Polynomial-chaos-based conditional statistics for probabilistic learning with heterogeneous data applied to atomic collisions of Helium on graphite substrate

Christian Soize, Quy-Dong To

▶ **To cite this version:**

Christian Soize, Quy-Dong To. Polynomial-chaos-based conditional statistics for probabilistic learning with heterogeneous data applied to atomic collisions of Helium on graphite substrate. Journal of Computational Physics, 2023, 496, pp.112582. 10.1016/j.jcp.2023.112582 . hal-04260804

HAL Id: hal-04260804
https://univ-eiffel.hal.science/hal-04260804

Submitted on 26 Oct 2023

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

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

Polynomial-chaos-based conditional statistics for probabilistic learning with heterogeneous data applied to atomic collisions of Helium on graphite substrate

Christian Soize[a,*], Quy-Dong To[a]

[a] *Université Gustave Eiffel, MSME UMR 8208, 5 bd Descartes, 77454 Marne-la-Vallée, France*

Abstract

A formulation and an algorithm are presented to construct a truncated polynomial chaos representation of a vector-valued random output. This representation depends on a vector-valued random input with a known probability measure and a vector-valued random latent variable with an unknown probability measure. The construction of this PCE representation relies solely on a training set comprising a small number of independent realizations of the non-Gaussian dependent random output and input vectors. The training set consists of heterogeneous data, which poses challenges in accurately estimating the chaos coefficients. Despite the heterogeneity of the data, the proposed formulation and algorithm allow for the construction of a highly accurate global surrogate model. Additionally, we propose an alternative approach by constructing a surrogate model based on prior separation of the heterogeneous dataset into subsets, each containing "quasi-homogeneous" data. The separation method is designed to account for a partial overlap of the probability measure supports associated with the subsets. The identification of the PCE is performed offline. By utilizing the PCE, a fast online surrogate model is obtained, enabling analysis of large dynamical systems beyond the computational capabilities currently available. An application to atomic collisions of Helium on a graphite substrate is presented, where the training set was generated by Molecular Dynamics simulations done in a previous paper. The obtained results demonstrate accuracy of the proposed approach.

Keywords: Polynomial chaos expansion, Statistical surrogate model, Probabilistic learning, Heterogeneous data, Uncertainty quantification, Atomic collisions.

1. Introduction

Physics problem addressed in the paper and its statistical surrogate model. To solve gas flow problems in engineering applications, a variety of simulation methods have been developed in the literature. They can be classified into two main types: continuum-based methods (Navier Stokes, moment equations, Burnette etc) and particle-based methods (Direct Simulation Monte Carlo, Lattice Boltzmann, Molecular Dynamics, etc.). In addition to the bulk behavior representing fluid-fluid interaction, it must be completed by the interaction between the fluid and the solid boundary. To avoid the huge computation cost relating to the modeling the solid phase, the interaction fluid-solid is usually substituted by statistical surrogate models. The construction of the latter can be done by studying separately the gas-wall collisions using the Molecular Dynamics. The presented application is the case of Helium and graphite at low temperature where complex phenomenon like adsorption and surface diffusion are present and dominant. However, the developed methodology is general and can be used for any complex dynamical systems and in particular, in the context of the application presented, can be applied to any gas-wall couple. The input vector to this statistical surrogate model consists of a realization (sample) $\mathbf{w}^j = \mathbf{v}_{\text{in}}^j$ of the random velocity vector $\mathbf{W}$ (the control variable) representing the velocity of the incident particle on the layer. The output vector is the corresponding realization $\mathbf{q}^j$ of the random vector $\mathbf{Q} = (\mathbf{V}_{\text{out}}, \Delta_t, D_x, D_y)$. This vector includes the random reflected velocity vector $\mathbf{V}_{\text{out}}$ (output) of the particle, the random absorption duration Δ_t (residence time), and the two displacement components D_x and

*Corresponding author: C. Soize, christian.soize@univ-eiffel.fr
Email addresses: christian.soize@univ-eiffel.fr (Christian Soize), quy-dong.to@univ-eiffel.fr (Quy-Dong To)

D_y in the (oxy) plane within the layer. There exist two primary physical regimes depending on the realization $\mathbf{w}^j$ of $\mathbf{W}$. In one regime, the particle undergoes quasi-reflection by the layer, resulting in a short absorption time. In the other regime, the particle is absorbed within the layer and then emerges after a more or less extended random duration and with random displacements in the plane of the layer. On the other hand, the dynamical system is extremely complex, and the state $\mathbf{Q}$ cannot be solely explained by $\mathbf{W}$. This implies that a deterministic mapping such as $\mathbf{Q} = \tilde{\mathbf{f}}(\mathbf{W})$ does not exist. Instead, there exists a random mapping $\mathbf{F}$ such that $\mathbf{Q} = \mathbf{F}(\mathbf{W})$. To construct the statistical surrogate model, it is necessary to introduce a hidden explanatory vector-valued random variable (latent variable). We can introduce a random vector $\mathbf{U}$ of unknown dimension, which is assumed to be statistically independent of $\mathbf{W}$. Thus, the relationship for $\mathbf{Q}$ can be expressed as $\mathbf{Q} = \mathbf{f}(\mathbf{W}, \mathbf{U})$. The only available information consists of the realizations $\{(\mathbf{q}_d^j, \mathbf{w}_d^j), j = 1, \ldots, n_d\}$ of $(\mathbf{Q}, \mathbf{W})$ obtained from the MD simulations, where n_d is relatively small. The objective is to construct a truncated polynomial chaos representation, $\mathbf{Q}^{\text{chaos}} = \mathbf{f}^{\text{chaos}}(\mathbf{W}, \mathbf{U})$, of $\mathbf{Q} = \mathbf{F}(\mathbf{W})$, which defines the statistical surrogate model for $\mathbf{Q}$. A probability measure $P_{\mathbf{U}}(d\mathbf{u})$ for $\mathbf{U}$ is then constructed to make the probability measure of $\mathbf{Q}^{\text{chaos}}$ as close as possible to the probability measure of $\mathbf{Q}$. This representation allows us to generate a realization $\mathbf{q}_0$ of $\mathbf{Q}^{\text{chaos}} \simeq \mathbf{Q}$ corresponding to a given realization $\mathbf{w}_0$ of $\mathbf{W}$, ensuring that $(\mathbf{q}_0, \mathbf{w}_0)$ is a consistent realization with respect to the probability measure of $(\mathbf{Q}, \mathbf{W})$. Therefore, computing the realization $\mathbf{q}_0$ of $\mathbf{Q}^{\text{chaos}}$, given $\mathbf{W} = \mathbf{w}_0$, is done quickly using $\mathbf{q}_0 = \mathbf{f}^{\text{chaos}}(\mathbf{w}_0, \mathbf{u}_0)$, where $\mathbf{u}_0$ is any realization of $\mathbf{U}$ (which is independent of $\mathbf{w}_0$).

Main difficulties related to the construction of the statistical surrogate model. There are three main difficulties. The first is that only a small training dataset $\{(\mathbf{q}_d^j, \mathbf{w}_d^j), j = 1, \ldots, n_d\}$ of dimension n_d is available. The number of points, n_d, is too small to construct the chaos representation $\mathbf{Q}^{\text{chaos}}$ of $\mathbf{Q}$. Therefore, it is necessary to generate a large learned dataset $\{(\mathbf{q}^\ell, \mathbf{w}^\ell), \ell = 1, \ldots, N\}$ consisting of $N \gg n_d$ learned realizations that follow the probability measure of $(\mathbf{Q}, \mathbf{W})$. This dataset will be created using only the information provided by the training dataset. The second problem is associated with the presence of the hidden random vector $\mathbf{U}$, of which we do not know the dimension and its probability measure. The third is related to the training dataset, which comprises heterogeneous data. Within this dataset, there is a combination of data related to two different regimes of physics processes. As we have explained, one regime corresponds to the scenario where the particle reflects off the layer, meaning there is no absorption of the particle. In this case, the duration Δ_t and the displacements (D_x, D_y) are small. The other regime corresponds to the possibility of particle absorption by the layer. Here, the particle moves inside the layer and emerges after a random duration Δ_t and random displacements (D_x, D_y), which can be large. Separating these two probabilistic phenomena (and therefore separating the points of the training dataset into distinct clusters) is challenging due to partial overlap in the supports of the probability measures associated with these two regimes. Consequently, the training dataset consists of heterogeneous data.

Rewriting the addressed physics problem within a broader context. We will present a general methodology that can be applied to other situations. To do this, we reformulate in a broader context, the problem we presented earlier for the specific physics problem in question. We consider a large-scale stochastic computational model that depends on an unknown (and therefore uncontrolled) random parameter, which is modeled by a random variable with value $\mathbb{R}^{n_u}$, denoted $\mathbf{U}$ (the latent variable). The dimension n_u and the probability measure $P_{\mathbf{U}}(d\mathbf{u})$ of $\mathbf{U}$ are both unknown. The control parameter is the random variable $\mathbf{W}$ with values in $\mathbb{R}^{n_w}$, whose probability measure is $P_{\mathbf{W}}(d\mathbf{w})$. We assume that $\mathbf{W}$ and $\mathbf{U}$ are independent, so we have $P_{\mathbf{W},\mathbf{U}}(d\mathbf{w}, d\mathbf{u}) = P_{\mathbf{W}}(d\mathbf{w}) \otimes P_{\mathbf{U}}(d\mathbf{u})$. The quantity of interest is the $\mathbb{R}^{n_q}$-valued random variable $\mathbf{Q} = \mathbf{F}(\mathbf{W})$. Let $\mathscr{C}_w \subset \mathbb{R}^{n_w}$ be the support of $P_{\mathbf{W}}(d\mathbf{w})$. The random mapping $\mathbf{F}$ is unknown. The joint probability measure of $\mathbf{Q}$ and $\mathbf{W}$ is denoted by $P_{\mathbf{Q},\mathbf{W}}(d\mathbf{q}, d\mathbf{w})$. In such a scenario, the random mapping $\mathbf{F}$ can be rewritten as $\mathbf{w} \mapsto \mathbf{F}(\mathbf{w}) = \mathbf{f}(\mathbf{w}, \mathbf{U})$ in which the deterministic mapping $\mathbf{f}$ is also unknown. Let $\mathbf{X} = (\mathbf{Q}, \mathbf{W})$ be the $\mathbb{R}^{n_x}$-valued random variable, where $n_x = n_q + n_w$, and its probability measure is denoted by $P_{\mathbf{X}}(d\mathbf{x})$. Random vector $\mathbf{X}$ is associated with the random manifold defined by the random graph $\{(\mathbf{F}(\mathbf{w}), \mathbf{w}), \mathbf{w} \in \mathscr{C}_w\}$ and is fully characterized by its probability measure $P_{\mathbf{X}}(d\mathbf{x})$, which represents the joint probability measure $P_{\mathbf{Q},\mathbf{W}}(d\mathbf{q}, d\mathbf{w})$ of $\mathbf{Q}$ and $\mathbf{W}$. For this problem, the only available information is the training set $\mathcal{D}_{\text{train}}(\mathbf{X}) = \{\mathbf{x}_d^j, j = 1, \ldots, n_d\}$ consisting of n_d independent realizations $\mathbf{x}_d^j = (\mathbf{q}_d^j, \mathbf{w}_d^j) \in \mathbb{R}^{n_x}$ of the $\mathbb{R}^{n_x}$-valued random variable $\mathbf{X} = (\mathbf{Q}, \mathbf{W})$. It is assumed that n_d is small. For each given realization $\mathbf{w}_d^j$ of $\mathbf{W}$, the corresponding realization $\mathbf{q}_d^j$ of $\mathbf{Q}$ is obtained from the computational model. However, there are no corresponding realizations $\mathbf{u}_d^\ell$ of $\mathbf{U}$ available for the given $\mathbf{x}_d^j$. In addition, as we have explained, it will be necessary to construct a large learned dataset $\mathcal{D}_{\text{learn}}(\mathbf{X}) = \{\mathbf{x}^\ell, \ell = 1, \ldots, N\}$ with $N \gg n_d$ in order

to build $\mathbf{f}^{\text{chaos}}$. The estimation of the probability measure $P_{\mathbf{X}}$ will be performed using the learned dataset. Note that the realization $\mathbf{u}^{\ell}$ of $\mathbf{U}$ corresponding to $\mathbf{x}^{\ell}$ is also not usable.

Short overview of the related works on polynomial chaos expansion (PCE) for developing a methodology to solve the problem. Given the statistical dependence between $\mathbf{Q}$ and $\mathbf{W}$, the joint probability measure $P_{\mathbf{Q},\mathbf{W}}(d\mathbf{q}, d\mathbf{w}) = P_{\mathbf{X}}(d\mathbf{x})$ can be accurately estimated using the large learned dataset $\mathcal{D}_{\text{learn}}(\mathbf{X})$ as well as the probability measure $P_{\mathbf{W}}$ of $\mathbf{W}$. If $\mathbf{Q}$ did not depend on $\mathbf{U}$, then the problem would easily be obtained by a projection on the chaos polynomials constructed with $P_{\mathbf{W}}$. When $\mathbf{Q}$ depends on $\mathbf{U}$, the coefficients obtained with such a projection become random coefficients yielding a truncated PCE $\mathbf{Q}^{\text{chaos}}$ with random coefficients for which a detailed analysis has been proposed in [1].

The concept of PCE for stochastic processes was first introduced by Wiener and Cameron in their seminal works [2, 3], while Ghanem and colleagues pioneered an effective Karhunen-Loève-based construction for random fields [4, 5]. The Wiener-Askey PCE was employed by Xiu [6], and the development of random fields in polynomial chaos for arbitrary probability measures was introduced by Soize [7]. The PCE with random coefficients were explored by [8, 1, 9], while Tipireddy presented a basis adaptation in homogeneous chaos spaces [10]. A compressed principal component analysis of non-Gaussian vectors using symmetric polynomial chaos was proposed by Mignolet [11]. Significant works are also devoted to the acceleration of stochastic convergence of PCE [12, 13, 14, 10, 15]. Polynomial chaos expansions have been and continue to be intensively used in both finite and infinite dimensions for uncertainty modeling and propagation [16, 17, 18, 19, 20, 1, 21, 15] (see also hereinafter the stochastic solvers and the stochastic finite elements).

After constructing probabilistic models of uncertainties, it becomes essential to investigate how these uncertainties propagate within systems. This requires the use of methods for solving stochastic equations. The initial set of methods is grounded on Monte Carlo numerical simulation techniques. The second set is based on spectral projection methods [22, 23, 24, 25, 26, 27, 28], such as those based on polynomial chaos expansions [4, 5, 29, 30] and called stochastic finite element method when the discretization method of the boundary value problems are performed using the finite element method [5, 31, 32, 33, 34, 35, 36] and also [37, 38, 39, 40, 41, 42].

Most often, the uncertainties probability model is a prior model. If targets are available for observations of the system, coming from experimental measurements or from more precise numerical simulations, a posterior probability model of uncertainties can be estimated by solving inverse statistical problems based on the maximum likelihood, the Bayesian inference, and machine learning. For general overviews on statistical inverse methods, see [43, 44, 45, 46, 47, 48, 49, 50], and for complements, see [51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]. The statistical identification of the coefficients of polynomial chaos representations of random fields can be found in [62, 63, 64], in particular in [65, 66, 67] for high dimension, and for representations of random vectors in [14, 68, 69, 70, 71, 72, 73]. The inverse identification of random matrices have been proposed in [74, 75, 76]. Statistical inverse methods are also used to perform model updating [77, 78, 79, 80], model selection [81, 82], and to construct surrogate models (or metamodels) [83, 84, 85, 86, 87, 88, 89].

The machine learning tools and artificial intelligence [90, 91, 92, 93] provide methods that make it possible to solve problems in UQ in the field of physics and engineering sciences. These problems could not be solved without these learning methods because the use of the usual methods would require computer resources, which are not available. Regarding these learning methods, let us cite, for example, learning with kernels [94, 95, 96], probabilistic and statistical learning [97, 98, 99, 100], learning on the manifolds [101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113], and probabilistic physics-based learning [114, 115, 116, 117, 118].

Heterogeneous data are ubiquitous in many scientific fields, and their importance continues to grow. As a result, research on heterogeneous data is important, leading to the emergence of new methods and techniques for analysis and modeling. Common methods for analyzing heterogeneous data at various levels of abstraction include regression methods that use polynomial chaos representations, kriging methods that are statistical interpolation techniques, and neural network and deep learning methods that can model complex relationships between input and output variables but require large training datasets. More advanced techniques comprise Hidden Markov Models (HMM) [119] that are particularly useful for modeling time series data with different regimes, Dirichlet Processes (DP) [120] that are nonparametric probabilistic models capable of clustering, hierarchical clustering methods that group similar data into clusters [121], Probabilistic Graphical Models (PGM)[122] that are effective for modeling systems with complex variable interactions, Hidden Markov Networks (HMN) [123] that are suitable for modeling complex systems with

unknown graph structures, and kernel algorithm [95].

Novelty. The novelty of this work is directly linked to the main challenges that we have identified. First, we present a formulation and an algorithm to build a surrogate statistical model in the form of a truncated polynomial chaos representation, $\mathbf{Q}^{\text{chaos}} = \mathbf{f}^{\text{chaos}}(\mathbf{W}, \mathbf{U})$, of the vector random variable $\mathbf{Q} = \mathbf{F}(\mathbf{W})$. In this context, $\mathbf{F}$ represents an unknown random mapping, and the available information solely consists of n_d independent realizations of the non-Gaussian dependent random vector $(\mathbf{Q}, \mathbf{W})$, where the value of n_d is small and the realizations correspond to heterogeneous data. Due to this limited heterogeneous dataset, accurately estimating the chaos coefficients becomes challenging. Furthermore, the vector-valued random variable $\mathbf{U}$ represents the latent random variables within $\mathbf{F}$, and it is necessary to determine its dimension and to chose its probability measure. Although the data is heterogeneous, the proposed formulation and algorithm enable the construction of a global surrogate model with excellent accuracy. Although the global surrogate model is highly accurate, we also propose an alternative approach by constructing a surrogate model based on a prior separation of the heterogeneous dataset into subsets, each consisting of "quasi-homogeneous" data. The proposed separation method is developed in the context of the existence of a partial overlap of the supports of probability measures associated with the subsets. It should be noted that in this case, no method can achieve an "exact" separation based solely on the available information.

