International Journal for Uncertainty Quantification, 5 (5): 393–414 (2015)

HIGH DIMENSIONAL SENSITIVITY ANALYSIS USING SURROGATE MODELING AND HIGH DIMENSIONAL MODEL REPRESENTATION

Martin Kubicek,[1,*] Edmondo Minisci,[1] & Marco Cisternino[2]

[1] *University of Strathclyde, James Weir Building 75 Montrose Street Glasgow, G1 IXJ, Scotland, United Kingdom*

[2] *OPTIMAD Engineering s.r.l, Via Giacinto Collegno 18, 10143 Torino, Italy*

[*] Address all correspondence to Martin Kubicek E-mail: Martin.Kubicek@strath.ac.uk

Original Manuscript Submitted: 9/8/2014; Final Draft Received: 5/5/2015

In this paper, a new non-intrusive method for the propagation of uncertainty and sensitivity analysis is presented. The method is based on the cut-HDMR approach, which is here derived in a different way and new conclusions are presented. The cut-HDMR approach decomposes the stochastic space into sub-domains, which are separately interpolated via a selected interpolation technique. This leads to a dramatic reduction of necessary samples for high dimensional spaces and decreases the influence of the Curse of Dimensionality. The proposed non-intrusive method is based on the coupling of an interpolation technique with the cut-HDMR (high dimension model representation) approach. The new conclusions obtained from the new derivation of the cut-HDMR approach allow one to interpolate each stochastic domain separately, including all stochastic variables and interactions between variables. Moreover, the same conclusions allow one to neglect non-important stochastic domains and therefore, drastically reduce the number of samples to detect and interpolate the higher order interactions. A new sampling strategy is introduced, which is based on a tensor product, but it uses the idea of Smoylak sparse grid for higher domains. For this work, the multi-dimensional Lagrange interpolation technique is selected and is applied for all parts of the cut-HDMR approach. However, the nature of the method allows one to use a combination of various interpolation techniques. The sensitivity analysis is performed on the surrogate model using the Monte Carlo sampling. In this work, the Sobol's approach is followed and sensitivity indices are established for each variable and interaction. Moreover, due to the obtained conclusions, the separate surrogate models allow one to visualize the uncertainty in the high dimensional space via histograms. The usage of a histogram for each stochastic domain allows one to establish full statistical properties of a given stochastic domain. This helps the user to better understand the stochastic propagation for the model of interest. The proposed interpolation technique and sensitivity analysis approach are tested on a simple example and applied on the well-known Borehole problem. Results of the proposed method are compared to the Monte Carlo sampling using the mean value and the standard deviation. Results of the sensitivity analysis of the Borehole case are compared to the literature results and the statistical visualization of each variable is provided.

KEY WORDS: *sensitivity analysis, uncertainty quantification, high dimensional space, cut-HDMR*

1. INTRODUCTION

Modern computers allow one to perform a simulation of complex physical, mathematical, and environmental processes. These simulations allow one to closely examine the physical phenomena and perform simulations under various conditions. Moreover, they allow one to perform a large number of simulations for a fraction of time compared to complex trials in testing chamber. Unfortunately, the complexity of modern physical models reached such a level

that a simple problem, like the sensitivity analysis of a given variable, can be hard to solve. Therefore, methods for estimation of sensitivity of input variables have been developed.

Most sensitivity analysis methods handle the complex code as a black box. This requires one to run the complex code many times with various setups. Among the most popular sensitivity analysis methods are the Monte Carlo (MC) based approaches. They randomly sample the stochastic domain with large number of samples and estimate the statistical properties of the given model.

The well-known Sobol sensitivity method [1, 2] is based on the MC approach. This method decomposes the output variance of the given model into parts attributable to input variables. The outcome of the method is the quantification of the influence of each variable and its interactions in the given model.

Another popular variance based method is the Fourier amplitude sensitivity testing (FAST) [3, 4]. The method performs the sensitivity analysis for a smaller number of required samples than the Sobol method. However, FAST is not capable of sensitivity analysis on interaction terms in a given model. Later, the FAST was extended to compute the total effect of a given variable. Despite the fact that FAST performs better (in terms of required sampling) than the Sobol method, the Sobol method is still the most widely used variance based method.

In order to get a better insight into the model and to visualize the influence of a variable, scatter plots [1, 2, 5, 6] can be used. Scatter plots are a very good way to estimate the behavior of a function in a given domain, but they require a great deal of experience in sensitivity analysis.

All the mentioned methods rely on a large number of samples, i.e., function calls. This implies a large computational burden and in many codes the time required to run sensitivity analysis would be infeasible. For the purpose of reducing the number of samples, the Method of Morris (MoM) [7] can be used. The MoM for global sensitivity analysis is a modification of one-step-at-a-time methods [2]. If one is interested in local sensitivity, the partial derivations can be used. Another well-used method is the linear regression [1, 2], but this approach requires that the model of interest is linear.

For the visualization of sensitivity, the CobWeb plots [8], also known as web diagrams, or the variable interaction network (VIN) diagrams [9] can be used. The Cobweb plots enable one to visualize the combination of the input variables, leading to a specific range of the output variable. The VIN diagrams can be used to track and to visualize additive structures in a function of interest.

A way to overcome the computational time of senstitivity analysis is to build a cheap surrogate model of the expensive function. Among the surrogate approaches, there are methods such as the stochastic collocation method (SC) [10, 11], the polynomial chaos (PC) [10–17], the Kriging surrogate model [4, 18, 19], and the Pade-Legendre approximation. These methods are non-intrusive by nature, i.e., they consider the code of interest as a black box. In the work of Sudret [20], the non-intrusive polynomial chaos was coupled with the Sobol method and it allows to one compute the Sobol indices directly from the PC expansion.

In the case of other surrogate models, the sensitivity indices are obtained by sampling these cheap models via MC approach. Unfortunately, one of the largest limitations of surrogate models is the so called Curse of Dimensionality (CoD), introduced by Richard Bellman [21]. This problem is still an issue and it limits the use of non-intrusive methods to a lower number of stochastic dimensions. Various sampling techniques were proposed to handle the CoD problem. The Latin Hyper-cube sampling (LHS) [4] was successfully used in various problems and some different approaches are available, such as LaPSO [22], uniform design (UD) [4], or Hammersley sampling [14, 23].

In the framework of uncertainty quantification (UQ) problems, Smolyak sparse grid [24–26] and its various modifications became very popular techniques. This sampling strategy combined with non-intrusive polynomial chaos (NIPC), gives very accurate results for a low number of samples. Unfortunately, even this approach is not affordable because of its high cardinality.

The cut-high dimensional model representation (cut-HDMR) [27, 28] was developed to decouple the interaction effects of chemical systems. It was successfully used in other fields such as uncertainty quantification [29, 30], sensitivity analysis [31, 32], and interpolation problems [33–35] and it proved to be a very efficient tool for high dimensional problems, especially for high dimensional integration. Unfortunately, only a limited number of papers was published on this topic.

In this work, all these problems together are addressed. The surrogate model is created and followed by a sensitivity analysis and a new approach used for UQ problems is developed. The approach is based on the cut-HDMR,

which is here derived in a different way. A new equation is established here and the cut-HDMR is derived from this equation. The proposed equation enlightens some important aspects of a high dimensional information propagation and important conclusions are derived here. Like the HDMR approach, the proposed approach allows one to decompose the stochastic space into sub-domains, which are then interpolated separately via a selected interpolation technique. Each interpolation technique is built accordingly to the conclusions obtained from a new derivation of the cut-HDMR. Since each domain is independent, also, the interpolation process is independent and therefore, only important domains are sampled and this dramatically reduces the necessary number of samples for high dimensional spaces. Although, the method can be coupled with any interpolation technique, only the multi-dimensional Lagrange interpolation (MLI) [11] was used here. The MC simulation is applied on each interpolated sub-domain to estimate not only the propagation of uncertainty, but also the sensitivity of a variable or a combination of variables. The outcome on each interpolated sub-domain can be visualized using histograms. This gives the user a completely new insight into the problem. A new sampling strategy for the given approach is introduced. The new sampling strategy is based on a Smolyak sparse grid idea and uses a lower number of samples in a high order stochastic space. This approach decreases the number of samples in high dimensional domains.

The paper is structured as follows. In the first part, a new derivation of the cut-HDMR approach is introduced and theoretical sensitivity indices are established. The second part is given to the numerical application of the cut-HDMR approach, its application to surrogate modeling and sensitivity analysis. Thr third part describes the necessary sampling strategy for surrogate models. The fourth part is given to applied examples of sensitivity analysis, and UQ approach. Lastly, the obtained results are discussed and conclusions are given.

2. THEORY

In order to derive the cut-HDMR, let us first introduce a new equation, the derivative equation (DE), which is derived from the analysis of variance (ANOVA) decomposition [9, 36]. Let us consider an integrable function, $f(\mathbf{x})$, which is defined on an n-dimensional unit hypercube—$[0,1]^n$ and $\mathbf{x} \in [0,1]^n$. The ANOVA representation of $f(\mathbf{x})$ can be

$$f(\mathbf{x}) = f_0 + \sum_{s=1}^{n} \sum_{i_1<..<i_s}^{n} f_{i_1 \ldots i_s}(\mathbf{x}) \tag{1}$$

where $1 \leq i_1 < i_2 < ... < i_s \leq n$ and $1 \leq s \leq n$. The explicit form of Eq. (1) is

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\leq i<j\leq n} f_{i,j}(x_i, x_j) + ... + f_{1,\ldots,n}(x_1, ..., x_n) \tag{2}$$

where f_0 is the constant term and represents the mean value of $f(\mathbf{x})$, the function $f_i(x_i)$ represents the contribution of variable x_i to function $f(\mathbf{x})$, the function $f_{i,j}(x_i, x_j)$ represents the pair correlated contribution to $f(\mathbf{x})$ by the input variables x_i and x_j, which are defined as $1 \leq i < j \leq n$, etc. The last term $f_{1,\ldots,n}(x_1, ..., x_n)$ contains the correlated contribution of all input variables and the total number of summands for Eq. (2) is 2^n.

