IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 April 2023, accepted 12 May 2023, date of publication 18 May 2023, date of current version 31 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3277521

RESEARCH ARTICLE

Data-Driven Method to Quantify Correlated Uncertainties

JEAHAN JUNG AND MINSEOK CHOI
Department of Mathematics, Pohang University of Science and Technology (POSTECH), Pohang, Gyeongbuk 37673, Republic of Korea
Corresponding author: Minseok Choi (mchoi@postech.ac.kr)

This work was supported by the National Research Foundation of Korea under Grant NRF-2021R1C1C1007875.

ABSTRACT Polynomial chaos (PC) has been proven to be an efficient method for uncertainty quantification, but its applicability is limited by two strong assumptions: the mutual independence of random variables and the requirement of exact knowledge about the distribution of the random variables. We describe a new data-driven method for dealing with correlated multivariate random variables for uncertainty quantification that requires only observed data of the random variables. It is based on the transformation of correlated random variables into independent random variables. We use singular value decomposition as a transformation strategy that does not require information about the probability distribution. For the transformed random variables, we can construct the PC basis to build a surrogate model. This approach provides an additional benefit of quantifying high-dimensional uncertainties by combining our method with the analysis-of-variance (ANOVA) method. We demonstrate in several numerical examples that our proposed approach leads to accurate solutions with a much smaller number of simulations compared to the Monte Carlo method.

INDEX TERMS Correlated random variables, high dimension, polynomial chaos expansion, uncertainty quantification.

I. INTRODUCTION

Many science and engineering models are subject to uncertainties from various sources including noisy data or the lack of information about the parameters in physical systems. To deal with these uncertainties, we can use regression-based methods [1], [2], [3] or sample-based methods depending on the sources of the uncertainties. Traditional sample-based methods such as Monte Carlo (MC) [4], [5], [6] requires a large number of model evaluations to investigate the impact of the uncertainties on the model output. This approach is not feasible for complex and large-scale models where a single simulation is computationally expensive. One approach to efficiently quantify uncertainty in such complex systems is to develop surrogate models, such as polynomial chaos (PC) expansions [7], [8], [9], [10], [11], [12], [13]. The PC expansion approximates the stochastic model output as a linear combination of orthogonal polynomial basis functions (i.e., polynomial chaos). The method for computing the coefficients of basis functions has been mainly studied in solving the stochastic differential equation, including the stochastic Galerkin method [9], [14], [15] and the stochastic collocation method [11], [16], [17], [18], [19], [20]. The number of simulations required in the PC method is an order of magnitude smaller than that required in the MC method for the same accuracy.

Despite the significant efficiency of the PC method, its application is limited by two strong assumptions. The first assumption is that it requires the exact knowledge of the probability distribution of uncertain parameters, which is rarely available in applied studies. To overcome this problem, several works [21], [22], [23], [24] have proposed an arbitrary PC method. This method requires only a finite number of moments of the random variables, which can be estimated from the available data. However, this method still relies on the second assumption of the PC method, the mutual independence of all random variables. This assumption leads to a straightforward construction of orthogonal polynomial basis functions, but the assumption may fail in practical problems where random variables are typically correlated. To account for correlations between random variables, a number of studies have been proposed using transformation of the random

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

variables [25], [26], [27], [28], dominating measures [29], [30], and the construction of orthogonal basis functions for correlated random variables [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]. However, these methods often require information about the probability distribution. Although the algorithms presented in [32], [34], [37] can be extended to cases where only observed data are available, the construction of orthogonal basis functions may be ill-conditioned when dealing with high-dimensional uncertainty. The authors in [41], [42] propose methods for constructing surrogate models when both of the above assumptions are absent, but they are only applicable to the stochastic Galerkin method.

In this paper, we describe a new method to construct a surrogate model for uncertainty quantification in the absence of both assumptions of the PC method. The method is based on the use of an invertible transformation to convert correlated random variables into independent random variables, followed by the construction of the PC basis for the converted variables. This approach has been used in several works [25], [26], [27], [28] with the well-known Rosenblatt [43] and Nataf [44] transformations, which require the exact knowledge of the probability density functions (PDFs) and cumulative distribution functions (CDFs) of the given random variables. However, closed forms of PDFs and CDFs are rarely available in practical applications. Therefore we use singular value decomposition (SVD) as a transformation strategy to construct a data-driven model. SVD provides a natural way to assimilate the data for transformation without requiring any information about the distribution of the random variables. The data transformed by SVD are assumed to represent the samples drawn from some independent random variables. We construct the PC basis for the transformed random variables using the arbitrary PC method [21], [22], [23], [24], which does not require exact knowledge of the PDFs and CDFs of the random variables. Our numerical examples demonstrate that the construction of our surrogate model requires much fewer simulations than those required by the MC method.

An additional benefit of our method is that it can cope with high dimensionality. When the stochastic model has high-dimensional multivariate random variables, the PC method suffers from the so-called curse of dimensionality because the number of required simulations grows rapidly as the dimension increases. There are many modifications of PC to alleviate this issue, such as the sparse-grid method [16], [19], [45], [46], [47], [48], [49], sparse PC method [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], and adaptive analysis-of-variance (ANOVA) method [63], [64], [65], [66]. Because these methods often rely on the tensor product structure of the model, they only work to handle independent random variables. In our framework, the methods are applicable after correlated random variables are transformed into independent random variables. In this study, we combine our proposed method with the adaptive ANOVA method to quantify the uncertainty in a stochastic model with high-dimensional correlated random variables.

Our main contributions can be summarized as:

- We propose a new data-driven method to deal with correlated multivariate random variables for uncertainty quantification that does not require the exact knowledge of the probability distribution of the random variables.
- The proposed method allows us to tackle the curse of dimensionality by combining our method with the adaptive ANOVA method.
- We demonstrate that our proposed method provides the same order of accuracy as the Monte Carlo method while requiring much fewer simulations.

The remainder of this paper is organized as follows. In Section II, we introduce the formulation of our stochastic problems and some extant methods for UQ. Section III is the core of this study, which explains a data-driven approach to deal with correlated random variables using proper transformation of them. In Section IV, we apply our framework to some examples of stochastic models and examine the efficacy of our method. We conclude this paper with some closing comments in Section V.

II. PRELIMINARIES

In this section, we present a formulation of the stochastic model of concern and introduce the existing UQ methods.

A. FORMULATION OF STOCHASTIC MODELS

We formulate a stochastic model based on the boundary value problem (BVP); however, the procedure is also applicable to general stochastic problems. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space where Ω, $\mathcal{F}$, and $\mathcal{P}$ are the sample space, σ-algebra, and probability measure, respectively. Consider a stochastic BVP defined in a physical domain $D \subset \mathbb{R}^n$,

$$\begin{aligned} \mathcal{L}(u(x,\omega), x, \omega) &= 0, \quad \text{in } \mathbb{R} \times D \times \Omega \\ \mathcal{B}(u(x,\omega), x, \omega) &= 0, \quad \text{in } \mathbb{R} \times \partial D \times \Omega \end{aligned} \tag{1}$$

with the solution $u : D \times \Omega \to \mathbb{R}$, where $\mathcal{L}$ is a differential operator and $\mathcal{B}$ is a boundary operator.

The important step prior to any procedure is to characterize the random input, ω, as finite random variables. The characterization is straightforward when the uncertainties of the model come from the finite physical parameters of the system since the parameters can be treated as finite random variables. However, this is nontrivial when the uncertainties include infinite-dimensional random processes. The common approach for dealing with this case is to employ the truncated Karhunen-Loève expansion to approximate the random process as a linear combination of finite random variables [7], [8], [14]. After the characterization, we can rewrite the original system (1) as

$$\begin{aligned} \mathcal{L}(u(x,\boldsymbol{\xi}), x, \boldsymbol{\xi}) &= 0, \quad \text{in } \mathbb{R} \times D \times S \\ \mathcal{B}(u(x,\boldsymbol{\xi}), x, \boldsymbol{\xi}) &= 0, \quad \text{in } \mathbb{R} \times \partial D \times S \end{aligned} \tag{2}$$

where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$ is the characterized finite random variables and $S \subset \mathbb{R}^d$ is a set of possible realizations of $\boldsymbol{\xi}$.

B. POLYNOMIAL CHAOS EXPANSION

This subsection introduces PC expansion [7], [8], [9], [10], [11], [12], [13], a popular surrogate model to quantify uncertainty. The MC method has been widely used for UQ due to the guaranteed convergence and being easy to implement. However, a crucial disadvantage of the MC method is its slow convergence rate, which requires many simulations to obtain reliable output statistics. The number of simulations required to construct PC expansion is usually orders of magnitude smaller than that required by the MC method for the same accuracy.

PC expansion approximates the stochastic solution, $u(x, \boldsymbol{\xi})$, of (2) as a linear combination of orthogonal polynomials (i.e., PC basis) in the function space L_w^2 with the inner product and norm

$$\begin{aligned}\langle g, h\rangle_{L_w^2} &= \mathbb{E}[g(\boldsymbol{\xi})h(\boldsymbol{\xi})] = \int_S g(z)h(z)w(z)dz,\\ \|g\|_{L_w^2} &= \langle g, g\rangle_{L_w^2}^{1/2}.\end{aligned}$$

where $w : S \to \mathbb{R}$ is a joint PDF of $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$. If we assume the mutual independence of $\xi_1, \cdots, \xi_d$, then the PC basis functions, denoted by $\Phi_{\boldsymbol{i}}$ with the multi-index $\boldsymbol{i} = (i_1, \cdots, i_d)$, are constructed by the tensor product of univariate orthogonal polynomials in each dimension, i.e.,

$$\Phi_{\boldsymbol{i}}(\xi_1, \cdots, \xi_d) = \phi_{i_1}^1(\xi_1)\phi_{i_2}^2(\xi_2)\cdots\phi_{i_d}^d(\xi_d). \tag{3}$$

Here, each $\phi_{i_j}^j$ is the i_j-th order univariate orthogonal monic polynomial associated with the random variable ξ_j, for $j = 1, \cdots, d$. Univariate monic orthogonal polynomials are well-known for some distributions [7], [8], [9]. They are also uniquely constructed for general distributions by the Stieltjes procedure [67].

