



**HAL**  
open science

# Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu, Dominique Lesselier, Bruno B Sudret, Joe Wiart

► **To cite this version:**

Zicheng Liu, Dominique Lesselier, Bruno B Sudret, Joe Wiart. Surrogate modeling based on resampled polynomial chaos expansions. Reliability Engineering and System Safety, 2020, 202, pp.107008. 10.1016/j.ress.2020.107008 . hal-01889651

**HAL Id: hal-01889651**

**<https://centralesupelec.hal.science/hal-01889651v1>**

Submitted on 22 Aug 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu<sup>a,b,\*</sup>, Dominique Lesselier<sup>b</sup>, Bruno Sudret<sup>c</sup>, Joe Wiart<sup>a</sup>

<sup>a</sup>Chaire C2M, LTCl, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France.

<sup>b</sup>Laboratoire des Signaux et Systèmes, UMR8506 (CNRS-CentraleSupélec-Université Paris-Sud), Université Paris-Saclay, Gif-sur-Yvette cedex 91192, France

<sup>c</sup>ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification, Stefano-Franscini-Platz 5, Zürich 8093, Switzerland

---

## Abstract

In surrogate modeling, polynomial chaos expansion (PCE) is popularly utilized to represent the random model responses, which are computationally expensive and usually obtained by deterministic numerical modeling approaches including finite-element and finite-difference time-domain methods. Recently, efforts have been made on improving the prediction performance of the PCE-based model and building efficiency by only selecting the influential basis polynomials (e.g., via the approach of least angle regression). This paper proposes an approach, named as resampled PCE (rPCE), to further optimize the selection by making use of the knowledge that the true model is fixed despite the statistical uncertainty inherent to sampling in the training. By simulating data variation via resampling ( $k$ -fold division utilized here) and collecting the selected polynomials with respect to all resamples, polynomials are ranked mainly according to the selection frequency. The resampling scheme (the value of  $k$  here) matters much and various configurations are considered and compared. The proposed resampled PCE is implemented with two popular selection techniques, namely least angle regression and orthogonal matching pursuit, and a combination thereof. The performance of the proposed algorithm is demonstrated on two analytical examples, a benchmark problem in structural mechanics, as well as a realistic case study in computational dosimetry.

## Keywords:

Surrogate modeling, Sparse polynomial chaos expansion, Resampled polynomial chaos expansion, Data resampling, Sensitivity analysis, Double cross validation

---

## 1. Introduction

Mathematical modeling is common practice nowadays for better understanding real-world phenomena. However, a closed-form solution of the governing equations is unavailable in general and numerical modeling schemes, such as finite-difference time-domain (FDTD) [1] and finite element method (FEM) [2], are commonly employed. The computational method can be considered as a black-box code that takes a vector of parameters as input and yields a vector of quantities of interest that can be further used to assess the system under consideration. However, the real-world system may not be accurately modeled, one critical factor being the uncertainty of input parameters [3], which can be taken into account by setting a probabilistic model of these parameters.

Describing inputs by random variables which follow specific probabilistic density functions (PDFs) [4], the propagation of such random inputs through the system yields random outputs and the investigation of such uncertainty propagation is **one of the major problems in** uncertainty quantification (UQ) [5]. Monte Carlo simulations (MCS) can be applied/used to run the UQ analysis, however, it becomes intractable when the computational cost of a single simulation is high (which corresponds with the cases **described** here). Surrogate

---

\*Corresponding author at Chaire C2M, LTCl, Télécom ParisTech, 46 rue Barrault, 75013 Paris. E-mail address: [zicheng.liu@telecom-paristech.fr](mailto:zicheng.liu@telecom-paristech.fr);

15 model (a.k.a. metamodel) is popularly utilized as a remedy to emulate the system response. Among various  
16 approaches, such as Gaussian process (Kriging method) [6], neural networks [7], etc., surrogate modeling  
17 based on polynomial chaos expansion (PCE) [8, 9, 10, 11, 12] is of interest here due to its advantages in  
18 both interpretation and versatility.

19 Representing the finite-variance random output on a Hilbert space spanned by multivariate basis poly-  
20 nomials orthogonal to the joint PDF of input variables, the numerical modeling of the system response is  
21 replaced by the computation of a PCE, while the expansion coefficients can be obtained by two different  
22 methodologies. For the so-called intrusive methods, taking the spectral finite element method [13] as an  
23 example, the classical FEM is combined with the Karhunen-Loève expansion of input random fields and  
24 the coefficients are obtained by a Galerkin scheme which results in a system of deterministic equations  
25 [14]. In contrast, without modifying the underlying code, hence as non-intrusive methods, coefficients can  
26 be obtained based on an experimental design (ED) by two popularly utilized approaches. While minimiz-  
27 ing the mean square error of data discrepancy leads to the solution of regression method [15], projection  
28 method [16, 17] exploits the orthogonality of basis functions, the expansion coefficient being the solution of  
29 multidimensional integrations which can be computed by quadrature methods.

30 A PCE, as an infinite series, should be truncated for computational purpose. How to perform this  
31 truncation optimally is the major issue, which is addressed in this paper. In the literature, a maximum  
32 value is commonly set to the total degree of multivariate polynomials [18]. However, the number of basis  
33 polynomials, as well as the required ED size, dramatically increases with the number of input variables,  
34 which is known as the curse-of-dimensionality. Thus, the so-called sparse PCE [19, 20, 21, 22] has been  
35 developed by only including the most influential polynomials in the truncation. Measuring this influence  
36 by correlation, the classical greedy algorithms, orthogonal matching pursuit (OMP) [23] and least angle  
37 regression (LARS) [24], have been utilized to rank the polynomials.

38 This contribution is aimed at stabilizing the constructed sparse PCE model with respect to small changes  
39 in the training data. Bagging (a.k.a. bootstrap aggregating) [25] is a popular approach, especially for  
40 decision tree methods, to stabilize the modeling approach by training multiple regression models based  
41 on bootstrap resamples [26] and taking the final prediction as the mean of all predictions. In the study  
42 of variable selection, rather than treating the resamples independently in the construction of regression  
43 models, the so-called *inclusion frequency* [27] (or *inclusion fraction* [28]) is computed as the criterion for the  
44 importance of a variable. With the knowledge that resamples are perturbed versions of the same original  
45 data, the truly important variables should be included in the built model for most bootstrap resamples since  
46 all models should reflect the same underlying data structure. The utilization of inclusion frequency improves  
47 the replication stability of selected variables [27, 29].

48 In this paper, the idea of inclusion frequency is applied to the construction of a sparse PCE model. Based  
49 on LARS or OMP, multiple PCE models are constructed based on resamples and involved basis polynomials  
50 are ranked according to the inclusion frequency. The replication stability of selected polynomials in the final  
51 model is expected to be enhanced. Since the PCE model is highly determined by the basis, the stability of  
52 the built model would be increased as well. Such construction method of a sparse PCE model is named as  
53 *resampled PCE* (rPCE).

54 Improvements and adjustments are made in rPCE based on the application procedure of inclusion fre-  
55 quency on variable selection. First, recent work in [30] shows subsampling [31] is superior to bootstrapping  
56 in the ability of distinguishing important and redundant variables and in the favor of sparse models. **Remark**  
57 **that subsampling consists of randomly drawing part of samples without replacement while bootstrapping**  
58 **approach generates observations of the same size as the original data but with replacement.** Here, an effi-  
59 cient subsampling technique,  $k$ -fold division, is applied, where the original data is divided into  $k$  parts and  
60 a resampling data set is composed of any  $k - 1$  parts. This procedure ensures the original data is fully  
61 explored with  $k$  resamples. In variable selection, variables are roughly labeled “important” or “redundant”  
62 by comparing the associated inclusion frequency with a cut-off value, the choice of which is still an open  
63 problem [32]. In rPCE, while the basis polynomials are ranked by inclusion frequency, the number of in-  
64 cluded polynomials in the final model is decided by cross validation. Moreover, for polynomials with the  
65 same inclusion frequency, the associated cross-validation errors are taken as an extra criterion for further  
66 ranking. Such ranking approach provides the possibility to combine different basis pursuit methods. Efforts

67 trying to merge the selection results of LARS and OMP are made.

68 This paper itself is organized as follows. A general framework of the PCE-based surrogate modeling  
69 is introduced in Section 2. Section 3 gives the concept of the full and sparse PCE truncation, where  
70 the building processes based on LARS and OMP are briefly described, respectively. The methodology of  
71 rPCE is illustrated in Section 4. Resampling data through the random division into  $k$  parts, based on the  
72 generated candidate polynomials by LARS and/or OMP, the importance of polynomials is evaluated through  
73 the inclusion frequency. The value of  $k$  matters and the determination strategy is discussed in Section 5,  
74 where the strategy to select the source of candidate polynomials (LARS, OMP, or their combination) is  
75 also presented. The improved performances in prediction and sensitivity analysis by rPCE are shown via  
76 application to two classical analytical functions, one finite-element model and one finite-difference-time-  
77 domain model in Section 6. Conclusions and perspectives follow in Section 7.

## 78 2. Surrogate model based on polynomial chaos expansion

### 79 2.1. Probabilistic modeling

80 Consider a physical model represented by a deterministic function  $\mathbf{y} = \mathcal{M}(\mathbf{x})$ , where  $\mathbf{x} \in \mathbb{R}^M$  and  
81  $\mathbf{y} \in \mathbb{R}^Q$ ,  $M, Q$  being the number of input and output quantities, respectively. The uncertainty of inputs and  
82 the propagation to responses lead to the description of  $\mathbf{x}$  and  $\mathbf{y}$  as random vectors,  $\mathbf{X}$  and  $\mathbf{Y}$ . Here, since  
83 each component of  $\mathbf{Y}$  can be separately analyzed in statistical learning, only cases with scalar response, i.e.,  
84  $Q = 1$ , are considered for simplicity.

Describing the random vector  $\mathbf{X}$  by the joint probability density function (PDF)  $p_{\mathbf{X}}$  and assuming that  
 $Y$  has a finite variance, the latter belongs to a Hilbert space  $L^2(\mathbb{R}^M, \mathcal{B}_M, \mathbb{P}_{\mathbf{X}})$ ,  $\mathcal{B}_M$  being the Borel  $\sigma$ -algebra  
of the event space  $\mathbb{R}^M$  and  $\mathbb{P}_{\mathbf{X}}$  being the probability measure of  $\mathbf{X}$ . The Hilbert space is equipped with the  
following inner product

$$\langle f, g \rangle = E[f(\mathbf{X})g(\mathbf{X})] = \int_{\mathbb{X}} f(\mathbf{x})g(\mathbf{x})p_{\mathbf{X}}(\mathbf{x})d\mathbf{x}, \quad (1)$$

85 and can be represented by a complete set of orthogonal basis functions.

### 86 2.2. Polynomial chaos expansion

Polynomial chaos expansion is a spectral representation of  $Y$  taking polynomials as basis functions,

$$Y = \sum_{\alpha \in \mathbb{N}^M} \beta_{\alpha} \psi_{\alpha}(\mathbf{X}), \quad (2)$$

87 where  $\alpha$  is a vector of non-negative integers indicating the order of multivariate polynomials  $\psi_{\alpha}$  and  $\beta_{\alpha}$   
88 is the corresponding expansion coefficient.

**The construction of  $\psi_{\alpha}(\mathbf{X})$  is briefly recalled now [8, 11].** Assuming that the input random variables are  
independent, the multivariate polynomials is a tensor product of univariate polynomials  $\pi_{\alpha_i}$ , i.e.,

$$\psi_{\alpha}(\mathbf{X}) = \pi_{\alpha_1}^{(1)}(X_1) \times \dots \times \pi_{\alpha_M}^{(M)}(X_M), \quad (3)$$

89 where  $\pi_{\alpha_i}^{(i)}$ 's are univariate orthonormal polynomials with respect to the PDF of the  $i$ -th parameter  $X_i$ ,  
90 with degree  $\alpha_i$  (e.g., Hermite polynomials for Gaussian distributions). This methodology is referred to as  
91 **generalized PCE (gPCE) [11, 12].** For PDFs not included in gPCE, a nonlinear mapping of input variables  
92 to the known ones can be made with the technique of isoprobabilistic transformation [33, 34] or specific  
93 orthogonal polynomials are computed numerically via the Stieltjes procedure [35].

The PCE coefficients  $\beta_{\alpha}$  are obtained in a non-intrusive way by the regression approach. A data set  
 $\{\mathbf{x}^{(n)}, n = 1, \dots, N\}$  sampled from the input PDF  $p_{\mathbf{X}}$  and the corresponding response  $\{y^{(n)} = \mathcal{M}(\mathbf{x}^{(n)})\}$   
compose altogether the ED. With notations of column vector  $\mathbf{y} = [y^{(n)}]$ ,  $\boldsymbol{\beta} = [\beta_{\alpha}]$  and matrix  $\boldsymbol{\psi} = [\psi_{\alpha}(\mathbf{x}^{(n)})]$ ,  
the PCE coefficients can be obtained from

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \boldsymbol{\psi}\boldsymbol{\beta}\|_2^2, \quad (4)$$

which yields the ordinary least square (OLS) [36] solution as **the normal equation**

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\psi}^T \boldsymbol{\psi})^{-1} \boldsymbol{\psi}^T \mathbf{y}, \quad (5)$$

the superscript “ $T$ ” denoting the transpose operation. **Remark that overfitting problems may be suffered for the OLS solution but can be avoided by implementing regularization [22, 37] techniques, as provided by the LARS algorithm described later on.**

Remark that, although only cases with independent inputs are considered in the above analysis, it is possible to describe the mutual dependence by a copula [38] and use Rosenblatt transformation [33] to cast the problem as a function of auxiliary independent variables.

### 2.3. Estimation of prediction performance

The model assessment is often performed by Monte Carlo simulations with a large test dataset, which is independent from the experimental design. Denote  $\widehat{\mathcal{M}}$  as the surrogate model, the input vector and response of the  $n$ -th test data as  $\mathbf{x}_{\text{test}}^{(n)}$  and  $y_{\text{test}}^{(n)}$ , respectively. The performance of the constructed model is assessed by computing the mean square error of data discrepancy

$$\epsilon_{\text{test}} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \left( \mathcal{M}(\mathbf{x}_{\text{test}}^{(n)}) - \widehat{\mathcal{M}}(\mathbf{x}_{\text{test}}^{(n)}) \right)^2. \quad (6)$$

For an easier interpretation of  $\epsilon_{\text{test}}$ , the associated coefficient of determination  $R_{\text{test}}^2$  is computed by

$$R_{\text{test}}^2 = 1 - \frac{\epsilon_{\text{test}}}{\text{Var}(\mathbf{y}_{\text{test}})}, \quad (7)$$

where  $\text{Var}(\mathbf{y}_{\text{test}}) = \sum_{n=1}^{N_{\text{test}}} (y_{\text{test}}^{(n)} - \bar{y}_{\text{test}})^2 / (N_{\text{test}} - 1)$  and  $\bar{y}_{\text{test}} = \sum_{n=1}^{N_{\text{test}}} y_{\text{test}}^{(n)} / N_{\text{test}}$ . Therefore, the closer  $R_{\text{test}}^2$  is to one, the more accurate is the prediction by  $\widehat{\mathcal{M}}$ .

However, in scenarios with high computational cost for a single simulation, it is usually intractable to have a large **test dataset**. Then, the same data as for training are often reused for **model assessment**. However, the underestimation of the generalization error is well-known in the case of overfitting [18]. **Cross-validation** was thus proposed and is commonly advocated [39, 40]. **Here, leave-one-out cross-validation (LOOCV) is applied and the corresponding cross-validation error reads:**

$$\epsilon_{\text{LOO}} = \frac{1}{N} \sum_{n=1}^N \left( \mathcal{M}(\mathbf{x}^{(n)}) - \widehat{\mathcal{M}}^{-(n)}(\mathbf{x}^{(n)}) \right)^2, \quad (8)$$

where  $\widehat{\mathcal{M}}^{-(n)}$  denotes the surrogate model trained by leaving the  $n$ -th data out. Remark that  $\epsilon_{\text{LOO}}$  is also known as predicted residual of squares (PRESS) or jackknife error [41] and it can be computed fast in single training process [18] by

$$\epsilon_{\text{LOO}} = \frac{1}{N} \sum_{n=1}^N \left( \frac{\mathcal{M}(\mathbf{x}^{(n)}) - \widehat{\mathcal{M}}(\mathbf{x}^{(n)})}{1 - h_n} \right)^2, \quad (9)$$

where  $h_n$  is the  $n$ -th diagonal element of the matrix  $\boldsymbol{\psi} (\boldsymbol{\psi}^T \boldsymbol{\psi})^{-1} \boldsymbol{\psi}^T$ .

