International Journal for Uncertainty Quantification, 11(4):1–23 (2021)

LOCALIZED STOCHASTIC GALERKIN METHODS FOR HELMHOLTZ PROBLEMS CLOSE TO RESONANCE

Guanjie Wang,[1] Fei Xue,[2] & Qifeng Liao[3,]*

[1]*School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, 201209, Shanghai, P. R. China*

[2]*Department of Mathematical Sciences, Clemson University, 29631, Clemson SC, USA*

[3]*School of Information Science and Technology, ShanghaiTech University, 201209, Shanghai, P. R. China*

*Address all correspondence to: Qifeng Liao, School of Information Science and Technology, ShanghaiTech University, 201209, Shanghai, P. R. China, E-mail: liaoqf@shanghaitech.edu.cn

Original Manuscript Submitted: mm/dd/yyyy; Final Draft Received: mm/dd/yyyy

Efficiently solving stochastic Helmholtz equations remains an open challenging problem, especially when the corresponding problems are close to resonant frequencies. For widely used stochastic Galerkin methods based on spectral stochastic finite element approximations, two main computational difficulties exist when solving this kind of problem: slow convergence rates of spectral approximation methods and efficiency degeneration of preconditioned iterative linear solvers. To address this issue, we focus on the multi-element generalized polynomial chaos (ME-gPC) for stochastic approximation and finite element methods for physical approximation. A novel localized stochastic Galerkin scheme based on the combination of ME-gPC finite element approximation and mean-based preconditioning is proposed and analyzed in this work. Theoretical analysis shows that the mean-based preconditioner can be efficient in this setting, and numerical studies demonstrate the overall efficiency of the localized stochastic Galerkin scheme to solve the stochastic Helmholtz equations close to resonance.

KEY WORDS: *Helmholtz equations, uncertainty quantification, multi-element generalized polynomial chaos, iterative solvers*

1. INTRODUCTION

During the last decade there has been a rapid development in numerical methods for stochastic Helmholtz equations. This explosion in interest has been driven by the need of conducting uncertainty quantification for ocean acoustic, optic, and electromagnetic problems [1–3]. The uncertainty sources for these problems typically come from lack of knowledge or measurement of wave numbers, source parameters, and boundary data. In the work of Elman et al. [4], random forcing functions and boundary conditions are studied, and multigrid solvers for the corresponding stochastic finite element approximation are developed. Uncertain scattering boundary shapes are studied by [5,6]. For random diffusion parameters and wave numbers, efficient domain decomposition solvers are developed in [7]. In these sources of uncertainty, uncertainty quantification for random wave numbers is very challenging, especially for these parameters close to resonant frequencies [8]. They rather lead to loss of coercivity of the Helmholtz operator, which in turn leads to linear systems with singular matrices to solve. This paper is devoted to stochastic Helmholtz problems close to resonance.

Among uncertainty quantification approaches, this work focuses on stochastic spectral methods, which have been shown to be effective in many disciplines [9,10]. Stochastic Galerkin [9,11] and stochastic collocation [10,12–17] are two kinds of widely used spectral approaches. The stochastic collocation method is studied for Helmholtz equations in [5,6] and related scalar hyperbolic equations in [16], while the results in [6] show that it remains an open challenging

2152–5080/21/$35.00

problem to develop an effective stochastic collocation procedure for high-frequency acoustic scattering problems. To efficiently solve stochastic Helmholtz problems close to resonance, it is essential to conduct adaptive refinements for stochastic approximations. Adaptive strategies have been actively developed for stochastic Galerkin methods, which include the multi-element generalized polynomial chaos (ME-gPC) method [18] and adaptive refinement processes based on a posteriori error estimators [19]. As these works have shown great potential in conducting adaptivity using stochastic Galerkin methods, we focus on ME-gPC [18] and the spectral stochastic finite element methods [20], which are based on stochastic Galerkin. In addition, we note that the well-posedness and a priori bounds for the solution of the Helmholtz equation with random coefficients are established in [21], and a stochastic Helmholtz preconditioning problem is analyzed in [22]. It is known that the overall efficiency of a stochastic Galerkin method relies on the following two aspects: choosing proper approximation spaces and designing suitable iterative solvers for the corresponding linear systems. The resonant frequencies cause difficulties for both aspects. First, when the wave numbers are close to resonant frequencies, variances of solutions become very large, and stochastic spectral approximation methods can have low convergence rates. Second, again due to large solution variances, the linear systems arising from spectral stochastic finite element methods can be ill-conditioned, and the efficiency of standard preconditioned iterative linear solvers can deteriorate.

To result in an overall efficient strategy for solving these Helmholtz problems, a novel localized stochastic Galerkin framework is proposed and analyzed in this work. We focus on the ME-gPC for stochastic approximation [18] and finite elements for physical approximation. We show that by dividing a stochastic domain into subdomains, ME-gPC provides efficient localized stochastic approximations for stochastic Helmholtz problems. At the same time, we propose and analyze a mean-based preconditioning scheme for the linear systems arising from the ME-gPC finite element approximations. Since the solution variance on each stochastic element is much smaller than that of the overall solution, the mean-based preconditioning is shown to be efficient in this setting. The main novelty of this work lies on the new combination of ME-gPC finite elements and the mean-based preconditioning scheme. To simplify the presentation, this paper is restricted to the situation where random inputs have uniform distributions. In Section 2, the detailed setting of stochastic Helmholtz equations is introduced, and the frameworks of stochastic Galerkin methods and ME-gPC are discussed. In Section 3, detailed formulations of the linear systems associated with the ME-gPC finite element approximation are presented, and the corresponding mean-based preconditioning scheme is proposed and analyzed. Numerical results are discussed in Sections 4 and 5 concludes the paper.

2. THE STOCHASTIC HELMHOLTZ EQUATION AND ITS DISCRETIZATION

Let $D \subset \mathbb{R}^d$ ($d = 2, 3$) denote a physical domain that is bounded, connected, with a polygonal boundary ∂D, and where $\boldsymbol{x} \in \mathbb{R}^d$ denote a physical variable. Let $\boldsymbol{\xi}$ be a vector that collects a finite number of real-valued random variables, and the dimension of $\boldsymbol{\xi}$ is denoted by N. We next write $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^T$ and assume that the random variables $\xi_1, \ldots, \xi_N$ are independently distributed on the bounded intervals $\Gamma_1, \ldots, \Gamma_N$, respectively. The image of $\boldsymbol{\xi}$ is then denoted by $\Gamma := \Gamma_1 \times \cdots \times \Gamma_N$ and the probability density function of $\boldsymbol{\xi}$ is denoted by $\rho(\boldsymbol{\xi})$. In this paper, we consider the following stochastic Helmholtz problem: find the unknown function $u(\boldsymbol{x}, \boldsymbol{\xi})$ satisfying

$$-\nabla^2 u(\boldsymbol{x}, \boldsymbol{\xi}) - \kappa^2(\boldsymbol{x}, \boldsymbol{\xi}) u(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}, \boldsymbol{\xi}) \quad \forall (\boldsymbol{x}, \boldsymbol{\xi}) \in D \times \Gamma, \tag{1}$$

$$u(\boldsymbol{x}, \boldsymbol{\xi}) = 0 \quad \forall (\boldsymbol{x}, \boldsymbol{\xi}) \in \partial D \times \Gamma. \tag{2}$$

In Eq. (1), $\kappa(\boldsymbol{x}, \boldsymbol{\xi})$ is the wave number and $f(\boldsymbol{x}, \boldsymbol{\xi})$ is the forcing term, which are assumed to have the following forms:

$$\kappa(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{m=0}^{N} \kappa_m(\boldsymbol{x}) \xi_m, \; f(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{m=0}^{N} f_m(\boldsymbol{x}) \xi_m, \tag{3}$$

where $\{\kappa_m(\boldsymbol{x})\}_{m=0}^N$ and $\{f_m(\boldsymbol{x})\}_{m=0}^N$ are real-valued deterministic functions, and we set $\xi_0 = 1$ for convenience. Note that only the homogeneous Dirichlet boundary condition [Eq. (2)] is considered in this work for simplicity, while our approach can be applied in the case of other boundary conditions, like Neumann or transparent boundary conditions.

To ensure the well-posedness of the problem, we assume that there exists a constant $\epsilon > 0$, such that all the eigenvalues associated with deterministic versions of Eqs. (1) and (2) have a modulus greater than ϵ. That is, for each realization of $\boldsymbol{\xi}$, considering the following deterministic Helmholtz eigenvalue problem (see [23–25])

$$-\nabla^2 u(\boldsymbol{x},\boldsymbol{\xi}) - \kappa^2(\boldsymbol{x},\boldsymbol{\xi})u(\boldsymbol{x},\boldsymbol{\xi}) = \lambda(\boldsymbol{\xi})u(\boldsymbol{x},\boldsymbol{\xi}) \tag{4}$$

with boundary condition [Eq. (2)], we collect all its eigenvalues [i.e., all values of $\lambda(\boldsymbol{\xi})$ in Eq. (4)] into a set denoted by $\Lambda_{\boldsymbol{\xi}}$, and assume that $|\lambda| > \epsilon$ for all $\lambda \in \cup_{\boldsymbol{\xi}\in\Gamma}\Lambda_{\boldsymbol{\xi}}$. Note that $\kappa(\boldsymbol{x},\boldsymbol{\xi})$ is called a resonant frequency if the corresponding eigenvalue problem [Eq. (4)] has a zero eigenvalue.

2.1 Variational Formulation

To introduce the variational form of Eqs. (1) and (2), some notations are required. Letting $g(\boldsymbol{\xi})$ be a function of the random vector $\boldsymbol{\xi}$, which maps Γ into $\mathbb{R}$, its expectation (mean value) is defined by

$$\mathbb{E}[g(\boldsymbol{\xi})] := \int_\Gamma \rho(\boldsymbol{\xi})g(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi},$$

where $\rho(\boldsymbol{\xi})$ is the probability density function of $\boldsymbol{\xi}$. Next, the Hilbert spaces $L^2(D)$ and $L^2_\rho(\Gamma)$ are defined as

$$L^2(D) := \left\{ v(\boldsymbol{x}) : D \to \mathbb{R} \;\middle|\; \int_D v^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty \right\},$$

$$L^2_\rho(\Gamma) := \left\{ g(\boldsymbol{\xi}) : \Gamma \to \mathbb{R} \;\middle|\; \int_\Gamma \rho(\boldsymbol{\xi})g^2(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} < \infty \right\},$$

which are equipped with the inner products

$$(v(\boldsymbol{x}),\hat{v}(\boldsymbol{x}))_{L^2} := \int_D v(\boldsymbol{x})\hat{v}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{5}$$

$$(g(\boldsymbol{\xi}),\hat{g}(\boldsymbol{\xi}))_{L^2_\rho} := \int_\Gamma \rho(\boldsymbol{\xi})g(\boldsymbol{\xi})\hat{g}(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}. \tag{6}$$

Following [26], we define the tensor space of $L^2(D)$ and $L^2_\rho(\Gamma)$ as

$$L^2(D)\otimes L^2_\rho(\Gamma) := \left\{ w(\boldsymbol{x},\boldsymbol{\xi}) \middle| w(\boldsymbol{x},\boldsymbol{\xi}) = \sum_{i=1}^{n} v_i(\boldsymbol{x})g_i(\boldsymbol{\xi}), v_i(\boldsymbol{x}) \in L^2(D),\, g_i(\boldsymbol{x}) \in L^2_\rho(\Gamma), n \in \mathbb{N}^+ \right\},$$

which is equipped with the tensor inner product

$$(w(\boldsymbol{x},\boldsymbol{\xi}),\hat{w}(\boldsymbol{x},\boldsymbol{\xi}))_{L^2\otimes L^2_\rho} = \sum_{i,j}(v_i(\boldsymbol{x}),\hat{v}_j(\boldsymbol{x}))_{L^2}(g_i(\boldsymbol{\xi}),\hat{g}_j(\boldsymbol{\xi}))_{L^2_\rho},$$

where it is clear that $(w(\boldsymbol{x},\boldsymbol{\xi}),\hat{w}(\boldsymbol{x},\boldsymbol{\xi}))_{L^2\otimes L^2_\rho} = \mathbb{E}\big[\int_D w(\boldsymbol{x},\boldsymbol{\xi})\hat{w}(\boldsymbol{x},\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{x}\big]$. Denoting the trace operator for functions in $L^2(D)$ by γ, we next define the solution and test function space

$$W := H^1_0(D)\otimes L^2_\rho(\Gamma) = \left\{ w(\boldsymbol{x},\boldsymbol{\xi}) \in L^2(D)\otimes L^2_\rho(\Gamma) \middle| \|w(\boldsymbol{x},\boldsymbol{\xi})\|_W < \infty \text{ and } w|_{\partial D\times\Gamma} = 0 \right\},$$

where

$$H^1_0(D) := \left\{ v \in L^2(D) \middle| \gamma(v) = 0,\, \partial v/\partial x_i \in L^2(D), i = 1,\ldots,d \right\},$$

and the norm $\|\cdot\|_W$ is defined by

$$\|w(\boldsymbol{x},\boldsymbol{\xi})\|_W^2 := \int_\Gamma \rho(\boldsymbol{\xi}) \int_D \nabla w \cdot \nabla w \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}.$$

Following [10,20,27], the variational formulation of Eqs. (1) and (2) with respect to the inner product $(\,\cdot\,,\,\cdot\,)_{L^2\otimes L^2_\rho}$ can be written as:

find $u(\boldsymbol{x},\boldsymbol{\xi}) \in W$ such that

$$\mathbb{E}\left[\int_D \left(\nabla u \cdot \nabla w - \kappa^2 uw\right) \mathrm{d}\boldsymbol{x}\right] = \mathbb{E}\left[\int_D fw \,\mathrm{d}\boldsymbol{x}\right],\ \forall w(\boldsymbol{x},\boldsymbol{\xi}) \in W. \tag{7}$$