Organization of the paper. It should be noted that Section 1 presents the physics problem addressed in the paper. It provides a definition of the statistical surrogate model of interest and presents the broader context in which the physics problem is posed. Section 2 focuses on the generation of a large learned dataset from a given small training dataset. To circumvent the numerical difficulties during the construction of the polynomial chaos representations, a normalization and a scaling of the learned dataset are carried out. This involves transforming $\mathbf{Q}$ and $\mathbf{W}$ into normalized/scaled random variables, denoted as $\mathbf{Y}$ and $\boldsymbol{\Xi}$ respectively. Section 3 deals with the construction of the non-separated multivariate polynomial chaos for $\boldsymbol{\Xi}$, while Section 4 addresses the construction of separated multivariate polynomial chaos for the latent random variable $\mathbf{U}$. In Section 5, we present the polynomial chaos expansion of random vector $\mathbf{Y}$. This expansion is obtained in tensorizing the two Hilbert bases related to $\boldsymbol{\Xi}$ and $\mathbf{U}$. Subsequently, we can construct the truncated polynomial chaos expansion of random vector $\mathbf{Y}$, which serves as the basis for constructing the statistical surrogate model. Section 7 is devoted to the identification of the unknown coefficients in the truncated PCE of $\mathbf{Y}$. The optimization problem involved is nonconvex, with a constraint, and its cost function is based on the Overlap criterion. To simplify the search for an optimal solution, we transform this problem into an unconstrained optimization problem. Section 8 finalizes the first part dedicated to the construction of the global statistical surrogate model, which is based on a representation of $\mathbf{Q}$ in polynomial chaos of $\mathbf{W}$ and $\mathbf{U}$. In section 9, we present an alternative approach, which consists in using the separation into clusters of the learned dataset, to construct the conditional statistics based on the polynomial chaos. The last section deals with the application to atomic collisions of Helium on graphite substrate.

Notations

x, η: lower-case Latin or Greek letters are deterministic real variables.
$\mathbf{x}, \boldsymbol{\eta}$: boldface lower-case Latin or Greek letters are deterministic vectors.
X: upper-case Latin letters are real-valued random variables.
$\mathbf{X}$: boldface upper-case Latin letters are vector-valued random variables.
$[x]$: lower-case Latin letters between brackets are deterministic matrices.
$[\mathbf{X}]$: boldface upper-case letters between brackets are matrix-valued random variables.

$\mathbb{C}$: set of all the complex numbers.
$\mathbb{M}_{n,m}$: set of the $(n \times m)$ real matrices.
$\mathbb{M}_n$: set of the square $(n \times n)$ real matrices.
$\mathbb{M}_n^+$: set of the positive-definite $(n \times n)$ real matrices.
$\mathbb{M}_n^{+0}$: set of the positive $(n \times n)$ real matrices.
N: number of points in the learned dataset.
n_d: number of points in the training dataset.
$\mathbb{N}$: set of all the integers $\{0, 1, 2, \ldots\}$.
$\mathbb{R}$: set of all the real number.
$\mathbb{R}^n$: Euclidean space of dimension n.
$[I_n]$: identity matrix in $\mathbb{M}_n$.
$\mathbf{x} = (x_1, \ldots, x_n)$: point in $\mathbb{R}^n$.
$[x]^T$: transpose of matrix $[x]$.
$\text{tr}\{[x]\}$: trace of the square matrix $[x]$.
$\boldsymbol{\alpha}$: multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ in $\mathbb{N}^n$.
δ_0: Dirac measure on $\mathbb{R}$ at point 0.
$\delta_{kk'}$: Kronecker's symbol.
$\delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$: $\delta_{\alpha_1\beta_1} \times \ldots \times \delta_{\alpha_n\beta_n}$.

E: mathematical expectation operator.
KDE: kernel density estimation.
PDF: probability density function.
PCE: polynomial chaos expansion.

Convention used for random variables. In this paper, for any finite integer $m \geq 1$, the Euclidean space $\mathbb{R}^m$ is equipped with the σ-algebra $\mathcal{B}_{\mathbb{R}^m}$. If $\mathbf{Y}$ is a $\mathbb{R}^m$-valued random variable defined on the probability space $(\Theta, \mathcal{T}, \mathcal{P})$, $\mathbf{Y}$ is a mapping $\theta \mapsto \mathbf{Y}(\theta)$ from Θ into $\mathbb{R}^m$, measurable from $(\Theta, \mathcal{T})$ into $(\mathbb{R}^m, \mathcal{B}_{\mathbb{R}^m})$, and $\mathbf{Y}(\theta)$ is a realization (sample) of $\mathbf{Y}$ for $\theta \in \Theta$. The probability measure of $\mathbf{Y}$ is the probability measure $P_{\mathbf{Y}}(d\mathbf{y})$ on the measurable set $(\mathbb{R}^m, \mathcal{B}_{\mathbb{R}^m})$ (we will simply say on $\mathbb{R}^m$). The Lebesgue measure on $\mathbb{R}^m$ is noted $d\mathbf{y}$ and when $P_{\mathbf{Y}}(d\mathbf{y})$ is written as $p_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}$, $p_{\mathbf{Y}}$ is the probability density function (PDF) on $\mathbb{R}^m$ of $P_{\mathbf{Y}}(d\mathbf{y})$ with respect to $d\mathbf{y}$.

2. Generation of a large learned dataset, its normalization and scaling

Generation of a large learned dataset. As described in Section 1, a large learned dataset $\mathcal{D}_{\text{learn}}(\mathbf{X}) = \{\mathbf{x}^\ell, \ell = 1, \ldots, N\}$ is generated from the training dataset $\mathcal{D}_{\text{train}}(\mathbf{X}) = \{\mathbf{x}_q^j, j = 1, \ldots, n_d\}$, where $N \gg n_d$. The learned dataset is generated using the PLoM algorithm under constraints [102, 108, 115, 110] to enforce the learned probability measure to match the given mean value and covariance matrix. From this, we obtain the learned realizations $\{\mathbf{q}^\ell, \ell = 1, \ldots, N\}$ and $\{\mathbf{w}^\ell, \ell = 1, \ldots, N\}$ for the random vectors $\mathbf{W}$ and $\mathbf{Q}$, respectively, where $(\mathbf{q}^\ell, \mathbf{w}^\ell) = \mathbf{x}^\ell \in \mathbb{R}^{n_q} \times \mathbb{R}^{n_w} = \mathbb{R}^{n_x}$.

Scaling the control random parameter $\mathbf{W}$ to obtain the random vector $\boldsymbol{\Xi}$. Since $\mathbf{W} = (W_1, \ldots, W_{n_w})$ is a $\mathbb{R}^{n_w}$-valued random variable, in order to avoid numerical difficulties while constructing its polynomial chaos, we introduce a scaled $\mathbb{R}^{n_w}$-valued random variable $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_{n_w})$ such that $\mathbf{W} = \mathbf{s}_w(\boldsymbol{\Xi})$ in which $\mathbf{s}_w = (s_{w,1}, \ldots, s_{w,n_w})$ is the mapping from $[-1\,, +1]^{n_w}$ into $\mathbb{R}^{n_w}$ such that, for all $k \in \{1, \ldots, n_w\}$, we have $w_k = s_{w,k}(\xi_k) = a_k\, \xi_k + b_k$ in which a_k and b_k are such that $s_{w,k}(-1) = \min_\ell w_k^\ell$ and $s_{w,k}(+1) = \max_\ell w_k^\ell$. The support of the probability measure $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$ of $\boldsymbol{\Xi}$ on $\mathbb{R}^{n_w}$ is chosen as the compact subset $\mathscr{C}_\xi = [-1\,, +1]^{n_w} \subset \mathbb{R}_w^n$. The N independent realizations $\{\boldsymbol{\xi}^\ell, \ell = 1, \ldots, N\}$ of $\boldsymbol{\Xi}$ are given by $\boldsymbol{\xi}^\ell = \mathbf{s}_w^{-1}(\mathbf{w}^\ell)$. With such a scaling, the support of the probability measure $P_{\mathbf{W}}(d\mathbf{w})$ is defined as the compact subset $\mathscr{C}_w = \mathbf{s}_w(\mathscr{C}_\xi)$. We then have introduced the mapping,

$$\boldsymbol{\xi} \mapsto \mathbf{w} = \mathbf{s}_w(\boldsymbol{\xi}) \,:\, \mathscr{C}_\xi \to \mathbb{R}^{n_w} \quad \text{such that} \quad \mathbf{W} = \mathbf{s}_w(\boldsymbol{\Xi})\,. \tag{2.1}$$

Normalization and scaling of the random vector $\mathbf{Q}$ to obtain random vector $\mathbf{Y}$. To address potential numerical difficulties in constructing the polynomial chaos expansion of $\mathbf{Q}$, we employ a normalization technique that involves a principal component analysis (PCA) followed by scaling. This process yields the $\mathbb{R}^{n_q}$-valued random variable $\mathbf{Y}$. Let $\underline{\mathbf{q}}$ be the empirical mean value of $\mathbf{Q}$ and $[\widehat{C}_{\mathbf{Q}}]$ its empirical covariance matrix, which are estimated with the independent realizations $\{\mathbf{q}^\ell, \ell = 1, \ldots, N\}$. Note that $n_q \ll N$ and we assume that $[\widehat{C}_{\mathbf{Q}}]$ belongs to $\mathbb{M}_{n_q}^+$. Let $[V] \in \mathbb{M}_{n_q}$ be the orthogonal matrix, such that $[V]^T[V] = [V]\,[V]^T = [I_{n_q}]$, constituted of the eigenvectors of matrix $[\widehat{C}_{\mathbf{Q}}]$ and let $[\zeta] \in \mathbb{M}_{n_q}^+$ be the diagonal matrix of the eigenvalues that are all positive. We thus have $[\widehat{C}_{\mathbf{Q}}] = [V]\,[\zeta]\,[V]^T$. We then define the normalized $\mathbb{R}^{n_q}$-valued random variable $\mathbf{R}$ such that $\mathbf{Q} = \underline{\mathbf{q}} + [V]\,[\zeta]^{1/2}\,\mathbf{R}$ and consequently, $\mathbf{R} = [\zeta]^{-1/2}[V]^T(\mathbf{Q} - \underline{\mathbf{q}})$. The realizations $\{\mathbf{r}^\ell, \ell = 1, \ldots, N\}$ of $\mathbf{R}$ are therefore computed by $\mathbf{r}^\ell = [\zeta]^{-1/2}[V]^T(\mathbf{q}^\ell - \underline{\mathbf{q}})$. Consequently, the empirical mean value $\underline{\mathbf{r}}$ and the empirical covariance matrix $[\widehat{C}_{\mathbf{R}}]$ estimated with $\{\mathbf{r}^\ell, \ell = 1, \ldots, N\}$ are such that $\underline{\mathbf{r}} = \mathbf{0}_{n_q}$ and $[\widehat{C}_{\mathbf{R}}] = [I_{n_q}]$. The random variable $\mathbf{R}$ is now scaled in a random variable $\mathbf{Y}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_{n_q})$ be the scaled $\mathbb{R}^{n_q}$-valued random variable $\mathbf{R} = (R_1, \ldots, R_{n_q})$ such that for all $i \in \{1, \ldots, n_q\}$, we have $R_i = s_i \times Y_i$ in which $s_i = (\max_\ell |r_i^\ell|) > 0$, yielding $Y_i = R_i/s_i$. The N independent realizations $\{\mathbf{y}^\ell, \ell = 1, \ldots, N\}$ of $\mathbf{Y}$ are such that $y_i^\ell = r_i^\ell/s_i$. The composition of these two transformations allows the bijective mapping $\mathbb{q}$ to be defined,

$$\mathbf{y} \mapsto \mathbf{q} = \mathbb{q}(\mathbf{y}) \,:\, \mathbb{R}^{n_q} \to \mathbb{R}^{n_q} \quad \text{such that} \quad \mathbf{Q} = \mathbb{q}(\mathbf{Y}) \quad \text{and} \quad \mathbf{Y} = \mathbb{q}^{-1}(\mathbf{Q})\,. \tag{2.2}$$

The empirical estimate of the covariance matrix $[\widehat{C}_{\mathbf{Y}}] \in \mathbb{M}_{n_q}^+$ of $\mathbf{Y}$ is such that $[\widehat{C}_{\mathbf{Y}}]_{ii'} = (s_i)^{-2}\delta_{ii'}$. Since $\mathbf{R}$ is centered, $\mathbf{Y}$ will be centered, and the second-order moment-matrix $[\mathcal{M}_{\mathbf{Y}}] = E\{\mathbf{Y}\,\mathbf{Y}^T\} \in \mathbb{M}_{n_q}^+$ estimated by $[\mathcal{M}_{\mathbf{Y}}] = (N-1)^{-1}\sum_{\ell=1}^N \mathbf{y}^\ell(\mathbf{y}^\ell)^T$ is equal to $[\widehat{C}_{\mathbf{Y}}]$. Matrix $[\mathcal{M}_{\mathbf{Y}}]$ can then be rewritten as $[\mathcal{M}_{\mathbf{Y}}] = [L_{\mathbf{Y}}]^2$ with $[L_{\mathbf{Y}}]_{ii'} = (s_i)^{-1}\delta_{ii'}$. The

mapping $\mathbf{f}$ from $\mathbb{R}^{n_w} \times \mathbb{R}^{n_u}$ into $\mathbb{R}^{n_q}$, which have defined in Section 1 and which is such that $\mathbf{Q} = \mathbf{f}(\mathbf{W}, \mathbf{U})$, is then transformed in a mapping $\mathbb{y}$ from $\mathbb{R}^{n_w} \times \mathbb{R}^{n_u}$ into $\mathbb{R}^{n_q}$ such that

$$\mathbf{Y} = \mathbb{y}(\mathbf{\Xi}, \mathbf{U}) = \mathbb{q}^{-1}(\mathbf{f}(\mathbf{s}_w(\mathbf{\Xi}), \mathbf{U})) \,. \tag{2.3}$$

3. Non-separated multivariate polynomial chaos for $\mathbf{\Xi}$

Hilbert space $\mathbb{H}$ associated with the random vector $\mathbf{\Xi}$. Let $\mathbf{\Xi} = (\Xi_1, \ldots, \Xi_{n_w})$ be the $\mathbb{R}^{n_w}$-valued random variable defined in Section 2. Let $\mathbb{H} = L^2_{P_{\mathbf{\Xi}}}(\mathbb{R}^{n_w}, \mathbb{R})$ be the Hilbert space of all the functions from $\mathbb{R}^{n_w}$ into $\mathbb{R}$, equipped with the inner product $\langle \mathbb{h} \,, \widetilde{\mathbb{h}} \rangle_{\mathbb{H}} = \int_{\mathbb{R}^{n_w}} \mathbb{h}(\boldsymbol{\xi})\, \widetilde{\mathbb{h}}(\boldsymbol{\xi})\, P_{\mathbf{\Xi}}(d\boldsymbol{\xi})$, and the associated norm $\| \mathbb{h} \|_{\mathbb{H}} = \langle \mathbb{h} \,, \mathbb{h} \rangle^{1/2}_{\mathbb{H}}$. For any $\mathbb{h}$ in $\mathbb{H}$, $H = \mathbb{h}(\mathbf{\Xi})$ is a second-order real-valued random variable such that $E\{H^2\} = E\{\mathbb{h}(\mathbf{\Xi})^2\} = \int_{\mathbb{R}^{n_w}} \mathbb{h}(\boldsymbol{\xi})^2\, P_{\mathbf{\Xi}}(d\boldsymbol{\xi}) < +\infty$.

Non-separated multivariate polynomial chaos in $\mathbb{H}$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_w}) \in \mathbb{N}^{n_w}$ be the multi-index, which includes $\boldsymbol{\alpha}^{(1)} = (0, \ldots, 0)$. For all $\boldsymbol{\alpha}$ in $\mathbb{N}^{n_w}$, let $\Psi_{\alpha_1,\ldots,\alpha_{n_w}}(\xi_1, \ldots, \xi_{n_w})$, rewritten as $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$, be the multivariate polynomials (non-separated with respect to $\xi_1, \ldots, \xi_{n_w}$), which are orthonormal in $\mathbb{H}$, $\langle \Psi_{\boldsymbol{\alpha}} \,, \Psi_{\boldsymbol{\beta}} \rangle_{\mathbb{H}} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, and such that $\Psi_{\boldsymbol{\alpha}^{(1)}}(\boldsymbol{\xi}) = 1$. It is known that $\{\Psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{N}^{n_w}\}$ is a Hilbert basis of $\mathbb{H}$.

Polynomial chaos expansion of $H = \mathbb{h}(\mathbf{\Xi})$. For any $\mathbb{h}$ in $\mathbb{H}$, the PCE of the second-order random variable H is written as $H = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^{n_w}} h^{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{\Xi})$, where the series of the PCE converges with respect to the norm of $\mathbb{H}$. The real coefficients $h^{\boldsymbol{\alpha}}$ are such that $\|H\|^2_{\mathbb{H}} = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^{n_w}} (h^{\boldsymbol{\alpha}})^2 < +\infty$ and can be calculated by $h^{\boldsymbol{\alpha}} = \langle H \,, \Psi_{\boldsymbol{\alpha}} \rangle_{\mathbb{H}} = E\{H\, \Psi_{\boldsymbol{\alpha}}(\mathbf{\Xi})\} = E\{\mathbb{h}(\mathbf{\Xi})\, \Psi_{\boldsymbol{\alpha}}(\mathbf{\Xi})\} = \int_{\mathbb{R}^{n_w}} \mathbb{h}(\boldsymbol{\xi})\, \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\, P_{\mathbf{\Xi}}(d\boldsymbol{\xi})$.

Truncated PCE $H^{\text{chaos}} = \mathbb{h}^{\text{chaos}}(\mathbf{\Xi})$ for representing $H = \mathbb{h}(\mathbf{\Xi})$. Let N_g be the maximum degree of the considered truncated polynomial chaos expansion. We then have $|\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_{n_w} \leq N_g$. The set $\{\boldsymbol{\alpha} \in \mathbb{N}^{n_w}, |\boldsymbol{\alpha}| \leq N_g\}$ of all the multi-indices in $\mathbb{N}^{n_w}$ such that $|\boldsymbol{\alpha}| \leq N_g$ is rewritten as $\{\boldsymbol{\alpha}^{(k)} = (\alpha^{(k)}_1, \ldots, \alpha^{(k)}_{n_w}) \in \mathbb{N}^{n_w}, k = 1, \ldots, \kappa\}$ with $\boldsymbol{\alpha}^{(1)} = (0, \ldots, 0)$ and where $\kappa = (N_g + n_w)!/(N_g!\, n_w!)$. To simplify the notation, the polynomial chaos of multi-index $\boldsymbol{\alpha}^{(k)}$ is rewritten as $\psi_k(\mathbf{\Xi}) = \Psi_{\boldsymbol{\alpha}^{(k)}}(\mathbf{\Xi})$. We then have,

$$\psi_1(\mathbf{\Xi}) = 1 \quad , \quad \langle \psi_k \,, \psi_{k'} \rangle_{\mathbb{H}} = \delta_{kk'} \quad , \quad E\{\psi_k(\mathbf{\Xi})\} = \delta_{1k} \,. \tag{3.1}$$

The truncated PCE of $H = \mathbb{h}(\mathbf{\Xi})$ is thus given by

$$H^{\text{chaos}} = \mathbb{h}^{\text{chaos}}(\mathbf{\Xi}) = \sum_{k=1}^{\kappa} h^k\, \psi_k(\mathbf{\Xi}) \quad , \quad h^k = \langle H \,, \psi_k \rangle_{\mathbb{H}} \,, \tag{3.2}$$

in which h^k is the rewriting of $h^{\boldsymbol{\alpha}^{(k)}}$.

