UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

Inferencia bayesiana asistida por modelos subrogados para el análisis de la cuantificación de la incertidumbre en problemas de ingeniería de cálculo intensivo

---

José Carlos García Merino

Programa de doctorado
Modelización y Experimentación en Ciencia y Tecnología

Conformidad de la directora de tesis:
Dra. Carmen Calvo Jurado

Conformidad del codirector de tesis:
Dr. Enrique García Macías

Esta tesis cuenta con la autorización de la directora y codirector de la misma, y de la Comisión Académica del programa. Dichas autorizaciones constan en el Servicio de la Escuela Internacional de Doctorado de la Universidad de Extremadura.

2024

UNIVERSIDAD DE EXTREMADURA

PHD THESIS

Surrogate modelling assisted bayesian inference for uncertainty quantification analysis of computationally intensive engineering problems

---

José Carlos García Merino

Supervisors:

Dra. Carmen Calvo Jurado

Dr. Enrique García Macías

Programa de doctorado
Modelización y Experimentación en Ciencia y Tecnología

2024

UNIVERSIDAD DE EXTREMADURA

DOCTORAL THESIS

Surrogate modelling assisted Bayesian inference for Uncertainty Quantification analysis of computationally intensive engineering problems

Author:
José Carlos García Merino

Supervisor:
Carmen Calvo Jurado
Enrique García Macías

March 14, 2024

Resumen

Esta tesis se enmarca en el ámbito de la Cuantificación de la Incertidumbre, un campo multidisciplinar con una marcada orientación práctica que integra conceptos provenientes de diversas disciplinas como la Matemática Aplicada, la Ingeniería, la Computación y la Estadística. La Cuantificación de la Incertidumbre puede definirse como el proceso de análisis de las incertidumbres asociadas a las predicciones basadas en modelos matemáticos. Una de las principales dificultades inherentes a la realización de tales análisis es que una buena parte de los modelos empleados utilizados en ingeniería son altamente demandantes computacionalmente, lo que resulta en que muchas técnicas comunes de análisis como la simulación Montecarlo o los algoritmos MCMC resulten inviables. Una estrategia capaz de sortear estas dificultades consiste en la sustitución los modelos originales por metamodelos, es decir, aproximaciones computacionalmente ligeras del modelo original. No obstante, en entornos prácticos son frecuentes los problemas mal condicionados y los comportamientos no lineales de los modelos involucrados, los cuales comprometen la efectividad del enfoque propuesto. Además, muchos fenómenos incorporan una incertidumbre intrínseca no directamente observable o medible, cuya naturaleza requiere el empleo de simuladores estocásticos, añadiendo una capa de dificultad al tratamiento estadístico del problema en cuestión. La presente investigación aborda el análisis de la incertidumbre en el marco de problemas de ingeniería complejos que enfrentan los desafíos mencionados anteriormente. Con este fin, se propondrán diferentes metamodelos capaces de reemplazar eficientemente los modelos computacionalmente intensivos originales. Además, se presentará una nueva técnica para la construcción de metamodelos en contextos estocásticos.

Abstract

This thesis is situated within the realm of Uncertainty Quantification, a multidisciplinary field with a pronounced practical orientation that integrates concepts from various disciplines such as Applied Mathematics, Engineering, Computer Science, and Statistics. Uncertainty Quantification can be defined as the process of analysing the uncertainties associated with predictions based on mathematical models. One of the primary challenges inherent in such analyses is that a significant part of the models employed in engineering are highly computationally demanding, rendering many common analysis techniques such as Monte Carlo simulation or MCMC algorithms unfeasible. A strategy capable of overcoming these difficulties involves substituting the forward models with metamodels, i.e., computationally lightweight approximations of the original model. However, in practical environments, ill-conditioned problems and nonlinear behaviours of the involved models are common, compromising the effectiveness of the proposed approach. Additionally, many phenomena incorporate intrinsic uncertainty that is not directly observable or measurable, whose nature requires the use of stochastic simulators, adding a layer of difficulty to the statistical treatment of the problem at hand. This research addresses the analysis of uncertainty within the framework of complex engineering problems that face the aforementioned challenges. To this end, various metamodels capable of efficiently replacing the computationally intensive original models will be proposed. Additionally, a new technique for constructing metamodels in stochastic contexts will be presented.

Acknowledgements

First and foremost, I owe my deepest gratitude to my thesis supervisors, Carmen and Enrique. Thank you Carmen for encouraging me to pursue a PhD, and for your patience and motivation in its toughest moments. Your dedication and commitment have made the completion of this thesis possible. Working hand by hand with you has been a privilege and an honour. Thank you also Enrique for sharing with me your passion for research and for being there to inspire me every time it seemed that we were not moving forward. Your extraordinary guidance have allowed me to acquire new knowledge and skills to grow as a scientist and, more importantly, as a person.

Secondly, I would like thank my colleagues in the Mathematics Department, both in Cáceres and in Badajoz. In particular, I would like to acknowledge Inés and Manolo Molina, who have always ensured my well-being in my teaching role, and Alberto, who has been with me since my undergraduate studies and whose support during these years has been fundamental to my growth as a mathematician. Special thanks to Lucía, whose kindness, generosity and patience over more than a decade I will be forever indebted to, and for which I cannot sufficiently express my appreciation for.

I can't but express my gratitude to all the people at LUT: both to Professor Heikki Haario, who kindly gave me the opportunity to work together with him and his team, and to the rest of the international crew: Hanz, Lam, Miracle, Rodrigo, Galdi.... And all the other wonderful people who helped me to make unforgettable memories.

I would also like to extend this acknowledgements to my parents and to all the people who went out of their way to assist me, but are not explicitly mentioned in this brief note. Last but not least, I would especially like to thank Elena for her understanding, sense of humour, unconditional affection and, above all, for allowing me to share a life project with her.

Contents

Chapter 1

Introduction

1.1 Motivation

One of the primary goals of mathematicians, scientists and engineers engaged in understanding physical phenomena has been to develop mathematical models capable of describing these realities. The accelerated development of computer science in recent years has made possible to explore many challenges in science and engineering through simulation, tackling problems that were once considered intractable. Computer simulations provide a comprehensive approach to study systems that have been effectively modelled, but whose solution is analytically unsolvable [1]. Computational models have found successful applications in a wide range of disciplines, including chemistry, economics, health, materials science, or civil engineering, just to mention a few [2–6]. Nonetheless, not all the parameters and conditions required for the simulations are always known nor experimentally measurable, and this uncertainty impacts on the simulation outputs. The process of quantifying and analysing the uncertainty in mathematical models constitutes the central concern of Uncertainty Quantification (UQ) [7].

From a general point of view, a mathematical model is nothing more than a 'black box' function $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^M \rightarrow \mathbb{R}^N$ mapping some input parameters (often referred to as *design* variables) from a parameter space $\mathcal{D}$ into certain quantities of interest (QoI) [8]. Within the realm of UQ, two major types of questions can be distinguished: forward and inverse problems. The former challenge, also known in the literature as Uncertainty Propagation (UP), involves analysing the propagation of parameter uncertainties through the computational model. During this process, uncertainties in the model input parameters are transformed into uncertainties of the output parameters. Theoretically, such problems are well solved by brute force through Monte Carlo Simulation (MCS) [1]. Nevertheless, as we discuss below, this is rarely feasible in practice due to computational constraints when dealing with complex numerical models. On the other hand, an inverse problem occurs when the QoI is known, but the model parameters responsible for that output are not. Moreover, in many instances, it is even unclear if the model is appropriate at all to represent the behaviour of the system. In these cases, model selection is an additional challenge [9]. It should be noted that inverse problems involve inferring potential causes by examining the observed or measured effect. This type of task is typically ill-posed and ill-conditioned, meaning that small changes in the effects can result

in significant changes in the associated solution, which can even lead to a loss of uniqueness of the solution. [10]. As a consequence, inverse problems are generally much more difficult to solve.

In practice, addressing UQ problems poses significant challenges, particularly when computationally intensive simulators are utilised to model the physical reality. These models have found widespread application in diverse fields such as structural mechanics [11] or materials science [12], employing popular techniques like the Finite Element Method (FEM), Boundary Element Methods (BEMs) or mesh-free numerical simulators. However, such simulators can take anywhere from a few minutes to several hours to run [13]. As a consequence, since MCS suffers from an extremely small convergence rate of $\mathcal{O}(1/\sqrt{n})$ [14], the employment of such complex simulators makes unfeasible the straightforward application of this technique.

A simple and efficient approach to overcome MCS difficulties is the utilisation of surrogate models. Roughly speaking, a surrogate model, or *metamodel* is a function that emulates the behaviour of a simulator but at negligible computational cost per run. Mathematically, given a model $\mathcal{M}$, a metamodel can be defined as a function $\widehat{\mathcal{M}} : \mathcal{D} \subset \mathbb{R}^M \rightarrow \mathbb{R}^N$ such that $\widehat{\mathcal{M}}(\mathbf{x}) \approx \mathcal{M}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{D}$. For the sake of simplicity, in this thesis, it will be assumed that $N = 1$. Note that this assumption holds without loss of generality, as fitting a surrogate model for each QoI would suffice otherwise. Currently, surrogate models have garnered substantial attention in scientific literature, owing to their flexibility and applicability to a wide variety of problems (see [15–19], to point out a few examples).

From the previous discussion, the main motivation of this thesis is to develop an efficient surrogate modelling framework capable of approximating computationally intensive engineering models, handling highly non-linear response surfaces, and accounting for intrinsic uncertainties. The computational efficiency of the surrogate modelling approach to be developed should allow conducting UP and Bayesian parameter inference analyses with computational times compatible with quasi real-time analysis.

1.2 State of the art review

In this section, an overview of the two primary areas investigated in this thesis is provided. Firstly, a classification of metamodelling techniques is presented, along with a discussion of their respective strengths and weaknesses. Subsequently, recent literature regarding Bayesian inference is examined, with a particular focus on the integrated application of surrogate models and Bayesian inference techniques

1.2.1 Surrogate modelling

As pointed out in Section 1.1, many classical statistical methodologies such as Monte Carlo analysis require the simulation of thousands of scenarios, becoming unfeasible when applied to computationally intensive models. In this light, surrogate modelling

stands out as a versatile strategy to circumvent the challenges associated with resource-intensive simulators. There exist a myriad of surrogate modelling techniques, each one based on a different mathematical approach. In a first step, metamodelling methods can be classified into three broad categories [20]: (i) reduced-order models (ROMs) [21]; (ii) multi-fidelity methods [22]; and (iii) data-driven or responsive surface methods (RSMs). ROMs techniques are developed by projecting the governing equations of the original model onto a low-dimensional subspace. Hence, this class of methods have the advantage of retaining the physics underlying the model. However, a major drawback of this approach in practical applications is its limited applicability, as it requires access to the model governing equations, which is often not available when using commercial software. Secondly, multi-fidelity methods are built by simplifying the underlying physics or reducing the numerical resolution. The main challenge of these techniques is their high case-dependency and the requirement for specialised expertise in order to balance prediction accuracy and computational efficiency. Finally, RSMs consist in employing a reduced number of realisations of the forward model to characterise behaviour of the system across the entire parameter space. It should be highlighted that, unlike multi-fidelity methods, they do not request expert knowledge of the system under study. Moreover, even if the QoI of the forward model is the output of a locked commercial software, RSMs only require the input/output relationship provided by the original model. As a consequence, the key advantage of these approaches lies in the fact that the forward model does not need to be modified and can be therefore treated as a black box [23]. Such features have fostered rapid developments of RSMs in recent years as a non-intrusive technique with great flexibility in a wide range of applications. The process of building a response surface surrogate model is depicted schematically in Fig. 1.1. After sampling the parameter space (a), the corresponding model evaluations are obtained by executing the forward model on selected points (b). This step typically represents the most computationally intensive phase of the process. It is worth noting that access to the source code of the forward model is not required; rather, the ability to iteratively run the forward model for all the sampling points suffices. Finally, the responses of the model at the sampling locations are used to construct the surrogate model (c).

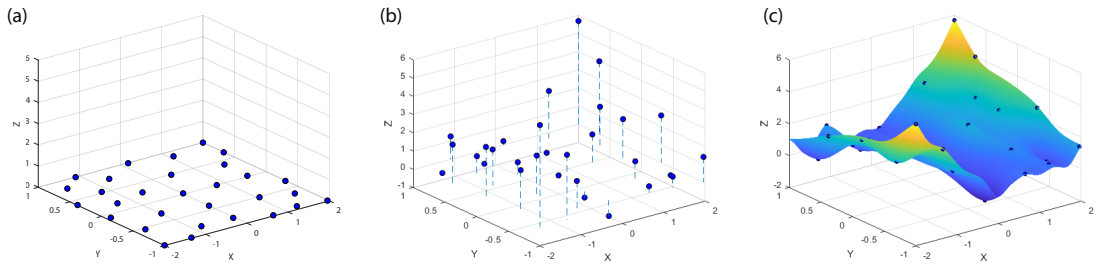

FIGURE 1.1: Surrogate model building process: exploration of parameter space (a), evaluation of forward model (b), and estimation of the response surface (c).

Some of the most popular response surface methodologies found in recent scientific literature include Kriging [12], Polynomial Chaos Expansion (PCE) [24], Radial Basis Functions (RBF) [25], Support Vector Machines (SVM) [26], and Neural Networks (NN) [27]. The selection of one technique over another depends on many factors, including the nature of the problem, the available computational resources, the objective of the analysis (prioritising speed or accuracy), or even the prior knowledge of the practitioner about surrogate modelling. For this reason, a substantial amount of research has been carried out over the last two decades comparing different metamodelling techniques. For instance, in the early 2000s, the works of Jin et al. [28, 29] were the first to compare various popular methodologies for building surrogate models. Their findings showed that RBF and Kriging outperformed the other alternatives. By contrast, the authors concluded that Kriging suffered from a slow tuning process, especially when large samples are used. This is explained by the need, during the tuning process, to repeatedly invert a matrix whose size corresponds to the number of observed realisations of the forward model. Posterior analyses persisted in emphasising Kriging as the most accurate method [30, 31]. In the same line, works by Van Gelder et al. [32] and Kroetz et al. [15] suggested Kriging and NN as the best approaches. Li and co-authors found SVM as the top technique, but with Kriging ranking a very close second [33]. In a recent comparison of popular metamodelling methods, Østergård and co-authors [16] further supported the superiority of Kriging, while highlighting the approach sensitivity to model options. After analysing a comprehensive and diverse set of problems, Kianifar and Campean [18] recommended the use of Kriging, stressing its robustness against problem scale or non-linearities in the response surface. Along the same lines, in an up-to-date review of more than 200 papers, Alizadeh et al. [17] pointed out that Kriging outperforms all other metamodelling techniques in terms of accuracy for different types of problems .

Another widespread methods employed in engineering practice for constructing surrogate models are PCEs. This family of techniques aims to provide a functional approximation of a computational model through its representation on a suitably built basis of polynomial functions [34]. PCE has been successfully applied in a wide variety of engineering contexts, including materials analysis [35, 36] and the assessment of civil structures like bridges [37] and dams [38, 39]. PCE stands out as an excellent choice for constructing surrogate models. Firstly, PCE, by its intrinsic design, enables the natural incorporation of various probability distributions (uniform, Gaussian, etc.) into the input parameter space. Secondly, PCE allows for the calculation of leave-one-out (LOO) error estimates analytically from the model coefficients themselves. This feature enables control over potential overfitting without requiring additional forward model evaluations. This is particularly advantageous in practical contexts where running a single simulation takes hours of computation. Moreover, a PCE model can be easily post-processed to obtain further information about the QoI, including the mean and variance of the model as well as a sensitivity analysis of the different model inputs [40]. Finally,

PCE can be naturally integrated into the Kriging framework, resulting in a more powerful meta-modelling technique, known as *Polynomial Chaos Kriging* (PCK) [41]. On the other hand, PCE faces limitations when applied to complicated problems. Overall, PCE approaches works better when applied to globally smooth problems [24]. Additionally, PCE suffers from the so-called 'curse of dimensionality', meaning that the number of forward model realisations required to construct an accurate metamodel increases exponentially with the dimension of the problem. As a consequence, the performance of the method drops dramatically in high-dimensional contexts (10 or more dimensions).

From the previous state-of-the-art, it is evident that there remain numerous open questions regarding the construction of surrogate models. One of the less explored topics is the approximation of highly nonlinear surfaces. In such problems, commonly used surrogate models such as PCE, Kriging, or even PCK suffer from very low convergence rates, rendering their use ineffective. The results presented in Section 3.2 demonstrate an efficient approach capable of overcoming these limitations. Another recent challenge in metamodelling literature is the emergence of stochastic simulators in engineering problems, which incorporate inherent randomness into their predictions. In this regard, techniques such as Stochastic Kriging (SK) [42] have emerged to approximate these types of models. However, only the simplest version of SK has been successfully implemented. This thesis proposes a novel approach combining SK and PCE, which outperforms standard methods available the literature.

1.2.2 Surrogate model based Bayesian inference

The problem of parameter estimation is a central concern in statistical theory. Inference problems typically emerge based on a set of data (observations of a random phenomenon) and a mathematical model dependent on specific parameters, aiming to shed light on the nature of the underlying process. The theory of statistical estimation addresses the methodologies necessary for obtaining accurate findings of the parameter values. Within statistical theory, there are two fundamental approaches that dominate the current scientific paradigm: the classical or *frequentist* approach and the *Bayesian* approach. The differences between these two perspectives are not only methodological, but fundamentally philosophical. For example, a key difference between frequentist and Bayesian statistics is the treatment of model parameters. Frequentist statisticians consider the true parameters of the model to be unknown but *fixed*, whereas in a Bayesian context these parameters are represented as random variables [43]. After decades of research, it is not possible to conclude that one approach is universally preferable to the other [44]. The primary criticism of Bayesian methodology concerns its susceptibility to the incorporation of subjective information into the data analysis process. This opens the risk of introducing errors or biases that would not otherwise occur [44]. On the other hand, Bayesian practitioners claim that Bayesian statistics enables the inclusion in the data analysis of all relevant information related to the problem, obtaining more natural and coherent results [43, 45].

Regardless of the philosophical aspects, Bayesian inference encountered other challenges in its early stages. Direct Bayesian inference is typically not possible, as the mathematical equations governing the problem are usually high-dimensional and analitically intractable. For much of the 20th century, drawing inferences in Bayesian statistics was limited to a few specific theoretical cases. This constraint played a significant role in leading many statisticians to adopt a frequentist point of view in practice [46]. However, the paradigm shifted in 1990 with the groundbreaking paper of Galfand and Smith [47], that quickly popularised Markov Chain Monte Carlo (MCMC) methods. MCMC are a family of simulation-based techniques that effectively address the issues above. The basic idea of every MCMC procedure is to exploit the law of large numbers by constructing a Markov chain with a stationary distribution proportional to the posterior distribution. As a result, a sample of the joint probability distribution function (PDF) of the model parameters can be obtained by collecting the states of the chain [48]. Sampling a PDF via MCMC algorithms has significant advantages, including their simplicity of implementation and their applicability to a wide range of problems. Furthermore, MCMC algorithms are guaranteed to converge if the sample size is large enough [49]. These features, together with the development of computing, favoured the widespread use of the first MCMC algorithms such as *Metropolis-Hastings* and *Gibbs sampler* [50]. However, since these methods are simulation-based and require a significant number of iterations to achieve success, they can suffer from slow convergence rates [51], especially if the simulator is computationally demanding.

In order to improve the efficiency of the most elementary MCMC techniques, several alternatives have been proposed in the literature in recent years. For instance, Hamiltonian Monte Carlo (HMC) takes advantage of Hamiltonian dynamics to incorporate information about the local geometry of the distribution, reducing the correlation between successive samples. This enables the PDF to be explored more effectively, thus reducing the total number of samples needed [52]; Particle MCMC (PMCMC) brings Sequential Monte Carlo (SMC) into the realm of MCMC methodology and is typically applied in the context of time series analysis where the likelihood of the observed data is not analytically tractable [53]; Adaptive Metropolis tackles the problem of misspecification of some user-defined settings in the Metropolis Hastings framework by allowing an automatic tuning of the algorithm [54]. As a result of this development, Bayesian inference via MCMC has become a common standard in a broad range of disciplines, such as structural identification [55], epidemiology [56], hydrology [57] or economics [58], just to mention a few.

Despite these advances, conducting Bayesian inference on sophisticated models that consume a large amount of computational resources remains challenging. In this context, surrogate models offer an efficient solution to perform Bayesian inference in a cost-effective manner while retaining the accuracy of high-fidelity models. For instance, PCE metamodels have been successfully applied for Bayesian model calibration in subsurface flow systems [59]. Analogously, Schneider and co-authors [60] employed PCE surrogate models to characterise the mechanical properties of a cross-laminated timber plate. In

a similar vein, Kriging has also been effectively employed in combination with MCMC techniques in multiple contexts. The work of Fursov et al. [61] demonstrates the joint use of Kriging and HMC to perform inference on reservoir models. Equally noteworthy, Ierimonti and co-authors [62] proposed a Kriging-based conjugate Bayesian identification methodology for online damage identification of an instrumented monumental building, the Consoli Palace in Gubbio (Italy). Finally, in the context of chemical engineering, Del val et al. [63] proposed an original approach employing a Kriging surrogate of the likelihood function in the Bayesian inference instead of approximating the forward model directly. Nonetheless, a common limitation of MCMC algorithms arises when exploring multimodal PDFs characterised by modes separated by regions of low probability density. This phenomenon often slows down convergence, leading to increased computational demands. To address this issue, this thesis proposes a novel approach that combines surrogate models with the Delayed Rejection Adaptive Metropolis (DRAM) algorithm developed by Haario et al. [64] for Bayesian parameter estimation of ill-posed problems.

1.3 Objectives

The overall purpose of this research is to provide an advance in the development of efficient surrogate models of computationally burdensome engineering problems together with their subsequent application to fast uncertainty quantification and Bayesian inference analyses. This general objective is divided into four sub-objectives:

Obj. 1: Review the state-of-the-art of response surface meta-modelling techniques, and development of an innovative surrogate model capable of handling highly non-linear response surfaces.

Obj. 2: Development of a novel surrogate-based Markov Chain Monte Carlo (MCMC) approach for parameter inference applications.

Obj. 3: Implementation of best suited surrogate model for the uncertainty quantification and parameter inference of computationally burdensome engineering problems, including the homogenisation of composite materials and the governing equations of hydrogen diffusion in metallic alloys.

Obj. 4: Extension of the surrogate modelling approach to be developed in Obj. 1 for the metamodelling of stochastic simulators.

1.4 Organisation of the thesis

This thesis has been structured in four chapters and three appendices as described below:

- Chapter 1 constitutes a brief overview of the state of the art of the fields of surrogate modelling and (surrogate model assisted) Bayesian inference, followed by a summary of the scientific contributions resulting from this work.

- Chapter 2 provides a comprehensive overview of the main theoretical results necessary for an easy understanding of the contributions. Specifically, this chapter includes a review of the fundamental concepts of surrogate modelling theory in Section 2.1, as well as a detailed presentation of the main metamodelling techniques employed throughout the thesis in Sections 2.2 and 2.3. In addition, a grounding in Bayesian statistics is provided in Section 2.4.

- Chapter 3 furnishes the main scientific findings achieved during the development of the thesis. Section 3.1 demonstrates the capabilities of surrogate models in order to perform UP analysis in engineering contexts, in particular within the realm of composite materials. Section 3.2 discusses an approach for approximating highly non-linear functions without compromising computational efficiency, and its further application to the hydrogen diffusion phenomenon in metallic alloys.

- Chapter 4 provides the conclusions of the conducted investigation, as well as a discussion of open questions for further research.

- Appendices A, B and C contain respectively the three full papers published in top-tier scientific journals during the development of the thesis.

1.5 Contributions

The pursuit of the objectives in this thesis has yielded several noteworthy contributions to the field of surrogate modelling, as evidenced by peer-reviewed scientific articles and communications at international conferences. In order to highlight their importance they are listed below:

- Peer-reviewed publications.

  - J.C. García-Merino, C. Calvo-Jurado, E. García-Macías, Surrogate modeling of the effective elastic properties of spherical particle-reinforced composite materials, Journal of Mathematical Chemistry 60(8) (2022) 1555-1570.
  - J.C. García-Merino, C. Calvo-Jurado, E. García-Macías, Polynomial chaos expansion for uncertainty propagation analysis in numerical homogenization of 2D/3D periodic composite microstructures, Composite Structures 300 (2022) 116130.
  - J.C. García-Merino, C. Calvo-Jurado, E. Martínez-Pañeda, E. García-Macías, Multielement polynomial chaos Kriging-based metamodelling for Bayesian inference of non-smooth systems, Applied Mathematical Modelling, 116 (2023) 510-531.
  - J.C. García-Merino, C. Calvo-Jurado, E. García-Macías, Sparse polynomial chaos expansion for universal Stochastic Kriging, Computational and Applied Mathematics 444 (2024) 115794.

- Communications at international conferences.
  - Kriging-based numerical approximation of effective elastic properties of reinforced anisotropic materials, *Conference on Mathematical Methods in Science and Engineering Conference (CMMSE)* 2021.
  - 2D and 3D surrogate model comparison for the effective elastic properties of particle-reinforced composite materials, *International Congress on Fundamental and Applied Sciences (ICFAS)* 2021.
  - Cost-efficient modeling for estimating the elastic properties of fiber-reinforced composites, *International Conference on Mechanical Models in Structural Engineering (CMMoST)* 2021*.
  - LAR-PCE based Stochastic Kriging for non parametric problems, *International Conference of Numerical Analysis and Applied Mathematics (ICNAAM)* 2022
  - An optimal sparse basis selection for Stochastic Kriging, *Conference on Mathematical Methods in Science and Engineering Conference (CMMSE)* 2023.
  - Stochastic surrogate modelling of short fiber-reinforced composite materials, *International Conference on Mechanical Models in Structural Engineering (CMMoST)* 2023**.

  *Awarded as second best predoctoral paper.

  **Awarded as best predoctoral paper.

- Research stays
  - Research stay at Lappeenranta-Lahti University of Technology (LUT) from June 12th, 2022 to September 16th, 2022.

Chapter 2

Theoretical background

This chapter provides a theoretical basis for the fundamental issues that are addressed throughout the thesis. The first part discusses the development of surrogate models, including both common features of all meta-modelling approaches as well as a detailed presentation of the main techniques employed in this work: PCE, Kriging, and Polynomial-Chaos based Kriging (PCK). The second part focuses on Bayesian inference, presenting some key theoretical concepts and an overview of the most popular algorithms.

2.1 Surrogate modelling: basics

The standard process for building an RSM surrogate consists of three basic steps. First of all, a set of points in the parameter space of the forward model, known as the *training set* or *Experimental Design* (ED), is selected. Secondly, the model is evaluated on these instances, and a metamodel is fitted based on the collected input/output data. Finally, the performance of the surrogate model is assessed [65]. A schematic representation of the process of fitting a metamodel is depicted in Fig. 1.1. The capabilities of a constructed surrogate model depends on several factors. Selecting an appropriate ED, employing a suitable metamodelling method, and providing adequate specifications for each particular technique can dramatically enhance its accuracy and efficiency.

Sampling the design space, with the aim of getting the maximum amount of information from a limited number of trials, has become a key issue in computer-based simulation [66]. The straightforward rationale behind this is that when the underlying model is complex and time consuming to evaluate, a proper ED becomes crucial to reduce the number of evaluations of the forward model and hence the computational cost of building the surrogate [67]. However, it should be highlighted that in some techniques, such as Kriging, both the fitting and evaluation time of the metamodel are strongly linked to the size of the ED [68]. This can become a potential limitation when working with large sample sizes or when the surrogate model is intended to be evaluated intensively (e.g. for solving successive inference problems with real-time collected data). Some of the most common methods reported in the literature for obtaining a training set are discussed below.

Two of the simplest ED that can be implemented are Monte Carlo (MC) sampling (also known as pure random sampling) and grid sampling [69]. Monte Carlo sampling

is based on the generation of independent random numbers according to the joint PDF of the forward model input parameters [70]. On the other hand, grid sampling consists of subdividing the area of interest using squares and taking samples from the nodes or points of intersection of the grid lines [71]. Fig. 2.1 (a) and (b) depict a grid sample and a MC sample on the 2D unit hypercube of 36 points each, respectively. These techniques offer easy implementation and minimal computational cost, but are often inefficient at exploring the parameter space. In a proper ED, the new points should be far away from existing locations to avoid redundant samples. This goal is referred to as the space-filling objective [72]. As it can be seen in Fig. 2.1, MC sampling concentrates several points in specific areas, while leaving large regions uncovered. On the other hand, grid methods locate a large number of samples at the margins of the parameter space, leaving the inner space less explored. However, it has been reported in the literature that, in general, it is preferable for the ED to cover the entire space rather than focusing on the boundaries [69].

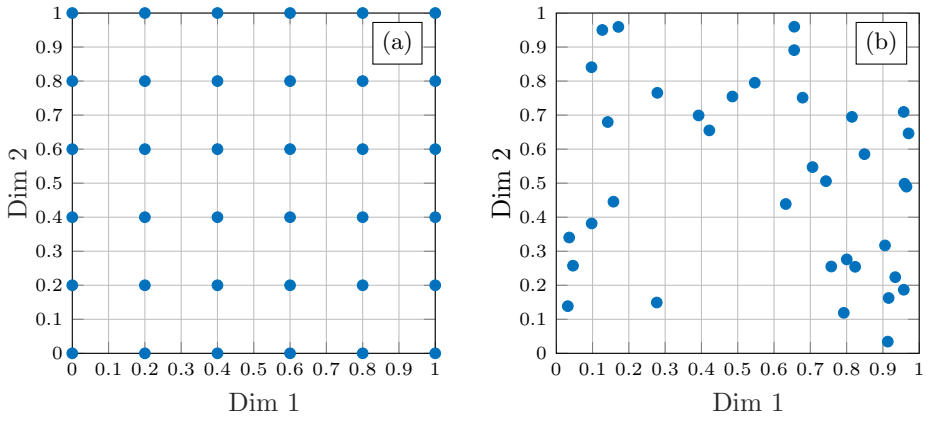

FIGURE 2.1: (a) A 36 points grid sample on 2D unit hypercube. (b) A pure random sample on the same parameter space.

A refined approach for space-filling sampling are *low-discrepancy sequences*, also known as *quasi-random methods*. Low-discrepancy sequences are sequences $\mathcal{S} \subset [0,1]^D$ of points with the property that any subset $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\} \subset \mathcal{S}$ has low discrepancy [73]. Derived from numerical integration, the discrepancy is a measure of the error in volume estimation by Monte Carlo methods. Formally, the discrepancy of a set $P \subset \mathcal{S}$ is defined as [74]:

$$D(P) = \sup_{E} \left| \frac{1}{N} \sum_{i=1}^{N} 1_{\{\mathbf{x}_i \in E\}} - \mathrm{Vol}(E) \right|,$$

where E represents the set of all rectangular boxes of the form $E = [0, t_1] \times ... \times [0, t_D)$, $t_k \in [0,1)$ for all $k \in \{1, ..., D\}$, and $\mathrm{vol}(E)$ is defined as $\mathrm{vol}(E) = \prod_{k=1}^{D} t_k$. Note that the definition can be easily extended without loss of generality to any sequence of points outside the unit hypercube.

For sampling purposes, it follows from the definition that a set of points P has a low

discrepancy if the number of points from P that fall into an arbitrary subset $Q \subseteq P$ is approximately proportional to the measure of that subset [73]. In other words, the discrepancy is a measure of how evenly a set of points fills a space [74]. Low-discrepancy methods provide samples with a high level of uniformity in a multidimensional space, but without statistical independence [66]. Some popular quasi-random designs are Sobol, Van der Corput and Halton sequences [73, 75]. Among these, Sobol's sequence is the most extended technique and its effectiveness has been demonstrated in metamodelling practice (see e.g. [67, 70]). The major drawbacks of these methods are their poor space coverage in high dimensions. Fig. 2.2 depicts the 2D projection of the first 120 points of the Sobol sequence on dimensions 1 and 2 (left) and 8 and 40 (right), respectively. It is clear that the space-filling Sobol's design works well on dimensions 1 and 2, but the projection to dimensions 8 and 40 reflects an extremely inefficient sampling. For a thorough discussion about low-discrepancy sequences, refer to references [73, 76, 77].

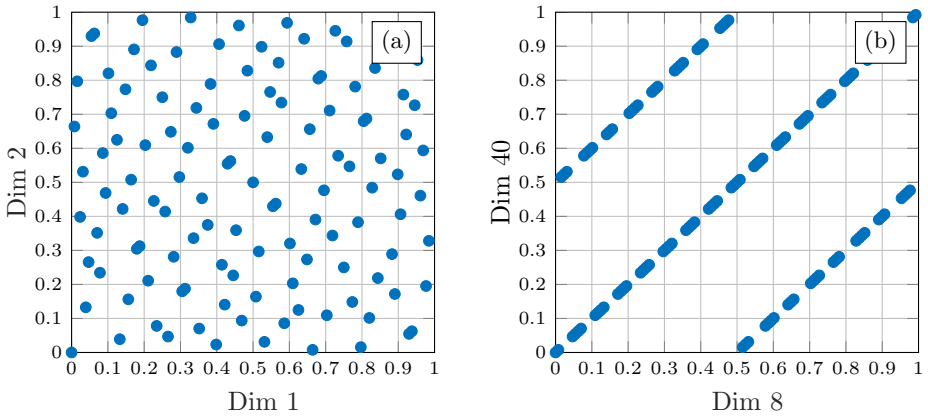

FIGURE 2.2: (a) Projection of a Sobol's sequence 120 points sample on 2D unit hypercube on dimensions 1 and 2. (b) Projection of the same sample on dimensions 8 and 40.

Beyond the techniques mentioned above, the most referred sampling method in the literature by far is the Latin Hypercube Sampling (LHS) [66, 69]. A Latin hypercube is constructed by dividing each dimension of the design space into n equal intervals and placing exactly one point in each interval for each dimension [73]. Intuitively, such configuration is similar to having n rooks on a $n \times n$ chess board without threatening each other. Fig. 2.3 (a) shows a 10-points ED generated by LHS on the unit square. LHS design can be easily generalised to evenly sample a non-uniform probability distribution of the input parameters. To do so, it is sufficient to divide the space in each dimension not into n equal intervals, but into n intervals each one containing $1/n$ of the probability mass of the PDF of the respective parameter [78]. However, despite its flexibility and ease of implementation, LHS is not exempt from limitations. A major drawback of the LHS method is that it does not have any restriction to prevent points from clustering together as the sample size increases. This means that the sample relies on chance to achieve correct space-filling properties. In an extreme scenario, the design could consist of a diagonal of points crossing the parameter space (see [73]). Moreover, LHS provides the entire sample all at once. If the performance of the metamodel is deemed unsatisfactory after

obtaining a training set through LHS, the gathered data and the surrogate model are typically discarded, and the whole process begins again with a new, larger sample from the parameter space [72]. Given that, in general, it is very difficult to determine in advance an appropriate sample size, sequential strategies that add points incrementally on a base ED have been recently proposed [72,79,80]. There are two common approaches to aggregate points to a given sample. The first one, purely geometric, consists of selecting new locations in the least explored regions of the parameter space [73]. In contrast, adaptive techniques aim to add new samples in regions where the surrogate model struggles to emulate the forward model, such as non-linearities or pronounced jumps. These positions, which are not necessarily opposed, are referred to in the literature as *exploration* and *exploitation* [79]. Therefore, an efficient strategy would be to start the sampling step with a reduced ED obtained by LHS and, in case the obtained metamodel accuracy is not sufficient, add new points to the training set applying a sequential method.

In a recent paper, Fugh and co-authors [80] reviewed the effectiveness of fourteen different sequential sampling techniques for calibrating Kriging surrogate models. To do so, the authors not only assessed the different approaches in terms of accuracy, but also took into account criteria such as the robustness of the method in high-dimensional situations, the computational cost of implementation, or the coding complexity, among others. The findings of those authors reflected that the best sampling method in terms of overall performance was the *Maximising Expected Prediction Error* (MEPE) proposed by Liu et al. [81], an approach combining both exploration and exploitation aspects. However, its capabilities were closely followed by the purely geometric exploration technique *Monte Carlo-Intersite-proj-th* (MIPT) presented in [73]. In addition, when other factors are taken into account, MEPE requires fitting a Kriging model for each new sample point, with the associated computational cost. In fact, MEPE is a sampling method only applicable within the Kriging framework and challenging to implement. Conversely, MIPT approach does not suffer from the aforementioned problems. MIPT method is based in sampling a new point by maximising the distance to the sample points already included in the ED. For that purpose, in each iteration, a large Monte Carlo sample $\mathcal{C}$ of candidates over the parameter space is collected and the point $\mathbf{x} \in \mathcal{C}$ that maximises the distance to the current training set is selected. Formally, given an ED $\{\mathbf{x}^1, ..., \mathbf{x}^n\}$ a new sample $\mathbf{x}^{n+1}_{\text{MIPT}}$ is chosen by solving the optimisation problem:

$$\mathbf{x}^{n+1}_{\text{MIPT}} = \arg\max_{\mathbf{x}^* \in \mathcal{C}} \left(\min_{\mathbf{x}^i \in \text{ED}} \left\| \mathbf{x}^* - \mathbf{x}^i \right\|_2 \right).$$

A 10-point ED generated by HS on the unit square is depicted in Fig. 2.3 (a). For the purpose of comparison, a 40-point ED generated from the ground up by LHS on the unit square is depicted in Fig. 2.3 (b). In contrast, Fig. 2.3 (c) illustrates an extension of the 10-point ED shown in Fig. 2.3 (a) generated according to the MIPT method comprising also 40 points. It is clear that the parameter space is more evenly filled after the application of MIPT. Consequently, for challenging metamodelling problems, MIPT is an easy-to-implement strategy capable of saving a lot of resources in sampling and posterior fitting

processes.

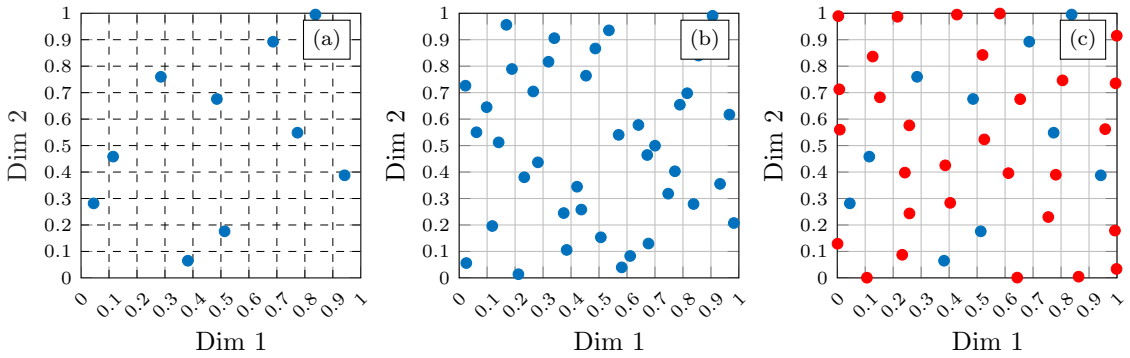

FIGURE 2.3: (a,b) LHS 10 points samples on the 2D unit hypercube with 10 and 40 points, respectively. (c) A 40 points sample generated from (a) according to MIPT method.

Another key aspect in the development of surrogate models is the assessment of model performance. There exist two main approaches to address this task: *cross-validation* (CV) methods and independent *validation set testing* [65]. The suitability of either procedure depends on factors inherent to the problem at hand, such as the computational cost of evaluating a single point in the parameter space or the time availability of the researcher. CV methods consist of dividing the ED into a collection of disjoint subsets, each of which is used to assess the predictive capability of the surrogate model trained by the remaining subsets [82]. Formally, given a mathematical model $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^{\mathbb{M}} \rightarrow \mathbb{R}$ and an ED $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\} \subset \mathcal{D}$, it is straightforward to define the set of pairs $\mathcal{T} = \{(\mathbf{x}^i, y^i),\ i = 1, ..., n\}$ where $y^i = \mathcal{M}(\mathbf{x}^i)$ for all $i \in \{1, ..., n\}$. Then, the set $\mathcal{T}$ is split into L exclusive and exhaustive subsets with approximately equal size such that [8]:

$$\bigcup_{\ell=1}^{L} \mathcal{T}_\ell = \mathcal{T}, \text{ and } \mathcal{T}_i \cap \mathcal{T}_j = \varnothing \quad \forall (i,j) \in \{1, ...L\}^2.$$

Afterwards, L surrogate models denoted by $\widehat{\mathcal{M}}^{-\ell}, k = 1, ..., L$ are fitted taking the union of these subsets as the training set letting the left-out subset for testing. Thus, L error measures are obtained by:

$$\varepsilon_\ell = \sum_{(\mathbf{x}^i, y^i) \in \mathcal{T}_\ell} \left(y^i - \widehat{\mathcal{M}}^{-\ell}(\mathbf{x}^i) \right)^2,\ \ell = 1, ..., L,$$

and the CV error is then straightforwardly defined as:

$$\varepsilon_{\text{CV}} = \frac{1}{L} \sum_{\ell=1}^{L} \varepsilon_\ell$$

When $L = n$, i.e. when each subset consists of only one sample, the CV error is referred to as *leave-one-out* (LOO) error. LOO validation has become a popular assessment method in the field of surrogate modelling due to its ability to provide approximately unbiased estimates of error without the need for obtaining new samples [82]. As a major

drawback, this method multiplies the computational time required for testing, as each of the metamodels is entirely independent of the others when different points are left out. Therefore, LOO is especially suitable in situations where running the forward model is very time-consuming, and, consequently, small training sets are utilised [83].

On the other hand, error assessment using a validation set consists of extracting a collection of points $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^K\}$ from the parameter space, independent of the ED and called *validation set*, and running both the model and the surrogate on it. Then, the difference between the values returned by the forward model and the metamodel is measured. Different error metrics can be employed to quantify this difference, distinguishing between global error metrics and others that are more focused on the maximum error committed locally. Among the former, *Normalised Root Mean-Square Error*, (NRMSE), the *Normalized Average Absolute Error* (NAAE) and the *Coefficient of Determination* (R^2) are particularly noteworthy whilst in the latter the most popular option is the *Normalised Maximum Absolute Error* (NMAE) [80, 84]. A detailed overview of these metrics is provided in Table 2.1.

TABLE 2.1: Error metrics for the accuracy assessment of surrogate models over a validation set of size K.

Normalised Root Mean-Square Error (NRMSE)	Coefficient of Determination (R^2)
$\mathrm{NRMSE} = \dfrac{\sqrt{\sum_{i=1}^{K}\left(\widehat{\mathcal{M}}(\mathbf{x}^i) - \mathcal{M}(\mathbf{x}^i)\right)^2}}{\max_i\left(\mathcal{M}(\mathbf{x}^i)\right) - \min_i\left(\mathcal{M}(\mathbf{x}^i)\right)}$	$\mathrm{R}^2 = 1 - \dfrac{\sum_{i=1}^{K}\left(\widehat{\mathcal{M}}(\mathbf{x}^i) - \mathcal{M}(\mathbf{x}^i)\right)^2}{\sum_{i=1}^{K}\left(\overline{\mathcal{M}} - \mathcal{M}(\mathbf{x}^i)\right)^2}$
Normalized Average Absolute Error (NAAE)	**Normalized Maximum Absolute Error (NMAE)**
$\mathrm{NAAE} = \dfrac{\sum_{i=1}^{K}\left\lvert\widehat{\mathcal{M}}(\mathbf{x}^i) - \mathcal{M}(\mathbf{x}^i)\right\rvert}{K\sigma_{\mathcal{M}}}$	$\mathrm{NMAE} = \dfrac{\max_i\left\lvert\widehat{\mathcal{M}}(\mathbf{x}^i) - \mathcal{M}(\mathbf{x}^i)\right\rvert}{K\sigma_{\mathcal{M}}}$

where $\sigma_{\mathcal{M}} = \sqrt{\dfrac{1}{K-1}\sum_{i=1}^{K}\left(\overline{\mathcal{M}} - \mathcal{M}(\mathbf{x}^i)\right)^2}$ and $\overline{\mathcal{M}}$ denotes the arithmetic mean of the model evaluations on the validation set.

2.2 Polynomial Chaos Expansions

Let $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^M \to \mathbb{R}$ be a computational model. In PCE framework, input parameters are represented by a M-dimensional random vector $\mathbf{X} = (X_1, ..., X_M)$ with known PDF $f_{\mathbf{X}}$. Consequently, the model output or QoI becomes also a random variable $Y = \mathcal{M}(\mathbf{X})$ whose properties are determined by the propagation of the uncertainty of the design variables through the computational model $\mathcal{M}$. The intuitive idea of a PCE metamodel is to represent the random variable Y as a series of random variables, that is, $Y = \sum_{j=0}^{\infty} y_j Z_j$. When Z_j are orthonormal polynomials on a random vector $\mathbf{X}$, this series is referred to as a *Polynomial Chaos Expansion* [85], naming the method. In order to rigorously define a surrogate PCE, it will be necessary to introduce a probabilistic setting.

Let $(\Omega, \mathcal{F}, P)$ a probability space, where Ω is the event space equipped with the σ-algebra $\mathcal{F}$ and P a probability measure defined on $\mathcal{F}$, and denote by $\mathcal{L}^2(\Omega, \mathcal{F}, P; \mathbb{R})$ the space of real random variables X with finite variance defined on Ω. Given $X_1, X_2 \in \mathcal{L}^2(\Omega, \mathcal{F}, P; \mathbb{R})$, an inner product can be defined as [85]:

$$< X_1, X_2 >_{\mathcal{L}_2} = \mathrm{E}[X_1 X_2] = \int_\Omega X_1(\omega) X_2(\omega) dP(\omega).$$

This inner product induces naturally the norm: $||X||_{\mathcal{L}^2} = \sqrt{\mathrm{E}[X^2]}$. It can be proved that, equipped with this norm, $\mathcal{L}^2(\Omega, \mathcal{F}, P; \mathbb{R})$ is a Hilbert space which admits a numerable basis $\left\{Z_j\right\}_{j=0}^{\infty}$ [85]. The first step to set up a PCE metamodel will be to construct a basis consisting of multivariate orthonormal polynomials in the input vector $\mathbf{X}$. For the sake of simplicity, it will be assumed from now on that the design variables are independent. Consequently, the joint PDF of input parameters $f_{\mathbf{X}}$ may be cast as $\prod_{i=1}^{M} f_{X_i}$, where f_{X_i} refers to the marginal PDF of X_i, $i = 1, ..., M$.

Consider a family $\left\{\psi_k^{(i)},\ k \in \mathbb{N}\right\}$ of orthonormal polynomials with respect to X_i, i.e., sattisfying:

$$\mathrm{E}[\psi_j^{(i)}(X_i)\ \psi_k^{(i)}(X_i)] = \delta_{jk}, \tag{2.1}$$

where subscript k denotes the degree of the polynomial $\psi_k^{(i)}$ and δ_{jk} refers to the Dirac delta function [34]. Such a family can be obtained by applying the Gram-Schmidt orthogonalization procedure to the collection $\{1, x, x^2, ...\}$ and normalising the resulting orthogonal polynomials [86]. For standard distributions, the associated family of orthogonal polynomials is well known. For example, if X_i follows an uniform distribution $U(-1, 1)$, the (normalised) Legendre polynomials satisfy Eq. (2.1) [87]. Some common distributions ans their associated families of orthonormal polynomials are reported in Table 2.2. It should be noted that, when design variables follow different distributions, it is still possible to use the orthogonal polynomial families listed in Table 2.2 through an isoprobabilistic transformation. For example, a lognormal variable can be transformed into a function of a standard normal variable, which can then be used in conjunction with Hermite polynomials [85].

Finally, a multivariate polynomial basis can be constructed from the previously defined univariate polynomials by tensor product. Defining a multi-index $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M)$ as an ordered list of integers, it is possible to associate a multivariate polynomial $\boldsymbol{\Psi}$ to any multi-index $\boldsymbol{\alpha}$ by:

$$\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X}) = \prod_{i=1}^{M} \psi_{\alpha_i}^{(i)}(X_i). \tag{2.2}$$

TABLE 2.2: Classical families of univariate polynomials for PCE metamodelling.

Type of variable	Orthogonal Polynomials	Orthonormal basis
Uniform $U(-1,1)$	Legendre $P_k(x)$	$P_k(x)/\sqrt{\frac{1}{2k+1}}$
Gaussian $N(0,1)$	Hermite $He_k(x)$	$He_k(x)/\sqrt{k!}$
Gamma $\Gamma(a,1)$	Laguerre $L_k^a(x)$	$L_k^a(x)/\sqrt{\frac{\Gamma(k+a+1)}{k!}}$
Beta $B(a,b)$	Jacobi $J_k^{a,b}(x)$	$J_k^{a,b}(x)/\mathcal{J}_{a,b,k}$
	$\mathcal{J}_{a,b,k} = \sqrt{\frac{2^{a+b+1}}{2k+a+b+1}\frac{\Gamma(k+a+1)\Gamma(k+b+1)}{\Gamma(k+a+b+1)\Gamma(k+1)}}$	

It is obvious that, by construction, such polynomials satisfy $E[\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X})\ \boldsymbol{\Psi}_{\boldsymbol{\beta}}(\mathbf{X})] = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, i.e are also orthonormal. Moreover, it can be proved that the set of all multivariate polynomials defined in Eq. (2.2) forms a basis of the Hilbert space $\mathcal{L}^2(\Omega, \mathcal{F}, P; \mathbb{R})$ [86]. As a result, the model response Y can be represented as:

$$Y = \mathcal{M}(\mathbf{X}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} \mathbf{a}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X}). \tag{2.3}$$

Although Eq. (2.3) is exact for an infinite number of polynomials, it is obvious that in practice only a finite number of terms can be computed. Once a polynomial basis has been determined, the next steps to construct a PCE surrogate model are to select a finite set of multi-indices $\mathcal{A} \subset \mathbb{N}^M$ and compute the corresponding coefficients $\mathbf{a}_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \mathcal{A}$. Many different strategies can be taken into account to truncate the polynomial series. One of the simplest approaches consists in selecting the set of all polynomials whose total degree $|\boldsymbol{\alpha}| = \sum_{i=1}^M \alpha_i$ is up to p, i.e., the set of multi-indices $\mathcal{A}^{M,p} = \{\boldsymbol{\alpha} \in \mathbb{N}^M : 0 \leq |\boldsymbol{\alpha}| \leq p\}$ [88]. Nevertheless, it should be noted that the cardinality of $\mathcal{A}^{M,p}$ is $\binom{M+p}{p}$. Therefore, in large-dimensional problems or non-linear systems where high degree polynomials are required to represent the response surface, the number of coefficients to be computed drastically increases, and so the subsequent computational burden of the method. As a partial solution, a more restrictive *hyperbolic truncation* scheme was proposed by Blatman and Sudret [89].

In practice, it is often observed that PCE surrogate models are mostly determined by terms where only few variables are involved. This phenomenon is referred to as the *sparsity-of-effects principle* [85]. To take advantage of it, the hyperbolic truncation scheme selects all the multi-indices belonging to $\mathcal{A}^{M,p}$ with q-norm $\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^M |\alpha_i|^q\right)^{\frac{1}{q}}$ less than or equal to p, namely:

$$\mathcal{A}^{M,p,q} = \{\boldsymbol{\alpha} \in \mathcal{A}^{M,p} : 0 \leq \|\boldsymbol{\alpha}\|_q \leq p\}, \quad 0 < q \leq 1. \tag{2.4}$$

Note that, the smaller q, the fewer polynomials will be considered in the final expansion, although all the univariate high-degree terms remain included. On the other hand, when $q = 1$, the hyperbolic truncation coincides with the standard one, i.e.,

$\mathcal{A}^{M,p,q} = \mathcal{A}^{M,p}$. Nonetheless, although substantial cost reductions can be achieved using the hyperbolic truncation scheme, the number of coefficients in the expansion may still be considerable. An adaptive technique based on the *least-angle regression* (LAR) algorithm for selecting a minimal basis will be explored in Section 3.3.

The last step to build a PCE metamodel is the computation of the coefficients $\mathbf{a}_{\boldsymbol{\alpha}}$. The aim is to find the set $\mathbf{a} = \{\mathbf{a}_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathcal{A}\}$ which minimise the mean-squared error, that is:

$$\mathbf{a} = \underset{\mathbf{a}\in\mathbb{R}^{\text{card}\,\mathcal{A}}}{\arg\min}\ \mathrm{E}\left[\left(\mathcal{M}(\mathbf{X}) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \mathbf{a}_{\boldsymbol{\alpha}}\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X})\right)^2\right]. \tag{2.5}$$

Given an ED $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \subset \mathcal{D}$ and the vector $\mathbf{y} = \left(y^1 = \mathcal{M}(\mathbf{x}^1), \ldots, y^n = \mathcal{M}(\mathbf{x}^n)\right)^{\mathrm{T}}$ collecting the realisations of the forward model on the ED, an estimated solution to the problem 2.5 is given by:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in\mathbb{R}^{\text{card}\,\mathcal{A}}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{M}(\mathbf{x}^i) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \mathbf{a}_{\boldsymbol{\alpha}}\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{x}^i)\right)^2. \tag{2.6}$$

The solution of the optimisation problem defined by Eq. (2.6) reads:

$$\hat{\mathbf{a}} = \left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{y} \tag{2.7}$$

where $\boldsymbol{\Psi}_{ij} = \left(\psi_{\boldsymbol{\alpha}_j}(\mathbf{x}^i)\right)_{i=1,\ldots,n}^{j=1,\ldots,\text{card}\,\mathcal{A}}$ is computed from the evaluation of the basis polynomials onto each point in the ED and referred to as the *information matrix* [86]. It should be noted that, in order to be well-posed, the number of unknown coefficients (i.e., card $\mathcal{A}$) in Eq. (2.7) must be smaller than the size of the ED. An empirical rue of thumb is to take an ED size from two to three times the number of unknowns [86]. In light of the above, the prediction of a PCE metamodel at an arbitrary point $\mathbf{x}$ of the domain $\mathcal{D}$ is given by:

$$\widehat{\mathcal{M}}^{PCE}(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{x}). \tag{2.8}$$

This formulation of a PCE metamodel presents some additional benefits. For example, the calculation of the LOO error presented in Section 2.1 is immediate, making it a good alternative when obtaining a validation set is not available. Through algebraic derivations, it can be proved that it is possible to compute the LOO error from a single expansion trained with the full original ED. Denoting by h_i the i-th component of the vector $\mathbf{h} = \text{diag}\left\{\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\right\}$, the metamodel LOO error estimate reads [85]:

$$\varepsilon_{LOO} = \sum_{i=1}^{k}\left(\frac{\mathcal{M}\left(\mathbf{x}^i\right) - \widehat{\mathcal{M}}^{PCE}(\mathbf{x}^i)}{1 - h_i}\right)^2. \tag{2.9}$$

As a consequence, since it is often difficult to know a priori the optimal degree of expansion, a suitable strategy is to construct different surrogates and to compare their LOO errors. A concurrent increase in the expansion degree and the LOO error is an indicator that the model is suffering from overfitting. Note that, for metamodel comparison,

a normalised version of the LOO error estimation is available by dividing the expression in Eq. (2.9) by the sample variance $\mathrm{Var}(\mathbf{y})$.

The ability to perform direct sensitivity analysis constitutes another strength of PCE models. The purpose of sensitivity analysis is to quantify which input parameters X_i, $i = 1, ...M$, or combinations thereof, best explain the variability of the QoI [40]. In this light, Sobol's indices are considered one of the most effective sensitivity measures for general computational models [90]. The idea behind the construction of Sobol's indices involves decomposing the response of a computational model $\mathcal{M}$ as the sum of a constant function $\mathcal{M}_0$, M univariate functions denoted by $\mathcal{M}_i$, $\binom{n}{2}$ bivariate functions $\mathcal{M}_{ij}$, and so forth, as shown in Eq. (2.10):

$$\mathcal{M}(\mathbf{x}) = \mathcal{M}_0 + \sum_{i=1}^{M} \mathcal{M}_i(x_i) + \sum_{1 \le i < j \le M} \mathcal{M}_{ij}(x_i, x_j) + \ldots + \mathcal{M}_{1,2,\ldots,M}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{D}. \tag{2.10}$$

Once the response of the model has been described following Eq. (2.10), the process continues decomposing the variance of the model output into fractions which can be attributed to inputs or sets of inputs. For example, given a model depending on two parameters, one might find that 70% of the output variance is attributable to changes in the first input, 20% by the variance in the second one, and 10% due to interactions between both. In this context, the first-order Sobol's indices are defined as [40]:

$$S_i = \frac{\mathrm{Var}[\mathcal{M}_i(X_i)]}{\mathrm{Var}[\mathcal{M}(X)]}. \tag{2.11}$$

Similarly, the second-order Sobol indices are defined by:

$$S_{ij} = \frac{\mathrm{Var}[\mathcal{M}_{ij}(X_i, X_j)]}{\mathrm{Var}[\mathcal{M}(X)]}. \tag{2.12}$$

Following the same logic, it is possible to define Sobol's indices for each set of input parameters $\mathbf{u} = \{i_1, ..., i_s\} \subseteq \{1, ..., M\}$. Consequently, each index $S_\mathbf{u}$ will be a sensitivity measure describing which amount of the total variance is due to the uncertainties in the set of input parameters $\mathbf{u}$. Finally, the total Sobol' index S_i^T, $i = 1, \ldots, M$ is defined as the sum of all the Sobol' indices involving the parameter i [86]:

$$S_i^T = \sum_{i \in \mathbf{u}} S_\mathbf{u}. \tag{2.13}$$

In this light, the index S_i^T quantifies the total impact of a given parameter X_i including all its interactions [40].

The Sobol's indices can be straightly approximated from a PCE surrogate. Indeed, defining for each set of variables $\mathbf{u} \subseteq \{1, \ldots, M\}$ the set of indices $\mathcal{A}_\mathbf{u}$ depending only on $\mathbf{u}$ as:

$$\mathcal{A}_\mathbf{u} = \{\alpha \in \mathcal{A} : \alpha_k \neq 0 \text{ if and only if } k \in \mathbf{u}\}, \tag{2.14}$$

Eq. (2.8) can be re-written as:

$$\widehat{\mathcal{M}}^{PCE}(\mathbf{x}) = \sum_{\mathbf{u}\subseteq\{1,\dots,M\}} \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\mathbf{u}}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{x}), \tag{2.15}$$

since the set of all $\mathcal{A}_{\mathbf{u}}$ defines a partition of $\mathcal{A}$. Therefore, due to the orthogonality of the PCE basis, it is verified that:

$$\mathrm{Var}\left[\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\mathbf{u}}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{x})\right] = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\mathbf{u}}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}^2. \tag{2.16}$$

Thus, the estimated value of the index S_u obtained via the PCE surrogate is given by:

$$\hat{S}_{\mathbf{u}} = \frac{\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\mathbf{u}}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}^2}{\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}^2}. \tag{2.17}$$

Analogously, the total indices S_i^T, $i = 1, \dots, M$, can be approximated by:

$$\hat{S}_i^T = \frac{\sum_{\boldsymbol{\alpha}\in\mathcal{A}_i^T} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}^2}{\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\mathbf{a}}_{\boldsymbol{\alpha}}^2}, \quad \text{with } \mathcal{A}_i^T = \{\boldsymbol{\alpha} \in \mathcal{A} : \alpha_i > 0\}. \tag{2.18}$$

This last property of PCE metamodels will play a crucial role to define the Multi-element Polynomial Chaos-based Kriging surrogate model, whose construction will be discussed in detail in Section 3.2 and paper B. For a more in-depth analysis of sensitivity analysis in the PCE framework, interested readers may refer to [40].

2.3 Kriging

Kriging (also known as *Gaussian process regression*) in one of the most popular methods for surrogate modelling in engineering practice [17]. First proposed by the South African mining engineer Danie G. Krige in the 1950s and further developed by the French mathematician Georges Matheron in the 1960s, Kriging was originally conceived as a geostatistical method for estimating ore grades in mining [65]. It has since evolved into a widely employed statistical technique for interpolation and spatial prediction in various fields such as geology, material science, and machine learning. The key mathematical assumption defining the method is that the output $\mathcal{M}(\mathbf{x})$ of the model in an arbitrary point $\mathbf{x} \in \mathcal{D}$ is a realisation of an underlying Gaussian random field. Given a probability space $(\Omega, \mathcal{F}, P)$ a random field is defined as a collection $\{F_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ of random variables indexed by an arbitrary set of indices [91]. For the sake of simplicity, only real random fields verifying that $\mathcal{X} \subset \mathbb{R}^M$ will be considered hereafter. Thus, it is possible to understand a random field as a random multivariate function $F : (\mathcal{X}, \Omega) \to \mathbb{R}$ which assigns a probability distribution $F(\mathbf{x}, \omega)$ to each point $\mathbf{x} \in \mathcal{X}$. A random field is said to be Gaussian if, for all $\lambda_1, ..., \lambda_k \in \mathbb{R}$ and $\mathbf{x}_1, ..., \mathbf{x}_k \in \mathcal{X}$, $\sum_{i=1}^k \lambda_i F_{\mathbf{x}_i}$ follows a multivariate Gaussian distribution [92].

Taking into account the preceding considerations, a Kriging model $\mathcal{M}^K$ for approximating $\mathcal{M}$ is defined by a Gaussian random field:

$$\mathcal{M}^K(\mathbf{x}) = \mathcal{T}(\mathbf{x}) + \mathcal{Z}(\mathbf{x}, \omega), \quad \mathbf{x} \in \mathcal{D} \tag{2.19}$$

where the term $\mathcal{T}(\mathbf{x})$ represents the mean of $\mathcal{M}^K$ and is referred to as *trend* and $\mathcal{Z}(\mathbf{x}, \omega)$ is an zero-mean field. As a consequence of Gaussian assumption, it can be proved that $\mathcal{Z}$ is *second-order stationary* [92]. This means that it can be fully characterised by its variance σ^2 and an auto-correlation function $\mathbf{R}(\mathbf{x}, \mathbf{x}') = \mathbf{R}(|\mathbf{x} - \mathbf{x}'|; \boldsymbol{\theta})$, depending only on the difference $|\mathbf{x} - \mathbf{x}'|$ and some hyper-parameters $\boldsymbol{\theta}$ to be determined. On the other hand, the trend term $\mathcal{T}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\beta} = \sum_{r=1}^{s} \beta_r f_r(\mathbf{x})$ is a linear regression model comprising s user-selected regression functions $f_r(\mathbf{x})$ and its corresponding (unknown) regression coefficients β_r (e.g, $f_r(\mathbf{x}) = \mathbf{x}^r$). This formulation corresponds to the most general version of Kriging, referred to as *Universal Kriging* (UK). However, the prevailing implementation of Kriging in the literature [17], labelled as Ordinary Kriging (OK), assumes that the trend term consists of a constant value β_0 to be estimated.

The auto-correlation function $\mathbf{R}$ is selected by the user and must satisfy $\mathbf{R}(\mathbf{0}; \boldsymbol{\theta}) = 1$ and $\mathbf{R}(|\mathbf{x} - \mathbf{x}'|; \boldsymbol{\theta}) \to 0$ when $|\mathbf{x} - \mathbf{x}'| \to +\infty$ [65]. Moreover, when the input dimension M is greater than one, it is common to restrict the possibilities to functions of the form $\mathbf{R}(|\mathbf{x} - \mathbf{x}'|; \boldsymbol{\theta}) = \prod_{i=1}^{M} R_i(x_i, x_i'; \theta_i)$, where x_i, x_i' and θ_i denote the i-th coordinate of the vectors $\mathbf{x}$, $\mathbf{x}'$ and $\boldsymbol{\theta}$, respectively [93]. The most frequent choice is the Gaussian model, although other popular alternatives can be found in Table 2.3.

TABLE 2.3: Frequently used correlation models for Kriging [93].

Correlation family	$R_i(x_i, x_i'; \theta_i)$
Linear	$\max\left(0, 1 - \dfrac{\lvert x_i - x_i'\rvert}{\theta_i^{-1}}\right)$
Exponential	$\exp\left(\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)$
Gaussian	$\exp\left(-\dfrac{1}{2}\left(\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)^2\right)$
Matérn 3/2	$\left(1 + \sqrt{3}\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)\exp\left(-\sqrt{3}\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)$
Matérn 5/2	$\left(1 + \sqrt{5}\dfrac{\lvert x_i - x_i'\rvert}{\theta_i} + \dfrac{5}{3}\left(\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)^2\right)\exp\left(-\sqrt{5}\dfrac{\lvert x_i - x_i'\rvert}{\theta_i}\right)$

The prediction $\widehat{\mathcal{M}}^K(\mathbf{x})$ of a Kriging model at any point of the parameter space is a direct consequence of the Gaussian assumption. Given an ED $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \subset \mathcal{D}$ and an auto-correlation function $\mathbf{R}$ satisfying the aforementioned conditions, and assuming for simplicity that the values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are known, it is verified that for any $\mathbf{x} \in \mathcal{X}$ the joint probability distribution of $\left(\widehat{\mathcal{M}}^K(\mathbf{x}), \mathbf{y}\right)^{\mathrm{T}}$ satisfies [94]:

$$\begin{pmatrix} \widehat{\mathcal{M}}^K(\mathbf{x}) \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}_{n+1}\left(\begin{bmatrix} \mathbf{f}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} \\ \mathbf{F}\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \mathbf{r}(\mathbf{x})^{\mathrm{T}} \\ \mathbf{r}(\mathbf{x}) & \mathbf{R} \end{bmatrix} \right), \tag{2.20}$$

where $\mathbf{r}(\mathbf{x})$ denotes the vector of cross-correlations between the prediction point $\mathbf{x}$ and each one of the observations of the ED, that is, $(\mathbf{r}_i(\mathbf{x}))_{i=1,\dots,n} = \mathbf{R}(\mathbf{x}, \mathbf{x}^i)$ and $\mathbf{y}$ is the vector containing the realisations of the forward model on the ED. Similarly, $\mathbf{F} = (\mathbf{F}_{ij}) = (f_j(\mathbf{x}^i))$ is the information matrix and $\mathbf{R}$ refers to the correlation matrix given by $(\mathbf{R}_{ij}) = (\mathbf{R}(\mathbf{x}^i, \mathbf{x}^j))$, $i, j = 1, \dots, n$. Conditioning this joint distribution on the training sample, it is obtained a Gaussian random variable referred to as *Kriging predictor*, whose expected value will be the prediction of the metamodel in an arbitrary point [94]. In addition, the variance value provides a measure of the prediction's uncertainty without requiring additional calculations. An interesting property of Kriging predictor is its interpolating nature: it is easy to verify that the variance of the prediction on training samples collapses to zero. Consequently, it holds that $\mathcal{M}^K(\mathbf{x}) = \mathcal{M}(\mathbf{x})$ for all $\mathbf{x} \in$ ED. Analytical expressions of the mean and variance of Kriging predictor are given in Eqs. (2.21) and (2.22), respectively [93]:

$$\mathrm{E}[\widehat{\mathcal{M}}^K(\mathbf{x})] = \mathbf{f}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}), \tag{2.21}$$

$$\mathrm{Var}[\widehat{\mathcal{M}}^K(\mathbf{x})] = \sigma^2\left(1 - \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) + \mathbf{u}(\mathbf{x})^{\mathrm{T}}(\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{u}(\mathbf{x})\right), \tag{2.22}$$

where $\mathbf{u}(\mathbf{x}) = \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) - \mathbf{f}(\mathbf{x})$. The main practical challenge in the implementation of a Kriging model is the estimation of the parameter $\boldsymbol{\beta}$ of the trend and the parameters $\boldsymbol{\theta}$ and σ^2 of the stochastic term $\mathcal{Z}$. Given an estimation $\hat{\boldsymbol{\theta}}$, optimal values of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2(\hat{\boldsymbol{\theta}})$ can be computed through the so called *best linear unbiased estimator* (BLUE) using Eqs. (2.23) and (2.24), respectively [41]:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) = \left(\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}\mathbf{R}^{-1}\mathbf{y}, \tag{2.23}$$

$$\hat{\sigma}^2(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\left(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\right), \tag{2.24}$$

As a consequence of the above, fitting a Kriging metamodel is reduced to the identification of the proper hyperparameters $\boldsymbol{\theta}$. There exist two main procedures in Kriging literature for obtaining a suitable estimation, based on either maximum likelihood (ML) or leave-one-out cross-validation (CV). However, both approaches require solving challenging minimisation problems, given respectively in Eqs. (2.25) and (2.26) [41, 95]:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{\theta}}\left[\frac{1}{n}\left(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}\right)|\mathbf{R}|^{1/n}\right], \tag{2.25}$$

$$\hat{\boldsymbol{\theta}}_{\mathrm{CV}} = \arg\min_{\boldsymbol{\theta}}\left[\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathrm{diag}(\mathbf{R}^{-1})\right)^{-1}\mathbf{R}^{-1}\right]. \tag{2.26}$$

It is unclear whether any of the approaches has superior capabilities, and the literature addressing the subject is scarce. It has been suggested that when the correlation function is misspecified, CV performs better than ML, while ML is optimal when the model is well specified [95]. Note that solving any of the minimisation problems (2.25) or (2.26) is the most computationally challenging step when fitting a Kriging surrogate. Two algorithm families are primarily employed to accomplish this task: local optimisers, typically gradient-based, and global methods, such as evolutionary algorithms. The former tends to converge faster and require fewer objective function evaluations, but they may perform poorly due to the possibility of becoming trapped in flat regions or local minima, especially with increasing input dimensions [93]. It should be noted that, in any case, it is required to repeatedly invert the matrix $\mathbf{R}$ for each combination of hyperparameters $\boldsymbol{\theta}$ considered. This process has a computational complexity $\mathcal{O}(n^3)$ [68], which makes fitting a Kriging metamodel a computational challenge when the sample size is large. Furthermore, as obtaining the Kriging prediction at a point $\mathbf{x} \in \mathcal{X}$ requires calculating the cross-correlation vector $\mathbf{r}(\mathbf{x})$, the evaluation time of the metamodel will also be affected negatively by large EDs. These issues will be further explored in Section 3.2.

