Surrogate modeling of indoor down-link human exposure based on sparse polynomial chaos expansion

Journal Article

Author(s):
Liu, Zicheng; Lesselier, Dominique; Sudret, Bruno (iD); Wiart, Joe

SURROGATE MODELING OF INDOOR DOWN-LINK HUMAN EXPOSURE BASED ON SPARSE POLYNOMIAL CHAOS EXPANSION

Z. Liu, D. Lesselier, B. Sudret, J. Wiart

CHAIR OF RISK, SAFETY AND UNCERTAINTY QUANTIFICATION
STEFANO-FRANSCINI-PLATZ 5
CH-8093 ZÜRICH

Data Sheet

Surrogate modeling of indoor down-link human exposure based on sparse polynomial chaos expansion

Zicheng Liu[1,2], Dominique Lesselier[2], Bruno Sudret[3], and Joe Wiart[1]

[1]Chaire C2M, LTCI, Télécom Paris, Paris 75013, France.

[2]Laboratoire des Signaux et Systèmes, UMR8506 (CNRS-CentraleSupelec-Université Paris-Sud), Université Paris-Saclay, Gif-sur-Yvette cedex 91192, France

[3]ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification, Stefano-Franscini-Platz 5, Zürich 8093, Switzerland

July 10, 2019

Abstract

Human exposure induced by wireless communication systems increasingly draws the public attention. Here, an indoor down-link scenario is concerned and the exposure level is statistically analyzed. The electromagnetic field (EMF) emitted by a WiFi box is measured and electromagnetic dosimetry features are evaluated from the whole-body specific absorption rate as computed with a Finite-Difference Time-Domain (a.k.a. FDTD) code. Due to computational cost, a statistical analysis is performed based on a surrogate model, which is constructed by means of so-called sparse polynomial chaos expansion (PCE), where the inner cross validation (ICV) is used to select the optimal hyperparameters during the model construction and assess the model performance. However, the ICV error is optimized and the model assessment tends to be overly optimistic with small data sets. The method of cross-model validation is used and outer cross validation is carried out for the model assessment. The effects of the data preprocessing are investigated as well. Based on the surrogate model, the global sensitivity of the exposure to input parameters is analyzed from Sobol' indices.

Keywords: Specific absorption rate, surrogate model, polynomial chaos expansion, least angle regression, orthogonal matching pursuit, cross-model validation, double cross validation, Sobol' indices, global sensitivity analysis, data preprocessing

1   Introduction

Human exposure [37] to radio-frequency (RF) electromagnetic fields (EMF) increasingly draws public attention [38] due to the rapid development of wireless communications and the Internet-of-Things (IoT).

The exposure by pervasive telecommunication devices and infrastructures needs to be evaluated. Research efforts have been carried out recently, in which the exposure level is evaluated by the computation of the specific absorption rate (SAR) [37], which quantifies the power absorbed by human tissues. The distribution of electric field inside the body is estimated by measurements with an equivalent phantom or by simulation with a numerical modeling approach, e.g., finite-difference in time-domain (FDTD) [29] and finite-element methods (FEM) [30].

According to the position of the human body under consideration, the investigation of electromagnetic dosimetry is often divided into indoor scenarios [13; 14; 12] and outdoor ones [7; 23], where near fields and far field are respectively considered. Particular attention is given to indoor scenarios here, since people spend about 70% of their time inside rooms [41]. So, in comparison with far-field scenarios, the emitting source can be placed quite close to the human body and the exposure level may be high. Despite related researches mainly on deterministic investigation of several representative cases [14; 13], a statistical analysis [37] is necessary to estimate the human exposure with uncertain factors (e.g., frequency, distance) and exhibit the key impacting factors, especially in highly varying scenarios.

Monte-Carlo simulations are often run for statistical analysis but may be intractable in our case due to the high computational/measurement cost to compute SAR. Recently, surrogate models based on the polynomial chaos expansions (PCE) have been constructed to approximate the complicated relation between the EMF exposure and the impacting factors [26; 25; 32]. Expanding the EMF exposure on orthogonal multivariate polynomials, the expansion coefficients are obtained as the ordinary least square (OLS) solution. To avoid the high variance of the constructed model and the overfitting problem, the cardinality of the basis should be constrained by setting a maximum value for the total degree of polynomials [28]. However, the curse-of-dimensionality, i.e., the fact that the cardinality of the basis polynomially increases with the number of considered factors and the total degree, prevents broad applications of ordinary PCE-based surrogate models. Then, to reduce the basis size, approaches to construct so-called *sparse PCE model* [5; 4; 28] have been proposed to only include the most influential polynomials in the PCE model. The influence of polynomials is usually measured by the correlation with the responses and the polynomials included in the sparse PCE model are selected by least angle regression (LARS) [11] or orthogonal matching pursuit (OMP) [31].

Two hyperparameters, the total degree of polynomials and the number of included polynomials in the final model, are involved in the construction of sparse PCE models [28]. Their selection is performed via cross validation (CV). The hyperparameters corresponding to the smallest CV error are treated as optimal. This criterion of smallest CV error is also used to assess the prediction performance of the final model. Therefore, CV plays the role for both model selection (hyperparameter selection) and model assessment. However, it has been recognized that if the final model is obtained via an extensive optimization by the CV and the performance is assessed by the same CV based on the same experimental design (ED), the model assessment tends to be optimistic [1].

In the mentioned above study, over-optimistic model assessments appear with small training data sets, which include tens to hundreds of samples. So-called *cross-model validation* (CMV) [1; 3] is used

to overcome the bias of the assessment. Regarding the CV for the model selection as the first-layer validation, a second-layer CV is introduced to assess the constructed model on data independent of the training process. With the assessment of models constructed with LARS and OMP by CMV, the best model is chosen for predictions and inferences. Meanwhile, data preprocessing (e.g. representation of inputs in Cartesian or spherical coordinate system) significantly impacts the performance of the surrogate model. Such effects are investigated here.

The problem statement describing the down-link dosimetry case study is given in Section 2. Sparse PCE models based on LARS and OMP are considered in Section 3. In Section 4, the idea of CMV is introduced in detail and the assessment accuracy is illustrated by results in Section 5, where the effect of data preprocessing is shown as well.

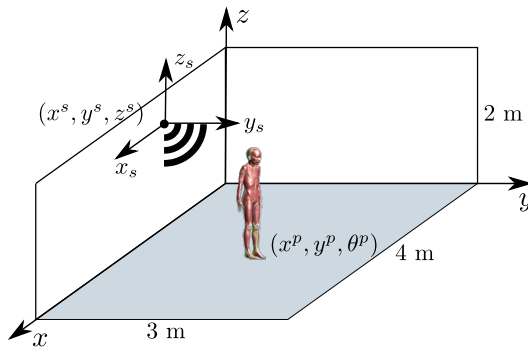## 2 Experimental design (ED)

Figure 1: Uncertain parameters considered in the indoor down-link scenario.

The scenario of interest is sketched in Fig. 1, where a WiFi box working at $2,400$ MHz is moving along the wall of a $4 \times 3 \times 2 \, \mathrm{m}^3$ room and a high-resolution female model, named as "Eartha" from the Virtual Family [8], is standing inside. The positions of the WiFi box and human model impact the electromagnetic exposure, which is quantified by the whole-body SAR, i.e., the averaged SAR over the whole body.

