Synthetic ground motions in heterogeneous geologies from various sources: the HEMEW^S-3D database

Fanny Lehmann[1,2], Filippo Gatti[2], Michaël Bertin[1], and Didier Clouteau[2]

[1]CEA, DAM, DIF, F-91297 Arpajon, France
[2]LMPS - Laboratoire de Mécanique Paris-Saclay, Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, Gif-sur-Yvette, France

Correspondence: Fanny Lehmann (fanny.lehmann@centralesupelec.fr)

Abstract. The ever-improving performances of physics-based simulations and the rapid developments of deep learning are offering new perspectives to study earthquake-induced ground motion. Due to the large amount of data required to train deep neural networks, applications have so far been limited to recorded data or two-dimensional simulations. To bridge the gap between deep learning and high-fidelity numerical simulations, this work introduces a new database of physics-based earthquake simulations.

The ~~HEMEW-3D~~ HEMEW^S-3D database comprises 30 000 simulations of elastic wave propagation in three-dimensional (3D) geological domains. Each domain is parametrized by a different geological model built from a random arrangement of layers augmented by random fields that represent heterogeneities. Elastic waves originate from a randomly located point-wise source parametrized by a random moment tensor. For each simulation, ground motion is synthetized at the surface by a grid of virtual sensors. The high frequency of waveforms (f_{max} = 5 Hz) allows extensive analyses of surface ground motion.

Existing and foreseen applications range from statistic analyses of the ground motion variability and machine learning methods on geological models, to deep learning-based predictions of ground motion depending on 3D heterogeneous geologies and source properties.

1 Introduction

Deep learning has a long tradition in seismology thanks to large networks of sensors recording earthquakes worldwide. Applications are extremely diverse, in terms of methods, data, and scientific goals (see e.g. Mousavi and Beroza (2023) for a review). Detecting earthquakes and discriminating them from other events such as explosions, quarry blasts, or seismic noise are the most common applications of deep learning in seismology (Mousavi and Beroza, 2023). A wide variety of methods are also devoted to characterizing earthquakes from ground motion recordings, for instance to estimate source mechanisms, earthquake location, and magnitude. The rapid improvements of deep learning in the last few years have even enabled its use in operational frameworks, thereby providing real-time predictions of earthquake parameters (Zhu et al., 2022).

However, all those methods rely on databases of seismic waveforms. While there exist several curated databases of recorded ground motion, they are sparse in regions with low-to-moderate seismicity or poor instrumental coverage (Bahrampouri et al., 2021; Michelini et al., 2021; Mousavi et al., 2019). In those cases, numerical simulations are a great opportunity to complement

existing databases. Simulations rely on computational schemes to solve the wave propagation equations from the earthquake source to the Earth surface; and provide synthetic waveforms at any spatial point of the simulation domain. Results of 3D physics-based simulations have been compiled for several past earthquakes in the BB-SPEEDset dataset (Paolucci et al., 2021) but the number of simulations is not appropriate for machine learning approaches.

In fact, physics-based simulations show several limitations. Firstly, they require a ~~detailled~~ detailed description of the ground properties that define the physical behaviour of the waves propagating in the Earth. Especially, ground properties should be given as three-dimensional (3D) geological models since 3D features have crucial effects that are not accounted for in two-dimensional (2D) settings (e.g. sedimentary basins leading to site effects) (Moczo et al., 2018; Smerzini et al., 2011; Zhu et al., 2020). Since extensive geophysical investigations are needed to obtain 3D geological models, they are rare, and when existing, they are still limited by epistemic uncertainties. Therefore, when trying to reproduce an earthquake with physics-based numerical simulations, uncertainties can be represented by random heterogeneities added to the reference model to introduce variability (Chaljub et al., 2021; Lehmann et al., 2022).

Quantifying the effects of 3D geological features is made more difficult by the second limitation of physics-based simulations, which is their high computational cost, especially when dealing with high frequencies and large spatial domains. Despite relying on high-performance computing (HPC) frameworks, seismic waves propagation simulations can reach tens to hundreds of thousands of equivalent CPU hours (Computational Processing Units, ~~Heinecke et al. (2014); Touhami et al. (2022)~~ Fu et al. (2017); Heinecke et al. (2014); Poursartip et al. (2020)). Since computational costs prevent statistical studies on synthetic waveforms due to a limited number of simulations per geological model, deep learning represents a promising alternative to obtain waveforms.

When predicting the surface ground motion generated by an earthquake, it is important to obtain time-series that describe the temporal evolution of shaking, and not only scalar features (such as Peak Ground Acceleration, Cumulative Absolute Velocity) that give useful but limited information. Physics-Informed Neural Networks (PINNs, Raissi et al. (2019)) successfully solved the wave equation ~~(Ding et al., 2023; Karimpouli and Tahmasebi, 2020; Moseley et al., 2020; Rasht-Behesht et al., 2022; ?; Song et al., 20~~ (Ding et al., 2023; Karimpouli and Tahmasebi, 2020; Moseley et al., 2020; Rasht-Behesht et al., 2022; Ren et al., 2024; Song et al., 2023; . However, applications are mainly limited to 2D domains and models cannot extrapolate to another geological configuration than the one used in the training phase. Alternatively, generative methods have been used to enhance existing numerical simulations, by increasing their spatial resolution ~~(e.g. ?)~~ (e.g. Gadylshin et al., 2021) or their frequency content (e.g. Gatti and Clouteau, 2020).

