Advanced methodology for uncertainty propagation in computer experiments with large number of inputs: *application to accidental scenario in a Pressurized water Reactor*

A. Marrel
CEA, DEN, DER, SESI, LEMS, 13108 Saint-Paul-lez-Durance, France

B. Iooss
EDF R&D, 6 Quai Watier, 78401 Chatou, France
Institut de Mathématiques de Toulouse, 31062 Toulouse, France.

ABSTRACT: Complex computer codes are often too time expensive to be directly used to perform uncertainty propagation or sensitivity analysis. A solution to cope with this problem consists in replacing the cpu-time expensive computer model by a cpu inexpensive mathematical function, called metamodel. Among the metamodels classically used in computer experiments, the Gaussian process (Gp) model has shown strong capabilities to solve practical problems. However, in case of high dimensional experiments (with typically several tens of inputs), the Gp metamodel building process remains difficult. To face this limitation, we propose a general methodology which combines several advanced statistical tools. First, an initial space-filling design is performed providing a full coverage of the high-dimensional input space (Latin hypercube sampling with optimal discrepancy property). From this, a screening based on dependence measures is performed. More specifically, the Hilbert-Schmidt independence criterion which builds upon kernel-based approaches for detecting dependence is used. It allows ordering the inputs by decreasing primary influence, for the purpose of the metamodeling. Furthermore, significance tests based either on asymptotic theory or permutation technique are performed to identify a group of potentially non-influential inputs. Then, a joint Gp metamodel is sequentially built with the group of influential inputs as explanatory variables. The residual effect of the group of non-influential inputs is captured by the dispersion part of the joint metamodel. Then, a sensitivity analysis based on variance decomposition can be performed through the joint Gp metamodel. The efficiency of the methodology is illustrated on a thermal-hydraulic calculation case simulating accidental scenario in a Pressurized water Reactor.

1 INTRODUCTION

Quantitative assessment of the uncertainties tainting the results of computer simulations is nowadays a major topic of interest in both industrial and scientific communities. One of the key issues in such studies is to get information about the output when the numerical simulations are expensive to run. For example, in nuclear engineering problems, one often faces up with cpu time consuming numerical models and, in such cases, uncertainty propagation, sensitivity analysis, optimization processing and system robustness analysis become difficult tasks using such models. In order to circumvent this problem, a widely accepted method consists in replacing cpu-time expensive computer models by cpu inexpensive mathematical functions, called metamodels (Fang et al. 2006). This solution has been applied extensively and has shown its relevance especially when simulated phenomena are related to a small number of random input variables (see Forrester et al. 2008 for example).

However, in case of high dimensional numerical experiments (with typically several tens of inputs), depending on the complexity of the underlying numerical model, the metamodel building process remains difficult, even unfeasible. For example, the Gaussian process (Gp) model (Santner et al. 2003) which has shown strong capabilities to solve practical problems, has some caveats when dealing with high dimensional problems. The main difficulty relies on the estimation of Gp hyperparameters. Manipulating pre-defined or well-adapted Gp kernels (as in Muehlenstaedt et al. 2012, Durrande et al. 2013) is a current research way, while coupling the estimation procedure with variable selection techniques has been proposed by several authors (Welch et al. 1992, Marrel et al. 2008, Woods and Lewis 2017).

In this paper, following the latter technique, we pro-

pose a rigorous and robust method for building a Gp metamodel with a high-dimensional vector of inputs before using it to perform variance-based sensitivity analysis.

To build this metamodel, we use a sequential methodology where the technical core are updated with more relevant statistical techniques. For example, the screening step is raised by the use of recent and powerful techniques in terms of variable selection using a small number of model runs. Second, contrary to the previous works, we do not remove the non-selected inputs from the Gp model, keeping the uncertainty caused by the dimension reduction by using the joint metamodel technique (Marrel et al. 2012). The integration of this residual uncertainty is important in terms of robustness of subsequent safety studies and sensitivity analysis. Finally, a sensitivity analysis based on variance decomposition is performed through the joint Gp metamodel, yielding both the estimation of the influence of each selected inputs and the total effect of the group of non-selected inputs.

Each step of our methodology is detailed in a dedicated section and illustrated on a guideline application, namely a thermal-hydraulic calculation case simulating accidental scenario in a nuclear reactor. This use-case is first described in the following section.

2 THERMAL-HYDRAULIC TEST-CASE

Our use-case consists in thermal-hydraulic computer experiments, typically used in support of regulatory work and nuclear power plant design and operation. Indeed, some safety analysis considers the so-called "Loss Of Coolant Accident" (LOCA), which takes into account a double-ended guillotine break with a specific size piping rupture. It is modeled with code CATHARE 2.V2.5 which simulated the thermal-hydraulic responses during a LOCA in a Pressurized water Reactor (Mazgaj et al. 2016).

