the model. When ε is not simulated, it is replaced by its mean (zero) and the WN-Ishigami function becomes $Y = \sin(X_1) + 7\sin(X_2)^2$. Therefore, the mix of the WN-Ishigami model and this new model perturbs the estimation of the sensitivity indices, even those unrelated to ε (like X_2).

This example confirms our expectation : the sensitivity indices derived from the "trigger" parameter method have not the same sense that the classical ones, i.e., the measure of the contribution of the

input parameter variability to the output variable variability. The sensitivity indices obtained with these two methods are unconnected because the "trigger" parameter method changes the structure of the model.

4.4.2 The joint modeling approach

We apply now the joint modeling approach which requires an initial input-output sample to fit the joint metamodel - the mean component Y_m and the dispersion component Y_d. For our application, a learning sample size of $n = 500$ was considered; i.e., n independent random samples of $(X_1, X_2, \varepsilon(\boldsymbol{u}))$ were simulated leading to n observations for Y. Let first remark that this method is extremely less cpu time consuming than the previous ones which needed a 10000-size sample.

Joint GLM and joint GAM fitting procedures are fully described in looss et al. [8] (see chapter 3). Some graphical residual analyses are particularly well suited to check the relevance of the mean and dispersion components of the joint models. In the following, we give the results of the joint models fitting on a learning sample $(X_1, X_2, \varepsilon(\boldsymbol{u}), Y)$. Let us recall that we fit a model to predict Y in function of (X_1, X_2).

Joint GLM fitting

For the joint GLM, fourth order polynomial for the parametric form of the model is considered. Moreover, only the explanatory terms are retained 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 appreciate the model goodness-of-fit. The mean component gives :

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-5.79193  -0.59880   0.03988   0.64202   3.51148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.92634    0.22128   8.706   <2e-16 ***
X1           4.74256    0.16198  29.278   <2e-16 ***
I(X2^2)      2.22879    0.14130  15.773   <2e-16 ***
I(X1^3)     -0.51398    0.02453 -20.951   <2e-16 ***
I(X2^4)     -0.28501    0.01588 -17.952   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 1.010101)

    Null deviance: 1901.0  on 499  degrees of freedom
Residual deviance:  500.0  on 495  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

The explained deviance of this model is $D_{expl} = 74\%$. It can be seen that it remains 26% of non explained deviance due to the model inadequacy and/or to the functional input parameter. The predictivity coefficient, i.e. coefficient of determination R^2 computed on a test sample, is $Q_2 = 71\%$. Q_2 is relatively coherent with the explained deviance.

For the dispersion component, using analysis of deviance techniques, none significant explanatory variable were found : the heteroscedastic character of the data has not been retrieved. Thus, the dispersion component is supposed to be constant ; and the joint GLM model is equivalent to a simple GLM - but with a different fitting process.

Joint GAM fitting

At present, we try to model the data by joint GAM. The resulting model is described by the following features ($s(.)$ denotes a penalized spline smoothing term) :

```
Mean component:
```

```
Family: quasi
Link function: identity
Formula:
y ~ X1 + s(X1) + s(X2)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19914    0.08727   48.12   <2e-16 ***
X1          -5.39131    0.34285  -15.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
       edf Est.rank     F p-value
s(X1) 5.503        8 144.1  <2e-16 ***
s(X2) 8.738        9 316.5  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.979   Deviance explained = 90.5%
GCV score = 1.0683   Scale est. = 1.0336    n = 500

Dispersion component:

Family: Gamma
Link function: log
Formula:
d ~ s(X1)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98812    0.07965   12.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
       edf Est.rank     F p-value
s(X1) 8.814        9 28.39  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0481   Deviance explained = 26.3%
GCV score = 3.2355   Scale est. = 3.172     n = 500
```

The explained deviance of the mean component is $D_{expl} = 90\%$ and the predictivity coefficient is $Q_2 = 77\%$. Therefore, the joint GAM approach outperforms the joint GLM one. Indeed, the proportion of explained deviance is clearly greater for the GAM model. Even if this is obviously related to an increasing number of parameters; this is also explained as GAMs are more adjustable than GLMs. This is confirmed by the increase of the predictivity coefficient - from 71% to 77%. Moreover, due to the GAMs flexibility, the explanatory variable X_1 is identified for the dispersion component. The interaction between X_1 and the functional input parameter $\varepsilon(\boldsymbol{u})$ which governs the heteroscedasticity of this model is therefore retrieved.