The ANOVA representation is well known and its properties are well described in various works [31, 32, 35]. In order to obtain the separated contributions of the function of interest, it is necessary to obtain the derivative of the function of interest according to its variables. The derivative represents a separate contribution to the function of interest for a given variable. Therefore, consider a function, $f(\mathbf{x})$, which is derivable and integrable. Accordingly, all terms in Eq. (2) are integrable and derivable too. Let us derive each term in Eq. (2) accordingly to its generic variable x_i and obtain the infinitesimal increment

$$df_i(x_i) = \frac{\partial f(\mathbf{x})}{\partial x_i} dx_i \tag{3}$$

which for the two-dimensional terms is

$$df_{i,j}(x_i, x_j) = \frac{\partial f(\mathbf{x})}{\partial x_i, x_j} dx_i dx_j \tag{4}$$

Higher order terms are derived accordingly. Since the first term on the right side of Eq. (2) is a constant, i.e., $f_0 = C$, hence $df_0 = 0$, then the infinitesimal increment of function $f(\mathbf{x})$ can be written as

$$df(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} dx_i + \sum_{1 \leq i < j \leq n} \frac{\partial f(\mathbf{x})}{\partial x_i, x_j} dx_i dx_j + ... + \frac{\partial f(\mathbf{x})}{\partial x_1, ..., x_n} dx_1 ... dx_n \tag{5}$$

Equation (5) is the basic form of DE and it relates the change of the function of interest on the change of input variables. Moreover, the equation describes the relationship between the stochastic spaces. In other words, the propagation of information from the lower order stochastic space to the higher order stochastic space depends on given partial derivative of the function of interest. However, the most important aspect is the independence of each derivative. This justifies the application of multiple surrogate techniques for the problem of interest. The DE is very hard to use in the practical applications and obtaining derivatives from a function of interest is in many cases a hard task and in some cases practically impossible. Moreover, Eq. (5) gives the infinitesimal increment of the function in a point. Therefore, the integral form is introduced. The integral form is a necessary step to obtain the well-known cut-HDMR model. In order to derive the integral form, let us integrate Eq. (5) in the same way as derivatives were applied

$$f_i(x_i) = \int \frac{\partial f(\mathbf{x})}{\partial x_i} dx_i \tag{6}$$

which for the two-dimensional terms is

$$f_{i,j}(x_i, x_j) = \int \int \frac{\partial f(\mathbf{x})}{\partial x_i, x_j} dx_i dx_j \tag{7}$$

Higher order terms are derived accordingly. Each term is handled as a separate function and therefore, it is integrated separately.

The left-hand side term in Eq. (5) is integrated in the following way

$$\int df(\mathbf{x}) = f(\mathbf{x}) + C \tag{8}$$

where $C = -f_0$, i.e., it is a constant. However, it is more practical to use definite integrals. The definite integral gives the equation more physical meaning and represents a finite increment to the quantity of interest. Therefore, let us rewrite Eq. (8) in the following form:

$$\int_{f(^{\mathbf{c}}\mathbf{x})}^{f(\mathbf{x})} df(\xi) = f(\mathbf{x}) - f(^{\mathbf{c}}\mathbf{x}) \tag{9}$$

Using Eq. (9) and a definite integral for each summand, the integrated form of Eq. (5) reads

$$\begin{aligned} f(\mathbf{x}) - f(^{\mathbf{c}}\mathbf{x}) = & \sum_{i=1}^{n} \int_{^c x_i}^{x_i} \frac{\partial f(\xi)}{\partial \xi_i} d\xi_i \\ & + \sum_{1 \leq i < j \leq n} \int_{^c x_i}^{x_i} \int_{^c x_j}^{x_j} \frac{\partial f(\xi)}{\partial \xi_i, \xi_j} d\xi_i d\xi_j + ... + \int_{^c x_1}^{x_1} ... \int_{^c x_n}^{x_n} \frac{\partial f(\xi)}{\partial \xi_1, ..., \xi_n} d\xi_1 ... d\xi_n \end{aligned} \tag{10}$$

where $^{\mathbf{c}}\mathbf{x}$ represents a central position in the stochastic space, called the central point. In the cut-HDMR model this point is called anchored point. In this case, the central point is considered as the statistical mean value of a given stochastic random variable, i.e., $^c x_i = mean(x_i)$. Note that the central point can be selected arbitrarily, but to obtain valid sensitivity analysis, the mean value of the considered variable has to be selected. This assures that the function value at the central point represents the assumed expected mean, while the partial mean represents its deviation.

This equation is similar to the essence of the cut-HDMR approach and to clarify the similarity, the transformation of the derivative equation into the cut-HDMR approach is later introduced in Section 3. In this work, each integral part

of Eq. (10) is called increment function. The notation for the increment function is dF_k and the subscript represents the given increment function, which is bounded as follows: $1 \leq k \leq 2^n - 1$. The increment function represents a finite increment to the function $f(\mathbf{x})$ and its physical meaning is the influence of the given stochastic domain to the function $f(\mathbf{x})$. The number of integrable variables represents the order of the increment function and higher order increment functions, i.e., $\geq$2, represent the influence of the interactions between variables.

In order to obtain the sensitivity analysis, it is necessary to define the mean value and the variance for a given function. Using the integral form of DE, the mean value for Eq. (2) can be obtained. Let us remind that the terms in Eq. (10) are orthogonal and therefore, each term can be integrated separately. The function (10) is integrated into the following form:

$$\begin{aligned}\mu = \int_{-\infty}^{\infty} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} = f(^{c}\mathbf{x}) + \sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{c_{x_i}}^{x_i}\frac{\partial f(\xi)}{\partial \xi_i}d\xi_i p_i(x_i)dx_i \\ + \sum_{1\leq i<j\leq n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{c_{x_i}}^{x_i}\int_{c_{x_j}}^{x_j}\frac{\partial f(\xi)}{\partial \xi_i, \xi_j}d\xi_i d\xi_j p_{ij}(x_i, x_j)dx_i dx_j \\ + ... + \int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\int_{c_{x_1}}^{x_1}...\int_{c_{x_n}}^{x_n}\frac{\partial f(\xi)}{\partial \xi_1, ..., \xi_n}d\xi_1...d\xi_n p_{1...n}(x_1, ..., x_n)dx_1...dx_n \end{aligned} \tag{11}$$

where $p_i(x_i)$ is the probability density function (PDF) for the given distribution. Note that the probability of each variable is handled separately too. This is a direct consequence of the orthogonality of terms in Eq. (10). From Eq. (11), the partial expected value, μ_i, can be defined. The partial expected value for the first-order terms is written in the following way:

$$\mu_i = \int_{-\infty}^{\infty}\int_{c_{x_i}}^{x_i}\frac{\partial f(\xi)}{\partial \xi_i}d\xi_i p_i(x_i)dx_i \tag{12}$$

and for the second-order terms the function reads

$$\mu_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{c_{x_i}}^{x_i}\int_{c_{x_j}}^{x_j}\frac{\partial f(\xi)}{\partial \xi_i, \xi_j}d\xi_i d\xi_j p_{ij}(x_i, x_j)dx_i dx_j \tag{13}$$

where $p_{ij}(x_i, x_j)$ is the joint PDF. If random variables are independent from each other, then the PDF is given by

$$p_{ij}(x_i, x_j) = p_i(x_i)p_j(x_j) \tag{14}$$

Higher order partial expected values are defined accordingly. Unfortunately, the same approach cannot be applied to the higher order momentum functions. The explanation is given in the Appendix. The partial variance can still be defined, however, it cannot be summed as the expected value. The partial variance represents a variance of the given increment function and it is a very good estimation of sensitivity for given increment function. The first order partial variance, $\sigma_i^2(x_i)$, reads

$$\sigma_i^2 = \int_{-\infty}^{\infty}\left(\int_{c_{x_i}}^{x_i}\frac{\partial f(\xi)}{\partial \xi_i}d\xi_i - \mu_i\right)^2 p_i(x_i)dx_i \tag{15}$$

and for the second-order terms the function reads

$$\sigma_{ij}^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\int_{c_{x_i}}^{x_i}\int_{c_{x_j}}^{x_j}\frac{\partial f(\xi)}{\partial \xi_i, \xi_j}d\xi_i d\xi_j - \mu_{ij}\right)^2 p_{ij}(x_i, x_j)dx_i dx_j \tag{16}$$

Higher order partial variances are defined accordingly. The work of Sobol [36] is followed and sensitivity indices are defined in the following way:

$$S_k = \frac{\sigma_k^2}{\sigma^2} \tag{17}$$

where k represents the selected increment function, i.e., the partial variance function, which is bounded as follow $1 \leq k \leq 2^n - 1$ and σ^2 represent the total variance defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (f(\mathbf{x}) - \mu)^2 p(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \left[\left(f(^{\mathbf{c}}\mathbf{x}) + \sum_{i=1}^{n} \int_{c_{x_i}}^{x_i} \frac{\partial f(\xi)}{\partial \xi_i} d\xi_i \right.\right.$$
$$\left.\left. + \sum_{1 \leq i < j \leq n} \int_{c_{x_i}}^{x_i} \int_{c_{x_j}}^{x_j} \frac{\partial f(\xi)}{\partial \xi_i, \xi_j} d\xi_i d\xi_j + ... + \int_{c_{x_1}}^{x_1} ... \int_{c_{x_n}}^{x_n} \frac{\partial f(\xi)}{\partial \xi_1, ..., \xi_n} d\xi_1 ... d\xi_n \right) - \mu \right]^2 p(\mathbf{x}) dx_1 ... dx_n \quad (18)$$

Note that

$$\sigma^2 \neq \sum_{k=1}^{2^n - 1} \sigma_k^2 \quad (19)$$

Sensitivity indices are all non-negative and by using them, the functional structure and the rank of variables for given function can be estimated. In most nonlinear functions, the theoretical approach is hard to apply and therefore, the derivative equation is transformed to the numerical approach leading to the well-known cut-HDMR model.