In this use-case, $d = 27$ scalar input variables of CATHARE are uncertain. They correspond to various system parameters as initial conditions, boundary conditions, some critical flowrates, interfacial friction coefficients, condensation coefficients, ... The output variable of interest is a single scalar which is the maximal peak cladding temperature during the accident transient.

In our problem, minimal and maximal values are defined for each uncertain input and, in the framework of probabilistic approach, their uncertainties are modeled by probability laws defined on the domain of variation (uniform, log-uniform, truncated normal and truncated log-normal laws). Moreover, the d inputs are supposed independent. Our first objective with this use-case is to provide a good metamodel for sensitivity analysis, uncertainty propagation and, more generally, safety studies. Indeed, the cpu-time cost of this computer code is too important to develop all the statistical analysis required in a safety study only using direct calculations of the computer code. A metamodel would allow to develop more complete and robust demonstration.

In what follows, the system under study is generically denoted

$$Y = g\left(X_1, \ldots, X_d\right) \tag{1}$$

where $g(\cdot)$ is the numerical model (also called the computer code), whose output Y and input parameters $X_1, \ldots, X_d$ belong to some measurable spaces $\mathcal{Y}$ and $\mathcal{X}_1, \ldots, \mathcal{X}_d$ respectively. $\mathbf{X} = (X_1, \ldots, X_d)$ is the input vector and we suppose that $\mathcal{X} = \prod_{k=1}^{d} \mathcal{X}_k \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. For a given value of the vector of inputs $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, a simulation run of the code yields an observed value $y = g(\mathbf{x})$.

3 STEP 1: INITIAL DESIGN OF EXPERIMENTS

The objective of the initial sampling step is to investigate the whole variation domain of the uncertain parameters in order to fit a predictive metamodel which approximates as accurately as possible the code in the whole domain of variation of the uncertain parameter. For this, we use a space-filling design (SFD) of a certain number n of experiments, providing a full coverage of the high-dimensional input space (Fang et al. 2006). This design enables to investigate the domain of variation of the uncertain parameters and provides a learning sample.

For the SFD type, a Latin Hypercube Sample (LHS) with optimal space-filling and good projection properties (Woods and Lewis 2017) would be well adapted. In particular, Fang et al. (2006, Damblin et al. (2013) have shown the importance of ensuring good low-order sub-projection properties. Maximum projection designs (Joseph et al. 2015) or low-centered L^2 discrepancy LHS (Jin et al. 2005) are then particularly well-suited.

Mathematically, this corresponds to the sample $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$ which is performed on the model g. This yields n model output values denoted $\left\{y^{(1)}, \ldots, y^{(n)}\right\}$ with $y^{(i)} = g(\mathbf{x}^{(i)})$. The obtained learning sample is denoted (X_s, Y_s) with $X_s = \left[\mathbf{x}^{(1)^T}, \ldots, \mathbf{x}^{(n)^T}\right]^T$ and $Y_s = \left[y^{(1)}, \ldots, y^{(n)}\right]^T$. The goal is to build an approximating model of g using the n-sample (X_s, Y_s).

The number n of simulations is a compromise between the CPU time required for each simulation and the number of input parameters. Some thumb rules propose to choose n at least as large as 10 times the dimension d of the input vector (Loeppky et al. 2009, Marrel et al. 2008).

To build the metamodel for the LOCA test case, $N = 500$ CATHARE simulations of this test case are performed following a space-filling LHS with good projection properties as the design of experiments. The obtained inputs-output sample constitutes the learning sample.

Remark 3.1 Note that the input values are sampled following their prior distributions defined on their variation ranges. Indeed, as we are not ensured to be able to build a sufficiently accurate metamodel, we prefer sample the inputs following the probabilistic distributions in order to have at least a probabilized sample of the uncertain output, on which statistical characteristics could be estimated. Moreover, as explained in the next section, dependence measures can be directly estimated on this sample, providing first usable results of sensitivity analysis.

4 STEP 2: INITIAL SCREENING BASED ON DEPENDENCE MEASURE

From the learning sample, a screening technique is performed in order to identify the primary influential inputs (PII) on the model output variability. It has been recently shown that screening based on dependence measures (Da Veiga 2015, De Lozzo and Marrel 2016) or on derivative-based global sensitivity measures (Kucherenko and Iooss 2017, Roustant et al. 2017) are very efficient methods which can be directly applied on a SFD. One of their great interest is that, additionally to their screening job, the sensitivity indices that they provide can be quantitatively interpreted and used to order the PII by decreasing influence, paving the way for a sequential building of metamodel.