2.2 Discretization with Generalized Polynomial Chaos Approximations

To obtain the discrete version of Eq. (7), we need to introduce a finite dimensional subspace of $L^2(D)\otimes L^2_\rho(\Gamma)$ and find an approximation located inside. First, the finite element approximation [28] is considered for physical approximation in this work. The finite element approximation space is denoted by $V_h := \mathrm{span}\{v_s(\boldsymbol{x})\}_{s=1}^{N_h} \subset H_0^1(D)$, where $\{v_s(\boldsymbol{x})\}_{s=1}^{N_h}$ are the standard trial (test) functions, h denotes the mesh size, and N_h is the finite element degrees of freedom. For stochastic approximation, we consider the generalized polynomial chaos (gPC) approximation [29] (and see [9,30] for polynomial chaos methods). The gPC space is denoted by $S_p := \mathrm{span}\{\Phi_i(\boldsymbol{\xi})\}_{i=1}^{N_p} \subset L^2_\rho(\Gamma)$, where $\{\Phi_i(\boldsymbol{\xi})\}_{i=1}^{N_p}$ includes orthonormal polynomials with respect to the inner product $(\,\cdot\,,\,\cdot\,)_{L^2_\rho}$; that is

$$(\Phi_i(\boldsymbol{\xi}),\Phi_j(\boldsymbol{\xi}))_{L^2_\rho} = \int_\Gamma \rho(\boldsymbol{\xi})\Phi_i(\boldsymbol{\xi})\Phi_j(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \delta_{ij},$$

where δ_{ij} is the Kronecker's delta. As usual, $\{\Phi_i(\boldsymbol{\xi})\}_{i=1}^{N_p}$ consists of orthonormal polynomials with total degrees up to p, and p is referred to as the gPC order. The finite dimensional subspaces of W are then defined as

$$W_{hp} := V_h \otimes S_p = \mathrm{span}\{v_s(\boldsymbol{x})\Phi_i(\boldsymbol{\xi}) | s = 1,\ldots,N_h, i = 1,\ldots,N_p\}.$$

To obtain the discrete version of Eqs. (1) and (2), we write the overall (gPC finite element) approximation of $u(\boldsymbol{x},\boldsymbol{\xi})$ in W_{hp} as

$$u_{hp}(\boldsymbol{x},\boldsymbol{\xi}) := \sum_{i=1}^{N_p} u_h^{(i)}(\boldsymbol{x})\Phi_i(\boldsymbol{\xi}) = \sum_{i=1}^{N_p}\sum_{s=1}^{N_h} u_{is} v_s(\boldsymbol{x})\Phi_i(\boldsymbol{\xi}). \tag{8}$$

In Eq. (8), the unknown coefficients u_{is} for $i = 1,\ldots,N_p$ and $s = 1,\ldots,N_h$, are determined by solving the discrete version of the stochastic Helmholtz problem Eqs. (1) and (2), which is formulated through substituting Eq. (8) into Eq. (7) and restricting the test functions to the subspace W_{hp}. Details of the corresponding linear system are given in Section 3.1.

2.3 Multi-Element Generalized Polynomial Chaos Methods

To achieve high stochastic accuracy for the gPC finite element approximation [Eq. (8)], there are two standard strategies. The first is to take a high gPC order p in Eq. (8); the second is to decompose the parameter space Γ into subspaces, and construct gPC approximations locally in each subspace with relatively low orders. As discussed in [18], the second strategy is especially efficient when the exact solution to approximate has discontinuity or singularity with respect to the random inputs. For this purpose, the studies [26,31] develop the stochastic Galerkin finite element method, and the works [18,32,33] develop the ME-gPC method. Since the stochastic Helmholtz problem considered in this paper is close to a resonant frequency (that is singularity), we consider the second strategy and especially focus on the ME-gPC method. We review the ME-gPC method following the presentation in [18,32,33].

Let $\{B_k\}_{k=1}^M$ be a partition of Γ, where M is the number of elements of the partition; that is, $\Gamma = \bigcup_{k=1}^M B_k$ and for $k_1, k_2 \in \{1, \ldots, M\}$, $B_{k_1} \bigcap B_{k_2} = \emptyset$ if $k_1 \neq k_2$. Following [18], each element B_k is assumed to be in the form of $B_k = [\, a_1^{(k)}, b_1^{(k)}\,) \times [\, a_2^{(k)}, b_2^{(k)}\,) \times \cdots \times [\, a_N^{(k)}, b_N^{(k)}\,]$, where $a_i^{(k)} < b_i^{(k)}$ for $i = 1, \ldots, N$ and $k = 1, \ldots, M$.

With the partition $\{B_k\}_{k=1}^M$, $u(\boldsymbol{x}, \boldsymbol{\xi})$ can be represented as

$$u(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^{M} u(\boldsymbol{x}, \boldsymbol{\xi})|_{\boldsymbol{\xi} \in B_k} = \sum_{k=1}^{M} I_k(\boldsymbol{\xi}) u(\boldsymbol{x}, \boldsymbol{\xi}),$$

where

$$I_k(\boldsymbol{\xi}) = \begin{cases} 1, & \boldsymbol{\xi} \in B_k; \\ 0, & \boldsymbol{\xi} \notin B_k. \end{cases} \tag{9}$$

For convenience, we next define $\boldsymbol{\zeta}_k := \boldsymbol{\xi}$ for $\boldsymbol{\xi} \in B_k$, and $u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k) := u(\boldsymbol{x}, \boldsymbol{\zeta}_k) = u(\boldsymbol{x}, \boldsymbol{\xi})|_{\boldsymbol{\xi} \in B_k}$. It is clear that the range of $\boldsymbol{\zeta}_k$ is B_k, and it can be viewed as a new random variable subject to the probability density function $\rho^{(k)}(\boldsymbol{\zeta}_k) := \rho(\boldsymbol{\zeta}_k) / \int_{B_k} \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$. With this probability density function, the restriction of the solution $u(\boldsymbol{x}, \boldsymbol{\xi})$ to B_k, that is, $u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k) = u(\boldsymbol{x}, \boldsymbol{\xi})|_{\boldsymbol{\xi} \in B_k}$, can be approximated through the standard gPC method. In the following, $u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ is also referred to as a local solution on B_k. Similar to Eq. (8), the gPC approximation of $u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ is written as

$$u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k) \approx u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k) := \sum_{i=1}^{N_p^{(k)}} u_h^{(k,i)}(\boldsymbol{x}) \Phi_i^{(k)}(\boldsymbol{\zeta}_k) = \sum_{i=1}^{N_p^{(k)}} \sum_{s=1}^{N_h} u_{is}^{(k)} v_s(\boldsymbol{x}) \Phi_i^{(k)}(\boldsymbol{\zeta}_k), \tag{10}$$

where p is the highest total degree of the gPC basis, $N_p^{(k)} = (N+p)!/(N!p!)$ is the number of basis functions on the random element B_k, $\{v_s(\boldsymbol{x})\}_{s=1}^{N_h}$ is a basis of the physical approximation space V_h, and $\{\Phi_i^{(k)}(\boldsymbol{\zeta}_k)\}_{i=1}^{\infty}$ is an orthonormal basis of the following Hilbert space:

$$L^2_{\rho^{(k)}}(B_k) := \left\{ g : B_k \to \mathbb{R} \;\middle|\; \int_{B_k} \rho^{(k)}(\boldsymbol{\zeta}_k) g^2(\boldsymbol{\zeta}_k) \, \mathrm{d}\boldsymbol{\zeta}_k < \infty \right\},$$

equipped with the inner product

$$(\Phi_i^{(k)}(\boldsymbol{\zeta}_k), \Phi_j^{(k)}(\boldsymbol{\zeta}_k))_{L^2_{\rho^{(k)}}} := \int_{B_k} \rho^{(k)}(\boldsymbol{\zeta}_k) \Phi_i^{(k)}(\boldsymbol{\zeta}_k) \Phi_j^{(k)}(\boldsymbol{\zeta}_k) \mathrm{d}\boldsymbol{\zeta}_k. \tag{11}$$

The multi-element gPC approximation of $u(\boldsymbol{x}, \boldsymbol{\xi})$ can be given by

$$u_{Mhp}(\boldsymbol{x}, \boldsymbol{\xi}) := \sum_{k=1}^{M} I_k(\boldsymbol{\xi}) u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^{M} I_k(\boldsymbol{\xi}) u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k), \tag{12}$$

where $I_k(\boldsymbol{\xi})$ and $u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ are defined in Eqs. (9) and (10), respectively.

As discussed in [18,32], the partition of the parameter domain Γ is adaptively constructed; the procedure can be summarized as follows. The first step is to initialize a partition of Γ. The second step is to select the important elements (subdomains of the parameter domain) that need to be split. Third, for each selected element, we find the sensitive dimensions, and split the element by equally dividing these dimensions in the parameter domain. The second and the third steps are repeated until no important element can be found for splitting.

To select the important elements, we follow the adaptive criterion developed in [18], which is to assess the local decay rate of relative errors of gPC approximations in each element $k = 1, \ldots, M$. While the presentation in [18] focuses on the case that the gPC coefficients are independent of the physical variable $\boldsymbol{x}$, we modify the criterion as follows. Let $\underline{\boldsymbol{u}}_i^{(k)} := [u_{i1}^{(k)}, \ldots, u_{iN_h}^{(k)}] \in \mathbb{R}^{N_h}$ denote the vector collecting coefficients to represent $u_h^{(k,i)}(\boldsymbol{x})$ [see Eq. (10)]. We define the local decay rate of relative error of gPC approximation on each element $k = 1, \ldots, M$ as

$$\eta_k := \frac{\left\| \sum_{i=N_{p-1}^{(k)}+1}^{N_p^{(k)}} \underline{\boldsymbol{u}}_i^{(k)} \circ \underline{\boldsymbol{u}}_i^{(k)} \right\|}{\left\| \sum_{i=1}^{N_p^{(k)}} \underline{\boldsymbol{u}}_i^{(k)} \circ \underline{\boldsymbol{u}}_i^{(k)} \right\|},$$

where $\circ$ denotes the Hadamard product and $\|\cdot\|$ denotes the standard Euclidean norm. The element B_k is labeled for splitting if

$$\Pr(\xi \in B_k)\eta_k^{\alpha} > \theta_1, \tag{13}$$

where $0 < \alpha < 1$ and θ_1 is a given threshold and α is a prescribed constant [18].

To select the sensitive dimensions, we compute the sensitivity of each random dimension. Let

$$\underline{\boldsymbol{u}}_{j,p}^{(k)} := \left[u_{(j,p)1}^{(k)}, \ldots, u_{(j,p)N_h}^{(k)} \right]^T \in \mathbb{R}^{N_h}$$

denote the vector collecting coefficients to represent $u_h^{(k,j,p)}(\boldsymbol{x})$, where $u_h^{(k,j,p)}(\boldsymbol{x})$ denotes the function $u_h^{(k,i)}(\boldsymbol{x})$ in Eq. (10) associated with the random dimension ξ_j with polynomial order p. The sensitivity of each random dimension r_j for $j = 1, \ldots, N$ is defined as

$$r_j := \frac{\left\| \underline{\boldsymbol{u}}_{j,p}^{(k)} \circ \underline{\boldsymbol{u}}_{j,p}^{(k)} \right\|}{\left\| \sum_{s=N_{p-1}^{(k)}+1}^{N_p^{(k)}} \underline{\boldsymbol{u}}_s^{(k)} \circ \underline{\boldsymbol{u}}_s^{(k)} \right\|}. \tag{14}$$

The sensitive dimensions are then defined by the dimensions $j \in \{1, \ldots, N\}$ satisfying

$$r_j > \theta_2 \cdot \max_{s=1,\cdots,N}(r_s), \tag{15}$$

where $\theta_2 \in (0, 1)$ is another threshold parameter. For Eq. (15), it is clear that there exists at least one sensitive dimension for each selected element. After that, we split the element by equally dividing these sensitive dimensions.

3. IMPLEMENTATION AND PRECONDITIONED ITERATIVE SOLVERS

In this section, we first give detailed formulations of the linear systems associated with the ME-gPC finite element approximation. After that, we propose and analyze the corresponding mean-based preconditioning scheme.

3.1 Linear Systems Associated with GPC and ME-GPC Methods

For the standard gPC finite element approximation as studied in detail in [34], we substitute Eq. (8) to the variational formulation Eq. (7), let the test function be $w(\boldsymbol{x}, \xi) = v_s(\boldsymbol{x})\Phi_i(\xi) \in W_{hp}$, and obtain the discretized Helmholtz equation, that is, the linear system

$$\boldsymbol{A}\underline{\boldsymbol{u}} = \underline{\boldsymbol{b}}, \tag{16}$$

where

$$\boldsymbol{A} = \boldsymbol{G}_{00} \otimes \boldsymbol{K} - \sum_{l=0}^{N} \sum_{m=0}^{N} \boldsymbol{G}_{lm} \otimes \boldsymbol{M}_{lm}, \tag{17}$$

$$\underline{\boldsymbol{b}} = \sum_{m=0}^{N} \underline{\boldsymbol{h}}_m \otimes \underline{\boldsymbol{f}}_m. \tag{18}$$

In Eqs. (17) and (18), $\otimes$ denotes the Kronecker tensor product and

$$\boldsymbol{K}(s,t) = \int_D \nabla v_s \cdot \nabla v_t \, \mathrm{d}\boldsymbol{x}, \tag{19}$$

$$\underline{\boldsymbol{f}}_m(t) = \int_D f_m v_t \,\mathrm{d}\boldsymbol{x},\ \boldsymbol{M}_{lm}(s,t) = \int_D \kappa_l \kappa_m v_s v_t \,\mathrm{d}\boldsymbol{x}, \tag{20}$$

$$\underline{\boldsymbol{h}}_m(i) = \mathbb{E}[\xi_m \Phi_i(\boldsymbol{\xi})],\ \boldsymbol{G}_{lm}(i,j) = \mathbb{E}[\xi_l \xi_m \Phi_i(\boldsymbol{\xi})\Phi_j(\boldsymbol{\xi})], \tag{21}$$

where $l, m = 0, 1, \ldots, N$; $j, k = 1, \ldots, N_p$, and $s, t = 1, \ldots, N_h$. For a uniformly distributed $\boldsymbol{\xi} \in [-1,1]^N$, $\underline{\boldsymbol{h}}_m$ and $\boldsymbol{G}_{lm}$ in Eq. (21) are given analytically in [34].