The N-th order PC expansion, denoted by $u_N(x, \boldsymbol{\xi})$, approximates the stochastic solution, $u(x, \boldsymbol{\xi})$, as

$$u(x, \boldsymbol{\xi}) \approx u_N(x, \boldsymbol{\xi}) = \sum_{|\boldsymbol{i}|\le N} \widehat{u}_{\boldsymbol{i}}(x)\Phi_{\boldsymbol{i}}(\boldsymbol{\xi}) \tag{4}$$

where $|\boldsymbol{i}| = i_1 + \cdots + i_d$. Each PC coefficient, $\widehat{u}_{\boldsymbol{i}}(x)$, is defined as

$$\widehat{u}_{\boldsymbol{i}}(x) = \frac{\langle u(x, \boldsymbol{\xi}), \Phi_{\boldsymbol{i}}(\boldsymbol{\xi})\rangle_{L_w^2}}{\langle \Phi_{\boldsymbol{i}}(\boldsymbol{\xi}), \Phi_{\boldsymbol{i}}(\boldsymbol{\xi})\rangle_{L_w^2}} = \frac{\int_S u(x, z)\Phi_{\boldsymbol{i}}(z)w(z)dz}{\int_S \Phi_{\boldsymbol{i}}(z)^2 w(z)dz} \tag{5}$$

for each $\boldsymbol{i}$, so that the PC expansion, u_N, is the optimal approximation of u in the sense that

$$\|u - u_N\|_{L_w^2} = \inf_{P\in\mathbb{P}_N} \|u - P\|_{L_w^2}$$

for any $x \in D$, where $\mathbb{P}_N$ is the subspace of L_w^2 spanned by $\{\Phi_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N}$. Moreover, it is known that the PC expansion u_N has spectral convergence if the solution $u \in L_w$ is smooth enough [8].

It is possible to compute various quantities of interest effectively using the PC approximation. For example, the mean and variance of $u(x, \boldsymbol{\xi})$ are easily approximated by exploiting the orthogonality of the PC basis functions:

$$\begin{aligned}\mathbb{E}[u](x, \boldsymbol{\xi}) &\approx \mathbb{E}[u_N](x, \boldsymbol{\xi})\\ &= \sum_{|\boldsymbol{i}|\le N} \widehat{u}_{\boldsymbol{i}}(x)\mathbb{E}[\Phi_i(\boldsymbol{\xi})] = \widehat{u}_{\boldsymbol{0}}(x),\\ \mathrm{Var}[u](x, \boldsymbol{\xi}) &\approx \mathrm{Var}[u_N](x, \boldsymbol{\xi})\\ &= \mathbb{E}\Big[\sum_{|\boldsymbol{i}|\le N} \widehat{u}_{\boldsymbol{i}}(x)\Phi_{\boldsymbol{i}}(\boldsymbol{\xi}) - \widehat{u}_{\boldsymbol{0}}(x)\Big]^2\\ &= \sum_{\boldsymbol{i}\neq\boldsymbol{0}, |\boldsymbol{i}|\le N} \widehat{u}_{\boldsymbol{i}}(x)^2\mathbb{E}[\Phi_{\boldsymbol{i}}(\boldsymbol{\xi})^2].\end{aligned} \tag{6}$$

Let us now consider the numerical computation of each PC coefficient, $\widehat{u}_{\boldsymbol{i}}(x)$. It is infeasible to use definition (5) because the solution, $u(x, \boldsymbol{\xi})$, is unknown. A typical approach to obtaining PC coefficients is the stochastic Galerkin method. The basic idea of the method is to find the solution in the $\mathbb{P}_N$ so that the residual of the governing equation is orthogonal to $\mathbb{P}_N$. To be concrete, one finds the solution, $u_N(x, \boldsymbol{\xi})$, of the form (4) that satisfies

$$\begin{aligned}&\mathbb{E}[\mathcal{L}(u_N(x, \boldsymbol{\xi}), x, \boldsymbol{\xi})\Phi_{\boldsymbol{i}}(\boldsymbol{\xi})] = 0, \quad \text{in } \mathbb{R} \times D\\ &\mathbb{E}[\mathcal{B}(u_N(x, \boldsymbol{\xi}), x, \boldsymbol{\xi})\Phi_{\boldsymbol{i}}(\boldsymbol{\xi})] = 0, \quad \text{in } \mathbb{R} \times \partial D\end{aligned}$$

for each $\boldsymbol{i}$ with $|\boldsymbol{i}| \le N$. Note that these equations are deterministic BVPs for $\widehat{u}_{\boldsymbol{i}}$'s. Hence, we can use the existing numerical methods to solve them.

Another widely used method for computing PC coefficients is the stochastic collocation method. This method is based on solving the stochastic equation at some realizations of $\boldsymbol{\xi}$, called the collocation points and denoted by $\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(Q)}$. Subsequently, the following deterministic equations are induced

$$\begin{aligned}&\mathcal{L}(v_q(x), x, \boldsymbol{\xi}^{(q)}) = 0, \quad \text{in } \mathbb{R} \times D\\ &\mathcal{B}(v_q(x), x, \boldsymbol{\xi}^{(q)}) = 0, \quad \text{in } \mathbb{R} \times \partial D\end{aligned}$$

for $q = 1, \cdots, Q$, and the solutions, $v_1, \cdots, v_q$, are obtained using well-known numerical schemes. A major approach to computing the PC coefficients in the stochastic collocation method is a pseudo-spectral approach, which approximates the integral in (5) using the solutions $v_1, \cdots, v_q$. In this approach, the collocation points are usually chosen as the tensor product of 1D Gaussian quadrature point for each ξ_j. Due to the mutual independence of $\xi_1, \cdots, \xi_d$, all the numerical properties of the univariate integration scheme are retained in tensor product construction.

C. ADAPTIVE ANOVA METHOD

In this subsection, we introduce an adaptive ANOVA method [63], [64], [65], [66] to deal with high dimensions. Note that the PC method suffers from expensive computational costs in high-dimensional problems. The number of deterministic equations to be solved is $(N + d)!/N!d! = O(d^N)$ and $O(M^d)$ in the stochastic Galerkin and stochastic collocation method, respectively, where N is the order of PC expansion and M is the number of quadrature points in each dimension. The adaptive ANOVA method solves this problem by representing a high-dimensional stochastic model as a sum of low-dimensional stochastic models.

ANOVA decomposition [68] is widely used in statistics, and it allows us to evaluate which set of variables is more important in the total variance of the function. Considering a square integrable function $f(x)$ of $x = (x_1, \cdots, x_d)$ on its domain $S = S_1 \times \cdots \times S_d \subset \mathbb{R}^d$, the ANOVA decomposition of f is

$$\begin{aligned} f(x) &= \sum_{u \subset \{1,\cdots,d\}} f_u \\ &= f_\emptyset + \sum_{1 \le j_1 \le d} f_{j_1}(x_{j_1}) + \sum_{1 \le j_1 < j_2 \le d} f_{j_1,j_2}(x_{j_1}, x_{j_2}) \\ &\quad + \cdots + f_{1,\cdots,d}(x_1, \cdots, x_d) \end{aligned}$$

where each f_u only depends on $\{x_j\}_{j \in u}$, and satisfies

$$f_\emptyset = \int_S f(x) d\nu(x), \quad \int_{S_j} f_u(x) d\nu_j(x_j) = 0$$

for all $j \in u$ where $\nu = \nu_1 \times \cdots \times \nu_d$ is a product probability measure on S. Then, the decomposition satisfies the orthogonality of its terms:

$$\int_S f_u(x) f_v(x) d\nu(x) = 0$$

for $u \neq v$. The orthogonality induces the property that the variance of f is the sum of the variances of all the terms in the decomposition:

$$\mathrm{Var}[f] = \sum_{u \subset \{1,\cdots,d\}} \mathrm{Var}[f_u].$$

To compute the ANOVA decomposition, we first define the operator P_u for each $u \subset \{1, \cdots, d\}$ as

$$\begin{aligned} P_j[f](x) &:= \int f(x_1, \cdots, x_j, \cdots, x_d) d\nu_j(x_j), \\ P_u &:= \prod_{j \in u} P_j. \end{aligned}$$

Subsequently, each f_u is computed by the following inductive methods

$$\begin{aligned} f_u &= \sum_{v \subset u} (-1)^{|u|-|v|} P_{\{1,\cdots,d\} \setminus v}[f], \\ f_u &= P_{\{1,\cdots,d\} \setminus u}[f] - \sum_{v \subsetneq u} f_v. \end{aligned} \tag{7}$$

For each f_u in the decomposition, the size, u, is called the order of f_u. In many physical and engineering problems, it has been assumed that the effects of the low order terms in the ANOVA decomposition are dominant [69]. This leads to the following approximation:

$$f \approx \sum_{u \in \{1,\cdots,d\}, |u| \le q} f_u$$

for some cutoff dimension, q. Computing the PC expansion of this truncated ANOVA decomposition instead of the original function, f, reduces the computational cost. Note that the number of terms in the truncated ANOVA decomposition is $O(d^q)$, and each term has an input dimension at most q. Therefore, the total number of deterministic equations to be solved is $O(d^q q^N)$ and $O(d^q M^q)$ in the stochastic Galerkin and stochastic collocation method, which is much smaller than $O(d^N)$ and $O(M^d)$, respectively.

The truncated ANOVA decomposition has two computational problems. First, the computation of the terms includes high-dimensional integrations. To avoid this, the product probability measure, ν, can be chosen as a point measure at some anchor, $c \in \mathbb{R}^d$ [63], [64], [66]. This choice transforms high-dimensional integration into a one-point evaluation. Another problem is that the truncated ANOVA decomposition still has many terms for a large d. In this case, we further reduce the terms by ignoring those having small variances. This is the adaptive ANOVA method, and its algorithm is summarized in Algorithm 1, proposed in [66].