Now, let us compare the DE to the well-known ANOVA decomposition. The basic concept behind HDMR is that many physical systems do not exhibit high-order cooperative behavior as it is proved by statistical evidence. The ANOVA decomposition was developed in order to quantify the influence of each variable or interaction of variables, but it does not show why the high-dimensional inputs in the various cases have null influence on the output. The DE clearly explains why this phenomenon is happening, i.e., in the most cases, the high order partial derivatives are smaller than the low order partial derivatives. Moreover, DE clearly shows why samples in the particular space have influence on the final output, while other samples do not. This cannot be deduced from the HDMR approach.

The integral form of the DE represents a variation of the cut-HDMR approach and represents the finite increment to the function of interest. The DE itself represents the infinitesimal increment to the function of interest and this is the main difference between the cut-HDMR and the DE. However, the cut-HDMR representation is a set of analytic equations (see Section 3) and does not allow one to deduce important conclusions discussed later in this section. The integral form of the DE shows in a clear way properties of the increment functions such as the zeroth value of the increment function if one of the integral limits is equal to the central point or the connection between derivatives and increment functions.

As discussed earlier, can be obtained several conclusions there from the integral form of DE and its integral variation. The first very important observation can be made from Eqs. (13) and (16). It can be seen that the absolute value of the integral part, e.g.,

$$\int_{c_{x_i}}^{x_i} \int_{c_{x_j}}^{x_j} \frac{\partial f(\xi)}{\partial \xi_i, \xi_j} d\xi_i d\xi_j$$

increases with the distance from the central point. At the same time, if the Gaussian distribution is assumed, the probability of occurrence, $p_{ij...k}$, decreases, i.e., the most samples are distributed around the central point, where interaction effects are negligible. Therefore, it can be concluded that tails of the output distribution are mainly given by higher order partial derivatives, i.e., if the output distribution has heavy tails and the input distributions are light tailed (most cases in real life), the interaction terms are very strong in the function of interest. This is also confirmed in a practical example in Section 6. The second important conclusion is the stochastic domain dependence. This aspect was already mentioned in Eq. (5) and it can be written as follows:

if

$$\frac{\partial f(x)}{\partial x_i} = 0 \quad \forall x \in R^n$$

then

$$\frac{\partial f(x)}{\partial x_j ... \partial x_k \partial x_i} = 0 \quad \forall x \in R^n$$

This statement can be extended to the increment functions in the following way:

$$\text{if}$$

$$dF_i(x) = 0 \quad \forall x \in R^n$$

$$\text{then}$$

$$dF_{ij\ldots k}(x) = 0 \quad \forall x \in R^n$$

The above assumption comes naturally from a basic integral calculus as an integration of 0 is always 0, i.e., integration of zeroth derivative is always 0. Moreover, the increment function will be either increasing, decreasing, or zero; it will never be a constant in the given domain. The second conclusion can be extended to the following statement. Each high order partial derivative contains information about the low order partial derivative and vice versa, i.e., $\partial f(x)/\partial x_1 \partial x_2$ contains information about $\partial f(x)/\partial x_1$ and about $\partial f(x)/\partial x_2$, while this information is not changed by integration. Therefore, it can be concluded: for the first case

$$\text{if}$$

$$\frac{\partial f(x)}{\partial x_i} \geq \frac{\partial f(x)}{\partial x_i \partial x_j} \quad \forall x \in R^n$$

$$\text{then}$$

$$dF_i(x) \geq dF_{ij}(x) \quad \forall x \in R^n$$

and for the second case

$$\text{if}$$

$$\frac{\partial f(x)}{\partial x_i} < \frac{\partial f(x)}{\partial x_i \partial x_j} \quad \forall x \in \left(R^n, \frac{\partial F(x)}{\partial x_i \partial x_j} \neq 0 \right)$$

$$\text{then}$$

$$dF_i(x) < dF_{ij}(x) \quad \forall x \in \left(R^n, \frac{\partial F(x)}{\partial x_i \partial x_j} \neq 0 \right)$$

Note that the second case was encountered only for rare problems and the first case is valid for the majority of problems. Therefore, for the first case, if the low order increment function converges using a certain number of samples, the higher order increment functions converge under less or the same number of samples for a given domain. This will lead to a simplification of the sampling space and dramatic reduction of necessary samples. Moreover, the sensitivity analysis has to hold the same condition. The same conclusion was experimentally confirmed in [31].

The last conclusion can be deduced from the independence of terms in Eq. (5). The independence of these terms allows one to use independent surrogate techniques and each of these techniques has a local maximum accuracy, which can be acquired by given surrogate technique. Therefore, if only a certain order of DE is selected, i.e., not all increment functions are included in final model, the maximum achievable accuracy is given by the influence of the neglected increment functions. In other words, adding higher orders of increment function will allow higher accuracy. This is well shown in work of [35].

3. NUMERICAL APPROACH—CUT-HDMR MODEL DERIVATION

The integral form of DE can be applied to real problems; however, the derivatives and integrals are not practical to compute. It is more convenient to transform the integral form into a set of equations which are easy to solve, i.e., transform the integral form to the analytic equation. The numerical approach is based on Eq. (10) and the well known First Fundamental Theorem of Calculus [37]. The theorem can be written in the following way:

$$f(b) - f(a) = \int_a^b B(x)dx \tag{20}$$

where $B(x)$ is a continuous function on the closed interval $[a,b]$. Using the Second Fundamental theorem of Calculus [37] and defining $B(x)$ in the following way:

$$B(x) = \frac{\partial f(x)}{\partial x}$$

Eq. (20) can be rewritten such that

$$f(b) - f(a) = \int_a^b \frac{\partial f(x)}{\partial x} dx \tag{21}$$

This simple formula can be applied to the integral form of DE [Eq. (10)] for all the first-order derivatives. This step allows one to cast off the derivative and the integral and replace them with a simple equation.

To closely explain the numerical application, let us consider a continuous function $f(x_1, x_2, x_3)$. The first term in the integral form of DE using Eq. (21) reads

$$dF_1(x_1) = \int_{c_{x_1}}^{x_1} \frac{\partial f(\xi_1, {}^c x_2, {}^c x_3)}{\partial \xi_1} d\xi_1 = f(x_1, {}^c x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) \tag{22}$$

where ${}^c x_i$ represents the position of the central point of the random variable x_i. The increment function, $dF_1(x)$, contains only one random variable and all other random variables are held constant at their central value. Increment functions for other variables (x_2, x_3) are created in a similar way.

Second-order increment functions are a bit more complex to handle. Let us consider the previous function and assume a point on a plane, e.g., $x_3 = {}^c x_3$. All integration parts of the integral DE, which are involving x_3, are zero and this assumption allows one to rewrite the integral form of DE as follows:

$$\begin{aligned} f(x_1, x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) = &\int_{c_{x_1}}^{x_1} \int_{c_{x_2}}^{x_2} \frac{\partial f(\xi_1, \xi_2, {}^c x_3)}{\partial \xi_1 \partial \xi_2} d\xi_1 d\xi_2 \\ &+ \int_{c_{x_1}}^{x_1} \frac{\partial f(\xi_1, {}^c x_2, {}^c x_3)}{\partial \xi_1} d\xi_1 + \int_{c_{x_2}}^{x_2} \frac{\partial f({}^c x_1, \xi_2, {}^c x_3)}{\partial \xi_2} d\xi_2 \end{aligned} \tag{23}$$

Clearly the one-dimensional integrals of Eq. (23) can be replaced by Eq. (22) and a similar equation for x_2. This leads to the following simplification:

$$\begin{aligned} f(x_1, x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) = &\int_{c_{x_1}}^{x_1} \int_{c_{x_2}}^{x_2} \frac{\partial f(\xi_1, \xi_2, {}^c x_3)}{\partial \xi_1 \partial \xi_2} d\xi_1 d\xi_2 \\ &+ f(x_1, {}^c x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) + f({}^c x_1, x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) \end{aligned} \tag{24}$$

which can be rewritten into the following form:

$$\begin{aligned} dF_{12}(x_1, x_2) = &\int_{c_{x_1}}^{x_1} \int_{c_{x_2}}^{x_2} \frac{\partial f(\xi_1, \xi_2, {}^c x_3)}{\partial \xi_1 \partial \xi_2} d\xi_1 d\xi_2 = f(x_1, x_2, {}^c x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) - f(x_1, {}^c x_2, {}^c x_3) \\ &+ f({}^c x_1, {}^c x_2, {}^c x_3) - f({}^c x_1, x_2, {}^c x_3) + f({}^c x_1, {}^c x_2, {}^c x_3) \end{aligned} \tag{25}$$

where the increment function, $dF_{12}(x_1, x_2)$, is treated as a function with only two variables and the rest is held constant. These steps allow one to cast off the double integral and replace it with a simple equation. The same approach can be applied to remaining parts of the integral form of DE. Equations (22) and (25) are the same as in the cut-HDMR approach; see [27, 30, 38]. From Eqs. (22) and (25) the following conclusion can be made:

$$\frac{\partial dF_{i\ldots k}(x)}{\partial x_i \ldots \partial x_k} = \frac{\partial f(x)}{\partial x_i \ldots \partial x_k} \tag{26}$$

The final step is to put all increment functions into the basic shape of the integral form of DE, which reads

$$\begin{aligned} f(x_1, x_2, x_3) - f({}^c x_1, {}^c x_2, {}^c x_3) = &dF_1(x_1) + dF_2(x_2) + dF_3(x_3) + dF_{12}(x_1, x_2) \\ &+ dF_{13}(x_1, x_3) + dF_{23}(x_2, x_3) + dF_{123}(x_1, x_2, x_3) \end{aligned} \tag{27}$$

Each increment function is then interpolated by an independent technique and Eq. (27) represents a sum of independent interpolation techniques. More importantly, the increment function can be sampled independently, which is a direct consequence of orthogonality of increment functions. Various techniques for interpolation of increment functions such as the Kriging surrogate model or SC approach can be used. Note that higher increment functions in Eq. (27) can be zero and, therefore, they can be easily neglected. This is a direct consequence of partial derivatives in different directions inside each increment function.

