A new method for stochastic analysis of structures under limited observations

Hongzhe Dai[a,*], Ruijing Zhang[a], Michael Beer[b,c,d]

[a]*School of Civil Engineering, Harbin Institute of Technology, Harbin 150090, China*

[b]*Institute for Risk and Reliability, Leibniz Universität Hannover, Callinstr. 34, Hannover, Germany*

[c]*Institute for Risk and Uncertainty and School of Engineering, University of Liverpool, Peach Street, Liverpool L69 7ZF, UK*

[d]*International Joint Research Center for Resilient Infrastructure & International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, 1239 Siping Road, Shanghai 200092, PR China*

(*Corresponding Author: Hongzhe Dai, E-mail: hzdai@hit.edu.cn)

Abstract

Reasonable modeling of non-Gaussian system inputs from limited observations and efficient propagation of system response are of great significance in uncertain analysis of real engineering problems. In this paper, we develop a new method for the construction of non-Gaussian random model and associated propagation of response under limited observations. Our method firstly develops a new kernel density estimation-based (KDE-based) random model based on Karhunen-Loeve (KL) expansion of observations of uncertain parameters. By further implementing the arbitrary polynomial chaos (aPC) formulation on KL vector with dependent measure, the associated aPC-based response propagation is then developed. In our method, the developed KDE-based model can accurately represent the input parameters from limited observations as the new KDE of KL vector can incorporate the inherent relation between marginals of input parameters and distribution of univariate KL variables. In addition, the aPC formulation can be effectively determined for uncertain analysis by virtue of the mixture representation of the developed KDE of KL vector. Furthermore, the system response can be propagated in a stable and accurate way with the developed D-optimal weighted regression method by the equivalence between the distribution of underlying aPC variables and that of KL vector. In this way, the current work provides an effective framework for the reasonable stochastic modeling and efficient response propagation of real-life engineering systems with limited observations. Two numerical examples, including the analysis of structures subjected to random seismic ground motion, are presented to highlight the effectiveness of the proposed method.

Keywords: Uncertain analysis; Random field modelling; PC-based response propagation; Limited observations; Kernel density estimation.

List of acronyms and abbreviations

aPC	arbitrary polynomial chaos	KDE	kernel density estimation
CDF	cumulative density function	KL	Karhunen-Loeve
DOF	degree of freedom	MCMC	Markov chain Monte Carlo
ED	experimental design	MCS	Monte Carlo simulation
IQR	interquartile range	PC	polynomial chaos
ISDE	Itô stochastic differential equation	PDF	probability density function

1. Introduction

In stochastic engineering problems, the proper consideration of uncertain input parameters is crucial to obtain an accurate and reliable solution [1-3]. Uncertain inputs are ubiquitous in engineering applications and include uncertainty in system parameters, material properties, source and interaction terms, boundary and initial conditions, etc [4-6]. A large number of practical problems involves uncertain input quantities with inherent spatio-temporal variability, and in such cases, random fields are commonly used for modelling spatial fluctuations as observed in various disciplines, for example, soil parameters and groundwater heights in geotechnical engineering, wind loads and earthquake excitations in structural engineering, and the amount of precipitation and evaporation in hydrology [7-12]. In real applications, it will often be the case that very few realizations are available regarding the uncertain input parameters, and only limited measurements can be obtained owing to limited storage capability of sensors or prohibitive cost in increasing observations, etc [13-16]. In this context, the Gaussian simplification is often made on the fields to empower their numerical simulation with any practical use owing to the fact that Gaussian fields are completely described by their second-order statistics. In fact, it has been evidenced by an ever-growing number of experimental databases that many physical phenomena are not Gaussian, and significant differences may arise in the estimation of system response if a Gaussian field is assumed. Although clearly more realistic in most instances, non-Gaussian models have had to contend with the scarcity of consistent mathematical theories for describing general infinite-dimensional probability measures [17-20]. More than ever, the goal then becomes to reasonably represent non-Gaussian input parameters from limited observations and to propagate the input uncertainty to satisfactorily quantify the effects on quantities of interest.

The problem of representing and propagating of non-Gaussian random inputs from the available observations to the desired results has attracted significant interest in the last decade. This research has spawned the development of two basic categories of methods. A first class of methods seeks to produce sample functions of the target non-Gaussian field according to its limited observations, and then to estimate random response of systems with Monte Carlo simulation (MCS). In this regard, Beer and Kougioumtzoglou reconstructed (spatio-)temporal non-Gaussian random model by recovering their (joint) power spectrum from limited measured data [11, 13, 21-23]. Wang *et al*. modeled uncertain input parameters from limited data by Karhunen-Loeve (KL) expansion in conjunction with Bayesian compressive sensing [24, 25]. While it provides an effective tool for reconstructing non-Gaussian fields through limited observations, this is the method of last resort since the attendant computational burden can be prohibitive for large-scale problems, and thereby rules out the method to be applicable in a wide range of engineering systems. As a promising alternative to sample-based method, the class of polynomial chaos-based (PC-based) methods has received increasing attention. The basic idea is firstly to synthesize the non-Gaussian field from limited data by KL expansion, and then to represent KL variables by PC expansion. By further using the well-established PC-based solution technique, the probabilistic information is efficiently propagated regarding the input parameters to the associated response of systems. The benefits of this class of method lies in the ability of PC expansion to characterize the non-Gaussian probabilistic behavior under limited measurements. In addition, as the capacity of PC expansion for the efficient propagation of uncertainty is naturally inherited, this class of method has the potential for addressing complex issues of general engineering interest.

While elegant, the utilization of the PC representation together with KL expansion poses a number of additional challenges in real applications. A first challenging issue is recognized as the reconstruction of PC-based model for faithfully representing non-Gaussian input parameters from limited data. This is because the joint probability density function (PDF) recovery of KL vector, which significantly affects the accuracy of input model, is quite challenging due to the nonlinear dependence of the non-Gaussian KL variables. Another aspect which deserves more attention is the fact that determination of the associated PC formulation is even further complicated, as a great number of high-dimensional integrals are involved in the multidimensional nonlinear transformation from KL variables to the underlying PC variables. In order to overcome these two difficulties, various efforts have been made in the last ten years. An early attempt is to pose the independent assumption on uncorrelated KL variables so that one-dimensional PC can be readily used to represent KL variables [26-29]. Although the above-mentioned two difficulties can be simultaneously circumvented, this scheme may lead to grossly inaccurate PC-based model of input field due to the ignorance of nonlinear dependence between KL variables. No wonder, the associated propagation of the system response would not have any significant meaning without an accurate input model. In order to capture the dependence of KL variables, the moment-constrained maximum entropy procedure was developed for estimating PDF of KL variable, Rosenblatt transformation was then employed for constructing the Hermite PC representation of KL variables [30]. However, since a large number of multivariate integrations have to be solved in both the maximization of entropy and Rosenblatt transformation, the computational cost of the method becomes intractable with respect to the number of KL variables. Although the histogram estimator was subsequently developed to convert the multivariate integrations in [30] to a set of univariate integrations by slicing the multivariate conditional PDFs in Rosenblatt transformation, the number of slices expands exponentially with the KL variables [31]. In fact, the integration scheme in [31] is equivalent to the tensor product quadrature to some extent, and thereby the method still suffers the curse of dimensionality. This is why the use of the methods in [30] and [31] has been limited to problems with low random dimensionalities. Very recently, [32] employed kernel density estimator (KDE) to recover PDF of KL variables, and then determined the PC coefficients of KL variables by MC integration, in which a new Ito stochastic differential equation (ISDE)-based sampler was developed to generate the MC integral points. Since KDE can be straightforwardly extended to high-dimensional cases without enormous computational burden, the curse of dimensionality encountered in above mentioned density estimators can be greatly alleviated [33]. Nevertheless, the use of multidimensional Silverman bandwidth in KDE inevitably results in an evident deviation in the estimation of marginal distribution of non-Gaussian KL variables, and thereby lead to an inappropriate PC-based input model. In addition, since the ISDE-based sampler is essentially a type of Markov Chain Monte Carlo (MCMC) sampler, the inherent deficiencies of MCMC, i.e., the autocorrelations of MCMC samples as well as the repeated evaluations of PDF of variables, severely decrease the efficiency for determining the PC formulation of KL variables [34].

Another significant challenge in the PC-based approach is the efficiency of the propagation of the response. This is clearly an important aspect, which affects the applicability of the method, i.e., its efficiency versus other types of propagation methods. It is acknowledged that the use of Hermite polynomials as PC bases may lead to optimum convergence for the Gaussian distributed input parameter. While for inputs with other common distribution types, Wiener-Askey orthogonal polynomials can also be used as PC bases to achieve the same convergence [35]. However, in the context of limited observations, the non-Gaussian KL variables may have broader distributions outside the Wiener–Askey family. In this case, the use of Wiener–Askey scheme may lead to a substantially slow convergence of PC expansion of system response, and the huge computational burden in response propagation prevents the application of the method in large-scale engineering problems. Therefore, the construction of proper PC basis is of crucial importance for the applicability of the method with respect to an optimal convergence purpose. On the other hand, given the optimal PC bases for uncertainty propagation, the determination of associated PC approximation of system response still remains to be a significant challenge due to the dependence of underlying PC variables. In fact, when the underlying PC variables are mutually independent, existing well-established methods can be readily used for response propagation, while there exists a dearth of algorithmic options for approximation of system response under dependent underlying PC variables [36, 37]. For this line of approach to be attractive in practice, two important objectives should be reached. Firstly, accurate construction of PC-based input model from limited observations should be able to achieved. This would be essential for faithfully capturing the non-Gaussian probabilistic behavior of input parameters, and would be particularly beneficial for constructing a general formulation for PC-based response analysis. Secondly, the associated response propagation should be suitable and efficient for high-dimensional and large-scale problems in terms of the computational demand so that the method can cover a wide range of applicability for a general purpose implementation.

The goal of this paper is to develop a new PC-based method for reasonably modeling of non-Gaussian system inputs as well as efficient propagation of associated system response under limited observations. Firstly, the limited observations of non-Gaussian uncertain input parameters are represented by KL expansion, resulting in a set of eigenpairs and corresponding KL random vector, followed by the development of a novel KDE for estimating the joint distribution of KL vector from their realizations, leading to the KDE-based random model of uncertain input parameters. In order to achieve the optimal convergence of associated response propagation, the aPC-based input model is then constructed by representing KL variables with aPC expansion weighted by their joint PDF. With the aPC representation of input parameters, we further develop a D-optimal weighted regression method for robust and accurate aPC approximation of system response. In our method, by incorporating the inherent relation between marginals of input field and distribution of univariate KL variables into the new KDE of KL vector, the developed KDE-based random model can accurately represent the input field from limited observations in terms of simultaneously reconstructing its marginals and second-order correlations. Furthermore, with the aid of the mixture representation of the developed KDE of KL vector, a new sample generator is developed for efficiently generating independent samples from KL vector, so that the enormous computational burden caused by repeated density evaluations as well as the inherent autocorrelations of generated samples in MCMC can be circumvented, and as a result, the aPC formulation of input parameters and stochastic system responses can be effectively determined. On the other hand, by virtue of the equivalence between the distribution of underlying aPC variables and that of KL vector, samples of underlying aPC variables are readily generated by the developed sampler for KL vector. With these samples, well-established PC-based solution techniques under independent PC variables are straightforwardly extended for the aPC-based response propagation by the developed D-optimal weighted regression method. In this way, the response is propagated in a robust and accurate way. With the reasonable stochastic modelling and efficient response propagation, the current work provides an effective framework for the stochastic analysis of practical engineering systems with limited observations.

The remainder of this paper is organized as follows. the novel KDE-based model construction technique for random field input parameter under limited observations is developed in Section 2. In Section 3, the associated aPC-based response propagation is developed. Two numerical examples are investigated to validate the effectiveness of proposed KDE-based model construction and aPC-based response propagation in Section 4.

2. A novel random field model of non-Gaussian input parameter with limited observations

Accurate representation of the non-Gaussian structural input parameters is the first essential step of the stochastic structural analysis with limited observations. As mentioned earlier, although various methods have been developed for this purpose, the KDE-based modelling technique is the most promising one among others because it permits to estimate multi-dimensional PDF of KL variables with reasonable computational demand. The main drawback of this method is that the resulting marginals may deviate from the true one due to the choice of bandwidth in KDE. This may lead to an inaccurate input model and thereby the untrustworthy estimation of its impact on engineering systems. In this section, we develop a new KDE-based model for accurately characterizing the non-Gaussian behavior of input parameters, and the KDE-based model in [32] is also presented for completeness.

2.1. Karhunen-Loeve expansion of non-Gaussian input parameter from limited observations

Consider a sequence of measurements of $w(x,\theta)$ at M locations over coordinates $x_1, x_2, \cdots, x_M$, named $w(x_i,\theta)$, $i=1,2,\cdots,M$ and there exists N independent observations of random variables $w(x_i,\theta)$ in each location. The observations of $w(x,\theta)$ can be summarized in an $N\times M$ matrix $\mathbf{W}=\left\{w(x_i,\theta_j)\right\}$, $i=1,\cdots,M$, $j=1,\cdots,N$.