Algorithm 1 Adaptive ANOVA Decomposition

Input: d-variate function f, threshold $s > 0$, cutoff dimension q

```
Compute $f_\emptyset$
for all $j = 1, \cdots, q$ do
    $S_j = \{u \in \{1, \cdots, d\} \mid |u| = j\}$
for all $j = 1, \cdots, q$ do
    for all $u \in S_j$ do
        Compute $f_u$ in (7)
    for all $u \in S_j$ do
        Compute
            $\theta_u = \frac{\mathrm{Var}[f_u]}{\sum_{l=1}^{j} \sum_{u \in S_l} \mathrm{Var}[f_u]}$
        if $\theta_u < s$ then
            for all $l = j + 1, \cdots, q$ do
                Delete $v$ containing $u$ in $S_l$
```

D. ARBITRARY POLYNOMIAL CHAOS

This subsection describes the arbitrary PC method [21], [22], [23], [24] for data-driven construction of the PC basis. To construct the PC basis, the arbitrary PC method only requires the data of random variables, and does not need the information about PDF of the random variables. This is based on the relationship between the statistical moments and the coefficients of the PC basis. This relationship is directly induced by the orthogonality of the PC basis. Let ζ be a random variable and denote the i-th order PC basis for ζ by

$$\psi_i(\zeta) = \sum_{k=0}^{i} p_{ik} \zeta^k.$$

Note that the N-th order PC basis, ψ_N, is orthogonal to every PC basis of degree less than N, and this implies that ψ_N is orthogonal to every monomial of degree less than N. This induces the following N equations:

$$\mathbb{E}\Big[\sum_{k=0}^{N} p_{Nk}\zeta^k\Big] = 0, \quad \mathbb{E}\Big[\sum_{k=0}^{N} p_{Nk}\zeta^{k+1}\Big] = 0, \quad \cdots, \quad \mathbb{E}\Big[\sum_{k=0}^{N} p_{Nk}\zeta^{k+N-1}\Big] = 0.$$

Because there are N equations of $N+1$ undetermined values, we add one more by assuming ψ_N is a monic polynomial. That is, we set $p_{NN} = 1$ to find those values. Using the notation $\mu_k = \mathbb{E}[\zeta^k]$, we obtain the following linear equation:

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_N \\ \mu_1 & \mu_N & \cdots & \mu_{N+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{N-1} & \mu_N & \cdots & \mu_{2N-1} \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} p_{N0} \\ \vdots \\ p_{NN} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{8}$$

It is proved that the square matrix in (8) is not singular under mild conditions [21]. Each μ_k can be directly computed by

$$\mu_k = \frac{1}{M}\sum_{m=1}^{M} (\zeta^{(m)})^k$$

where $\zeta^{(1)}, \cdots, \zeta^{(M)}$ are samples of the ζ.

III. DATA-DRIVEN SURROGATE MODEL OF CORRELATED RANDOM VARIABLES

In this section, we introduce a method for constructing a data-driven surrogate model of correlated random variables. The main idea is to convert the correlated random variables, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$, into independent random variables, $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_d)$, using an appropriate data-driven invertible transformation, T. If the transformation is available, the stochastic solution, $u(x, \boldsymbol{\xi})$, of (2) can be represented as a PC expansion of $\boldsymbol{\zeta}$:

$$u(x, \boldsymbol{\xi}) = u(x, T^{-1}(\boldsymbol{\zeta})) =: v(x, \boldsymbol{\zeta}) \approx \sum_{|\boldsymbol{i}| \le N} \widehat{v}_{\boldsymbol{i}}(x)\Psi_{\boldsymbol{i}}(\boldsymbol{\zeta}) \tag{9}$$

where the PC basis function $\Psi_{\boldsymbol{i}}(\boldsymbol{\zeta})$ is constructed using a tensor product approach as in (3). The coefficient $\widehat{v}_{\boldsymbol{i}}$ can be obtained by stochastic Galerkin or stochastic collocation method. By substituting $\boldsymbol{\zeta}$ with $T(\boldsymbol{\xi})$, an approximate solution can be obtained:

$$u(x, \boldsymbol{\xi}) \approx \sum_{|\boldsymbol{i}| \le N} \widehat{v}_{\boldsymbol{i}}(x)\Psi_{\boldsymbol{i}}(T(\boldsymbol{\xi})). \tag{10}$$

We employ SVD to construct an invertible transformation T in a data-driven manner. Two commonly used transformations for T are the Rosenblatt and Nataf transformations, which rely on knowledge of the CDF and PDF of the random variables $\boldsymbol{\xi}$. However, closed forms of CDFs and PDFs of the random variables are rarely given in real applications, which makes the above two transformations less practical. In contrast, our approach using SVD does not require prior knowledge of the CDF and PDF of $\boldsymbol{\xi}$. It can even be used when the random variables have an arbitrary distribution, such as a discrete probability measure. The only requirement is the data of the random variables, making our approach practical and applicable in many real-world scenarios.

Turning our attention to the construction of the transformation T, suppose we have M data items of random variables, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$, denoted by $\Xi = \{\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(M)}\}$. We can assume that the random variables have a mean of zero because if their mean is nonzero, we can make them with zero mean by shifting. Let A be the $M \times d$ matrix, whose m-th row represents the component of $\boldsymbol{\xi}^{(m)}$, and the j-th row represents the data of ξ_j. That is,

$$A = \begin{bmatrix} \xi_1^{(1)} & \xi_2^{(1)} & \cdots & \xi_d^{(1)} \\ \xi_1^{(2)} & \xi_2^{(2)} & \cdots & \xi_d^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \xi_1^{(M)} & \xi_2^{(M)} & \cdots & \xi_d^{(M)} \end{bmatrix}. \tag{11}$$

We then consider SVD of the matrix A,

$$A = U\Sigma V^T = \sum_{s=1}^{r} \sigma_s U_s V_s^T \tag{12}$$

where $U \in \mathbb{R}^{M\times M}$, $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, $\Sigma \in \mathbb{R}^{M\times d}$ is a rectangular diagonal matrix with diagonal entries $\sigma_1 \ge \cdots \ge \sigma_d \ge 0$, and U_s, V_s are the s-th column vectors of matrices U, V, respectively. Because we only need first r columns to represent the SVD, we assume that U, Σ, and V refer to the matrices of size $M \times r, r \times r, d \times r$, respectively, by discarding unnecessary columns. Equation (12) represents the data of each random variable, ξ_j, as a linear combination of the vectors, $U_1, \cdots, U_r$. If we regard elements in U_j as generated samples of some random variable, ζ_j, for $j = 1, \cdots, d$, then $\boldsymbol{\xi}$ and $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_r)$ have the following relationship:

$$\begin{aligned} \boldsymbol{\xi} &= \begin{bmatrix}\xi_1 \cdots \xi_d\end{bmatrix} = \sum_{s=1}^{r} \sigma_s \zeta_s V_s^T \\ &= \begin{bmatrix}\zeta_1 \cdots \zeta_r\end{bmatrix} \Sigma V^T =: T^{-1}(\boldsymbol{\zeta}), \\ \boldsymbol{\zeta} &= \begin{bmatrix}\zeta_1 \cdots \zeta_r\end{bmatrix} = \begin{bmatrix}\zeta_1 \cdots \zeta_r\end{bmatrix} \Sigma V^T V \Sigma^{-1} \\ &= \begin{bmatrix}\xi_1 \cdots \xi_d\end{bmatrix} V \Sigma^{-1} =: T(\boldsymbol{\xi}). \end{aligned} \tag{13}$$

The mean of $\boldsymbol{\zeta}$ is zero because it is a linear transformation of $\boldsymbol{\xi}$, whose mean is assumed to be zero. Each covariance component of $\boldsymbol{\zeta}$ is

$$\mathbb{E}[(\zeta_j - \mathbb{E}\zeta_j)(\zeta_k - \mathbb{E}\zeta_k)] = \mathbb{E}[\zeta_j\zeta_k] \approx \frac{1}{M}\sum_{m=1}^{M} \zeta_j^{(m)}\zeta_k^{(m)} = \frac{1}{M}\delta_{jk}$$

where $\zeta_j^{(m)}$ and $\zeta_k^{(m)}$ are samples of ζ_j and ζ_k, identical to the entries of orthonormal column vectors U_j and U_k. Therefore,

Algorithm 2 Surrogate Model of Correlated Random Variables

```
Input: Data of random variables $\boldsymbol{\xi}$, simulator of an unknown stochastic process $u(x, \boldsymbol{\xi})$
Construct the matrix $A$ defined in (11)
From SVD $A = U\Sigma V^T$, identify the samples of $\boldsymbol{\zeta}$ and the invertible transformation $T$
Obtain the PC basis $\Psi_{\boldsymbol{i}}$, for $\boldsymbol{\zeta}$ by solving (8)
if $\boldsymbol{\zeta}$ is high-dimensional then
    Use the adaptive ANOVA decomposition to reduce the basis
Compute the PC coefficients $\widehat{v}_{\boldsymbol{i}}(x)$ in (9).
Substitute $\boldsymbol{\zeta}$ with $T(\boldsymbol{\xi})$ as in (10).
```

T is an invertible transformation converting correlated variables into uncorrelated random variables. For convenience, we assume T denotes $\sqrt{M}T$ instead of the definition (13) so that $\zeta_1, \cdots, \zeta_r$ are uncorrelated random variables with mean zero and unit variance. We assume these uncorrelated random variables to be mutually independent. This assumption has been adopted in some literature [64], [70], [71] for practical purposes and has been demonstrated to provide considerable accuracy. Under the assumption of mutual independence of $\boldsymbol{\zeta}$, it is possible to construct the PC basis, $\Psi_{\boldsymbol{i}}$, in (9) using the tensor product as in (3). To find the univariate PC basis for each ζ_j, we use the arbitrary PC method. The overall procedure is summarized in Algorithm 2.

Remark 1: Our approach allows us to tackle the curse of dimensionality by combining our method with the adaptive ANOVA method. If $\boldsymbol{\xi}$ is high-dimensional, it may result in a high-dimensional $\boldsymbol{\zeta}$, which makes it inefficient to simply apply the PC method to $\boldsymbol{\zeta}$. To overcome this challenge, we suggest applying the adaptive ANOVA method to reduce the number of basis and the cost of computing each PC coefficient. This approach enables us to effectively manage high-dimensional correlated uncertainties, without being overly burdened by computational demands.