The recent emergence of Scientific Machine Learning (SciML) is offering a new paradigm to predict physics-based ground motion parametrized by 3D ground properties and source parameters, with intrinsic generalization ability to various resolutions and geological configurations. SciML has led to significant scientific developments in communities with large, reliable, and freely available databases. For instance, in numerical weather prediction, Bonev et al. (2023) and Pathak et al. (2022) took advantage of the ERA5 dataset provided by the European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020). In seismology, Mousavi and Beroza (2023) pointed out that "the limitations on training data and generalization are the main challenges in solving inverse and forward problems using supervised [Deep Neural Networks]."

In this work, we describe the first open database of seismic simulations associated with 3D heterogeneous geological models. The ~~HEMEW-3D~~ HEMEW^S-3D (HEterogeneous Materials and Elastic Waves with Source variability in 3D) database contains 30 000 high-fidelity simulations in 3D domains of size $9.6\,\mathrm{km} \times 9.6\,\mathrm{km} \times 9.6\,\mathrm{km}$~~Lehmann (2023)~~. This represents a ~~rather~~ challenging computational task ~~(1.6 million CPU hours in total)~~accounting for 9×10^5 CPU hours and $4.4\,\mathrm{MWh}$ in total. Ground motion was synthetized at the surface of the simulation domain for ~~20 s~~ $8\,\mathrm{s}$ on a grid of ~~16 × 16~~ 32×32 virtual sensors. ~~This database was used to produce the first SciML model predicting 3D ground motion (Lehmann et al., 2023a) and this study exemplifies numerous other applications~~Data are available at https://doi.org/10.57745/LAI6YU.

In the following, Section 2 provides an overview of existing datasets in related fields, Section 3 describes the geological models, sources, and surface wavefields in the database, Section 4 ~~illustrates~~ analyses physical characteristics, Section 5 illustrates applications, and Section 6 discusses limitations and ~~Section ?? discusses applications and~~ perspectives.

2 Related Work

Datasets of recorded ground motion ~~are incredibly important for~~ have enabled major deep learning applications in seismology but ~~their limitations have already been mentioned~~they have several limitations in data scarse regions. In this section, we focus on datasets with 2D or 3D data used in geophysics and seismology ~~.~~ with SciML applications. Due to the mathematical similarities between wave propagation and fluid flow (both are governed by hyperbolic equations), related studies are reviewed beyond the field of seismology. This highlights the challenges of high-fidelity numerical simulations for deep learning applications.

2.1 3D datasets

Due to the high computational costs of solving 3D Partial Differential Equations (PDEs), only very few 3D datasets are available. CO2 underground storage has been explored with SciML based on 3D numerical simulations (Grady et al., 2023; Wen et al., 2023; Witte et al., 2023). To support the study of Witte et al. (2023), Annon (2022) provided 4,000 simulation results for 3D CO2 flow through geological models based on the Sleipner dataset complemented by random fields (Equinor, 2020). The Kimberlina dataset also contains 6,000 CO2 leakage rates simulations (Mansoor et al., 2020). However, the geological models in both databases are all variants of the geological model carefully estimated for a given region, thereby limiting the reproducibility in other areas.

2.2 Geophysical datasets

A few datasets of realistic geological units have been developed, such as the Noddyverse dataset of 3D geological models (Jessell et al., 2022). ~~This datasetdoes not provide any ground motion. For~~ In this dataset, geological models result from the deformation of horizontal layers by successive geological events (folds, faults, unconformities, dykes, plugs, shear zones, and tilts) but no associated ground motion is provided. Along the same line of geological deformation, the OpenFWI database combines geological models with associated waveforms, and targets 2D ~~geophysical inversion, databases combining velocity models and associated waveforms have been inspired by the deformations of geological layers (Deng et al., 2022; Liu et al., 2021)~~

Table 1. Summary of datasets providing geological models and seismic wavefields. ~~Dimension of geological models~~Domain: size of the physical domain with the number of grid points given in parenthesis, in (width, depth) for 2D datasets, in (width, length, depth) for 3D datasets. ~~Domain: size of the physical domain.~~ Dimension of seismic wavefields: (receivers along width, time steps) for 2D datasets, (receivers along width, receivers along length, time steps) for 3D datasets.~~Components: number of velocity components for each sensor (1 means that the acoustic wave equation is solved, 2 or 3 means that the elastic wave equation is solved)~~

Dataset	
Noddyverse (Jessell et al., 2022)	
OpenFWI (Deng et al., 2022)	
OpenFWI ~~(Deng et al., 2022)~~ Kimberlina CO2 (Deng et al., 2022)	
OpenFWI ~~(Deng et al., 2022)~~ Kimberlina 3D (Deng et al., 2022)	
$\mathbb{E}^{\mathbf{FWI}}$ ~~?~~ (Feng et al., 2023)	
~~Ovadia et al. (2023)*~~ **HEMEW^{S}-3D (this study)**	~~16k/4k ? $[0, \pi]^2$ 1300; 1600 m/s sinusoidal fluctuations ? 1 2 sources with random location and amplitude Yang et al. (2023)* 18k/2k 64 × 64 1~~