For the random field $\mathbf{W}(\theta)=\left\{w(x_i,\theta)\right\}$, the truncated KL expansion of the original observations yields the following approximation

$$\hat{\mathbf{W}}(\theta)=\bar{\mathbf{W}}+\sum_{i=1}^{m}\sqrt{\lambda_i}\phi_i\xi_i(\theta) \tag{1}$$

where $\bar{\mathbf{W}}=[\bar{W}_1,\cdots,\bar{W}_M]$ is the mean of $\mathbf{W}(\theta)$, m is the truncation order related to the ratio of retained energies. Accordingly, the pairs $\{\lambda_i,\phi_i\}$ are the first m eigenvalues and eigenvectors of the covariance matrix $\mathbf{C}_{\mathbf{W}}$ of the field $\mathbf{W}(\theta)$. Generally, the number of retained terms is adopted such that $\sum_{m+1}^{M}\lambda_i \Big/ \sum_{i=1}^{M}\lambda_i \ll 1$. The second-order KL vector $\boldsymbol{\xi}(\theta)=[\xi_1(\theta),\cdots,\xi_m(\theta)]^{\mathrm{T}}$ has zero mean and unit covariance matrix, i.e.,

$$\mathrm{E}\left[\boldsymbol{\xi}(\theta)\right]=\mathbf{0},\quad \mathrm{E}\left[\boldsymbol{\xi}(\theta)\boldsymbol{\xi}^T(\theta)\right]=\mathbf{I}_m \tag{2}$$

where $\mathbf{I}_m$ is an $m\times m$ identity matrix. In the context of limited observations, the mean $\bar{W}_i\simeq\frac{1}{N}\sum_{j=1}^{N}w\left(x_i,\theta_j\right)$, $i=1,\cdots,M$ and the covariance matrix $\mathbf{C}_{\mathbf{W}}$ can be estimated as

$$\hat{\mathbf{C}}_{\mathbf{W}}=\frac{\mathbf{W}^{\mathrm{T}}\mathbf{W}}{N-1}-\frac{\mathbf{W}^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{W}}{N(N-1)} \tag{3}$$

where **U** is an *N*-dimensional vector whose entries are all one, and the KL variables $\xi_i(\theta)$ are characterized by their limited realizations as

$$\Xi_{ij}^{\mathrm{obs}}=\lambda_i^{-1/2}\left[\mathbf{W}_j-\bar{W}_i\right]\phi_i \tag{4}$$

for $i=1,\cdots,m$, $j=1,\cdots,N$, where $\mathbf{W}_j=[w\left(x_1,\theta_j\right),\cdots,w\left(x_M,\theta_j\right)]$, $\{\lambda_i\}$ and $\{\phi_i\}$ are the first m eigenvalues and eigenvectors of matrix $\hat{\mathbf{C}}_{\mathbf{W}}$, respectively.

It is known that, if the random field is Gaussian distributed, then $\xi_i(\theta)$, $i=1,\cdots,m$ are independent standard Gaussian variables [38]. While for non-Gaussian field $\mathbf{W}(\theta)$, the associated KL variables $\xi_i(\theta)$, $i=1,\cdots,m$ are not Gaussian and hence not independent, and in this case, joint density of $\xi_i(\theta)$, $i=1,\cdots,m$ has to be reconstructed from their limited realizations in Eq. (4) to capture the nonlinear dependence of KL variables. In [32], the KDE is used to estimate the distribution of KL vector $\boldsymbol{\xi}(\theta)$ as

$$\hat{p}_{\boldsymbol{\Xi}}(\boldsymbol{\xi})=\frac{1}{N}\sum_{j=1}^{N}\mathbf{K}_m\left(\frac{\hat{s}}{s}\boldsymbol{\Xi}_j^{\mathrm{obs}},\hat{s}^2\mathbf{I}_m\right) \tag{5}$$

where $\mathbf{K}_m\left(\frac{\hat{s}}{s}\boldsymbol{\Xi}_j^{\text{obs}}, \hat{s}^2\mathbf{I}_m\right)$ denotes an *m*-dimensional normal distribution with mean $\frac{\hat{s}}{s}\boldsymbol{\Xi}_j^{\text{obs}}$ and covariance matrix $\hat{s}^2\mathbf{I}_m$, s is a multidimensional Silverman bandwidth defined by $s=\left\{4/\left[N\left(2+m\right)\right]\right\}^{1/(m+4)}$, $\hat{s}$ is a positive parameter adopting as $\hat{s}=s\big/\sqrt{s^2+\left(N-1\right)/N}$, and $\boldsymbol{\Xi}_j^{\text{obs}}=[\Xi_{1j}^{\text{obs}},\cdots,\Xi_{mj}^{\text{obs}}]$. Once the joint distribution of KL variables has been estimated by Eq. (5), the model of random field $\mathbf{W}(\theta)$ can then be readily approximated by Eq. (1). Compared with other types of density estimator, e.g. maximum entropy or histogram estimator, since the KDE in Eq. (5) can be straightforwardly extended to high-dimensional cases without enormous computational burden, it provides an general scheme for KL-based random field reconstruction from limited observations [33]. Unfortunately, the choice of multi-dimensional Silverman bandwidth *s* in Eq. (5) inevitably leads to the deviation of the marginals of non-Gaussian fields, and as a result, the model in [32] is incapable of accurately capturing the non-Gaussian features of input parameters. This challenge regarding the accuracy of model of input parameters significantly hinders the practical application of the method.

2.2. A novel KDE-based random model of input parameters

In order to accurately model the non-Gaussian input parameters from limited observations, we develop a novel KDE for estimating the joint PDF of KL variables so that the most two critical statistics of a general non-Gaussian input field, i.e., marginal distribution as well as second-order correlations, can be simultaneously reconstructed. For the purpose of matching marginals of input parameters, we firstly choose the bandwidth s according to univariate KDE, rather than the multi-dimensional case as in [32]. This is because marginal of $\mathbf{W}(\theta)$ is actually synthesized by the linear combination of the associated univariate KL variables $\xi_i(\theta)$, $i=1,\cdots,m$, whose distribution can be obtained by marginalizing the distribution of KL vector $\boldsymbol{\xi}(\theta)$ as

$$\hat{p}_{\Xi_i}\left(\xi_i\right)=\int_{R^{m-1}}\hat{p}_{\boldsymbol{\Xi}}\left(\boldsymbol{\xi}\right)d\xi_1\cdots d\xi_{i-1}d\xi_{i+1}\cdots d\xi_m=\frac{1}{N}\sum_{j=1}^{N}K_1\left(\frac{\hat{s}}{s}\Xi_{ij}^{\text{obs}},\hat{s}^2\right) \tag{6}$$

In this way, the capacity of the KDE-based model for modelling marginals of $\mathbf{W}(\theta)$ is essentially improved, when compared with that in [32]. In the context of the developed univariate KDE, we further determine the univariate bandwidth as

$$s^{(i)}=0.9\min\left\{\sigma_i,\ \text{IQR}_i/1.34\right\}N^{-1/5},\quad i=1,\cdots,m \tag{7}$$

instead of Silverman bandwidth in [32], where IQR_i is the interquartile range (IQR) of realizations $\boldsymbol{\Xi}_i=\left\{\Xi_{i1}^{\text{obs}},\cdots,\Xi_{iN}^{\text{obs}}\right\}$. This is because the Silverman bandwidth works well only when the underlying density to be estimated is normally distributed. While for non-Gaussian variables, especially for those following long-tailed and skew distribution or multimodal distribution, the use of Silverman bandwidth may lead to an oversmoothed estimation. It is known that the KL variables of the random field model are commonly far from Gaussian under limited observations and thereby outliers are prone to be occurred in the realizations $\boldsymbol{\Xi}_i$. Since the IQR is more insensitive to outliers of samples of non-Gaussian KL variables, the incorporation of the interquartile range into Eq. (7) can produce a more robust estimate of the bandwidth $s^{(i)}$ when compared with the use of Silverman bandwidth. As a direct consequence, the non-Gaussianality of each KL variable can be effectively captured, and thereby the resulted input model can accurately characterize the non-Gaussian behavior. Note that since the IQR_i in Eq. (7) generally produces different bandwidths $s^{(i)}$, $i=1\cdots m$ for the associated KL variables, the resulting computational complexity may significantly decrease the efficiency for the construction of KDE-based model. In order to decrease the computational complexity, we further suggest that all KL variables share the same bandwidth s_{sh} as

$$s_{\text{sh}} = \sum_{i=1}^{m} w_i s^{(i)}, \quad w_i = \lambda_i^{1/2} \left(\sum\nolimits_{j=1}^{m} \lambda_j^{1/2} \right)^{-1} \tag{8}$$

where w_i is the weight of bandwidth $s^{(i)}$. As shown in Eq. (7), the value of w_i decreases with the index *i* of KL variables, indicating that the $s^{(i)}$ with smaller *i* contributes more to the proposed bandwidth s_{sh}. This is consistent with the fact that the KL variable with larger eigenvalue contributes more to the marginals of a random field [39]. In this sense, the shared bandwidth in Eq. (8) would be particularly beneficial in terms of the computational demand in KDE with reasonable accuracy.

Based on the bandwidth s_{sh} determined by Eq. (7) and Eq. (8), a new KDE is developed for estimating the joint distribution of $\boldsymbol{\xi}(\theta)$ as

$$\hat{p}_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) = \frac{1}{N} \sum_{j=1}^{N} K_m \left(\frac{\hat{s}_{\text{sh}}}{s_{\text{sh}}} \boldsymbol{\Xi}_j^{\text{obs}}, \hat{s}_{\text{sh}}^2 \mathbf{I}_m \right) \tag{9}$$

where positive parameter is adopted as $\hat{s}_{\text{sh}} = s_{\text{sh}} \Big/ \sqrt{s_{\text{sh}}^2 + (N-1)/N}$. Once the KL variables have been determined by Eq. (9), the non-Gaussian model of input field $w(x,\theta)$ can be accordingly synthesized with the KL expansion in Eq. (1).

2.2.1. Properties of the developed KDE-based model

It is acknowledged that marginal distributions as well as the second-order correlations are the two most concerned probabilistic characteristics of a general non-Gaussian random field. In the following, these two properties of the new KDE-based non-Gaussian model are investigated.

By using the relation between a random field $\mathbf{W}(\theta)$ and its associated KL variables $\boldsymbol{\xi}(\theta)$, the characteristic function of the marginal, $\hat{W}_k(\theta)$, $k = 1, \cdots, M$ of developed KDE-based model is formulated as

$$\begin{aligned} \varphi_{\hat{W}_k}(u) &= \mathrm{E}\left[\exp\left(\mathrm{i} u \hat{W}_k \right) \right] = \int_{R^m} \exp\left(\mathrm{i} u \bar{W}_k + \mathrm{i} u \sum_{j=1}^{m} \sqrt{\lambda_j} \phi_{jk} \xi_j \right) p_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi} \\ &= \exp\left(\mathrm{i} u \bar{W}_k \right) \frac{1}{N} \sum_{l=1}^{N} \int_{R^m} \exp\left(\mathrm{i} u \sum_{j=1}^{m} \sqrt{\lambda_j} \phi_{jk} \xi_j \right) \mathbf{K}_m \left(\frac{\hat{s}_{\text{sh}}}{s_{\text{sh}}} \boldsymbol{\Xi}_j^{\text{obs}}, \hat{s}_{\text{sh}}^2 \mathbf{I}_m \right) d\boldsymbol{\xi} \end{aligned} \tag{10}$$

Based on the property of multivariate normal distribution $\mathbf{K}_m(\cdot,\cdot)$, Eq. (10) is rewritten as

$$\varphi_{\hat{W}_k}(u) = \frac{1}{N} \sum_{l=1}^{N} \exp\left[iu \left(\bar{W}_k + \frac{\hat{s}_{\text{sh}}}{s_{\text{sh}}} \sum_{j=1}^{m} \sqrt{\lambda_j} \phi_{jk} \Xi_{lj}^{\text{obs}} \right) - \frac{1}{2} u^2 \hat{s}_{\text{sh}}^2 \sum_{j=1}^{m} \lambda_j \phi_{jk}^2 \right] \tag{11}$$

With the derived characteristic function $\varphi_{\hat{W}_k}(u)$ in Eq. (11), the mean and variance of $\hat{W}_k(\theta)$ are respectively calculated as

$$\begin{aligned} \mathrm{E}\left[\hat{W}_k \right] &= \mathrm{i}^{-1} \left. \frac{d\varphi_{\hat{W}_k}(u)}{du} \right|_{u=0} = \bar{W}_k \\ \mathrm{var}\left[\hat{W}_k \right] &= \mathrm{E}\left[\hat{W}_k^2 \right] - \mathrm{E}\left[\hat{W}_k \right]^2 = \mathrm{i}^{-2} \left. \frac{d^2\varphi_{\hat{W}_k}(u)}{du^2} \right|_{u=0} - \bar{W}_k^2 = \sum_{j=1}^{m} \lambda_j \phi_{jk}^2 \end{aligned} \tag{12}$$

The results in Eq. (12) are consistent with the counterparts of truncated random field $\hat{\mathbf{W}}(\theta)$, implying the capacity of the developed KDE-based model for reconstructing the first two order statistics of marginals of the field $\mathbf{W}(\theta)$. By further taking the Fourier transform on characteristic function $\varphi_{\hat{W}_k}(u)$, the PDF of $\hat{W}_k(\theta)$ is readily obtained as

$$\hat{p}\left(\hat{W}_k\right) = \mathbf{F}\left[\varphi_{\hat{W}_k}(u)\right] = \frac{1}{N}\sum_{l=1}^{N}\mathbf{K}_1\left(\bar{W}_k + \frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\sum_{j=1}^{m}\lambda_j^{1/2}\phi_{jk}\Xi_{lj}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\sum_{j=1}^{m}\lambda_j\phi_{jk}^2\right) \tag{13}$$

Since the positive parameter $\hat{s}_{\mathrm{sh}}$ in Eq. (9) is adopted as $\hat{s}_{\mathrm{sh}} = s_{\mathrm{sh}}\Big/\sqrt{s_{\mathrm{sh}}^2 + (N-1)/N}$, relation $\hat{s}_{\mathrm{sh}}/s_{\mathrm{sh}} \to 1$ holds as $N \to +\infty$. Given the consistency of KDE, it is natural that the PDF of $\hat{W}_k(\theta)$ in Eq.(13) converges to the true marginal density $p\left(\hat{W}_k\right)$ in probability as $N \to +\infty$ [33]. Therefore, the developed KDE-based model is capable of accurately reconstructing marginals of the field $\mathbf{W}(\theta)$.