To compute the partial expected value and the partial variance, the established theoretical partial mean and variance are not suitable. Therefore, a numerical estimation of partial mean and variance follows

$$\mu_k = \frac{1}{z}\sum_{j=1}^{z} dF_k(\mathbf{x}_j) \tag{28}$$

$$\sigma_k^2 = \frac{1}{z-1}\sum_{j=1}^{z}(dF_k(\mathbf{x}_j) - \mu_k)^2 \tag{29}$$

where z represents the number of samples of the MC simulation applied on the surrogate model of the increment function, $dF_k(x)$. The expected value is established as a sum of partial expected values and reads

$$\mu = f({}^{\mathbf{c}}\mathbf{x}) + \sum_{k=1}^{2^n-1} \mu_k \tag{30}$$

where n represents a number of random variables. The total variance is defined as follows:

$$\sigma^2 = \frac{1}{z-1}\sum_{j=1}^{z}\left(\sum_{k=1}^{2^n-1} dF_k(\mathbf{x}_j) - \mu\right)^2 \tag{31}$$

where, again, n represents the number of random variables. One fundamental problem arises from the estimation of the total variance. The total variance requires all increment functions to be included and the number of samples required grows exponentially with increasing number of random variables. Practically, higher order increment functions are neglected if they have zero or very low influence on the final result and the functions (30) and (31) become

$$\mu \approx f({}^{c}x) + \sum_{k=1}^{s_D} \mu_k \tag{32}$$

$$\sigma^2 \approx \frac{1}{z-1}\sum_{j=1}^{z}\left(\sum_{k=1}^{s_D} dF_k(\mathbf{x}_j) - \mu\right)^2 \tag{33}$$

where s_D represents a selected number of increment functions, which is limited to $s_D \leq 2^n - 1$. The sensitivity indices are obtained via Eq. (17).

3.1 The Interpolation Process

The proposed method is based on the interpolation of the considered increment function on the given stochastic domain. However, the interpolation process is done in a new and more efficient way. As mentioned in previous sections, each increment function is handled as a separate problem and the final model comes from the sum of various interpolation techniques. However, in this work, only one type of interpolation technique is used for all increment functions and the selected interpolation technique is the multi-dimensional Lagrange interpolation. The interpolation technique is applied in the following way. Let us consider the general increment function $dF_k(x_i, ..., x_j)$, the interpolation model of the increment function is created using only samples from the given stochastic domain and the rest is held constant,

i.e., only samples $[^c x_1, ...,^c x_{i-1}, x_i, ..., x_j,^c x_{j+1}, ...,^c x_n]$ are considered. Other samples are completely neglected in the process of interpolation, i.e., they have null influence on the process. The interpolation model is created and stored. The statistical properties for the given increment function are obtained using the Monte Carlo approach applied directly to this surrogate model. The final model is created as a sum of these models [Eq. (27)] and overall statistical properties are obtained directly from the final model. The sampling strategy for given interpolation techniques is discussed in the next section.

4. SAMPLING STRATEGY

The sensitivity analysis using the derivative equation approach is based on an interpolation approach and requires a special sampling strategy. This problem comes from the increment function as each increment function requires samples from all involved sub-spaces. Therefore, random sampling strategies such as LHS or UD cannot be used and samples cannot be completely randomly spread around the stochastic domain.

The first step in the sampling strategy is to sample the first-order increment functions, i.e., $dF_i(x_i)$. To illustrate the sampling process, consider a function with three random stochastic variables. The first order increment functions [function (22)] require samples only on the abscissas, i.e., x_1, x_2, x_3. The sampling nodes for the 1D increment functions are selected as Chebyshev-Gauss nodes [39] in the case of normal distribution or Clenshaw-Curtis nodes [40] in the case of other distribution. The example of samples around abscissas is shown in Fig. 1. The derivative method requires a sample at the central position of the stochastic space, which is selected at the mean value for a given distribution of a random variable, i.e., $mean(x_i) = {}^c x_i$. A 1D interpolation model can be constructed for each variable using the central point and samples on the corresponding abscissa. When the first-order increment functions are established, the MC simulation can be applied on each surrogate model. Therefore, partial expected values and variances can be computed for each increment function. The total variance for the first-order can be computed via Eq. (33) and sensitivity indices via Eq. (17). Note that if the function is known to be additive, the sampling process is stopped.

Let us now focus on the higher derivative case. The higher increment functions require increment functions from all lower stochastic spaces, e.g., dF_{123} requires dF_1, dF_2, dF_3, dF_{12}, dF_{13}, and dF_{23}. Each one of these increment functions requires samples from the relative stochastic domain, e.g., dF_1 requires samples on abscissa x_1, dF_{12} requires samples on plane x_1, x_2, etc... In other words, each increment function requires samples corresponding to all lower stochastic domains. Moreover, geometrically speaking, the samples have to lie on the cross section of the given sub-domain (see Fig. 2 on left). This is later on called cross-section condition.

In the same way, the surrogate model can be created from samples on a plane, a volume, or a hyper-volume in the case of a high dimensional space. When the surrogate model is established, the MC simulation is applied and partial expected values and variances are obtained.

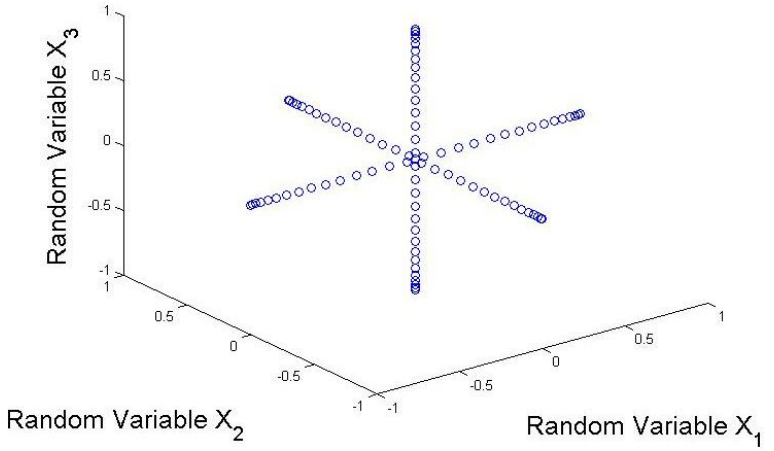

FIG. 1: Example of samples on the abscissas for x_1, x_2, x_3—a symmetrical Guassian distribution.

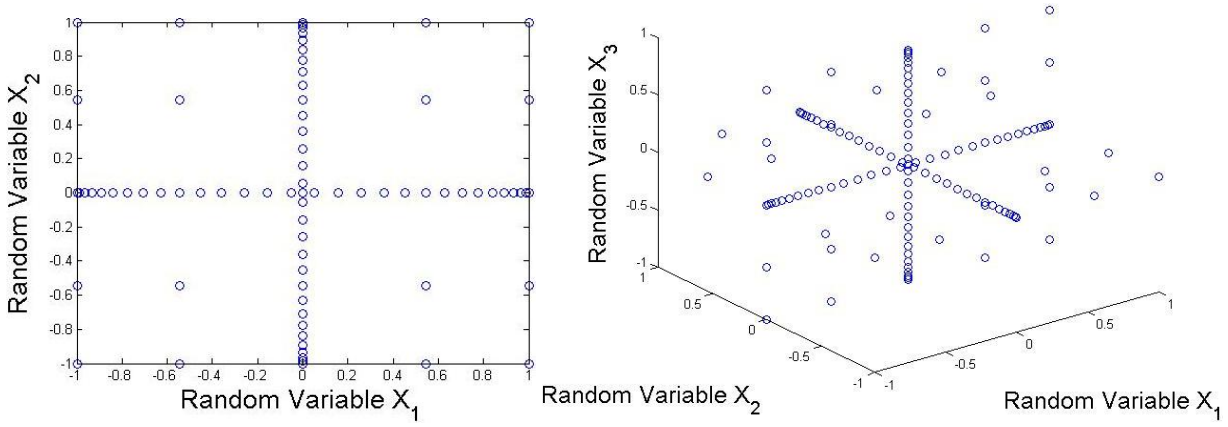

FIG. 2: Example of the cross section condition in the 2D and the 3D stochastic space.

4.1 Design Technique of Non-Complete Tensor Product and Selection of Higher Order Derivatives

For higher order derivatives, it is not necessary to have the same number of samples as in lower order derivatives as shown in Fig. 2. From conclusions given in the Section 2, it is obvious that higher order interactions can play a non-significant role and they can be interpolated with a low number of samples (see Borehole example). Also, note that the higher order increment function [Eq. (25)] can be zero and can be neglected for the interpolation process. Therefore, the process of sampling in higher dimensional domains can be divided into two steps: The first step is the selection of non-zero increment functions, which influence the final model, and the second step is the sampling in higher stochastic domains.

The selection of increment functions is based on a cross-validation approach, i.e., few samples are randomly sampled on the borders of the stochastic domain and the value of the increment function is estimated at each sample point. Recalling the second conclusion given in Section 2, the increment functions with null derivative can be neglected without affecting the final result. Therefore, if the value of the increment functions corresponding to all sampled points is lower than ϵ, the increment function is neglected. The value of ϵ is a numerical threshold and, in this work, is set to $<10^{-14}$. The higher order increment functions including the neglected increment function are neglected too. This leads to a reduction of samples necessary to estimate the high-order increment functions and it is a specially effective in any case of a high dimensional stochastic space.