The averaged SAR over the body $\mathbb{V}$ is computed according to

$$\mathrm{SAR} = \frac{1}{V} \int_{\mathbb{V}} \frac{|\boldsymbol{E}(\boldsymbol{r})|^2 \sigma(\boldsymbol{r})}{\rho(\boldsymbol{r})} d\boldsymbol{r}, \tag{1}$$

where V is the volume of $\mathbb{V}$, $|\boldsymbol{E}(\boldsymbol{r})|^2$ is the root-mean-square of electric field, and $\sigma(\boldsymbol{r})$, $\rho(\boldsymbol{r})$ denote conductivity and density at the sampling position $\boldsymbol{r}$. From Eq.(1), one sees that the uncertain factors influencing the SAR computation may come from the accuracy of measured/simulated electric fields, the physical properties of tissues, and the numerical computation of the integral in Eq. (1). Here, only the uncertainty of the electric field due to the relative position of the source and the human model is considered.

In the global Cartesian coordinate system, as shown in Fig. 1, the position of the WiFi box and the one of the human model are denoted by (x^s, y^s, z^s) and (x^p, y^p, z^p), respectively. The human model is supposed

3

Source (notation)	Range	Human model (notation)	Range
x^s (X_1)	$[0.3, 3.7]$ m	x^p (X_4)	$[0.05, 3.95]$ m
y^s (X_2)	$[0.3, 2.7]$ m	y^p (X_5)	$[0.05, 2.95]$ m
z^s (X_3)	$[0.25, 2]$ m	θ^p (X_6)	$[0, 360)$ degree

Table 1: Range of the uncertain variables.

to stand in the room, thus z^p is considered a constant and not a uncertain variable. Its orientation may matter and is considered through the parameter θ^p. Therefore, six uncertain variables are involved in the uncertainty analysis. Assuming the variables follow uniform distributions, the input space is defined by the ranges given in Table 1, where 0.25 m is the minimum height of source, 0.3 m, 0.05 m are the minimum distance of human model and source to the wall, respectively. To efficiently characterize the response of the physical system with a limited number of samples, Latin hypercube sampling (LHS) [39] is used to determine the input samples while the corresponding whole-body SAR is computed according to the following procedures.

When the human body is far from the emitting source, the front of the incident wave is approximately planar and the source is often treated as planar waves [20; 9]. However, this approximation is not applicable in our case since the source may be quite close to the human body and the near field should be analyzed too. To characterize the near field scattered by the box, the measurement system StarLab by MVG® is used to measure the field. The field is expressed as a spherical wave expansion [18], with which an in-house FDTD code can be used to compute the field inside a high-resolution (1 mm³) human model. However, the room size is too large to be directly considered in the FDTD code. The analyzed region is then reduced by computing a Huygens box enclosing the human model [32]. As a result, the computational burden becomes bearable with the utilization of GPU techniques and about 3 hours are required for each single deterministic computation. However, the computation cost is still too high to envisage statistical analysis by Monte-Carlo simulations. A surrogate model (or metamodel) is built in the next section to mimic the nonlinear relation between the whole-body SAR and six input parameters.

3   Surrogate modeling based on polynomial chaos expansions

To facilitate the illustration, let us denote the six parameters gathered in Table 1 by $\{X_i, i = 1, 2, \ldots, 6\}$, the computed whole-body SAR by Y, and the resulting SAR $Y = \mathcal{M}(\boldsymbol{X})$. Assuming $E[Y^2] < +\infty$, the polynomial chaos expansion of the random SAR reads:

$$Y = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^6} \beta_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{2}$$

The construction of the basis follows the theory of generalized PCE (gPCE) [40; 35]. $\Psi_{\boldsymbol{\alpha}}$ is a multivariate polynomial, which is a tensor product of univariate polynomials, i.e.,

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \pi_{\alpha_1}(X_1) \times \ldots \times \pi_{\alpha_6}(X_6), \tag{3}$$

4

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_6]$, α_i as the degree of the univariate polynomial in variable X_i. The polynomial family π_{α_i} depends on the probability density function (PDF) of the random variable X_i. Since all input parameters have uniform distributions, Legendre polynomials are used to construct the basis polynomials in our case study. The hyperparameter $\boldsymbol{\alpha}$ and the expansion coefficient $\beta_{\boldsymbol{\alpha}}$ are learned from the available input samples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ and corresponding output values $\{y^{(1)}, \ldots, y^{(N)}\}$, N being the size of the experimental design (ED).

In practice, the infinite series of Eq. (2) shall be truncated and the most influential basis polynomials are to be included in the truncation. The commonly utilized approach is setting a maximum value to the total degree of basis polynomials [4], i.e., $\sum_{i=1}^{6} \alpha_i \leq p$. Such a setting yields the so-called *full PCE* model. The value of p decides for the flexibility of the PCE model. With a large p, the constructed model tends to be unbiased but with a high variance (especially when only a small set of data are available) and vice versa. An optimization is necessary to select the p corresponding to the best model, given a particular set of runs. Here, "best" is defined in terms of prediction accuracy, since the prediction ability is of interest.

The assessment of prediction performance is often carried out by cross validation (CV), which includes the technique of leaving-many-out [17], bootstrapping [10] and k-fold CV [2]. The latter is performed by dividing the whole set of data points into k subsets with approximately the same size and by evaluating the surrogate model built from $(k-1)$ subsets on the remaining k-th subset. With $k = N$, the k-fold CV becomes the so-called *leave-one-out CV* and the validation error is computed by

$$\epsilon^{\text{LOO}} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{M}(\boldsymbol{x}^{(n)}) - \widehat{\mathcal{M}}^{-(n)}(\boldsymbol{x}^{(n)}) \right)^2, \tag{4}$$

where the superscript "(n)" denotes the n-th sample, $n = 1, 2, \ldots, N$, and $\widehat{\mathcal{M}}^{-(n)}$ denotes the surrogate model trained by leaving the n-th sample out. This way, the value of p corresponding to the minimum CV error is selected.

Denote $\mathbb{A}^{\text{full}}$ as the set of selected $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}} = [\Psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{A}^{\text{full}}]$, and $\boldsymbol{\beta} = [\beta_{\boldsymbol{\alpha}}, , \boldsymbol{\alpha} \in \mathbb{A}^{\text{full}}]^T$, "$T$" the transposition operator. The truncated PCE of Eq. (2) is written as

$$\mathcal{M}^{\text{full}}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{A}^{\text{full}}} \beta_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}} \boldsymbol{\beta}_{\mathbb{A}^{\text{full}}}. \tag{5}$$

The ordinary-least-square (OLS) solution to the expansion coefficients is

$$\hat{\boldsymbol{\beta}}_{\mathbb{A}^{\text{full}}} = \left(\boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}}^T \boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}} \right)^{-1} \boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}}^T \boldsymbol{y}, \tag{6}$$

where $\boldsymbol{y}$ is the vector gathering the SAR values computed for each input vector $\boldsymbol{x}^{(i)}$, $i = 1, \ldots, N$. With $\boldsymbol{\Psi}_{\mathbb{A}^{\text{full}}}$ and $\hat{\boldsymbol{\beta}}_{\mathbb{A}^{\text{full}}}$, the full PCE model is constructed with Eq. (5). However, the cardinality of $\mathbb{A}^{\text{full}}$ equals $\binom{p+6}{p}$, which increases polynomially with p. As a result, the construction process of full PCE models suffers from the curse-of-dimensionality [15], which means that the required ED size shall become large to avoid the high variance (or the overfitting problem) of the constructed model. Then, so-called *sparse polynomial chaos expansions* [4; 5] have been proposed to reduce the basis.