~~. The OpenFWI database (Deng et al., 2022) has been recently used with Neural Operators (Li et al., 2022). The above-mentioned wavefieldswere simulated with the 2D~~ geophysical inversion as the main application (Deng et al., 2022). OpenFWI contains geological models made of horizontal and non-horizontal layers with various folds. It also includes real geological models from field survey areas and models of CO2 geological storage. To generate the wavefields, the acoustic wave equation ~~(i.e. with 1-component waveforms), that saw several Neural Operators applications (Li et al., 2022; Ovadia et al., 2023; Yang et al., 2021) .~~ is solved in the 2D domains. Waves originate from a line of sources at the surface and wavefields are acquired on a line of receivers at depth. The $\mathbb{E}^{FWI}$ database is an extension of OpenFWI to the elastic wave equation, providing two-component ground motion time series (Feng et al., 2023).

Several other studies have computed simulation outputs for the acoustic or elastic wave equation but data are not public (e.g. Liu et al., 2021; Ovadia et al., 2023; Zhang et al., 2023). Table 1 summarizes the characteristics of ~~those datasets .~~ the public datasets and shows that no database provides solutions of the elastic wave equation in 3D domains. Our HEMEW^{S}-3D database intends to fill this gap.

3 Dataset creation

3.1 The elastic wave equation

Elastodynamics describes reversible wave propagation phenomena in solid and fluid domains. In solid mechanics, the solution is represented by a displacement field $\boldsymbol{u} \in \mathbb{R}^3$ propagating in a 3D Euclidean space. We consider a truncated propagation domain $\Omega = [0; L]^3$ with absorbing boundary conditions all around, except the traction-free top surface; and a solution $\boldsymbol{u} : \Omega \times [0, T] \rightarrow \mathbb{R}^3$. The domain length is fixed to L = 9600 m and the total time is T = ~~20 s~~8 s. In its most general form, the elastic wave equation writes

$$\rho \frac{\partial^2 \boldsymbol{u}}{\partial t^2} = \nabla \lambda (\nabla \cdot \boldsymbol{u}) + \nabla \mu \left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T \right] + (\lambda + 2\mu) \nabla (\nabla \cdot \boldsymbol{u}) - \mu \nabla \times \nabla \times \boldsymbol{u} + \boldsymbol{f} \tag{1}$$

where $\rho : \Omega \rightarrow \mathbb{R}$ is the material unit mass density, $\lambda : \Omega \rightarrow \mathbb{R}$, $\mu : \Omega \rightarrow \mathbb{R}$ are the Lamé parameters, characterizing the thermodynamically reversible mechanical behaviour of the material, and $\boldsymbol{f}$ is the body force distribution. In geomechanics, properties ρ, λ, and μ are rarely independently characterized due to a lack of measurements. Therefore, it is legitimate to assume that there is a single informative variable from which all parameters can be deduced. In this work, the velocity of shear waves V_S is the informative variable. Equation 1 can then be rewritten under the general form

$$\mathcal{L}(V_S, \boldsymbol{u}) = \boldsymbol{f} \tag{2}$$

3.2 Earthquake source

In our database, the forcing term $\boldsymbol{f}(\boldsymbol{x}, t) = \operatorname{div} \boldsymbol{m}(\boldsymbol{x}) \cdot s(t)$ is the divergence of a moment tensor density $\boldsymbol{m}$, localized at a ~~fixed~~ point-wise location~~for all simulations~~. $\boldsymbol{m}$ encodes the source radiation patterns as a double couple representing a point-wise kinematic discontinuity in the media. ~~The~~

The source position is represented by the coordinates $(x_s, y_s, z_s) \in \Omega$, not too close from the boundaries to avoid numerical issues due to absorbing boundary conditions. The position is chosen from a Latin Hypercube Sampling (LHS) with

$$x_s \in [1.2; 8.4\,km]$$
$$y_s \in [1.2; 8.4\,km]$$
$$z_s \in [-9.0; -0.6\,km] \quad (3)$$

In addition to the source ~~parameters correspond to the estimation of the Le Teil earthquake (moment magnitude M_w 4.9, France, 2019). The radiation pattern of the double-couple source is described by three angles: strike = 48°, dip = 45°, and rake = 88° (Delouis et al., 2021).~~ position, the source is parametrized by the symmetric 3×3 moment tensor. An equivalent formulation is obtained with the three angles (strike, dip, rake) (Aki and Richards, 1980). With this representation, the angles are sampled from a LHS with a strike between 0° and 360°, dip between 0° and 90°, and rake between 0° and 360°.