In the considered LOCA test case, the adjoint model is not available and the derivatives of the model output are therefore not computed because of their costs. The screening step will then be based only on HSIC dependence measures, directly estimated from the inputs-output sample.

4.1 *Screening based on HSIC dependence measure*

Da Veiga (2015) and more recently De Lozzo and Marrel (2016) have proposed to use dependence measures for screening purpose, by applying them directly on a SFD. These sensitivity indices are not the classical ones variance-based measures (see Iooss and Lemaître 2015 for a global review). They consider higher order information about the output behavior in order to provide more detailed information. Among them, the Hilbert-Schmidt independence criterion (HSIC) introduced by Gretton et al. (2005) builds upon kernel-based approaches for detecting dependence, and more particularly on cross-covariance operators in reproducing kernel Hilbert spaces (RKHS).

If we consider two RKHS $\mathcal{F}_k$ and $\mathcal{G}$ of functions $\mathcal{X}_k \to \mathbb{R}$ and $\mathcal{Y} \to \mathbb{R}$ respectively, the crossed-covariance $C_{X_k,Y}$ operator associated to the joint distribution of (X_k, Y) is the linear operator defined for every $f_{X_k} \in \mathcal{F}_k$ and $g_Y \in \mathcal{G}$ by:

$$\langle f_{X_k}, C_{X_k,Y} g_Y \rangle_{\mathcal{F}_k} = \mathrm{Cov}\left(f_{X_k}, g_Y\right). \tag{2}$$

$C_{X_k,Y}$ generalizes the covariance matrix by representing higher order correlations between X_k and Y through nonlinear kernels. The HSIC criterion is then defined by the Hilbert-Schmidt norm of the cross-covariance operator:

$$\mathrm{HSIC}(X_k, Y)_{\mathcal{F}_k,\mathcal{G}} = \|C_k\|^2_{HS}. \tag{3}$$

From this, Da Veiga (2015) introduces a normalized version of the HSIC which provides a sensitivity index of X_k:

$$R^2_{\mathrm{HSIC},k} = \frac{\mathrm{HSIC}(X_k, Y)}{\sqrt{\mathrm{HSIC}(X_k, X_k)\mathrm{HSIC}(Y, Y)}}. \tag{4}$$

Gretton et al. (2005) also propose a Monte Carlo estimator of $\mathrm{HSIC}(X_k, Y)$ and a plug-in estimator can be deduced for $R^2_{\mathrm{HSIC},k}$. Note that Gaussian kernel functions with empirical estimations of the variance parameter are used in our application (see Gretton et al. 2005 for details).

Then, from the estimated R^2_{HSIC}, independence tests can be performed for a screening purpose. The objective is to separate the inputs into two sub-groups, the significant ones and the non-significant ones. For a given input X_k, it aims at testing the null hypothesis "$\mathcal{H}_0^{(k)}$: X_k and Y are independent", against its alternative "$\mathcal{H}_1^{(k)}$: X_k and Y are dependent". The significance level[1] of these tests is hereinafter noted α. Several statistical hypothesis tests are available: asymptotic versions, spectral extensions and bootstrap versions for non-asymptotic case. All these tests are described and compared in De Lozzo and Marrel (2016); a guidance to use them for a screening purpose is also proposed. At the end of the screening step, the inputs selected as significant are also ordered by decreasing R^2_{HSIC}. This order will be used for the sequential metamodel building in step 3.

4.2 *Application on LOCA test case*

From the learning sample of $N = 500$ simulations, R^2_{HSIC} dependence measures are estimated and bootstrap tests with $\alpha = 0.1$ are performed. Eleven inputs are selected as significantly influential. Ordering them by decreasing R^2_{HSIC} reveals the predominance influence of X_{10} ($R^2_{\mathrm{HSIC}} \approx 0.39$), followed by X_2, X_{12} and X_{22}, ($R^2_{\mathrm{HSIC}} \approx 0.04$, 0.02 and 0.02 respectively). X_{15},

[1]The significance level of a statistical hypothesis test is the rate of the type I error which corresponds to the rejection of the null hypothesis $\mathcal{H}_0$ when it is true.

X_{13}, X_9, X_5, X_{14}, X_{26} and X_{27} have a lower influence (R^2_{HSIC} around 0.01)) and the others variables are considered as negligible by statistical tests.