In order to better explain the process, let us consider an increment function $dF_{12}(x_1, x_2)$. The test sample would be positioned at $x_1 = 1, x_2 = 1$ (Fig. 2) and the value of increment function, $dF_{12}(x_1, x_2)$, at that sample is estimated. The testing samples can be positioned anywhere in the stochastic domain, but it is proposed to position samples in the corners of the selected stochastic domain because the interaction terms, represented by higher order increment functions, get stronger, when the sample is positioned further away from abscissas. In this work, only one sample to estimate the influence of increment function is used. Note that all samples have to be governed by the cross-section condition.

Unfortunately, the above approach has a drawback in some particular applications. For example, the function, $e^{(x_1 x_2)}$, with the central point $x_1 = 0, x_2 = 0$, is an example of this problem. In this case, the increment function along x_1 or x_2 will be zero and the above statement fails. However, the conditions required for the algorithm to neglect an influential increment function are extremely narrow such as an exact central point and a specific function. Moreover, both conditions have to be satisfied and therefore, the probability of false neglection is extremely low.

The second step of the sampling process concerns the sampling strategy in higher dimensions. The choice of the sampling strategy is an aspect as problematic as it is fundamental in high dimensional interpolation. In this work, an engineering approach for the estimation of samples in higher dimensions is used. First, let us again recall the second conclusion from Section 2. The high-order partial derivative contains information about the low-order partial derivative and vice versa. Therefore, instead of using the pure tensor product, only a portion of samples can be used

to accurately describe the higher domain. Therefore, the basic equation for establishing the number of samples used in higher dimensions reads

$$N = \frac{\left(n^{-(1/k)}O\right)^{-k}}{2} \tag{34}$$

where n is the number of samples in 1D quadrature for the given direction, O is the selected order of derivation, and k is the coefficient of growth (COG). The COG is pre-selected by the user at the beginning of the interpolation process and each direction can have different COG. The final number is rounded to integer. Figure 3 shows how the number of samples, N, used for any given order of the increment function varies with the order of the increment function, O. It can be seen that for the first-order increment functions, the number of samples is 15, for the second order increment functions, the number of samples is 6, etc. Figure 3 was created using $k = 1.5$ and $n = 30$. The process of application on the random variable follows. The given stochastic domain, i.e. the random variable space, is separated into two intervals, where the central point is the separating point. The number of samples, N, is applied for both intervals. This step is taken because of the unknown nature of an input distribution and it works well for non-standard distributions such as Gumbel or Weibull. The positioning of new samples follows. The first sample is positioned on the boundary of a given domain. Then each following sample is positioned into the middle of a larger free space of given interval, i.e., the sampled points create intervals and the new point is positioned into the middle of the largest interval. Note that the new point is forced to hold the cross-section condition. The process is repeated for the other side of the stochastic domain. The process of selection and sampling of higher increment functions is written in Algorithm 1.

It can be understood that this approach is closely related to the Smoylak sparse grid approach. However, the Smoylak sparse grid grows too fast in higher dimensions and it was found that this growth is not necessary. Also note that increment functions are independent from each other and therefore, the number of samples in each increment function can differ. For each increment function, a different surrogate model can be used, but in this work, the MLI technique was used for each increment function.

The selected algorithm is based on an empirical experience and reflects a pure engineering approach. The user starts with a high value of COG (around 2) and observes the change of the final PDF, while decreasing the COG number. The tuning process stops when the output PDF is fully converged for a given order of the increment function. Then the maximum order of the increment function can be increased or the problem can be claimed as a fully converged.

The proposed approach has few very important properties. The first one: It is easy to implement and it can be easily used in practical applications. The user does not have to think about higher dimensions or run time-consuming algorithms for a high dimensional sampling strategy. Simple modification of COG can decrease or increase the number of samples for a given order of increment function. The second advantage is that samples are nested and the change of the COG coefficient leads to new samples added to the stochastic domain, while preserving the old ones. Therefore, the analysis does not have to be repeated. This simplifies the convergence process as simply changing COG and

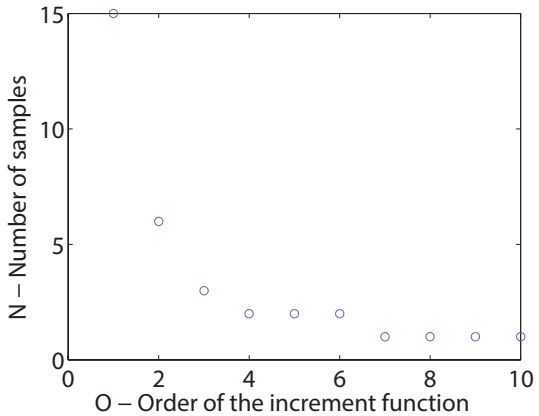

FIG. 3: Number of samples for the higher order increment functions.

Algorithm 1: Sampling strategy for the higher order increment functions

for $k = 1$ *to Number of increment functions* **do**
- 1. Randomly sample given domain on its border;
- 2. Compute the value of $dF_k(x)$ for all sampled points;
- **if** *For all values of* $dF_k(x) < \epsilon$ **then**
  - 1. Delete the increment function, $dF_k(x)$, from the final model;
  - 2. Delete all increment functions from the final model, which includes current increment function;
- **else**
  - **for** $j = 1$ *to Number of variables in current increment function* **do**
    - 1. Compute Eq. (34) and round it;
    - 2. Separate interval x_j into two intervals by the mean value of x_j;
    - 3. Sample the stochastic domain (hold the cross section condition) on both intervals;
  - **end**
  - 1. Create a tensor product from created samples;
  - 2. Compute values of $dF_k(x)$ in each of the samples;
  - 3. Create a surrogate model from obtained values;
  - 4. Include the surrogate model in the final model, see Section 3;
- **end**

end

observing the final result can bring the desired result. More importantly, each dimension is treated independently. This allows one to set different number of samples on each dimension. This is very useful as the most real-life cases converge faster under different type of quadrature.

5. VISUALIZATION OF HIGHER ORDER SENSITIVITY AND SENSITIVITY ESTIMATION

The process of sensitivity visualization is easy and straightforward. At first, let us recall the first important property of DE, the independence of the increment function. This allows one to visualize each increment function separately and observe its influence on the model. For visualization purposes, histograms are used.

The first-order increment function represents the sensitivity of one particular variable as it can be seen in Eq. (22). Using the Monte Carlo sampling, the statistical properties of this particular variable can be observed and its influence can be estimated. From the obtained samples, the partial mean value and the partial variance can be estimated. The partial mean value represents the influence of the input distribution on the final mean value of the output distribution and the partial variance represents the variability for the given distribution.

Let us now assume a higher order increment function. Note that the lower stochastic domains are null for given domain, e.g., the increment function dF_{12} have zeroth value at abscissas x_1 and x_2 and therefore, this function represents a pure interaction effect, i.e., how combination of variables influence the final model. In the engineering world, this represents for example chemical reactions in hypersonic flows.

In order to explain the above description on an example, let us consider the following function:

$$F(x_1, x_2, x_3) = x_1^3 + 0.1x_2^2 - e^{x_1 x_3} \tag{35}$$

where x_i is a random variable with a uniform distribution and defined on interval [0, 2]. On this function, the properties of DE can be easily demonstrated and the visualization of high-order increment functions can be showed. A histogram of each increment function is shown in Fig. 4 and the histogram of the function F [Eq. (35)] is shown in Fig. 5. From these histograms, we can observe the influence of each random variable on the output. It can be seen that the increment function, dF_1, is moving the mean value to higher values and slightly influencing the variance. The second increment function, dF_2, is influencing the mean value and the variance in a negligible way. However, it can be easily understood from the histogram that the function is monotonic in a given range. From the histogram of the increment function, dF_3,

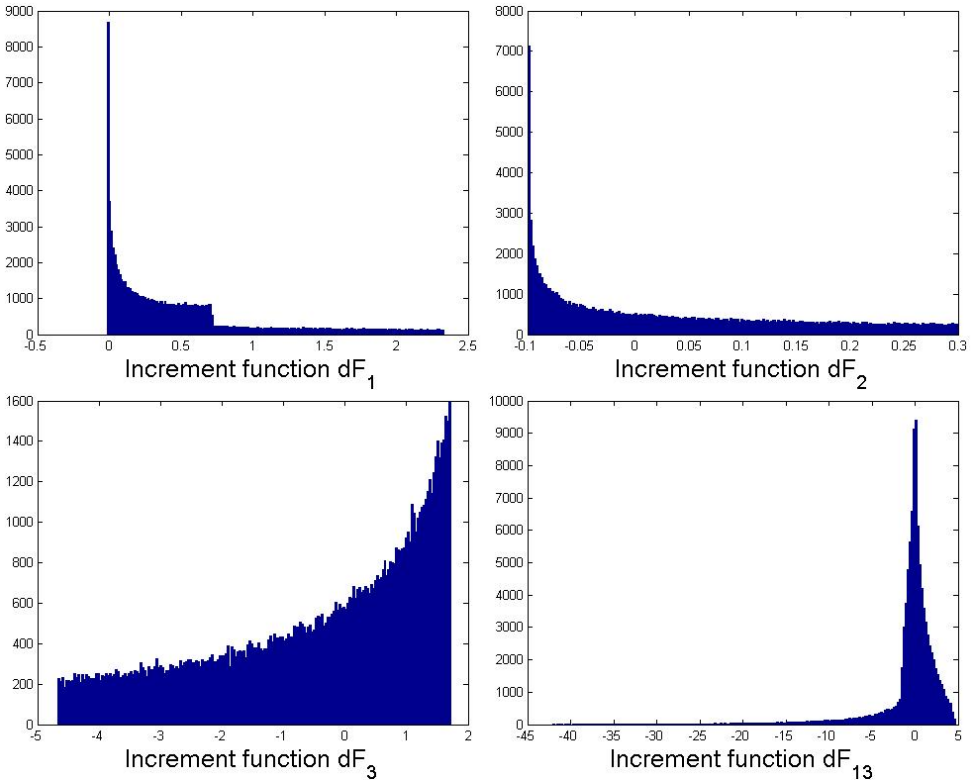

FIG. 4: Visualization of histograms for each increment function.