Measuring the influence of multivariate polynomials by the correlation with the sampled output data, only the most correlated polynomials are included in the sparse PCE model. Matching pursuit algorithms,

For $p = 1, 2, \ldots, p_{\max}$,

1. Initialization: residual $\boldsymbol{R}_0 = \boldsymbol{y}$, active set $\mathbb{A}_0^a = \varnothing$, candidate set $\mathbb{A}_0^c = \mathbb{A}_p^{\text{full}}$.

2. For $j = 1, 2, \ldots, P_{\max} = \min\{N - 1, \text{card}(\mathbb{A}_p^{\text{full}})\}$,

   1) Find the basis element most correlated with $\boldsymbol{R}_{j-1}$: $\boldsymbol{\alpha}_j = \arg\max_{\boldsymbol{\alpha} \in \mathbb{A}_{j-1}^c} \left| \boldsymbol{R}_{j-1}^T \boldsymbol{\psi}_{\boldsymbol{\alpha}} \right|$.

   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) With $\boldsymbol{\psi}_{\mathbb{A}_j^a}$, compute $\boldsymbol{\beta}_j$ as the OLS solution.

   4) Update residual $\boldsymbol{R}_j = \boldsymbol{y} - \boldsymbol{\psi}_{\mathbb{A}_j^a}^T \boldsymbol{\beta}_j$.

   End

3. Based on $\boldsymbol{\psi}_{\mathbb{A}_j^a}$ and $\boldsymbol{\beta}_j$, compute ϵ_j^{LOO}, $j = 1, 2, \ldots, P_{\max}$.

4. $\epsilon_p^{\min} = \min\{\epsilon_j^{\text{LOO}}\}$, $P = \arg\min_j \epsilon_j^{\text{LOO}}$ and the PCE model corresponding to $\boldsymbol{\psi}_{\mathbb{A}_P^a}$ is selected.

5. For $p > 2$, if $\epsilon_p^{\min} > \epsilon_{p-1}^{\min} > \epsilon_{p-2}^{\min}$, stop.

   End

Table 2: Sparse PCE model based on orthogonal matching pursuit.

e.g., orthogonal matching pursuit (OMP) [31] and least angle regression (LARS) [11], have been applied to rank and select the basis polynomials. The procedures to construct sparse PCE models based on OMP are summarized in Table 2 [27], where the number of included polynomials is smaller than $N - 1$ to avoid overfitting modeling and the optimal number is decided by cross validation. The maximum value of p is denoted as $p_{\max}$ and the optimal value is decided by an early-stop criterion, i.e., when the CV error increases for two subsequent values of p, the algorithm will stop to avoid the overfitting problem. The model construction with LARS follows a similar procedure and is shown less greedy, since the data residual $\boldsymbol{R}$ evolves with equiangular directions of the basis instead of the basis itself [27; 11].

4 Cross-model validation

As mentioned in Section 3, a number of candidate surrogate models is constructed with the training data set by tuning the hyperparameters of the training process, i.e., the number of included polynomials in the final model P and the total degree of multivariate polynomials p. With the model selection by cross validation, the surrogate model with the smallest validation error (ϵ^{LOO} here) is chosen as the final model.

The best model in terms of a given training process has the smallest validation error in ideal cases. However, the possibility for a candidate PCE model getting underestimated assessments may lead to a validation error even smaller than the error with the "true" model. Consequently, this candidate model would be selected despite that it is suboptimal. Moreover, if this smallest validation error is used as the metric of model assessment, which is the case for some applications of sparse PCE models [32; 22], the finally selected surrogate model would be over-optimistically assessed. This phenomenon is the so-called

model-selection bias and can be detected by comparing the above validation error for model selection and the testing error which is obtained from the prediction on a data set independent of the training process. The bias is detected if the validation error is much smaller than the testing error.

The phenomenon of model-selection bias happens in the research of Quantitative Structure-Activity Relationship (QSAR) [34], where a suitable subset of a high number of molecular descriptors is selected to optimize the model representation. The technique of cross-model validation (CMV) [1] (or double cross validation [3]) was proposed with the knowledge that the data for model assessment should not be involved in the model construction or selection.

CMV introduces a second-layer cross validation, which is so-called *outer CV*, while the CV for model selection is *inner CV*. In the outer CV, the experimental design is randomly divided into two disjoint data sets, one for model assessment, the other for model construction and selection. This way, the data set for model assessment is independent of the training process. Thus, the outer CV error is an unbiased metric for model assessment.

Approaches including the holdout method [2], leave-many-out, and k-fold CV are commonly utilized to split ED. The leave-one-out CV is applied in the outer CV considering the following aspects. First, most of available data are used for training in LOOCV to improve the surrogate-modeling performance, especially considering that cases with a quite small ED are of concern here. In addition, the outer CV actually assesses surrogate models constructed from subsets of the ED. Strictly speaking, the final model trained with the whole set of available data is not validated and the assessment by outer CV tends to be pessimistic due to less training data. Since only a single sample is left out from the training process, outer LOOCV can well approximate the performance of the final model.
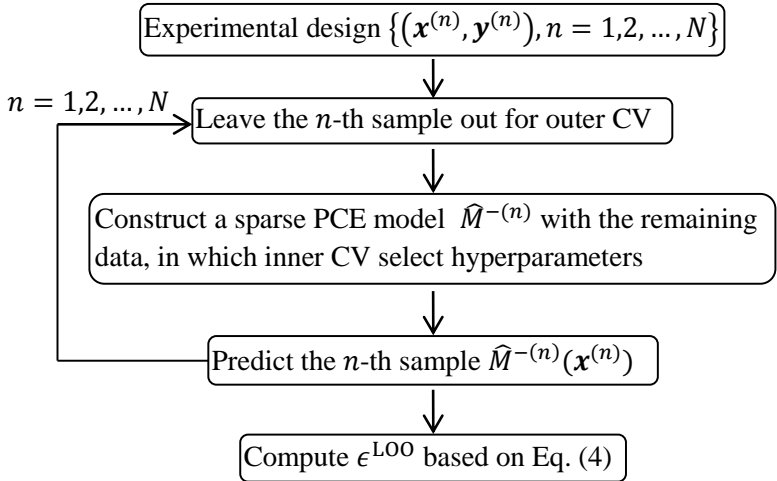

Figure 2: Flow chart of outer leave-one-out cross validation.

Fig. 2 gives the flow chart performing the outer CV by leaving a single sample out. Remark that, although each sample is sequentially left out for outer CV, actually no sample is totally independent of the training process. However, the mode-selection bias is absent since the sample for outer CV is not involved in the model construction and selection in every single data division.

Due to the introduction of outer LOOCV, N surrogate models are built to compute the outer CV error and about $N-1$ times more computations are required. Since cases with small ED are of concern, the construction of a single surrogate model is usually efficient and the running of outer CV is tractable. Otherwise, after the division of ED, since the model construction with different training data sets is independent of each other, parallel computational techniques (e.g., distributed computation) can be used.

5 Results

The sparse PCE models are constructed with the open-source software UQLab V1.2.0 [28], in which LARS or OMP is used to select influential basis polynomials. Default configurations of UQLab for sparse PCE models are adopted, except that the value of $p_{\max}$ in Table 2 is set at 10. This section shows the model assessment by inner CV, CMV and independent testing. The improved performance based on pre-processing the data and the results of sensitivity analysis are shown in the second subsection.