To obtain the ME-gPC finite element approximation $u_{Mhp}(\boldsymbol{x}, \boldsymbol{\xi})$ in Eq. (12) [note that $\boldsymbol{\xi}$ is involved in the problem through Eq. (3)], approximations of the local solutions $u^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ for $k = 1, \ldots, M$ need to be computed. The framework for computing each local approximation, $u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ in Eq. (10), is almost the same as that for computing the standard gPC finite element approximation Eq. (8), except for the following modifications. We denote the linear system for computing $u_{hp}^{(k)}(\boldsymbol{x}, \boldsymbol{\zeta}_k)$ by

$$\boldsymbol{A}_k \underline{\boldsymbol{u}}_k = \underline{\boldsymbol{b}}_k, \tag{22}$$

where

$$\boldsymbol{A}_k = \boldsymbol{G}_{00}^{(k)} \otimes \boldsymbol{K} - \sum_{l=0}^{N}\sum_{m=0}^{N} \boldsymbol{G}_{lm}^{(k)} \otimes \boldsymbol{M}_{lm}, \tag{23}$$

$$\underline{\boldsymbol{b}}_k = \sum_{m=0}^{N} \underline{\boldsymbol{h}}_m^{(k)} \otimes \underline{\boldsymbol{f}}_m. \tag{24}$$

In Eqs. (23) and (24), $\boldsymbol{K}$, $\boldsymbol{M}_{lm}$, and $\underline{\boldsymbol{f}}_m$ are defined in Eqs. (19) and (20), but $\underline{\boldsymbol{h}}_m^{(k)}$ and $\boldsymbol{G}_{lm}^{(k)}$ are modified as

$$\underline{\boldsymbol{h}}_m^{(k)}(i) = \mathbb{E}\Big[\zeta_m^{(k)} \Phi_i^{(k)}(\boldsymbol{\zeta}_k)\Big], \quad \boldsymbol{G}_{lm}^{(k)}(i,j) = \mathbb{E}\Big[\zeta_l^{(k)} \zeta_m^{(k)} \Phi_i^{(k)}(\boldsymbol{\zeta}_k)\Phi_j^{(k)}(\boldsymbol{\zeta}_k)\Big],$$

where we note that $\boldsymbol{\zeta}_k = \left[\zeta_1^{(k)}, \ldots, \zeta_N^{(k)}\right]^T$, and for an arbitrary random function $g(\boldsymbol{\zeta}_k)$, $\mathbb{E}[g(\boldsymbol{\zeta}_k)] := \int_{B_k} \rho^{(k)}(\boldsymbol{\zeta}_k)\, g(\boldsymbol{\zeta}_k)\mathrm{d}\boldsymbol{\zeta}_k$.

When considering independent uniform random inputs $\boldsymbol{\xi}$, we can take the advantage of the Legendre-chaos on every random element B_k to result in efficient implementation, which is discussed in the following. To simplify the presentation, in this section we set $\Gamma = [-1,1]^N$, $\rho(\boldsymbol{\xi}) = (1/2)^N$, $B_k = \left[a_1^{(k)}, b_1^{(k)}\right) \times \left[a_2^{(k)}, b_2^{(k)}\right) \times \cdots \times \left[a_N^{(k)}, b_N^{(k)}\right]$, $\boldsymbol{a}_k = \left[a_1^{(k)}, \ldots, a_N^{(k)}\right]^T$, $\boldsymbol{b}_k = \left[b_1^{(k)}, \ldots, b_N^{(k)}\right]^T$, and recall that the inner products $(\,\cdot\,,\,\cdot\,)_{L^2_\rho}$ and $(\,\cdot\,,\,\cdot\,)_{L^2_{\rho^{(k)}}}$ are defined in Eqs. (6) and (11), respectively. In this setting, we have the following theorem.

Theorem 1. *Assume that $\boldsymbol{\xi}$ is uniformly distributed in $\Gamma = [-1,1]^N$. Given an orthonormal basis $\left\{\Phi_j^{(k)}(\boldsymbol{\zeta}_k)\right\}_{j=1}^{\infty}$ of the Hilbert space $L^2_{\rho^{(k)}}(B_k)$, $\left\{\Phi_j^{(k)}(g_k(\boldsymbol{\xi}))\right\}_{j=1}^{\infty}$ forms an orthonormal basis of $L^2_\rho(\Gamma)$, where g_k is a bijection from Γ to the closure of B_k and is defined as*

$$g_k(\boldsymbol{\xi}) := \frac{1}{2}\mathrm{diag}(\boldsymbol{b}_k - \boldsymbol{a}_k)\boldsymbol{\xi} + \frac{1}{2}(\boldsymbol{b}_k + \boldsymbol{a}_k),\ \forall \boldsymbol{\xi} \in \Gamma.$$

Proof. Our proof proceeds through the following two steps: the first is to show that $\left\{\Phi_j^{(k)}(g_k(\boldsymbol{\xi}))\right\}_{j=1}^{\infty}$ is an orthonormal system, and the second is to show that $\left\{\Phi_j^{(k)}(g_k(\boldsymbol{\xi}))\right\}_{j=1}^{\infty}$ is complete.

We first prove the orthogonality of $\left\{\Phi_j^{(k)}(g_k(\boldsymbol{\xi}))\right\}_{j=1}^{\infty}$. As given, $\left\{\Phi_j^{(k)}(\boldsymbol{\zeta}_k)\right\}_{j=1}^{\infty}$ is an orthonormal basis with respect to the inner product $(\,\cdot\,,\,\cdot\,)_{L^2_{\rho^{(k)}}}$:

$$\int_{B_k} \rho^{(k)}(\boldsymbol{\zeta}_k)\Phi_i^{(k)}(\boldsymbol{\zeta}_k)\Phi_j^{(k)}(\boldsymbol{\zeta}_k)\mathrm{d}\boldsymbol{\zeta}_k = \delta_{ij}. \tag{25}$$

Substituting the coordinate transformation $\zeta_k = g_k(\xi)$ into Eq. (25), we have

$$\begin{aligned}\delta_{ij} &= \int_{B_k} \rho^{(k)}(\zeta_k)\Phi_i^{(k)}(\zeta_k)\Phi_j^{(k)}(\zeta_k)\mathrm{d}\zeta_k \\ &= \int_{\Gamma} \rho^{(k)}(g_k(\xi))\Phi_i^{(k)}(g_k(\xi))\Phi_j^{(k)}(g_k(\xi))\det\left(\frac{\partial\zeta_k}{\partial\xi}\right)\mathrm{d}\xi.\end{aligned}$$

Since $\det(\partial\zeta_k/\partial\xi) = \prod_{k=1}^{N}[(b_k - a_k)/2]$ and $\rho^{(k)}(\zeta_k) = \prod_{k=1}^{N}[1/(b_k - a_k)]$, we have

$$\int_{\Gamma} \rho(\xi)\Phi_i^{(k)}(g_k(\xi))\Phi_j^{(k)}(g_k(\xi))\,\mathrm{d}\xi = \delta_{ij},$$

which means that $\left\{\Phi_j^{(k)}(g_k(\xi))\right\}_{j=1}^{\infty}$ is an orthonormal system in $L^2_\rho(\Gamma)$.

We next prove the completeness of $\left\{\Phi_j^{(k)}(g_k(\xi))\right\}_{j=1}^{\infty}$. The Fourier coefficients of $f(\xi) \in L^2_\rho(\Gamma)$ are defined as

$$C^f_\rho(j) := (f(\xi), \Phi_j^{(k)}(g_k(\xi)))_{L^2_\rho}\ j = 1, 2, \cdots.$$

If the Parseval's identity holds for any $f(\xi) \in L^2_\rho(\Gamma)$,

$$(f(\xi), f(\xi))_{L^2_\rho} = \sum_{j=1}^{\infty} |C^f_\rho(j)|^2, \tag{26}$$

then $\{\Phi_j^{(k)}(g_k(\xi))\}_{j=1}^{\infty}$ is complete [35]. By changing variables, $\xi = g_k^{-1}(\zeta_k)$, we have

$$\begin{aligned}(f(\xi), f(\xi))_{L^2_\rho} &= \int_{\Gamma} \rho(\xi) f^2(\xi)\,\mathrm{d}\xi \\ &= \int_{B_k} \rho(g_k^{-1}(\zeta_k)) f^2(g_k^{-1}(\zeta_k))\det\left(\frac{\partial\xi}{\partial\zeta_k}\right)\mathrm{d}\zeta_k.\end{aligned}$$

As $\det(\partial\xi/\partial\zeta_k) = \prod_{k=1}^{N}[2/(b_k - a_k)]$ and $\rho(\xi) = \prod_{k=1}^{N}[1/2]$, we then have

$$\begin{aligned}(f(\xi), f(\xi))_{L^2_\rho} &= \int_{B_k} \rho^{(k)}(\zeta_k) f^2(g_k^{-1}(\zeta_k))\,\mathrm{d}\zeta_k \\ &= \left(f\left(g_k^{-1}(\zeta_k)\right), f\left(g_k^{-1}(\zeta_k)\right)\right)_{L^2_{\rho^{(k)}}}.\end{aligned} \tag{27}$$

On the other hand, each Fourier coefficient $C^f_\rho(j)$ can be rewritten as

$$\begin{aligned}C^f_\rho(j) &= (f(\xi), \Phi_j^{(k)}(g_k(\xi)))_{L^2_\rho} \\ &= \int_{\Gamma} \rho(\xi) f(\xi)\Phi_j^{(k)}(g_k(\xi))\,\mathrm{d}\xi \\ &= \int_{B_k} \rho(g_k^{-1}(\zeta_k)) f(g_k^{-1}(\zeta_k))\Phi_j^{(k)}(\zeta_k)\det\left(\frac{\partial\xi}{\partial\zeta_k}\right)\mathrm{d}\zeta_k \\ &= \int_{B_k} \rho^{(k)}(\zeta_k) f(g_k^{-1}(\zeta_k))\Phi_j^{(k)}(\zeta_k)\,\mathrm{d}\zeta_k \\ &= \left(f(g_k^{-1}(\zeta_k)), \Phi_j^{(k)}(\zeta_k)\right)_{L^2_{\rho^{(k)}}}.\end{aligned}$$

This means $C_{\rho}^{f}(j)$ for $j = 1, \ldots,$ are the Fourier coefficients of $f(g_k^{-1}(\zeta_k)) \in L^2_{\rho^{(k)}}(B_k)$. It is obvious that $\{\Phi_j^{(k)}(\zeta_k)\}_{j=1}^{\infty}$ is complete, since it is an orthonormal basis. By the Parseval's identity, we have

$$\big(f\big(g_k^{-1}(\zeta_k)\big), f\big(g_k^{-1}(\zeta_k)\big)\big)_{L^2_{\rho^{(k)}}} = \sum_{j=1}^{\infty} |C_{\rho}^{f}(j)|^2. \tag{28}$$

Combining Eqs. (27) and (28) establishes the Parseval's identity [Eq. (26)], which means that $\{\Phi_j^{(k)}(g_k(\boldsymbol{\xi}))\}_{j=1}^{\infty}$ is complete. □

For uniformly distributed random inputs $\boldsymbol{\xi}$, using Theorem 1, $\underline{\boldsymbol{h}}_m^{(k)}$ and $\boldsymbol{G}_{lm}^{(k)}$ can be constructed as

$$\underline{\boldsymbol{h}}_0^{(k)} = \underline{\boldsymbol{h}}_0; \tag{29}$$

$$\underline{\boldsymbol{h}}_m^{(k)} = \frac{b_m^{(k)} - a_m^{(k)}}{2}\underline{\boldsymbol{h}}_m + \frac{b_m^{(k)} + a_m^{(k)}}{2}\underline{\boldsymbol{h}}_0,\ m \neq 0; \tag{30}$$

$$\boldsymbol{G}_{00}^{(k)} = \boldsymbol{G}_{00}; \tag{31}$$

$$\boldsymbol{G}_{0m}^{(k)} = \boldsymbol{G}_{m0}^{(k)} = \frac{b_m^{(k)} - a_m^{(k)}}{2}\boldsymbol{G}_{m0} + \frac{b_m^{(k)} + a_m^{(k)}}{2}\boldsymbol{G}_{00},\ m \neq 0; \tag{32}$$

$$\begin{aligned}\boldsymbol{G}_{lm}^{(k)} = {} & \frac{(b_m^{(k)} - a_m^{(k)})(b_l^{(k)} - a_l^{(k)})}{4}\boldsymbol{G}_{ml} + \frac{(b_m^{(k)} - a_m^{(k)})(b_l^{(k)} + a_l^{(k)})}{4}\boldsymbol{G}_{m0} \\ & + \frac{(b_m^{(k)} + a_m^{(k)})(b_l^{(k)} - a_l^{(k)})}{4}\boldsymbol{G}_{l0} + \frac{(b_m^{(k)} + a_m^{(k)})(b_l^{(k)} + a_l^{(k)})}{4}\boldsymbol{G}_{00},\ l, m \neq 0,\end{aligned} \tag{33}$$

where $\underline{\boldsymbol{h}}_m$ and $\boldsymbol{G}_{lm}$ are defined in Eq. (21), and their analytic expressions are given in [34].

To summarize the above procedure, when the random inputs $\boldsymbol{\xi}$ are independently and uniformly distributed in $\Gamma = [-1, 1]^N$, we can first compute the matrices and vectors defined in Eqs. (20) and (21) for the standard gPC finite element approximation, and then for each linear system [Eq. (22)] associated with each local random element B_k for $k \in \{1, \ldots, M\}$ in the ME-gPC setting, we can cheaply assemble it using Eqs. (23) and (24) and Eqs. (29)–(33).

3.2 Mean-Based Preconditioning

To obtain the ME-gPC finite element approximation [Eq. (12)], linear systems $\boldsymbol{A}_k\underline{\boldsymbol{u}}_k = \underline{\boldsymbol{b}}_k$ for $k = 1, \ldots, M$ need to be solved. Since these linear systems are large but sparse, we consider iterative methods to solve them, and we in particular use the induced dimension reduction (IDR(s)) method [36,37]. To result in a small number of iterations, preconditioners are typically required in iterative methods. That is, instead of solving the original linear system [Eq. (22)], we solve the following right preconditioned linear system:

$$\boldsymbol{A}_k\boldsymbol{P}_k^{-1}\tilde{\underline{\boldsymbol{u}}}_k = \underline{\boldsymbol{b}}_k, \quad \text{with} \quad \underline{\boldsymbol{u}}_k = \boldsymbol{P}_k^{-1}\tilde{\underline{\boldsymbol{u}}}_k \tag{34}$$

where the nonsingular matrix $\boldsymbol{P}_k$ is called a preconditioner.