## 3. Surrogate modeling based on full PCE and sparse PCE

The accurate PCE of the true model is an infinite series and needs a truncation for the sake of computation. From Eq. (2), one sees that truncating a PCE is actually selecting a subset of  $\mathbb{N}^M$  for  $\boldsymbol{\alpha}$  such that the system response can be represented by the associated polynomials at a sufficient accuracy. Assuming the selected  $\boldsymbol{\alpha}$  vectors compose the set  $\mathbb{A}$ , the truncated PCE can be written as

$$\widehat{\mathcal{M}}(\mathbf{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{A}} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathbf{X}). \quad (10)$$

105 **Setting a maximum value to the total degree of polynomials** leads to the so-called full PCE model, which  
 106 suffers from the curse-of-dimensionality [42], meaning that the cardinality of  $\mathbb{A}$  sharply increases with the  
 107 number of input parameters, as explained below. **While the problem of curse-of-dimensionality can be**  
 108 **moderated by the algorithm of Smolyak sparse quadrature [43]**, recently least angle regression (LARS)  
 109 [20, 44] and orthogonal matching pursuit (OMP) [23, 44] have been used to downsize the truncation and  
 110 achieve the so-called sparse PCE model.

### 111 3.1. Full PCE model

112  $\mathbb{A}$  is commonly selected by setting a maximum to the total degree of multivariate polynomials, i.e.,  
 113  $\mathbb{A}_{\text{full}} = \{\boldsymbol{\alpha} \in \mathbb{N}^M, \sum_{i=1}^M \alpha_i \leq p\}$ ,  $p$  a positive integer. The PCE-based surrogate model with this setup  
 114 is named in the sequel as the full PCE model. However, the cardinality of  $\mathbb{A}_{\text{full}}$ , denoted by  $P_{\text{full}}$ , equals  
 115  $\binom{p+M}{p}$  and polynomially increases with the value of  $p$  and  $M$ . Moreover, to ensure the well-conditioning of  
 116 the information matrix  $\boldsymbol{\psi}$  in Eq. (5), the ED size  $N$  should be larger than  $P_{\text{full}}$ . As a result, the resulting  
 117 curse-of-dimensionality prevents the application of the full PCE model in scenarios with large  $p$  and  $M$ .

### 118 3.2. Sparse PCE model

119 The problem of curse-of-dimensionality during the construction of full PCE models is addressed by  
 120 constructing the so-called sparse PCE models, where  $\mathbb{A}$  is downsized through the use of greedy algorithms,  
 121 so that only the most influential polynomials are included in the truncated PCE.

---

For  $p = 1, \dots, p_{\text{max}}$ ,

1.  $\mathbb{A}_{\text{full}} = \{\boldsymbol{\alpha} \in \mathbb{N}^M, \sum_{i=1}^M \alpha_i \leq p\}$  and set active set  $\mathbb{A}_0^a = \emptyset$ ;
2. Rank basis polynomials in  $\{\boldsymbol{\psi}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{A}_{\text{full}}^p\}$  by LARS, OMP, or rPCE. The  $\boldsymbol{\alpha}$  corresponding with the first  $J_{\text{max}}$  most influential basis polynomials compose the set  $\{\boldsymbol{\alpha}_j, j = 1, \dots, J_{\text{max}}\}$ .
3. For  $j = 1, \dots, J_{\text{max}}$ ,  
 Update  $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$ . Based on  $\boldsymbol{\psi}_{\mathbb{A}_j^a}$ , compute  $\boldsymbol{\beta}_j$  as the OLS solution and associated  $\epsilon_{\text{LOO}}^j$ .  
 End
4.  $J = \arg \min_j \{\epsilon_{\text{LOO}}^j\}$  and  $\epsilon_{\text{LOO}}^{p, \min} = \epsilon_{\text{LOO}}^J$ . When  $p \geq 3$ , if  $\epsilon_{\text{LOO}}^{p, \min} > \epsilon_{\text{LOO}}^{p-1, \min} > \epsilon_{\text{LOO}}^{p-2, \min}$ , stop the model-construction process and output the PCE model corresponding with  $\boldsymbol{\psi}_{\mathbb{A}_J^a}$ .

End

---

Table 1: Procedures of constructing a sparse PCE model,  $J_{\text{max}} = \min\{N - 1, \text{card}(\mathbb{A}_{\text{full}})\}$ .

122 Table 1 presents the procedures to construct a sparse PCE model. Based on candidate  $\boldsymbol{\alpha}$  from the full  
 123 PCE model, i.e.,  $\boldsymbol{\alpha} \in \mathbb{A}_{\text{full}}$ , the associated basis polynomials are ranked (e.g., by correlation with response  
 124 data  $\boldsymbol{y}$  for OMP in Table A.1) and the first  $J_{\text{max}}$  most influential ones are selected by OMP (refer to algorithm  
 125 in Table A.1) or LARS (refer to Table A.2),  $J_{\text{max}}$  being the maximum number of included polynomials in  
 126 the final constructed PCE model and set as  $\min\{N - 1, \text{card}(\mathbb{A}_{\text{full}})\}$  (otherwise the least-square problem  
 127 becomes ill-posed).

128 Assessing the model performance by leave-one-out cross-validation, the optimal number of selected poly-  
 129 nomials,  $J$ , corresponds with the PCE model with the minimal  $\epsilon_{\text{LOO}}$ , the computation of which follows (9),  
 130 where only the surrogate model constructed with the whole set of data is required.

131 The optimal value for the total degree of polynomials follows an early-stopping criterion. Setting the  
 132 maximum value for  $p$ , a progressive increase stops when the minimal  $\epsilon_{\text{LOO}}$  increase with two consequent  $p$ .

## 133 4. Surrogate modeling based on resampled PCE

134 During replications with resampled training data, different PCE truncations are obtained by LARS or  
 135 OMP and the inclusion frequency of involved polynomials can be computed. Resampled PCE (rPCE) is

136 proposed to refine standard PCE truncation schemes by making use of the inclusion frequency. Cross-  
 137 validation error associated with each polynomial is an additional factor to further rank polynomials with  
 138 the same inclusion frequency. Efforts to combine selection results by LARS and OMP to further improve  
 139 the performance of rPCE are also presented.

#### 140 4.1. Resampled PCE based on LARS or OMP

141 Inclusion frequency is defined as the percentage of replications [27] in which a given basis polynomial is  
 142 selected by LARS or OMP. **The variation of training data is simulated by the subsampling technique,  $k$ -fold**  
 143 **division**, considering its efficiency in exploiting the information of original data, i.e., the training process  
 144 makes use of all data in  $k$  replications.

145 Dividing the whole set of data into  $k$  subsets, all with approximately same size. Of  $k$  subsets, the  $l$ -th  
 146 subset is left out and the remaining  $k - 1$  subsets are used for the PCE construction. Varying  $l$  from 1  
 147 to  $k$ , one **has**  $k$  PCE models built by LARS/OMP and the associated active sets are denoted by  $\mathbb{A}_{P,(l)}^a$ ,  
 148  $l = 1, \dots, k$ . The subscript “ $P$ ” and superscript “ $a$ ” are ignored in  $\mathbb{A}_{P,(l)}^a$  to be  $\mathbb{A}_{(l)}$  in the followings.

149 To search for the most frequent  $\alpha$  indices within the  $k$  different sets  $\mathbb{A}_{(l)}$ ,  $l = 1, \dots, k$ , one can merge the  
 150 latter into a **multiset**  $\mathbb{A}^{\text{Mul}} = \{\mathbb{A}_{(1)}, \dots, \mathbb{A}_{(k)}\}$ , **the superscript “Mul” denoting a multiset (rather than set),**  
 151 **which allows for multiple instances for each  $\alpha$ .** Then the selection frequency of  $\alpha$  in the  $k$  building processes  
 152 is equal to the number of its duplicates in  $\mathbb{A}^{\text{Mul}}$ . **Denote  $\mathbb{A}$  as the set (thus no duplicate elements) composed**  
 153 **of elements in  $\mathbb{A}^{\text{Mul}}$ . The selection frequency corresponding with each of element in  $\mathbb{A}$  is an integer in the**  
 154 **interval  $[1, k]$  and saved in the vector  $\mathbf{s}_f$ .** The inclusion frequency is computed as the normalized frequency,  
 155 i.e.,  $\mathbf{s}_f/k$ .

156 For applications wherein the idea of inclusion frequency has been applied, the final model keeps compo-  
 157 nents (e.g., **influential variables for the variable-selection problem [27]**) for which the inclusion frequency  
 158 exceeds the cutpoint  $\nu$ . The value of  $\nu$  impacts much on the stability and complexity of the final model but  
 159 is usually arbitrarily taken [30, 32], and no conclusive method seems available for an optimal choice of  $\nu$   
 160 [27, 29]. To avoid this problem, based on the ranked polynomials, the total number of included polynomials  
 161 (rather than the cutpoint  $\nu$ ) in the final model is chosen by cross validation **following the procedures in**  
 162 **Table 1.**

163 However, during the running of rPCE, different **multi-indices**  $\alpha$  might have the same frequency, which  
 164 introduces some uncertainty in the ranking of polynomials. To avoid this uncertainty, one more factor,  
 165 namely the effect of each basis polynomial on  $\epsilon_{\text{LOO}}$ , is considered.

166 From the LARS/OMP procedures, one can see that the correlated polynomials are sequentially added  
 167 into the active set, thus the increment of  $\epsilon_{\text{LOO}}$  by adding  $\alpha_j$  into  $\mathbb{A}_{j-1}^a$  equals  $\Delta\epsilon_{\text{LOO}}^j = \epsilon_{\text{LOO}}^j - \epsilon_{\text{LOO}}^{j-1}$  for  
 168  $j \geq 1$ , where  $\epsilon_{\text{LOO}}^0$  is set as 0. **Thus, each  $\alpha$  in  $\mathbb{A}^{\text{Mul}}$  corresponds with a  $\Delta\epsilon_{\text{LOO}}$ .**

**Add the superscript “ $(l)$ ” to the notation standing for the quantity obtained by leaving the  $l$ -th subset**  
**out from model construction.** Then, the so-called **error score**  $\mathbf{s}_e$  can be computed as the mean of all terms  
 $\Delta\epsilon_{\text{LOO}}^{(l),j}$  mapping to the same element of  $\mathbb{A}$ , i.e.,

$$169 \quad s_e^i = \frac{1}{s_f^i \Delta\epsilon_{\text{LOO}}^{\max}} \sum_{\{(l),j|\alpha^{(l),j}=\alpha^i\}} \Delta\epsilon_{\text{LOO}}^{(l),j}, \quad i = 1, \dots, \text{card}\{\mathbb{A}\}. \quad (11)$$

where the superscript “ $i$ ” stands for the  $i$ -th element of a vector or set. The normalization by  $\Delta\epsilon_{\text{LOO}}^{\max}$ , the  
 maximum element of  $|\Delta\epsilon_{\text{LOO}}^{(l),j}|$ , is to confine the value of  $s_e^i$  between  $-1$  and  $1$  such that the ranking of  
 polynomials by the total score

$$170 \quad \mathbf{s} = \mathbf{s}_f + \mathbf{s}_e, \quad (12)$$

169 is mainly affected by  $\mathbf{s}_f$  in rPCE. Remark that  $\mathbf{s}_f$  is used instead of inclusion frequency (the normalized  $\mathbf{s}_f$ )  
 170 and is **subsequently named frequency score.**

#### 171 4.2. Resampled PCE combining LARS and OMP

172 The way to rank polynomials in rPCE allows the possibility to combine the results by LARS and OMP.  
 173 Following the procedures in Section 4.1,  $\mathbb{A}^{\text{Mul}}$  and  $\mathbb{E}^{\text{Mul}}$  (**multiset of  $\Delta\epsilon_{\text{LOO}}^{(l),j}$** ) can be obtained by LARS

174 and OMP separately, denoted by  $\mathbb{A}^{\text{Mul,LARS}}$ ,  $\mathbb{E}^{\text{Mul,LARS}}$  and  $\mathbb{A}^{\text{Mul,OMP}}$ ,  $\mathbb{E}^{\text{Mul,OMP}}$ , respectively. Then,  
 175 merging results by LARS and OMP into a single **multiset**,  $\mathbb{A}^{\text{Mul}} = \{\mathbb{A}^{\text{Mul,LARS}}, \mathbb{A}^{\text{Mul,OMP}}\}$  and  $\mathbb{E}^{\text{Mul}} =$   
 176  $\{\mathbb{E}^{\text{Mul,LARS}}, \mathbb{E}^{\text{Mul,OMP}}\}$ , from which  $\mathbb{A}$  and the associated total score  $\mathbf{s}$  can be computed. Then, **the basis**  
 177 **polynomials associated with  $\mathbb{A}$  are ranked according to  $\mathbf{s}$  and the construction of a sparse PCE model follows**  
 178 **procedures in Table 1.**

## 179 5. Parameter settings

### 180 5.1. Resampling scheme

181 The  $k$ -fold division is used to simulate the data variation in rPCE and the value of  $k$  matters on the  
 182 performance. A tradeoff lies behind the determination of  $k$ . With a small  $k$  (e.g.,  $k = 2$ ), a large portion  
 183 (half) of data is apart from the building process. As a result, some information of the true system might  
 184 be lost or not accurately learned by the surrogate model and the selected polynomials may not be truly  
 185 influential. On the other side, a large  $k$ , (e.g.,  $k = N$ ) cannot sufficiently simulate the data statistical  
 186 variation and the selected polynomials in the construction of the  $k$  different PCEs might have a high  
 187 correlation. This way, the polynomials selected by rPCE would be almost the same as those with LARS or  
 188 OMP and the prior knowledge, from which rPCE is to benefit, cannot be well exploited.

189 The proposed strategy is to merge  $\mathbb{A}^{\text{Mul}}$  obtained for different values of  $k$ . Considering that the validation  
 190 error on the data left out is used to estimate the prediction performance in Section 5.2 and values of  
 191 3, 5, 10, 20,  $N$  (leave-one-out), are usually recommended [39, 45, 46] for  $k$ -fold cross validation, rPCE will  
 192 run based on the **multiset**  $\mathbb{A}^{\text{Mul}} = [\mathbb{A}_3^{\text{Mul}}, \mathbb{A}_5^{\text{Mul}}, \mathbb{A}_{10}^{\text{Mul}}, \mathbb{A}_{20}^{\text{Mul}}, \mathbb{A}_N^{\text{Mul}}]$ , where the subscript of  $\mathbb{A}_q^{\text{Mul}}$  corresponds  
 193 with the value of  $k$ . Data variation is fully simulated via  $k = 3, 5$  and the bias error is small considering  
 194 that in average about  $0.86N$  resamples (without replacement) are used to generate candidate polynomials.  
 195 It seems not easy to optimize the setting of  $k$ , especially considering that the optimal value may differ w.r.t.  
 196 scenarios. However, the proposed setting is revealed robust in the various application examples.

With respect to a set of  $k$  values, i.e.,  $k = \{3, 5, 10, 20, N\}$ , the total score can be computed based on  
 $\mathbf{s}_{f,k}$  and  $\mathbf{s}_{e,k}$ , the subscript “ $k$ ” indicating the quantity for a specific value of  $k$ . Denote  $\mathbb{A}$  as the copy of  
 $\mathbb{A}^{\text{Mul}}$  but without element duplication. For each  $\alpha$  in  $\mathbb{A}$ , its selection frequency can be computed by

$$f^i = \sum_{k=\{3,5,10,20,N\}} s_{f,k}^i, \quad i = 1, \dots, \text{card}(\mathbb{A}), \quad (13)$$

197 where the superscript “ $i$ ” stands for the  $i$ -th element of a vector and  $s_{f,k}^i$  equals zero if the  $i$ -th  $\alpha$  of  $\mathbb{A}$  is not  
 198 in  $\mathbb{A}_k^{\text{Mul}}$ . Since  $s_{f,k}^i$  is upper bounded by  $k$ , the polynomials selected with small values of  $k$  (e.g., elements  
 199 in  $\mathbb{A}_3^{\text{Mul}}$ ) will have small **values of  $f^i$**  and be less likely to have high ranks in rPCE.

To solve this problem, instead of (13), the frequency score is computed as a summation of weighted  $s_{f,k}^i$ :

$$s_f^i = \sum_{k=\{3,5,10,20,N\}} s_{f,k}^i \frac{\text{lcm}(3, 20, N)}{k}, \quad i = 1, \dots, \text{card}(\mathbb{A}), \quad (14)$$

200 where  $\text{lcm}(3, 20, N)$  computes the least common multiple of 3, 20,  $N$  (same for 3, 5, 10, 20,  $N$ ). The weights  
 201 give rise to the same maximum value of the summands in (14). Consequently, the candidate polynomials  
 202 w.r.t. different values of  $k$  are equally considered in rPCE.

203 Finally, the set of  $k$  values, i.e.,  $\{3, 5, 10, 20, N\}$ , needs an adjustment for a small  $N$ . For instance,  $k$  can  
 204 only be 3, 5, 10,  $N$  when  $N = 15$ .