5.1 Validation

The accuracy of model assessments by inner CV and outer CV (or CMV) is checked by the validation on independent testing data, the size of which should be large enough to ensure the testing accuracy. For the concerned scenario, since the size of the available data is limited, a subset (e.g., half) of the data is arbitrarily drawn for independent testing. The over-optimistic assessment by inner CV is independent of the considered scenarios. Two analytical functions, namely the Ishigami function and borehole function, are analyzed. In these two cases, a large set of testing data is easily achievable.

According to the above illustration, one has a three-layer validation including inner CV, outer CV and independent testing. Inner CV is commonly used to do both hyperparameter selection and model assessment. In the framework of cross-model validation, the model assessment is performed by outer CV instead. The assessment accuracy is examined by the independent testing.

The validation error corresponding with the three-layer validation is denoted by ϵ_{type}, where the subscript is "ICV" fo inner CV, "OCV" for outer CV or "test" for testing. Here, ϵ_{ICV} is withdrawn from UQLab, in which the LOOCV is performed and the validation error is corrected [6] by

$$\epsilon_{\text{ICV}} = \epsilon^{\text{LOO}} \left(1 - \frac{\text{card}(\mathbb{A})}{N} \right)^{-1} \left(1 + \text{tr}((\boldsymbol{\Psi}^T \boldsymbol{\Psi})^{-1}) \right), \tag{7}$$

where "card" and "tr" means the cardinality of a set and the trace of a matrix. Based on the validation error, the determination coefficient is computed as

$$Q^2_{\text{type}} = 1 - \frac{\epsilon_{\text{type}}}{\text{var}(\boldsymbol{y})}, \tag{8}$$

which indicates that with a better surrogate model, the associated Q^2 is closer to 1. If the inner and outer CV can properly assess the PCE model, we would expect that $Q^2_{\text{ICV}} \approx Q^2_{\text{test}}$ and $Q^2_{\text{OCV}} \approx Q^2_{\text{test}}$.

5.1.1 Ishigami function

The Ishigami function [21], which is widely used for benchmarking in uncertainty quantification and sensitivity analysis, is defined by

$$Y = \sin X_1 + a \sin^2 X_2 + b X_3^4 \sin X_1, \tag{9}$$

where $a = 7$, $b = 0.1$ here. All random variables X_i, $i = 1, 2, 3$, are independent and uniformly distributed over $[-\pi, \pi]$. Legendre polynomials are used to compose the basis according to the principle of the generalized PCE.
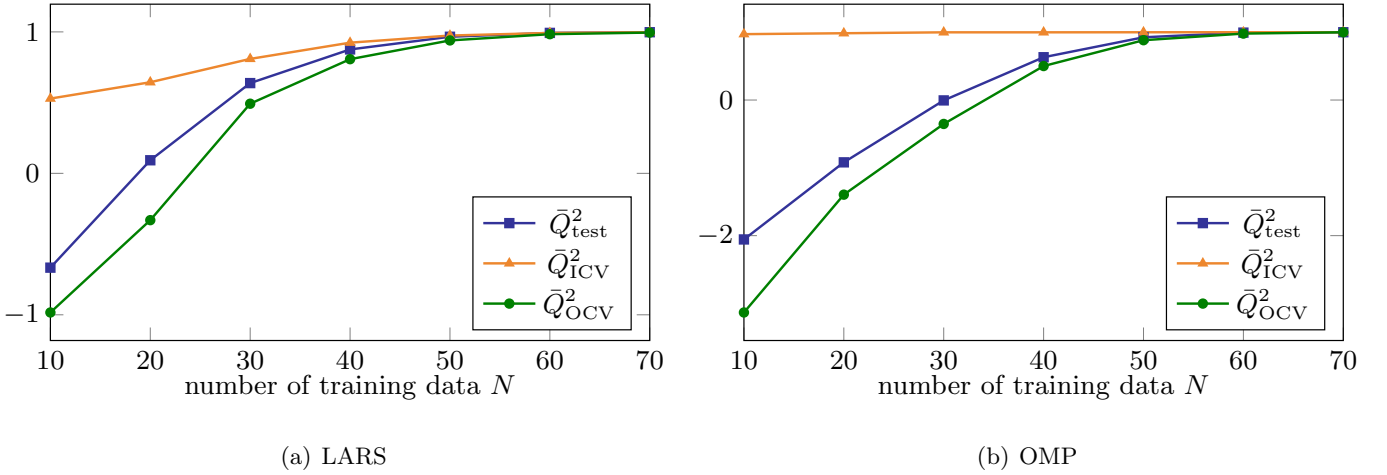

(a) LARS

(b) OMP

Figure 3: Ishigami function - mean of Q^2 over 100 replications versus the increasing number of training data.

Samples in the input space are obtained by Latin hypercube sampling (LHS). The testing data are composed of 10^4 samples, while the number of training data N increases from 10 to 70 and 100 replications are performed for each N. The obtained validation results are shown in Fig. 3, where $\bar{Q}^2$ denotes the averaged quantity over the replications. As observed, no matter with LARS or OMP, the inner CV tends to optimistically assess the constructed model and over-optimistic assessments happen when $N \leq 40$, where the variance of the surrogate modeling is high and the possibility for a candidate model having smaller (underestimated) validation error than the true model is large. In contrast, the outer CV is likely to have slightly pessimistic assessments and the bias extent decreases as the performance of the surrogate model improves.

5.1.2 Borehole function

The Borehole funtion [19] models the water flow through a borehole and is defined as

$$Y = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\left(1 + T_u/T_l\right) + 2L T_u/r_w^2 K_w} \tag{10}$$

The description and distribution of the involved eight independent variables are given in Table 3. While Legendre polynomials are used for the variables with uniform distributions, Hermite polynomials are used for the other variables after the isoprobabilistic transformation into standard normal variables [24].

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^2/yr)	Uniform	$[63070, 115600]$	transmissivity of upper aquifer
H_u (m)	Uniform	$[990, 1110]$	potentiometric head of upper aquifer
T_l (m^2/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	$[9855, 12045]$	hydraulic conductivity of borehole

Table 3: Borehole function - description and distribution of input variables.
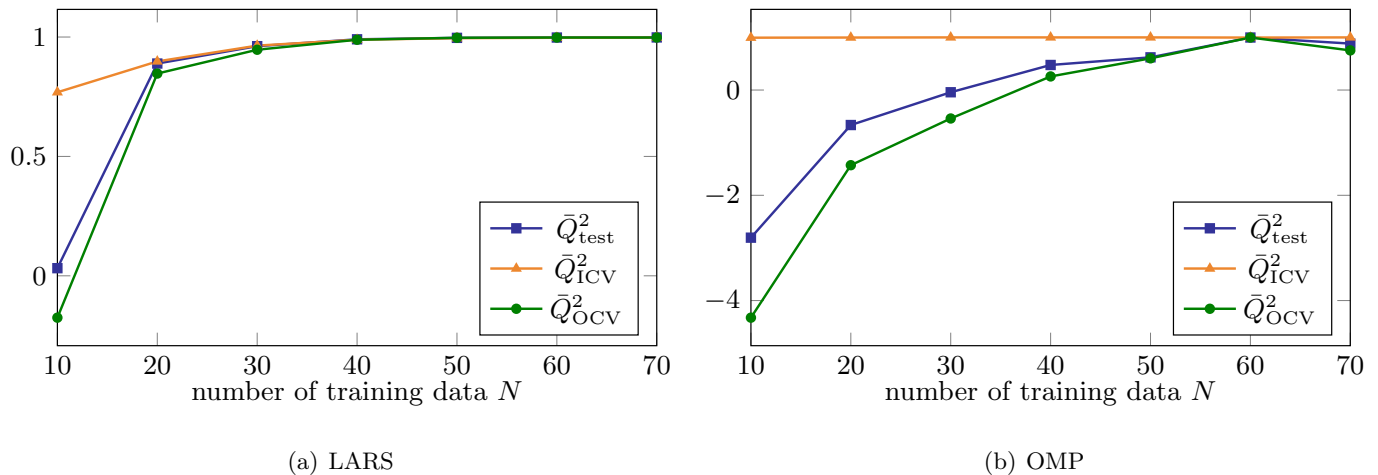

(a) LARS

(b) OMP

Figure 4: Borehole function - mean of Q^2 over 100 replications versus the increasing number of training data.