Following the mean-based preconditioning scheme originally proposed for polynomial chaos methods [20,38, 39], we design a mean-based preconditioner for these linear systems arising from the ME-gPC finite element approximation for stochastic Helmholtz problems. The mean-based preconditioner is constructed through the discrete version of Eqs. (1) and (2) corresponding to the mean of the input. We denote the expectation of ζ_k by $\zeta_k^{(0)}$, $\zeta_k^{(0)} := \mathbb{E}[\zeta_k] = \int_{B_k} \rho^{(k)}(\zeta_k)\zeta_k \,\mathrm{d}\zeta_k$. Then the mean-based preconditioner is given by

$$\boldsymbol{P}_k = \boldsymbol{G}_{00} \otimes (\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}),$$

where $\boldsymbol{G}_{00}$, $\boldsymbol{K}$ are defined in Eqs. (19) and (21), and $\boldsymbol{M}_{\text{pre}}$ is computed through

$$\boldsymbol{M}_{\text{pre}}(s,t) = \int_D \kappa^2\Big(\boldsymbol{x}, \zeta_k^{(0)}\Big) v_s v_t \,\mathrm{d}\boldsymbol{x},\ s, t = 1, \ldots, N_h.$$

Note that $\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}$ is the coefficient matrix associated with the discrete version of Eqs. (1) and (2) with $\xi = \zeta_k^{(0)}$. In addition, $\boldsymbol{P}_k$ is the same as $\boldsymbol{A}_k$ if the variances of the random inputs ζ_k are zeros; detailed analysis of how $\boldsymbol{P}_k$ approximates $\boldsymbol{A}_k$ is discussed in Section 3.3. In particular, when the random inputs ξ are independently and uniformly distributed in $[-1,1]^N$, the mean-based preconditioner can be constructed as

$$\boldsymbol{M}_{\text{pre}} = \boldsymbol{M}_{00} + \sum_{m=1}^{N} (b_m^{(k)} + a_m^{(k)}) \boldsymbol{M}_{0m} + \sum_{l=1}^{N} \sum_{m=1}^{N} \frac{(b_l^{(k)} + a_l^{(k)})(b_m^{(k)} + a_m^{(k)})}{4} \boldsymbol{M}_{lm}.$$

In addition, to start the iterative solving procedure, we set an initial guess $\underline{\boldsymbol{u}}_k^{(0)} = \boldsymbol{P}_k^{-1} \underline{\boldsymbol{b}}_k$.

At each iteration step, for the purpose of preconditioning, we solve a linear system of the form

$$\boldsymbol{P}_k \hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}, \tag{35}$$

where

$$\hat{\boldsymbol{x}} = \begin{bmatrix} \hat{\boldsymbol{x}}_1 \\ \hat{\boldsymbol{x}}_2 \\ \vdots \\ \hat{\boldsymbol{x}}_{N_p} \end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{\boldsymbol{y}}_1 \\ \hat{\boldsymbol{y}}_2 \\ \vdots \\ \hat{\boldsymbol{y}}_{N_p} \end{bmatrix}, \text{ where } \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i \in \mathbb{R}^{N_h} \text{ for } i = 1, \ldots, N_p.$$

Instead of forming $\boldsymbol{P}_k$ explicitly and solving the linear system [Eq. (35)] directly, we can take advantages of the structure of Kronecker tensor product. Since $\boldsymbol{G}_{00}$ is symmetric, solving Eq. (35) is equivalent to solving the following problem:

$$(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}) \hat{\boldsymbol{X}} \boldsymbol{G}_{00} = \hat{\boldsymbol{Y}}, \tag{36}$$

where

$$\hat{\boldsymbol{X}} = \left[\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{N_p}\right] \text{ and } \hat{\boldsymbol{Y}} = \left[\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_{N_p}\right].$$

To compute the solution of Eq. (36), we only need to solve N_p linear systems with size $N_h \times N_h$, which is much cheaper than solving Eq. (22) (whose size is $N_h N_p \times N_h N_p$). See [20,34] for more details.

3.3 Improved Efficiency of the Iterative Linear Solvers for ME-GPC Systems

As studied in [34], the mean-based preconditioner for the linear system of the gPC discretization [i.e., Eq. (16)] can become inefficient when the underlying stochastic Helmholtz problem is close to a resonant frequency. This is because the coefficient matrix of such a linear system, either unpreconditioned or preconditioned, have some rather small eigenvalues, such that Krylov subspace methods converge slowly. To tackle this difficulty, we show that the linear systems arising from the ME-gPC discretization are much easier to solve by a Krylov subspace method with mean-based preconditioning than their counterpart from the standard gPC discretization.

To analyze such a difference, consider the linear system $\boldsymbol{A}_k \underline{\boldsymbol{u}}_k = \underline{\boldsymbol{b}}_k$ obtained from the ME-gPC discretization. Let $\boldsymbol{A}_k = \boldsymbol{P}_k + \mathbf{Q}_{\zeta_k}$, where $\boldsymbol{P}_k$ is the "mean" part of $\boldsymbol{A}_k$ used as the mean-based preconditioner, and $\mathbf{Q}_{\zeta_k}$ is the "stochastic" part of $\boldsymbol{A}_k$, which vanishes if all random variables ζ_k take their mean values with probability 1.

Note that the moduli of eigenvalues $\lambda(\xi)$ in Eq. (4) for all $\xi \in \Gamma$ are assumed to be greater than ϵ, where $\epsilon > 0$ is a constant, and recall that

$$\boldsymbol{P}_k = \boldsymbol{G}_{00} \otimes (\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}), \tag{37}$$

where $\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}$ is the coefficient matrix associated with the discrete version of Eqs. (1) and (2) with $\xi = \zeta_k^{(0)}$. For the eigenvalue problem [Eq. (4)] associated with $\xi = \zeta_k^{(0)}$, it can be discretized into

$$(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}) \boldsymbol{u} = \lambda \boldsymbol{B} \boldsymbol{u}, \tag{38}$$

where $\boldsymbol{B}(s,t) = \int_D v_s v_t \, \mathrm{d}\boldsymbol{x}, \ s,t = 1, \ldots, N_h$. Since the eigenvalues of Eq. (38) are approximations of Eq. (4) associated with $\xi = \zeta_k^{(0)}$, we assume $\sigma_{\min}\left(\boldsymbol{B}^{-1}(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}})\right) > \epsilon$. With this assumption, our main insight in this section is stated in the lemma and the theorem in the following.

Lemma 1. *Assume that the eigenvalue problem [Eq. (4)] associated with $\xi = \zeta_k^{(0)}$ is discretized into Eq. (38), and suppose that $\sigma_{min}\big(\boldsymbol{B}^{-1}(\boldsymbol{K} - \boldsymbol{M}_{pre})\big) > \epsilon$, where σ_{min} refers to the smallest singular value of the matrix under discussion. Let $\boldsymbol{P}$ and $\boldsymbol{P}_k$ be the mean-based preconditioners for the original system $\boldsymbol{A}\underline{\boldsymbol{u}} = \underline{\boldsymbol{b}}$ based on standard gPC and the system $\boldsymbol{A}_k\underline{\boldsymbol{u}}_k = \underline{\boldsymbol{b}}_k$ based on ME-gPC, suppose that both of them are nonsingular. Then, there exists a constant $C = \sigma_{min}(\boldsymbol{P})/(\sigma_{min}(\boldsymbol{G}_{00})\sigma_{min}(\boldsymbol{B})\epsilon)$, such that*

$$\sigma_{min}(\boldsymbol{P}) \le C\sigma_{min}(\boldsymbol{P}_k).$$

Proof. Since $\boldsymbol{P}_k$ is nonsingular and note Eq. (37), the matrix $\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}$ is invertible. Thus, we have

$$\|(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}})^{-1}\|_2 = \|(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}})^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\|_2 \le \|(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}})^{-1}\boldsymbol{B}\|_2\|\boldsymbol{B}^{-1}\|_2,$$

where $\|\cdot\|_2$ is the spectral norm. That is

$$\sigma_{\min}(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}) \ge \sigma_{\min}\big(\boldsymbol{B}^{-1}(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}})\big)\sigma_{\min}(\boldsymbol{B}) > \sigma_{\min}(\boldsymbol{B})\epsilon.$$

Thus

$$\sigma_{\min}(\boldsymbol{P}_k) = \sigma_{\min}(\boldsymbol{G}_{00})\sigma_{\min}(\boldsymbol{K} - \boldsymbol{M}_{\text{pre}}) > \sigma_{\min}(\boldsymbol{G}_{00})\sigma_{\min}(\boldsymbol{B})\epsilon,$$

and then there exists a constant $C > 0$ such that

$$\sigma_{\min}(\boldsymbol{P}) \le C\sigma_{\min}(\boldsymbol{P}_k),$$

where

$$C = \frac{\sigma_{\min}(\boldsymbol{P})}{\sigma_{\min}(\boldsymbol{G}_{00})\sigma_{\min}(\boldsymbol{B})\epsilon}.$$

□

Theorem 2. *Under the assumption of Lemma 1, we further assume that each random variable ξ_l obeys a uniform distribution on $\Gamma_l = [-1, 1]$ for $1 \le l \le N$ and consequently ζ_k obeys a uniform distribution on each $\left[a_l^{(k)}, b_l^{(k)}\right]$ for $1 \le k \le M$. For a given k, suppose that the stochastic element B_k is sufficiently small; that is, $\max_{l=1}^{m}\left(b_l^{(k)} - a_l^{(k)}\right)$ is sufficiently small. Then, $\|\boldsymbol{A}_k\boldsymbol{P}_k^{-1} - I\|_2 < \|\boldsymbol{A}\boldsymbol{P}^{-1} - I\|_2$ where $\|\cdot\|_2$ is the spectral norm.*

Proof. For the uniform distribution, from Eqs. (23) and (31)–(33), we have

$$\begin{aligned}
\boldsymbol{A}_k &= \boldsymbol{G}_{00}\otimes\boldsymbol{K} - \boldsymbol{G}_{00}\otimes\boldsymbol{M}_{00} - \sum_{m=1}^{N}\left(b_m^{(k)} + a_m^{(k)}\right)\boldsymbol{G}_{00}\otimes\boldsymbol{M}_{m0} - \sum_{m=1}^{N}\left(b_m^{(k)} - a_m^{(k)}\right)\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{m0} \\
&\quad - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} - a_m^{(k)}\right)\left(b_l^{(k)} - a_l^{(k)}\right)}{4}\boldsymbol{G}_{ml}\otimes\boldsymbol{M}_{lm} - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} - a_m^{(k)}\right)\left(b_l^{(k)} + a_l^{(k)}\right)}{4}\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{lm} \\
&\quad - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} + a_m^{(k)}\right)\left(b_l^{(k)} - a_l^{(k)}\right)}{4}\boldsymbol{G}_{l0}\otimes\boldsymbol{M}_{lm} - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} + a_m^{(k)}\right)\left(b_l^{(k)} + a_l^{(k)}\right)}{4}\boldsymbol{G}_{00}\otimes\boldsymbol{M}_{lm} \\
&= \boldsymbol{P}_k + \boldsymbol{Q}_{\zeta_k},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{Q}_{\zeta_k} &= -\sum_{m=1}^{N}\left(b_m^{(k)} - a_m^{(k)}\right)\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{m0} - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} - a_m^{(k)}\right)\left(b_l^{(k)} - a_l^{(k)}\right)}{4}\boldsymbol{G}_{ml}\otimes\boldsymbol{M}_{lm} \\
&\quad - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} - a_m^{(k)}\right)\left(b_l^{(k)} + a_l^{(k)}\right)}{4}\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{lm} - \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\left(b_m^{(k)} + a_m^{(k)}\right)\left(b_l^{(k)} - a_l^{(k)}\right)}{4}\boldsymbol{G}_{l0}\otimes\boldsymbol{M}_{lm}.
\end{aligned}$$

For a given k, let us denote $\vartheta_1 = \max_{m=1}^{N}\left(b_m^{(k)} - a_m^{(k)}\right)$, and $\vartheta_2 = \max_{m=1}^{N}\left|b_m^{(k)} + a_m^{(k)}\right|$. Note that $\vartheta_2 \le 2\max_{m=1}^{N}$ $\max\left\{|a_m^{(k)}|, |b_m^{(k)}|\right\} \le 2\max_{m=1}^{N}\max\{|\xi_m| : \xi_m \in \Gamma_m\}$ is uniformly bounded for all k.

It follows that

$$\begin{aligned}\|\boldsymbol{Q}_{\zeta_k}\|_2 &\le \vartheta_1 \sum_{m=1}^{N}\|\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{m0}\|_2 + \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\vartheta_1^2}{4}\|\boldsymbol{G}_{ml}\otimes\boldsymbol{M}_{lm}\|_2 \\ &+ \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\vartheta_1\vartheta_2}{4}\|\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{lm}\|_2 + \sum_{l=1}^{N}\sum_{m=1}^{N}\frac{\vartheta_1\vartheta_2}{4}\|\boldsymbol{G}_{l0}\otimes\boldsymbol{M}_{lm}\|_2 \\ &= \vartheta_1 \sum_{m=1}^{N}\|\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{m0}\|_2 + \frac{\vartheta_1^2}{4}\sum_{l=1}^{N}\sum_{m=1}^{N}\|\boldsymbol{G}_{ml}\otimes\boldsymbol{M}_{lm}\|_2 + \frac{\vartheta_1\vartheta_2}{2}\sum_{l=1}^{N}\sum_{m=1}^{N}\|\boldsymbol{G}_{m0}\otimes\boldsymbol{M}_{lm}\|_2,\end{aligned}$$

Since $\mathrm{Var}\left(\zeta_m^{(k)}\right) = (b_m - a_m)^2/12$, we have $\|\boldsymbol{Q}_{\zeta_k}\|_2 = O(\vartheta_1) = O\left(\left(\max_{i=1}^{N}\mathrm{Var}\left(\zeta_m^{(k)}\right)\right)^{1/2}\right)$, thanks to the (uniform) boundedness of ϑ_2.Therefore, $\lim_{|B_k|\to 0}\|\mathbf{Q}_{\zeta_k}\|_2 = \lim_{\mathrm{Var}(\zeta_k)\to 0}\|\mathbf{Q}_{\zeta_k}\|_2 = 0$. This pattern also holds for the linear system $\boldsymbol{A}\underline{\boldsymbol{u}} = \underline{\boldsymbol{b}}$ arising from the standard gPC discretization (a special case by setting $B_k = \Gamma$), where $\boldsymbol{A} = \boldsymbol{P} + \mathbf{Q}_{\boldsymbol{\xi}}$. Since $\boldsymbol{\xi} \in \Gamma = \cup_{k=1}^{M} B_k$, we have $\mathrm{Var}(\boldsymbol{\zeta}_k) \ll \mathrm{Var}(\boldsymbol{\xi})$ if $|B_k| \ll |\Gamma|$ and if $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}_k$ obey the same type of distribution. Here, we assume that $|B_k|$ is sufficiently small, and the smallest singular value of $\mathbf{Q}_{\boldsymbol{\xi}}$ is bounded away from zero. Since $\lim_{|B_k|\to 0}\|\mathbf{Q}_{\zeta_k}\|_2 = 0$, for a sufficiently small $|B_k|$, there is a sufficiently small $c > 0$, such that $\|\mathbf{Q}_{\zeta_k}\|_2 = \sigma_{\max}(\mathbf{Q}_{\zeta_k}) \le c\sigma_{\min}(\mathbf{Q}_{\boldsymbol{\xi}})$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and the smallest singular values of the relevant matrices, respectively.