In order to further determine second-order correlation of the developed KDE-based model, second-order properties of the KL variables in Eq. (9) is firstly investigated.

Proposition 1. *The set of random variables* $\boldsymbol{\xi}(\theta)$ *with the joint distribution defined by Eq.* (9) *are uncorrelated, and have zero means and unit variances, i.e.,* $\mathrm{E}[\boldsymbol{\xi}] = \mathbf{0}$, $\mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}] = \mathbf{I}_m$.

Proof. By using the properties of KDE, the joint distribution of $\boldsymbol{\xi}(\theta)$ in Eq. (9) is rewritten as

$$p_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) = \frac{1}{N}\sum_{j=1}^{N}\mathbf{K}_m\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\boldsymbol{\Xi}_j^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\mathbf{I}_m\right) = \frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{m}\mathbf{K}_1\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\Xi_{ij}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\right) \tag{14}$$

Thus, for $k = 1,\cdots,m$, the mean of $\xi_k(\theta)$ is readily calculated as

$$\mathrm{E}\left[\xi_k\right] = \int_{\boldsymbol{\Xi}}\xi_k p_{\boldsymbol{\Xi}}(\boldsymbol{\xi})d\boldsymbol{\xi} = \frac{1}{N}\sum_{j=1}^{N}\int_{\Xi_k}\xi_k\mathbf{K}_1\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\Xi_{kj}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\right)d\xi_k = \frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\frac{1}{N}\sum_{j=1}^{N}\Xi_{kj}^{\mathrm{obs}} \tag{15}$$

Since the relation $\frac{1}{N}\sum_{j=1}^{N}\Xi_{kj}^{\mathrm{obs}} = 0$ holds for all *k*, we have $\mathrm{E}[\boldsymbol{\xi}] = \mathbf{0}$. Further, $\mathrm{E}\left[\xi_k\xi_l\right]$, $1 \le k,l \le m$ can be formulated as

$$\mathrm{E}\left[\xi_k\xi_l\right] = \frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{\Xi}}\xi_k\xi_l\prod_{i=1}^{m}\mathbf{K}_1\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\Xi_{ij}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\right)d\boldsymbol{\xi} \tag{16}$$

For $k = l$, we have

$$\mathrm{E}\left[\xi_k\xi_l\right] = \frac{1}{N}\sum_{j=1}^{N}\left[\int_{\Xi_k}\xi_k^2 K_1\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\Xi_{kj}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\right)d\xi_k\right] = \frac{\hat{s}_{\mathrm{sh}}^2}{s_{\mathrm{sh}}^2}\frac{N-1}{N}\left(\frac{1}{N-1}\sum_{j=1}^{N}\left(\Xi_{kj}^{\mathrm{obs}}\right)^2\right) + \hat{s}_{\mathrm{sh}}^2 \tag{17}$$

While for $k \ne l$, we have

$$\mathrm{E}\left[\xi_k\xi_l\right] = \frac{1}{N}\sum_{j=1}^{N}\left(\prod_{r=k,l}\int_{\Xi_r}\xi_r\mathbf{K}_1\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\Xi_{rj}^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2\right)d\xi_r\right) = \frac{\hat{s}_{\mathrm{sh}}^2}{s_{\mathrm{sh}}^2}\frac{N-1}{N}\left(\frac{1}{N-1}\sum_{j=1}^{N}\left(\Xi_{kj}^{\mathrm{obs}}\Xi_{lj}^{\mathrm{obs}}\right)\right) \tag{18}$$

With the relation $\frac{1}{N-1}\sum_{j=1}^{N}\left(\Xi_{kj}^{\mathrm{obs}}\Xi_{lj}^{\mathrm{obs}}\right) = \delta_{kl}$, it is easy to verify $\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\right] = \mathbf{I}_m$ by following Eq. (17) and Eq. (18). This completes the proof.

With the first two order properties of KL variables in Proposition 1, the second-order correlation of the developed KDE-based model is immediately calculated as

$$\begin{aligned}\operatorname{cov}\left(W_k(\theta),W_l(\theta)\right)&=\mathrm{E}\left[\left(W_k(\theta)-\mathrm{E}[W_k(\theta)]\right)\left(W_l(\theta)-\mathrm{E}[W_l(\theta)]\right)\right]\\&=\mathrm{E}\left[\sum_{i=1}^{m}\sqrt{\lambda_i}\phi_{ik}\xi_i(\theta)\sum_{j=1}^{m}\sqrt{\lambda_j}\phi_{jl}\xi_j(\theta)\right]\\&=\sum_{i=1}^{m}\sum_{j=1}^{m}\sqrt{\lambda_i}\sqrt{\lambda_j}\phi_{ik}\phi_{jl}\mathrm{E}[\xi_i(\theta)\xi_j(\theta)]=\sum_{i=1}^{m}\lambda_i\phi_{ik}\phi_{jl}\end{aligned}\tag{19}$$

As shown in Eq.(19), the optimal approximation of correlations of the field $\mathbf{W}(\theta)$ in mean-square sense can be achieved since KL variables are uncorrelated [4]. In this way, the developed KDE-based model is capable of simultaneously reconstructing both the non-Gaussian marginals and the second-order correlations of the field $\mathbf{W}(\theta)$ from limited observations. Therefore, the developed new KDE-based model can accurately characterize the non-Gaussian behavior of input parameters.

2.2.2. Example

In this section, an illustrative example is presented to demonstrate the capacity of the developed KDE-based model for accurately modelling non-Gaussian input field with limited observations. Without loss of generality, let $a(x,\theta)$, $x\in(-0.5,0.5)$ be a spatial/temporal uncertain parameter of a practical engineering system. The field $a(x,\theta)$ can be the conductivity in diffusion problems, Young's modulus of materials in mechanical problems, etc. In practice, it is impossible to get access to the complete probabilistic characteristics of field $a(x,\theta)$, and the available information can only be a set of nodal realizations on the spatial/temporal domain, which is directly or indirectly identified from limited specimens. Since the aim of this section is to investigate the capacity of the developed KDE-based model with limited observations, rather than identification techniques, the limited observations of $a(x,\theta)$ are artificially generated from Algorithm 1. With the obtained limited observations of field $a(x,\theta)$, the performance of proposed KDE-based model is examined through comparing with the conventional KDE model in section 2.1. The accuracy of these two models is assessed by comparing with observations of the field $a(x,\theta)$.

Algorithm 1 The artificial generation of sample realizations **A** of random conductivity parameter $a(x,\theta)$.

1: Select a total of $N=21$ observation points equidistantly in the definition domain of $a(x,\theta)$, i.e., $\mathbf{X}=\{X_1=-0.5, X_2=-0.45,\cdots,X_{21}=0.5\}$.

2: Calculate the $N\times N$ symmetric positive matrix $\mathbf{C}_G$ by $C_G(i,j)=\exp\left(-\left|X_i-X_j\right|\right)$, $X_i,X_j\subset\mathbf{X}$ on the observed points, and decompose matrix $\mathbf{C}_G$ into *N* eigenvalues and corresponding eigenvectors $\{\phi_i^G\}_{i=1}^N$.

3: Generate samples $\mathbf{\Psi}_i$, $i=1,\cdots,N$ of *N* mutual independent standard normal variables, and the samples size of each $\mathbf{\Psi}_i$ is $M=250$.

4: Calculate the **G** by $\mathbf{G}=\sum_{i=1}^{N}\sqrt{\lambda_i^G}\phi_i^G\mathbf{\Psi}_i$.

5: Synthesize the realizations **A** of $a(x,\theta)$ by $\mathbf{A}=\exp(\mathbf{G})+k$, where $k=6$.

Fig.1 shows the eigenvalues $\{\lambda_i\}_{i=1}^{21}$ of covariance matrix $\mathbf{C_A}$ of the observations **A**. The truncated parameter *m* in Eq. (1) is adopted as $m=8$ such that a total of 97.74% energy are retained. Fig. 2 depicts the relative errors between the original covariance matrix $\mathbf{C_A}$ and the predicted covariance matrices from conventional KDE model and the proposed model. It is clear that the covariance matrices obtained from both conventional KDE model and proposed model are in good accordance with that of observations. Fig. 3 displays the marginal cumulative density functions (CDFs) of field $a(x,\theta)$ at $x=-0.5$, $x=0$ and $x=0.5$. Evidently, the marginals generated by conventional KDE model deviate from the observations, this is because the conventional KDE model does not incorporate the inherent relation between marginals of input field and distribution of univariate KL variables. In contrast, the marginal distributions from proposed model agrees well with the observations, indicating the effectiveness of proposed model for accurately reconstructing the field $a(x,\theta)$ from limited observations in terms of simultaneously reconstructing its marginals and second-order correlations.

3. Arbitrary polynomial chaos-based system response analysis with the developed KDE-based model

The second step in the analysis of uncertain systems under limited observation is the propagation of uncertainty through the system and the assessment of its stochastic response. As mentioned earlier, although the PC-based methods have been developed for this purpose, the use of Wiener-Askey scheme may lead to a low computational efficiency especially in the case of high dimensionality, which significantly hinder the application of the method for practical engineering systems of interest. In this section, we develop an aPC-based propagation method for efficient stochastic response analysis by constructing aPC bases according to the KL variables in KDE-based model. The general form of the existing PC-based uncertain analysis framework is also reviewed [40].

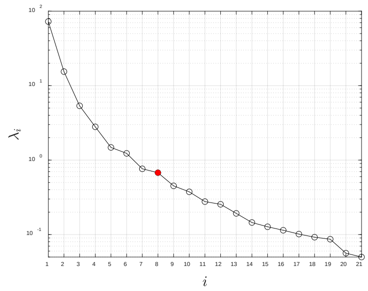


Figure 1: The eigenvalues of covariance matrix $\mathbf{C_A}$.

Figure. 2: Relative errors between the original covariance matrix $\mathbf{C_A}$ and the predicted covariance matrices (Left: Conventional KDE model; right: Proposed model).

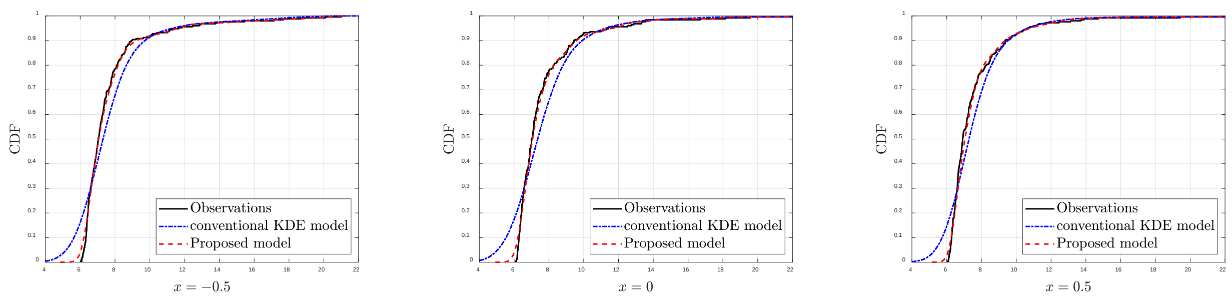


Figure. 3: The marginal distributions of observations at $x = -0.5$, $x = 0$ and $x = 0.5$.

3.1. Framework of PC-based uncertain analysis

In the framework of PC-based stochastic analysis, system response is generally projected into the same PC subspace as the uncertain input parameters. In the context of KL-based representation of input parameters, the PC expansion of the KL variable is generally formulated as

$$\xi_i(\theta) = \sum_{j=0}^{P} \alpha_{ij} \Psi_j[\boldsymbol{\eta}(\theta)], \quad 1 \le i \le m, 0 \le j \le P \tag{20}$$

where $\boldsymbol{\eta}(\theta)$ are m-dimensional underlying variables of PC expansion, P is the number of truncated terms, calculated as $P+1=(m+p)!/(m!p!)$, p is the order of m-dimensional normalized orthogonal polynomials $\Psi_j[\cdot]$, and α_{ij} are the PC coefficients to be determined. By virtue of the orthogonality of $\Psi_j[\cdot]$, α_{ij} in Eq. (20) are calculated by

$$\alpha_{ij} = \frac{\mathrm{E}\left[\xi_i \Psi_j[\boldsymbol{\eta}]\right]}{\mathrm{E}\left[\Psi_j^2[\boldsymbol{\eta}]\right]} = \int_{\mathbf{H}} G_i(\boldsymbol{\eta}) \Psi_j[\boldsymbol{\eta}] p_{\mathbf{H}}(\boldsymbol{\eta}) d\boldsymbol{\eta} \tag{21}$$

for $1 \le i \le m$, $0 \le j \le P$, where $p_{\mathbf{H}}(\boldsymbol{\eta})$ is the density of $\boldsymbol{\eta}(\theta)$, and $\boldsymbol{\xi} = \mathbf{G}(\boldsymbol{\eta}) = (G_1(\boldsymbol{\eta}), \cdots, G_m(\boldsymbol{\eta}))$ can be determined by following Rosenblatt transformation [8,30,31]

$$\begin{aligned} \eta_1 &= P_{\mathrm{H}_1}^{-1}\left[P_{\Xi_1}(\xi_1)\right] \\ \eta_i &= P_{\mathrm{H}_i}^{-1}\left[P_{\Xi_i|\Xi_{i-1}\cdots\Xi_1}\left(\xi_i \middle| \xi_{i-1}\cdots\xi_1\right)\right], \quad i = 2, \cdots, m \end{aligned} \tag{22}$$

where $P_{\mathrm{H}_i}^{-1}[\cdot]$ is the inverse CDF of the PC variable η_i. $P_{\Xi_i|\Xi_{i-1}\cdots\Xi_1}\left(\xi_i \middle| \xi_{i-1}\cdots\xi_1\right)$, $1 \le i \le m$ is the conditional CDF of $\xi_i(\theta)$. With the PC representation of KL variables $\xi_i(\theta)$, the PC-based representation of non-Gaussian input field can be constructed by substituting Eq. (20) into Eq. (1),

$$\hat{\mathbf{W}}(\theta) = \bar{\mathbf{W}} + \sum_{i=1}^{m} \sum_{j=0}^{P} \sqrt{\lambda_i} \phi_i \alpha_{ij} \Psi_j\left[\boldsymbol{\eta}(\theta)\right] \tag{23}$$

Given the PC representation of stochastic input $\hat{\mathbf{W}}(\theta)$ of an engineering system, the stochastic system response $Y(\theta)$ can be projected into the same PC subspace $\{\Psi_j[\boldsymbol{\eta}(\theta)]\}_{j=1}^{\infty}$, yielding the identical PC representation of $Y(\theta)$, i.e., $Y(\boldsymbol{\eta})$. For practical implementation, system response $Y(\boldsymbol{\eta})$ is generally approximated by the truncated PC expansion as

$$Y(\boldsymbol{\eta}) \simeq \sum_{k=0}^{P} \alpha_k \Psi_k(\boldsymbol{\eta}) \tag{24}$$

which can also be written using a vector notation as $Y(\boldsymbol{\eta}) \simeq \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\Psi}(\boldsymbol{\eta})$.

