Model Reduction for Large-Scale Earthquake Simulation in an Uncertain 3D Medium

P. Sochala[a,*], F. De Martin[a], O. Le Maître[b]

[a] *Bureau de Recherches Géologiques et Minières, 45060 Orléans, France*
[b] *Centre de Mathématiques Appliquées, CNRS, Inria, Ecole Polytechnique, 91120 Palaiseau, France*

Abstract

In this paper, we are interested in the seismic wave propagation into an uncertain medium. To this end, we performed an ensemble of 400 large-scale simulations that requires 4 million core-hours of CPU time. In addition to the large computational load of these simulations, solving the uncertainty propagation problem requires dedicated procedures to handle the complexities inherent to large data set size and the low number of samples. We focus on the peak ground motion at the free surface of the 3D domain, and our analysis utilizes a surrogate model combining two key ingredients for complexity mitigation: i) a dimension reduction technique using empirical orthogonal basis functions and ii) a functional approximation of the uncertain reduced coordinates by polynomial chaos expansions. We carefully validate the resulting surrogate model by estimating its predictive error using bootstrap, truncation, and cross-validation procedures. The surrogate model allows us to compute various statistical information of the uncertain prediction, including marginal and joint probability distributions, interval probability maps, and 2D fields of global sensitivity indices.

Keywords: seismic wave propagation; uncertainty quantification; empirical orthogonal functions; polynomial chaos; interval probability; global sensitivity analysis

1. Introduction

This paper concerns the study of earthquake ground motions, in the presence of uncertainties in the seismic wave velocity, using large-scale, physics-based simulations. The site of interest is a sedimentary basin [1–3] whose three-dimensional geological model has been built from the aggregation of multiple geophysical campaigns. Despite the colossal work done to build a model extending along more than 60×30 km, uncertainties naturally remain. In this context, the goal of Uncertainty Quantification (UQ) is to estimate the uncertainties in the prediction of the model (output) as induced by the uncertainties of its inputs. We rely on a probabilistic approach in which the uncertain inputs are equipped with a given probability law, such that the uncertain predictions can be characterized statistically through, for instance, the estimation of their statistical moments, confidence intervals, probability distribution, quantiles, sensitivity indices.

While there is an extensive literature devoted to stochastic soil dynamics (see [4] for a review) and uncertainties in structural reliability [5], especially for earthquake engineering (see for instance [6–8]), the effect of an uncertain geological structure on the seismic wave propagation using model reduction technique has received scant attention. Here, we consider a model with seven independent random parameters describing the uncertain geological properties in the domain, and we focus on the prediction of the Peak Ground Motion (PGM), a quantity that is particularly relevant and useful to earthquake engineering. For a given earthquake, the ground motion varies from a place to another, and one must estimate the spatial dependences of

*Corresponding author
Email addresses: p.sochala@brgm.fr (P. Sochala), f.demartin@brgm.fr (F. De Martin), olivier.le-maitre@polytechnique.edu (O. Le Maître)

the PGM to produce seismic hazard maps. The challenge is then to compute the dependences on the seven random model parameters of the spatial PGM field. In other words, we have to estimate a random field.

In the present work, the UQ analysis follows a non-intrusive implementation, meaning that a deterministic solver is used several times to evaluate the seismic wave propagation for different values (or realizations) of the uncertain inputs parametrizing the geological structure of the soil. This ensemble of numerical simulations of three-dimensional seismic wave propagation is then exploited to perform the uncertainty analyses. The main problems that we face to achieve accurate uncertainty characterizations are threefold: i) the computational burden of the simulations ensemble, ii) the large size of the dataset associated with the ensemble of solutions (2D fields), and iii) the limited number of members that can be computed.

High performance computing is essential in this study to handle the computational load stemming from the size of the domain of study, the space and time resolutions, and the number of ensemble runs. Specifically, for each member of the ensemble, the size of the computational domain is $64 \times 42 \times 5\,\mathrm{km}^3$ and the time of simulation is 30 s, discretized with a time step size equal to 2×10^{-4} s. The space and time steps are chosen to prevent aliasing for waves up to 3 Hz and respecting the stability condition of the explicit time scheme. As a result, the computational mesh involves a total of 7,053,889 unstructured hexahedron elements with close to 500 million spatial nodes and 150,000 time steps. Although our solver is efficiently parallelized, as shown by the strong scalability test presented later in section 2.3, the cost of a single simulation amounts roughly to 10,000 core-hours. Therefore, limitations on the computational budget imposed an ensemble of 400 direct model evaluations only, for a total of 4 millions core-hours of computations, that have been executed on the supercomputer Shaheen II of the Extreme Computing Research Center at the King Abdullah University of Science and Technology (KAUST, KSA). Regarding the storage requirement, the ensemble of 400 simulations produced up to 60 TB of results, corresponding to the three components of the dynamical ground motion fields at the free surface nodes only and after applying time filtering and decimation. This procedure leads to the storage at 1,000 times (instead of the 150,000 computational time steps), consistently with the maximum frequency precisely captured on the mesh. This dataset was subsequently post-treated to extract the PGM (*i.e.* ignoring time-evolutions to retain the local maximum absolute value over the three displacement components and the 1,000 recorded times), reducing the dataset to a more manageable size of 20 GB.

Since the ensemble size was limited to 400 members, straightforward UQ methods, such as the Monte-Carlo (MC) simulation method, typically requiring tens of thousand or more members, could not be employed. Instead, we exploit in this work the relatively low dimensionality (7) of the uncertain inputs to construct a functional approximation of the (smooth) dependencies of the uncertainty PGM fields with the uncertain inputs. The advantage here is that provided the approximation, or surrogate model, is accurate enough, it can replace the original high-fidelity model to perform the various uncertainty analyses. As the free surface of the computational model is discretized with roughly 10 millions degrees of freedom, the construction of a surrogate model for each of the degrees of freedom to obtain hazard maps would be a computationally heavy task because of the millions of functional approximations to be formed. Following [9], we then rely on a preliminary dimensionality reduction procedure to represent the PGM fields with a low number of basis functions. Specifically, the spatial reduction is achieved using Empirical Orthogonal Functions (EOF) [10] associated with the spatial covariance matrix assembled from the ensemble of model outputs. As discussed later, the EOF decomposition not only provides a reduction but also gives physical insight on the main processes impacted by the uncertainty; in the present work, the dominant EOF mode concerns the body wave inside the sedimentary basin, while the following modes capture the uncertainty caused by the surface wave along the boundary of the basin.

Following the spatial reduction, the functional representation of the uncertain PGM field proceeds with the approximation of the dependencies of the EOF coordinates with the uncertain inputs. Polynomial Chaos (PC) expansions [5, 11] are used to this end. Because of the small size ensemble, we use a compressed sensing technique [12–14] in order to identify the coefficients of the PC series. Validation results show a negligible cross-validation error reflecting the faithful prediction capacity of the resulting EOF-PC surrogate. The surrogate is able to predict the PGM marginal and joint probability density functions correctly and therefore to produce probabilistic seismic hazard maps. The interval probability maps allow us to quantify the maximum of the PGM values and to identify the most hazardous areas inside the sedimentary basin.

The global sensitivity analysis based on the surrogate shows that the damping coefficient of the medium is the parameter which mainly controls the PGM.

The structure of the paper is as follows. Section 2 introduces the seismic wave propagation model, the case study, and the parallel solver used to perform the numerical simulations. Section 3 describes the uncertainty setting including the uncertain inputs with their parametrization, and the first moments of the PGM. Section 4 describes the construction of the surrogate model, with the dimension reduction technique and the polynomial chaos method. Section 5 presents the results of the interval probability maps and sensitivity analyses. Finally, in Section 6, we draw the significant conclusions of this work.

2. Geophysical framework

This section describes briefly the seismic wave propagation physical model, the geography and properties of the basin consider in this study, and finally the numerical code used to generate the simulation results.

2.1. Seismic wave propagation model

The general form of the equation of motion in a continuous medium is

$$\rho \frac{\partial^2 \boldsymbol{s}}{\partial t^2} = \nabla \cdot \boldsymbol{\sigma} + \boldsymbol{f}, \tag{1}$$

where ρ is the bulk density of the medium, $\boldsymbol{s} := (s_x, s_y, s_z)^\top$ is the displacement field, $\boldsymbol{\sigma}$ is the stress tensor, and $\boldsymbol{f}$ is the body force (per unit mass). The governing equation (1) is completed by a modified Hooke's law (stress-strain relationship) modeling the intrinsic attenuation [15]. The stress tensor is decomposed into an unrelaxed elastic tensor $\boldsymbol{\sigma}_{\mathrm{U}}$ and an attenuation tensor $\boldsymbol{\sigma}_{\mathrm{A}}$, such that

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_{\mathrm{U}} - \boldsymbol{\sigma}_{\mathrm{A}}. \tag{2}$$

The stress-strain relationship for an unrelaxed linear elastic isotropic material reads

$$\boldsymbol{\sigma}_{\mathrm{U}} = 2M_{\mathrm{s}}\boldsymbol{\epsilon} + (M_{\mathrm{p}} - 2M_{\mathrm{s}})\mathrm{Tr}(\boldsymbol{\epsilon})\boldsymbol{I}, \tag{3}$$

where, in the small deformation theory with negligible rigid-body rotation, the strain tensor $\boldsymbol{\epsilon}$ is defined as $\left(\nabla \boldsymbol{s} + \nabla^\top \boldsymbol{s}\right)/2$. The anelastic tensor is modeled by a sum of R memory variables $\{\boldsymbol{\sigma}^k\}$ [16, 17]

$$\boldsymbol{\sigma}_{\mathrm{A}} = \sum_{k=1}^{R} \boldsymbol{\sigma}^k, \tag{4}$$

that are governed by a system of first-order differential equations [18]

$$\tau^k \frac{\partial \boldsymbol{\sigma}^k}{\partial t} + \boldsymbol{\sigma}^k = 2M_{\mathrm{s}}\boldsymbol{\epsilon} w_{\mathrm{s}}^k + \left(M_{\mathrm{p}} w_{\mathrm{p}}^k - 2M_{\mathrm{s}} w_{\mathrm{s}}^k\right)\mathrm{Tr}(\boldsymbol{\epsilon})\boldsymbol{I}, \tag{5}$$

with the unrelaxed compression M_{p} and shear M_{s} moduli defined as

$$M_{\mathrm{p,s}} = \frac{\rho v_{\mathrm{p,s}}^2}{\left|1 - \sum_{k=1}^{R} \dfrac{w_{\mathrm{p,s}}^k}{1 + i\omega_r \tau^k}\right|}. \tag{6}$$

In the previous relations, we have denoted Tr() the trace operator, $\boldsymbol{I}$ the identity matrix, v_{p} and v_{s} the compressional and shear wave velocities, w_{s}^k and w_{p}^k some weight coefficients, τ^k the relaxation times, ω_r a reference angular frequency and i the imaginary unit ($i^2 = -1$).

The two wave velocities v_{s} and v_{p} can be expressed as functions of the Lamé coefficients λ and μ of the material, through the relations $v_{\mathrm{s}} = \sqrt{\mu/\rho}$ and $v_{\mathrm{p}} = \sqrt{(\lambda + 2\mu)/\rho}$. The weight coefficients w_{p}^k and w_{s}^k

depend on the compressional wave quality factor Q_p and on the shear wave quality factor Q_s, respectively. The method proposed by [19] is adopted for deriving these weight coefficients. They are calculated using the interpolation formula (indices p, s are dropped for simplification)

$$w^k = \chi(\chi\alpha^k + \beta^k), \tag{7}$$

where α^k and β^k are fixed interpolation coefficients. The factor χ depends only on Q and is estimated by

$$\chi = \frac{3.071 + 1.433Q^{-1.158}\ln(Q/5)}{1 + 0.415Q}, \quad 5 \leqslant Q \leqslant 5000. \tag{8}$$

The numerical implementation used in this study for solving Eq. (1) in an attenuating medium has been verified in [20]. The definition of the uncertain parameters v_p, v_s, Q_p and Q_s are further discussed in Section 3.

2.2. Case study description

Our study considers a tectonically active Mygdonian basin located in Greece, close to the city of Thessaloniki. We are interested here in quantifying the uncertainty in the prediction of the PGM in the sedimentary basin. The extend of the domain simulated is $64 \times 42 \times 5$ km. The topography of the site and the basin shape are depicted Fig. 1. Since 1993, this site is monitored, through the multi-purpose physical laboratory called EUROSEISTEST, with long-term measurements and data on the ground motion, site effects, soil and site characterization, wave propagation and soil-foundation-structure interaction [21, 22]. The definition of our uncertain velocity model (see Section 3.1) is inspired from the data presented in [3]; it consists of two layers overlaying the half-space as depicted in Fig. 2 showing a vertical cross-section passing through the so-called TST station. In this figure, the interface between the two layers corresponds to the geological Mygdonian/Pre-Mygdonian limit (MPL) boundary. The depth of the MPL being considered as uncertain, its minimum and maximum location are depicted in Fig. 2 as dash-dotted lines. It will be further discussed in Section 3.1.

A seismic excitation can be implemented using either a force-based or a displacement-based approach. In our model, the medium is excited by a shear plane wave with vertical incidence obtained by imposing a uniform displacement in the y-direction, over the horizontal xy plane located at constant depth $z_s = -3744$ m (see the schematic representation in Fig. 6),

$$\boldsymbol{s}(x, y, z_s) = (0, s(t), 0)^\top, \tag{9}$$

where $s(t)$ denotes the time source function. In practice, this displacement is enforced during during 1 s. According to the displacement-based approach and ignoring the effects of self-gravitation, the body force $\boldsymbol{f}$ is set to zero in Eq. (1). The source function $s(t)$ is a pseudo-unit impulse,*i.e.* a Dirac delta function filtered by a Butterworth low-pass filter, depicted in Fig. 3 in time and frequency domains. The degrees of freedom on the vertical boundaries are locked except in the direction of the shear motion. An absorbing condition is applied at the bottommost boundary of the domain to absorb the downgoing waves. The temporal structure of the seismic source $s(t)$ can be changed a posteriori due to the linearity of the model in Eq. (1) to Eq. (5). Indeed, one can reconstruct any earthquake with similar excitation in the same plane, but a different time function, by a simple convolution of the computed unit impulse response with the desired source spectrum. However, the case of a non-uniform seismic source is not accessible from the present configuration, and future studies will have to explore the effects of localized sources with uncertain characteristics.