2.3.1 Polynomial-chaos based Kriging

The aforementioned PCE and Kriging surrogate modelling methods can be combined in a natural way to produce a completely new technique referred to as *Polynomial Chaos based Kriging* (PCK). The basic idea PCK consists of approximating the global behaviour of the computational model $\mathcal{M}$ employing a PCE surrogate model whereas Kriging manages the local variability of the model output. To this end, a polynomial expansion will be incorporated as the trend term into the Universal Kriging formulation presented in Section 2.3.

Let $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^M \to \mathbb{R}$ be the computational model to be approximated, and suppose that an ED $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \subset \mathcal{D}$ and the corresponding realisations $\mathbf{y}$ of the forward model on it are available. Additionally, knowledge of the joint PDF of the M-dimensional input of the model is assumed, as established in Section 2.2. The construction of a PCK metamodel involves two main steps: first, the determination of the set of polynomials contained in the regression part, and secondly, the calibration of the correlation hyperparameters $\boldsymbol{\theta}$ as well as the Kriging parameters σ^2, $\boldsymbol{\beta}$. Synthesising the above, and employing the notations of Sections 2.2 and 2.3, the PCK surrogate model is given by Eq. (2.27) [41]:

$$\mathcal{M}^{PCK}(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{x}) + \mathcal{Z}(\mathbf{x}, \omega). \tag{2.27}$$

Consequently, the PCK prediction in any given point $\mathbf{x} \in \mathcal{D}$) reads:

$$\mathrm{E}[\widehat{\mathcal{M}}^{PCK}(\mathbf{x})] = \boldsymbol{\Psi}(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\hat{\boldsymbol{\beta}}) \tag{2.28}$$

It should be noted that the selection of the polynomial basis $\mathcal{A}$ is particularly important in the context of PCK. An excessive number of polynomials can lead to overfitting problems. Therefore, taking into account adaptive algorithms such as LAR to create a sparse basis becomes essential when constructing a PCK metamodel [41]. This point will be further discussed in Section 3.3.

2.4 Bayesian Inference

Flexibility and generality characterise Bayesian inference. A key advantage of the Bayesian approach over frequentist inference is its ability to quantify uncertainty directly, making it possible to fit complex models with with numerous parameters in a natural way within a common framework [96]. This section provides an overview of Bayesian inference to set the results of Section 3.2 in context.

2.4.1 Basic concepts

Statistical inference can be defined as the process of drawing conclusions about unobserved quantities from observations. Any inference process begins by defining a model, denoted by f, from which to extract information to a certain data set. In general terms, we can write:

$$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}, \tag{2.29}$$

where $\mathbf{y} = \{y_1, ..., y_n\}$ represents the observations of the model, $\boldsymbol{\theta} = \{\theta_1, ..., \theta_m\}$ a set of unknown parameters and $\mathbf{x} = \{x_1, ..., x_n\}$ the conditions, set by the experimenter, under which the model is evaluated. Finally, $\boldsymbol{\varepsilon}$ describes the model error. Within the Bayesian framework, Eq. (2.29) can be interpreted to signify that each observation y_i is the realisation of a random variable whose probability distribution depends on x_i and the unknown parameters set $\boldsymbol{\theta}$. In this context, the goal of statistical inference is to obtain information about $\boldsymbol{\theta}$. Since from a Bayesian point of view $\{\theta_1, ..., \theta_m\}$ are random variables with its own PDFs, it is possible to to take advantage of Bayes' theorem by writing:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}, \quad p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{2.30}$$

Each term in Eq. (2.30) has its own interpretation. $p(\boldsymbol{\theta})$ is known as the *prior* distribution of parameters $\boldsymbol{\theta}$, representing what is known about the parameters of the model before seeing any data. $p(\mathbf{y}|\boldsymbol{\theta})$ is referred to as the *likelihood*, and represents what is known after learning from the data. The *posterior* distribution of parameters, $p(\boldsymbol{\theta}|\mathbf{y})$, denotes the conditional probability distribution of the unobserved parameters, given the observed data. Posterior distribution can be seen as a compromise between data and prior information. Finally, $p(\mathbf{y})$ is known as the *evidence* or *marginal likelihood*. Evidence is a normalising constant calculated by integrating the likelihood of the data over the

entire parameter space of the model, weighted by the prior distribution of the parameters. Although this quantity plays a central role in Bayesian model selection, computing such an integral is typically very challenging, especially in contexts where the parameter space is multi-dimensional. As a result, most inference algorithms seek to avoid an explicit computation [97].

Equation (2.30) provides a simple illustration about the differences between the frequentist and Bayesian viewpoints. The former paradigm relies solely on data or experience, assigning probabilities to events based on empirical evidence from repetitions of the experiment of interest. In contrast, the Bayesian process is based on subjective assessments of the relative likelihood of events [44]. Prior distributions represent the (subjective) beliefs of the experimenter about the probability distribution of the unknown parameters before observing any data. These assumptions are updated as more data is collected, resulting ultimately in a posterior distribution that reflects the beliefs of the practitioner at the end of the inference process. As a consequence, Bayesians and frequentists interpret their estimates of the parameters of interest differently. From the frequentist point of view, given a model f dependent on a set of parameters $\boldsymbol{\theta}$ and a data set $\mathbf{y}$, the outcome of the inference process will be a single interval for each parameter θ_i (referred to as *confidence interval*), containing its true value with a sufficiently high given probability. It is noteworthy that these intervals are random before the observation of the data, but fixed afterwards. Moreover, in the frequentist framework, practitioners consider the parameters θ_i as unknown but fixed values. As a result, the statement 'The interval $[a, b]$ contains the true value of θ_i with a 95% probability', should be understood in terms of the relative frequency. That is, if the process of obtaining the confidence interval is repeated a sufficiently large number of times, it is expected that approximately 95% of these replications will yield an interval containing the true value of the parameter [98]. On the other hand, from a Bayesian perspective, the result of an inference process is the full posterior distribution of parameters. In contrast to the frequentist standpoint, the statement 'The interval $[a, b]$ contains the true value of θ_i with a 95% probability', can be now understood much more literally, in the sense that 95% of the probability mass of the posterior distribution lies within the interval $[a, b]$ [98]. Within Bayesian framework a region $[a, b]$ fulfilling such property is referred to as a *credible interval* rather than a confidence interval. However, it is important to note some differences with the classical concept. In the first place, a credible interval does not necessarily have to be an interval in multimodal distributions (see Fig. 2.4). Thus, it is better to refer to them as *credible regions*. Furthermore, the Bayesian credible interval is not fully determined after observing the data. This is due to two reasons: firstly, the choice of prior distribution can impact on the resulting posterior; in addition, there exist an infinite number of intervals containing a given percentage of the posterior probability mass. Therefore, equal-tailed credible intervals and high-density regions (HDR) are employed the most in practice. An equal-tailed credible interval at level (1-α)% is a region that encompasses (α)% of the probability mass of the posterior distribution around the median. Equal-tailed credible intervals are easy to compute and are useful when the PDF of the parameters is

symmetric and unimodal. On the other hand, an HDR is a region C in the parameter space that contains $100(1-\alpha)\%$ of the posterior mass satisfying that for all $\theta_1 \in C$ and $\theta_2 \notin C$, $g(\theta_1|x) \geq g(\theta_2|x)$, being g is the posterior PDF of parameter θ [99]. It follows immediately from the definition that the HDR is the smallest of all regions of probability coverage (1-α). Fig. 2.4 depicts the equal-tailed credible interval and HDR at 95% for a given Gaussian mixture, illustrating that only the latter is able to capture the bimodal nature of the distribution.

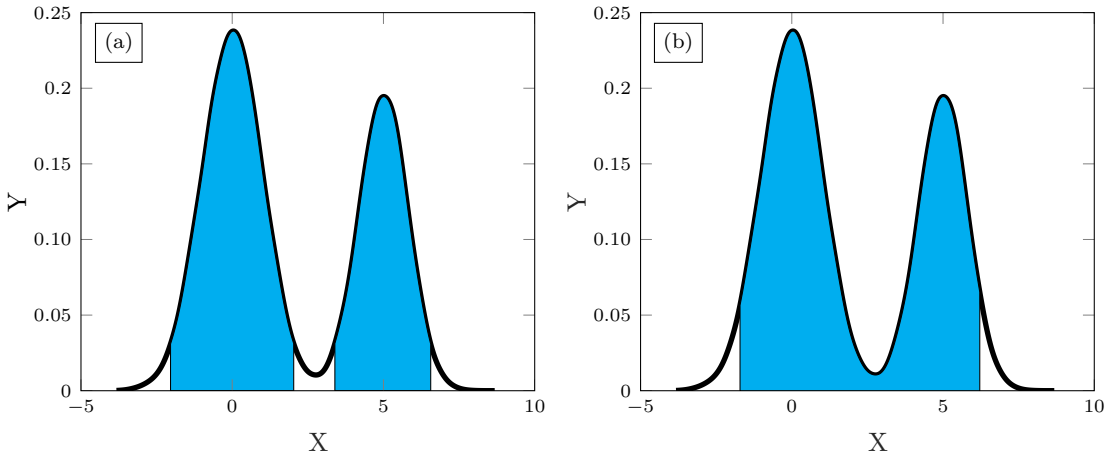

FIGURE 2.4: Bayesian estimates: 95% Highest density region (a) and 95% Equal-tailed credible interval (b)

2.4.2 Markov Chain Monte Carlo: the Metropolis-Hastings algorithm

Computing the posterior distribution analytically or numerically according to Eq. (2.30) is a very challenging task. For this reason, an alternative procedure should be sought. To circumvent the problem of calculating the model evidence, it is important to note that this factor is independent of the $\boldsymbol{\theta}$ parameters of interest. Therefore, an interesting strategy is to consider methods that allow drawing samples from the posterior directly, without explicit computation. This can be done through the MCMC family of algorithms. To gain an understanding of MCMC, let us begin with some fundamental concepts.

A *stochastic process* is a sequence of random variables $\{X_0, X_1, ...\}$, where the indices $0, 1, 2...$ represent successive time states. Given any time $t \geq 0$, it is possible to represent the conditional probability of a future event from all the events already observed. In this context, it is said that $\{X_0, X_1, ...\}$ is a *Markov process* (or *chain*) if the conditional probability of a future event depends only on the immediately preceding state. That is, if [100]:

$$P(X_{t+1}|X_t, X_{t-1}, ..., X_1, X_0) = P(X_{t+1}|X_t), \quad \forall t \geq 0. \tag{2.31}$$

A Markov process is fully determined by its initial distribution and its transition kernel [101]. The initial distribution is simply the marginal distribution of X_0. Denoting by $\mathcal{X}$ the *state space* of the Markov chain, i.e., the set of all possible realisations of the terms of the chain, a transition Kernel is the conditional distribution of X_{t+1} given X_t. That is, a function $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ verifying:

1. $\forall x \in \mathcal{X}, \quad K(x, \cdot)$ is a probability measure,

2. $\forall A \in \mathcal{B}(\mathcal{X}), \quad K(\cdot, A)$ is measurable,

where $\mathcal{B}(\mathcal{X})$ is the Borel set on $\mathcal{X}$. Note that it is assumed that the transition kernel does not depend on t. This property, known as *time-homogeneity*, is common in MCMC literature.

The intuitive idea of MCMC methods is to construct a Markov chain such that, after a certain time, it stabilises on a distribution that does not depend on time or the initial conditions. This distribution is referred to as the *stationary* or *invariant* distribution of the process. By selecting a suitable transition kernel, it is possible to ensure that the stationary distribution matches a given distribution of interest (e.g. the posterior $p(\boldsymbol{\theta}|\mathbf{y})$), even if a expression of the latter is unknown.

There are two ways to understand MCMC: from the perspective of a practitioner or from a mathematical standpoint. From a purely practical point of view, MCMC is usually seen as a 'black box': the Markov chain is a tool that produces some output when executed, which is the required invariant distribution, and whose inside cannot be seen [102]. Beyond the obvious, since from the black box point of view nothing is considered apart from the above, some problems such as pseudo-convergence can arise: a Markov chain may appear to have converged to its equilibrium distribution even though it has not. This can be the case if the distribution of interest is strongly multimodal (see for example, the one depicted in the Fig. 2.4), requiring many iterations to move from one mode to another. When the time required to switch between these parts is much longer than the length of the simulated Markov chain, it may appear to have converged, but this 'equilibrium' is highly dependent on the initial point of the chain [102]. From a theoretical perspective, one can consider questions such as the existence and uniqueness of the stationary distribution of a given Markov process or whether it is possible to ensure that this distribution coincides with the distribution of interest, provided it exists. Answering these questions is generally difficult, and related results are scattered throughout the mathematics, statistics, and engineering literature, making it challenging to get a full overview of MCMC foundations [103]. A synthesis between the two approaches is presented below: firstly, some key results from the theory of Markov processes that allow to ensure convergence to a stationary distribution of a particular chain are presented. Then, the Metropolis-Hastings (MH) algorithm, one of the most elementary MCMC methods, is described and proved that fulfils the convergence criteria. Finally, some guidelines for the practical implementation of MCMC algorithms are provided.

Given a Markov process $\{X_0, X_1, ...\}$ characterised by the transition kernel $K(\cdot, \cdot)$, $\pi : \mathcal{B}(\mathcal{X}) \to [0, 1]$ is said to be a stationary distribution of the chain when it is verified that:

$$\pi(B) = \int_{\mathcal{X}} K(x, B)\pi(x)\, \mathrm{d}x, \quad \forall B \in \mathcal{B}(\mathcal{X}). \tag{2.32}$$

A Markov process does not necessarily need to have a stationary distribution. However, there are simple sufficient conditions that can be checked in most MCMC algorithms

to ensure the existence of at least one stationary distribution. One of these conditions is referred to as *detailed balance*. A Markov chain with transition kernel K satisfies the detailed balance condition if there exists a probability measure π satisfying [101]:

$$K(y,x)\pi(y) = K(x,y)\pi(x), \quad \forall x, y \in \mathcal{X}. \tag{2.33}$$

Therefore, it can be easily verified that π is a stationary distribution of the chain. In fact, for each measurable set $B \subseteq \mathcal{X}$, it holds that:

$$\begin{aligned}\int_{\mathcal{X}} K(y,B)\pi(y)\mathrm{d}y &= \int_{\mathcal{X}}\int_{B} K(y,x)\pi(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathcal{X}}\int_{B} K(x,y)\pi(x)\mathrm{d}x\mathrm{d}y = \int_{B}\pi(x)\mathrm{d}x = \pi(B)\end{aligned}$$

since $\int_{\mathcal{X}} K(x,y)\ \mathrm{d}y = 1$.
One question that arises from this result is whether the measure π is unique. The answer lies in the concept of *irreducibility*. A Markov chain is said to be irreducible if, for every set A with positive probability (respect to π), the chain has some chance of visiting A in finite time, independent from the initial state [103]. That is, a Markov chain is irreducible if there exists an $n \in \mathbb{N}$ such that $K^n(x,A) > 0$ for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ such that $\pi(A) > 0$, where K^n is defined recursively by:

$$K^n(x,A) = \int_{\mathcal{X}} K^{n-1}(y,A)K(x,\mathrm{d}y), \quad n > 1, \quad K^1(x,A) = K(x,A). \tag{2.34}$$

Note that the notation $K(x,\mathrm{d}y)$ in Eq. (2.34) indicates the transition of the chain from a state x to some arbitrary state $y \in \mathcal{X}$. Irreducibility is a sufficient condition to ensure the uniqueness of the stationary distribution, provided it exists [103]. Moreover, it can be proved [104] that an irreducible chain that admits a stationary distribution has some chance of visiting any set A with positive probability an infinite number of times and, furthermore, that the expected total number of visits to any such set A is infinite. When the latter condition is fulfilled, the Markov process is said to be *recurrent*, and convergence is assured regardless of the initial value of the chain [100].

Considering the foregoing, the MH algorithm is presented below. MH relies on two essential components: 1) a *target* distribution $\pi(x)$ known except for a constant of proportionality (i.e. $\pi(x) = \frac{1}{Z}p(x)$ for some given function $p(x)$ and some unknown constant Z which normalises p to make it a probability distribution) and 2) a conditional density $q(x,y)$ easy to simulate and referred to as the *proposal* distribution. In order to construct the corresponding Markov chain, at each iteration of the algorithm, a candidate y from the state space $\mathcal{X}$ is proposed and an α acceptance probability of that state calculated. If the candidate is accepted, the chain moves to that state; otherwise, it remains in the current state and a new candidate is proposed in the next iteration. The algorithm will ultimately deliver samples from the target distribution [100]. The whole process is detailed in Algorithm 1.

Algorithm 1 Metropolis-Hastings Algorithm

Initialize x_0 ▷ Initial state
for $t = 1$ to T **do** ▷ Number of iterations
  Propose a candidate y from proposal distribution $q(x_{t-1}, \cdot)$
  Calculate acceptance probability:

$$\alpha(x_{t-1}, y) = \min\left\{1, \frac{p(y)q(y, x_{t-1})}{p(x_{t-1})q(x_{t-1}, y)}\right\}$$

  Generate a uniform random number $u \sim \text{Uniform}(0, 1)$
  if $u \leq \alpha(x_{t-1}, y)$ **then**
    Accept the candidate: $x_t = y$
  else
    Reject the candidate: $x_t = x_{t-1}$
  end if
end for

Note that the ratio $p(y)/p(x)$ is the same as $\pi(y)/\pi(x)$. The requirement that $p(x)$ must only be proportional to the density, rather than exactly equal to it, makes the MH algorithm particularly useful for Bayesian inference. Recalling Eq. (2.30), since the posterior distribution satisfies $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$, that is, it is known except for a proportionality constant $p(\mathbf{y})$ hard to compute, it follows immediately that MH algorithm allows to bypass the difficulties of evidence calculation.

The proposal distribution q has to be specified by the researcher. A common choice is a Gaussian distribution centred at x_{t-1} in each step t. An interesting property of this pick is that $q(x, y) = q(y, x)$, which simplifies the calculation of the acceptance probability α. Under these conditions, it will now be verified that the MH algorithm converges to the desired distribution. It is easy to check that that the transition kernel of the process defined in Algorithm 1 is given by:

$$K(x_t, y) = q(x_t, y)\alpha(x_t, y) + \delta_{x_t}(y)r(x_t),$$

where δ_{x_i} denotes the delta-Dirac mass located at x_i and $r(x_t) = 1 - \int_{\mathcal{X}} \alpha(x_t, s)q(x_t, s)\mathrm{d}s$ stands for the probability of rejecting the sampled candidate from the proposal distribution. Now, it is straightforward to verify the detailed balance condition given by Eq. (2.33):

The case where $x_{i+1} = x_i$ is trivial. On the other hand, if $x_{i+1} \neq x_i$ we have:

$$\begin{aligned} &K(x, y)p(x) = \alpha(x, y)q(x, y)p(x) = \\ &\min\left\{1, \frac{p(y)q(y, x_{t-1})}{p(x_{t-1})q(x_{t-1}, y)}\right\} p(x, y)p(x) = \\ &\min\left\{1, q(x, y)p(x), p(y)q(y, x)\right\} = K(y, x)p(y). \end{aligned}$$

Note that, if a value y belongs to the support of the target distribution but is never extracted from the proposal distribution, then $q(x, y) = 0$. Denoting by $\mathcal{E}$ the support of the target density, a simple condition that guarantees that all the values belonging to $\mathcal{E}$ can be generated as proposals is that $q(x, y) > 0$ for all $x, y \in \mathcal{E}$, which is satisfied with Gaussian proposals. Furthermore, this property is also a sufficient condition for establishing the irreducibility of the chain [101]. Summing all of the above, it is concluded that the MH algorithm converges to the target distribution $\pi(x)$.

2.4.3 MCMC in practice

Focusing on the paradigmatic case of Gaussian proposals, despite having established theoretical convergence to a stationary distribution, not all possible proposals will explore it with equal efficiency [54]. If the variance of the Gaussian proposal is too small, then the acceptance probability α is close to 1 because points that are close to each other have similar densities. In this case, the distribution target is explored at a very slow pace. On the other hand, when the proposal variance is too large, once the chain reaches a point in a high probability region the new candidates tend to be far from this region towards areas of low density. Therefore, the proposals are often rejected and the chain remains stuck at high-density points for long periods of time. As a result, an inadequate proposal can incur in higher computational costs and yield poorer estimates. Fig. 2.5 (a,b) and (c,d) demonstrate an example of chains and their estimation of the target distribution for a too small and too large proposal, respectively. An optimal acceptance rate is about 44% in one-dimensional problems and 23% in the general case [105, 106]. Fig. 2.5 (e,f) depicts the same example but selecting an appropriate proposal. However, it is worth noting that as long as the acceptance rate is not too low or too high, the algorithm will perform well overall, so these values should be understood as general guidelines.

Another relevant consideration in practice is the question of burn-in. Under irreducibility conditions, the Markov chain defined by the MH algorithm converges to the stationary distribution regardless of the initial state. However, if this point is in an area of very low probability density, the chain may need a period of iterations to reach the invariant distribution. Burn-in period is defined as a set of early samples that are discarded because the chain has not yet converged [107]. An important consideration regarding burn-in is its post-hoc nature. Decisions about burn-in must be made after sampling and observing the chains. It is often a good idea to be conservative: if the chain is long enough, discarding additional samples is safe, as the remaining ones are drawn from the stationary distribution [107]. Fig. 2.5 (a) shows that, since the initial state of the chain ($X_0 = 0$) is outside the high density area of the target distribution and the proposal is very small, the chain takes a few hundred iterations to converge. Therefore a conservative burn-in period of 500 samples has been discarded.

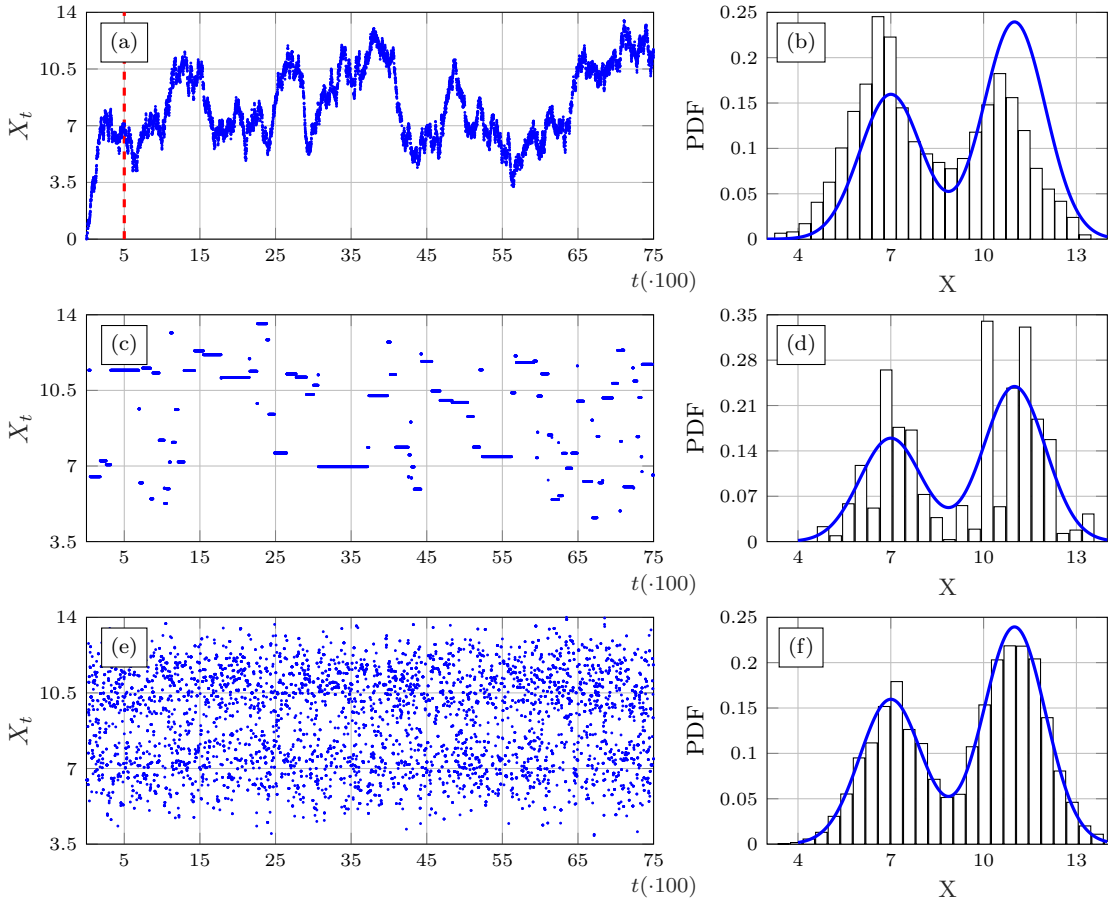

FIGURE 2.5: MH sampling from a Gaussian mixture with: a too small proposal (a,b), a too large proposal (c,d) and an appropriate proposal (e,f). A burn-in period of 500 samples has been considered in the first scenario.

In addition to the aforementioned difficulties, a bad proposal can lead to the pseudo-convergence problem described in Section 2.4.2, especially in distributions whose modes are separated by areas of low probability density. A simple and commonly employed trial-and-error diagnostic for this behaviour consists in running multiple chains with different initial states, and observe whether they converge to different modes. Another possibility is to employ adaptive algorithms such as Adaptive Metropolis (AM) or Delay Rejection Adaptive Metropolis (DRAM) proposed by Haario et al. (see ref. [54, 64]). These alternative algorithms are based on similar theoretical principles but allow for adjusting the proposal as the chain progresses. Fig. 2.6 (a,b) illustrates the pseudo-convergence phenomenon through the exploration of a bimodal Gaussian mixture with a poorly-tuned MH. Fig. 2.6 (c,d) demonstrates the performance of the AM algorithm using the same proposal and initial state of the chain. Refer to [108] for a general discussion about adaptive MCMC algorithms.

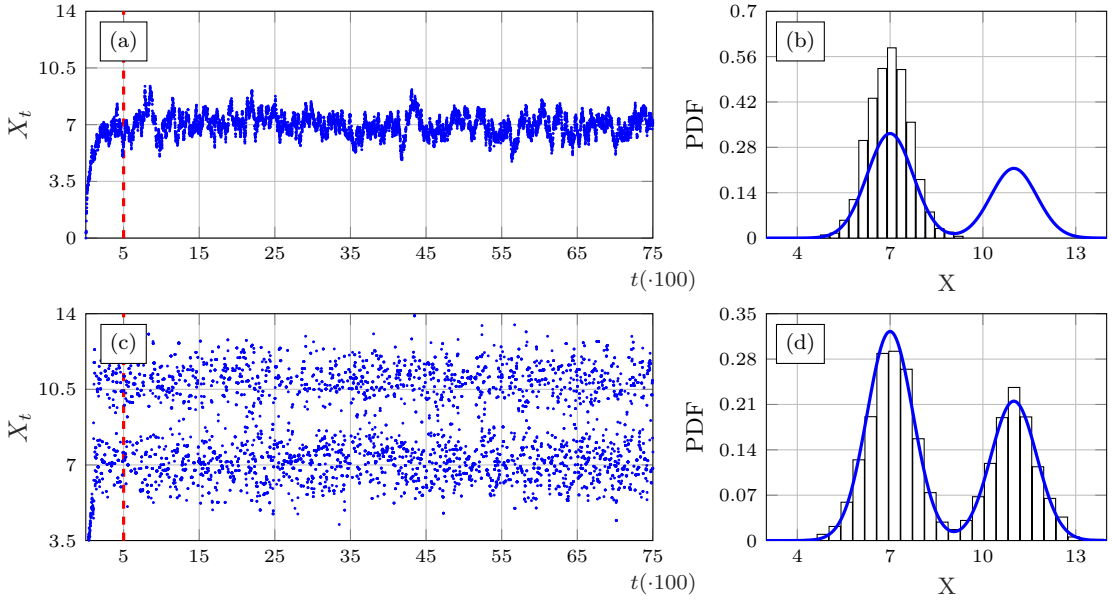

FIGURE 2.6: Comparison of MCMC sampling from a Gaussian mixture with modes separated by a region of low probability density, using the MH algorithm (a,b) and the AM algorithm (c,d). A burn-in period of 500 samples has been considered in both scenarios.

Chapter 3

Results and Discussion

This chapter presents the main findings obtained during the development of the thesis, which culminated in the three publications included as appendices.

3.1 PCE for estimation of the effective elastic properties of composite materials

The first case study involves the development and subsequent UP analysis of a polymeric material through computational models via PCE. Two different 2D and 3D models of a composite material consisting of an epoxy matrix doped with glass particles are investigated, and the results are compared with experimental data extracted from the literature. The presented results demonstrate the ability of the proposed methodology to perform UP analyses with minimum computational effort.

3.1.1 Composite materials

Composite materials, often simply referred to as *composites*, are defined as the artificial combination of two or more materials, known as *phases* [109]. Composites encompass a wide range of materials, from classical reinforced concrete or masonry to the latest nano-modified materials. There are several reasons why a new material resulting from the combination of various components can outperform conventional ones. Typical examples include materials that are cheaper, lighter, stronger or more durable than classical alternatives [110]. Consequently, there exist a myriad of composites. Common examples are particle-reinforced [111], fibre-reinforced [112] or laminated materials [113], just to mention a few. SEM micrographs of some particle-reinforced and fibre-reinforced microstructures observed are depicted in Fig. 3.1. The present case study focuses on predicting the effective mechanical properties of a composite consisting of glass particle-reinforced epoxy matrix. Epoxy polymers exhibit several desirable mechanical properties, including high bending/compression strength, wear resistance, and coefficient of friction, making them suitable for numerous engineering applications, particularly as adhesives [114]. However, their inherent brittleness and susceptibility to crack propagation constitute significant limitations. To address this limitation, rigid glass and silica particles are commonly employed to enhance the toughness of epoxy polymers [114].

The resulting composite has found widespread applications across various fields, such as automotive and aeronautics [115].

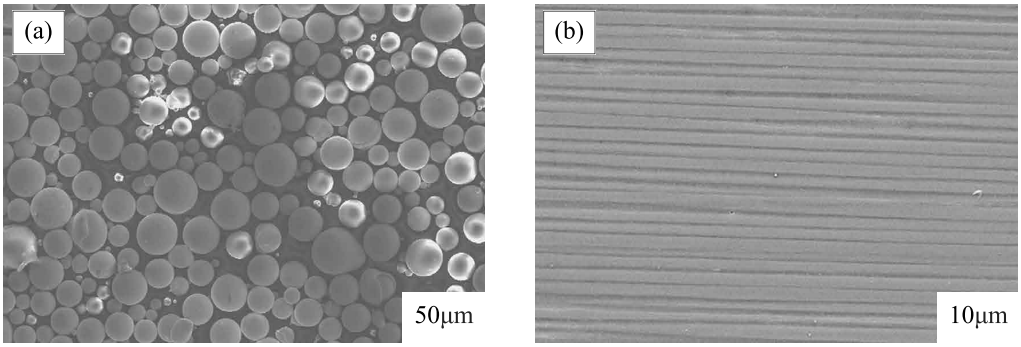

FIGURE 3.1: (a): $Zr_48Cu_36Ag_8Al_8$ metallic glass particulate reinforcement (reproduced from [116]). (b) Cross sectional uniform distribution of carbon fibres in epoxy matrix (reproduced from [117])

There are three approaches for addressing the challenge of predicting the effective mechanical properties of a composite, namely experimental, analytical and numerical procedures. Experimental methods provides accurate results, but requires appropriate materials and means to perform laboratory tests. Moreover, while experimental data are accurate in particular scenarios, their predictive capabilities in the presence of changes in experimental conditions are limited. On the other hand, a multitude of analytical approaches have been proposed in the literature, such as the Mori-Tanaka [118] method or the Hashin-Shtrikman [112] bounds, just to name a few. However, these methods are limited by the need to rely on assumptions and simplifications of the underlying microstructure. In view of the above, numerical methods have become particularly popular owing to their ability to represent the exact geometry of complex composites [119]. However, the development of high-fidelity models of material microstructures is a challenging task involving extraordinary computational burden [120]. In this light, surrogate modelling techniques have opened vast new possibilities to overcome such limitations.

3.1.2 Representative volume elements

One of the most popular numerical approaches for determining the effective properties of composites consists on virtually testing a *representative volume element* (RVE) of the material using Finite Element (FE) methods. An RVE is defined as a sample of a composite that is structurally typical on average of the entire material and contains a sufficient number of inclusions for the apparent overall moduli to be effectively independent of the surface values of traction and displacement [121]. Consider a composite of interest occupying a domain with boundary $\Gamma = \partial\Omega$. Let Ω_1 denote the region occupied by the filler material embedded within the matrix Ω_0, each with their respective boundaries Γ_1 and Γ_0 verifying $\Gamma_0 \cup \Gamma_1 = \Gamma$ and $\Gamma_0 \cap \Gamma_1 = \emptyset$. The elastic response of the material to a certain boundary force field $\mathbf{t}$, an imposed displacement field at the boundary $\bar{\mathbf{u}}$, and a volume force field $\mathbf{g}$ is governed by the elliptic steady-state problem described by Eq.

(3.1):

$$\begin{cases} -\text{div}\ [\mathbf{C}(\mathbf{x})\boldsymbol{\varepsilon}(\mathbf{x})] = \mathbf{g} & \text{in } \Omega \\ \mathbf{u}(\mathbf{x}) = \bar{\mathbf{u}} & \text{on } \Gamma_0 \\ \mathbf{C}(\mathbf{x})\boldsymbol{\varepsilon}(\mathbf{x}) \cdot \mathbf{n} = \mathbf{t} & \text{on } \Gamma_1 \end{cases} \tag{3.1}$$

where $\mathbf{u}(\mathbf{x})$ denotes the displacement field and $\mathbf{n}$ the outer unit normal to the domain boundary Γ. In this context, the strain tensor $\boldsymbol{\varepsilon}$ can be obtained through the displacement field $\mathbf{u}(\mathbf{x})$ by:

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{3.2}$$

and related to the stress tensor by the generalized Hooke's law $\sigma(\mathbf{x}) = \mathbf{C}(\mathbf{x})\varepsilon(\mathbf{x})$, where $\mathbf{C}(\mathbf{x})$ refers to the stiffness tensor of the composite at a point $\mathbf{x}$ within the domain. Specifically, $\mathbf{C}(\mathbf{x})$ takes the form of either $\mathbf{C}^1(\mathbf{x})$ or $\mathbf{C}^0(\mathbf{x})$, depending on whether the point $\mathbf{x}$ lies within Ω_1 or Ω_0, with $\mathbf{C}^1$ and $\mathbf{C}^0$ denoting the elastic tensors of the filler and the matrix, respectively.

In this light, the elastic properties of the composite are determined by the stiffness tensor $\mathbf{C}$. However, extracting the exact solution of the displacement field from Eq. (3.1) is a challenging task, and closed-form solutions can only be found for simplified or ideal microstructures. Consequently, it suffices to determine the overall behaviour of the composite represented by the homogenised elastic tensor $\mathbf{C}^*$ satisfying the problem

$$\begin{cases} -\text{div}\ [\mathbf{C}^*\boldsymbol{\varepsilon}(\mathbf{x})] = \mathbf{g} & \text{in } \Omega \\ \mathbf{u}(\mathbf{x}) = \bar{\mathbf{u}} & \text{on } \Gamma_0 \\ \mathbf{C}^*\boldsymbol{\varepsilon}(\mathbf{x}) \cdot \mathbf{n} = \mathbf{t} & \text{on } \Gamma_1. \end{cases} \tag{3.3}$$

In order to solve the problem given by Eq. (3.3), once defined a suitable RVE of the investigated composite, it is subjected to a series of virtual tests by imposing certain boundary conditions on its external faces. In this work periodic boundary conditions given by Eq. (3.4) are adopted due to their acknowledged higher accuracy for low-to medium-scale RVEs [122]

$$u_i = \hat{\varepsilon}_{ij}x_j + v_i, \tag{3.4}$$

where $\hat{\varepsilon}_{ij}$ denote the volume average strains, and v_i represents the local periodic part of the displacement components u_i on the boundary surfaces. As a result, the proposed RVE shall satisfy geometrical periodicity. For better understanding, Fig. 3.2 (a) depicts the modelling of the material as the periodic replication of a suitable RVE. In a similar manner, Fig. 3.2 (b) shows the constraint equations to be defined for a pair of nodes located on the opposite surfaces A^- and A^+, where S_{ij} stands for the components of the compliance matrix in Hooke's law and U_i is the is the displacement tensor in the i-direction.

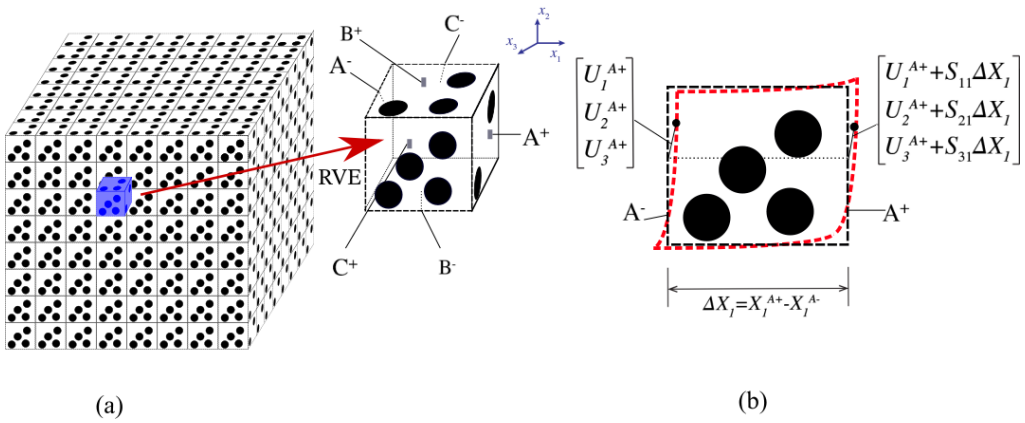

FIGURE 3.2: (a) Representation of a periodic composite and definition of the RVE. (b) Periodic boundary conditions for a pair of nodes located on the opposite surfaces A^- and A^+.

In view of the preceding discussion, the volume average stresses $\hat{\sigma}_{ij}$ and strains $\hat{\varepsilon}_{ij}$ in the RVE can be computed as:

$$\hat{\sigma}_{ij} = \frac{1}{V}\int_V \sigma_{ij} \mathrm{d}V, \quad \hat{\varepsilon} = \frac{1}{V}\int_V \varepsilon_{ij} \mathrm{d}V, \tag{3.5}$$

where V denotes the volume of the considered RVE. Finally, the ij-th component of the elastic tensor can be directly estimated as $\mathbf{C}^*_{ij} = \hat{\sigma}_{ij}/\hat{\varepsilon}_{ij}$. Interested readers can refer to reference [123] for an extended discussion on numerical homogenisation.

In order to ensure realistic modelling of the material microstructure, the proposed RVE must allow for a degree of randomness in the distribution of the doping fillers. However, developing an RVE with a high volume proportion of particle reinforcement is a challenging task, as classical algorithms for randomly placing spheres in a cubic model only achieve a maximum filler volume fraction of 30% [124]. For this reason, a *dropping and rolling* algorithm has been employed in this work for the development of the RVE [125]. This process consists of simulating a physics system for the particles, where inclusions are sequentially introduced in the RVE under a gravitational field, and the stacking process is governed by rigid contacts and sliding with pre-existing particles. The ultimate filler volume fraction ξ_1 is controlled via an inter-particle parameter describing the minimal distance between particles. To ensure the periodicity of the microstructure, the spheres are allowed to bypass the walls and reappear on the opposite side of the box. A schematic description of the algorithm is provided in the Fig. 3.3. Examples of 2D and 3D RVEs developed for this work are depicted in Fig. 3.4.

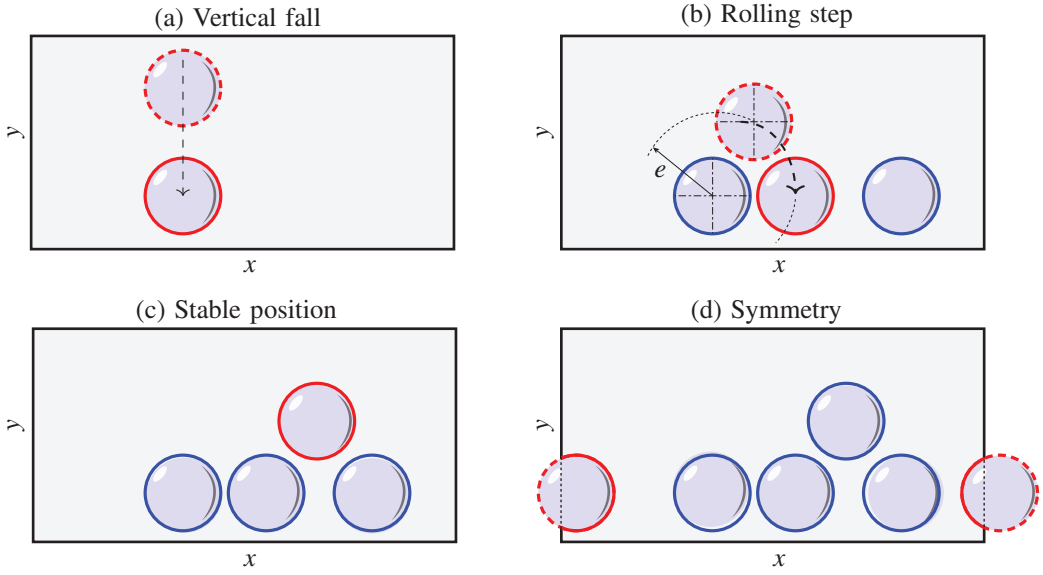

FIGURE 3.3: Outline of the dropping and rolling algorithm for 2D RVEs.

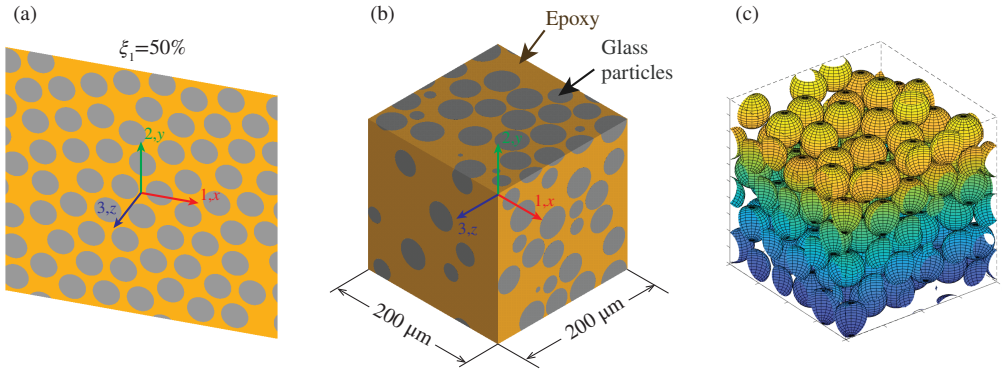

FIGURE 3.4: (a) Example of a periodic 200 µm edge 2D RVE matrix doped with $\xi_1 \approx 50\%$ of glass particles with radius 10 µm. (b) RVE doped with $\xi_1 \approx 50\%$ volume fraction of glass microspheres with radius 20 µm. (c) Example of a 3D RVE generated particle distribution by means of the rolling and dropping procedure with $\xi_1 \approx 50\%$.

For the construction and homogenisation of the microstructure, a combination of scripts generated in MATLAB environment and the commercial FEM tool ANSYS [126] is used. Once the geometry is constructed in Matlab, it is discretised in ANSYS using plane-strain 8-nodes quadratic elements (PLANE 183) for 2D RVEs, while 4-nodes linear tetrahedral solid elements (SOLID 285) are used for 3D RVEs. At the end, a numerical homogenisation script is executed in ANSYS. It is important to emphasise that the homogenisation process takes only a few minutes for 2D models but can exceed 6 hours in the case of 3D models. This renders conducting traditional Monte Carlo analyses infeasible, making necessary the use of surrogate models.

3.1.3 PCE-assisted Uncertainty Propagation analysis

Following the theoretical framework introduced in Section 2.2, two PCE surrogate models, namely $\hat{\mathcal{M}}^{2D}$ and $\hat{\mathcal{M}}^{3D}$, have been constructed in order to compare how uncertainty

is propagated through different material models. The variables of interest under consideration were the filler volume fraction and the mechanical properties (Young's moduli and Poisson's ratios) of the composite constituents. The former was assumed uniformly distributed and ranging from 0 to 50%. On the other hand, the elastic properties of the phases were considered to be normally distributed with the statistical properties reported in Table 3.1 following Ref. [127]. Note that the standard deviations of the mechanical properties have been obtained by imposing a total uncertainty (3σ) of $\pm 10\%$ the mean values of each distribution [128]. After establishing the distribution of input variables, an ED of 50 points and an independent validation set of 25 points were sampled by means of the LHS method described in Section 2.1. Subsequently, RVE construction and homogenisation scripts were executed for each combination of parameters considered. For the concluding step of PCE model fitting, polynomial expansions of degree up to 4 were selected, and $q = 0.95$ was taken in Eq. (2.4). In order to enhance the accuracy and robustness of the meta-model, the LAR method for selecting an adaptive basis was also applied to this case study.

TABLE 3.1: Statistical distributions of the material parameters.

Variable	**Notation**	**Distribution**	
Young's modulus of epoxy matrix [GPa]	E_0	Gaussian	$\mathcal{N}(3, 0.1)$
Poisson's ratio of epoxy matrix [-]	ν_0	Gaussian	$\mathcal{N}(0.4, 0.0133)$
Young's modulus of glass reinforcement [GPa]	E_1	Gaussian	$\mathcal{N}(76, 2.5333)$
Poisson's ratio of glass reinforcement [-]	ν_1	Gaussian	$\mathcal{N}(0.23, 0.0076)$
Volume fraction of inclusions [-]	ξ_1	Uniform	$\mathcal{U}(0, 0.5)$

It can be observed in Fig. 3.5 that the surrogate models $\hat{\mathcal{M}}^{2D}$ and $\hat{\mathcal{M}}^{3D}$ faithfully reflect the behaviour of the corresponding models $\mathcal{M}^{2D}$ and $\mathcal{M}^{3D}$ on the independent validation set, with $R^2 > 0.99$ in both cases.

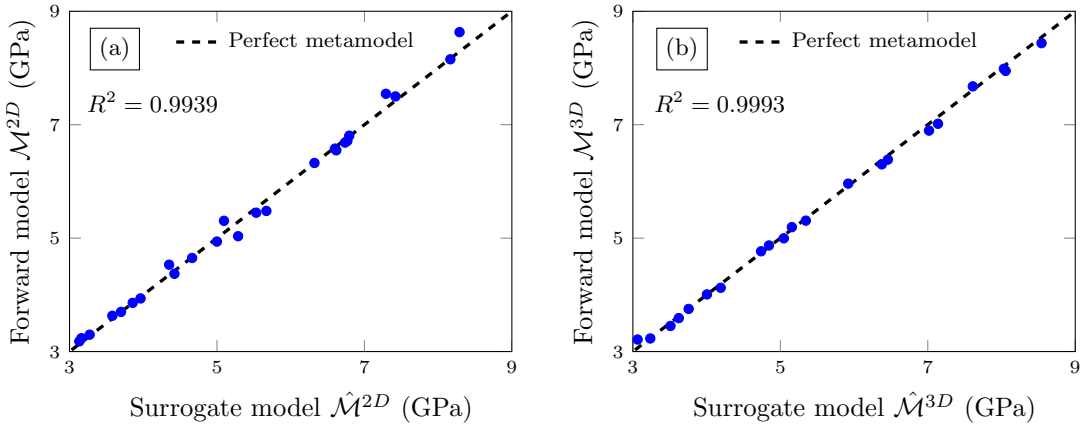

FIGURE 3.5: Comparison of the numerical estimates by FEM model and predictions of the PCE surrogates in terms of the elastic moduli in (a) 2D and (b) 3D cases.

For UP analyses, in both 2D and 3D scenarios, 25 000 Monte carlo samples were drawn from the parameter space described in Table 3.1. The use of the metamodel enables this procedure to take less than a second. Results of the process are depicted in Fig. 3.6, where red dashed lines represent the 95% confidence levels. In order to validate the numerical predictions, the characterisation data reported by Smith in reference [129] for the same composite investigated herein are included in the figure. It can be observed that the 3D model fits the data slightly better, especially in the high volume fraction region. Interestingly, it is noted that the uncertainty in the Young's modulus increases for larger filler contents (higher that 25%) in the 2D scenario, while remains almost constant when considering 3D RVEs. The results demonstrate the usefulness of simplified 2D models for rapid analysis, but also reveal their limitations in capturing the behaviour of three-dimensional composites reinforced with doped particles at high filler volume fractions.

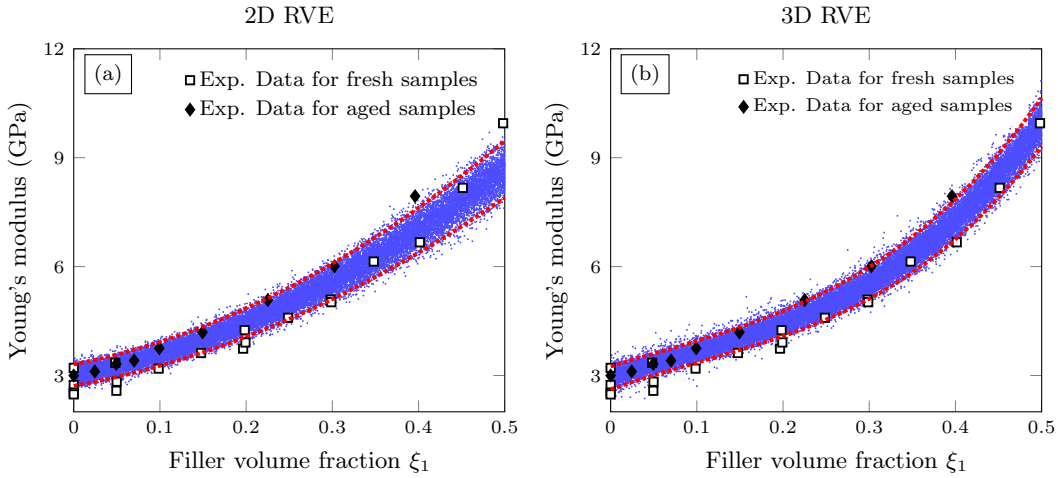

FIGURE 3.6: Comparison between the prediction of the Young's moduli of epoxy/glass composite versus the filler volume fraction ξ_1 in the 2D (a) and 3D (b) cases. 95% confidence levels are indicated with red dashed lines. Experimental data retrieved from reference [129].

Finally, it is worth noting that the presented approach provides additional benefits. Particularly interesnting is the possibility of performing uncertainty propagation analysis for each of the variables separately by simply keeping fixed the other parameters to their mean value. For example, Fig. 3.7 depicts the estimated PDFs from the Monte Carlo samples of the composite of interest doped with 15% and 45% filler respectively, after varying each of the mechanical properties reported in Table 3.1. It can be seen that for low filler volume fractions uncertainties in E_0 dominate the effective Young's modulus of the composite. In fact, almost no effect is observed for variations in the elastic modulus of the glass inclusions and the remaining parameters. Nonetheless, for higher filler volume fractions, significant influence is found due to uncertainty in the Poisson's ratio of the host matrix.

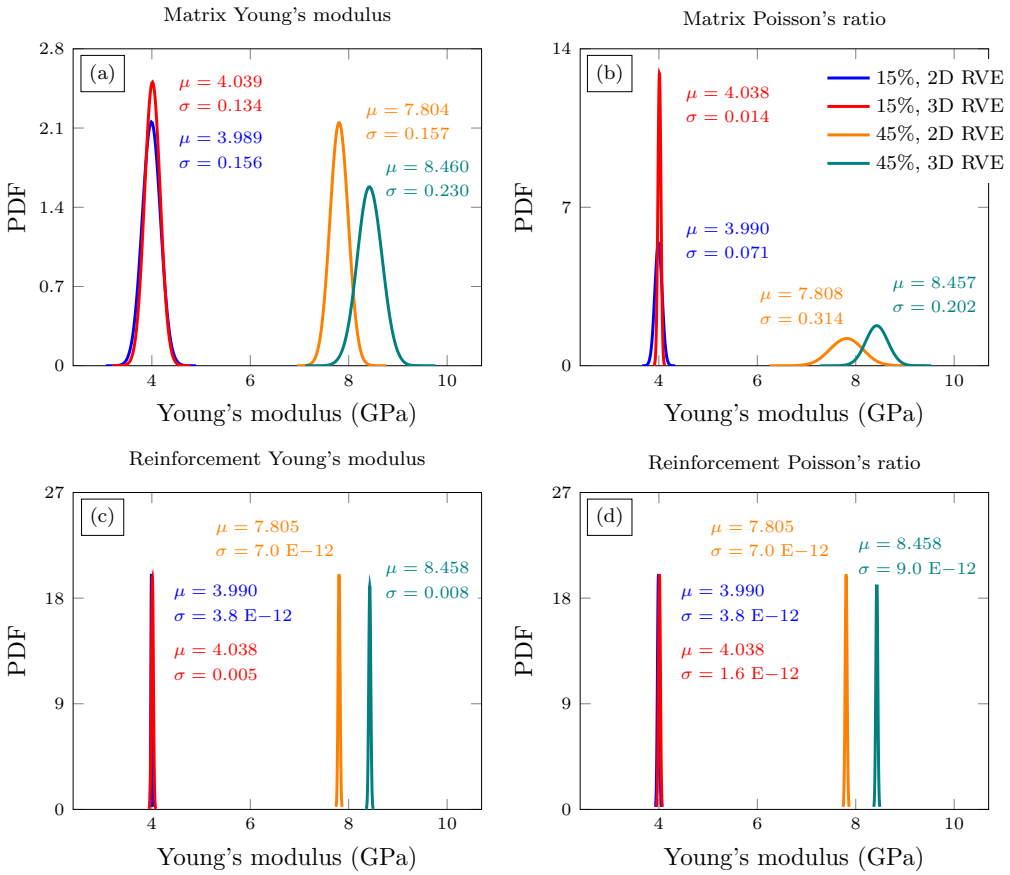

FIGURE 3.7: PDFs of elastic properties of 2D and 3D composites made of epoxy doped with 15% and 45% of glass inclusions considering uncertainties in the matrix Young's modulus (a), matrix Poisson's ratio (b), Young's modulus (c) and reinforcement Poisson's modulus (d).

3.2 Surrogate assisted MCMC: Multi-element PCK applied to TDS flux curves

PCE and Kriging metamodelling techniques presented in Sections 2.2 and 2.3 constitute both excellent options to approximate relatively simple models exhibiting smooth behaviour. Nevertheless, great challenges emerge when the forward model presents substantial non-linearities. The approach outlined in paper B consists of a simple regular block partitioning approach. On this basis, the space domain is partitioned into distinct subregions with splitting directions chosen on the basis of a preliminary sensitivity analysis, prioritising the division of those parameters exhibiting the highest sensitivities. This approach allows to accommodating nonlinear behaviours in the forward model while retaining flexibility and minimal computational cost.

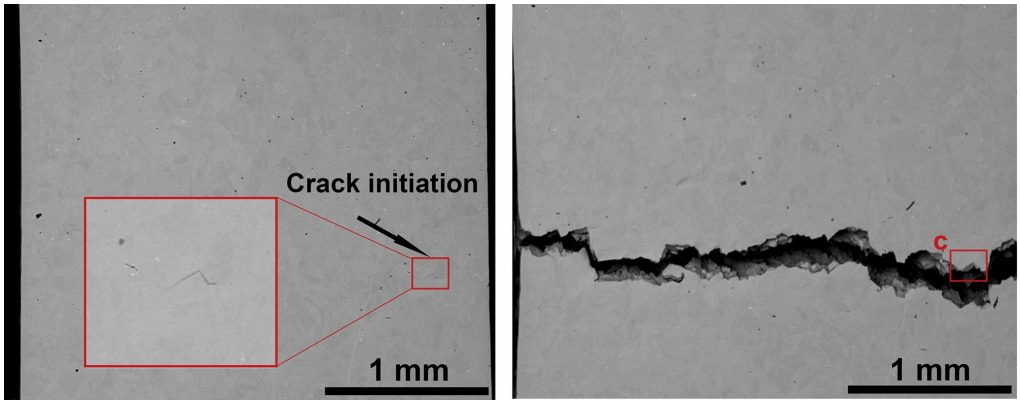

FIGURE 3.8: SEM images of the evolution of an hydrogen-assisted crack in a nickel-based metallic alloy sample during in-situ tension (reproduced from [130]).

A formidable challenge for surrogate modelling is given by the Thermal Desorption Spectroscopy (TDS) governing diffusion equation. TDS is an experimental technique employed to study the desorption of molecules from a solid surface as a function of temperature, being key for characterising the Hydrogen embrittlement (HE) phenomenon in metals. HE involves the loss of ductility and toughness of metallic alloys induced by hydrogen atoms deposited at micro-structural defects as dislocations, grain boundaries or vacancies [131, 132], which may lead to the appearance of cracks with the subsequent risks of unpredictable and premature failures in systems exposed to hydrogenated environments (Fig. 3.8). In this context, a TDS experiment comprises exposing a metallic specimen to a gas of interest at a known pressure and temperature, allowing gas molecules to be adsorbed onto the surface of the sample. Then, the sample is heated at a controlled rate and, as the temperature increases, the adsorbed molecules gain enough energy to overcome the binding forces holding them to the surface, causing desorption. Finally, the flux of desorbing molecules is monitored using a mass spectrometer [133]. In light of the above, an hydrogen trap can be defined as a site on the surface of a metal where hydrogen atoms are captured due the aforementioned defects. When a metallic sample is heated during a TDS analysis, the trapped hydrogen atoms are released from the surface and detected. For a better understanding, the kinetics of hydrogen trapping

and detrapping in metals is commonly described with a two-level system as sketched in Fig. 3.11 (b) for the case of a single trap.

The interpretation of TDS data for the characterisation of hydrogen traps is of pivotal importance for the design of HE-resistant alloys. However, the process of characterising these traps via Bayesian inference is a very challenging task due to two primary reasons: firstly, the significant computational burden required to solve the partial differential equation (PDE) governing TDS experiments, and secondly, the presence of the multi-modality phenomenon described in Section 2.4.3. In order to circumvent these difficulties, a metamodel-assisted version of the adaptive MCMC DRAM algorithm developed by Haario et al. [64] was implemented for maximum efficiency.

The rest of the section is divided into two parts: firstly, the detailed development of the metamodelling approach is presented, followed by its application for Bayesian analysis in the TDS context.

3.2.1 Multielement Polynomial Chaos based Kriging

The PCK metamodel technique presented in Section 2.3.1 suffers from low convergence rates when the forward model $\mathcal{M}$ exhibits non-smoothness [134]. Thus, achieving accurate predictions often demands significantly larger ED sizes. This aspect undermines the computational efficiency of the PCE-based Kriging model, which is dominated by the $\mathcal{O}(N^3)$ complexity of the Kriging predictor. As a result, long construction times or even memory overflow issues emerge when solving the hyperparameters optimisation step described in Eqs. (2.25) and (2.26). Furthermore, the larger the size of the ED, the slower the evaluation of the corresponding metamodel, reducing or vanishing the advantages of the surrogate approach. To address this issue, a multi-element PCK (ME-PCK) model inspired by the ME-gPC method by Wan and Karniadakis [135] is proposed. This technique consists in partitioning the random input space into a finite set of non-overlapping subdomains. Then, a local PCK surrogate model is constructed in each subdomain following the formulation in Section 2.3.1. Finally, these local surrogates are assembled into a global metamodel, forming a piece-wise function. In order to determine the direction of the partitions, an approach relying on sensitivity analysis based on the Sobol's indices has been adopted. Note that, as referred in Section 2.2, the Sobol's indices can be readily computed as a by-product of the PCE. Therefore, priority in the partitioning is given to the direction of those parameters with highest sensitivity, i.e., those with the greatest effect on the variability of the QoI. Regarding the total number of partitions, it can be a priori determined depending on the available computational resources. Increasing the number of partitions will always imply improvements in the metamodel tuning and evaluation speed, but its effect on the approximation accuracy is unclear.

On this basis, given a computational model $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^M \to \mathbb{R}$, a decomposition of the domain $\mathcal{D}$ is defined as a collection $\{\mathcal{D}_j \subset \mathcal{D}, j \in \mathcal{J}\}$ satisfying

$$\mathcal{D} = \bigcup_{j \in \mathcal{J}} \mathcal{D}_j, \quad \mathcal{D}_j \cap \mathcal{D}_{j'} = \emptyset, \text{ if } j \neq j', \tag{3.6}$$

After setting a domain partition, for each region $\mathcal{D}_j$ a local ED is sampled and a PCK surrogate -denoted by $\hat{\mathcal{M}}^j$- is subsequently constructed. Thus, denoting by $I_{\mathcal{D}_j}$ the indicator variable:

$$I_{\mathcal{D}_j}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{D}_j \\ 0 & \text{otherwise} \end{cases}$$

the global Multielement PCK metamodel is defined as:

$$\hat{\mathcal{M}}^{ME-PCK}(\mathbf{x}) = \sum_{j \in \mathcal{J}} I_{\mathcal{D}_j}(\mathbf{x}) \hat{\mathcal{M}}^j(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{D}. \tag{3.7}$$

The efficiency and accuracy of the proposed ME-PCK surrogate have been validated first by approximating an exceptionally complex analytical benchmark function and then by replicating the flux curves resulting from solving the governing diffusion equation characterising TDS experiments. The first case study explores the Drop-Wave function, also known as the Salomon function [136], given by:

$$f : [-10, 10]^2 \to \mathbb{R}, \quad f(x_1, x_2) = 1 - \cos\left(2\pi\sqrt{x_1^2 + x_2^2}\right) + 0.1\sqrt{x_1^2 + x_2^2}. \tag{3.8}$$

Twelve surrogate models were constructed for illustrative purposes, varying the size of the ED and the number of domain subdivisions. In particular, four experimental design sets $\text{ED}_i \subset \mathcal{D}$, $i = 1, \ldots, 4$ containing 360, 720, 1440 and 2880 samples respectively have been defined. Additionally, three different domain partition schemes, namely $\mathcal{P}^j$, $j = 1, \ldots, 3$, have been considered. These include $\mathcal{P}^1 = [-10, 10]^2$, $\mathcal{P}^2 = ([-10, 0) \cup [0, 10])^2$, and $\mathcal{P}^3 = \left(\left[-10, \frac{-10}{3}\right) \cup \left[\frac{-10}{3}, \frac{10}{3}\right) \cup \left[\frac{10}{3}, 10\right]\right)^2$, leading to a total of two, four and nine sub-domains, respectively. The ME-PCK metamodels developed for this example are denoted by $\mathcal{M}_i^j$, where the subscript indicates the size of the ED and the superscript the partition under consideration. The landscape of the investigated function as well as its approximation by five of the constructed surrogate models is depicted in Fig. 3.9.

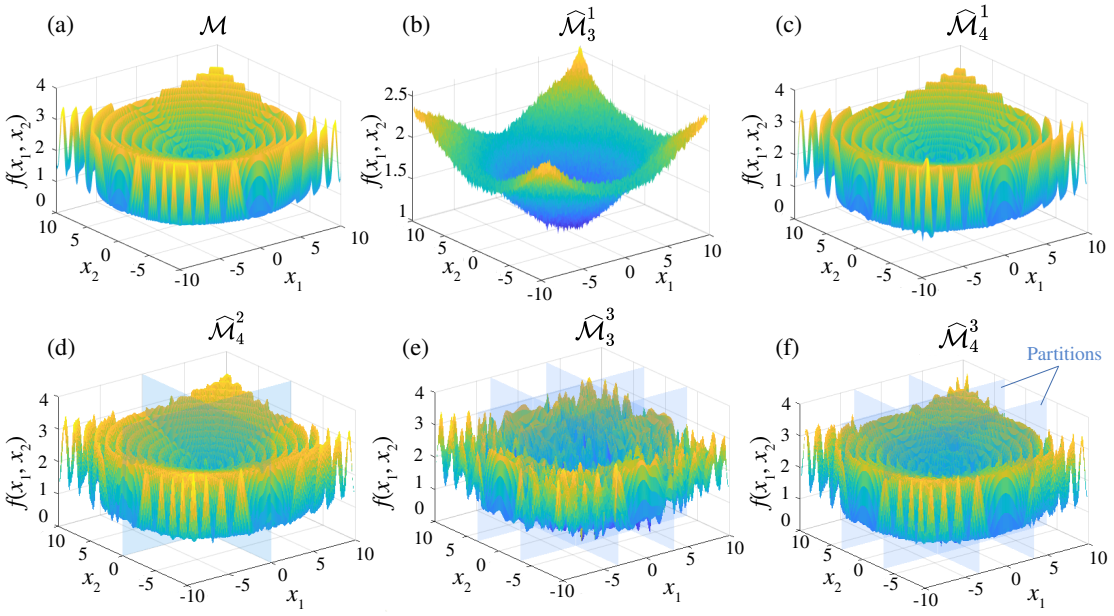

FIGURE 3.9: Exact response surface of the Drop-Wave function (a), and predictions by surrogate models $\hat{\mathcal{M}}_1^3$ (b), $\hat{\mathcal{M}}_4^1$ (c), $\hat{\mathcal{M}}_4^2$ (d), $\hat{\mathcal{M}}_3^3$ (e) and $\hat{\mathcal{M}}_4^3$ (f).