The source amplitude corresponds to ~~the real~~ a seismic moment $M_0 = 2.47 \times 10^{16}\,\mathrm{N\,m}$. ~~The~~, and the source time evolution is ~~given by $t \mapsto 1 - \left(1 + \frac{t}{\tau}\right) e^{-\frac{t}{\tau}}$~~ a spice bench given by $s(t) = 1 - \left(1 + \frac{t}{\tau}\right) e^{-\frac{t}{\tau}}$ with $\tau = 0.1$ s ~~.~~ (Fig. A1). Due to the linearity of the elastic wave equation (Eq. 1), it is important to notice that the choice of the source time function in the HEMEW^S-3D database does not constrain the variability of resulting ground motions.

First, any magnitude can be obtained by applying a scalar factor to the ground motion wavefields. Second, the response to any source time function can be computed from the Green function $\boldsymbol{G}(\boldsymbol{x}, t)$, which is the fundamental solution of the elastic wave equation 1 when the source is an impulse point force located at $\boldsymbol{x}_s$ and occurring at $t = t_0$. For the reference source time function $s(t)$, the solution $\boldsymbol{u}(\boldsymbol{x}, t)$ provided in HEMEW^S-3D can be written as the convolution of the Green function with the source time function, $\boldsymbol{u}(\boldsymbol{x}, t) = \boldsymbol{G}(\boldsymbol{x}, t) * s(t)$

Computing the solution u_1 for a new source time function s_1 (provided that the moment tensor density $\boldsymbol{m}$ and the geological parameters a remain the same), is straightforward following the steps below:

1. compute the Fourier transform of the reference source time function $\hat{s} := \mathcal{F}(s)$ and the solution $\hat{u} := \mathcal{F}(u)$
2. derive the Green function in the frequency domain $\hat{G} = \dfrac{\hat{u}}{\hat{s}}$
3. compute the Fourier transform of the new source time function $\hat{s}_1$
4. compute the new solution in the frequency domain $\hat{u}_1 = \hat{G} * \hat{s}_1$
5. deduce the new solution in the temporal domain $u_1 = \mathcal{F}^{-1}(\hat{u}_1)$

From these remarks, one should remember that ground motion wavefields in the HEMEW^S-3D database originate from point-wise sources with different locations $\boldsymbol{x}_s \in \mathbb{R}^3$ and orientations $\boldsymbol{\theta}_s \in \mathbb{R}^3$ but the same source time function.

3.3 Heterogeneous geological models

The ~~HEMEW-3D database contains pairs $\{V_{S,i}, \dot{\boldsymbol{u}}_i\}_i$~~ HEMEW^{S}-3D database contains samples $\{V_S^{(i)}, \boldsymbol{x}_s^{(i)}, \boldsymbol{\theta}_s^{(i)}, \dot{\boldsymbol{u}}^{(i)}\}_i$ that satisfy equation 2 ($\dot{\boldsymbol{u}}$ denotes the velocity field obtained as the time derivative of the displacement field $\boldsymbol{u}$). The 3D geological models $V_S(\boldsymbol{x})$ are non-stationary random fields defined as a mean stair function (horizontal homogeneous layers) to which fluctuations are added, as illustrated in Figure 1.

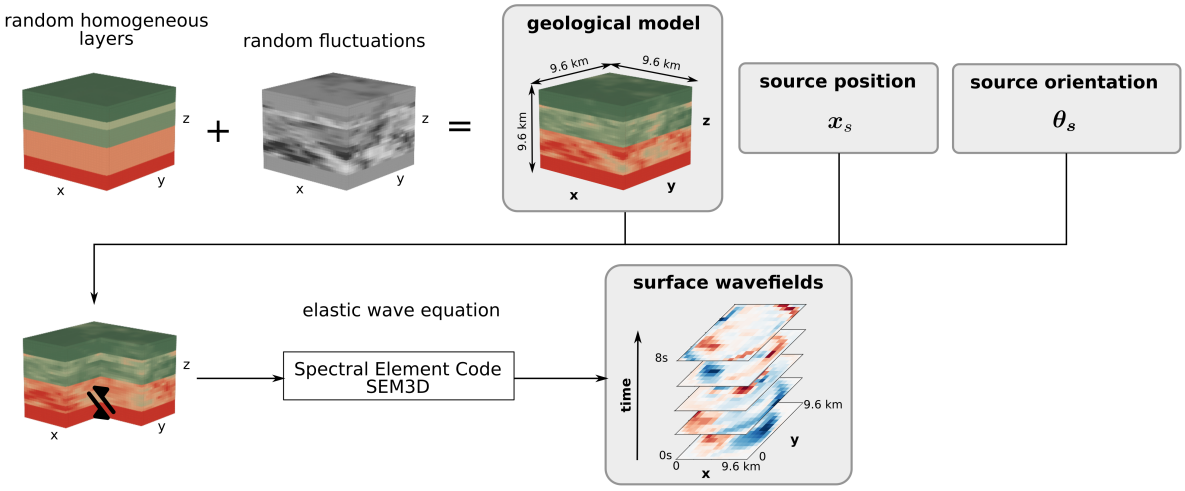


Figure 1. Geological models are built by adding heterogeneities to randomly chosen horizontal layers. Then, elastic waves are propagated from ~~the~~ a source with a random position and random orientation to the surface, where velocity wavefields are synthetized.