The preconditioned coefficient matrices associated with the standard and the ME-gPC discretization are

$$\boldsymbol{A}\boldsymbol{P}^{-1} = (\boldsymbol{P} + \mathbf{Q}_{\boldsymbol{\xi}})\boldsymbol{P}^{-1} = I + \mathbf{Q}_{\boldsymbol{\xi}}\boldsymbol{P}^{-1},$$

and

$$\boldsymbol{A}_k\boldsymbol{P}_k^{-1} = (\boldsymbol{P}_k + \mathbf{Q}_{\boldsymbol{\zeta}_k})\boldsymbol{P}_k^{-1} = I + \mathbf{Q}_{\boldsymbol{\zeta}_k}\boldsymbol{P}_k^{-1},$$

respectively. By Lemma 1, we have $\sigma_{\min}(\boldsymbol{P}) \le C\sigma_{\min}(\boldsymbol{P}_k)$, that is, $\|\boldsymbol{P}_k^{-1}\|_2 \le C\|\boldsymbol{P}^{-1}\|_2$, where $C = \sigma_{\min}(\boldsymbol{P})$ $/(\sigma_{\min}(\boldsymbol{G}_{00})\sigma_{\min}(\boldsymbol{B})\epsilon)$. Thus, we have

$$\|\mathbf{Q}_{\zeta_k}\boldsymbol{P}_k^{-1}\|_2 \le \|\mathbf{Q}_{\zeta_k}\|_2\|\boldsymbol{P}_k^{-1}\|_2 \le C\|\mathbf{Q}_{\zeta_k}\|_2\|\boldsymbol{P}^{-1}\|_2 \le cC\sigma_{\min}(\mathbf{Q}_{\boldsymbol{\xi}})\|\boldsymbol{P}^{-1}\|_2 \le cC\|\mathbf{Q}_{\boldsymbol{\xi}}\boldsymbol{P}^{-1}\|_2.$$

Here, since $|B_k|$ is assumed to be sufficiently small, c is also sufficiently small, such that $cC < 1$. The theorem is thus established. □

The above theorem states that, under reasonable assumptions, the preconditioned coefficient matrix $\boldsymbol{A}_k\boldsymbol{P}_k^{-1}$ associated with ME-gPC is closer to the identity matrix than $\boldsymbol{A}\boldsymbol{P}^{-1}$ arising from standard gPC, such that Krylov subspace iterative linear solvers are likely to converge more rapidly for solving $\boldsymbol{A}_k\boldsymbol{P}_k^{-1}\tilde{\underline{\boldsymbol{u}}}_k = \underline{\boldsymbol{b}}_k$; see, for example, [40, Lecture 40]. In the next section, we provide numerical evidence to show that $\boldsymbol{A}_k\boldsymbol{P}_k^{-1}$ is indeed much closer to identity if $|B_k|$ is sufficiently small.

4. NUMERICAL STUDY

In this section, two test problems are studied. For both test problems, the physical domain considered is $[-1, 1]^2$, and we discretize in physical space using a bilinear rectangular finite element approximation [28,41] with a uniform 33×33 grid. The results of the ME-gPC finite element approximation [Eq. (12)] are presented in this section, while results of the standard gPC finite element approximation [Eq. (8)] are also presented for comparison. In the following, the gPC orders for standard gPC and ME-gPC finite element approximations are referred to as the global gPC order and the local gPC order, respectively.

Linear systems resulting from both gPC and ME-gPC are solved by the IDR(1) iterative method with the mean-based preconditioning scheme. The stopping criterion for the iterative methods is based on the relative residual $\|\boldsymbol{A}_k\underline{\boldsymbol{u}}_k^{(i)} - \underline{\boldsymbol{b}}_k\|/\|\underline{\boldsymbol{b}}_k\|$, where the superscript i denotes the iteration number. The iteration terminates when the relative residual is smaller than 10^{-8}.

4.1 One-Dimensional Random Input

In this test problem, the wave number and forcing term are given by

$$\kappa(\boldsymbol{x},\xi) = \kappa_0(\boldsymbol{x}) + \kappa_1(\boldsymbol{x})\xi,\ f(\boldsymbol{x},\xi) = \cos\Big(\frac{\pi x_1}{2}\Big)\cos\Big(\frac{\pi x_2}{2}\Big),$$

where $\kappa_0(\boldsymbol{x}) = \pi/\sqrt{2} + 0.401$, $\kappa_1 = 0.4$, and ξ is uniformly distributed on $\Gamma = [-1,1]$.

As discussed in Section 2, the stochastic Helmholtz problem Eqs. (1) and (2) is well posed if all eigenvalues of Eq. (4) are bounded away from zero, while the zero eigenvalue causes resonance. This paper focuses on the cases that have eigenvalues close to zero. For this test problem, the minimum absolute value of the eigenvalues is very small—that is, $\min_{\xi\in[-1,1]}|\lambda(\xi)| \approx 4.4\times 10^{-3}$. As discussed in our earlier work [34], directly applying the standard gPC with the mean-based preconditioning scheme is not efficient for this test problem, and in the following we present the results of ME-gPC combined with the mean-based preconditioning scheme.

To assess the accuracy of the ME-gPC finite element approximation, we define the following quantities to measure the relative errors in mean and variance function estimates,

$$E_{\text{mean}} := \|\mathbb{E}[u_{Mhp}] - \mathbb{E}[u_{\text{ref}}]\|_{L^2}/\|\mathbb{E}[u_{\text{ref}}]\|_{L^2}, \tag{39}$$

$$E_{\text{variance}} := \|\text{Var}(u_{Mhp}) - \text{Var}(u_{\text{ref}})\|_{L^2}/\|\text{Var}(u_{\text{ref}})\|_{L^2}, \tag{40}$$

where the reference solution u_{ref} is obtained through the global gPC method with total degree $p = 700$. Moreover, we also investigate the relative L^2 error, which is defined as

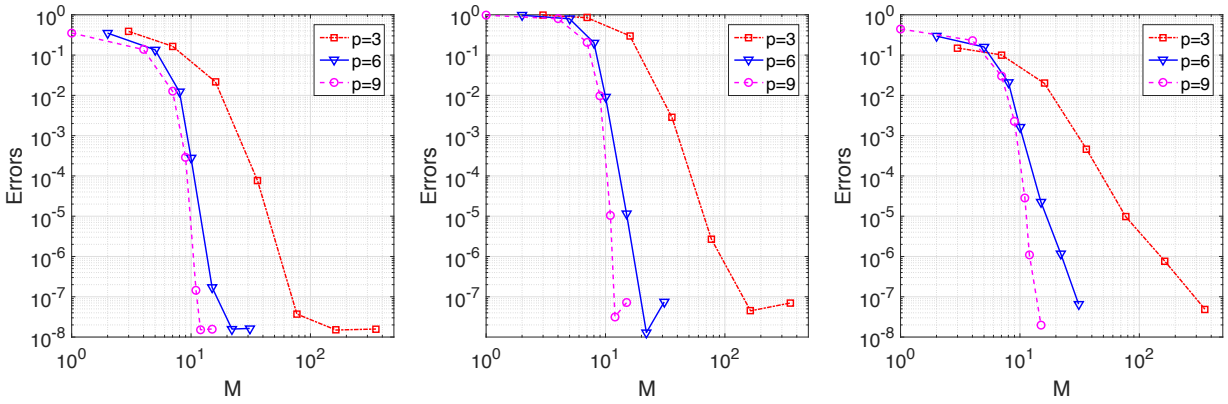
$$E_{L^2} := \left(\int_D\int_\Gamma (u_{Mhp} - u_{\text{ref}})^2\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}\right)^{1/2}\Big/\left(\int_D\int_\Gamma (u_{\text{ref}})^2\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}\right)^{1/2}. \tag{41}$$

The parameter α in Eq. (13) is set to 0.5 for all numerical studies in this paper (note that since the input parameter is one-dimensional, the threshold θ_2 for selecting dimensions is not involved in this test problem). In the same way, we assess the errors of the standard gPC through Eqs. (39)–(41) with u_{Mhp} replaced by u_{hp} [see Eq. (8)].

The errors in mean and variance function estimates and the relative L^2 errors of the ME-gPC finite element approximation are shown in Fig. 1. It is clear that, for a given local gPC order p, the errors decrease quickly as the number of stochastic elements M increases (the increasing in M is achieved through decreasing the threshold θ_1). In addition, for the same number of stochastic elements, it is not surprising that higher local gPC orders result in smaller errors.

To look more closely at the errors of ME-gPC, we show the mean, the variance, and the relative L^2 errors in Table 1, where different values of the threshold are considered. As shown in Table 1, for a given local gPC order, as the value of the threshold θ_1 decreases, the number of stochastic elements increases, which leads to the decrease of errors in mean and variance function estimates and the relative L^2 errors.

In Fig. 2, lengths of the stochastic elements generated in ME-gPC are shown by the heights of the rectangles. Since for $\xi = -1$, the minimum absolute value of the eigenvalues of Eq. (4) is about 4.4×10^{-3}; this small value leads to the Helmholtz problem close to resonance. It is clear that, there are a large number of stochastic elements close to $\xi = -1$, and the elements become smaller as ξ goes to -1, which are expected for resolving problems close to resonance. Figure 2 also shows the number of iterations to solve each linear system [Eq. (22)] of ME-gPC using IDR(1) with mean-based preconditioning. Note that the dark colors near the corner $[-1,0]^T$ in Figs. 2(c)–2(e) are caused by the dense rectangular boundaries to show the stochastic element sizes. In addition, it is not surprising that the smaller the threshold θ_1, the finer the partition of the stochastic domain in these pictures.



(a) (b) (c)

FIG. 1: Errors of ME-gPC, one-dimensional random input. (a) Mean errors, (b) variance errors, and (c) relative L^2 errors.

TABLE 1: Errors of ME-gPC in mean and variance function estimates and the relative L^2 errors, $\alpha = 0.5$, one-dimensional random input

		$\theta_1 = 10^{-2}$		$\theta_1 = 10^{-3}$		$\theta_1 = 10^{-4}$		$\theta_1 = 10^{-5}$	
		M	**Error**	M	**Error**	M	**Error**	M	**Error**
Mean errors	$p = 3$	7	1.66×10^{-1}	16	2.17×10^{-2}	36	7.71×10^{-5}	77	3.72×10^{-8}
	$p = 6$	5	1.35×10^{-1}	8	1.22×10^{-2}	10	2.76×10^{-4}	15	1.71×10^{-7}
	$p = 9$	4	1.37×10^{-1}	7	1.26×10^{-2}	9	2.87×10^{-4}	11	1.47×10^{-7}
Variance errors	$p = 3$	7	8.47×10^{-1}	16	3.00×10^{-1}	36	2.84×10^{-3}	77	2.72×10^{-6}
	$p = 6$	5	7.96×10^{-1}	8	2.03×10^{-1}	10	9.22×10^{-3}	15	1.15×10^{-5}
	$p = 9$	4	8.00×10^{-1}	7	2.08×10^{-1}	9	9.63×10^{-3}	11	1.04×10^{-5}
Relative L^2 errors	$p = 3$	7	1.01×10^{-1}	16	2.00×10^{-2}	36	4.68×10^{-4}	77	9.68×10^{-6}
	$p = 6$	5	1.59×10^{-1}	8	2.12×10^{-2}	10	1.60×10^{-3}	15	2.24×10^{-5}
	$p = 9$	4	2.31×10^{-1}	7	3.04×10^{-2}	9	2.27×10^{-3}	11	2.80×10^{-5}

Figure 3 shows the results of standard gPC for this test problem. From Fig. 3(a), to achieve an accuracy with small mean and variance errors, the standard gPC requires a high global gPC order p. For example, to obtain an approximation with variance error smaller than 10^{-7}, the standard gPC needs a global gPC order more than 400, while ME-gPC can also achieve the same accuracy with a local gPC order less than 10 as shown in Fig. 1(b) (although partitioning of the stochastic domain is required). From Fig. 3(b), the number of iterations for IDR(1) with mean-based preconditioning typically increases as the global gPC order increases, and for $p \geq 400$, more than 400 iterations are required. On the contrary, the local gPC order in ME-gPC is typically small (we only take at most $p = 9$ in this test problem), and the numbers of iterations are much smaller (at most 10 iterations for IDR(1) iterative method). We have also tested the IDR(1) method without preconditioning, which is referred to as plain IDR. We found that the relative residual of plain IDR can not research the tolerance 10^{-8} for 2000 iterations for gPC. For ME-gPC, plain IDR typically requires hundreds of iterations to reach the stopping criterion. As these numbers of iterations are much larger than IDR(1) with mean-based preconditioning, we report the results of only the preconditioned method in detail.

Figure 4 shows the six smallest magnitude eigenvalues of the coefficient matrix and the preconditioned coefficient matrix arising from gPC. It can be seen that the six smallest magnitude eigenvalues of the coefficient matrix become larger after preconditioning, which is expected for a preconditioned system. However, after preconditioning, the smallest magnitudes of the eigenvalues are still very small—around 10^{-4} to 10^{-2}, which implies that the mean-based preconditioner is not very efficient for the gPC system.