It can be observed in Fig. 3.9 (a) that the Drop-Wave function is highly non-linear. Among the presented surrogates, those trained with larger EDs, namely, $\hat{\mathcal{M}}_4^1$, $\hat{\mathcal{M}}_4^2$ and $\hat{\mathcal{M}}_4^3$, clearly outperforms the others as expected. However, the number of subdomains do not seem to systematically improve or degrade prediction accuracy. Comparing models $\hat{\mathcal{M}}_3^1$ y $\hat{\mathcal{M}}_3^3$, the latter is undoubtedly superior to the former. Conversely, when ED_4 is considered, the global PCK surrogate outperforms the Multielement approach in terms of accuracy. A more detailed analysis including evaluation times of the different metamodels is depicted in Fig. 3.10. It is important to highlight that partitioning in a higher number of regions leads to lower evaluation times for all the considered EDs. This reduction is mainly due to the reductions in the computational cost involved in the construction of the Kriging predictor in Eq. (2.21). Additionally, it should be noted that the accuracy of metamodels trained with a larger number of subdomains is higher compared to those with no partitions for limited to moderate EDs, but this effect is inverted for the larger EDs. Nevertheless, taking both metrics into account, the ME-PCK clearly provides a cost-effective option for integrating the surrogate model into a Bayesian inference scheme where evaluation time is crucial. Furthermore, the ME-PCK offers additional benefits. For instance, adjusting the $\hat{\mathcal{M}}_4^1$ model takes over 2900 seconds, almost twelve times longer than the 245 seconds required to construct $\hat{\mathcal{M}}_4^3$. An in-depth analysis of the accuracy and the evaluation and construction times of each metamodel can be found in paper B.

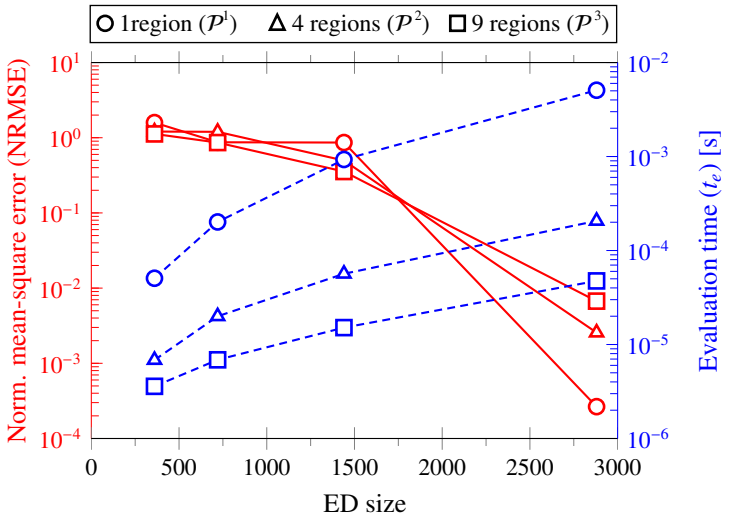

FIGURE 3.10: Performance assessment of PCK surrogate models for the Drop-Wave function (validation set of 20 000 samples).

Having established the capabilities of the ME-PCK approach, the metamodelling of TDS governing equation represents an exceptional case study to justify the use of the presented approach in real engineering applications. Consider a one-dimensional specimen of length L as sketched in Fig. 3.11 (a). Following the work by Raina et al. [137], a non-dimensional version of the governing PDE describing the hydrogen diffusion in a TDS test is given by:

$$\frac{\partial \overline{\theta}_L}{\partial \overline{t}}\left[1+\sum_{i=1}^{N_t}\frac{\overline{K}_i\,\overline{N}_i}{\left(1+\overline{K}_i\,\theta_L^0\,\overline{\theta}_L\right)^2}\right]+\frac{\overline{\theta}_L}{\overline{T}^2}\sum_{i=1}^{N_t}\frac{\overline{K}_i\,\overline{N}_i\,\overline{\Delta H}_i\,\overline{\phi}}{\left(1+\overline{K}_i\,\theta_L^0\,\overline{\theta}_L\right)^2}=\overline{D}_L\frac{\partial^2\overline{\theta}_L}{\partial\overline{x}^2}, \tag{3.9}$$

which represents an extended diffusion Fick's law with sink terms. In this model, N_t stands for the number of traps, θ_L^0 refers to the lattice occupation of hydrogen atoms at the beginning of the process. The remaining non-dimensional variables involved in Eq. (3.9) are described in Table 3.2.

The initial and boundary conditions of the PDE in Eq. (3.9) are schematically presented in Fig. 3.11 (a). Denoting by $\overline{\theta}_L = \theta_L/\theta_L^0$, at $t = 0$ it is assumed an initial uniform lattice occupancy $\overline{\theta}_L\left(\overline{x}, \overline{t}=0\right) = 1$. Thereafter, the hydrogen lattice occupancy is assumed zero at the boundaries, that is $\overline{\theta}_L\left(\overline{x}=\pm 1/2, \overline{t}>0\right)=0$. As temperature raises, the lattice occupancy evolves spatially and temporally as sketched in Fig. 3.11 (c), and the flux of hydrogen atoms $J(t)$ diffusing out at boundaries is measured as presented in Fig. 3.11 (d). This flux can be obtained in non-dimensional terms after solving Eq. (3.9) as [137]:

$$\overline{J} = -\overline{D}_L\,\theta_L^o\,\frac{\partial\overline{\theta}_L}{\partial\overline{x}}. \tag{3.10}$$

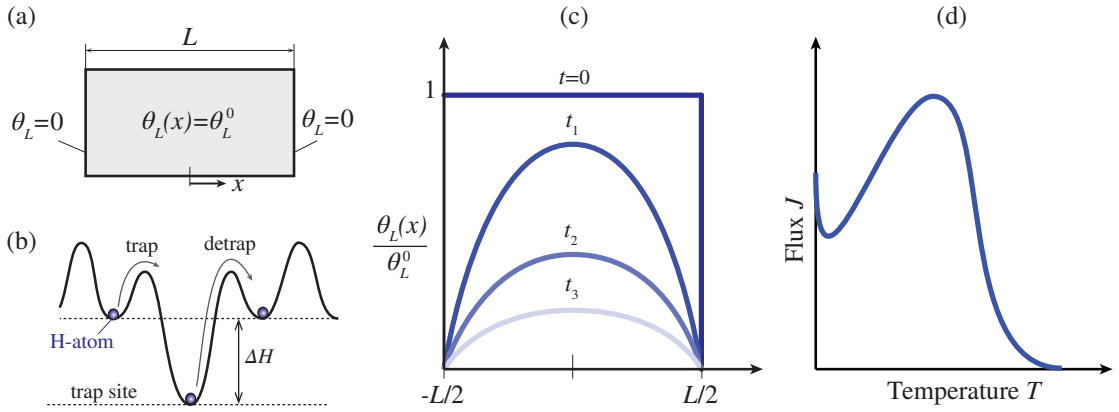

FIGURE 3.11: (a) A schematic illustration of initial and boundary conditions in a TDS test. (b) Schematic definition of binding energy in a one-dimensional diffusion path. (c) Transient solution curves of the normalised lattice occupancy fraction θ_L/θ_L^0 at different times t along the specimen's thickness. (d) A schematic of typical hydrogen desorption flux versus temperature curves obtained in a TDS test.

TABLE 3.2: Non-dimensional variables used in the hydrogen diffusion PDE employed for the TDS tests.

Spatial coordinate	$\overline{x}$	Lattice activation energy	$\overline{Q}$
Time coordinate	$\overline{t}$	Trap binding energy	$\overline{\Delta H}_i$
Heating rate	$\overline{\phi}$	Lattice diffusion coefficient	$\overline{D}_L$
Trap density	$\overline{N}_i$	Local equilibrium constant	$\overline{K}$
Temperature	$\overline{T}$	Fractional lattice occupancy	$\overline{\theta}_L$

In general, most of the variables involved in Eq. (3.9) are known constants dependent on the material under study or inputs of the system controlled by the researcher. Therefore, given specific experimental conditions, a hydrogen flux curve resulting from a TDS analysis is determined as a function of the trap binding energies $\overline{\Delta H}$ and trap densities $\overline{N}$ of the existing hydrogen traps. Depending upon the hydrogen trap configuration, several different regimes can be observed, as previously discussed by Raina et al. [137]. Specifically, their results for the case of metals containing a single trap showed that no peak flux is attained for low trap densities and binding energies. Alternatively, when a peak flux is found, those authors identified two distinct regimes (I and II) originated by two types of microstructural defects, referred to as shallow and deep traps. Shallow traps are characterised by large trap densities, and give origin to peak fluxes that are highly sensitive to both $\overline{N}$ and $\overline{\Delta H}$. On the other hand, deep traps are characterised by low trap densities, resulting in peak fluxes that are insensitive to the trap binding energy. The existence of these different regimes turns the construction of a surrogate model covering the whole domain of the traps into an notably challenging task. However, it is possible for two or more hydrogen traps to coexist and interact within the specimen being investigated, resulting in the formation of new types of curves and sharply increasing metamodelling difficulty. For the development of this case study, the possible presence of up to two traps has been considered. In light of the preceding discussion, in the surrogate modelling step, temperature and the trap densities and binding energies of each trap are considered as input variables, which amounts to five design variables. That

is, $\mathbf{x} = \left[\overline{T}, \overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2)\right]^{\mathrm{T}} \subset \mathbb{R}^5$. The variation of trap binding energies and trap densities are selected as the physically meaningful ranges $-40 \leq \overline{\Delta H}_i \leq -10$ and $10^{-7} \leq \overline{N}_i \leq 10^{-2}$. Temperature is set to range between 293 and 1200 Kelvin degrees. The values of the remaining physical quantities have been chosen to illustrate the behaviour of a typical ferritic steel sample following Ref. [137]

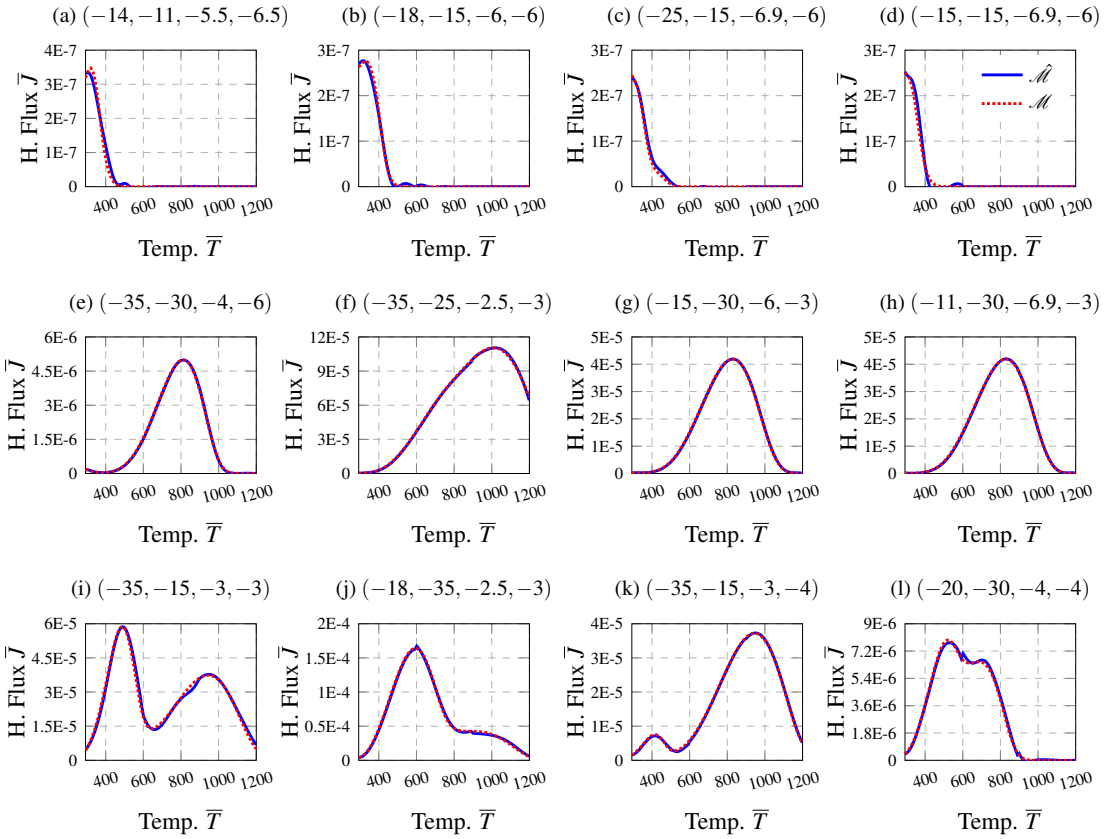

FIGURE 3.12: Surrogate modelling of the Hydrogen flux curves obtained by TDS of metals with different values of trap binding energies and concentrations. Quantities in parenthesis represent the parameters of the traps $\left(\overline{\Delta H}_1, \overline{\Delta H}_1, \log(\overline{N}_1), \log(\overline{N}_2)\right)$.

The ME-PCK metamodel employed to approximate the TDS curves for posterior Bayesian inference was constructed by splitting the activation energy domains ($\overline{\Delta H}_i$, $i = 1, 2$) in two, while three segments were considered for the partition of the domain of the trap densities ($\log(\overline{N}_i)$, $i = 1, 2$) and the temperature $\overline{T}$. An ED consisting of 1000 sample points per region was selected using the MIPT method described in Section 2.1, resulting in a total size of 108 000 samples. The final metamodel was assessed using an independent validation set of 36 000 points. In addition, the evaluation time of the metamodel was 2 ms per point, compared to 280 ms for the forward model, a reduction of the computational burden of 99.3%. To illustrate the effectiveness of the surrogate model in representing the different stages observed in the TDS test, Fig. 3.12 shows the comparison of the forward model and the surrogate predictions for a variety of combinations of traps, including the case of fluxes without peak, one single peak, and two peaks. It is clearly observed that the proposed PCK model can accurately reproduce all the different

regimes observable in the TDS test.

It should be noted that the technical details regarding the mathematical development, as well as the experimental conditions of the TDS analysis, have been omitted for briefness in the main text. However, complete derivations and in-depth analyses are included in paper B.

3.2.2 MCMC for TDS governing diffusion equation

In this last subsection, a Bayesian parameter inference analysis using the adaptive DRAM algorithm and the previously defined ME-PCK surrogate model is presented. The unknown parameters on which the inference process was performed were the trap binding energies $(\overline{\Delta H}_1, \overline{\Delta H}_2)$ and densities $(\log(\overline{N}_1), \log(\overline{N}_2))$ of the two trap system previously established in Section 3.2.1. Two experiments, namely EI and EII, were designed to assess the performance of the proposed approach in inferring the parameters of hydrogen traps across the two regions described in reference [137]. In each experiment, synthetic experimental data were generated from the forward model. The first one, corresponding to Fig. 3.12 (f), consists of a two-trap system with properties $\boldsymbol{\theta} = (-35, -25, -2.5, -3)$. On the other hand, the experiment EII is defined by $\boldsymbol{\theta} = (-15, -30, -6, -3)$. The flow curve resulting from this configuration is depicted in Fig. 3.12 (f). In this scenario, the trap the pair $(-15, -6)$ characterising one of the traps is located within the non-flux area identified by Raina and and co-authors [137]. As a result, only the trap defined by the pair $(-30, -3)$ will be identifiable in the posterior inference analysis. Note (see Fig. 3.12 (h)) that modifying the pair $(-15, -3)$ has no effect on the corresponding flux curve if it remains within the non-flux area. Additionally, a second analysis of the EII experiment was performed after affecting the flux curve with a significant zero-mean Gaussian white noise (see Fig. 3.15) in order to demonstrate that the presence of measurement noise does not substantially alter the inference outcome. In both experiments, since the TDS governing equation does not distinguish the order of the traps, the problem is ill-posed and the posterior distribution is expected to exhibit two modes corresponding to two symmetric solutions. Consequently, the challenge for MCMC algorithms will be to explore the PDF of the parameters of interest without getting stuck around one of the theoretical solutions as it is usually the case when implementing standard MCMC methods.

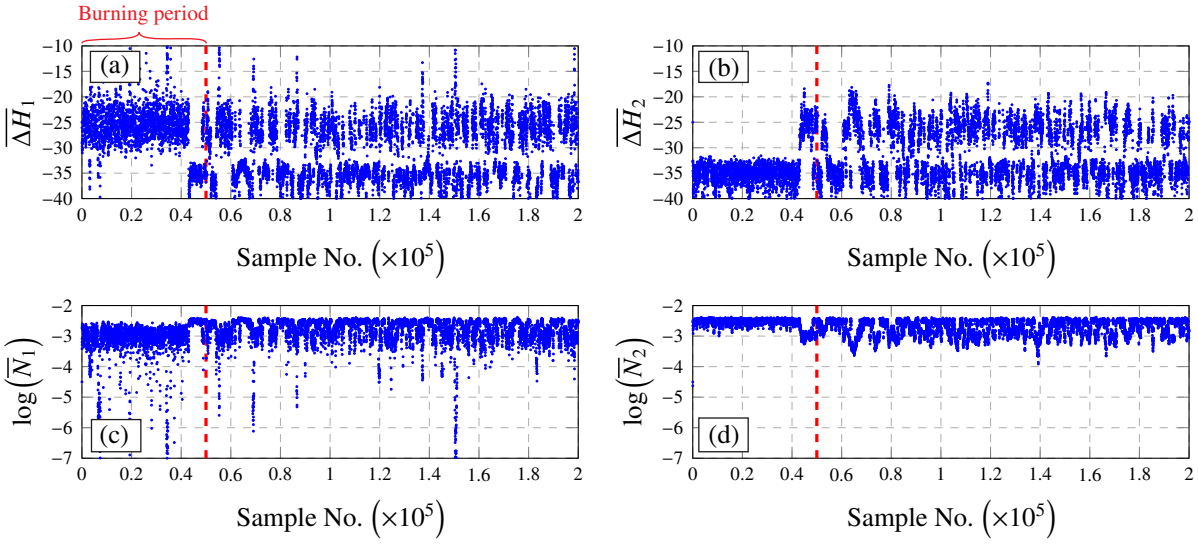

FIGURE 3.13: Markov chains generated by DRAM MCMC of trap parameters $\overline{\Delta H}_1$, $\overline{\Delta H}_2$, $\log(\overline{N}_1)$, $\log(\overline{N}_2)$ for TDS Experiment EI.

The numerical results demonstrate the effectiveness of the proposed method for inferring the characteristic parameters of a hydrogen trap. The Markov chain and the joint posterior PDF obtained for Experiment EI are presented in Figs. 3.13 and 3.14, respectively. A chain containing a total number of 200 000 samples with a burning period of 50 000 samples were drawn utilising the adaptive DRAM algorithm. The significant number of required samples highlights both the necessity of a surrogate model-based approach and the difficulty of addressing multimodal distributions during the Bayesian inference process. As expected, the problem exhibits a bimodal behaviour because of the existence of the two possible solutions $\boldsymbol{\theta} = (-25, -35, -3, -2.5)$ and $\boldsymbol{\theta} = (-35, -25, -2.5, -3)$. Note that during the initial few thousand iterations, the chain remains close to the first solution and subsequently alternates between both. For validation purposes, the posterior PDF has been also computed by direct integration of the forward solution. To do so, the evidence of the model has been computed over a mesh of 60^4 elements. This required forty five hours of parallel computation on ten cores, while the MCMC approach only required about four hours on a single core. The close fittings between the exact marginal PDFs and those predicted by the surrogate model-based Bayesian inference in Fig. 3.14 demonstrate the accuracy of the developed approach.

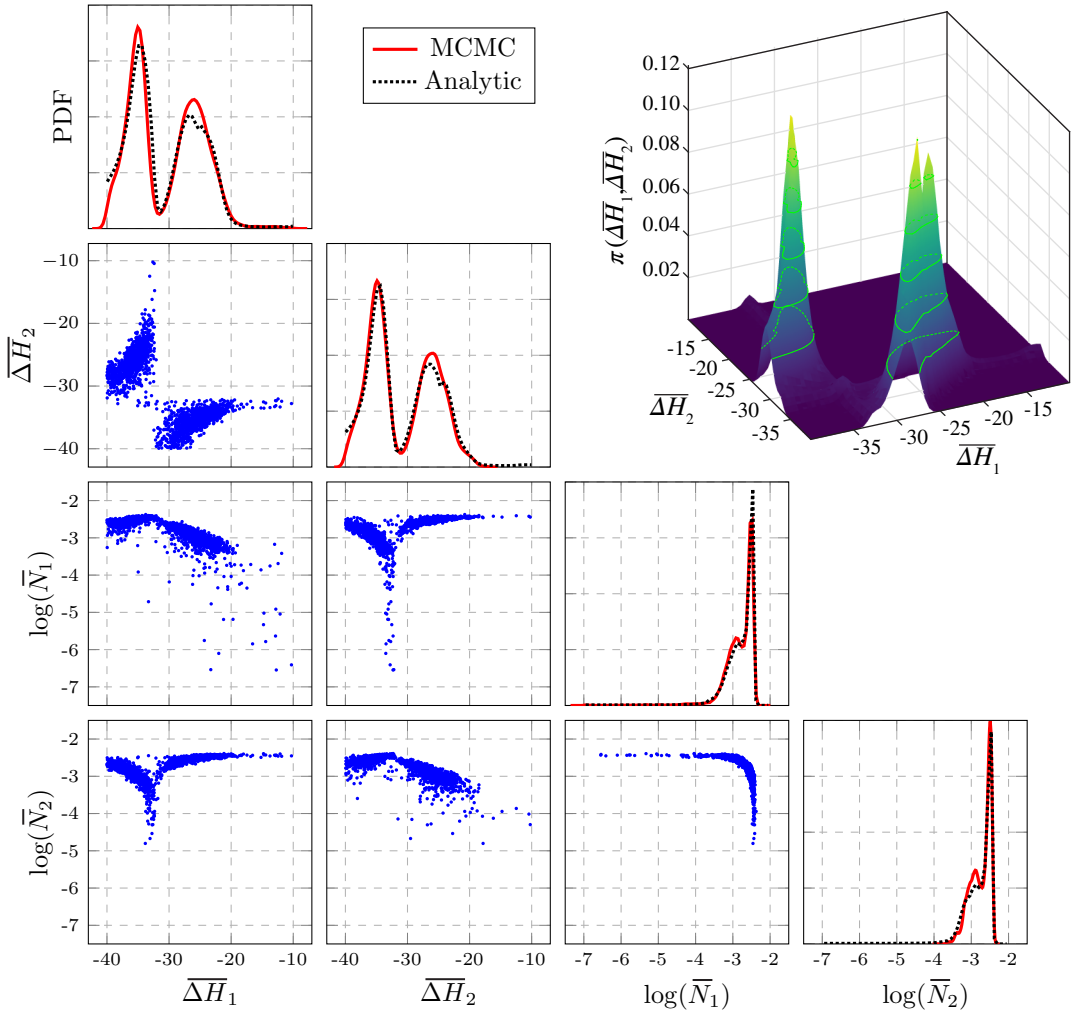

FIGURE 3.14: Bayesian identification results of the trap parameters $\boldsymbol{\theta} = (\overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2))$ of TDS Experiment EI. The surface plot in the top right corner corresponds to the marginal PDF over $(\overline{\Delta H}_1, \overline{\Delta H}_2)$ obtained by numerical integration.

The sampling of the posterior PDF in Experiment EII was more challenging given the existence of large regions of low probability, requiring up to 480 000 samples with a burning period of the first 160 000 to achieve convergence. Interestingly, this phenomenon attenuates when the flux curve is affected by noise, only requiring a chain of 120 000 samples with a burning period of 30 000 to attain convergence. This is expectable since the noise-induced lower probability concentration around the exact true solutions makes it easier for the chain to span from one solution to the symmetric one. The posterior PDF obtained for the TDS experiment EII in both noisy and noiseless scenarios is depicted in Fig. 3.15

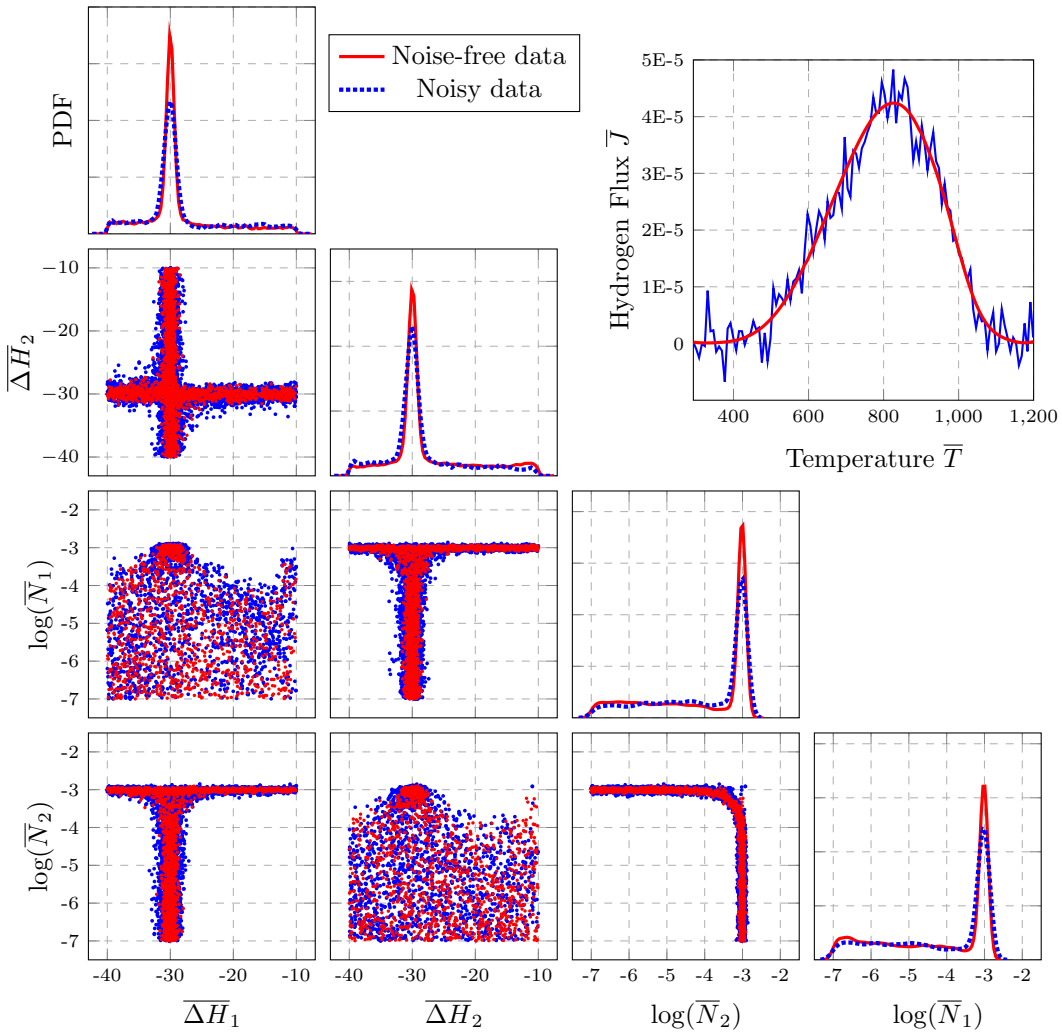

FIGURE 3.15: Bayesian identification results of the trap parameters $\boldsymbol{\theta} = (\overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2))$ of experiment EII with noise unpolluted (red) and polluted data (blue).

An extended discussion including a complete description of DRAM algorithm, technical details of the implemented approach, and further analysis of the obtained posterior PDFs in experiments EI and EII can be found in paper B.

3.3 Sparse Polynomial Chaos for Stochastic Kriging

The last case study included in this thesis involves the development of an extension of Stochastic Kriging (SK) surrogate modelling technique capable of addressing computer stochastic simulations. Firstly, the importance of stochastic simulators to manage uncertainty present in engineering problems is discussed. The remainder of the section presents the theoretical principles of SK as a generalisation of Kriging, introduces the LAR algorithm, and discusses the performance of the proposed approach on some typical case studies from the SK literature.

3.3.1 Beyond deterministic simulators

Throughout this thesis, a computational model has been consistently treated as a function $\mathcal{M} : \mathcal{D} \subset \mathbb{R}^M \rightarrow \mathbb{R}$ mapping some input parameters into a certain QoI to be determined. This approach, referred to as deterministic, is suitable for a wide range of problems where systems or processes are assumed to be predictable and devoid of inherent randomness. However, such a course of action is inappropriate for facing real-world problems characterised by intrinsic or non-parameterisable random behaviour, for which the adoption of stochastic simulators becomes necessary for proper modelling. A stochastic simulator incorporates an intrinsic variance or uncertainty in its predictions. Thus, while a deterministic simulator always produces the same prediction for a fixed input, in stochastic models each simulation run may produce different outputs as a result of its inherent randomness. Such models have gained significant popularity with manifold applications in diverse fields such as epidemiology [4, 138], materials science [139, 140], or reliability analysis [141], enabling researchers to make robust predictions to assist decision-making for complex systems.

Despite the introduction of numerous surrogate modelling techniques in the literature over the past few years to replace intensive numerical models, their success mostly limits to deterministic simulators, while the consideration of stochasticity remains an open research topic. In this light, Stochastic Kriging (SK) proposed by Ankenman et al. [42] has gained significant recognition as a general metamodelling tool for representing stochastic response surfaces (see for example [142, 143]). Defined as an extension of Kriging, SK accounts for heteroscedastic noise in the response by estimating the sample variance at each point of the ED using multiple runs of the simulator for each specific input. Consequently, one of the primary challenges often encountered in SK is the requirement of a significant number of replications at each ED point to accurately estimate the sample variance [144], which often leads to serious computational demands. For this reason, most research within SK realm has been directed towards alleviating that limitation through the optimising the ED sampling process [145–147]. Secondly, many works [148, 149] have focused on improving the computational efficiency of the approach proposed by Ankenman. Nevertheless, there are some aspects of SK that remain not fully addressed in the literature. Notably, while most applications of SK in the literature are limited to the use of constant trend models (Ordinary Kriging), the use of

more advanced models as Universal SK has remained largely unexplored. The higher complexity of UK compared to OK might suggest that it would be expected to provide more accurate estimates. However, in most instances, the opposite phenomenon holds. It has been reported by many authors that a trend misspecification may UK not to improve the accuracy in its estimates [150] or even worsen its performance [151]. This limitation can be attributed to the fact that the estimation of additional parameters within SK framework can potentially induce over-fitting problems [152]. Therefore, to harness the advantages of UK while mitigating its potential drawbacks, it is crucial to choose a minimal trend basis that effectively captures the complexity of the response surface. Motivated by these considerations, paper C discusses an optimal choice of the trend model for SK metamodelling based on the principles of PCK method presented in Section 2.3.1. In particular, the proposed metamodel involves a Universal SK model with an adaptive sparse PCE model as the trend term.

3.3.2 Kriging vs SK

For a better understanding of the conceptual differences between Kriging and SK approaches, the M/M/1 queue problem is presented below as a first case study. A M/M/1 queue is a simple yet powerful model employed to analyse the behaviour of a wide range of real-world queuing systems, from call centres to manufacturing lines [153]. The model represents a system where there is a single server and a line of entities to be served (e.g., packets, customers, etc.). In this context, entities arrive randomly at a certain rate, called *arrival rate*, and wait in a queue until the server becomes available. The waiting time is determined by the aforementioned arrival rate and by the *service rate*, which refer to the time it takes for the server to process an entity and complete its service, respectively. The example under consideration consists of the simulation of an M/M/1 queue where the service rate is fixed at one and the arrival rate ranges from 0.3 to 0.9. The response surface y represents the expected waiting time as a function of the arrival rate x. The accuracy of the simulation at each design point is determined by its time run-length T, that is, the period of time over which the behaviour of the system is observed and analysed. The M/M/1 model is one of the most studied examples in the SK literature [145, 154, 155]. It is particularly notable for its distinct and progressively increasing intrinsic variance across the parameter space. This characteristic naturally provides a formidable yet simple example to understand the potential benefits of the proposed approach. Despite the stochastic nature of the model, analytical expressions for the response surface and its intrinsic variance are known [145, 154]:

$$y(x) = \frac{1}{1-x}, \quad V(x) = \frac{2x\,(1+x)}{T\,(1-x)^4}. \tag{3.11}$$

Fig. 3.16 depicts a comparison between the deterministic PCK approach introduced in Section 2.3.1 and the LAR-SK proposed in paper C. To address this problem, 30 simulations of the M/M/1 queue with $T = 2000$ have been considered. As can be observed, their outcomes reveal a significant instability in the region $x > 0.7$. Focusing in Fig. 3.16

(a), the PCK model struggles to cope with significant amounts of noise from stochastic simulation as a consequence of the interpolating nature of Kriging metamodels discussed in Section 2.3. In contrast, a more suitable fit provided by the proposed LAR-SK model is shown in Fig. 3.16 (b). SK models filter out uncertainty by balancing interpolation and regression. Simplifying, in those regions where the intrinsic variance due to the stochastic nature of the simulation is small compared to the changes in the response surface produced by the modification of the input parameters, SK behaves like Kriging. Conversely, when the intrinsic variance is the cause of most of the changes in the response surface, SK behaves more like a regression model [151]. This phenomenon is key in explaining the flexibility and ability of SK to address a wide range of problems.

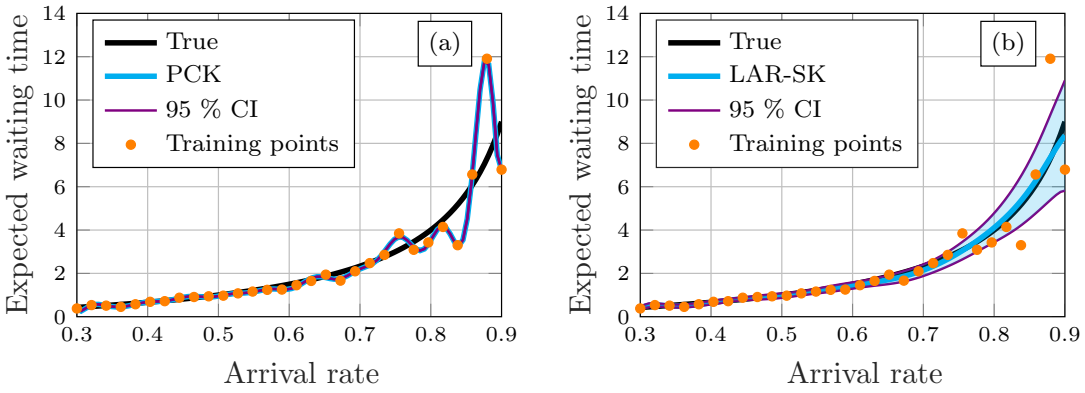

FIGURE 3.16: Surrogate modelling of M/M/1 queue comparison (runlength = 2000): PCK (a) vs LAR-SK (b).

The theoretical basis of an SK model is similar to that of Kriging. Firstly, in order to account for intrinsic uncertainty, Ankenman et al. [42] proposed to run multiple simulations at the same training point, so that the ED must be redefined as ED $= \left\{\left(\mathbf{x}^i, n_i\right)\right\}_{i=1}^{k}$, with $n_i \in \mathbb{N}$ being the number of replications at the design point $\mathbf{x}^i$. Then, analogously to Eq. 2.19, the output of the l-th simulation replication at $\mathbf{x}^i$, $i = 1, \ldots, k$ can be modelled as:

$$\mathcal{M}_l^{SK}(\mathbf{x}^i) = \mathcal{T}(\mathbf{x}^i) + \mathcal{Z}(\mathbf{x}^i, \omega) + \varepsilon_l(\mathbf{x}^i), \quad l = 1, \ldots, n_i. \tag{3.12}$$

Note that the terms $\mathcal{Z}\left(\mathbf{x}^i, \omega\right)$ and $\varepsilon_l\left(\mathbf{x}^i\right)$ in Eq. (3.12) represent the model *extrinsic uncertainty* and the *intrinsic uncertainty* for the l-th replication, respectively. It can be proved (see [42]) that the intrinsic uncertainty can be characterised by the $k \times k$ covariance matrix $\boldsymbol{\Sigma}_\varepsilon = (\boldsymbol{\Sigma}_\varepsilon)_{ij} = \mathrm{Cov}\left(\bar{\varepsilon}\left(\mathbf{x}^i\right), \bar{\varepsilon}\left(\mathbf{x}^j\right)\right)$, with $\bar{\varepsilon}\left(\mathbf{x}^i\right) = \frac{1}{n_i}\sum_{l=1}^{n_i} \varepsilon_l\left(\mathbf{x}^i\right)$ for all $\mathbf{x}^i \in$ ED.

Assuming the matrix $\boldsymbol{\Sigma}_\varepsilon$ is known, the calibration of a SK model involves maximising the likelihood function [42]:

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}; \overline{\mathcal{M}}\right) = \frac{1}{\sqrt{|\mathbf{C}|\ (2\pi)^k}} \exp\left(-\frac{1}{2}\left(\overline{\mathcal{M}} - \mathbf{F}\boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{C}^{-1}\left(\overline{\mathcal{M}} - \mathbf{F}\boldsymbol{\beta}\right)\right) \tag{3.13}$$

where $\mathbf{C} = \boldsymbol{\Sigma}_\mathcal{Z} + \boldsymbol{\Sigma}_\varepsilon$ and $\overline{\mathcal{M}}$ denotes the vector formed by the averages of all the simulations performed at each design point, that is, $\overline{\mathcal{M}} = \left[\overline{\mathcal{M}}\left(\mathbf{x}^1\right), \ldots, \overline{\mathcal{M}}\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}$, with

$\overline{\mathcal{M}}\left(\mathbf{x}^i\right) = \frac{1}{n_i}\sum_{l=1}^{n_i}\mathcal{M}_l(\mathbf{x}^i)$.

It is important to highlight that, when $\boldsymbol{\Sigma}_\varepsilon = \mathbf{0}$, the optimisation problem stated in Eq. (3.13) can be simplified to the optimisation of the hyperparameters $\boldsymbol{\theta}$ given by Eq. (2.25), and then obtaining the optimal estimates for $\boldsymbol{\beta}$ and σ through Eqs. (2.23) and (2.24). However, such a simplification is not possible within the SK framework and all parameters require to be optimised simultaneously. As a result, fitting an SK metamodel becomes significantly more challenging when the trend term is composed of a large number of functions. This fact reinforces the necessity to look for a minimal basis when implementing the SK method.

After conducting the parameter estimation following Eq. (3.13), the derivation of the SK predictor follows a similar rationale to that outlined in Section 2.3 for deterministic Kriging. Specifically, the mean and variance of the SK predictor $\widehat{\mathcal{M}}^{SK}(\mathbf{x})$ are given by [145]:

$$\mathrm{E}[\widehat{\mathcal{M}}^{SK}(\mathbf{x})] = \mathbf{f}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{c}(\mathbf{x})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) \tag{3.14}$$

$$\mathrm{Var}[\widehat{\mathcal{M}}^{SK}(\mathbf{x})] = \sigma^2 - \mathbf{c}(\mathbf{x})^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}) + \mathbf{u}_{\mathbf{c}}^{\mathrm{T}}(\mathbf{x})(\mathbf{F}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{u}_{\mathbf{c}}(\mathbf{x}) \tag{3.15}$$

where $\mathbf{c}(\mathbf{x}) = \sigma^2\mathbf{r}(\mathbf{x})$ and $\mathbf{u}_{\mathbf{c}}(\mathbf{x}) = \mathbf{F}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}) - \mathbf{f}(\mathbf{x})$.

For the above discussion, it has been assumed that the matrix $\boldsymbol{\Sigma}_\varepsilon$ is known. However, this is not often the case. In practice, $\boldsymbol{\Sigma}_\varepsilon$ should be estimated and then incorporated into Eq. (3.13). For this purpose it is common to consider that, for each $\mathbf{x}^i \in$ ED, $\varepsilon_l(\mathbf{x}^i)$, $l = 1, 2, ..., n_i$ are i.i.d. Gaussians with zero mean and variance $V(\mathbf{x}^i)$, and independent as well from $\varepsilon_h(\mathbf{x}^j)$ for all h and $j \neq i$. Since the quantities $V(\mathbf{x}^i)$ are not observable, they will be estimated by: $\hat{V}(\mathbf{x}^i) = \frac{1}{n_i - 1}\sum_{l=1}^{n_i}\left(\mathcal{M}_l\left(\mathbf{x}^i\right) - \overline{\mathcal{M}}(\mathbf{x}^i)\right)^2$ [42]. As a result, the matrix $\hat{\boldsymbol{\Sigma}}_\varepsilon$ approximating $\boldsymbol{\Sigma}_\varepsilon$ will be the diagonal matrix $\hat{\boldsymbol{\Sigma}}_\varepsilon = \mathrm{diag}\{\hat{V}(\mathbf{x}^1)/n_1, ..., \hat{V}(\mathbf{x}^k)/n_k\}$. It has been reported (see [42]) that this approach does not introduce biases in the SK prediction.

3.3.3 LAR-PCE SK

The LAR-PCE SK developed paper C is introduced in this section. For this purpose, the adaptive *Least Angle Regression* (LAR) algorithm for the determination of a sparse polynomial basis is firstly discussed. For the rest of the section, $\mathcal{A}$ will denote a preselected multi-indices set by means of Eq. (2.4).

The LAR algorithm [156] is a method for selecting the subset of input parameters that are most relevant for predicting a target variable. In the context of PCE metamodelling, LAR can be employed to identify the most influential input parameters for constructing a sparse polynomial basis. In particular, LAR provides an efficient procedure to implement the *Least Absolute Shrinkage and Selection Operator* (LASSO) method introduced by Tibshirani [157]. LASSO core insight consists on the imposition of a penalty on the total magnitude of the coefficients of a regression model, promoting sparse models where many coefficients are zero. Specifically, LASSO aims to solve the least squares problem

given by:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{\mathcal{P}}}\left(\sum_{i=1}^{k}\left(\left\|\boldsymbol{\Psi}(\mathbf{x}^i)\boldsymbol{\beta}-\mathcal{M}(\mathbf{x}^i)\right\|_2\right)+\eta\|\boldsymbol{\beta}\|_1\right),\quad \eta\in\mathbb{R},\quad \eta\geq 0. \tag{3.16}$$

Where $\mathcal{P}$ denotes the cardinality of the set of predictors $\mathcal{A}$ and η is a predetermined constant: larger values of η lead to sparser models. The computation of the solution of the optimisation problem (3.16) is a convex quadratic programming problem with linear inequality constraints, which can be solved very efficiently by means of the LAR procedure [156].

The key idea behind LAR algorithm consists of building a regression model that incorporates, in each iteration, one additional predictor selected from the polynomial basis given by $\mathcal{A}$. The result of the process will be a set of PCE metamodels of decreasing sparsity. Finally, the optimal PCE surrogate can be selected by a cross-validation scheme using Eq. (2.9). Denoting by $\mathcal{A}_j$ the set of active basis functions, i.e. the set of polynomials considered in the regression model in the j-th iteration, the LAR algorithm can be summarily described by the following steps:

1. Set the initial residual $\mathbf{r}_0$ equal to the vector of observations $\mathbf{y}$ and initialise the coefficients vector $\boldsymbol{\beta}_{\boldsymbol{\alpha}_j}=0$, $j=1,\ldots,\mathcal{P}-1$. Set all the predictors $\Psi_{\boldsymbol{\alpha}_j}$ standardised to have mean zero and unit norm.

2. Find the predictor $\Psi_{\boldsymbol{\alpha}_j}$ most correlated with the current residual $\mathbf{r}_j=\mathbf{y}-\boldsymbol{\Psi}_{\mathcal{A}_j}\boldsymbol{\beta}_{\mathcal{A}_j}$, where $\boldsymbol{\Psi}_{\mathcal{A}_j}\boldsymbol{\beta}_{\mathcal{A}_j}$ denotes the predictions of the current model on the ED points.

3. Move $\boldsymbol{\beta}_{\boldsymbol{\alpha}_j}$ from $\mathbf{0}$ towards its least-squares coefficient on $\Psi_{\boldsymbol{\alpha}_j}$, until another predictor $\Psi_{\boldsymbol{\alpha}_s}$ has as much correlation with the current residual $\mathbf{r}_j$ as $\Psi_{\boldsymbol{\alpha}_j}$.

4. Move $\{\boldsymbol{\beta}_{\boldsymbol{\alpha}_j},\boldsymbol{\beta}_{\boldsymbol{\alpha}_s}\}$ in the direction defined by their joint least squares coefficient of the current residual on $\{\Psi_{\boldsymbol{\alpha}_j},\Psi_{\boldsymbol{\alpha}_s}\}$, until another predictor $\Psi_{\boldsymbol{\alpha}_t}$ has the same correlation with the current residual.

5. If a non zero coefficient hits zero, discard it from the current metamodel and recompute the current joint least-square direction.

6. Continue the process until m= min $(\mathcal{P},k-1)$ predictors have been added.

Step 3 and 4 involve 'moving' the active coefficients in the direction of their least squares value. It corresponds to an updating of model coefficients of the form $\boldsymbol{\beta}_{\mathcal{A}_{j+1}}=\boldsymbol{\beta}_{\mathcal{A}_j}+\zeta_j\mathbf{u}_{\mathcal{A}_j}$, where $\boldsymbol{\beta}_{\mathcal{A}_j}$ is the vector of coefficients for the set of the active basis functions $\mathcal{A}_j$ at the j-th iteration and ζ_j and $\mathbf{u}_{\mathcal{A}_j}$ are referred to as the LAR *step* and *descent direction*, respectively. Both quantities can be derived algebraically, as discussed in [85]. It can be proved (see [85]) than the procedure described above provides the complete trajectories of LASSO solution coefficients at once as the tuning parameter η in Eq. (3.16) is increased from 0 up to a maximum value.

In light of the above discussion, it is possible to define the LAR-PCE SK predictor as a particular case of the Eqs. (3.14) and (3.15). Specifically, the the mean and variance of the proposed optimal LAR-PCE based SK predictor reads:

$$\mathrm{E}[\widehat{\mathcal{M}}^{PCE-SK}(\mathbf{x})] = \boldsymbol{\Psi}(\mathbf{x})\boldsymbol{\beta} + \mathbf{c}(\mathbf{x})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta}) \tag{3.17}$$

$$\mathrm{Var}[\widehat{\mathcal{M}}^{PCE-SK}(\mathbf{x})] = \sigma^2 - \mathbf{c}(\mathbf{x})^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}) + \mathbf{u}_{\mathbf{c}}^{\mathrm{T}}(\mathbf{x})(\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Psi})^{-1}\mathbf{u}_{\mathbf{c}}(\mathbf{x}) \tag{3.18}$$

where $\mathbf{c}(\mathbf{x})$ and $\mathbf{u}_{\mathbf{c}}(\mathbf{x})$.

3.3.4 Numerical examples

Two numerical examples are presented below in order to illustrate the capabilities of the proposed LAR-PCE SK approach. The assessment of LAR-PCE SK includes a detailed comparison against ordinary SK considering different EDs. The purpose of defining different sampling scenarios is to demonstrate that the proposed LAR-PCE SK method consistently outperforms ordinary SK, regardless of whether a sparse and deep ED (with few points but high accuracy) or a denser but shallower ED is chosen. In all the analyses, the model performance is assessed by means of two error metrics, namely the Empirical Root Mean Squared Error (ERMSE) and the Normalised Maximum Absolute Error (NMAE).

The first case study consists of the simulation of an M/M/1 queue described in Section 3.3.2. For this analysis, the intrinsic variance is assumed as known following Eq. (3.11), and therefore just one single simulation is run for each design point. That is, $\overline{\mathcal{M}}$ is directly taken as $\left(\mathcal{M}(\mathbf{x}^1), ..., \mathcal{M}\left(\mathbf{x}^k\right)\right)$ and $\boldsymbol{\Sigma}_\varepsilon = \mathrm{diag}\{V(\mathbf{x}^1), ..., V(\mathbf{x}^k)\}$. Three different scenarios with varying number of design points and run-lengths but with constant computational resources have been considered. The first one consists of 10 equispaced design points with a run-length of 6000; the second one includes of 30 design points with a run-length of 2000; and the third scenario encompasses 50 design points with a run-length of 1200. In the case of LAR-PCE SK, following the recommendations provided in reference [86], the maximum degree of the polynomials was set at 5, 10 and 16 for the scenarios one to three, respectively.

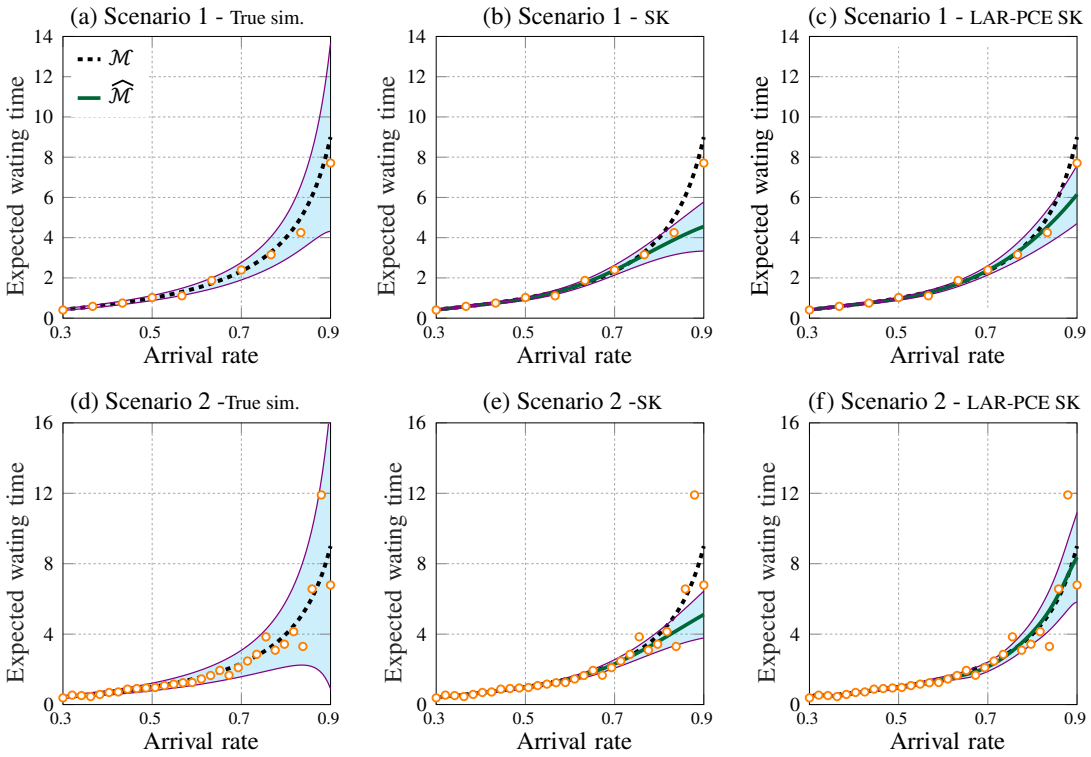

FIGURE 3.17: M/M/1 example. Comparison between the true simulator (a,d) against the predictions by the proposed LAR-PCE SK (c,f) and Ordinary SK (b,e) under the assumption of known variance in Scenarios 1 (a,b,c) and 2 (d,e,f). Scatter points denote the ED samples, and light blue shaded areas represent the 95% confidence interval.

Figs. 3.17 (a,d) depict the true response surface described by Eq. (3.11) along with the analytically computed 95% confidence intervals represented as $\left(y(x) \pm 1.96\sqrt{V(x)}\right)$ for Scenarios 1 and 2, respectively. Similarly, Figs. 3.17 (b,e) and (c,f) illustrate the estimates produced by Ordinary SK and LAR-PCE SK, respectively. It can be observed that for low values of x, where the intrinsic uncertainty is small, both meta-models present a similar behaviour. In contrast, for high values of x where both the QoI and the intrinsic variance experience a rapid growth, ordinary SK fails to capture the variation in the response, whereas LAR-PCE SK exhibits a significantly better performance. This is to be expected since, as reported in [151], Ordinary SK virtually ignores those simulation outputs whose estimated intrinsic variance are large compared to the estimated extrinsic variance. With the aim of assessing the consistency of the proposed LAR-PCE SK approach, the results of the fitting of one hundred independent experiments are reported in Fig. 3.18. For this purpose, the metamodel performance is assessed by means of two error metrics which evaluates respectively the global quality of the meta-model and the maximum the local error. The first one, referred to as ERMSE, is defined as $\text{ERMSE} = \sqrt{\sum_{i=1}^{K}\left(\widehat{\mathcal{M}}(\mathbf{x}^i) - \mathcal{M}(\mathbf{x}^i)\right)^2}$. Note that this metric coincides with the numerator of NRMSE described in Table 2.1.

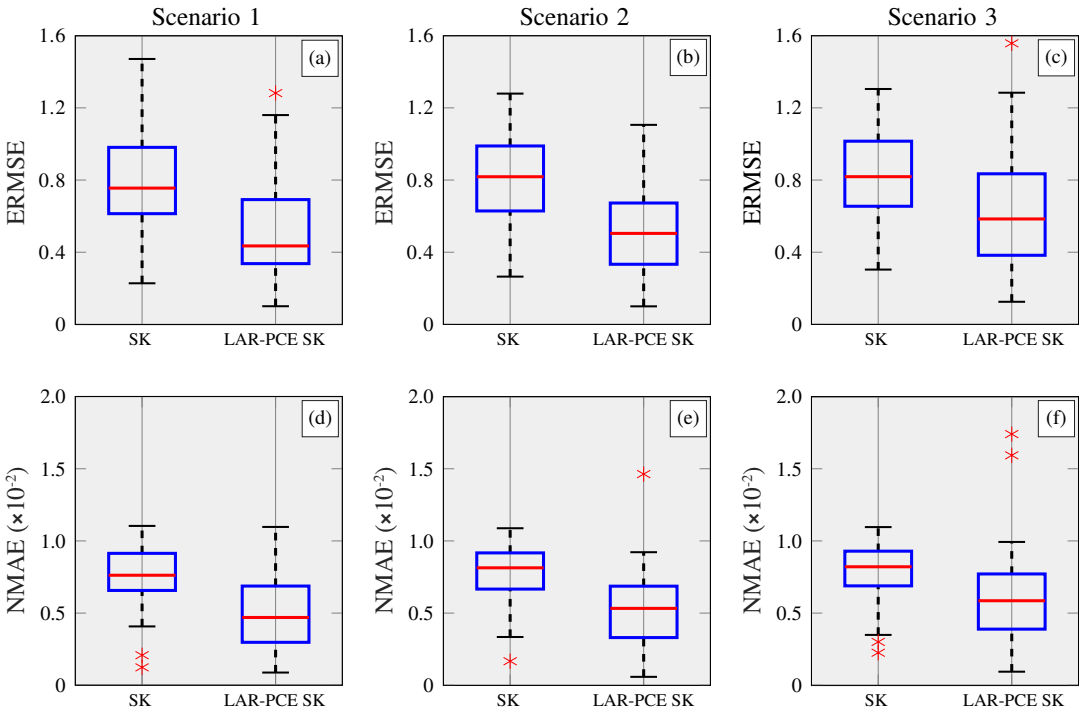

FIGURE 3.18: Performance analysis of Ordinary SK and LAR-PCE SK in Case Study I in terms of ERMSE and NMAE error metrics using 100 known intrinsic noise M/M/1 experiments and sampling Scenarios 1 to 3.

It is clear that significant enhancements have been achieved for both ERMSE and NMAE metrics. These results demonstrate that introducing a proper trend model effectively mitigates some of the drawbacks of ordinary SK. Nevertheless, it should be noted that, with respect to variance estimation, both approaches similarly fail to capture the true variability, as can be observed in Fig. 3.17. This issue has been recognised and documented in the SK literature [148, 152]. Although there are methods available to address this problem (refer e.g. to [152]), they are beyond the scope of the present study, in which the primary focus is on the importance of properly selecting the trend model in SK.

The second case study examines the Ishigami function, a classical benchmark function commonly investigated in literature [41]. This function is characterised by its non-linear and non-monotonic behaviour, making it a challenging test for meta-modelling techniques. Specifically, the Ishigami function takes as input a three-dimensional vector $\mathbf{x} = [x_1, x_2, x_3]^t$ with each component x_i ranging from $-\pi$ to π, and outputs a scalar value $\mathcal{M}(\mathbf{x})$ given by the non-linear sum of three sinusoidal functions. For the purpose of this work, a stochastic version of the Ishigami function is defined by incorporating a normally distributed random noise term $\varepsilon_l(\mathbf{x})$ with zero mean and standard deviation proportional to the absolute value of the function. This results in the following stochastic Ishigami function:

$$\begin{aligned} \mathcal{M}_l(x_1, x_2, x_3) &= \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1) + \varepsilon_l(x_1, x_2, x_3), \\ \varepsilon_l &\sim \mathcal{N}\left(0, \sqrt{\left|\sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1)\right|}\right), \; x_1, x_2, x_3 \in [-\pi, \pi]. \end{aligned} \tag{3.19}$$

A similar analysis to that conducted for the simulation of the M/M/1 queuing model has been performed for the stochastic Ishigami function described in Eq. (3.19). Additionally, with the purpose of stressing the importance of selecting an appropriate sparse basis of polynomials in the trend term, a full PCE-SK surrogate without LAR has been also included in the comparison. This model is defined by the inclusion of all polynomials specified by the hyperbolic truncation set defined in Eq. (2.4) up to a certain degree. In contrast to the M/M/1 case study, in this analysis the intrinsic variance $\mathbf{\Sigma}_\varepsilon$ is supposed to be unknown and it is estimated through a series of replications at each design point. In particular, a total computational budget C of 2560 replications to be distributed along the training sites has been considered. Consistently with reference [155], the number of replications assigned to each design point x_i is determined in a single step as proportional to the model variance following $n_i = \frac{\sqrt{V(\mathbf{x}_i)}}{\sum_{i=1}^{k}\sqrt{V(\mathbf{x}_i)}}C$. This procedure for distributing the computational budget has been selected for the sake of simplicity. However, alternative methods which employ a multi-step approach for this purpose such as those described in reference [145] may be considered for bypassing this issue.

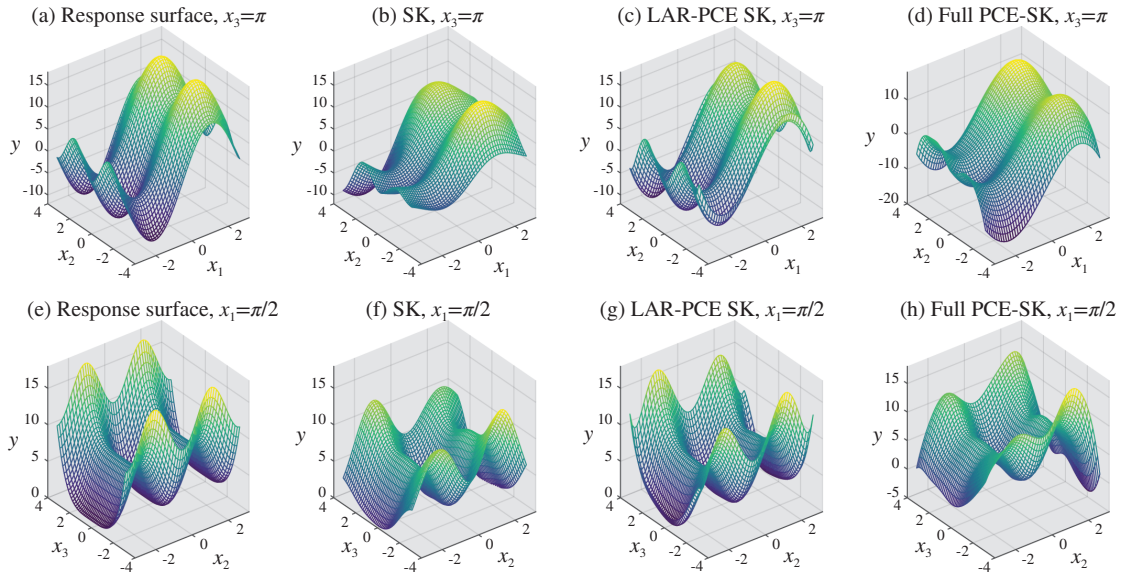

FIGURE 3.19: Response surfaces of the Ishigami function (a,e), and predictions by Ordinary SK (b,f), LAR-PCE SK (c,g) and full PCE-SK (d,h) metamodels with $k = 64$ design points. The top and bottom row panels denote the response surfaces obtained for $x_3 = \pi$ and $x_1 = \pi/2$, respectively.

Analogously to the M/M/1 example, three distinct scenarios with $k = 64$, $k = 128$ and $k = 256$ sampling locations determined by LHS method have been considered. With respect to the degree of the polynomial basis to be considered, the maximum degree of the polynomial basis for the full PCE-SK case has been set at 5, 8, and 10 for Scenarios 1, 2, and 3, correspondingly. On the other hand, degrees of 9, 11, and 12 have been considered in Scenarios 1 to 3, respectively. It is worth noting that, due to the sparse nature of the basis selected by the LAR algorithm, it is possible to consider polynomials of slightly higher degree for LAR-PCE SK. Finally, a value of 0.8 is specified for the q-norm for all the considered scenarios. The outcome of an arbitrary replication of Scenario 1 is depicted in Fig. 3.19. In order to assess the consistency of outperformance of LAR-SK

model, the results of the comparison between the different approaches discussed across a macro-replication of one hundred independent experiments are presented in Fig. 3.20. In view of Fig. 3.20, it can be stated that the presented LAR-PCE approach is able to approximate complex surfaces more accurately than the classical SK metamodel proposed in Ref. [42]. Secondly, it is noteworthy that blindly setting an inadequate polynomial basis as the trend model resulted in notable errors and instability in the results across all the scenarios, highlighting the critical importance of a proper basis selection.

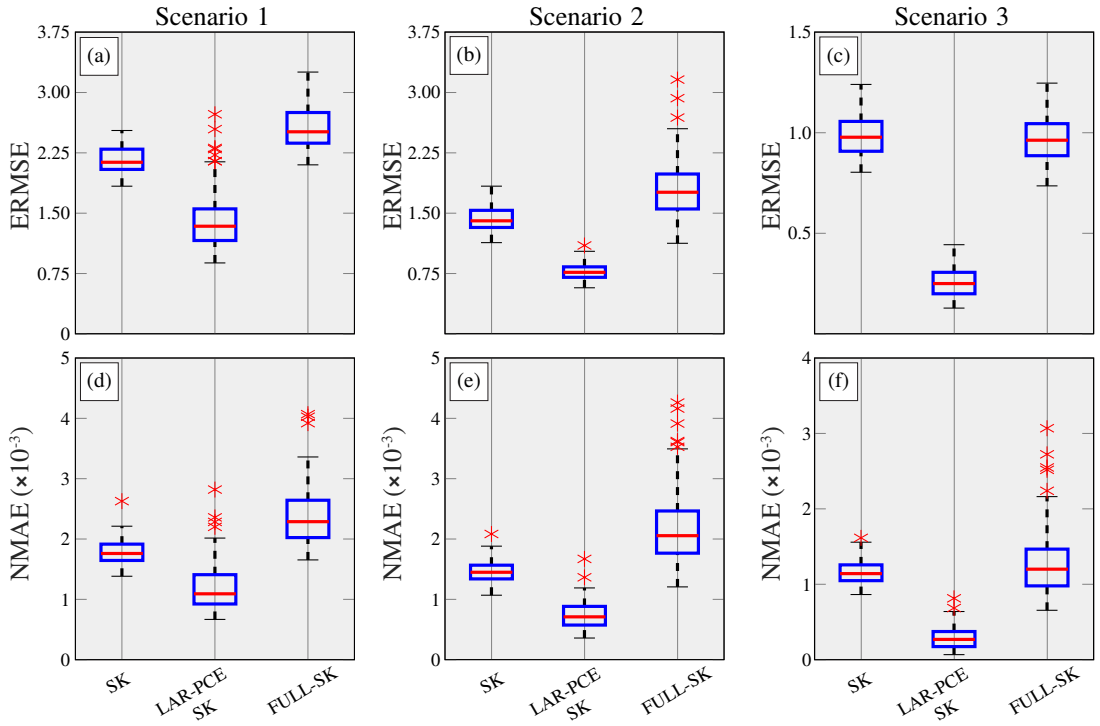

FIGURE 3.20: Performance analysis of Ordinary SK, LAR-PCE SK and full PCE SK in terms of ERMS and NMAE error metrics using 100 experiments of the Ishigami function and sampling Scenarios 1 to 3.

Chapter 4

Conclusions and future developments

Accurately modelling the behaviour of real-world systems poses a challenging task, often requiring extraordinary computational resources. In this light, the key achievement of this thesis has been the development of a suitable general framework for conducting efficiently UQ analyses in a wide range of computationally intensive engineering problems. To this end, the effectiveness of the metamodel-based approach has been demonstrated in both forward UP contexts and Bayesian analysis for inverse problems. Furthermore, the development of surrogate models capable of approximating stochastic response surfaces opens up vast possibilities for future research.

Overall, the main contributions of this thesis have are summarised in the following points:

- A novel application of a PCE surrogate model to a real engineering problem. In particular, an UP analysis of a particle-reinforced composite with volume fraction up to 50%. The presented results have showed their effectiveness to conduct fast uncertainty propagation analysis of epoxy composites doped with glass particles but suggest the ability of the proposed methodology to be adapted to a large variety of composite materials.

- The development of a variation of the PCK technique to deal with ill-conditioned and highly nonlinear problems. In particular, the ME-PCK based on Sobol indices provides computational time savings of several orders of magnitude in both surrogate model construction and evaluation. This computational efficiency facilitates the agile execution of MCMC algorithms for Bayesian parameter inference, being especially relevant in multimodal problems requiring long chains to converge to the posterior distribution. The capabilities of the proposed approach have been applied to a challenging engineering problem, the characterisation of hydrogen flow curves as a function of temperature resulting from TDS experiments.

- Fill a literature gap within the realm of stochastic surrogate models. The development and application of surrogate models to stochastic simulators is a topic with great potential in the field of uncertainty propagation. The results presented in

section 3.3 demonstrate that the inclusion of suitable trend models has the potential to greatly increase the accuracy of SK metamodels without compromising its computational efficiency. For this purpose, the LAR algorithm is crucial to avoid overfitting problems.

- The recent advancements mentioned above have resulted in numerous publications in high-impact journals, which can be found as appendices of this report.

Despite the aforementioned achievements, further exploration on surrogate models and their applications to Bayesian inference is necessary. A future research field involves exploring the potential of the evidence term in Equation 2.30 for Bayesian model selection. The ratio of evidences from two competing statistical models, known as the Bayes factor, can be utilised to quantify the support for one model over the other. However, in problems of moderate or large dimensionality, numerical integration becomes extremely expensive for the computation of model evidence. In this line, the Nested Sampling (NS) algorithm is a method for approximating multidimensional integrals with allow to address the problem. NS allows for the estimation of Bayesian evidence by transforming the multi-dimensional evidence integral into an easy one-dimensional problem which can be solved by simple methods such as the trapezium rule. Although the algorithm is widely used in cosmology and particle physics [97], its application in other fields is still relatively unexplored. In this line, promising developments are expected within the realm of Structural Health Monitoring (SHM). For instance, once the presence of damage has been detected in a structure, a common approach consists of proposing a FE model and finding the optimal parameters of the model. Instead, it is more effective to consider a family of competing models representing the different damage mechanisms that a structure may experience. Thus, the challenge of identifying the model that most accurately represents the effects induced by damage naturally arises. The NS algorithm would thus provide an efficient tool for tackling this problem. It is important to note that surrogate models are crucial in the described approach: NS requires to evaluate thousands of times the likelihood of each of the proposed models in order to compute the model evidence. Such a process is infeasible to carry out directly due to the enormous computational burden of solving a FE model.