Sampling the input space by LHS method and taking 10^4 data for independent testings, the determination coefficients computed by the three-layer validation are given in Fig. 4, where the optimistic assessments by ICV and pessimistic ones by OCV are observed. The over-optimistic assessment by ICV occurs when the constructed surrogate model has a poor performance, which happens when $N = 10$ with LARS and $N = 10, \ldots, 50$ with OMP. In contrast, the outer CV slightly underestimates poor surrogate models but can properly assess the ones with high Q^2. Thus, one may conclude that the outer CV fits the practical requirements for the model assessment, since poor surrogate models can be prevented from the final application.

5.1.3 Specific absorption rate

Constructing the sparse PCE model with the whole set of data, which includes 350 samples, the prediction of the training data is shown in Fig. 5. As seen, the surrogate models based on LARS and OMP seem

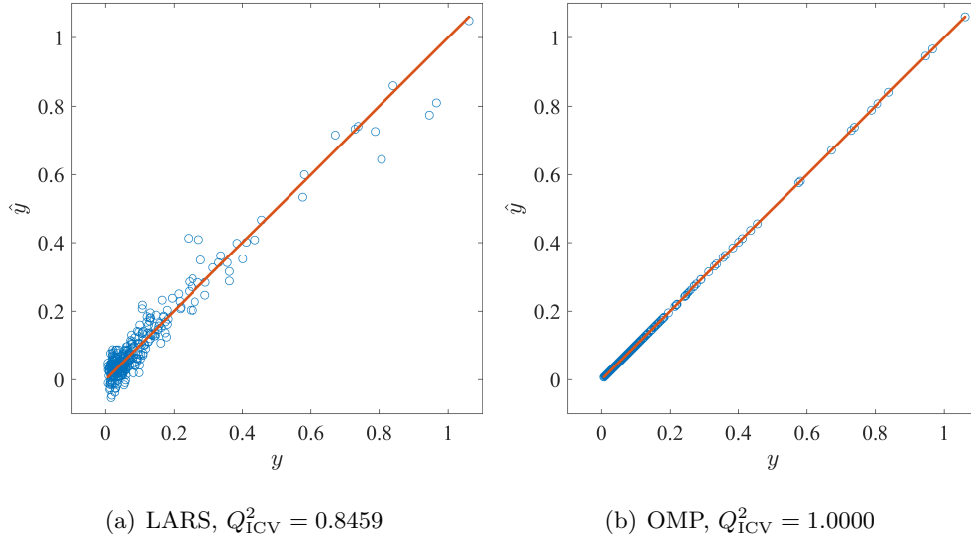(a) LARS, $Q^2_{\mathrm{ICV}} = 0.8459$ (b) OMP, $Q^2_{\mathrm{ICV}} = 1.0000$

Figure 5: Specific absorption rate - prediction of training data with the sparse PCE model based on the whole set of data.

to have excellent prediction performance. Especially, the model based on OMP can accurately predict the training samples with $Q^2_{\mathrm{ICV}} = 1.0000$, i.e., the validation error is zero. In some applications of sparse PCE models, since the inner validation error is so small, the constructed model is to be used to predict unknown inputs and perform inferences. However, since the data for ICV already participated in the training process, the performance assessment should be checked further in case of over-optimistic assessments.

(a) LARS, $\bar{Q}^2_{\mathrm{ICV}} = 0.7490$ and $\bar{Q}^2_{\mathrm{test}} = 0.4203$ (b) OMP, $\bar{Q}^2_{\mathrm{ICV}} = 1.0000$ and $\bar{Q}^2_{\mathrm{test}} = -0.1262$
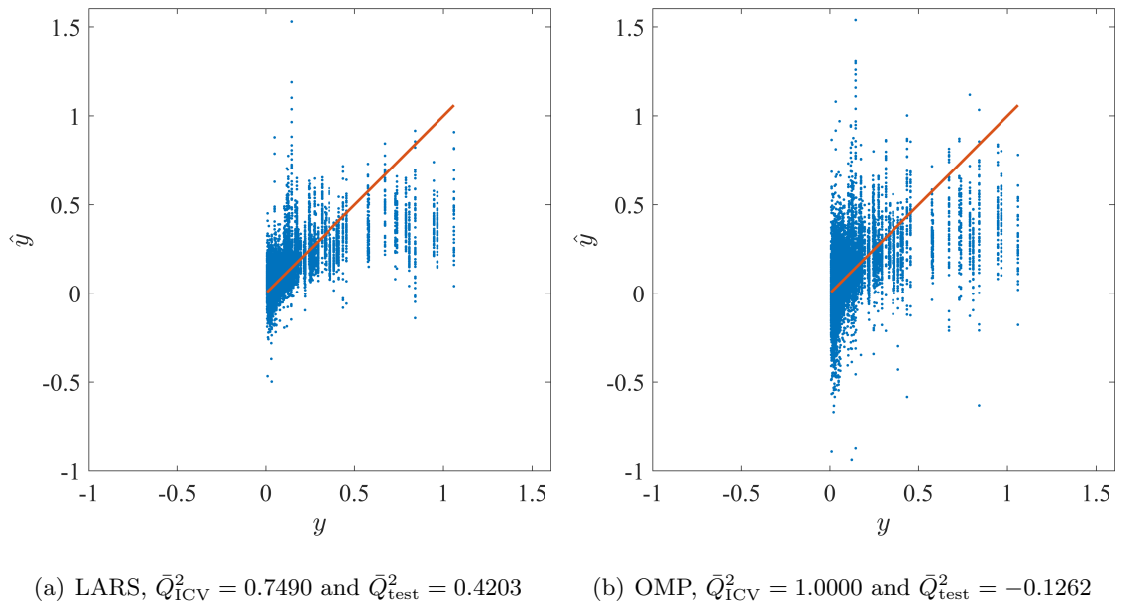
Figure 6: Specific absorption rate - prediction of independent testing data with the sparse PCE model based on an half set of data