3.2. Construction of arbitrary PC bases for uncertain system analysis

Although various types of PC approximation in Eq. (24) enable the stochastic response converge to the true one as $P \to \infty$, the convergent rate and thereby the efficiency of the propagation heavily depends on the choice of PC bases $\Psi_j[\boldsymbol{\eta}]$. It is known that, only when KL variables follow Gaussian or other common distributions, the use of Wiener–Askey scheme may provide the optimal convergence [35]. While in the context of limited observations, KL variables of the developed KDE-based model generally have much broader distributions outside the Wiener–Askey scheme. In order to propagate the input uncertainty as efficiently as possible, the aPC formulation is adopted in this study. The aPC expansion is a generalization of Wiener-Askey chaos and enables to construct orthogonal PC bases with respect to arbitrary distribution [8, 41, 42]. By constructing multidimensional orthogonal polynomials weighted by the measure of KL variables in Eq. (9) as aPC bases, the optimal convergence of the system response analysis could be achieved.

The construction of multidimensional orthogonal polynomials starts from specifying a set of linearly independent multi-index monomials as

$$\boldsymbol{\varphi}(\boldsymbol{\xi}) = [\varphi_0(\boldsymbol{\xi}), \cdots, \varphi_P(\boldsymbol{\xi})]^{\mathrm{T}} = \{\xi_1^{\alpha_1} \times \cdots \times \xi_m^{\alpha_m}\}, \alpha_1 + \cdots + \alpha_m \le p \tag{25}$$

where $P+1 = (m+p)!/(m!p!)$. With the multivariate polynomial bases $\boldsymbol{\varphi}(\boldsymbol{\xi})$, the multivariate orthonormal polynomials $\boldsymbol{\Psi}(\boldsymbol{\xi}) = (\Psi_0(\boldsymbol{\xi}), \cdots, \Psi_P(\boldsymbol{\xi}))^{\mathrm{T}}$ with respect to probability measure $p_{\Xi}(\boldsymbol{\xi})$ in Eq. (9) is accordingly constructed by the following Cholesky decomposition

$$\boldsymbol{\Psi}(\boldsymbol{\xi}) = \boldsymbol{\varphi}(\boldsymbol{\xi})\mathbf{L}^{-1} \tag{26}$$

where $\mathbf{L}$ is an upper triangular matrix from Cholesky decomposition on a $P \times P$ matrix $\mathbf{G}$, where its *ij*-th element is defined as

$$G_{ij} = \int_{\Xi} \varphi_i(\boldsymbol{\xi})\varphi_j(\boldsymbol{\xi}) p_{\Xi}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \mathrm{E}[\varphi_i(\boldsymbol{\xi})\varphi_j(\boldsymbol{\xi})] \tag{27}$$

Clearly, the core of constructing orthonormal polynomials $\boldsymbol{\Psi}(\boldsymbol{\xi})$ lies in the evaluation of the multivariate integration in Eq. (27). Different from conventional Wiener-Askey scheme, the multivariate integration in Eq. (27) can not be transformed to multiplication of univariate integrals because of the dependence of KL variables $\boldsymbol{\xi}(\theta)$ in Eq. (9). As a result, MC integration has to be employed for evaluating Eq. (27) as

$$G_{ij} = \mathrm{E}[\varphi_i(\boldsymbol{\xi})\varphi_j(\boldsymbol{\xi})] \approx \frac{1}{K}\sum_{k=1}^{K} \varphi_i\left(\boldsymbol{\Xi}^{(k)}\right)\varphi_j\left(\boldsymbol{\Xi}^{(k)}\right) \tag{28}$$

where $\boldsymbol{\Xi}^{(k)}$ is the *k*-th sample realizations of multi-dimensional KL vector $\boldsymbol{\xi}(\theta)$. We note that the evaluation of Eq. (27) is hindered by the challenge in generating sample realizations $\boldsymbol{\Xi}^{(k)}$ from dependent KL variables. Although MCMC sampler has been developed for this purpose in [32], a huge number of repeated density evaluations yields enormous computational burden. Moreover, the inherent autocorrelation in the resulting MCMC samples dramatically reduces the efficiency of MC integration [34]. These two inherent deficiencies inevitably decrease the effectiveness of MCMC sampler for evaluating matrix $\mathbf{G}$, especially in the case of high dimensionality. Therefore, the most challenging issue in the construction of aPC bases is the generation of samples from multi-dimensional KL variables in an effective way.

3.2.1. Generator of independent samples of KL variables

In order to circumvent the deficiencies encountered in MCMC sampler, we develop a new sampler for generating independent realizations from multi-dimensional KL variables, so that Eq. (27) can be accurately evaluated in an efficient way. By formulating the joint PDF in Eq.(9) as

$$p_{\Xi}(\boldsymbol{\xi}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{K}_m\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\boldsymbol{\Xi}_i^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2 \mathbf{I}_m\right) = \int_X p_X(x)\, p_{\Xi}(\boldsymbol{\xi} \mid x) dx \tag{29}$$

where $p_X(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x-i)$, and $\delta(\cdot)$ is Dirac delta function, joint distribution of KL variables can be further rewritten as

$$p_{\Xi}(\boldsymbol{\xi}) = \int_X p_X(x)\, p_{\Xi}(\boldsymbol{\xi} \mid x) dx = \sum_{i=1}^{N} \frac{1}{N} p_{\Xi}(\boldsymbol{\xi} \mid X = i) \tag{30}$$

where $p_{\Xi}(\boldsymbol{\xi} \mid X = i) \sim \mathbf{K}_m\left(\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\boldsymbol{\Xi}_i^{\mathrm{obs}}, \hat{s}_{\mathrm{sh}}^2 \mathbf{I}_m\right)$ is an *m*-dimensional Gaussian distribution with mean $\frac{\hat{s}_{\mathrm{sh}}}{s_{\mathrm{sh}}}\boldsymbol{\Xi}_i^{\mathrm{obs}}$ and covariance $\hat{s}_{\mathrm{sh}}^2 \mathbf{I}_m$. From Eq. (30), it can be found that $p_{\Xi}(\boldsymbol{\xi})$ is essentially a type of mixture of distribution, in which each component is the multivariate normal distribution. In view of this, samples of multi-dimensional KL variables can be accordingly obtained by firstly choosing $p_{\Xi}(\boldsymbol{\xi} \mid X = i)$, $i = 1, \cdots, N$ with probability $1/N$, and then generating Gaussian-distributed samples from $p_{\Xi}(\boldsymbol{\xi} \mid X = i)$.

Algorithm 2 Generating samples from multi-dimensional KL variables in Eq. (9).

Input: parameters s_{sh}, $\hat{s}_{\text{sh}}$; realizations $\boldsymbol{\Xi}^{\text{obs}}$; the total number of samples K to be generated.

Output: K samples of KL variables $\boldsymbol{\xi}(\theta)$, i.e., $\boldsymbol{\Xi}$.

1: Define $\boldsymbol{\Xi} = \varnothing$

2: **for** $k = 1$ to K **do**

3: $i \sim \text{Uniform}(1, N)$

4: $\boldsymbol{\Xi}^{\text{Mix}} \sim K_m\left(\dfrac{\hat{s}_{\text{sh}}}{s_{\text{sh}}}\boldsymbol{\Xi}_i^{\text{obs}}, \hat{s}_{\text{sh}}^2 \mathbf{I}_m\right)$

5: $\boldsymbol{\Xi} = \boldsymbol{\Xi} \cup \boldsymbol{\Xi}^{\text{Mix}}$

6: **end for**

The resulting procedure for the generation of independent realizations from multi-dimensional KL variables $\boldsymbol{\xi}(\theta)$ is summarized in Algorithm 2. Since the uniformly distributed variables i in Step 3 and the normally distributed variables $\boldsymbol{\Xi}^{\text{Mix}}$ in Step 4 can both be readily generated, enormous computational burden resulting from the repeated density evaluations in MCMC sampler are no longer required. More importantly, since the independent samples of each component in mixture distribution can be generated in Step 4, samples of KL vector from Algorithm 2 are mutually independent. This property would be particularly beneficial in terms of the accuracy for estimating elements of matrix **G** in Eq. (28), because the inherent autocorrelations in MCMC samples is bypassed. These two distinguished properties in the developed sampler in Algorithm 2 guarantee the effective evaluation of matrix **G**, and as a result, the aPC bases $\boldsymbol{\Psi}(\boldsymbol{\xi})$ can be readily constructed.

It should be noted that, since the KL vector $\boldsymbol{\xi}(\theta)$ admits $\text{E}[\boldsymbol{\xi}] = \mathbf{0}$, $\text{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\text{T}}] = \mathbf{I}_m$ as proved in Proposition 1, the first $m+1$ elements of aPC bases $\boldsymbol{\Psi}(\boldsymbol{\xi})$ are then $\{\Psi_0(\boldsymbol{\xi}), \cdots, \Psi_m(\boldsymbol{\xi})\} = \{1, \xi_1, \cdots, \xi_m\}$. Therefore, PC coefficients α_{ij} of input fields in Eq. (23) become one for $i = j = 1, \cdots, m$, and the remaining coefficients α_{ij} reduce to zero.

3.3. Arbitrary Polynomial chaos expansion of system responses

With the aPC representation of input parameters, the next step is to approximate the system response Y by determining the aPC coefficients α_{ij} in Eq. (24). Although various intrusive and non-intrusive methods can be employed for this purpose, regression-based method is adopted in this study as it allows to use the third party software in a *black-box* fashion [43, 44]. It is known that accuracy and stability of this type of method heavily depends on the choice of collocation points, i.e., experimental design (ED) of underlying PC variables. In fact, most available ED schemes are evolved from the crude MC sampling, regardless of the dependence of PC variables [36]. This is why the ED schemes of independent PC variables have been quite well-established, while there exists a dearth of algorithmic options for ED of dependent PC variables. In this sense, the most critical issue in the aPC-based response analysis is the sample generation of dependent aPC variables so that the according ED of aPC variables can be further developed to achieve an accurate response propagation.

By representing the mapping $\boldsymbol{\xi} = \mathbf{G}(\boldsymbol{\eta})$ in Eq. (21) as $\xi_i = G_i(\boldsymbol{\eta}) = \sum_{k=1}^{\infty} g_{ik}\Psi_k[\boldsymbol{\eta}]$, and substituting this series into Eq. (21), we further reformulate aPC coefficients of input fields as

$$\alpha_{ij} = \sum_{k=1}^{\infty} g_{ik} \int_{\mathbf{H}} \Psi_k[\boldsymbol{\eta}]\Psi_j[\boldsymbol{\eta}] p_{\mathbf{H}}(\boldsymbol{\eta}) d\boldsymbol{\eta} = \sum_{k=1}^{\infty} g_{ik}\delta_{kj} \tag{31}$$

As mentioned above, since the relation $\alpha_{ij} = \delta_{ij}$ holds, we have $g_{ik} = \delta_{ik}$. Thus, with the constructed aPC formulation in Eqs. (25)-(27), the mapping $\boldsymbol{\xi} = \mathbf{G}(\boldsymbol{\eta})$ reduces to $\xi_i = \sum_{k=1}^{\infty} \delta_{ik}\Psi_k[\boldsymbol{\eta}] = \Psi_i[\boldsymbol{\eta}] = \eta_i$, $1 \le i \le m$. implying that the distribution of underlying aPC variables is equivalent to that of KL vector in Eq. (9). By this equivalence, independent samples of aPC variables can be readily generated from Algorithm 2, and as a consequence, available ED techniques under independent PC variables can be straightforwardly extended to those under dependent aPC variables. Thus, the challenge in the ED of dependent aPC variables is overcome.

Based on the obtained samples of aPC variables, we further develop a D-optimal weighted regression method for a more robust and accurate aPC approximation of system responses, in which the collocation points are determined by maximizing the determinant of information matrix.