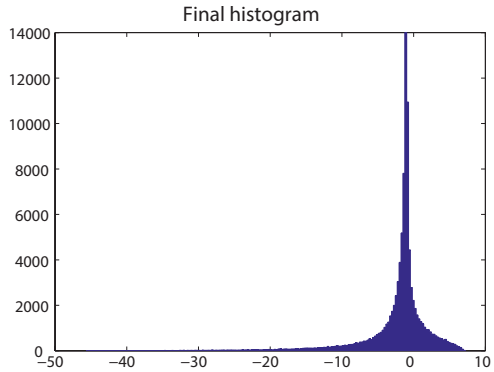

FIG. 5: Visualization of the final histogram.

it can be observed that the function is monotonic and with slight influence on the final PDF. However, if the variable x_3 is coupled with the variable x_1, the increment function, dF_{13}, plays the major role on the shape of the final PDF and it is responsible for tails of the final distribution. In physical meaning, the increment function, dF_{13} represents an interaction between inputs (x_1, x_3) such as, previously mentioned, chemical reactions. Note that all other increment functions are zero and therefore, they are not shown.

The integral form of the DE equation allows one to visualize the behavior of a function in the stochastic domain. Note that using DE is different than just simple sampling. The derivative process separates variables of interest from non-interesting variables and following integration reshapes the function of interest, i.e., creates a function of increment.

6. ANALYTIC CASE STUDIES

For the proposed method, a Borehole model [11, 13] is selected, which is a well-known test case for sensitivity analysis methods. It consist of 8D physical problem and the function reads

$$F(x) = \frac{2\pi T_u(H_u - H_l)}{\ln(r/r_w)[1 + (2LT_u/\ln(r/r_w)r_w^2K_w) + (T_u/T_l)]} \tag{36}$$

where r_w is the radius of borehole (m), r is the radius of influence (m), T_u is the transmissivity of upper aquifer (m2/yr), H_u is the potentiometric head of upper aquifer (m), T_l is the transmissivity of lower aquifer (m2/yr), H_l is the potentiometric head of lower aquifer (m), L is the length of borehole (m), and K_w the hydraulic conductivity of borehole (m/yr). Distributions associated with each random variable are summarized in Table 1. Results for important increment functions are summarized in Table 2. Note that the sensitivity of other increment functions were less than 10^{-6} and therefore, they are not listed here. Histograms of important increment functions are listed in Figs. 6 and 7.

TABLE 1: Distributions for the 8D Borehole model

Random Variable	Distribution type	Mean	Standard deviation
$r_w(x_1)$	Normal	0.10	0.0161812
$r(x_2)$	Log-normal	7.71	1.0056
—	—	Min	Max
$T_u(x_3)$	Uniform	63,070	115,600
$H_u(x_4)$	Uniform	990	1,110
$T_l(x_5)$	Uniform	63.1	11.6
$H_l(x_6)$	Uniform	700	820
$L(x_7)$	Uniform	1,120	1,680
$K_w(x_8)$	Uniform	9,855	12,045

TABLE 2: Distributions for the Borehole model

Increment function	Partial mean	Partial variance	Mean sensitivity	Variance sensitivity
1	1.8747	540.4055	0.6598	0.6741
4	–0.0191	72.3027	–0.0067	0.0901
6	–0.0024	72.6236	–0.0008	0.0905
7	0.9942	71.2503	0.3499	0.0888
8	0.0128	16.8106	0.0045	0.0209
78	0.0024	0.2382	0.0008	0.0002
68	–0.0007	0.2409	–0.0002	0.0209
67	–0.0052	1.0413	–0.0018	0.0012
48	–0.0009	0.2390	–0.0003	0.0002
47	–0.0039	1.0315	–0.0013	0.0012
18	0.0098	1.7948	0.0034	0.0022
17	0.0203	7.6782	0.0071	0.0095
16	–0.0215	7.8591	–0.0075	0.0098
14	–0.0133	7.8159	–0.0047	0.0097
178	0.0003	0.0253	0.0001	3.1682e-05
168	–5.9561e-06	0.0258	–2.0964e-06	3.2260e-05
167	–9.8284e-05	0.1144	–3.4594e-05	0.0001
148	–0.0010	0.0254	–0.0003	3.1803e-05
147	–0.0011	0.1136	–0.0004	0.0001

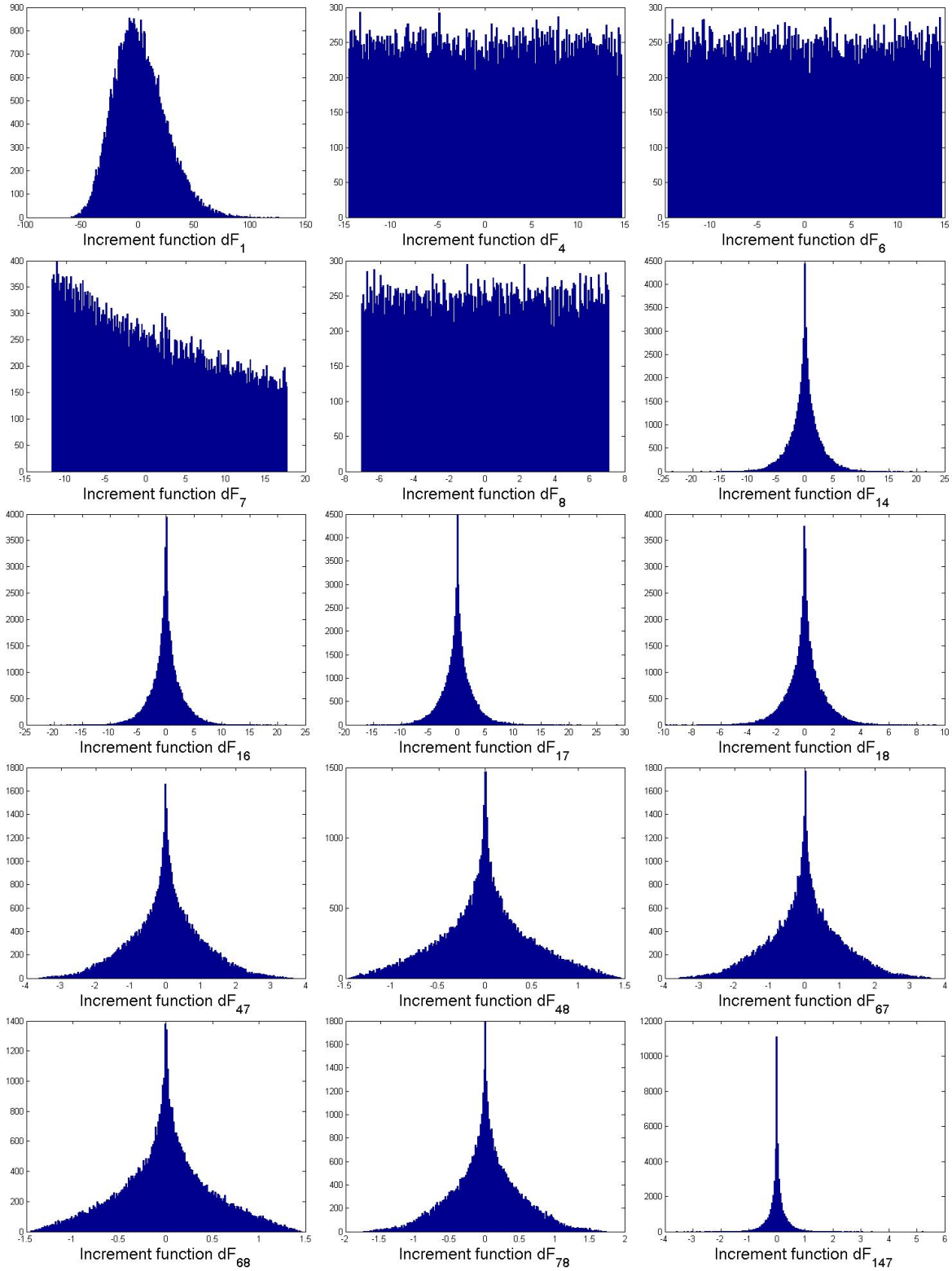

FIG. 6: Visualization of histograms for each increment function.

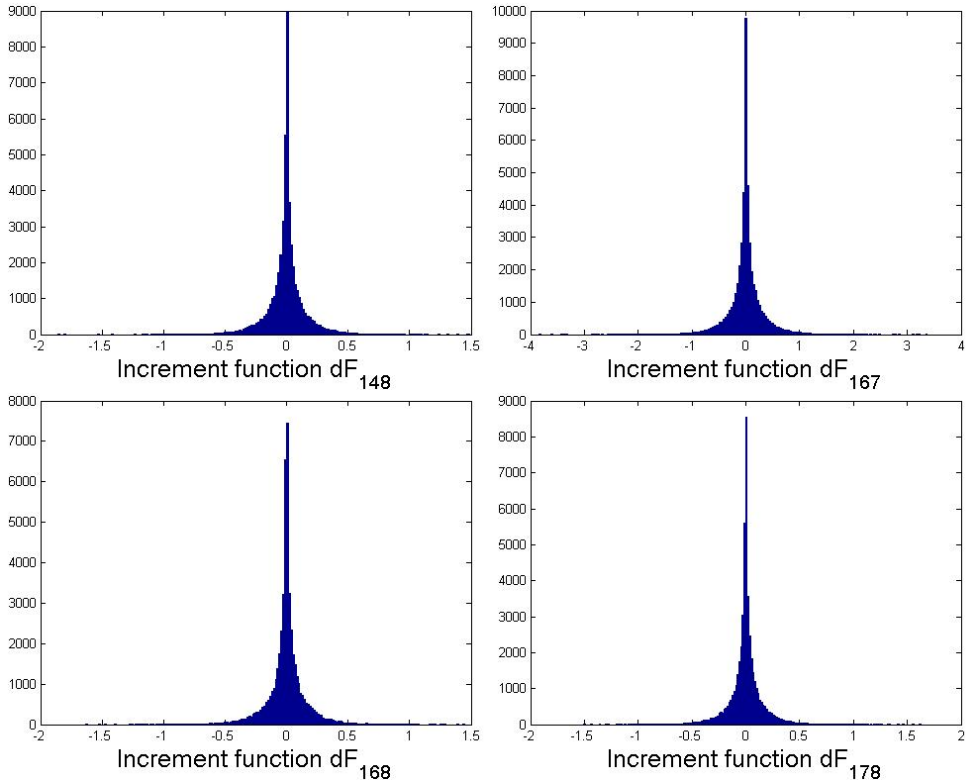

FIG. 7: Visualization of histograms for each increment function.