4. Separated multivariate polynomial chaos for $\mathbf{U}$

As previously mentioned, the vector-valued random variable $\mathbf{U}$ is a latent variable for which no information is available. Consequently, in the context of PCE, we choose $\mathbf{U}$ as a normalized $\mathbb{R}^{n_u}$-valued Gaussian random variable whose probability measure is written as $P_{\mathbf{U}}(d\mathbf{u}) = (2\pi)^{-n_u/2} \exp(-\|\mathbf{u}\|^2/2)\, d\mathbf{u}$. The dimension $n_u \geq 1$ is unknown and needs to be determined. We use notations similar to those introduced in Section 3.

Hilbert space $\mathbb{G}$ associated with the random vector $\boldsymbol{U}$. Let $\mathbb{G} = L^2_{P_{\mathbf{U}}}(\mathbb{R}^{n_u}, \mathbb{R})$ be the Hilbert space of all the functions from $\mathbb{R}^{n_u}$ into $\mathbb{R}$, equipped with the inner product $\langle \mathbb{g} \,, \widetilde{\mathbb{g}} \rangle_{\mathbb{G}} = \int_{\mathbb{R}^{n_u}} \mathbb{g}(\mathbf{u})\, \widetilde{\mathbb{g}}(\mathbf{u})\, P_{\mathbf{U}}(d\mathbf{u})$, and the associated norm $\| \mathbb{g} \|_{\mathbb{G}} = \langle \mathbb{g} \,, \mathbb{g} \rangle^{1/2}_{\mathbb{G}}$. For any $\mathbb{g}$ in $\mathbb{G}$, $G = \mathbb{g}(\mathbf{U})$ is a second-order real-valued random variable such that $E\{G^2\} = E\{\mathbb{g}(\mathbf{U})^2\} = \int_{\mathbb{R}^{n_u}} \mathbb{g}(\mathbf{u})^2\, P_{\mathbf{U}}(d\mathbf{u}) < +\infty$.

Normalized multivariate Hermite polynomials as the separated polynomial chaos in $\mathbb{G}$. Let $\mathbf{a} = (a_1, \ldots, a_{n_u}) \in \mathbb{N}^{n_u}$ be the multi-index, which includes the zero multi-index $a^{(1)} = (0, \ldots, 0)$. For all $\mathbf{a}$ in $\mathbb{N}^{n_u}$, let $\Phi_{a_1,\ldots,a_{n_u}}(u_1, \ldots, u_{n_u}) =$

$\Phi_{a_1}(u_1)\times\ldots\times\Phi_{a_{n_u}}(u_{n_u})$, rewritten as $\Phi_{\mathbf{a}}(\mathbf{u})$, such that $\Phi_{\mathbf{a}^{(1)}}(\mathbf{u})=1$, and where Φ_{a_i} are the normalized Hermite polynomials on $\mathbb{R}$. Therefore, $\{\Phi_{\mathbf{a}}, \mathbf{a}\in\mathbb{N}^{n_u}\}$ is an orthonormal family in $\mathbb{G}$, $\langle\Phi_{\mathbf{a}}\,,\Phi_{\mathbf{b}}\rangle_{\mathbb{G}}=\delta_{\mathbf{ab}}$, and constitutes a Hilbert basis of $\mathbb{G}$.

Polynomial chaos expansion of $G=\mathbb{g}(\boldsymbol{U})$. For any $\mathbb{g}$ in $\mathbb{G}$, the PCE of the second-order random variable G is written as $G=\sum_{\mathbf{a}\in\mathbb{N}^{n_u}} g^{\mathbf{a}}\Phi_{\mathbf{a}}(\mathbf{U})$, where the series of the PCE converges with respect to the norm of $\mathbb{G}$. The real coefficients $g^{\mathbf{a}}$ are such that $\|G\|^2_{\mathbb{G}}=\sum_{\mathbf{a}\in\mathbb{N}^{n_u}}(g^{\mathbf{a}})^2<+\infty$ and can be calculated by $g^{\mathbf{a}}=\langle G\,,\Phi_{\mathbf{a}}\rangle_{\mathbb{G}}=E\{G\,\Phi_{\mathbf{a}}(\mathbf{U})\}=E\{\mathbb{g}(\mathbf{U})\,\Phi_{\mathbf{a}}(\mathbf{U})\}=\int_{\mathbb{R}^{n_u}}\mathbb{g}(\mathbf{u})\,\Phi_{\mathbf{a}}(\mathbf{u})\,P_{\mathbf{U}}(d\mathbf{u})$.

Truncated PCE $G^{\text{chaos}}=\mathbb{g}^{\text{chaos}}(\boldsymbol{U})$ *for representing* $G=\mathbb{g}(\boldsymbol{U})$. Let n_g be the maximum degree of the considered truncated polynomial chaos expansion. We then have $|\mathbf{a}|=a_1+\ldots+a_{n_u}\leq n_g$. The set $\{\mathbf{a}\in\mathbb{N}^{n_u},|\mathbf{a}|\leq n_g\}$ of all the multi-indices in $\mathbb{N}^{n_u}$ such that $|\mathbf{a}|\leq n_g$ is rewritten as $\{\mathbf{a}^{(m)}=(a_1^{(m)},\ldots,a_{n_u}^{(m)})\in\mathbb{N}^{n_u}, m=1,\ldots,\mu\}$ with $\mathbf{a}^{(1)}=(0,\ldots,0)$ and where $\mu=(n_g+n_u)!/(n_g!\,n_u!)$. To simplify the notation, the polynomial chaos of multi-index $\mathbf{a}^{(m)}$ is rewritten as $\varphi_m(\mathbf{U})=\Phi_{\mathbf{a}^{(m)}}(\mathbf{U})$. As in Section 3, we have,

$$\varphi_1(\mathbf{U})=1\quad,\quad\langle\varphi_m\,,\varphi_{m'}\rangle_{\mathbb{G}}=\delta_{mm'}\quad,\quad E\{\varphi_m(\mathbf{U})\}=\delta_{1m}\,. \tag{4.1}$$

The truncated PCE of $G=\mathbb{g}(\mathbf{U})$ is thus given by

$$G^{\text{chaos}}=\mathbb{g}^{\text{chaos}}(\mathbf{U})=\sum_{m=1}^{\mu}g^m\,\varphi_m(\mathbf{U})\quad,\quad g^m=\langle G\,,\varphi_m\rangle_{\mathbb{G}}\,, \tag{4.2}$$

in which g^m is the rewriting of $g^{\mathbf{a}^{(m)}}$.

5. Polynomial chaos expansion of random vector Y

Hilbert space $\mathbb{F}$ *associated with the random variable* $(\boldsymbol{\Xi},\boldsymbol{U})$. Let $\mathbb{F}=\mathbb{H}\otimes\mathbb{G}$ be the Hilbert space defined according to the universal property of the tensor product of $\mathbb{H}$ and $\mathbb{G}$, which has to be understood as the completion $\mathbb{H}\widehat{\otimes}\mathbb{G}$ of space $\mathbb{H}\otimes\mathbb{G}$. Hilbert space $\mathbb{F}$ is equipped with the inner product

$$\langle\mathbb{f}\,,\widetilde{\mathbb{f}}\rangle_{\mathbb{F}}=\int_{\mathbb{R}^{n_w}}\int_{\mathbb{R}^{n_u}}\mathbb{f}(\boldsymbol{\xi},\mathbf{u})\,\widetilde{\mathbb{f}}(\boldsymbol{\xi},\mathbf{u})\,P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})\otimes P_{\mathbf{U}}(d\mathbf{u})\,, \tag{5.1}$$

and the associated norm $\|\,\mathbb{f}\,\|_{\mathbb{G}}=\langle\mathbb{f}\,,\mathbb{f}\rangle_{\mathbb{F}}^{1/2}$.

Multivariate polynomial chaos in $\mathbb{F}$. The family of functions $\{\Gamma_{\boldsymbol{\alpha}\mathbf{a}}=\Psi_{\boldsymbol{\alpha}}\otimes\Phi_{\mathbf{a}}\,,\,\boldsymbol{\alpha}\in\mathbb{N}^{n_w}\,,\,\mathbf{a}\in\mathbb{N}^{n_u}\}$, in which $\Psi_{\boldsymbol{\alpha}}$ and $\Phi_{\mathbf{a}}$ are defined in Sections 3 and 4, is a Hilbert basis of $\mathbb{F}$. We then have

$$\langle\Gamma_{\boldsymbol{\alpha}\mathbf{a}},\Gamma_{\boldsymbol{\beta}\mathbf{b}}\rangle_{\mathbb{F}}=\langle\Psi_{\boldsymbol{\alpha}},\Psi_{\boldsymbol{\beta}}\rangle_{\mathbb{H}}\times\langle\Phi_{\mathbf{a}},\Phi_{\mathbf{b}}\rangle_{\mathbb{G}}=\delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}\,\delta_{\mathbf{ab}}\,, \tag{5.2}$$

and $\Gamma_{\boldsymbol{\alpha}^{(1)}\mathbf{a}^{(1)}}(\boldsymbol{\xi},\mathbf{u})=\Psi_{\boldsymbol{\alpha}^{(1)}}(\boldsymbol{\xi})\times\Phi_{\mathbf{a}^{(1)}}(\mathbf{u})=1$.

Polynomial chaos expansion of $\boldsymbol{Y}=\mathbb{y}(\boldsymbol{\Xi},\boldsymbol{U})$. Let $\mathbb{F}_{n_q}=L^2_{P_{\boldsymbol{\Xi}}\otimes P_{\mathbf{U}}}(\mathbb{R}^{n_w}\times\mathbb{R}^{n_u},\mathbb{R}^{n_q})$ denote the Hilbert space of the square-integrable functions on $\mathbb{R}^{n_w}\times\mathbb{R}^{n_u}$ with values in $\mathbb{R}^{n_q}$, with respect to the probability measure $P_{\boldsymbol{\Xi}}\otimes P_{\mathbf{U}}$. Since $\mathbb{F}_{n_q}=L^2_{P_{\boldsymbol{\Xi}}\otimes P_{\mathbf{U}}}(\mathbb{R}^{n_w}\times\mathbb{R}^{n_u})\otimes\mathbb{R}^{n_q}$, it can be deduced that $\mathbb{F}_{n_q}=\mathbb{F}\otimes\mathbb{R}^{n_q}$. Considering the introduced hypotheses, the mapping $(\boldsymbol{\xi},\mathbf{u})\mapsto\mathbb{y}(\boldsymbol{\xi},\mathbf{u})$ defined by Eq. (2.3) belongs to $\mathbb{F}_{n_q}$. Hence, the second-order $\mathbb{R}^{n_q}$-valued random variable $\mathbf{Y}=\mathbb{y}(\boldsymbol{\Xi},\mathbf{U})$ admits the polynomial chaos expansion,

$$\mathbf{Y}=\sum_{\boldsymbol{\alpha}\in\mathbb{N}^{n_w}}\sum_{\mathbf{a}\in\mathbb{N}^{n_u}}\mathbf{z}^{\boldsymbol{\alpha}\mathbf{a}}\,\Gamma_{\boldsymbol{\alpha}\mathbf{a}}(\boldsymbol{\Xi},\mathbf{U})\,, \tag{5.3}$$

whose coefficients $\mathbf{z}^{\boldsymbol{\alpha}\mathbf{a}}$ in $\mathbb{R}^{n_q}$ are such that $\sum_{\boldsymbol{\alpha}\in\mathbb{N}^{n_w}}\sum_{\mathbf{a}\in\mathbb{N}^{n_u}}\|\mathbf{z}^{\boldsymbol{\alpha}\mathbf{a}}\|^2<+\infty$. The series of the PCE is convergent for the norm of $\mathbb{F}_{n_q}$.

Remark on the impossibility of using projection to compute the coefficients. It may seem that coefficients $\mathbf{z}^{\alpha\mathbf{a}}$ in $\mathbb{R}^{n_q}$ could be computed by the projection $\mathbf{z}^{\alpha\mathbf{a}} = E\{\mathbf{Y}\,\Gamma_{\alpha\mathbf{a}}(\boldsymbol{\Xi},\mathbf{U})\}$. However, it is not possible due to the fact that the random vector $\mathbf{U}$, introduced as a latent random variable, is independent of $\mathbf{Y}$ despite the statistical dependence between $\mathbf{Y}$ and $\boldsymbol{\Xi}$. Consequently, we have $E\{\mathbf{Y}\,\Psi_\alpha(\boldsymbol{\Xi})\,\Phi_\mathbf{a}(\mathbf{U})\} = E\{\mathbf{Y}\,\Psi_\alpha(\boldsymbol{\Xi})\}\,E\{\Phi_\mathbf{a}(\mathbf{U})\} = \mathbf{0}_{n_q}$ for all $\mathbf{a} \in \mathbb{N}^{n_u}$, except for $\mathbf{a} = \mathbf{a}^{(1)} = (0,\ldots,0)$.

6. Truncated polynomial chaos expansion of random vector Y

Using the index renumbering introduced in Eqs. (3.2) and (4.2), the truncated PCE of $\mathbf{Y} = \mathbb{y}(\boldsymbol{\Xi},\mathbf{U}))$ is written as,

$$\mathbf{Y}^{\text{chaos}} = \mathbb{y}^{\text{chaos}}(\boldsymbol{\Xi},\mathbf{U})) = \sum_{k=1}^{\kappa}\sum_{m=1}^{\mu} \mathbf{z}^{km}\,\gamma_{km}(\boldsymbol{\Xi},\mathbf{U})\,, \tag{6.1}$$

in which, for $k = 1,\ldots,\kappa$ and $m = 1,\ldots,\mu$, the vector-valued coefficient $\mathbf{z}^{km}$ is a rewriting of $\mathbf{z}^{\alpha^{(k)}\mathbf{a}^{(m)}}$ and where

$$\gamma_{km}(\boldsymbol{\Xi},\mathbf{U}) = \Gamma_{\alpha^{(k)}\mathbf{a}^{(m)}}(\boldsymbol{\Xi},\mathbf{U}) = \psi_k(\boldsymbol{\Xi})\,\varphi_m(\mathbf{U}) \quad , \quad \langle\gamma_{km},\gamma_{k'm'}\rangle_{\mathbb{F}} = \langle\psi_k\,,\psi_{k'}\rangle_{\mathbb{H}} \times \langle\varphi_m\,,\varphi_{m'}\rangle_{\mathbb{G}}\,\delta_{kk'}\,\delta_{mm'}\,. \tag{6.2}$$

Equation (6.2) with Eqs. (3.1) and (4.1) yields $\gamma_{11}(\boldsymbol{\Xi},\mathbf{U}) = 1$, $\gamma_{k1}(\boldsymbol{\Xi},\mathbf{U}) = \psi_k(\boldsymbol{\Xi})$, $\gamma_{1m}(\boldsymbol{\Xi},\mathbf{U}) = \varphi_m(\mathbf{U})$, and Eq. (6.2) yields $E\{\gamma_{km}(\boldsymbol{\Xi},\mathbf{U})\} = \delta_{1k}\,\delta_{1m}$. In addition, it can easily be seen that $E\{\|\mathbf{Y}^{\text{chaos}}\|^2\} = \sum_{k=1}^{\kappa}\sum_{m=1}^{\mu}\|\mathbf{z}^{km}\|^2 < +\infty$. From Eqs. (6.1) and (6.2), it can be deduced that

$$E\{\mathbf{Y}^{\text{chaos}}\} = \mathbf{z}^{11} \quad , \quad E\{\mathbf{Y}^{\text{chaos}}(\mathbf{Y}^{\text{chaos}})^T\} = \sum_{k=1}^{\kappa}\sum_{m=1}^{\mu}\mathbf{z}^{km}\,(\mathbf{z}^{km})^T\,. \tag{6.3}$$

As explained at the end of Section 5, the projection of $\mathbf{Y}$ yields $\mathbf{z}^{km} = E\{\mathbf{Y}\,\gamma_{km}(\boldsymbol{\Xi},\mathbf{U})\}$. However, only the vector-valued coefficients $\{\mathbf{z}^{k,1}, k = 1,\ldots,\kappa\}$ can be calculated through this projection. By rewriting $\mathbf{z}^{k,1}$ as $\underline{\mathbf{z}}^k$, these coefficients are given by

$$\underline{\mathbf{z}}^k = E\{\mathbf{Y}\,\psi_k(\boldsymbol{\Xi})\}\,. \tag{6.4}$$

We then obtain the PCE $\mathbf{Y}^{\text{chaos}}_{\text{proj}}$ of $\mathbf{Y}$ through a projection without considering the latent random variable $\mathbf{U}$. In other words,

$$\mathbf{Y}^{\text{chaos}}_{\text{proj}} = \sum_{k=1}^{\kappa}\underline{\mathbf{z}}^k\,\psi_k(\boldsymbol{\Xi})\,. \tag{6.5}$$

The error, $\|\mathbf{Y} - \mathbf{Y}^{\text{chaos}}_{\text{proj}}\|_{\mathbb{F}_{n_q}}$, between $\mathbf{Y} = \mathbb{y}(\boldsymbol{\Xi},\mathbf{U})$ (see Eq. (2.3)) and $\mathbf{Y}^{\text{chaos}}_{\text{proj}}$, is significant and can only be reduced by including the latent random vector $\mathbf{U}$.