We first formulate the estimation of aPC coefficients in Eq. (24) as the following weighted least squares form

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\hat{\boldsymbol{\alpha}} \in R^{P+1}} \left\| \mathbf{V}_{\mathrm{ED}} \boldsymbol{\Psi}_{\mathrm{ED}} \hat{\boldsymbol{\alpha}} - \mathbf{V}_{\mathrm{ED}} \mathbf{Y}_{\mathrm{ED}} \right\|^2 \tag{32}$$

and the PC coefficients $\hat{\boldsymbol{\alpha}}$ are determined by

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{\Psi}_{\mathrm{ED}}^{\mathrm{T}} \mathbf{V}_{\mathrm{ED}}^2 \boldsymbol{\Psi}_{\mathrm{ED}})^{-1} \boldsymbol{\Psi}_{\mathrm{ED}}^{\mathrm{T}} \mathbf{V}_{\mathrm{ED}}^2 \mathbf{Y}_{\mathrm{ED}} \tag{33}$$

where $\boldsymbol{\Psi}_{\mathrm{ED}}$ is an $N_{\mathrm{ED}} \times (P+1)$ matrix defined by $\boldsymbol{\Psi}_{\mathrm{ED}}(i,j) = \Psi_j(\boldsymbol{\Xi}_{\mathrm{ED}}^{(i)})$, $i = 1, \cdots, N_{\mathrm{ED}}$, $j = 0, \cdots, P$, $\mathbf{V}_{\mathrm{ED}}$ is an $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ diagonal matrix with the *i*-th element v_i adopted as the root inverse of Christoffel function, i.e., $v_i = [\sum_{j=0}^{P} \Psi_j^2(\boldsymbol{\Xi}_{\mathrm{ED}}^{(i)})]^{-1/2}$, and $\mathbf{Y}_{\mathrm{ED}} = [Y(\boldsymbol{\Xi}_{\mathrm{ED}}^{(1)}), \cdots, Y(\boldsymbol{\Xi}_{\mathrm{ED}}^{(N_{\mathrm{ED}})})]^{\mathrm{T}}$. $N_{\mathrm{ED}} = r(P+1)$ is the number of D-optimal collocation points $\boldsymbol{\Xi}_{\mathrm{ED}} = [\boldsymbol{\Xi}_{\mathrm{ED}}^{(1)}, \cdots, \boldsymbol{\Xi}_{\mathrm{ED}}^{(N_{\mathrm{ED}})}]$, where $r > 1$ is the oversampling ratio, and $\boldsymbol{\Xi}_{\mathrm{ED}}$ are determined by solving the following D-optimal optimization problem [45]

$$\boldsymbol{\Xi}_{\mathrm{ED}} = \underset{\dim(\boldsymbol{\Xi}_{\mathrm{ED}}) = m \times N_{\mathrm{ED}}}{\arg\max} \det \left| \tilde{\mathbf{V}}(\boldsymbol{\xi}) \tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi}) \tilde{\boldsymbol{\Psi}}^{\mathrm{T}}(\boldsymbol{\xi}) \tilde{\mathbf{V}}(\boldsymbol{\xi}) \right| \tag{34}$$

where $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ matrix $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi}) = [\Psi_0(\boldsymbol{\xi}), \cdots, \Psi_P(\boldsymbol{\xi}), \cdots, \Psi_{N_{\mathrm{ED}}}(\boldsymbol{\xi})]^{\mathrm{T}}$ is the enrichment of $N_{\mathrm{ED}} \times (P+1)$ matrix $\boldsymbol{\Psi}(\boldsymbol{\xi})$ according to total-degree graded reverse lexicographic ordering such that the dimension of orthogonal polynomials increases from $P+1$ to N_{ED}, and the entities of $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ diagonal matrix $\tilde{\mathbf{V}}(\boldsymbol{\xi})$ are $\tilde{v}_i = [\sum_{j=1}^{N_{\mathrm{ED}}} \Psi_j^2(\boldsymbol{\Xi}_{\mathrm{ED}}^{(i)})]^{-1/2}$. Algorithm 3 describes the details of determining set $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi})$.

Algorithm 3 The determination of the set $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi})$ in Eq. (34).

Input: The maximum degree *p* of aPC bases $\boldsymbol{\Psi}(\boldsymbol{\xi})$; the number of collocation points N_{ED}; the size $(P+1)$ of $\boldsymbol{\Psi}(\boldsymbol{\xi})$; the set $\boldsymbol{\varphi}(\boldsymbol{\xi})$ in Eq. (25).

Output: Enriched aPC bases $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi})$.

Compute the multi-index monomials set $\boldsymbol{\varphi}^{(p+1)}(\boldsymbol{\xi}) = \{\xi_1^{\alpha_1} \times \cdots \times \xi_m^{\alpha_m}\}$, $\alpha_1 + \cdots + \alpha_m = p+1$.

Impose the reverse lexicographic ordering on $\boldsymbol{\varphi}^{(p+1)}(\boldsymbol{\xi})$.

Create the set $\tilde{\boldsymbol{\varphi}}(\boldsymbol{\xi})$ by appending the first $(N_{\mathrm{ED}} - P - 1)$ elements of $\boldsymbol{\varphi}^{(p+1)}(\boldsymbol{\xi})$ to $\boldsymbol{\varphi}(\boldsymbol{\xi})$.

Construct aPC bases $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi})$ based on the set $\tilde{\boldsymbol{\varphi}}(\boldsymbol{\xi})$ via Eq. (28) and Eq. (26).

Given a set of $N^{\mathrm{C}} \gg N^{\mathrm{ED}}$ independent realizations of aPC variables $\boldsymbol{\Xi}_{\mathrm{C}} = [\boldsymbol{\Xi}_{\mathrm{C}}^{(1)}, \cdots, \boldsymbol{\Xi}_{\mathrm{C}}^{(N_{\mathrm{C}})}]$ generated by Algorithm 2 as the candidate pool, the optimization problem in Eq. (34) is approximated by choosing the set $\boldsymbol{\Xi}_{\mathrm{ED}}$ from the candidate set $\boldsymbol{\Xi}_{\mathrm{C}}$ via column-pivoted QR decomposition of matrix $\tilde{\mathbf{V}}_{\mathrm{C}} \tilde{\boldsymbol{\Psi}}_{\mathrm{C}}$, i.e.,

$$(\tilde{\mathbf{V}}_{\mathrm{C}} \tilde{\boldsymbol{\Psi}}_{\mathrm{C}})^{\mathrm{T}} \mathbf{P} = \mathbf{Q} [\mathbf{R}_1 \quad \mathbf{R}_2] \tag{35}$$

where $\tilde{\mathbf{V}}_{\mathrm{C}}$ is an $N_{\mathrm{C}} \times N_{\mathrm{C}}$ diagonal matrix with *i*-th entities $\tilde{\mathbf{V}}_{\mathrm{C}}(i,i) = [\sum_{j=1}^{N_{\mathrm{ED}}} \Psi_j^2(\boldsymbol{\Xi}_{\mathrm{C}}^{(i)})]^{-1/2}$, and the *ij*-th element of $N_{\mathrm{C}} \times N_{\mathrm{ED}}$ matrix $\tilde{\boldsymbol{\Psi}}_{\mathrm{C}}$ is $\tilde{\boldsymbol{\Psi}}_{\mathrm{C}}(i,j) = \Psi_j(\boldsymbol{\Xi}_{\mathrm{C}}^{(i)})$. $\mathbf{Q}$ is an $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ orthogonal matrix, $\mathbf{R}_1$ is an $N_{\mathrm{ED}} \times N_{\mathrm{ED}}$ nonsingular upper-triangular matrix, and $\mathbf{P}$ is an $N_{\mathrm{C}} \times N_{\mathrm{C}}$ permutation matrix that permutes the columns of $(\tilde{\mathbf{V}}_{\mathrm{C}} \tilde{\boldsymbol{\Psi}}_{\mathrm{C}})^{\mathrm{T}}$ such that the absolute value of the diagonal entries of $\mathbf{R}_1$ are in the descending order. Let $\boldsymbol{\pi} = \mathbf{P}^{\mathrm{T}} \times [1, \cdots, N_{\mathrm{C}}]$ be a vector that converts the pivots encoded in matrix $\mathbf{P}$ to the specific rows of $\tilde{\mathbf{V}}_{\mathrm{C}} \tilde{\boldsymbol{\Psi}}_{\mathrm{C}}$, the collocation points can thus be determined by

$$\boldsymbol{\Xi}_{\mathrm{ED}} = \boldsymbol{\Xi}_{\mathrm{C}}\left(\boldsymbol{\pi}_{\mathrm{ED}},:\right) \tag{36}$$

where $\boldsymbol{\pi}_{\mathrm{ED}} = \boldsymbol{\pi}\left(1:N_{\mathrm{ED}}\right)$ is the first entities of $\boldsymbol{\pi}$.

With the obtained collocation points $\boldsymbol{\Xi}_{\mathrm{ED}}$ in Eq. (36), the aPC coefficients of the system response can be readily determined by Eq. (33). For clarity, the procedure for aPC expansion of stochastic response is summarized as follows:

(a) Specify the maximum degree p of multidimensional polynomials $\boldsymbol{\Psi}(\boldsymbol{\xi})$ and the oversampling ratio r, and determine the number of collocation points N_{ED}.

(b) Construct the corresponding N_{ED} orthonormal polynomials $\tilde{\boldsymbol{\Psi}}(\boldsymbol{\xi}) = [\Psi_1(\boldsymbol{\xi}),\cdots,\Psi_{N_{ED}}(\boldsymbol{\xi})]^{\mathrm{T}}$ by Algorithm 3.

(c) Generate $N_{\mathrm{C}} \gg N_{\mathrm{ED}}$ samples $\boldsymbol{\Xi}_{\mathrm{C}} = [\boldsymbol{\Xi}_{\mathrm{C}}^{(1)},\cdots,\boldsymbol{\Xi}_{\mathrm{C}}^{(N_{\mathrm{C}})}]$ of KL variables by Algorithm 2 as the candidate pool, and then determine collocation points of aPC variables $\boldsymbol{\Xi}_{\mathrm{ED}} = [\boldsymbol{\Xi}_{\mathrm{ED}}^{(1)},\cdots,\boldsymbol{\Xi}_{\mathrm{ED}}^{(N_{\mathrm{ED}})}]$ by Eq.(35) and Eq.(36).

(d) Synthesize N_{ED} samples of input random field $\hat{\mathbf{W}}(\theta)$ by Eq.(1) and evaluate the deterministic model on N_{ED} points, and then estimate the aPC coefficients of system response by Eq. (33) and approximate system response by Eq. (24).

With the aPC expansion of the system response, the PDF of system response can be accordingly obtained by the simulation of aPC approximation of response based on the aPC samples generated by Algorithm 2.

3.4. An efficient uncertain analysis method of engineering systems with limited observations

The flowchart of the proposed method for the stochastic analysis of engineering system with limited observations of uncertain input parameters is sketched in Fig. 4. As shown in Fig. 4, steps 1 and 2 devote to construct the novel KDE-based random model for input parameters from limited observations. In order to efficiently determine the subsequent stochastic responses, KL variables in the developed model are further represented by aPC expansion in Step 3. The associated aPC-based stochastic response analysis are developed in Step 4, leading to a unified framework for stochastic modelling and the subsequent response propagation of engineering systems, in which only limited observations are available. It is worth mentioning that, by incorporating the inherent relation between marginals of input field and distribution of univariate KL variables, the new KDE of KL vector developed for modelling uncertain inputs in Step 2 can overcome the inaccurate reconstruction of marginals in conventional KDE-based model and thereby can accurately capture the non-Gaussian characteristics of an input field in terms of simultaneously reconstructing its marginals and second-order correlations. In this way, the developed KDE-based random model provides an effective tool for non-Gaussian uncertain input parameters representation of general engineering interest from limited observations. We also note that, with the aid of the mixture representation of the developed KDE of KL vector in Eq. (9), a new sample generator is developed for efficiently generating independent samples from KL vector in Algorithm 2, so that the enormous computational burden caused by repeated density evaluations as well as the inherent autocorrelations of generated samples in MCMC can be circumvented. In this way, the enormous computational burden in the MC-based construction of aPC bases can be greatly alleviated, and thereby aPC formulation of input parameters and stochastic system responses can be effectively determined. In addition, by virtue of the equivalence between the distribution of underlying aPC variables and that of KL vector, we generate samples of underlying aPC variables by Algorithm 2 once again. With these samples, the challenges regarding the ED of mutually dependent aPC variables and the subsequent aPC-based response analysis can be addressed by developing a D-optimal weighted regression method. In this way, the response is propagated in a robust and accurate way. With the reasonable stochastic modelling and efficient response propagation, the current work provides an effective framework for the stochastic analysis of practical engineering systems with limited observations.

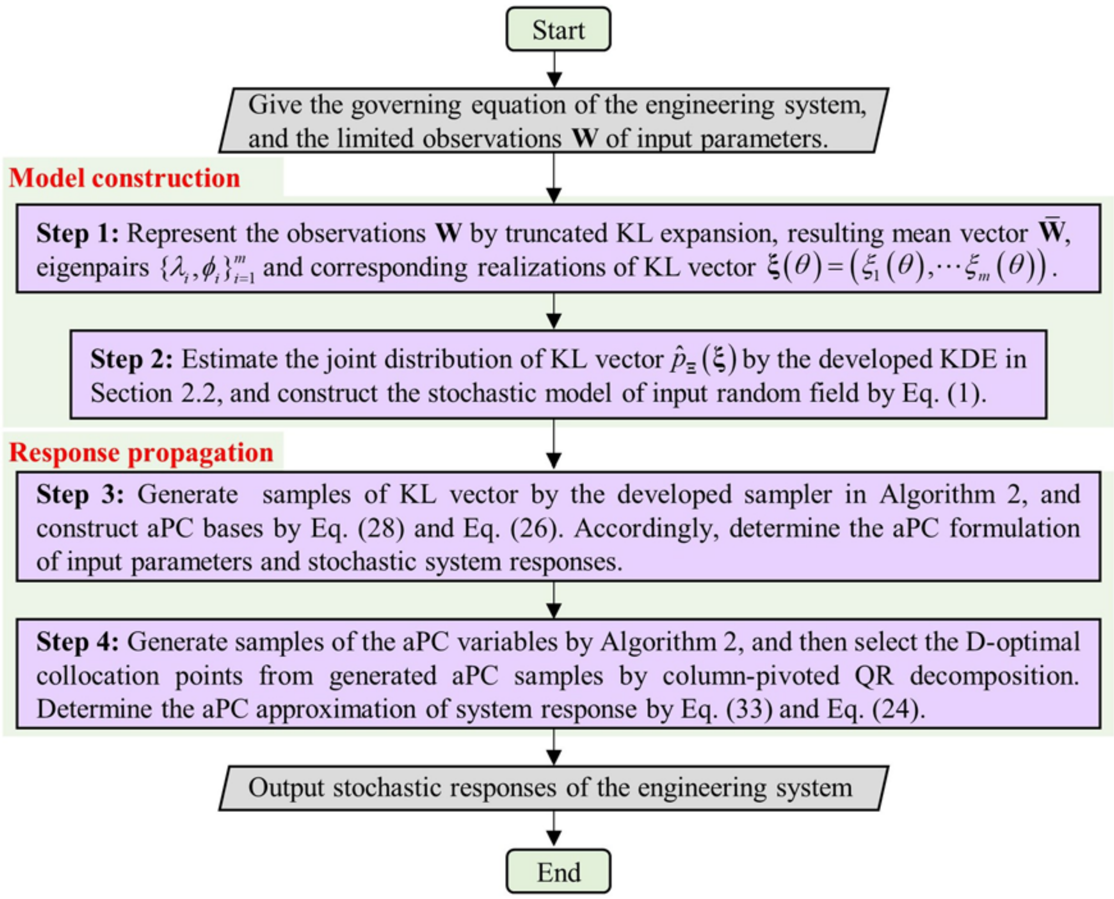


Figure.4: Flowchart of the proposed method.