Note that the estimated HSIC and the results of significant tests are relatively stable when the learning sample size varies from $N = 300$ to $N = 500$. Only two or three selected variables with a very low HSIC (R^2_{HSIC} around 0.01) can differ. This confirms the robustness of the HSIC indices and the associated significance tests for qualitative sorting and screening purpose.

In the next steps, the eleven significant inputs are considered as the explanatory variables, denoted PII, in the joint metamodel and will be successively included in the building process. The other sixteen variables will be joined in a so-called *uncontrollable* parameter.

5 STEP 3: JOINT GP METAMODEL WITH SEQUENTIAL BUILDING PROCESS

Among all the metamodel-based solutions (polynomials, splines, neural networks, etc.), we focus our attention on the Gaussian process (Gp) regression, which extends the kriging principles of geostatistics to computer experiments by considering the correlation between two responses of a computer code depending on the distance between input variables. The Gp-based metamodel presents some real advantages compared to other metamodels: exact interpolation property, simple analytical formulations of the predictor, availability of the mean squared error of the predictions and the proved efficiency of the model (Santner et al. 2003).

However, for its application to complex industrial problems, developing a robust implementation methodology is required. Indeed, fitting a Gp model implies the estimation of several hyperparameters involved in the covariance function. In complex situations (e.g. large number of inputs), some difficulties can arise from the parameter estimation procedure (instability, high number of hyperparameters, see Marrel et al. 2008 for example). To tackle this issue, we propose a progressive estimation procedure which combines the result of the previous screening step and a joint Gp approach (Marrel et al. 2012).

5.1 *Sequential building process based on successive inclusion of explanatory variables*

At the end of the screening step, the inputs selected as significant (group of PII) are ordered by decreasing influence. The sorted PII are successively included in the metamodel explanatory inputs while the other inputs (remaining PII and the sixteen non-selected inputs) are joined in a single macro-parameter which is considered as an uncontrollable parameter (i.e. a stochastic parameter, notion detailed in section 5.2). Thus, at the j^{th} iteration, a joint Gp metamodel is built with, as explanatory inputs, the j sorted PII. The definition and building procedure of a joint Gp is fully described in Marrel et al. (2012) and summarized in the next subsection.

However, building a Gp or a joint Gp involves to perform a numerical optimization in order to estimate all the parameters of the metamodel (covariance hyperparameters and variance parameter). As we usually consider in computer experiments anisotropic (stationary) covariance, the number of hyperparameters linearly increases with the number of inputs. In order to improve the robustness of the optimization process and deal with a large number of inputs, the estimated hyperparameters obtained at the $(j-1)^{th}$ iteration are used, as starting points for the optimization algorithm. This procedure is repeated until the inclusion of all the PII. Note that this sequential estimation process is directly adapted from the one proposed by Marrel et al. (2008).

5.2 *Joint Gp metamodel*

In the framework of stochastic computer codes, Zabalza et al. (1998) proposed to model the mean and dispersion of the code output by two interlinked Generalized Linear Models (GLM), called "joint GLM". Marrel et al. (2012) extends this approach to several nonparametric models and obtains the best results with two interlinked Gp models, called "joint Gp". In this case, the stochastic input is considered as an uncontrollable parameter denoted $\mathbf{X}_\varepsilon$ (i.e. governed by a seed variable).

We extend this approach to a group of non-explanatory variables. More precisely, the input variables $\mathbf{X} = (X_1, \ldots, X_d)$ are divided in two subgroups: the explanatory ones denoted $\mathbf{X}_{\mathbf{exp}}$ and the others denoted $\mathbf{X}_\varepsilon$. The output is thus defined by $y = g(\mathbf{X}_{\mathbf{exp}}, \mathbf{X}_\varepsilon)$. Under this hypothesis, the joint metamodeling approach yields building two metamodels, one for the mean Y_m and another for the dispersion component Y_d:

$$Y_m(\mathbf{X}_{\mathbf{exp}}) = \mathbb{E}(Y|\mathbf{X}_{\mathbf{exp}}) \tag{5}$$

$$Y_d(\mathbf{X}_{\mathbf{exp}}) = \text{Var}(Y|\mathbf{X}_{\mathbf{exp}}) = \mathbb{E}\left[(Y - Y_m(\mathbf{X}_{\mathbf{exp}}))^2|\mathbf{X}_{\mathbf{exp}}\right]. \tag{6}$$

To fit these mean and dispersion components, we propose to use the methodology proposed by Marrel et al. (2012). First, an initial Gp denoted $Gp_{m,1}$ is estimated for the mean component with homoscedastic nugget effect. A nugget effect is required to relax the interpolation property of the Gp metamodel, which would yield zero residuals for the whole learning sample. Then, a second Gp, denoted $Gp_{v,1}$, is

built for the dispersion component with, here also, an homoscedastic nugget effect. $Gp_{v,1}$ is fitted on the squared residuals from the predictor of $Gp_{m,1}$. Its predictor is considered as an estimator of the dispersion component. The predictor of $Gp_{v,1}$ provides an estimation of the dispersion at each point, which is considered as the value of the heteroscedastic nugget effect. The homoscedastic hypothesis is so removed and a new Gp, denoted $Gp_{m,2}$, is fitted on data, with the estimated heteroscedastic nugget. Finally, the Gp on the dispersion component is updated from $Gp_{m,2}$ following the same methodology as for $Gp_{v,1}$.