The computation of error score follows as:

$$s_e^i = \frac{1}{f^i} \sum_{k=\{3,5,10,20,N\}} s_{e,k}^i, \quad i = 1, \dots, \text{card}(\mathbb{A}), \quad (15)$$

205 where  $s_{e,k}^i$  equals zero if the  $i$ -th  $\alpha$  of  $\mathbb{A}$  is not in  $\mathbb{A}_k^{\text{Mul}}$ .



206 *5.2. Source of candidate polynomials*

207 Section 4 presents the rPCE based on candidate polynomials generated by three sources, LARS, OMP or  
 208 their combination, and one needs to decide which source is the optimal option. The polynomials commonly  
 209 and frequently selected by two different approaches are believed influential and more likely to be included  
 210 in rPCE. However, if one approach has a much worse performance than the other, the combination scheme  
 211 would not be recommended, since the candidate polynomials generated by the worse approach might deteri-  
 212 orate the performance of rPCE. Therefore, if LARS is much better than OMP, only candidate polynomials  
 213 by LARS participate into the ranking in rPCE, and vice versa. Otherwise, the combination scheme is used.

214 The criterion of “much better” should be properly set. Assuming a large set of validation data is available,  
 215 as illustrated in Section 2.3,  $R_{\text{test}}^2$  can be computed as the unbiased estimation of the prediction performance.  
 216 Here, the comparison of two building approaches is conducted with the analysis of the distribution of  $R_{\text{test}}^2$ .  
 217 Varying the training data, a sequence of surrogate models is built and the associated  $R_{\text{test}}^2$  values are  
 218 computed. Representing  $\mathbb{R}_{\text{test,LARS}}^2$  and  $\mathbb{R}_{\text{test,OMP}}^2$  as the sets of  $R_{\text{test}}^2$  values obtained by LARS and OMP  
 219 respectively, the first and third quartile of these two sets are computed and denoted by  $Q_1^{\text{LARS}}$ ,  $Q_1^{\text{OMP}}$ ,  
 220  $Q_3^{\text{LARS}}$ ,  $Q_3^{\text{OMP}}$ . Then, if  $Q_1^{\text{LARS}} > Q_3^{\text{OMP}}$ , one considers that LARS is much better than OMP, and vice  
 221 versa. Otherwise, LARS and OMP are considered with similar performances and the combination scheme  
 222 would be adopted.

223 However, again a large set of validation data is usually not available due to the high computational costs.  
 224 Here,  $R_{\text{test}}^2$  is approximated through the validation on the data left out in the  $k$ -fold division. With different  
 225 values of  $k$  and  $l$ , the validations generate a set of determination coefficient  $R_{k,(l)}^2$  as the approximations to  
 226  $R_{\text{test}}^2$ ,  $l = 1, \dots, k$ ,  $k \in \{3, 5, 10, 20, N\}$ . Denoting  $\mathbb{R}_{\text{LARS}}^2$  and  $\mathbb{R}_{\text{OMP}}^2$  as the sets of  $R_{k,(l)}^2$  values obtained by  
 227 LARS and OMP, the distribution of sets  $\mathbb{R}_{\text{test}}^2$  is then simulated by  $\mathbb{R}_{\text{LARS}}^2$  and  $\mathbb{R}_{\text{OMP}}^2$ . Remark that two  
 228 layers of cross validations now have been operated in rPCE. The outer cross validation is just illustrated to  
 229 simulate the distribution of  $\mathbb{R}_{\text{LARS}}^2$  and  $\mathbb{R}_{\text{OMP}}^2$ . The inner one is embedded in the running of LARS and OMP  
 230 to compute  $\epsilon_{\text{LOO}}$  in Table A.1 and A.2. The two-layer cross validation here is indeed an realization of the  
 231 known *double-cross-validation* (DCV) [47] or *cross model validation* (CMV) [45, 48]. The related literature  
 232 shows the unbiased estimation of  $R_{\text{test}}^2$  by the determination coefficient from the outer cross-validation errors,  
 233 i.e.,  $R_{k,(l)}^2$ .

234 The procedures to rank basis polynomials by rPCE are summarized in Fig. 1. Then, the construction  
 235 of sparse PCE models follows the steps in Table 1.

236 Benefiting from the obtained PCE model, the global sensitivity analysis, which measures the impacts  
 237 of input variables to the response, can be conducted via the computation of Sobol’ indices [49, 50] for  
 238 independent variables or Kucherenko indices [51] for dependent cases by Monte-Carlo simulations. Note  
 239 that in the case of independent inputs, Sobol’ indices are readily available from PCE coefficients, as shown  
 240 in [52].

241 **6. Application examples**

242 The knowledge that the influential polynomials are to be frequently selected during replications is first  
 243 checked on a specially designed function, the true basis polynomials of which are known. Then, to present the  
 244 performance of surrogate modeling based on rPCE and the comparisons to LARS and OMP, two benchmark  
 245 functions (with dimension  $M = 3$  and  $M = 8$ , respectively), a finite-element model (with  $M = 10$ ) and  
 246 a finite-difference-time-domain model (with  $M = 4$ ) are analyzed. The PCE models based on LARS and  
 247 OMP are obtained with the Matlab package UQLab ([www.uqlab.com](http://www.uqlab.com)) [53, 54], where the maximum degree  
 248 of multivariate polynomials  $p$  is set as 20. Using resampling, UQLab provides the candidate polynomials  
 249 to rPCE. Remark that if no specific configurations are given in the following examples, resampled PCE is  
 250 performed with the suggested configurations in Section 5, i.e., optimized source (LARS, OMP, or both) of  
 251 candidate polynomials and candidate polynomials from  $k = \{3, 5, 10, 20, N\}$ .

252 Latin-Hypercube sampling [55] is used to sample the input random variables. Since cases with a small  
 253 ED are concerned in this paper, the size of ED  $N$  is chosen between 10 and 50 here. As mentioned in Section  
 254 2.2, dependent variables can be analyzed after the transformation into the corresponding independent ones

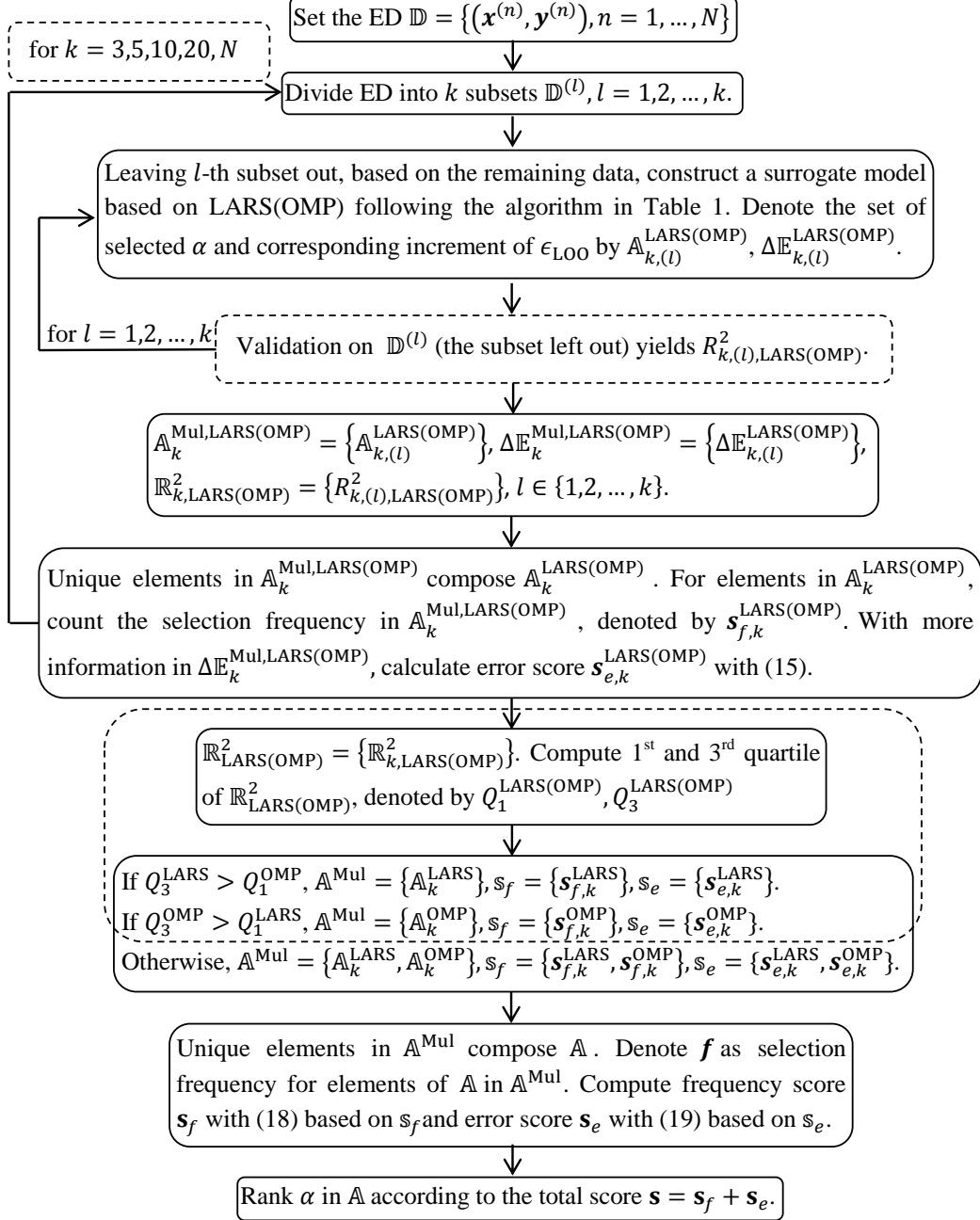


Fig. 1: Flow chart for ranking basis polynomials based on resampled PCE, where steps enclosed by dashed lines are with the suggested configurations in Section 5.

255 through the generalized Nataf transformation, so only examples with independent variables are presented  
 256 in this section and the global sensitivity is analyzed with the computation of Sobol' indices.

257 *6.1. Summation of multivariate polynomials*

To show that the influential polynomials associated with the true model are frequently selected, the surrogate modeling of the following expression,

$$Y = 1 + X_1 + X_1 X_2 + X_1 X_2^2 + X_1 X_2^3, \quad (16)$$

258 which is a summation of five multivariate polynomials (including the constant term), is conducted.  $X_1$  and  
 259  $X_2$  are independent variables that follow the Gaussian distributions  $\mathcal{N}(0, 1)$  and  $\mathcal{N}(6, 1)$ , respectively. OMP  
 260 is used to build a sparse PCE model with 12 data points for training and  $10^4$  data for independent testing.  
 261 A total of 100 PCE constructions are made to check the selection frequency of polynomials.

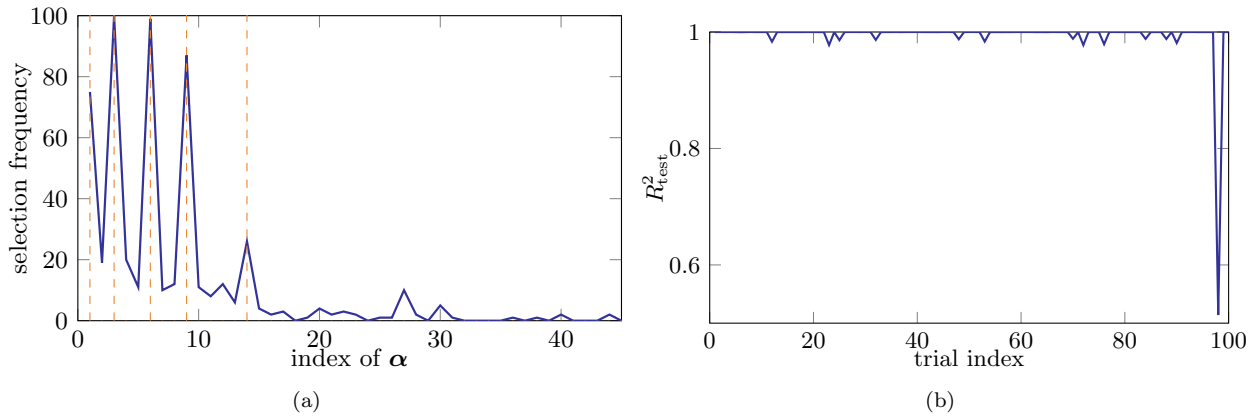


Fig. 2: Example 1: Summation of multivariate polynomials - (a) the selection frequency of  $\alpha$  by OMP and (b) the associated  $R_{\text{test}}^2$  in all replications

262 Due to the Gaussian distribution of input variables, Hermite polynomials are used to compose the  
 263 basis, where the bivariate polynomials are indexed by  $\alpha = (\alpha_1, \alpha_2)$ . The constant term corresponds with  
 264  $\alpha = (0, 0)$ , while the other four terms in Eq. (16) are with  $(1, 0), (1, 1), (1, 2), (1, 3)$ , respectively. Labeling  
 265  $\alpha$  by integers, the selection frequency during the 100 PCE constructions is plotted in Fig. 2(a), where the  
 266 dashed lines indicate the five true  $\alpha$  indices. Remark that, the selection frequency is smaller than 2 when  
 267 the labels are larger than 45 and only the results with labels  $\leq 45$  are displayed for a better visualization.  
 268 As observed, although the true indices of  $\alpha$  are not always selected, they are the most frequent ones during  
 269 replications. Making use of this knowledge and selecting the most frequent  $\alpha$  (also the associated polynomial)  
 270 may improve the performance of the obtained PCE model and avoid the outliers (for example the 98-th  
 271 replication with  $R_{\text{test}}^2 = 0.51$  in Fig. 2(b), where  $X_2, X_1, X_1^3$  are selected as the basis).

272 *6.2. Ishigami function*

The Ishigami function, which is defined by

$$Y = \sin X_1 + a \sin^2 X_2 + b X_3^4 \sin X_1, \quad (17)$$

273 is widely used for benchmarking in uncertainty and sensitivity analysis. The parameters are set to  $a = 7$ ,  
 274  $b = 0.1$  and the input random variables  $X_i, i = 1, 2, 3$ , are independent and uniformly distributed over  
 275  $[-\pi, \pi]$ . Legendre polynomials are thus used as the basis according to the principle of the generalized PCE.

276 First, 50 data points are used for building the surrogate model and  $10^4$  points for estimating the prediction  
 277 performance. The analysis is repeated 100 times in order to investigate the statistical uncertainty of different  
 278 modeling approaches. The prediction of all validation data ( $10^6$  data over 100 replications) by the surrogate  
 279 models built based on LARS, OMP and rPCE is shown in Fig. 3, where  $y$  stands for the true value,  $\hat{y}$  for the  
 280 predicted one, and the solid line indicates the case when  $\hat{y}$  exactly equals  $y$ . As observed, although rPCE

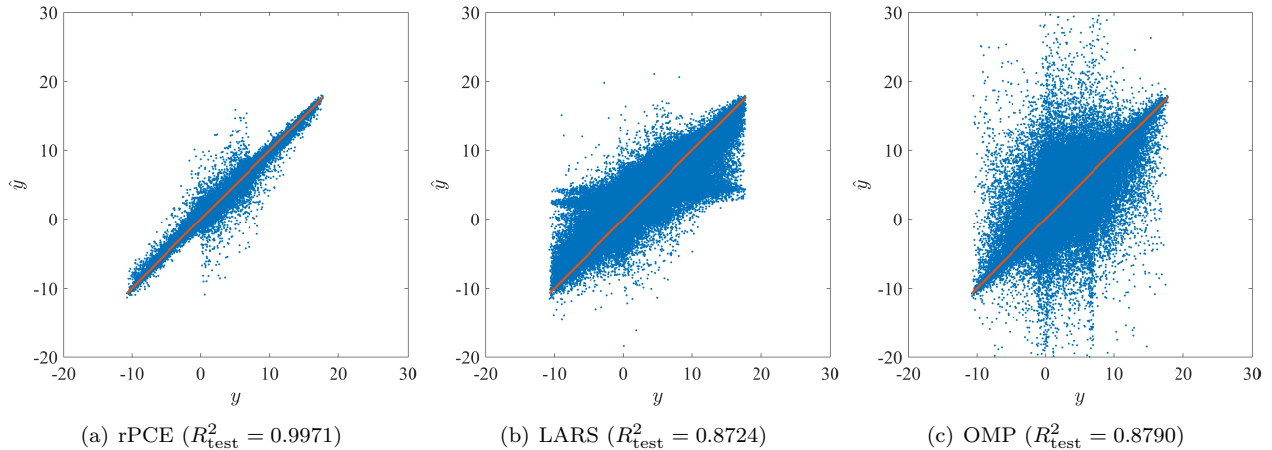


Fig. 3: Ishigami function - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 50 data points (100 replications).