Sobol indices

From the joint GLM and the joint GAM, Sobol sensitivity indices can be computed using equations (9) and (10) - see Table 4.2. The reference values are extracted from the macroparameter method and Saltelli's algorithm in table 4.1. The standard deviation estimates (sd) are obtained from 100 repetitions of the Monte-Carlo estimation procedure - which uses $N = 10000$ model computations for one index estimation. The joint GLM and joint GAM gives good estimations of S_1 and S_2. Despite the joint GLM leads to an accurate estimation for S_{T_ε}, we will see later that it is a lucky break. A problem occurs with the estimation of S_{T_ε} with joint GAM. In fact, an efficient modeling of $\mathrm{Var}(Y|\boldsymbol{X})$ is difficult, which is a common statistical difficulty in heteroscedastic regression problems (Antoniadis & Lavergne [1]). Another way to estimate the total sensitivity index S_{T_ε} is to compute the unexplained variance of the mean component model given directly by $1 - Q_2$, with Q_2 the predictivity coefficient of the mean component model. In practical applications, Q_2 can be estimated via leave-one-out or

cross validation procedures. In our analytical case, the index estimated with this method and the joint GAM gives a correct estimation - 0.23 instead of 0.25.

Indices	Reference	Joint GLM			Joint GAM		
	Values	Values	sd	Method	Values	sd	Method
S_1	0.551	0.572	4e-3	MC	0.569	5e-3	MC
S_2	0.207	0.179	8e-3	MC	0.233	7e-3	MC
S_{T_ε}	0.248	0.250	2e-3	MC	0.197	1e-3	MC
		0.29	—	Q_2	0.23	—	Q_2
S_{12}	0	0	—	Eq	0	—	Eq
$S_{1\varepsilon}$	0.248	0	—	Eq	> 0	—	Eq
$S_{2\varepsilon}$	0	0	—	Eq	0	—	Eq
$S_{12\varepsilon}$	0	0	—	Eq	0	—	Eq
S_{T_1}	0.808	0.832	4e-3	Eq	—	—	—
S_{T_2}	0.212	0.179	8e-3	Eq	0.233	7e-3	Eq
S_ε	0	0.250	2e-3	Eq	—	—	—

TAB. 4.2 – **Sobol sensitivity indices (with standard deviations) for the WN-Ishigami function : exact and estimated values from joint GLM and joint GAM (fitted with a 500-size sample). "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.**

For the other sensitivity indices, the conclusions draw from the GLM formula are completely erroneous : as the dispersion component is constant, the interaction indices are null. Thus, $S_\varepsilon = S_{T_\varepsilon} = 0.25$ while $S_\varepsilon = 0$ in reality. In contrary, the deductions draw from GAM formulas are correct : (X_1, ε) interaction sensitivity is positive, $S_{2\varepsilon} = S_{12\varepsilon} = 0$, $S_{T_2} = S_2$, $S_{12} = S_{23} = S_{123} = 0$. The drawback of this method is that some indices (S_{T_1} and S_ε) remain unknown due to the non separability of the dispersion component effects.