IV. NUMERICAL EXAMPLES

In this section, we present several numerical examples to verify the effectiveness of the proposed method. First, we investigate the convergence of our surrogate model. Let $u(x, \boldsymbol{\xi})$ be a stochastic process of physical variable, $x \in D$, and characterized random variables, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$, with a joint PDF, $w : S \to \mathbb{R}$. Let $u_N(x, \boldsymbol{\xi})$ be the N-th order data-driven approximation using the dataset, $\Xi = \{\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(M)}\}$, including M data. To evaluate our surrogate model, we consider relative errors of u_N, $\mathbb{E}[u_N]$, $\mathrm{Var}[u_N]$, approximated by

$$
\begin{aligned}
&\frac{\mathbb{E}\|u - u_N\|_{L^2(D)}}{\mathbb{E}\|u\|_{L^2(D)}} \\
&\quad = \left(\frac{\int_S \int_D \left(u(x, z) - u_N(x, z)\right)^2 dx w(z) dz}{\int_S \int_D u(x, z)^2 dx w(z) dz} \right)^{1/2} \\
&\quad \approx \left(\frac{\sum_{l,m=1,1}^{M_D, M_S} \left(u(x^{(l)}, \boldsymbol{\xi}^{(m)}) - u_N(x^{(l)}, \boldsymbol{\xi}^{(m)})\right)^2}{\sum_{l=1,1}^{M_D, M_S} u(x^{(l)}, \boldsymbol{\xi}^{(m)})^2} \right)^{1/2}, \\
&\frac{\|\mathbb{E}[u] - \mathbb{E}[u_N]\|_{L^2(D)}}{\|\mathbb{E}[u]\|_{L^2(D)}} \\
&\quad = \left(\frac{\int_D \left(\mathbb{E}[u](x, \cdot) - \mathbb{E}[u_N](x, \cdot)\right)^2 dx}{\int_D \mathbb{E}[u](x, \cdot)^2 dx} \right)^{1/2} \\
&\quad \approx \left(\frac{\sum_{l=1}^{M_D} \left(\mathbb{E}u(x^{(l)}, \cdot) - \mathbb{E}[u_N](x^{(l)}, \cdot)\right)^2}{\sum_{l=1}^{M_D} \mathbb{E}[u](x^{(l)}, \cdot)^2} \right)^{1/2}, \\
&\frac{\|\mathrm{Var}[u] - \mathrm{Var}[u_N]\|_{L^2(D)}}{\|\mathrm{Var}[u]\|_{L^2(D)}} \\
&\quad = \left(\frac{\int_D \left(\mathrm{Var}[u](x, \cdot) - \mathrm{Var}[u_N](x, \cdot)\right)^2 dx}{\int_D \mathrm{Var}[u](x, \cdot)^2 dx} \right)^{1/2} \\
&\quad \approx \left(\frac{\sum_{l=1}^{M_D} \left(\mathrm{Var}[u](x^{(l)}, \cdot) - \mathrm{Var}[u_N](x^{(l)}, \cdot)\right)^2}{\sum_{l=1}^{M_D} \mathrm{Var}[u](x^{(l)}, \cdot)^2} \right)^{1/2},
\end{aligned}
$$

respectively, using M_D discretization points on D, denoted by $x^{(1)}, \cdots, x^{(M_D)}$, and $M_S = 90{,}000$ reference samples drawn from $\boldsymbol{\xi}$, denoted by $\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(M_S)}$. $\mathbb{E}[u]$ and $\mathrm{Var}[u]$ are computed by MC integration using the reference samples, i.e.,

$$
\begin{aligned}
\mathbb{E}[u](x, \cdot) &= \frac{1}{M_S} \sum_{m=1}^{M_S} u(x, \boldsymbol{\xi}^{(m)}), \\
\mathrm{Var}[u](x, \cdot) &= \frac{1}{M_S} \sum_{m=1}^{M_S} (u(x, \boldsymbol{\xi}^{(m)}) - \mathbb{E}[u](x, \cdot))^2.
\end{aligned}
$$

Meanwhile, $\mathbb{E}[u_N]$ and $\mathrm{Var}[u_N]$ are computed as in (6).

We further investigate the number of required simulations to construct surrogate models that yield mean and variance errors similar to those obtained by the MC method. The results are summarized in Table 1, which shows that our model requires significantly fewer simulations than the MC method to obtain errors of the same order. For example, for the three-dimensional stochastic BVP problem, our model achieves the same order of mean and variance error as the MC method of 1,000 simulations with 37 times fewer simulations. Each example in the table is discussed in more detail in the following subsections.

A. SMOOTH FUNCTION

We first investigate the convergence properties of our method by applying it to real-valued smooth functions of random variables, $\boldsymbol{\xi}$. This example will demonstrate the ability of our method to construct data-driven surrogate models that quantify correlated uncertainties. In addition, we illustrate how our approach significantly reduces the number of required simulations compared to the MC method when computing mean and variance.

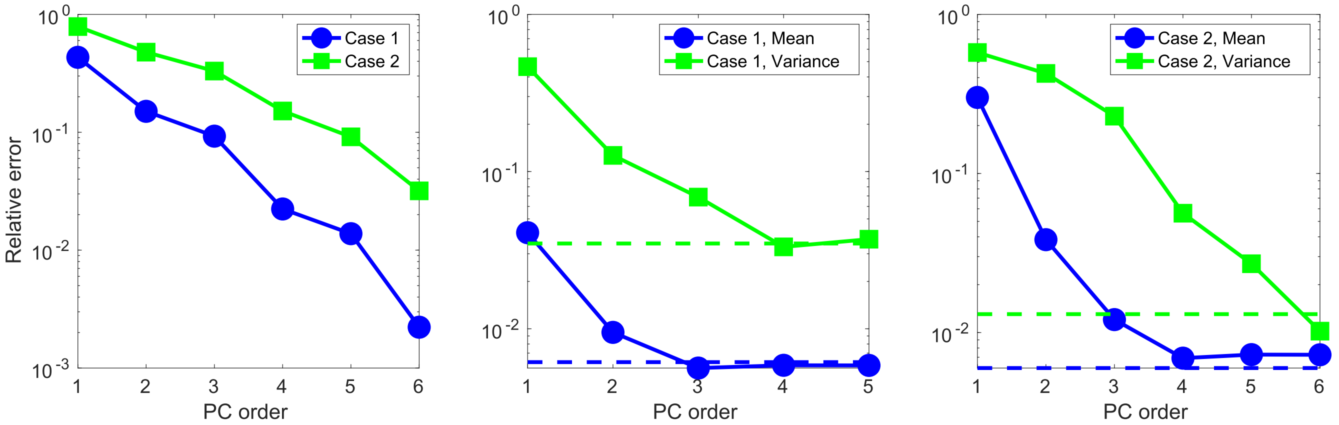

FIGURE 1. Errors of our surrogate models for oscillatory Genz functions. (Left) Relative errors of u_N for Case 1 and 2. The errors show spectral convergence. (Middle) Relative errors of mean and variance for Case 1. The 4th order PC expansion is required for both errors to reach the dashed line. (Right) Relative errors of mean and variance for Case 2. The 6th order PC expansion is required for both errors to reach the dashed line. (Dashed lines) Relative errors of mean and variance obtained by the MC method with 10,000 samples in the dataset, Ξ. This gives error limits of our data-driven model arising from Ξ.

TABLE 1. The number of simulations required using MC and our method for each numerical example.

Example	Required number of simulation (relative error of mean / variance)			
	MC		Our method	
Function, Case 1	1,000	(3.16% / 6.83%)	**16**	(3.20% / 6.48%)
Function, Case 1	10,000	(0.61% / 3.49%)	**25**	(0.59% / 3.32%)
Function, Case 2	1,000	(2.60% / 3.72%)	**25**	(3.23% / 2.82%)
Function, Case 2	10,000	(0.60% / 1.31%)	**49**	(0.73% / 1.02%)
Damped pendulum	1,000	(1.87% / 4.33%)	**64**	(2.18% / 5.08%)
Damped pendulum	10,000	(0.80% / 1.43%)	**81**	(0.64% / 1.48%)
Chemical reaction	1,000	(2.05% / 9.31%)	**9**	(1.97% / 7.75%)
Chemical reaction	10,000	(0.07% / 1.41%)	**25**	(0.07% / 0.77%)
SPDE, $d = 3$	1,000	(1.03% / 9.37%)	**27**	(0.71% / 7.03%)
SPDE, $d = 3$	10,000	(0.30% / 0.66%)	**125**	(0.25% / 0.60%)
SPDE, $d = 45$	1,000	(1.85% / 4.82%)	**190**	(1.66% / 4.08%)
SPDE, $d = 45$	10,000	(0.35% / 1.64%)	**2,101**	(0.26% / 1.45%)

To observe the effect of function oscillation on convergence, we consider the oscillatory Genz function [72]:

$$u(\boldsymbol{\xi}) = \cos(2\pi b + \sum_{i=1}^{d} a_i \xi_i)$$

where a_i and b are constants, and $d = 2$. In this case, u only depends on random variables, and the physical variable is not considered. Note that the Genz function exhibits more oscillation when the constants, a_i, come from a wide range. We consider two cases of Genz functions where the a_i's come from the intervals [0,1] and [0,1.5], referred to as Case 1 and 2, respectively. The value of b, which does not significantly affect the difficulty of integration, is selected randomly in the range [0,1].

To create a correlated probability distribution of the random variables $\boldsymbol{\xi}$, we define a Gaussian mixture distribution defined as

$$\sum_{i=1}^{2} w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{L}_i \boldsymbol{L}_i^T + \boldsymbol{I})$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^d$, $\boldsymbol{L}_i \in \mathbb{R}^{d\times d}$ is a lower triangular matrix, and $\boldsymbol{I} \in \mathbb{R}^d$ is an identity matrix. The constants, w_i, are randomly chosen so that their sum is one, and the entries in $\boldsymbol{\mu}_i$ and $\boldsymbol{L}_i$ are randomly drawn from uniform distributions on [0, 1], [−1, 1]. It should be noted that we define the PDF to produce a random sample of $\boldsymbol{\xi}$, and our approach does not require the exact knowledge of the PDF in practice. The number of data is set to $M = 10{,}000$. We employ the pseudo-spectral approach to find coefficients of the N-th order PC model for the transformed variables, $\boldsymbol{\zeta}$, for $N = 1, \cdots, 6$. Collocation points are constructed by the tensor product of $N+1$ quadrature points in each dimension, as obtained by the Golub-Welsch formula [73].

The results of our method are shown in Fig. 1. In the left graph, the accuracy of the approximation, u_N, improves as PC order increases for both cases of Genz functions. Since the Genz function in Case 2 has more oscillation than in Case 1, the approximation in Case 2 requires a higher PC order for the same error. This shows the spectral convergence of our surrogate model, which demonstrates that the proposed method extends the PC method to situations where only data of correlated random variables are available.

In the middle and right graphs of Fig. 1, the accuracies of approximate mean and variance, $\mathbb{E}[u_N]$ and $\mathrm{Var}[u_N]$, respectively, also improve as the PC order increases until they reach the dashed lines. The dashed lines in the middle and right graphs signify the relative errors of mean and variance obtained by the MC method with $M = 10,000$ samples in the dataset, Ξ. Because our data-driven model originates from the dataset, the dashed lines give the error limits of our data-driven model. For both mean and variance errors to reach the dashed lines, the 4th and 6th order PC expansions are respectively required in Case 1 and 2. Note that $(N+1)^d$ function evaluations are needed to construct the N-th order PC expansion. This means that our method needs only 25 and 49 function evaluations to achieve the same accuracy as the MC method with 10,000 function evaluations in Case 1 and 2, respectively. This shows the computational efficiency of our method, which requires significantly fewer simulations compared to the MC method when computing the mean and variance.