3.3.1 Homogeneous models

A $1.8\,\mathrm{km}$-thick homogeneous layer is imposed at the bottom of each geological model, with a V_S value of $V_{S,max} = 4500\,\mathrm{m/s}$. The minimum S-wave velocity is $V_{S,min} = 1071\,\mathrm{m/s}$. Above the bottom layer, the number of horizontal layers and their thickness are randomly chosen for each sample ~~$V_{S,i}$~~ $V_S^{(i)}$, with the sole constraint to fill the total depth with 2 to 7 layers. Then, the mean layer-wise value is drawn from the uniform distribution $\mathcal{U}([\mu_1;\mu_2])$. All layer-wise values are chosen independently. Values of $\mu_1 = V_{S,min}/0.6 = 1785\,\mathrm{m/s}$ and $\mu_2 = V_{S,max}/1.4 = 3214\,\mathrm{m/s}$ were determined to ensure that most values remain bounded within $[V_{S,min};V_{S,max}]$ after the addition of random fields in each layer (see Table 2 for a summary of the parameters).

To recover the other geological properties, the ratio of P- to S-wave velocity was fixed to $V_P/V_S = 1.7$. The density ρ is computed as a function of the P-wave velocity (Molinari and Morelli, 2011)

$$\rho = 1.6612V_P - 0.4721V_P^2 + 0.0671V_P^3 - 0.0043V_P^4 + 0.000106V_P^5 \tag{4}$$

Parameter	Statistical distribution
Number of heterogeneous layers N_ℓ	$\mathcal{U}(\{1,2,3,4,5,6\})$
Layers' thickness $h_1, \cdots, h_{N_\ell}$	$\mathcal{U}(\{(h_1, \cdots, h_{N_\ell}) > 0 \vert h_1 + \cdots + h_{N_\ell} = 7.8\})$
Mean V_S value per layer	$\mathcal{U}([1785, 3214])$
Layer-wise coefficient of variation	$\vert\mathcal{N}(0.2, 0.1)\vert$
Layer-wise correlation length along x	$\mathcal{U}(\{1.5, 3, 4.5, 6\ \mathrm{km}\})$
Layer-wise correlation length along y	$\mathcal{U}(\{1.5, 3, 4.5, 6\ \mathrm{km}\})$
Layer-wise correlation length along z	$\mathcal{U}(\{1.5, 3, 4.5, 6\ \mathrm{km}\})$

Table 2. Statistical distribution of each parameter describing the geological models. Mean V_S values, coefficients of variation, and correlation lengths are chosen independently in each layer. Since the bottom layer has a constant thickness of $1.8\,\mathrm{km}$, it is not included in these parameters.

Attenuation factors for P-waves (Q_P) and S-waves (Q_S) are computed as

$$Q_P = \max\left(\frac{V_P}{20}, \frac{V_S}{5}\right); Q_S = \frac{V_S}{10} \tag{5}$$

3.3.2 Addition of heterogeneities

The layers' thicknesses and mean values describe the general structure of the propagation domain and they correspond to the prior physical information usually available. However, geomaterials of the Earth's crust contain much variability, especially along the horizontal directions. This heterogeneity can be represented by random fields, characterized by their correlation length and coefficient of variation. Following previous studies on geological heterogeneity (~~see Khazaie et al. (2016); Scalise et al. (2021) among others~~e.g. Hartzell et al. (2010); Imperatori and Mai (2013); Khazaie et al. (2016); Scalise et al. (2021); Thompson et al. (2007)), we drew random fields with a von Karman correlation kernel and a Hurst exponent of 0.1 (Chernov, 1960) (marginal distributions are log-normal to preserve positive values).

In order to provide a sufficient dataset variability, the choice of correlation lengths and coefficients of variation is tricky yet crucial (Colvez, 2021). The correlation length gives an idea of the distance above which two points x_A and x_B have independent geological properties ~~$V_{s,i}(x_A)$ and $V_{S,i}(x_B)$~~$V_S^{(i)}(x_A)$ and $V_S^{(i)}(x_B)$. We chose correlation lengths randomly in $\{1.5, 3, 4.5, 6\}$ km, to mix samples with small- and large-scale heterogeneity. In addition, large coefficients of variation were chosen to provide high geological contrasts, following the ~~normal distribution $\mathcal{N}(0.2, 0.1)$~~folded normal distribution $\vert\mathcal{N}(0.2, 0.1)\vert$ with mean 0.2 and coefficient of variation 0.1. Coefficients of variation around $20\,\%$ are common at the surface (Arroucau, 2020), while it is known that values up to $40\,\%$ can be found locally (El Haber et al., 2021).

The 3D random fields computation is made highly efficient by the use of the spectral representation (Shinozuka and Deodatis, 1991; de Carvalho Paludo et al., 2019). With this formulation, a centered Gaussian random field V_S determined by its auto-covariance function $\mathcal{R}$ can be decomposed as a sum of independent identically distributed random variables $(V_{S,n})_{-N \leq n \leq N}$,

with uniform distribution over $[0, 2\pi]$

$$V_S(x) = \sum_{n=-N}^{N} \sqrt{2\hat{\mathcal{R}}(n\Delta k)} \cos(n\Delta k \cdot x + V_{S,n})$$

where $\hat{\mathcal{R}}$ is the Fourier transform of the autocovariance function $\mathcal{R}$ and Δk is the unit volume in $\mathbb{R}^3$.