Randomly choosing half of the data set for independent testing, sparse PCE models are constructed with the remaining data. Repeating this process 100 times, the prediction of $17,500$ resampled samples is given in Fig. 6. Although both approaches perform poorly, surrogate models based on LARS seem better than those based on OMP in the prediction of small whole-body SARs. The averaged Q^2_{test} supports the same conclusion. However, since $\bar{Q}^2_{\text{ICV}} = 1$, the OMP-based PCE model would be considered as an accurate representation of the original FDTD model. One may conclude that the assessment by inner CV can be quite misleading since it can be over-optimistic.
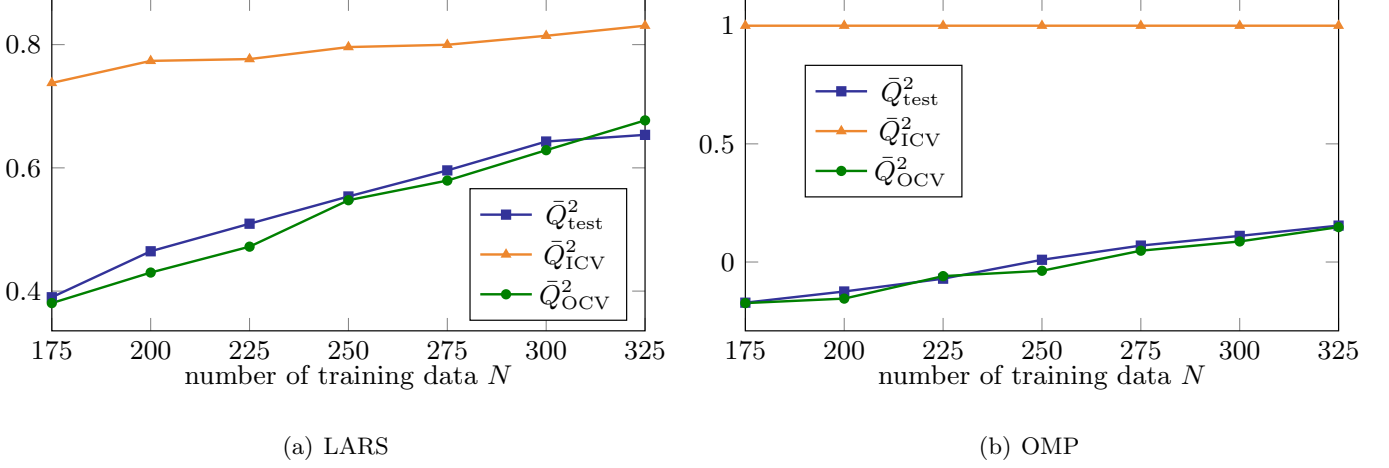

(a) LARS

(b) OMP

Figure 7: Specific absorption rate - mean of Q^2 over 100 replications versus the increasing number of training data.

Since the size of the available data set is usually small, the number of data for model assessment is preferred to be small so that most of the data is used in the training process. Therefore, the outer CV in cross-model validation is operated by leave-one-out CV. To check the accuracy of outer CV, a portion of data is used for testing. Repeating the training and testing process 100 times, the average determination coefficients are plotted in Fig. 7 as a function of the size of training data set, N, which increases from 175 to 325 while the size of testing data set equals $350 - N$. As observed from the curve of $\bar{Q}^2_{\text{test}}$, the performance of the constructed surrogate model is improved when the size of the experimental design increases. The sparse PCE models constructed by OMP lead to poor prediction performances with $\bar{Q}^2_{\text{test}} < 0.16$ while LARS is a better choice to built the sparse PCE model in this case study. While the inner CV always assesses the PCE models in an overoptimistic way, the assessment by outer CV is close to what is obtained by independent testing. Moreover, the outer CV based on LOOCV slightly underestimates the determination coefficient for most cases meaning that it tends to give a conservative error estimate. Then, the whole set of data is put in the training (including outer CV) process and the obtained $Q^2_{\text{OCV}} = 0.6812$ with LARS, which is smaller than Q^2_{ICV} in Fig. 5(a).
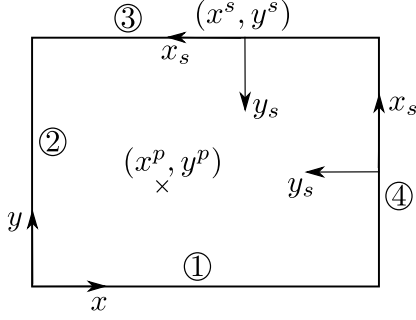
Figure 8: Specific absorption rate - formulation of input uncertainties in the local polar coordinate system of the WiFi box.

5.2 Data preprocessing and refined surrogate model based on sensitivity analysis

The reflections by walls, ground and ceiling are ignored for the computation of SAR. Therefore, the SAR distribution is determined by the *relative position and orientation* of the human model vs. the WiFi box. Then, the considered input uncertainties can be represented in the local coordinate system of the WiFi box, as sketched in Fig. 8, where the y axis of the local coordinate system, denoted by y_s, is always orthogonal to the box surface and directed towards the inner space. As a result, the positive y_s axis is opposite to the y axis (x axis) when the box is placed on the 3rd wall (4th wall). Denote (x^s, y^s), (x^p, y^p) as Cartesian coordinates of the box and human model in the x, y plane of the global coordinate system. Keeping z^s to represent the height of the box, the other five uncertain variables can be reduced to three through the following formulas:

$$
\begin{aligned}
&x_s^p = x^p - x^s, y_s^p = y^p - y^s, \theta_s^p = \theta^p, \text{ if the box on the 1st wall,} \\
&x_s^p = y^s - y^p, y_s^p = x^p - x^s, \theta_s^p = (\theta^p + 90°)\,\text{mod.}\,(360°), \text{ if the box on the 2nd wall,} \\
&x_s^p = x^s - x^p, y_s^p = y^s - y^p, \theta_s^p = (\theta^p + 180°)\,\text{mod.}\,(360°), \text{ if the box on the 3rd wall,} \\
&x_s^p = y^p - y^s, y_s^p = x^s - x^p, \theta_s^p = (\theta^p + 270°)\,\text{mod.}\,(360°), \text{ if the box on the 4th wall,}
\end{aligned}
\tag{11}
$$

where "mod." denotes the modulo operator. (r_s^p, ψ_s^p), as the polar coordinates of (x_s^p, y_s^p), denote the relative distance and angle, respectively. Thus, four variables r_s^p, ψ_s^p, θ_s^p, z^s can fully represent the uncertainty of inputs.

Performing outer CV by LOOCV, the accuracy of model assessment is checked by independent testings. Varying the size of training data (the size of the testing data as well), the averaged determination coefficients over 100 replications are shown in Fig. 9. Model assessment by inner CV reveals optimistic (especially with OMP) whereas the outer CV shows unbiasedness. Taking the whole set of data for training, the outer CV yields $Q_{\text{OCV}}^2 = 0.9024$ with LARS and 0.8337 with OMP, which are much larger than the values obtained with six input variables.