Results for the interpolation part are summarized in Table 3 and quadratures used for each variable are summarized in Table 4. The interpolation part was compared to MC simulation and results of MC simulation are given in Table 5. Histograms of final uncertainty obtained by DUQ and MC are shown in Fig. 8.

TABLE 3: DUQ approach for the 8D Borehole model

Case	Function calls	Mean	Standard deviation	Relative error of mean	Relative error of S. D.	Max. order	COG
1	35	73.9792	27.6947	2.0733e-04	0.0357	1	—
2	147	73.9657	28.7215	2.5849e-05	8.5043e-05	2	1.5
3	595	73.9639	28.7183	1.7129e-06	2.4218e-05	3	1

TABLE 4: DUQ approach for the 8D Borehole model, CC(x)—Clenshaw-Curtis nodes (number of nodes), Ch(x)—Gauss-Chebyshev nodes (number of nodes)

Case	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	Ch(5)	Ch(5)	CC(5)	CC(5)	CC(5)	CC(5)	CC(7)	CC(5)
2	Ch(5)	Ch(5)	CC(5)	CC(5)	CC(5)	CC(5)	CC(7)	CC(5)
3	Ch(5)	Ch(5)	CC(5)	CC(5)	CC(5)	CC(5)	CC(7)	CC(5)

TABLE 5: Monte Carlo simulation for the 8D Borehole function

Function calls	Mean	Standard deviation
100000	73.9639	28.7191

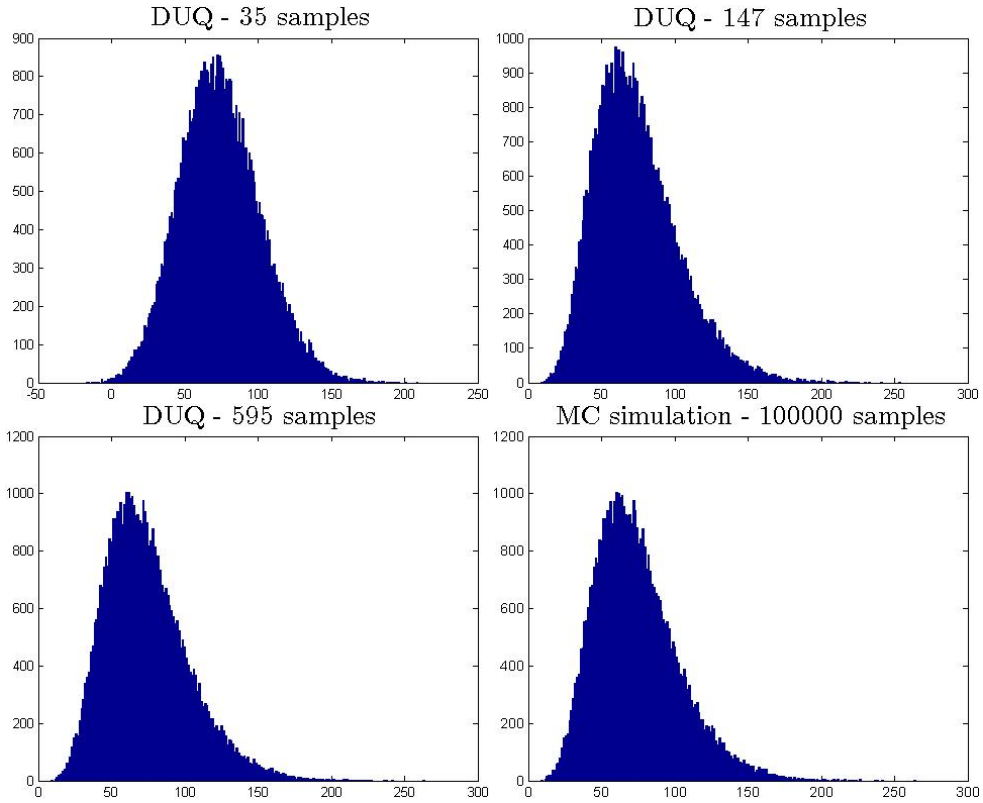

FIG. 8: Histograms of final uncertainty in the Borehole model.

6.1 Discussion about Borehole Model

The sensitivity analysis via DUQ method estimates the same sensitivity as was obtained in [41]. Moreover, a visualization of each variable's influence is provided. In Fig. 6, it can be seen that the far largest influence is given by variable x_1 and the shape of final PDF (Fig. 8) is mainly given by this variable. The behavior of variable x_7 is interesting: this variable does not influence the variance of the final uncertainty, but it has a high influence on the final expected value (around 35%). This can be seen in Fig. 6, where the increment is biased to the right. The sensitivity analysis performed in this paper is in agreement with [41]. The sensitivity analysis was done using the interpolation process, which was created with 595 samples and 1st, 2nd, and 3rd order increment functions.

The sensitivity analysis is based on the quality of the interpolation process. It can be seen in Fig. 8 that for 35 samples, the PDF is relatively well captured and for 147 samples, the shape of the PDF is very well captured. For 595 samples, the right tail end of the PDF is captured and only small differences can be recognized. Higher order (>3) increment functions are neglected as adding higher orders would create very small changes and even if it gives a higher accuracy, it would require a much higher number of samples.

7. DISCUSSION ABOUT SENSITIVITY ANALYSIS, INTERPOLATION PROCESS, AND DERIVATIVE EQUATION

The new approach to the interpolation brings a new way to interpolate a high dimensional problems. Followed by the sensitivity analysis, the approach brings a completely new insight into high dimensional uncertainty.

The independence of increment functions on the lower order stochastic space, i.e., the increment function is zero in lower order stochastic space, allows one to visualize interaction terms alone. This is very useful to estimate the PDF of interaction terms. Moreover, the non-standard influence of each variable, such as variables x_1 and x_7 in the case

of the Borehole model, can be estimated and further research can focus on these variables. The influence of the input distribution for each variable can be established. This can be done by application of a MC sampling with different input distributions to the created surrogate model. Therefore, the influence of a different input distribution can be established with no additional calling of the expensive code. Note that the central point has to be selected in the mean of given random variables.

The interpolation process is slightly more complex to understand. In this work, each increment function is handled separately, i.e., each increment function is interpolated with a possible different surrogate technique. It was found that it is easier and more accurate to interpolate each problem separately than trying to interpolate all together. At the same time, each interpolation technique represents a new challenge and errors in one interpolation technique will propagate to the whole model.

In terms of convergence, the process is separated into two steps. The first step leads to a local convergence, where only convergence of a surrogate model on a given increment function is observed. It is performed by adding samples to a given domain and observing the behavior of the interpolation technique. The second part leads to a global convergence, i.e., how each increment function influences the final model. It is performed by adding fully converged increment functions to the final model. In other works [30, 35], it is suggested to stop at a prescribed level and add samples in an isotropic way. In this work, the non-isotropic approach is adopted and it was found that it leads to a faster and more accurate interpolation. Moreover, the two steps convergence process assures the simplicity of a high dimensional interpolation.

In the Borehole example, it was found that the first-order increment functions converge under a low number of samples, e.g., 5 or 7, which is the same conclusion given in [31]. The higher order increment function requires even less samples for a dimension, i.e., one sample in each corner of a given stochastic domain was sufficient to let the surrogate model converge to the true function. This can be done due to the independence of the increment functions. Moreover, from the analytic approach, the higher order increment functions have to converge under the same or lower number of samples for the first case. If the second case is valid, observing only the convergence of the low-order increment functions is not enough. This case can be recognized by a large change in the final PDF, when the second-order fully converged increment functions are added to the final model. It was found that observing convergence of each increment function separately is enough to make a full model completely converged. This was confirmed in the Borehole example. Note that each increment function will be zero at the central point and all its sub-domains. This simplifies the interpolation process, but this observation was not included into the interpolation process, see Eq. (23). Moreover, it was observed that the convergence of the right tail in Fig. 8 is given mainly by the higher order increment functions, i.e., if the function has high interaction terms, the final distribution have long tails. Naturally, this comes from the DE equation as it was shown through the analytic approach.

8. CONCLUSION AND FUTURE WORK

In this work, a new method for the estimation of uncertainty propagation is presented. A new way of obtaining the cut-HDMR model via the derivative equation is presented. From this equation important conclusions can be made, which lead to a significant reduction of samples for a high dimensional space. However, the most important aspect of the DE is that it shows how the information propagates from the low-order stochastic space to the higher order stochastic space. This knowledge could be fundamental for future research on high dimensional interpolation. The DE shows in understandable way the independence of each stochastic domain; this can be important for research in various fields such as optimization, multi-fidelity interpolation, and high dimensional interpolation. Moreover, from the obtained conclusions, it is shown that the tails of the output distribution are mainly given by the high-order interaction terms. This is confirmed in the considered example. Using the conclusions obtained from the DE, a high dimensional efficient interpolation technique is built. For each part of the cut-HDMR, an independent interpolation technique is used and the final model is constructed as a sum of these models. The parts with null influence on the final model are neglected and each part of the cut-HDMR is constructed using different number of samples. It was shown that the interpolation of a low-order increment functions with various number of samples leads to a satisfactory accuracy and a dramatic reduction of necessary samples for a high dimensional interpolation. The interpolation accuracy is shown on an applied example, which is well known for being hard to interpolate, and compared to direct MC sampling.