B. DAMPED PENDULUM

This section aims to highlight the ability of our model to accurately capture solutions that are highly dependent on the model parameters. In addition, we investigate the impact of the number of data in a given dataset, Ξ, on the convergence of our model. For this purpose, we consider a nonlinear stochastic ordinary differential equation (SODE) system of a damped pendulum described as

$$\begin{aligned} \frac{d\theta}{dt} &= \omega \\ \frac{d\omega}{dt} &= -\xi_1^2\omega - \xi_2^2 \sin\theta, \quad \begin{bmatrix}\theta(0)\\ \omega(0)\end{bmatrix} = \begin{bmatrix}0\\ 4\end{bmatrix}, \quad 0 \le t \le 5. \end{aligned} \tag{14}$$

The solution of this system exhibits oscillatory behavior; therefore it is sensitive to changes in parameters. The correlated random parameters $\boldsymbol{\xi} = (\xi_1, \xi_2)$ follow another Gaussian mixture distribution defined as

$$\sum_{i=1}^{4} w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{L}_i\boldsymbol{L}_i^T + 0.1\boldsymbol{I}) \tag{15}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{L}_i \in \mathbb{R}^{d\times d}$ are the lower triangular matrices. Entries in $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{L}_i$ are randomly drawn from uniform distributions on the intervals $[0.5, 1.5]$, $[5, 6]$, and $[-0.2, 0.2]$, respectively, for $i = 1, \cdots, 4$. Parameters ξ_1 and ξ_2 are affected by gravitational acceleration, the length of the string, and the drag coefficient. We square ξ_1 and ξ_2 to constrain the coefficients of ω and $\sin\theta$ to negative values. We use the pseudo-spectral approach to determine the PC coefficients and a time integrator at each collocation point is the 4th order Runge-Kutta method with a step size of 0.1.

Fig. 2 illustrates the impact of the number of data, M, on the convergence of our model. The left graph shows the errors of the surrogate models for $M = 1{,}000$ and $M = 10{,}000$. We can observe that the slope of spectral convergence is steeper for $M = 10{,}000$ than for $M = 1{,}000$ because both the basis and coefficients of the PC model are more accurately computed. On the other hand, in the middle and right graphs, the errors for the mean and variance of the solution decrease as the PC order increases until they reach the dashed line. Since the error of the MC method is lower for $M = 10{,}000$ than for $M = 1{,}000$, achieving the same order of error as the MC method for $M = 10{,}000$ requires a higher order and thus more simulations to generate a surrogate model. Therefore, we can conclude that larger datasets lead to more accurate surrogate models, but more simulations are necessary to achieve the same order of accuracy as the MC method.

Next, we demonstrate that our model precisely and computationally efficiently estimates the mean and variance of the solution. Due to the oscillatory pattern of the solution, it requires higher order than Section IV-A for the error to reach the dashed line. To be precise, for both mean and variance errors to reach the dashed lines, the 7th and 8th order PC expansions are respectively required for $M = 1{,}000$ and $M = 10{,}000$. This means that to obtain the same level of accuracy as the MC method using 1,000 and 10,000 equations, only $(7+1)^2 = 64$ and $(8+1)^2 = 81$ deterministic equations need to be solved, respectively. The mean and variance of the 8th order PC approximate solution trajectories using $M = 10{,}000$ data are depicted in Fig. 3. This shows that our surrogate model adequately captures the solution's oscillatory behavior, despite its strong dependence on random parameters.

C. CHEMICAL REACTION MODEL

In this example, we emphasize the practical applicability of our method by applying it to a chemical reaction model that describes competing species absorbing onto a surface from a gas phase [37], [74]. We also compare our approach to existing methods that handle correlated random variables, to demonstrate the superiority of our method. The model is described by the following ODE system with correlated random parameters α and β:

$$\begin{aligned} \frac{du_1}{dt} &= \alpha s - \gamma u_1 - 4u_1u_2 \\ \frac{du_2}{dt} &= 2\beta s^2 - 4u_1u_2 \\ \frac{du_3}{dt} &= \lambda s - \eta u_3 \end{aligned} \tag{16}$$

where u_i represents the fraction of adsorption sites occupied by species i, and the vacant site fraction is given by $s = 1-u_1-u_2-u_3$. The correlated random parameters are defined by $\alpha = 2 + 2\xi_1/3$ and $\beta = 20 + 15\xi_2/4$ for the proper scaling, where $\boldsymbol{\xi} = (\xi_1, \xi_2)$ follows the Gaussian mixture distribution described in (15), but entries in $\boldsymbol{\mu}_i$ and $\boldsymbol{L}_i$ are randomly drawn from uniform distribution on the interval $[-0.5, 0.5]$. The other parameters values used to construct the example are $\gamma = 0.04$, $\lambda = 0.36$, and $\eta = 0.016$. We focus on approximating the predictive distribution of the mass fraction of the second species u_2 at time $t = 50$ where the initial condition is $(u_1(0), u_2(0), u_3(0)) = (0.3, 0.3, 0.3)$.

We will compare our method to two existing methods that quantify correlated random uncertainties. To ensure a fair comparison, we assume that all methods are provided with the same dataset, Ξ, including $M = 10{,}000$ data. The method proposed in [34] uses the Gram-Schmidt approach to construct the orthogonal PC basis of the surrogate model. Pseudo-spectral method is used to compute PC coefficients and the quadrature points and weights are constructed by using weighted complete-linkage (WCL) clustering and block coordinate descent (BCD) solver. Another method is proposed in [37], which also uses the Gram-Schmidt approach to construct the PC basis, but PC coefficients are

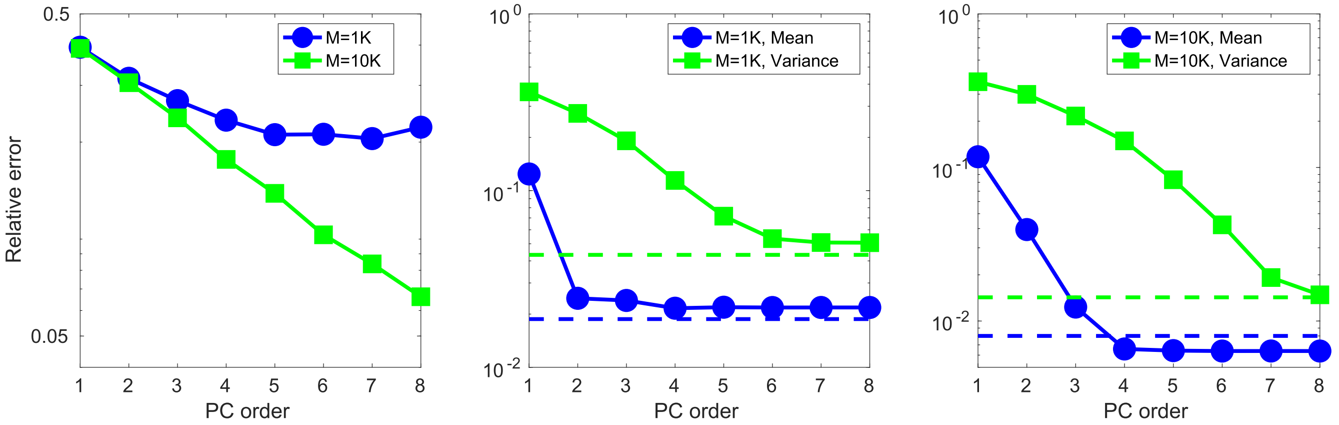

FIGURE 2. Errors of approximate solution to the SODE system (14). (Left) Relative error of u_N. The slope of spectral convergence is steeper for $M = 10{,}000$ than for $M = 1{,}000$ because both the basis and coefficients of the PC model are more accurately computed. (Middle) Relative errors of the solution mean and variance for $M = 1{,}000$. The 7th order PC expansion is required for both errors to reach the dashed line. (Right) Relative errors of the solution mean and variance for $M = 10{,}000$. The 8th order PC expansion is required for both errors to reach the dashed line. (Dashed lines) Relative errors of mean and variance obtained by MC method using the dataset, Ξ. This gives error limits of our data-driven model arising from Ξ.

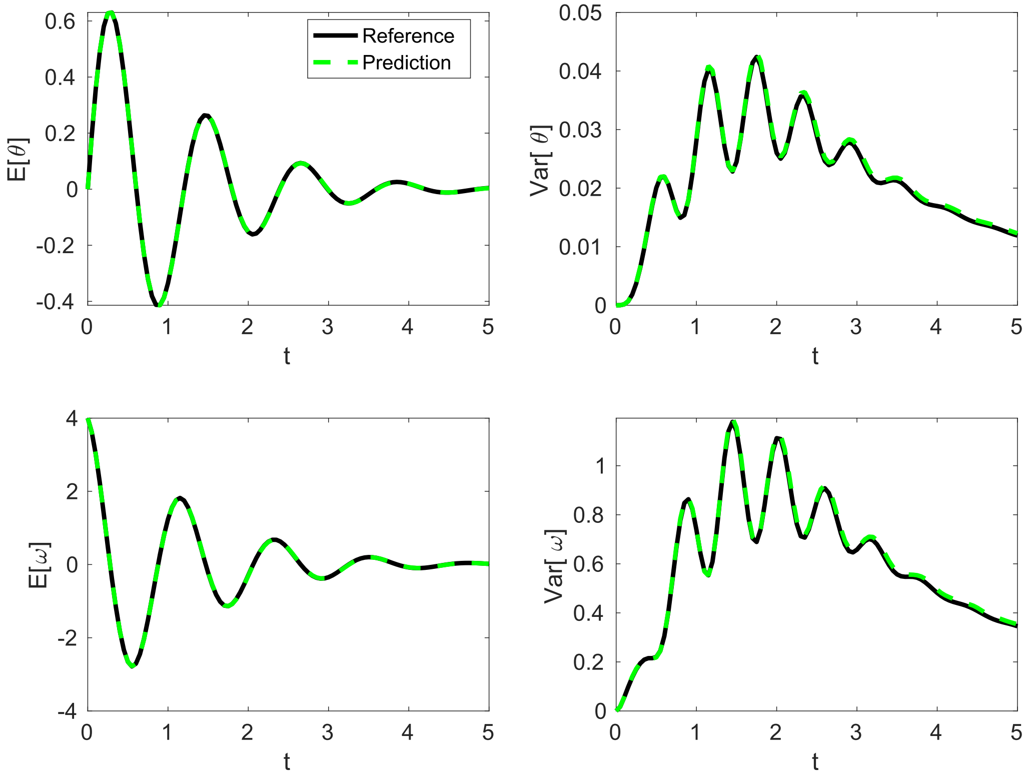

FIGURE 3. Mean and variance of the reference and the 8th order PC approximate solution trajectories to the SODE system (14). Our surrogate model adequately captures the oscillatory behavior, despite its strong dependence on random parameters.

obtained by interpolation approach [8] on the Leja sequence (LS) generated from the given dataset.