Remark 5.1 Note that some parametric choices are made for all the Gp metamodels: a constant trend and a Matérn stationary anisotropic covariance are chosen. All the hyperparameters (covariance parameters) and the nugget effect (when homoscedastic hypothesis is done) are estimated by maximum likelihood optimization process.

5.3 *Assessment of metamodel accuracy*

To evaluate the accuracy of the metamodel, we use the predictivity coefficient Q^2:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{n_{\text{test}}} \left(y^{(i)} - \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} y^{(i)}\right)^2} \tag{7}$$

where $(x^{(i)})_{1 \leq i \leq n_{\text{test}}}$ is a test sample, $(y^{(i)})_{1 \leq i \leq n_{\text{test}}}$ are the corresponding observed outputs and $(\hat{y}^{(i)})_{1 \leq i \leq n_{\text{test}}}$ are the metamodel predictions. Q^2 corresponds to the coefficient of determination in prediction and can be computed on a test sample independent from the learning sample or by cross-validation on the learning sample. The closer to one the Q^2, the better the accuracy of the metamodel.

5.4 *Application on LOCA test case*

The joint Gp metamodel is built from the learning sample of $N = 500$: the eleven PII identified at the end of the the screening step are considered as the explanatory variables while the sixteen others are considered as the uncontrollable parameter. Gps on mean and dispersion components are built using the sequential building process described in section 5.1 where PII ordered by decreasing R^2_{HSIC} are successively included in Gp. Q^2 coefficient of mean component Gp_m is computed by cross validation at each iteration of the sequential building process. The results which are given by Table 5.4 show an increasing predictivity until its stabilization around 0.87, which illustrates the robustness of building process. The first four PII make the major contribution yielding a Q^2 around 0.8, the four following ones yield minor improvements (increase of 0.02 on average for each input) while the three last PII does not improve the Gp predictivity.

Table 1: Evolution of Gp_m metamodel predictivity during the sequential process building, for each new additional PII.

Additional PII	X_{10}	X_2	X_{12}	X_{22}	X_{15}	X_{13}
Q^2	0.60	0.64	0.70	0.79	0.81	0.83
Additional PII	X_9	X_5	X_{14}	X_{26}	X_{27}	
Q^2	0.85	0.85	0.87	0.87	0.87	

Thus, only 13% of the output variability remains not explained by Gp_m, this includes both the inaccuracy of the Gp_m (part of Y_m not fitted by Gp) and the total effect of the uncontrollable parameter, i.e. the group of non-selected inputs.

6 STEP 4: VARIANCE-BASED SENSITIVITY ANALYSIS

Sensitivity Analysis (SA) methods allow to answer the question "How do the input parameters variations contribute, qualitatively or quantitatively, to the variation of the output?" (Saltelli et al. 2008). These tools can detect non-significant input parameters in a screening context, determinate the most significant ones, measure their respective contributions to the output or identify an interaction between several inputs which impacts strongly the model output. From this, engineers can guide the characterization of the model by reducing the output uncertainty: for instance, they can calibrate the most influential inputs and fix the non-influential ones to nominal values. Many surveys on SA exist in the literature, such as Kleijnen (1997), Frey and Patil (2002) or Helton et al. (2006). SA can be divided into two sub-domains: the Local SA (LSA) and the Global SA (GSA). The first one studies the effects of small input perturbations around nominal values on the model output (Cacuci 1981) while the second one considers the impact of the input uncertainty on the output over the whole variation domain of uncertain inputs (Saltelli et al. 2008). We focus here on one of the most widely used GSA indices, namely Sobol' indices which are based on output variance decomposition.