Another gap in the literature that is expected to be filled in the near future is the appropriate selection of user-selected kriging parameters. Given that in engineering practice the majority of metamodel development time is dedicated to evaluating the forward model, ensuring correct design is key for metamodel performance. Many studies comparing metamodelling techniques have been carried out in the last two decades. Nevertheless, existing research on how to properly fit a Kriging model is scarce. As explained in Section 2.3, in order to develop a Kriging surrogate, users must specify a sampling procedure, a trend, an autocorrelation function, and an algorithm for estimating the model's hyperparameters. Any misspecification can result in a substantial decrease in accuracy, implying the need for a larger sample and, consequently, more computational resources. To address this challenge, a comprehensive benchmark of analytical functions collected

from the metamodelling literature has been constructed. The referred benchmark encompasses linear and non-linear problems spanning a large range of dimensionalities. Preliminary results suggest that Kriging models are very sensitive to the choice of correlation model, with Matérn functions providing the most robust results when the response surface is unknown. In a similar vein, the adoption of an appropriate sampling procedure is also of major importance. It was observed that for small ED sizes LHS produces better surrogates. However, as the sample size increases, MIPT sampling results in more robust and accurate fits. A possible explanation for this phenomenon is that for small EDs, MIPT method locates a large number of samples close to the boundaries of the parameter space, resulting in less accurate metamodels. This behaviour would be more intense in high dimensional problems than in low dimensional ones.

Bibliography

[1] P. Honarmandi, R. Arróyave, Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design, Integrating Materials and Manufacturing Innovation 9 (2020) 103–143.

[2] J. Mikelson, M. Khammash, Likelihood-free nested sampling for parameter inference of biochemical reaction networks, PLoS computational biology 16 (10) (2020) e1008264.

[3] L. Tesfatsion, Agent-Based Computational Economics: Overview and Brief History, Springer International Publishing, 2023, Ch. 4, pp. 41–58.

[4] L. J. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, Infectious Disease Modelling 2 (2) (2017) 128–142.

[5] M. Marino, B. Hudobivnik, P. Wriggers, Computational homogenization of polycrystalline materials with the Virtual Element Method, Computer Methods in Applied Mechanics and Engineering 355 (2019) 349–372.

[6] E. García-Macías, F. Ubertini, Least Angle Regression for early-stage identification of earthquake-induced damage in a monumental masonry palace: Palazzo dei Consoli, Engineering Structures 259 (2022) 114119.

[7] R. C. Smith, Uncertainty quantification: theory, implementation, and applications, Vol. 12, Siam, 2013.

[8] B. Sudret, S. Marelli, J. Wiart, Surrogate models for uncertainty quantification: An overview, in: 2017 11th European conference on antennas and propagation (EUCAP), IEEE, 2017, pp. 793–797.

[9] R. C. Aster, B. Borchers, C. H. Thurber, Parameter estimation and inverse problems, Elsevier, 2018.

[10] W. Dörfler, M. Hochbruck, J. Köhler, A. Rieder, R. Schnaubelt, C. Wieners, What Is an Inverse and Ill-Posed Problem?, Springer International Publishing, 2023.

[11] E. García-Macías, F. Ubertini, Real-time Bayesian damage identification enabled by sparse PCE-Kriging meta-modelling for continuous SHM of large-scale civil engineering structures, Journal of Building Engineering 59 (2022) 105004.

[12] R. Rodríguez-Romero, V. Compán, A. Sáez, E. García-Macías, Hierarchical metamodelling for fast prediction of the elastic properties of stone injected with CNT/cement mortar, Construction and Building Materials 408 (2023) 133725.

[13] K. M. Hamdia, M. Silani, X. Zhuang, P. He, T. Rabczuk, Stochastic analysis of the fracture toughness of polymeric nanoparticle composites using polynomial chaos expansions, International Journal of Fracture 206 (2017) 215–227.

[14] G. Grimmett, D. Stirzaker, Probability and random processes, Oxford university press, 2020.

[15] H. M. Kroetz, R. K. Tessari, A. T. Beck, Performance of global metamodeling techniques in solution of structural reliability problems, Advances in Engineering Software 114 (2017) 394–404.

[16] T. Østergård, R. L. Jensen, S. E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Applied Energy 211 (2018) 89–103.

[17] R. Alizadeh, J. K. Allen, F. Mistree, Managing computational complexity using surrogate models: a critical review, Research in Engineering Design 31 (2020) 275–298.

[18] M. R. Kianifar, F. Campean, Performance evaluation of metamodelling methods for engineering problems: towards a practitioner guide, Structural and Multidisciplinary Optimization 61 (1) (2020) 159–186.

[19] J. V. S. do Amaral, J. A. B. Montevechi, R. de Carvalho Miranda, W. T. de Sousa Junior, Metamodel-based simulation optimization: A systematic literature review, Simulation Modelling Practice and Theory 114 (2022) 102403.

[20] M. J. Asher, B. F. W. Croke, A. J. Jakeman, L. J. M. Peeters, A review of surrogate models and their application to groundwater modeling, Water Resour. Res. 51 (8) (2015) 5957–5973.

[21] B. M. de Gooijer, J. Havinga, H. J. M. Geijselaers, A. H. van den Boogaard, Evaluation of POD based surrogate models of fields resulting from nonlinear FEM simulations, Adv. Model. Simul. Eng. Sci. 8 (1) (2021) 1–33.

[22] L. Zhang, Y. Wu, P. Jiang, S. K. Choi, Q. Zhou, A multi-fidelity surrogate modeling approach for incorporating multiple non-hierarchical low-fidelity data, Adv. Eng. Inform. 51 (2022) 101430.

[23] Y. Shi, Z. Lu, L. Xu, S. Chen, An adaptive multiple-Kriging-surrogate method for time-dependent reliability analysis, Appl. Math. Model. 70 (2019) 545–571.

[24] N. Lüthen, S. Marelli, B. Sudret, Sparse polynomial chaos expansions: Literature survey and benchmark, SIAM/ASA Journal on Uncertainty Quantification 9 (2) (2021) 593–649.

[25] C. Wang, A. Wang, S. Chen, G. Zhang, B. Zhu, Optimal operation of microgrids based on a radial basis function metamodel, IEEE Systems Journal 16 (3) (2022) 4756–4767.

[26] A. Roy, S. Chakraborty, Seismic Reliability Analysis of Structures by an Adaptive Support Vector Regression-Based Metamodel, Journal of Earthquake Engineering (2023) 1–25.

[27] F. Dunke, S. Nickel, Neural networks for the metamodeling of simulation models with online decision making, Simulation Modelling Practice and Theory 99 (2020) 102016.

[28] R. Jin, W. Chen, T. W. Simpson, Comparative studies of metamodelling techniques under multiple modelling criteria, Structural and multidisciplinary optimization 23 (2001) 1–13.

[29] R. Jin, X. Du, W. Chen, The use of metamodeling techniques for optimization under uncertainty, Structural and Multidisciplinary Optimization 25 (2003) 99–116.

[30] E. N. Ben-Ari, D. M. Steinberg, Modeling data from computer experiments: an empirical comparison of kriging with MARS and projection pursuit regression, Quality Engineering 19 (4) (2007) 327–338.

[31] B.-S. Kim, Y.-B. Lee, D.-H. Choi, Comparison study on the accuracy of metamodeling technique for non-convex functions, Journal of Mechanical Science and Technology 23 (2009) 1175–1181.

[32] L. Van Gelder, P. Das, H. Janssen, S. Roels, Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners, Simulation Modelling Practice and Theory 49 (2014) 245–257.

[33] Y.-F. Li, S. H. Ng, M. Xie, T. Goh, A systematic comparison of metamodeling techniques for simulation optimization in decision support systems, Applied soft computing 10 (4) (2010) 1257–1273.

[34] S. Marelli, N. Lüthen, B. Sudret, UQLab user manual – Polynomial chaos expansions, Tech. rep., Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, report UQLab-V2.0-104 (2022).

[35] S. Dey, T. Mukhopadhyay, H. H. Khodaparast, S. Adhikari, Fuzzy uncertainty propagation in composites using Gram–Schmidt polynomial chaos expansion, Applied Mathematical Modelling 40 (7-8) (2016) 4412–4428.

[36] X. Peng, D. Li, H. Wu, Z. Liu, J. Li, S. Jiang, J. Tan, Uncertainty analysis of composite laminated plate with data-driven polynomial chaos expansion method under insufficient input data of uncertain parameters, Composite Structures 209 (2019) 625–633.

[37] P. Ni, Y. Xia, J. Li, H. Hao, Using polynomial chaos expansion for uncertainty and sensitivity analysis of bridge structures, Mechanical Systems and Signal Processing 119 (2019) 293–311.

[38] M. A. Hariri-Ardebili, B. Sudret, Polynomial chaos expansion for uncertainty quantification of dam engineering problems, Engineering Structures 203 (2020) 109631.

[39] L. YiFei, C. MaoSen, H. Tran-Ngoc, S. Khatir, M. A. Wahab, Multi-parameter identification of concrete dam using polynomial chaos expansion and slime mould algorithm, Computers & Structures 281 (2023) 107018.

[40] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability engineering & system safety 93 (7) (2008) 964–979.

[41] R. Schobi, B. Sudret, J. Wiart, Polynomial-chaos-based Kriging, International Journal for Uncertainty Quantification 5 (2).

[42] B. Ankenman, B. L. Nelson, J. Staum, Stochastic kriging for simulation metamodeling, in: 2008 Winter simulation conference, IEEE, 2008, pp. 362–370.

[43] C. P. Robert, et al., The Bayesian choice: from decision-theoretic foundations to computational implementation, Vol. 2, Springer, 2007.

[44] F. J. Samaniego, A comparison of the Bayesian and frequentist approaches to estimation, Vol. 24, Springer, 2010.

[45] A. Gelman, C. R. Shalizi, Philosophy and the practice of Bayesian statistics, British Journal of Mathematical and Statistical Psychology 66 (1) (2013) 8–38.

[46] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, et al., Bayesian statistics and modelling, Nature Reviews Methods Primers 1 (1) (2021) 1.

[47] A. E. Gelfand, A. F. Smith, Sampling-based approaches to calculating marginal densities, Journal of the American statistical association (1990) 398–409.

[48] M. Van Oijen, Bayesian compendium, Springer, 2020.

[49] A. Barbu, S.-C. Zhu, A. Barbu, S.-C. Zhu, Convergence Analysis of MCMC, Monte Carlo Methods (2020) 189–209.

[50] D. B. Hitchcock, A history of the Metropolis-Hastings algorithm, American Statistician (2003) 254–257.

[51] Y. Che, X. Wu, G. Pastore, W. Li, K. Shirvan, Application of Kriging and Variational Bayesian Monte Carlo method for improved prediction of doped UO2 fission gas release, Ann. Nucl. Energy. 153 (2021) 108046.

[52] S. Brooks, A. Gelman, G. Jones, X.-L. Meng, Handbook of markov chain monte carlo, CRC press, 2011.

[53] A. Endo, E. Van Leeuwen, M. Baguelin, Introduction to particle Markov-chain Monte Carlo for disease dynamics modellers, Epidemics 29 (2019) 100363.

[54] H. Haario, E. Saksman, J. Tamminen, An adaptive Metropolis algorithm, Bernoulli (2001) 223–242.

[55] Y. Huang, C. Shao, B. Wu, J. L. Beck, H. Li, State-of-the-art review on Bayesian inference in structural system identification and damage assessment, Advances in Structural Engineering 22 (6) (2019) 1329–1351.

[56] M. Amadi, A. Shcherbacheva, H. Haario, Agent-based modelling of complex factors impacting malaria prevalence, Malaria journal 20 (1) (2021) 1–15.

[57] D. H. Nguyen, S.-H. Kim, H.-H. Kwon, D.-H. Bae, Uncertainty Quantification of Water Level Predictions from Radar-based Areal Rainfall Using an Adaptive MCMC Algorithm, Water Resources Management 35 (7) (2021) 2197–2213.

[58] A. Guerini, A. Marziali, G. De Nicolao, MCMC calibration of spot-prices models in electricity markets, Applied Stochastic Models in Business and Industry 36 (1) (2020) 62–76.

[59] A. H. Elsheikh, I. Hoteit, M. F. Wheeler, Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates, Computer Methods in Applied Mechanics and Engineering 269 (2014) 515–537.

[60] F. Schneider, I. Papaioannou, D. Straub, C. Winter, G. Müller, Bayesian parameter updating in linear structural dynamics with frequency transformed data using rational surrogate models, Mech. Syst. Signal Process. 166 (2022) 108407.

[61] I. Fursov, M. Christie, G. Lord, Applying kriging proxies for Markov chain Monte Carlo in reservoir simulation, Computational Geosciences 24 (2020) 1725–1746.

[62] L. Ierimonti, N. Cavalagli, I. Venanzi, E. García-Macías, F. Ubertini, A transfer Bayesian learning methodology for structural health monitoring of monumental structures, Eng. Struct. 247 (2021) 113089.

[63] A. del Val, O. P. Le Maître, T. E. Magin, O. Chazot, P. M. Congedo, A surrogate-based optimal likelihood function for the Bayesian calibration of catalytic recombination in atmospheric entry protection materials, Appl. Math. Model. 101 (2022) 791–810.

[64] H. Haario, M. Laine, A. Mira, E. Saksman, DRAM: efficient adaptive MCMC, Statistics and computing 16 (2006) 339–354.

[65] A. Forrester, A. Sobester, A. Keane, Engineering design via surrogate modelling: a practical guide, John Wiley & Sons, 2008.

[66] R. Yondo, E. Andrés, E. Valero, A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses, Progress in aerospace sciences 96 (2018) 23–61.

[67] S. E. Davis, S. Cremaschi, M. R. Eden, Efficient surrogate model development: impact of sample size and underlying model dimensions, in: Computer Aided Chemical Engineering, Vol. 44, Elsevier, 2018, pp. 979–984.

[68] S. Lophaven, H. Nielsen, J. Søndergaard, Aspects of the Matlab toolbox DACE, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2002.

[69] J. Luo, X. Ma, Y. Ji, X. Li, Z. Song, W. Lu, Review of machine learning-based surrogate models of groundwater contaminant modeling, Environmental Research (2023) 117268.

[70] G. Blatman, B. Sudret, M. Berveiller, Quasi random numbers in stochastic finite element analysis, Mechanics & Industry 8 (3) (2007) 289–297.

[71] R. O. Hussain, H. H. Hussain, Natural Occurring Radionuclide Materials, in: Radioisotopes-Applications in Physical Sciences, IntechOpen, 2011, pp. 3–20.

[72] J. Eason, S. Cremaschi, Adaptive sequential sampling for surrogate model generation with artificial neural networks, Computers & Chemical Engineering 68 (2014) 220–232.

[73] K. Crombecq, E. Laermans, T. Dhaene, Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling, European Journal of Operational Research 214 (3) (2011) 683–696.

[74] A. Savine, Modern computational finance: AAD and parallel simulations, John Wiley & Sons, 2018.

[75] M. Baudin, Low Discrepancy Toolbox Manual, Tech. rep., Électricité de France (2013).

[76] H. Chi, M. Mascagni, T. Warnock, On the optimal Halton sequence, Mathematics and computers in simulation 70 (1) (2005) 9–21.

[77] L. Kocis, W. J. Whiten, Computational investigations of low-discrepancy sequences, ACM Transactions on Mathematical Software (TOMS) 23 (2) (1997) 266–294.

[78] A. Navid, S. Khalilarya, M. Abbasi, Diesel engine optimization with multi-objective performance characteristics by non-evolutionary Nelder-Mead algorithm: Sobol sequence and Latin hypercube sampling methods comparison in DoE process, Fuel 228 (2018) 349–367.

[79] H. Liu, Y. S. Ong, J. Cai, A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design, Structural and Multidisciplinary Optimization 57 (2018) 393–416.

[80] J. N. Fuhg, A. Fau, U. Nackenhorst, State-of-the-art and comparative review of adaptive sampling methods for kriging, Archives of Computational Methods in Engineering 28 (2021) 2689–2747.

[81] H. Liu, J. Cai, Y. S. Ong, An adaptive sampling approach for Kriging metamodeling by maximizing expected prediction error, Computers & Chemical Engineering 106 (2017) 171–182.

[82] Y. Pang, Y. Wang, X. Lai, S. Zhang, P. Liang, X. Song, Enhanced Kriging leave-one-out cross-validation in improving model estimation and optimization, Computer Methods in Applied Mechanics and Engineering 414 (2023) 116194.

[83] H. Liang, M. Zhu, Z. Wu, Using cross-validation to design trend function in Kriging surrogate modeling, AIAA Journal 52 (10) (2014) 2313–2327.

[84] M. Moustapha, J. M. Bourinet, B. Guillaume, B. Sudret, Comparative study of Kriging and support vector regression for structural engineering applications, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 4 (2) (2018) 04018005.

[85] G. Blatman, Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis, Ph.D. thesis, Clermont-Ferrand 2 (2009).

[86] B. Sudret, Polynomial chaos expansions and stochastic finite element methods, in: J. C. Kok-Kwang Phoon (Ed.), Risk and Reliability in Geotechnical Engineering, CRC Press, 2015, pp. 265–300.

[87] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM journal on scientific computing 24 (2) (2002) 619–644.

[88] N. Fajraoui, S. Marelli, B. Sudret, Sequential design of experiment for sparse polynomial chaos expansions, SIAM/ASA Journal on Uncertainty Quantification 5 (1) (2017) 1061–1085.

[89] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, Journal of Computational Physics 230 (6) (2011) 2345–2367.

[90] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global sensitivity analysis: the primer, John Wiley & Sons, 2008.

[91] E. Vanmarcke, Random fields: analysis and synthesis, World Scientific, 2010.

[92] H. Biermé, Introduction to random fields and scale invariance, Stochastic Geometry: Modern Research Frontiers (2019) 129–180.

[93] C. Lataniotis, D. Wicaksono, S. Marelli, B. Sudret, UQLab user manual – Kriging (Gaussian process modeling), Tech. rep., Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, report UQLab-V2.0-105 (2022).

[94] A. Marrel, B. Iooss, F. Van Dorpe, E. Volkova, An efficient methodology for modeling complex computer codes with Gaussian processes, Computational Statistics & Data Analysis 52 (10) (2008) 4731–4744.

[95] F. Bachoc, Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, Computational Statistics & Data Analysis 66 (2013) 55–69.

[96] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis, Chapman and Hall/CRC, 1995.

[97] G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Csányi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, et al., Nested sampling for physical scientists, Nature Reviews Methods Primers 2 (1) (2022) 39.

[98] G. Casella, R. Berger, Statistical Inference, Duxbury Resource Center, 2001.

[99] R. J. Hyndman, Computing and graphing highest density regions, The American Statistician 50 (2) (1996) 120–126.

[100] W. R. Gilks, S. Richardson, D. Spiegelhalter, Markov chain Monte Carlo in practice, CRC press, 1995.

[101] C. P. Robert, G. Casella, Monte Carlo statistical methods, Vol. 2, Springer, 1999.

[102] C. J. Geyer, Introduction to Markov Chain Monte Carlo, in: Handbook of Markov Chain Monte Carlo, CRC Press, 2011, pp. 3–48.

[103] S. D. Hill, J. C. Spall, Stationarity and convergence of the metropolis-hastings algorithm: Insights into theoretical aspects, IEEE Control Systems Magazine 39 (1) (2019) 56–67.

[104] S. P. Meyn, R. L. Tweedie, Markov chains and stochastic stability, Springer Science & Business Media, 2012.

[105] A. Gelman, W. R. Gilks, G. O. Roberts, Weak convergence and optimal scaling of random walk Metropolis algorithms, The annals of applied probability 7 (1) (1997) 110–120.

[106] G. O. Roberts, J. S. Rosenthal, Optimal scaling for various Metropolis-Hastings algorithms, Statistical science 16 (4) (2001) 351–367.

[107] D. Van Ravenzwaaij, P. Cassey, S. D. Brown, A simple introduction to Markov Chain Monte–Carlo sampling, Psychonomic bulletin & review 25 (1) (2018) 143–154.

[108] C. Andrieu, J. Thoms, A tutorial on adaptive MCMC, Statistics and computing 18 (2008) 343–373.

[109] M. A. J. Mazumder, H. Sheardown, A. Al-Ahmed, Functional polymers, Springer, 2019.

[110] D. Nepal, S. Kang, K. M. Adstedt, K. Kanhaiya, M. R. Bockstaller, L. C. Brinson, M. J. Buehler, P. V. Coveney, K. Dayal, J. A. El-Awady, et al., Hierarchically structured bioinspired nanocomposites, Nature materials 22 (1) (2023) 18–35.

[111] A. Pegoretti, A. Dorigato, Polymer composites: Reinforcing fillers, Encyclopedia of Polymer Science and Technology (2002) 1–72.

[112] W. J. Parnell, C. Calvo-Jurado, On the computation of the Hashin-Shtrikman bounds for transversely isotropic two-phase linear elastic fibre-reinforced composites, Journal of Engineering Mathematics 95 (2015) 295–323.

[113] S. Rana, R. Fangueiro, Advanced composite materials for aerospace engineering: processing, properties and applications, Woodhead Publishing, 2016.

[114] T. H. Hsieh, M. Y. Shen, Y. S. Huang, Q. Q. He, H. C. Chen, Mechanical properties of glass bead-modified polymer composite, Polymers and Polymer Composites 26 (1) (2018) 35–44.

[115] S. Bhatia, S. Angra, S. Khan, Mechanical and wear properties of epoxy matrix composite reinforced with varying ratios of solid glass microspheres, in: Journal of Physics: Conference Series, Vol. 26, IOP Publishing, 2019, pp. 35–44.

[116] T. He, O. Ertuğrul, N. Ciftci, V. Uhlenwinkel, K. Nielsch, S. Scudino, Effect of particle size ratio on microstructure and mechanical properties of aluminum matrix composites reinforced with Zr48Cu36Ag8Al8 metallic glass particles, Materials Science and Engineering: A 742 (2019) 517–525.

[117] V. Srinivasa, V. Shivakumar, V. Nayaka, S. Jagadeeshaiaih, M. Seethram, R. Shenoy, A. Nafidi, Fracture morphology of carbon fiber reinforced plastic composite laminates, Materials Research 13 (2010) 417–424.

[118] E. García-Macías, L. Rodríguez-Tembleque, A. Sáez, Bending and free vibration analysis of functionally graded graphene vs. carbon nanotube reinforced composite plates, Composite Structures 186 (2018) 123–138.

[119] H. Ahmadimoghaddamseighalani, A Stochastic Finite Element Analysis Framework for the Multiple Physical Modeling of Filler Modified Polymers, Ph.D. thesis, University of Alberta (2021).

[120] S. David Müzel, E. P. Bonhin, N. M. Guimarães, E. S. Guidi, Application of the finite element method in the analysis of composite materials: A review, Polymers 12 (4) (2020) 818.

[121] R. Hill, Elastic properties of reinforced solids: some theoretical principles, Journal of the Mechanics and Physics of Solids 11 (5) (1963) 357–372.

[122] P. Henyš, L. Čapek, J. Březina, Comparison of current methods for implementing periodic boundary conditions in multi-scale homogenisation, European Journal of Mechanics-A/Solids 78 (2019) 103825.

[123] C. K. Yan, On homogenization and de-homogenization of composite materials, Ph.D. thesis, Drexel University (2003).

[124] J. Segurado, J. Llorca, A numerical approximation to the elastic properties of sphere-reinforced composites, Journal of the Mechanics and Physics of Solids 50 (10) (2002) 2107–2121.

[125] T. Gentieu, A. Catapano, J. Jumel, J. Broughton, Computational modelling of particulate-reinforced materials up to high volume fractions: Linear elastic homogenisation, Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications 233 (6) (2019) 1101–1116.

[126] ANSYS, Ansys® Academic Research Mechanical, Release 18.1.

[127] J. Ju, K. Yanase, Micromechanics and effective elastic moduli of particle-reinforced composites with near-field particle interactions, Acta Mechanica 215 (2010) 135–153.

[128] D. Onalo, O. Oloruntobi, S. Adedigba, F. Khan, L. James, S. Butt, Static Young's modulus model prediction for formation evaluation, Journal of Petroleum Science and Engineering 171 (2018) 394–402.

[129] J. C. Smith, Experimental values for the elastic constants of a particulate-filled glassy polymer, Journal of Research of the National Bureau of Standards. Section A, Physics and Chemistry 80 (1) (1976) 45.

[130] Z. Zhang, G. Obasi, R. Morana, M. Preuss, In-situ observation of hydrogen induced crack initiation in a nickel-based superalloy, Scripta Materialia 140 (2017) 40–44.

[131] R. P. Gangloff, Hydrogen-assisted Cracking, in: I. Milne, R. Ritchie, B. Karihaloo (Eds.), Comprehensive Structural Integrity Vol. 6, Elsevier Science, New York, NY, 2003, pp. 31–101.

[132] S. K. Dwivedi, M. Vishwakarma, Hydrogen embrittlement in different materials: A review, International Journal of Hydrogen Energy 43 (46) (2018) 21603–21616.

[133] F. J. Castro, G. Meyer, Thermal desorption spectroscopy (TDS) method for hydrogen desorption characterization (I): theoretical aspects, Journal of alloys and Compounds 330 (2002) 59–63.

[134] G. Pellegrini, Polynomial Chaos Expansion with applications to PDEs, Ph.D. thesis, University of Verona (2014).

[135] X. Wan, G. E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM Journal on Scientific Computing 28 (3) (2006) 901–928.

[136] R. Salomon, Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions. A survey of some theoretical and practical aspects of genetic algorithms, BioSystems 39 (3) (1996) 263–278.

[137] A. Raina, V. S. Deshpande, N. A. Fleck, Analysis of thermal desorption of hydrogen in metallic alloys, Acta Materialia 144 (2018) 777–785.

[138] A. Mourad, F. Mroue, Z. Taha, Stochastic mathematical models for the spread of COVID-19: a novel epidemiological approach, Mathematical Medicine and Biology: A Journal of the IMA 39 (1) (2022) 49–76.

[139] R. Thiedmann, O. Stenzel, A. Spettl, P. R. Shearing, S. J. Harris, N. P. Brandon, V. Schmidt, Stochastic simulation model for the 3D morphology of composite materials in Li–ion batteries, Computational Materials Science 50 (12) (2011) 3365–3376.

[140] S. Meyer, A. Brückner-Foit, A. Möslang, A stochastic simulation model for microcrack initiation in a martensitic steel, Computational materials science 26 (2003) 102–110.

[141] G. K. Sakki, I. Tsoukalas, P. Kossieris, C. Makropoulos, A. Efstratiadis, Stochastic simulation-optimization framework for the design and assessment of renewable energy systems under uncertainty, Renewable and Sustainable Energy Reviews 168 (2022) 112886.

[142] X. Chen, K. K. Kim, Efficient VaR and CVaR measurement via stochastic kriging, INFORMS Journal on Computing 28 (4) (2016) 629–644.

[143] X. Ruan, Q. Zhou, L. Shu, J. Hu, L. Cao, Accurate prediction of the weld bead characteristic in laser keyhole welding based on the stochastic kriging model, Metals 8 (7) (2018) 486.

[144] E. Baker, P. Barbillon, A. Fadikar, R. B. Gramacy, R. Herbei, D. Higdon, J. Huang, L. R. Johnson, P. Ma, A. Mondal, et al., Analyzing stochastic computer models: A review with opportunities, Statistical Science 37 (1) (2022) 64–89.

[145] X. Chen, Q. Zhou, Sequential experimental designs for stochastic Kriging, in: Proceedings of the Winter Simulation Conference 2014, IEEE, 2014, pp. 3821–3832.

[146] X. Chen, Q. Zhou, Sequential design strategies for mean response surface metamodeling via stochastic kriging with adaptive exploration and exploitation, European Journal of Operational Research 262 (2) (2017) 575–585.

[147] Z. Wang, M. Ierapetritou, Constrained optimization of black-box stochastic systems using a novel feasibility enhanced Kriging-based method, Computers & Chemical Engineering 118 (2018) 210–223.

[148] P. Hao, S. Feng, H. Liu, Y. Wang, B. Wang, B. Wang, A novel Nested Stochastic Kriging model for response noise quantification and reliability analysis, Computer Methods in Applied Mechanics and Engineering 384 (2021) 113941.

[149] Y. Che, Z. Guo, C. Cheng, Generalized polynomial chaos-informed efficient stochastic Kriging, Journal of Computational Physics 445 (2021) 110598.

[150] S. D. Tajbakhsh, E. Del Castillo, J. L. Rosenberger, A Bayesian approach to sequential optimization based on computer experiments, Quality and Reliability Engineering International 31 (6) (2015) 1001–1012.

[151] J. Staum, Better simulation metamodeling: The why, what, and how of stochastic kriging, in: Proceedings of the 2009 Winter Simulation Conference (WSC), IEEE, 2009, pp. 119–133.

[152] J. P. Kleijnen, E. Mehdad, Estimating the variance of the predictor in stochastic Kriging, Simulation Modelling Practice and Theory 66 (2016) 166–173.

[153] S. Zheng, A. F. Seila, Some well-behaved estimators for the M/M/1 queue, Operations Research Letters 26 (5) (2000) 231–235.

[154] X. Chen, B. E. Ankenman, B. L. Nelson, Enhancing stochastic kriging metamodels with gradient estimators, Operations Research 61 (2) (2013) 512–528.

[155] G. Xie, X. Chen, Uniform error bounds for Stochastic Kriging, in: 2020 Winter Simulation Conference (WSC), IEEE, 2020, pp. 361–372.

[156] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least Angle Regression, The Annals of Statistics 32 (2) (2004) 407–451.

[157] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B: Statistical Methodology 58 (1) (1996) 267–288.

Appendix A

Paper A: Polynomial chaos expansion for uncertainty propagation analysis in numerical homogenization of 2D/3D periodic composite microstructures

Composite Structures 300 (2022) 116130

Contents lists available at ScienceDirect

Composite Structures

journal homepage: www.elsevier.com/locate/compstruct

Response to comments on

Polynomial chaos expansion for uncertainty propagation analysis in numerical homogenization of 2D/3D periodic composite microstructures

J.C. García-Merino [a,*], C. Calvo-Jurado [a], E. García-Macías [b]

[a] Department of Mathematics, School of Technology, 10003, Cáceres, Spain
[b] Department of Structural Mechanics and Hydraulic Engineering, Campus Universitario de Fuentenueva (Edificio Politécnico), 18071, Granada, Spain

ARTICLE INFO

MSC:
35Q74
35Q82
74Q15
74B99

Keywords:
Composites
Homogenization
Linear elasticity
Polynomial chaos expansion
Surrogate modelling
Uncertainty propagation

ABSTRACT

This paper proposes the use of adaptive polynomial chaos expansion (PCE) for uncertainty propagation analysis of the numerical homogenization of polymer composites doped with random dispersions of spherical inclusions. The developed PCE acts as a surrogate model bypassing the computationally intense numerical homogenization of the elastic properties of 2D/3D representative volume elements (RVEs) loaded with moderate to high filler contents. Numerical results and discussion are presented to assess the accuracy and computational efficiency of 2D and 3D homogenization meta-models. The ability of the developed approaches to perform uncertainty propagation analyses with minimum computational effort represents the main contribution of this work, which holds vast potential for the stochastic design of macroscopic composite structural elements.

1. Introduction

The development of composite materials constitutes an issue of great interest since the 1960s due to their numerous applications in Science and Engineering [1]. Predicting the effective mechanical properties of composite materials is often a challenging problem, involving complex microstructures with multi-scale and highly heterogeneous properties. To address this issue, a large variety of different homogenization methods have been proposed in the literature [2]. These range from analytical mean-field homogenization approaches [3,4] to numerical techniques using finite element modelling (FEM) [5], boundary elements methods (BEM) [6–8], atomistic-based continuum mechanics [9], molecular dynamics simulations (MDS) [10], Virtual FEM (VEM) [11–13] and Fast Fourier Transform techniques [14], just to mention a few. Numerical methods are particularly popular owing to their ability to represent the exact geometry of complex microstructures [15], keeping minimal the number of assumptions and simplifications of the underlying microstructure [5]. Nonetheless, these approaches suffer from large computational demands due to the elevated mesh densities usually required to discretize RVEs. This represents a major limitation in various applications where numerous model evaluations are required, namely optimization, inverse calibration, uncertainty propagation or reliability analyses.

The simplest numerical homogenization approach consists in the use of periodic unit cells [16] containing a canonical arrangement of particles which are assumed to replicate periodically throughout the composite material. However, most composite materials present a certain degree of randomness in the dispersion of the doping fillers. The problem of the determination of the effective behaviour of random composite materials has been extensively treated in the literature. Theoretical approaches span from classical bounds for two-phase composites [17] to analytical [18,19] or numerical [20,21] homogenization techniques. Moreover, considerable research efforts have been exerted toward the definition of accurate RVEs for heterogeneous materials through the analysis of convergence rates in computational homogenization approaches (see e.g. [22,23]). In general, algorithms for the generation of random microstructures can be firstly classified according to the dimensions of the RVE, namely 2D or 3D. Two-dimensional models represent a cost-efficient approach particularly well-suited for composite materials with uni-axial symmetry such as fully aligned fibre-reinforced composites [24]. Nonetheless, many composite materials present complex microstructures that can hardly be represented by 2D models, being imperative to implement more computationally intense 3D RVEs [25]. In this regard, several works in the literature have reported about the applicability and limitations of two- and three-dimensional RVEs. It is worth noting the work by Nil and

* Corresponding author.
E-mail address: jcgarcia@unex.es (J.C. García-Merino).

https://doi.org/10.1016/j.compstruct.2022.116130
Received 24 May 2022; Received in revised form 1 August 2022; Accepted 13 August 2022
Available online 20 August 2022

co-authors [26] who presented a comparison between 3D and 2D homogenized RVEs of meso-scale concrete. Their results showed that the effective diffusivity of the 3D model is about 1.4 times that of the 2D model. The effects of aggregate/void areas on the homogenized elastic and tensile fracture behaviours of concrete using 2D and 3D RVEs was also investigated by Hua et al. [27]. In that work, a 3D model was constructed from 2D images by a bottom-up stacking algorithm, therefore the resultant 3D model is strongly correlated to the 2D. In Ref. [28], the stresses of 2D and 3D RVEs of porous polymer material were investigated and compared. Those authors concluded that if 2D model reliably represents the 3D geometry, then the deformation behaviour can be analysed using the 2D RVE as the difference between stresses red obtained by the 2D and 3D RVEs do not exceed 10%. On the other hand, algorithms are also often classified according to the procedure used to define the filler distribution, namely dynamic (i), constructive (ii), non-sequential (iii), and discrete element simulations (iv). Dynamic algorithms are often considerably time consuming, since they need to update the position and interactions of the particles. Popular dynamic approaches are the movement and mechanical contraction or growth algorithms [29,30]. While easily implementable, the convergence rate of these algorithms for moderate to large filler contents is very limited, being hard to achieve filler contents above 30%. On the other hand, constructive algorithms have lower computational burden since the position of the particles are sequentially defined. Among this class of algorithms, one of the most intuitive is the Random Sequential Algorithm [31] (RSA), in which the location of any particle is randomly defined and accepted if it does not overlap with any of the previously located ones. Nonetheless, one of the major drawbacks of RSA regards the limited filler volume fractions that can be achieved, which usually limits to about 30% in 3D geometries [32]. Alternatively, constructive algorithms using dropping and rolling rules allows to evaluate the elastic properties of highly loaded composite materials. These algorithms simulate the process of spheres dropping into a dimension-specified cell when subjected to a gravitational field [33], allowing one to achieve high filler concentrations with limited computational costs. Non-sequential algorithms simulate a very dense distribution of large overlapping spheres that can move to reduce the degree of overlapping [34]. Finally, discrete element methods simulate filler packing as a dynamic process where inter-particle forces are explicitly accounted for [35].

While numerical homogenization methods provide a faithful representation of microstructural composites, their inherent computational demands limit their applicability in resource-intensive applications such as optimization or probabilistic design. In this light, recent developments in the realm of surrogate modelling have opened vast new possibilities to overcome such limitations. Roughly speaking, meta-models or surrogate models represent computationally lighter representations of parent models [36,37]. Once constructed, a surrogate model may be considered as a black-box representation of the original model, in many applications with computational costs several orders of magnitude lower. Despite their obvious potential, only a few research works in the literature report about the use of surrogate models for fast homogenization of composite structures. A noteworthy contribution was done by Dey and co-authors [38], who proposed a fuzzy membership function to identify parameter uncertainties in the characteristics analysis of noise in laminated composites. Peng et al. [39] proposed an uncertainty analysis method for composite laminated plates using PCE under insufficient input data related to uncertain design parameters. To that aim, the parameters space was divided according to the observation significance level of the variables. PCE methods are particularly well-suited for uncertainty propagation analysis since the expansion functions are intrinsically defined in stochastic terms [40–42]. In addition, PCE offers a direct framework for Global Sensitivity Analysis [43], allowing to compute Sobol's indices with the relative importance of the input random variables with no additional function evaluations [43]. Furthermore, PCE methods represent a non-intrusive technique, which results particularly useful when the parent model is defined through commercial software where access to the core scripts is not granted. In this line, Tha et al. [36] developed a stochastic multi-scale model for the uncertainty quantification of carbon fibre reinforced composites. In that work, red a PCE model was trained using data sets constructed by Monte Carlo simulations of a numerical homogenization model covering the design space of the parameters and the so called quantities of interest (QoIs), i.e. the corresponding model evaluations on the experimental design (ED).

In light of the discussion above, this work proposes a computationally efficient surrogate model based on adaptive PCE for fast evaluation and uncertainty propagation analysis of the elastic properties of composite materials obtained by 2D/3D numerical homogenization. To minimize the computational cost and maximize the robustness of the meta-model, adaptive sparse expansions are implemented based on the Least Angle Regression (LAR) algorithm [44]. This technique automatically identifies the optimal order of the polynomials in the PCE by a model selection technique for sparse linear models. This work focuses on the elastic properties of epoxy doped with glass fibre spherical inclusions, although the presented approach is general for any composite material. The training datasets are generated using 2D and 3D RVEs discretized by FEM with periodic boundary conditions. The location of the particles are defined using dropping and rolling rules, which allows to obtain filler volume fractions up to 50%. Once constructed, the surrogate models are used to assess the effects of uncertainties in the properties of the micro-constituents on the overall elastic properties of the composite. Finally, numerical results and discussion are presented to appraise the uncertainty propagation effects in 2D and 3D RVEs.

The remainder of this work is organized as follows. Section 2 presents the mathematical formulation of the elastic homogenization of composite materials, including the theoretical fundamentals of the random packing of particles. Section 3 outlines the theoretical formulation of PCE, and Section 4 presents the proposed surrogate model-based uncertainty propagation analysis of composite materials. Section 5 presents the numerical results and discussion and, finally, Section 6 concludes the paper.

2. Statement of the problem

Working in Cartesian coordinates and denoting $\mathbf{x} = (x, y, z)$, we restrict our attention in this work to a two-phase composite occupying a domain $\Omega \subset \mathbb{R}^3$. Within this domain, Ω_1 is identified as the inclusion phase or filler (glass particles) embedded into the host phase or matrix (epoxy) Ω_0, with respective boundaries Γ_1, Γ_0, and satisfying $\Omega = \Omega_1 \cup \Omega_0$, $\Omega_1 \cap \Omega_0 = \emptyset$. The volume fractions ξ_1 and ξ_0 of the inclusions and matrix phases, respectively, are defined as:

$$\xi_s = \frac{|\Omega_s|}{|\Omega|} = \frac{1}{|\Omega|}\int_{\Omega} \chi_s(\mathbf{x})\mathrm{d}\mathbf{x}, \quad s = 0,\ 1 \quad \text{subject to} \quad \xi_1 + \xi_0 = 1, \tag{1}$$

where $|\cdot|$ relates the volume of the constituent phases, and $\chi_s(\mathbf{x})$ is the characteristic function taking a value of one when $\mathbf{x} \in \Omega_s$ and zero otherwise. The governing equations of the elastic response of the body to a certain boundary force field $\mathbf{t}$, a imposed displacement field at the boundary $\bar{\mathbf{u}}$, and a volume force field $\mathbf{g}$ is given by an elliptic steady-state problem:

$$\begin{cases} -\mathrm{div}\,[\mathbf{C}(\mathbf{x})\boldsymbol{\epsilon}(\mathbf{x})] = \mathbf{g} & \text{in } \Omega \\ \mathbf{u}(\mathbf{x}) = \bar{\mathbf{u}} & \text{on } \Gamma_0 \\ \mathbf{C}(\mathbf{x})\boldsymbol{\epsilon}(\mathbf{x}) \cdot \mathbf{n} = \mathbf{t} & \text{on } \Gamma_1 \end{cases} \tag{2}$$

where $\mathbf{u}(\mathbf{x})$ is the displacement field, $\Gamma_0 \cup \Gamma_1 = \Gamma$ with $\Gamma_0 \cap \Gamma_1 = \emptyset$, and $\mathbf{n}$ the outer unit normal to the domain boundary Γ. The strain tensor $\boldsymbol{\epsilon}(\mathbf{x})$ is a fourth-order tensor, whose components in Mandel–Voigt notation can be derived from the displacement field $\mathbf{u}(\mathbf{x})$ as:

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{3}$$

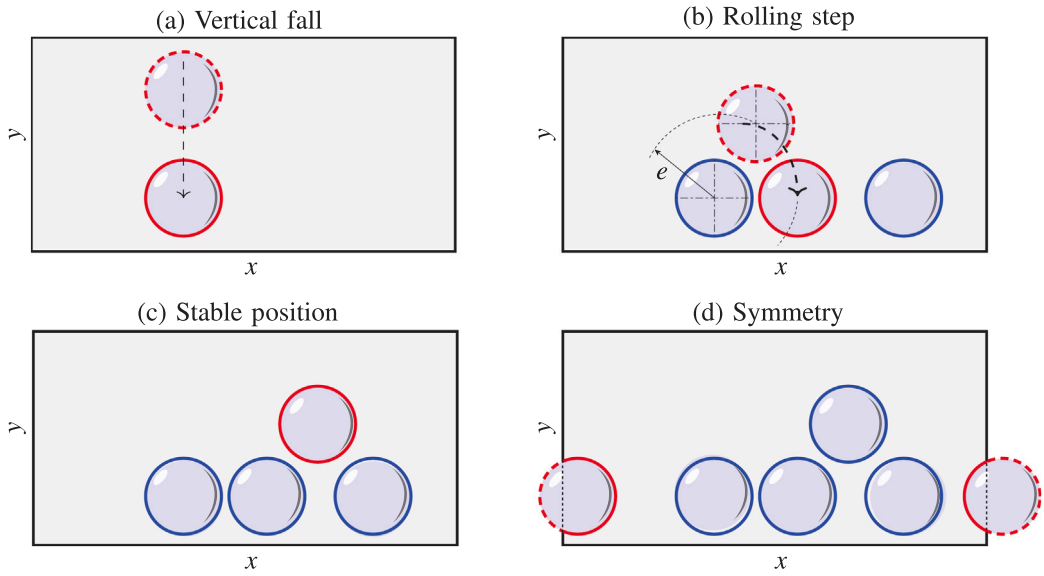

Fig. 1. Outline of the dropping and rolling algorithm for 2D RVEs.

and it is related to the stress tensor by the generalized Hooke's law $\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{C}(\mathbf{x})\boldsymbol{\epsilon}(\mathbf{x})$. The stiffness tensor $\mathbf{C}(\mathbf{x})$ of a material point $\mathbf{x}$ is given by:

$$\mathbf{C}(\mathbf{x}) = \begin{cases} \mathbf{C}^0 & \text{if } \mathbf{x} \in \Omega_0 \\ \mathbf{C}^1 & \text{if } \mathbf{x} \in \Omega_1 \end{cases} \tag{4}$$

with $\mathbf{C}^1$ and $\mathbf{C}^0$ the elastic tensors of the filler and the matrix, respectively. Extracting the exact solution of the displacement field from Eq. (2) is a challenging task, and closed-form solutions can only be found for simplified or ideal composite microstructures. Nevertheless, in practical applications, it is usual to characterize the macroscopic effective properties of the composite, represented by a fourth order elastic tensor $\mathbf{C}^*$ satisfying the homogenized problem of Eq. (2) given by:

$$\begin{cases} -\text{div}\,[\mathbf{C}^*\boldsymbol{\epsilon}(\mathbf{x})] = \mathbf{g} & \text{in } \Omega \\ \mathbf{u}(\mathbf{x}) = \bar{\mathbf{u}} & \text{on } \Gamma_0 \\ \mathbf{C}^*\boldsymbol{\epsilon} \cdot \mathbf{n} = \mathbf{t} & \text{on } \Gamma_1 \end{cases} \tag{5}$$

A variety of techniques have been proposed in the literature to solve the homogenization problem in Eq. (5), being numerical homogenization techniques particularly popular owing to their ability to faithfully represent the microstructure of the composite. These approaches are based on the definition of a RVE containing a sufficient number of inclusions to statistically represent the composite as a whole. Then, the RVE is subjected to a series of virtual tests by imposing certain boundary conditions on its external faces. The most commonly adopted approaches are Dirichlet (imposed displacements), Neumann (imposed forces), and periodic boundary conditions. In this work, periodic boundary conditions are adopted due to their recognized higher accuracy for low-to medium-scale RVEs [45], thereby geometrical periodicity will be imposed on the RVE. Let us focus in this work on spherical inclusions of radius r, and cubic (square) RVEs of edge length L. In order to reach high filler volume fractions, the dropping and rolling rules are considered in this work. This algorithm describes an iterative process where inclusions are sequentially introduced in a RVE under a gravitational field, and rigid contacts and sliding with pre-existing particles define the stacking process [33]. The filler volume fraction is controlled via the inter-particle parameter e describing the minimal distance between particles. In a 2D context, the algorithm is organized into four main steps as sketched in Fig. 1:

1. Considering gravity along the y-axis, generate the coordinate centre (x, y) of a particle, where x is randomly chosen and y is located above the top face of the RVE. Then, the particle is moved sequentially with linear trajectory of decreasing y (Fig. 1(a)).
2. If the particle contacts the floor of the box, it is assumed that it reaches its final position, and we return to Step 1 (Fig. 1 (a), (d)).
3. Otherwise, rolling rules minimizing the sphere gravity potential are applied. This means that the generated sphere rolls down over its contacting particle/particles until it reaches a stable position (with the minimum gravity potential) or the floor (Fig. 1 (b), (c)).
4. Then, return to Step 1 and repeat the process until the RVE is filled with particles.

In the 3D framework the process is similar, but in the rolling step different possibilities must be taken into account, depending on whether the falling particle comes into contact with one, two or three spheres [46]. In the first case, the sphere rolls in the same way as in the two-dimensional case. In the second case, the particle rolls in the direction defined by the plane normal to the vector joining the centres of the contact spheres. In the last one, the incoming sphere moves on the two spheres on which it rolls down most steeply until it reaches a stable position. Additionally, to impose periodic boundary conditions as explained hereafter, the geometric periodicity of the RVE is imposed. To do so, two different approaches are commonly adopted in the literature either inclusions are allowed or not to cut the boundaries of the RVE. The second approach is adopted in this work since, while the implementation becomes more involved, the resulting RVEs are more realistic. To this aim, any particle that intersects any of the vertical walls or corners of the RVE is replicated at the opposite wall or corner (see Figs. 1(d) and 2). Such a process can be described in the more complex case of 3D RVEs by the following procedure:

1. Consider a particle in stable position with centre's coordinates (x, y, z). This centre is allowed to partially fall outside of the RVE.
2. Check if the particle intersects any of the surfaces of the cell. If so, symmetric particles-one, three or seven, depending if the original one cuts one, two or three cell faces, respectively-are generated in order to impose periodicity (see Fig. 1(d)).
3. Check if any of the generated particles in the previous step overlap the previous existing ones. If this is not the case, the proposal particle is accepted, otherwise return to Step 1.
4. Repeat until the desired filler volume fraction is achieved.

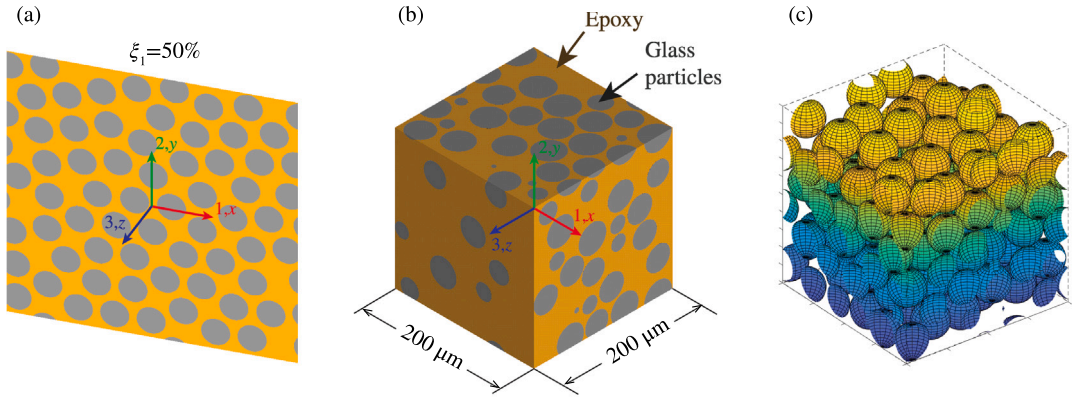

Fig. 2. (a) Example of a periodic 200 μm edge 2D RVE matrix doped with $\xi_1 \approx 50\%$ of glass particles with radius 10 μm. (b) RVE doped with $\xi_1 \approx 50\%$ volume fraction of glass microspheres with radius 20 μm. (c) Example of a 3D RVE generated particle distribution by means of the rolling and dropping procedure with $\xi_1 \approx 50\%$.

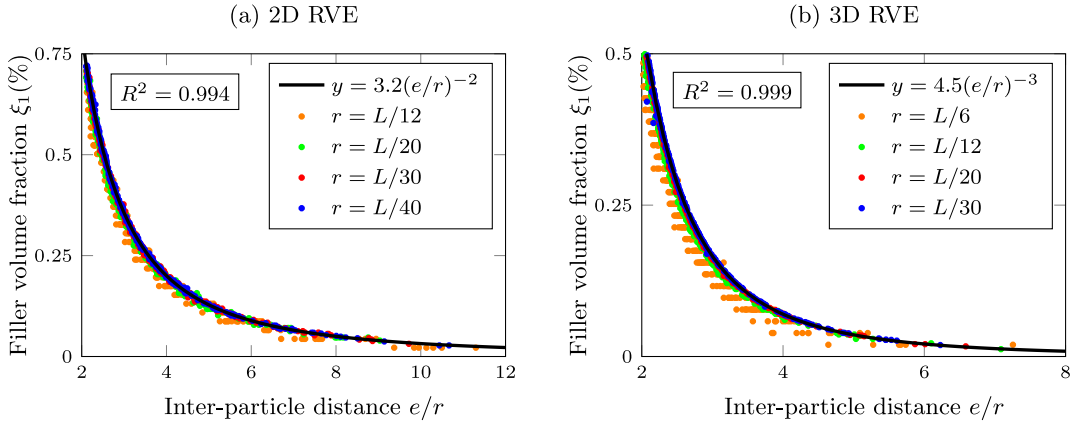

Fig. 3. Relationship between the filler volume fraction ξ_1 and the ratio between the inter-particle distance parameter e and the radius of the particles r for 2D and 3D RVEs.

For the 2D specimen $\Omega \subset \mathbb{R}^2$, a completely analogous procedure can be implemented. In either case, the final filler volume fraction ξ_1 is directly determined by the inter-particle distance parameter e and the diameter of the particles. In order to provide a simple expression for the relationship between the ratio e/r and the resulting volume fraction ξ_1, a Monte Carlo analysis with 300 different geometries has been conducted in a RVE with $L = 300$ μm with four different particles' radii r, namely $r = 7.5, 10, 15$, and 25 μm for 2D case and $r = 10, 15, 25$, and 50 μm for 3D. Fig. 3 shows the obtained volume fractions versus the ratio e/r. It is noted in this figure that a common trend is found for all the cases thanks to the non-dimensionalization of the results. Only some instabilities are observed for large particle sizes. In this light, new expressions for the filler volume fraction are obtained by curve fitting of the results in Fig. 3 as $\xi_1 = 3.2\,(e/r)^{-2}$ and $\xi_1 = 4.5\,(e/r)^{-3}$ for 2D and 3D RVEs, respectively. Note that the implemented algorithm achieves filler volume fractions up to 50%.

Once the RVE is defined, its effective elastic properties are obtained in this work by applying periodic boundary conditions as anticipated above. The enforcement of geometrical periodicity implies that the composite material can be conceived as the periodic replication of the RVE (Fig. 4(a)). The general periodic boundary conditions on the cell faces of a RVE are given by [47]:

$$u_i = \hat{\varepsilon}_{ij} x_j + v_i, \tag{6}$$

where $\hat{\varepsilon}_{ij}$ denote the volume average strains, and v_i represents the local periodic part of the displacement components u_i on the boundary surfaces. The latter displacement components are generally unknown and depend upon the applied loading. Indices i and j denote the global Cartesian directions. In the case of square RVEs like the ones used in this work and sketched in Fig. 4(a), Eq. (6) takes a more explicit expression. Consider the notation of the cell surfaces A^-/A^+, B^-/B^+, and C^-/C^+ shown in Fig. 4(a).

Then, the displacements on a pair of opposite boundary cells (with their normal along the x_j axis) read:

$$u_i^{K+} = \hat{\varepsilon}_{ij} x_j^{K+} + v_i^{K+}, \quad u_i^{K-} = \hat{\varepsilon}_{ij} x_j^{K-} + v_i^{K-}, \tag{7}$$

where indexes $K+$ and $K-$ indicate the displacements along the positive and negative x_j directions, respectively. Local fluctuations v_i^{K-} and v_i^{K+} must be identical on every two opposing faces to comply with the periodicity of the RVE. Therefore, the local displacement components can be dropped from the formulation by the difference between the expressions in Eq. (7), leading to:

$$u_i^{K+} - u_i^{K-} = \hat{\varepsilon}_{ij} \left(x_j^{K+} - x_j^{K-} \right). \tag{8}$$

Therefore, the RVE can be subjected to a desired strain state by imposing proper displacements on its boundary surfaces. Then, the volume average stresses $\hat{\sigma}_{ij}$ and strains $\hat{\varepsilon}_{ij}$ in the RVE can be computed as:

$$\hat{\varepsilon}_{ij} = \frac{1}{\mathcal{V}} \int_{\mathcal{V}} \varepsilon_{ij} \mathrm{d}\mathcal{V}, \quad \hat{\sigma}_{ij} = \frac{1}{\mathcal{V}} \int_{\mathcal{V}} \sigma_{ij} \mathrm{d}\mathcal{V}, \tag{9}$$

with $\mathcal{V}$ being the volume of the RVE. Then, the ij-th component of the elastic tensor can be directly estimated as $\mathbf{C}^*_{ij} = \hat{\sigma}_{ij}/\hat{\varepsilon}_{ij}$.

3. Adaptive PCE surrogate modelling

Denote by $\mathcal{M}^*$, with $*= 2D, 3D$, the numerical homogenization models of representative 2D/3D RVEs. Assuming the microstructural

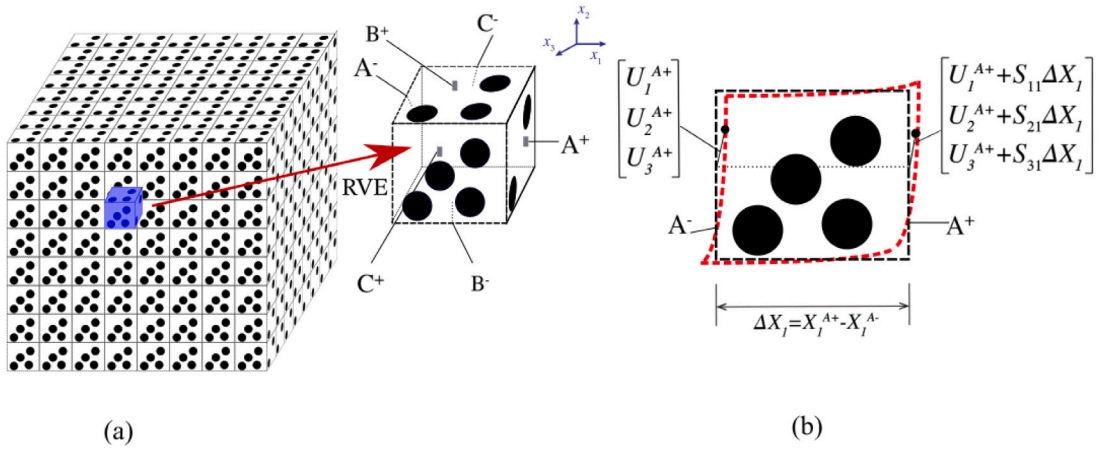

Fig. 4. (a) Representation of a periodic composite and definition of the RVE. (b) Periodic boundary conditions for a pair of nodes located on the opposite surfaces A^- and A^+.

properties of interest as independent random variables arranged in a vector $\mathbf{X} = (X_1, X_2, \dots, X_M)$, $\mathcal{M}^*$ maps a M-dimensional input parameter space to a 1-dimensional output space $\mathcal{M}^* : \mathbb{R}^M \to \mathbb{R}$. Due to the presence of uncertainties in the input vector, $\mathbf{X}$ is considered a random variable with known probability distribution. Then, a surrogate model $\hat{\mathcal{M}}^*$ aims to emulate the original homogenization model $\mathcal{M}^*$, but at a lower computational cost. The general procedure to construct a surrogate model can be summarized in the following steps:

1. Sample two independent input data sets: the *training set* (TS) T and the *validation set* (VS) V covering the parameter design space:

$$T = \left\{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}\right\} \subset \mathbb{R}^{M\times N}, \quad V = \left\{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}\right\} \subset \mathbb{R}^{M\times K}. \tag{10}$$

2. Evaluate model $\mathcal{M}^*$ on set T.
3. Solve an optimization problem to identify the parameters of the surrogate model.
4. Assess the metamodel's accuracy by evaluating $\mathcal{M}^*$ and $\hat{\mathcal{M}}^*$ on set V.

In this study, PCE is adopted for its convenience to represent stochastic magnitudes of interest. In particular, PCE is used to represent the effective elastic properties y obtained by numerical homogenization through an expansion onto an orthogonal multivariable polynomial basis [16] as:

$$y = \mathcal{M}^*(\mathbf{X}) = \sum_{\alpha\in\mathbb{N}^M} a_\alpha \boldsymbol{\Psi}_\alpha(\mathbf{X}), \qquad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M),\ \alpha_i \in \mathbb{N}, \tag{11}$$

where:

- $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X}) = \prod_{i=1}^{M} \psi_{\alpha_i}^{(i)}(X_i)$ are multivariate polynomials.
- $\psi_{\alpha_i}^{(i)}$ are orthogonal polynomials depending on the stochastic nature of the input variables $\mathbf{X}$. In this work, Legendre and Hermite polynomials will be used to represent uniform and Gaussian distributions.

Although expression (11) can be proved exact for an infinite number of polynomials, in practice only a finite number of terms in $\sum_{\alpha\in\mathbb{N}^M} a_\alpha \boldsymbol{\Psi}_\alpha(\mathbf{X})$ can be computed. As a consequence, different strategies can be taken into account to truncate the polynomial series. The simplest one consists in selecting all the polynomials whose total degree $|\boldsymbol{\alpha}| = \sum_{i=1}^{M} \alpha_i$ belongs to the set:

$$\mathcal{A}^{M,p} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^M : 0 \le |\boldsymbol{\alpha}| \le p\right\}, \tag{12}$$

where the cardinality of $\mathcal{A}^{M,p}$ is equal to:

$$\binom{M+p}{p} = \frac{(M+p)!}{M!p!}. \tag{13}$$

Nonetheless, when M and p are large enough, this procedure of polynomial selection may lead to computing a large number of coefficients and the subsequent computational burden. As an alternative, an hyperbolic truncation scheme can be employed. This approach consists in selecting all multi-indices with q-norm [48]

$$\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{\frac{1}{q}},$$

less or equal to p, i.e.:

$$\mathcal{A}^{M,p,q} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^M : \|\boldsymbol{\alpha}\|_q \le p\right\}. \tag{14}$$

This strategy has been proved efficient due to the fact that high interaction terms often have coefficients close to zero. Nonetheless, although substantial cost reductions can be achieved using the hyperbolic truncation scheme, the number of coefficients in the expansion may still be considerable. A better cost-efficient solution can be obtained by using the adaptive LAR algorithm [49]. LAR constructs a set of expansions incorporating an increasing number of basis polynomials $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$, from 1 to $\mathcal{P} = \mathrm{card}\left(\mathcal{A}^{M,p,q}\right)$. Then, the resulting sequence of index sets is used to construct different expansions and the best metamodel is selected by a cross validation procedure. Finally, the expansion coefficients $\boldsymbol{a} = \{\boldsymbol{a}_{\boldsymbol{\alpha}},\ \ \boldsymbol{\alpha} \in \mathcal{A}^{M,p} \subset \mathbb{N}^M\}$ are obtained by minimizing the expectation of the least squares errors:

$$\hat{\boldsymbol{a}} = \arg\min_{a\in\mathbb{R}^{\mathcal{P}}} \frac{1}{N} \sum_{i=1}^{M} \left[\mathcal{M}^*(\mathbf{X}^{(i)}) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \boldsymbol{a}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X}^{(i)})\right]^2, \tag{15}$$

where $\mathbf{X}^{(i)} \in T$, $i = 1, \dots, N$. Once Eq. (15) has been solved, the resulting PCE surrogate model can be written as follows:

$$y = \mathcal{M}^*(\mathbf{X}) \approx \hat{\mathcal{M}}^*(\mathbf{X}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\boldsymbol{a}}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\mathbf{X}). \tag{16}$$

4. Surrogate model-based uncertainty propagation analysis of composite materials

The main contribution of this section regards the application of the previously introduced surrogate modelling approach to the analysis of the propagation of uncertainties in the material parameters through the homogenization of both 2D/3D composites with high filler volume fraction. The general methodology is sketched in Fig. 5 and includes five main steps, namely (i) geometry construction using the dropping and rolling method, (ii) design of proper RVEs, (iii) construction of the metamodels, (iv) validation and (v) uncertainty propagation analysis.

In a first step, general 2D/3D quadratic and cubic models allowing particle intersection with the cell boundaries (Fig. 2) are implemented. For the construction and homogenization of the microstructure, a combination of scripts generated in MATLAB environment [50] and the

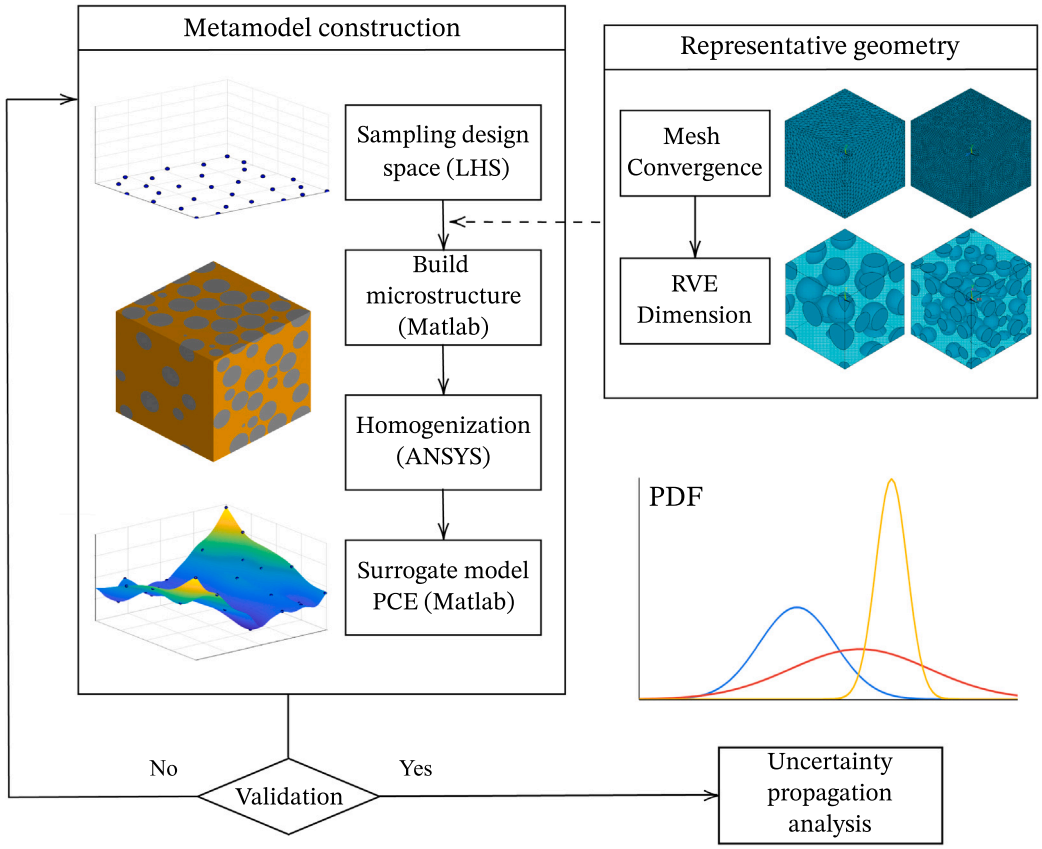

Fig. 5. Flowchart of the proposed meta-model approach to perform uncertainty propagation analysis in 2D/3D homogenization.

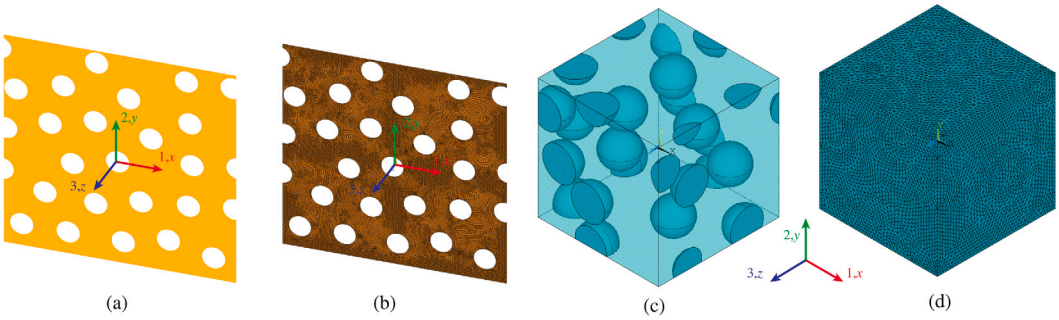

Fig. 6. (a) Matrix and (b) meshed matrix of an example of a periodic 200 μm edge 2D RVE of epoxy doped with $\xi_1 \approx 20\%$ volume fraction of spherical particles. (c) RVE with a length side of 200 μm doped with $\xi_1 \approx 15\%$ volume fraction of glass microspheres (d) and corresponding FEM mesh (90 443 nodes).

commercial FEM tool ANSYS is used [51]. Specifically, the adopted methodology for defining the geometry of cubic RVEs involves the steps previously given in Section 2. Once the geometry is constructed, it is discretized in ANSYS using plane-strain 8-nodes quadratic elements (PLANE 183) for 2D RVEs, while 4-nodes linear tetrahedral solid elements (SOLID 285) are used for 3D RVEs. Two samples of the generated RVEs are shown in Fig. 6 (a), (c) along with their corresponding meshes in Fig. 6 (b), (d). Once constructed, the numerical homogenization model is used to draw the samples of the training and validation sets for the construction of the surrogate model.

To assess the precision of the developed surrogate models, both local and global error metrics are considered. These include the coefficient of determination R^2 and the Normalized Average Absolute Error (NAAE) for the global accuracy and the Normalized Maximum Absolute Error (NMAE) for the local case, are reported in Table 1, where $\bar{y}$ denotes the arithmetic mean of the validation output set. These global metrics are evaluated over two data sets $\left\{y^{(1)} = \mathcal{M}(\mathbf{X}^{(1)}), \ldots, y^{(K)} = \mathcal{M}(\mathbf{X}^{(K)})\right\}$ and $\left\{\hat{y}^{(1)} = \hat{\mathcal{M}}^*(\mathbf{X}^{(1)}), \ldots, \hat{y}^{(K)} = \hat{\mathcal{M}}^*(\mathbf{X}^{(K)})\right\}$ obtained, respectively, by the evaluation of the numerical homogenization models and their corresponding meta-models over the VS with K realizations.