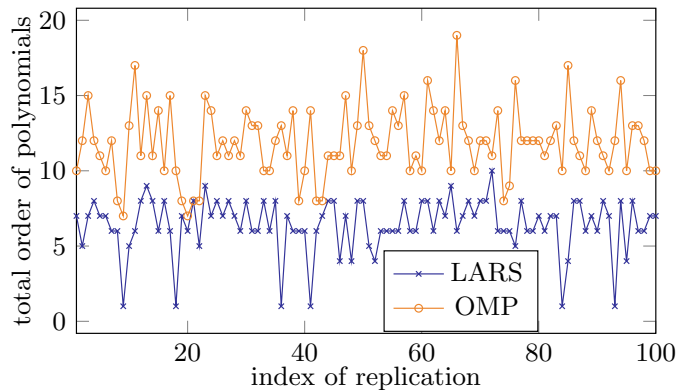


Fig. 4: Ishigami function - optimal total order of polynomials selected by LARS and OMP in 100 replications.

281 and OMP provide unbiased estimations of the Ishigami function, OMP suffers from more outliers and a  
 282 higher variance. LARS tends to have larger predictions (relative to the true values) when  $y < 0$  and smaller  
 283 predictions when  $y > 8$ . Meanwhile, the prediction variance of LARS is not as small as rPCE.

284 The reason for different performances of surrogate modeling based on LARS and OMP may be seen from  
 285 Fig. 4. The PCE models are with higher orders when constructed based on OMP than based on LARS.  
 286 The larger value of the total order  $p$  leads to a bigger polynomial basis and thus a more flexible surrogate  
 287 model, which tends to have less biased but high-variance predictions.

288 As mentioned in Section 5, statistical uncertainty is emulated via the  $k$ -fold division in rPCE and the  
 289 value of  $k$  matters. The suggested configuration of rPCE is combining the polynomial-selection results with  
 290  $k = \{3, 5, 10, 20, N\}$ . To show the effects of  $k$ ,  $R^2_{\text{test}}$  is computed at each replication and 100 values of  $R^2_{\text{test}}$   
 291 yield the box plots of Fig. 5, where  $k = 1$  indicates the surrogate modeling with the whole set of training  
 292 data but without the refinement by rPCE and “all  $k$ ” denotes the rPCE results by combining results with  
 293 different values of  $k$ . As observed, when  $k = 1$ , although the interquartile range (IQR), i.e., the span between  
 294 the first quartile to the third quartile, of LARS is larger than that of OMP, more outliers appear with OMP  
 295 and the minimum  $R^2_{\text{test}}$  is even smaller than  $-1.5$ . With rPCE, except the case of  $k = 3$ , improvements can  
 296 be observed from the reduced outliers and/or prediction variance. The combination of LARS and OMP,  
 297 denoted by “LARS+OMP” (see Section 4.2), seems to have advantages over the rPCE based on LARS or  
 298 OMP and the advantages are more obvious with cases  $k = 3$  and 5.

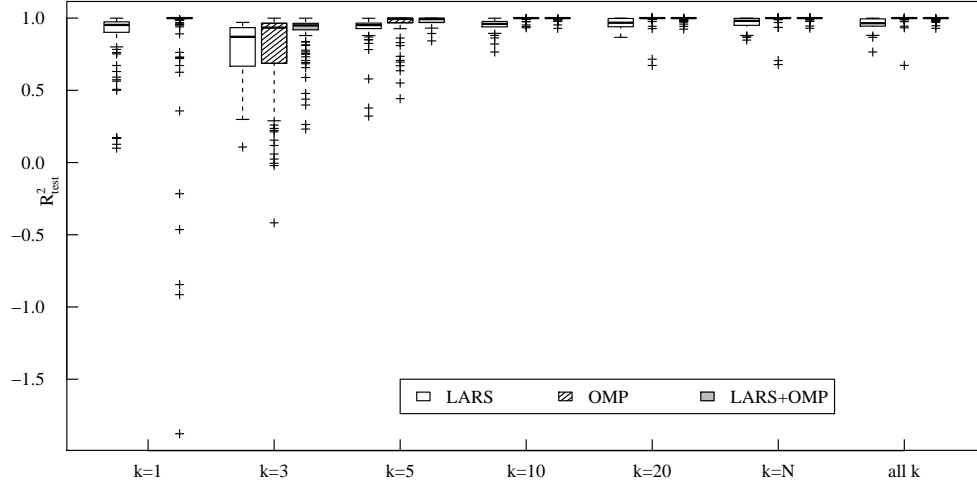


Fig. 5: Ishigami function - box plots of  $R_{\text{test}}^2$  using different values of  $k$  in  $k$ -fold division with 50 data points (100 replications).

Table 2: Ishigami function - mean of  $R_{\text{test}}^2$  over 100 replications with 50 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.8723	0.8788	
$k = 3$	0.7890	0.7734	0.8935
$k = 5$	0.9281	0.9566	0.9817
$k = 10$	0.9542	0.9972	0.9974
$k = 20$	0.9630	0.9919	0.9969
$k = N$	0.9686	0.9918	0.9978
all $k$	0.9619	0.9947	0.9971

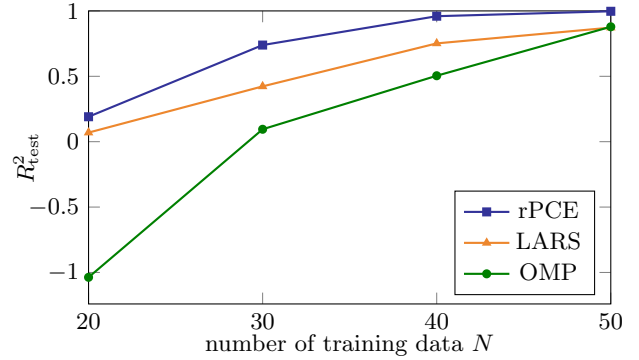


Fig. 6: Ishigami function - mean of  $R_{\text{test}}^2$  versus different values of  $N$  (100 replications).

299 As quantitative comparisons, Table 2 gives the mean of  $R_{\text{test}}^2$  over 100 replications. Generally, OMP is  
300 better than LARS. However, the advantage of OMP is not large and, as a result, the combination of LARS  
301 and OMP in rPCE generates better surrogate models. Remark that the means in Table 2 are obtained by  
302 fixing the value of  $k$  and the source of candidate polynomials (LARS, OMP, or LARS+OMP) during all  
303 replications. Selecting the “all  $k$ ” option and optimizing the polynomial source at each replication with the  
304 suggested configuration in Section 5, the obtained mean of  $R_{\text{test}}^2$  equals 0.9972, only  $6 \times 10^{-4}$  smaller than  
305 the highest value when  $k = N$  with LARS+OMP.

306 Simulations with  $N = 20, 30, 40$  are also operated with the same configurations and the means of  $R_{\text{test}}^2$   
307 are plotted as the line graph in Fig. 6, which shows the better performance of rPCE compared to LARS  
308 and OMP in the cases with small EDs.

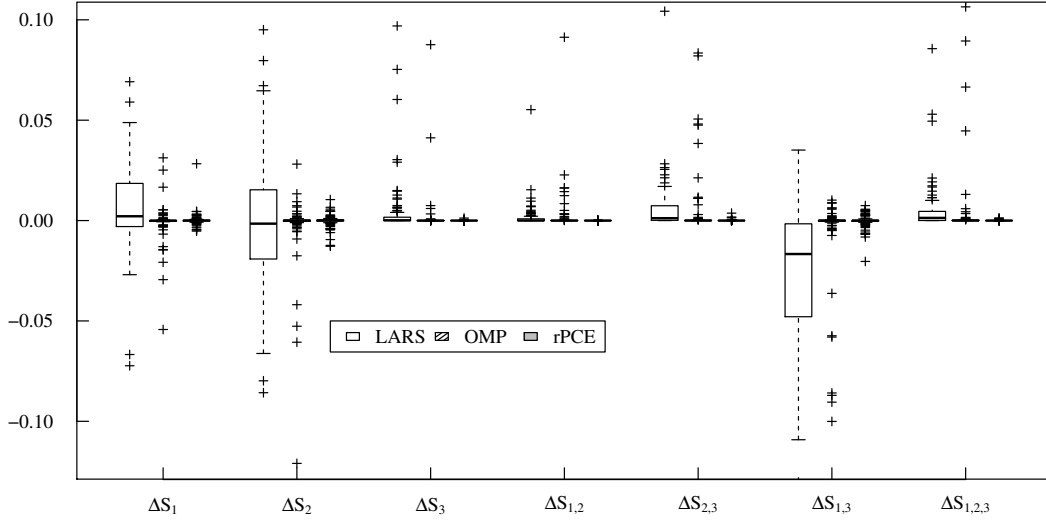


Fig. 7: Ishigami function - the estimation error of Sobol' indices with 50 data points (100 replications).

	Reference	rPCE	LARS	OMP
$S_1$	0.3139	0.3141	0.3553	0.3017
$S_2$	0.4424	0.4422	0.4152	0.4239
$S_3$	0.0000	0.0000	0.0114	0.0028
$S_{1,2}$	0.0000	0.0000	0.0017	0.0052
$S_{2,3}$	0.0000	0.0001	0.0096	0.0042
$S_{1,3}$	0.2437	0.2435	0.2019	0.2363
$S_{1,2,3}$	0.0000	0.0001	0.0049	0.0258

Table 3: Ishigami function - mean of Sobol' indices 50 data points (100 replications).

The Sobol' sensitivity indices can be analytically computed according to

$$\begin{aligned}
 D &= \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{18} + \frac{1}{2}, \\
 D_1 &= \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50} + \frac{1}{2}, \\
 D_2 &= \frac{a^2}{8}, \quad D_{1,3} = \frac{8b^2\pi^8}{225}, \\
 D_3 &= D_{1,2} = D_{2,3} = D_{1,2,3} = 0.
 \end{aligned} \tag{18}$$

Taking the analytical solution as the reference, the estimation error of the Sobol' indices by the PCE-based surrogate model is computed by

$$\Delta S_i = S_i^{\text{PCE}} - S_i^{\text{ref}}, \tag{19}$$

309 where the superscripts of  $S$  indicate the generation approach. With  $N = 50$  and 100 replications, the box  
 310 plots of all  $\Delta S_i$  are shown in Fig. 7, where only values between  $-0.12$  and  $0.1$  are presented for a better view  
 311 and several outliers are absent. The variance of  $\Delta S_i$  is relatively large with LARS when the Sobol' indices  
 312 are non zero, i.e.,  $\Delta S_1$ ,  $\Delta S_2$ ,  $\Delta S_{1,3}$ , and the outliers are efficiently avoided by rPCE. The mean of  $S_i$  is  
 313 given by Table 3, from which the superiority of rPCE in the sensitivity analysis of the Ishigami function is  
 314 obviously observed. The accuracy of rPCE for estimating Sobol' indices is in the order of  $10^{-4}$  when using  
 315 50 data points in the experimental design.

Name	Distribution	Bounds	Description
$r_w$ (m)	$\mathcal{N}(0.10, 0.0161812)$	[0.05, 0.15]	radius of borehole
$r$ (m)	Lognormal(7.71, 1.0056)	[100, 50000]	radius of influence
$T_u$ (m <sup>2</sup> /yr)	Uniform	[63070, 115600]	transmissivity of upper aquifer
$H_u$ (m)	Uniform	[990, 1110]	potentiometric head of upper aquifer
$T_l$ (m <sup>2</sup> /yr)	Uniform	[63.1, 116]	transmissivity of lower aquifer
$H_l$ (m)	Uniform	[700, 820]	potentiometric head of lower aquifer
$L$ (m)	Uniform	[1120, 1680]	length of borehole
$K_w$ (m/yr)	Uniform	[1500, 15000]	hydraulic conductivity of borehole

Table 4: Borehole function - description and distribution of input variables [56].

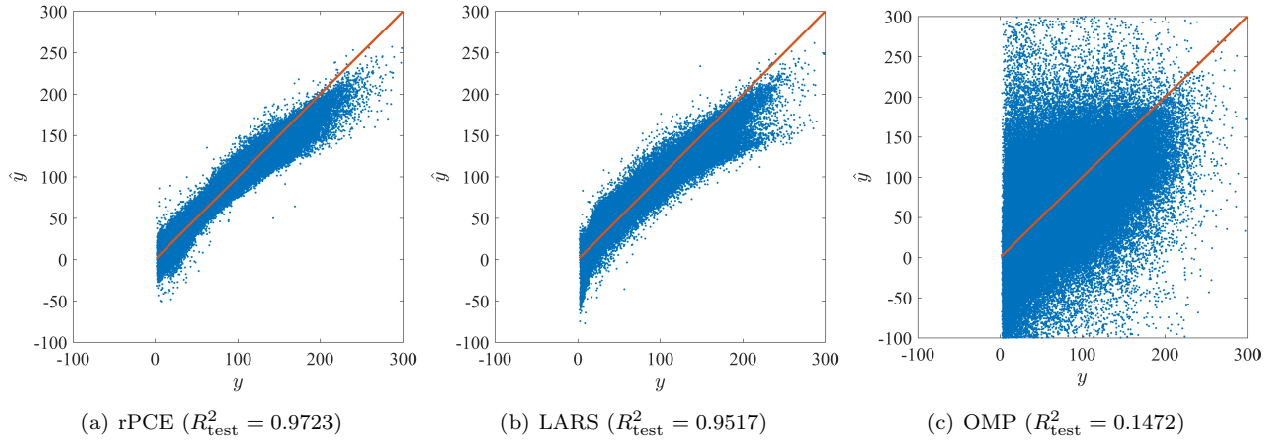


Fig. 8: Borehole function - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 40 data points (100 replications).

The Borehole function with expression

$$Y = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w) (1 + T_u/T_l) + 2LT_u/r_w^2 K_w} \quad (20)$$

models the water flow through a borehole and is a benchmark for emulation and prediction tests. This function has 8 independent variables, the description and distribution of which are presented in Table 4, where the range of  $k_w$  is set as [1 500, 15 000], rather than the usual [9 855, 12 045], to make this function more nonlinear and non-additive. For the construction of multivariate polynomials, Hermite polynomials are used for  $r_w$  and  $r$  (after an isoprobabilistic transformation into a standard normal variable) whereas Legendre polynomials are used for the other variables.

Making use of 40 data points in the model training and  $10^4$  points for validation at each replication, the prediction of validation data obtained from 100 replications is shown in Fig. 8. As seen, the PCE models constructed by the three methods are unbiased approximations of the Borehole function when  $y < 150$ . The underestimation when  $y > 150$  is due to the small portion (1.63 percents for all replications) of data in this range. In prediction variance, rPCE is much better than OMP and slightly superior to LARS. The latter may be explained by observing the box plots in Fig. 9.

The poor performance with OMP may be explained from the procedures for constructing sparse PCE models in Table 1. One sees that leave-one-out cross validation error  $\epsilon_{\text{LOO}}$  is used for tuning two hyperparameters, the total order  $p$  and the number of included polynomials  $J$ .  $p$  is decided by an early-stopping

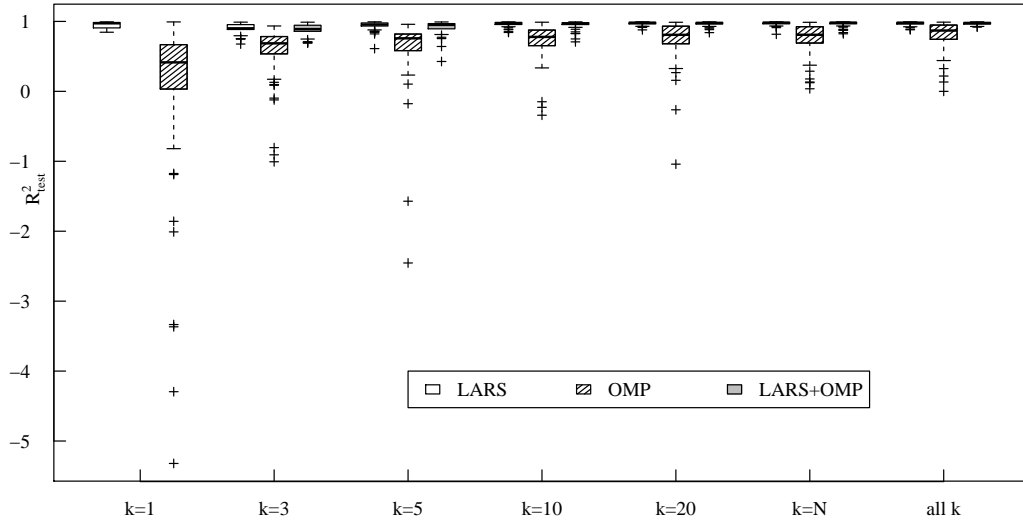


Fig. 9: Borehole function - box plots of  $R^2_{\text{test}}$  using different values of  $k$  with 40 data points (100 replications).