For the analysis of the response, we shall consider the PGM at each point at the free surface. Denoting here $u(\boldsymbol{x})$ the PGM field, it is defined as the maximum over time of the displacement in the direction of the imposed motion, that is

$$u(\boldsymbol{x}) = \max_t |s_y(\boldsymbol{x}, t)|$$

due to the fact that the displacement (9) is imposed along the y-direction (corresponding to the geographic North pole). Note the maximum over time is substituted with the maximum over the simulation time, since for the considered excitation the medium returns asymptotically to rest and $\boldsymbol{s}(t) \to 0$ when $t \to \infty$ everywhere in the domain.

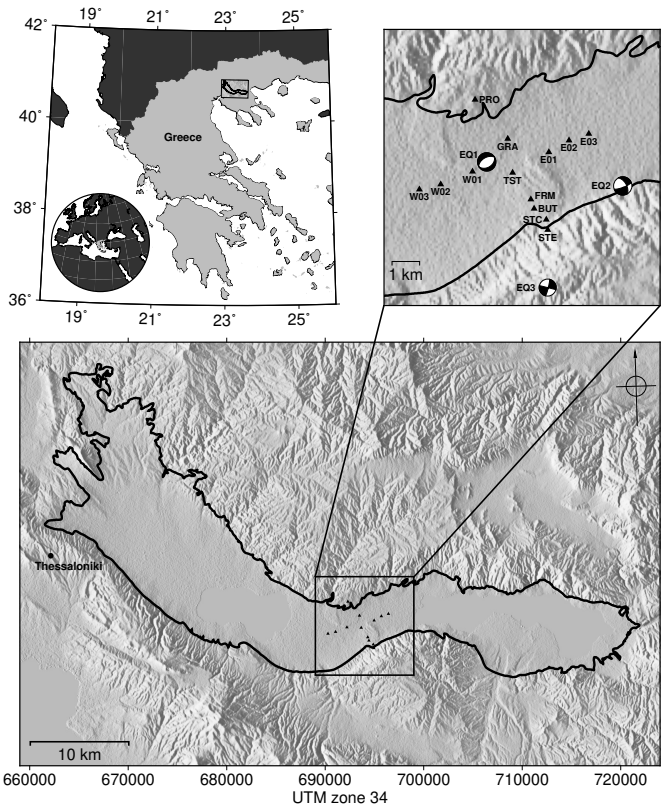

Figure 1: Location of the domain of study in Greece close to the city of Thessaloniki. The free surface of the domain is depicted together with the internal boundary between the sedimentary basin and the bedrock (solid line).

2.3. Numerical code

The governing equations of the ground motion are solved using EFISPEC3D [23], a three-dimensional open-source parallel computer program. This numerical code has been extensively validated on semi-analytical solutions [20, 24, 25]. In more details, EFISPEC3D solves the weak form of Eq. (1) using a continuous Galerkin spectral finite element method [26–28] with an explicit time marching scheme [29] constrained by a CFL condition.

The computational domain is meshed with 7,053,889 unstructured hexahedron elements, to form the mesh depicted in Fig. 4. The size of the elements is adapted over the domain of simulation to propagate a minimal wavelength of 35 m without spatial aliasing. The wavelength of 35 m corresponds to a shear wave velocity of 104 m/s at 3 Hz, 104 m/s being the minimum possible shear wave velocity (see Tab. 1). The procedure for the mesh generation is as follows: i) A hexahedral mesh $\mathcal{M}_1$ of the domain of simulation with a flat free surface is generated by CUBIT [30]. The length of the edges of the hexahedron elements varies from 315 m at the bottom to 105 m at the free surface. Such a refinement is done using the tripling technique implemented in CUBIT. ii) A mesh $\mathcal{M}_2$ is obtained by a home-made code deforming the flat free surface of $\mathcal{M}_1$ to fit the topography provided by a discrete elevation model. iii) The final mesh $\mathcal{M}$ is produced by CUBIT by refining with the tripling technique the hexahedron elements of the mesh $\mathcal{M}_2$ located within the sedimentary basin. The mesh $\mathcal{M}$ is shown in Fig. 4.

The spectral finite element method used to solve the weak form of equation of motion is based on $\mathbb{Q}_4$ Lagrange basis functions yielding 476,426,167 Gauss–Lobatto–Legendre points in the computational domain. The computational cost of one simulation with the mesh $\mathcal{M}$ is about 10,000 core-hours and the storage size of the space and time-dependent displacement field at the free surface is 175 GB. The efficiency of EFISPEC3D in terms of scalability can be appreciated from Fig. 5 where typical results of a strong

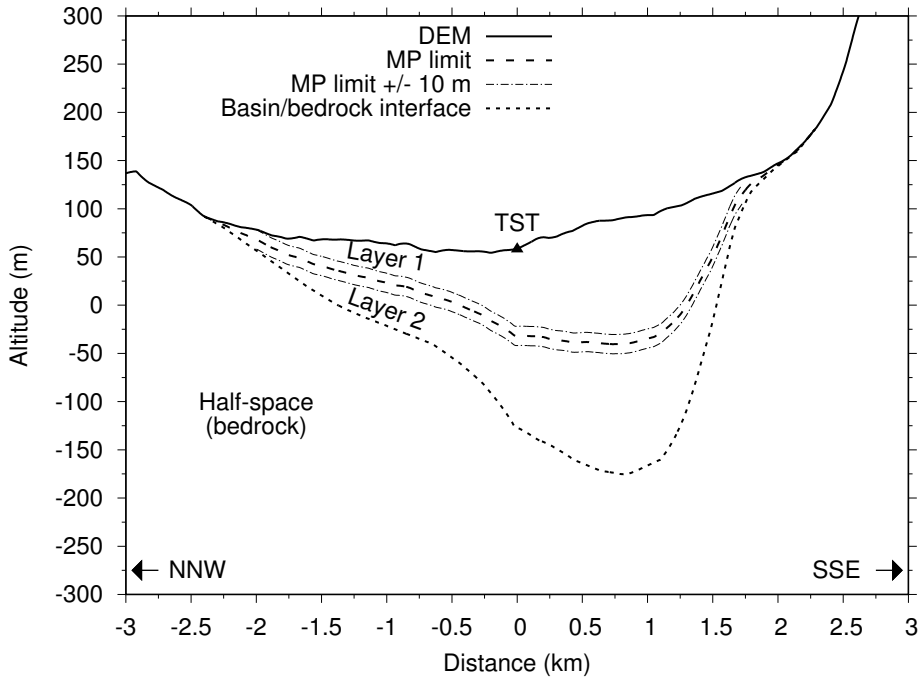

Figure 2: Vertical cross-section of the 3D model at the TST station.

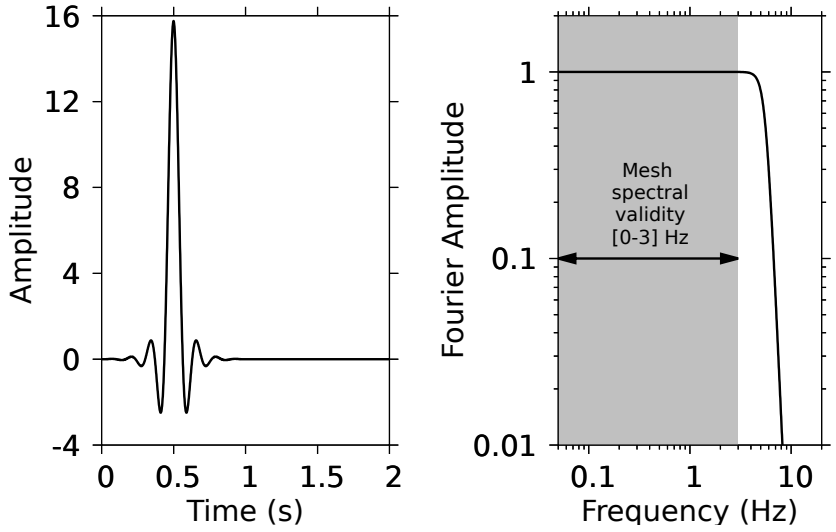

Figure 3: Source function $s(t)$ represented in time and frequency domains.

scaling test are plotted. This particular scalability test was designed to fill 60% of the 128 GB memory of a single node. The associated mesh was made of 1,906,624 hexahedron elements. An acceptable strong scaling speedup of 5,376 is observed up to 16,384 cores using MPI non-blocking communications. The simulations performed for this study have been launched on 2,048 cores.

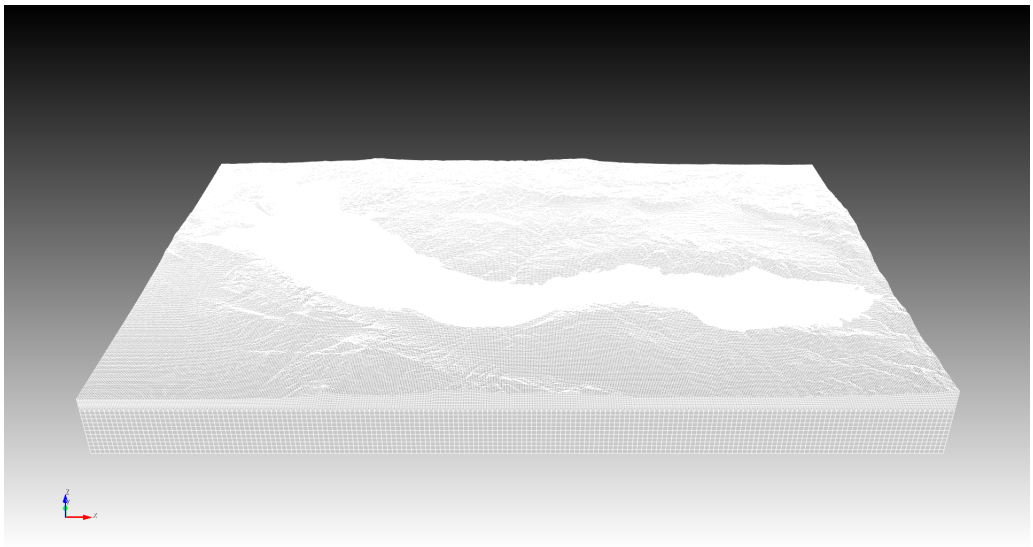

Figure 4: Mesh $\mathcal{M}$ of the computational domain ($64 \times 42 \times 5$ km) composed by 7,053,889 hexahedron elements. The two levels of mesh refinement are noticeable at the top part of the domain (tripling technique) and inside the sedimentary basin.

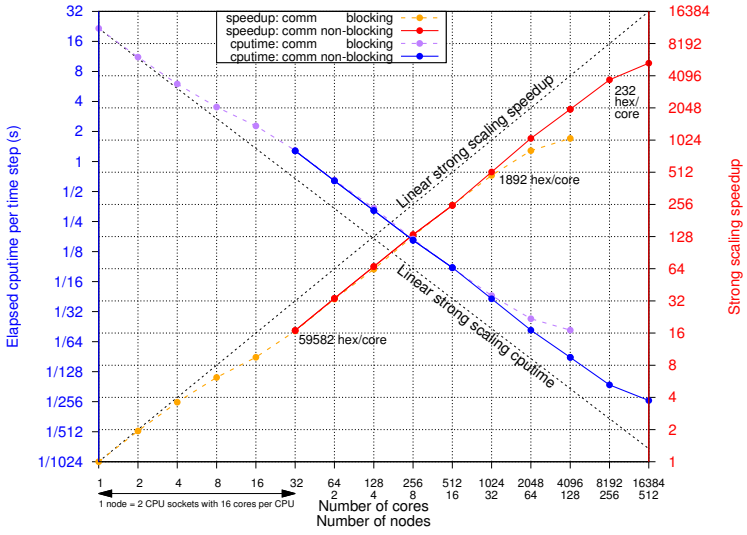

Figure 5: Strong scaling test of EFISPEC3D on ShaheenII (KAUST) performed on a 1,906,624 hexahedron elements mesh: speedup and cputime are reported for both MPI blocking communications (dashed lines) and MPI non-blocking communications.

3. Uncertain settings

In this section, we detail the uncertainty model selected for the analysis, and we propose a preliminary study of the PGM field uncertainty based on the quasi-Monte Carlo (QMC) set of 400 simulations.

3.1. Uncertain inputs

In a priori uncertainty quantification studies, as performed in this work, the definition of the uncertainty model is crucial. Typically, its definition combines some a priori about the level of knowledge concerning the problem definition (topology, geological structure, material properties,...) and the importance of these imprecisions on the quantity of interest (PGM). When the "important" uncertain inputs are selected, the

next step is to define their probabilistic model which includes the prescription of the uncertainty ranges and distributions as well as the possible correlation between inputs, etc. In practice, in the selection step, one has to balance the complexity of the resulting analysis with the completeness of possible uncertainty sources. These aspects are particularly important in the context of large scale models with high computational cost, where the affordable number of simulations is limited.

In the present work, we decided to favor an uncertainty model with a low number of uncertain inputs, in order to maintain a reasonable computational cost while ensuring sufficient accuracy and simple enough analysis to draw meaningful conclusions regarding the respective effects considered. Indeed, characterizing and understanding the individual effects of multiple uncertainty sources is usually demanding, both computationally and in term of analysis. Neglecting the body force source in Eq. (1), the main parameters governing the wave propagation through the stress-strain relationship are the unrelaxed moduli in Eq. (3) and both the unrelaxed moduli and the quality factors (*via* the weigth coeffficients w^k) in Eq. (5). The unrelaxed moduli defined by Eq. (6) depends on the wave velocities, of the bulk density and of the quality factors. We consider, a priori, the wave velocities v_s and v_p as influent parameters because the square of these velocities intervene in the numerator of the unrelaxed moduli (see Eq. (6)). Note that the bulk density ρ also intervenes in the numerator of the unrelaxed moduli, but without the square exponentiation. Therefore, we neglect it to reduce the number of uncertain parameters. The quality factors being present both in the denominator of the unrelaxed moduli and in the differential equations (5), we consider them as influent and therefore uncertain.