By estimating Sobol indices with those obtained from other learning samples, we observe that the estimates are rather dispersed : it seems that the estimates are not robust according to different learning samples for the joint models. To examine this effect, we propose to study two different sample sizes ($n = 200$ and $n = 500$). For each sample size, we repeat 25 times, the fitting process on different learning samples, and we compute Sobol indices as previously. In fact, for each sample size, we obtain 25 estimates of each sensitivity index. The variability of the indices is due at present to the learning sample variability. Figures 4.1 and 4.2 show the results of this investigation, which are particularly convincing. The boxplots are based on the 25 different estimates. From these figures, several conclusion can be drawn :

- for the joint GAM, the boxplot interquartile interval of each index contains its reference value. In contrary, the joint GLM fails to obtain correct estimates : except for S_1, the sensitivity reference values are outside the interquartile intervals of the obtained boxplots ;
- the superiority of the joint GAM with respect to the joint GLM is corroborated, especially for S_2 and S_{T_ε} ;
- the increase of the learning sample size has no effect on the joint GLM results (due to the parametric form of this model). However, for the joint GAM, boxplots widths are strongly reduced

from $n = 200$ to $n = 500$. In addition, the mean estimates seem to converge to the reference values ;

- as explained before, the estimation of S_{T_ε} using the predictivity coefficient Q_2 is markedly better than through the dispersion component model. This is not the case for the joint GLM.