We note that, the computational burden of proposed method is dominated by the response propagation since even the most time-consuming step in the model construction, i.e., the determination of first $m<M$ eigenpairs of the observations **W** in Step 1, can be performed with low computational cost. In the developed response propagation, the total CPU running time T_{total} consists of the time taken by the aPC formulation of the stochastic system, denoted by T_1, and the time needed in repeated evaluations of the deterministic system, denoted by T_2. For most stochastic analysis of structures of practical interest, the majority of computational cost is expended on the repeated evaluations of deterministic structures, and the CPU time T_1 needed for aPC formulation can be negligible in comparison with T_2 for performing repeated evaluations of deterministic structures, especially for large-scale engineering system.

4. Numerical examples

In this section, two numerical examples illustrating the application of the developed method are presented. The first example is a one-dimensional diffusion problem with random conductivity parameter, in which the realizations of random parameter are generated from Algorithm 1. Since the KDE-based model of conductivity parameter has been constructed in Section 2.2.2, the obtained results are directly applied for the associated response propagation. In example 2, a set of recorded natural ground motion time histories, which are chosen according to some site-specific criteria from the NGA strong-motion database established by the Pacific Earthquake Engineering Research Center (PEER), are investigated. The performance of proposed KDE-based model for seismic ground motion is examined in the same way as in example 1. With the random model of seismic ground motion, the response propagation of an eight degree-of freedom (DOF) linear structure and a twenty DOF nonlinear structure subjected to seismic ground motion are further performed to validate the proposed method for complex problems. In both examples, the number of samples for numerically constructing aPC bases in section 3.2 is chosen as $K = 10^4$, the oversampling ratio is adopted as $r = 1.25$, the number of candidate samples for performing D-optimal ED in section 3.3 is chosen as $N^{\mathrm{C}} = 10^4$, and the accuracy of aPC-based response approximation is examined through comparing with the references given by 10^5 MCS. To implement, all computer programs have been run on a notepad (core i7-11800H CPU and 32 GB RAM).

4.1. one-dimensional diffusion problem

The first example considers a simple one-dimensional diffusion problem governed by

$$-\frac{d}{dx}\left[a(x,\theta)\frac{du}{dx}(x,\theta)\right] = 0, x \in (-0.5, 0.5) \tag{37}$$

with boundary conditions $u(-0.5,\theta) = 0, u(0.5,\theta) = 1$. With the constructed KDE-based model of field $a(x,\theta)$ in section 2.2.2, the associated response propagation is accordingly performed to validate the accuracy of proposed method.

Fig. 5 shows the aPC approximation of $u(x,\theta)$ with different polynomial orders at locations $x = -0.25$, $x = 0$ and $x = 0.25$. The MCS results are also given as references to check the developed aPC-based response propagation. It is evident that a high precision approximation can be reached with a quite low order, i.e., $p = 2$, illustrating the high accuracy of the proposed aPC-based response propagation.

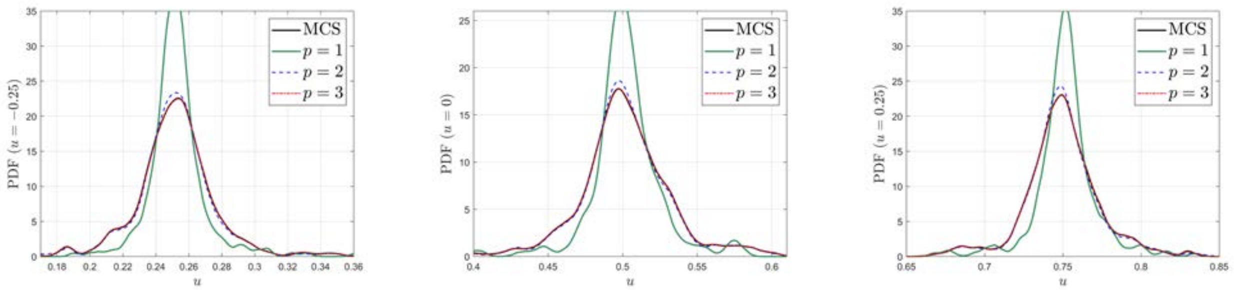


Figure.5: The stochastic response of diffusion system in Eq. (35). (Left: $u = -0.25$; middle: $u = 0$; right: $u = 0.25$).

4.2. Application to linear and nonlinear structures subjected to non-Gaussian seismic ground motion

In this example, the practical application of developed method for the uncertain analysis of engineering structures subjected to seismic ground motions is demonstrated. It is acknowledged that seismic ground motion is one of the typical natural hazards and should be modeled as a random process. Although a few techniques, e.g. spectral representation method, stochastic harmonic function representation, etc. provide convenient frameworks for characterizing non-Gaussian non-stationary seismic ground motions, obtained time histories cannot necessarily reconstruct all features of natural accelerograms. Although using recorded accelerograms can straightforwardly overcome this problem, the available ground motion time histories for a given scenario and site-condition are generally too limited to carry out subsequent response analysis and system assessment. In this case, the significant role of model construction consistent with limited time histories in the assessment of seismic safety of engineering structures is highlighted.

Table 1

The site-specific criteria for selecting the natural ground motion time histories.

Earthquake magnitude	Focal distance	Soil type
$5 \le M \le 6$	$1\text{km} \le D \le 20\text{km}$	Medium to hard soil with $V_s \ge 600\, m/s$

In this example, the natural accelerograms are selected from the NGA strong motion database with the site-specific criteria in Table 1. The purpose of specifying values of *M*, *D* and V_s in Table 1 as intervals rather than deterministic values is to incorporate the uncertain and imperfect knowledge of these site-specific ground motion parameters. According to the criteria in Table 1, a total of 102 ground motion time histories are selected, and each time history of 18s is discretized into 1801 points with step size $\Delta t = 0.01s$, as shown in Fig. 6. Fig. 7 shows eigenvalues of covariance matrix of the observations, and the first fifty eigenmodes are used for model construction such that a total of 95.23% energy is retained.

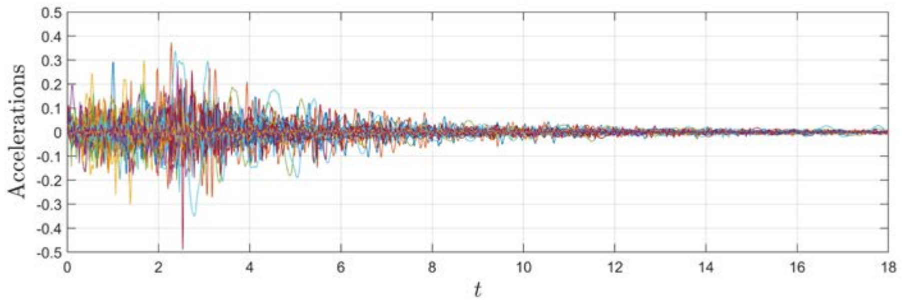


Figure 6: A total of 102 observed time histories selecting according to the site-specific criteria in Table 1.

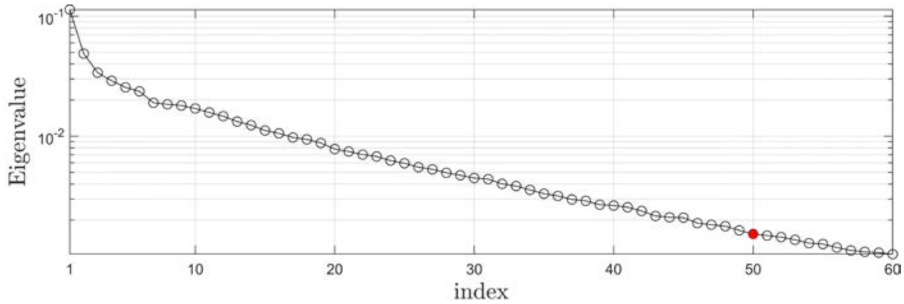
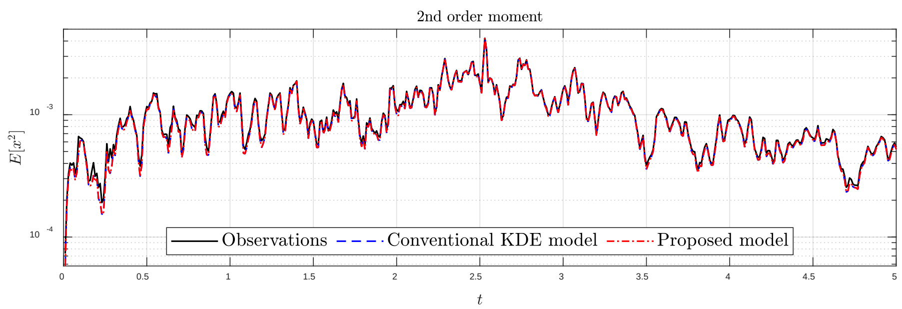


Figure 7: The first sixty eigenvalues of time histories

The proposed KDE-based model of seismic ground motion is constructed in only 0.71s. Fig. 8 shows the second-order to sixth-order moments of the marginal distributions of the conventional KDE model and proposed model. Since the scales of moment values varies greatly in the whole time histories, only the moments from 0s to 5s are displayed for the sake of clarity. It is clear that the conventional KDE model only matches the second order moment of the observations, while the proposed model enables to reconstruct the first six-order statistics in a high precision. Fig. 9 further illustrate probabilistic characteristic of marginal distributions, in which the marginal cumulative density functions (CDFs) of selected time histories at $t = 1.5s$, 6.5s, 11.5s and 16.5s are displayed. It is clear that, since the marginals generated by conventional KDE model evidently deviate from the observations, it is incapable of capturing the non-Gaussian features of seismic ground motion. In contrast, the marginals from proposed model agrees well with the observations, illustrating the effectiveness of the proposed method for modelling the non-Gaussian seismic ground motions.



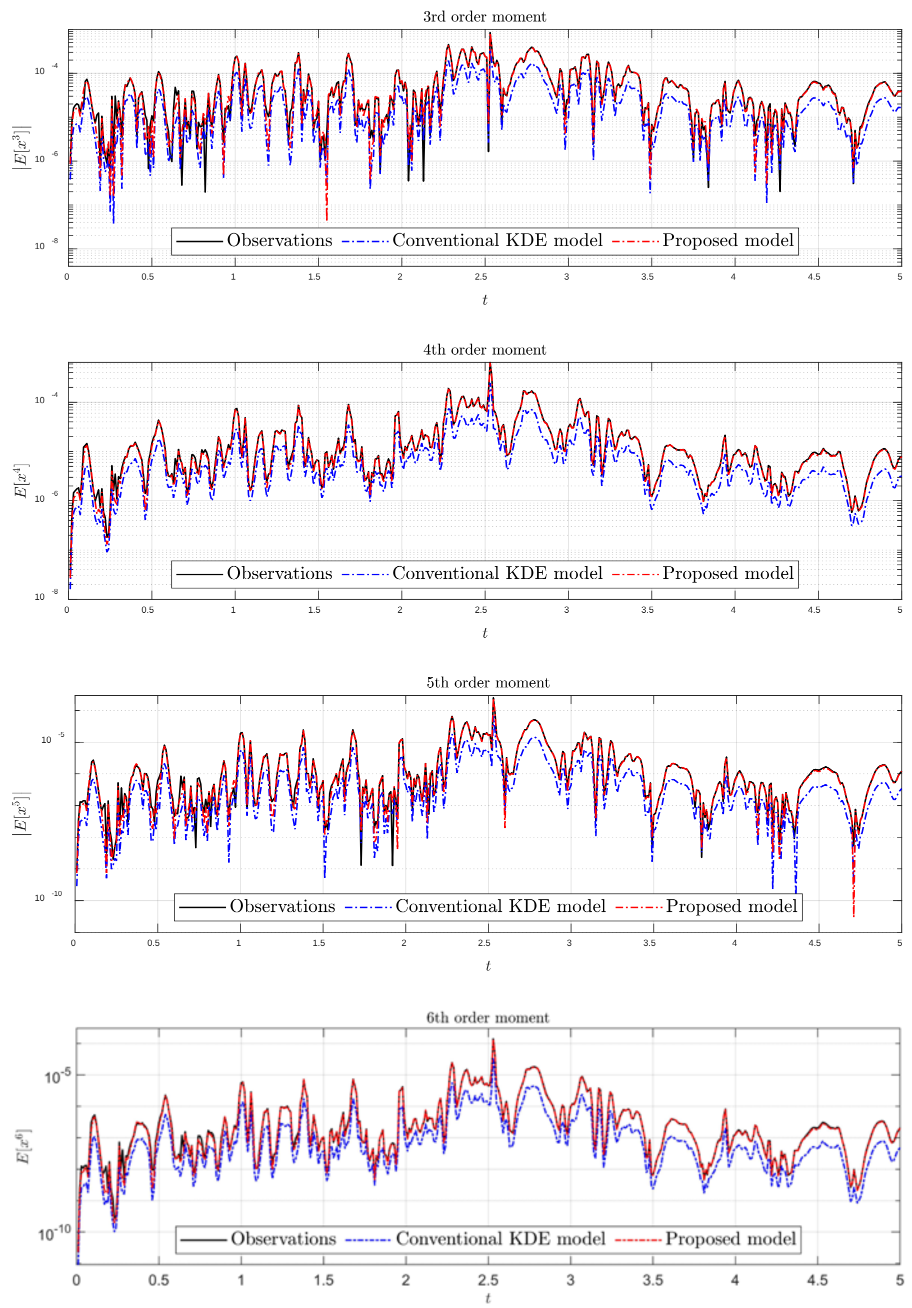
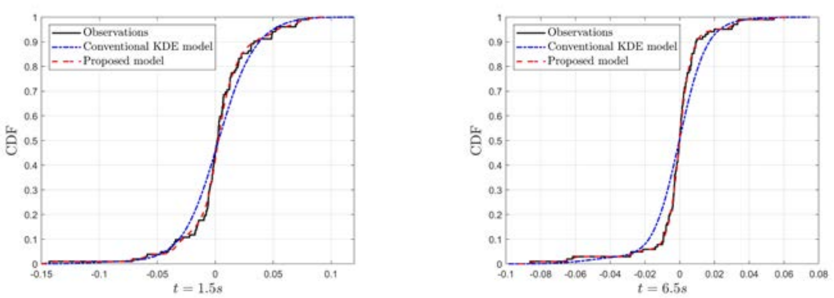


Figure. 8: First six order statistical moments of the marginal distributions.