In our simulations, we only consider uncertainties in the shear wave velocity profile and completely neglect any uncertainty in the compressional wave velocity associated with the diagonal part of the unrelaxed and anelastic tensors. This simplified view is supported by the dominant role of the shear waves in the dynamics of the present problem, because of the nature of the excitation imposed on the medium. Further, by neglecting uncertainties in the compressional wave velocity, we avoid the need to specify a dependence structure between the uncertainties of the two wave velocities. A schematic view of the shear wave velocity structure is shown in Fig. 6. We assume affine vertical profiles within each sedimentary layer defined by top and bottom values. Specifically, we denote $v_{\mathrm{s},1}^{\mathrm{top}}$, $v_{\mathrm{s},1}^{\mathrm{bot}}$ (resp. $v_{\mathrm{s},2}^{\mathrm{top}}$, $v_{\mathrm{s},2}^{\mathrm{bot}}$) the shear wave velocity values at the top and bottom of first sedimentary layer (resp. second layer) and impose a linear dependence on the vertical coordinates z. Within the bedrock, we assume a constant value ($v_{\mathrm{s,bed}}$) with a uniform distribution. This representation induces vertical gradients in v_s, that increase when the thicknesses of the layers decrease, and then horizontal gradients since the layers have non-flat boundaries. Regarding the uncertain values of $v_{\mathrm{s},1,2}^{\mathrm{top,bot}}$, we again assumed uniform distributions in the ranges listed in Table 1. Inspecting these ranges, we see that they are set to ensure that the velocity in the bedrock, $v_{\mathrm{s,bed}}$ is almost surely higher than in the sediment and that the velocity in the second (lower) layer is in turn always higher than the velocity in the first layer. Besides, the uncertain ranges ensure that the velocity is increasing with the distance to the free surface.

Although the shear wave velocities can be measured *in situ* by different geophysical methods, the quality factors are more difficult to determine. A practical assumption is to define them proportional to the wave velocities as

$$Q = \frac{v}{q} \tag{10}$$

with q a wave velocity scaling factor. In this study, we consider the q scaling factor as uncertain. To limit the number of uncertain parameters, we set this factor constant within the two layers (*i.e.* not constant for each layer), and we assume a uniform distribution within the range $[5, 20]$.

Whence the velocities and the quality factors are specified within each layers, the shape of the geometrical model is entirely determined from the geometries of the layers. Here, we consider a seventh and last uncertainty source, assuming an imprecision on the knowledge of the location of the MPL between the two layers. For simplicity, we assume an uncertainty model with just one parameter, δd, measuring a global shift in the vertical direction from the MPL geometry provided by the numerical terrain model [3]. For δd we set an uncertainty range of ± 10 m with a uniform distribution. We observe that displacing the interface (MPL) of the two sedimentary layers can induce significant changes at the boundary of the sedimentary basin, where the thickness is small, with the possible disappearance of a layer close to the basin edges depending on δd.

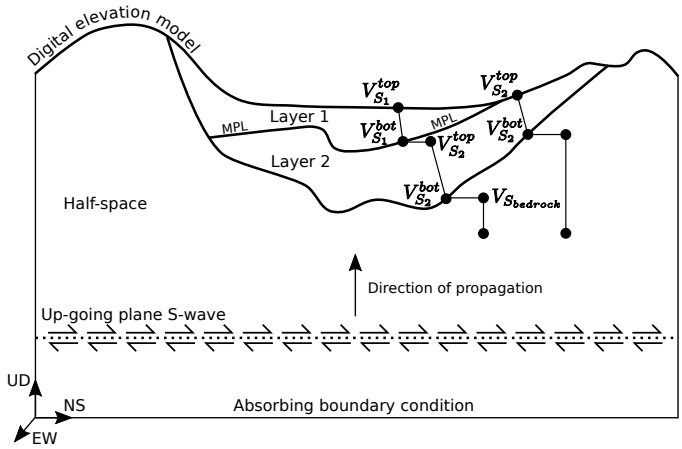

Figure 6: Schematic representation of the shear wave velocity structure; in the sedimentary basis the profile is piecewise linear and defined by values at the top and bottom of the layers 1 and 2; in the bedrock the value is constant.

The numerical schemes of the seismic code EFISPEC3D can naturally accommodate these situations. We also remark that the range of δd is smaller than the element size used for the spatial discretization (roughly set to 35 m). The seemingly small displacement of the interface, relative to the size of the elements, is not a problem for the EFISPEC3D code which perceives any change in the interface location, owing to the exact integration of the (smoothed) velocity model [31].

Random samples of the shear wave velocity profile v_s inside the sedimentary layers below the TST station are reported in the left plot of Fig. 7 for illustration purpose. The right plot shows the corresponding realizations of the shear wave quality factor Q_s computed from (10). These profiles correspond to the 400 QMC realization considered in the following.

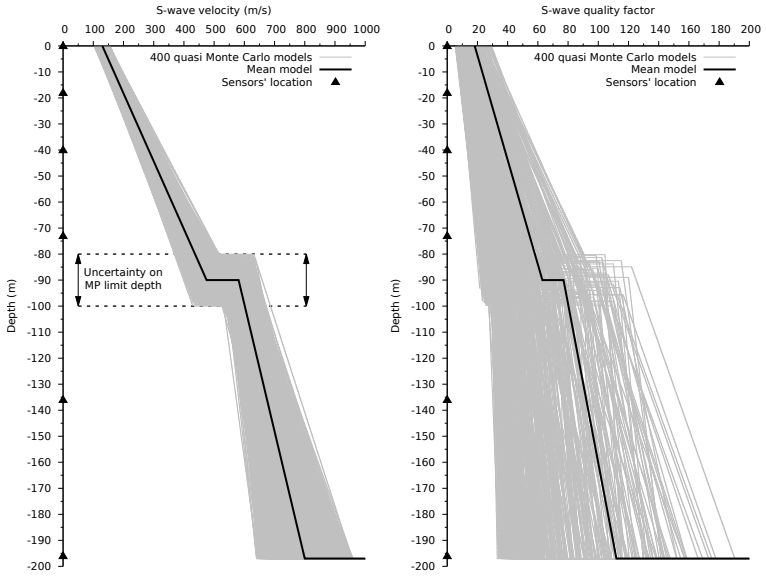

Figure 7: 400 QMC samples of the shear wave velocity v_s (left) and shear quality factor Q_s (right) profiles at the vertical of the TST station. The discontinuities corresponds to the uncertain MPL location. Also plotted are the nominal profiles.

Table 1 summarizes the resulting seven inputs of our uncertainty settings, while Table 2 reports, for completeness, the remaining deterministic inputs of the model. In particular, we notice that the compres-

sional wave velocity model also relies on piecewise affine profiles in the sedimentary layers. As a result, the uncertainty in the MPL (δd) indirectly induces some changes in the compressional wave velocity model within the sedimentary basin, even though the values at the top and bottom boundaries of the layers are deterministic.

	$v_{\mathrm{s},1}^{\mathrm{top}}$	$v_{\mathrm{s},1}^{\mathrm{bot}}$	$v_{\mathrm{s},2}^{\mathrm{top}}$	$v_{\mathrm{s},2}^{\mathrm{bot}}$	$v_{\mathrm{s,bed}}$	q_{s}	δd
Nominal	130 (m/s)	475 (m/s)	581 (m/s)	800 (m/s)	2,100 (m/s)	12.5	0 (m)
Range	$[104, 156]$	$[428, 523]$	$[523, 639]$	$[640, 960]$	$[1920, 2280]$	$[5, 20]$	$[-10, 10]$

Table 1: The nominal values and ranges of the 7 uncertain inputs retained in the uncertainty analysis.

	Layer 1		Layer2		Bedrock
	top	bot	top	bot	
ρ (kg/m^3)	2,075	2,130	2,130	2,250	2,500
v_{p} (m/s)	1,500	2,100	2,100	2,700	4,270
$Q_{\mathrm{s}}, Q_{\mathrm{p}}$ (m/s)	–	–	–	–	5,000

Table 2: Deterministic inputs of the seismic problem.

3.2. Canonical parametrization

Consider an abstract probability space $\mathcal{P} = (\Theta, \Sigma, d\mu)$, where Θ is the set of random events, Σ the associated σ-algebra, and $d\mu$ the probability measure. The seven uncertain inputs are defined in $\mathcal{P}$ and are assumed to be independent and uniformly distributed within the variability ranges reported in Table 1. For convenience, each of the random input can be parametrized using a canonical (standard) real-valued random variable, by means of suitable transformations. Let $\boldsymbol{\xi}(\theta) = (\xi_1(\theta), \ldots, \xi_7(\theta))$ be the random vector gathering all the canonical random variables, Ξ the domain of $\boldsymbol{\xi}$ and $p_{\boldsymbol{\xi}}$ its probability density function; we define the image probability space $\mathcal{P}_{\boldsymbol{\xi}} = (\Xi, \mathcal{B}, p_{\boldsymbol{\xi}})$, where $\mathcal{B}$ is the Borel set of Ξ. Owing to the uniform distribution for the input parameters, a straightforward choice is to use $\xi_i \sim \mathcal{U}([0, 1])$ for all the canonical variables, together with an affine mapping to recover the (uniform) distribution of the random input associated to the ξ_i, namely $v_{\mathrm{s},1,2}^{\mathrm{top,bot}}$, $v_{\mathrm{s,bed}}$, q_{s}, and δd. In this case, we have $\Xi = [0, 1]^7$ and

$$p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \begin{cases} 1, \boldsymbol{\xi} \in \Xi, \\ 0, \text{otherwise.} \end{cases} \tag{11}$$

Then a real-valued random output $u(\theta)$ of the model, is seen as a functional of the canonical variables $\boldsymbol{\xi}$, and we shall write $u(\boldsymbol{\xi})$. Furthermore, the expectation operator $\mathbb{E}(\cdot)$ can be defined either from the abstract probability space or the image space through the transfer theorem,

$$\mathbb{E}(u) = \int_{\Theta} u(\boldsymbol{\xi}(\theta)) d\mu(\theta) = \int_{\Xi} u(y) p_{\boldsymbol{\xi}}(y) dy. \tag{12}$$

Similarly, the variance operator $\mathbb{V}(\cdot)$ expresses as

$$\mathbb{V}(u) = \mathbb{E}\left((u - \mathbb{E}(u))^2\right) = \int_{\Xi} (u(y) - \mathbb{E}(u))^2 p_{\boldsymbol{\xi}}(y) dy. \tag{13}$$

In the following, we shall assume that all quantities of interest are second-order random quantity, that is have a finite variance.

3.3. QMC sample set and empirical mean and variance

As mentioned in the introduction, our UQ study relies on direct evaluations of EFISPEC3D associated with a sample set of uncertain inputs. Following the discussion above, we have to sample the parametric domain Ξ to generate the sample set of inputs. Specifically, we use in the present paper a QMC procedure based on a Sobol sequence with scrambling [32–34] to generate $N = 400$ samples from $p_{\boldsymbol{\xi}}$.

Let $\mathcal{X} = \left\{\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(N)}\right\}$ be the QMC sample set of inputs, and $\mathcal{U} = \left\{u^{(1)}, \ldots, u^{(N)}\right\}$ the corresponding ensemble of model values of the quantity of interest. The ensemble $\mathcal{U}$ allows to compute the empirical estimates of the mean and variance of $u(\boldsymbol{\xi})$:

$$\widehat{\mathbb{E}}_{\mathcal{X}}(u) = \frac{1}{N}\sum_{i=1}^{N} u^{(i)}, \quad \widehat{\mathbb{V}}_{\mathcal{X}}(u) = \frac{1}{N-1}\sum_{i=1}^{N}\left(u^{(i)} - \widehat{\mathbb{E}}_{\mathcal{X}}(u)\right)^2. \tag{14}$$

A confidence interval on the empirical estimators of $\mathbb{E}(m(u))$ (for instance, the mean or variance) can be obtained by bootstrapping. The bootstrap method [35] consists in i) generating a set of R replicas $\{\mathcal{X}_i\}$ from resampling with replacement of $\mathcal{X}$, and ii) computing the bootstrap error $\mathcal{E}(m(u))$ using these replicas through

$$\mathcal{E}(m(u)) = \left[\frac{1}{R-1}\sum_{i=1}^{R}\left(\widehat{\mathbb{E}}_{\mathcal{X}_i}(m(u)) - \widehat{\mathbb{E}}_{\mathcal{X}}(m(u))\right)^2\right]^{1/2}. \tag{15}$$

This procedure estimates the sampling error of an estimator. In other words, the bootstrap error measures the impact of the finite sampling on the estimator.

Figure 8 shows the empirical estimators of the mean and variance of the PGM field, together with their respective bootstrap errors (15). Here, the estimators are evaluated at every node of the free surface mesh to produce the spatial structure. The empirical mean highlights that the amplification (PGM > 1) of the motion occurs predominantly inside the sedimentary basin, with PGM levels close to 3 on average; along the basin boundary, the averaged PGM attains much lower values. The topography also emerges, outside the sedimentary basin, where a value lower than one (cool colors) represents an actual attenuation of the ground motion. The empirical variance shows roughly the same behavior as the mean, with high values inside the sedimentary basin, and vanishing values outside. The highest variance appears in the eastern part of the basin, denoting an area more susceptible to the input variability. Finally, we observe that the bootstrap errors on the empirical estimators of the mean and variance are not significant, meaning that the QMC sample set $\mathcal{X}$, with $N = 400$ samples, yields reliable estimators of the two first moments of the PGM. This last point brings confidence in the fact that the sample set size is enough to have appropriately captured most of the events.