Matrix representation of the realizations of the PCE $\boldsymbol{Y}^{\text{chaos}}$ of $\boldsymbol{Y}$. The N realizations of $\mathbf{Y}^{\text{chaos}}$, as defined by Eq. (6.1), are given by

$$\mathbf{y}^{\text{chaos},\ell} = \sum_{k=1}^{\kappa}\sum_{m=1}^{\mu}\mathbf{z}^{km}\,\gamma_{km}(\boldsymbol{\xi}^\ell,\mathbf{u}^\ell)\,. \tag{6.6}$$

Instead of using indices k and m, we introduce the global index j such that

$$j = (k,m) \in \{1,\ldots,J\} \quad \text{for} \quad (k,m) \in \{1,\ldots,\kappa\}\times\{1,\ldots,\mu\} \quad \text{with} \quad J = \kappa\times\mu\,.$$

In the following, it is assumed that $N \gg J$. By employing the global index j, Eq. (6.6) can be rewritten in the following matrix form,

$$[y^{\text{chaos}}] = [z]\,[\gamma] \in \mathbb{M}_{n_q,N} \quad , \quad [z] \in \mathbb{M}_{n_q,J} \quad , \quad [\gamma] \in \mathbb{M}_{J,N}\,, \tag{6.7}$$

in which the entries of matrices $[y^{\text{chaos}}]$, $[z]$, and $[\gamma]$ are

$$[y^{\text{chaos}}]_{i\ell} = y_i^{\text{chaos},\ell} \quad , \quad [z]_{ij} = z_i^{km} \quad , \quad [\gamma]_{j\ell} = \gamma_{km}(\boldsymbol{\xi}^\ell,\mathbf{u}^\ell)\,. \tag{6.8}$$

Using Eqs. (6.6) to (6.8), Eq. (6.2) can be rewritten as

$$\frac{1}{N-1}\,[\gamma]\,[\gamma]^T = [I_J]\,. \tag{6.9}$$

It should be noted that the factor $\frac{1}{N-1}$, which is used instead of $\frac{1}{N}$, originates from the statistical estimator employed to compute the realizations of polynomial chaos (refer to Page 122 of [124]).

Computation of matrix $[\gamma]$. Since $[\gamma]_{j\ell} = \gamma_{km}(\boldsymbol{\xi}^\ell, \mathbf{u}^\ell) = \psi_k(\boldsymbol{\xi}^\ell)\,\varphi_m(\mathbf{u}^\ell)$ (see Eq. (6.2)), by introducing the matrices $[\psi] \in \mathbb{M}_{\kappa,N}$ and $[\varphi] \in \mathbb{M}_{\mu,N}$ such that $[\psi]_{k\ell} = \psi_k(\boldsymbol{\xi}^\ell)$ and $[\varphi_{m\ell}] = \varphi_m(\mathbf{u}^\ell)$, the entries of matrix $[\gamma] \in \mathbb{M}_{J,N}$ can be written as $[\gamma]_{j\ell} = [\psi]_{k\ell}\,[\varphi_{m\ell}]$ with $j = (k, m)$. Matrix $[\psi]$, which is associated with the N realizations of the non-separated Hilbert basis constructed with $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$ on $\mathbb{R}^{n_w}$, and matrix $[\varphi]$, related to the N realizations of the separated normalized-Hermite-based Hilbert basis constructed with $P_{\mathbf{U}}(d\mathbf{u})$ on $\mathbb{R}^{n_u}$, are computed using the algorithm detailed on Page 122 of [124]) (see also [125, 67]).

Constraint on matrix $[z]$ *defined by the second-order moment-matrix* $[\mathcal{M}_\mathbf{Y}]$ *of* $\boldsymbol{Y}$. Similarly to the introduced notation $[y^{\text{chaos}}]$, we define $[y] \in \mathbb{M}_{n_q,N}$ such that $[y]_{i\ell} = y_i^\ell$. The estimate of $[\mathcal{M}_\mathbf{Y}] = E\{\mathbf{Y}\mathbf{Y}^T\}$ is then expressed using the same notation: $[\mathcal{M}_\mathbf{Y}] = [y]\,[y]^T/(N-1)$ (note that $[\mathcal{M}_\mathbf{Y}]$ is a given data derived from the learned dataset, as explained in Section 2). Hence, by imposing the equation $E\{\mathbf{Y}^{\text{chaos}}(\mathbf{Y}^{\text{chaos}})^T\} = [\mathcal{M}_\mathbf{Y}]$, we derive the following constraint,

$$[z]\,[z]^T = [\mathcal{M}_\mathbf{Y}] \quad , \quad [z] \in \mathbb{M}_{n_q,J}\,. \tag{6.10}$$

Relationship between the matrix $[y^{\text{chaos}}]$ *representing the realizations of* $\boldsymbol{Y}^{\text{chaos}}$ *and its counterpart* $[q^{\text{chaos}}]$ *representing the realizations of* $\boldsymbol{Q}^{\text{chaos}}$. Using the mapping $\mathbf{y} \mapsto \mathbf{q} = \mathbb{q}(\mathbf{y})$ defined by Eq. (2.2), we have for the random variables,

$$\mathbf{Q} = \mathbb{q}(\mathbf{Y}) \quad , \quad \mathbf{Q}^{\text{chaos}} = \mathbb{q}(\mathbf{Y}^{\text{chaos}})\,. \tag{6.11}$$

From the N realizations of $\mathbf{Y}^{\text{chaos}}$, represented by the matrix $[y^{\text{chaos}}] \in \mathbb{M}_{n_q,N}$, we can derive the N corresponding realizations $\{\mathbf{q}^{\text{chaos},\ell}, \ell = 1, \ldots, N\}$ as expressed in the matrix $[q^{\text{chaos}}] \in \mathbb{M}_{n_q,N}$. In this matrix, each entry $[q^{\text{chaos}}]_{i\ell}$ represents $q_i^{\text{chaos},\ell} = \mathbb{q}_i(\mathbf{y}^{\text{chaos},\ell})$. This relationship can be written in matrix form as:

$$[q^{\text{chaos}}] = [\mathbb{Q}([y^{\text{chaos}}])] \quad , \quad \mathbb{Q} : \mathbb{M}_{n_q,N} \to \mathbb{M}_{n_q,N}\,. \tag{6.12}$$

7. Optimization problem for estimating the coefficients of the truncated PCE

Optimal value $[z^{\text{opt}}]$ *of matrix* $[z]$. To estimate the matrix $[z] \in \mathbb{M}_{n_q,J}$ that contains the coefficients of the truncated PCE of $\mathbf{Y}$, as defined by Eq. (6.7), various approaches can be employed, in particular the maximum likelihood method (see the references given in Section 1 regarding the statistical identification of PCE coefficients). Among all the methods, we propose using the OVL_i (Overlap) indicator, which quantifies the overlap between the PDF of the components Q_i of $\mathbf{Q}$ and the PDF of the components Q_i^{chaos} of $\mathbf{Q}^{\text{chaos}}$. The OVL_i indicator is associated with the L^1-norm of functions $q_i \mapsto p_{Q_i}(q_i) - p_{Q_i^{\text{chaos}}}(q_i;[z])$ on $\mathbb{R}$, where p_{Q_i} and $p_{Q_i^{\text{chaos}}}(\cdot\,;[z])$ are the probability density functions of the real-valued random variables Q_i and Q_i^{chaos}, respectively. These PDFs are estimated using Gaussian KDE applied to the realizations $\{q_i^\ell, \ell = 1, \ldots, N\}$ for Q_i and to the realizations $\{q_i^{\text{chaos},\ell}, \ell = 1, \ldots, N\}$ for Q_i^{chaos}, which correspond to the columns of matrix $[q^{\text{chaos}}] = [\mathbb{Q}([z]\,[\gamma])]$ (see Eq. (6.12) with Eq. (6.7)). For $i \in \{1, \ldots, n_q\}$, $\text{OVL}_i([z])$ is written as

$$\text{OVL}_i([z]) = 1 - \frac{1}{2}\int_{\mathbb{R}} |p_{Q_i}(q_i) - p_{Q_i^{\text{chaos}}}(q_i;[z])|\,dq_i\,, \tag{7.1}$$

and the cost function is defined by

$$\mathcal{J}([z]) = \frac{1}{n_q}\sum_{i=1}^{n_q} \text{OVL}_i([z])\,. \tag{7.2}$$

It can be seen that $0 \le \mathcal{J}([z]) \le 1$ and the upper bound is reached when $p_{Q_i} = p_{Q_i^{\text{chaos}}}(\cdot\,;[z])$ for all i. Hence, the optimization problem is written as

$$[z^{\text{opt}}] = \arg\max_{[z]\in\mathscr{C}_{\text{ad}}} \mathcal{J}([z])\,, \tag{7.3}$$

in which the admissible set $\mathscr{C}_{\text{ad}} \subset \mathbb{M}_{n_q,J}$ allows the constraint defined by Eq. (6.10) to be taken into account,

$$\mathscr{C}_{\text{ad}} = \{\, [z] \in \mathbb{M}_{n_q,J} \;,\; [z]\,[z]^T = [\mathcal{M}_{\mathbf{Y}}] \,\}\,. \tag{7.4}$$

Transforming the constrained optimization problem into an unconstrained optimization problem. The optimization problem defined by Eq. (7.3) is nonconvex. We propose to solve it using an algorithm designed for unconstrained optimization problem. Due to the nonconvex nature of the problem, the estimated solution will strongly depend on the initial point $[z_0]$ chosen for the initialization of the optimization algorithm. We then need to transform the optimization problem on $\mathscr{C}_{\text{ad}}$ into an unconstrained optimization problem on $\mathbb{M}_{n_q,J}$, and also carefully choose the initial point. To do this, we need to introduce a $[\hat{z}] \mapsto [z]$ transformation to eliminate the constraint, which will then be automatically satisfied. Additionally, we have to introduce a second transformation $[\tilde{z}] \mapsto [\hat{z}]$ to search for an optimal solution in the vicinity of $[z_0] = [\underline{z}] \in \mathbb{M}_{n_q,J}$. The columns of $[\underline{z}]$, denoted $\underline{\mathbf{z}}^1, \ldots, \underline{\mathbf{z}}^J$, are defined as the projection on the polynomial chaos (see Section 6).

(a) Transformation $[\hat{z}] \mapsto [z]$ from $\mathbb{M}_{n_q,J}$ into $\mathbb{M}_{n_q,J}$. Let $[\hat{z}]$ be any unconstrained matrix given in $\mathbb{M}_{n_q,J}$. Let $[c]$ in $\mathbb{M}_{n_q}$ be the upper triangular matrix resulting from the Cholesky factorization of the matrix $[\hat{z}]\,[\hat{z}]^T \in \mathbb{M}^+_{n_q}$. We have $[c]^T\,[c] = [\hat{z}]\,[\hat{z}]^T$, which implies the existence of $[c]^{-1}$. Utilizing the decomposition $[\mathcal{M}_{\mathbf{Y}}] = [L_{\mathbf{Y}}]^2$ introduced in Section 2, the desired transformation is expressed as

$$[\hat{z}] \mapsto [z] = [L_{\mathbf{Y}}]\,[c]^{-T}\,[\hat{z}] \;:\; \mathbb{M}_{n_q,J} \to \mathbb{M}_{n_q,J}\,. \tag{7.5}$$

It can easily be verified that for any $[\hat{z}]$ in $\mathbb{M}_{n_q,J}$, we have $[z]\,[z]^T = [\mathcal{M}_{\mathbf{Y}}]$.

(b) Transformation $[\tilde{z}] \mapsto [\hat{z}]$ from $\mathbb{M}_{n_q,J}$ into $\mathbb{M}_{n_q,J}$. Matrix $[\underline{z}]$ being constructed as the projection of $[y]$ onto the subspace spanned by $[\gamma]$, using Eqs. (6.7) and (6.9) yields

$$[\underline{z}] = \frac{1}{N-1}\,[y]\,[\gamma]^T\,. \tag{7.6}$$

It is important to note that $[\underline{z}]\,[\underline{z}]^T = [y]\,[\chi]\,[y]^T$ where $[\chi] = [\gamma]^T\,[\gamma]/(N-1) \neq [I_N]$. Consequently, $[\underline{z}]\,[\underline{z}]^T \neq [\mathcal{M}_{\mathbf{Y}}]$. For $[\underline{z}] \in \mathbb{M}_{n_q,J}$ defined by Eq. (7.6), the transformation $[\tilde{z}] \mapsto [\hat{z}]$ from $\mathbb{M}_{n_q,J}$ into $\mathbb{M}_{n_q,J}$ is defined by

$$[\hat{z}]_{ij} = [\underline{z}]_{ij}\,(1 + [\tilde{z}]_{ij}) \quad , \quad i \in \{1, \ldots, n_q\} \quad , \quad j \in \{1, \ldots, J\}\,. \tag{7.7}$$

This transformation shows that as $[\tilde{z}]$ explores $\mathbb{M}_{n_q,J}$ in the vicinity of $[\tilde{z}_0] = [0]$, $[\hat{z}]$ also explores $\mathbb{M}_{n_q,J}$ in the vicinity of $[\underline{z}]$, and $[z] = [L_{\mathbf{Y}}]\,[c]^{-T}\,[\hat{z}]$ satisfies the constraint $[z]\,[z]^T = [\mathcal{M}_{\mathbf{Y}}]$.

(c) Transformation $[\tilde{z}] \mapsto [z] = [\mathbb{z}([\tilde{z}])]$ from $\mathbb{M}_{n_q,J}$ into $\mathbb{M}_{n_q,J}$. The composition of transformation $[\tilde{z}] \mapsto [\hat{z}]$ defined by Eq. (7.5) and transformation $[\hat{z}] \mapsto [z]$ defined by Eq. (7.7) is a well-defined transformation $[\tilde{z}] \mapsto [z] = [\mathbb{z}([\tilde{z}])]$.

(d) Reformulation of the optimization problem and algorithm. The optimization problem defined by Eq. (7.3) can be rewritten as,

$$[z^{\text{opt}}] = [\mathbb{z}([\tilde{z}^{\text{opt}}])] \quad , \quad [\tilde{z}^{\text{opt}}] = \arg \max_{[\tilde{z}] \in \mathbb{M}_{n_q,J}} \widetilde{\mathcal{J}}([\tilde{z}]) \quad , \quad \widetilde{\mathcal{J}}([\tilde{z}]) = \mathcal{J}([\mathbb{z}([\tilde{z}])])\,. \tag{7.8}$$

This nonconvex optimization problem can be solved using various algorithms. As we are searching for a solution $[\tilde{z}^{\text{opt}}]$ in the vicinity of $[0_{n_q,J}]$, an appropriate choice for the algorithm may be the unconstrained quasi-Newton algorithm, initialized with $[\tilde{z}_0] = [0_{n_q,J}]$, where the gradient is not explicitly provided.

Optimal truncated PCE of $\boldsymbol{Y}$. The Gaussian probability measure $P_{\mathbf{U}}(d\mathbf{u})$ of the latent random variable $\mathbf{U}$ depends on its dimension n_u and the polynomial chaos expansion in $\mathbf{U}$ depends on the maximum degree n_g (see Section 4). Consequently, the PCE $\mathbf{Y}^{\text{chaos}} = \mathbb{y}^{\text{chaos}}(\boldsymbol{\Xi}, \mathbf{U})$ of $\mathbf{Y} = \mathbb{y}(\boldsymbol{\Xi}, \mathbf{U})$, which depends on N_g, also depends on n_u and n_g. The presented methodology allows for estimating an optimal value $[z^{\text{opt}}(N_g, n_u, n_g)]$ that depends on N_g, n_u, and n_g. For a given value of each of these three integers, the error between $\mathbf{Y}$ and its chaos representation $\mathbf{Y}^{\text{chaos}}$ can be quantified by evaluating

$$\mathcal{J}^{\text{opt}}(N_g, n_u, n_g) = \mathcal{J}([z^{\text{opt}}(N_g, n_u, n_g)])\,, \tag{7.9}$$

in which $\mathcal{J}([z])$ is defined by Eq. (7.2). Consequently, the optimal truncated PCE of $\mathbf{Y}$ is obtained by using the optimal values N_g^{opt}, n_u^{opt}, and n_g^{opt} of N_g, n_u, and n_g, respectively, such that

$$(N_g^{\text{opt}}, n_u^{\text{opt}}, n_g^{\text{opt}}) = \arg \max_{N_g \geq 2\,,\, n_u \geq 1\,,\, n_g \geq 1} \mathcal{J}^{\text{opt}}(N_g, n_u, n_g)\,, \tag{7.10}$$

Comments about the algorithm for solving the optimal values N_g^{opt}, n_u^{opt}, and n_g^{opt}. These optimal values are obtained by solving the optimization problem defined by Eq. (7.10). A simple and direct approach, although it can be computationally expensive, is to compute the value of $\mathcal{J}^{\text{opt}}(N_g, n_u, n_g)$ at each point in a three-dimensional grid. This grid corresponds to a discretization of the domain $[2, N_g^{\max}] \times [1, n_u^{\max}] \times [1, n_g^{\max}] \subset \mathbb{N}^3$, in which $N_g^{\max}$, $n_u^{\max}$, and $n_g^{\max}$ are set to sufficiently large values. Another approach, which is less computationally expensive, is to use the assumptions introduced in section 1, which are related to the underlying physical problem that generated the data. In this context, if it were not necessary to introduce the latent variable $\mathbf{U}$, then the solution would be $[z^{\text{opt}}(N_g)] = [\underline{z}(N_g)]$ given by Eq. (7.6). The corresponding value of the overlap can be expressed as follows,

$$\underline{\mathcal{J}}(N_g) = \mathcal{J}([\underline{z}(N_g)])\,. \tag{7.11}$$

Since the latent variable $\mathbf{U}$ is essential, at convergence with respect to N_g, the overlap $\underline{\mathcal{J}}(N_g)$ will be less than 1. The difference $1 - \underline{\mathcal{J}}(N_g)$ makes it possible to quantify the error induced by the projection method. We can then quickly estimate an optimal value N_g^{opt} for N_g using a one-dimensional grid over $[2, N_g^{\max}] \subset \mathbb{N}$. In the presence of the latent variable $\mathbf{U}$, we set N_g to this optimal value N_g^{opt} and then we search for the optimal values n_u^{opt} and n_g^{opt} on a two-dimensional grid over $[1, n_u^{\max}] \times [1, n_g^{\max}] \subset \mathbb{N}^2$.

8. Polynomial-chaos-based statistical surrogate model

Presenting the problem to be solved. The problem at hand involves computing the realization $\mathbf{q}_0^{\text{chaos}}$ of $\mathbf{Q}^{\text{chaos}}$ given $\boldsymbol{\Xi} = \boldsymbol{\xi}_0$, where $\boldsymbol{\xi}_0$ represents a realization of $\boldsymbol{\Xi}$ following the probability measure $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$, and $\mathbf{u}_0$ represents a realization of $\mathbf{U}$ following the probability measure is $P_{\mathbf{U}}(d\mathbf{u})$. In fact, a realization $\mathbf{w}_0$ of $\mathbf{W}$ is given, then the realization $\boldsymbol{\xi}_0$ is calculated by $\boldsymbol{\xi}_0 = \mathbf{s}_w^{-1}(\mathbf{w}_0)$ (see Eq. (2.1)).