Once constructed, PCE sensitivity analysis (SA) is used to investigate the contribution (relative importance) of each random input parameter of the mathematical model $\mathcal{M}$ to its stochastic response. The non-linearity of the proposed model implies that not only the variables individually affect the response, but also the interactions between them. Therefore, the joint-effect of parameter variation must be also quantified. Sobol' indices [43] represent a global SA method as they provide the sensitivity of the predictions of the surrogate model to variations in the selected input variables. The Sobol' index $S_{\boldsymbol{u}}$ associated to the subset $\boldsymbol{u} = \{i_1, \ldots, i_M\} \subset \{1, \ldots, M\}$ is defined as the ratio between the contribution to the model variance given by the interaction among the

Table 1
Error metrics for the accuracy assessment of surrogate models over a validation set of size K.

Coefficient of Determination (R^2)	Normalized average absolute error (NAAE)	Normalized maximum absolute error (NMAE)
$R^2 = 1 - \frac{\sum_{i=1}^{K}\left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_{i=1}^{K}\left(\bar{y} - y^{(i)}\right)^2}$	$\text{NAAE} = \frac{\sum_{i=1}^{K}\lvert\hat{y}^{(i)} - y^{(i)}\rvert}{K\sqrt{\frac{1}{k-1}\sum_{i=1}^{K}\left(\bar{y} - y^{(i)}\right)^2}}$	$\text{NMAE} = \frac{\max_{i=1}^{K}\lvert\hat{y}^{(i)} - y^{(i)}\rvert}{K\sqrt{\frac{1}{k-1}\sum_{i=1}^{K}\left(\bar{y} - y^{(i)}\right)^2}}$

Table 2
Mechanical properties of the constituent phases of epoxy doped with glass fibre micro-spheres [52].

	Epoxy polymer	Glass fibre particles
Young's Modulus [GPa]	3.0	76.0
Poisson's ratio [–]	0.4	0.23
Radius [μm]	–	8–20

components of $\boldsymbol{u}$ and the total variance. Thanks to the orthogonality of the PCE, it can be written as:

$$S_{\boldsymbol{u}} = \left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\boldsymbol{u}}} \boldsymbol{a}_{\boldsymbol{\alpha}}^2\right) \Big/ \left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}}^{\mathcal{P}} \boldsymbol{a}_{\boldsymbol{\alpha}}^2\right), \qquad \mathcal{A}_{\boldsymbol{u}} = \{\boldsymbol{\alpha}\in\mathcal{A} : \alpha_k \neq 0 \text{ when } k\in\boldsymbol{u}\}. \tag{17}$$

5. Numerical results and discussion

This section presents the numerical results and discussion obtained by applying the previously introduced methodology for the uncertainty propagation analysis of epoxy composites doped with glass fibre spherical particles with a wide range of filler concentrations (from 0% up to 50%). The presented results are organized as follows: Section 5.1 presents preliminary convergence analyses conducted to identity the minimum RVE's sizes and mesh densities. Section 5.2 reports the construction of the surrogate models and their quality assessment. Finally, Section 5.3 reports the surrogate model-based uncertainty propagation analyses by introducing uncertainties in the elastic properties of the micro-constituents. In the following analyses, the material properties of epoxy and glass spheres are taken from Ref. [52] as listed in Table 2.

5.1. Convergence analysis of the RVEs' sizes and mesh density

To guarantee that a RVE statistically represents the composite material, its dimension and the number of embedded particles must be carefully selected. In addition, since the microstructure is then discretized using FEM, another important aspect to determine regards the mesh density. In general, the dimensions of the RVE and its mesh density must be chosen as a trade-off between representativity and computational cost. Firstly, the discretization of the RVEs is selected fine enough to ensure that the homogenized constitutive tensor $\mathbf{C}^*$ is mesh independent. On the other hand, the dimensions of the RVE are selected according to the degree of isotropy of the obtained constitutive tensor. Given that both matrix phase and reinforcing particles are isotropic, spherical, and randomly dispersed, the resulting effective constitutive tensor must also present isotropic symmetry. In this light, Fig. 7 reports the convergence analysis of two 2D meshes with edge lengths of 120 μm and 240 μm. In both cases, a filler volume fraction of 20% has been selected. In this figure, the effective Young's moduli E_x, E_y are presented versus the total number of nodes in the mesh (xyz reference frame indicated in Fig. 6). Five different meshes with increasing densities have been considered, including 2116, 7592, 11 555, 22 304, 43 776 nodes for the 120 × 120 μm RVE and 15 086, 27 515, 43 388, 83 531, 111512 for the 240 × 240 μm one. It is noted in this figure that convergence is approximately achieved between the third and the fourth discretization densities (22 000–83 500 nodes, mean element edge size 1.32% the RVE size). Therefore, the forth mesh density is selected as a conservative solution. These analyses were also conducted for 3D RVEs, achieving similar conclusions (mean element edge size 1.11% the RVE size).

Once the mesh density has been defined, the size of the RVEs is selected by convergence analysis of the isotropy degree of the effective constitutive tensor. To do so, five geometries with different RVE sizes have been studied, namely 80, 120, 160, 200 and 240 μm (80, 120, 160 and 200 μm in 3D) have been studied. The convergence analysis of the effective Young's moduli E_i obtained for 2D and 3D RVEs are shown in Figs. 8 and 9, respectively. Moreover, the evolution of the mean Poisson's ratios $\{\nu_{xy}\}$ and $\{\nu_{xy}, \nu_{xz}, \nu_{yz}\}$ obtained by 2D and 3D RVEs are also reported in Fig. 10(a) and (b), respectively. In the ideal case of perfect isotropy, ratios E_i/E_j should be exactly equal to 1. Therefore, following Figs. 8 (a) and 9 (a,b,c), it can be concluded that convergence is approximately achieved for RVEs with an edge length of 240 μm in 2D and 200 μm in 3D. In these cases, it is noted that the ratios between the elastic moduli are very close to 1.

The previous analyses are further investigated in Fig. 11, which reports the mean values of the main components of the effective constitutive tensor $\mathbf{C}^*$ obtained for five realizations of the RVEs with increasing sizes. In light with the previous analyses, it is observed in this figure that the composites exhibit a higher degree of isotropy as the RVE size increases. For instance, note in Fig. 11 (a) and (c) that the components C_{ii} (diagonal terms of the constitutive tensor) tend to a common value as expected for isotropic materials. The residual anisotropy in the effective constitutive tensor is eliminated hereafter through integration over all possible orientations in the Euler space, also referred to as the orientational average. To do so, tensor $\mathbf{C}^*$ is rotated to any orientation in the Euler space as $C_{ijkl} = a_{ip}a_{jq}a_{kr}a_{ls}C^*_{pqrs}$, where $\boldsymbol{a}$ is the transformation matrix consisting of θ, ϕ, and ψ rotation angles. Then, the orientational average of $\mathbf{C}^*$ can be obtained as [53]:

$$\langle\mathbf{C}^*\rangle = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\int_0^{\pi/2}\mathbf{C}^*(\theta,\phi,\psi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}\psi. \tag{18}$$

On this basis, the effective elastic modulus E and Poisson's ratio ν can be obtained from the compliance tensor $\mathbf{S} = (\mathbf{C}^*)^{-1}$ in Mandel–Voigt notation as

$$E = \frac{1}{S_{11}}, \quad \nu = -ES_{12}. \tag{19}$$

5.2. Surrogate model construction. Accuracy, sensitivity analysis and computational efficiency

Following the theoretical framework introduced in Section 3, to approximate the 2D/3D homogenized Young's moduli, two PCE surrogate models $\hat{\mathcal{M}}^{2D}$ and $\hat{\mathcal{M}}^{3D}$ have been constructed taking $p = 0.95$ in Eq. (14) and model orders between 1 and 4. To this purpose, 50 training points and 25 independent validation points have been considered by means of the Latin Hypercube Sampling (LHS) [54] procedure. Five input variables have been included in both models, namely the Young's moduli and Poisson's ratios of the matrix and the inclusions, as well as the filler volume fraction. Uncertainties in the elastic properties of the constituents are assumed normally distributed with mean values from Table 2 and standard deviations corresponding to 1/30 of mean values (corresponding deviations of up to ±10% of them). On the other hand, the filler volume fraction is assumed uniformly distributed within the range from 0 to 50%. The selected probability distributions are collected in Table 3.

The performance assessment of the developed surrogate models through the error metrics from Table 1 as well as the computational

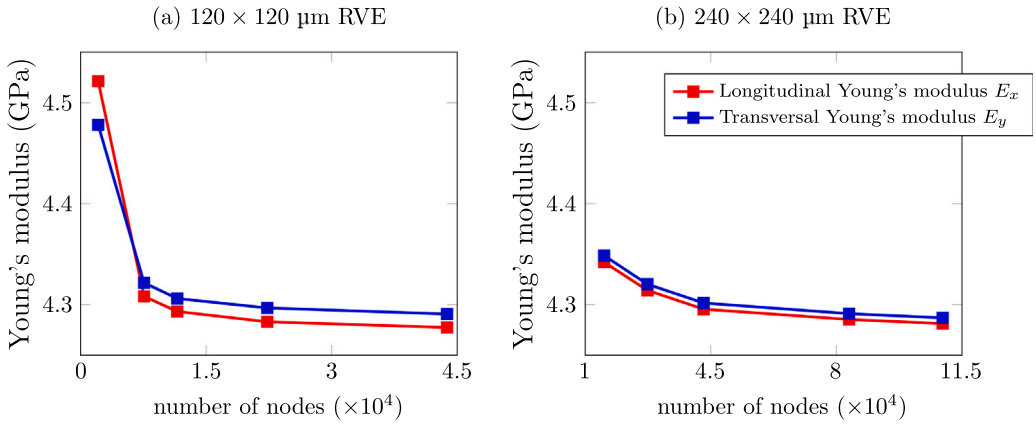

Fig. 7. Convergence analysis of the mesh density for a 2D RVE of dimensions 120 × 120 µm (a) and 240 × 240 µm (b) doped with 20% volume fraction of spherical particles with radius 8 µm.

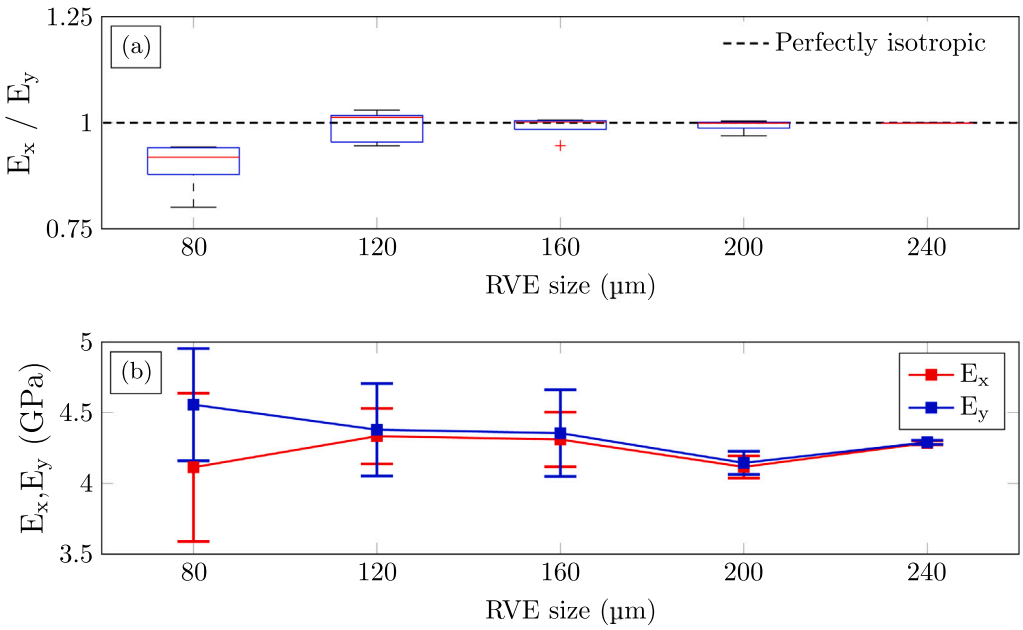

Fig. 8. Analysis of the isotropy of the composite (a) and evolution of the elastic Young's moduli E_x, E_y with error bars denoting the standard deviation (b) against the length side of the 2D RVE. (Filler volume fraction $\xi_1 = 20\%$).

Table 3
Statistical distributions of the material parameters.

Variable	Notation	Distribution
Young's modulus of epoxy matrix [GPa]	E_0	Gaussian $\mathcal{N}(3, 0.1)$
Poisson's ratio of epoxy matrix [–]	ν_0	Gaussian $\mathcal{N}(0.4, 0.0133)$
Young's modulus of glass reinforcement [GPa]	E_1	Gaussian $\mathcal{N}(76, 2.5333)$
Poisson's ratio of glass reinforcement [–]	ν_1	Gaussian $\mathcal{N}(0.23, 0.0076)$
Volume fraction of inclusions [–]	ξ_1	Uniform $\mathcal{U}(0, 0.5)$

Table 4
Surrogate model accuracy.

	R^2	NAAE (10^{-2})	NMAE (10^{-2})
$\hat{\mathcal{M}}^{2D}$	0.9939	5.34	0.84
$\hat{\mathcal{M}}^{3D}$	0.9993	2.15	0.40

time savings are reported in Table 4. It can noted in this table that the surrogate models $\hat{\mathcal{M}}^{2D}$ and $\hat{\mathcal{M}}^{3D}$ faithfully reflect the behaviour of the corresponding parent models $\mathcal{M}^{2D}$ and $\mathcal{M}^{3D}$. This coincides with Fig. 12 which shows a comparison between them and the forward

Table 5
Sobol' indices.

	S_{ξ_1}	S_{E_0}	S_{ν_0}	S_{E_1}	S_{ν_1}
$\hat{\mathcal{M}}^{2D}$	0.9913	0.0075	0.0031	$1.7021E{-}5$	0
$\hat{\mathcal{M}}^{3D}$	0.9831	0.0086	0.0130	0	0

FEM over the 25 samples of the validation set. The low scatters of the points around the diagonal lines and the coefficients of determination (very close to 1) corroborate that the surrogate models are formed with accuracy. With respect to the computational times, it is important to remark that, while the numerical homogenization models $\mathcal{M}^{3D}$ and $\mathcal{M}^{2D}$ take around 6 h and several minutes, respectively, the evaluation of the respective meta-models ($\hat{\mathcal{M}}^{3D}$, $\hat{\mathcal{M}}^{2D}$) takes less than 1 millisecond (99.7% reduction). Based on the metrics in Table 4 and the fact that 2D and 3D models provide similar results for moderate volume fractions, 2D approximations may be adopted in these cases for fast evaluations with reasonably good accuracy. Finally, Table 5 reports the Sobol' indices of the constructed expansions. As it is expected, the

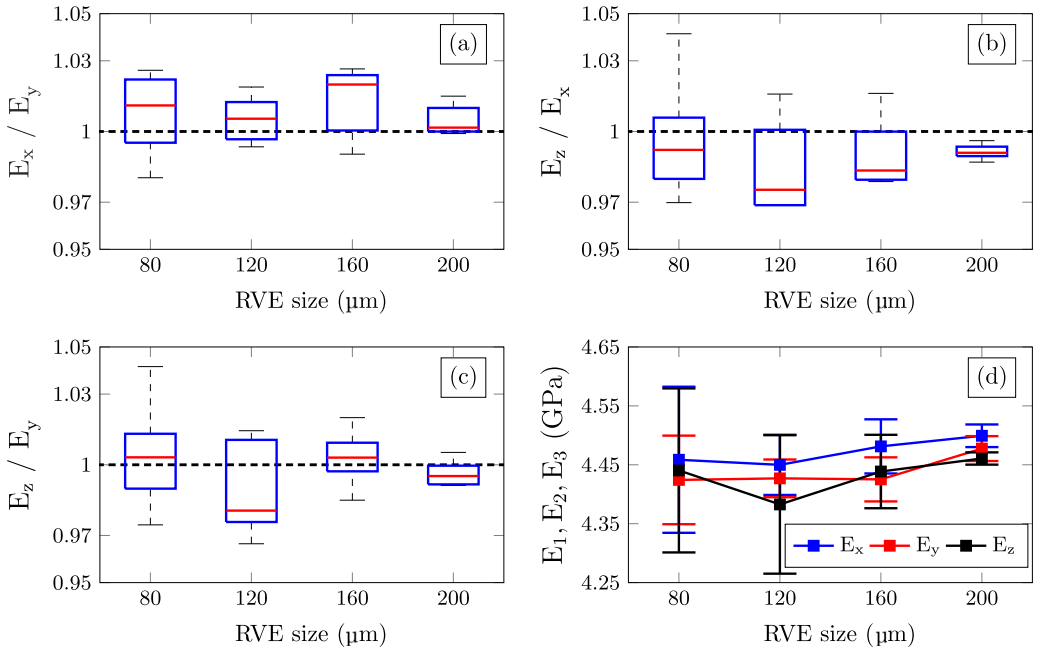

Fig. 9. Analysis of the isotropy of the composite in terms of the ratio between the Young moduli (a) E_x/E_y, (b) E_z/E_x, (c) E_z/E_y and (d) evolution of the mean Young's moduli E_x, E_y, E_z against the length side of the 3D RVE for a volume fraction $\xi_1 = 0.2$ (error bars in (d) denote standard deviation).

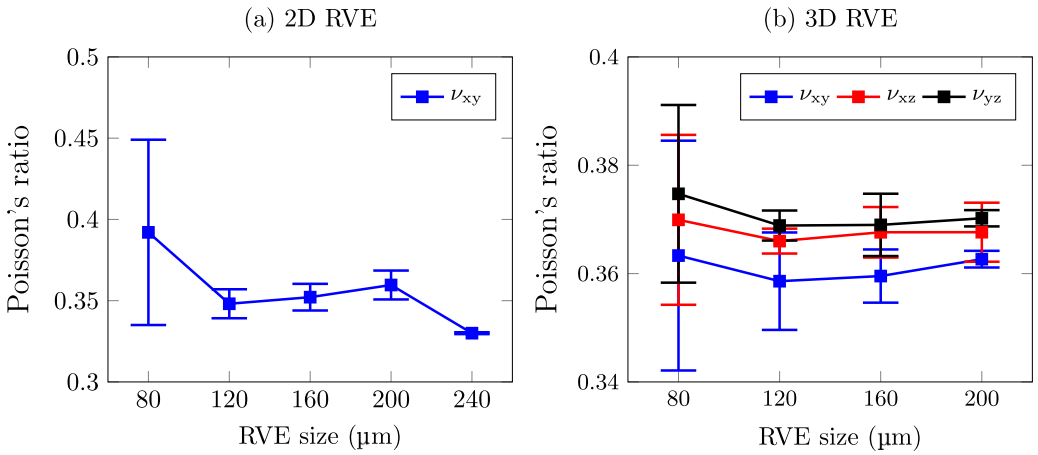

Fig. 10. Convergence analysis of the mean effective Poisson's ratios $\{\nu_{xy}\}$ and $\{\nu_{xy}, \nu_{xz}, \nu_{yz}\}$ obtained by 2D (a) and 3D RVEs (b) against the RVE size for a filler volume fraction of $\xi_1 = 0.2$. Error bars stand for standard deviations.

homogenization results are particularly sensitive to the filler's volume fraction contribution followed by the properties of the matrix.

5.3. Uncertainty propagation analyses. Comparison between the 2D and 3D surrogate approximations

The accuracy and computational efficiency of the developed surrogate models are leveraged herein to analyse the propagation of uncertainties in the microstructural properties of epoxy/glass composites upon the effective elastic properties. Firstly, uncertainty propagation analyses are conducted by considering isolated uncertainties in every of the considered microstructural parameters. In each analysis, 25 000 Monte Carlo individuals are drawn by sampling from the statistical distribution of the parameter under analysis as reported in Table 3, while the remaining parameters are fixed to their mean values. It is important to remark that the simulations took less than 1 s, while conducting such analyses using the numerical homogenization model would be simply infeasible. Once the Monte Carlo samples are obtained, the statistical distributions of the effective elastic properties are estimated in frequentist terms. Following this, Fig. 13 represents the probability distribution functions (PDFs) of the considered parameters in the expansions for the 2D/3D surrogate models. In both 2D and 3D scenarios, Fig. 13 shows that for low filler volume fractions, uncertainties in E_0 dominate the uncertainty of the effective Young's modulus. Nonetheless, almost no effect is observed for variations in the elastic modulus of the glass inclusions and the remaining parameters. It is also noted that for higher filler volume fractions, some influence is found due to uncertainty in the Poisson's ratio of the host matrix.

Finally, Fig. 14 presents the statistical probability distribution functions (PDFs) of the effective elastic moduli of 2D/3D glass fibre/epoxy composites when considering simultaneous uncertainty in all the considered microstructural parameters. The analyses are conducted for two filler volume fractions, namely 15% and 45%. It is shown in this figure that, for the filler volume fraction of 15%, the mean values of the elastic moduli predicted by the 2D and the 3D models are almost identical. However, as the volume fraction increases, increasing differences arise. For example, it is noted in Fig. 14 that for a filler volume fraction of

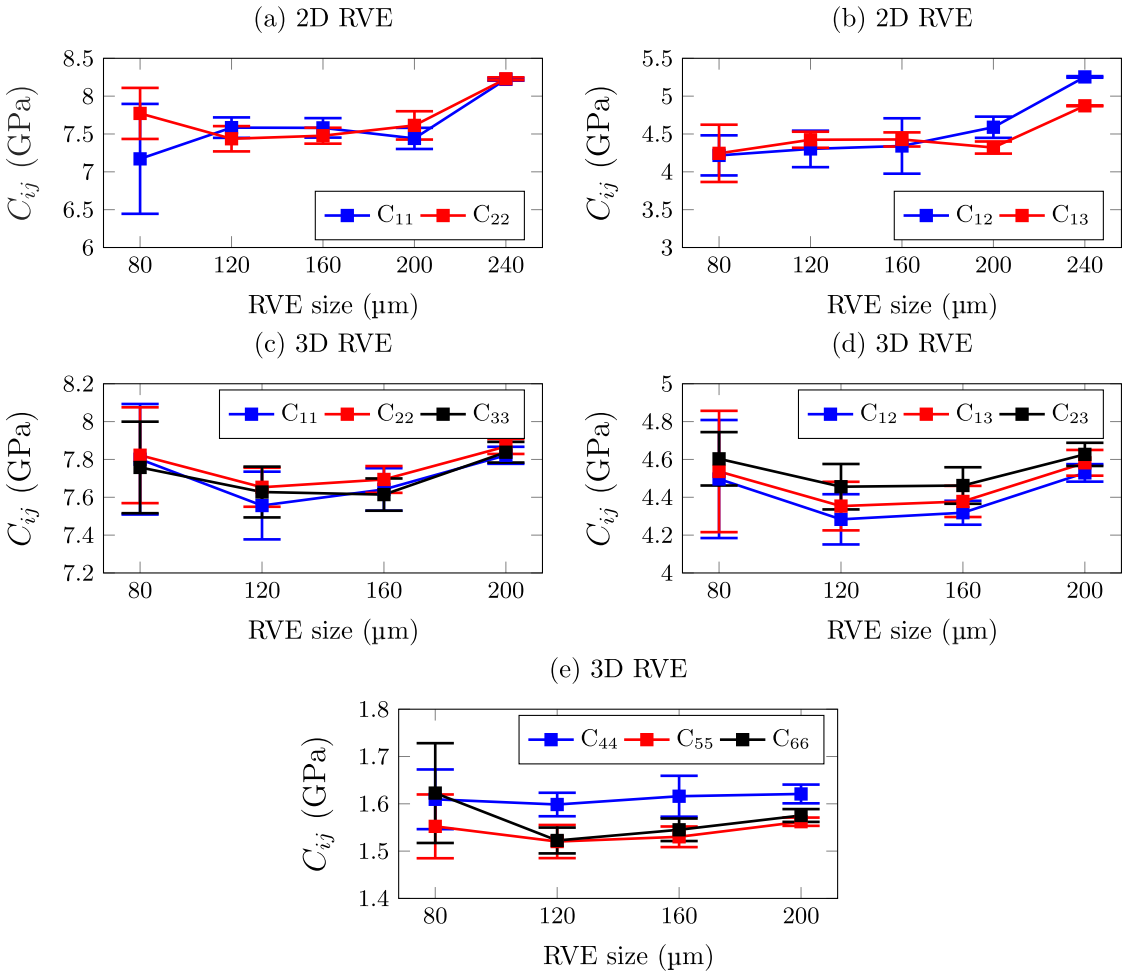

Fig. 11. Convergence analysis of the mean effective constitutive tensor components $\{C_{ij}\}$ by 2D (a,b) and 3D RVEs (c,d,e) against the RVE size for a filler volume fraction of $\xi_1 = 0.2$. Error bars stand for standard deviations.

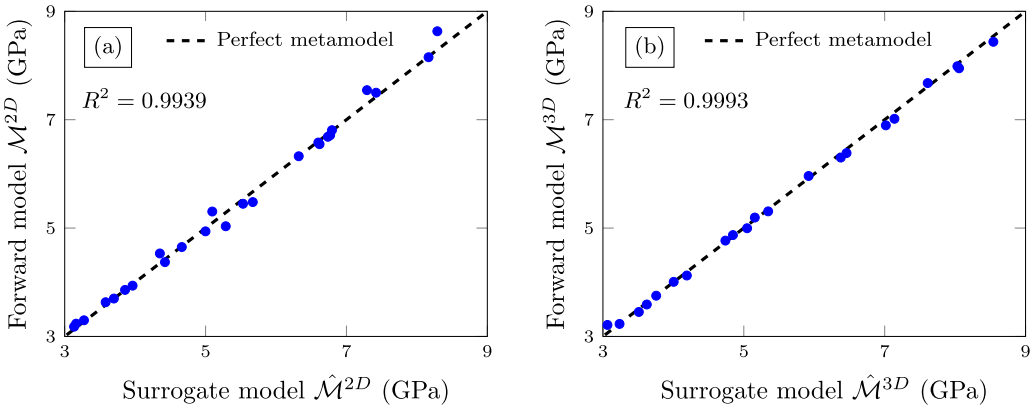

Fig. 12. Comparison of the numerical estimates by FEM and the predictions of the PCE surrogate models in terms of the elastic moduli in (a) 2D and (b) 3D cases.

45%, the mean values of the Young's moduli predicted by the 2D and the 3D model are around 7.807 GPa and 8.459 GPa, respectively.

In order to provide further insight into the uncertainty propagation characteristics of the elastic moduli of glass fibre/epoxy composites, Fig. 15 reports a scatter plot of the Monte Carlo samples of the effective elastic modulus for filler volume fractions up to 50%. In this figure, the 95% confidence levels are indicated with red dashed lines. In order to validate the numerical predictions, the experimental characterization data reported by Smith in Ref. [55] for the same composite material investigated herein are included in the figure. It can be seen that the 3D model fits the data slightly better, especially in the high volume fraction region. Interestingly, it is noted that the uncertainty in the Young's modulus increases for larger filler contents (higher that 25%) in the 2D case, while the uncertainty in the elastic modulus when considering 3D RVEs remains almost constant. This fact evidences the limitations related to the use of simplified 2D RVEs to reproduce the randomness in the three-dimensional distribution of particle-reinforced composites doped at moderate to high filler volume fractions.

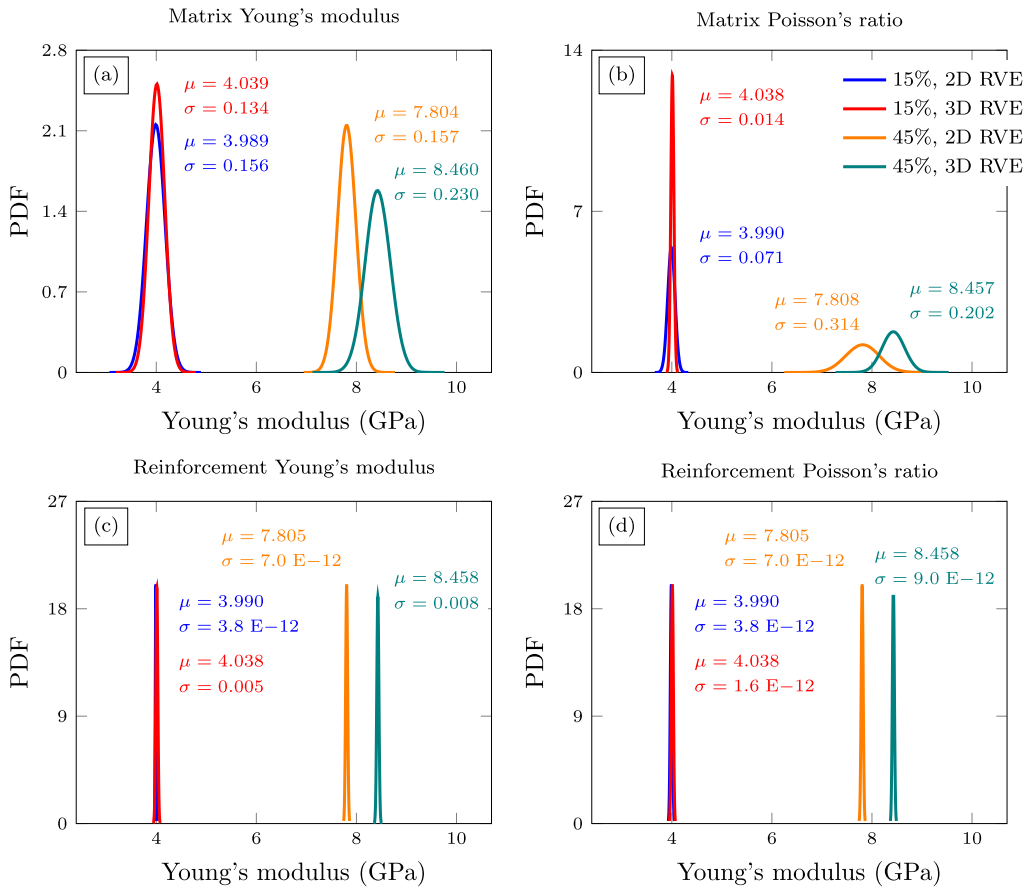

Fig. 13. PDFs of elastic properties of 2D and 3D composites made of epoxy doped with 15% and 45% of glass fibre inclusions considering uncertainties in the matrix Young's modulus (a), in the matrix Poisson's ratio (b), in the glass reinforcement Young's modulus (c) and in the glass reinforcement Poisson's modulus (d).

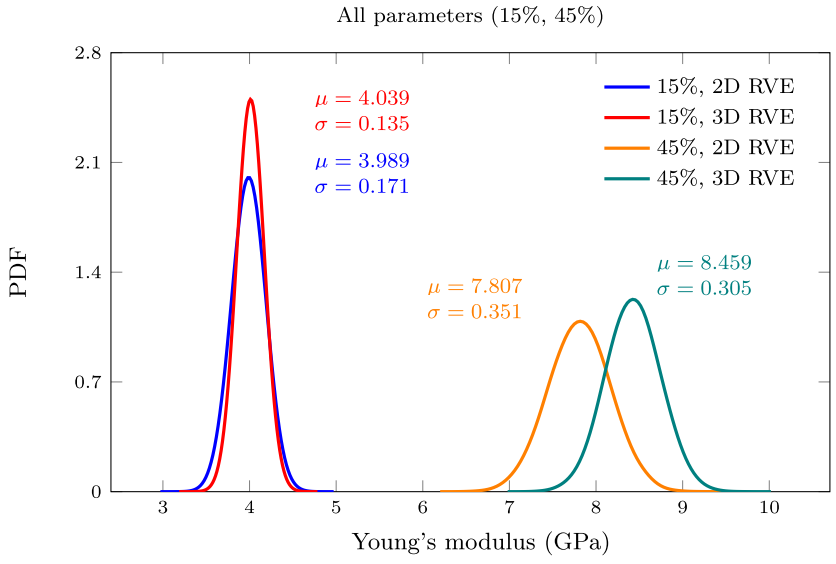

Fig. 14. PDFs of the Young's modulus of epoxy doped with 15% and 45% glass particles considering 2D (blue lines) and 3D (red lines) RVEs for simultaneous uncertainties in all the parameters.

6. Conclusions

This work has investigated the surrogate modelling of 2D/3D highly computationally demanding numerical homogenization problems of glass particles-reinforced epoxy composites. To this aim, PCE meta-modelling technique has been employed to bypass such computational intensive approaches and so perform uncertainty propagation analyses. Specifically, PCE-based uncertainty quantification analyses of the Young's modulus of epoxy/glass composites have been conducted for filler volume fractions up to 50%, using both 2D and 3D models. The statistical distributions of the micromechanical parameters of interest have been predefined, namely the filler volume fraction, Young's

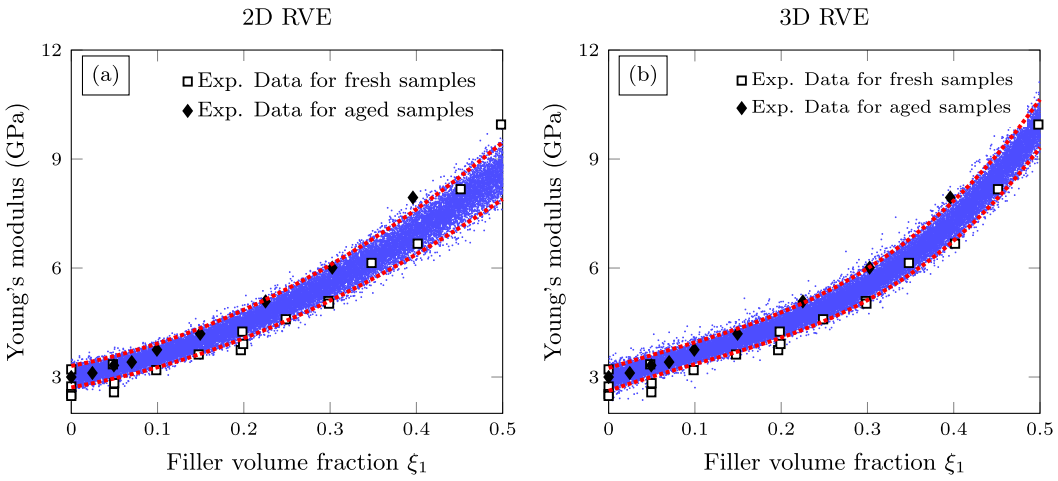

Fig. 15. Comparison between the stochastic prediction of the Young's moduli of epoxy/glass fibre versus the filler volume fraction ξ_1 in the 2D (a) and 3D (b) cases. The 95% confidence levels are indicated with red dashed lines. Experimental data retrieved from Ref. [55].

modulus and the Poisson's ratio of the composite phases. Then, the design space has been sampled by means of Latin Hypercube Sampling in order to construct the PCE surrogate model. Finally, direct Monte Carlo simulations over the parameter space have been performed. The numerical results and discussion have proved the appropriateness of the developed schemes to conduct efficient uncertainty propagation analyses of the homogenization parameters. The key conclusions of this research include:

- Optimal PCE surrogate models have been developed to approximate both 2D/3D homogenization of epoxy composites doped with up to 50% volume fraction of glass filler particles using the dropping and rolling rules. Relationships between the minimal inter-particle distance and the filler volume fraction has been proposed for 2D and 3D RVEs, which represents a useful tool for the algorithm implementation.
- PCE meta-models have been defined through preliminary parametric analyses accounting for prediction accuracy and computational cost. To this aim, a set of local and global error metrics has been presented.
- The computing advantages of the developed meta-models have proved crucial to perform computationally intensive applications. Specifically, the presented results have showed their effectiveness to conduct fast uncertainty propagation analysis of epoxy composites doped with glass particles.
- The numerical results and discussion have evidenced the potential of the proposed approach for stochastic design and material selection applications. These open great opportunities for applications in material model design.
- For low and moderate filler volume fractions, it has been observed that both 2D and 3D models provide similar results, being the dispersion slightly higher when the 2D model is employed. On the other hand, for higher filler volume fractions, the discrepancies between the two models increase.
- In terms of sensitivity analysis, Sobol' indices of both models show that for a fixed volume fraction, uncertainties in the matrix properties dominate the uncertainty of the whole composite. Thus, following the above statement, for lower filler volume fractions, the material properties of the reinforcement can be set to their mean values and the homogenization could be done using the 2D surrogate model without major changes in the final result.

The presented numerical results and discussion suggest the ability of the proposed methodology to be adapted to a large variety of composite materials. In this light, future developments regard the extension of the presented formulation to other composites with more complex microstructures, multi-physics homogenization problems (e.g. piezoelectric materials), as well as the explicit modelling of the variability in the effective properties due to the randomness of the microstructure by developing a stochastic version of the presented surrogate model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work has been partially supported through Ministerio Ciencia e Innovación (Spain) [PID2020-116809GB-I00] and by the Junta de Extremadura (Spain) through Research Group Grants [GR18023]. This work was also supported by the Consejería de Transformación Económica, Conocimiento, Empresas *y* Universidades de la Junta de Andalucía (Spain) [P18-RT-3128].

References

[1] Oladele IO, Omotosho TF, Adediran AA. Polymer-based composites: An indispensable material for present and future applications. Int J Polym Sci 2020.
[2] Raju B, Hiremath SR, Mahapatra DRD. A review of micromechanics based models for effective elastic properties of reinforced polymer matrix composites. Compos Struct 2018;204:607–19.
[3] Hill R. A self-consistent mechanics of composite materials. J Mech Phys 1965;13(4):213–22.
[4] Mori T, Tanaka K. Average stress in matrix and average elastic energy of materials with misfitting inclusions. Acta Metall 1973;21(5):571–4.
[5] Guo Q, Yao W, Li W, Gupta N. Constitutive models for the structural analysis of composite materials for the finite element analysis: A review of recent practices. Compos Struct 2020;113267.
[6] Hashin Z, Shtrikman S. A variational approach to the theory of the elastic behaviour of multiphase materials. J Mech Phys Solids 1963;11:127–40.
[7] Reuss A. Calculation of the flow limits of mixed crystals on the basis of the plasticity of mono-crystals. Z Angew Math Mech 1929;9:49–58.
[8] Voigt W. Ueber die Beziehung zwischeden beiden Elasticitatsconstanten. Ann Phys 1889;38:573–87.
[9] Clayton JD, Chung PW. An atomistic-to-continuum framework for nonlinear crystal mechanics based on asymptotic homogenization. J Mech Phys Solids 2006;54(8):1604–39.

[10] Pace T, Rahmaninejad H, Sun B, Kekenes-Huskey PM. Homogenization of continuum-scale transport properties from molecular dynamics simulations: An application to aqueous-phase methane diffusion in silicate channels. J Phys Chem B 2021;125(41):11520–33.

[11] Artioli E, Beirã da Veiga L, Verani M. An adaptive curved virtual element method for the statistical homogenization of random fibre-reinforced composites. Finite Elem Anal Des 2020;177:103418.

[12] Pingaro M, Reccia E, Trovalusci P, Masiani R. Fast statistical homogenization procedure (FSHP) for particle random composites using virtual element method. Comput Mech 2019;64(1):197–210.

[13] Marino Michele, Hudobivnik Blaž, Wriggers Peter. Computational homogenization of polycrystalline materials with the virtual element method. Comput Methods Appl Mech Engrg 2019;355:349–72.

[14] To QD, Bonnet G. Fft based numerical homogenization method for porous conductive materials. Comput Methods Appl Mech Engrg 2020;368:113160.

[15] Ahmadi MH. A stochastic finite element analysis framework for the multiple physical modeling of filler modified polymers. (Ph.D. thesis), University of Alberta; 2021.

[16] Reddy AC. Two dimensional (2D) RVE-based modeling of interphase separation and particle fracture in graphite/5050 particle reinforced composites. In: 3rd national conference on materials and manufacturing processes, Hyderabad, India. 2002, p. 179–83.

[17] Hill R. Elastic properties of reinforced solids: some theoretical principles. J Mech Phys Solids 1963;11:357–72.

[18] Ostoja-Starzewski M. Material spatial randomness: From statistical to representative volume elemen. Probab Eng Mech 2006;21:112–32.

[19] García-Macías Enrique, Ubertini Filippo. Mathematical modeling and simulation. Smart Nanoconcr Cement-Based Mater 2020;101–56.

[20] Stroven M, Askes H, Sluys LJ. Numerical determination of representative volumes for granular materials. Comput Mehtods Appl Eng 2004;193:3221–38.

[21] Naili C, Doghri I, Kanit T, Sukiman MS, Aissa-Berraies A, Imad A. Short fiber reinforced composites: Unbiased full-field evaluation of various homogenization methods in elasticity. Compos Sci Technol 2020;187:107942.

[22] Teradaa K, Horib M, Kyoyac T, Kikuchid N. Simulation of the multi-scale convergence in computational homogenization approaches. Int J Solids Struct 2000;37:2285–311.

[23] Kanit T, Forest S, Galliet I, Mounoury V, Jeulin D. Determination of the size of the representative volume element for random composites: statistical and numerical approach. Int J Solids Struct 2003;40:3647–79.

[24] Gentieu T, Catapano A, Jumel J, Broughton J. Computational modelling of particulate-reinforced materials up to high volume fractions: Linear elastic homogenisation. Proc Inst Mech Eng 2019;233(6):1101–16.

[25] Ma S, Zhuang X, Wang X. 3D micromechanical simulation of the mechanical behavior of an in-situ Al3Ti/A356 composite. Compos B 2019;176:107115.

[26] Nilenius F, Larsson F, Lundgren K, Runesson K. A 3d/2d comparison between heterogeneous mesoscale models of concrete. multi-scale modeling and characterization of infrastructure material. Dordrecht: Springer; 2013, p. 249–59.

[27] Huang Y, Yan D, Yang Z, Liu G. 2D and 3D homogenization and fracture analysis of concrete based on in-situ X-ray computed tomography images and Monte Carlo simulations. Eng Fract Mech 2016;163:37–54.

[28] Zeleniakiene D, Griškevičiu P, Leišis V. The comparative analysis of 2D and 3D microstructural models stresses of porous polymer materials. Mechanics 2005;53(3):22–6.

[29] Kristiansen L, Alan Wouterse K, Philipse A. Simulation of random packing of binary sphere mixtures by mechanical contraction. Phys A 2005;(2–4):249–62.

[30] Lubachevsky BD, Stillinger FH, Pinson EN. Disks vs, spheres: Contrasting properties of random packings. J Stat Phys 1991;64(3):501–24.

[31] Rintoul MD, Torquato S. Reconstruction of the structure of dispersions. J Colloid Interface Sci 1997;186:467–76.

[32] Segurado J, Llorca J. A numerical approximation to the elastic properties of sphere-reinforced composites. J Mech Phys 2002;50(10):2107–21.

[33] Shi Y, Zhang Y. Simulation of random packing of spherical particles with different size distributions. Appl Phys A 2008;92(3):621–6.

[34] Jodrey WS, Elmer MT. Simulation of random packing of spheres. Simulation 1979;32(1):1–12.

[35] Cundall PA, Otto DLS. A discrete numerical model for granular assemblies. Geotechnique 1979;29(1):47–65.

[36] Thapa M, Mulani SB, Walters RW. Stochastic multi-scale modeling of carbon fiber reinforced composites with polynomial chaos. Compos Struct 2019;213:82–97.

[37] Mohamedou M, Zulueta K, Chung CN, Rappel H, Beex L, Adam L, Noels L. Bayesian identification of mean-field homogenisation model parameters and uncertain matrix behavior in non-aligned short fiber composites. Compos Struct 2019;220:64–80.

[38] Dey S, Mukhopadhyay T, Khodaparast HH, Adhikari S. Fuzzy uncertainty propagation in composites using Gram–Schmidt polynomial chaos expansion. Appl Math Model 2016;407(8):4412–28.

[39] Peng X, Li D, Wu H, Liu Z, Li J, Jiang S, Tan J. Uncertainty analysis of composite laminated plate with data-driven polynomial chaos expansion method under insufficient input data of uncertain parameters. Compos Struct 2019;209:625–33.

[40] Østergård T, Jensen RL, Maagaard SE. A comparison of six metamodeling techniques applied to building performance simulations. Appl Energy 2018;211:89–103.

[41] Sun X, Choi YY, Choi JI. Global sensitivity analysis for multivariate outputs using polynomial chaos based surrogate models. Appl Math Model 2020;82:867–87.

[42] Xiu D, Karniadakis GE. Modeling uncertainty in flow simulations via generalized polynomial chaos. J Comput Phys 2003;1871:137–67.

[43] Smith RC. Uncertainty quantification. 1st ed.. Philadelphia: SIAM; 2014.

[44] B. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32(2):407–99.

[45] Henyš Petr, Čapek Lukáš, Březina Jan. Comparison of current methods for implementing periodic boundary conditions in multi-scale homogenisation. Eur J Mech A Solids 2019;78:103825.

[46] Zhou J, Zhang Y, Chen JK. Numerical simulation of random packing of spherical particles for powder-based additive manufacturing. J Manuf Sci Eng Trans ASME 2009;131(3):031004.

[47] Berger H, Kari S, Gabbert U, Rodríguez-Ramos R, Bravo-Castillero J, Guinovart-Díaz RA. Comprehensive numerical homogenisation technique for calculating effective coefficients of uniaxial piezoelectric fibre composites. Mater Sci Eng A 2005;412(1–2):53–60.

[48] Sudret B. Polynomial chaos expansions and stochastic finite element methods. Risk Reliab Geotech Eng 2014;624:265–300.

[49] Blatman G, Sudret B. Adaptive sparse polynomial chaos expansion based on least angle regression. J Comput Phys 2011;230:2345–67.

[50] MATLAB and statistics toolbox release. Natick, Massachusetts, United States: The MathWorks, Inc.; 2012b.

[51] Ansys® academic research mechanical. Release 18.1, help system, mechanical APDL guide. ANSYS, Inc..

[52] Ju J, Yanase K. Micromechanics and effective elastic moduli of particle-reinforced composites with near-field particle interactions. Acta Mech 2010;215(1):135–53.

[53] Schjødt-Thomsen J, Ryszard P. The Mori–Tanaka stiffness tensor: diagonal symmetry, complex fibre orientations and non-dilute volume fractions. Mech Mater 2001;33(10):531–44.

[54] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 2000;42(1):55–61.

[55] Smith JC. Experimental values for the elastic constants of a particulate-filled glassy polymer. J Res Natl Bur Stand (U. S.) 1976;80(1):45–9.

Appendix B

Paper B: Multielement polynomial chaos Kriging-based metamodelling for Bayesian inference of non-smooth systems

Applied Mathematical Modelling 116 (2023) 510–531

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

Multielement polynomial chaos Kriging-based metamodelling for Bayesian inference of non-smooth systems

J.C. García-Merino[a,*], C. Calvo-Jurado[a], E. Martínez-Pañeda[b], E. García-Macías[b,c]

[a] Department of Mathematics, School of Technology, Cáceres, 10003 Spain
[b] Department of Civil and Environmental Engineering, Imperial College, London SW7 2AZ, UK
[c] Department of Structural Mechanics and Hydraulic Engineering, Campus Universitario de Fuentenueva (Edificio Politécnico) 18071 Granada Spain

ARTICLE INFO

Article history:
Received 12 March 2022
Revised 28 November 2022
Accepted 30 November 2022
Available online 5 December 2022

Keywords:
Bayesian inference
Model calibration
Surrogate modelling
Hydrogen embrittlement
Thermal desorption spectroscopy
Polynomial chaos expansion Kriging

ABSTRACT

This paper presents a surrogate modelling technique based on domain partitioning for Bayesian parameter inference of highly nonlinear engineering models. In order to alleviate the computational burden typically involved in Bayesian inference applications, a multielement Polynomial Chaos Expansion based Kriging metamodel is proposed. The developed surrogate model combines in a piecewise function an array of local Polynomial Chaos based Kriging metamodels constructed on a finite set of non-overlapping subdomains of the stochastic input space. Therewith, the presence of non-smoothness in the response of the forward model (e.g. nonlinearities and sparseness) can be reproduced by the proposed metamodel with minimum computational costs owing to its local adaptation capabilities. The model parameter inference is conducted through a Markov chain Monte Carlo approach comprising adaptive exploration and delayed rejection. The efficiency and accuracy of the proposed approach are validated through two case studies, including an analytical benchmark and a numerical case study. The latter relates the partial differential equation governing the hydrogen diffusion phenomenon of metallic materials in Thermal Desorption Spectroscopy tests.

1. Introduction

The widespread use of high performance computing (HPC) technologies has enabled the increasingly frequent adoption of computationally intensive numerical models in a myriad of disciplines in academia and industry. Such high fidelity models allow conducting virtual testing of engineering systems, avoiding the technical limitations and minimizing the costs associated with traditional experimental testing. Nevertheless, formidable challenges still arise when implementing these models into computationally demanding studies such as optimization [1], sensitivity analysis [2], model identification and calibration [3], reliability analysis [4], or robust design [5]. As a solution, a variety of surrogate models or metamodels have been proposed in the literature in recent years. Nevertheless, despite the considerable advances in the field, there remain open research challenges for their extensive use such as handling highly nonlinear model responses, limited training datasets and,

* Corresponding author.
E-mail address: jcgarcia@unex.es (J.C. García-Merino).

https://doi.org/10.1016/j.apm.2022.11.039

in general, the online implementation of surrogate models to enable decision-making in engineering systems [6,7]. In particular, the development of real time surrogate model-based parameter estimation approaches draws high interest in frontier research fields such as digital twins development [8] and smart maintenance of engineering systems [9].

In their broadest sense, surrogate models are computationally light black box representations of resource-intensive models. Surrogate modelling methods can be generally classified into three categories [10]: (i) projection-based methods or reduced-order models (ROMs) [11]; (ii) multi-fidelity methods [12]; and (iii) data-driven or responsive surface methods (RSMs). Projection-based techniques project the governing equations of the original model onto a low-dimensional subspace. Hence, although ROMs have the advantage of retaining the physics underlying the model, these are limited to situations where access to the governing equations is granted, which is not the case in many practical applications when using commercial software. Multi-fidelity methods are built by simplifying the underlying physics or reducing the numerical resolution. These methods may also render some difficulties, being highly case-dependent and requiring specialized expertise to find a suitable trade-off between prediction accuracy and computational burden. Such difficulties have fostered rapid developments of RSMs in recent years as a non-intrusive technique with great flexibility in a wide range of applications. The key advantage of these methods relies in the fact that the forward model does not need to be modified and, therefore, it can be essentially treated as a black box [13]. Among the broad variety of RSMs available in the literature, some of the most popular ones are Kriging [14], radial basis functions (RBF) [15], support vector regression (SVR) [16], artificial neural networks (ANN) [17], Gaussian process (GP) regression [18], polynomial chaos expansions (PCE) [2] and Polynomial Chaos Expansion based Kriging (PCK) [19]. These models are trained by exploiting a set of realizations of parameters of interest of the model, called the experimental design (ED), and the corresponding model evaluations, also called quantities of interest (QoI). Therefore, the accuracy of data-driven surrogate models is highly determined by the dimensions of the design space and the number and distribution of the training samples in the ED [20]. In engineering practice, obtaining the ED constitutes the most time-consuming part since it requires the evaluation of the computationally intensive forward model at each sample point. Choosing high quality EDs is thus critical to achieve high accuracy in the surrogate model construction with the least possible number of training samples. Sampling approaches can generally be divided into static (one-shot) and sequential methods. One-shot sampling generates the training sample points in one single step, and common approaches include fractional designs and orthogonal arrays [21]. While these techniques offer easy implementation and minimal computational cost, the determination of the optimal sample size may be troublesome when the behaviour of the forward model is unknown. To minimize such difficulties, a number of sequential sampling strategies have been introduced, including adaptive and space-filling sequential methods [22]. On one hand, space-filling sequential designs such as Latin Hypercube Sampling (LHS), Sobol, Hammersley and Halton sampling [23] generate samples iteratively to attain good coverage of the parametric domain. On the other hand, adaptive sampling techniques iteratively draw new samples in regions of the parametric space with large prediction errors, enabling to account for local refinements in the ED (refer to references [22,24] for a thorough state-of-the-art review).

A second major challenge of non-intrusive surrogate models regards the difficulties involved in the fitting of non-smooth models exhibiting unsteadiness, sparseness or large perturbations. In these cases, adaptive solutions accommodating local relevant refinements of the response surface are required. A large volume of research has been conducted in the last decade to address this issue, giving origin to a number of advanced surrogate modelling techniques such as multi-resolution generalized PCE [25], domain partitioning [26], sparse grid collocation [27], Voronoi tesselations [28], local search at trust regions [29], clustering-based partitioning [30], ensembles of surrogate models [31], multi-element generalized polynomial chaos (ME-gPC) method [27] and multi-element probabilistic collocation methods (ME-PCM) [32]. In this light, different partitioning techniques can be found in the literature in the context of the pure GPs [33,34], PCE [35], and PCE-based mapping of likelihood functions in parameter inference applications [25,36]. In those works, different domain partition criteria were proposed based on dissimilarity measurements between regions [34], data density [33], or maximum residual differences [35,36]. In general, these approaches usually generate series of subdomains where the non-smooth response surface can be assumed locally smooth, thus enabling the definition of local surrogate models contributing to the global response in a piecewise fashion.

The development of cost-efficient surrogate models opens vast new opportunities for real-time parameter estimation applications. Probabilistic Bayesian approaches are particularly attractive owing to their ability to assess the effects of uncertainties on the model parameters and the derived response predictions, as well as their robustness to noise pollution and efficiency to deal with ill-conditioning and ill-posedness. Such excellent features have fostered their implementation in multiple fields such as structural identification [37], geotechnical problems [38], material characterization [39], and bioengineering [40], just to mention a few. In general, Bayesian parameter estimation approaches exploit experimental data to infer the posterior probability distribution functions (PDFs) of certain unknown model parameters through the Bayes' theorem. Nonetheless, the direct evaluation of the posterior PDFs requires solving the possibly high-dimensional integral related to the evidence of the model. Therefore, except for some trivial cases, posterior PDFs often need to be approximated numerically. Markov chain methods constitute the most widespread set of techniques to extract series of samples to estimate the posterior PDFs, allowing to sample from a large class of high-dimensional distributions. Popular procedures for Markov chain Monte Carlo (MCMC) sampling are the Metropolis-Hastings [41] and Gibbs algorithms [42]. The basic idea of these techniques is to construct a Markov chain with a stationary distribution resembling the posterior distribution, in such a way that a sample of the joint PDF of the model parameters can be obtained by collecting the states of the chain. These methods guarantee asymptotic convergence to the exact PDF, although a considerably large number of iterations are

typically required to achieve convergence, which compromises the computational efficiency of the inference [43]. To alleviate the computational burden in classical MCMC techniques, a variety of more efficient sampling algorithms have been proposed in recent years, including Sequential Monte Carlo (SMC) [44], Transitional MCMC [45], and Bayesian broad learning [46]. Notwithstanding these advances, their elevated computational cost remains a critical limitation when high-fidelity models are considered in the inference problem. Herein is where surrogate models offer an efficient solution to conduct cost-efficient Bayesian inference while retaining the accuracy of high-fidelity models. The enormous potentials of this approach are evidenced by the increasing number of research studies reported in recent years. It is worth noting the work by Schneider et al. [47] who proposed a Bayesian procedure using rational PCE metamodels of the response of dynamic systems in the frequency domain, and demonstrated its effectiveness for the identification of a cross-laminated timber plate. Xing et al. [48] developed an additive GP model for multi-fidelity surrogate modelling inserted into a non-parametric Bayesian approach with a closed-form solution for the predictive posterior PDF. The accuracy and flexibility of their approach were validated with several benchmark problems, including the identification of a solid oxide fuel cell model and an elbow-shape pipe under turbulent mixing flow conditions. Ierimonti and co-authors [49] proposed a Kriging-based conjugate Bayesian identification methodology for online damage identification of an instrumented monumental building, the Consoli Palace in Gubbio (Italy). Del Val et al. [50] implemented an MCMC approach to infer the catalytic recombination parameters of reusable thermal protection materials exploiting plasma wind tunnel measurements. Interestingly, instead of bypassing a high-fidelity model, those authors approximated the likelihood function in the Bayesian inference using a GP surrogate model to accelerate the parameters estimation.

In light of the literature review above, it is apparent that the fields of surrogate modelling and its applications for fast parameter estimation have experienced considerable advances in recent years. Nonetheless, light metamodels capable of bypassing highly nonlinear models for online parameter estimation are yet to be fully developed. In this context, the present work presents a novel multielement surrogate model extending the PCK model proposed in Schobi et al. [19] for online Bayesian parameter inference of non-smooth models exhibiting highly nonlinear spatial variability. To do so, a simple but efficient and flexible regular block partitioning approach is adopted. On this basis, the space domain is partitioned into distinct subregions with splitting directions defined after a preliminary sensitivity analysis. The algorithm prioritizes the partition of those variables providing highest sensitivities in terms of the Sobol indices, which can be readily extracted from the coefficients of a global PCE [51]. Regarding the construction of the local PCK surrogate models, the optimal order of the polynomials in the PCE is automatically identified by a model selection technique for sparse linear models, the least-angle regression (LAR) algorithm set out by Efron et al. [52]. Then, the optimal PCE is inserted into a Kriging predictor as the trend term, while the stochastic term is fitted through a genetic algorithm (GA) global optimization approach. The global model response is obtained by combining the local surrogate models in a piecewise fashion. In order to minimize the computational burden in the construction of the surrogate models, the ED is obtained with the adaptive Monte Carlo-Intersite-proj-th (MIPT) approach [22]. Finally, the surrogate model is used for Bayesian parameter estimation using a cost-efficient adaptive MCMC with delayed rejection (DRAM) algorithm developed by Haario et al. [53]. The effectiveness of the proposed approach is validated through two benchmark case studies: (i) an analytical benchmark; (ii) and a partial differential equation (PDE) for thermal desorption spectroscopy (TDS) experiments of hydrogen in metals. The latter represents a formidable example of a model exhibiting a non-smooth behaviour in the shape of nonlinearities and unsteadiness. The presented results and discussion demonstrate the robustness of the proposed approach to the presence of measurement noise and multimodality in the posterior distributions. Overall, the developed surrogate model assisted Bayesian inference approach represents a comprehensive solution for fast parameter inference of computationally intensive mathematical models, offering wide applicability across disciplines such as the identification of structural systems, geotechnical problems, Bioengineering, or Astrophysics, among others.

The remainder of this paper is organized as follows. Section 2 outlines the theoretical formulation of the proposed approach. Section 3 presents the numerical results and discussion. In particular, two case studies are investigated, namely a benchmark analytical model and a numerical model of the TDS analysis of hydrogen desorption in metals. Finally, Section 4 discusses the contributions of this work and presents the main concluding remarks.

2. Theoretical formulation

The main purpose of this section is to present the theoretical fundamentals of the proposed surrogate model-based Bayesian inference approach. The general methodology is sketched in Fig. 1 and comprises four main steps, namely (i) domain partitioning; (ii) construction of the local surrogate models; (iii) surrogate models assemblage; and (iv) validation.

2.1. Surrogate modelling: polynomial chaos expansion based Kriging

Let (Ω, Σ, μ) be a probability space, with $\Omega \subset \mathbb{R}^M$ denoting the event space equipped with a σ-algebra Σ of subsets of Ω and a probability measure μ such that $\mu(\Omega) = 1$. Let $\mathcal{M} : \Omega \subset \mathbb{R}^M \rightarrow \mathbb{R}$ be a computational model mapping a vector of input variables $\mathbf{x} = [x_1, \ldots, x_M]^{\mathrm{T}}$ into the output variable or quantity of interest $y \in \mathbb{R}$. The goal of surrogate modelling is to approximate the forward model $\mathcal{M}$, which is typically computationally intensive, by a computationally inexpensive function $\hat{\mathcal{M}}$. In this paper, PCK metamodels combining PCE and Kriging are adopted. In this light, PCE is used to approximate the global behaviour of the computational model $\mathcal{M}$ while the Kriging metamodel captures its local behaviour.

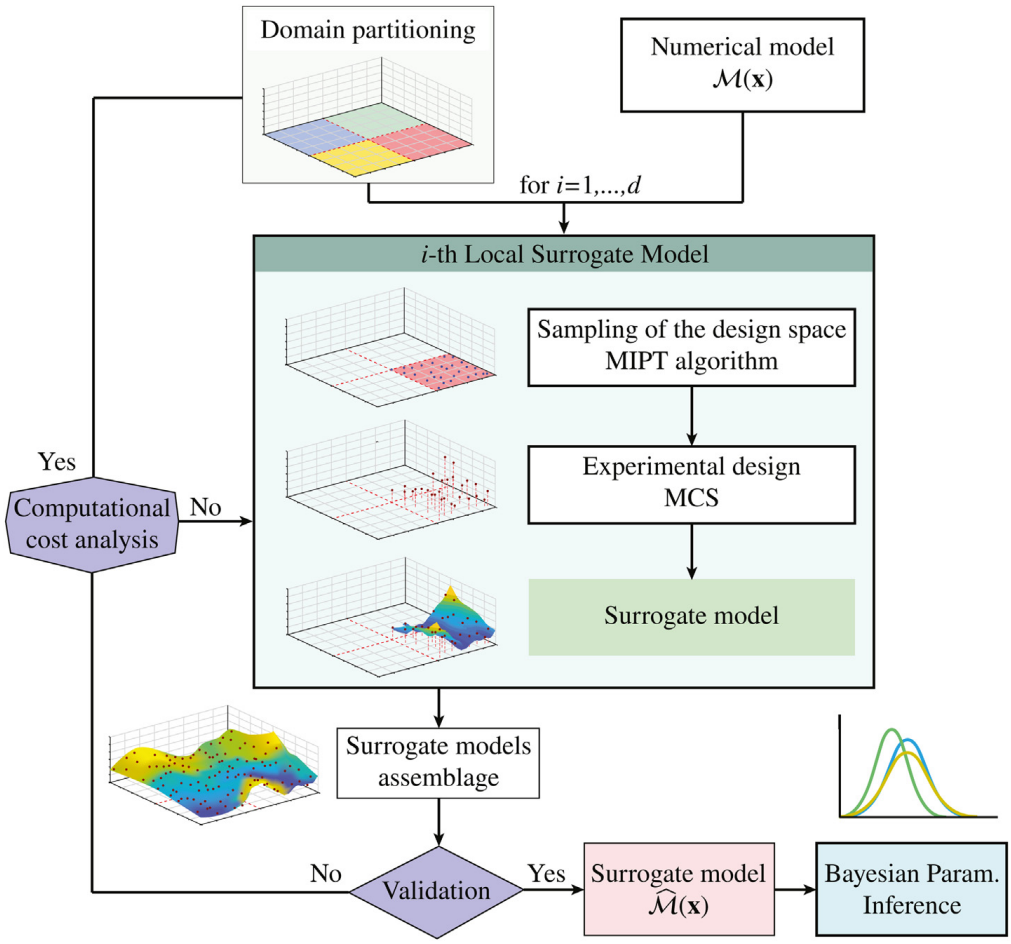

Fig. 1. Flowchart of the proposed PCK metamodel to perform Bayesian Inference.