Finally, V_S values are clipped between $V_{S,min} = 1071\,\mathrm{m/s}$ and $V_{S,max} = 4500\,\mathrm{m/s}$. These bounds correspond to the velocity of shear-waves in hard sediments and at the bottom of the continental crust (Molinari and Morelli, 2011).

It should be noted that all layers have distinct coefficients of variation and correlation lengths, meaning that different random fields are drawn inside each layer. Also, random fields are drawn only once for each set of parameters.

3.3.3 Representation in the database

Geological realizations ~~$V_{S,i}$~~ $V_S^{(i)}$ are discretized over a grid of $32 \times 32 \times 32$ elements (corresponding to x, y, z axes) and are provided as `.npy` arrays. The total size of the geological dataset is 3.9 GB, split in 15 files of 2000 geological models for easier data management. Additionally, metadata give ~~the minimum, mean~~ parameters of each layer: the mean V_S value, the thickness, the coefficient of variation, and the correlation lengths along x, ~~maximum, and standard deviation of the pixel-wise geological values~~y, and z.

3.4 Solutions of the wave equation

The elastic wave equation was solved in each domain i by means of the open source code SEM3D[1] (Touhami et al., 2022) based on the Spectral Element Method (Faccioli et al., 1997; Komatitsch and Tromp, 1999). The dimension of the simulation mesh is prescribed by the maximum frequency f_{max} one aims at exactly resolving. In this study, f_{max} was fixed at 5 Hz, which is relatively high for this type of simulations. Many simulations have been conducted so far with an accuracy up to 1 or 2 Hz (Rekoske et al., 2023; Rosti et al., 2023), while high-fidelity simulations for local realistic earthquake scenarios extend up to 10 Hz (Castro-Cruz et al., 2021; De Martin et al., 2021; Heinecke et al., 2014) (and exceptionally up to 18 Hz, such as in Fu et al. (2017)). Then, the smallest wavelength $\lambda_{min} = V_{S,min}/f_{max}$ must be described on the mesh by at least 5 quadrature points. With 7 Gauss-Lobatto-Legendre quadrature points per mesh element, this leads to elements of size $h = \frac{7}{5} \cdot \frac{V_{S,min}}{f_{max}} = 300\,\mathrm{m}$. This explains that 32 elements in each direction amount to a domain size of $L = 9600\,\mathrm{m}$.The time-marching scheme is a leap-frog second-order accurate explicit scheme, solved for velocity fields.

To maintain reasonable computational loads and reflect real-life situations, velocity fields were recorded only at the surface of the propagation domain. A regular grid of ~~16×16~~ 32×32 sensors was placed between $150\,\mathrm{m}$ and $9450\,\mathrm{m}$ in both horizontal directions (implying a distance of 300 m between two neighbouring sensors). At each monitoring point, the three-component velocity field is synthetized with a 100 Hz sampling frequency between 0 s and 8 s. Figure 2 illustrates velocity waveforms at ~~five~~ eight virtual sensors.