Polynomial chaos-based algorithm for computing the conditional realization $\boldsymbol{q}_0^{\text{chaos}}$ of $\boldsymbol{Q}^{\text{chaos}}$ given $\boldsymbol{\Xi} = \boldsymbol{\xi}_0$. Using the optimal truncated PCE $\mathbf{Q}^{\text{chaos}}$ of $\mathbf{Q}$, we compute the corresponding realization $\mathbf{q}_0^{\text{chaos}}$ of $\mathbf{Q}^{\text{chaos}}$ according to Eqs. (6.11), (6.7), and (6.8), by

$$\mathbf{q}_0^{\text{chaos}} = \mathbb{q}(\mathbf{y}_0^{\text{chaos}}) \quad , \quad \mathbf{y}_0^{\text{chaos}} = [z^{\text{opt}}(N_g^{\text{opt}}, n_u^{\text{opt}}, n_g^{\text{opt}})]\, \gamma_0^{\text{opt}} \quad , \quad \gamma_0^{\text{opt}} \in \mathbb{R}^J\,. \tag{8.1}$$

Validation. The verification of Eq. (8.1) is performed in a probability framework as follows. We generate $N_v \sim N$ additional realizations $\boldsymbol{\xi}_0$ of $\boldsymbol{\Xi}$ and $\mathbf{u}_0$ of $\mathbf{U}$, which are distinct from the realizations $\{(\boldsymbol{\xi}^\ell, \mathbf{u}^\ell), \ell = 1, \ldots, N\}$. For each realization $(\boldsymbol{\xi}_0, \mathbf{u}_0)$, we compute the realization $\mathbf{q}_0^{\text{chaos}}$ of $\mathbf{Q}^{\text{chaos}}$ using Eq. (8.1). We quantify the error between the reference $\mathbf{Q}$ (defined by the N realizations from the learned dataset) and $\mathbf{Q}^{\text{chaos}}$ (defined by the N_v realizations generated using Eq. (8.1)), by computing $\mathcal{J}^{\text{opt}}$ using Eq. (7.9) for the optimal values N_g^{opt}, n_u^{opt}, and n_g^{opt}. Additionally, we compare the probability density functions of Q_i and Q_i^{chaos} for $i = 1, \ldots, n_q$.

9. Cluster separation of the learned dataset constituted of heterogeneous data

As we explained in Section 1, in the case of heterogeneous data, it can be interesting to use an adapted approach based on the formulation presented in Sections 3 to 8. This approach involves performing a prior cluster separation of the learned dataset into distinct clusters, each consisting of "quasi-homogeneous" data. It is important to note that the objective remains the same as before, which is to construct a global polynomial chaos representation $\mathbf{Q}^{\text{chaos}} = \mathbf{f}^{\text{chaos}}(\mathbf{W}, \mathbf{U})$ of $\mathbf{Q} = \mathbf{f}(\mathbf{W}, \mathbf{U})$ based on the polynomial chaos representation constructed for each distinct cluster. This is particularly useful when the training dataset is generated by physical processes that exhibit multiple regimes simultaneously. Although the numerical offline computation may increase, the online prediction with the PCE $\mathbf{Q}^{\text{chaos}}$ given $\boldsymbol{\Xi} = \boldsymbol{\xi}_0$ remains unaffected. It should be noted that the cluster separation is performed on the learned

dataset rather than the training dataset, which is too small. This choice offers an advantage because the PLoM algorithm used to generate the learned dataset from the training dataset preserves the concentration of the probability measure in the vicinity of the manifolds while enhancing the available information contained in the points of the training dataset. The use of prior separation into distinct clusters, allows us to better understand how the learned dataset is structured and facilitates the construction of the representation in polynomial chaos for each cluster.

Hypothesis used for building the cluster separation of the heterogeneous learned dataset. It is assumed that there are multiple physical regimes, which may involve a cluster separation of the learned dataset $\mathcal{D}_{\text{learn}}(\mathbf{X})$ into K distinct clusters based on their proximity or similarity.

Comments on existing algorithms for constructing cluster separation into distinct clusters of points. There are various algorithms available for constructing a cluster separation of a set of points. The principal methods include K-means [126, 127], Hierachical Agglomerative Clusturing [128], Divisive Clusturing [129], Density-Based Spectral Clusturing with its variants [130, 131, 132], Mean-Shift Clusturing [133], Spectral Clusturing [134], Gaussian Mixture Model [135], and more. In the context of the previously introduced hypothesis, our objective is to determine the number K of clusters while considering the existence of K physical regimes. Therefore, algorithms based on Density-Based Spectral Clustering are not well suited for this purpose. For large datasets, algorithms such as Agglomerative Clustering, Mean-Shift Clustering, or Spectral Clustering (which is well adapted to image segmentation) require significant computational resources in terms of RAM and CPU. The Gaussian mixture model could be a candidate approach. However, this method is sensitive to dimensionality and outliers, and gives for each identified cluster a Gaussian measure, which may not be suitable for the non-Gaussian case. For the application presented in this paper, three clustering algorithms have been tested: Divisive Clustering, Hierarchical Agglomerative Clustering, and K-means with the squared Euclidean distance. In K-means, each centroid is calculated as the mean of the points in its respective cluster.

Notation and formulation related to the cluster separation. The chosen algorithm is applied to perform the partition into K clusters of $\mathcal{D}_{\text{learn}}(\mathbf{Q}) = \{\mathbf{q}^\ell, \ell = 1, \ldots, N\}$. For $\sigma = 1, \ldots, K$, let $\mathcal{I}_\sigma = \{\ell_1^{(\sigma)}, \ldots, \ell_{N_\sigma}^{(\sigma)}\} \subset \{1, \ldots, N\}$ be the subset of indices resulting from this cluster separation. We then have $N = \sum_{\sigma=1}^K N_\sigma$, and

$$\cup_{\sigma=1}^K \mathcal{I}_\sigma = \{1, \ldots, N\} \quad , \quad \cap_{\sigma=1}^K \mathcal{I}_\sigma = \{\emptyset\} \, . \tag{9.1}$$

Consequently, the resulting partition of $\mathcal{D}_{\text{learn}}(\mathbf{X})$ can be expressed as,

$$\mathcal{D}_{\text{learn}}(\mathbf{X}) = \cup_{\sigma=1}^K \mathcal{D}_\sigma(\mathbf{X}^{(\sigma)}) \quad , \quad \cap_{\sigma=1}^K \mathcal{D}_\sigma(\mathbf{X}^{(\sigma)}) = \{\emptyset\} \quad , \quad \mathcal{D}_\sigma(\mathbf{X}^{(\sigma)}) = \{\mathbf{x}^\ell = (\mathbf{q}^\ell, \mathbf{w}^\ell) \, , \ell \in \mathcal{I}_\sigma\} \, , \tag{9.2}$$

where $\mathbf{X}^{(\sigma)} = (\mathbf{Q}^{(\sigma)}, \mathbf{W}^{(\sigma)})$ is the $\mathbb{R}^{n_x}$-valued random variable for which $\{\mathbf{x}^\ell, \ell \in \mathcal{I}_\sigma\}$ are N_σ independent realizations. Furthermore, we introduce a discrete-valued random variable B_d with values in $\{1, 2, \ldots, K\}$, which is statistically dependent on $\mathbf{W}$. The realizations $\{b_d^\ell, \ell = 1, \ldots, N\}$ of B_d, corresponding to $\{\mathbf{w}^\ell, \ell = 1, \ldots, N\}$, are defined as follows,

$$\text{if} \quad \ell \in \mathcal{I}_\sigma \quad \text{then} \quad b_d^\ell = 2\sigma - (K+1) \quad , \quad \sigma \in \{1, \ldots, K\} \, . \tag{9.3}$$

For instance, for $K = 2$, if $\ell \in \mathcal{I}_1$, then $b_d^\ell = -1$ while if $\ell \in \mathcal{I}_2$, then $b_d^\ell = +1$. It is important to note that it is impossible to obtain a "perfect" separation, regardless of the algorithm used within the framework of probability theory, because of the potential partial overlap of the supports of the probability measures for the random variables $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$.

Smoothing the discrete random variable B_d into a real-valued random variable B. In order to use the same algorithm as the one presented in Sections 3 to 8 for constructing the truncated PCE $B^{\text{chaos}} = \mathbb{b}^{\text{chaos}}(\mathbf{W}, \mathbf{U})$ of $B_d = \mathbb{b}_d(\mathbf{W}, \mathbf{U})$, we introduce a real-valued random variable B that corresponds to smoothing B_d on $\mathbb{R}$. The probability density function of $B = \mathbb{b}(\mathbf{W}, \mathbf{U})$ is constructed using the Gaussian KDE with the realizations $\{b_d^\ell, \ell = 1, \ldots, N\}$ defined by Eq. (9.3). It can be expressed as $p_B(b\,;\beta) = \frac{1}{N} \sum_{\ell=1}^N \frac{1}{\sqrt{2\pi}\beta} \exp\{-\frac{1}{2\beta^2}(b - b_d^\ell)^2\}$ where β represents the bandwidth. As β approaches zero, $p_B(b, ;\beta), db$ converges to the discrete probability measure $P_{B_d}(db) = \frac{1}{N} \sum_{\ell=1}^N \delta_0(b - b_d^\ell)$. The smoothing of B_d with B involves reducing the Silverman bandwidth, which is accomplished by selecting $\beta = \frac{1}{2}(\frac{4}{3N})^{1/5} \sigma_{B_d}$, where σ_{B_d}

is the standard deviation of B_d estimated using the realizations $\{b_d^\ell, \ell = 1, \ldots, N\}$.

Identifying the cluster number σ based on a given realization of the real-valued random variable B. The presented formulation for cluster separation enables the construction of the truncated PCE, $B^{\text{chaos}} = \mathbb{b}^{\text{chaos}}(\mathbf{W}, \mathbf{U})$, of $B = \mathbb{b}(\mathbf{W}, \mathbf{U})$. For a given realization $\mathbf{w}_0$ of $\mathbf{W}$ and $\mathbf{u}_0$ of $\mathbf{U}$, we need to identify the σ-index in $\{1, \ldots, K\}$ of the associated cluster, as explained earlier. Considering the generation of realizations $\{b_d^\ell, \ell = 1, \ldots, N\}$ defined by Eq. (9.3), along with the construction of the smoothed random variable B, which will be represented by B^{chaos}, we propose the following method for identifying the σ-index. Let $\mathcal{B}_\sigma$ be the subset of $\mathbb{R}$ defined by $\mathcal{B}_1 =]-\infty, 2-K]$, $\mathcal{B}_K =]K-2, +\infty[$, and if $K \geq 3$, $\mathcal{B}_\sigma =]2\sigma - K - 2, 2\sigma - K]$ for $2 \leq \sigma < K$. Let $b_0 = \mathbb{b}(\mathbf{w}_0, \mathbf{u}_0)$ be the corresponding realization of B, that will be approximated by $b_0^{\text{chaos}} = \mathbb{b}^{\text{chaos}}(\mathbf{w}_0, \mathbf{u}_0)$. We identify the value of the σ-index in $\{1, \ldots, K\}$ as the index σ of interval $\mathcal{B}_\sigma$ to which b_0 belongs.

Truncated PCE of B and of $\{\boldsymbol{X}^{(\sigma)}, \sigma = 1, \ldots, K\}$, computation of the PCE-based statistical surrogate model, realizations, and validation. The truncated PCE of B and $\mathbf{X}^{(\sigma)}$ for all $\sigma = 1, \ldots, K$ are constructed using the methodology outlined in Sections 3 to 8. Following the approach described in Section 8, we consider any realization $\boldsymbol{\xi}_0$ of $\boldsymbol{\Xi}$ from $P_{\boldsymbol{\Xi}}(d\boldsymbol{\xi})$ and any realization $\mathbf{u}_0$ of $\mathbf{U}$ from $P_{\mathbf{U}}(d\mathbf{u})$. By utilizing the surrogate model based on cluster separation, we compute the conditional realization $\mathbf{q}_0$ of $\mathbf{Q}^{\text{chaos}}$ given $\boldsymbol{\Xi} = \boldsymbol{\xi}_0$. The validation process follows the methodology presented in Section 8.

10. Application to atomic collisions of Helium on graphite substrate

10.1. Description of the physical system and its Molecular Dynamics simulations for generating the training dataset

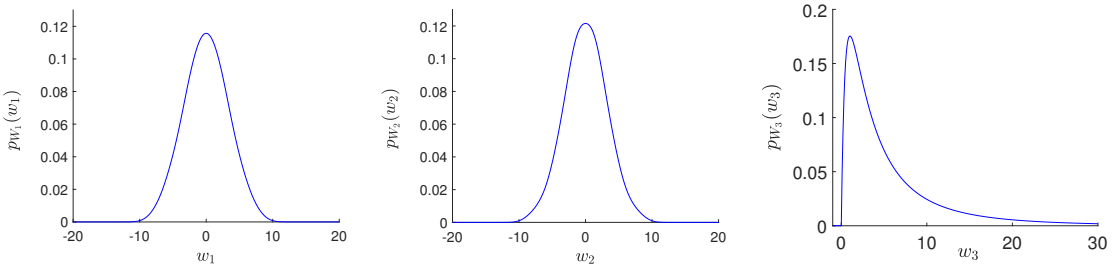
The details of the dynamical system and its analysis can be found in [136]. Here we provide a brief overview. The wall model consists of three layers of graphene with dimensions 17.04×17.22 Å^2 in the xOy plane and is periodically replicated in the x and y directions. The z coordinates are orthogonal to the xOy plane. The wall contains 336 carbon atoms and has a width of $6.8, AA$. All molecular dynamics simulations involving graphite, which is composed of carbon atoms (C), and helium (He) gas atoms, were performed using the LAMMPS (Large-scale Atomic Molecular Massively Parallel Simulator) package. The adaptive intermolecular reactive bond order potential was used to model the interactions between the carbon atoms in graphite. The 12-6 Lennard Jones potential was employed to describe the interactions between carbon and helium atoms. A cutoff distance r_{cut} of 12 Å has been chosen for these interactions. A graphitic wall served as the lower boundary, and a reflective plane was located at a distance of 18.8 Å from the surface, serving as the upper boundary. A control plane was placed at a distance of $z = r_{\text{cut}}$ from the carbon surface, where information about the crossing atoms was recorded, including their incoming/outgoing times, velocities, and displacements during their residence time. During the simulation, the bottom sheet was fixed, while the other two layers were kept at a constant temperature of $50°K$ using the Nosé-Hoover thermostat in the NVT ensemble, with a relaxation temperature parameter set to 40 time steps. The layers were free to interact with the helium atoms. Approximately 1×10^6 time steps of $1\,fs$ were spent equilibrating the system at the desired temperature before proceeding with subsequent statistical analysis. The velocity-Verlet algorithm was used for time integration.

In the following, the three components of the control parameter $\mathbf{W}$ correspond to the components $(V_{x,\text{in}}, V_{y,\text{in}}, V_{z,\text{in}})$ of the velocity $\mathbf{V}_{\text{in}}$ of the incident particle on the layer. The random output vector $\mathbf{Q} = (\mathbf{V}_{\text{out}}, log(\Delta_t), D_x, D_y)$ consists of the random output velocity vector $\mathbf{V}_{\text{out}} = (V_{x,\text{out}}, V_{y,\text{out}}, V_{z,\text{out}})$ (reflected velocity) of the particle, the random logarithm duration of absorption $log(\Delta_t)$ (residence time), and the two displacement components D_x and D_y in the xOy plane within the layer. In this application, all distances are measured in Angstroms (Å), and time is measured in picoseconds (ps).

10.2. Polynomial-chaos-based surrogate model

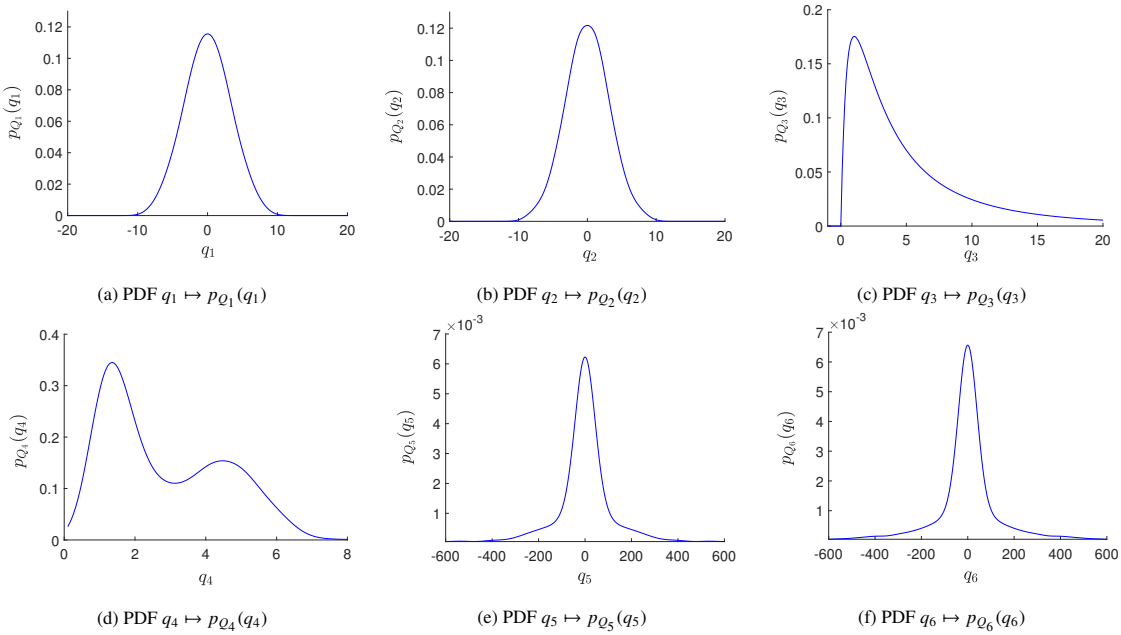
Generation of the large learned dataset and probability density functions of $\boldsymbol{W}$ and $\boldsymbol{Q}$. For the application under consideration, all the presented results correspond to a temperature of $50°K$. The large learned dataset is generated as explained in Section 2, with $N = 200\,000$ learned realizations on the basis of the training dataset consisting of $n_d = 2\,000$ points. The training dataset was generated as explained in Section 10.1. The probability density functions of the 3 components of $\mathbf{W}$, estimated using the Gaussian KDE with the N learned realizations, are shown in Figures 1a

to 1c. The corresponding probability density functions of the 6 components of $\mathbf{Q}$ are shown in Figures 2a to 2f. These figures define the reference and will be used for comparing with the PDFs given by the PCE. Figure 2d, which dis-



(a) PDF $w_1 \mapsto p_{W_1}(w_1)$ (b) PDF $w_2 \mapsto p_{W_2}(w_2)$ (c) PDF $w_3 \mapsto p_{W_3}(w_3)$

Figure 1: For $N = 200\,000$ learned realizations, PDF of the components W_1, W_2, and W_3 of the random control parameter $\mathbf{W}$, which represents the incident velocity vector $\mathbf{V}_{\text{in}} = (V_{x,\text{in}}, V_{y,\text{in}}, V_{z,\text{in}})$.