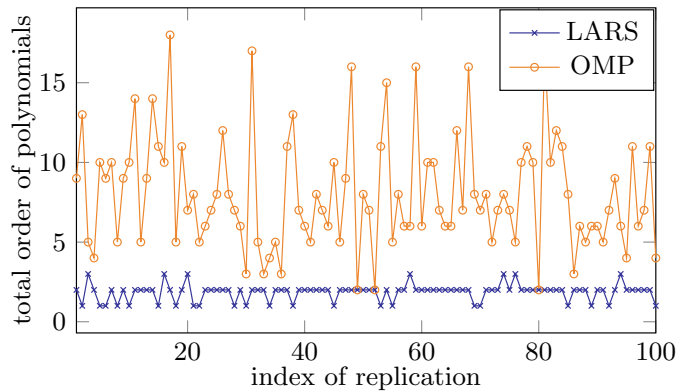


Fig. 10: Borehole function - optimal total order of polynomials selected by LARS and OMP in 100 replications.

332 strategy based on the minimum value of  $\epsilon_{\text{LOO}}$ , i.e.,  $\epsilon_{\text{LOO}}^{\min}$ , which is obtained during the tuning of  $J$ . Since the  
 333 estimation of leave-one-out cross validation is with a high variance [47] (compared with e.g., 10-fold cross-  
 334 validation) and small values of  $\epsilon_{\text{LOO}}$  are preferred, the resulted  $\epsilon_{\text{LOO}}^{\min}$  tends to be optimistic for the model  
 335 assessment. However, the optimistic  $\epsilon_{\text{LOO}}^{\min}$  is then used to tune the value of  $p$ . The biased model-assessment  
 336 by  $\epsilon_{\text{LOO}}^{\min}$  may lead to an improper value for  $p$ . With OMP, obviously the optimal value for  $p$  is overvalued  
 337 as shown by Fig. 10 and thus the prediction is with a high variance as already seen in Fig. 8 and 9.

338 Building the PCE model with the whole set of data, i.e.,  $k = 1$ , the third quartile of  $R^2_{\text{test}}$  with LARS is  
 339 obviously larger than the first quartile with OMP. As explained in Section 5.2, the candidate polynomials will  
 340 be generated by LARS in rPCE, rather than OMP and LARS+OMP, and the results in Fig. 9 provide good  
 341 arguments for this strategy. As seen, the performance of OMP is remarkably improved after the refinement  
 342 by rPCE. However, no matter the value of  $k$ , OMP is still the worst polynomial selection scheme for rPCE  
 343 and LARS seems to be the best option, except that LARS+OMP is slightly better than LARS when taking  
 344 the “all  $k$ ” option.

345 Similar phenomena maybe more clearly observed from Table 5. The mean of  $R^2_{\text{test}}$  with LARS is signif-  
 346 icantly larger than the one with OMP and consequently the rPCE based on LARS is preferred. Applying  
 347 this strategy and automatically selecting the candidate-polynomial source at each replication, the obtained  
 348 mean of  $R^2_{\text{test}} = 0.9724$ . Fig. 11 provides more results when  $N \in \{20, 30, 40, 50\}$ . Since OMP has been



Table 5: Borehole function - mean of  $R_{\text{test}}^2$  with 40 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.9517	0.1467	
$k = 3$	0.9072	0.5852	0.8859
$k = 5$	0.9451	0.6434	0.9239
$k = 10$	0.9673	0.7293	0.9587
$k = 20$	0.9736	0.7506	0.9704
$k = N$	0.9743	0.7633	0.9697
all $k$	0.9719	0.8112	0.9723

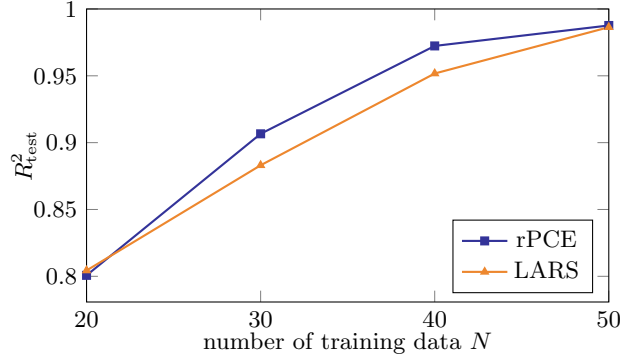


Fig. 11: Borehole function - mean of  $R_{\text{test}}^2$  versus different values of  $N$  (100 replications).

349 shown much worse than rPCE and LARS, its associated line graph is not displayed for a clear view of the  
 350 comparison between LARS and rPCE. The improvements are reached with rPCE in general except for the  
 351 case of  $N = 20$ .

	Reference	rPCE	LARS	OMP
$r_w$	0.3127	0.3072	0.2962	0.4127
$r$	0.0000	0.0010	0.0023	0.1967
$T_u$	0.0000	0.0010	0.0015	0.1635
$H_u$	0.0487	0.0418	0.0420	0.1995
$T_l$	0.0000	0.0011	0.0018	0.1802
$H_l$	0.0487	0.0431	0.0427	0.1751
$L$	0.0472	0.0423	0.0427	0.2026
$K_w$	0.6369	0.6376	0.6322	0.6259
$\Sigma$	1.0942	1.0751	1.0614	2.1562

Table 6: Borehole function - mean of the total Sobol' indices with 40 data points (100 replications).

352 Global sensitivity analysis is then considered and the total Sobol' indices are computed from the various  
 353 PC expansions. The reference values are obtained by the Monte Carlo method with  $10^7$  data and presented  
 354 in Table 6. The importance of variables  $r$ ,  $T_u$  and  $T_l$  can be neglected and the response uncertainty mainly  
 355 comes from the variation of  $r_w$  and  $K_w$ . The same conclusions can be drawn from the estimation results by  
 356 rPCE and LARS. The summation of reference values is close to 1, which indicates weak variable interactions.  
 357 However, the estimation by OMP leads to the opposite conclusion. The stochastic property of the estimation  
 358 deviation  $\Delta S^T$  is revealed by Fig. 12. The estimation variance by OMP is large, especially when the true  
 359 value of  $S^T$  is small, and rPCE outperforms LARS in terms of the estimation variance and the control of  
 360 outliers.

#### 361 6.4. Maximum deflection of a truss structure

362 In Fig. 13, six vertical loads denoted by  $P_1 \sim P_6$  are put on a truss structure composed of 23 bars,  
 363 the cross-sectional area and Young's modulus of which are respectively denoted by  $A$  and  $E$ , the subscripts  
 364 "h" and "o" standing for the horizontal and oblique bars. The response quantity of interest, the mid-span  
 365 deflection  $V$ , is computed with the finite-element method.

366 To analyze the uncertainty of the response, the input parameters are modeled by ten independent random  
 367 variables following the distributions in Table 7. Transforming the input variables into standard normal ones

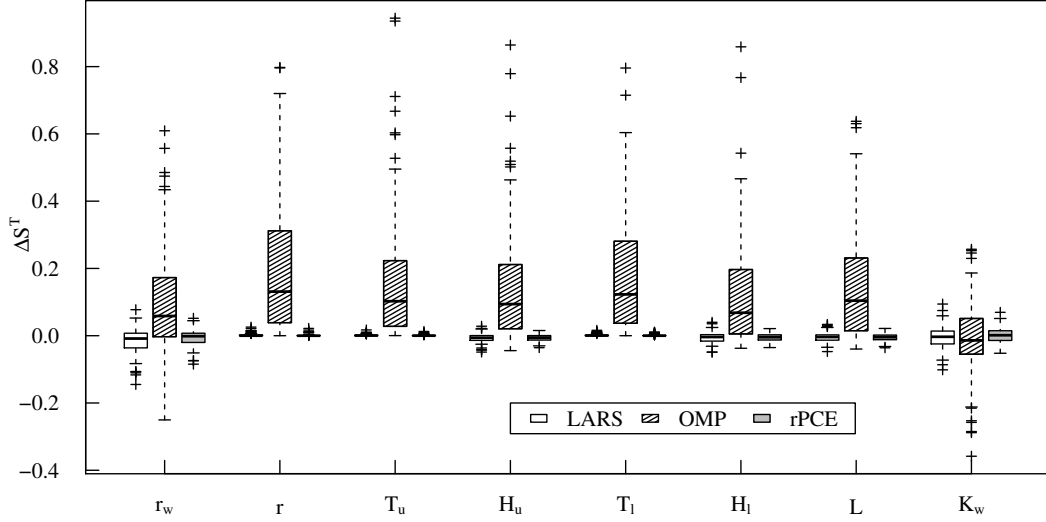


Fig. 12: Borehole function - the estimation error of total Sobol' indices with 40 data points (100 replications).

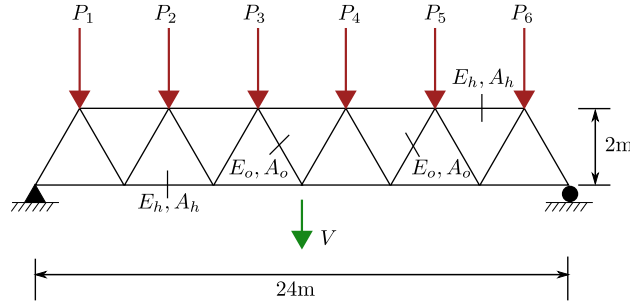


Fig. 13: Sketch of a truss structure made of 23 bars [20].

Variable	Distribution	Mean	Std	Description
$E_h, E_o$ (Pa)	Lognormal	$2.1 \times 10^{11}$	$2.1 \times 10^{10}$	Young's moduli
$A_h$ (m <sup>2</sup> )	Lognormal	$2.0 \times 10^{-3}$	$2.0 \times 10^{-4}$	cross-section area of horizontal bars
$A_o$ (m <sup>2</sup> )	Lognormal	$1.0 \times 10^{-3}$	$1.0 \times 10^{-4}$	cross-section area of oblique bars
$P_1 \sim P_6$ (N)	Gumbel	$5.0 \times 10^4$	$7.5 \times 10^3$	vertical loads

Table 7: Truss deflection - description and distribution of input variables [20].

368 with the isoprobabilistic transformation, LARS, OMP and rPCE surrogate models are built with basis  
 369 composed of Hermite polynomials.

370 With  $N = 50$  and  $10^4$  data for validation at each replication, Fig. 14 shows the prediction results by the  
 371 surrogate models over 100 replications and the solid line indicates the true values of  $V$ . OMP definitely fails  
 372 in this scenario. Although the predictions are unbiased, the variance is high due to the too much flexibility  
 373 of the PCE model built by OMP. In contrast, LARS and rPCE achieve a much better trade-off between  
 374 the variance and bias. Moreover, rPCE is slightly superior to LARS in variance and the number of outliers.  
 375 The poor prediction performance when  $V < -0.11$  is a consequence of a small portion (0.78 percent for all  
 376 replications) of data in this range.

377 Based on the validation data,  $R_{\text{test}}^2$  is computed at each replication and the distribution of  $R_{\text{test}}^2$  over 100  
 378 replications is given in Fig. 15. The results with  $k = 1$  indicate the running of LARS and OMP with the  
 379 whole set of data, thus no refinement of the basis by rPCE and “all  $k$ ” means that rPCE is run based on

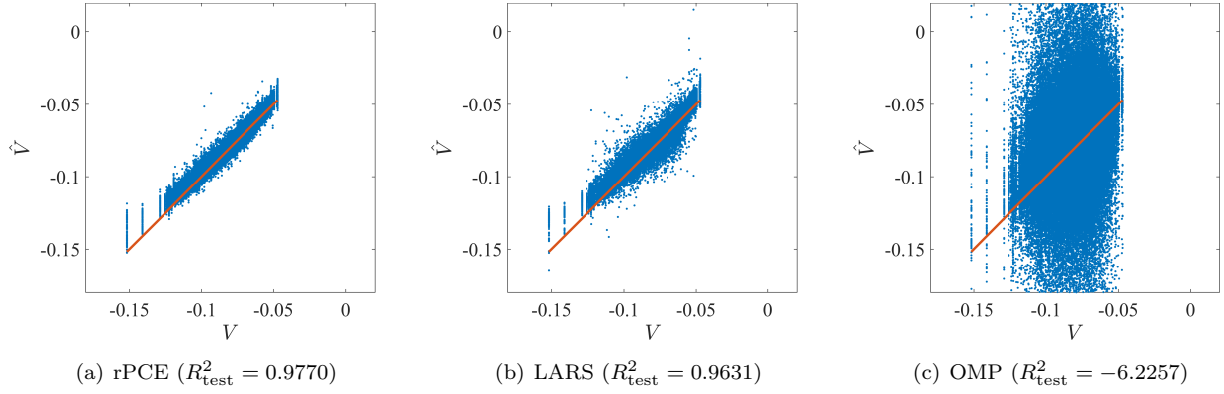


Fig. 14: Truss deflection - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 50 data points (100 replications).

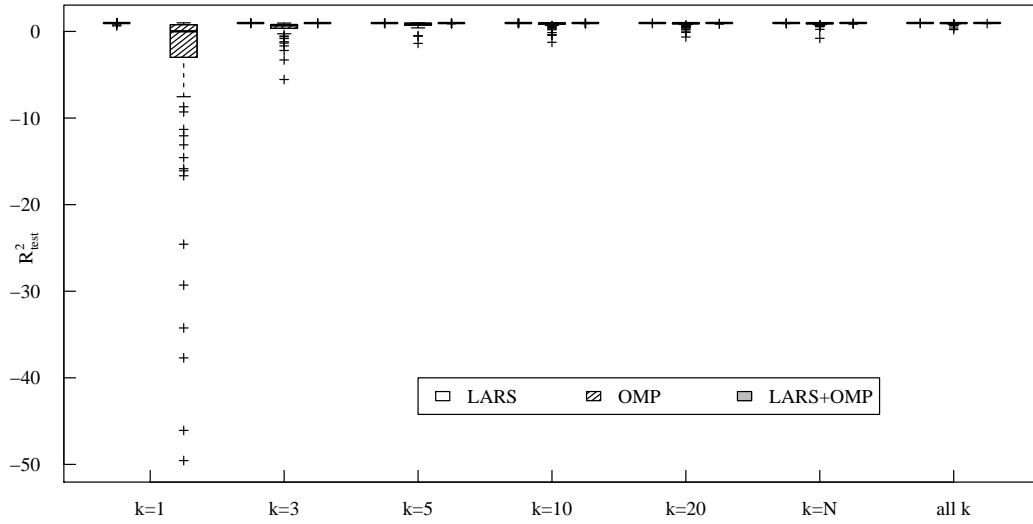


Fig. 15: Truss deflection - box plots of  $R^2_{\text{test}}$  using different values of  $k$  with 50 data points (100 replications).

Table 8: Truss deflection - mean of  $R^2_{\text{test}}$  with 50 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.9631	-6.2248	
$k = 3$	0.9651	0.3873	0.9641
$k = 5$	0.9658	0.7915	0.9660
$k = 10$	0.9692	0.8273	0.9693
$k = 20$	0.9726	0.8721	0.9735
$k = N$	0.9735	0.8974	0.9741
all $k$	0.9744	0.9315	0.9762

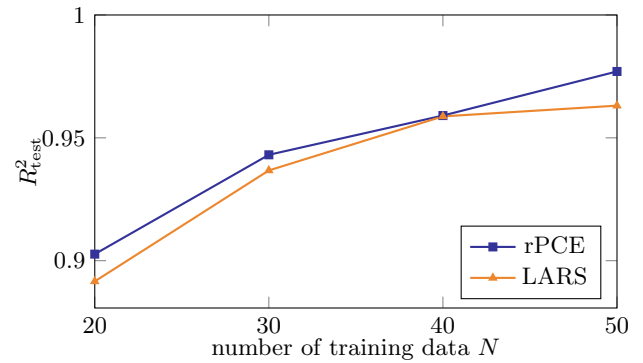


Fig. 16: Truss deflection - mean of  $R^2_{\text{test}}$  versus different values of  $N$  (100 replications).

380 the combination of candidate polynomials generated with  $k = [3, 5, 10, 20, N]$ . Although the performance  
381 of OMP is much enhanced with the application of rPCE, LARS is still better than OMP, whatever the  
382 value of  $k$ . The rPCE model combining LARS and OMP seems to have the same performance with the  
383 rPCE model based on LARS itself. Table 8 presents the associated mean of  $R_{\text{test}}^2$ . As seen, the highest  
384 mean appears with LARS+OMP when all  $k$  values are considered, but, with the same configurations, the  
385 difference between LARS and LARS+OMP is only 0.0018. Optimizing the selection of candidate polynomials  
386 at each replication, as displayed in Fig. 16, the mean value reaches 0.9770 for the “all  $k$ ” option. The slight  
387 superiority of rPCE to LARS is also seen with  $N = 20, 30, 40$ .

	Reference	rPCE	LARS	OMP
$E_h$	0.367	0.3713	0.3748	0.4295
$E_o$	0.010	0.0121	0.0135	0.2290
$A_h$	0.388	0.3695	0.3715	0.4037
$A_o$	0.014	0.0127	0.0135	0.2291
$P_1$	0.004	0.0046	0.0057	0.2105
$P_2$	0.031	0.0359	0.0365	0.2251
$P_3$	0.075	0.0750	0.0759	0.2808
$P_4$	0.079	0.0756	0.0751	0.2557
$P_5$	0.035	0.0355	0.0361	0.2271
$P_6$	0.005	0.0048	0.0061	0.1891
$\Sigma$	1.008	0.9969	1.0086	2.6795

Table 9: Truss deflection - mean of the total Sobol’ indices with 50 data points (100 replications).