4. Model reduction

In this paper, we are interested in the uncertain PGM at the free surface of the three-dimensional domain. The two main issues encountered in the analysis of the PGM field are the size of the data and the low number of samples available. To overcome these difficulties, we rely on a surrogate model based on i) a reduction of data leading to a tractable set of spatial basis functions to represent the PGM field, ii) a functional approximation of the uncertain coordinates in the reduced basis of the field. The resulting surrogate model can then be used to perform the uncertainty analysis.

In what follows, the random PGM field at the free surface denoted $u(\boldsymbol{x}, \boldsymbol{\xi})$, is defined at the computational nodes lying on the free surface and we denote $\mathrm{u}(\boldsymbol{\xi})$ the random vector of the PGM at the N_{x} nodes of the free surface. Let $\widehat{\mathrm{u}}$ be the mean value of $\mathrm{u}(\boldsymbol{\xi})$; we set

$$\mathrm{u}(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \mathrm{u}'(\boldsymbol{\xi}), \tag{16}$$

such that the fluctuation vector u' is centered. The key idea of the reduction is to represent the fluctuating part of u as a random linear combination of few deterministic vectors.

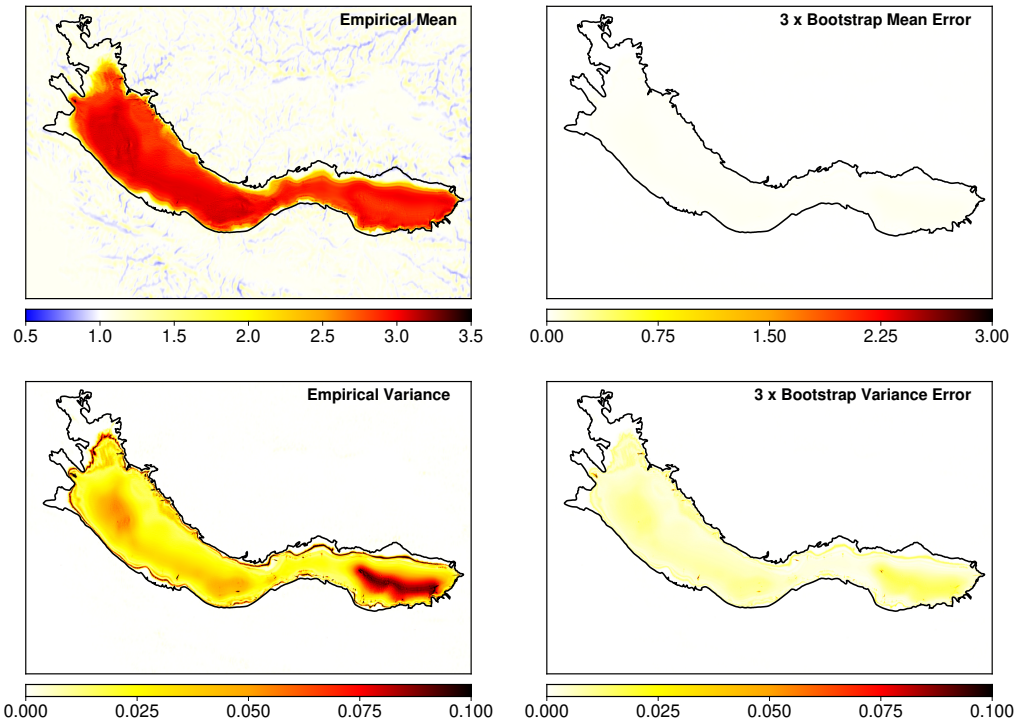

Figure 8: Empirical estimators and bootstrap errors of the PGM field mean (top line) and variance (bottom line). The bootstrap errors are based on $R = 10$ replicas.

4.1. Empirical Orthogonal Functions

Several approaches have been proposed to represent random fields or vectors using series expansions: the Karhunen-Loève Expansion (KLE) [36, 37], the Expansion Optimal Linear Estimation (EOLE) method [38], and the Orthogonal Series Expansion (OSE) [39]. Here, we rely on the KLE for its optimal convergence property: the truncated KLE minimizes in the mean squared sense the representation error. More specifically, we use the algebraic version of KLE through the Empirical Orthogonal Functions (EOF) [10]. These functions are the (normalized) eigenvectors u_k solving

$$C\mathrm{u}_k = \lambda_k \mathrm{u}_k, \quad \|\mathrm{u}_k\|_{\ell_2} = (\mathrm{u}_k^\top \mathrm{u}_k)^{1/2} = 1, \tag{17}$$

where $C \in \mathbb{R}^{N_\mathrm{x},N_\mathrm{x}}$ is the empirical covariance matrix of the random vector $\mathrm{u}'(\boldsymbol{\xi})$. The eigenvectors (or eigenmodes) $\mathrm{u}_k \in \mathbb{R}^{N_\mathrm{x}}$ form an orthonormal basis, that is $\mathrm{u}_k \cdot \mathrm{u}_l = \delta_{kl}$ with δ_{kl} the Kronecker delta. The covariance is not explicitly known and would have to be estimated from the ensemble $\{\mathrm{u}^{(1)}, \ldots, \mathrm{u}^{(N)}\}$. However, this approach is not practical for large N_x (recall that in our case $N_\mathrm{x} \approx 10^7$). Besides, the empirical estimate of C would have a maximal rank equal to $N - 1$. To avoid these issues, we use the snapshot method [40] which is based on the decomposition of the ensemble covariance matrix. Specifically, let $M = \left[\mathrm{u}^{(1)'}, \cdots, \mathrm{u}^{(N)'}\right] \in \mathbb{R}^{N_\mathrm{x},N}$ be the matrix of the ensemble fluctuations, and set the reduced covariance $\widetilde{C} = M^\top M \in \mathbb{R}^{N,N}$. The snapshot method proceeds as follow: i) solve the (reduced) eigenvalue problem

$$\widetilde{C}\mathrm{w}_k = \lambda_k \mathrm{w}_k, \tag{18}$$

for $\mathrm{w}_k \in \mathbb{R}^N$, ii) compute the desired eigenvectors using

$$\mathrm{u}_k = \frac{M\mathrm{w}_k}{\|M\mathrm{w}_k\|_{\ell_2}}.$$

The indexation of the eigenmodes can be made such that $\lambda_1 \geq \lambda_2 \geq \dots \lambda_N \geq 0$ (the positivity of the eigenvalues comes from the definition of the covariance matrix). The truncated EOF approximation of the random PGM vector $\mathrm{u}(\boldsymbol{\xi})$ is finally obtained by retaining the first r dominant eigenmodes, leading to

$$\mathrm{u}^r(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \sum_{k=1}^{r} \sqrt{\lambda_k}\mathrm{u}_k\zeta_k(\boldsymbol{\xi}). \tag{19}$$

In Eq. 19, the EOF random coordinates ζ_k can be determined by the projection of the random vector on the corresponding component, $\zeta_k(\boldsymbol{\xi}) := \mathrm{u}'(\boldsymbol{\xi})^\top \mathrm{u}_k/\sqrt{\lambda_k}$, because the eigenfunctions are orthonormal. It is known that this procedure leads to the optimal representation error in the Mean Squared Error (MSE) sense. In fact, the error is given by the sum of the disregarded eigenvalues

$$\mathbb{E}\left(\|\mathrm{u}(\boldsymbol{\xi}) - \mathrm{u}^r(\boldsymbol{\xi})\|_{\ell_2}^2\right) = \sum_{k>r} \lambda_k.$$

Figure 9 shows the decay of the first eigenvalues λ_k for the snapshots covariance matrix estimated from the ensemble of the PGM. The decay of the eigenvalues is fast (note the logarithmic scale of the vertical axis): the first mode contains 71% of the total fluctuation energy, the second 20% and the third 4%. As a result, with just the first three dominant modes, the global MSE is less than 5%.

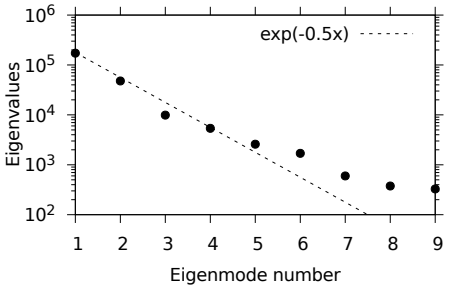

Figure 9: First eigenvalues λ_k of the EOF decomposition.

The three dominant eigenmodes $\mathrm{u}_{k=1,2,3}$ (scaled by $\sqrt{\lambda_k}$ respectively) are plotted in Fig. 10 using the computational mesh. It is first observed that the modes are non-zero only inside the basin, that is only in areas where the variance is non-zero (see the empirical variance field in Fig. 8). The first eigenmode has a constant sign over the whole basin and can be associated with the body wave. The second mode takes significant values along the internal boundary of the basin (corresponding to the surface waves) and exhibits different signs at the interior of the domain. The third eigenmode, which corresponds to lower energy fluctuation, is also essentially supported in the interior of the basin and involves a finer spatial structure accounting for differences between the western and eastern part of the basin. The last plot, in the bottom-right panel, details the MSE at mesh point, where the averaged is taken as the sample averaged based on $\mathcal{X}$. One can observe that the error level is low compared to the variance level in Fig. 8, with a color range ten times less than for the empirical variance. Regarding the spatial structure of the reduction error, it suggests that the neglected eigenmodes containing the remaining 5% energy are only affecting the edge of the basin where surface waves relexion occurs.

4.2. Polynomial chaos expansion

Once the main spatial eigenmodes $\{\mathrm{u}_k\}_{1\leq k\leq r}$ are set, the next step is to represent the dependence on $\boldsymbol{\xi}$ of each EOF coordinate $\zeta_k(\boldsymbol{\xi})$. To approximate $\zeta_k(\boldsymbol{\xi})$ we choose here to rely on Polynomial Chaos (PC) expansions [5, 11]. Other choices of a non-intrusive surrogate model method are possible, such as Gaussian processes [41], but several reasons motivate the use of PC expansions. First, PC expansions are known to be reliable and very accurate for the approximation of smooth quantities of interest in low to moderate

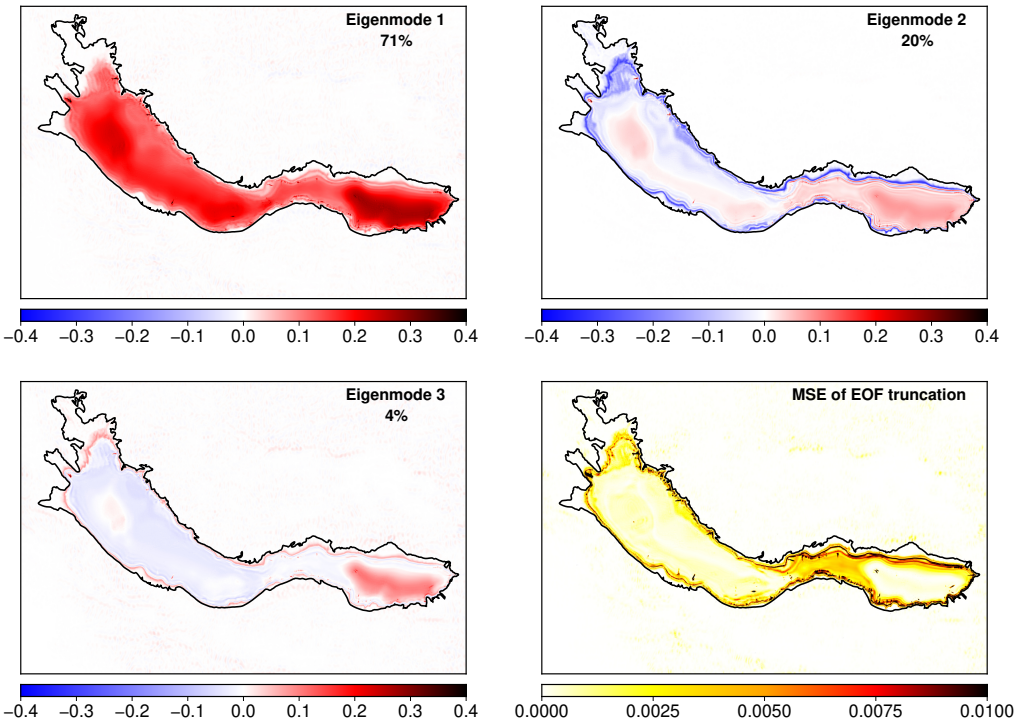

Figure 10: Dominant rescaled eigenmodes $\mathrm{u}_k\sqrt{\lambda_k}$ of the EOF decomposition, and (componentwise) empirical $\mathrm{MSE}(\mathrm{u}, \mathrm{u}^r)$ for $r = 3$.

dimensions. The EOF coordinates have a limited smoothness in the present problem. It is due to the hyperbolic nature of the model and the non-linearity of the quantity of interest. However, given the reduced set of observations available for the construction, it is reasonable to seek for smooth approximations of the $\zeta_k(\boldsymbol{\xi})$. Second, the PC expansions are just polynomial approximations relying on a particular choice of orthogonal basis. Therefore, well-established procedures are applicable for model selection and the adaption of model sparsity to the available observations set. Finally, PC expansions can be easily re-sampled, at a little cost, to generate new replicas and therefore carry virtually any statistical characterization that would be too costly or demanding for other methods. For instance, we exploit in the next section the orthogonality of the PC basis to derive exact expressions for the global sensitivity indices from the PC expansion of the EOF coordinates.