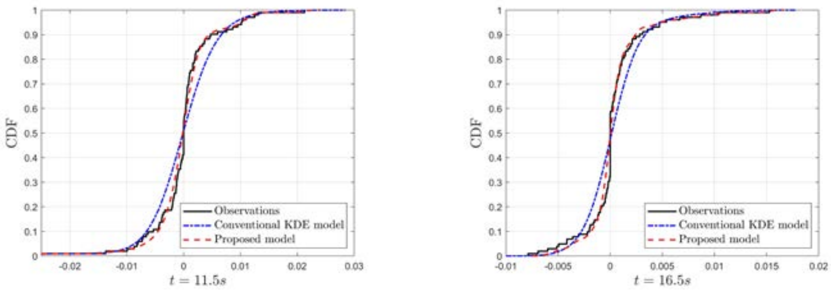


Figure. 9: The marginal CDFs of seismic ground motion at $t = 1.5s$, $6.5s$, $11.5s$ and $16.5s$.

In order to further validate the associated aPC-based response propagation, the stochastic response analysis of an 8-DOF linear system and a 20-DOF nonlinear shear-frame structure driven by the constructed seismic ground motions model are investigated in section 4.2.1 and section 4.2.2, respectively.

4.2.1. An 8-DOF linear structure subjected to seismic ground motion

The 8-DOF frame structure shown in Fig. 10 is subjected to the constructed stochastic ground motion [46]. The lumped masses from bottom to top are 3.442, 3.278, 3.056, 2.756, 2.739, 2.739, 2.739, 2.739 $(\times 10^5\,kg)$, the lateral inter-story stiffness from bottom to top are 1.92, 1.85, 1.63, 1.62, 1.60, 1.60, 0.96, 0.89 $(\times 10^8\,N/m)$. The Rayleigh damping is adopted such that $\mathbf{C} = a\mathbf{M} + b\mathbf{K}$, where $a = 0.2463s^{-1}$, $b = 0.0071s$.

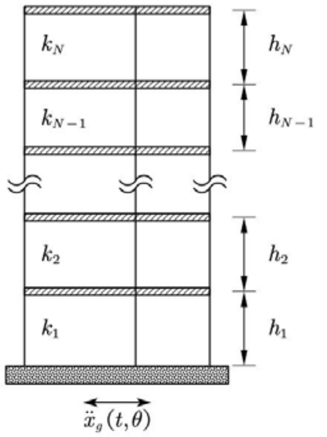
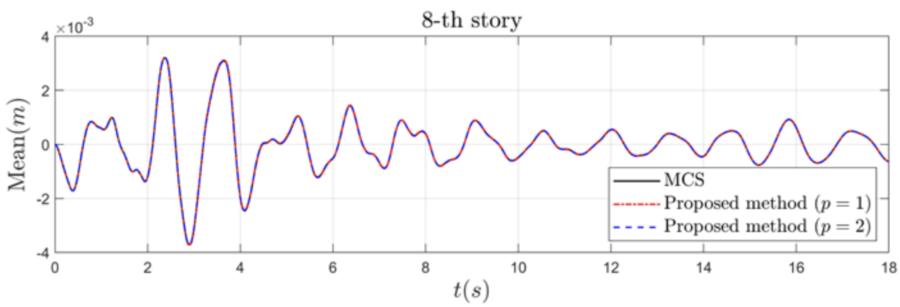


Figure. 10: Diagram of the shear-frame structure.

Fig. 11 depicts the mean values and standard deviations of stochastic response of 8-th story obtained by developed one and two orders aPC expansion. The MCS results are also displayed for validating the method. In Fig. 12, the probabilistic distribution of seismic response of 8-th story at typical time points, i.e., $t = 1.5s$, $t = 6.5s$, $t = 11.5s$ and $t = 16.5s$ are plotted. It is evident that the one-order aPC expansion is enough to produce an excellent approximation of response. In this case, only $N_{\mathrm{ED}} = 1.25 \times (50+1) = 64$ evaluations of the deterministic system are required, illustrating the high efficiency of developed method, as also evidenced by the CPU time of developed method depicted in Table 2.



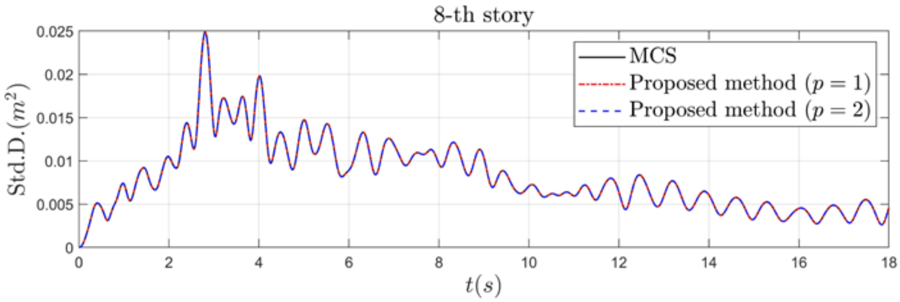


Figure. 11: The mean values and standard deviations of the stochstic response of 8-th story.

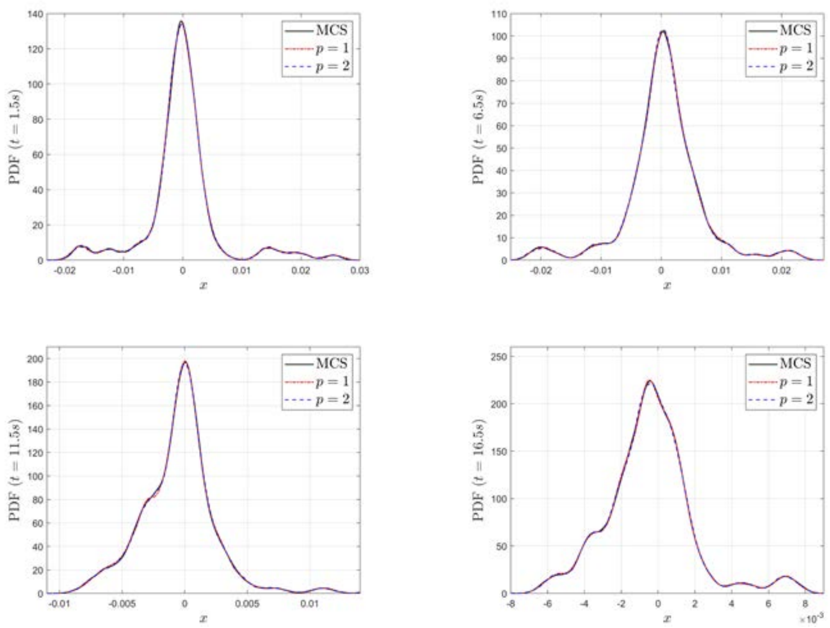


Figure. 12: The PDF curves of the seismic response of 8-th story at some typical time points.

Table 2
Comparison of CPU times of the developed method and MCS.

Methods	T_1	T_2	T_{total}
Developed method ($p=1$)	0.268s	0.631s	0.899s
1×10^5 MCS	-	848.919s	848.919s

4.2.2. A 20-DOF nonlinear structure subjected to seismic ground motion

In this section, a 20-DOF nonlinear frame structure is further investigated. The lumped mass and corresponding inter-story stiffness of the structure are displayed in Table 3. The nonlinear behavior is described by the Bouc-Wen hysteresis model, and the governing equations are formulated as [46]

$$\mathbf{M}\ddot{\mathbf{X}} + \mathbf{C}\dot{\mathbf{X}} + \alpha\mathbf{K}\mathbf{X} + (1-\alpha)\mathbf{K}\mathbf{Z} = -\mathbf{M}\mathbf{I}\ddot{x}_g(t,\theta) \tag{38}$$

where $\mathbf{X}$, $\dot{\mathbf{X}}$ and $\ddot{\mathbf{X}}$ are the displacement, velocity and lateral acceleration vector, respectively. $\mathbf{M} = \mathrm{diag}(m_1, m_2, \cdots, m_{20})$ denotes the mass matrix, $\mathbf{K}$ indicates the initial stiffness matrix, and the Rayleigh damping is adopted such that $\mathbf{C} = a\mathbf{M} + b\mathbf{K}$, where $a = 0.2463s^{-1}$, $b = 0.0071s$. $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_{20})^{\mathrm{T}}$ means the hysteresis displacement. The parameters in Bouc-Wen model take the values $\alpha = 0.04$, $A = 1$, $n = 1$, $q = 0.3$, $p = 10$, $d_\psi = 5$, $\lambda = 0.5$, $\psi = 0.05$, $\beta = 100$, $\gamma = 180$, $d_\nu = 1000$, $d_\eta = 1000$ and $\xi = 0.2$.

Table 3

The values of lumped mass and inter-story stiffness of shear-frame structure in example 2.

Story	Lumped mass $(\times 10^5\,\mathrm{kg})$	Inter-story stiffness $(\times 10^8\ \mathrm{N/m})$
1-2	4.5	3.5
3-12	4.3	3.2
13-17	4.1	3.0
18-20	3.9	2.8

Fig. 13 shows the random displacements of the 15-th story at four typical time points from proposed method with one and two-order aPC expansion, the MCS results are also depicted for comparison. Different from the linear structure case in section 4.2.1, the one-order aPC is not adequately to approximate the system response due to the strong nonlinearity of the system. However, the approximation accuracy rapidly increases to a reasonable level when the order of aPC reaches to two, i.e., $p = 2$, as also evidenced by the probability density surface of displacement of 20-th story demonstrated in Fig. 14. In this case, only $N_{\mathrm{ED}} = 1.25\times(50+2)!/(50!2!) = 1658$ evaluations of the deterministic hysteresis system are required. Table 4 depicts the CPU times of developed aPC-based method and MCS. Clearly, the MCS takes much more time than developed method, and this trend will be more apparent with the increasing of complexity of systems. By comparing with the results from MC method, it is clear that the proposed method enables to provide an excellent response approximation with justified computational cost, illustrating the potential of proposed method in the applications of large-scale engineering systems.

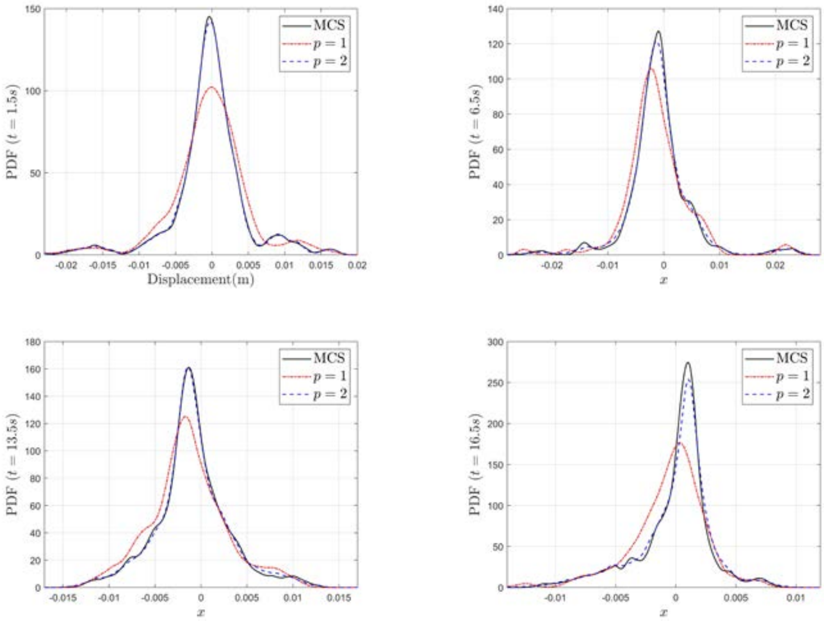
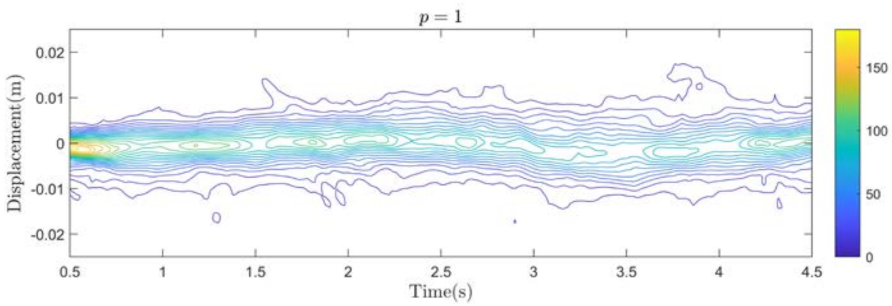


Figure. 13: Random resposnes at four typical time points.