(a) PDF $q_1 \mapsto p_{Q_1}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3}(q_3)$

(d) PDF $q_4 \mapsto p_{Q_4}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6}(q_6)$

Figure 2: For $N = 200\,000$ learned realizations, PDF of the components Q_1 to Q_6 of the random output vector $\mathbf{Q}$, which represents $(V_{x,\text{out}}, V_{y,\text{out}}, V_{z,\text{out}}, log(\Delta_t), D_x, D_y)$.

plays the PDF of the logarithm of the residence time (duration of absorption) $Q_4 = \log(\Delta_t)$, reveals the existence of two distinct physical regimes. The first regime corresponds to a short residence time, with a PDF peak at approximatively $\exp(1.38) \simeq 4\,ps$. Conversely, the peak of the long residence-time regime is around $\exp(4.5) \simeq 90\,ps$. It should be noted that these two residence-time regimes are not entirely separated, as the PDF exhibits a local minimum at approximately $\exp(3) \simeq 20\,ps$, where the PDF has a significant value of 0.11. The presence of these two physical regimes is also evident in the PDF of $Q_5 = D_x$ (Figure 2e) and $Q_6 = D_y$ (Figure 2f). These PDFs exhibit strong non-Gaussian behavior and display two distinct patterns: one for the distance in the range of $[-120, 120]$, Å, and the other for distances in the intervals $]-\infty, -120[\cup]120, +\infty[$ Å. For a more comprehensive analysis of the underlying physics, we refer the reader to [136].

PCE constructed through projection and quantification of the induced error. As explained in Section 7, the projection-

based truncated PCE of $\mathbf{Y}$ is constructed in estimating the coefficients-matrix $[z]$ in Eq. (6.7) by using the projection method yielding $[z] = [\underline{z}]$ defined by Eq. (7.6). Figure 3 displays the graph of function $N_g \mapsto \underline{\mathcal{J}}(N_g)$ defined by Eq. (7.11). Since, with a such projection, the random latent variable $\mathbf{U}$ is not taken into account (corresponding to $(n_u, n_g) = (0, 0)$), the projection method induces a significant error of $1 - 0.7425$ for $N_g = 12$, which is approximately a 25% error. We intentionally limited the degree N_g to 12, yielding $\kappa = 455$ polynomial chaos. In fact, increasing the degree leads to a higher computational cost with little improvement in convergence. Introducing the latent variable $\mathbf{U}$ is necessary to reduce the error. Figures 4a to 4f show the PDFs of the 6 components of $\mathbf{Q}$. These PDFs correspond to

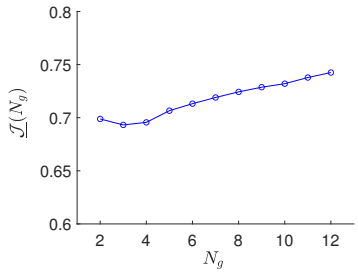
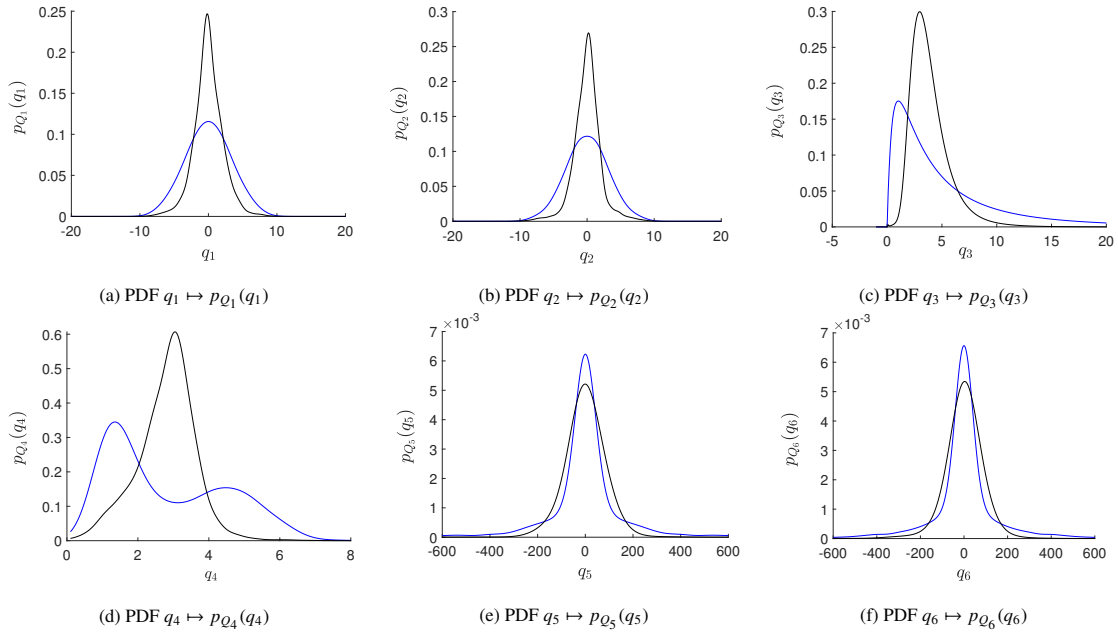


Figure 3: Graph of the function $N_g \mapsto \underline{\mathcal{J}}(N_g)$ allowing the error induced by the projection method to be quantified.

the reference values (estimated using the learned dataset) and the values estimated using the truncated PCE with the projection method and $N_g = 12$. All computation are performed using $N = 200\,000$ learned realizations. As expected, there is a significant error between the reference values and the PCE constructed through projection. This comparison is interesting because it demonstrates that the presence of the two physical regimes can only be reproduced with the PCE by introducing the latent variable $\mathbf{U}$. The control variable $\mathbf{W}$, which represents the incident velocity vector alone, is insufficient to explain the phenomena.



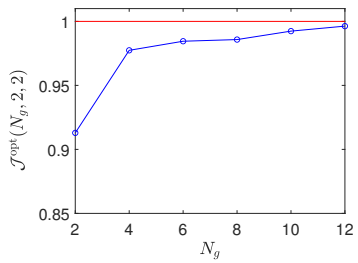
(a) PDF $q_1 \mapsto p_{Q_1}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3}(q_3)$

(d) PDF $q_4 \mapsto p_{Q_4}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6}(q_6)$

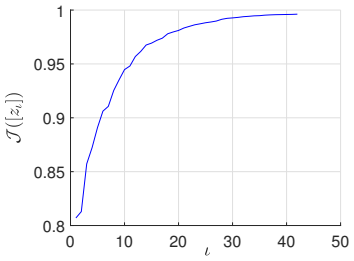
Figure 4: For $N = 200\,000$ learned realizations, PDF of the components Q_1 to Q_6 of the random output vector $\mathbf{Q}$, corresponding to the reference (blue thick line) and estimated with the truncated PCE by the projection method with $N_g = 12$ (black thin line).

Optimization problem for estimating the coefficients of the truncated PCE and convergence analysis with respect to n_u

and n_g. For a total of 200,000 learned realizations, the convergence with respect to N_g, n_u, and n_g, has been analyzed by studying the function $(N_g, n_u, n_g) \mapsto \mathcal{J}^{\text{opt}}(N_g, n_u, n_g)$ defined by Eq. (7.9). As previously explained, $N_d = 12$ is a suitable value for N_g^{opt}, which will be confirmed below. We have studied the function $(n_u, n_g) \mapsto \mathcal{J}^{\text{opt}}(12, n_u, n_g)$ for $N_g = 12$. The calculation gives $\mathcal{J}^{\text{opt}}(12, 0, 0) = 0.7425$, and for the points $(1, 1)$, $(1, 2)$, and $(2, 2)$, we have got 0.996. This indicates that convergence is reached when $n_u = 1$ and $n_g = 1$. However, looking at the PDFs of the components of $\mathbf{Q}^{\text{chaos}}$, we see a slightly better match to the PDF of $\mathbf{Q}$ when $n_u = 2$ and $n_g = 2$, corresponding to $\mu = 6$ polynomial chaos. Figure 5a displays the graph of the function $N_g \mapsto \mathcal{J}^{\text{opt}}(N_g, 2, 2)$. This figure shows that $N_g^{\text{opt}} = 12$ is an optimal choice, and gives excellent convergence. As previously mentioned, the optimization problem defined by Eq. (7.8) is solved using the quasi-Newton algorithm. For $N_g = 12$, $n_u = 2$, and $n_g = 2$, yielding $J = \kappa \times \mu = 455 \times 6 = 2\,730$ coefficients with values in $\mathbb{R}^6$ for the truncated PCE, Figure 5b displays the graph of the function $\iota \mapsto \mathcal{J}([z_\iota])$, where ι represents the iteration number in the quasi-Newton algorithm. This graph illustrates the rapid convergence rate.



(a) Graph of function $N_g \mapsto \mathcal{J}^{\text{opt}}(N_g, 2, 2)$.



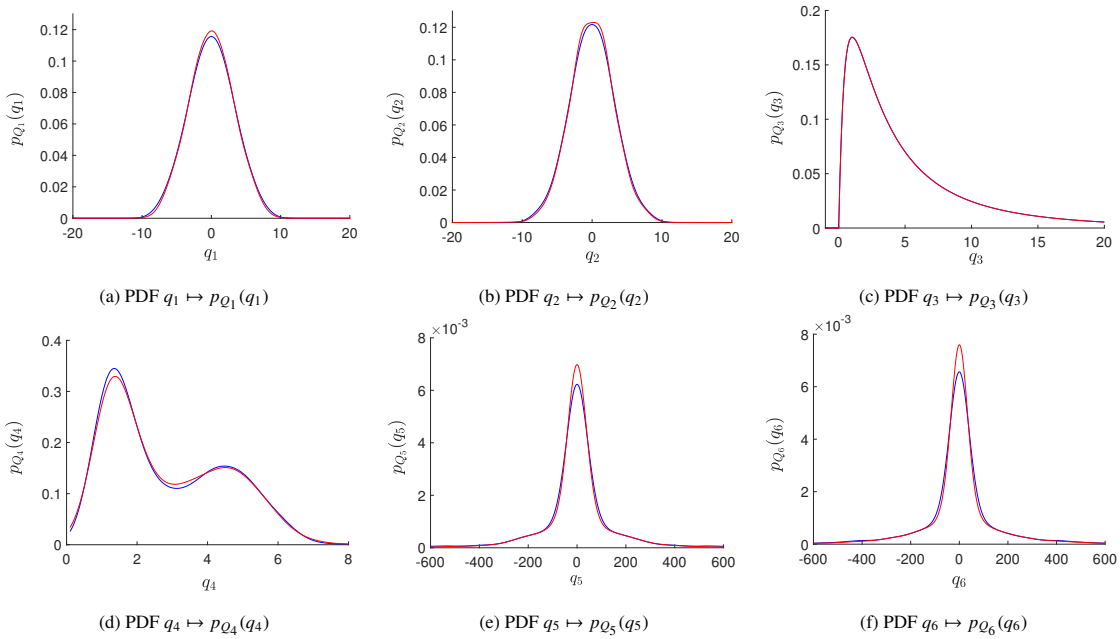
(b) Graph of function $\iota \mapsto \mathcal{J}([z_\iota])$.

Figure 5: Convergence analysis of the truncated PCE with respect to N_g for $n_u = n_g = 2$ (a) and convergence of the quasi-Newton algorithm as a function of the iteration number ι (b).

Validation of the polynomial-chaos based conditional statistics. Once the optimal coefficients of the truncated PCE have been estimated as described above, the validation process is carried out following the explanation in Section 8.To do so, $N_v = 200\,000$ new realizations of random control variable $\mathbf{W}$ are generated according to the probability measure $P_{\mathbf{W}}(d\mathbf{w})$. Then, N_v corresponding realizations of $\mathbf{Q}^{\text{chaos}}$ are generated using Eq. (8.1). Figures 6a to 6f display the PDFs of the 6 components of $\mathbf{Q}$. These PDFs correspond to the reference values (estimated with the learned dataset) and the values estimated using the optimal truncated PCE with $N_g = 12$, $n_u = 2$, and $n_g = 2$. The predictions obtained with this statistical substitution model based on the truncated PCE are very good.

10.3. Polynomial-chaos-based surrogate model using a cluster separation

As we have explained in Section 10.2 (also see Figure 2d, the heterogeneous data is generated by two distinct physical regimes that cannot be perfectly separated. To identify a good cluster-separation algorithm adapted to the considered heterogeneous data, we tested three algorithms: K-means, Divisive Clusturing, and Hierachical Agglomerative Clusturing, with $K = 2$ and $K = 3$ clusters. The best separation was achieved with K-means for $K = 2$ clusters. All computations were performed using the $N = 200\,000$ points of $\mathcal{D}_{\text{learn}}(\mathbf{Q})$. Using the notation introduced in Eqs. (9.1) to (9.3), we obtained a first cluster $\sigma = 1$ with $\mathcal{I}_1$ consisting of $N_1 = 21\,059$ points, and the second cluster $\sigma = 2$ with $\mathcal{I}_2$ comprising $N_2 = 178\,941$ points. Cluster $\mathcal{D}_1(\mathbf{X}^{(1)})$ corresponds to a physical regime with a long residence time, while cluster $\mathcal{D}_2(\mathbf{X}^{(2)})$ corresponds to a mixture of the two physical regimes, primarily containing short residence time but also medium and long residence time. Certainly, a better separation could be obtained "manually", based on physics expertise without using a clustering algorithm and based on the analysis of the single component Q_4 (see [136] in the context of this application). However, we wish to present here a cluster separation using a separation algorithm which acts "simultaneously" on all the components of the random vector $\mathbf{Q}$ (because these components are statistically dependent). The goal is to illustrate the algorithm we propose to construct realizations using a statistical surrogate model constructed from the PCE of each cluster and the PCE of the random variable B. In all the figures of Section 10.3 displaying PDFs, the thin blue lines correspond to the reference estimated by the Gaussian KDE using $N_1 = 21\,059$ realizations for $\mathbf{Q}^{(1)}$, $N_2 = 178\,941$ realizations for $\mathbf{Q}^{(2)}$, $N = 200\,000$ realizations for both B and $\mathbf{Q}$.



(a) PDF $q_1 \mapsto p_{Q_1}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3}(q_3)$

(d) PDF $q_4 \mapsto p_{Q_4}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6}(q_6)$

Figure 6: PDF of the components Q_1 to Q_6 of the random output vector $\mathbf{Q}$, corresponding to the reference (blue thin line) and estimated with the optimal truncated PCE with $N_g = 12$, $n_u = 2$, and $n_g = 2$ (red thick line). In figures (a), (b), and (c), the curves are almost identical.

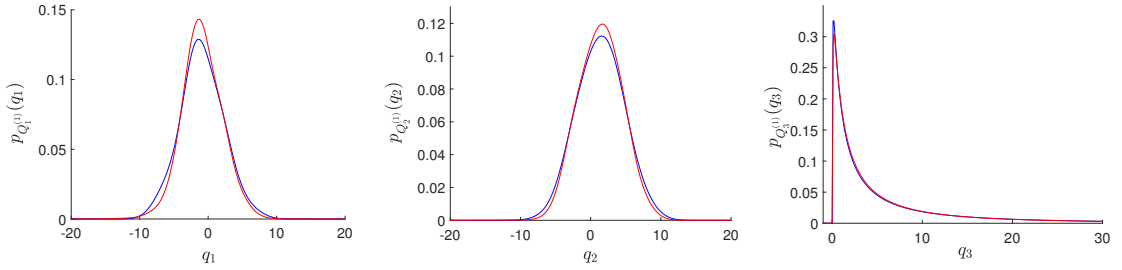
These realizations are obtained from the learned dataset.

Optimal PCE of $\boldsymbol{Q}^{(1)}$ corresponding to cluster $\sigma = 1$. Convergence is reached for $N_g^{\text{opt}} = 4$, $n_u^{\text{opt}} = 2$, and $n_g^{\text{opt}} = 2$. The corresponding value of $\mathcal{J}^{\text{opt}}(N_g^{\text{opt}}, n_u^{\text{opt}}, n_g^{\text{opt}})$ is 0.990, indicating a very good convergence. Figures 7a to 7f display the PDFs of the 6 components of $\mathbf{Q}^{(1)}$ (the reference) compared to the estimated PDFs of $\mathbf{Q}^{(1),\text{chaos}}$ using the optimal truncated PCE with N_1 points. The predictions obtained through this approach for the first cluster are accurate. Figure 7d shows that this cluster, $\mathcal{D}_1(\mathbf{X}^{(1)})$, corresponds to a long residence time, as the peak of the PDF of $Q_4^{(1)}$ occurs at $\Delta_t = exp(4.8) \simeq 120\, ps$. The peak of the PDF of $Q_5 = D_x$ is reached at $D_x = -123$ Å, and the peak of the PDF of $Q_6 = D_y$ is reached at $D_y = 190$ Å.

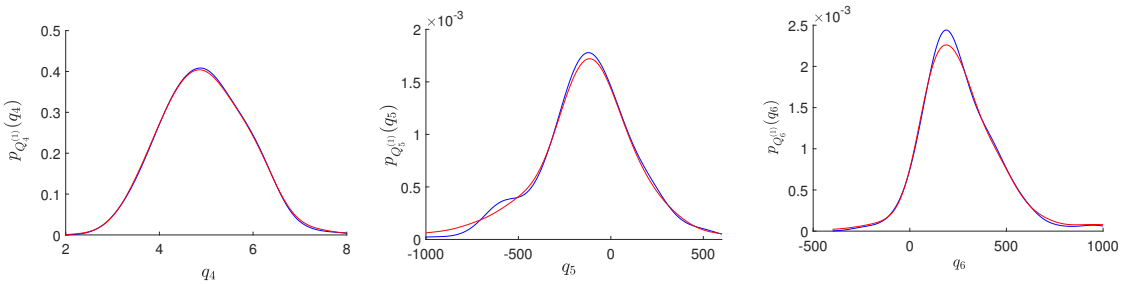
Optimal PCE of $\boldsymbol{Q}^{(2)}$ corresponding to cluster $\sigma = 2$. Convergence is reached for $N_g^{\text{opt}} = 8$, $n_u^{\text{opt}} = 2$, and $n_g^{\text{opt}} = 2$. The corresponding value of $\mathcal{J}^{\text{opt}}(N_g^{\text{opt}}, n_u^{\text{opt}}, n_g^{\text{opt}})$ is 0.987, indicating a good convergence. Figures 8a to 8f display the PDFs of the 6 components of $\mathbf{Q}^{(2)}$ (the reference) compared to the estimated PDFs of $\mathbf{Q}^{(2),\text{chaos}}$ using the optimal truncated PCE with N_2 points. The predictions obtained through this approach for the second cluster are accurate. Figure 8d shows a mixture of the two physical regimes in this cluster, $\mathcal{D}_2(\mathbf{X}^{(2)})$, with the dominant regime of short residence time, as the first peak of the PDF of $Q_4^{(2)}$ occurs at $\Delta_t = exp(1.33) \simeq 3.8\, ps$.