2.1.1. Polynomial chaos expansion

Assume that the input vector $\mathbf{x} \in \Omega$ is constituted by M independent random variables components $\{x_i\}_{i=1}^{M}$ with probability density functions μ_{X_i}. Then, PCE represents the output response $y \in \mathbb{R}$ as the infinite expansion of $\mathcal{M}$ over an orthonormal basis of multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$ as:

$$y = \mathcal{M}(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}), \tag{1}$$

where $a_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$, $\alpha_i \in \mathbb{N}$ are the coefficients of the expansion. Orthonormal basis families associated with a variety of standard distributions can be found in reference [54]. Multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$ can be expressed in terms of a family of univariate polynomials $\left\{\psi_j^{(i)},\ j \in \mathbb{N}\right\}$ as $\Psi_{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^{M} \psi_{\alpha_i}^{(i)}(x_i)$. Polynomials $\psi_j^{(i)}$ are also orthonormal with respect to the marginal distribution, that is:

$$\mathbb{E}\left[\psi_j^{(i)}(x_i),\ \psi_k^{(i)}(x_i)\right] = \int \psi_j^{(i)}(u)\, \psi_k^{(i)}(u)\, \mu_{x_i}(u)\, \mathrm{d}u = \delta_{j,k}, \tag{2}$$

with $\delta_{j,k}$ being the Dirac Delta function. Random variable $\mathbf{x}$ induces the probability measure $\mu_{\mathbf{x}}$ on $\left(\mathbb{R}^M, \mathcal{B}\left(\mathbb{R}^M\right)\right)$, with $\mathcal{B}(\mathbb{R})$ denoting the Borel σ-algebra $\mu_{\mathbf{x}}(B) = \mu\{\omega \in \Omega : \mathbf{x}(\omega) \in B\}$, $B \in \mathcal{B}$. The PDF of $\mathbf{x}$ is given by the probability measure $\mu_{\mathbf{x}}$ as $\mu_{\mathbf{x}}(\mathbf{u}) = \mu(\mathbf{x} \leq \mathbf{u})$, $\mathbf{u} \in \mathbb{R}^M$. Note that, since the components of $\mathbf{x}$ are independent, the joint probability function $\mu_{\mathbf{x}}(\mathbf{u})$ is given by the product of the marginal PDFs of x_i, i.e. $\mu_{\mathbf{x}}(\mathbf{u}) = \prod_{i=1}^{M} \mu_{x_i}(u_i)$, $\mathbf{u} = (u_1, \ldots, u_M) \in \mathbb{R}^M$. Therefore, Eq. (2) can be written in a more compact form as:

$$\mathbb{E}\left[\Psi_{\boldsymbol{\alpha}}(\mathbf{x}),\ \Psi_{\boldsymbol{\beta}}(\mathbf{x})\right] = \int \Psi_{\boldsymbol{\alpha}}(\mathbf{u}) \Psi_{\boldsymbol{\alpha}}(\mathbf{u})\, \mu_{\mathbf{x}}(\mathbf{u})\ \mathrm{d}\mathbf{u} = \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \quad \boldsymbol{\alpha},\ \boldsymbol{\beta} \in \mathbb{N}^M. \tag{3}$$

Although the expression in Eq. (1) is exact for an infinite number of terms, in practice only a finite number can be computed and a certain truncation scheme needs to be adopted. One of the simplest approach consists in selecting all the

polynomials whose total degree $|\boldsymbol{\alpha}| = \sum_{i=1}^{M} \alpha_i$ belongs to the set:

$$\mathcal{A}^{M,p} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^M : 0 \leq |\boldsymbol{\alpha}| \leq p\right\}, \quad \text{with} \quad \operatorname{card} \mathcal{A}^{M,p} = \binom{M+p}{p} = \frac{(M+p)!}{M!\,p!}. \tag{4}$$

For high-dimensional and non-linear problems, this truncation procedure usually leads to large numbers of polynomial coefficients and considerable computational burdens. Nevertheless, it is often observed in many practical applications that coefficients corresponding to high interaction terms between the input variables are close to zero, a phenomenon that is also known as the *sparsity-of-effect principle* [51]. To alleviate this, an hyperbolic truncation scheme can be adopted. This approach selects all multi-indices with q-norm $\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{\frac{1}{q}}$ less than or equal to a certain model order p, i.e. $\mathcal{A}^{M,p,q} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^M : \|\boldsymbol{\alpha}\|_q \leq p\right\}$. Note that the expansion tends to only maintain univariate polynomials as the q-norm decreases, thus achieving a reduction in the computational cost of the metamodel. Nonetheless, the number of terms may remain elevated since the potential sparseness in the coefficients is not being actually assessed. Following the work by Blatman and Sudret [51], the LAR algorithm [52] is adopted to further reduce the number of polynomial coefficients. In the context of PCE, LAR constructs a set of expansions incorporating an increasing number of basis polynomials $\Psi_{\boldsymbol{\alpha}}$, from 1 to $\mathcal{P} = \operatorname{card}\left(\mathcal{A}^{M,p,q}\right)$. The resulting sequence of index sets is used to construct a family of expansions with decreasing sparseness. Finally, a cross-validation procedure can be implemented to select the best metamodel among the obtained family of expansions. In particular, the Bayesian Information Criterion (BIC) is adopted in this work. Once the optimal index set is selected, the expansion coefficients $\boldsymbol{a} = \{\boldsymbol{a}_\alpha, \boldsymbol{\alpha} \in \mathcal{A}^{M,p} \subset \mathbb{N}^M\}$ are obtained by minimizing the expectation of the least squared error:

$$\boldsymbol{a} = \arg\min_{\boldsymbol{a} \in \mathbb{R}^{\mathrm{P}}} \mathbb{E}\left[\left(\mathcal{M}(\mathbf{x}) - \sum_{\boldsymbol{\alpha} \in \mathcal{A}^{M,p,q}} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x})\right)^2\right]. \tag{5}$$

In practice, Eq. (5) is calibrated on a set of N realizations $\Xi = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\}$ of the input variable $\mathbf{x}$ forming the ED. In order to obtain a representative ED, the adaptive MIPT sampling method [22] is adopted in this work. Then, the expectation operator in Eq. (5) is replaced by its discretized version using the ED realizations as follows:

$$\boldsymbol{a} = \arg\min_{\boldsymbol{a} \in \mathbb{R}^{\mathrm{P}}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{M}\left(\mathbf{x}^{(i)}\right) - \sum_{\boldsymbol{\alpha} \in \mathcal{A}^{M,p,q}} \boldsymbol{a}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}\left(\mathbf{x}^{(i)}\right)\right)^2. \tag{6}$$

Denoting the realizations of the output variable y by $\mathbf{y} = \{y^{(1)} = \mathcal{M}(\mathbf{x}^{(1)}), \ldots, y^{(N)} = \mathcal{M}(\mathbf{x}^{(N)})\}^{\mathrm{T}}$, the solution of the optimization problem in Eq. (6) reads:

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{y}, \qquad \boldsymbol{\Theta} = (\boldsymbol{\Theta}_{ij}) = \left[\psi_j(\boldsymbol{x}^{(i)})\right]_{i=1,\ldots,N}^{j=1,\ldots,P}, \tag{7}$$

where $\boldsymbol{\Theta}$ denotes the information matrix calculated from the evaluation of the basis polynomials on Ξ. For the least-square minimization problem in Eq. (6) to be well posed, the size of the ED is usually selected according to the heuristic rule $N \approx 2 \cdot P$ or $3 \cdot P$ [51]. Once Eq. (6) has been solved, the predictions of the PCE surrogate model can be obtained as:

$$\hat{y} = \hat{\mathcal{M}}_{PCE}(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \hat{\boldsymbol{a}}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}). \tag{8}$$

2.1.2. Polynomial chaos expansion based Kriging (PCK)

The Kriging method assumes that the response of a computational model $\mathcal{M}(\mathbf{x})$ is modelled by the sum of a stochastic random process $\mathcal{Z}(\mathbf{x})$ and a regression model $\mathcal{T}(\mathbf{x})$, also called trend, in the form Fuhg et al. [22]:

$$\hat{\mathcal{M}}(\mathbf{x}) = \mathcal{T}(\mathbf{x}) + \mathcal{Z}(\mathbf{x}). \tag{9}$$

The stochastic component in Eq. (9) is fully determined by the covariance function [55]:

$$\operatorname{Cov}\left(\mathcal{Z}(\mathbf{x}), \mathcal{Z}\left(\mathbf{x}'\right)\right) = \mathbb{E}\left[\mathcal{Z}(\mathbf{x})\mathcal{Z}(\mathbf{x}')\right] = \sigma^2 R\left(\left|\mathbf{x} - \mathbf{x}'\right|; \boldsymbol{\theta}\right), \tag{10}$$

with σ^2 being the process variance, and $R\left(\left|\mathbf{x} - \mathbf{x}'\right|; \boldsymbol{\theta}\right)$ an auto-correlation function [56] between two input sample points $\mathbf{x}$ and $\mathbf{x}$' that depends on certain hyper-parameters $\boldsymbol{\theta}$ to be computed. In this work, the Gaussian correlation function is adopted as:

$$R\left(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}\right) = \prod_{\ell=1}^{M} \exp\left[-\theta_\ell \left(x_\ell - x'_\ell\right)^2\right]. \tag{11}$$

The trend term of the Kriging model in Eq. (9) interpolates the forward model evaluations at the ED, while the local variability is captured by the stochastic process. Depending on the form of the trend, three different versions of Kriging are typically referred to in the literature [22], including simple, ordinary and universal Kriging, which respectively correspond to polynomials of degrees 0, 1 and N. In this work, with the aim of combining the excellent global approximation capabilities of the PCE previously introduced in Section 2.1.1, the sparse PC expansion obtained by LAR is introduced in the shape of the trend term in Eq. (9). The resulting PCK metamodel reads:

$$\hat{\mathcal{M}}_{PCK}(\mathbf{x}) = \hat{\mathcal{M}}_{PCE}(\mathbf{x}) + \mathcal{Z}(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \hat{\boldsymbol{a}}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}) + \mathcal{Z}(\mathbf{x}). \tag{12}$$

The construction of the PCK metamodel in Eq. (12) consists in two steps. Firstly, the optimal set of orthonormal polynomials $\boldsymbol{\Psi}_{\boldsymbol{\alpha}}$ (for $\boldsymbol{\alpha}\in\mathcal{A}$ the truncation set) is obtained by LAR as indicated in Section 2.1.1. Secondly, the calculation of hyperparameters $\hat{\boldsymbol{\theta}}$ and the polynomial coefficients and the process variance $\{\boldsymbol{a}(\hat{\boldsymbol{\theta}}), \sigma^2(\hat{\boldsymbol{\theta}})\}$ are obtained. The optimal correlation parameters $\hat{\boldsymbol{\theta}}$ can be determined by the Maximum-Likelihood-Estimation (ML) through the following minimization problem [57]:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{N} (\mathbf{y} - \boldsymbol{\Theta}\boldsymbol{a})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \boldsymbol{\Theta}\boldsymbol{a}) (\det \mathbf{R})^{1/N} \right]. \tag{13}$$

In order to solve the optimization problem in Eq. (13), local optimization algorithms such as gradient-based methods are often used. Nonetheless, a major drawback of these techniques relates the troublesome identification of global maxima/minima, being possible to get stuck in local maxima/minima. To avoid this, a global genetic algorithm optimization procedure is used in this work. Since the correlation matrix is symmetric and positive definite, its inverse in Eq. (13) is computed by Cholesky decomposition. Then, once $\hat{\boldsymbol{\theta}}$ is computed, the polynomial coefficients and the process variance $\{\boldsymbol{a}(\hat{\boldsymbol{\theta}}), \sigma^2(\hat{\boldsymbol{\theta}})\}$ are calculated using the Empirical Best Linear Unbiased Estimator (BLUE) as [58]:

$$\boldsymbol{a}\left(\hat{\boldsymbol{\theta}}\right) = \left(\boldsymbol{\Theta}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{\Theta}\right)^{-1} \boldsymbol{\Theta}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{y}, \quad \sigma^2\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{N} (\mathbf{y} - \boldsymbol{\Theta}\boldsymbol{a})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \boldsymbol{\Theta}\boldsymbol{a}), \tag{14}$$

where $R_{ij} = R\left(\left|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right|; \hat{\boldsymbol{\theta}}\right)$ is the correlation matrix and $\boldsymbol{\Theta}_{ij} = \psi_j\left(\mathbf{x}^{(i)}\right)$ the information matrix evaluated at all the samples of the ED.

Effective explorative sampling

With the aim of generating representative EDs, the MIPT algorithm is adopted as a computationally efficient and easily implementable adaptive sampling technique. The main advantage of this technique compared to space-filling techniques such as LHS regards its ability to avoid local clustering of points which may consequently lead to numerical instabilities in the inverse of the Kriging correlation matrix in Eq. (14). This exploration distance-based sampling method iteratively augments the ED by adding new sampling points with maximum distance with respect to the data population in the ED among a large set of N_c random Monte-Carlo candidates. Specifically, among the candidates set $\mathcal{C} = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N_c)}\}$, a new sample $\mathbf{x}^{(N+1)}$ is chosen by solving the following optimization problem:

$$\mathbf{x}^{(N+1)} = \arg\max_{\boldsymbol{\xi}^*\in\mathcal{C}} \left(\min_{\mathbf{x}^{(i)},\, i=1,\ldots,N} \left\| \boldsymbol{\xi}^* - \mathbf{x}^{(i)} \right\|_2 \right), \tag{15}$$

with $\|\cdot\|_2$ the euclidean norm, i.e. $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{M} x_i^2\right)^{1/2}$.

2.1.3. Multi-element surrogate model approach

The previously presented PCK metamodel suffers from low convergence rates when the forward model $\mathcal{M}$ exhibits non-smoothness [59]. Thus, considerably large ED sizes are often required to achieve accurate predictions. This aspect undermines the computational efficiency of the PCE-based Kriging model, which is dominated by the $\mathcal{O}(N^3)$ complexity of the Kriging predictor. In turn, this implies long construction times or even memory overflow issues when solving the optimization problem in Eq. (13). Moreover, the larger the size of the ED, the slower the evaluation of the corresponding metamodel, which reduces or vanishes the advantages of the surrogate approach. To address this issue, a multi-element PCK model inspired by the ME-gPC method by Wan and Karniadakis [60] is proposed in this work. This approach consists in the partitioning of the random input space into a finite set of non-overlapping subdomains, the construction of a local PCK surrogate model in each one following the formulation in Section 2.1.2 and, finally, assembling them into a piecewise function to obtain a global metamodel, as sketched in Fig. 1.

To systematize the partition of the domain, a simple but efficient regular block partitioning approach has been adopted. To do so, the domain partitions are conducted one variable at a time based on a preliminary sensitivity analysis. To this aim, the algorithm prioritizes the partition of those variables providing highest sensitivities in terms of the Sobol indices, which can be readily extracted from the coefficients of a global PCE [51]. In this way, priority in the partitioning is given to the direction of those parameters with highest sensitivity, i.e., those with the greatest effect on the variability of the quantity of interest y. Over a set a surrogate models considering an increasing number of domain partitions, the optimal number of

Table 1
Error metrics for the accuracy assessment of surrogate models over a validation set (VS) of size K (Ref. [61]).

Normalized mean-square error (NRMSE)	Normalized average absolute error (NAAE)
$\text{NRMSE} = \sum_{i=1}^{K}\left(\hat{\upsilon}^{(i)} - \upsilon^{(i)}\right)^2 \Big/ \sum_{i=1}^{K}\left(\bar{\boldsymbol{\Upsilon}} - \upsilon^{(i)}\right)^2$	$\text{NAAE} = (K\sigma_{\boldsymbol{\Upsilon}})^{-1}\sum_{i=1}^{K}\left\lvert\hat{\upsilon}^{(i)} - \upsilon^{(i)}\right\rvert$
Coefficient of Determination (R^2)	Normalized maximum absolute error (NMAE)
$R^2 = \sigma^2_{\boldsymbol{\Upsilon}\hat{\boldsymbol{\Upsilon}}} \Big/ \sigma^2_{\boldsymbol{\Upsilon}}\sigma^2_{\hat{\boldsymbol{\Upsilon}}}$	$\text{NMAE} = (K\sigma_{\boldsymbol{\Upsilon}})^{-1}\max_{i=1}^{K}\left\lvert\hat{\upsilon}^{(i)} - \upsilon^{(i)}\right\rvert$

partitions is determined in terms of the quality metrics reported hereafter in Table 1. In general, for the random variable $\mathbf{x}: \Omega \to \mathcal{D}_{\mathbf{x}} \subset \mathbb{R}^M$, a decomposition is defined as:

$$\mathcal{D}_{\mathbf{x}} = \bigcup_{j\in\mathcal{J}} \mathcal{D}_j, \quad \mathcal{D}_j \cap \mathcal{D}_{j'} = \emptyset, \quad \text{if } j \neq j', \tag{16}$$

where $\chi_{\mathcal{D}_j}: \Omega \to \mathbb{R}$ denotes the indicator random variable:

$$\chi_{\mathcal{D}_j}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{D}_j \\ 0 & \text{otherwise} \end{cases}.$$

In this way, the global model is defined in a piecewise fashion as:

$$\hat{\mathcal{M}}_{PCK}(\mathbf{x}) = \sum_{j\in\mathcal{J}} \chi_{\mathcal{D}_j}(\mathbf{x})\hat{\mathcal{M}}^j(\mathbf{x}). \tag{17}$$

As aforementioned, the number of partitions in this work is defined after a parametric analysis. Nevertheless, the previous formulation may be readily automated as follows. The splitting criterion of the domain is determined by a certain user-defined accuracy goal and a minimum number of samples $\mathcal{N}$ per region. Afterwards, the splitting process is performed iteratively from a PCK model built over the full parameter space Ω. In case the target accuracy has not been reached, the space is split into two regions and the ED is enriched in each of these subdomains by the MIPT algorithm until there are $\mathcal{N}$ samples in each one. Note that, given the sequential nature of MIPT, the information of the previously extracted samples is not lost. If the accuracy goal is not reached yet, a new division of the space and a new enrichment of the ED are performed.

2.1.4. Surrogate model accuracy. Complexity analysis of the algorithm

To evaluate accuracy of the developed metamodel, both local and global error metrics are considered. These metrics are computed by considering a validation set (VS) $\boldsymbol{\Lambda} = \{\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(K)}\}$, $K \in \mathbb{N}$, of the parameters space (independent of the ED). Denote by $\boldsymbol{\Upsilon} = \{\upsilon^{(1)} = \mathcal{M}(\boldsymbol{\xi}^{(1)}), \ldots, \upsilon^{(K)} = \mathcal{M}(\boldsymbol{\xi}^{(K)})\}$ and $\hat{\boldsymbol{\Upsilon}} = \{\hat{\upsilon}^{(1)} = \hat{\mathcal{M}}_{PCK}(\boldsymbol{\xi}^{(1)}), \ldots, \hat{\upsilon}^{(K)} = \hat{\mathcal{M}}_{PCK}(\boldsymbol{\xi}^{(K)})\}$ the outputs of the VS estimated by the forward model and the metamodel, respectively. Then, the accuracy of the surrogate model can be assessed by using the error metrics like those collected in Table 1. In this table, $\bar{\boldsymbol{\Upsilon}}$ and $\sigma_{\boldsymbol{\Upsilon}} = \sqrt{\left(\sum_{i=1}^{K}\left(\bar{\boldsymbol{\Upsilon}} - \upsilon^{(i)}\right)^2\right)/(K-1)}$ denote the arithmetic mean and the quasi standard deviation of $\boldsymbol{\Upsilon}$, respectively. Term $\sigma_{\boldsymbol{\Upsilon}\hat{\boldsymbol{\Upsilon}}}$ represents the covariance of $(\boldsymbol{\Upsilon}, \hat{\boldsymbol{\Upsilon}})$, and $\sigma^2_{\boldsymbol{\Upsilon}}$ and $\sigma^2_{\hat{\boldsymbol{\Upsilon}}}$ indicate the variance of $\boldsymbol{\Upsilon}$ and $\hat{\boldsymbol{\Upsilon}}$, respectively. Note that the error metric NMAE in Table 1 provides a local estimation of accuracy, while NRMSE, NAAE, and R^2 represent global accuracy measures.

In addition to the error metrics shown in Table 1, and to verify the whole rate of convergence of the proposed model to the unknown function on untried points, we are interested in bounding the *maximum PCK-predictive error* over the domain $\mathcal{D}_{\mathbf{x}} \subset \mathbb{R}^M$:

$$\sup_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}} |\mathcal{M}(\mathbf{x}) - \hat{\mathcal{M}}_{PCK}(\mathbf{x})|, \tag{18}$$

where

$$\hat{\mathcal{M}}_{PCK}(\mathbf{x}) = \sum_{j\in\mathcal{J}} \chi_{\mathcal{D}_j}(\mathbf{x})\hat{\mathcal{M}}^j_{PCK}(\mathbf{x}) =$$
$$\sum_{j\in\mathcal{J}} \chi_{\mathcal{D}_j}(\mathbf{x})\left[\mathbf{r}^{j\mathrm{T}}(\mathbf{x})\mathbf{R}^{j-1}\boldsymbol{y} - \left(\boldsymbol{\Theta}^{j\mathrm{T}}\mathbf{R}^{j-1}\mathbf{r}^j(\mathbf{x}) - \boldsymbol{\Psi}^j(\mathbf{x})\right)^{\mathrm{T}}\left(\boldsymbol{\Theta}^{j\mathrm{T}}\mathbf{R}^{j-1}\boldsymbol{\Theta}^j\right)^{-1}\boldsymbol{\Theta}^{j\mathrm{T}}\mathbf{R}^{j-1}\boldsymbol{y}\right],$$

is the best linear unbiased predictor (BLUP) of the model $\mathcal{M}$ response at any untried point $\mathbf{x} \in \mathcal{D}_{\mathbf{x}}$, with $\mathbf{r}^j(\mathbf{x}) = [R^j(|\mathbf{x} - \mathbf{x}^{(1)}|), \ldots R^j(|\mathbf{x} - \mathbf{x}^{(N_j)}|)]^{\mathrm{T}}$ the vector of correlations between the design sites $\Xi_j = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_j)}\right\} \subset \mathcal{D}_j$ and $\mathbf{x}$, and R^j the selected correlation function particularized in the jth subregion. Note that the uniform bound in Eq. (18) covers the worst case for the prediction error of the PCK model.

It has been reported in the literature [62,63] that the prediction error of the universal Kriging converges to zero under uniform metric. Adapting Theorem 2 in Wang et al. [63] to the multielement PCE-Kriging model proposed in this work, the prediction error can be stated to satisfy:

$$\mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}}|\mathcal{M}(\mathbf{x})-\hat{\mathcal{M}}_{PCK}(\mathbf{x})|\right]\leq \mathcal{J}\cdot\max_{j\in\mathcal{J}}\mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{D}_j}|\mathcal{M}(\mathbf{x})-\hat{\mathcal{M}}_{PCK}(\mathbf{x})|\right]=\mathcal{O}\left(\mathcal{J}P_{\Xi_j}\left(\mathcal{P}^j A+\log^{\frac{1}{2}}P_{\Xi_j}^{-1}\right)\right), \tag{19}$$

where $\mathcal{J}$ is the number of subdomains, $\mathcal{P}^j=\text{card}\left(\mathcal{A}^{M,p,q}\right)$, and A is a constant depending on the eigenvalues of $\boldsymbol{\Theta}_j$. Term $P_{\Xi_j}(\mathbf{x})$ denotes the *power function* given by $P^2_{\Xi_j}(\mathbf{x}):=1-\mathbf{r}^{j^{\mathrm{T}}}(\mathbf{x})\mathbf{R}^{j^{-1}}\mathbf{r}^j(\mathbf{x})$, and $P_{\Xi}:=\sup_{\mathbf{x}\in\mathcal{D}_{\mathbf{x}}}P_{\Xi_j}(\mathbf{x})$ is the supremum of the pointwise predictive standard deviation. It is thus reasonable to look for EDs minimizing P_{Ξ_j}. Note that the rate of convergence in Eq. (19) is a deterministic function dependent on the experimental design Ξ_j and decreasing with P_{Ξ_j}. In fact, when $N_j=\text{card}(\Xi_j)$ increases, P_{Ξ_j} tends to zero and so does the multielement PCK prediction error under the uniform metric in Eq. (19).

On the other hand, the algorithm for finding an optimizer of Eq. (13) is an iterative process involving the calculation of the inverse and determinant of a large $N\times N$ covariance matrix $R_{ij}=R\left(\left|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right|;\hat{\boldsymbol{\theta}}\right)$. Thus, the computational effort to obtain the solution may become impractical for large numbers N of training data points in Ξ. Note that the PCK model requires $\mathcal{O}(N^3)$ operations and has a memory complexity of the order of N^2 [34]. In this light, the splitting technique presented in Section 2.1.3 leads to substantial reductions in the computational effort. Specifically, taking $N_j=\text{card}(\Xi_j)$, with $N_j\ll N$, $\mathcal{J}\ll N_j$, the algorithm effort and the memory storage reduces to $\mathcal{J}\cdot\mathcal{O}(N_j^3)\sim\mathcal{O}(N_j^3)\ll\mathcal{O}(N^3)$ and $\mathcal{J}\cdot\mathcal{O}(N_j^2)\sim\mathcal{O}(N_j^2)\ll\mathcal{O}(N^2)$, respectively. On the other hand, the optimal order of the polynomials in the PCE is automatically identified by the LAR algorithm. It is reported in reference [52] that the LAR algorithm with M variables requires $\mathcal{O}(M^3+N_jM^2)$ computations in any subdomain $\mathcal{D}_j$. Therefore, in our case where $M\ll N_j$, it follows that $M^3<N_jM^2$ and, thus, $\mathcal{O}(N_jM^2)\sim\mathcal{O}(N_j)$. Hence, the computational complexity of PCE when inserted as the trend term is marginal with respect to the overall construction of the Kriging model, thereby we can deduce that the efficiency of the proposed PCE-Kriging metamodel is $\mathcal{O}(N_j^3)\ll\mathcal{O}(N^3)$.

2.2. Bayesian parameter inference via MCMC

In the Bayesian inference framework, model parameters $\boldsymbol{\theta}$ are conceived as a random variable with a certain posterior PDF π described by Bayes' theorem:

$$\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)=\frac{p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{p(\boldsymbol{y})},\quad p(\boldsymbol{y})=\int_{\Omega}p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}, \tag{20}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})=\mathcal{L}(\boldsymbol{\theta})$ denotes the likelihood function, $p(\boldsymbol{\theta})$ the prior distribution of the model parameters, and $p(\boldsymbol{y}|\mathcal{M})$ a normalizing constant, also called evidence. In the context of this work, $\boldsymbol{y}$ and $\boldsymbol{\theta}$ represent a set of n experimental observations and the model parameters of the metamodel to be calibrated, respectively. Errors $\boldsymbol{\varepsilon}$ between the experiment and the predictions of the surrogate model are assumed to be normally distributed with zero mean and standard deviation $\sigma_{\boldsymbol{\varepsilon}}$, that is $\boldsymbol{y}=\widehat{\mathcal{M}}(\boldsymbol{\theta})+\boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}\sim\mathcal{N}(0,\sigma_{\boldsymbol{\varepsilon}}\boldsymbol{I})$. Then, the likelihood function $\mathcal{L}(\boldsymbol{\theta})$ can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta})=\frac{\exp\left(-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\sum_{i=1}^{n}|y_i-\widehat{\mathcal{M}}(\boldsymbol{\theta})|^2\right)}{\sqrt{2\pi}\sigma_{\boldsymbol{\varepsilon}}}. \tag{21}$$

Obtaining π from Eq. (20) in analytical closed-form is infeasible in most practical applications, being MCMC methods the most popular approach to numerically characterize the PDF of the model parameters. This approach allows one to draw samples from π without computing the model evidence, which is independent from the model parameters $\boldsymbol{\theta}$. In this work, the DRAM algorithm developed by Haario et al. [53] is implemented. This approach combines delayed rejection (DR) [64] and adaptive Metropolis (AM) [65], which enhances the sampling efficiency of the sampling and enables the identification of multi-modal PDFs. Given a set of observed data samples in vector $\mathbf{d}$, the working principle of the DRAM approach can be outlined as follows:

1. Initialize the parameter set $\boldsymbol{\theta}_c=\boldsymbol{\theta}_0$ and the number T of desired samples. Set an initial point from the parameter space and the covariance of the proposal distribution $\Sigma_p=\Sigma_0$. The proposal distribution is chosen as a multivariate Gaussian distribution with mean $\boldsymbol{\theta}_c$ and covariance matrix Σ_p. Select the initial non-adaptation period n_o and set $i=1$.
2. Propose a new parameter value $\boldsymbol{\theta}_{p,1}$ by sampling from a proposal PDF $S_1(\boldsymbol{\theta},\boldsymbol{\theta}_c)$. Accept $\boldsymbol{\theta}_{p,1}$ with probability:

$$\alpha_1\left(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{p,1}\right)=\min\left(1,\frac{p\left(\mathbf{d}\,|\boldsymbol{\theta}_{p,1}\right)S_1(\boldsymbol{\theta}_{p,1},\boldsymbol{\theta}_c)}{p\left(\mathbf{d}\,|\boldsymbol{\theta}_c\right)S_1(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{p,1})}\right), \tag{22}$$

and go to step (4). If rejected, propose a second stage move in step (3).

3. Propose a second stage move $\boldsymbol{\theta}_{p,2}$ sampling from $S_2(\boldsymbol{\theta}, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_c)$. This second stage proposal depends not only on the current position of the chain but also on the candidate that has just been proposed and rejected. Accept or reject $\boldsymbol{\theta}_{p,2}$ by setting:

$$\boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}_{p,2}, & \text{with probability } \alpha_2\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_{p,2}\right), \\ \boldsymbol{\theta}_c, & \text{with probability } 1-\alpha_2\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_{p,2}\right), \end{cases} \tag{23}$$

with

$$\alpha_2\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_{p,2}\right) = \min\left\{1, \frac{p\left(\mathbf{d}\,|\boldsymbol{\theta}_{p,2}\right) S_1\left(\boldsymbol{\theta}_{p,2}, \boldsymbol{\theta}_{p,1}\right) S_2\left(\boldsymbol{\theta}_{p,2}, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_c\right)\left[1-\alpha_1\left(\boldsymbol{\theta}_{p,2}, \boldsymbol{\theta}_{p,1}\right)\right]}{p\left(\mathbf{d}\,|\boldsymbol{\theta}_c\right) S_1\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}\right) S_2\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}, \boldsymbol{\theta}_{p,1}\right)\left[1-\alpha_1\left(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{p,1}\right)\right]}\right\}. \tag{24}$$

4. Update the covariance matrix Σ_p as:

$$\Sigma_p = \begin{cases} \Sigma_0 & i \le n_0 \\ s_d \mathrm{cov}\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i\right) & i > n_0 \end{cases} \tag{25}$$

with s_d a scaling parameter. Following [66], $s_d = 2.4^2/d$ with d being the number of fitting parameters is recommended as a good default value in most applications.

5. Go to step 2, until the desired number of samples T are obtained.

3. Numerical results and discussion

This section presents two application case studies to demonstrate the effectiveness of the proposed surrogate model-based Bayesian parameter estimation. These include a two-dimensional benchmark function and the PDE for TDS testing of hydrogen desorption in metals. The previous formulation has been implemented in Matlab environment, and all the numerical tests have carried out in a computer Intel(R) Core(TM) i9-10900X CPU@3.70 GHz with 64 GB of RAM memory. In the remainder of this section, for simplicity in the notation, the predictions of the PCK metamodels $\hat{\mathcal{M}}_{PCK}$ are noted as $\hat{\mathcal{M}}$. A q-norm value of 0.95 and Legendre polynomials of orders ranging from 2 to 6 are selected to build the PCEs in all the analyses hereafter. For the generation of the EDs, the number of random Monte Carlo candidate samples in the MIPT algorithm introduced in Section 2.1.2 is set to 25,000.

3.1. Two-dimensional Drop-Wave function

This first case study investigates the Drop-Wave function, also known as the Salomon's function [67], given by $f: \mathcal{D}_{\mathbf{x}} = [-10,\ 10]^2 \in \mathbb{R}^2 \rightarrow \mathbb{R}$:

$$f(x_1, x_2) = 1 - \cos\left(2\pi\sqrt{x_1^2 + x_2^2}\right) + 0.1\sqrt{x_1^2 + x_2^2}. \tag{26}$$

This function is commonly used for benchmarking optimization algorithms. Owing to its highly non-linear character, the Drop-Wave function represents an ideal case study to validate the proposed multi-element PCK metamodel. Note that the surrogate modelling of this function using conventional techniques is extremely challenging given its fast-varying gradients and irregular response as observed in Fig. 2 (a). Following Section 2.1.3, four experimental design sets $ED_i \subset \mathcal{D}_{\mathbf{x}}$, $i = 1, \ldots, 4$ containing 360, 720, 1440 and 2880 samples have been defined. In addition, three different number of domain partition schemes $\mathcal{P}^j$, $j = 1, \ldots, 3$, have been considered. These include $\mathcal{P}^1 = [-10,\ 10]^2$, $\mathcal{P}^2 = ([-10,\ 0) \cup [0,\ 10])^2$, and $\mathcal{P}^3 = \left(\left[-10,\ \frac{-10}{3}\right) \cup \left[\frac{-10}{3},\ \frac{10}{3}\right) \cup \left[\frac{10}{3},\ 10\right]\right)^2$, leading to a total of two, four and nine sub-domains, respectively. The number of samples has been chosen with the aim of obtaining a wide range of errors to correctly identify the convergence of the prediction error. For instance, if one takes the R^2 error metric, note that the constructed metamodels exhibit a wide range of R^2 values from 0.011 to 0.999. Additionally, the predictions by a previously reported multielement approach, the Stochastic Spectral Embedding (SSE) model proposed by Sudret and Marelli [35], are also presented as a reference solution. The SSE model is a PCE-based technique consisting of constructing a sequence of residual spectral expansions of the target model in subdomains of the input space. The implementation included in the UQLab software [68] has been used to carry out the analyses. Four different surrogate models have been built, one for each considered ED. As parameters, a q-norm value of 0.95, polynomials ranging from degree 2 to 10, and a minimum size of points per region equal to the size(ED)/120 have been selected. To sample the ED, the sequential experimental design based on the LHS implemented in UQLab has been chosen.

Hence, a total of sixteen surrogate models have been constructed. For ease in the discussion, the PCK surrogate models are specified with sub- and super-indexes denoting the size of the ED and the number of partitions, respectively, as reported in Table 2. All the surrogate models have been validated using a reasonably large VS of 20,000 samples, and the accuracy of the models has been evaluated through the accuracy metrics reported in Table 3. The computational times involved in the construction t_c of the surrogate models, as well as their evaluation times t_e and t_e(VS) for a single point and the full VS

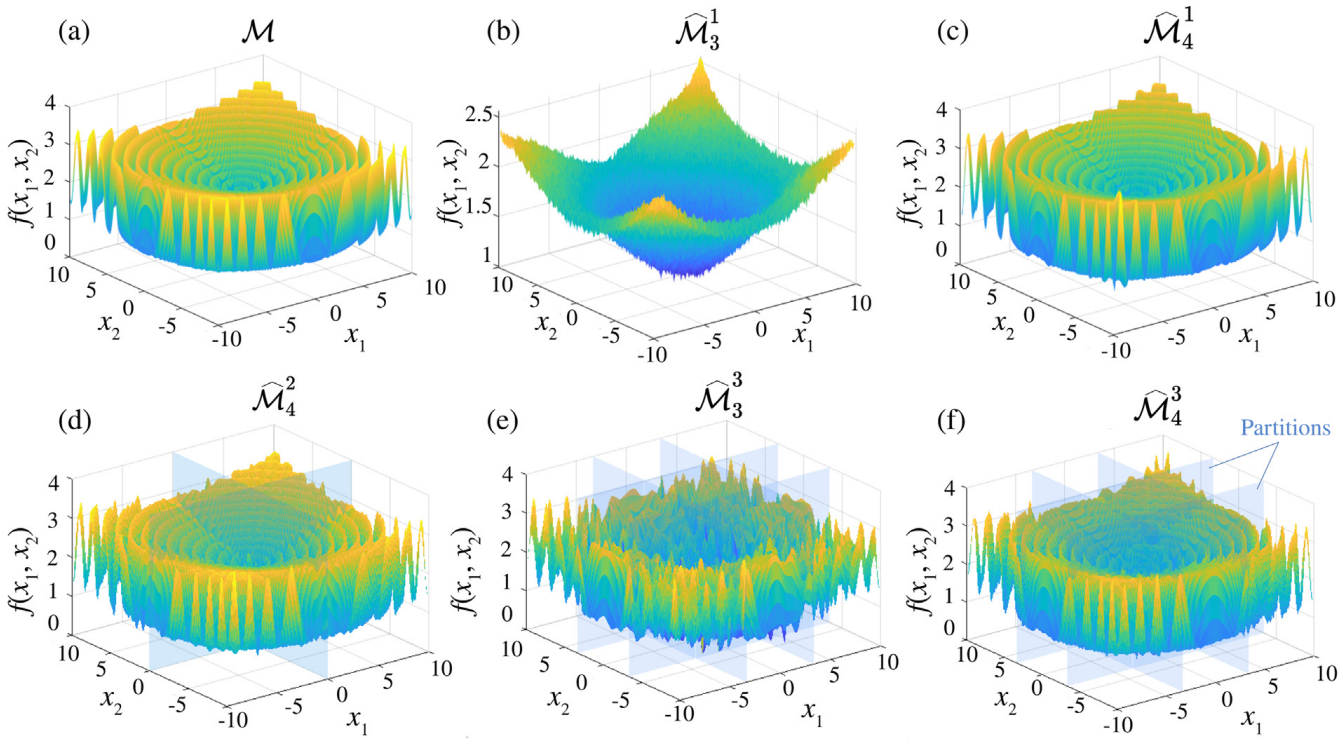

Fig. 2. Exact response surface of the Drop-Wave function (a), and predictions by surrogate models $\hat{\mathcal{M}}_1^3$ (b), $\hat{\mathcal{M}}_4^1$ (c), $\hat{\mathcal{M}}_4^2$ (d), $\hat{\mathcal{M}}_3^3$ (e) and $\hat{\mathcal{M}}_4^3$ (f).

Table 2
PCK surrogate-models constructed for the Drop-Wave function considering increasing EDs (ED_i) with varying numbers of domain partitions ($\mathcal{P}^j$) (VS of 20,000 samples).

No. of sub-domains	ED_1 (360 samples)	ED_2 (720 samples)	ED_3 (1440 samples)	ED_4 (2880 samples)
$\mathcal{P}^1$ (1 partition)	$\hat{\mathcal{M}}_1^1$	$\hat{\mathcal{M}}_2^1$	$\hat{\mathcal{M}}_3^1$	$\hat{\mathcal{M}}_4^1$
$\mathcal{P}^2$ (4 partitions)	$\hat{\mathcal{M}}_1^2$	$\hat{\mathcal{M}}_2^2$	$\hat{\mathcal{M}}_3^2$	$\hat{\mathcal{M}}_4^2$
$\mathcal{P}^3$ (9 partitions)	$\hat{\mathcal{M}}_1^3$	$\hat{\mathcal{M}}_2^3$	$\hat{\mathcal{M}}_3^3$	$\hat{\mathcal{M}}_4^3$
SSE (reference)	$\hat{\mathcal{M}}_1^{SSE}$	$\hat{\mathcal{M}}_2^{SSE}$	$\hat{\mathcal{M}}_3^{SSE}$	$\hat{\mathcal{M}}_4^{SSE}$

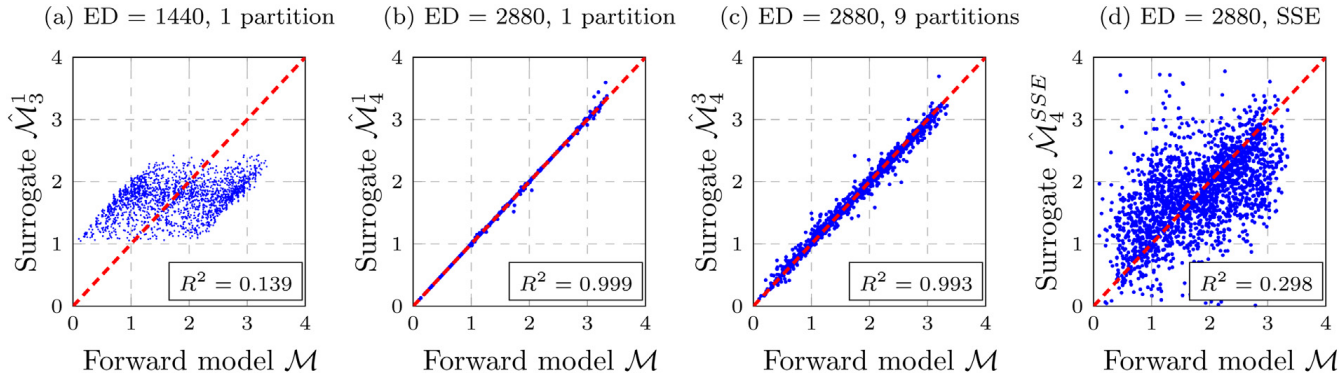

Fig. 3. Forward model evaluations versus surrogate model predictions for the Drop-Wave function. $\hat{\mathcal{M}}_3^1$ (a), $\hat{\mathcal{M}}_4^1$ (b), $\hat{\mathcal{M}}_4^3$ (c) and $\hat{\mathcal{M}}_4^{SSE}$ (d) (VS of 20,000 samples).

have been also computed and collected to compare their effectiveness. The comparison of the exact response surface of the Drop-Wave function and the predicted ones by five of the constructed surrogate models is depicted in Fig. 2.

Fig. 3 shows the scatter plots of the forward model evaluated on the VS versus four metamodels, $\hat{\mathcal{M}}_3^1$ (a), $\hat{\mathcal{M}}_4^1$ (b), $\hat{\mathcal{M}}_4^3$ (c) and $\hat{\mathcal{M}}_4^{SSE}$ (d). The first two metamodels consider the whole design space, while the last two account for partition approaches. It is noted in this figure that the best approximations are found for $\hat{\mathcal{M}}_4^1$ (no partitions) and $\hat{\mathcal{M}}_4^3$ (9 partitions with EDs of 2880 samples). The predictions by these models exhibit low scatter around the diagonal line (perfect metamodel) with coefficients of determination R^2 very close to 1. The limited efficiency of the SSE model in this case study is evidenced by the large scatter of its predictions along the diagonal in Fig. 3(d). Interestingly, note that the predictions by $\hat{\mathcal{M}}_4^1$ slightly

Table 3
Accuracy and computational efficiency analysis of proposed surrogate models applied to the Drop-Wave function. Terms t_c, t_e and t_e(VS) denote the time of construction, average point evaluation, and evaluation on the full validation set, respectively.

Model	NAAE	NMAE	NRMSE	R^2	t_c [s]	t_e [s]	t_e(VS) [s]
$\hat{\mathcal{M}}_1^1$	1.030E + 0	1.736E−4	1.589E + 0	0.024	34.1	5.069E−5	1.013
$\hat{\mathcal{M}}_2^1$	8.340E−1	1.001E−4	8.701E−1	0.134	151.3	2.014E−4	4.028
$\hat{\mathcal{M}}_3^1$	8.321E−1	8.159E−5	8.627E−1	0.139	661.8	9.301E−4	18.601
$\hat{\mathcal{M}}_4^1$	2.876E−3	2.707E−5	2.651E−4	0.999	2905.7	5.078E−3	101.540
$\hat{\mathcal{M}}_1^2$	9.162E−1	1.399E−4	12.080E−1	0.043	20.3	6.840E−6	0.138
$\hat{\mathcal{M}}_2^2$	8.978E−1	1.412E−4	11.980E−1	0.823	44.4	1.996E−5	0.399
$\hat{\mathcal{M}}_3^2$	5.448E−1	1.355E−4	4.974E−1	0.551	131.8	5.656E−5	1.131
$\hat{\mathcal{M}}_4^2$	2.256E−2	3.976E−5	2.559E−3	0.997	594.7	2.066E−4	4.132
$\hat{\mathcal{M}}_1^3$	8.757E−1	2.044E−4	11.230E−1	0.079	28.9	3.590E−6	0.071
$\hat{\mathcal{M}}_2^3$	7.765E−1	1.231E−4	8.622E−1	0.193	41.8	6.875E−6	0.138
$\hat{\mathcal{M}}_3^3$	4.404E−1	1.297E−4	3.551E−1	0.645	86.4	1.513E−5	0.303
$\hat{\mathcal{M}}_4^3$	4.345E−2	4.667E−5	6.712E−3	0.993	244.7	4.750E−5	0.950
$\hat{\mathcal{M}}_1^{SSE}$	1.249E + 0	6.463E−4	2.684E + 0	0.011	42.9	3.529E−5	0.706
$\hat{\mathcal{M}}_2^{SSE}$	1.191E + 0	1.302E−3	2.733E + 0	0.024	33.9	3.401E−5	0.681
$\hat{\mathcal{M}}_3^{SSE}$	8.838E−1	1.714E−3	1.535E + 0	0.113	30.6	3.266E−5	0.653
$\hat{\mathcal{M}}_4^{SSE}$	6.748E−1	5.694E−4	1.041E + 0	0.298	32.5	3.101E−5	0.620

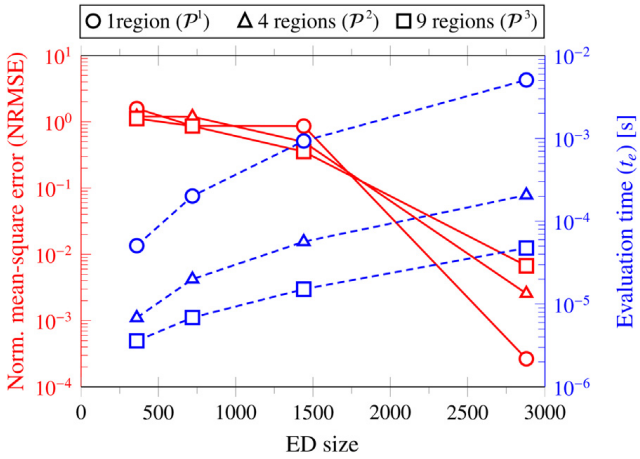

Fig. 4. Performance assessment of PCK surrogate models for the Drop-Wave function (VS of 20 000 samples).

outperform those obtained with $\hat{\mathcal{M}}_4^3$, while higher numbers of partitions do not seem not to systematically improve the prediction accuracy. Nevertheless, the computational times involved in the construction and evaluation of $\hat{\mathcal{M}}_4^1$ are, respectively, about 10 and 100 times those required by $\hat{\mathcal{M}}_4^3$ (see Table 3). It is extracted from this analysis that the selection of the optimal surrogate model must be conducted by balancing the computational burden and the fitting accuracy. In this regard, Fig. 4 investigates the computational efficiency in terms of t_e versus prediction accuracy (NRMSE) for all the considered surrogate models. In this figure, it is trivially observed that as the size of the ED increases, both computational time and accuracy of the metamodels increase. It is important to highlight that the consideration of higher number of subdomains leads to lower evaluation times for all the considered EDs. This is explained by the implementation of the LAR algorithm to extract optimal sets of polynomials in the PCE, and in particular, thanks to the reductions in the computational cost involved in the construction of the Kriging predictor (see Section 2.1.4). When inspected in a partition-wise fashion, the forward model exhibits a smoother behaviour, in such a way that the PCE requires less high-order polynomials to reproduce its behaviour. This results in more compact expansions, which also decreases the cost in the computation of the correlation matrix in the Kriging metamodel. On the other hand, note that the higher the order in the PCE, the larger the number of samples that are required in the ED to fit the expansion with accuracy. In this light, to reach a comparable accuracy to the one achieved by the Multi-Element PCK ($R^2 > 0.99$) through SSE, it is necessary to increase the degree of the polynomial expansion up to 14 and to sample a ED with more than 45,000 training points. These results strengthen the comparatively superior convergence rate of the proposed approach for this class of problems with highly non-linear spatial variability. As previously detailed in Section 2.1.4, the computational complexity of the proposed PCK model is $\mathcal{O}(N_j^3)$. This is due to the Cholesky decomposition of the correlation matrix $\mathbf{R}$ in Eq. (13). Therefore, as the ED increases, the computational cost involved in the determination of the stochastic hyper-parameters of the Kriging model and its evaluation rises dramatically. On the other hand, the dependence of the accuracy of the metamodel with the number of partitions is not so clear. It is noted that the accuracy of metamodels trained with a larger number of subdomains is higher compared to those with no

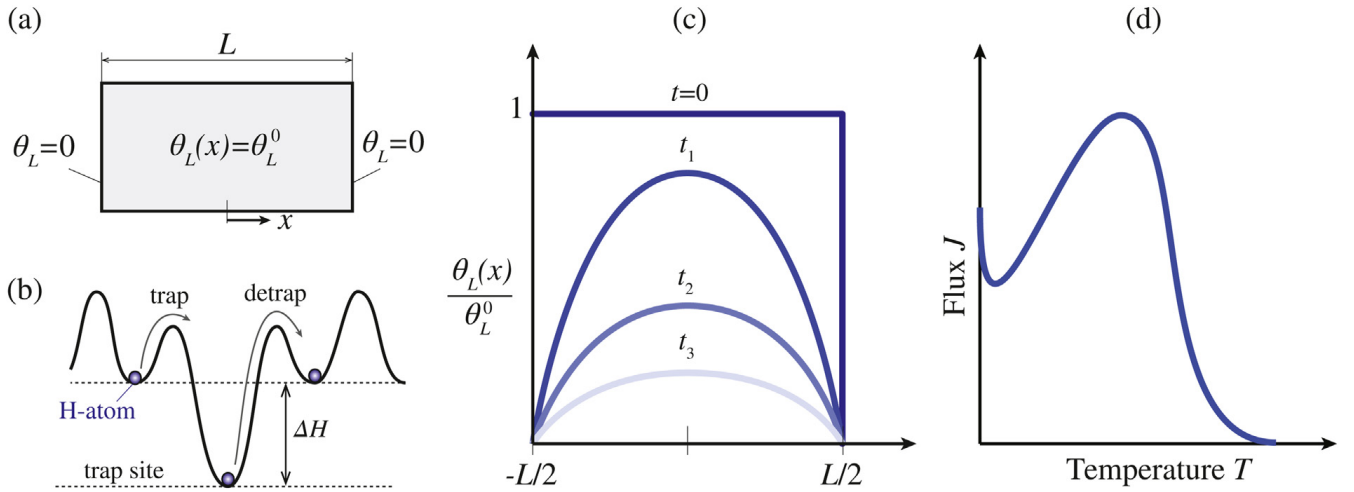

Fig. 5. (a) A schematic illustration of initial and boundary conditions in a TDS test. (b) Schematic definition of binding energy in a one-dimensional diffusion path. (c) Transient solution curves of the normalised lattice occupancy fraction θ_L/θ_L^0 at different times t along the specimen's thickness. (d) A schematic of typical hydrogen desorption flux versus temperature curves obtained in a TDS test.

partitions for limited to moderate EDs. Nevertheless, when the size of the ED goes from moderate to large, the accuracy diminishes.

3.2. Thermal desorption spectroscopy (TDS) of hydrogen in metals

This last section reports the use of the proposed surrogate model for the Bayesian identification of hydrogen desorption and trapping characteristics in metallic materials. Hydrogen embrittlement (HE) refers to the loss of ductility and toughness of metallic alloys induced by hydrogen atoms deposited at lattice sites and micro-structural defects such as dislocations, grain boundaries or vacancies [69,70]. Although this phenomenon has been extensively documented since the 19th century [71], the growing trend towards a hydrogen-based economy as a means of mitigating CO_2 emissions and fossil fuel dependency has generated unprecedented interest on HE research. Micro-structural defects in metals act as 'trap' sites, which sequester hydrogen and govern the susceptibility to HE [72,73]. Their characterization is thus of pivotal importance for the understanding of HE and the design of HE-resistant alloys, and this is generally achieved using TDS experiments [74]. The TDS test involves several stages [75]: charging a sample with hydrogen, heating the sample at a fixed rate, and detecting the flux of desorbing hydrogen as a function of temperature by using a mass spectrometer. The hydrogen flow curve of desorbing hydrogen as a function of temperature defines the TDS spectrum, whose peaks can be associated with the presence of diverse micro-structural defects. Nonetheless, the formation of peaks in the TDS spectrum may be induced by the combined action of manifold hydrogen traps, being necessary to use simulation models and inverse calibration for their identification. Previous investigations on the modelling of the TDS test evidenced the existence of non-smooth relationships between the flux curves and the parameters characterizing micro-structural defects (see e.g. [76]), making this application a formidable benchmark case study for the formulation presented in this work. In the remainder of this section, the PDE governing the hydrogen diffusion in materials tested by TDS is introduced in Section 3.2.1. The construction of the surrogate model and its performance evaluation is reported in Section 3.2.2 and, finally, Section 3.2.3 presents the Bayesian parameter identification results.

3.2.1. TDS governing diffusion equation

Consider a one-dimensional specimen of length L as sketched in Fig. 5(a). The specimen is subjected to increasing temperatures T, starting from T_o and increasing at a constant heating rate ϕ. Hydrogen atoms occupy normal intersticial lattice sites (NILS) and additionally can reside at trapping sites such as interfaces or dislocations. The kinetics of hydrogen trapping and detrapping in metals is commonly described with a two-level system as sketched in Fig. 5(b) for the case of a single trap. The potential landscape in this figure describes the diffusion path of hydrogen in metals, the trap binding energy ΔH being the difference between detrapping and trapping energies. Let us assume that the number of hydrogen traps in the specimen amounts to N_t. Then, let $C_L(x,t)$ and $C_{T,i}(x,t)$, $i=1,\dots,N_t$, denote the hydrogen concentration in the lattice and in the i-th trap, respectively, with $x\in[-L/2,L/2]$ and t respectively denoting space and time. On this basis, the Fickian diffusion equation needs to be enriched with source and sink terms as [76]:

$$\frac{\partial C_L}{\partial t}+\sum_{i=1}^{N_t}\frac{\partial C_{T,i}}{\partial t}=D_L\frac{\partial^2 C_L}{\partial x^2}, \tag{27}$$

with $D_L=D_o\exp(-Q/RT)$ being the lattice diffusion coefficient, which is expressed in terms of the lattice activation energy Q, diffusion pre-exponential factor D_o, and the universal gas constant R. It is convenient to introduce the lattice and

Table 4
Non-dimensional variables used in the hydrogen diffusion PDE employed for the TDS tests.

Spatial coordinate	$\overline{x} = x/L$	Lattice activation energy	$\overline{Q} = Q/(RT_o)$
Time coordinate	$\overline{t} = tD_o/L^2$	Trap binding energy	$\overline{\Delta H_i} = \Delta H_i/(RT_o)$
Heating rate	$\overline{\phi} = \left(\phi L^2\right)/(T_o D_o)$	Lattice diffusion coefficient	$\overline{D}_L = D_L/D_o$
Trap density	$\overline{N}_i = (\alpha N_{T,i})/(\beta N_L)$	Local equilibrium constant	$\overline{K} = \exp\left\{-\frac{\overline{\Delta H_i}}{\overline{T}}\right\}$
Temperature	$\overline{T} = T/T_o$	Fractional lattice occupancy	$\overline{\theta}_L = \theta_L/\theta_L^o$

trap occupancy fractions θ_L and $\theta_{T,i}\left(\left\{\theta_L, \theta_{T,i}\right\} \in [0,1]\right)$, respectively, by rewriting the corresponding concentrations in the form $C_L = \theta_L \beta N_L$ and $C_{T,i} = \theta_{T,i} \alpha N_{T,i}$. Here, β is the number of NILS per unit volume, α is the number of atoms sites per trap, N_L is the number of lattice atoms per unit volume, and $N_{T,i}$ is the number of trap sites per unit volume. Therefore, Eq. (27) can be rewritten as:

$$\frac{\partial \theta_L}{\partial t} + \sum_{i=1}^{N_t} \left(\frac{\alpha N_{T,i}}{\beta N_L}\right) \frac{\partial \theta_{T,i}}{\partial t} = D_L \frac{\partial^2 \theta_L}{\partial x^2}. \tag{28}$$

The PDE in Eq. (28) needs to be complemented with trap kinetic equations. To this aim, the formulations by MacNabb and Foster [77] and Oriani [78] are commonly adopted. The latter represents a simplification of the former by assuming that a local equilibrium exists between the hydrogen atoms at the lattice sites and the i-th trap such that, for $\theta_L \ll 1$,

$$\theta_{T,i} = \frac{K_i \theta_L}{(1 + K_i \theta_L)}, \tag{29}$$

with K_i being the local equilibrium constant for the i-th trap:

$$K_i = \exp\left\{-\frac{\Delta H_i}{RT}\right\}. \tag{30}$$

Introducing Eq. (29) into (28), and following the non-dimensional formulation developed by Raina et al. [76], the governing PDE describing hydrogen diffusion in the TDS test can be recast in a compact form as:

$$\frac{\partial \overline{\theta}_L}{\partial \overline{t}} \left[1 + \sum_{i=1}^{N_t} \frac{\overline{K}_i \overline{N}_i}{\left(1 + \overline{K}_i \theta_L^0 \overline{\theta}_L\right)^2}\right] + \frac{\overline{\theta}_L}{\overline{T}^2} \sum_{i=1}^{N_t} \frac{\overline{K}_i \overline{N}_i \overline{\Delta H}_i \overline{\phi}}{\left(1 + \overline{K}_i \theta_L^0 \overline{\theta}_L\right)^2} = \overline{D}_L \frac{\partial^2 \overline{\theta}_L}{\partial \overline{x}^2}, \tag{31}$$

with θ_L^0 being the initial lattice occupancy. The non-dimensional variables employed are listed in Table 4.

The initial and boundary conditions of the PDE in Eq. (31) are schematically presented in Fig. 5(a). At $t = 0$, it is assumed an initial uniform lattice occupancy $\overline{\theta}_L\left(\overline{x}, \overline{t} = 0\right) = 1$. Thereafter, the hydrogen lattice occupancy is assumed zero at the boundaries, that is $\overline{\theta}_L\left(\overline{x} = \pm 1/2, \overline{t} > 0\right) = 0$. As temperature raises, the lattice occupancy evolves spatially and temporally as sketched in Fig. 5(c), and the flux of hydrogen atoms $J(t)$ diffusing out at boundaries is measured as presented in Fig. 5(d). This flux can be obtained in non-dimensional terms after solving Eq. (31) as [76]:

$$\overline{J} = -\overline{D}_L \theta_L^o \frac{\partial \overline{\theta}_L}{\partial \overline{x}}. \tag{32}$$

Generally, the magnitudes of Q, D_0 and θ_L^0 are known, and the heating rate ϕ is an input to the TDS system. Therefore, the TDS spectrum can be used to map the microstructural hydrogen traps, as characterised by their trap densities ($\overline{N}_i$) and binding energies ($\overline{\Delta H}_i$). These can be obtained for a given flux curve $\overline{J}$ by the inverse calibration of the PDE in Eq. (31).

The surrogate modelling of the flux curves obtained after solving Eq. (31) represents a formidable problem due to the strong nonlinearities of these curves. Specifically, depending upon the hydrogen trap configuration, several different regimes can be observed, as previously discussed by Raina *et al.* [76]. Specifically, their results for the case of metals containing a single trap showed that no peak flux is attained for low trap densities and binding energies. Alternatively, when a peak flux is found, those authors identified two distinct regimes (I and II) originated by two types of microstructural defects, referred to as shallow and deep traps. Shallow traps are characterized by large trap densities, and give origin to peak fluxes that are highly sensitive to both N and ΔH. On the other hand, deep traps are characterized by low trap densities, resulting in peak fluxes that are insensitive to the trap binding energy. The existence of these different regimes turns the construction of a surrogate model covering the whole domain of the traps into an notably challenging task. Note that a large number of high-order polynomials and a dense ED need to be included in the PCE to accurately represent the whole global behaviour of the hydrogen flux. Such large EDs may severely compromise the computational efficiency of the surrogate model since, as indicated above, the complexity of the Cholesky decomposition of the correlation matrix $\mathbf{R}$ in Eq. (13) is $\mathcal{O}\left(N^3\right)$. The metamodeling of TDS experiments thus represents an exceptional case study to justify the use of the domain partitioning approach presented in Section 2.1.3.

Table 5
Accuracy and computational efficiency analysis of PCK metamodels developed for the surrogate modelling of hydrogen diffusion flux curves obtained by TDS (VS = 36,000).

	$\hat{\mathcal{M}}_1$ 72 subdomains ED = 72,000	$\hat{\mathcal{M}}_2$ 72 subdomains ED = 108,000	$\hat{\mathcal{M}}_3$ 108 subdomains ED = 72,036	$\hat{\mathcal{M}}_4$ 108 subdomains ED = 108,000
t_e [ms]	1.97	4.86	0.83	2.02
NAAE	1.0637E−02	7.540E−03	1.115E−02	7.736E−03
NMAE	2.517E−05	2.239E−05	2.363E−05	1.656E−05
NRMSE	9.452E−04	5.259E−04	1.208E−03	5.229E−04
R^2	0.9991	0.9995	0.9988	0.9995

3.2.2. Surrogate modelling of TDS flux curves for metals with two traps

The PDE in Eq. (31) is solved numerically by using the *pdepe* solver in MATLAB. A space discretization of 201 elements along x was found to provide mesh-independent results. To illustrate the behaviour of a ferritic steel sample, representatitive model parameters from reference [76] have been adopted herein, including a lattice activation energy $Q = 6.7$ kJmol^{-1}, diffusion pre-exponential factor $D_o = 2 \times 10^{-7}$ m^{2}s^{-1}, heating rate $\overline{\phi} = 0.1$ and lattice density $N_L = 8.46 \times 10^{28}$ atoms m^{-3}, with $\alpha = \beta = 1$. The initial temperature and occupancy fraction are chosen as $T_o = 293$ K and $\theta_L^0 = 10^{-6}$, respectively, and the thickness of the specimen is chosen as $L = 5$ mm. The variation of trap binding energies and trap densities are selected as the physically meaningful ranges $-40 \leq \overline{\Delta H}_i \leq -10$ and $10^{-7} \leq \overline{N}_i \leq 10^{-2}$. In the present study, we limit to the modelling of metals with two hydrogen traps, i.e. $N_t = 2$. Therefore, in the surrogate modelling, temperature and the trap densities and binding energies are considered as input variables, which amounts to 5 design variables, i.e. $\mathcal{M}(\boldsymbol{x}) = \bar{J}$ with $\boldsymbol{x} = \left[\overline{T}, \overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2)\right]^{\mathrm{T}} \subset \mathbb{R}^5$.

After some preliminary sensitivity analyses, two partitions of $\mathcal{D}_{\mathbf{x}}$ have been considered, namely $\mathcal{P}^1$, $\mathcal{P}^2$. Partition $\mathcal{P}^1$ has been defined by splitting the temperature $\overline{T}$ and activation energy domains ($\overline{\Delta H}_i$, $i = 1, 2$) in two, while three segments were considered for the partition of the domain of the trap densities ($\log(\overline{N}_i)$, $i = 1, 2$). On the other hand, the partitions in $\mathcal{P}^2$ remain identical except for the temperature domain which is divided in three sub-domains. In order to define the optimal surrogate model, EDs of 1000 and 1500 sampling points per-subdomain have been considered for $\mathcal{P}^1$, while EDs of 667 and 1000 points per sub-domain have been defined for $\mathcal{P}^2$. This amounts to four different surrogate models labelled with $\hat{\mathcal{M}}_i$, $i = 1, \ldots, 4$. In order make a fair comparison between the different proposals, models $\hat{\mathcal{M}}_1$ and $\hat{\mathcal{M}}_3$ are trained with EDs of 72,000 and 72,036 samples respectively, while $\hat{\mathcal{M}}_2$ and $\hat{\mathcal{M}}_4$ are trained with 108,000 samples. The comparison of the metamodels in terms of accuracy and computational efficiency is reported in Table 5 over a VS of 36,000 samples. Similarly to the results in the previous case study, the consideration of domain partitioning leads to considerable computational time reductions and moderate reductions in prediction accuracy. Note that the evaluation time of the forward model is about 280 ms, so all the metamodels achieve reductions between 98.3%–99.7%. The computation time of the metamodel depends upon the size of the ED in each region, which explains why models $\hat{\mathcal{M}}_3$ and $\hat{\mathcal{M}}_2$ are the fastest and slowest ones, respectively. On the other hand, the accuracy of the metamodel increases as so does the size of the ED. Indeed, models $\hat{\mathcal{M}}_2$ and $\hat{\mathcal{M}}_4$ exhibit significantly lower errors compared to models $\hat{\mathcal{M}}_1$ and $\hat{\mathcal{M}}_3$. Therefore, in view of these results, $\hat{\mathcal{M}}_4$ provides a good trade-off between computational efficiency and accuracy, and it is selected in the subsequent Bayesian model parameter inference. To illustrate the effectiveness of the surrogate model in representing the different stages observed in the TDS test, Fig. 7 shows the comparison of the forward model and the predictions by $\hat{\mathcal{M}}_4$ for a variety of combinations of traps, including the case of fluxes without peak, one single peak, and two peaks. It is observed that the proposed PCK model can accurately reproduce all the different regimes observable in the TDS test. Only some minor errors are observed in the no flux regime, given the imposed limitation on the order of the polynomials in the PCE for the sake of computational efficiency. Finally, in order to highlight the superior performance of the proposed multi-element PCK metamodel, Fig. 6 furnishes the comparison of the predictions by standard LAR-PCE (trained with 76,000 samples) and $\hat{\mathcal{M}}_4$. These results clearly evidence the superior performance of the proposed approach with respect to LAR-PCE, whose predictions versus the forward model exhibits a large scatter around the diagonal line with a low coefficient of determination of $R^2 = 0.46$.

3.2.3. Bayesian inference of the trapping sites from a TDS experiment

In this last subsection, the previous surrogate model $\hat{\mathcal{M}}_4$ is used to conduct Bayesian parameter inference following the MCMC algorithm in Section 2.2. The trap binding energies and densities of the two trap system are chosen as the inference parameters $\boldsymbol{\theta} = \left(\overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2)\right)$ in Eq. (20). With the purpose of assessing the performance of the implemented DRAM MCMC approach to infer the parameters of hydrogen traps covering the two different regions described in reference [76], two different trap configurations are considered to generate synthetic experimental data from the forward model. A two-trap system (EI) with properties $\boldsymbol{\theta} = (-25, -35, -3, -2.5)$ is considered first. The second one (EII) instead is defined by $\boldsymbol{\theta} = (-15, -30, -6, -3)$. The flux curves obtained in EI and EII correspond to those previously shown in Fig. 7(f) and (g), respectively. In addition, to evaluate the sensitivity of the model parameter inference to the presence of noise pollution in the experiment, a second analysis of the EII experiment was performed after affecting the flux curve with a zero-mean Gaussian white noise with a standard deviation equal to 0.4 times the mean value of the unpolluted flux curve

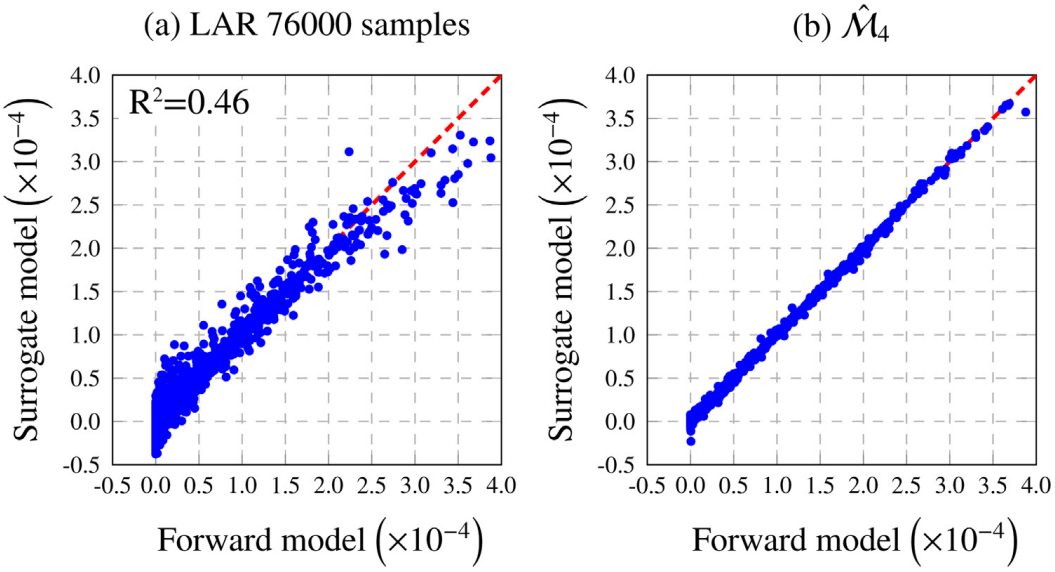

Fig. 6. Scatter plots of Hydrogen flux curves obtained by the forward solution of the PDF of the TDS test versus the predictions by standard LAR-PCE (a) and by the proposed multi-element PCK metamodel $\hat{\mathcal{M}}_4$ (b) (VS of 36,000 samples).

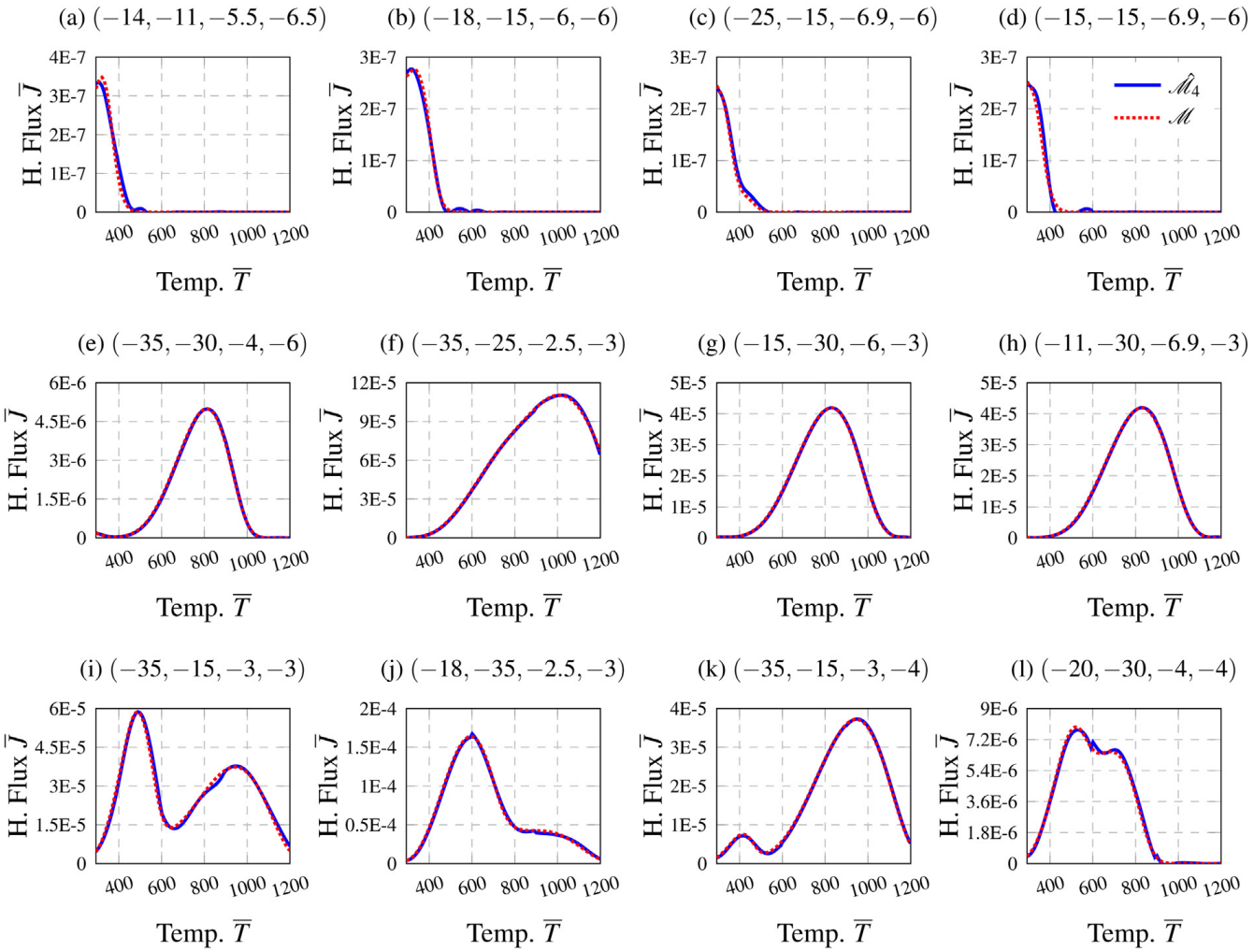

Fig. 7. Surrogate modelling of the Hydrogen flux curves obtained by TDS of metals with different values of trap binding energies and concentrations. Quantities in parenthesis represent the parameters of the traps $\left(\overline{\Delta H}_1, \overline{\Delta H}_1, \log(\overline{N}_1), \log(\overline{N}_2)\right)$.