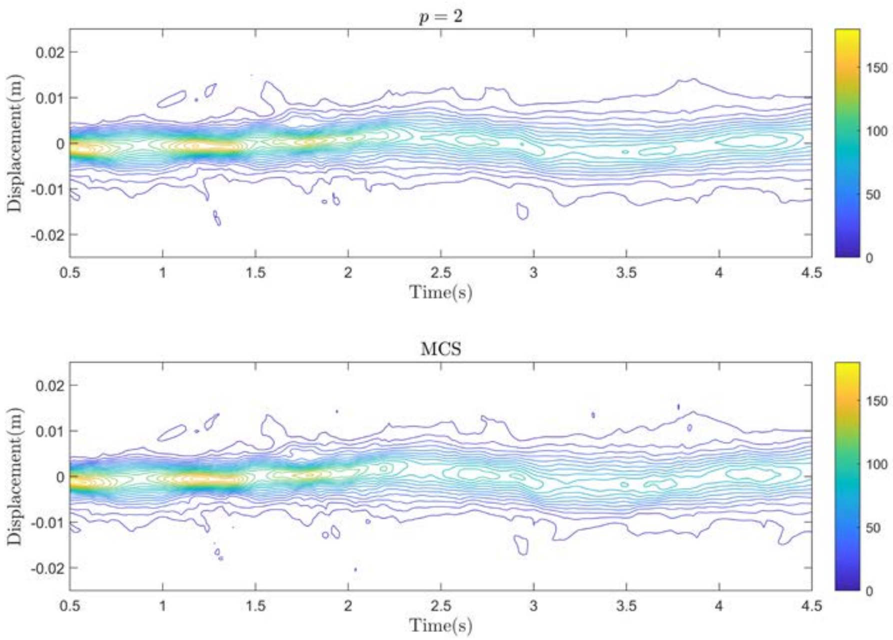


Figure. 14: The contour of PDF surface.

Table 4
Comparison of CPU times of the developed method and MCS.

Methods	T_1	T_2	T_{total}
Developed method ($p = 2$)	18.322s	141.654s	159.976s
1×10^5 MCS	-	7315.572s	7315.572s

5. Conclusions

This paper develops a new method for reasonably modeling of non-Gaussian system inputs as well as efficient propagation of associated system response under limited observations. The developed method firstly represents the limited non-Gaussian observations by KL expansion in terms of a set of KL variables. Followed by the development of a novel KDE for estimating the joint distribution of KL vector from their realizations, leading to the KDE-based random model of uncertain input parameters. In order to achieve the optimal convergence of associated response propagation, the aPC-based input model is further constructed by representing KL variables with aPC expansion weighted by their joint PDF. With the aPC representation of input parameters, a D-optimal weighted regression method is finally developed for robust and accurate aPC approximation of system response. In our method, by incorporating the inherent relation between marginals of input field and distribution of univariate KL variables into the new KDE of KL vector, the developed KDE-based random model can accurately represent the input field from limited observations in terms of simultaneously reconstructing its marginals and second-order correlations. Furthermore, with the aid of the mixture representation of the developed KDE of KL vector, a new sample generator is developed for efficiently generating independent samples from KL vector, so that the aPC formulation can be effectively constructed. On the other hand, by virtue of the equivalence between the distribution of underlying aPC variables and that of KL vector, samples of underlying aPC variables are readily generated by the developed sampler for KL vector. With these samples, well-established ED techniques under independent PC variables are straightforwardly extended for the estimation of aPC coefficients by further developing a D-optimal weighted regression method. In this way, the response can be propagated in a robust and accurate way. Two numerical examples, including a one-dimensional diffusion problem and the analysis of structures subjected to random seismic ground motion, have been studied to illustrate the effectiveness of developed method. In both examples, the developed KDE-based random model enables to reasonably capture the probabilistic characteristics of uncertain input parameters, and the developed aPC-based response propagation can efficiently determine the stochastic response of systems. The current work provides an effective framework for the stochastic analysis of practical engineering systems with limited observations.

We point out that, since the proposed model construction is based on KL expansion of one-dimensional random fields, the proposed stochastic modelling can be readily extended to the reconstruction of multidimensional and/or cross-correlated random fields with limited observations by introducing the existing generalized KL expansion for multidimensional and/or cross-correlated fields. In the future work, the current framework will be further generalized to the stochastic analyses of engineering systems involving multidimensional and/or cross-correlated random field parameters under limited observations by combining the generalized KL expansion developed by the present authors in [3].

Acknowledgments

This research was supported by Grant from the National Natural Science Foundation of China (Project 11972009). This support is gratefully acknowledged.

References

[1] Jianbing Chen, Fan Kong, and Yongbo Peng. A stochastic harmonic function representation for non-stationary stochastic processes. *Mechanical Systems and Signal Processing*, 96:31–44, 2017.

[2] Zhangjun Liu and Zenghui Liu. Random function representation of stationary stochastic vector processes for probability density evolution analysis of wind-induced structures. *Mechanical Systems and Signal Processing*, 106:511–525, 2018.

[3] Hongzhe Dai, Ruijing Zhang, and Michael Beer. A new perspective on the simulation of cross-correlated random fields. *Structural Safety*, 96:102201, 2022.

[4] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Dover Publications, INC, 2003.

[5] Jie Li and Jianbing Chen. The principle of preservation of probability and the generalized density evolution equation. *Structural Safety*, 30(1):65–77, 2008.

[6] Guohai Chen and Dixiong Yang. Direct probability integral method for stochastic response analysis of static and dynamic structural systems. *Computer Methods in Applied Mechanics and Engineering*, 357:112612, 2019.

[7] Yu Wang, Tengyuan Zhao, and Kok Kwang Phoon. Statistical inference of random field auto-correlation structure from multiple sets of incomplete and sparse measurements using bayesian compressive sampling-based bootstrapping. *Mechanical Systems and Signal Processing*, 124(JUN.1):217–236, 2019.

[8] Ruijing Zhang and Hongzhe Dai. Independent component analysis-based arbitrary polynomial chaos method for stochastic analysis of structures under limited observations. *Mechanical Systems and Signal Processing*, 173:109026, 2022.

[9] Fan Kong, Renjie Han, Shujin Li, andWei He. Non-stationary approximate response of non-linear multi-degree-of-freedom systems subjected to combined periodic and stochastic excitation. *Mechanical Systems and Signal Processing*, 166:108420, 2022.

[10] Jun Xu, Zhikang Wu, and Zhao-Hui Lu. An adaptive polynomial skewed-normal transformation model for distribution reconstruction and reliability evaluation with rare events. *Mechanical Systems and Signal Processing*, 169:108589, 2022.

[11] George D Pasparakis, Ketson RM dos Santos, Ioannis A Kougioumtzoglou, and Michael Beer. Wind data extrapolation and stochastic field statistics estimation via compressive sampling and low rank matrix recovery

methods. *Mechanical Systems and Signal Processing*, 162:107975, 2022.

[12] Ruijing Zhang, Liang Li, and Hongzhe Dai. A copula-based Gaussian mixture closure method for stochastic response of nonlinear dynamic systems. *Probabilistic Engineering Mechanics*, 59:103015, 2020.

[13] Ioannis A. Kougioumtzoglou, Ketson R. M. Dos Santos, and Liam Comerford. Incomplete data based parameter identification of nonlinear and time-variant oscillators with fractional derivative elements. *Mechanical Systems and Signal Processing*, 94(SEP.):279–296, 2017.

[14] Yu Wang, Tengyuan Zhao, and Kok Kwang Phoon. Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. *Canadian Geotechnical Journal*, 55(6):862–880, 2018.

[15] Fabrice Poirion and Irmela Zentner. Stochastic model construction of observed random phenomena. *Probabilistic Engineering Mechanics*, 36:63–71, 2014.

[16] Ruijing Zhang and Hongzhe Dai. A non-Gaussian stochastic model from limited observations using polynomial chaos and fractional moments. *Reliability Engineering & System Safety*, page 108323, 2022.

[17] Xufang Zhang, Qian Liu, and He Huang. Numerical simulation of random fields with a high-order polynomial based Ritz–Galerkin approach. *Probabilistic Engineering Mechanics*, 55:17–27, 2019.

[18] Ming-Na Tong, Yan-Gang Zhao, and Zhao Zhao. Simulating strongly non-Gaussian and non-stationary processes using Karhunen–Loève expansion and L-moments-based Hermite polynomial model. *Mechanical Systems and Signal Processing*, 160:107953, 2021.

[19] Zhibao Zheng, Hongzhe Dai, Yuyin Wang, and Wei Wang. A sample-based iterative scheme for simulating non-stationary non-Gaussian stochastic processes. *Mechanical Systems and Signal Processing*, 151:107420, 2021.

[20] Fabrice Poirion and Irmela Zentner. Non-Gaussian non-stationary models for natural hazard modeling. *Applied Mathematical Modelling*, 37(8):5938–5950, 2013.

[21] Ioannis A Kougioumtzoglou, Ioannis Petromichelakis, and Apostolos F Psaros. Sparse representations and compressive sampling approaches in engineering mechanics: A review of theoretical concepts and diverse applications. *Probabilistic Engineering Mechanics*, 61:103082, 2020.

[22] Liam Comerford, Ioannis A Kougioumtzoglou, and Michael Beer. Compressive sensing based stochastic process power spectrum estimation subject to missing data. *Probabilistic Engineering Mechanics*, 44:66–76, 2016.

[23] Liam Comerford, Ioannis A Kougioumtzoglou, and Michael Beer. An artificial neural network approach for stochastic process power spectrum estimation subject to missing data. *Structural Safety*, 52:150–160, 2015.

[24] Tengyuan Zhao and Yu Wang. Non-parametric simulation of non-stationary non-Gaussian 3D random field samples directly from sparse measurements using signal decomposition and Markov Chain Monte Carlo (MCMC) simulation. *Reliability Engineering & System Safety*, 203:107087, 2020.

[25] Silvana Montoya-Noguera, Tengyuan Zhao, Yue Hu, Yu Wang, and Kok-Kwang Phoon. Simulation of non-stationary non-Gaussian random fields from sparse measurements using bayesian compressive sampling and Karhunen-Loève expansion. *Structural Safety*, 79:66–79, 2019.

[26] Roger G Ghanem and Alireza Doostan. On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data. *Journal of Computational Physics*, 217(1):63–81, 2006.

[27] Irmela Zentner and Fabrice Poirion. Enrichment of seismic ground motion databases using Karhunen–Loève expansion. *Earthquake Engineering & Structural Dynamics*, 41(14):1945–1957, 2012.

[28] Christophe Desceliers, Roger Ghanem, and Christian Soize. Maximum likelihood estimation of stochastic chaos representations from experimental data. *International Journal for Numerical Methods in Engineering*, 66(6):978–1001, 2006.

[29] Loujaine Mehrez, Alireza Doostan, David Moens, and Dirk Vandepitte. Stochastic identification of composite

material properties from limited experimental databases, part ii: Uncertainty modelling. *Mechanical Systems and Signal Processing*, 27:484–498, 2012.

[30] Sonjoy Das, Roger Ghanem, and James 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):2207–2234, 2008.

[31] Sonjoy Das, Roger Ghanem, and Steven Finette. Polynomial chaos representation of spatio-temporal random fields from experimental measurements. *Journal of Computational Physics*, 228(23):8726–8751, 2009.

[32] Christian Soize. Polynomial chaos expansion of a multimodal random vector. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):34–60, 2015.

[33] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[34] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[35] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM journal on scientific computing*, 24(2):619–644, 2002.

[36] Nora Luthen, Stefano Marelli, and Bruno Sudret. Sparse polynomial chaos expansions: Literature survey and benchmark. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):593–649, 2021.

[37] Lukáš Novák, Miroslav Vořechovsk`y, Václav Sadílek, and Michael D Shields. Variance-based adaptive sequential sampling for polynomial chaos expansion. *Computer Methods in Applied Mechanics and Engineering*, 386:114105, 2021.

[38] Kok-Kwang Phoon, SP Huang, and Ser Tong Quek. Simulation of second-order processes using Karhunen-Loève expansion. *Computers & structures*, 80(12):1049–1060, 2002.

[39] K.K. Phoon, H.W. Huang, and S.T. Quek. Simulation of strongly non-Gaussian processes using Karhunen-Loève expansion. *Probabilistic Engineering Mechanics*, 20(2):188–198, April 2005.

[40] Christian Soize and Roger Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26(2):395–410, 2004.

[41] Jeroen AS Witteveen and Hester Bijl. Modeling arbitrary uncertainties using Gram-Schmidt polynomial chaos. In *44th AIAA aerospace sciences meeting and exhibit*, page 896, 2006.

[42] Sharif Rahman. A polynomial chaos expansion in dependent random variables. *Journal of Mathematical Analysis and Applications*, 464(1):749–775, 2018.

[43] Marc Berveiller, Bruno Sudret, and Maurice Lemaire. Stochastic finite element: a non-intrusive approach by regression. *European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique*, 15(1-3):81–92, 2006.

[44] Géraud Blatman and Bruno Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of computational Physics*, 230(6):2345–2367, 2011.

[45] Ling Guo, AkilNarayan, LiangYan, and Tao Zhou. Weighted approximate fekete points: sampling for least-squares polynomial approximation. *SIAM Journal on Scientific Computing*, 40(1): A366–A387, 2018.

[46] Jun Xu. A new method for reliability assessment of structural dynamic systems with random parameters. *Structural Safety*, 60:130–143, 2016.