4.2.1. PC basis

Any second order random functional $f : \Xi \subset \mathbb{R}^d \mapsto \mathbb{R}$ has a convergent PC expansion of the form

$$f(\boldsymbol{\xi}) = \sum_{\boldsymbol{l}} f_{\boldsymbol{l}} \phi_{\boldsymbol{l}}(\boldsymbol{\xi}), \tag{20}$$

where the function $\phi_{\boldsymbol{l}}$ are orthonormal polynomials in $\boldsymbol{\xi}$, such that

$$\int_{\Xi} \phi_{\boldsymbol{l}}(y)\phi_{\boldsymbol{k}}(y)p_{\boldsymbol{\xi}}(y)dy = \begin{cases} 1, & \boldsymbol{l} = \boldsymbol{k} \\ 0, & \text{otherwise.} \end{cases}$$

Classically, the $\phi_{\boldsymbol{l}}$ are d-variate polynomials defined by the tensorization of univariate polynomial families, whose particular choice depends on the distribution of the corresponding ξ_i [42]. In this context, the component of the multi-index $\boldsymbol{l} = (l_1, \ldots, l_d) \in \mathbb{N}^d$ refers to the individual degrees of the univariate polynomials

in each of the ξ_i. Therefore $\|\boldsymbol{l}\|_1 = \sum_{i=1}^d l_i$ is the total degree of $\phi_{\boldsymbol{l}}$, when $\|\boldsymbol{l}\|_\infty = \max_{i=1,\dots,d} l_i$ is the partial degree. In our case, where $d = 7$ and $\Xi = [0,1]^7$ with a uniform measure, the polynomials are simply the multivariate Legendre polynomials.

The PC approximation of the EOF coordinates $\zeta_k(\boldsymbol{\xi})$ reads

$$\zeta_k^{\mathcal{L}}(\boldsymbol{\xi}) = \sum_{\boldsymbol{l}\in\mathcal{L}} \zeta_{k,\boldsymbol{l}} \phi_{\boldsymbol{l}}(\boldsymbol{\xi}), \tag{21}$$

where $\{\zeta_{k,\boldsymbol{l}}\}_{\boldsymbol{l}\in\mathcal{L}}$ are *deterministic* coefficients, and $\mathcal{L}$ is the finite set of the expansion multi-indices. In the following we denote $\mathrm{P} = |\mathcal{L}|$ the dimension of the PC basis. We shall restrict ourself to the same set $\mathcal{L}$ to search the PC expansions the EOF coordinates, also regularization and sparsity promotion with eventually result in different expansion bases (see below). Plugging the PC expansions of the EOF coordinates in (19) yields the EOF-PC surrogate,

$$\mathrm{u}(\boldsymbol{\xi}) \approx \mathrm{u}^r(\boldsymbol{\xi}) \approx \mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \sum_{k=1}^r \sqrt{\lambda_k}\mathrm{u}_k \zeta_k^{\mathcal{L}}(\boldsymbol{\xi}). \tag{22}$$

4.2.2. Determination of the PC coefficients

Let us write the MSE resulting from the EOF-PC approximation of the random vector $\mathrm{u}(\boldsymbol{\xi})$. Since both the EOF modes u_k and EOF coordinates ζ_k are orthonormal, it comes

$$\begin{aligned}
\mathcal{R}^2 &= \mathbb{E}\left(\|\mathrm{u}(\boldsymbol{\xi}) - \mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi})\|_{\ell_2}^2\right) \\
&= \mathbb{E}\left(\|\mathrm{u}^r(\boldsymbol{\xi}) - \mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi})\|_{\ell_2}^2\right) + \mathbb{E}\left(\|\sum_{k>r}\sqrt{\lambda_k}\mathrm{u}_k(\boldsymbol{\xi})\|_{\ell_2}^2\right) \\
&= \mathbb{E}\left((\mathrm{u}^r(\boldsymbol{\xi}) - \mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi}))^\top(\mathrm{u}^r(\boldsymbol{\xi}) - \mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi}))\right) + \sum_{k>r}\lambda_k \\
&= \mathbb{E}\left(\sum_{k=1}^r \lambda_k(\mathrm{u}_k^\top \mathrm{u}_k)(\zeta_k(\boldsymbol{\xi}) - \zeta_k^{\mathcal{L}}(\boldsymbol{\xi}))^2\right) + \sum_{k>r}\lambda_k \\
&= \sum_{k=1}^r \lambda_k \mathbb{E}\left((\zeta_k(\boldsymbol{\xi}) - \zeta_k^{\mathcal{L}}(\boldsymbol{\xi}))^2\right) + \sum_{k>r}\lambda_k.
\end{aligned} \tag{23}$$

This result shows that the global MSE has two distinct contributions: the MSE associated to the PC approximation of the coordinates $\zeta_k(\boldsymbol{\xi})$ and the EOF truncation error with is fixed when r is set. Then, only the first contribution remains to be minimized. However, finding the PC approximations $\zeta_k^{\mathcal{L}}$ by the direct minimization of the MSE error in (23) is not possible as $\mathrm{u}(\boldsymbol{\xi})$ is not fully known. Instead, equation (23) suggests to consider the minimization problem discretized using the sample set $\mathcal{X}$,

$$\mathcal{R}_{\mathcal{X}}^2 = \sum_{j=1}^N \sum_{k=1}^r \lambda_k \left[\zeta_k^{(j)} - \zeta_k^{\mathcal{L}}(\boldsymbol{\xi}^{(j)})\right]^2, \tag{24}$$

where the we have denoted

$$\zeta_k^{(j)} \doteq (\mathrm{u}^{(j)} - \widehat{\mathrm{u}})^\top \mathrm{u}_k, \tag{25}$$

the projection of the (centered) j-th element of the PGM sample set $\mathcal{U}$ onto the k-th EOF eigenmode.

A regularized version of (24) is preferred, because of the limited number of samples available [43], and we solve for $\gamma > 0$ the following minimization problem

$$\underset{\boldsymbol{\zeta}_1,\cdots,\boldsymbol{\zeta}_r}{\arg\min} \left(\sum_{k=1}^r \lambda_k \|\boldsymbol{M}\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_k^{\mathrm{obs}}\|_{\ell_2}^2 + \gamma \sum_{k=1}^r \|\boldsymbol{\zeta}_k\|_{\ell_1}\right), \tag{26}$$

where $\boldsymbol{M} \in \mathbb{R}^{N,P}$ is the matrix with entries $\phi_{\boldsymbol{l}}(\boldsymbol{\xi}^{(i)})$, $\boldsymbol{\zeta}_k \in \mathbb{R}^P$ is the vector of the PC coefficients $\zeta_{k,\boldsymbol{l}}$, and $\boldsymbol{\zeta}_k^{\text{obs}} \in \mathbb{R}^N$ is the vector of EOF coordinates observations $\zeta_k^{(j)}$. The role of the ℓ_1-norm penalty term is to promote the sparsity of the PC expansion and prevent overfitting of the observations. Problem 26 can be solved with the basis pursuit denoising algorithm [12–14]. In practice, we implemented a two stages strategy. In a first stage, a constrained form of Problem 26 is solved for 40 values (logarithmically spaced over the interval $[10, 10^3]$) of the parameter that controls the fitting error. A K-fold cross-validation (CV) procedure is then employed to estimate the RMSE of the approximations and select the best value of the parameter. For each EOF coordinate, the corresponding set $\mathcal{L}_k \subseteq \mathcal{L}$ of active PC coefficients (*i.e.* non zero) is recorded. In a second stage we solved r ordinary least squares problems to determine the final PC expansion coefficients, say $\boldsymbol{\zeta}_k^*$, of $\zeta_k(\boldsymbol{\xi})$ in the span of the active modes $\{\phi_{\boldsymbol{l}}, \boldsymbol{l} \in \mathcal{L}_k\}$.

4.2.3. Validation

We now assess the predictive capacity of the PGM EOF-PC surrogate. We first focus on the PC approximation of the EOF coordinates which is analysed in Table 3. For each of the first three ζ_k, the table reports ℓ_1 and ℓ_0-norm of their vector of PC coefficients, for approximations in PC bases of increasing polynomial order p (from top to bottom). Specifically, the table reports for all degree p, the dimension of the PC basis $\mathrm{P} = |\mathcal{L}|$ (second column) as well as the number of active PC coefficients of ζ_k measured by its ℓ_0-norm. The initial multi-index set $\mathcal{L}$ uses on the total order truncation, so

$$\mathrm{P} + 1 = \frac{(d+p)!}{p!d!},$$

where $d = 7$ in our case. Therefore comparing P and $\|\boldsymbol{\zeta}_k^*\|_{\ell_0}$ is a direct measure of the parsimony. The table also reports the ℓ_1-norm of the PC approximations of the EOF coordinates. We observe that, as p increases, the number of active coefficients increases and eventually levels up. It is also remarked that more active coefficients are needed for ζ_1, compared to ζ_2, while even less coefficients are retained for ζ_3: $\|\boldsymbol{\zeta}_1^*\|_0 > \|\boldsymbol{\zeta}_2^*\|_0 > \|\boldsymbol{\zeta}_3^*\|_0$. This behavior is consistent with the higher energy content associated to ζ_1 compared to ζ_3. However, we observe that all the three coordinates demand a comparable PC degree. Similarly, the ℓ_1-norms of the final coordinate expansions are seen to converge, up to fluctuations caused by the validation procedure, when p increases, providing further confidence in the convergence of the approximation with p.

The active coefficients corresponding to the univariate polynomials $\phi_{\boldsymbol{l}}$ are plotted in Fig. 11 for $p = 10$. The missing dots indicate non-active coefficients, thus reflecting the sparsity pattern. The approximations of the three ζ_k show less and less univariate polynomials: 50 for the first EOF coordinate, 41 for the second one, and 34 for the third one. As a closing remark, we mention that if one selects the active coefficients of the ζ_k independently, considering for instance the k minimization problems

$$\arg\min_{\boldsymbol{\zeta}_k} \left(\|\boldsymbol{M}\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_k^{\text{obs}}\|_{\ell_2}^2 + \gamma_k \|\boldsymbol{\zeta}_k\|_{\ell_1} \right), \tag{27}$$

the resulting PC expansions $\zeta_k^{\mathcal{L}}$ have similar precision (overall), but are less parsimonious with more active coefficients compared to the global approach in (26). Indeed, the decoupled approach does not perceive the relative importance of the EOF modes and tend to retain more active modes when $k = 2, 3$.

To assess the convergence of the EOF-PC surrogate we present the cross-validation (CV) errors in the last columns of Table 3. We follow the classical K-fold CV technique [44, chap. 7], which is common when validation data are missing, to compute the CV-errors. The table reports the total CV-error $\mathcal{E}^2$ defined by the weighted sum

$$\mathcal{E}^2 = \sum_{k=1}^{r} \lambda_k \mathcal{E}_k^2,$$

where $\mathcal{E}_k^2$ is the CV-error on the k-th coordinate $(\zeta_k(\boldsymbol{\xi}) - \zeta_k^{\mathcal{L}}(\boldsymbol{\xi}))^2$. We have repeated the CV-error estimation for different numbers of folds, $K = 5$, 10 and 20, and found that the estimation of $\mathcal{E}^2$ was robust w.r.t. to K (not shown). All CV-errors reported hereafter use the CV procedure with $K = 10$ folds. Table 3 reports the $\mathcal{E}^2$ in the form of mean value and standard deviation (following the $\pm$ sign) estimated using

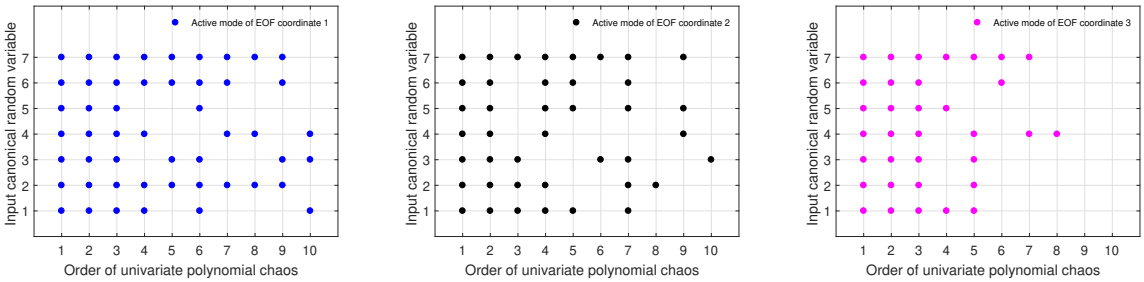

Figure 11: Subsets of active PC coefficients associated to the univariate polynomials. Results are shown for the 3 EOF coordinates $\zeta_k^{\mathcal{L}}$ using a PC degree $p = 10$.

20 random replicas of the validation sets. We observe that the dispersion of the CV-errors remain less than 6% of their averaged values, and are not diverging as the polynomial degree is increasing. This fact demonstrates the convergence of the approximation and the effectiveness of the regularization in preventing overfitting. The relative total CV-errors, $\mathcal{E}_{\text{rel}} = \sqrt{\mathcal{E}^2/(\sum_{k=1}^{r} \lambda_k)}$, are also reported in the last column of the table to appreciate the importance of the CV-errors. As expected, the CV-errors initially decrease with the polynomial degree p and eventually stagnate for the highest PC degrees (9 and 10). From these degrees, more simulations would be necessary to improve further the EOF-PC surrogate. For the ensemble of simulations available to construct the approximation, it is seen that the minimal relative CV-error on the PGM is estimated to be lower or about 1%. Of course, the CV-errors are not sharp estimates of the actual surrogate errors, which are expected to be larger. However, the relative CV-error $\mathcal{E}_{\text{rel}}^2$ is around 10^{-5}, so it is significantly lower than the EOF truncation error (10^{-2}). Having a CV-error which is much less than the EOF truncation error, we can safely state that the latter dominates the PC error on the coordinates. This aspect is further illustrated in Fig. 12, which depicts the spatial distributions of the empirical MSE due to the EOF and EOF-PC approximations. The empirical MSE is estimated using the simulations set. It shows that the two errors are comparable on the elements of the samples set. Because the PC construction method prevents overfitting, it suggests that the PC truncation error is indeed negligible compared to the EOF truncation.