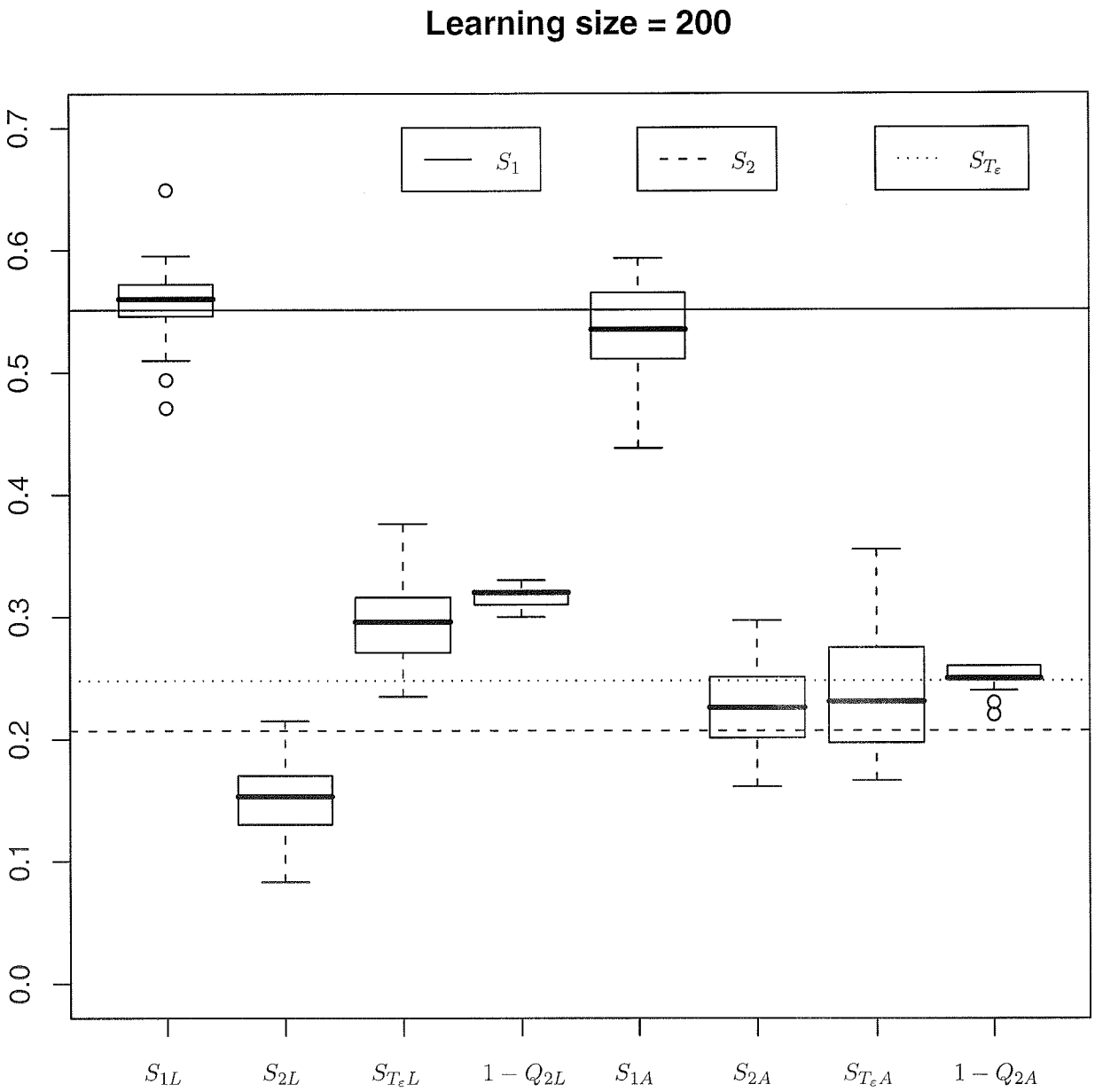

Fig. 4.1 – **WN-Ishigami application. Comparison of Sobol indices estimates. Reference values :** $S_1, S_2, S_{T_\varepsilon}$**. Joint GLM :** $S_{1L}, S_{2L}, S_{T_\varepsilon L}, 1 - Q_{2L}$**. Joint GAM :** $S_{1A}, S_{2A}, S_{T_\varepsilon A}, 1 - Q_{2A}$**.** $1 - Q_{2L}$ **(resp.** $1 - Q_{2A}$**) is the estimation of** S_{T_ε} **via the joint GLM (resp. joint GAM)** Q_2 **coefficient. Learning sample size :** $n = 200$**.**

In conclusion, this example shows that the joint models, and specially the joint GAM, can adjust complex heteroscedastic situations. Moreover, the joint models offer a theoretical basis to compute efficiently global sensitivity indices for models with functional input parameter.