6.1 *Sobol' indices*

A classical approach in GSA consists of computing the first-order and total Sobol' indices which are based on the output variance decomposition (Sobol 1993, Homma and Saltelli 1996). If the variables $X_1, \ldots, X_d$ are independent and if $\mathbb{E}[g^2(X)] < +\infty$, we can apply the Hoeffding decomposition to the random variable $g(X)$ (Hoeffding 1948):

$$g(X) \quad = \quad \sum_{u \subset \{1,\ldots,d\}} g_u(X_u) \tag{8}$$

where $g_\emptyset = \mathbb{E}[g(X)]$, $g_i(X_i) = \mathbb{E}[g(X)|X_i] - g_\emptyset$ and $g_u(X_u) = \mathbb{E}[g(X)|X_u] - \sum_{v \subset u} g_v(X_v)$, with $X_u = (X_i)_{i \in u}$, for all $u \subset \{1, \ldots, d\}$. All the 2^d terms in (8) have zero mean and are mutually uncorrelated with each other. This decomposition is unique and leads

to the Sobol' indices. These are the elements of the $g(X)$ variance decomposition according to the different groups of input parameter interactions in (8). More precisely, for each $u \subset \{1, \ldots, d\}$, the first-order and total Sobol' sensitivity indices of X_u are defined by:

$$S_u = \frac{Var\left[g_u(X_u)\right]}{Var\left[g(X)\right]} \text{ and } S_u^T = \sum_{v \supset u} S_v.$$

S_u represents the part of the output variance explained by X_u, independently from the other inputs, and S_u^T is the part of the output variance explained by X_u considered separately and in interaction with the other input parameters.

In practice, we are usually interested in the first-order sensitivity indices $S_1, \ldots, S_d$, the total ones $S_1^T, \ldots, S_d^T$ and sometimes in the second-order ones S_{ij}, $1 \leq i < j \leq d$. The model g is devoid of interactions if $\sum_{i=1}^d S_i \approx 1$.

Sobol' indices are widely used in GSA because they are easy to interpret and directly usable in a dimension reduction approach. However, their estimation (based on Monte-Carlo methods for example) requires a large number of model evaluations, which is intractable for time expensive computer codes. An common solution consists in using a metamodel to compute these indices. Note that, when the Q^2 of the metamodel is estimated on a probabilized sample of the inputs, it provides an estimation of the part of variance unexplained by the metamodel. This can be kept in mind when interpreting the Sobol' indices estimated with the metamodel.

6.2 *Sobol' indices with a joint Gp metamodel*

In the case where a joint Gp metamodel is used to take into account an uncontrollable input X_ε, we have shown in Marrel et al. (2012) how to deduce Sobol' sensitivity indices from this joint metamodel. Indeed, the variance of the output variable $Y(\mathbf{X_{exp}}, X_\varepsilon)$ can be rewritten and deduced from the two metamodels:

$$\text{Var}Y(\mathbf{X_{exp}}, X_\varepsilon)\text{Var}_{\mathbf{X_{exp}}}\left[Y_m(\mathbf{X_{exp}})\right] + \mathbb{E}_{\mathbf{X_{exp}}}\left[Y_d(\mathbf{X_{exp}})\right] \tag{9}$$

where $\mathbb{E}_X$ (resp. Var_X) denotes the mean (resp. variance) operator with respect to the pdf of X. Furthermore, the variance of Y is the sum of the contributions of all the d controllable inputs $\mathbf{X_{exp}} = (X_1, \ldots, X_d)$ and the uncontrollable one X_ε:

$$\text{Var}(Y) = V_\varepsilon(Y) + \sum_{i=1}^{d} \sum_{|J|=i} \left[V_J(Y) + V_{J\varepsilon}(Y)\right] \tag{10}$$

where $V_\varepsilon(Y) = \text{Var}_{X_\varepsilon}[\mathbb{E}_{\mathbf{X_{exp}}}(Y|X_\varepsilon)]$, $V_i(Y) = \text{Var}_{\mathbf{X_i}}[\mathbb{E}_{\mathbf{X_{-i}}}(Y|X_i)]$, $V_{i\varepsilon}(Y) = \text{Var}_{X_i X_\varepsilon}[\mathbb{E}_{\mathbf{X_{exp,-i}}}(Y|X_i X_\varepsilon)] - V_i(Y) - V_\varepsilon(Y)$, $V_{ij}(Y) = \text{Var}_{X_i X_j}[\mathbb{E}_{\mathbf{X_{-i,-j}}}(Y|X_i X_j)] - V_i(Y) - V_j(Y) \ldots$