		Coordinate 1		Coordinate 2		Coordinate 3		CV Errors	
p	P	$\|\boldsymbol{\zeta}_1^*\|_{\ell_1}$	$\|\boldsymbol{\zeta}_1^*\|_{\ell_0}$	$\|\boldsymbol{\zeta}_2^*\|_{\ell_1}$	$\|\boldsymbol{\zeta}_2^*\|_{\ell_0}$	$\|\boldsymbol{\zeta}_3^*\|_{\ell_1}$	$\|\boldsymbol{\zeta}_3^*\|_{\ell_0}$	$\mathcal{E}^2$ $(\pm STD)$	$\mathcal{E}_{\text{rel}}$
5	792	5.35	165	5.02	99	6.66	56	22.4 $\pm$ 1.00	0.99%
6	1716	5.50	251	5.35	158	6.83	75	12.0 $\pm$ 0.32	0.72%
7	3432	5.43	240	5.40	179	6.87	83	7.85 $\pm$ 0.31	0.58%
8	6435	5.39	272	5.50	232	6.97	104	6.20 $\pm$ 0.29	0.52%
9	11440	5.43	298	5.55	254	7.06	123	5.58 $\pm$ 0.28	0.49%
10	19448	5.41	289	5.55	257	7.05	119	5.22 $\pm$ 0.31	0.48%

Table 3: Initial degree p, dimension of PC basis (P), ℓ_1 and ℓ_0-norms of the PC approximations of the EOF coordinates, total ($\mathcal{E}^2$) and relative ($\mathcal{E}_{\text{rel}}$) CV-errors. CV-errors are estimated using 10 folds and 20 replicas of the validation sets.

5. Results

In this section, we highlight the statistical information that can be retrieved from the surrogate. Because the variance outside the sedimentary basin is negligible, as shown previously in Fig. 8, we carry out the subsequent analysis inside the basin only. First, the estimation of PGM probability density function can be particularly useful in soil-structure interaction in which the ground motion directly affect the acceleration

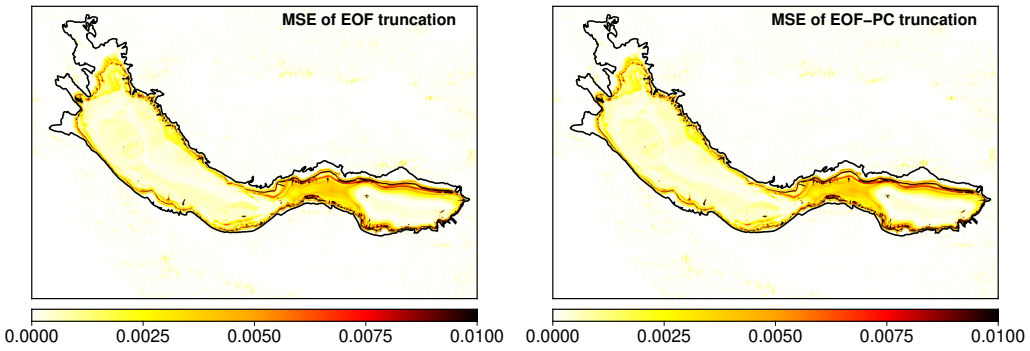

Figure 12: Empirical $\mathrm{MSE}(\mathrm{u}, \mathrm{u}^r)$ and $\mathrm{MSE}(\mathrm{u}, \mathrm{u}^{r,\mathcal{L}})$.

experienced by the structure. Second, interval probability maps provide information on the most vulnerable areas subjected to an earthquake. Third, the sensitivity analysis allows us to rank the contribution of the inputs parameters, and their interaction with others, onto the variance of the output quantity.

5.1. Probability density functions

Figure 13 shows the locations of three stations S_1, S_2, and S_3, where the probability density function (PDF) of the PGM is estimated by the EOF-PC surrogate (of order 4). To assess the impact of the EOF reduction, the PC expansions of the PGM are also directly computed at these stations, extracting the values at the corresponding locations for the 400 QMC simulations. Figure 14 depicts the PDFs estimated by the sampling of the EOF-PC and PC surrogates with 10^4 realizations, as well as the histograms of the 400 QMC realizations. For each station, the PDFs of the EOF-PC and direct PC surrogates are in close agreement, denoting a low impact of the EOF reduction on the uncertain prediction. Also, the agreement with the empirical histogram based on the 400 QMC samples is satisfying. However, the ability to sample the PC surrogates with as many realizations as needed provides much smoother and detailed PDFs. Finally, the shape of the PDF depends significantly on the location of the station, reflecting the strong spatial dependence of the susceptibility of the PGM on the uncertain seismic wave velocity model. One can speculate that this result extends everywhere in the basin, meaning that using only three EOF modes is enough to estimate the PDFs of the PGM correctly at all the mesh nodes (about 10 million).

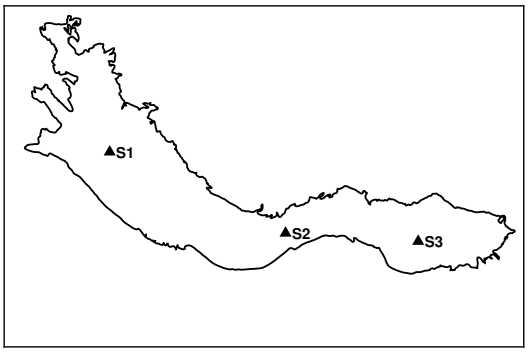

Figure 13: Location of stations S_1, S_2 and S_3.

To stress further the advantage of the surrogate based approaches, we report in Figure 15 the joint densities of the PGM for the three couples of stations (S_1, S_2), (S_1, S_3) and (S_2, S_3). The plots show the iso-values of the joint PDFs obtained using the 400 QMC realizations (top row) and 10^5 realizations drawn from the EOF-PC surrogate (bottom row). Here, we relied on a standard Kernel Density Estimation

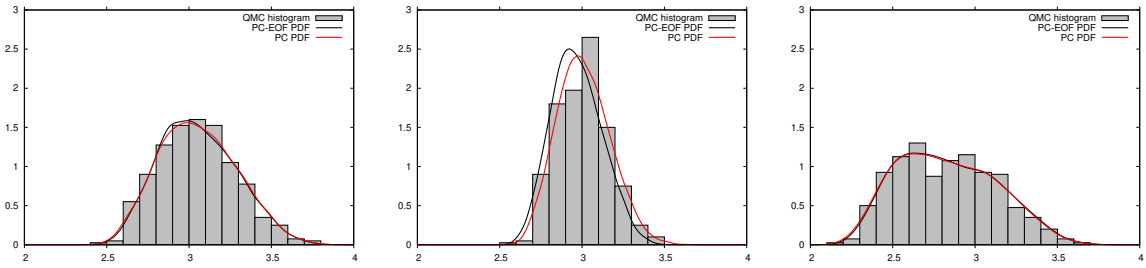

Figure 14: PDFs and ensemble histograms of the PGM at stations S_1 (left), S_2 (middle), and S_3 (right).

(KDE) method [45] to estimate the joint densities. We again observe the qualitative agreement of the two estimations, for all the couples of stations, although the width of the densities appears overestimated for the direct estimation from the 400 QMC realizations. This effect is explained by the limited number of samples that requires a wider kernel size in the KDE, and it would increase when considering the estimation of joint-density function between an increasing number of stations. This behavior underlines the importance of being capable of sampling the surrogates sufficiently to obtain converged estimations. Further, the EOF-PC construction provides the advantage of sampling the whole PGM field at once using a limited number of modes. In contrast, to estimate the joint densities between a large number of couples of stations, one would need to construct the PC expansions of the PGM at *all* the stations in the direct PC approach.

As a final note on the joint-densities, we remark that the isovalues reported in Fig. 15 highlight the correlations between the PGM at the stations: the coefficients of determination (equal to the part of the variance part explained by a linear relation) are 0.9 for the couple (S_1, S_2), 0.88 for (S_1, S_3) and 0.73 for (S_2, S_3).

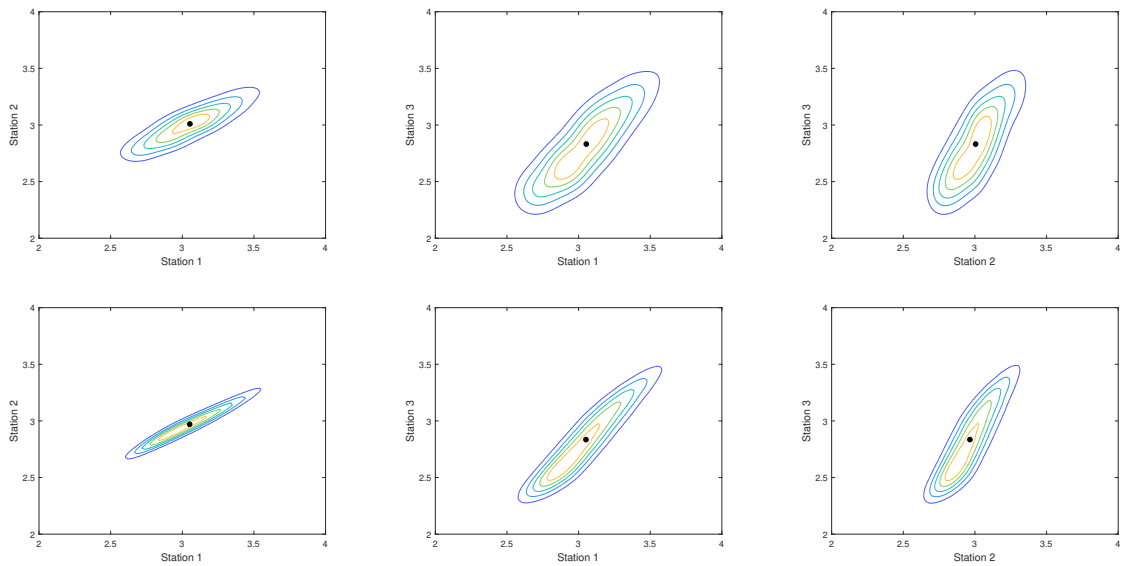

Figure 15: Five iso-values of joint PDFs of the PGM at the couple of stations (S_1, S_2) (left), (S_1, S_3) (middle), and (S_2, S_3) (right). Densities are estimated by bivariate KDE. The top row corresponds to the estimation using the 400 QMC realizations while the bottom row is derived from 10^5 samples of the EOF-PC surrogate.

5.2. Interval probability maps

Another information of interest that one can retrieve cheaply from the EOF-PC surrogate is the localization of the most vulnerable areas, where the buildings are at risk of breaking and collapsing under an earthquake event. These areas can be identified using interval probability maps. Let I be a (bounded or unbounded) interval; the interval probability at the i-th node $\boldsymbol{x}_i$ of the mesh is the probability that the PGM at this node belongs to I. It is expressed as

$$\mathbb{P}\left(u(\boldsymbol{x}_i, \boldsymbol{\xi}) \in I\right) = \int_{\Xi} \mathbb{1}_{u(\boldsymbol{x}_i,\boldsymbol{\xi})\in I}\; p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \tag{28}$$

where $\mathbb{1}_{\mathcal{S}}$ is the indicator function of the set $\mathcal{S}$. The integral (28) is estimated by sampling the surrogate with M random realizations $\{\boldsymbol{\xi}^{(j)}\}_{1\leq j\leq M}$ drawn from the distribution $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, and estimating the interval probability by

$$\mathbb{P}\left(\mathrm{u}_i^{r,\mathcal{L}}(\boldsymbol{\xi}) \in I\right) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\mathrm{u}_i^{r,\mathcal{L}}(\boldsymbol{\xi}^{(j)})\in I}. \tag{29}$$

Figure 16 presents the interval probabilities given by Eq. (29) for the following four intervals:

$$I_1 = (-\infty, 2.5), \quad I_2 = (2.5, 3), \quad I_3 = (3, 3.5), \quad \text{and} \quad I_4 = (3.5, \infty).$$

These maps have been obtained with $M = 10^4$ LHS realizations, a value that was selected from a convergence study not shown for briefness. The top left panel shows that the PGM is almost always lower than 2.5 at the boundary of the basin, while the bottom right panel indicates that the probability of the PGM to exceed 3.5 is low in all the basin. These findings are consistent with the mean and standard deviation fields shown previously in Fig. 8. The probabilities of the intervals I_2 and I_3 bring more detailed information on the spatial response measured by the PGM. The probability that the PGM is within the range $(2.5, 3)$ is higher than $1/2$ everywhere inside the basin, except in the vicinity of its boundary. It is, however, lower (around 0.5) in the central area of the basin, while it peaks to 1 in a stripe along the edge of the basin, where the PGM is unlikely to exceed 3. The probability that the PGM ranges over $(3, 3.5)$ globally decreases with values not exceeding 60%.