Global sensitivity analysis is performed through the computation of total Sobol' indices, which measure the total contribution (summation of individual and interacted ones) of an input variable to the response uncertainty. Based on the PCE model constructed with the whole set of data, the total Sobol' indices are
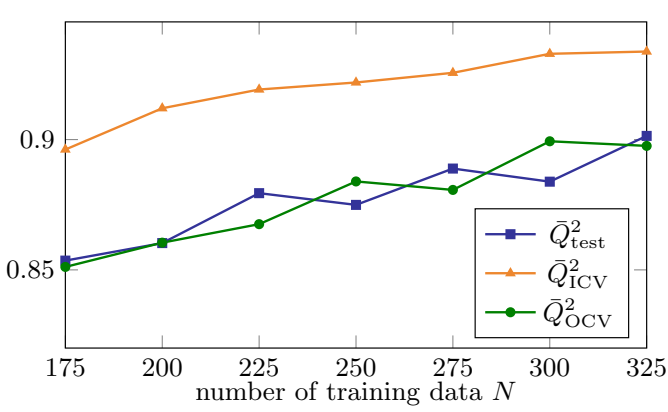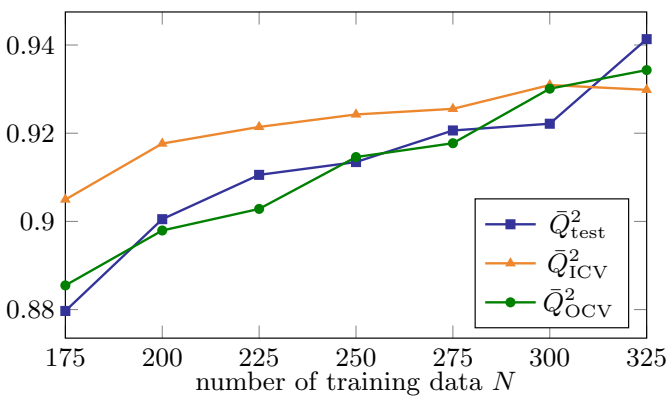
(a) LARS
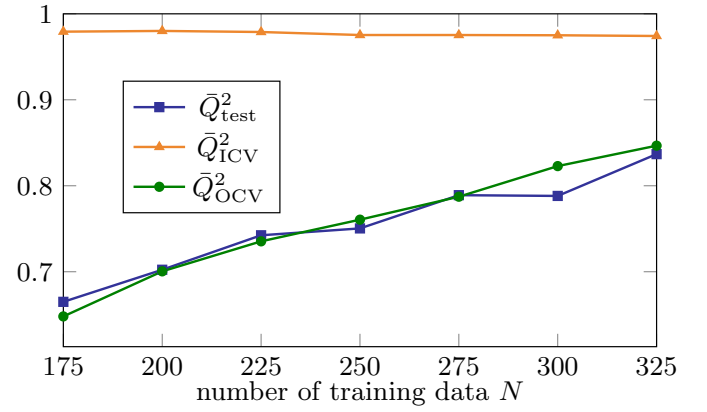
(b) OMP

Figure 9: Specific absorption rate - mean of Q^2 over 100 replications versus N with the reformulated data.

	r_s^p	ψ_s^p	z^s	θ_s^p
LARS	0.9772	0.0405	0.2039	0.0404
OMP	0.9773	0.0677	0.2324	0.0555

Table 4: Specific absorption rate - total Sobol' indices computed with the PCE model based on the whole set of reformulated data.

computed from PCE coefficients [36] and provided in Table 4. The PCE models trained with LARS and OMP support the same conclusion that the uncertainty of response is mainly contributed to by r_s^p and z^s and the effects of ψ_s^p and θ_s^p can be neglected.

(a) LARS
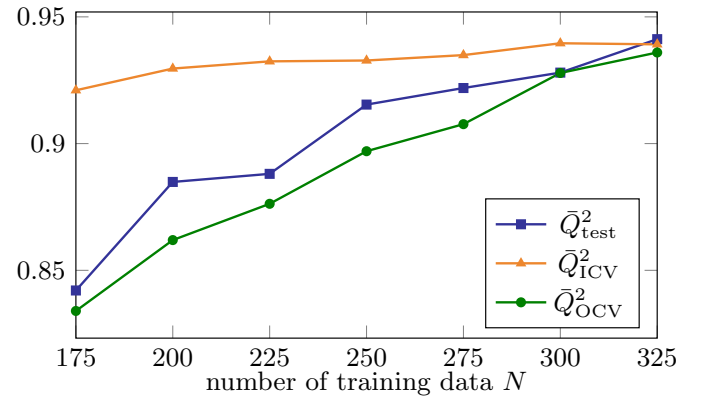
(b) OMP

Figure 10: Specific absorption rate - mean of Q^2 over 100 replications versus N with surrogate models only considering variables r_s^p and z^s.

Following the result of the sensitivity analysis, surrogate models are constructed only considering two
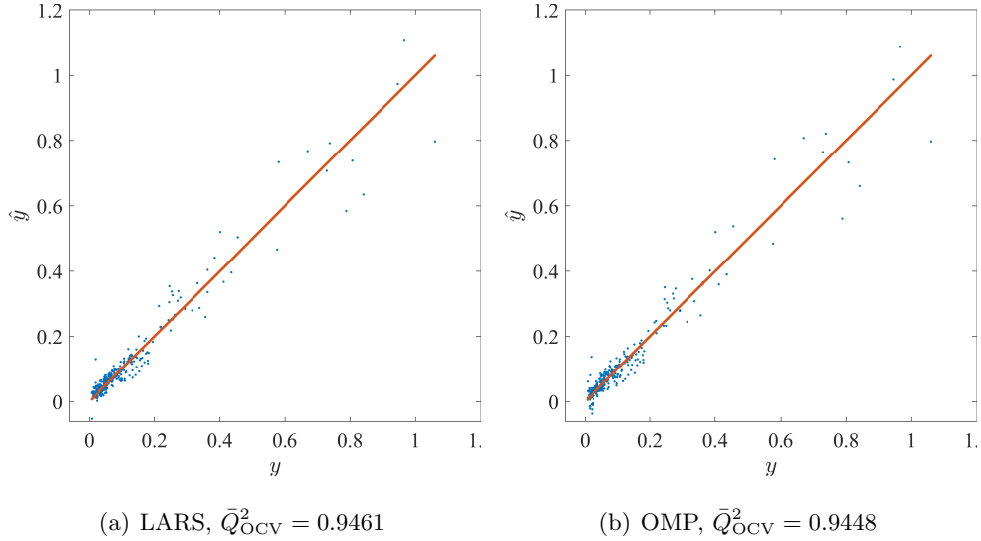
(a) LARS, $\bar{Q}^2_{\mathrm{OCV}} = 0.9461$ (b) OMP, $\bar{Q}^2_{\mathrm{OCV}} = 0.9448$

Figure 11: Specific absorption rate - outer cross validation by leaving one sample out when only r^p_s and z^s are considered as uncertain variables.

uncertain variables, r^p_s and z^s. The accuracy of mode assessment by inner and outer CV is checked by independent testings, as shown in Fig. 10. With LARS, while inner CV underestimates the CV error for most cases, the deviation of assessments by inner CV is smaller than 0.03. With OMP, the deviation for assessments by inner CV becomes larger (e.g. the deviation ≈ 0.08 when $N = 175$) but outer CV still closely follows the reference value (provided by independent testings). Making use of the whole set of data to construct the sparse PCE model and running outer CV by leaving one sample out, the prediction of all samples left out is shown in Fig. 11, where unbiased estimations are observed. The associated determination coefficient equals 0.9461 with LARS and 0.9448 with OMP, which are larger than the values obtained with four input variables, as described before.

6 Conclusions

The human exposure by a WiFi box inside a room is statistically analyzed by constructing a surrogate model based on polynomial chaos expansion. While the samples of the input space are determined by Latin hypercube sampling, the response is generated by combining measurements of scattered fields and an in-house FDTD code through the computation of a Huygens box. Based on a limited sample set, the sparse PCE model is constructed and assessed by cross validation (CV) (including inner CV and outer CV). After the pre-processing of the input variables and the sensitivity analysis, it is shown that the performance of the surrogate models can be improved.

Inner CV is commonly used to perform model selection and assessment during the construction of sparse PCE models. However, since the CV error is optimized during the model selection, the model assessment by inner CV is found over-optimistic (especially when the surrogate model does not perform well due to a very small experimental design). The introduction of outer CV in the methodology of

cross-model validation allows one to get more accurate estimation of the validation error.