388 Global sensitivity analysis is conducted by computing the total Sobol’ indices based on the PCE coef-  
389 ficients. The reference values listed in Table 9 are obtained with  $5.5 \times 10^6$  Monte Carlo simulations [20].  
390 Since the characteristics of the horizontal bars impact more the displacement at midspan than the oblique  
391 ones, the total Sobol’ indices of  $E_h$  and  $A_h$  are much larger than those of  $E_o$  and  $A_o$ . Moreover, due to the  
392 same type of probabilistic distribution and the fact that the products  $E_h A_h$  (resp.  $E_o A_o$ ) are the physically  
393 meaningful quantities in the analysis,  $E_h$  and  $A_h$  (resp.  $E_o$  and  $A_o$ ) have similar importance to the response.  
394 Considering the variables of  $P_i$ ,  $i = 1, \dots, 6$ ,  $P_i$  and  $P_{7-i}$  play the same role due to the geometric symmetry  
395 of the structure and greater sensitivities are observed for loads closer to the midspan. The above conclusions  
396 are clearly supported by the estimations of rPCE and LARS. In contrast, the largely biased estimation by  
397 OMP might give a wrong understanding of the physical phenomena. For instance, one may falsely conclude  
398 that the actually negligible interactions among inputs have great effects on the midspan deflection, since the  
399 sum of the total Sobol’ indices obtained by OMP is much larger than 1.

400 The distribution of the prediction error of total Sobol’ indices  $\Delta S^T$  is given in Fig. 17. In addition to  
401 the largely biased and scattered OMP, rPCE and LARS has similar  $\Delta S^T$  distribution with relatively small  
402 variances.

### 403 6.5. Estimation of specific absorption rate

404 The population is surrounded by a increasing number of wireless local area networks (WLAN) and the  
405 electromagnetic exposure of human body by WLAN access points needs to be estimated to make sure the  
406 exposure level is under the limit [57]. Here, an indoor down-link scenario is considered, as sketched in  
407 Fig. 18. A high-resolution model of a 8-year girl (1.36 m high), named as “Eartha”, from the Virtual  
408 Classroom [58], is standing inside a  $4 \times 3 \times 2$  m<sup>3</sup> room, which is equipped with a WLAN source operating  
409 at 2.4 GHz. The field emitted by the source is measured using the StarLab near-field-measurement system,  
410 which is based on spherical wave expansion [59], by Microwave Vision Group (MVG<sup>®</sup>). With an in-house  
411 finite-difference-time-domain (FDTD) code, the whole-body specific absorption rate (SAR) [60], which is  
412 the system response here, is computed as the ratio of the total power absorbed in the body to the mass of  
413 the human model and with the unit mW/kg.

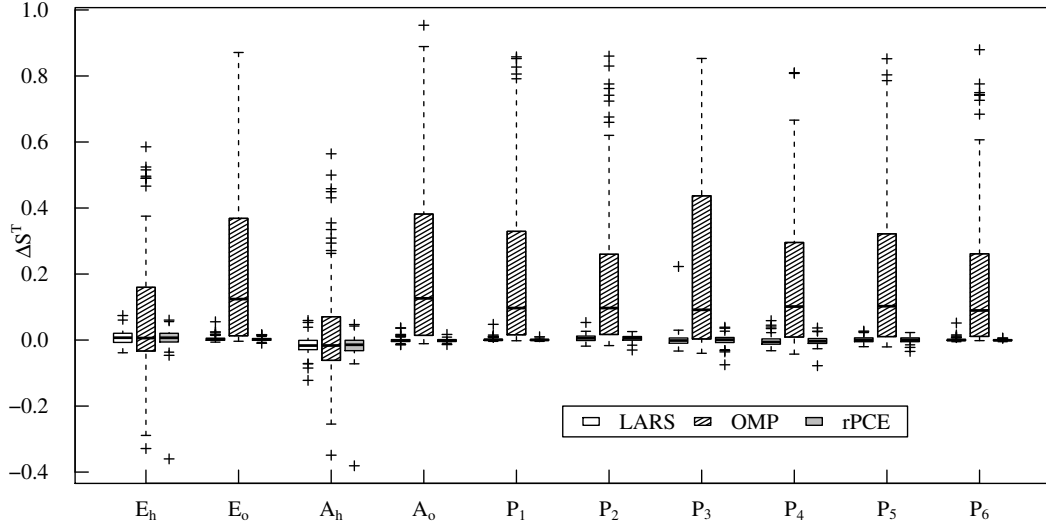


Fig. 17: Truss deflection - the estimation error of total Sobol' indices with 50 data points (100 replications).

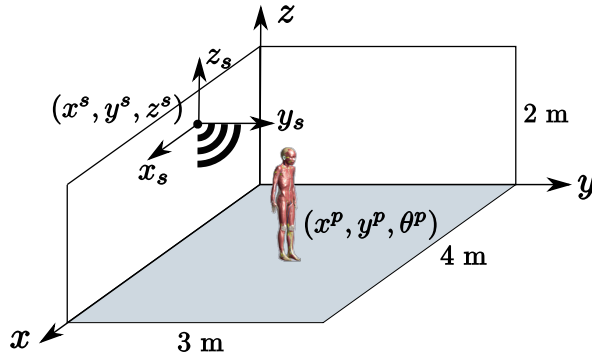


Fig. 18: Sketch of the human-exposure estimation in an indoor down-link scenario.

414 The parameters considered are the position of the emitting source and the human model, whose coordi-  
 415 nates are denoted by  $(x^s, y^s, z^s)$  and  $(x^p, y^p, z^p)$ , respectively.  $z^p$  is set as 0, since we consider that the  
 416 human model is standing on the ground. The human orientation  $\theta^p$ , which is defined as the angle between  
 417 the direction faced by the human model and  $x$ -axis, may matter and is taken into account. The reflection by  
 418 the walls, ceiling and ground is neglected in the simulation and the WLAN source is attached to the walls.  
 419 Thus, six parameters are involved.  $x^s, y^s, z^s, x^p, y^p$  are assumed to be uniformly distributed over  $[0.3, 3.7]$ ,  
 420  $[0.3, 2.7]$ ,  $[0.25, 2]$ ,  $[0.05, 3.95]$ ,  $[0.05, 2.95]$  in meters and  $\theta^p$  over  $[0, 360]$  in degrees, where the lower bound  
 421 value 0.3 m is the minimum distance between the human model and the wall, 0.25 m is the minimum height  
 422 of the source and 0.05 m is the minimum distance of the WLAN source to the wall.

423 The number of input variables can be reduced via a coordinate transformation. Without the reflection  
 424 by the walls, the system response is actually driven by the relative position between the source and the  
 425 human model. The relative position is represented in the  $(x, y)$  plane. In the local coordinate system of the  
 426 source, as shown in Fig. 18, position and orientation of the human model are denoted by polar coordinates  
 427  $(r_s^p, \phi_s^p)$  and  $\theta_s^p$ . Thus, four parameters  $r_s^p, \phi_s^p, \theta_s^p$ , and  $z^s$  are used in the following uncertainty analysis.

428 Sampling 350 points from the input space with the Latin-Hypercube sampling method, the prediction  
 429 performance of the obtained surrogate models is estimated with the leave-many-out approach, where 10  
 430 data are randomly chosen from the experimental design for validation and an approximation of  $R_{\text{test}}^2$  is  
 431 yielded by repeating this process 100 times. Consequently, with the remaining 340 data, surrogate models

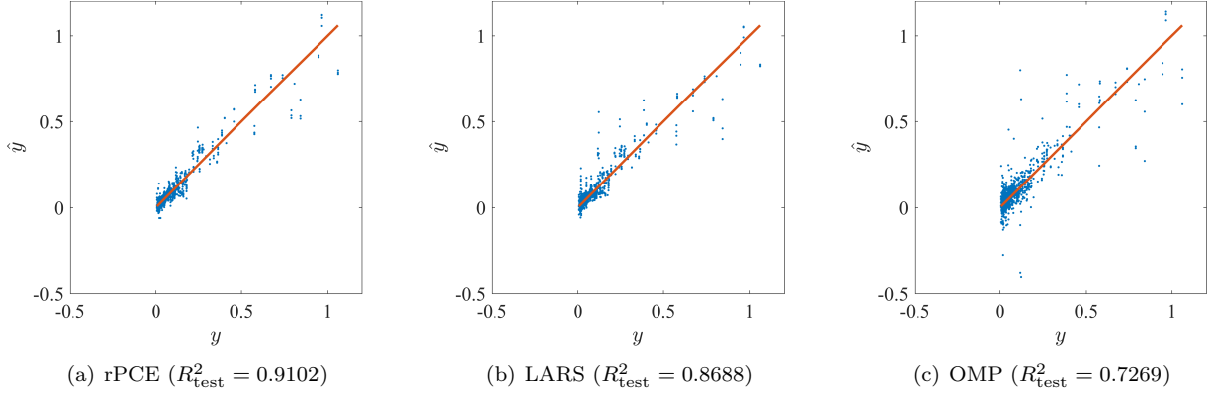


Fig. 19: SAR estimation - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 340 data points (100 replications).

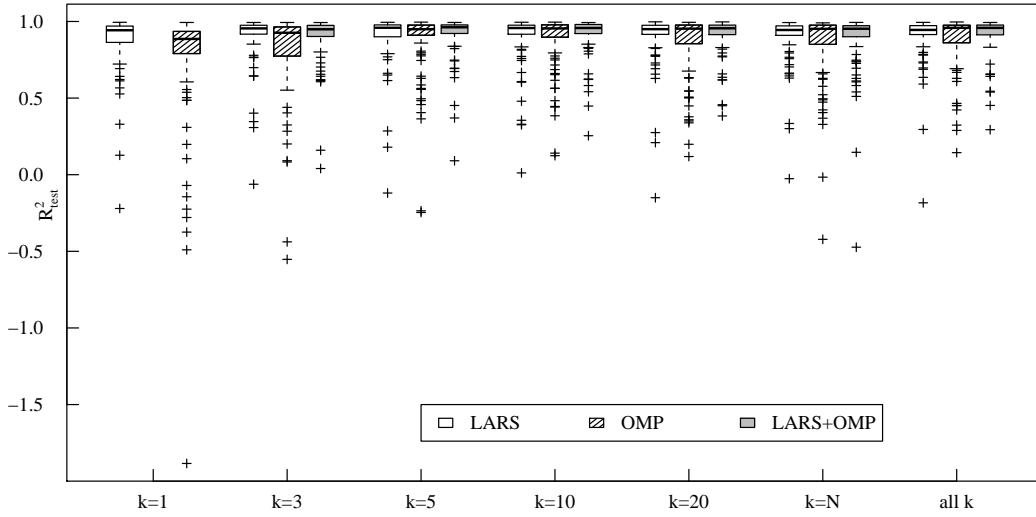


Fig. 20: SAR estimation - box plots of  $R^2_{\text{test}}$  with different values of  $k$  (100 replications).

432 are obtained with LARS, OMP and rPCE. Then, a validation set of size  $10^3$  is computed and the results  
 433 are shown in Fig. 19. As seen, the whole-body SAR is smaller than 0.2 for most of cases (90 percents for  
 434 all replications) in this scenario. However, the three approaches can provide unbiased estimations when the  
 435 SAR value is larger than 0.2, in addition to the the superiority of rPCE to LARS and OMP in variance  
 436 and suppression of outliers. The associated box plots of  $R^2_{\text{test}}$  is given in Fig. 20. The refinement by rPCE  
 437 reduces the variance of modeling by LARS and OMP with different values of  $k$ , except for the case with  
 438 OMP and  $k = 3$ . The combination of LARS and OMP seems to be the best option for rPCE and actually is  
 439 selected by the suggested scheme in Section 4.2 during all replications (although three options are available  
 440 at each replication), since LARS has the same-level performances with OMP. Table 10 shows the mean of  
 441  $R^2_{\text{test}}$ .

442 The total Sobol' indices are computed based on the PCE coefficients and the mean values are presented  
 443 in Table 11. As seen, the whole-body human exposure is mainly impacted by the relative distance  $r_s^p$  and  
 444 the height of the source  $z^s$  has a smaller influence. The small value w.r.t. the relative angle between the  
 445 human model and the source,  $\phi_s^p$ , might be explained by looking at the contours of electric-field intensity in  
 446 Fig. 21, where the WLAN source locates at the center of a wall and field values are sampled in the  $(x_s, y_s)$   
 447 plane with  $z_s = 0$ . As observed, the dependency of wave strength on radiation directions is weak. The

Table 10: SAR estimation - mean of  $R_{\text{test}}^2$  with 340 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.8799	0.7500	
$k = 3$	0.9085	0.8186	0.9046
$k = 5$	0.9067	0.8771	0.9182
$k = 10$	0.8995	0.8854	0.9171
$k = 20$	0.9033	0.8628	0.9157
$k = N$	0.8995	0.8521	0.8893
all $k$	0.9068	0.8794	0.9178

Table 11: SAR estimation - mean of the total Sobol' indices with 340 data points (100 replications).

	rPCE	LARS	OMP
$r_s^p$	0.9809	0.9714	0.9761
$\phi_s^p$	0.0128	0.0357	0.0984
$z^s$	0.2175	0.1954	0.2925
$\theta_s^p$	0.0098	0.0316	0.0743
$\Sigma$	1.2210	1.2341	1.4412

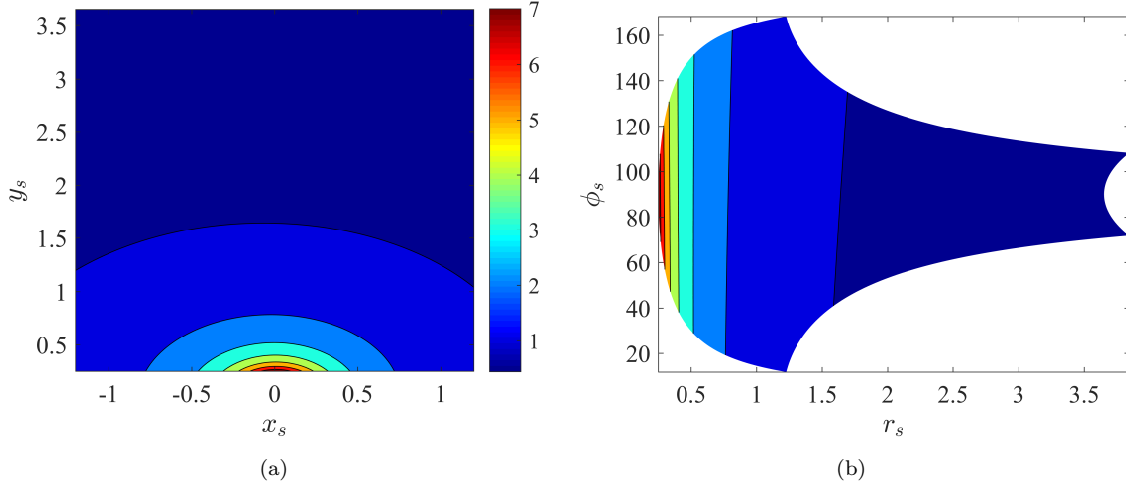


Fig. 21: SAR estimation - contour of electric-field intensity (a) in the  $(x, y)$  plane and (b) its representation in the polar coordinate system,  $z_s = 0$ .

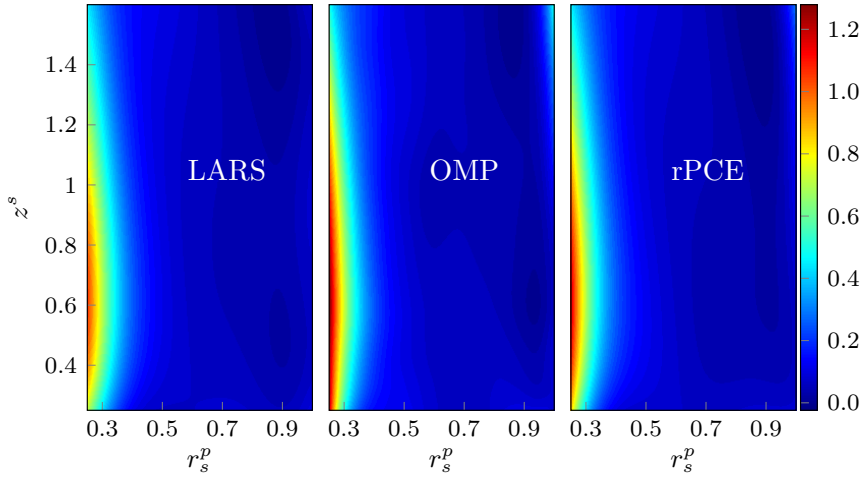


Fig. 22: SAR estimation - the prediction of whole-body SAR with 340 data points (100 replications).