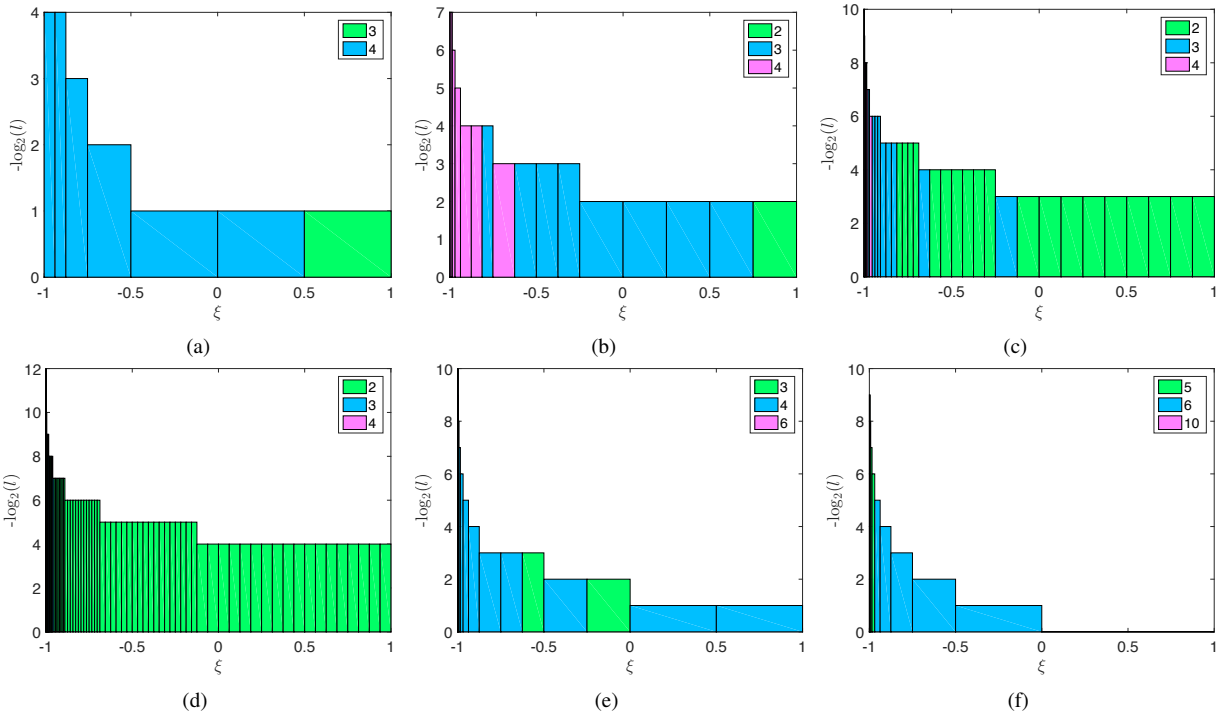


FIG. 2: Stochastic element sizes for different p and θ_1, where l is the length of the stochastic element and the numbers in the legend represent the number of iteration steps, one-dimensional random input. The dark colors near the color $[-1, 0]^T$ in (c), (d), and (e) are caused by the dense rectangle boundaries. (a) $p = 3, \theta_1 = 10^{-2}$, (b) $p = 3, \theta_1 = 10^{-3}$, (c) $p = 3, \theta_1 = 10^{-4}$, (d) $p = 3, \theta_1 = 10^{-5}$, (e) $p = 6, \theta_1 = 10^{-5}$, and (f) $p = 9, \theta_1 = 10^{-5}$

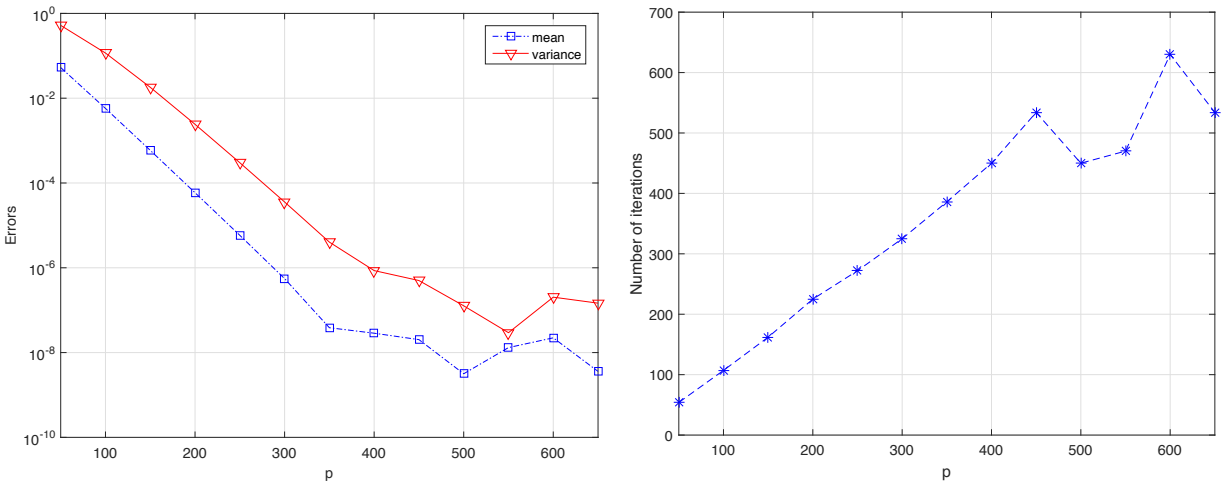


FIG. 3: Results of the standard gPC finite element approximation, one-dimensional random input. (a) Errors of standard gPC and (b) preconditioned IDR(1).

For comparison, we consider the eigenvalues of the coefficient matrix arising from ME-gPC. For each random element $k = 1, \ldots, M$, the six smallest magnitude eigenvalues are denoted by $\lambda_1^{(k)}, \ldots, \lambda_6^{(k)}$ with $|\lambda_1^{(k)}| \leq \ldots \leq |\lambda_6^{(k)}|$. Next, the suspicion element is defined to be

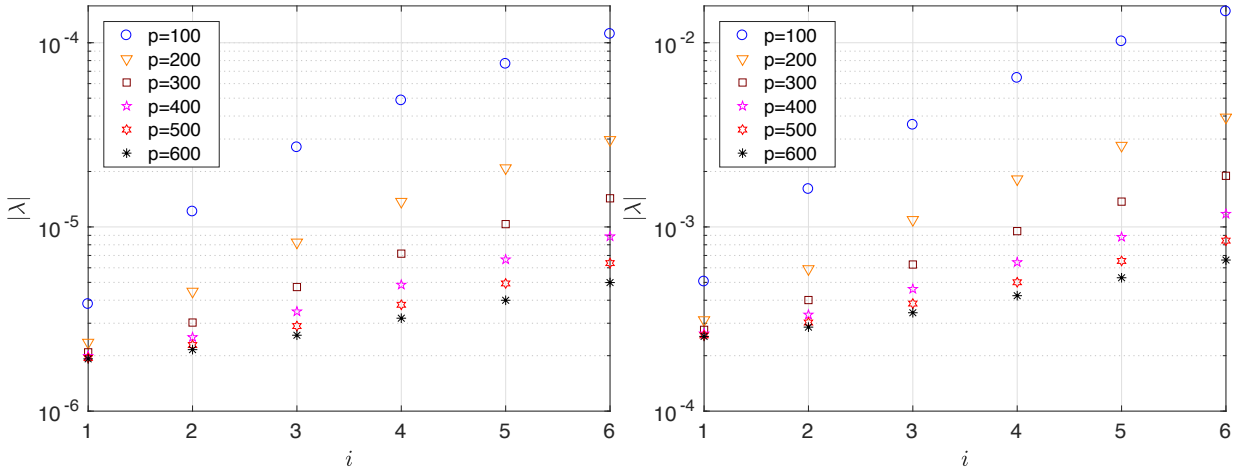


FIG. 4: The six smallest magnitude eigenvalues of the coefficient matrix (left) and the preconditioned coefficient matrix (right) arising from gPC, one-dimensional random input

$$k^* = \arg\min_k \sum_{i=1}^{6} |\lambda_i^{(k)}|.$$

Figure 5 shows the six smallest magnitude eigenvalues for this suspicion element k^*. As we have shown in Section 3.3, the preconditioned coefficient matrix arising from ME-gPC is closer to identity than that arising from standard gPC. As a result, the six smallest magnitude eigenvalues of the preconditioned coefficient matrix arising form ME-gPC are much farther from zero than those associated with standard gPC. This implies the mean-based preconditioned IDR(1) method can converge in much fewer iterations with ME-gPC (at most 10 iterations) than it does with standard gPC.

To compare overall performances of the standard gPC and the ME-gPC, we consider the CPU times of the implementations, and our results are obtained in MATLAB on a workstation with 2.10 GHz Intel(R) Xeon(R) Gold 6130 CPU. Figure 6 shows the times of generating the standard gPC and the ME-gPC finite element approximations, with respect to mean, variance, and relative L^2 errors. For the standard gPC, the CPU times considered in this paper refer to the time of solving the linear system $\boldsymbol{A}\underline{\boldsymbol{u}} = \underline{\boldsymbol{b}}$ [i.e., Eq. (16)] using IDR(1) with the mean-based preconditioner; for ME-gPC, the CPU times are the sum of the times for solving the linear systems $\boldsymbol{A}_k\underline{\boldsymbol{u}}_k = \underline{\boldsymbol{b}}_k$ [i.e., Eq. (22)] for

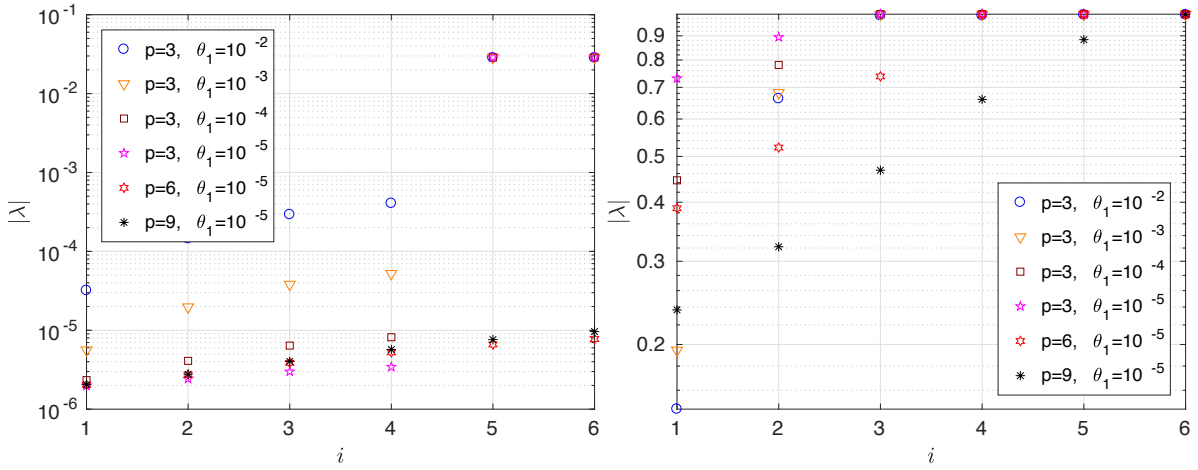


FIG. 5: The six smallest magnitude eigenvalues of the coefficient matrix arising from ME-gPC associated with the suspicion element (left) and the corresponding preconditioned coefficient matrix (right), one-dimensional random input

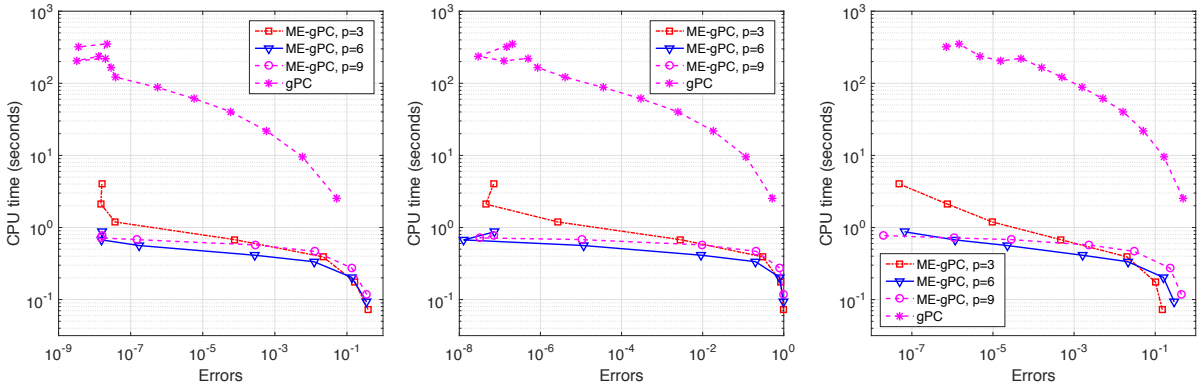


FIG. 6: Comparison of CPU times for gPC and ME-gPC, one-dimensional random input. (a) CPU times w.r.t. mean errors, (b) CPU times w.r.t. variance errors, and (c) CPU times w.r.t. relative L^2 errors.

$k = 1, \ldots, M$ (also using IDR(1) with the mean-based preconditioner). It is clear that, to achieve the same accuracy in mean and variance estimates, the CPU times required by ME-gPC are less than those of the standard gPC.

4.2 Two-Dimensional Random Input

In this test problem, the wave number is given by

$$\kappa(\boldsymbol{x}, \boldsymbol{\xi}) = k_0(\boldsymbol{x}) + \kappa_1(\boldsymbol{x})\xi_1 + \kappa_2(\boldsymbol{x})\xi_2,$$

where $\kappa_0 = 0.41$ and

$$\kappa_1(\boldsymbol{x}) = 0.24 \times (1.1 + 0.1\cos(x_1)),\ \kappa_2(\boldsymbol{x}) = 0.08 \times (1.1 + 0.1\sin(x_2)).$$

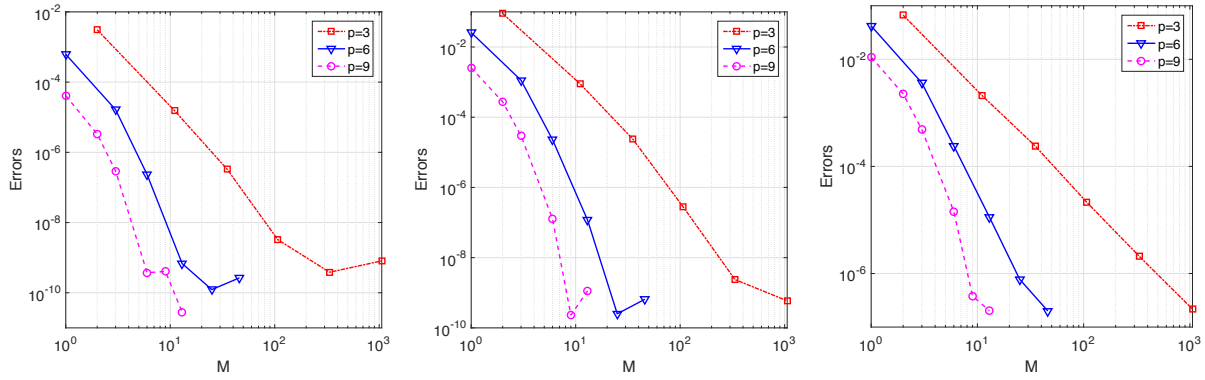
The forcing term is given by

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})\xi_1 + f_2(\boldsymbol{x})\xi_2, \tag{42}$$

where $f_0(\boldsymbol{x}) = 2\big(0.5 - x_1^2 - x_2^2\big)$, and $f_i(\boldsymbol{x}) = 0.5 \cdot \sqrt{3} f_0(\boldsymbol{x}), i = 1, 2$. The parameter α in Eq. (13) is set to 0.5 and the threshold θ_2 in Eq. (15) is set to 0.2 in this test problem. The reference solution is obtained through the global gPC method with $p = 40$.