Variance of the mean component $Y_m(\mathbf{X})$ denoted hereafter Y_m can be also decomposed:

$$\text{Var}(Y_m) = \sum_{i=1}^{d} \sum_{|J|=i} V_J(Y_m)\ . \tag{11}$$

As $V_i(Y_m) = \text{Var}_{\mathbf{X_i}}\mathbb{E}_{\mathbf{X_{exp,-i}}}[\mathbb{E}_{X_\varepsilon}(Y|\mathbf{X_{exp}})|X_i] = V_i(Y)$, Sobol' indices according to input variables $\mathbf{X}_{exp} = (X_i)_{i=1\ldots d}$ can be derived and estimated from Y_m:

$$S_J = \frac{V_J(Y_m)}{\text{Var}(Y)} \text{ for any } J \subset \mathbf{X}_{exp}. \tag{12}$$

Similarly, the total sensitivity index of X_ε is given by:

$$S_\varepsilon^{tot} = \frac{V_\varepsilon(Y) + \sum_{i=1}^{d} \sum_{|J|=i} V_{J\varepsilon}(Y)}{\text{Var}(Y)} = \frac{\mathbb{E}_{\mathbf{X}_{exp}}[Y_d(\mathbf{X}_{exp})]}{\text{Var}(Y)}\ . \tag{13}$$

Note that, as $Y_d(\mathbf{X_{exp}})$ is a positive random variable, positivity of S_ε^{tot} is guaranteed. In practice, $\text{Var}(Y)$ can be estimated from the data or from simulations of the fitted joint model, using equation (9).

S_ε^{tot} is interpreted as the total sensitivity index of the uncontrollable process. The limitation of this approach is that only the total part of uncertainty related to X_ε is estimated; its individual effect is not distinguished from its interaction with the other parameters. However, these potential interactions could be pointed out, considering all the primary and total effects of all the other parameters. The SA of Y_d can also be a relevant indicator: if an input variable X_i is not influential on Y_d, we can deduce that $S_{i\varepsilon}$ is equal to zero.

6.3 *Results on LOCA test case*

From the joint Gp built in section 5.4, Sobol' indices of PII are estimated from Gp_m metamodel using equation (12), $\text{Var}(Y)$ being estimated with Gp_m and Gp_d using equation (9). For this, intensive Monte Carlo methods are used (see e.g. pick-and-freeze estimator of Gamboa et al. 2016). The first Sobol' indices of PII are given by Table 6.3 and represent 85 % of the total variance of the output. X_{10} remains the major influential input with 59 % of explained variance, followed to a lesser extend by X_{12} and X_{22} with for each

Table 2: First Sobol' indices of PII (in %), estimated with Gp_m metamodel.

Input	X_{10}	X_2	X_{12}	X_{22}	X_{15}	X_{13}
1st Sobol' index	59	3	8	8	2	1
Input	X_9	X_5	X_{14}	X_{26}	X_{27}	
1st Sobol' index	2	0	2	0	0	

of them 8% of variance. The *partial* total Sobol' indices involving only PII and derived from Gp_m show that additional 4 % of variance is due to interaction between X_{10}, X_{12} and X_{22}. The other PII have negligible influence. Lastly, all PII explain around 89 % of the output variance, of which 79 % is only due to X_{10}, X_{12} and X_{22}. From Gp_d metamodel and using equation (13),the total effect of the uncontrollable parameter, i.e. the group of the sixteen not-explanatory inputs, is estimated to 9.7 %. This includes the effect of the uncontrollable parameter alone and in interaction with the PII. To further investigate these interactions, Sobol'-based SA and HSIC-based statistical dependence tests are applied on Y_d and reveal that only X_{10}, X_{14}, X_2, X_{22} and X_4 potentially interact with the uncontrollable parameter.

7 CONCLUSION AND PROSPECTS

Using an efficient sequential building process, we fitted a predictive joint Gp metamodel on a high dimensional thermal-hydraulic test case simulating accidental scenario in a Pressurized water Reactor (LOCA test case). An initial screening step based on advanced dependence measures and associated statistical tests enabled to identify a group of significant inputs, allowing dimension reduction. The efforts of optimization when fitting the metamodel fitting can be concentrated on the main influential inputs and the robustness of metamodeling is thus increased. Moreover, thanks to the joint metamodel approach, the non-selected inputs are not completely removed: the residual uncertainty due to dimension reduction is integrated in the metamodel and the global influence of non-selected inputs is so controlled.

From this joint Gp metamodel, several statistical analyses, not feasible with the numerical model due to its computational cost, become accessible. Thus, on LOCA application, a sensitivity analysis based on variance decomposition is performed using the joint Gp: Sobol' indices are computed and reveal that the output is mainly explained by four uncertain inputs: one input is strongly influential with around 60% of output variance explained, the three others being of minor influence. The quite less influence of all the other inputs is also confirmed.