5.3. Sensitivity fields

We provide a global sensitivity analysis of the PGM using a decomposition of the variance. The analysis uses the Sobol–Hoeffding (SH) decomposition of the output, leading to normalized elementary contributions to the variance, the so-called Sobol indices (see Appendix A). We further exploit the representation of the PGM field and start with the SH decomposition of the EOF coordinates (A.1). Introducing the PC expansion of the EOF coordinates, we get [46]

$$\zeta_k(\boldsymbol{\xi}) = \sum_{\boldsymbol{i}\subseteq\mathcal{D}} \zeta_k^{(\boldsymbol{i})}(\boldsymbol{\xi}_{\boldsymbol{i}}) = \sum_{\boldsymbol{i}\subseteq\mathcal{D}}\sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}} \zeta_{k,\boldsymbol{l}}\phi_{\boldsymbol{l}}(\boldsymbol{\xi}), \tag{30}$$

where $\mathcal{D} = \{1, 2, \ldots, d\}$ and $\mathcal{L}_{\boldsymbol{i}}$ is the subset of PC multi-indexes associated to $\boldsymbol{i}$ defined as

$$\mathcal{L}_{\boldsymbol{i}} := \{\boldsymbol{l} \in \mathcal{L}(p),\ l_{i\in\boldsymbol{i}} > 0 \text{ and } l_{i\in\boldsymbol{i}_\sim} = 0\}$$

with $\boldsymbol{i}_\sim := \mathcal{D}\setminus\boldsymbol{i}$. Plugging the SH decomposition (30) into the EOF-PC surrogate, and arranging, we obtain

$$\mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \sum_{\boldsymbol{i}\subseteq\mathcal{D}}\sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}}\left(\sum_{k=1}^{r}\sqrt{\lambda_k}\mathrm{u}_k\zeta_{k,\boldsymbol{l}}\right)\phi_{\boldsymbol{l}}(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \sum_{\boldsymbol{i}\subseteq\mathcal{D}}\sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}}\mathrm{v}_{\boldsymbol{l}}\phi_{\boldsymbol{l}}(\boldsymbol{\xi}). \tag{31}$$

Observing that the $\mathcal{L}_{\boldsymbol{i}}$ form a partition of the PC multi-index set $\mathcal{L}$, it is seen that the vectors $\mathrm{v}_{\boldsymbol{l}}$ are actually the PC modes of reduced stochastic field $\boldsymbol{u}^{r,\mathcal{L}}(\boldsymbol{\xi})$. For a given multi-index $\boldsymbol{i}$, the estimation of the associated partial variance field $\mathbb{V}_{\boldsymbol{i}}(\boldsymbol{u}^{r,\mathcal{L}})$ from (31) therefore requires to compute the N_{x} components of the PC modes

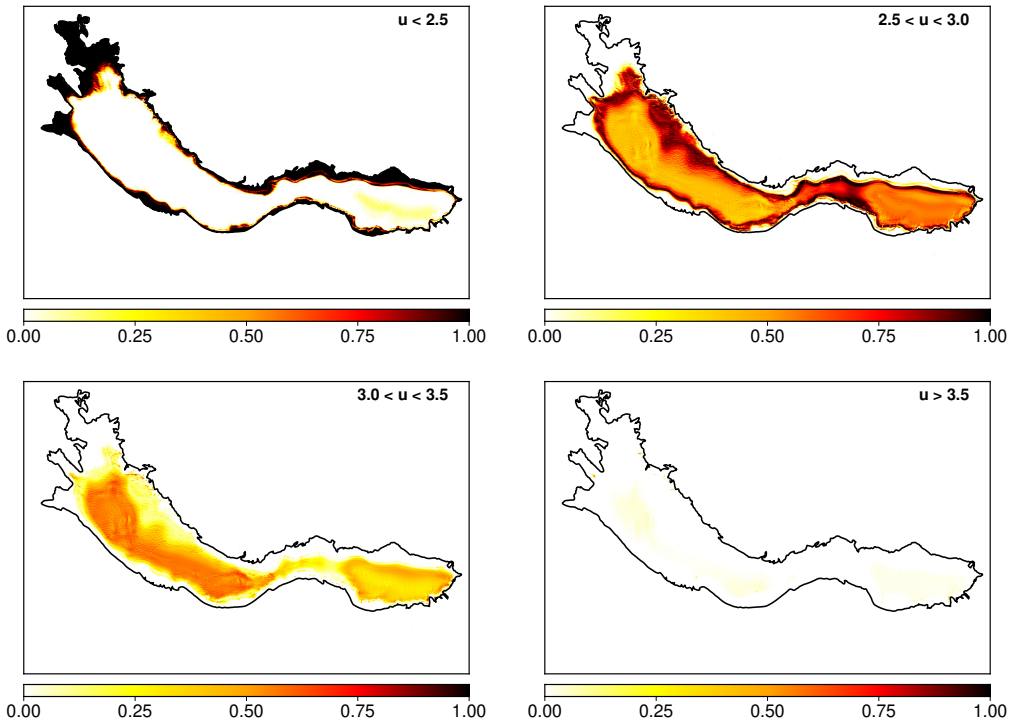

Figure 16: PGM interval probability maps for different intervals obtained with a LHS of 10^4 realizations.

$\mathrm{v}_{\boldsymbol{l}} = \sum_{k=1}^{r} \sqrt{\lambda_k}\mathrm{u}_k\zeta_{k,\boldsymbol{l}}$, for all the multi-indices $\boldsymbol{l} \in \mathcal{L}_{\boldsymbol{i}}$. When a low number r of EOF modes is selected, as it is the case here, we can compute more efficiently the partial variance fields. To this end, we rewrite (31) to

$$\mathrm{u}^{r,\mathcal{L}}(\boldsymbol{\xi}) = \widehat{\mathrm{u}} + \sum_{\boldsymbol{i}\subseteq\mathcal{D}} \left(\sum_{k=1}^{r} \sqrt{\lambda_k}\mathrm{u}_k \sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}} \zeta_{k,\boldsymbol{l}}\phi_{\boldsymbol{l}}(\boldsymbol{\xi}) \right) = \widehat{\mathrm{u}} + \sum_{\boldsymbol{i}\subseteq\mathcal{D}} \mathrm{u}^{(\boldsymbol{i})}(\boldsymbol{\xi}), \tag{32}$$

where $\mathrm{u}^{(\boldsymbol{i})}(\boldsymbol{\xi}) \doteq \sum_{k=1}^{r} \sqrt{\lambda_k}\mathrm{u}_k \sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}} \zeta_{k,\boldsymbol{l}}\phi_{\boldsymbol{l}}(\boldsymbol{\xi})$. In fact, the stochastic vectors $\mathrm{u}^{(\boldsymbol{i})}$ are the elements of the SH decomposition of u. The determination of these elements has the same complexity as the computation of the PC modes of the reduced representation. However, expressing the variance and relying on the variance sum law, the variance of $\mathrm{u}^{(\boldsymbol{i})}$ can be reformulated in

$$\mathbb{V}(\mathrm{u}^{(\boldsymbol{i})}) = \sum_{k,m=1}^{r} \sqrt{\lambda_k\lambda_m}\mathrm{u}_k * \mathrm{u}_m \mathbb{C}\mathrm{ov}_{\boldsymbol{i}} \left(\zeta_k^{\mathcal{L}}(\boldsymbol{\xi}), \zeta_m^{\mathcal{L}}(\boldsymbol{\xi}) \right), \tag{33}$$

where $*$ denotes the component-wise product, and the partial covariance $\mathbb{C}\mathrm{ov}_{\boldsymbol{i}}$ between the two PC expansions reads

$$\mathbb{C}\mathrm{ov}_{\boldsymbol{i}} \left(\zeta_k^{\mathcal{L}}(\boldsymbol{\xi}), \zeta_m^{\mathcal{L}}(\boldsymbol{\xi}) \right) = \sum_{\boldsymbol{l}\in\mathcal{L}_{\boldsymbol{i}}} \zeta_{k,\boldsymbol{l}}\zeta_{m,\boldsymbol{l}} \|\phi_{\boldsymbol{l}}\|^2. \tag{34}$$

Then, for moderate r, the low size partial covariance matrices can be computed once for all and reused subsequently to evaluate the component of the partial variance vector (33).

We report in Fig. 17 the first-order sensitivity indices and the global measure of the interaction effects between the random inputs of the model (see Appendix A). For clarity, the plots of the first-order sensitivity fields have been ordered by they maximal intensity over the basin. In Fig. 17-(a), we report the fraction of the PGM variance within the sedimentary basin, which is due to the uncertainty in the location of the MP interface. One sees that this input uncertainty is responsible for all the variability at the edges of the

basin. The dominance of the uncertain location is because the positioning of the MP interface strongly impacts the velocity profile below the sedimentary basin/bedrock limit; further, the first or the second layer can even vanish close to the basin boundary for some realizations. In contrast, the impact of the location of the interface quickly drops to zero as the considered point moves inside the basin. Recalling that the uncertainty (variability) in the PGM value is low at the boundary of the basin, the effect of the uncertainty in the MP interface location is overall quite limited.

Figure 17-(b) shows the first-order sensitivity index associated with the scaling factor. The uncertainty in this factor is seen to induce a significant fraction of the PGM variance within the whole of the domain, including in the areas where the variance is the largest. In particular, the index reaches values as high as 0.8 on the east side of the domain. The importance of this parameter is not surprising as it modifies the damping characteristics of the medium, with a direct impact on the amplitude of the motion. A previous study [47] reached a similar conclusion in the one-dimensional case.

The next two contributions are the top velocity of the first layer (Fig. 17-(c)) and the velocity of the bedrock (Fig. 17-(d)). We see that the first-order sensitivity fields associated with these two inputs are similar in their magnitude and spatial structure. They are both responsible for 20-25% of the variance within the bulk of the basin, except on the eastern part where their influence is less but not vanishing. The case of the velocity at the bottom of the second layer, shown in Fig. 17-(e) is similar, but exhibits slightly lower values for the first-order indices in the basin. In contrast, in Fig. 17-(f), the bottom velocity of the first layer is seen to induce much less variability within the basin and to affect only the immediate neighborhood of the basin boundaries, where it is never dominant but only accounts for a small fraction of variance. Finally, the top velocity of the second layer has a vanishingly small first-order sensitivity index everywhere in the basin (see Fig. 17-(g)). Since the total interaction effect (shown in Fig. 17-(h)) is seen to be also very small, it is possible to conclude that i) the effects of the different uncertain input are additive, and ii) the uncertainty in the top velocity of the second layer (and possibly the bottom velocity of the first one) could entirely disregard in subsequent investigations.

.

6. Conclusion

The article presented an uncertainty study for large-scale seismic wave propagation assuming a poorly known velocity model. A three-dimensional spectral finite element parallelized solver, EFISPEC3D, was used to perform 400 ensemble simulations launched on a supercomputer to supply the computational load of about 4 million core-hours of CPU time. To tackle with the massive data size and the low number of samples, the analysis of the peak ground motion is based on a surrogate approach that combines dimension reduction technique with empirical orthogonal functions and functional approximation with polynomial chaos expansion. The results in the sedimentary basin emphasize three findings. First, the surrogate reproduces faithfully and efficiently the spatial dependence of the probability density functions of the peak ground motion, as well as joint densities between arbitrary locations. Second, the interval probability maps show that most of the peak ground motion realizations are lower than 3.5, with the lowest values at the bedrock-basin interface and highest values inside the basin. Last, a global sensitivity analysis highlights that the scaling factor is the main contributor to the peak ground motion uncertainty.

In future works, the methodology will have to be applied to more realistic and complete uncertainty models for the wave velocities, yielding a larger number of stochastic dimension to sample. We are also planning to take into account the time variation of the ground motion, and not only the peak values. Starting with the analysis at selected spatial locations (stations), the ultimate goal will be the analysis of the uncertain dynamical ground motion in the whole domain. This objective will require to deal with the 60 TB of data (for the same number of simulations), calling for additional reduction methods.

7. Acknowledgment

The work of P. Sochala and F. De Martin have been supported by internal funding of BRGM. The authors are grateful to D. Keyes and P. Thierry for supporting the project " Earthquake Ground Motion

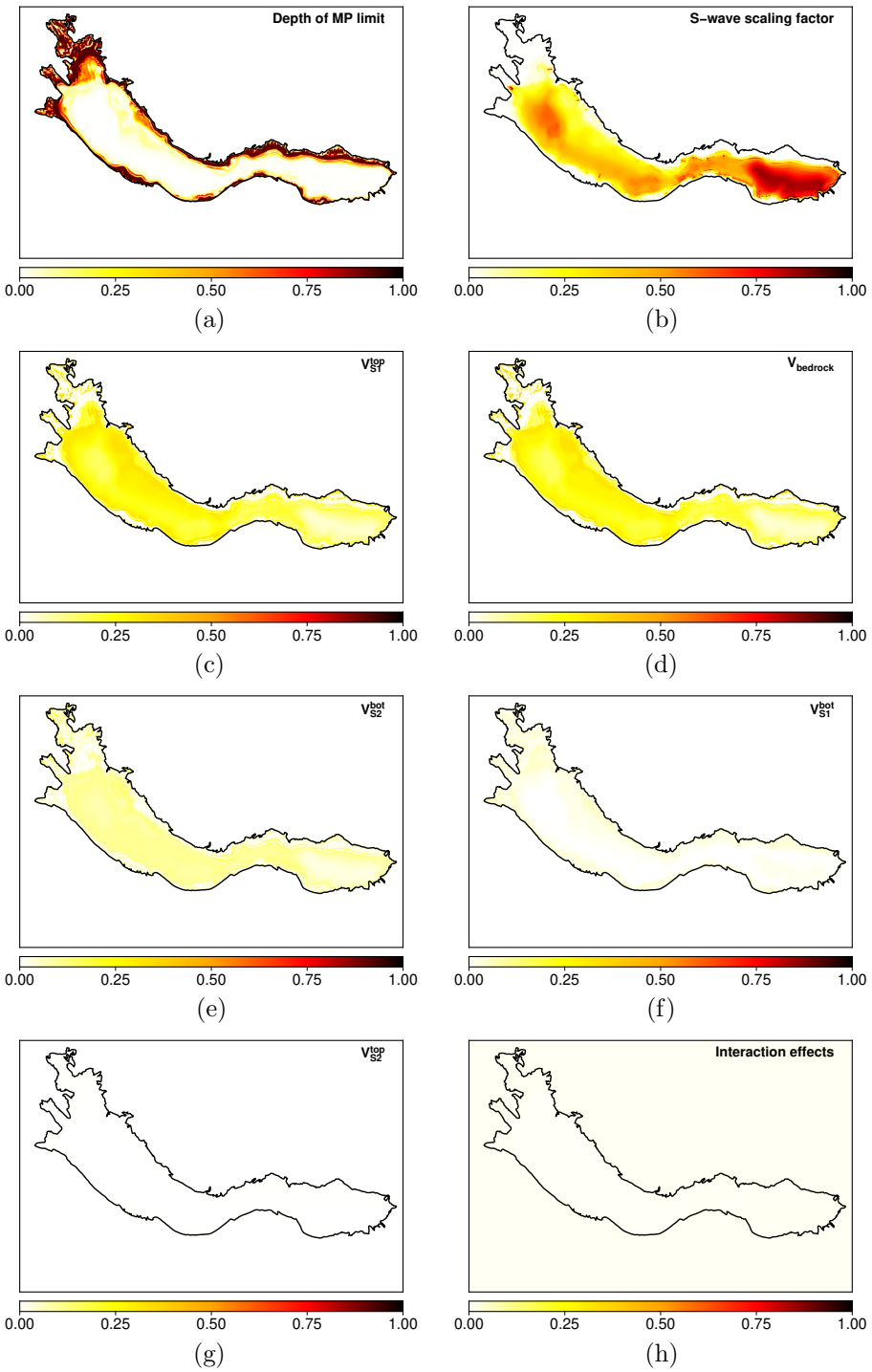

Figure 17: First-order sensitivity indices and total interaction effects of the uncertain PGM. (a) MP interface location. (b) S-wave scaling factor. (c) First layer top velocity. (d) Bedrock velocity. (e) Second layer bottom velocity. (f) First layer bottom velocity. (g) Second layer top velocity. (h) Total interactions

Analysis and extreme computing on multi-Petaflops machine" at the KAUST Extreme Computing Research

Center. The authors are also thankful to E. Chaljub, E. Maufroy, F. Hollender and P.-Y. Bard for providing the velocity model data and for the fruitful discussions about the definition of the Mygdonian basin model.