Errors in mean and variance function estimates and the relative L^2 error of the ME-gPC finite element approximation are shown in Fig. 7. Similarly to the test problem with one-dimensional random input, for a given local gPC order, the errors decrease quickly as the number of stochastic elements M increases (the increase in M is also achieved through decreasing the threshold θ_1). Again, for the same number of stochastic elements, higher local gPC orders result in smaller errors. To look more closely at the errors of ME-gPC, we show the mean, the variance, and the relative L^2 errors in Table 2, where different values of the threshold θ_1 are considered. As shown in Table 2, for a given local gPC order, as the value of the threshold θ_1 decreases, the number of stochastic elements increases, which leads to the decrease of the errors. Compared with the test problem with one-dimensional random input, the number of random elements for this test problem increases more rapidly as the value of the threshold θ_1 decreases.

Figure 8 shows partitions of the stochastic domain by the ME-gPC method. As expected, as the threshold θ_1 decreases, ME-gPC generates more stochastic elements. It is clear that the sizes of stochastic elements become smaller as $\boldsymbol{\xi}$ goes to the corner $(-1, -1)^T$, which reveals the fact that the solution $u(\boldsymbol{x}, \boldsymbol{\xi})$ changes rapidly as $\boldsymbol{\xi}$ gets close to $(-1, -1)^T$. In addition, from Fig. 8, there are more partitions in the ξ_1 direction, which implies that the solution is more sensitive with respect to ξ_1. Figure 8 also shows the numbers of iterations to solve each linear system [Eq. (22)] of ME-gPC using IDR(1) with mean-based preconditioning. It can be seen that, the numbers of iterations are again all very small—they are at most eight for different values of the local gPC order p and the threshold θ_1.



(a) (b) (c)

FIG. 7: Errors of ME-gPC with $\alpha = 0.5$, $\theta_2 = 0.2$, two-dimensional random input. (a) Mean errors, (b) variance errors, and (c) relative L_2 errors.

TABLE 2: Errors of ME-gPC in mean and variance function estimates and the relative L^2 errors, $\alpha = 0.5$, $\theta_2 = 0.2$, two-dimensional random input

		$\theta_1 = 10^{-2}$		$\theta_1 = 10^{-3}$		$\theta_1 = 10^{-4}$		$\theta_1 = 10^{-5}$	
		M	Error	M	Error	M	Error	M	Error
Mean errors	$p = 3$	11	1.54×10^{-5}	35	3.32×10^{-7}	107	3.26×10^{-9}	335	3.85×10^{-10}
	$p = 6$	3	1.64×10^{-5}	6	2.34×10^{-7}	13	6.80×10^{-10}	25	1.23×10^{-10}
	$p = 9$	2	3.32×10^{-6}	3	2.93×10^{-7}	6	3.73×10^{-10}	9	4.05×10^{-10}
Variance errors	$p = 3$	11	9.00×10^{-4}	35	2.44×10^{-5}	107	2.78×10^{-7}	335	2.37×10^{-9}
	$p = 6$	3	1.13×10^{-3}	6	2.31×10^{-5}	13	1.17×10^{-7}	25	2.52×10^{-10}
	$p = 9$	2	2.74×10^{-4}	3	2.96×10^{-5}	6	1.28×10^{-7}	9	2.29×10^{-10}
Relative L^2 errors	$p = 3$	11	2.13×10^{-3}	35	2.41×10^{-4}	107	2.14×10^{-5}	335	2.08×10^{-6}
	$p = 6$	3	3.62×10^{-3}	6	2.42×10^{-4}	13	1.12×10^{-5}	25	7.79×10^{-7}
	$p = 9$	2	2.27×10^{-3}	3	4.85×10^{-4}	6	1.43×10^{-5}	9	3.78×10^{-7}

Figure 9 shows the results of standard gPC for this test problem. From Fig. 9(a), to achieve an accuracy with small mean and variance errors, the standard gPC requires a high global gPC order. For example, to obtain an approximation with variance error smaller than 10^{-6}, the standard gPC needs a global gPC order around 20, while ME-gPC can also achieve the same accuracy with a local gPC order less than five as shown in Figs. 7(b) and 8 (although partitioning of the stochastic domain is required). From Fig. 9(b), the number of iterations for IDR(1) with mean-based preconditioning typically increases as the global gPC order increases, and for $p \approx 20$, around 25 iterations are required, while the iteration numbers required by ME-gPC are at most 8 as shown in Fig. 8.

In Fig. 10, the six smallest magnitude eigenvalues of the coefficient matrix and the preconditioned coefficient matrix arising from gPC are shown. From the figure we can see that the six smallest magnitude eigenvalues of the coefficient matrix become larger after preconditioning.

For comparison, we plot the six smallest magnitude eigenvalues of the coefficient matrix arising from ME-gPC in Fig. 11. Again, the eigenvalues of the preconditioned coefficient matrix arising from ME-gPC are much farther from zero than those associated with standard gPC. As a result, the mean-based preconditioned IDR(1) method can converge in fewer iterations with ME-gPC (according to Fig. 8, at most eight iterations) than it does with standard gPC.

Finally, we compare the CPU times of the standard gPC and the ME-gPC for this test problem. Again, the CPU times of the standard gPC refer to the time of solving the linear system [Eq. (16)] using IDR(1) with the mean-based

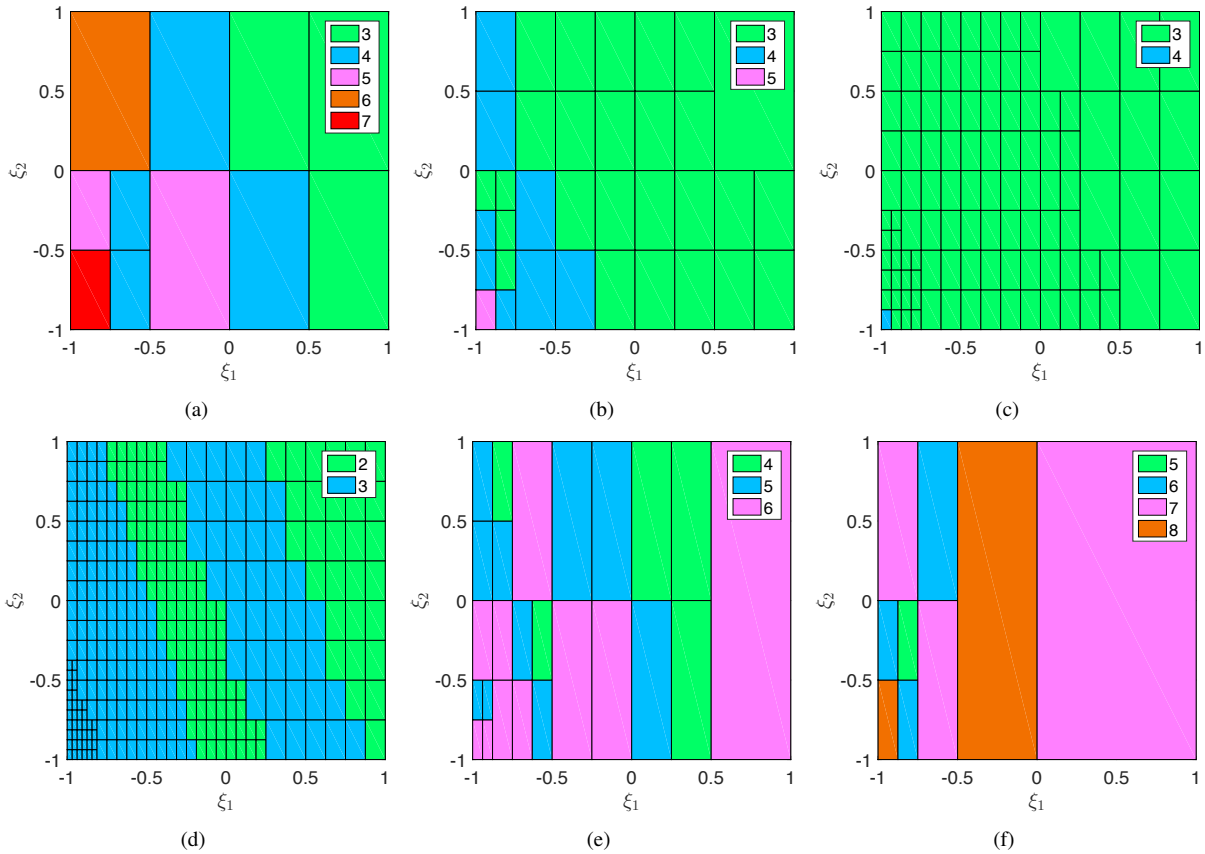


FIG. 8: Preconditioned IDR(1) iterations on each stochastic element generated by ME-gPC, two-dimensional random input. The numbers in the legend represent the number of iteration steps. (a) $p = 3, \theta_1 = 10^{-2}$, (b) $p = 3, \theta_1 = 10^{-3}$, (c) $p = 3, \theta_1 = 10^{-4}$, (d) $p = 3, \theta_1 = 10^{-5}$, (e) $p = 6, \theta_1 = 10^{-5}$, and (f) $p = 9, \theta_1 = 10^{-5}$.

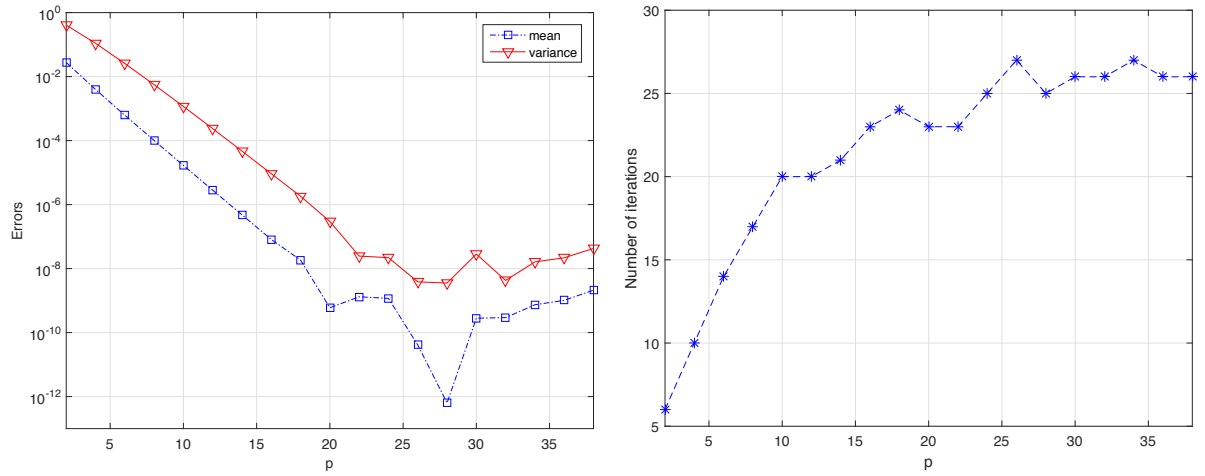


FIG. 9: Results of the standard gPC finite element approximation, two-dimensional random input. (a) Errors of standard gPC and (b) preconditioned IDR(1) iterations.

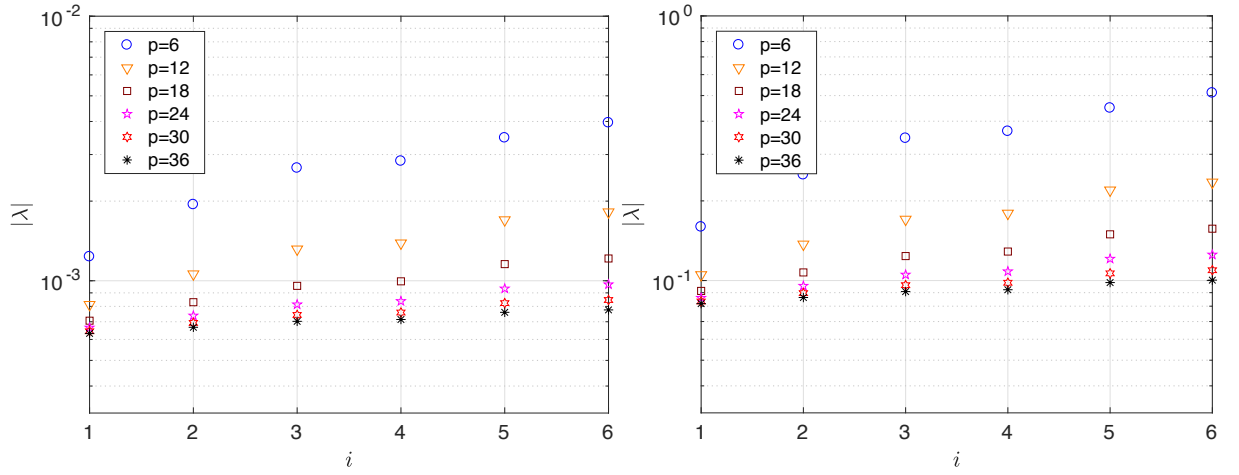


FIG. 10: The six smallest magnitude eigenvalues of the coefficient matrix (left) and the preconditioned coefficient matrix (right) arising from gPC, two-dimensional random input.