The next step is to use the joint Gp metamodel to perform uncertainty propagation for the estimation of failure probabilities and quantiles. In the LOCA test case, we are particularly interested by the estimation of high quantile (at the order of 95% to 99%) of the model output temperature. In nuclear safety, methods of conservative computation of quantiles (Nutt and Wallis 2004) have been largely studied. However, several complementary information are often useful and are not accessible in a high-dimensional context. Then, we expect that the joint Gp metamodel could help to access this information: the uncertainty of the influential inputs will be directly and accurately propagated through the mean component of the joint metamodel while a confidence bound could be derived from the dispersion component in order to take into account the residual uncertainty of the other inputs. On this last point, the interest of heteroscedastic approach in joint Gp could also be illustrated and compared with its homoscedastic version.

8 ACKNOWLEDGMENTS

We are grateful to Henri Geiser and Thibault Delage who performed the computations of the CATHARE code.

REFERENCES

Cacuci, D. (1981). Sensitivity theory for nonlinear systems. I. Nonlinear functional analysis approach. *Journal of Mathematical Physics 22*, 2794.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation 85*, 1283–1305.

Damblin, G., M. Couplet, & B. Iooss (2013). Numerical studies of space filling designs: Optimization of Latin hypercube samples and subprojection properties. *Journal of Simulation 7*, 276–289.

De Lozzo, M. & A. Marrel (2016). New improvements in the use of dependence measures for sensitivity analysis and screening. *Journal of Statistical Computation and Simulation 86*, 3038–3058.

Durrande, N., D. G. O., Roustant, & L. Carraro (2013). ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis 155*, 57–67.

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

Forrester, A., A. Sobester, & A. Keane (Eds.) (2008). *Engineering design via surrogate modelling: a practical guide*. Wiley.

Frey, H. & S. Patil (2002). Identification and review of sensitivity analysis methods. *Risk Analysis 22*, 553–578.

Gamboa, F., A. Janon, T. Klein, A. Lagnoux, & C. Prieur (2016). Statistical inference for sobol pick freeze Monte Carlo methods. *Statistics 50*, 881–902.

Gretton, G., O. Bousquet, A. Smola, & B. Schölkopf (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings Algorithmic Learning Theory*, pp. 63–77. Springer-Verlag.

Helton, J., J. Johnson, C. Salaberry, & C. Storlie (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety 91*, 1175–1209.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics 19*, 293–325.

Homma, T. & A. Saltelli (1996). Importance measures in global sensitivity analysis of non linear models. *Reliability Engi-*

neering and System Safety 52, 1–17.

Iooss, B. & P. Lemaître (2015). A review on global sensitivity analysis methods. In C. Meloni and G. Dellino (Eds.), *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*. Springer.

Jin, R., W. Chen, & A. Sudjianto (2005). An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference 134*, 268–287.

Joseph, V., E. Gul, & S. Ba (2015). Maximum projection designs for computer experiments. *Biometrika 102*, 371–380.

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

Kucherenko, S. & B. Iooss (2017). Derivative-based global sensitivity measures. In R. Ghanem, D. Higdon, and H. Owhadi (Eds.), *Springer Handbook on Uncertainty Quantification*. Springer.

Loeppky, J., J. Sacks, & W. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*, 366–376.

Marrel, A., B. Iooss, S. Da Veiga, & M. Ribatet (2012). Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing 22*, 833–847.

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

Mazgaj, P., J.-L. Vacher, & S. Carnevali (2016). Comparison of CATHARE results with the experimental results of cold leg intermediate break LOCA obtained during ROSA-2/LSTF test 7. *EPJ Nuclear Sciences & Technology 2*(1).

Muehlenstaedt, T., O. Roustant, L. Carraro, & S. Kuhnt (2012). Data-driven Kriging models based on FANOVA-decomposition. *Statistics & Computing 22*, 723–738.

Nutt, W. & G. Wallis (2004). Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. *Reliability Engineering and System Safety 83*, 57–77.

Roustant, O., F. Barthe, & B. Iooss (2017). Poincaré inequalities on intervals - application to sensitivity analysis. *submitted https://hal.archives-ouvertes.fr/hal-01388758*.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Salsana, & S. Tarantola (2008). *Global sensitivity analysis - The primer*. Wiley.

Santner, T., B. Williams, & W. Notz (2003). *The design and analysis of computer experiments*. Springer.

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

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

Woods, D. & S. Lewis (2017). Design of experiments for screening. In R. Ghanem, D. Higdon, and H. Owhadi (Eds.), *Springer Handbook on Uncertainty Quantification*. Springer.

Zabalza, I., J. Dejean, & D. Collombier (1998, september). Prediction and density estimation of a horizontal well productivity index using generalized linear models. In *ECMOR VI, Peebles*.