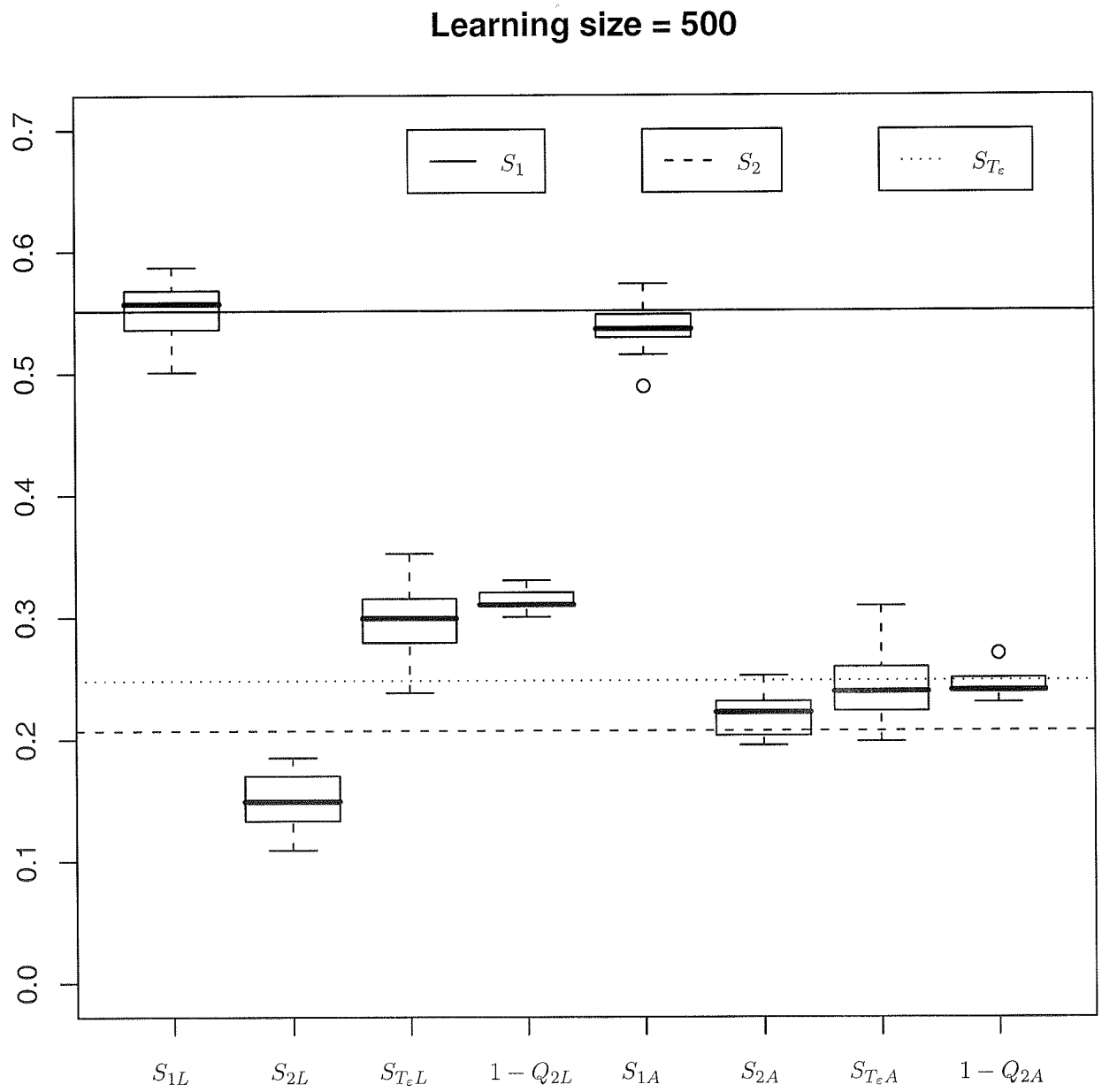

Fig. 4.2 – **WN-Ishigami application. Comparison of Sobol indices estimates. Reference values :** $S_1, S_2, S_{T_\varepsilon}$**. Joint GLM :** $S_{1L}, S_{2L}, S_{T_\varepsilon L}, 1 - Q_{2L}$**. Joint GAM :** $S_{1A}, S_{2A}, S_{T_\varepsilon A}, 1 - Q_{2A}$**.** $1 - Q_{2L}$ **(resp.** $1 - Q_{2A}$**) is the estimation of** S_{T_ε} **via the joint GLM (resp. joint GAM)** Q_2 **coefficient. Learning sample size :** $n = 500$**.**

4.5 APPLICATION TO A NUCLEAR FUEL IRRADIATION SIMULATION

The METEOR computer code, developed within the Fuel Studies Department in CEA Cadarache, studies the thermo-mechanical behavior of the fuel rods under irradiation in a nuclear reactor core. In particular, it computes the fission gas swelling and the cladding creep (Garcia et al. [3]). These two output variables are considered in our analysis. These variables are of fundamental importance for the physical understanding of the fuel behavior and for the monitoring of the nuclear reactor core.

Input parameters of such mechanical models can be evaluated either by database analyses, arguments invoking simplifying hypotheses, expert judgment. All these considerations lead to assign to each input parameter a nominal value associated with an uncertainty. In this study, six uncertain input parameters are considered : the initial internal pressure X_1, the pellet and cladding radius X_2, X_3, the microstructural fuel grain diameter X_4, the fuel porosity X_5 and the time-dependent irradiation power $P(t)$. $X_1, \ldots, X_5$ are all modeled by Gaussian independent random variables with the following coefficient of variations : cv$(X_1) = 0.019$, cv$(X_2) = 1.22 \times 10^{-3}$, cv$(X_3) = 1.05 \times 10^{-3}$, cv$(X_4) = 0.044$, cv$(X_5) = 0.25$. The last variable $P(t)$ is a temporal function (discretized in 3558 values) and its uncertainty $\varepsilon(t)$ is modelled like a stochastic process. For simplicity, a temporal white noise (of uniform law ranging between -5% and $+5\%$) was introduced.