[1] https://github.com/sem3d/SEM

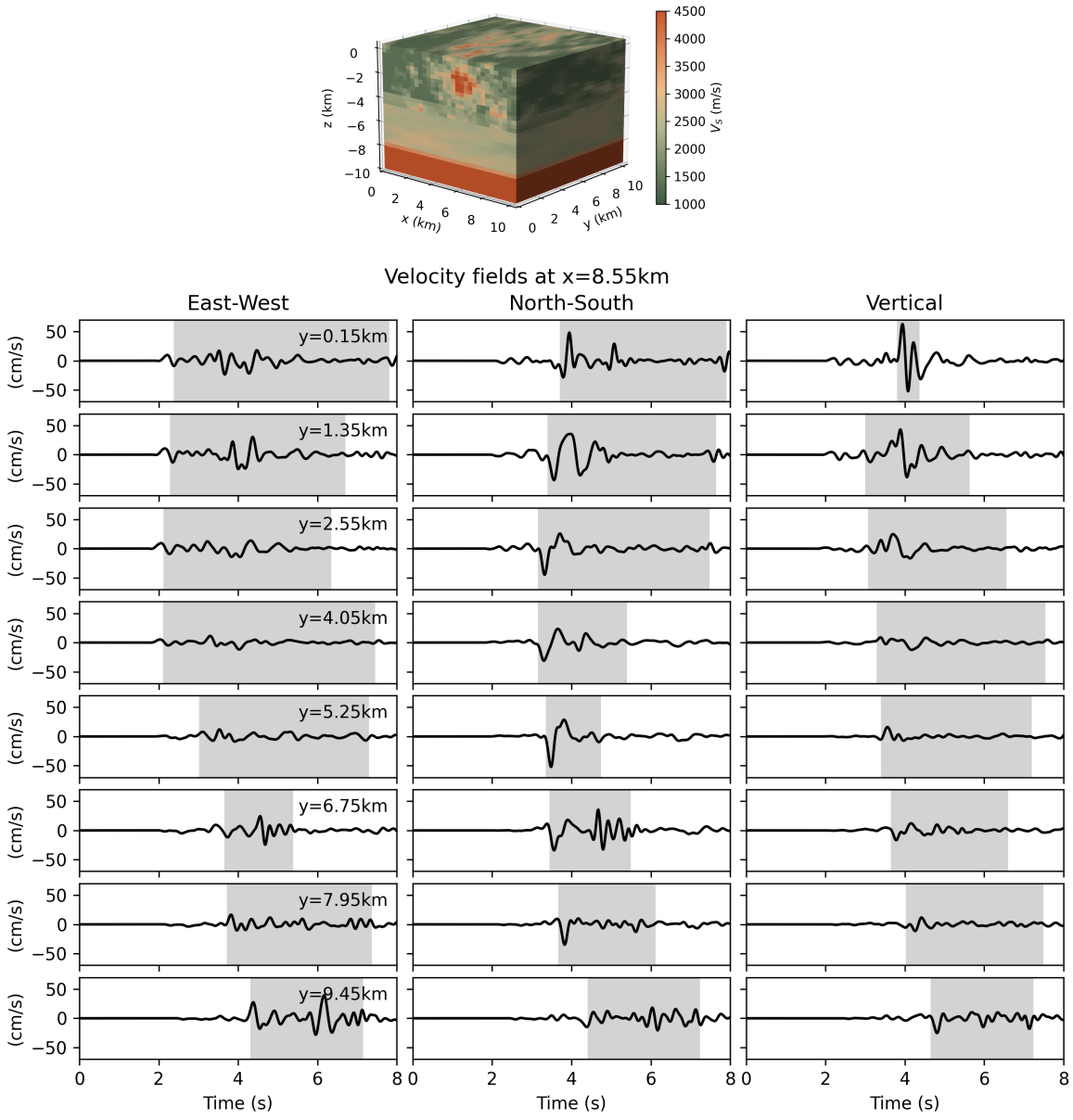


Figure 2. 3-component velocity waveforms ~~syntehtized~~ synthetized at ~~5~~ eight virtual sensors ~~with x-coordinate between 1.39 km and 3.87 km,~~ on a line parallel to the y axis at ~~y-coordinate equal 7.59 km~~x=8.85 km. The shaded area extends from 5 % to 95 % of the Arias intensity, hence its length equals the Relative Significant Duration (RSD). The corresponding geological model is shown at the top. The source is located at (2.04 km, 3.64 km, −2.17 km).

3.4.1 Representation in the database

Velocity fields are provided as `feather` ~~dataframes, easily readable with the common~~ `pandas` ~~library in~~ `Python`~~. Each dataframe has $16 \times 16 \times 3$ rows~~ `.h5` files. Each file contains three keys: `uE`, `uN`, `uZ` corresponding to the ~~16×16 sensors and the 3 velocity components . The 2005 columns contain the row's attributes (index i of the corresponding geological model $V_{S,i}$, sensor's coordinates, velocity component)and time steps $0, 0.01, 0.02, \cdots, 19.99$. The velocity wavefields database amounts to 369.9 GB, split in 300 files of 100 simulations each. In addition, metadata associated with each sample give the first wave arrival time at the surface, and the maximum amplitude over time for different sensors~~three components of ground motion (East-West, North-South, Vertical). Each velocity field is of shape $32 \times 32 \times 800$ where the first index corresponds to the y axis, the second index to the x axis, and the third index to the temporal axis.

Files are gathered in `.zip` archives containing 100 simulation results. The 300 .zip files amount to 263.4GB. They are downloadable individually (0.87GB per file).

4 Dataset analysis

4.1 Descriptive statistics of the temporal evolution

Since most of the geological parameters are chosen uniformly randomly (Table 2), the geological dataset is well-balanced: geological models with 1 to 6 layers are equipartitioned and all random fields parameters have approximately the same frequency. Mean ~~V_S~~ V_S values range from $1756\,\mathrm{m/s}$ to $3145\,\mathrm{m/s}$~~(Figure ??)~~.

The first wave arrival time is a crucial parameter for earthquake early warning and seismic phase picking is a common task with deep learning models. Arrival time depends on the distance between the earthquake source and the monitoring sensor, as well as the geological properties on the propagation path. Wave arrival times are usually determined from recordings, either manually by experts, or with machine learning methods. However, it is possible to compute ~~approximated~~ almost exact arrival times from synthetic velocity fields since ground motion is almost zero before the first wave arrival. Therefore, we obtained the wave arrival times for the P-waves as the first time where the ~~velocity amplitude exceeds 10^{-4}m/s. The~~ amplitude exceeds $0.1\,\%$ of the maximum amplitude. Due to the variability in the source depth and the different wave velocities in the geological models, first wave arrival times ~~range from $1.08\,\mathrm{s}$ to $2.66\,\mathrm{s}$. Since waves propagate faster in geological domains with large V_S, the~~ vary significantly between samples and between sensors. Figure 3a shows that $10\,\%$ of velocity time series are initiated before $0.65\,\mathrm{s}$ while $10\,\%$ of time series are still null after $2.17\,\mathrm{s}$.