448 human orientation  $\theta_s^p$  affects the distribution of SAR in the human body. However, as the mean value of  
 449 this distribution, the whole-body SAR is not much affected by  $\theta_s^p$ . The sum of the total Sobol' indices in

450 Table 11 is larger than 1 and the excess values indicate that  $z^s$  impacts the response mainly through its  
451 interaction with  $r_s^p$ . Such an interaction can be viewed from the map of predicted SAR in Fig. 22, where  
452  $\phi_s^p, \theta_s^p$  are fixed to zero and  $r_s^p, z^s$  are uniformly sampled over  $[0.25, 1], [0.25, 2]$ , respectively. The amplitude  
453 of each pixel in the map is a mean of 100 predictions by the built PCE models during all replications. The  
454 three approaches provide similar results.

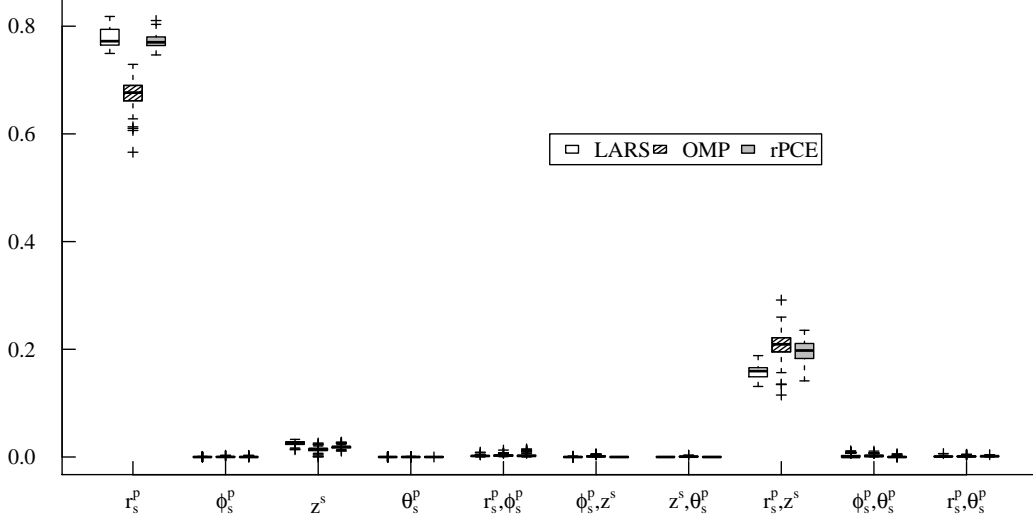


Fig. 23: SAR estimation - the estimation of first-order and second-order Sobol' indices with 340 data points (100 replications).

455 Considering the height of the human model is 1.36 m, tissues mainly locate at the heights between 0.4 m  
456  $\leq z^s \leq 1.0$  m. One observes that the whole-body SAR is rather small when the source is farther from this  
457 influential region of the human model.  $r_s^p$  and  $z^s$  model the distance between the source and this influential  
458 region together and their interactions happen. The distribution of the estimated first-order and second-order  
459 Sobol' indices is proposed in Fig. 23, which presents that  $r_s^p$  and its interaction with  $z^s$  contribute the most  
460 to the uncertainty of the response.

#### 461 6.6. Example with varied input dimension

To investigate the effects of the dimension of input on the modeling performance, the following test  
function [53] is used,

$$y = 3 + \frac{1}{M} \sum_{k=1}^M k(x_k^3 - 5x_k) + \ln \left( \frac{1}{3M} \sum_{k=1}^M k(x_k^2 + x_k^4) \right) + x_1x_2^2 - x_3x_5 + x_2x_4 + x_{M-4} + x_{M-4}x_M^2, \quad (21)$$

462 where  $M$  denotes the number of variables, which are independent and uniformly distributed in the range of  
463  $[1, 2]$ . To increase the non-linearity, the range of  $x_{20}$  (when  $M \geq 20$ ) is changed as  $[1, 3]$ .

464 The value of  $M$  changes from 11 to 41 with a step 5 and the size of experimental design  $N$  is fixed as 200  
465 independently of  $M$ . For a statistical assessment of the modeling performance (accuracy and efficiency), 50  
466 replications are performed and  $10^3$  data are used for the independent test at each replication. Remark that,  
467 due to randomness of the LHS method, different training and test datasets are used in replications.

468 From the methodology of rPCE, one knows that the computational cost is proportional to the number  
469 of resampled datasets. In Section 5.1, the suggested (not obliged) setting of  $k$  is a set of values, i.e.,  
470  $k = \{3, 5, 10, 20, N\}$ . As a result, the corresponding computational cost would be high when the ED size  $N$   
471 is large. Here, a lighter setting of  $k$ ,  $k = \{3, 5, 10, 20\}$ , is applied and improved modeling performances are  
472 still observed as presented by the following results. For the configuration of UQLab, which the running of  
473 LARS and OMP is based on, the maximum value of total degree  $p$  is set as 5.



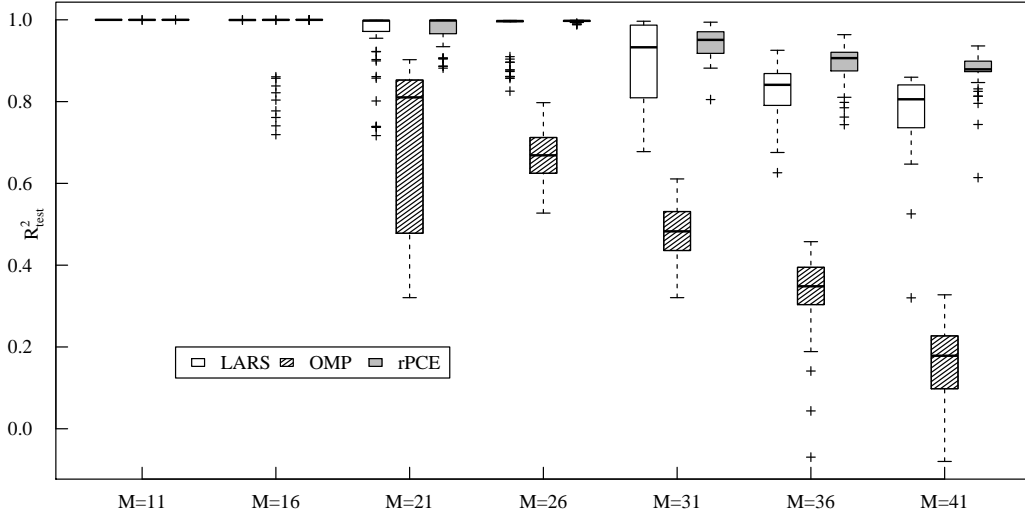


Fig. 24: Example with varied dimension - box plots of  $R^2_{\text{test}}$  with different values of  $M$  (50 replications).

474 The effects of  $M$  on the modeling accuracy can be observed from the distribution of  $R^2_{\text{test}}$  in Fig. 24.  
 475 When  $M$  equals 11 and 16, accurate models are constructed with the three approaches, although outliers  
 476 appear with OMP and  $M = 16$ . As the dimension of input increases, the modeling accuracy becomes  
 477 poorer in both variance and bias. While LARS performs much better than OMP when  $M \geq 21$ , substantial  
 478 advantages of rPCE (with suggested configurations) are observed. When  $M \in \{31, 36, 41\}$ , the mean value  
 479 of  $R^2_{\text{test}}$  with rPCE is larger than the value with LARS and OMP, and the variance of rPCE is also superior  
 480 to the other two approaches.

		M=11	M=16	M=21	M=26	M=31	M=36	M=41
$k = 1$	LARS	0.9998	0.9995	0.9573	0.9679	0.8985	0.8260	0.7761
	OMP	0.9998	0.9634	0.6940	0.6679	0.4832	0.3308	0.1536
$k = 3$	LARS	0.9997	0.9996	0.9422	0.9249	0.8646	0.8322	0.8125
	OMP	0.9998	0.8072	0.7810	0.7737	0.6514	0.5358	0.3870
	L+O	0.9998	0.9996	0.8929	0.8771	0.7810	0.7262	0.6805
$k = 5$	LARS	0.9998	0.9995	0.9600	0.9726	0.8899	0.8574	0.8351
	OMP	0.9999	0.9552	0.8171	0.7915	0.6935	0.5894	0.4826
	L+O	0.9999	0.9996	0.9511	0.9651	0.8630	0.8110	0.7681
$k = 10$	LARS	0.9999	0.9995	0.9714	0.9945	0.9316	0.8724	0.8445
	OMP	0.9999	0.9963	0.8395	0.8194	0.7252	0.6239	0.5340
	L+O	0.9999	0.9998	0.9668	0.9937	0.9210	0.8557	0.8193
$k = 20$	LARS	0.9999	0.9995	0.9824	0.9971	0.9523	0.8947	0.8714
	OMP	0.9999	0.9999	0.8392	0.8195	0.7197	0.6149	0.5165
	L+O	0.9999	0.9999	0.9784	0.9971	0.9404	0.8692	0.8391
all $k$	LARS	0.9999	0.9996	0.9765	0.9965	0.9437	0.8904	0.8725
	OMP	0.9999	0.9987	0.8423	0.8248	0.7316	0.6191	0.5011
	L+O	0.9999	0.9998	0.9738	0.9961	0.9371	0.8790	0.8604

Table 12: Example with varied dimension - mean of  $R^2_{\text{test}}$  with varied values of  $k$  and  $M$  (50 replications), “L+O” denoting the combination of LARS and OMP.

481 From the mean value of  $R^2_{\text{test}}$  (w.r.t. 50 replications) in Table 12, one finds the effects of varied  $M$  on  
 482 rPCE with different configurations. Remark that the cases with  $k = 1$  correspond with the modeling results

483 based on LARS or OMP without the refinement by rPCE and the other cases are results of rPCE with  
484 different configurations. “all  $k$ ” here means the combination of selection results with  $k = \{3, 5, 10, 20\}$ . As  
485 seen, when  $M \geq 21$  and  $k$  is configured as 10, 20, or “all  $k$ ”, significant improvements are observed. When  
486 only OMP is applied, the mean value of  $R_{\text{test}}^2$  increases from 0.6940 to 0.8423 when  $M = 21$ . With LARS,  
487 the mean value increases from 0.7761 to 0.8725 when  $M = 41$ . Remark that these two peak values are  
488 reached with the configuration of “all  $k$ ”. Concerning on the source of candidate polynomials, since LARS  
489 performs much better than OMP, rPCE makes use of the selection results by LARS more often than OMP  
490 or their combination. Consequently, rPCE with the suggested setting (denoted by “L+O” in Table 12), is  
491 slightly (not significantly) inferior to the setting only based on LARS in modeling accuracy.

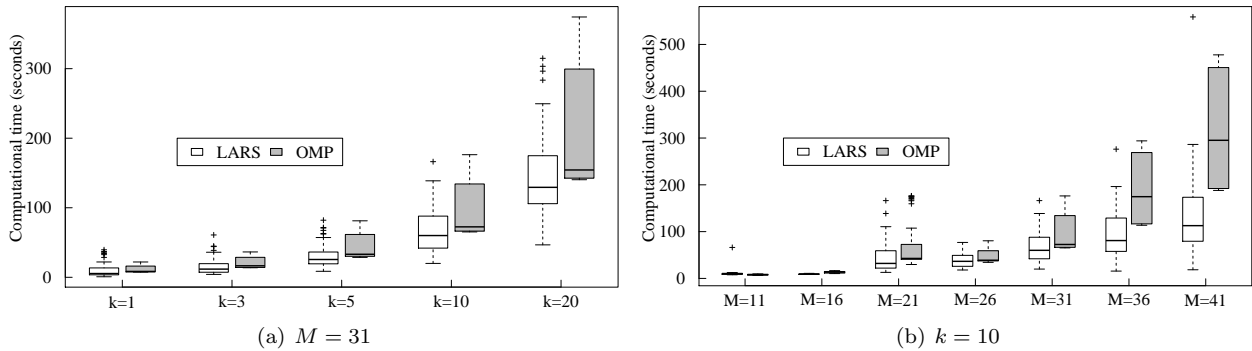


Fig. 25: Example with varied dimension - computational time versus varied values of  $k$  and  $M$  (50 replications).

492 With a laptop (dual cores, clock speeds 2.6 GHz, memory 16 GB), the computational time of rPCE is  
493 shown in Fig. 25. Effects of  $k$  and  $M$  are studied by fixing  $M = 31$  and  $k = 10$ , respectively. Remark that  
494 this figure gives the total time cost for each configuration, i.e., for a specific value of  $k$ , the total time for  
495 constructing  $k$  PCE models is given. As observed, since about two thirds of the dataset is used for model  
496 construction, the computation time with  $k = 3$  is only slightly longer than (rather than three times as) with  
497  $k = 1$ . As the value of  $k$  increases, since the size of training datasets and the number of model constructions  
498 gets larger, the computational time increases fast. Moreover, in general the computational cost with OMP  
499 is higher than that with LARS. Similar phenomena are observed when the value of  $M$  increases. When  
500  $M = 41$ , the maximum time cost on running both LARS and OMP is below 18 minutes which is usually  
501 much shorter than the time cost on getting new samples (e.g.,  $\approx 3$  hours are required for getting a new  
502 sample in the example of Section 6.5).

## 503 7. Conclusions

504 A new polynomial selection approach, called resampled PCE, has been investigated herein to refine the  
505 ranking of importance of candidate polynomials in the context of sparse polynomial chaos expansions. Based  
506 on the selected polynomials by LARS and OMP, with the simulation of data variation by resampling, both  
507 the selection frequency and the increment on cross-validation error associated with each basis polynomial  
508 are arguments in the computation of a total score used in the ranking process. With the PCE model based  
509 on rPCE, sensitivity analysis is conveniently performed via the analytical computation of the Sobol’ indices  
510 based on the expansion coefficients.

511 Two factors impact the performance of rPCE. First, the data resampling is conducted by dividing the  
512 whole set of data into  $k$  similar-sized subsets. The value of  $k$  needs to be optimized and set as a combination  
513 of good candidates  $\{3, 5, 10, 20, N\}$ . Second, the candidate polynomials can be generated by LARS, OMP  
514 or both. If LARS performs much better than OMP, the resulting selection of polynomials is based on  
515 LARS, and vice versa. Otherwise, both the polynomials selected by LARS and OMP would all be treated  
516 as candidates in rPCE.

517 The performance of rPCE, LARS and OMP is tested on two analytical functions, the maximum deflection  
518 of a truss structure and the estimation of the whole-body SAR (specific absorption rate). In terms of  
519 prediction and sensitivity analysis, OMP-based PCE modeling seems the worst among these three methods,  
520 especially when the size of ED is small. In contrast, the LARS-based approach generally generates a better  
521 model and the refinements by rPCE are obvious in terms of prediction variance and the number of outliers.  
522 In any case, rPCE performs as least as well as LARS for global sensitivity analysis.

523 Although the size of ED is fixed here, the samples can be automatically enriched to reach a certain  
524 accuracy in a specific estimation (e.g., moments) [61, 20, 62, 63]. Moreover, since the building processes  
525 with multiple resamples are independent in rPCE, the technique of parallel computations can be applied to  
526 ensure the building efficiency of rPCE at the same level with LARS or OMP.

527 In resampled PCE, a high computational cost may be suffered, especially with the suggested setting  
528 of  $k$ , i.e.,  $k = \{3, 5, 10, 20, N\}$  and a large  $N$ . As shown by the flow chart in Fig. 1, the possible high  
529 computational cost is due to the loop about  $k$  and  $l$ . However, one should realize that the loop about  $k$  and  
530  $l$  are not necessarily performed in sequence and can proceed in parallel. All resampled datasets with different  
531 values of  $k$  can be first easily generated with LHS method. Since the surrogate modeling with respect to  
532 different sets of resamples is separable, techniques of parallel computation (e.g., computation with GPU and  
533 distributed computation) can be applied. Moreover, the suggested setting is not compulsory in the running  
534 of resampled PCE. Actually, from the results in Section 6, we can see the improved modeling performance  
535 has been observed with  $k = 10, 20$ , or  $k = \{3, 5, 10, 20\}$ , even better performances can be obtained if with  
536 the suggested setting. Thus, if the additional computational cost is considered high (especially relative to  
537 the cost of obtaining new data), 10 or 20 would be suggested for the setting of  $k$ .

538 In forthcoming investigations, more complex scenarios (e.g., electromagnetic dosimetry for human models  
539 in the telecommunications network [64, 65, 66]) are to be analyzed, where a high-order PCE model is often  
540 required and the classical approaches easily sink into the overfitting problem. Resampled PCE has the  
541 potential to avoid this problem. The refined selection of polynomials reduces the possibility of including  
542 redundant or irrelevant basis polynomials in the expansion, thus would have better chances to reach a model  
543 with a proper complexity. Here, rPCE combines two forward basis pursuit approaches and the improvements  
544 may be slight due to similar selected polynomials. The combination of different kinds of approaches (e.g.  
545 forward selection, backward elimination [67, 19] and sparsity-based approach [22]) is open to investigation.