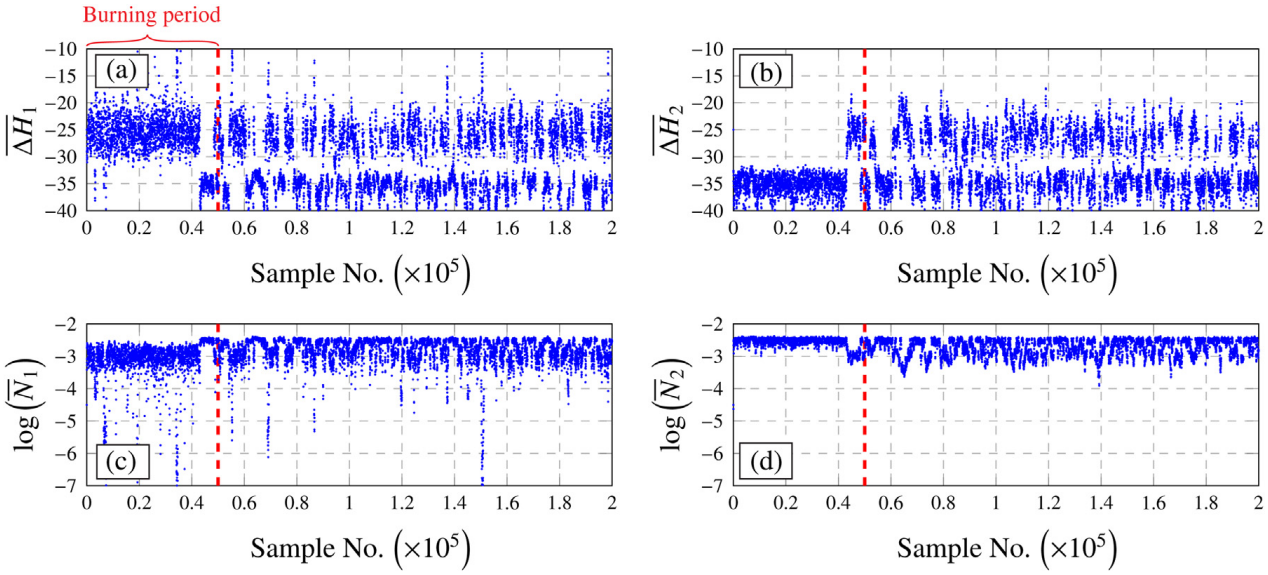

Fig. 8. Markov chains generated by DRAM MCMC of trap parameters $\overline{\Delta H}_1$, $\overline{\Delta H}_2$, $\log(\overline{N}_1)$, $\log(\overline{N}_2)$ for TDS Experiment EI.

(note later in Fig. 11 that such a noise level represents a considerably low signal-to-noise ratio). The experiment EI was defined to illustrate the potentials of the implemented DRAM MCMC to draw samples from a multi-modal distribution. Note that the PDE in Eq. (31) does not differentiate the order of the traps, thereby the problem is ill-posed and the posterior distribution is expected to exhibit two modes corresponding to two symmetric solutions. Instead, the experiment EII was designed to account for a trap ($\overline{\Delta H}_1 = -15$, $\log(\overline{N}_1) = -6$) in the regime with no flux as identified by Raina and et al. [76], while the second trap ($\overline{\Delta H}_2 = -30$, $\log(\overline{N}_2) = -3$) represents a deep trap. Therefore, the PDF in this case should be uni-modal.

In the inference analyses, uninformative uniform priors $\mathcal{U}(-40, -10)$ and $\mathcal{U}(-7, -2)$ are selected for $\overline{\Delta H}_i$ and $\log(\overline{N}_i)$ $(i = 1, 2)$, respectively. A total number of 200 000 samples with a burning time of 50 000 samples were drawn by the previously introduced Bayesian inference approach for EI. The sampling of the posterior PDF in Experiment EII was more challenging given its uni-modal nature with large regions of low probability, requiring up to 480,000 samples with a burning period of 160,000 samples to achieve convergence. Interestingly, this phenomenon attenuates when the flux curve is affected by noise, only requiring a chain of 120,000 samples with a burning period of 30,000 to attain convergence. This is expectable since the noise-induced lower probability concentration around the exact true solutions makes it easier for the chain to span from one solution to the symmetric one. The initial location state was defined as $\boldsymbol{\theta}_0 = (-25, -25, -4.5, -4.5)$, while the prediction error was set to $\sigma_{\boldsymbol{\varepsilon}} = 1E-9$ and $1.6E-5$ for the noise unpolluted and polluted cases, respectively. After some initial calibration by visual inspection of the chain traces, a diagonal covariance matrix with entries equal $\left(0.05 \cdot \boldsymbol{\theta}_0\right)^2$ was initially defined for the Gaussian proposal. In the AM step the proposal distribution was scaled by a factor $s_d = 2.4^2/d$ and the non-adaptation period n_0 was set to 500 and 4000 for the EI and EII experiments, respectively. On the other hand, in the DR step the proposal is scaled down by a factor of 0.2.

The Markov chain and the joint posterior PDF obtained for Experiment EI are presented in Figs. 8 and 9, respectively. As anticipated, the problem is ill-posed and there exist two potential solutions, namely $\boldsymbol{\theta} = (-25, -35, -3, -2.5)$ and $\boldsymbol{\theta} = (-35, -25, -2.5, -3)$. This manifests in the marginal PDFs in Fig. 9. Specifically, the PDFs corresponding to parameters $\overline{\Delta H}_1$ and $\overline{\Delta H}_2$ have two identical modes at -35 and -25, and parameters $\log(\overline{N}_1)$ and $\log(\overline{N}_2)$ have two modes at -3 and -2.5. It is observed in Fig. 9 that, indeed, the implemented DRAM algorithm is capable of exploring the two modes in the distribution, without getting stuck around one of them as it is usually the case when implementing standard MCMC methods. For validation purposes, the posterior PDF has been also computed by direct integration of the forward solution. To do so, the evidence of the model has been computed over a mesh of 60^4 elements. This required forty five hours of parallel computation on ten cores, while the MCMC approach only required about four hours on a single core. The Highest Density Regions (HDRs) at the 80% and 50% level of both distributions are reported in Table 6 The close fittings between the exact marginal PDFs and those predicted by the surrogate model-based Bayesian inference in Fig. 9 demonstrate the accuracy of the developed approach, as it is also evident from the computed HDRs in Table 6. Finally, the Markov chain, and the posterior PDF obtained for the TDS experiment EII are reported in Figs. 10 and 11, respectively, and the posterior HDR values are reported in Table 7. In this case, the PDFs exhibit one single mode as previously anticipated. This corresponds to the shallow trap ($\overline{\Delta H}_1 = -30$, $\log(\overline{N}_1) = -3$), while the trap in the no-flux regime goes unnoticed. From a Bayesian perspective, this represents an observability limitation of the experiment, being the model of one single trap more likely to represent the material given the experimental evidence. Furthermore, it is noted that the presence of measurement noise does not substantially alter the inference outcome. The modes of the posteriors for the trap densities parameters $\overline{\Delta H}_1$ and

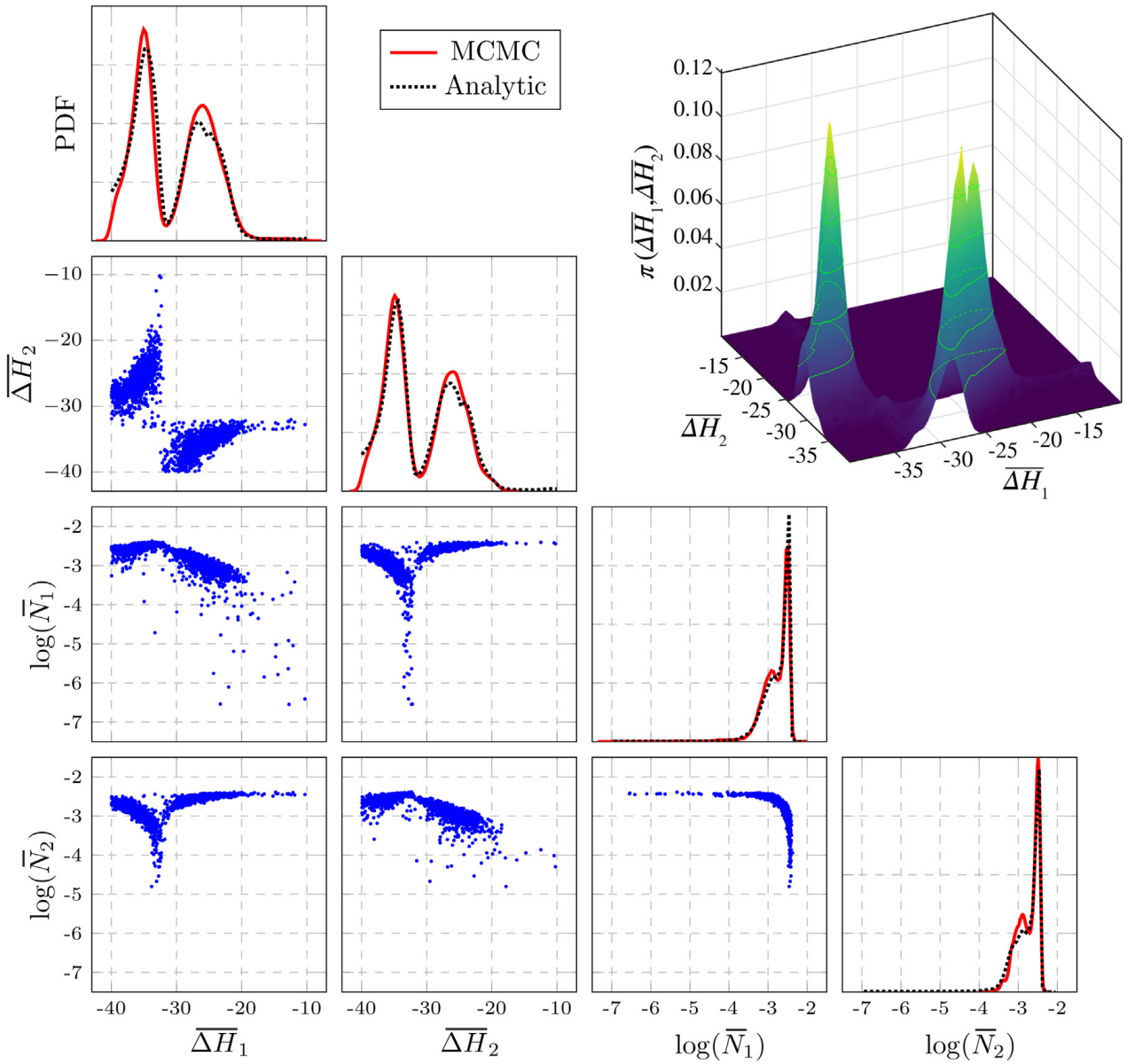

Fig. 9. Bayesian identification results of the trap parameters $\boldsymbol{\theta} = \left(\overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2)\right)$ of TDS Experiment EI. The surface plot in the top right corner corresponds to the marginal PDF over $\left(\overline{\Delta H}_1, \overline{\Delta H}_2\right)$ obtained by numerical integration.

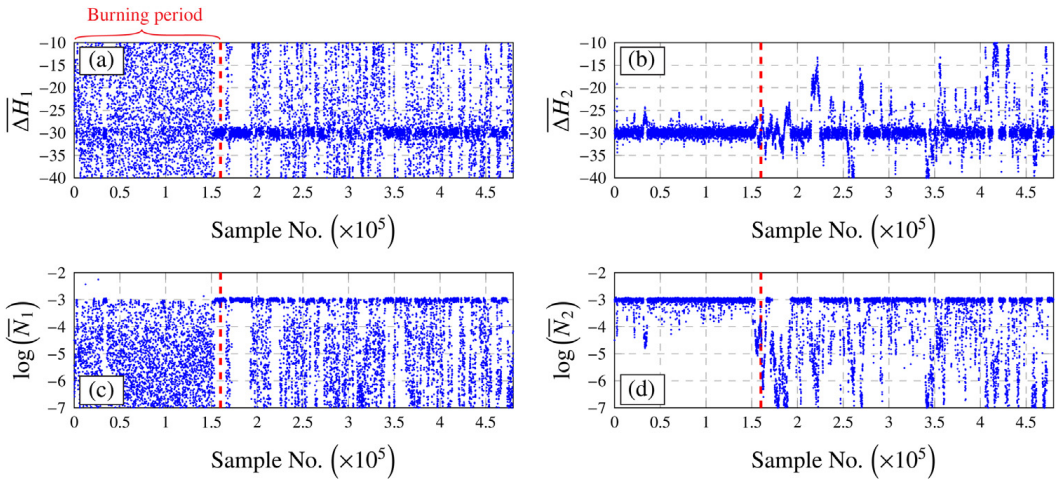

Fig. 10. Markov chains generated by DRAM MCMC of trap parameters $\overline{\Delta H}_1$, $\overline{\Delta H}_2$, $\log(\overline{N}_1)$, $\log(\overline{N}_2)$ for TDS Experiment EII under noise-free data conditions.

Table 6
HDR at 80% and 50% of the PDFs obtained by direct integration and by MCMC for experiment EI.

HDR	$\overline{\Delta H}_1$	$\overline{\Delta H}_2$	$\log(\overline{N}_1)$	$\log(\overline{N}_2)$
80% HDR (MCMC)	$(-38.004,-33.084) \cup (-28.931,-22.606)$	$(-37.833,-32.686) \cup (-28.628,-23.134)$	$(-3.075,-2.418)$	$(-3.059,-2.736) \cup (-2.683,-2.409)$
80% HDR (Analytical)	$(-38.515,-32.253) \cup (-29.121,-22.729)$	$(-38.179,-32.247) \cup (-29.184,-22.991)$	$(-3.069,-2.386)$	$(-3.099,-2.389)$
50% HDR (MCMC)	$(-36.726,-33.659) \cup (-27.398,-24.842)$	$(-36.546,-33.280) \cup (-27.143,-25.164)$	$(-2.925,-2.869) \cup (-2.653,-2.427)$	$(-2.898,-2.868) \cup (-2.637,-2.416)$
50% HDR (Analytical)	$(-36.884,-32.840) \cup (-27.295,-24.816)$	$(-36.680,-32.899) \cup (-27.424,-25.208)$	$(-2.702,-2.406)$	$(-2.703,-2.399)$

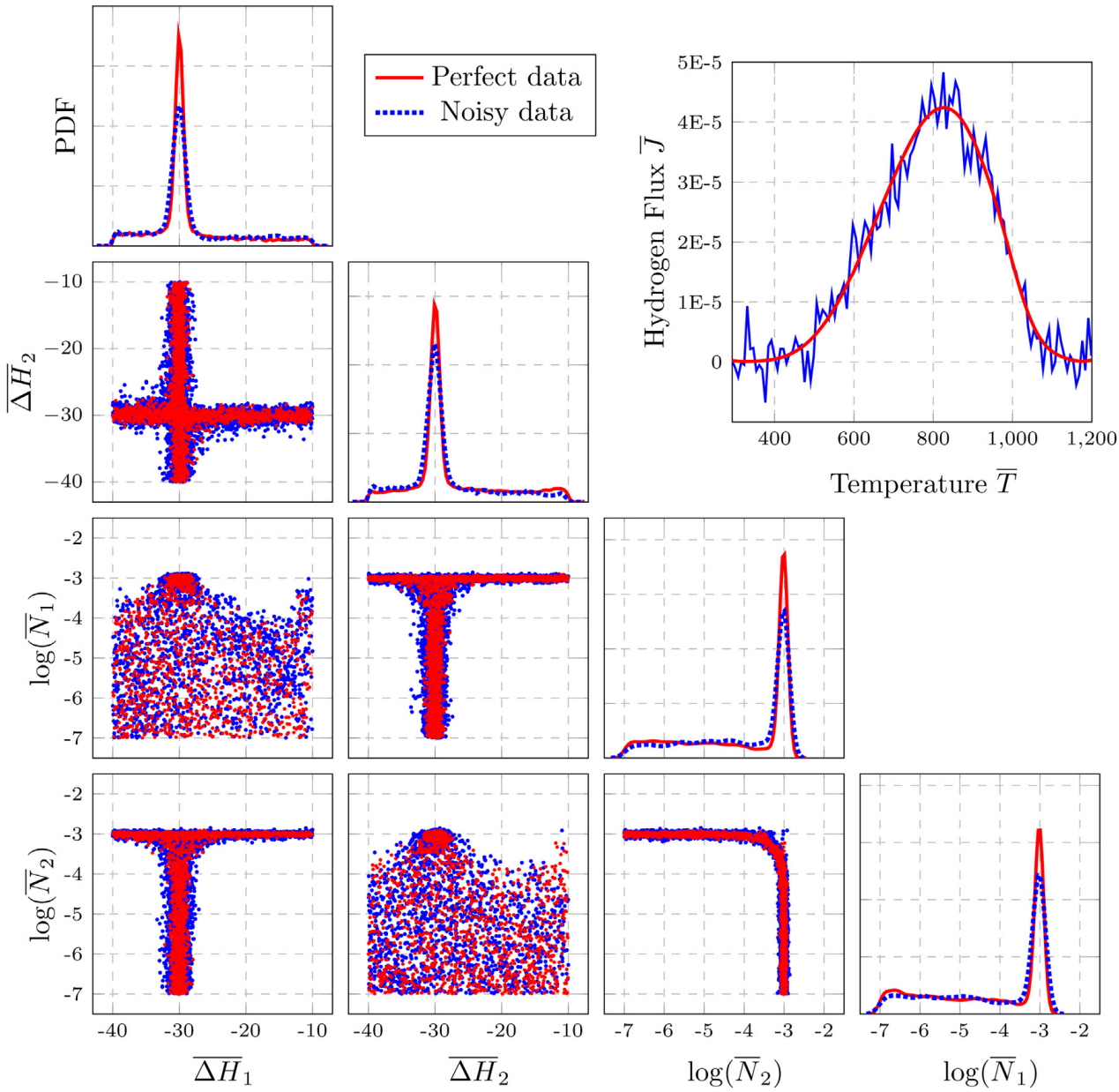

Fig. 11. Bayesian identification results of the trap parameters $\boldsymbol{\theta} = \left(\overline{\Delta H}_1, \overline{\Delta H}_2, \log(\overline{N}_1), \log(\overline{N}_2)\right)$ of experiment EII with noise unpolluted (red) and polluted data (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
HDR at 80% and 50% of the PDFs obtained from noisy and noise-free data by MCMC for experiment EII.

HDR	$\overline{\Delta H}_1$	$\overline{\Delta H}_2$	$\log(\overline{N}_1)$	$\log(\overline{N}_2)$
80% HDR (noise-free data)	(−39.872, −22.145)	(−34.206,−25.128)	(−5.887,−2.927)	(−5.346,−2.920)
80% HDR (noisy data)	(−39.970,−26.221)	(−39.314,−24.378)	(−5.410,−2.900)	(−5.733,−2.889)
50% HDR (noise-free data)	(−31.526,−28.406)	(−30.841,−29.199)	(−3.886,−2.924)	(−3.088,−2.946)
50% HDR (noisy data)	(−31.534,−28.698)	(−31.576,−28.656)	(−3.307,−2.896)	(−3.286,−2.883)

$\overline{\Delta H}_2$ of the noise-free scenario are −29.966 and −30.002, while for the noisy scenario the values −29.989 and −29.961 are obtained, which represents a difference of 0.077% and 0.137%, respectively. On the other hand, for parameters $\log(\overline{N}_1)$ and $\log(\overline{N}_2)$ the modes of the posteriors in the noise-free case are −3.009 and −3.016, whereas in the noisy scenario they take values −3.017 and −3.019, meaning a difference of 0.266% and 0.010%, respectively. This confirms that the proposed approach is robust to the presence of measurement noise. Overall, these results illustrate the potential of the developed approach for model selection and information gain analysis of TDS results, which are left for future developments.

4. Conclusions

This work presents the development of a multi-element PCK meta-model for surrogate model-based Bayesian parameter inference of highly nonlinear engineering models. The proposed metamodel combines adaptive sparse PCE and Kriging metamodelling to attain both global and local prediction capabilities. The optimal order of the polynomials in the PCE is automatically identified by the LAR algorithm. Then, the optimal PCE is inserted into a Kriging predictor as the trend term, while the stochastic term is fitted through GA optimization. With the aim of tackling non-smoothness in the forward model, a simple regular block partitioning approach has been implemented. On this basis, the space domain is split into a discrete number of subsets where local surrogate models are constructed. Then, the global model response is obtained by combining the local metamodels in a piecewise fashion. Finally, the surrogate model is used for Bayesian parameter estimation using a cost-efficient DRAM MCMC with DR and AM capabilities. The effectiveness of the proposed approach has been validated through two benchmark case studies: (i) the analytical Drop-Wave function; (ii) and a PDE for TDS tests. Key findings and contributions of this work include:

- Optimal surrogate models ought to be defined by preliminary parametric analyses accounting for prediction accuracy and computational cost. The latter is particularly critical when performing computationally intense applications such as Bayesian parameter estimation. To this aim, this work has presented a set of error metrics and a methodological discussion through two validation case studies.
- The analysis of the Salomon function has shown that the proposed multi-element PCK model with regular block partitioning provides similar accuracy ($R^2 > 0.99$, NMAE $< 10^{-4}$) as the (classical) PCK approach, while achieving 100 and 10 times shorter evaluation and construction times, respectively. Moreover, the presented results have shown that the proposed method outperforms the SSE technique for the analysis of such a highly nonlinear surface, requiring 20 times fewer samples to achieve a comparable accuracy.
- The size of the ED and the number of domain partitions critically determine the computational cost of the developed sparse PCE-Kriging metamodel. Specifically, the partition of non-smooth problems into a finite set of sub-domains allows the sparse adaptive PCE to eliminate a considerable number of high-order components through LAR, so achieving important savings in the construction of the Kriging model and the evaluation of the resulting metamodel.
- The developed surrogate model-based DRAM MCMC approach allows to conduct fast Bayesian parameter inference. In particular, the proposed approach has been applied to the identification of micro-structural traps in metallic alloys subject to TDS. The hydrogen fluxes obtained in TDS test represent a considerable challenge in surrogate modelling due to the presence of diverse regimes depending on the configuration of the hydrogen traps. In terms of R^2, the proposed approach is capable of reproducing more than 99.9% of the hydrogen diffusion TDS model with computational time savings of 99.3% with respect to the forward numerical model.
- The presented analyses evidence the potential of the developed approach for conducting inverse characterisation of hydrogen-metal interactions. The accuracy of the proposed PCK surrogate model in conjunction with DRAM MCMC opens vast possibilities for future applications in model selection, and information gain analysis of TDS hydrogen desorption tests.

Despite its simplicity, the adopted regular block partitioning model has demonstrated significant performance in terms of computational savings. In this respect, future research will involve the development of more efficient partitioning algorithms that would allow the sampling effort to be localised where the forward model presents greater non-linearities, thus achieving similar accuracies with smaller sample sizes. Another interesting goal for future work consists in the development of multielement surrogate PCK-based models capable of dealing with discontinuities in the response surface.

Data availability

Data will be made available on request.

Acknowledgements

This work has been partially supported through the Ministerio de Ciencia e Innovación [PID2020-116809GB-I00] of Spain and from the Junta de Extremadura through the Research Group Grant [GR18023]. E. García-Macías was supported by the Consejería de Transformación Económica, Conocimiento Empresas y Universidades de la Junta de Andalucía (Spain) through the reseacrh project [P18-RT-3128]. E. Martínez-Pañeda acknowledges financial support from the EPSRC [Grant EP/V009680/1] and from UKRI's Future Leaders Fellowship programme [Grant MR/V024124/1].

References

[1] Q. Chen, C. Yang, Hybrid algorithm for multi-objective optimization design of parallel manipulators, Appl. Math. Model. 98 (2021) 245–265.
[2] X. Sun, Y.Y. Choi, J.I. Choi, Global sensitivity analysis for multivariate outputs using polynomial chaos-based surrogate models, Appl. Math. Model. 82 (2020) 867–887.
[3] E. García-Macías, I. Venanzi, F. Ubertini, Metamodel-based pattern recognition approach for real-time identification of earthquake-induced damage in historic masonry structures, Autom. Constr. 120 (2020) 103389.

[4] L. Hong, H. Li, K. Peng, A combined radial basis function and adaptive sequential sampling method for structural reliability analysis, Appl. Math. Model. 90 (2021) 375–393.
[5] Y. Yuan, A. Jones, R. Setchfield, C.W. Schwingshackl, Robust design optimisation of underplatform dampers for turbine applications using a surrogate model, J. Sound Vib. 494 (2021) 115528.
[6] J. Stork, M. Friese, M. Zaefferer, T. Bartz-Beielstein, A. Fischbach, B. Breiderhoff, B. Naujoks, T. Tušar, et al., Open issues in surrogate-assisted optimization, in: T. Bartz-Beielstein, et al. (Eds.), High-Performance Simulation-Based Optimization, Springer, 2020, pp. 225–244.
[7] J. Hao, W. Ye, L. Jia, G. Wang, J. Allen, Building surrogate models for engineering problems by integrating limited simulation data and monotonic engineering knowledge, Adv. Eng. Inform. 49 (2021) 101342.
[8] S. Chakraborty, S. Adhikari, R. Ganguli, The role of surrogate models in the development of digital twins of dynamic systems, Appl. Math. Model. 90 (2021) 662–681.
[9] H. Sun, H.V. Burton, H. Huang, Machine learning applications for building structural design and performance assessment: state-of-the-art review, J. Build. Eng. 33 (2021) 101816.
[10] M.J. Asher, B.F.W. Croke, A.J. Jakeman, L.J.M. Peeters, A review of surrogate models and their application to groundwater modeling, Water Resour. Res. 51 (8) (2015) 5957–5973.
[11] B.M. de Gooijer, J. Havinga, H.J.M. Geijselaers, A.H. van den Boogaard, Evaluation of pod based surrogate models of fields resulting from nonlinear FEM simulations, Adv. Model. Simul. Eng. Sci. 8 (1) (2021) 1–33.
[12] L. Zhang, Y. Wu, P. Jiang, S.K. Choi, Q. Zhou, A multi-fidelity surrogate modeling approach for incorporating multiple non-hierarchical low-fidelity data, Adv. Eng. Inform. 51 (2022) 101430.
[13] Y. Shi, Z. Lu, L. Xu, S. Chen, An adaptive multiple-Kriging-surrogate method for time-dependent reliability analysis, Appl. Math. Model. 70 (2019) 545–571.
[14] J.P.C. Kleijnen, Kriging metamodeling in simulation: a review, Eur. J. Oper. Res. 192 (3) (2009) 707–716.
[15] M.D. Buhmann, Radial basis functions, Acta Numer. 9 (2000) 1–38.
[16] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[17] G.B. Kingston, M. Rajabalinejad, B.P. Gouldby, P.H.A.J.M. Van Gelder, Computational intelligence methods for the efficient reliability analysis of complex flood defence structures, Struct. Saf. 33 (1) (2011) 64–73.
[18] E. Schulz, M. Speekenbrink, A. Krause, A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions, J. Math. Psychol. 85 (2018) 1–16.
[19] R. Schobi, B. Sudret, J. Wiart, Polynomial-chaos-based Kriging, Int. J. Uncertain. Quantif. 5 (2) (2015) 59–63.
[20] R. Alizadeh, J.K. Allen, F. Mistree, Managing computational complexity using surrogate models: a critical review, Res. Eng. Des. 31 (3) (2020) 275–298.
[21] N.V. Queipo, R.T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P.K. Tucker, Surrogate-based analysis and optimization, Prog. Aerosp. Sci. 41 (1) (2005) 1–28.
[22] J.N. Fuhg, A. Fau, U. Nackenhorst, State-of-the-art and comparative review of adaptive sampling methods for Kriging, Arch. Comput. Methods Eng. 28 (2021) 2689–2747.
[23] S.S. Garud, I.A. Karimi, M. Kraft, Design of computer experiments: a review, Comput. Chem. Eng. 106 (2017) 71–95.
[24] H. Liu, Y.S. Ong, J. Cai, A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design, Struct. Multidiscip. Optim. 57 (1) (2018) 393–416.
[25] O. Le Maître, H.N. Najm, R.G. Ghanem, O.M. Knio, Multi-resolution analysis of wiener-type uncertainty propagation schemes, J. Comput. Phys. 197 (2) (2004) 502–531.
[26] A. Menafoglio, G. Gaetani, P. Secchi, Random domain decompositions for object-oriented Kriging over complex domains, Stoch. Environ. Res. Risk Assess. 32 (12) (2018) 3421–3437.
[27] A. Resmini, J. Peter, D. Lucor, Sparse grids-based stochastic approximations with applications to aerodynamics sensitivity analysis, Int. J. Numer. Methods. Eng. 106 (1) (2016) 32–57.
[28] S.A. Mattis, T. Butler, Enhancing piecewise-defined surrogate response surfaces with adjoints on sets of unstructured samples to solve stochastic inverse problems, Int. J. Numer. Methods. Eng. 119 (10) (2019) 923–940.
[29] Y.S. Ong, P.B. Nair, A.J. Keane, Evolutionary optimization of computationally expensive problems via surrogate modeling, AIAA J. 41 (4) (2003) 687–696.
[30] R.P. Liem, C.A. Mader, J.R.R.A. Martins, Surrogate models and mixtures of experts in aerodynamic performance prediction for aircraft mission analysis, Aerosp. Sci. Technol. 43 (2015) 126–151.
[31] R. Teixeira, M. Nogal, A. O'Connor, Adaptive approaches in metamodel-based reliability analysis: a review, Struct. Saf. 89 (2021) 102019.
[32] J. Foo, X. Wan, G.E. Karniadakis, The multi-element probabilistic collocation method (ME-PCM): error analysis and applications, J. Comput. Phys. 227 (22) (2008) 9572–9595.
[33] C. Rasmussen, Z. Ghahramani, Infinite mixtures of gaussian process experts, Adv. Neural Inf. Process. Syst. 14 (2001) 881–888.
[34] B.A. Konomi, A.A. Hanandeh, M. Pulong, E.L. Kang, Computationally efficient nonstationary nearest-neighbor gaussian process models using data-driven techniques, Environmetrics 30 (8) (2019) e2571.
[35] S. Marelli, P.R. Wagner, C. Lataniotis, B. Sudret, Stochastic spectral embedding, Int. J. Uncertain. Quantif. 11 (2) (2021).
[36] P. Wagner, S. Marelli, B. Sudret, Bayesian model inversion using stochastic spectral embedding, J. Comput. Phys. 436 (2021) 110141.
[37] Y. Huang, C. Shao, B. Wu, J.L. Beck, H. Li, State-of-the-art review on Bayesian inference in structural system identification and damage assessment, Adv. Struct. Eng. 22 (6) (2019) 1329–1351.
[38] W. Liu, X. Luo, F. Huang, M. Fu, Prediction of soil water retention curve using Bayesian updating from limited measurement data, Appl. Math. Model. 76 (2019) 380–395.
[39] J.M. Emery, M.D. Grigoriu, R.V. Field Jr., Bayesian methods for characterizing unknown parameters of material models, Appl. Math. Model. 40 (13–14) (2016) 6395–6411.
[40] P. Hauseux, J.S. Hale, S. Cotin, S.P.A. Bordas, Quantifying the uncertainty in a hyperelastic soft tissue model with stochastic parameters, Appl. Math. Model. 62 (2018) 86–102.
[41] R. Dwivedi, Y. Chen, M.J. Wainwright, B. Yu, Log-concave sampling: metropolis-hastings algorithms are fast!, J. Mach. Learn. Res. 20 (183) (2019) 1–42.
[42] S.H. Cheung, S. Bansal, A new Gibbs sampling based algorithm for Bayesian model updating with incomplete complex modal data, Mech. Syst. Signal Process. 92 (2017) 156–172.
[43] Y. Che, X. Wu, G. Pastore, W. Li, K. Shirvan, Application of Kriging and variational Bayesian Monte Carlo method for improved prediction of doped UO2 fission gas release, Ann. Nucl. Energy 153 (2021) 108046.
[44] P. Del Moral, A. Doucet, A. Jasra, Sequential monte carlo samplers, J. R. Stat. Soc. Ser. B Methodol. 68 (3) (2006) 411–436.
[45] J. Ching, Y.C. Chen, Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, J. Eng. Mech. 133 (7) (2007) 816–832.
[46] T. Yin, H. Zhu, An efficient algorithm for architecture design of Bayesian neural network in structural model updating, Computer-Aided Civ. Infrastruct. Eng. 35 (4) (2020) 354–372.
[47] F. Schneider, I. Papaioannou, D. Straub, C. Winter, G. Müller, Bayesian parameter updating in linear structural dynamics with frequency transformed data using rational surrogate models, Mech. Syst. Signal Process. 166 (2022) 108407.
[48] W.W. Xing, A.A. Shah, P. Wang, S. Zhe, Q. Fu, R.M. Kirby, Residual Gaussian process: a tractable nonparametric Bayesian emulator for multi-fidelity simulations, Appl. Math. Model. 97 (2021) 36–56.
[49] L. Ierimonti, N. Cavalagli, I. Venanzi, E. García-Macías, F. Ubertini, A transfer Bayesian learning methodology for structural health monitoring of monumental structures, Eng. Struct. 247 (2021) 113089.

[50] A. del Val, O.P. Le Maître, T.E. Magin, O. Chazot, P.M. Congedo, A surrogate-based optimal likelihood function for the Bayesian calibration of catalytic recombination in atmospheric entry protection materials, Appl. Math. Model. 101 (2022) 791–810.
[51] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.
[52] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[53] H. Haario, M. Laine, A. Mira, E. Saksman, DRAM: efficient adaptive MCMC, Stat. Comput. 16 (4) (2006) 339–354.
[54] K.K. Phoon, J. Ching, Risk and Reliability in Geotechnical Engineering, CRC Press, Boca Raton, FL, USA, 2015.
[55] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Stat. Sci. 4 (4) (1989) 409–423.
[56] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[57] H. Cham, E. Reshetnyak, B. Rosenfeld, W. Breitbart, Full information maximum likelihood estimation for latent variable interactions with incomplete indicators, Multivar. Behav. Res. 52 (1) (2008) 12–30.
[58] J. Kleffe, J.N.K. Rao, Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model, J. Multivar. Anal. 43 (1) (1992) 1–15.
[59] G. Pellegrini, Polynomial Chaos Expansion with Applications to PDEs, University of Verona, 2014 Ph.D. thesis.
[60] G.K. W. Xiaoliang, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM J. Sci. Comput. 28 (3) (2006) 901–928.
[61] M. Moustapha, J.M. Bourinet, B. Guillaume, B. Sudret, Comparative study of Kriging and support vector regression for structural engineering applications, ASCE-ASME J. Risk Uncertain. Eng. Syst. A 4 (2) (2018) 04018005.
[62] K. Ritter, Average-Case Analysis of Numerical Problems, Springer, 2000.
[63] W. Wang, R. Tuo, C.F. Jeff Wu, On prediction properties of Kriging: uniform error bounds and robustness, J. Am. Stat. Assoc. 115 (530) (2020) 920–930.
[64] A. Mira, On metropolis-hastings algorithms with delayed rejection, Metron 59 (3–4) (2001) 231–241.
[65] H. Haario, E. Saksman, J. Tamminen, An adaptive metropolis algorithm, Bernoulli 7 (2) (2001) 223–242.
[66] D.C. Stenning, D.A. van Dyk, Bayesian Statistical Methods For Astronomy Part II: Markov Chain Monte Carlo, STATS Astrophysics. EDP Sciences, 2021.
[67] R. Salomon, Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions. a survey of some theoretical and practical aspects of genetic algorithms, Biosystems 39 (3) (1996) 263–278.
[68] S. Marelli, B. Sudret, Uqlab: a framework for uncertainty quantification in matlab, in: The 2nd International Conference on Vulnerability and Risk Analysis and Management, 2014, pp. 2554–2563.
[69] R.P. Gangloff, Hydrogen-assisted Cracking, in: I. Milne, R. Ritchie, B. Karihaloo (Eds.), Comprehensive Structural Integrity, vol. 6, Elsevier Science, New York, NY, 2003, pp. 31–101.
[70] S.K. Dwivedi, M. Vishwakarma, Hydrogen embrittlement in different materials: a review, Int. J. Hydrog. Energy. 43 (46) (2018) 21603–21616.
[71] W.H. Johnson, On some remarkable changes produced in iron and steel by the action of hydrogen and acids, Proc. R. Soc. Lond. 23 (1875) 168–179.
[72] R. Fernández-Sousa, C. Betegón, E. Martínez-Pañeda, Analysis of the influence of microstructural traps on hydrogen assisted fatigue, Acta Mater. 199 (2020) 253–263.
[73] M. Isfandbod, E. Martínez-Pañeda, A mechanism-based multi-trap phase field model for hydrogen assisted fracture, Int. J. Plast. 144 (2021) 103044.
[74] A. Zafra, Z. Harris, C. Sun, E. Martinez-Pañeda, Comparison of hydrogen diffusivities measured by electrochemical permeation and temperature-programmed desorption in cold-rolled pure iron, J. Nat. Gas Sci. Eng. 98 (2022) 104365.
[75] F.J. Castro, G. Meyer, Thermal desorption spectroscopy (TDS) method for hydrogen desorption characterization (I): theoretical aspects, J. Alloys Compd. 330 (2002) 59–63.
[76] A. Raina, V.S. Deshpande, N.A. Fleck, Analysis of thermal desorption of hydrogen in metallic alloys, Acta Mater. 144 (2018) 777–785.
[77] A. McNabb, P.K. Foster, A new analysis of diffusion of hydrogen in iron and ferrite, Trans. Met. Soc. 227 (3) (1963) 618–627.
[78] R.A. Oriani, The diffusion and trapping of hydrogen in steel, Acta Metall. 18 (1) (1970) 147–157.

Appendix C

Paper C: Sparse polynomial chaos expansion for universal stochastic kriging

Journal of Computational and Applied Mathematics 444 (2024) 115794

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Sparse polynomial chaos expansion for universal stochastic kriging

J.C. García-Merino [a,*], C. Calvo-Jurado [a], E. García-Macías [b]

[a] Departamento de Matemáticas, Escuela Politécnica, Universidad de Extremadura, 10003, Cáceres, España
[b] Department of Structural Mechanics and Hydraulic Engineering, University of Granada, Campus de Fuentenueva s/n, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Kriging
Least angle regression
Polynomial chaos expansion
Stochastic simulation
Surrogate modelling
Uncertainty quantification

ABSTRACT

Surrogate modelling techniques have opened up new possibilities to overcome the limitations of computationally intensive numerical models in various areas of engineering and science. However, while fundamental in many engineering applications and decision-making, the incorporation of uncertainty quantification into meta-models remains a challenging open area of research. To address this issue, this paper presents a novel stochastic simulation approach combining sparse polynomial chaos expansion (PCE) and Stochastic Kriging (SK). Specifically, the proposed approach adopts adaptive sparse PCE as the trend model in SK, achieving both global and local prediction capabilities and maximizing the role of the stochastic term to conduct uncertainty quantification. To maximize the generalization and computational efficiency of the meta-model, the Least Angle Regression (LAR) algorithm is adopted to automatically select the optimal polynomial basis in the PCE. The computational effectiveness and accuracy of the proposed approach are appraised through a comprehensive set of case studies and different quality metrics. The presented numerical results and discussion demonstrate the superior performance of the proposed approach compared to the classical ordinary SK model, offering high flexibility for the characterization of both extrinsic and intrinsic uncertainty for a wide variety of problems.

1. Introduction

Computer models, commonly referred to as simulators, have become ubiquitous assets in most areas of engineering and science. In cases where systems or processes are assumed to be predictable and devoid of inherent randomness, a deterministic simulator can be adopted. Nonetheless, deterministic approaches appear inadequate for facing real-world problems characterized by intrinsic or non-parameterizable randomness, for which the adoption of stochastic simulators becomes necessary for proper modelling. A stochastic simulator incorporates randomness or uncertainty in its predictions. Thus, while a deterministic simulator always produces the same prediction for a fixed input, in stochastic simulators, each simulation run may produce different results as a result of the inherent randomness in the model. Such models have gained significant popularity with manifold applications in diverse fields such as epidemiology [1,2], materials science [3,4], or reliability analysis [5], enabling researchers to make robust predictions to assist decision-making for complex systems. However, for both deterministic and stochastic computationally demanding simulators, formidable challenges arise when implemented in iterative procedures such as optimization [6], sensitivity analysis [7], parameter inference [8], or uncertainty propagation [9]. Despite the introduction of numerous surrogate models or meta-models in the literature over the past few years to replace intensive numerical models and alleviate the computational burden, their success mostly limits to deterministic simulators, while the consideration of stochasticity remains an open research topic.

* Corresponding author.
E-mail address: jcgarcia@unex.es (J.C. García-Merino).

https://doi.org/10.1016/j.cam.2024.115794
Received 24 July 2023; Received in revised form 18 December 2023
Available online 26 January 2024

There is a vast literature on the development of surrogate models for deterministic simulators [10,11]. In particular, non-intrusive models such as polynomial chaos expansion (PCE) [12], radial basis functions [13] or Kriging [14] have received significant attention from the scientific community due to their effectiveness in addressing resource-intensive simulations. The construction of these meta-models is usually conducted on the basis of a training dataset, also referred to as the experimental design (ED) [14]. Such a training dataset contains the evaluations of the forward model, representing the quantity of interest (QoI), for a set of combinations of the input parameters covering the parameter space. In the context of stochastic simulators, the QoI usually presents a certain level of uncertainty, stemming from the stochastic nature of the model parameters and/or the presence of epistemic uncertainties in the model predictions [15]. A first approach to meta-model such processes involves estimating the probability density function (PDF) of the simulator's predictions through functional decomposition [16]. Alternatively, Stochastic Kriging (SK) [17] has gained significant recognition as a general meta-modelling tool for representing stochastic simulation response surfaces. Defined as an extension of Kriging, SK accounts for heteroscedastic noise in the response by estimating the sample variance at each point of the ED using multiple runs of the simulator for each specific input. The effectiveness of SK has been demonstrated on a broad range of applications, including risk and reliability analysis [18,19], industrial welding processes [20], and game theoretic simulations [21], to mention a few.

One of the primary challenges encountered in SK is the requirement of a significant number of replications at each ED point to accurately estimate the sample variance [15], which often leads to serious computational demands. One approach to alleviate such a limitation regards the optimal definition of the ED. In the case of stochastic simulators, sampling efforts are divided into two parts: (i) sampling new ED locations, and (ii) acquiring additional replications at existing points. In this light, numerous sequential sampling techniques have been presented (refer to [22,23] for an extensive state-of-the-art review). On the other hand, various extensions of the classical SK have been proposed in the literature. It is worth noting the work by Binois et al. [24], who proposed modelling the input-dependent noise as a separate Gaussian process. Additionally, they employed computationally efficient techniques such as the Woodbury identity to address the challenges encountered in SK for large EDs. In a similar vein, Hao et al. [19] introduced a Nested SK model, which decouples in an iterative way the response noise parameters from the deterministic Kriging parameters used for calibrating the meta-model. In a similar context, Cheng et al. [25] proposed a generalized polynomial chaos-stochastic Kriging (gPCE SK) approach, which combines the benefits of PCE and SK. The gPCE SK method proposed by those authors utilizes a local PCE surrogate model at each design point, which is calibrated using a small number of simulation runs. By relying on these local surrogate models, numerous replications can be performed at each training point, significantly reducing the computational burden involved in the calibration of the SK model. However, it is important to note that this model assumes a distinction between controllable design variables and non-controllable noise variables, the latter causing the variations between replicates at each fixed design point. Furthermore, it relies on knowledge of the PDFs of the uncontrollable variables, which may be a strong assumption in many real-life applications.

In light of the discussion above, it is apparent that the development and application of surrogate modelling approaches for stochastic simulators have experienced considerable advances in recent years. Nevertheless, there are some aspects of SK that remain not fully addressed in the literature. Notably, while most applications of SK in the literature are limited to the use of constant trend models (Ordinary Kriging), the use of more advanced models for Universal SK remains largely unexplored in the literature, to the best of the author's knowledge. To address this gap, this paper presents an extension of the deterministic Polynomial-Chaos Kriging (PCK) method originally presented in [26] for the meta-modelling of stochastic simulators with predictions contaminated by noise. The PCK meta-model is a powerful technique that combines PCE with Kriging to develop highly efficient surrogate models for complex systems. Notably, a prior work by the authors [27] demonstrated the ability of PCK meta-modelling to successfully capture the behaviour of highly non-linear functions. Drawing from this positive outcome, the current study aims to apply the principles of PCK to the realm of stochastic simulators. Specifically, the proposed meta-model involves a Universal SK model with an adaptive sparse PCE model as the trend term. The orthonormal polynomial basis of the PCE is optimally selected by a model selection technique for sparse linear models, the least-angle regression (LAR) algorithm set out by Efron [28]. Afterwards, the adjusted LAR-PC model is inserted as the trend term in a Universal SK model, and the hyper-parameters of the resulting meta-model are calibrated through a genetic algorithm (GA) global optimization approach. The effectiveness of the proposed approach is validated through a set of well-known benchmark case studies in the surrogate modelling literature. The presented results and discussion demonstrate the potential of the proposed approach and the importance of selecting a suitable trend model in the context of SK.

The remainder of this paper is organized as follows. Section 2 outlines the theoretical formulation of Kriging and SK respectively. Section 3 provides a concise review of PCE and LAR and illustrates the proposed LAR-PCE SK approach. Section 4 presents the numerical results and discussion and, finally, Section 5 closes the paper with concluding remarks and potential future developments.

2. Surrogate modelling: Deterministic and Stochastic Kriging

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, where Ω denotes the event space equipped with a σ-algebra $\mathcal{F}$ of subsets of Ω and the probability measure μ. Given $\Omega \subset \mathbb{R}^M$ as a connected set, let $\mathcal{M} : \Omega \subset \mathbb{R}^M \rightarrow \mathbb{R}$ be a computational model mapping an M-dimensional vector of input (design) variables $\boldsymbol{x} = \left[x_1, \ldots, x_M\right]^{\mathrm{T}}$ into the output variable $y \in \mathbb{R}$ (response surface or QoI). The aim of a surrogate model is to replace the original model $\mathcal{M}$ by an approximated one $\mathcal{M}^*$, whose evaluation needs lower computational cost. When defined in a non-intrusive framework, the hyper-parameters of $\mathcal{M}^*$ must be calibrated on a basis of a certain training dataset. Such a training dataset consists in the realization of the forward model $\mathcal{M}$ over a set ED $= \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}\right\}$ of k samples of the input variable $\boldsymbol{x}$ covering the design space, often referred to as the *experimental design* (ED). Note that, when the model is deterministic, the purpose of the surrogate model is simply to approximate the response surface $y = \mathcal{M}(\boldsymbol{x})\ \forall \boldsymbol{x} \in \Omega$. Conversely, when

the model $\mathcal{M}$ exhibits a stochastic behaviour and its response is affected by a zero-mean random noise, the goal is to approximate the mean surface $\mathbb{E}[\mathcal{M}(\Omega)]$ and, possibly, provide a certain characterization of the uncertainty [17]. In this context, Kriging or *Gaussian process* meta-modelling has proved to represent a powerful meta-modelling approach commonly used in engineering or geostatistics. In a deterministic computer experiment, the Kriging model assumes that the model response $\mathcal{M}(\boldsymbol{x})$ can be conceived as the sum of a deterministic and a stochastic model as:

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}^K(\boldsymbol{x}) = \mathcal{T}(\boldsymbol{x}) + \mathcal{Z}(\boldsymbol{x}), \tag{1}$$

where $\mathcal{Z}(\boldsymbol{x})$ is a zero-mean stochastic process and the trend $\mathcal{T}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta} = \sum_{r=1}^{s}\beta_r f_r(\boldsymbol{x})$ is a linear regression model with s regression coefficients β_r and user-selected regression functions $f_r(\boldsymbol{x})$ (e.g $f_r(\boldsymbol{x}) = \boldsymbol{x}^r$). The formulation in Eq. (1) corresponds to the most general version of Kriging, often referred to as *Universal Kriging* (UK). Simpler expressions known as *Simple Kriging* (SK) and *Ordinary Kriging* (OK) are deduced when the trend term is assumed to be constant throughout the design space. In SK, it is assumed that the mean of the variable being predicted is known and constant across the entire area of study, whereas (OK) estimates the mean based on the available sample data. An interpretation of Eq. (1) is that deviations from the regression model may resemble a sample path of a properly chosen stochastic process [29]. The stochastic process $\mathcal{Z}(\boldsymbol{x})$ is determined by the process variance σ and an auto-correlation function $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{R}(|\boldsymbol{x} - \boldsymbol{x}'|; \boldsymbol{\theta})$, which depends on the spatial distance $\left|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right|$ and some hyper-parameters $\boldsymbol{\theta}$ to be determined. The auto-correlation function $\mathbf{R}$ must satisfy $\mathbf{R}(\mathbf{0}; \boldsymbol{\theta}) = 1$ and $\mathbf{R}\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}; \boldsymbol{\theta}\right) \to 0$ when $\left|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right| \to +\infty$. Thus, $\mathcal{Z}$ is characterized by the covariance matrix [30,31]:

$$\boldsymbol{\Sigma}_{\mathcal{Z}} := \left\{\mathrm{Cov}\left(\mathcal{Z}\left(\boldsymbol{x}^{(i)}\right), \mathcal{Z}\left(\boldsymbol{x}^{(j)}\right)\right)\right\}_{ij} = \sigma^2 \mathbf{R}\left(\left|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right|; \boldsymbol{\theta}\right). \tag{2}$$

Among the possible choices of the function $\mathbf{R}$, a common alternative and the one used in the present work is the Gaussian correlation function $\mathbf{R}(|\boldsymbol{x} - \boldsymbol{x}'|; \boldsymbol{\theta}) = \exp\left(-\boldsymbol{\theta}|\boldsymbol{x} - \boldsymbol{x}'|^2\right)$ [29].

The Kriging model is thus fully determined by the hyper-parameters $\{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma\}$, which are usually estimated by maximizing the likelihood function over the k samples of the ED [32]:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma, \boldsymbol{\beta}; \mathbf{y}) = \frac{1}{\sqrt{(2\pi\sigma^2)^k |\mathbf{R}|}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})\right], \tag{3}$$

where $\mathbf{y} = \left[y^{(1)} = \mathcal{M}\left(\boldsymbol{x}^{(1)}\right), \ldots, y^{(k)} = \mathcal{M}\left(\boldsymbol{x}^{(k)}\right)\right]^{\mathrm{T}}$ and $\mathbf{F}_{ij} = f_j(\boldsymbol{x}^{(i)})$ is referred to as the *information* matrix. Nevertheless, it is possible to extract estimates in closed form of the hyper-parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ as functions of the estimates of $\hat{\boldsymbol{\theta}}$ by the best linear unbiased estimator (BLUE) as:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) = \left(\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}, \quad \hat{\sigma}^2\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{k}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}). \tag{4}$$

Substituting Eq. (4) into (3) and taking logarithm, a simplified optimization problem that depends only on $\boldsymbol{\theta}$ is derived as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}\left[\frac{1}{k}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})|\mathbf{R}|^{1/k}\right]. \tag{5}$$

Therefore, the Kriging mean and mean-squared error (MSE) at any point $\boldsymbol{x}$ in the parameter space can be computed as [31]:

$$\begin{aligned} \hat{\mathcal{M}}^K(\boldsymbol{x}) &= \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \boldsymbol{r}(\boldsymbol{x})^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right), \\ \widehat{\mathrm{MSE}}(\boldsymbol{x}) &= \sigma^2\left[1 - \boldsymbol{r}(\boldsymbol{x})^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) + \boldsymbol{u}(\boldsymbol{x})^{\mathrm{T}}\left(\mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\boldsymbol{u}(\boldsymbol{x})\right], \end{aligned} \tag{6}$$

where $\boldsymbol{u}(\boldsymbol{x}) = \mathbf{F}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})$ and $r_i(\boldsymbol{x}) = \mathbf{R}\left(\left|\boldsymbol{x} - \boldsymbol{x}^{(i)}\right|; \hat{\boldsymbol{\theta}}\right)$, $i = 1, \ldots, k$.

To accommodate the existence of stochasticity in the response surface model $\mathcal{M}(x)$, Ankenman et al. [17] proposed Stochastic Kriging as an extension of the classical Kriging scheme specified in Eq. (1) to define the response surface $\mathcal{M}(\boldsymbol{x})$ in stochastic terms. In this context, several simulation runs in the same design point are performed, so the experimental design must be redefined as $\mathrm{ED} = \left\{\left(\boldsymbol{x}^{(i)}, n_i\right)\right\}_{i=1}^{k}$, where $n_i \in \mathbb{N}$ is the number of replications at the design point $\boldsymbol{x}^{(i)}$. In particular, the output of the lth simulation replication at $\boldsymbol{x}^{(i)}$, $i = 1, \ldots, k$ is given by:

$$\mathcal{M}_i\left(\boldsymbol{x}^{(i)}\right) = \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta} + \mathcal{Z}\left(\boldsymbol{x}^{(i)}\right) + \varepsilon_l\left(\boldsymbol{x}^{(i)}\right), \quad l = 1, \ldots, n_i, \tag{7}$$

where the terms $\mathcal{Z}\left(\boldsymbol{x}^{(i)}\right)$ and $\varepsilon_l\left(\boldsymbol{x}^{(i)}\right)$ represent the model *extrinsic uncertainty* and the *intrinsic uncertainty* for the lth replication, respectively. In other words, the changes in the model response are due to the simulator evaluation at different points in the parameter space, or due to the stochastic nature of the model itself, respectively. Analogously to Eq. (2), the intrinsic uncertainty can be characterized by the $k \times k$ covariance matrix:

$$\boldsymbol{\Sigma}_{\varepsilon} = \left\{\mathrm{Cov}\left(\bar{\varepsilon}\left(\boldsymbol{x}^{(i)}\right), \bar{\varepsilon}\left(\boldsymbol{x}^{(j)}\right)\right)\right\}_{ij}, \quad \bar{\varepsilon}\left(\boldsymbol{x}^{(i)}\right) = \frac{1}{n_i}\sum_{l=1}^{n_i}\varepsilon_l\left(\boldsymbol{x}^{(i)}\right), \; i = 1, \ldots, k. \tag{8}$$

In this light, similarly to Kriging, the calibration of the surrogate model involves maximizing a likelihood function, whose expression is the natural consequence of incorporating the intrinsic uncertainty in the Kriging methodology:

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}; \overline{\mathcal{M}}\right) = \frac{1}{\sqrt{\left|\boldsymbol{\Sigma}_{\mathcal{Z}} + \boldsymbol{\Sigma}_{\varepsilon}\right|(2\pi)^k}} \exp\left(-\frac{1}{2}\left(\overline{\mathcal{M}} - \mathbf{F}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\mathcal{Z}} + \boldsymbol{\Sigma}_{\varepsilon}\right)^{-1}\left(\overline{\mathcal{M}} - \mathbf{F}\boldsymbol{\beta}\right)\right), \tag{9}$$

where

$$\overline{\mathcal{M}} = \left[\overline{\mathcal{M}}\left(\boldsymbol{x}^{(1)}\right), \ldots, \overline{\mathcal{M}}\left(\boldsymbol{x}^{(k)}\right)\right]^{\mathrm{T}}, \quad \overline{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right) = \frac{1}{n_i}\sum_{l=1}^{n_i}\mathcal{M}_l(\boldsymbol{x}^{(i)}), \tag{10}$$

and $\boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)$ represents the $k\times 1$ *cross-correlation vector* $\left\{\mathrm{Cov}\left(\mathcal{Z}(\boldsymbol{x}),\mathcal{Z}\left(\boldsymbol{x}^{(j)}\right)\right)\right\}_{j=1}^{k}$ between the prediction point $\boldsymbol{x}$ and each one of the ED points. Upon the computation of the SK parameters, the SK predictor along with its associated mean squared error (MSE) can be derived as follows [33]:

$$\begin{aligned}\hat{\mathcal{M}}^{SK}(\boldsymbol{x}) &= \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)\left[\boldsymbol{\Sigma}_{\mathcal{Z}}+\boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\left(\overline{\mathcal{M}}-\mathbf{F}\hat{\boldsymbol{\beta}}\right),\\ \widehat{\mathrm{MSE}}(\boldsymbol{x}) &= \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\boldsymbol{\Sigma}_{\mathcal{Z}}+\boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot) + \boldsymbol{\zeta}^{\mathrm{T}}\left(\mathbf{F}^{\mathrm{T}}\left[\boldsymbol{\Sigma}_{\mathcal{Z}}+\boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\mathbf{F}\right)^{-1}\boldsymbol{\zeta},\end{aligned} \tag{11}$$

with $\boldsymbol{\zeta} = \boldsymbol{f}(\boldsymbol{x}) - \mathbf{F}^{\mathrm{T}}\left[\boldsymbol{\Sigma}_{\mathcal{Z}}+\boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)$. It is worth mentioning that the last term in Eq. (11) represents the variability inflation of the SK predictor resulting from the estimation of $\hat{\boldsymbol{\beta}}$. Additionally, it is important to highlight that, in deterministic scenarios where the intrinsic uncertainty is negligible, $\boldsymbol{\Sigma}_{\varepsilon}$ vanishes and Eq. (11) naturally converges to Eq. (6). In other words, SK reduces to the standard Kriging metamodel when $\boldsymbol{\Sigma}_{\varepsilon}$ tends to zero.

3. LAR-PCE based Stochastic Kriging

The higher complexity of UK compared to OK might suggest that it leads to improved estimates. However, as reported in the literature, a trend misspecification can lead UK not to improve the accuracy in its estimates [34] or even worsen its performance [35]. This limitation can be attributed to the fact that UK involves the estimation of additional parameters ($\boldsymbol{\beta}$), which can potentially induce over-fitting [36]. Therefore, to harness the advantages of UK while mitigating its potential drawbacks, it is crucial to choose a minimal trend basis that effectively captures the complexity of the response surface. Motivated by these considerations, the objective of this section is to identify an optimal choice of the trend model $\boldsymbol{f}$ in Eq. (7) by combining PCE and LAR. Hereafter, the fundamentals of PCE and the LAR algorithm are briefly presented, followed by their combination into Universal SK.

3.1. Polynomial-chaos expansion

The PCE model represents the model output response $y \in \mathbb{R}$ as its expansion through a basis of orthonormal multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$ as:

$$y = \mathcal{M}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}), \tag{12}$$

where $a_{\boldsymbol{\alpha}}$ are the coefficients of the expansion, and the multi-dimensional index notation $\boldsymbol{\alpha} = \left[\alpha_1,\ldots,\alpha_M\right]^{\mathrm{T}}$, $\alpha_i \in \mathbb{N}$, has been adopted. Depending on the nature of the input variables, classical families of orthonormal polynomials $\Psi_{\boldsymbol{\alpha}}$ include the Hermite or Legendre polynomials [37]. Although the expression (12) is exact on deterministic simulators for an infinite number of terms, only a finite number of terms can be computed in practice. Consequently, different truncation schemes have been proposed in the literature. One of the simplest approaches consists in selecting the set of all polynomials whose total degree $|\boldsymbol{\alpha}| = \sum_{i=1}^{M}\alpha_i$ is up to p, that is $\mathcal{A}^{M,p} = \{\boldsymbol{\alpha}\in\mathbb{N}^M : 0 \le |\boldsymbol{\alpha}| \le p\}$ [38]. This leads to a total number of $\mathcal{P}$ terms in the expansion:

$$\mathcal{P} := \mathrm{card}\left(\mathcal{A}^{M,p}\right) = \binom{M+p}{p}. \tag{13}$$

Note however that, for high-dimensional or non-linear problems, this truncation scheme leads to computing a large number of coefficients. As a solution, a more restrictive *hyperbolic truncation scheme* was proposed by Blatman and Sudret [12]. Based on the common observation in practice that coefficients associated with high-order interaction terms are usually close to zero (often referred to as the *sparsity-of-effect principle*), this truncation scheme selects all the multi-indices belonging to the set $\mathcal{A}^{M,p,q} = \{\boldsymbol{\alpha}\in\mathbb{N} : 0 \le \|\boldsymbol{\alpha}\|_q \le p\}$ with q-norm $\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M}|\alpha_i|^q\right)^{\frac{1}{q}}$. Note that, as the q-norm decreases, the number of mixed-order terms in the expansion also decreases until, ultimately leaving only the univariate polynomials remain.

Once the set of candidate polynomials and an ED of k samples are defined, the expansion coefficients $\boldsymbol{a} = \left\{\boldsymbol{a}_{\alpha},\ \boldsymbol{\alpha}\in\mathcal{A}^{M,p}\subset\mathbb{N}^M\right\}$ are obtained by minimizing the expectation of the least squares residuals:

$$\boldsymbol{a} = \arg\min_{\boldsymbol{a}\in\mathbb{R}^{\mathcal{P}}}\frac{1}{k}\sum_{i=1}^{k}\left[\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \sum_{\boldsymbol{\alpha}\in\mathcal{A}^{M,p}}\boldsymbol{a}_{\boldsymbol{\alpha}}\boldsymbol{\Psi}_{\boldsymbol{\alpha}}\left(\boldsymbol{x}^{(i)}\right)\right]^2. \tag{14}$$

Collecting the realizations of the QoI in the ED in vector form as $\mathbf{y} = \left[y^{(1)}=\mathcal{M}(\boldsymbol{x}^{(1)}),\ldots,y^{(k)}=\mathcal{M}(\boldsymbol{x}^{(k)})\right]^{\mathrm{T}}$, the solution of the optimization problem (14) is given by:

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{y}, \quad \boldsymbol{\Psi} = \left(\Psi_{ij}\right) = \left[\Psi_{\boldsymbol{\alpha}_j}\left(\boldsymbol{x}^{(i)}\right)\right]_{i=1,\ldots,k}^{j=1,\ldots,\mathcal{P}}, \tag{15}$$

where the *information matrix* $\boldsymbol{\Psi}$ contains the evaluation of the basis polynomials on the ED. Once Eq. (14) has been solved, the estimates by the resulting PCE surrogate model read:

$$\mathcal{M}^{PC}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}^{M,p}} \hat{\boldsymbol{a}}_{\boldsymbol{\alpha}} \boldsymbol{\Psi}_{\boldsymbol{\alpha}}(\boldsymbol{x}). \tag{16}$$

The quality of the resulting expansion is typically assessed through cross-validation. Among the different metrics available in the literature, the leave-one-out (LOO) error ϵ_{LOO} is particularly interesting since it can be estimated in closed form for linear models. In the case of the PCE model as given in Eq. (16), ϵ_{LOO} can be expressed as follows:

$$\epsilon_{LOO} = \frac{\sum_{i=1}^{k} \left(\frac{\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \mathcal{M}^{PC}(\boldsymbol{x}^{(i)})}{1 - h_i} \right)^2}{\sum_{i=1}^{k} \left(\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \hat{\mu}_{\boldsymbol{y}} \right)^2}, \tag{17}$$

where h_i is the ith component of the vector $\mathbf{h} = \operatorname{diag}\left\{ \boldsymbol{\Psi} \left(\boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Psi}^{\mathrm{T}} \right\}$ and $\hat{\mu}_{\boldsymbol{y}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{M}\left(\boldsymbol{x}^{(i)}\right)$.

3.2. Least angle regression (LAR)

The LAR algorithm is a model selection approach [28] that provides a very efficient procedure to implement the *Least Absolute Shrinkage and Selection Operator* (LASSO) method introduced by Tibshirani [39]. LASSO is a procedure for shrinking some regression coefficients to zero in a regression model by adding a restriction or penalty term to the least squares problem as:

$$\min_{\boldsymbol{\beta} \in \mathbf{R}^{\mathcal{P}}} \left(\|\boldsymbol{\Psi}(\boldsymbol{x})\,\boldsymbol{\beta} - \mathcal{M}(\boldsymbol{x})\|_2 + \eta \|\boldsymbol{\beta}\|_1 \right), \quad \eta \in \mathbb{R}, \quad \eta \geq 0. \tag{18}$$

The norm $\|\cdot\|_1$ adds a penalty equal to the absolute value of the magnitude of the coefficients. In other words, it limits the size of the coefficients leading to sparse models, i.e. models with few coefficients. The computation of the solution of the optimization problem in Eq. (18) is a convex quadratic programming problem with linear inequality constraints, which can be solved very efficiently by means of the LAR algorithm [28]. It is based on a variable selection procedure called *Incremental Forward Stepwise Regression* [40] that builds the model sequentially, adding one basis polynomial function $\Psi_{\boldsymbol{\alpha}_i}$ at a time to the set $\mathcal{A}_i$ of active basis functions, i.e. the set of predictors most correlated with the current residual $r_i = y - \boldsymbol{\Psi}_{\mathcal{A}_i} \boldsymbol{\beta}_{\mathcal{A}_i}$ at the ith iteration. The LAR algorithm can be summarily described by the following steps:

1. Set the vector residuals $r_0 = y$ and initialize the coefficients vector $\boldsymbol{\beta}_{\boldsymbol{\alpha}_i} = 0$, $i = 1, \ldots, \mathcal{P} - 1$. Set all the predictors $\Psi_{\boldsymbol{\alpha}_i}$ standardized to have mean zero and unit norm.
2. Find the predictor $\Psi_{\boldsymbol{\alpha}_i}$ most correlated with the current residual $r_i = y - \boldsymbol{\Psi}_{\mathcal{A}_i} \boldsymbol{\beta}_{\mathcal{A}_i}$ (with $\boldsymbol{\beta}_{\mathcal{A}_i}$ the vector of coefficients for the set of the active basis functions $\mathcal{A}_i$ at the ith iteration).
3. Move $\boldsymbol{\beta}_{\boldsymbol{\alpha}_i}$ from $\mathbf{0}$ towards its least-squares coefficient on $\Psi_{\boldsymbol{\alpha}_i}$, until another predictor $\Psi_{\boldsymbol{\alpha}_j}$ has as much correlation with the current residual r_i as $\Psi_{\boldsymbol{\alpha}_i}$.
4. Move $\{\boldsymbol{\beta}_{\boldsymbol{\alpha}_i}, \boldsymbol{\beta}_{\boldsymbol{\alpha}_j}\}$ in the direction defined by their joint least squares coefficient of the current residual on $\{\Psi_{\boldsymbol{\alpha}_i}, \Psi_{\boldsymbol{\alpha}_j}\}$, until another predictor $\Psi_{\boldsymbol{\alpha}_l}$ has the same correlation with the current residual.
5. Continue the process until $m = \min(\mathcal{P}, k - 1)$ predictors have been added.