Optimal PCE of random variable B for identifying the cluster number σ. Convergence is reached for $N_g^{\text{opt}} = 14$, $n_u^{\text{opt}} = 2$, and $n_g^{\text{opt}} = 2$. The corresponding value of $\mathcal{J}^{\text{opt}}(N_g^{\text{opt}}, n_u^{\text{opt}}, n_g^{\text{opt}})$ is 0.900. It should be noted that the rate of convergence is not critical as we are dealing with two clusters ($K = 2$), where $\mathcal{B}_1 =]-\infty, 0]$ and $\mathcal{B}_2 =]0, +\infty[$. Thus, we only need to determine if, for each realization $b_0^{\text{chaos}} = \mathbb{b}^{\text{chaos}}(\mathbf{w}_0, \mathbf{u}_0)$ of B^{chaos}, we have $b_0^{\text{chaos}} \leq 0$ or $b_0^{\text{chaos}} > 0$. Figure 9 displays the PDF of B (the reference) compared to the estimated PDF of B^{chaos} using the optimal truncated PCE with N points. The prediction obtained through this approach for the second cluster is sufficiently accurate.

Validation of the PCE-based statistical surrogate model using the cluster separation. Once the optimal coefficients of the truncated PCEs for $\mathbf{Q}^{(1),\text{chaos}}$, $\mathbf{Q}^{(2),\text{chaos}}$, and B^{chaos} have been estimated as described above, the validation process is carried out following the explanation in Section 9. To do so, $N_v = 200\,000$ new realizations of random control

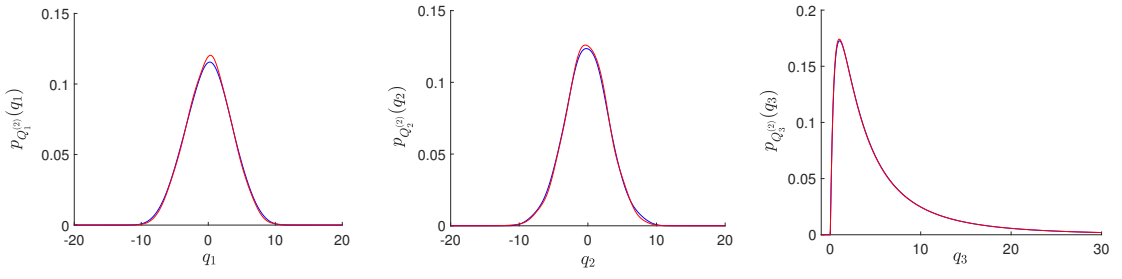


(a) PDF $q_1 \mapsto p_{Q_1^{(1)}}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2^{(1)}}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3^{(1)}}(q_3)$

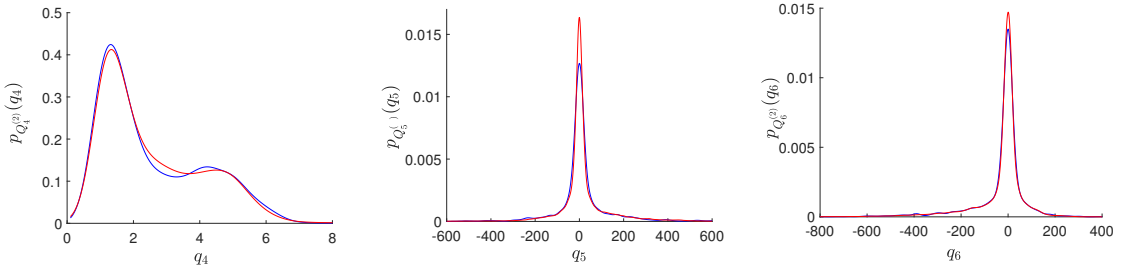


(d) PDF $q_4 \mapsto p_{Q_4^{(1)}}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5^{(1)}}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6^{(1)}}(q_6)$

Figure 7: PDF of the components $Q_1^{(1)}$ to $Q_6^{(1)}$ of the random output vector $\mathbf{Q}^{(1)}$, corresponding to the reference (blue thin line) and $Q_1^{(1,\text{chaos})}$ to $Q_6^{(1,\text{chaos})}$ of the random output vector $\mathbf{Q}^{(1,\text{chaos})}$, estimated with the optimal truncated PCE with $N_g = 4$, $n_u = 2$, and $n_g = 2$ (red thick line).



(a) PDF $q_1 \mapsto p_{Q_1^{(2)}}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2^{(2)}}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3^{(2)}}(q_3)$



(d) PDF $q_4 \mapsto p_{Q_4^{(2)}}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5^{(2)}}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6^{(2)}}(q_6)$

Figure 8: PDF of the components $Q_1^{(2)}$ to $Q_6^{(2)}$ of the random output vector $\mathbf{Q}^{(2)}$, corresponding to the reference (blue thin line) and $Q_1^{(2,\text{chaos})}$ to $Q_6^{(2,\text{chaos})}$ of the random output vector $\mathbf{Q}^{(2,\text{chaos})}$, estimated with the optimal truncated PCE with $N_g = 8$, $n_u = 2$, and $n_g = 2$ (red thick line). In figures (b) and (c), the curves are almost identical.

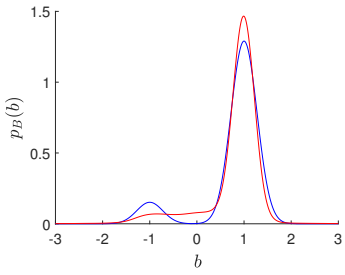
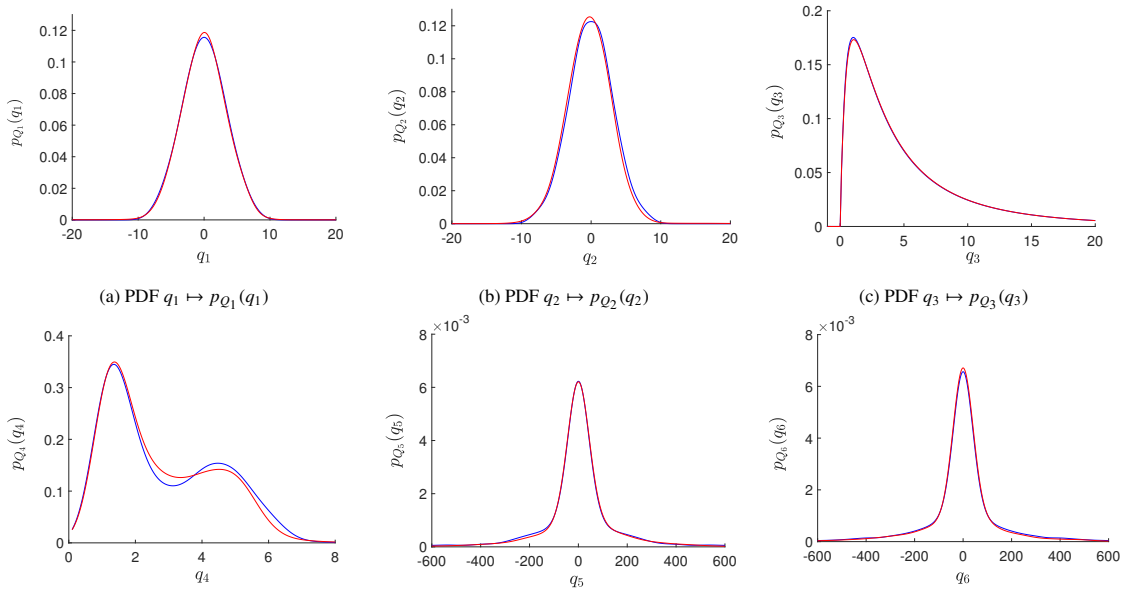


Figure 9: Graph of the function $N_g \mapsto \underline{\mathcal{J}}(N_g)$ representing the relationship between the degree N_g of the PCE constructed using the projection method, enabling quantification of the induced error.

variable $\mathbf{W}$ are generated according to the probability measure $P_{\mathbf{W}}(d\mathbf{w})$. Then, N_v corresponding realizations of $\mathbf{Q}^{\text{chaos}}$ are generated. Figures 10a to 10f display the PDFs of the 6 components of $\mathbf{Q}$. These PDFs correspond to the reference values (estimated with the learned dataset) and the values estimated using the algorithm based on the three PCEs. The predictions obtained through this PCE approach are highly accurate.



(a) PDF $q_1 \mapsto p_{Q_1}(q_1)$ (b) PDF $q_2 \mapsto p_{Q_2}(q_2)$ (c) PDF $q_3 \mapsto p_{Q_3}(q_3)$

(d) PDF $q_4 \mapsto p_{Q_4}(q_4)$ (e) PDF $q_5 \mapsto p_{Q_5}(q_5)$ (f) PDF $q_6 \mapsto p_{Q_6}(q_6)$

Figure 10: PDF of the components Q_1 to Q_6 of the random output vector $\mathbf{Q}$, corresponding to the reference (blue thin line) and estimated on the base of the optimal truncated PCEs of the clusters (red thick line).

11. Conclusion

We have presented a formulation and an algorithm developed for a polynomial chaos representation of a vector-valued random quantity of interest (the output) as a function of an input composed of a part of a vector-valued random control parameter with known probability measure and another part of a vector-valued random latent variable with unknown probability measure. The training dataset consists of heterogeneous data, which poses challenges for accurately estimating the chaos coefficients. The proposed approach enables the construction of a global, accurate, and highly efficient statistical surrogate model for evaluating an output realization given an input realization. As an alternative, we have also proposed a construction based on the clustering of the learned dataset, which can facilitate

offline construction in the case of heterogeneous data. The application presented, which concerns the collisions of helium atoms on a graphite substrate, demonstrates the accuracy and efficiency of the proposed approach. This approach, which allows the construction of a rapid online surrogate model, is general and should allow the analysis of large complex systems beyond the computational capabilities currently available.

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.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors do not use Generative AI and AI-assisted technologies in the writing process and take full responsibility for the content of the publication.

References

[1] C. Soize, R. Ghanem, Reduced chaos decomposition with random coefficients of vector-valued random variables and random fields, Computer Methods in Applied Mechanics and Engineering 198 (21-26) (2009) 1926–1934. doi:10.1016/j.cma.2008.12.035.

[2] N. Wiener, The homogeneous chaos, American Journal of Mathematics 60 (1) (1938) 897–936.

[3] R. H. Cameron, W. T. Martin, The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals, Annals of Mathematics 48 (2) (1947) 385–392. doi:10.2307/1969178.

[4] R. Ghanem, P. Spanos, Polynomial chaos in stochastic finite elements, Journal of Applied Mechanics - Transactions of the ASME 57 (1) (1990) 197–202.

[5] R. Ghanem, P. D. Spanos, Stochastic Finite Elements: a Spectral Approach, Springer-Verlag, New York, 1991.

[6] D. Xiu, G. E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24 (2) (2002) 619–644. doi:10.1137/S1064827501387826.

[7] C. Soize, R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM Journal on Scientific Computing 26 (2) (2004) 395–410. doi:10.1137/S1064827503424505.

[8] D. Lucor, C.-H. Su, G. E. Karniadakis, Generalized polynomial chaos and random oscillators, International Journal for Numerical Methods in Engineering 60 (3) (2004) 571–596. doi:10.1002/nme.976.

[9] S. Dolgov, B. N. Khoromskij, A. Litvinenko, H. G. Matthies, Polynomial chaos expansion of random coefficients and the solution of stochastic partial differential equations in the tensor train format, SIAM/ASA Journal on Uncertainty Quantification 3 (1) (2015) 1109–1135.

[10] R. Tipireddy, R. Ghanem, Basis adaptation in homogeneous chaos spaces, Journal of Computational Physics 259 (2014) 304–317. doi:10.1016/j.jcp.2013.12.009.

[11] M. Mignolet, C. Soize, Compressed principal component analysis of non-Gaussian vectors, SIAM/ASA Journal on Uncertainty Quantification 8 (4) (2020) 1261–1286. doi:10.1137/20M1322029.

[12] D. Ghosh, R. Ghanem, Stochastic convergence acceleration through basis enrichment of polynomial chaos expansions, International journal for numerical methods in engineering 73 (2) (2008) 162–184. doi:10.1002/nme.2066.

[13] V. Keshavarzzadeh, R. Ghanem, S. Masri, O. Aldraihem, Convergence acceleration of polynomial chaos solutions via sequence transformation, Computer Methods in Applied Mechanics and Engineering 271 (2014) 167–184. doi:10.1016/j.cma.2013.12.003.

[14] Y. M. Marzouk, H. N. Najm, Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, Journal of Computational Physics 228 (6) (2009) 1862–1902. doi:10.1016/j.jcp.2008.11.024.

[15] C. Soize, Polynomial chaos expansion of a multimodal random vector, SIAM-ASA Journal on Uncertainty Quantification 3 (1) (2015) 34–60. doi:10.1137/140968495.

[16] B. J. Debusschere, H. N. Najm, P. P. Pébay, O. M. Knio, R. Ghanem, O. P. Le Maître, Numerical challenges in the use of polynomial chaos representations for stochastic processes, SIAM journal on scientific computing 26 (2) (2004) 698–719. doi:10.1137/S1064827503427741.

[17] X. Wan, G. E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, Journal of Computational Physics 209 (2) (2005) 617–642. doi:10.1016/j.jcp.2005.03.023.

[18] X. Wan, G. E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM Journal on Scientific Computing 28 (3) (2006) 901–928. doi:10.1137/050627630.

[19] G. Blatman, B. Sudret, Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, Comptes Rendus Mécanique 336 (6) (2008) 518–523. doi:10.1016/j.crme.2008.02.013.

[20] S. Das, R. Ghanem, S. Finette, Polynomial chaos representation of spatio-temporal random fields from experimental measurements, Journal of Computational Physics 228 (23) (2009) 8726–8751. doi:10.1016/j.jcp.2009.08.025.

[21] O. G. Ernst, A. Mugler, H.-J. Starkloff, E. Ullmann, On the convergence of generalized polynomial chaos expansions, ESAIM: Mathematical Modelling and Numerical Analysis 46 (2) (2012) 317–339. doi:10.1051/m2an/2011045.

[22] I. Babuska, R. Tempone, G. E. Zouraris, Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation, Computer Methods in Applied Mechanics and Engineering 194 (12-16) (2005) 1251–1294.

[23] H. Matthies, A. Keese, Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering 194 (12-16) (2005) 1295–1331.
[24] I. Babuska, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Journal on Numerical Analysis 45 (3) (2007) 1005–1034.
[25] A. Nouy, A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering 196 (45-48) (2007) 4521–4537.
[26] A. Nouy, Generalized spectral decomposition method for solving stochastic finite element equations: Invariant subspace problem and dedicated algorithms, Computer Methods in Applied Mechanics and Engineering 197 (51-52) (2008) 4718–4736.
[27] A. Nouy, O. P. L. Maitre, Generalized spectral decomposition for stochastic nonlinear problems, Journal of Computational Physics 228 (1) (2009) 202–235.
[28] A. Nouy, Proper generalized decomposition and separated representations for the numerical solution of high dimensional stochastic problems, Archives of Computational Methods in Engineering 17 (4) (2010) 403–434.
[29] O. Le Maître, O. M. Knio, Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics, Springer Science & Business Media, 2010.
[30] H. N. Najm, Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics, Annual review of fluid mechanics 41 (2009) 35–52. doi:10.1146/annurev.fluid.010908.165248.
[31] R. Ghanem, R. M. Kruger, Numerical solution of spectral stochastic finite element systems, Computer Methods in Applied Mechanics and Engineering 129 (1996) 289–303.
[32] R. Ghanem, J. Red-Horse, Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach, Physica D 133 (1-4) (1999) 137–144.
[33] R. Ghanem, Ingredients for a general purpose stochastic finite elements formulation, Computer Methods in Applied Mechanics and Engineering 168 (1-4) (1999) 19–34. doi:10.1016/S0045-7825(98)00106-6.
[34] M. F. Pellissetti, R. G. Ghanem, Iterative solution of systems of linear equations arising in the context of stochastic finite elements, Advances in engineering software 31 (8-9) (2000) 607–616. doi:10.1016/S0965-9978(00)00034-X.
[35] M. Deb, I. Babuska, J. Oden, Solution of stochastic partial differential equations using galerkin finite element techniques, Computer Methods in Applied Mechanics and Engineering 190 (2001) 6359–6372.
[36] R. Ghanem, P. Spanos, Stochastic Finite Elements: A spectral Approach, (revised edition) Dover Publications, New York, 2003.
[37] P. Frauenfelder, C. Schwab, R. Todor, Finite elements for elliptic problems with stochastic coefficients, Computer Methods in Applied Mechanics and Engineering 194 (2-5) (2005) 205–228.
[38] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique 15 (1-3) (2006) 81–92. doi:10.3166/remn.15.81-92.
[39] X. Xu, A multiscale stochastic finite element method on elliptic problems involving uncertainties, Computer Methods in Applied Mechanics and Engineering 196 (25-28) (2007) 2723–2736.
[40] H. G. Matthies, Stochastic finite elements: Computational approaches to stochastic partial differential equations, Zamm-Zeitschrift Fur Angewandte Mathematik Und Mechanik 88 (11) (2008) 849–873.
[41] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, Journal of Computational Physics 230 (6) (2011) 2345–2367. doi:10.1016/j.jcp.2010.12.021.
[42] N. Luthen, S. Marelli, B. Sudret, Sparse polynomial chaos expansions: Literature survey and benchmark, SIAM/ASA Journal on Uncertainty Quantification 9 (2) (2021) 593–649. doi:10.1137/20M1315774.
[43] J. Kaipio, E. Somersalo, Statistical and Computational Inverse Problems, Vol. 160, Springer Science & Business Media, 2005. doi:10.1007/b138659.
[44] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3) (2001) 425–464. doi:10.1111/1467-9868.00294.
[45] A. Tarantola, Inverse Problem Theory And Methods For Model Parameter Estimation, Vol. 89, SIAM, Philadelphia, 2005.
[46] A. M. Stuart, Inverse problems: a Bayesian perspective, Acta Numerica 19 (2010) 451–559. doi:10.1017/S0962492910000061.
[47] H. Owhadi, C. Scovel, T. Sullivan, On the brittleness of Bayesian inference, SIAM Review 57 (4) (2015) 566–582. doi:10.1137/130938633.
[48] H. G. Matthies, E. Zander, B. V. Rosić, A. Litvinenko, O. Pajonk, Inverse problems in a Bayesian setting, in: Computational Methods for Solids and Fluids, Vol. 41, Springer, 2016, pp. 245–286. doi:10.1007/978-3-319-27996-1_10.
[49] M. Dashti, A. M. Stuart, The Bayesian approach to inverse problems, in: R. Ghanem, D. Higdon, O. Houman (Eds.), Handbook of Uncertainty Quantification, Springer, Cham, Switzerland, 2017, Ch. 10, pp. 311–428. doi:10.1007/978-3-319-12385-1_7.
[50] C. Soize, R. Ghanem, C. Desceliers, Sampling of Bayesian posteriors with a non-Gaussian probabilistic learning on manifolds from a small dataset, Statistics and Computing 30 (5) (2020) 1433–1457. doi:10.1007/s11222-020-09954-6.
[51] P. Fearnhead, Exact and efficient Bayesian inference for multiple changepoint problems, Statistics and Computing 16 (2) (2006) 203–213. doi:10.1007/s11222-006-8450-8.
[52] A. Golightly, D. J. Wilkinson, Bayesian sequential inference for nonlinear multivariate diffusions, Statistics and Computing 16 (4) (2006) 323–338. doi:10.1007/s11222-006-9392-x.
[53] N. Zabaras, B. Ganapathysubramanian, A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach, Journal of Computational Physics 227 (9) (2008) 4697–4735.
[54] X. Ma, N. Zabaras, An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method, Inverse Problems 25 (3) (2009) Article Number: 035013.
[55] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, O. Ghattas, Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations, SIAM Journal on Scientific Computing 33 (1) (2011) 407–432. doi:10.1137/090780717.
[56] T. A. El Moselhy, Y. M. Marzouk, Bayesian inference with optimal maps, Journal of Computational Physics 231 (23) (2012) 7815–7850. doi:10.1016/j.jcp.2012.07.022.