The electromagnetic reflections by walls, ceiling and ground are not considered in the present work. As a result, the six input variables describing the problem can be fully represented by four variables, which largely reduce the complexity of the surrogate model. Next, the wave reflections will be taken into account. Due to mutual scattering, the physics behind may become too complicated to approximate the physical system by a single surrogate model. Then, the input space may be divided into subregions, for each of which a surrogate model can be built [16; 33].

Acknowledgments

References

[1] Anderssen, E., K. Dyrstad, F. Westad, and H. Martens (2006). Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst. 84* (1-2), 69–74.

[2] Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Stat. Surv. 4*, 40–79.

[3] Baumann, D. and K. Baumann (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminf. 6* (1), 47.

[4] Blatman, G. and B. Sudret (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Prob. Eng. Mech. 25* (2), 183–197.

[5] Blatman, G. and B. Sudret (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *J. Comput. Phys. 230* (6), 2345–2367.

[6] Chapelle, O., V. Vapnik, and Y. Bengio (2002). Model selection for small sample regression. *Mach. Learn. 48* (1-3), 9–23.

[7] Chobineh, A., Y. Huang, T. Mazloum, E. Conil, and J. Wiart (2019). Statistical model of the human RF exposure in small cell environment. *Ann. Telecommun. 74* (1-2), 103–112.

[8] Christ, A., W. Kainz, E. G. Hahn, K. Honegger, M. Zefferer, E. Neufeld, W. Rascher, R. Janka, W. Bautz, J. Chen, et al. (2009). The virtual family - development of surface-based anatomical models of two adults and two children for dosimetric simulations. *Phys. Med. Biol. 55* (2), N23–38.

[9] Conil, E., A. Hadjem, A. Gati, M.-F. Wong, and J. Wiart (2011). Influence of plane-wave incidence angle on whole body and local exposure at 2100 MHz. *IEEE Trans. Electromagn. Compat. 53* (1), 48–52.

[10] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat. 7*(1), 1–26.

[11] Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *Ann. Stat. 32*(2), 407–499.

[12] Ferreira, J. B., Á. A. A. De Salles, and C. E. Fernández-Rodriguez (2015). SAR simulations of EMF exposure due to tablet operation close to the user's body. In *SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC 2015)*, Porto de Galinhas, Brazil, pp. 1–5.

[13] Findlay, R. and P. Dimbylow (2010). SAR in a child voxel phantom from exposure to wireless computer networks (Wi-Fi). *Phys. Med. Biol. 55*(15), N405.

[14] Findlay, R. and P. Dimbylow (2012). SAR in children from exposure to wireless local area networks (WLAN). In *Asia-Pacific Symposium on Electromagnetic Compatibility (APEMC 2012)*, Singapore, pp. 733–736.

[15] Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*, Volume 1. New York, USA:: Springer series in statistics.

[16] Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *Ann. Stat. 19*(1), 1–67.

[17] Geisser, S. (1975). The predictive sample reuse method with applications. *J. Am. Stat. Assoc. 70*(350), 320–328.

[18] Hald, J. (1988). *Spherical Near-Field Antenna Measurements*, Volume 26. London, UK: Peter Peregrinus Ltd.

[19] Harper, W. V. and S. K. Gupta (1983). Sensitivity/uncertainty analysis of a borehole scenario comparing Latin hypercube sampling and deterministic sensitivity approaches. Technical report, Battelle Memorial Inst.

[20] Hirata, A., S. Kodera, J. Wang, and O. Fujiwara (2007). Dominant factors influencing whole-body average SAR due to far-field exposure in whole-body resonance frequency and GHz regions. *Bioelectromagnetics 28*(6), 484–487.

[21] Ishigami, T. and T. Homma (1990). An importance quantification technique in uncertainty analysis for computer models. In *1st International Symposium on Uncertainty Modeling and Analysis*, Maryland, USA, pp. 398–403.

[22] Kersaudy, P., B. Sudret, N. Varsier, O. Picon, and J. Wiart (2015). A new surrogate modeling technique combining kriging and polynomial chaos expansions–application to uncertainty analysis in computational dosimetry. *J. Comput. Phys. 286*, 103–117.

[23] Lacroux, F., E. Conil, A. C. Carrasco, A. Gati, M.-F. Wong, and J. Wiart (2008). Specific absorption rate assessment near a base-station antenna (2,140 MHz): Some key points. *Ann. Telecommun. 63*(1-2), 55–64.

[24] Lebrun, R. and A. Dutfoy (2009). A generalization of the Nataf transformation to distributions with elliptical copula. *Prob. Eng. Mech. 24*(2), 172–178.

[25] Liorni, I., M. Parazzini, S. Fiocchi, and P. Ravazzani (2015). Study of the influence of the orientation of a 50-Hz magnetic field on fetal exposure using polynomial chaos decomposition. *Int. J. Environ. Res. Public. Health 12*(6), 5934–5953.

[26] Liorni, I., M. Parazzini, N. Varsier, A. Hadjem, P. Ravazzani, and J. Wiart (2016). Exposure assessment of one-year-old child to 3G tablet in uplink mode and to 3G femtocell in downlink mode using polynomial chaos decomposition. *Phys. Med. Biol. 61*(8), 3237.

[27] Liu, Z., D. Lesselier, B. Sudret, and J. Wiart (2018). Surrogate modeling based on resampled polynomial chaos expansions. *arXiv:1810.09116*.

[28] Marelli, S. and B. Sudret (2019). UQLab user manual–polynomial chaos expansions, Report UQLab-V1.2-104. *Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich*.

[29] Martínez-Búrdalo, M., A. Martin, A. Sanchis, and R. Villar (2009). FDTD assessment of human exposure to electromagnetic fields from WiFi and bluetooth devices in some operating situations. *Bioelectromagnetics 30*(2), 142–151.

[30] Meyer, F. J., D. B. Davidson, U. Jakobus, and M. A. Stuchly (2003). Human exposure assessment in the near field of GSM base-station antennas using a hybrid finite element/method of moments technique. *IEEE Trans. Biomed. Eng. 50*(2), 224–233.

[31] Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *27th Asilomar Conference on Signals, Systems and Computers (ACSSC 1993)*, Pacific Grove, USA, pp. 40–44.

[32] Pinto, Y. and J. Wiart (2017). Statistical analysis and surrogate modeling of indoor exposure induced from a WLAN source. In *11th European Conference on Antennas and Propagation (EUCAP 2017)*, pp. 806–810.

[33] Quinlan, J. R. et al. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, Volume 92, Hobart, Tasmania, pp. 343–348.

[34] Roy, K., S. Kar, and R. N. Das (2015). *A primer on QSAR/QSPR modeling: Fundamental concepts*. Springer.

[35] Sudret, B. (2007). 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*.

[36] Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Syst. Safety 93*(7), 964–979.

[37] Wiart, J. (2016). *Radio-Frequency Human Exposure Assessment: From Deterministic to Stochastic Methods.* John Wiley & Sons.

[38] World Health Organization and others (2010). WHO research agenda for radiofrequency fields.

[39] Wyss, G. D. and K. H. Jorgensen (1998). A users guide to LHS: Sandias Latin hypercube sampling software, Report SAND98-0210,. *Sandia National Laboratories, Albuquerque, USA.*

[40] Xiu, D. and G. E. Karniadakis (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput. 24*(2), 619–644.

[41] Zeghnoun, A., F. Dor, and A. Grégoire (2010). Description du budget espace-temps et estimation de l'exposition de la population française dans son logement. *Institut de Veille Sanitaire.*