As in the previous application, additionally to its scalar random variables, the model includes an input functional variable $P(t)$. To compute Sobol indices of this model, we have first tried to use the macroparameter method. We have succedeed to perform the calculations with $N = 1000$ (for the Monte-Carlo sample sizes of Eqs. (6a), (6b) and (6c)). The sensitivity indices estimates have been obtained after 10 computation days and were extremely imprecise, with strong variations between 0 and 1. Because of the required cpu time, an increase of the sample size N to obtain acceptable sensitivity estimates was unconceivable. Therefore, the goal of this section is to show how the use of the joint modeling approach allows to estimate the sensitivity indices of the METEOR model and, in particular, to quantify the functional input variable influence.

500 METEOR calculations were carried out using 500 Monte-Carlo sampling of the input parameters. As expected, the white noise on $P(t)$ generates an increase in the standard deviation of the output variables (compared to simulations without a white noise) : 6% increase for the variable *fission gas swelling* and 60% for the variable *cladding creep*.

4.5.1 Gas swelling

We start by studying the gas swelling model output. With a joint GLM, the following result for Y_m and Y_d were obtained :

$$\begin{cases} Y_m & = -76 - 0.4X_1 + 20X_2 + 8X_4 + 134X_5 + 0.02X_4^2 - 2X_2X_4 - 6X_4X_5 \\ \log(Y_d) & = -2.4X_1 \end{cases} \quad (14)$$

The explained deviance of the mean component is $D_{expl} = 86\%$. As the residual analyses of mean and dispersion components do not show any biases, the resulting model seems satisfactory. The joint GAM was also fitted on these data and led to similar results. Thus, it seems that spline terms are useless and that a joint GLM model is suited.

Table 4.3 shows the results for the Sobol indices estimation using Monte-Carlo methods applied on the metamodel (14). The standard deviation (sd) estimates are obtained from 100 repetitions of the Monte-Carlo estimation procedure -which uses 10^5 model computations for one index estimation. It is useless to perform the Monte-Carlo estimation for some indices because they can be deduced from the joint model equations. For example, $S_3 = 0$ (resp. $S_{\varepsilon 2} = 0$) because X_3 (res. X_2) is not involved in the mean (resp. dispersion) component in equation (14). Moreover, we know that $S_{\varepsilon 1} > 0$ because X_1 is an explanatory variable inside the dispersion component Y_d. However, this formulation does not allow to have any idea about S_ϵ which reflects the first order effect of ε. Therefore, some indices are not accessible, such as S_ε and $S_{\varepsilon 1}$ non distinguishable inside the total sensitivity index S_{T_ε}. Finally,

we can check that $\sum_{i=1}^{5} S_i + \sum_{i,j=1,i<j}^{5} S_{ij} + S_{T_\varepsilon} = 1$ holds - up to numerical approximations.

It can be seen that X_4 (grain diameter) and X_5 (fuel porosity) are the most influent factors (each one having 40% of influence), and do not interact with the irradiation power $P(t)$ (representing by its uncertainty ε). In addition, the effect of $P(t)$ is not negligible (14%) and parameter X_1 (internal pressure) acts only with its interaction with $P(t)$. A sensitivity analysis by fixing X_1 could allow us to obtain some information about the first order effect of ε in the model.

4.5.2 Cladding creep

We study now the cladding creep model output. With a joint GLM, the model for Y_m and Y_d is :

$$\begin{cases} Y_m & = -2.75 + 1.05X_2 - 0.15X_3 - 0.58X_5 \\ \log(Y_d) & = 156052 - 76184X_2 + 9298X_2^2 \end{cases} \tag{15}$$

The explained deviance of the mean component is $D_{expl} = 26\%$. As the residual analyses of mean and dispersion components show some biases, the resulting model is not satisfactory.