The proposed approach requires a special sampling strategy, which is introduced in this manuscript. The samples in higher domains are distributed in a similar way as it is used in the Smoylak sparse grid approach; however, in this work, an empirical approach is used. The proposed sampling approach is fast, easy to use, and allows one to sample efficiently the high dimensional space. Moreover, the provided algorithm is nested, which has a significant advantage in real engineering applications. It means that the samples have not to be repeated and a fast convergence can be obtained.

A new way to sensitivity analysis is presented too. The analysis is based on the Sobol approach and, due to the nature of the proposed method, a visualization of each variable and its interaction via histograms can be done. The use of visualization on a simple example was shown and results obtained for the Borehole model were in agreement with other references. Moreover, on the Borehole model, it was described in detail how each variable influences the final model. The visualization of sensitivity of each variable helps the user to better understand the influence of the considered variable.

In future work, the focus will be aimed to interpolation techniques including polynomial chaos, neural network or Kriging model, and their various combinations. Since, the cut-HDMR approach represents an interesting opportunity for a multi-fidelity modeling, the multi-fidelity implications of the cut-HDMR approach will be investigated. The sampling strategy represents another interesting aspect to consider to improve the efficiency of the interpolation process.

REFERENCES

1. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S., *Global Sensitivity Analysis: The Primer*, John Wiley and Sons, New York, 2008.
2. Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M., *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, John Wiley and Sons, New York, 2004.
3. Saltell, A. and Bolado, R., An alternative way to compute fourier amplitude sensitivity test, *Comput. Stat. Data Anal.*, 26:445–460, 1998.
4. Fang, K. T., Li, R., and Sudjianto, A., *Design and Modeling for Computer Experiments*, Chapman and Hall/CRC Press, New York, 2006.
5. Kleijnen, J. P. C. and Helton, J. C., Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques, *Reliability Eng. System Safety*, 65:147–185, 1999.
6. Kleijnen, J. P. C. and Helton, J. C., Statistical analyses of scatterplots to identify important factors in large-scale simulations, 2: Robustness of techniques, *Reliability Eng. System Safety*, 65:187–197, 1999.
7. Morris, M. D., Factorial sampling plans for preliminary computational experiments, *Technometrics*, 33(2):161–174, 1991.
8. Weisstein, E. W., Web diagram, http://mathworld.wolfram.com/WebDiagram.html, 2014.
9. Hooker, G., Discovering additive structure in black box functions, In *KDD 04 Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ACM New York, NY, USA, pp. 575–580, 2004.
10. Eldred, M. and Burkardt, J., Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification, In *47th AIAA Aerospace Sciences Meeting including The New Horizons Forum and Aerospace Exposition*, American Institute of Aeronautics and Astronautics, Orlando, FL, 2009.
11. Eldred, M., Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design, in *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conf.*, Palm Springs, CA, 2009.
12. Wiener, N., The homogeneous chaos, *Am. J. Math.*, 60(4):897–936, 1938.
13. Togawa, K., Benigni, A., and Monti, A., Advantages and challenges of non intrusive polynomial chaos theory, in *Proc. of the 2011 Grand Challenges on Modeling and Simulation Conf.*, Crosbie, R. (Ed.), Society for Modeling and Simulation International, Vista, CA, pp. 30–35, 2011.
14. Cheng, H. and Sandu, A., Collocation least-squares polynomial chaos method, in *SpringSim'10, Proc. of the 2010 Spring Simulation Multiconference*, Society for Computer Simulation International, San Diego, CA, 2010.

15. Hosder, S., Walters, R., and Perez, R., A non-intrusive polynomial chaos method for uncertainty propagation in cfd simulations, in *44th AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, 2006.
16. Branicki, M. and Majda, A., Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities, *Commun. Math. Sci.*, 11:55–103, 2013.
17. Eldred, M. S., Webster, C. G., and Constantine, P. G., Evaluation of non-intrusive approaches for Wiener-Askey generalized polynomial chaos, In *The 49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conf.*, AIAA-2008-1892, Schaumburg, IL, 2008.
18. Lee, J. and Kwon, J. H., On the use of kriging in the interpolation of fluid-structure interaction analysis, *Jpn. Soc. Comput. Fluid Dyn.*, 16:294–299, 2008.
19. Forrester, A. I. J., Sobester, A., and Keane, A. J., *Engineering Design via Surrogate Modelling*, John Wiley and Sons, New York, 2008.
20. Sudret, B., Global sensitivity analysis using polynomial chaos expansions, *Reliability Eng. System Safety*, 93(7):964–979, 2008.
21. Bellman, R. E., *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, 1961.
22. Chen, R.-B., Hsieh, D.-N., Hung, Y., and Wang, W., Optimizing latin hypercube designs by particle swarm, *Stat. Comput.*, 23:663–676, 2013.
23. Hosder, S., Walters, R. W., and Balch, M., Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables, in *48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conf.*, Honolulu, HI, 2007.
24. Pfluger, D., *Spatially Adaptive Sparse Grids for High-Dimensional Problem*, Verlag Dr. Hut, Munich, 2010.
25. Gerstner, T. and Griebel, M., Numerical integration using sparse grids, *Numer. Algorithms*, 18:209–232, 1998.
26. Barthelmann, V., Novak, E., and Ritter, K., High dimensional polynomial interpolation on sparse grids, *Adv. Comput. Math.*, 12:273–288, 2000.
27. Rabitz, H., Alis, O. F., Shorter, J., and Shim, K., Efficient input-output model representations, *J. Phys. Chem.*, 117:11–20, 1999.
28. Shorter, J. A., Ip, P. C., and Rabitz, H. A., An efficient chemical kinetics solver using high dimensional model representation, *J. Phys. Chem.*, 103:7192–7198, 1999.
29. Li, G., Wang, S.-W., Rabitz, H., Wang, S., and Jaffe, P. R., Global uncertainty assessments by high dimensional model representations (HDMR), *Chem. Eng. Sci.*, 57(21):4445–4460, 2002.
30. Ma, X., *An Efficient Computational Framework for Uncertainty Quantification in Multiscale Systems*, PhD Thesis, Cornell University, 2011.
31. Tang, K., Congedo, P. M., and Abgral, R., Sensitivity analysis using anchored ANOVA expansion and high order moments computation, *Int. J. Num. Methods Eng.*, 102:1554–1584, 2015.
32. Li, G., Artamonov, M., Rabitz, H., Wang, S. W., Georgopoulos, P. G., and Demiralp, M., High-dimensional model representations generated from low order TERMSLP-RS-HDMR, *J. Comput. Chem.*, 24(5):647–656, 2003.
33. Tunga, M. A., An approximation method to model multivariate interpolation problems: Indexing HDMR, *Math. Comput. Model.*, 53:1970–1982, 2011.
34. Tunga, M. A. and Demiralp, M., Hybrid high dimensional model representation (HHDMR) on the partitioned data, *J. Comput. Appl. Math.*, 185:107–132, 2006.
35. Zhang, Z., Choi, M., and Karniadakis, G. E., Error estimates for the ANOVA method with polynomial chaos interpolation: Tensor product functions, *SIAM J. Sci. Comput.*, 34(2):1165–1186, 2012.
36. Sobol, I. M., Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Math. Comput. Simul.*, 55:271–280, 2001.
37. Weisstein, E. W., Fundamental theorems of calculus, http://mathworld.wolfram.com/FundamentalTheoremsofCalculus.html, 1999.
38. Li, G., Wang, S.-W., and Rabitz, H., High dimensional model representation (HDMR): Concepts and applications, *J. Phys. Chem.*, 117:11–20, 1999.

39. Gil, A., Segura, J., and Temme, N. M., *Numerical Methods for Special Functions*, Society for Industrial Mathematics, 2007.
40. Clenshaw, C. W. and Curtis, A. R., A method for numerical integration on an automatic computer, *Numer. Math.*, 2:197–205, 1960.
41. Joseph, V. R., Hung, Y., and Sudjianto, A., Blind Kriging: A new method for developing metamodels, *ASME J. Mech. Des.*, 130:031102, 2008.

APPENDIX

Let us assume a function $f(x_1, x_2)$ with two random variables. The variance is computed in the following way:

$$\sigma^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\left(\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1 + \int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2 + \int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\right) - \mu\right]^2 p(x_1,x_2)dx_1dx_2 \quad \text{(A.1)}$$

Now, let us closely look at the inner part of Eq. (A.1). The inner part can be expanded in the following way:

$$\begin{aligned}
&\left[\left(\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1 + \int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2 + \int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\right) - \mu\right]^2 \\
&= \left(\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\right)^2 + 2\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2 + \left(\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\right)^2 \\
&+ 2\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2 + 2\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2 \\
&+ \left(\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2\right)^2 - 2\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\mu - 2\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\mu - 2\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2\mu + \mu^2
\end{aligned} \quad \text{(A.2)}$$

If the same approach as in Section 2 is followed, the inner part of Eq. (A.1) is expanded in the following way

$$\begin{aligned}
&\left[\left(\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1 - \mu_1\right)^2 + \left(\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2 - \mu_2\right)^2 + \left(\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\right) - \mu_{12}\right]^2 \\
&= \left(\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\right)^2 + \left(\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\right)^2 + \left(\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2\right)^2 \\
&- 2\int_{c_{x_1}}^{x_1}\frac{\partial f(\xi)}{\partial \xi_1}d\xi_1\mu_1 - 2\int_{c_{x_1}}^{x_1}\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_1,\xi_2}d\xi_1 d\xi_2\mu_{12} - 2\int_{c_{x_2}}^{x_2}\frac{\partial f(\xi)}{\partial \xi_2}d\xi_2\mu_2 + \mu_1^2 + \mu_2^2 + \mu_{12}^2
\end{aligned} \quad \text{(A.3)}$$

Equation (A.2) clearly differs from Eq. (A.3). Therefore, a sum of partial variances is not the total variance as it is mentioned in Section 2.