[57] G. Perrin, C. Soize, D. Duhamel, C. Funfschilling, Karhunen–loève expansion revisited for vector-valued random fields: Scaling, errors and optimal basis., Journal of Computational Physics 242 (2013) 607–622. doi:10.1016/j.jcp.2013.02.036.
[58] H. N. Najm, K. Chowdhary, Inference given summary statistics, in: R. Ghanem, D. Higdon, O. Houman (Eds.), Handbook of Uncertainty Quantification, Springer, Cham, Switzerland, 2017, Ch. 3, pp. 33–67.
[59] P. Tsilifis, R. Ghanem, Bayesian adaptation of chaos representations using variational inference and sampling on geodesics, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 474 (2217) (2018) 20180285. doi:10.1098/rspa.2018.0285.
[60] Q. Zhou, W. Liu, J. Li, Y. Marzouk, An approximate empirical Bayesian method for large-scale linear-Gaussian inverse problems, Inverse Problems (2018).
[61] G. Perrin, C. Soize, Adaptive method for indirect identification of the statistical properties of random fields in a Bayesian framework, Computational Statistics 35 (1) (2020) 111–133. doi:10.1007/s00180-019-00936-5.
[62] C. Desceliers, R. Ghanem, C. Soize, Maximum likelihood estimation of stochastic chaos representations from experimental data, International Journal for Numerical Methods in Engineering 66 (6) (2006) 978–1001. doi:10.1002/nme.1576.
[63] S. Das, R. Ghanem, J. C. Spall, Asymptotic sampling distribution for polynomial chaos representation from data: a maximum entropy and fisher information approach, SIAM Journal on Scientific Computing 30 (5) (2008) 2207–2234. doi:10.1137/060652105.
[64] M. Arnst, R. Ghanem, C. Soize, Identification of Bayesian posteriors for coefficients of chaos expansions, Journal of Computational Physics 229 (9) (2010) 3134–3154. doi:10.1016/j.jcp.2009.12.033.
[65] C. Soize, Identification of high-dimension polynomial chaos expansions with random coefficients for non-Gaussian tensor-valued random fields using partial and limited experimental data, Computer methods in applied mechanics and engineering 199 (33-36) (2010) 2150–2164. doi:10.1016/j.cma.2010.03.013.
[66] C. Soize, A computational inverse method for identification of non-Gaussian random fields using the Bayesian approach in very high dimension, Computer Methods in Applied Mechanics and Engineering 200 (45-46) (2011) 3083–3099. doi:10.1016/j.cma.2011.07.005.
[67] G. Perrin, C. Soize, D. Duhamel, C. Funfschilling, Identification of polynomial chaos representations in high dimension from a set of realizations, SIAM Journal on Scientific Computing 34 (6) (2012) A2917–A2945. doi:10.1137/11084950X.
[68] R. Madankan, P. Singla, T. Singh, P. D. Scott, Polynomial-chaos-based Bayesian approach for state and parameter estimations, Journal of Guidance, Control, and Dynamics 36 (4) (2013) 1058–1074. doi:10.2514/1.58377.
[69] B. Chen-Charpentier, D. Stanescu, Parameter estimation using polynomial chaos and maximum likelihood, International Journal of Computer Mathematics 91 (2) (2014) 336–346. doi:10.1080/00207160.2013.809069.
[70] 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. doi:10.1016/j.cma.2013.11.001.
[71] J. B. Nagel, B. Sudret, Spectral likelihood expansions for Bayesian inference, Journal of Computational Physics 309 (2016) 267–294.
[72] I. Sraj, O. P. Le Maître, O. M. Knio, I. Hoteit, Coordinate transformation and polynomial chaos for the Bayesian inference of a Gaussian process with parametrized prior covariance function, Computer Methods in Applied Mechanics and Engineering 298 (2016) 205–228. doi:10.1016/j.cma.2015.10.002.
[73] Q. Shao, A. Younes, M. Fahs, T. A. Mara, Bayesian sparse polynomial chaos expansion for global sensitivity analysis, Computer Methods in Applied Mechanics and Engineering 318 (2017) 474–496. doi:10.1016/j.cma.2017.01.033.
[74] C. Soize, E. Capiez-Lernout, J.-F. Durand, C. Fernandez, L. Gagliardini, Probabilistic model identification of uncertainties in computational models for dynamical systems and experimental validation, Computer Methods in Applied Mechanics and Engineering 198 (1) (2008) 150–163. doi:10.1016/j.cma.2008.04.007.
[75] C. Desceliers, C. Soize, S. Naili, G. Haïat, Probabilistic model of the human cortical bone with mechanical alterations in ultrasonic range, Mechanical Systems and Signal Processing 32 (-) (2012) 170–177. doi:10.1016/j.ymssp.2012.03.008.
[76] M. Arnst, C. Soize, Identification and sampling of Bayesian posteriors of high-dimensional symmetric positive-definite matrices for data-driven updating of computational models, Computer Methods in Applied Mechanics and Engineering 352 (2019) 300–323. doi:10.1016/j.cma.2019.04.025.
[77] J. L. Beck, L. S. Katafygiotis, Updating models and their uncertainties. i: Bayesian statistical framework, Journal of Engineering Mechanics 124 (4) (1998) 455–461.
[78] J. Beck, S. Au, Bayesian updating of structural models and reliability using markov chain monte carlo simulation, Journal of Engineering Mechanics - ASCE 128 (4) (2002) 380–391.
[79] J. Ching, J. Beck, K. Porter, Bayesian state and parameter estimation of uncertain dynamical systems, Probabilistic Engineering Mechanics 21 (1) (2006) 81–96.
[80] S. Cheung, J. Beck, Calculation of posterior probabilities for bayesian model class assessment and averaging from posterior samples based on dynamic system data, Computer-Aided Civil and Infrastructure Engineering 25 (5) (2010) 304–321.
[81] L. Parussini, D. Venturi, P. Perdikaris, G. E. Karniadakis, Multi-fidelity Gaussian process regression for prediction of random fields, Journal of Computational Physics 336 (2017) 36–50.
[82] W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Ročková, T. Group, Adaptive bayesian SLOPE: model selection with incomplete data, Journal of Computational and Graphical Statistics 31 (1) (2022) 113–137. doi:10.1080/10618600.2021.1963263.
[83] J. P. Kleijnen, Kriging metamodeling in simulation: A review, European Journal of Operational Research 192 (3) (2009) 707–716. doi:10.1016/j.ejor.2007.10.013.
[84] P. G. Constantine, E. Dow, Q. Wang, Active subspace methods in theory and practice: applications to kriging surfaces, SIAM Journal on Scientific Computing 36 (4) (2014) A1500–A1524. doi:10.1137/130916138.
[85] P. Kersaudy, B. Sudret, N. Varsier, O. Picon, J. Wiart, A new surrogate modeling technique combining kriging and polynomial chaos expansions–application to uncertainty analysis in computational dosimetry, Journal of Computational Physics 286 (2015) 103–117.
[86] J. P. Kleijnen, Regression and kriging metamodels with their experimental designs in simulation: a review, European Journal of Operational Research 256 (1) (2017) 1–16. doi:10.1016/j.ejor.2016.06.041.
[87] D. G. Giovanis, M. D. Shields, Data-driven surrogates for high dimensional models using Gaussian process regression on the Grassmann manifold, Computer Methods in Applied Mechanics and Engineering 370 (2020) 113269. doi:10.1016/j.cma.2020.113269.

[88] Z. Liu, D. Lesselier, B. Sudret, J. Wiart, Surrogate modeling based on resampled polynomial chaos expansions, Reliability Engineering & System Safety 202 (2020) 107008. doi:10.1016/j.ress.2020.107008.
[89] Y. Zhou, Z. Lu, J. Hu, Y. Hu, Surrogate modeling of high-dimensional problems via data-driven polynomial chaos expansions and sparse partial least square, Computer Methods in Applied Mechanics and Engineering 364 (2020) 112906. doi:10.1016/j.cma.2020.112906.
[90] K. B. Korb, A. E. Nicholson, Bayesian artificial intelligence, CRC press, Boca Raton, 2010.
[91] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT press, 2012.
[92] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (7553) (2015) 452–459. doi:10.1038/nature14541.
[93] S. Russel, P. Norvig, Artifical Intelligence, A Modern Approach, Third Edition, Pearson, Harlow, 2016.
[94] T. Hofmann, B. Schölkopf, A. J. Smola, Kernel methods in machine learning, The Annals of Statistics 36 (3) (2008) 1171–1220. doi:10.1214/009053607000000677.
[95] B. Scholkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2018.
[96] J.-L. Akian, L. Bonnet, H. Owhadi, É. Savin, Learning best kernels from data in gaussian process regression. with application to aerodynamics, Journal of Computational Physics 470 (2022) 111595. doi:10.1016/j.jcp.2022.111595.
[97] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 2000. doi:10.1007/978-1-4757-3264-1.
[98] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Vol. 112, Springer, 2013.
[99] J. Taylor, R. J. Tibshirani, Statistical learning and selective inference, Proceedings of the National Academy of Sciences 112 (25) (2015) 7629–7634. doi:10.1073/pnas.1507583112.
[100] R. Swischuk, L. Mainini, B. Peherstorfer, K. Willcox, Projection-based model reduction: Formulations for physics-based machine learning, Computers & Fluids 179 (2019) 704–717. doi:10.1016/j.compfluid.2018.07.021.
[101] A. C. Öztireli, M. Alexa, M. Gross, Spectral sampling of manifolds, ACM Transactions on Graphics (TOG) 29 (6) (2010) 1–8. doi:10.1145/1882261.1866190.
[102] C. Soize, R. Ghanem, Data-driven probability concentration and sampling on manifold, Journal of Computational Physics 321 (2016) 242–258. doi:10.1016/j.jcp.2016.05.044.
[103] G. Perrin, C. Soize, S. Marque-Pucheu, J. Garnier, Nested polynomial trends for the improvement of Gaussian process-based predictors, Journal of Computational Physics 346 (2017) 389–402. doi:10.1016/j.jcp.2017.05.051.
[104] C. Soize, R. Ghanem, Polynomial chaos representation of databases on manifolds, Journal of Computational Physics 335 (2017) 201–221. doi:10.1016/j.jcp.2017.01.031.
[105] G. Perrin, C. Soize, N. Ouhbi, Data-driven kernel representations for sampling with an unknown block dependence structure under correlation constraints, Computational Statistics & Data Analysis 119 (2018) 139–154. doi:10.1016/j.csda.2017.10.005.
[106] C. Soize, R. Ghanem, C. Safta, X. Huan, Z. P. Vane, J. C. Oefelein, G. Lacaze, H. N. Najm, Q. Tang, X. Chen, Entropy-based closure for probabilistic learning on manifolds, Journal of Computational Physics 388 (2019) 528–533. doi:10.1016/j.jcp.2018.12.029.
[107] Y. Kevrekidis, Manifold learning for parameter reduction, Bulletin of the American Physical Society 65 (2020). doi:10.1016/j.jcp.2019.04.015.
[108] C. Soize, R. Ghanem, Probabilistic learning on manifolds, Foundations of Data Science 2 (3) (2020) 279–307. doi:10.3934/fods.2020013.
[109] K. Kontolati, D. Loukrezis, K. R. dos Santos, D. G. Giovanis, M. D. Shields, Manifold learning-based polynomial chaos expansions for high-dimensional surrogate models, International Journal for Uncertainty Quantification 12 (4) (2022). doi:10.1615/Int.J.UncertaintyQuantification.2022039936.
[110] C. Soize, R. Ghanem, Probabilistic learning on manifolds (PLoM) with partition, International Journal for Numerical Methods in Engineering 123 (1) (2022) 268–290. doi:10.1002/nme.6856.
[111] J. O. Almeida, F. A. Rochinha, A probabilistic learning approach applied to the optimization of wake steering in wind farms, Journal of Computing and Information Science in Engineering 23 (1) (2023) 011003. doi:10.1115/1.4054501.
[112] K. Zhong, J. G. Navarro, S. Govindjee, G. G. Deierlein, Surrogate modeling of structural seismic response using Probabilistic Learning on Manifolds, Earthquake Engineering and Structural Dynamics Online (2023) 1–22.
[113] J. O. Almeida, F. A. Rochinha, Uncertainty quantification of waterflooding in oil reservoirs computational simulations using a probabilistic learning approach, Journal of Computing and Information Science in Engineering 13 (4) (2023) 1–22. doi:10.1615/Int.J.UncertaintyQuantification.2023041042.
[114] S. Pan, K. Duraisamy, Physics-informed probabilistic learning of linear embeddings of nonlinear dynamics with guaranteed stability, SIAM Journal on Applied Dynamical Systems 19 (1) (2020) 480–509. doi:10.1137/19M1267246.
[115] C. Soize, R. Ghanem, Physics-constrained non-Gaussian probabilistic learning on manifolds, International Journal for Numerical Methods in Engineering 121 (1) (2020) 110–145. doi:10.1002/nme.6202.
[116] C. Soize, R. Ghanem, Probabilistic learning on manifolds constrained by nonlinear partial differential equations for small datasets, Computer Methods in Applied Mechanics and Engineering 380 (2021) 113777. doi:10.1016/j.cma.2021.113777.
[117] C. Soize, Probabilistic learning inference of boundary value problem with uncertainties based on Kullback-Leibler divergence under implicit constraints, Computer Methods in Applied Mechanics and Engineering 395 (2022) 115078. doi:10.1016/j.cma.2022.115078.
[118] C. Soize, Probabilistic learning constrained by realizations using a weak formulation of fourier transform of probability measures, Computational Statistics (2022) 1–30, published online 23 December 2022doi:10.1007/s00180-022-01300-w.
[119] L. Rabiner, B. Juang, An introduction to hidden Markov models, ieee assp magazine 3 (1) (1986) 4–16. doi:10.1109/MASSP.1986.1165342.
[120] A. Rodriguez, D. B. Dunson, A. E. Gelfand, The nested Dirichlet process, Journal of the American statistical Association 103 (483) (2008) 1131–1154. doi:10.1198/016214508000000553.
[121] Y. Jung, H. Park, D.-Z. Du, B. L. Drake, A decision criterion for the optimal number of clusters in hierarchical clustering, Journal of Global Optimization 25 (1) (2003) 91–111. doi:10.1023/A:1021394316112.
[122] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.
[123] Z. Ghahramani, An introduction to hidden Markov models and Bayesian networks, International journal of pattern recognition and artificial intelligence 15 (01) (2001) 9–42. doi:10.1142/S0218001401000836.
[124] C. Soize, Uncertainty Quantification. An Accelerated Course with Advanced Applications in Computational Engineering, Springer, New

York, 2017. doi:10.1007/978-3-319-54339-0.
[125] C. Soize, C. Desceliers, Computational aspects for constructing realizations of polynomial chaos in high dimension, SIAM Journal on Scientific Computing 32 (5) (2010) 2820–2831. doi:10.1137/100787830.
[126] J. A. Hartigan, M. A. Wong, et al., A K-means clustering algorithm, Royal Statistical Society,Applied statistics 28 (1) (1979) 100–108. doi:10.2307/2346830.
[127] K. P. Sinaga, M.-S. Yang, Unsupervised K-means clustering algorithm, IEEE access 8 (2020) 80716–80727. doi:10.1109/ACCESS.2020.2988796.
[128] A. Lukasová, Hierarchical agglomerative clustering procedure, Pattern Recognition 11 (5-6) (1979) 365–381. doi:10.1016/0031-3203(79)90049-9.
[129] S. M. Savaresi, D. L. Boley, S. Bittanti, G. Gazzaniga, Cluster selection in divisive clustering algorithms, in: Proceedings of the 2002 SIAM International Conference on Data Mining, SIAM, 2002, pp. 299–314.
[130] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, Density-based spatial clustering of applications with noise, in: Int. Conf. knowledge discovery and data mining, Vol. 240, 1996.
[131] J. Sander, M. Ester, H.-P. Kriegel, X. Xu, Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, Data mining and knowledge discovery 2 (1998) 169–194. doi:10.1023/A:1009745219419.
[132] H. Bäcklund, A. Hedblom, N. Neijman, A density-based spatial clustering of application with noise, Data Mining TNM033 33 (2011) 11–30.
[133] K.-L. Wu, M.-S. Yang, Mean shift-based clustering, Pattern Recognition 40 (11) (2007) 3035–3052. doi:10.1016/j.patcog.2007.02.006.
[134] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (2007) 395–416. doi:10.1007/s11222-007-9033-z.
[135] C. Maugis, G. Celeux, M.-L. Martin-Magniette, Variable selection for clustering with gaussian mixture models, Biometrics 65 (3) (2009) 701–709. doi:10.1111/j.1541-0420.2008.01160.x.
[136] P. Magnico, Q.-D. To, Collisions, adsorption and self diffusion of gas in nanometric channels by molecular dynamics and stochastic simulation and the case of helium gas in graphitic slit pore, International Journal of Heat and Mass Transfer 214 (2023) 124371. doi:10.1016/j.ijheatmasstransfer.2023.124371.