Steps 3 and 4 update the set of *active* coefficients towards their least-square value in the form $\boldsymbol{\beta}_{\mathcal{A}_{i+1}} = \boldsymbol{\beta}_{\mathcal{A}_i} + \zeta_i \boldsymbol{u}_{\mathcal{A}_i}$. The factor ζ_i is the smallest positive scalar such that some new index j joins the active set $\mathcal{A}_{i+1} = \mathcal{A}_i \cup \{j\}$, and the vector $\boldsymbol{u}_{\mathcal{A}_i}$ is known as the *descent direction* (refer to [28] for further details). The name *least angle* arises from a geometric interpretation of this process, whereby $\boldsymbol{u}_{\mathcal{A}_i}$ provides the smallest (and equal) angle with each of the predictors in $\mathcal{A}_i$. Using the LAR scheme, a collection of m PCE meta-models of decreasing sparsity is extracted. Then, the optimal meta-model can be selected by a cross-validation scheme, to which the LOO error from Eq. (17) has been adopted in this work. Typically, the optimal model maximizing prediction accuracy and minimizing overfitting presents a cardinality, i.e. the number of active basis functions, lower that the cardinality $\mathcal{P}$ of the whole candidate basis. It is known that LAR only requires $\mathcal{O}(\mathcal{P}^3 + k\mathcal{P}^2)$ computations [28]. Therefore, when $\mathcal{P} \ll k$, it follows that $\mathcal{P}^3 < k\mathcal{P}^2$ and $\mathcal{O}(k\mathcal{P}^2) \sim \mathcal{O}(k)$. Thus, the computational complexity added by the PCE model when inserted as the trend term in the SK model is marginal with respect to the optimization problem related to the determination of the hyper-parameters in Eq. (5), which is $\mathcal{O}(k^3)$.

3.3. LAR-PCE Stochastic Kriging

This section presents the proposed PCE SK surrogate modelling approach, synthetically described in Fig. 1. One of the most remarkable features of the proposed model regards the automatic optimal definition of the trend model $\boldsymbol{f}$ for Universal SK. This is achieved by the PCE meta-model introduced in Eq. (16) using the hyperbolic truncation and the LAR algorithm overviewed in Sections 3.1 and 3.2, respectively. Note that the resulting approach provides a trend model with high flexibility in virtue of the

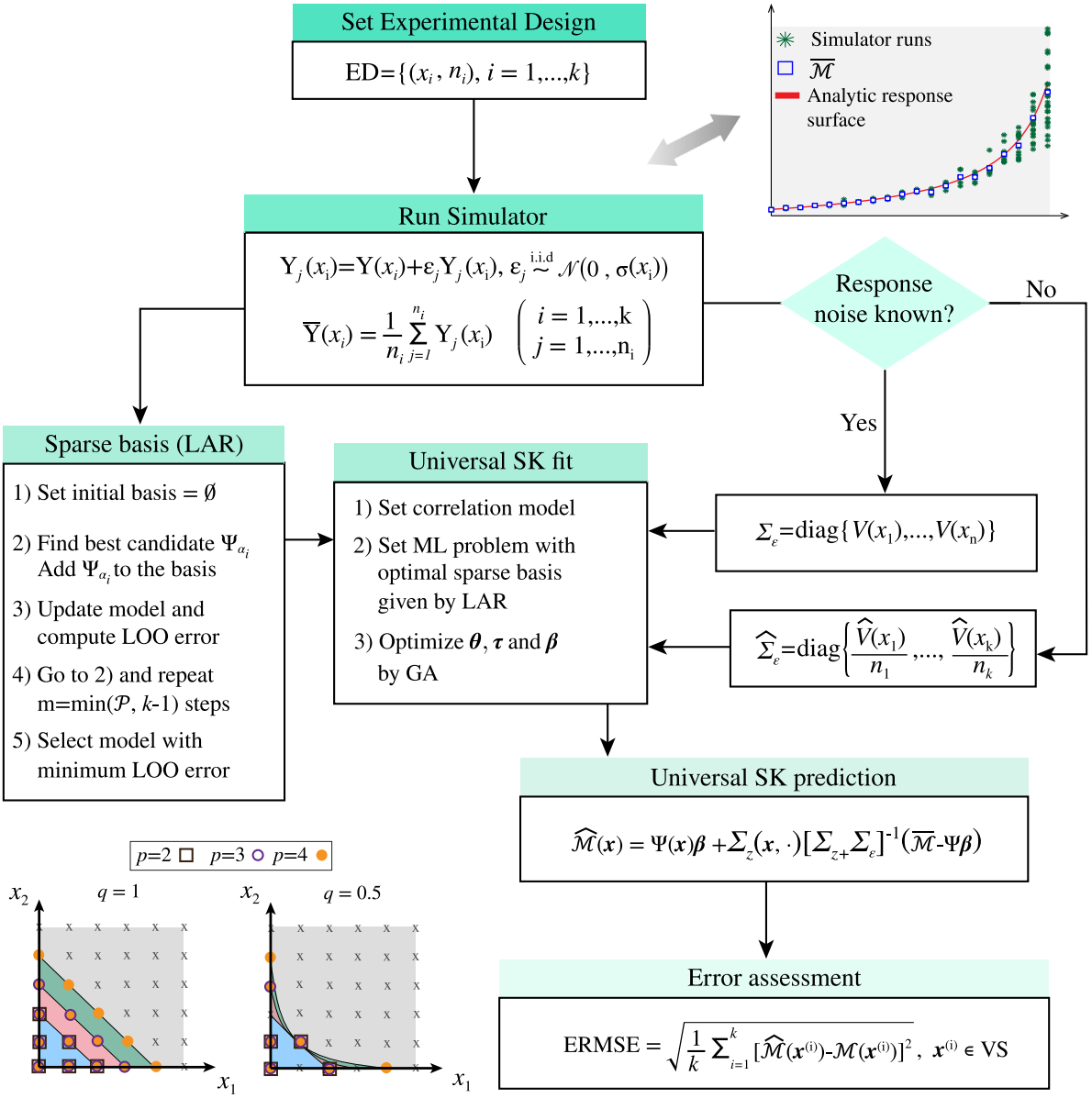

Fig. 1. Flowchart of the proposed PCE SK surrogate modelling approach for stochastic simulators.

LAR algorithm, in such a way that the terms in the PCE expansion are automatically tuned depending on the specific problem at hand. For the development of the proposed meta-model, once the optimal set $\boldsymbol{\Psi}$ is identified, a universal SK surrogate model is fitted (see Eq. (9)) through a GA global optimization approach to avoid stagnation in local extrema. Then, the prediction of the response surface and the associated MSE are extracted from Eqs. (25) and (26) respectively.

Under the common assumption in the literature that $\{\varepsilon_l(\boldsymbol{x}^{(i)})\}_{l=1}^{n_i}$ are i.i.d. and independent of the random field $\mathcal{Z}$ [17,25,41], the proposed optimal LAR-PCE based SK predictor reads:

$$\hat{\mathcal{M}}^{PCESK}(\boldsymbol{x}) = \boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\boldsymbol{\Sigma}_{\mathcal{Z}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\left(\overline{\mathcal{M}} - \boldsymbol{\Psi}\boldsymbol{\beta}\right), \tag{19}$$

and the corresponding optimal MSE reads

$$\mathrm{MSE}\left(\hat{\mathcal{M}}^{PCESK}\right) = \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\boldsymbol{\Sigma}_{\mathcal{Z}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot). \tag{20}$$

It is straightforward to check that the estimator proposed above is the best linear estimator of the form $\hat{\mathcal{M}}^{PCESK}(\boldsymbol{x}) = \lambda_0(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\overline{\mathcal{M}}$, where λ_0 and $\boldsymbol{\lambda}$ are weights depending on $\boldsymbol{x}$ and chosen in the sense of minimizing the MSE. Moreover, the

estimator is unbiased. This can be deduced by following the general reasoning given in [42] as follows:

$$\mathrm{MSE} = \mathbb{E}\left[\left(\mathcal{M}^K(\boldsymbol{x}) - \hat{\mathcal{M}}^{PCESK}(\boldsymbol{x})\right)^2\right] = \mathbb{E}\left[\left(\boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} + \mathcal{Z}(\boldsymbol{x}) - \lambda_0(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\overline{\mathcal{M}} \pm \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right)^2\right]$$

$$= \left(\boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} - \lambda_0(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right)^2 + \mathrm{Var}\left(\mathcal{Z}(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\overline{\mathcal{M}} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right) + \mathbb{E}\left[\left(\mathcal{Z}(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\overline{\mathcal{M}} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right)\right]^2$$

$$= \left(\boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} - \lambda_0(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right)^2 + \Sigma_{\mathcal{Z}}(\boldsymbol{x},\boldsymbol{x}) + \sum_{i,j=1}^{k}\lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\mathrm{Cov}\left(\mathcal{Z}\left(\boldsymbol{x}^{(i)}\right),\mathcal{Z}\left(\boldsymbol{x}^{(j)}\right)\right) + \sum_{i,j=1}^{k}\lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\mathrm{Cov}\left(\overline{\varepsilon}\left(\boldsymbol{x}^{(i)}\right),\overline{\varepsilon}\left(\boldsymbol{x}^{(j)}\right)\right)$$

$$-2\sum_{i=1}^{k}\lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\mathrm{Cov}\left(\mathcal{Z}\left(\boldsymbol{x}^{(i)}\right),\mathcal{Z}(\boldsymbol{x})\right) - 2\sum_{i=1}^{k}\lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\mathrm{Cov}\left(\mathcal{Z}\left(\boldsymbol{x}^{(i)}\right),\overline{\varepsilon}\left(\boldsymbol{x}^{(i)}\right)\right) =$$

$$= \left(\boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} - \lambda_0(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}\right)^2 + \Sigma_{\mathcal{Z}}(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\left[\Sigma_{\mathcal{Z}} + \Sigma_{\varepsilon}\right]\boldsymbol{\lambda}(\boldsymbol{x}) - 2\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\Sigma_{\mathcal{Z}}(\boldsymbol{x},\cdot),$$

where $\lambda_i(\boldsymbol{x})$ denotes the ith component of $\boldsymbol{\lambda}(\boldsymbol{x})$. Minimizing the MSE with respect to λ_0 and $\boldsymbol{\lambda}$, i.e. taking $\partial\mathrm{MSE}/\partial\lambda_0 = \partial\mathrm{MSE}/\partial\boldsymbol{\lambda} = 0$, extract the solutions $\lambda_0(\boldsymbol{x}) = \boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Psi}\boldsymbol{\beta}$ and $\boldsymbol{\lambda}(\boldsymbol{x}) = \left[\Sigma_{\mathcal{Z}} + \Sigma_{\varepsilon}\right]^{-1}\Sigma_{\mathcal{Z}}(\boldsymbol{x},\cdot)$ can be readily extracted. With these weights, the optimal linear estimator and MSE are directly given by Eqs. (19) and (20), respectively. Furthermore, following that $\mathcal{Z}$ is a zero-mean stochastic process, i.e. $\mathbb{E}[\mathcal{Z}(\boldsymbol{x})] = 0$, and $\mathbb{E}\left[\overline{\mathcal{M}} - \boldsymbol{\Psi}\boldsymbol{\beta}\right] = 0$, it follows that:

$$\mathbb{E}\left[\hat{\mathcal{M}}^{PCESK}(\boldsymbol{x}) - \mathcal{M}^K(\boldsymbol{x})\right] = \boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} + \Sigma_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\Sigma_{\mathcal{Z}} + \Sigma_{\varepsilon}\right]^{-1}\mathbb{E}\left[\overline{\mathcal{M}} - \boldsymbol{\Psi}\boldsymbol{\beta}\right] - \boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta} - \mathbb{E}[\mathcal{Z}(\boldsymbol{x})] = 0, \tag{21}$$

and, therefore, it is extracted that $\hat{\mathcal{M}}^{PCESK}(\boldsymbol{x})$ is an unbiased estimation for $\mathcal{M}^K(\boldsymbol{x})$. The estimation of the hyper-parameters of the proposed LAR-PCE SK surrogate model requires maximizing the likelihood function described in Eq. (9), where the information matrix $\mathbf{F}$ is simply replaced by the optimal trend matrix $\boldsymbol{\Psi}$. However, some considerations must be made. Firstly, note that the covariance matrix of the noise in the response surface (intrinsic bias between the surrogate model and the true simulator's response), Σ_{ε}, in Eq. (9) is unknown in general. Nonetheless, ε is commonly assumed in practice to be normally distributed, so one can write:

$$\left\{\varepsilon_l\left(\boldsymbol{x}^{(i)}\right)\right\}_{l=1}^{n_i} \sim \mathcal{N}\left(0, V_i\right), \quad \text{with } V_i \text{ estimated as} \quad V_i = \frac{1}{n_i - 1}\sum_{l=1}^{n_i}\left(\mathcal{M}_l\left(\boldsymbol{x}^{(i)}\right) - \overline{\mathcal{M}}_l\left(\boldsymbol{x}^{(i)}\right)\right)^2, \tag{22}$$

leading to a positive definite diagonal covariance matrix Σ_{ε} (see Theorem 1 in [17]). Secondly, it should be noted that, when calibrating a Kriging meta-model for deterministic simulators, Eq. (3) can be simplified as a function of $\boldsymbol{\theta}$ by leveraging on the BLUE, while in SK this fact is no longer valid. Instead, $\boldsymbol{\beta}$ must be first approximated as $\hat{\boldsymbol{a}}$ in Eq. (15). This entails limited computational overhead because the trend term has been optimally adjusted beforehand by LAR-PCE. Then, the extrinsic variance σ^2 needs to be simultaneously optimized with the hyper-parameters $\boldsymbol{\theta}$ as:

$$\hat{\boldsymbol{\theta}}, \hat{\sigma}^2 = \underset{\boldsymbol{\theta},\,\sigma^2}{\arg\max}\,\log\left(\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\right)\right). \tag{23}$$

Finally, and based on Eq. (23), the optimal estimate for $\boldsymbol{\beta}$ as a function of $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ can be written similarly to Eq. (4) as:

$$\hat{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \varepsilon\right) = \left[\boldsymbol{\Psi}^{\mathrm{T}}\left(\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right)^{-1}\boldsymbol{\Psi}\right]^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\left(\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right)^{-1}\overline{\mathcal{M}}. \tag{24}$$

Once the optimal parameters have been derived, the optimal predictor can be re-written as:

$$\hat{\mathcal{M}}^{PCESK}(\boldsymbol{x}) = \boldsymbol{\Psi}(\boldsymbol{x})\hat{\boldsymbol{\beta}} + \hat{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right]^{-1}\left(\overline{\mathcal{M}} - \boldsymbol{\Psi}\hat{\boldsymbol{\beta}}\right). \tag{25}$$

with the MSE estimator

$$\widehat{\mathrm{MSE}}(\boldsymbol{x}) = \hat{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\boldsymbol{x}) - \hat{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}}\left[\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right]^{-1}\hat{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot) + \boldsymbol{\gamma}^{\mathrm{T}}\left(\boldsymbol{\Psi}^{\mathrm{T}}\left[\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\gamma}, \tag{26}$$

and $\boldsymbol{\gamma} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\Psi}^{\mathrm{T}}\left[\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}\right]^{-1}\hat{\Sigma}_{\mathcal{Z}}(\boldsymbol{x},\cdot)$. The expression (26) follows from the development in [42] for standard Kriging by replacing the covariance matrix of the random field by $\hat{\Sigma}_{\mathcal{Z}} + \hat{\Sigma}_{\varepsilon}$.

Let us lastly focus on the common assumption of the stochastic process $\mathcal{Z}$ following a Gaussian distribution. In this case, when considering an ED of k samples, the hypotheses in Eq. (2) imply that $\left(\mathcal{M}(\boldsymbol{x}), \overline{\mathcal{M}}\left(\boldsymbol{x}^{(1)}\right), \ldots, \overline{\mathcal{M}}\left(\boldsymbol{x}^{(k)}\right)\right) \sim \mathcal{N}_{k+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ follow a multivariate Gaussian distribution with mean vector and covariance matrix given by:

$$\boldsymbol{\mu} = (\mathcal{M}(\boldsymbol{x}), \boldsymbol{\Psi}(\boldsymbol{x})\boldsymbol{\beta}) \in \mathbb{R}^{k+1} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \Sigma_{\mathcal{Z}}(\boldsymbol{x},\cdot)^{\mathrm{T}} \\ \Sigma_{\mathcal{Z}}(\boldsymbol{x},\cdot) & \Sigma_{\mathcal{Z}} + \Sigma_{\varepsilon} \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}, \tag{27}$$

and Eq. (19) coincides with the conditional mean of $\mathcal{M}(\boldsymbol{x})$ given $\overline{\mathcal{M}}$, i.e. $\mathbb{E}\left[\mathcal{M}(\boldsymbol{x})|\,\overline{\mathcal{M}}\right]$ (see [30]).

To conclude, the quality of the proposed metamodel must be assessed. In practical contexts where the computational burden is high, cross-validation techniques are popular approaches for evaluation the accuracy of stochastic simulators (refer to reference [43] for an extensive discussion). Nevertheless, in this work, to conduct a robust assessment of the proposed PCE SK metamodel, its accuracy has been evaluated on the basis of a dense validation set $\mathrm{VS} := \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$, with $K \gg k$, covering the parameter space, and using diverse quality metrics reported hereafter.

4. Numerical results and discussion

This section presents an extensive numerical assessment of the proposed LAR-PCE SK meta-model. The evaluation includes a detailed comparison against ordinary SK through various well-known benchmark examples from the literature. These presented numerical studies cover both single (Section 4.1) and multi-dimensional functions (Sections 4.2 and 4.3) contaminated by noise. For each case study, given a fixed computational budget, several sampling strategies to obtain the ED are discussed. These approaches range from (i) using a low number of sampling points with many replications to represent the uncertainty to (ii) strategies with numerous sampling points with less budget devoted to the characterization of the uncertainty. Additionally, intermediate strategies that fall between these two extremes are also explored. The purpose of defining different sampling scenarios is to demonstrate that the proposed LAR-PCE SK method consistently outperforms ordinary SK, regardless of whether a sparse and deep ED (with few points but high accuracy) or a denser but shallower ED is chosen. The computational effort allocated at each point is proportional to its corresponding noise variance in every instance. In all the analyses, the model performance is assessed by means of two error metrics, namely the Empirical Root Mean Squared Error (ERMSE) and the Normalized Maximum Absolute Error (NMAE) over a large VS given by:

$$\text{ERMSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\widehat{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right) - \mathcal{M}\left(\boldsymbol{x}^{(i)}\right)\right)^2}, \quad \text{NMAE} = \max_{1\le i\le K}\frac{\left|\widehat{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right) - \mathcal{M}\left(\boldsymbol{x}^{(i)}\right)\right|}{K\sigma_{VS}}, \; x^{(i)} \in VS, \tag{28}$$

where:

$$\sigma_{VS} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\overline{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right) - \mathcal{M}\left(\boldsymbol{x}^{(i)}\right)\right)^2}. \tag{29}$$

Note that the ERMSE metric evaluates the global quality of the meta-model, while the NMAE metric is an indicator of the local error.

4.1. Case study I: M/M/1 queue

The M/M/1 queue model is a simple yet powerful model used to analyse the behaviour of a wide range of real-world queuing systems [44], from call centres to manufacturing lines. The model represents a system where there is a single server and a line of entities to be served (e.g., packets, customers, etc.). In this system, entities arrive randomly at a certain rate, called arrival rate, and wait in a queue until the server becomes available. The waiting time is determined by the arrival rate and by the service rate, which refer to the time it takes for the server to process an entity and complete its service, respectively.

The present case study consists of the simulation of an M/M/1 queue where the service rate is fixed at one and the arrival rate ranges from 0.3 to 0.9. The response surface y represents the expected waiting time as a function of the arrival rate x. The accuracy of the simulation at each design point is determined by its time run-length T, that is, the period of time over which the system's behaviour is observed and analysed. This system is one of the most studied examples in the SK literature [17,33,45]. It is particularly notable for its distinct and progressively increasing intrinsic variance across the parameter space. This characteristic naturally provides a formidable yet simple example to understand the potential benefits of the proposed approach. Despite the stochastic nature of the model, analytical expressions for the response surface and its intrinsic variance are known [17,33]:

$$y(x) = \frac{1}{1-x}, \; V(x) = \frac{2x(1+x)}{T(1-x)^4}. \tag{30}$$

In the following, we introduce two variations of the experiment. In the first one in Section 4.1.1, the intrinsic variance is assumed as known following Eq. (30), and one single simulation is run for each design point. In the second variation in Section 4.1.2, the intrinsic variance of the simulator is estimated by performing multiple replications at each of the ED points. For both cases, the results obtained by the classical Ordinary SK are compared with those obtained by the proposed LAR-PCE SK approach.

4.1.1. Known intrinsic noise

For this first numerical investigation, we will consider the intrinsic variance as $\Sigma_\epsilon = \text{diag}\left(V\left(\boldsymbol{x}^{(i)}\right)\right)_{i=1}^{k}$. Additionally, $\overline{\mathcal{M}}$ is directly taken as $\left(\mathcal{M}\left(x^{(i)}\right)\right)_{i=1}^{k}$ since one single replication at each design point is considered in this first numerical study. It is important to note that, although this situation may not reflect the complexity of most real-world applications, it serves as a useful analysis to shed light on the characteristics of the proposed LAR-PCE SK approach. Three different scenarios with varying number of design points and run-lengths but with constant computational resources have been considered. The first one involves 10 equispaced design points with a run-length of 6000; the second one consists of 30 design points with a run-length of 2000; and the third scenario encompasses 50 design points with a run-length of 1200. In the case of LAR-PCE SK, following the recommendations provided in Ref. [46], the maximum degree of the polynomials is considered as 5, 10 and 16 for the scenarios one to three, respectively. It is worth noting that, given the one-dimensional nature of the problem, the LAR algorithm simply selects the optimal degree for the uni-variate regression model.

Fig. 2 (a,d) depict the true response surface along with the analytically computed 95% confidence intervals represented as $\left(y(x) \pm z_\alpha\sqrt{V(x)}\right)$ for Scenarios 1 and 2, respectively. Similarly, Fig. 2 (b,e) and (c,f) illustrate the estimates produced by Ordinary SK and LAR-PCE SK, respectively. Estimated 95% confidence intervals in Fig. 2 (b,e) and (c,f) have been computed as

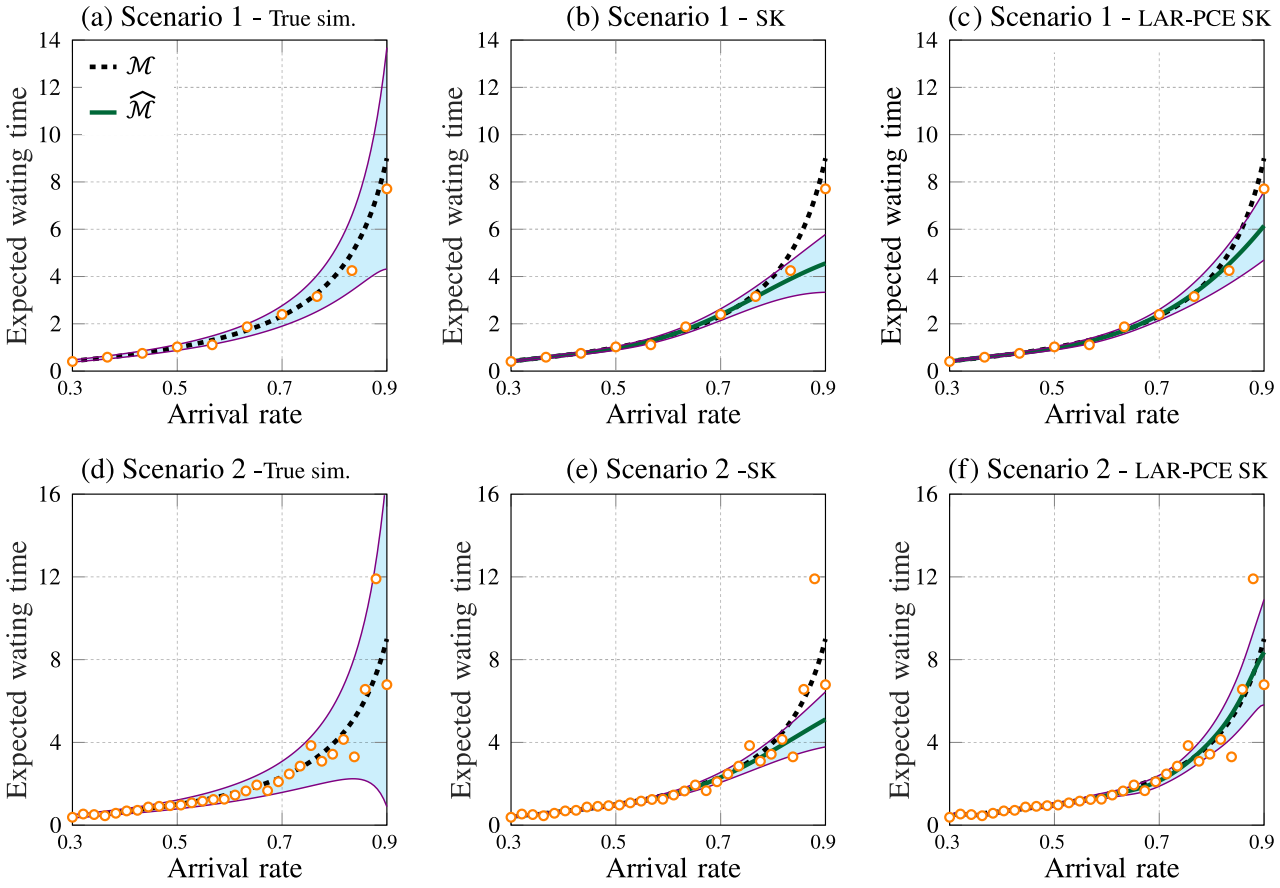

Fig. 2. Case Study I: M/M/1 example. Comparison between the true simulator (a,d) against the predictions by the proposed LAR-PCE SK (c,f) and Ordinary SK (b,e) under the assumption of known variance (one replication per ED sample) in Scenarios 1 (a,b,c) (10 ED samples and run-length = 6000) and 2 (d,e,f) (30 ED samples and run-length = 2000). Scatter points denote the ED samples, and light blue shaded areas represent the 95% confidence interval.

Table 1
Summary of the accuracy in the estimates of the M/M/1 model with known intrinsic noise variance by ordinary SK and LAR-PCE SK adopting sampling scenarios 1 to 3.

	SK		LAR-PCE SK	
	Mean ERMSE	Mean NMAE	Mean ERMSE	Mean NMAE
Scenario 1	0.787	0.0076	0.516	0.0049
Scenario 2	0.810	0.0079	0.519	0.0052
Scenario 3	0.832	0.0080	0.650	0.0074

$\left(\hat{\mathcal{M}}^{SK} \pm z_\alpha \sqrt{\widehat{MSE}(x)}\right)$ and $\left(\hat{\mathcal{M}}^{PCESK} \pm z_\alpha \sqrt{\widehat{MSE}(x)}\right)$, respectively. It can be seen that for low values of x, where the intrinsic uncertainty is small, both meta-models exhibit a similar behaviour. In contrast, for high values of x where both the QoI and the intrinsic variance experience a rapid growth, Ordinary SK fails to capture the variation in the response, whereas LAR-PCE SK exhibits a significantly better performance. This is to be expected since, as reported in [35], Ordinary SK virtually ignores those simulation outputs whose estimated intrinsic variance are large compared to the estimated extrinsic variance. Thus, these results demonstrate that introducing a proper trend model effectively mitigates this issue.

To provide a further comparison in performance of LAR-PCE SK with respect to Ordinary SK, the results of the fitting of one hundred independent experiments are reported in Fig. 3 and Table 1. Significant enhancements have been achieved in terms of ERMSE, with improvements ranging from 20% to 35% on average. Similarly, the NMAE showed average improvements ranging from 8% to 35%. It is noteworthy that, when considering the median values, the improvements are even more significant, ranging from 28% to 42% for ERMSE and from 28% to 38% for NMAE. Finally, it can be noted in Fig. 2 that, in terms of variance estimation, both approaches similarly fail to capture the true variability. This issue has been recognized and documented in the SK literature [19,36]. Although there are methods available to address this problem (refer e.g. to [36]), they are beyond the scope of our study, in which the primary focus is on the evaluation of the importance of properly selecting the trend model in SK.

4.1.2. Unknown intrinsic noise

This section introduces a modified version of the previous investigation, where the intrinsic variance is assumed to be unknown. Consequently, following Eq. (22), the estimation of the Σ_ϵ matrix is carried out through a series of replications at each ED point.

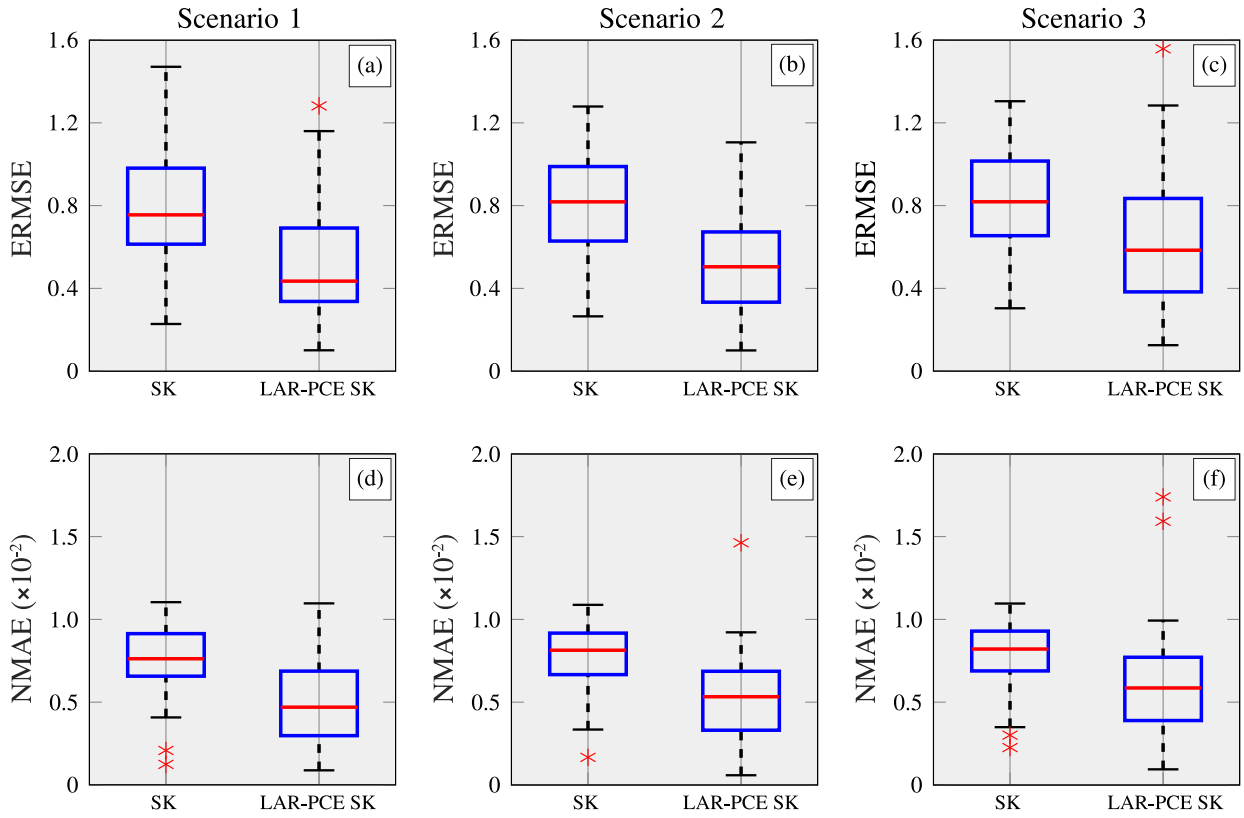

Fig. 3. Performance analysis of ordinary SK and LAR-PCE SK in case study I in terms of ERMS and NMAE error metrics using 100 known intrinsic noise M/M/1 experiments and sampling Scenarios 1 to 3.

Table 2
Summary of the accuracy in the estimates of the M/M/1 model with unknown intrinsic noise variance by ordinary SK and LAR-PCE SK adopting sampling scenarios 1 to 3.

	SK		LAR-PCE SK	
	Mean ERMSE	Mean NMAE	Mean ERMSE	Mean NMAE
Scenario 1	0.073	5.16E−04	0.067	4.67E−04
Scenario 2	0.107	1.26E−03	0.084	7.53E−04
Scenario 3	0.142	1.80E−03	0.127	1.34E−03

In this particular analysis, a total computational budget C of 500 replications to be distributed along the training sites has been considered. Consistently with Ref. [45], the number of replications assigned to each design point x_i is determined in a single step as proportional to the model variance following $n_i = \frac{\sqrt{V(\boldsymbol{x}_i)}}{\sum_{i=1}^{k}\sqrt{V(\boldsymbol{x}_i)}}C$. It should be noted that the use of the actual variance to determine the number of replications at each point in our analysis has been chosen for the sake of simplicity. However, alternative methods which employ a multi-step approach for this purpose such as those described in Ref. [33] may be considered for bypassing this issue.

The same three scenarios as discussed in Section 4.1.1 have been considered over a macro-replication of one hundred experiments. The results are summarized in Table 2 and Fig. 4. As in the previous case, a relevant improvement in performance can be observed in each situation, ranging from 8% to 21% on average with respect to ERMSE, and from 9% to 40% with respect to NMAE. Considering the median values, similar conclusions to the previous analyses can be extracted, as can be seen in Fig. 4.

4.2. Case study II: two dimensional example

This second case study investigates the egg-box shaped surface suggested in Ref. [45]. The simulation output at a design point $\boldsymbol{x} = \left[x_1, x_2\right]^{\mathrm{T}} \in [-1,1]^2$ on the jth replication is given by:

$$\mathcal{M}_j\left(x_1, x_2\right) = \sin\left(9x_1^2\right) + \sin\left(9x_2^2\right) + \varepsilon_j\left(x_1, x_2\right), \tag{31}$$

with $\varepsilon_j \sim \mathcal{N}\left(0, 2 + \cos\left(\pi + \left(x_1 + x_2\right)/2\right)\right)$.

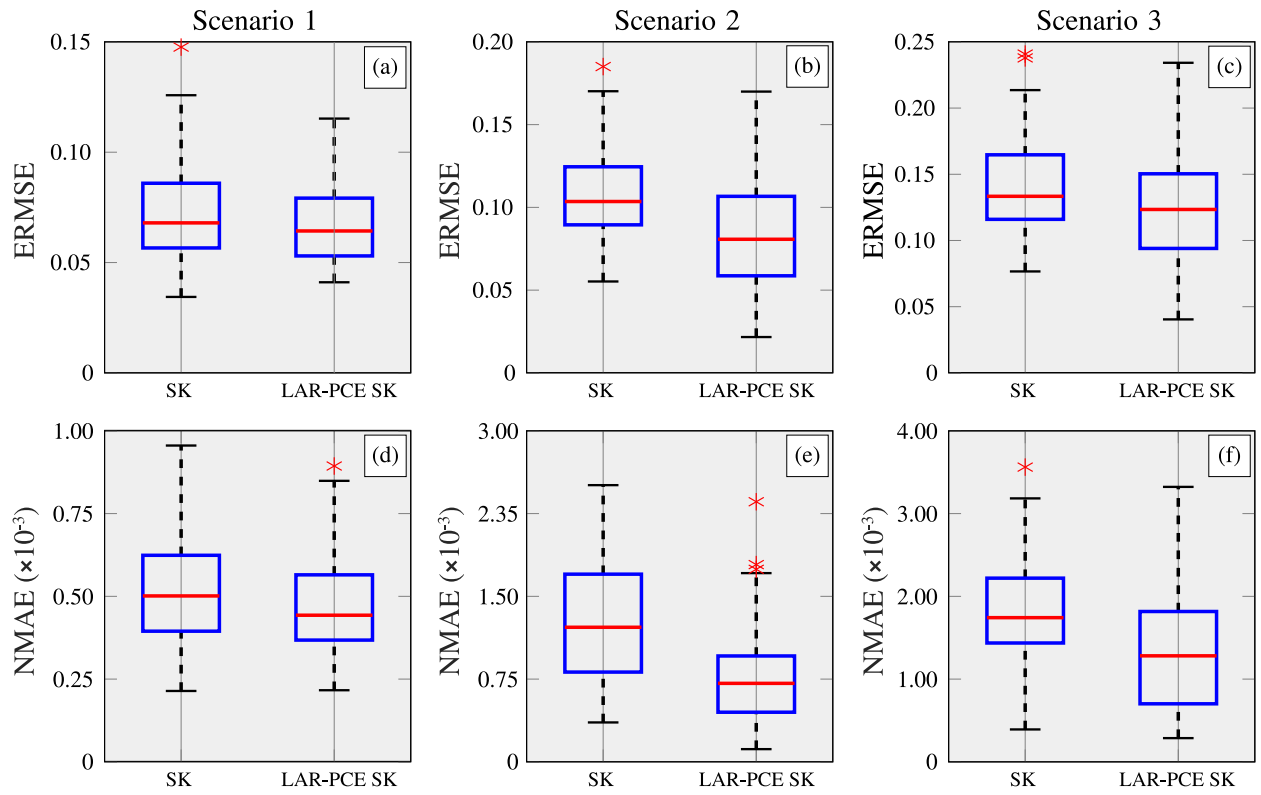

Fig. 4. Performance analysis of ordinary SK and LAR-PCE SK in case study I in terms of ERMS and NMAE error metrics using 100 unknown intrinsic noise M/M/1 experiments and sampling Scenarios 1 to 3.

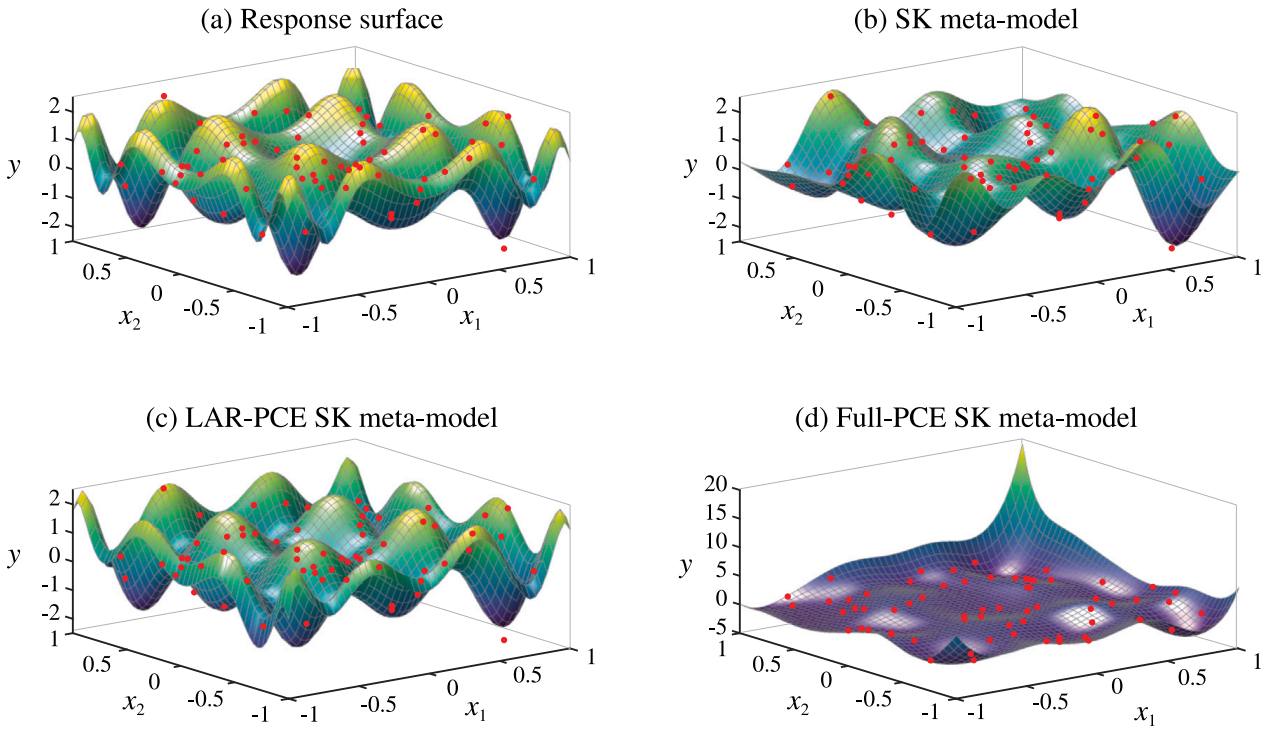

Fig. 5. Response surface of the case study II (a), and predictions by SK (b), LAR-PCE SK (c) and full PCE SK (d) surrogate models with $k = 64$ design points.

Likewise in the previous case study, three different scenarios have been considered by varying the number of design points but keeping a fixed computational budget C of 1280 replications to be distributed among them. Specifically, the three different settings are characterized by an ED size of k = 32, 64, and 128 points, respectively. The locations of the design points were determined by

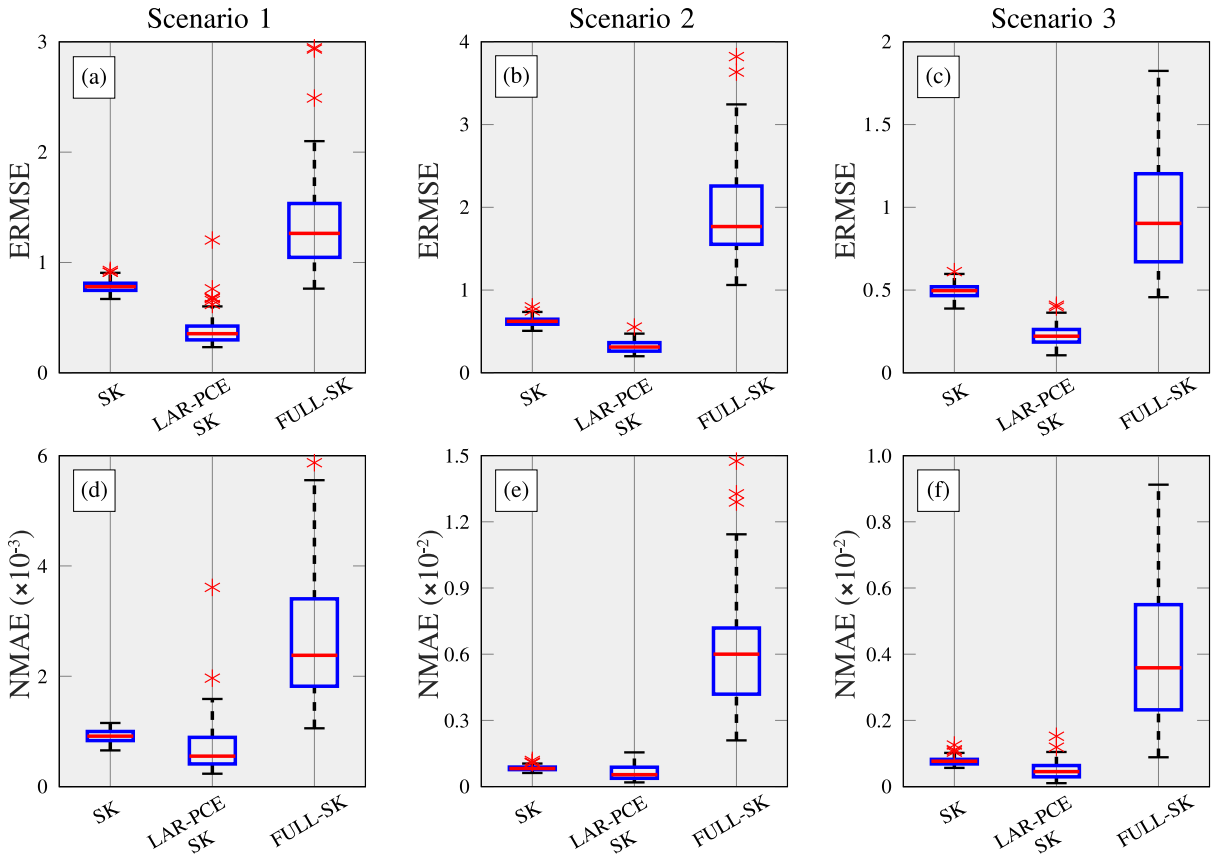

Fig. 6. Performance analysis of ordinary SK, LAR-PCE SK and full PCE SK in terms of ERMS and NMAE error metrics using 100 experiments of Case Study II and sampling Scenarios 1 to 3.

adopting the Latin Hypercube Sampling (LHS) algorithm, a commonly used technique for generating diverse and evenly distributed design points in multi-dimensional space. The number of replications n_i at each design point has been adjusted proportionally to its standard deviation as in Section 4.1.2. For each scenario, the performance of Ordinary SK is benchmarked against the proposed LAR-PCE SK meta-model. Additionally, to stress the importance of choosing an appropriate sparse basis of polynomials in the trend model, a full PCE SK model without LAR has been also included in the comparison. In this case, the entire polynomial basis is included up to a certain degree given by the hyperbolic truncation defined in Section 3.1. Regarding the polynomial basis settings, a q-norm value of 0.8 has been employed in all cases. With respect to the degree of the polynomial basis to be considered, in compliance with the recommendations by Sudret [46], the maximum degree of the polynomial basis for the full PCE SK case has been set at 5, 8, and 10 for Scenarios 1, 2, and 3, correspondingly. On the other hand, due to the sparse nature of the basis selected by the LAR algorithm, it is possible to consider polynomials of slightly higher degree for LAR-PCE SK. Specifically, degrees of 9, 11, and 12 have been considered in Scenarios 1 to 3, respectively.

Fig. 5 illustrates the outcomes of an arbitrary replication of Scenario 2. It depicts the true response surface to be approximated (a), alongside the meta-models constructed using Ordinary SK (b), LAR-PCE SK (c), and full PCE SK (d) with a total of $k = 64$ design points. Fig. 6 and Table 3 present the results of the comparison between the different approaches discussed in the considered scenarios over a macro-replication of one hundred independent experiments. In terms of performance, LAR-PCE SK consistently outperforms the other methods, while full PCE SK showed the poorest performance based on both ERMSE and NMAE metrics. Comparing Ordinary SK with the proposed LAR-PCE SK, average reductions around 50% in ERMSE are found in all the considered scenarios. With respect to the local accuracy, NMAE enhancements vary from 23% in Scenarios 1 and 2, to 36% in Scenario 3. Taking into account the median, similar conclusions can be inferred. On the other hand, it is noteworthy that blindly setting an inadequate polynomial basis as the trend model resulted in notable errors and instability in the results across all the scenarios, highlighting the crucial importance of a proper basis selection. Remarkably, LAR-PCE SK included a modest number of terms in the basis with a median of 9, 11, and 14 terms for Scenarios 1 to 3, respectively. In contrast, the complete expansion included 17, 32, and 49 polynomials in the basis for each respective scenario.

4.3. Case study III: Ishigami function

This last case study investigates the Ishigami function [26], a classical benchmark function widely used in the literature. This function is characterized by its non-linear and non-monotonic behaviour, making it a challenging test for meta-modelling techniques.

Table 3
Summary of the accuracy in case study II adopting sampling scenarios 1 to 3.

	SK		LAR-PCE SK	
	Mean ERMSE	Mean NMAE	Mean ERMSE	Mean NMAE
Scenario 1	0.782	9.13E−04	0.385	6.98E−04
Scenario 2	0.623	8.32E−04	0.316	6.42E−04
Scenario 3	0.495	7.69E−04	0.228	4.90E−04

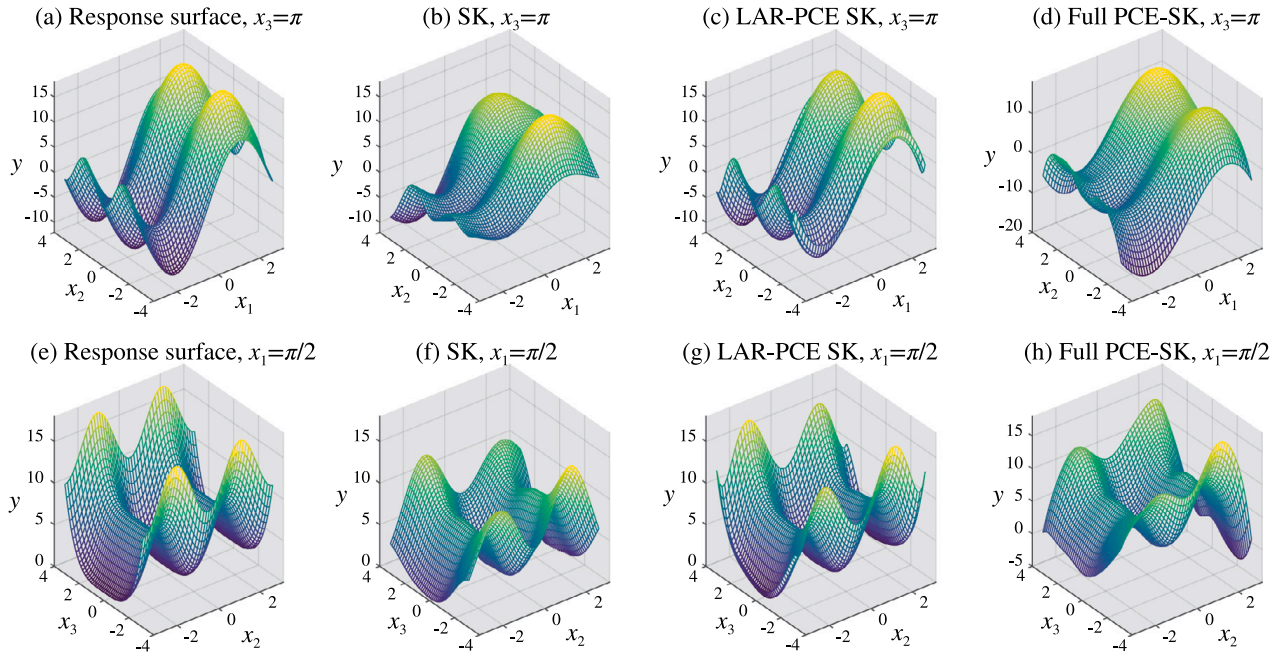

Fig. 7. Response surfaces of the Ishigami function (a,e), and predictions by ordinary SK (b,f), LAR-PCE SK (c,g) and full PCE SK (d,h) meta-models with $k = 64$ design points. The top and bottom row panels denote the response surfaces obtained for $x_3 = \pi$ and $x_1 = \pi/2$, respectively.

The function takes as input a three-dimensional vector $\boldsymbol{x} = \left[x_1, x_2, x_3\right]^{\mathrm{T}}$ with each component x_i ranging from $-\pi$ to π, and outputs a scalar value $\mathcal{M}_j(\boldsymbol{x})$ given by the non-linear sum of three sinusoidal functions. For the purpose of this work, a stochastic version of the Ishigami function is defined by incorporating a normally distributed random noise term $\varepsilon_j(\boldsymbol{x})$ with zero mean and standard deviation proportional to the absolute value of the function. This results in the following stochastic Ishigami function:

$$\begin{aligned}\mathcal{M}_j\left(x_1, x_2, x_3\right) &= \sin\left(x_1\right) + 7\sin^2\left(x_2\right) + 0.1x_3^4\sin\left(x_1\right) + \varepsilon_j\left(x_1, x_2, x_3\right),\ x_1, x_2, x_3 \in [-\pi, \pi], \\ \varepsilon_j &\sim \mathcal{N}\left(0, \sqrt{\left|\sin\left(x_1\right) + 7\sin^2\left(x_2\right) + 0.1x_3^4\sin\left(x_1\right)\right|}\right).\end{aligned} \tag{32}$$

The ED has been defined with a total computational budget of $C = 2560$. Three distinct scenarios considering different ED sizes have been investigated: $k = 64$ (Scenario 1), $k = 128$ (Scenario 2), and $k = 256$ (Scenario 3). The allocation of replications has been defined proportionally to the standard deviation, as outlined in Section 4.2. On this basis, Ordinary SK, LAR-PCE SK, and full-PCE SK meta-models have been constructed and compared in terms of accuracy. The calibrations for Scenarios 1 to 3 have been performed with different maximum degrees for the polynomial basis. Specifically, the full-PCE SK meta-models have been calibrated with maximum degrees of 4, 6, and 7, respectively. On the other hand, the LAR-PCE SK meta-models have been calibrated with slightly higher maximum degrees of 6, 8, and 9 for Scenarios 1 to 3, respectively. Additionally, a value of 0.8 is specified for the q-norm for all the considered scenarios.

Fig. 7 shows an arbitrary replication of Scenario 1 ($k = 64$ design points), displaying the true response surface (a) and the predictions of the meta-models constructed using SK (b), LAR-PCE SK (c), and full PCE SK (d). The comparison among the different approaches based on a replication of 100 independent experiments is presented in Fig. 8 and Table 4. It can be observed that LAR-PCE SK consistently outperforms the other methods in terms of both global (ERMSE) and local (NMAE) accuracy. In this case study, similarly to Case Study II, full PCE SK exhibits the poorest performance. Nonetheless, in this particular case study, the difference between full PCE SK and Ordinary SK is not as pronounced. Instead, when comparing SK with the proposed LAR-PCE SK, significant enhancements are noticeable in terms of ERMSE, with average improvements ranging from 35% in Scenario 1 to 74% in Scenario 3. Similar enhancements are also observed when considering the NMAE metric or the improvement in median terms. Lastly, it is worth mentioning the sparsity of the polynomial bases used in the trend model. While the full-PCE expansions consist of a larger number of terms (23, 50, and 68 in Scenarios 1 to 3), the LAR-selected polynomials are significantly more sparse, with only 10, 14, and 17 terms on average, respectively. This allows capturing the trend without encountering over-fitting issues.

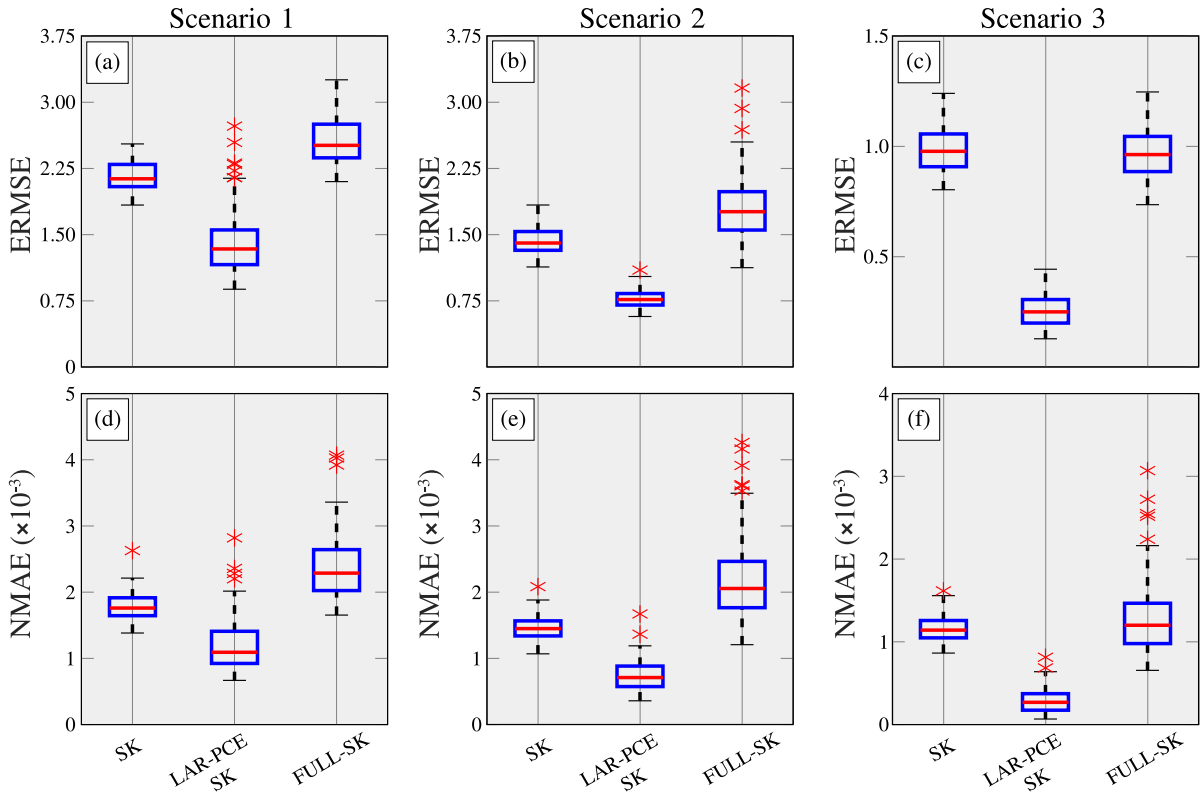

Fig. 8. Performance analysis of ordinary SK, LAR-PCE SK and full PCE SK in terms of ERMS and NMAE error metrics using 100 experiments of the Ishigami function and sampling Scenarios 1 to 3.

Table 4
Summary of the accuracy in the estimates of the Ishigami function for sampling scenarios 1 ($k = 64$), 2 ($k = 128$), and 3 ($k = 256$).

	SK		LAR-PCE SK	
	Mean ERMSE	Mean NMAE	Mean ERMSE	Mean NMAE
Scenario 1	2.169	1.77E−03	1.418	1.20E−03
Scenario 2	1.421	1.46E−03	0.772	7.32E−04
Scenario 3	0.983	1.11E−03	0.257	2.92E−04

5. Conclusions

This study has proposed the use of Universal Stochastic Kriging for constructing surrogate models for stochastic simulators. Specifically, the proposed approach, LAR-PCE SK, adopts adaptive sparse PCE as the trend model in a Universal SK model. With the aim of minimizing over-fitting, the LAR algorithm is adopted to automatically identify the optimal polynomial basis in the PCE, providing the resulting model with local/global prediction abilities and high flexibility for a broad variety of problems. Lastly, the hyper-parameters of the resulting Universal SK model are fitted through a global GA optimization. The effectiveness of the proposed approach has been validated through three classical benchmark case studies, namely: (i) the M/M/1 queue; (ii) a noisy eggbox-shaped surface; and (iii) a stochastic version of the Ishigami function. The numerical results and discussion have proved the appropriateness of the developed scheme to analyse and develop surrogate models in a wide range of situations, mitigating some of the deficiencies of Ordinary SK when dealing with stochastic processes with high variability. The key findings and contributions of this work include the following:

- The proposed LAR-PCE SK approach has demonstrated superior performance compared to Ordinary SK across various scenarios, including both single and multiple dimensions, with known and unknown intrinsic variance. Benefiting from the model selection capabilities of the LAR algorithm, LAR-PCE SK has shown particularly promising for multidimensional applications.
- Since the computational complexity added by the PCE trend model in Universal SK is marginal, the enhancements in accuracy by the proposed LAR-PCE SK do not entail significant increases in the computational burden compared to Ordinary SK.
- The analysis of Case Studies II and III has revealed that the consideration of a full PCE as a trend basis leads to unacceptable results due to over-overfitting. This stresses the need to consider an optimal sparse basis to effectively enhance the performance of the ordinary SK.

The presented numerical results and discussion suggest the ability of the proposed methodology to be adapted to a large variety of problems in science and engineering. In this light, future developments regard the extension of the presented formulation to other procedures for sparse basis selection. Another interesting objective for future work involves addressing the inherent underestimation of the variance within the SK methodology. A well-selected trend term could have the potential to improve the ability of the stochastic component in SK to accurately capture the overall stochastic nature of the model. Thus, developing a strategy to effectively handle this underestimation could further enhance the performance and robustness of the surrogate modelling approach.

Data availability

Data will be made available on request.

Acknowledgements

This work has been partially supported by the Ministerio de Ciencia e Innovación (Spain) [PID2020-116809GB-I00], by the Junta de Extremadura (Spain) through Research Group Grants [GR18023] by the European Regional Development Fund (ERDF) and by Junta de Extremadura, Spain (Ref. IB20040). E. García-Macías was partially supported by the Ministerio de Ciencia e Innovación (Spain) through the research project "BRIDGEXT - Life-extension of ageing bridges: Towards a long-term sustainable Structural Health Monitoring" (Ref. PID2020-116644RB-I00).

References

[1] L.J. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, Infect. Dis. Model. 2 (2) (2017) 128–142.
[2] A. Mourad, F. Mroue, Z. Taha, Stochastic mathematical models for the spread of COVID-19: a novel epidemiological approach, Math. Med. Biol. 39 (1) (2021) 49–76.
[3] R. Thiedmann, O. Stenzel, A. Spettl, P.R. Shearing, S.J. Harris, N.P. Brandon, V. Schmidt, Stochastic simulation model for the 3d morphology of composite materials in li–ion batteries, Comput. Mater. Sci. 50 (12) (2011) 3365–3376.
[4] S. Meyer, A. Brückner-Foit, A. Möslang, A stochastic simulation model for microcrack initiation in a martensitic steel, Comput. Mater. Sci. 26 (2003) 102–110.
[5] G.K. Sakki, I. Tsoukalas, P. Kossieris, C. Makropoulos, A. Efstratiadis, Stochastic simulation–optimization framework for the design and assessment of renewable energy systems under uncertainty, Renew. Sustain. Energy Rev. 168 (2022) 112886.
[6] C. Luo, S.L. Zhang, C. Wang, Z. Jiang, A metamodel–assisted evolutionary algorithm for expensive optimization, J. Comput. Appl. Math. 236 (5) (2011) 759–764.
[7] [7], K. Khaledi, E. Mahmoudi, M. Datcheva, D. König, T. Schanz, Sensitivity analysis and parameter identification of a time dependent constitutive model for rock salt, J. Comput. Appl. Math. 293 (2016) 128–138.
[8] E. García-Macíias, I. Venanzi, F. Ubertini, Metamodel-based pattern recognition approach for real–time identification of earthquake–induced damage in historic masonry structures, Autom. Constr. 120 (2020) 103389.
[9] J. García-Merino, C. Calvo-Jurado, E. García-Macías, Polynomial chaos expansion for uncertainty propagation analysis in numerical homogenization of 2d/3d periodic composite microstructures, Compos. Struct. 300 (2022) 116130.
[10] R. Alizadeh, J.K. Allen, F. Mistree, Managing computational complexity using surrogate models: a critical review, Res. Eng. Des. 31 (2020) 275–298.
[11] J. Kudela, R. Matousek, Recent advances and applications of surrogate models for finite element method computations: a review, Soft Comput. 26 (24) (2022) 13709–13733.
[12] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.
[13] L. Nuñez, R.G. Regis, K. Varela, Accelerated random search for constrained global optimization assisted by radial basis function surrogates, J. Comput. Appl. Math. 340 (2018) 276–295.
[14] J.P. Kleijnen, Kriging metamodeling in simulation: A review, European J. Oper. Res. 192 (3) (2009) 707–716.
[15] E. Baker, P. Barbillon, A. Fadikar, R.B. Gramacy, R. Herbei, D. Higdon, J. Huang, L.R. Johnson, P. Ma, A. Mondal, et al., Analyzing stochastic computer models: A review with opportunities, Stat. Sci. 37 (1) (2022) 64–89.
[16] V. Moutoussamy, S. Nanty, B. Pauwels, Emulators for stochastic simulation codes, ESAIM: Proc. Surv. 48 (2015) 116–155.
[17] X. Chen, B.E. Ankenman, B.L. Nelson, Enhancing stochastic Kriging metamodels with gradient estimators, Oper. Res. 61 (2) (2013) 512–528.
[18] X. Chen, K.K. Kim, Efficient var and cvar measurement via stochastic Kriging, INFORMS J. Comput. 28 (4) (2016) 629–644.
[19] P. Hao, S. Feng, H. Liu, Y. Wang, B. Wang, B. Wang, A novel nested stochastic Kriging model for response noise quantification and reliability analysis, Comput. Methods Appl. Mech. Engrg. 384 (2021) 113941.
[20] X. Ruan, Q. Zhou, L. Shu, J. Hu, L. Cao, Accurate prediction of the weld bead characteristic in laser keyhole welding based on the stochastic Kriging model, Metals 8 (7) (2018) 486.
[21] J. Pousi, J. Poropudas, K. Virtanen, Game theoretic simulation metamodeling using stochastic Kriging, in: Proceedings of the 2010 Winter Simulation Conference, IEEE, 2010, pp. 1456–1467.
[22] X. Chen, Q. Zhou, Sequential design strategies for mean response surface metamodeling via stochastic Kriging with adaptive exploration and exploitation, European J. Oper. Res. 262 (2) (2017) 575–585.
[23] Z. Wang, M. Ierapetritou, Constrained optimization of black-box stochastic systems using a novel feasibility enhanced Kriging-based method, Comput. Chem. Eng. 118 (2018) 210–223.
[24] M. Binois, R.B. Gramacy, M. Ludkovski, Practical heteroscedastic Gaussian process modeling for large simulation experiments, J. Comput. Graph. Stat. 27 (4) (2018) 808–821.
[25] Y. Che, Z. Guo, C. Cheng, Generalized polynomial chaos-informed efficient stochastic Kriging, J. Comput. Phys. 445 (2021) 110598.
[26] R. Schobi, B. Sudret, J. Wiart, Polynomial–chaos-based Kriging, Int. J. Uncertain. Quant. 5 (2) (2015).
[27] J. García-Merino, C. Calvo-Jurado, E. Martínez-Pañeda, E. García-Macías, Multielement polynomial chaos Kriging–based metamodelling for Bayesian inference of nonsmooth systems, Appl. Math. Model. 116 (2023) 510–531.
[28] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.
[29] S.N. Lophaven, H.B. Nielsen, J. Sondergaard, et al., DACE: A Matlab Kriging Toolbox, Vol. 2, Citeseer, 2002.
[30] T.J. Santner, B.J. Williams, W.I. Notz, B.J. Williams, The Design and Analysis of Computer Experiments, Vol. 1, Springer, 2003.
[31] C.E. Rasmussen, C.K. Williams, et al., Gaussian Processes for Machine Learning, Vol. 1, Springer, 2006.

[32] A. Marrel, B. Iooss, F. Van Dorpe, E. Volkova, An efficient methodology for modeling complex computer codes with Gaussian processes, Comput. Stat. Data Anal. 52 (10) (2008) 4731–4744.
[33] X. Chen, Q. Zhou, Sequential experimental designs for stochastic Kriging, in: Proceedings of the Winter Simulation Conference 2014, IEEE, 2014, pp. 3821–3832.
[34] S.D. Tajbakhsh, E. Del Castillo, J.L. Rosenberger, A Bayesian approach to sequential optimization based on computer experiments, Qual. Reliab. Eng. Int. 31 (6) (2015) 1001–1012.
[35] J. Staum, Better simulation metamodeling: The why, what, and how of stochastic Kriging, in: Proceedings of the 2009 Winter Simulation Conference, WSC, IEEE, 2009, pp. 119–133.
[36] J.P. Kleijnen, E. Mehdad, Estimating the variance of the predictor in stochastic Kriging, Simul. Model. Pract. Theory 66 (2016) 166–173.
[37] D. Xiu, G.E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.
[38] N. Fajraoui, S. Marelli, B. Sudret, Sequential design of experiment for sparse polynomial chaos expansions, ASA J. Uncertain. Quant. 5 (1) (2017) 1061–1085.
[39] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1) (1996) 267–288.
[40] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 2, Springer, 2009.
[41] X. Chen, B.E. Ankenman, B.L. Nelson, The effects of common random numbers on stochastic Kriging metamodels, ACM Trans. Model. Comput. Simul. 22 (2) (2012) 1–20.
[42] M.L. Stein, Interpolation of Spatial Data: Some Theory for Kriging, Springer Science and Business Media, 1999.
[43] J.P.C. Kleijnen, Design and Analysis of Simulation Experiments, Springer, 2018.
[44] S. Zheng, A.F. Seila, Some well–behaved estimators for the m/m/1 queue, Oper. Res. Lett. 26 (5) (2000) 231–235.
[45] G. Xie, X. Chen, Uniform error bounds for stochastic Kriging, in: 2020 Winter Simulation Conference, WSC, IEEE, 2020, pp. 361–372.
[46] [46], B. Sudret, Polynomial chaos expansions and stochastic finite–element methods, in: Risk and Rel. Geoth. Eng., CRC Press, 2018, pp. 295–330.