References

[1] M. V. Manakou, D. G. Raptakis, F. J. Chávez-García, P. I. Apostolidis, K. D. Pitilakis, 3d soil structure of the mygdonian basin for site response analysis, Soil Dyn. Earthq. Eng. 30 (11) (2010) 1198–1211.
[2] C. Guyonnet-Benaize, Coauthors, Imaging 3d geological structure of the mygdonian basin (northern greece) with geological numerical modeling and geophysical methods., in: EGU General Assembly Conference Abstracts, Vol. 15, 2013.
[3] E. Maufroy, E. Chaljub, N. P. Theodoulidis, Z. Roumelioti, F. Hollender, P.-Y. Bard, F. De Martin, C. Guyonnet-Benaize, L. Margerin, Source-related variability of site response in the mygdonian basin (greece) from accelerometric recordings and 3d numerical simulations, Bull. Seism. Soc. Am. 107 (2) (2017) 787–808.
[4] G. D. Manolis, Stochastic soil dynamics, Soil Dyn. Earthq. Eng. 22 (1) (2002) 3 – 15. doi:http://dx.doi.org/10.1016/S0267-7261(01)00055-0.
[5] R. G. Ghanem, S. D. Spanos, Stochastic Finite Elements: a Spectral Approach, Springer Verlag, 1991.
[6] G. Ahmadi, Stochastic earthquake response of structures on sliding foundation, Int. J. Eng. Sci. 21 (2) (1983) 93 – 102. doi:https://doi.org/10.1016/0020-7225(83)90001-0.
[7] C. Desceliers, C. Soize, S. Cambier, Non-parametric–parametric model for random uncertainties in non-linear structural dynamics: application to earthquake engineering, Earthq. Eng. Struct. D. 33 (3) (2004) 315–327. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/eqe.352, doi:10.1002/eqe.352.
[8] B. R. Ellingwood, K. Kinali, Quantifying and communicating uncertainty in seismic risk assessment, Struct. Saf. 31 (2) (2009) 179 – 187. doi:https://doi.org/10.1016/j.strusafe.2008.06.001.
[9] G. Li, M. Iskandarani, M. Le Hénaff, J. Winokur, O. P. Le Maître, O. M. Knio, Quantifying initial and wind forcing uncertainties in the gulf of Mexico, Comput. Geosci. 20 (5) (2016) 1133–1153. doi:10.1007/s10596-016-9581-4.
[10] E. N. Lorenz, Empirical Orthogonal Functions and Statistical Weather Prediction, Scientific report / MIT. Statistical Forecasting Project, Massachusetts Institute of Technology, Department of Meteorology, 1956.
[11] O. P. Le Maître, O. M. Knio, Spectral Methods for Uncertainty Quantification, Scientific Computation, Springer, 2010.
[12] S. S. Chen, D. L. Donoho, M. A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comp. 20 (1998) 33–61.
[13] E. Van Den Berg, M. P. Friedlander, SPGL1: A solver for large-scale sparse reconstruction (June 2007).
[14] E. Van Den Berg, M. P. Friedlander, Probing the pareto frontier for basis pursuit solutions, SIAM J. Sci. Comp. 31 (2) (2008) 890–912. doi:10.1137/080714488.
[15] H.-P. Liu, D. L. Anderson, H. Kanamori, Velocity dispersion due to anelasticity; implications for seismology and mantle composition, Geophys. J. Int. 47 (1) (1976) 41–58.
[16] S. M. Day, J. B. Minster, Numerical simulation of attenuated wavefields using a Padé approximant method, Geophys. J. Int. 78 (1984) 105–118.
[17] H. Emmerich, M. Korn, Incorporation of attenuation into time-domain computations of seismic wave fields, Geophysics 52 (9) (1987) 1252–1264. doi:10.1190/1.1442386.
[18] S. Ma, P. Liu, Modeling of the perfectly matched layer absorbing boundaries and intrinsic attenuation in explicit finite-element methods, Bull. Seism. Soc. Am. 96 (1) (2006) 1779–1794.
[19] P. Liu, R. J. Archuleta, Efficient modeling of Q for 3D numerical simulation of wave propagation, Bull. Seism. Soc. Am. 96 (4) (2006) 1352–1358.
[20] E. Chaljub, Coauthors, 3-D numerical simulations of earthquake ground motion in sedimentary basins: testing accuracy through stringent models, Geophys. J. Int. 201 (1) (2015) 90–111.
[21] EUROSEISTEST web site, http://euroseis.civil.auth.gr.
[22] K. Pitilakis, Z. Roumelioti, D. Raptakis, M. Manakou, K. Liakakis, A. Anastasiadis, D. Pitilakis, The euroseistest strong-motion database and web portal, Seismol. Res. Lett. 84 (5) (2013) 796–804.
[23] EFISPEC3D web site, http://efispec.free.fr.
[24] F. De Martin, Verification of a spectral-element method code for the Southern California Earthquake Center LOH.3 viscoelastic case, Bull. Seism. Soc. Am. 101 (6) (2011) 2855–2865. arXiv:http://www.bssaonline.org/content/101/6/2855.full.pdf+html, doi:10.1785/0120100305.
[25] S. Matsushima, T. Hirokawa, F. De Martin, H. Kawase, F. J. Sánchez-Sesma, The effect of lateral heterogeneity on horizontal-to-vertical spectral ratio of microtremors inferred from observation and synthetics, Bull. Seism. Soc. Am. 104 (1) (2014) 381–393.
[26] A. T. Patera, A spectral element method for fluid dynamics: laminar flow in a channel expansion, J. Comput. Phys. 54 (1984) 468–488.
[27] Y. Maday, A. T. Patera, Spectral element methods for the incompressible navier-stokes equations, State of the art survey in computational mechanics (1989) 71–143.
[28] D. Komatitsch, J. P. Vilotte, The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures, Bull. Seism. Soc. Am. 88 (2) (1998) 368–392.
[29] N. M. Newmark, A method for computation of structural dynamics, in: Proc. Am. Soc. Civ. Eng., 1959, pp. 67–94.
[30] T. D. Blacker, W. J. Bohnhoff, T. L. Edwards, CUBIT mesh generation environment. volume 1: Users manual (1994).
[31] Y. Capdeville, L. Guillot, J.-J. Marigo, 1-d non-periodic homogenization for the seismic wave equation, Geophys. J. Int. 181 (2) (2010) 897–910.

[32] I. M. Sobol, Uniformly distributed sequences with an additional uniform property, USSR Comput. Math. Math. Phys. 16 (1976) 236–242.
[33] C. Dutang, P. Savicky, randtoolbox: Generating and testing random numbers, r package version 1.30.0 (2019).
[34] H. Faure, S. Tezuka, Another random scrambling of digital (t,s)-sequences, in: K.-T. Fang, H. Niederreiter, F. J. Hickernell (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 242–256.
[35] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
[36] K. Karhunen, Über lineare methoden in der wahrscheinlichkeitsrechnung, Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys. 37 (1947) 1–79.
[37] M. Loève, Probability Theory, The university series in higher mathematics, affiliated East-West-Press Pvt. Limited, 1968.
[38] C. Li, A. Der Kiureghian, Optimal discretization of random fields, J. Eng. Mech. 119 (6) (1993) 1136–1154. doi:10.1061/(ASCE)0733-9399(1993)119:6(1136).
[39] J. Zhang, B. Ellingwood, Orthogonal series expansions of random fields in reliability analysis, J. Eng. Mech. 120 (12) (1994) 2660–2677. doi:10.1061/(ASCE)0733-9399(1994)120:12(2660).
[40] L. Sirovich, Turbulence and the Dynamics of Coherent Structures. Part 1 : Coherent Structures, Quarterly of Applied Mathematics 45 (3) (1987) 561–571.
[41] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2005.
[42] D. B. Xiu, G. E. Karniadakis, The Wiener-Askey Polynomial Chaos for stochastic differential equations, SIAM J. Sci. Comp. 24 (2002) 619–644.
[43] M. Iskandarani, S. Wang, A. Srinivasan, W. C. Thacker., J. Winokur, O. Knio, An overview of uncertainty quantification techniques with application to oceanic and oil-spill simulations, J. Geophys. Res.: Oceans 121 (4) (2016) 2789–2808. doi:10.1002/2015JC011366.
[44] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York Inc., 2001.
[45] A. Bowman, A. Azzalini, Applied smoothing techniques for data analysis, New York: Oxford University Press, 1997.
[46] T. Crestaux, O. Le Maître, J.-M. Martinez, Polynomial chaos expansion for sensitivity analysis, Reliab. Eng. and Syst. Saf. 94 (7) (2009) 1161–1172.
[47] P. Sochala, F. De Martin, Surrogate combining harmonic decomposition and polynomial chaos for seismic shear waves in uncertain media, Comput. Geosci.doi:https://doi.org/10.1007/s10596-017-9677-5.
[48] W. Hoeffding, A class of statistics with asymptotically normal distribution, The annals of Mathematical Statistics 19 (1948) 293–325.
[49] I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp. 1 (1993) 407–414.

Appendix A. Sobol–Hoeffding decomposition

Let $\boldsymbol{\xi}$ be a set of d independent random variables with value in $\Xi = \Xi_1 \times \cdots \times \Xi_d$, and let $\mathcal{D} = \{1, 2, \ldots, d\}$. Then $f : \Xi \mapsto \mathbb{R} \in L^2(\Xi, p_{\boldsymbol{\xi}})$ admits the so-called Sobol–Hoeffding (SH) decomposition [48, 49] into 2^d functionals,

$$f(\boldsymbol{\xi}) := \sum_{\boldsymbol{i} \subseteq \mathcal{D}} f^{(\boldsymbol{i})}(\boldsymbol{\xi}_{\boldsymbol{i}}), \tag{A.1}$$

where $\boldsymbol{\xi}_{\boldsymbol{i}}$ denotes the sub-vector of $\boldsymbol{\xi}$ having for components $\xi_{i \in \boldsymbol{i}}$. The SH decomposition is *hierarchical* since it encompasses univariate, bivariate, ... and d-variate functionals as well as a constant term $f^{(\emptyset)}$ (equal to the mean value of f). The decomposition (A.1) is also *orthogonal* in the sense that the functionals $f^{(\boldsymbol{i})}$ verify the following relations,

$$\forall \boldsymbol{i} \subseteq \mathcal{D}, j \in \boldsymbol{i}, \quad \int_{\Xi_j} f^{(\boldsymbol{i})}(\boldsymbol{\xi}_{\boldsymbol{i}}) p(\xi_j) d\xi_j = 0, \tag{A.2}$$

$$\forall \boldsymbol{i}, \boldsymbol{j} \subseteq \mathcal{D}, \boldsymbol{i} \neq \boldsymbol{j}, \quad \int_{\Xi} f^{(\boldsymbol{i})}(\boldsymbol{\xi}_{\boldsymbol{i}}) f^{(\boldsymbol{j})}(\boldsymbol{\xi}_{\boldsymbol{j}}) p(\boldsymbol{\xi}) d\boldsymbol{\xi} = \mathbb{E}(f^{(\boldsymbol{i})} f^{(\boldsymbol{j})}) = 0. \tag{A.3}$$

By applying the variance operator (13) onto (A.1), and by using the orthogonality property (A.3), we obtain the decomposition of the *total variance* of f,

$$\mathbb{V}(f) = \sum_{\boldsymbol{i} \subseteq \mathcal{D}^*} \mathbb{V}(f^{(\boldsymbol{i})}), \tag{A.4}$$

where $\mathcal{D}^* := \mathcal{D} \setminus \emptyset$. In the variance decomposition (A.4), the *partial variance* $\mathbb{V}(f^{(\boldsymbol{i})}) = \mathbb{E}((f^{(\boldsymbol{i})})^2)$ is interpreted as the contribution to the total variance $\mathbb{V}(f)$ of the interaction between parameters $\xi_{i \in \boldsymbol{i}}$. Of particular interest are the so-called first order indices for $\boldsymbol{i} = \{i\}$, which measure the effect of the i-th

uncertain input on the variance. Typically, this effect is normalized by the total variance defining the first-order sensitivity indices by

$$S_i = \frac{\mathbb{V}(f^{(i)})}{\mathbb{V}(f)}.$$

The total sensitivity index, instead, aims at measuring all the effect of ξ_i, including through its interactions with other random inputs. It comes,

$$T_i = \frac{\sum_{\boldsymbol{i} \in \mathcal{D}, i \in \boldsymbol{i}} \mathbb{V}(f^{(\boldsymbol{i})})}{\mathbb{V}(f)}.$$

It follows from these definitions that

$$\sum_{i=1}^{d} S_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{d} T_i \geq 1.$$

The effect of the random variables are additive if the sum of the first-order indices is equal to one; otherwise, a part of the variance is due to interactions between the inputs and these interactions can be measured globally by $1 - \sum_i S_i$.