Fig. 4 illustrates the relationship between the number of required simulations for constructing the surrogate models and the error of mean and variance of the surrogate models. Our method outperforms the other two methods by reaching the dashed line in only 25 simulations. The dashed line represents the error achieved by the MC method using 10,000 data in the given dataset. Therefore, our approach achives the same accuracy as the MC method of 10,000 simulations using only 25 simulations, which demonstrates the computational efficiency of our method in practical applications. Although the combination of WCL and BCD also reaches the dashed line with much less simulations than the MC method, it requires 42 simulations, which is more than our method. On the other hand, the error of LS method does not eventually reach the

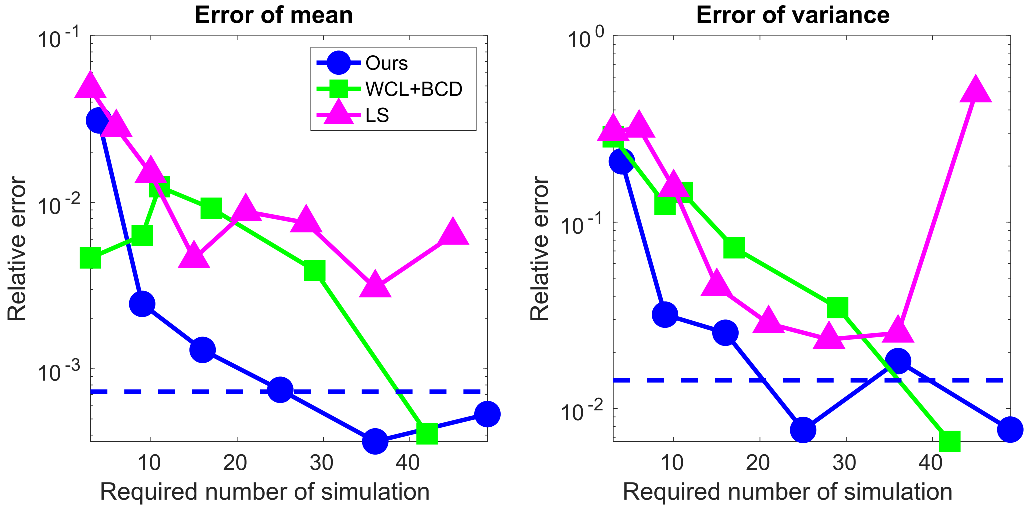

FIGURE 4. Errors of our method and the existing methods in the chemical reaction model. (Left) The relationship between the number of simulations required for each method and the relative error of the mean. (Right) The relationship between the number of simulations required for each method and the relative error of the variance. Our method outperforms the other two methods, achieving the same order of error as the MC method with only 25 simulations. (Dashed lines) Relative errors of mean and variance obtained by MC method using the dataset, Ξ. This gives error limits of our data-driven model arising from Ξ.

dashed line. These findings demonstrate the superiority of our method over the other two methods.

D. STOCHASTIC ELLIPTIC BOUNDARY VALUE PROBLEM

In our final example, we consider a stochastic elliptic BVP that has various practical applications such as electrical potential in conductive materials and flow of a fluid in porous media in oil and gas production. Furthermore, we will show that our method efficiently quantifies high-dimensional correlated uncertainty in a data-driven manner, which is the unique contribution of our method.

The stochastic elliptic BVP is described as

$$\begin{aligned} -\nabla \cdot (a(x,y,\omega)\nabla u(x,y,\omega)) &= f(x,y) \quad \text{in } D \times \Omega \\ u(x,y,\omega) &= 0 \quad \text{on } \partial D \times \Omega \end{aligned} \tag{17}$$

where $D = [-1, 1]^2$ and the random process a is characterized by correlated random variables $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_d)$. The random variables $\boldsymbol{\xi}$ are set to follow the Gaussian mixture distribution defined in (15), but entries in $\boldsymbol{\mu}_i$ are randomly drawn from a uniform distribution on the interval $[-0.2, 0.2]$ for each i. The random process a has the form of

$$\log a(x,y,\boldsymbol{\xi}) = \cos \pi x \cos \pi y + \sum_{k=1}^{d} \frac{1}{k^3} a_k(x,y)\xi_k^3$$

where a_k's are Fourier basis on $D = [-1, 1]^2$. The motivation behind cubing ξ_k's is to enhance the influence of the random variables on the resulting solution.

We first observe the convergence and computational efficiency of our method in the case of three-dimensional $\boldsymbol{\xi}$. We use the pseudo-spectral approach to determine the PC coefficients. The relative error of u_N for $d = 3$ in the left graph of Fig. 5 decreases as the PC order N increases, which shows the spectral convergence of our method. In the middle graph, the errors for the solution mean and variance also decrease until they reach the dashed lines. For both errors to reach the dashed lines, the 4th order PC expansion is required. This means that our method only needs to solve $(4 + 1)^3 = 125$ deterministic equations to achieve the same accuracy as the MC method using 10,000 equations, showing the computational efficiency of our approach. Fig. 6 shows that the 4th order surrogate model provides a good approximation to the mean and variance of the reference solution. These result demonstrate that our model works well in practical problems described by stochastic elliptic BVPs.

Next, we consider the case having a 45-dimensional $\boldsymbol{\xi}$ to describe the effectiveness of our model for handling high-dimensional correlated uncertainties. We exclude dimension reduction via SVD itself, which means the transformed variables $\boldsymbol{\zeta}$ is also 45-dimensional. To deal with the high dimensionality of $\boldsymbol{\zeta}$, we use the pseudo-spectral method combined with the adaptive ANOVA method described in Algorithm 1 to deal with high dimensionality. The anchor point is set to the mean of $\boldsymbol{\zeta}$ as proposed in [63], which is zero in our case. The stochastic solution is decomposed into one constant term, 45 one-dimensional terms, and 120 two-dimensional terms by the adaptive ANOVA method with a cutoff dimension $q = 2$ and a threshold $s = 0.003$. By obtaining the PC approximation of each term, we significantly reduce the amount of computation compared to obtaining the 45-dimensional PC approximation of the original stochastic solution.

The convergence results of our method are shown in Fig. 5. In the left graph, the accuracy of u_N for 45-dimensional $\boldsymbol{\xi}$ improves as PC order increases. The errors for the solution mean and variance in the right graph also decrease as the

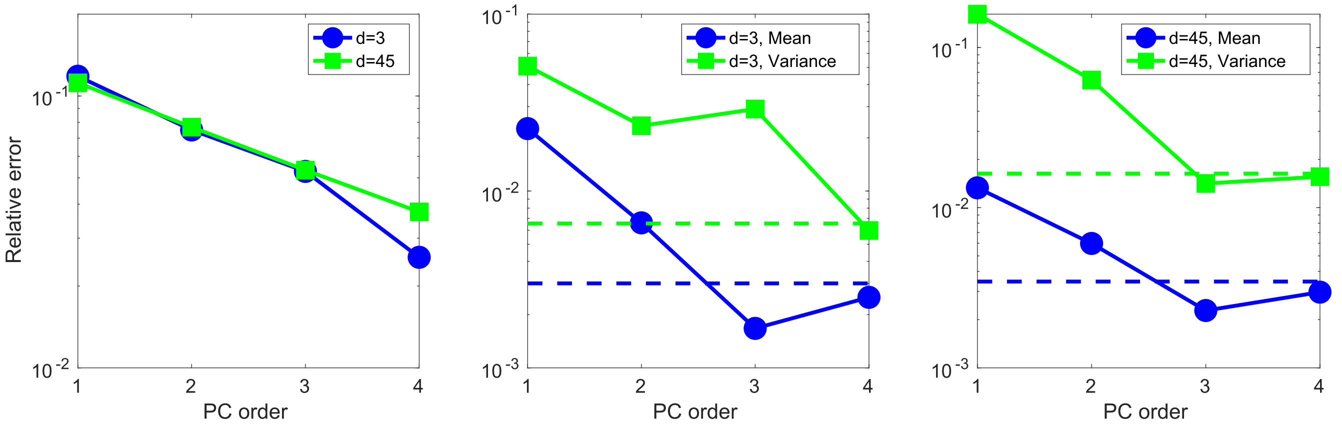

FIGURE 5. Errors of approximate solution to the BVP (17). (Left) Relative error of u_N for $d = 3, 45$. The errors are reduced as PC order increases and shows spectral convergence. (Middle) Relative errors of the solution mean and variance for $d = 3$. The 4th order PC expansion is required for both errors to reach the dashed line. (Right) Relative errors of the solution mean and variance for $d = 45$. The 3rd order PC expansion is required for both errors to reach the dashed line. (Dashed lines) Relative errors of mean and variance obtained by the MC method with 10,000 samples in the dataset, Ξ. This gives error limits of our data-driven model arising from Ξ.

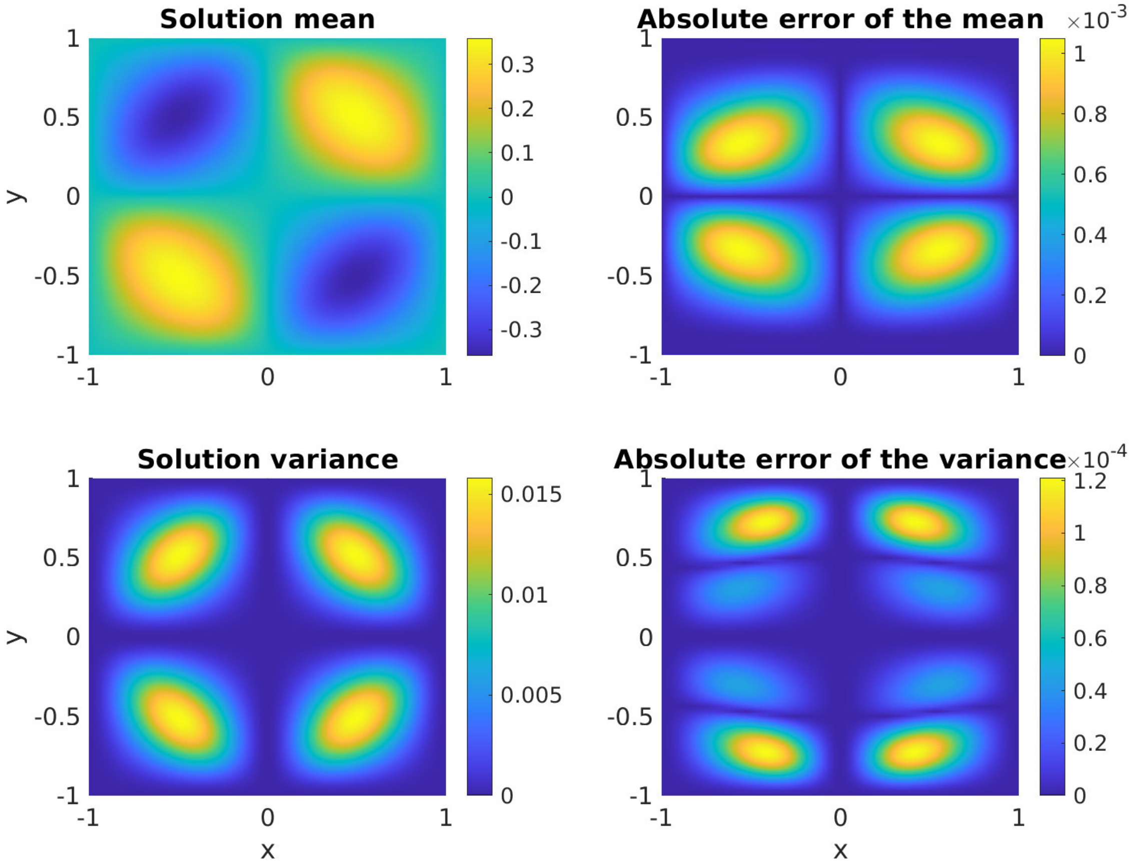

FIGURE 6. Mean and variance of the 4th order approximate solution to the BVP (17) with $d = 3$ and their absolute errors.