For the joint GAM, the spline terms $\{s(X_2), s(X_3), s(X_5)\}$ and $s(X_2)$ are added within the mean component and the dispersion component respectively. The explained deviance of the mean component is $D_{expl} = 29\%$ which is not significantly greater than 26%. However, as the mean component residual biases of the joint GAM are smaller than those observed for the joint GLM, the joint GAM seems to be more relevant than the joint GLM.

Table 4.3 shows the Sobol indices estimates using Monte-Carlo methods and deductions from the joint model equations. For the joint GLM and joint GAM of the cladding creep, $\sum_{i=1}^{5} S_i + \sum_{i,j=1,i<j}^{5} S_{ij} + S_{T_\varepsilon} = 1$ holds – up to numerical imprecisions. Due to the proximity of the two joint models, results are similar. This analysis shows that the parameter X_2 (pellet radius) explains 28% of the uncertainty of the cladding creep phenomenon, while the other scalar parameters have negligible influence. The greater part of the cladding creep variance (70%) is explained by the irradiation power uncertainty. Physicists may be interested in quantifying the interaction influence between the pellet radius and the irradiation power. Unfortunately, this interaction is not available for the moment in our analysis.

Indices	Gas swelling			Cladding creep					
	Joint GLM			Joint GLM			Joint GAM		
	Values	sd	Method	Values	sd	Method	Values	sd	Method
S_1	0.029	6e-3	MC	0.000	1e-3	MC	0.000	1e-3	MC
S_2	0.024	5e-3	MC	0.294	1e-4	MC	0.282	2e-4	MC
S_3	0	—	Eq	0.006	1e-3	MC	0.007	1e-3	MC
S_4	0.394	5e-3	MC	0.000	1e-3	MC	0.000	1e-3	MC
S_5	0.409	6e-3	MC	0.006	1e-3	MC	0.006	1e-3	MC
S_{24}	0.002	5e-3	MC	0	—	Eq	0	—	Eq
S_{45}	0.000	9e-3	MC	0	—	Eq	0	—	Eq
other S_{ij}	0	—	Eq	0	—	Eq	0	—	Eq
S_{T_ε}	0.143	1e-4	MC	0.694	1e-4	MC	0.704	3e-4	MC
S_ε	—	—	—	—	—	—	—	—	—
$S_{\varepsilon 1}$	—	—	—	0	—	Eq	0	—	Eq
$S_{\varepsilon 2}$	0	—	Eq	—	—	—	—	—	—
other $S_{\varepsilon i}$	0	—	Eq	0	—	Eq	0	—	Eq
S_{T_1}	—	—	—	0.000	1e-3	Eq	0.000	4e-3	Eq
S_{T_2}	0.026	7e-3	Eq	—	—	—	—	—	—
S_{T_3}	0	—	Eq	0.006	1e-3	Eq	0.007	4e-3	Eq
S_{T_4}	0.396	7e-3	Eq	0.000	1e-3	Eq	0.000	4e-3	Eq
S_{T_5}	0.409	0.011	Eq	0.006	1e-3	Eq	0.006	4e-3	Eq

TAB. 4.3 – **Sobol sensitivity indices (with standard deviations sd) from joint models fitted on the outputs of the METEOR code. "Method" indicates the estimation method : MC for the Monte-Carlo procedure and Eq for a deduction from the joint model equation. "—" indicates that the value is not available.**