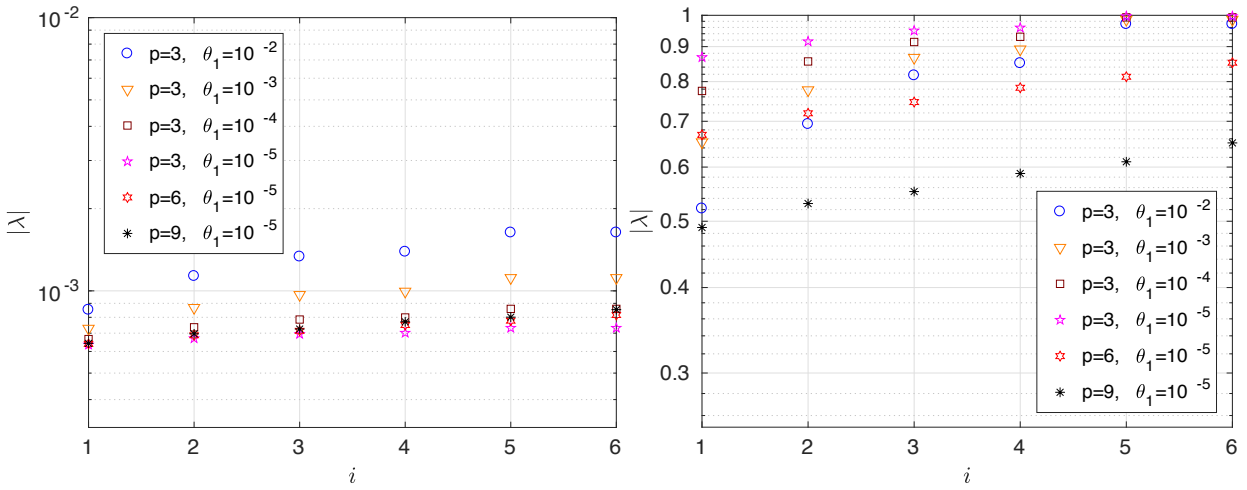


FIG. 11: The six smallest magnitude eigenvalues of the coefficient matrix arising from ME-gPC associated to the suspicion element (left) and the corresponding preconditioned coefficient matrix (right), two-dimensional random input.

preconditioner; for ME-gPC, the CPU times are the sum of the times for solving [Eq. (22)] for $k = 1, \ldots, M$ (also using IDR(1) with the mean-based preconditioner). Figure 12 shows the times of generating the standard gPC and the ME-gPC finite element approximations, with respect to mean, variance, and relative L^2 errors. It is clear that to achieve the same accuracy in mean and variance estimates, the CPU times required by ME-gPC are typically less than those of the standard gPC.

To summarize, by using the ME-gPC method, we turn a problem (typically hard to solve) to subproblems that can be solved easily. The sizes of the linear systems arising from the subproblems are smaller than those of the original problem. Besides, the mean-based preconditioner is more efficient for the subproblems than the original problem. Moreover, the subproblems can be solved simultaneously on different processors.

5. CONCLUSIONS

Conducting adaptive localized procedures is a fundamental concept for solving PI stems that are close to singular. For the open problem of stochastic Helmholtz equations close to resonance, in vork we propose and analyze

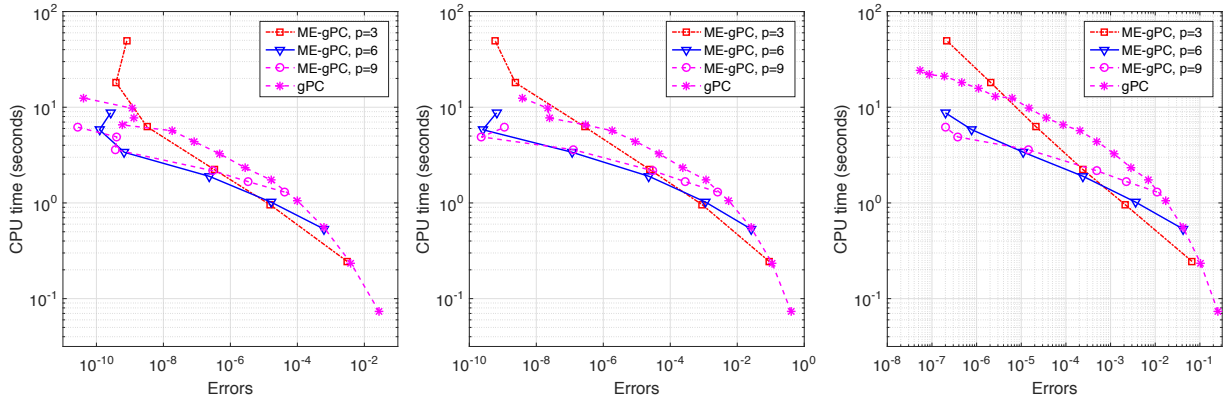


FIG. 12: Comparison of CPU times for gPC and ME-gPC, two-dimensional random input. (a) CPU times w.r.t. mean errors, (b) CPU times w.r.t. variance errors, and (c) CPU times w.r.t. relative L^2 errors.

a novel localized stochastic Galerkin framework based on the combination of ME-gPC finite element approximation and mean-based preconditioning. Through this localized approach, a global solution with large variances is decomposed into a union of local solutions with small variances. This reduction in variance results in not only an efficient stochastic spectral approximation, but also an efficient mean-based preconditioning scheme that is newly proposed and analyzed in this work for the linear systems arising from the ME-gPC finite element approximation. Our analysis and numerical results demonstrate that this new combination of ME-gPC and the mean-based preconditioner can provide an efficient strategy for solving the stochastic Helmholtz equations close to resonance. However, as a limitation of ME-gPC, this new strategy can only be applied to problems with low-dimensional random parameters. For stochastic Helmholtz problems with high-dimensional parameterization, current efforts are focused on decomposing high-dimensional random inputs into a combination of low-dimensional inputs, which include the analysis of variance decomposition [42–47] and dimension reduction based on active subspaces [48]. Implementing such strategies will be the focus of our future work.

ACKNOWLEDGMENTS

G. Wang and Q. Liao are supported by the Natural Science Foundation of Shanghai (No. 20ZR1436200), the Science Challenge Project (No. TZ2018001), and the National Natural Science Foundation of China (No. 11601329). F. Xue is supported by the National Science Foundation under Grant Nos. DMS-1719461 and DMS-1819097.

REFERENCES

1. Jensen, F.B., Kuperman, W.A., Porter, M.B., and Schmidt, H., *Computational Ocean Acoustics*, New York: Springer Science & Business Media, 2011.
2. März, R., *Integrated Optics: Design and Modeling*, Boston: Artech House on Demand, 1995.
3. Vassallo, C., *Optical Waveguide Concepts*, Amsterdam: Elsevier, 1991.
4. Elman, H.C., Ernst, O.G., O'Leary, D.P., and Stewart, M., Efficient Iterative Algorithms for the Stochastic Finite Element Method with Application to Acoustic Scattering, *Comput. Methods Appl. Mech. Eng.*, **194**(9):1037–1055, 2005.
5. Xiu, D. and Shen, J., An Efficient Spectral Method for Acoustic Scattering from Rough Surfaces, *Commun. Comput. Phys.*, **2**(1):54–72, 2007.
6. Tsuji, P., Xiu, D., and Ying, L., A Fast Method for High-Frequency Acoustic Scattering from Random Scatterers, *Int. J. Uncertainty Quantif.*, **1**(2):99–117, 2011.
7. Jin, C. and Cai, X.C., A Preconditioned Recycling GMRES Solver for Stochastic Helmholtz Problems, *Commun. Comput. Phys.*, **6**(2):342–353, 2009.

8. Eigel, M., Gittelson, C., Schwab, C., and Zander, E., Adaptive Stochastic Galerkin FEM, *Comput. Methods Appl. Mech. Eng.*, **270**:247–269, 2014.
9. Ghanem, R.G. and Spanos, P.D., *Stochastic Finite Elements: A Spectral Approach*, North Chelmsford: Courier Corporation, 2003.
10. Xiu, D., *Numerical Methods for Stochastic Computations: a Spectral Method Approach*, Princeton: Princeton University Press, 2010.
11. Babuška, I., Tempone, R., and Zouraris., G., Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations, *SIAM J. Numer. Anal.*, **42**(2):800–825, 2004.
12. Xiu, D. and Hesthaven, J., High-Order Collocation Methods for Differential Equations with Random Inputs, *SIAM J. Sci. Comput.*, **27**(3):1118–1139, 2005.
13. Babuška, I. and Tempone, N.R., A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data, *SIAM J. Numer. Anal.*, **45**(3):1005–1034, 2007.
14. Foo, J., Wan, X., and Karniadakis, G.E., The Multi-Element Probabilistic Collocation Method (ME-PCM): Error Analysis and Applications, *J. Comput. Phys.*, **227**(22):9572–9595, 2008.
15. Ma, X. and Zabaras, N., An Adaptive Hierarchical Sparse Grid Collocation Algorithm for the Solution of Stochastic Differential Equations, *J. Comput. Phys.*, **228**(8):3084–3113, 2009.
16. Tang, T. and Zhou, T., Convergence Analysis for Stochastic Collocation Methods to Scalar Hyperbolic Equations with a Random Wave Speed, *Commun. Comput. Phys.*, **8**(1):226–248, 2010.
17. Narayan, A. and Xiu, D., Stochastic Collocation Methods on Unstructured Grids in High Dimensions via Interpolation, *SIAM J. Sci. Comput.*, **34**(3):A1729–A1752, 2012.
18. Wan, X. and Karniadakis, G.E., An Adaptive Multi-Element Generalized Polynomial Chaos Method for Stochastic Differential Equations, *J. Comput. Phys.*, **209**(2):617–642, 2005.
19. Bespalov, A. and Silvester, D., Efficient Adaptive Stochastic Galerkin Methods for Parametric Operator Equations, *SIAM J. Sci. Comput.*, **38**(4):A2118–A2140, 2016.
20. Powell, C.E. and Elman, H.C., Block-Diagonal Preconditioning for Spectral Stochastic Finite-Element Systems, *IMA J. Numer. Anal.*, **29**(2):350–375, 2009.
21. Pembery, O.R. and Spence, E.A., The Helmholtz Equation in Random Media: Well-Posedness and A Priori Bounds, *SIAM/ASA J. Uncertainty Quantif.*, **8**(1):58–87, 2020.
22. Graham, I.G., Pembery, O.R., and Spence, E.A., Analysis of a Helmholtz Preconditioning Problem Motivated by Uncertainty Quantification, *Numer. Anal.*, arXiv:2005.13390, 2020.
23. Karchevskii, E. and Solov'ev, S., Investigation of a Spectral Problem for the Helmholtz Operator on the Plane, *Differ. Eq.*, **36**(4):631–634, 2000.
24. Kartchevski, E., Nosich, A., and Hanson, G., Mathematical Analysis of the Generalized Natural Modes of an Inhomogeneous Optical Fiber, *SIAM J. Appl. Math.*, **65**(6):2033–2048, 2005.
25. Frolov, A. and Kartchevskiy, E., Integral Equation Methods in Optical Waveguide Theory, in *Inverse Problems and Large-Scale Computations*, New York: Springer, pp. 119–133, 2013.
26. Babuška, I., Tempone, R.l., and Zouraris, G.E., Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations, *SIAM J. Numer. Anal.*, **42**(2):800–825, 2004.
27. Feng, X., Lin, J., and Lorton, C., An Efficient Numerical Method for Acoustic Wave Scattering in Random Media, *SIAM/ASA J. Uncertainty Quantif.*, **3**(1):790–822, 2015.
28. Elman, H.C., Silvester, D.J., and Wathen, A.J., *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford: Oxford University Press, 2014.
29. Xiu, D. and Karniadakis, G.E., The Wiener--Askey Polynomial Chaos for Stochastic Differential Equations, *SIAM J. Sci. Comput.*, **24**(2):619–644, 2002.
30. Marzouk, Y.M. and Najm, H.N., Dimensionality Reduction and Polynomial Chaos Acceleration of Bayesian Inference in Inverse Problems, *J. Comput. Phys.*, **228**(6):1862–1902, 2009.
31. Deb, M.K., Babuška, I.M., and Oden, J.T., Solution of Stochastic Partial Differential Equations Using Galerkin Finite Element Techniques, *Comput. Methods Appl. Mech. Eng.*, **190**(48):6359–6372, 2001.

32. Wan, X. and Karniadakis, G.E., Multi-Element Generalized Polynomial Chaos for Arbitrary Probability Measures, *SIAM J. Sci. Comput.*, **28**(3):901–928, 2006.
33. Augustin, F. and Rentrop, P., Stochastic Galerkin Techniques for Random Ordinary Differential Equations, *Numer. Math.*, **122**(3):399–419, 2012.
34. Wang, G. and Liao, Q., Efficient Spectral Stochastic Finite Element Methods for Helmholtz Equations with Random Inputs, *East Asian J. Appl. Math.*, **9**(3):601–621, 2019.
35. Rudin, W., *Real and Complex Analysis*, New York: Tata McGraw-Hill Education, 2006.
36. Sonneveld, P. and Van Gijzen, M.B., IDR(S): A Family of Simple and Fast Algorithms for Solving Large Nonsymmetric Systems of Linear Equations, *SIAM J. Sci. Comput.*, **31**(2):1035–1062, 2008.
37. Du, L., Sogabe, T., Yu, B., Yamamoto, Y., and Zhang, S.L., A Block IDR(s) Method for Nonsymmetric Linear Systems with Multiple Right-Hand Sides, *J. Comput. Appl. Math.*, **235**(14):4095–4106, 2011.
38. Ghanem, R.G. and Kruger, R.M., Numerical Solution of Spectral Stochastic Finite Element Systems, *Comput. Methods Appl. Mech. Eng.*, **129**(3):289–303, 1996.
39. Pellissetti, M.F. and Ghanem, R.G., Iterative Solution of Systems of Linear Equations Arising in the Context of Stochastic Finite Elements, *Adv. Eng. Software*, **31**(8):607–616, 2000.
40. Trefethen, L.N. and Bau, D., *Numerical Linear Algebra*, Philadelphia: Society for Industrial and Applied Mathematics, 1997.
41. Braess, D., *Finite Elements*, London: Cambridge University Press, 1997.
42. Gao, Z. and Hesthaven, J.S., On ANOVA Expansions and Strategies for Choosing the Anchor Point, *Appl. Math. Comput.*, **217**:3274–3285, 2010.
43. Ma, X. and Zabaras, N., An Adaptive High-Dimensional Stochastic Model Representation Technique for the Solution of Stochastic Partial Differential Equations, *J. Comput. Phys.*, **229**(10):3884–3915, 2010.
44. Zhang, Z., Choi, M., and Karniadakis, G., Error Estimates for the ANOVA Method with Polynomial Chaos Interpolation: Tensor Product Functions, *SIAM J. Sci. Comput.*, **34**(2):A1165–A1186, 2012.
45. Yang, X., Choi, M., Lin, G., and Karniadakis, G., Adaptive ANOVA Decomposition of Stochastic Incompressible and Compressible Flows, *J. Comput. Phys.*, **231**(4):1587–1614, 2012.
46. Liao, Q. and Lin, G., Reduced Basis ANOVA Methods for Partial Differential Equations with High-Dimensional Random Inputs, *J. Comput. Phys.*, **317**:148–164, 2016.
47. Cho, H. and Elman, H.C., An Adaptive Reduced Basis Collocation Method Based on PCM ANOVA Decomposition for Anisotropic Stochastic PDEs, *Int. J. Uncertainty Quantif.*, **8**:193–210, 2018.
48. Constantine, P.G., *Active Subspaces*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2015.