PC order increases until they reach the dashed lines. The 3rd order approximation is needed for both errors to reach the dashed lines, and this requires solving 2,101 deterministic equations using the adaptive ANOVA method. That is, our method solves only 2,101 equations and achieves the same accuracy as the MC method using 10,000 equations, which demonstrates the computational effectiveness of our method in high-dimensional problems. The results in Fig. 7 indicate that the mean and variance of the reference solution are accurately approximated by the 3rd order surrogate model.

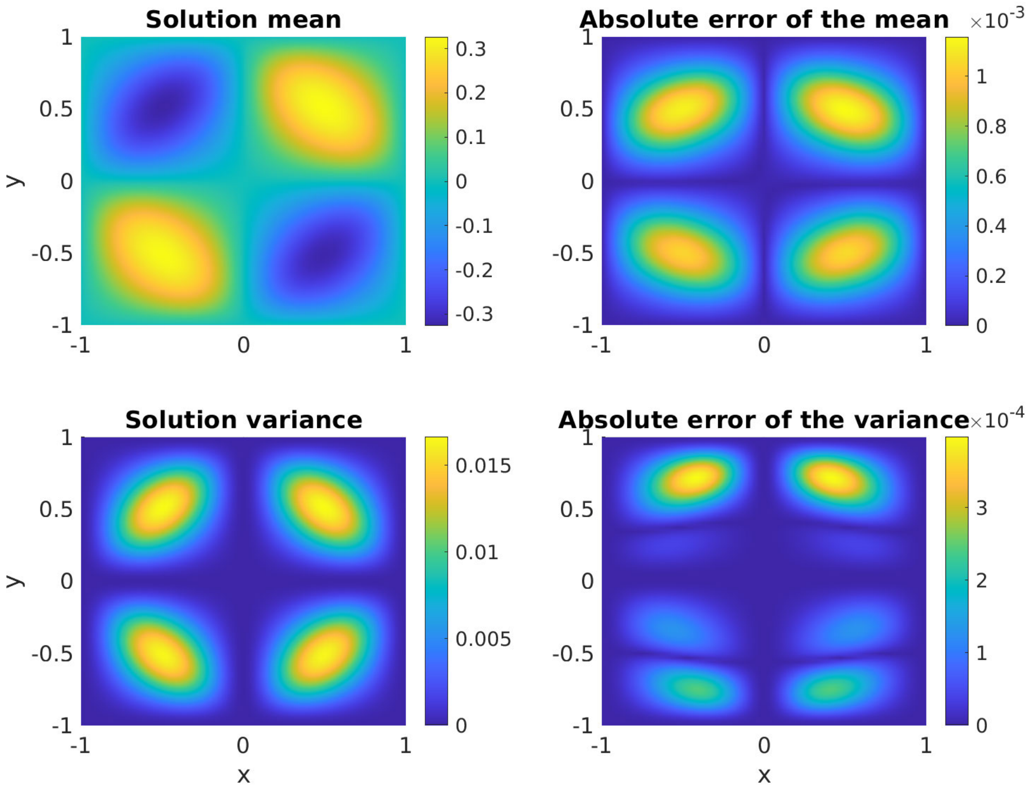

FIGURE 7. Mean and variance of the 3rd order approximate solution to the BVP (17) with $d = 45$ and their absolute errors.

This example shows the potential of our method to quantify high-dimensional correlated uncertainties.

V. CONCLUSION

We propose a new data-driven framework for dealing with stochastic models with correlated random variables. Our method is based on a transformation from correlated random variables to independent random variables. The Rosenblatt and Nataf transformations are not suitable for constructing a data-driven model because they require information about the distribution of random variables. Instead, we use SVD as the transformation, which provides a natural way to assimilate the data in building the transformation. We then create an orthogonal polynomial basis for transformed random variables using arbitrary PC. Our framework provides an additional benefit of dealing with high-dimensional correlated uncertainties by combining constructed PC basis with the adaptive ANOVA method. Numerical results show that our methods accurately propagate the moments of the states for both low and high dimensional stochastic systems with much smaller number of simulations compared to the MC method.

The theoretical convergence properties of our method are not covered in this paper. However, we recognize the importance of studying the theoretical part to improve the understanding and applicability of our method. We will study these theoretical properties in future work. We believe that these additional studies will contribute to the advancement of state-of-the-art models in various fields of computational science and engineering.

REFERENCES

[1] L. Martino and J. Read, “A joint introduction to Gaussian processes and relevance vector machines with connections to Kalman filtering and other kernel smoothers,” *Inf. Fusion*, vol. 74, pp. 17–38, Oct. 2021.

[2] J. Q. Candela and L. K. Hansen, “Learning with uncertainty-Gaussian processes and relevance vector machines,” Tech. Univ. Denmark, Copenhagen, Denmark, Tech. Rep. 2004, pp. 1–152. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9964090

[3] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[4] G. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*. Berlin, Germany: Springer, 2013.

[5] W.-L. Loh, “On Latin hypercube sampling,” *Ann. Statist.*, vol. 24, no. 5, pp. 2058–2080, 1996.

[6] M. Stein, “Large sample properties of simulations using Latin hypercube sampling,” *Technometrics*, vol. 29, no. 2, pp. 143–151, May 1987.

[7] D. Xiu, “Fast numerical methods for stochastic computations: A review,” *Commun. Comput. Phys.*, vol. 5, nos. 2–4, pp. 242–272, 2009.

[8] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

[9] D. Xiu and G. E. Karniadakis, “The wiener-askey polynomial chaos for stochastic differential equations,” *SIAM J. Sci. Comput.*, vol. 24, no. 2, pp. 619–644, Jan. 2002.

[10] D. Xiu and G. Em Karniadakis, “Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos,” *Comput. Methods Appl. Mech. Eng.*, vol. 191, no. 43, pp. 4927–4948, Sep. 2002.

[11] M. Eldred, “Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design,” in *Proc. 50th AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn., Mater. Conf.*, May 2009, pp. 1–10, doi: 10.2514/6.2009-2274.

[12] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann, "On the convergence of generalized polynomial chaos expansions," *ESAIM, Math. Model. Numer. Anal.*, vol. 46, no. 2, pp. 317–339, Mar. 2012.

[13] B. Sudret, "Polynomial chaos expansions and stochastic finite element methods," *Risk Rel. Geotechnical Eng.*, pp. 265–300, 2014.

[14] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*. North Chelmsford, MA, USA: Courier Corporation, 2003.

[15] D. Xiu and G. E. Karniadakis, "Modeling uncertainty in flow simulations via generalized polynomial chaos," *J. Comput. Phys.*, vol. 187, no. 1, pp. 137–167, May 2003.

[16] D. Xiu and J. S. Hesthaven, "High-order collocation methods for differential equations with random inputs," *SIAM J. Scientific Comput.*, vol. 27, no. 3, pp. 1118–1139, Jan. 2005.

[17] D. Xiu, "Efficient collocational approach for parametric uncertainty analysis," *Commun. Comput. Phys.*, vol. 2, no. 2, pp. 293–309, Apr. 2007.

[18] I. Babuška, F. Nobile, and R. Tempone, "A stochastic collocation method for elliptic partial differential equations with random input data," *SIAM J. Numer. Anal.*, vol. 45, no. 3, pp. 1005–1034, Jan. 2007.

[19] F. Nobile, R. Tempone, and C. G. Webster, "An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data," *SIAM J. Numer. Anal.*, vol. 46, no. 5, pp. 2411–2442, Jan. 2008.

[20] A. Narayan and D. Xiu, "Stochastic collocation methods on unstructured grids in high dimensions via interpolation," *SIAM J. Scientific Comput.*, vol. 34, no. 3, pp. A1729–A1752, Jan. 2012.

[21] S. Oladyshkin and W. Nowak, "Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion," *Rel. Eng. Syst. Saf.*, vol. 106, pp. 179–190, Oct. 2012.

[22] S. Oladyshkin and W. Nowak, "Incomplete statistical information limits the utility of high-order polynomial chaos expansions," *Rel. Eng. Syst. Saf.*, vol. 169, pp. 137–148, Jan. 2018.

[23] M. Ashraf, S. Oladyshkin, and W. Nowak, "Geological storage of CO_2: Application, feasibility and efficiency of global sensitivity analysis and risk assessment using the arbitrary polynomial chaos," *Int. J. Greenhouse Gas Control*, vol. 19, pp. 704–719, Nov. 2013.

[24] I. Kröker and S. Oladyshkin, "Arbitrary multi-resolution multi-wavelet-based polynomial chaos expansion for data-driven uncertainty quantification," *Rel. Eng. Syst. Saf.*, vol. 222, Jun. 2022, Art. no. 108376.

[25] Y. Noh, K. K. Choi, and L. Du, "Reliability-based design optimization of problems with correlated input variables using a Gaussian copula," *Structural Multidisciplinary Optim.*, vol. 38, no. 1, pp. 1–16, Mar. 2009.

[26] D. Li, Y. Chen, W. Lu, and C. Zhou, "Stochastic response surface method for reliability analysis of rock slopes involving correlated non-normal variables," *Comput. Geotechnics*, vol. 38, no. 1, pp. 58–68, Jan. 2011.

[27] Y. Zhang and N. V. Sahinidis, "Uncertainty quantification in CO_2 sequestration using surrogate models from polynomial chaos expansion," *Ind. Eng. Chem. Res.*, vol. 52, no. 9, pp. 3121–3132, Mar. 2013.