## 546 Appendix A. Ranking basis polynomials based on LARS or OMP

- 
1. Initialization: residual  $\mathbf{R}_0 = \mathbf{y}$ , active set  $\mathbb{A}_0^a = \emptyset$ , candidate set  $\mathbb{A}_0^c = \mathbb{A}_{full}$ .
  2. For  $j = 1, \dots, P_{max} = \min\{N - 1, \text{card}(\mathbb{A}_{full})\}$ ,
    - 1) Find the basis most correlated with  $\mathbf{R}_{j-1}$ ,  $\alpha_j = \arg \max_{\alpha \in \mathbb{A}_{j-1}^c} |\mathbf{R}_{j-1}^T \psi_\alpha|$ .
    - 2) Update  $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \alpha_j$  and  $\mathbb{A}_j^c = \mathbb{A}_{j-1}^c \setminus \alpha_j$ .
    - 3) With  $\psi_{\mathbb{A}_j^a}$ , compute  $\beta_j$  as the OLS solution.
    - 4) Update residual  $\mathbf{R}_j = \mathbf{y} - \psi_{\mathbb{A}_j^a}^T \beta_j$ .
- End
- 

Table A.1: Ranking basis polynomials based on orthogonal matching pursuit (OMP).

547 The PCE model based on orthogonal matching pursuit (OMP) is iteratively built and the iterative  
548 procedure is summarized in Table A.1. At each iteration, the influence of each polynomial term  $\psi_\alpha$   
549 is measured by its correlation with the data residual  $\mathbf{R}$  (the initial value being  $\mathbf{y}$ ). The  $\alpha$  corresponding  
550 with the most correlated basis term  $\psi_\alpha$  becomes a member of the active set  $\mathbb{A}^a$ . Then, computing the  
551 basis function  $\psi_{\mathbb{A}^a}$  supported by the active set, the associated coefficients are obtained by minimizing the  
552 least-square error and  $\mathbf{R}$  is updated as the new residual. The most influential polynomials are sequentially  
553 selected by repeating the procedure above.

- 
1. Initialization: residual  $\mathbf{R}_0 = \mathbf{y}$ , active set  $\mathbb{A}_0^a = \emptyset$ , candidate set  $\mathbb{A}_0^c = \mathbb{A}_{full}$ ;
  2. For  $j = 1, \dots, P_{max} = \min\{N - 1, \text{card}(\mathbb{A}_{full})\}$ ,  
 If  $j$  equals 1, define  $\mathbf{u}_1 = \boldsymbol{\psi}_{\boldsymbol{\alpha}_1}$ ,  $\boldsymbol{\alpha}_1 = \arg \max_{\boldsymbol{\alpha} \in \mathbb{A}_0^c} |\mathbf{R}_0^T \boldsymbol{\psi}_{\boldsymbol{\alpha}}|$ , and update  $\mathbb{A}_1^a = \{\boldsymbol{\alpha}_1\}$ ,  $\mathbb{A}_1^c = \mathbb{A}_0^c \setminus \boldsymbol{\alpha}_1$ .  
 Otherwise,  
 1) update  $\mathbf{R}_{j-1} = \mathbf{R}_{j-2} + \gamma_{j-1} \mathbf{u}_{j-1}$ ,  $\gamma_{j-1}$  the smallest step length when  $\mathbf{R}_{j-1}$  has the same correlation with a basis polynomial (denoted by  $\boldsymbol{\psi}_{\boldsymbol{\alpha}_j}$ ,  $\boldsymbol{\alpha}_j \in \mathbb{A}_{j-1}^c$ ) as those with all polynomials in  $\boldsymbol{\psi}_{\mathbb{A}_{j-1}^a}$ .  
 2) update  $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$  and  $\mathbb{A}_j^c = \mathbb{A}_{j-1}^c \setminus \boldsymbol{\alpha}_j$ .  
 3) compute the equiangular vector of all polynomials in  $\boldsymbol{\psi}_{\mathbb{A}_j^a}$  as  $\mathbf{u}_j$ .  
 End
- 

Table A.2: Ranking basis polynomials based on least angle regression (LARS).

Least angle regression (LARS) is a less greedy version of traditional forward selection methods. It is known that different flavors of LARS yield efficient solutions of LASSO [37] (which constrains both the data discrepancy by ordinary least square and the sparsity of regression coefficients by  $\ell_1$ -norm) and forward stagewise linear regression [68] (another promising model-selection method), respectively.

The iterative algorithm of sparse PCE modeling based on LARS (originally proposed in [20]) is given in Table A.2, where details on how to compute step length  $\gamma_{j-1}$  and equiangular vector  $\mathbf{u}_j$  can be found in [24]. As seen from this short summary, the building process is similar with the one based on OMP, except that from the second iteration since the residual  $\mathbf{R}$  evolves along the equiangular directions of basis functions other than along basis functions themselves.

## Acknowledgments

The first author has been supported in part by the Emergence programme of the Science and Technologies of Information and Communication (STIC) department, University Paris-Saclay.

## References

- [1] A. Taflove, S. C. Hagness, Computational Electrodynamics: the Finite-Difference Time-Domain Method, Artech House, 2005.
- [2] K.-J. Bathe, E. L. Wilson, Numerical Methods in Finite Element Analysis, Prentice-Hall, 1976.
- [3] R. R. Barton, Tutorial: Input uncertainty in output analysis, in: Proc. Winter Simulation Conference, WSC2012, Berlin, Germany, 2012.
- [4] A. Kolmogorov, Foundations of the Theory of Probability: Second English Edition, Dover Publications, 1956.
- [5] R. L. Iman, J. C. Helton, An investigation of uncertainty and sensitivity analysis techniques for computer models, Risk Anal. 8 (1) (1988) 71–90.
- [6] J. P. Kleijnen, Kriging metamodeling in simulation: A review, Eur. J. Oper. Res. 192 (3) (2009) 707–716.
- [7] D. J. MacKay, Bayesian methods for adaptive models, Ph.D. thesis, California Institute of Technology, CA, USA (1992).
- [8] B. Sudret, Uncertainty propagation and sensitivity analysis in mechanical models—contributions to structural reliability and stochastic spectral methods, Habilitation à diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France.
- [9] K. Sepahvand, S. Marburg, H.-J. Hardtke, Uncertainty quantification in stochastic systems using polynomial chaos expansion, Int J. Appl. Mech. 2 (02) (2010) 305–353.
- [10] P. Kersaudy, S. Mostarshedi, B. Sudret, O. Picon, J. Wiart, Stochastic analysis of scattered field by building facades using polynomial chaos, IEEE Trans. Antennas Propag. 62 (12) (2014) 6382–6393.
- [11] C. Soize, R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM J. Sci. Comput. 26 (2) (2004) 395–410.
- [12] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.
- [13] R. G. Ghanem, P. D. Spanos, Stochastic Finite Elements: A Spectral Approach, Dover Publications, 2003.
- [14] B. Sudret, M. Berveiller, M. Lemaire, A stochastic finite element method in linear mechanics, CR Mécanique 332 (7) (2004) 531–537.
- [15] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, Eur. J. Comput. Mech. 15 (1-3) (2006) 81–92.

- 592 [16] O. P. Le Maître, M. T. Reagan, H. N. Najm, R. G. Ghanem, O. M. Knio, A stochastic projection method for fluid flow:  
593 II. Random process, *J. Comput. Phys.* 181 (1) (2002) 9–44.
- 594 [17] L. Gilli, D. Lathouwers, J. Kloosterman, T. Van der Hagen, A. Koning, D. Rochman, Uncertainty quantification for  
595 criticality problems using non-intrusive and adaptive polynomial chaos techniques, *Ann. Nucl. Energy* 56 (2013) 71–80.
- 596 [18] G. Blatman, Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis, Ph.D.  
597 thesis, Université Blaise Pascal, Clermont-Ferrand, France (2009).
- 598 [19] G. Blatman, B. Sudret, An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element  
599 analysis, *Probab. Eng. Mech.* 25 (2) (2010) 183–197.
- 600 [20] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, *J. Comput. Phys.*  
601 230 (6) (2011) 2345–2367.
- 602 [21] A. Doostan, H. Owhadi, A non-adapted sparse approximation of PDEs with stochastic inputs, *J. Comput. Phys.* 230 (8)  
603 (2011) 3015–3034.
- 604 [22] J. D. Jakeman, M. S. Eldred, K. Sargsyan, Enhancing  $\ell_1$ -minimization estimates of polynomial chaos expansions using  
605 basis selection, *J. Comput. Phys.* 289 (2015) 18–34.
- 606 [23] J. A. Tropp, A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, *IEEE Trans.*  
607 *Inf. Theory* 53 (12) (2007) 4655–4666.
- 608 [24] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, *Ann. Stat.* 32 (2) (2004) 407–499.
- 609 [25] L. Breiman, Bagging predictors, *Mach. Learn.* 24 (2) (1996) 123–140.
- 610 [26] B. Efron, R. J. Tibshirani, *An Introduction to The Bootstrap*, CRC Press, 1994.
- 611 [27] W. Sauerbrei, M. Schumacher, A bootstrap resampling procedure for model building: application to the Cox regression  
612 model, *Stat. Med.* 11 (16) (1992) 2093–2109.
- 613 [28] P. Royston, W. Sauerbrei, et al., Bootstrap assessment of the stability of multivariable models, *Stata J.* 9 (4) (2009) 547.
- 614 [29] E. Anna, Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso  
615 and bootstrap approach, Master’s thesis, Umeå University (2017).
- 616 [30] R. De Bin, S. Janitza, W. Sauerbrei, A.-L. Boulesteix, Subsampling versus bootstrapping in resampling-based model  
617 selection for multivariable regression, *Biometrics* 72 (1) (2016) 272–280.
- 618 [31] S. Geisser, The predictive sample reuse method with applications, *J. Am. Stat. Assoc.* 70 (350) (1975) 320–328.
- 619 [32] M. Walschaerts, E. Leconte, P. Besse, Stable variable selection for right censored data: comparison of methods, arXiv  
620 preprint arXiv:1203.4928.
- 621 [33] R. Lebrun, A. Dutfoy, A generalization of the Nataf transformation to distributions with elliptical copula, *Probab. Eng.*  
622 *Mech.* 24 (2) (2009) 172–178.
- 623 [34] M. Lemaire, *Structural Reliability*, John Wiley & Sons, 2013.
- 624 [35] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press on Demand, 2004.
- 625 [36] C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, 1973.
- 626 [37] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal Stat. Soc. Series B* (1996) 267–288.
- 627 [38] R. B. Nelsen, *An Introduction to Copulas*, Springer Science & Business Media, 2007.
- 628 [39] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: *Proc. 14th*  
629 *International Joint Conference on Artificial Intelligence, IJCAI1995*, Vol. 14, Montreal, Canada, 1995, pp. 1137–1145.
- 630 [40] K. Konakli, B. Sudret, Polynomial meta-models with canonical low-rank approximations: numerical insights and compar-  
631 ison to sparse polynomial chaos expansions, *J. Comput. Phys.* 321 (2016) 1144–1169.
- 632 [41] B. Efron, *The Jackknife, The Bootstrap, and Other Resampling Plans*, SIAM, 1982.
- 633 [42] J. Friedman, T. Hastie, R. Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics, 2001.
- 634 [43] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, *Dokl. Akad.*  
635 *Nauk SSSR* 148 (5) (1963) 1042–1045.
- 636 [44] S. Marelli, B. Sudret, UQLab user manual–polynomial chaos expansions, Report UQLab-V1.1-104, Chair of Risk, Safety  
637 & Uncertainty Quantification, ETH Zürich.
- 638 [45] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Reducing over-optimism in variable selection by cross-model validation,  
639 *Chemom. Intell. Lab. Syst.* 84 (1-2) (2006) 69–74.
- 640 [46] T. Fushiki, Estimation of prediction error by using k-fold cross-validation, *Statistics and Computing* 21 (2) (2011) 137–146.
- 641 [47] D. Baumann, K. Baumann, Reliable estimation of prediction errors for QSAR models under model uncertainty using  
642 double cross-validation, *J. Cheminf.* 6 (1) (2014) 47.
- 643 [48] L. Gidskehaug, E. Anderssen, B. K. Alsberg, Cross model validation and optimisation of bilinear regression models,  
644 *Chemom. Intell. Lab. Syst.* 93 (1) (2008) 1–10.
- 645 [49] I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, *Mathem. Mod. Comput. Exp.* 1 (4) (1993) 407–414.
- 646 [50] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, *Reliab. Eng. Syst. Safe.*  
647 52 (1) (1996) 1–17.
- 648 [51] S. Kucherenko, S. Tarantola, P. Annoni, Estimation of global sensitivity indices for models with dependent variables,  
649 *Comput. Phys. Commun.* 183 (4) (2012) 937–946.
- 650 [52] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, *Reliab. Eng. Syst. Safe.* 93 (7) (2008) 964–979.
- 651 [53] S. Marelli, B. Sudret, UQLab: A framework for uncertainty quantification in Matlab, in: *Proc. 2nd International Confer-*  
652 *ence on Vulnerability, Risk Analysis and Management, ICVRAM2014*, Liverpool, United Kingdom, 2014, pp. 2554–2563.
- 653 [54] *Chair of Risk, Safety and Uncertainty Quantification of ETH Zurich*, [online] Available at: <https://www.uqlab.com>  
654 [Accessed Aug. 18, 2019].
- 655 [55] M. D. McKay, R. J. Beckman, W. J. Conover, Comparison of three methods for selecting values of input variables in the  
656 analysis of output from a computer code, *Technometrics* 21 (2) (1979) 239–245.

- 657 [56] S. Xiong, P. Z. Qian, C. J. Wu, Sequential design and analysis of high-accuracy and low-accuracy computer codes,  
658 *Technometrics* 55 (1) (2013) 37–46.
- 659 [57] E. Van Deventer, E. Van Rongen, R. Saunders, WHO research agenda for radiofrequency fields, *Bioelectromagnetics* 32 (5)  
660 (2011) 417–421.
- 661 [58] M.-C. Gosselin, E. Neufeld, H. Moser, E. Huber, S. Farcito, L. Gerber, M. Jedensjoe, I. Hilber, F. Di Gennaro, B. Lloyd,  
662 et al., Development of a new generation of high-resolution anatomical models for medical device evaluation: the virtual  
663 population 3.0, *Phys. Med. Biol.* 59 (18) (2014) 5287.
- 664 [59] J. E. Hansen, *Spherical Near-Field Antenna Measurements*, Peter Peregrinus Ltd, 1988.
- 665 [60] I. Liorni, M. Parazzini, N. Varsier, A. Hadjem, P. Ravazzani, J. Wiart, Exposure assessment of one-year-old child to 3G  
666 tablet in uplink mode and to 3G femtocell in downlink mode using polynomial chaos decomposition, *Phys. Med. Biol.*  
667 61 (8) (2016) 3237.
- 668 [61] V. Picheny, D. Ginsbourger, O. Roustant, R. T. Haftka, N.-H. Kim, Adaptive designs of experiments for accurate approx-  
669 imation of a target region, *J. Mech. Des.* 132 (7) (2010) 071008.
- 670 [62] S. Dubreuil, M. Berveiller, F. Petitjean, M. Salaün, Construction of bootstrap confidence intervals on sensitivity indices  
671 computed by polynomial chaos expansion, *Reliab. Eng. Sys. Safety* 121 (2014) 263–275.
- 672 [63] N. Fajraoui, S. Marelli, B. Sudret, Sequential design of experiment for sparse polynomial chaos expansions, *SIAM/ASA*  
673 *J. Unc. Quant.* 5 (1) (2017) 1061–1085.
- 674 [64] I. Liorni, M. Parazzini, S. Fiocchi, P. Ravazzani, Study of the influence of the orientation of a 50-Hz magnetic field on  
675 fetal exposure using polynomial chaos decomposition, *Int. J. Environ. Res. Public. Health* 12 (6) (2015) 5934–5953.
- 676 [65] P. Kersaudy, B. Sudret, N. Varsier, O. Picon, J. Wiart, A new surrogate modeling technique combining Kriging and  
677 polynomial chaos expansions—application to uncertainty analysis in computational dosimetry, *J. Comput. Phys.* 286 (2015)  
678 103–117.
- 679 [66] Y. Huang, J. Wiart, Simplified assessment method for population RF exposure induced by a 4G network, *IEEE J.*  
680 *Electromagn. RF Microw. Med. Biol.* 1 (1) (2017) 34–40.
- 681 [67] N. Mantel, Why stepdown procedures in variable selection, *Technometrics* 12 (3) (1970) 621–625.
- 682 [68] S. Weisberg, *Applied Linear Regression*, John Wiley & Sons, 2005.