4.6 CONCLUSION

This chapter has proposed a solution to perform global sensitivity analysis for time consuming computer models which depend on functional input parameters, such as a stochastic process or a random field. Our purpose concerned the computation of variance-based importance measures of the model output according the the uncertain input parameters. We have discussed a first natural solution which consists in integrating the functional input parameter inside a macroparameter, and using standard Monte-Carlo algorithms to compute sensitivity indices. This solution is not applicable for time consuming computer code. We have discussed another solution, used in previous studies, based on the replacement of the functional input parameter by a "trigger" parameter that governs the integration or not of the functional input uncertainties. However, the estimated sensitivity indices are strongly biased due to changes in the model structure carrying out by the method itself. Finally, we have proposed an innovative solution, the joint modeling method, based on a preliminary step of double (and joint) metamodel fitting, which resolves the large cpu time problem of Monte-Carlo methods. It consists in rejecting the functional input parameters in noisy input variables. Then, two metamodels depending only on the scalar random input variables are simultaneously fitted : one for the mean function and one for the dispersion (variance) function.

Tests on an analytical function have shown the relevance of the joint modeling method, which provides all the sensitivity indices of the scalar input parameters and the total sensitivity index of the functional input parameter. In addition, it reveals in a qualitative way the influential interactions between the functional parameter and the scalar input parameters. A research way for the future would be to distinguish the contributions of several functional input parameters, who are at the moment totally mixed in one sensitivity index. This is the main drawback of the proposed method.

In an industrial application, the feasibility and usefulness of our methodology was established. Indeed, other methods are not applicable in this application because of large cpu time of the computer code. To a better understanding of the model behavior, the information brought by the global sensitivity analysis can be very useful to the physicist or the modeling engineer. The joint model can also serve in propagation uncertainty and reliability studies of complex models, containing input random functions, to obtain some mean predictions with their confidence intervals.

4.7 ACKNOWLEDGMENTS

This work has been done with Mathieu Ribatet (CEMAGREF Lyon).

4.8 REFERENCES

[1] A. Antoniadis and C. Lavergne. Variance function estimation in regression by wavelet methods. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and statistics*. Springer, 1995.

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

[3] P. Garcia, C. Struzik, M. Agard, and V. Louche. Mono-dimensional mechanical modelling of fuel rods under normal and off-normal operating conditions. *Nuclear Science and Design*, 216 :183–201, 2002.

[4] J.E. Gentle. *Random number generation and Monte Carlo methods*. Springer, second edition, 2003.

[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, M. Ribatet, and A. Marrel. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics*, submitted, 2008. Available at URL : http ://fr.arxiv.org/abs/0802.0443v1.

[9] 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.

[10] J. Jacques, C. Lavergne, and N. Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering and System Safety*, 91 :1126–1134, 2006.

[11] C. Lantuéjoul. *Geostatistical simulations - Models and algorithms.* Springer, 2002.

[12] M. Marquès, J.F. Pignatel, P. Saignes, F. D'Auria, L. Burgazzi, C. Müller, R. Bolado-Lavin, C. Kirchsteiger, V. La Lumia, and I. Ivanov. Methodology for the reliability evaluation of a passive system and its integration into a probabilistic safety assesment. *Nuclear Engineering and Design*, 235 :2612–2631, 2005.

[13] M. Marseguerra, R. Masini, E. Zio, and G. Cojazzi. Variance decomposition-based sensitivity analysis via neural networks. *Reliability Engineering and System Safety*, 79 :229–238, 2003.

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

[15] P. Ruffo, L. Bazzana, A. Consonni, A. Corradi, A. Saltelli, and S. Tarantola. Hydrocarbon exploration risk evaluation through uncertainty and sensitivity analysis techniques. *Reliability Engineering and System Safety*, 91 :1155–1162, 2006.

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

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

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

[19] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Salsana, and S. Tarantola. *Global sensitivity analysis - The primer.* Wiley, 2008.

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

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

[22] I.M. Sobol. Global sensitivity indices for non linear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55 :271–280, 2001.

[23] 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.

[24] 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*, 22 :17–31, 2008.

[25] S. Wood. *Generalized Additive Models : An Introduction with R.* CRC Chapman & Hall, 2006.

[26] 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.