[28] M. Abdelmalak and M. Benidris, "A polynomial chaos-based approach to quantify uncertainties of correlated renewable energy sources in voltage regulation," *IEEE Trans. Ind. Appl.*, vol. 57, no. 3, pp. 2089–2097, May 2021.

[29] J. Jakeman, M. Eldred, and D. Xiu, "Numerical approach for quantification of epistemic uncertainty," *J. Comput. Phys.*, vol. 229, no. 12, pp. 4648–4663, Jun. 2010.

[30] X. Chen, E.-J. Park, and D. Xiu, "A flexible numerical approach for quantification of epistemic uncertainty," *J. Comput. Phys.*, vol. 240, pp. 211–224, May 2013.

[31] C. Soize and R. Ghanem, "Physical systems with random uncertainties: Chaos representations with arbitrary probability measure," *SIAM J. Sci. Comput.*, vol. 26, no. 2, pp. 395–410, Jan. 2004.

[32] Q. Lin, F. Xiong, F. Wang, and X. Yang, "A data-driven polynomial chaos method considering correlated random variables," *Structural Multidisciplinary Optim.*, vol. 62, no. 4, pp. 2131–2147, Oct. 2020.

[33] J. A. S. Witteveen and H. Bijl, "Modeling arbitrary uncertainties using gram-schmidt polynomial chaos," in *Proc. 44th AIAA Aerosp. Sci. Meeting Exhib.*, Reno, Nevada, 2006, pp. 1–9.

[34] C. Cui and Z. Zhang, "Stochastic collocation with non-Gaussian correlated process variations: Theory, algorithms, and applications," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 9, no. 7, pp. 1362–1375, Jul. 2019.

[35] C. Cui and Z. Zhang, "High-dimensional uncertainty quantification of electronic and photonic IC with non-Gaussian correlated process variations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 8, pp. 1649–1661, Aug. 2020.

[36] J. Feinberg, V. G. Eck, and H. P. Langtangen, "Multivariate polynomial chaos expansions with dependent variables," *SIAM J. Scientific Comput.*, vol. 40, no. 1, pp. A199–A223, Jan. 2018.

[37] J. D. Jakeman, F. Franzelin, A. Narayan, M. Eldred, and D. Plfüger, "Polynomial chaos expansions for dependent random variables," *Comput. Methods Appl. Mech. Eng.*, vol. 351, pp. 643–666, Jul. 2019.

[38] S. Rahman, "A polynomial chaos expansion in dependent random variables," *J. Math. Anal. Appl.*, vol. 464, no. 1, pp. 749–775, Aug. 2018.

[39] S. Rahman, "Uncertainty quantification under dependent random variables by a generalized polynomial dimensional decomposition," *Comput. Methods Appl. Mech. Eng.*, vol. 344, pp. 910–937, Feb. 2019.

[40] D. Lee and S. Rahman, "Practical uncertainty quantification analysis involving statistically dependent random variables," *Appl. Math. Model.*, vol. 84, pp. 324–356, Aug. 2020.

[41] J. A. Paulson, E. A. Buehler, and A. Mesbah, "Arbitrary polynomial chaos for uncertainty propagation of correlated random variables in dynamic systems," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3548–3553, Jul. 2017.

[42] G. Wang, H. Xin, D. Wu, P. Ju, and X. Jiang, "Data-driven arbitrary polynomial chaos-based probabilistic load flow considering correlated uncertainties," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 3274–3276, Jul. 2019.

[43] A. Nataf, "Determination des distribution don't les marges sont donnees," *Comp. Rendus de l'Academie des Sci.*, vol. 225, pp. 42–43, Aug. 1962. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7015608

[44] M. Rosenblatt, "Remarks on a multivariate transformation," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 470–472, Sep. 1952.

[45] V. Barthelmann, E. Novak, and K. Ritter, "High dimensional polynomial interpolation on sparse grids," *Adv. Comput. Math.*, vol. 12, no. 4, pp. 273–288, Mar. 2000.

[46] F. Nobile, R. Tempone, and C. G. Webster, "A sparse grid stochastic collocation method for partial differential equations with random input data," *SIAM J. Numer. Anal.*, vol. 46, no. 5, pp. 2309–2345, Jan. 2008.

[47] X. Ma and N. Zabaras, "An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations," *J. Comput. Phys.*, vol. 228, no. 8, pp. 3084–3113, May 2009.

[48] P. G. Constantine, M. S. Eldred, and E. T. Phipps, "Sparse pseudospectral approximation method," *Comput. Methods Appl. Mech. Eng.*, vols. 229–232, pp. 1–12, Jul. 2012.

[49] P. R. Conrad and Y. M. Marzouk, "Adaptive smolyak pseudospectral approximations," *SIAM J. Scientific Comput.*, vol. 35, no. 6, pp. A2643–A2670, Jan. 2013.

[50] G. Blatman and B. Sudret, "Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach," *Comp. Rendus Mécanique*, vol. 336, no. 6, pp. 518–523, Jun. 2008.

[51] H. Bazargan, M. Christie, A. H. Elsheikh, and M. Ahmadi, "Surrogate accelerated sampling of reservoir models with complex structures using sparse polynomial chaos expansion," *Adv. Water Resour.*, vol. 86, pp. 385–399, Dec. 2015.

[52] P. Diaz, A. Doostan, and J. Hampton, "Sparse polynomial chaos expansions via compressed sensing and D-optimal design," *Comput. Methods Appl. Mech. Eng.*, vol. 336, pp. 640–666, Jul. 2018.

[53] R. Baptista, V. Stolbunov, and P. B. Nair, "Some greedy algorithms for sparse polynomial chaos expansions," *J. Comput. Phys.*, vol. 387, pp. 303–325, Jun. 2019.

[54] N. Lüthen, S. Marelli, and B. Sudret, "Sparse polynomial chaos expansions: Literature survey and benchmark," *SIAM/ASA J. Uncertainty Quantification*, vol. 9, no. 2, pp. 593–649, Jan. 2021.

[55] G. Blatman and B. Sudret, "An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis," *Probabilistic Eng. Mech.*, vol. 25, no. 2, pp. 183–197, Apr. 2010.

[56] G. Blatman and B. Sudret, "Adaptive sparse polynomial chaos expansion based on least angle regression," *J. Comput. Phys.*, vol. 230, no. 6, pp. 2345–2367, Mar. 2011.

[57] A. Doostan and H. Owhadi, "A non-adapted sparse approximation of PDEs with stochastic inputs," *J. Comput. Phys.*, vol. 230, no. 8, pp. 3015–3034, Apr. 2011.

[58] L. Mathelin and K. A. Gallivan, "A compressed sensing approach for partial differential equations with random input data," *Commun. Comput. Phys.*, vol. 12, no. 4, pp. 919–954, Oct. 2012.

[59] X. Yang and G. E. Karniadakis, "Reweighted ℓ1-minimization method for stochastic elliptic differential equations," *J. Comput. Phys.*, vol. 248, pp. 87–108, Sep. 2013.

[60] J. Peng, J. Hampton, and A. Doostan, "A weighted ℓ1-minimization approach for sparse polynomial chaos expansions," *J. Comput. Phys.*, vol. 267, pp. 92–111, Jun. 2014.

[61] J. D. Jakeman, M. S. Eldred, and K. Sargsyan, "Enhancing ℓ1-minimization estimates of polynomial chaos expansions using basis selection," *J. Comput. Phys.*, vol. 289, pp. 18–34, May 2015.

[62] J. Hampton and A. Doostan, "Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies," *J. Comput. Phys.*, vol. 280, pp. 363–386, Jan. 2015.

[63] X. Ma and N. Zabaras, "An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations," *J. Comput. Phys.*, vol. 229, no. 10, pp. 3884–3915, May 2010.

[64] X. Yang, M. Choi, G. Lin, and G. E. Karniadakis, "Adaptive ANOVA decomposition of stochastic incompressible and compressible flows," *J. Comput. Phys.*, vol. 231, no. 4, pp. 1587–1614, Feb. 2012.

[65] Z. Zhang, M. Choi, and G. E. Karniadakis, "Error estimates for the ANOVA method with polynomial chaos interpolation: Tensor product functions," *SIAM J. Sci. Comput.*, vol. 34, no. 2, pp. A1165–A1186, Jan. 2012.

[66] Z. Zhang, X. Yang, I. V. Oseledets, G. E. Karniadakis, and L. Daniel, "Enabling high-dimensional hierarchical uncertainty quantification by ANOVA and tensor-train decomposition," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 1, pp. 63–76, Jan. 2015.

[67] W. Gautschi, "On generating orthogonal polynomials," *SIAM J. Sci. Stat. Comput.*, vol. 3, no. 3, pp. 289–317, 1982.

[68] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski, "On decompositions of multivariate functions," *Math. Comput.*, vol. 79, no. 270, pp. 953–966, Nov. 2009.

[69] H. Rabitz, Ö. F. Aliş, J. Shorter, and K. Shim, "Efficient input—Output model representations," *Comput. Phys. Commun.*, vol. 117, nos. 1–2, pp. 11–20, Mar. 1999.

[70] Z. Zhang, X. Hu, T. Y. Hou, G. Lin, and M. Yan, "An adaptive ANOVA-based data-driven stochastic method for elliptic PDEs with random coefficient," *Commun. Comput. Phys.*, vol. 16, no. 2, pp. 571–598, Aug. 2014.

[71] D. Ayres and M. D. Eaton, "Uncertainty quantification in nuclear criticality modelling using a high dimensional model representation," *Ann. Nucl. Energy*, vol. 80, pp. 379–402, Jun. 2015.

[72] A. Genz, "A package for testing multiple integration subroutines," in *Numerical Integration*. Cham, Switzerland: Springer, 1987, pp. 337–340.

[73] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Math. Comput.*, vol. 23, no. 106, pp. 221–230, 1969.

[74] R. D. Vigil and F. T. Willmore, "Oscillatory dynamics in a heterogeneous surface reaction: Breakdown of the mean-field approximation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 54, no. 2, pp. 1225–1231, Aug. 1996.

JEAHAN JUNG received the B.S. degree in mathematics from the Pohang University of Science and Technology, Pohang, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in mathematics. His current research interests include uncertainty quantification and physics-informed machine learning.

MINSEOK CHOI received the B.S. degree in mechanical engineering and mathematics and the M.S. degree in mechanical engineering from Seoul National University, Seoul, South Korea, in 2002 and 2007, respectively, and the Ph.D. degree in applied mathematics from Brown University, Providence, USA, in 2014. He was a Postdoctoral Researcher with Princeton University, USA, until 2017. He is currently an Assistant Professor in mathematics with the Pohang University of Science and Technology (POSTECH), South Korea. His research interests include physics-informed machine learning, uncertainty quantification, and related areas of applied mathematics.