As expected, the P-wave arrival time is strongly correlated with the hypocentral distance (Fig. 4a) since shorter hypocentral distances are associated with shorter propagation paths. The mean velocity on the propagation path also influences the first wave arrival time ~~is~~ but variability is higher. Figure 4b indeed shows that the P-wave arrival time is negatively correlated with the mean ~~V_S value (Figure ??).~~ S-wave velocity in the whole domain. It confirms that waves propagate slower when the mean velocity is lower. Computing the mean velocity of the domain is only an approximation of the velocity seen by the waves on the different propagation paths. In particular, the mean velocity does not depend on the sensor position with this approximation.

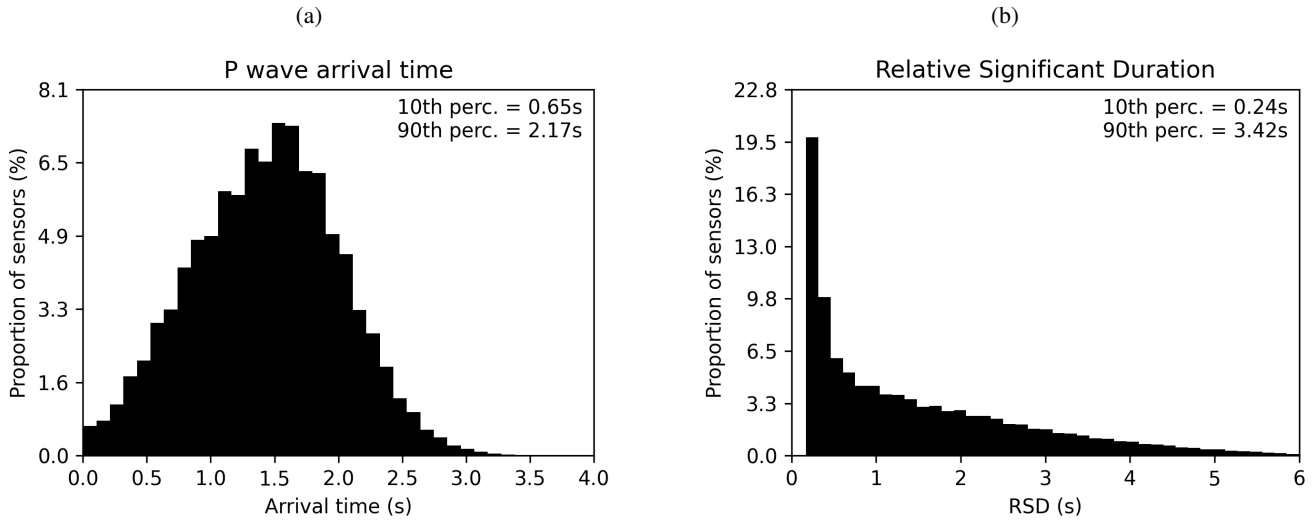


Figure 3. Distributions of the temporal features of velocity time series at each monitoring sensor and for 30 000 samples. (a) the first P-wave arrival time is computed on the vertical component (b) the Relative Significant Duration (RSD) is shown for the East-West component, results are very similar for the two other components

~~Seismic velocity fields can~~ The temporal evolution of ground motion can also be characterized by ~~several intensity measures.~~ its Relative Significant Duration (RSD). It corresponds to the duration of the signal between 5% and 95% of the Arias intensity (I_A) (Arias, 1970)

$$I_A = \frac{\pi}{2g} \int_0^T a^2(t) dt \tag{6}$$

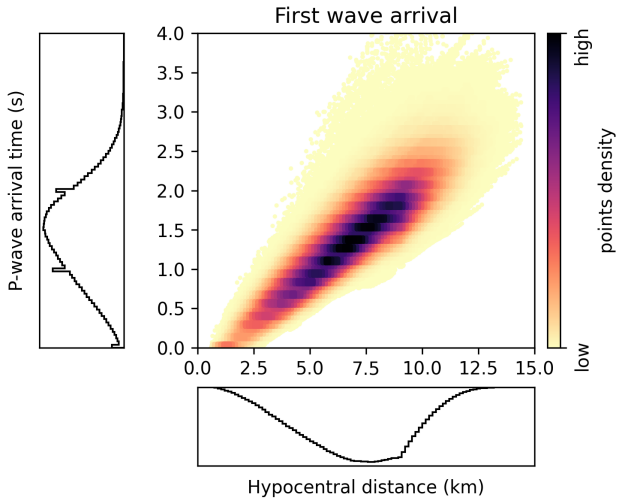
where $a(t)$ is the acceleration field and T is the total duration of the signal. Figure 3b shows that RSD covers a large variation range, from 0.17 s to 7.60 s. This variability is illustrated in Fig. 2 where the grey areas represent the RSD. One can especially notice that samples with a strong pulse have a small RSD. Indeed, most of the energy is concentrated around the pulse.

Quantitatively, the HEMEW^S-3D database contains short ground motions since 10 % of time series have a RSD lower than 0.24 s, as well as longer ground motions where 10 % of time series have a RSD longer than 3.42 s. The median RSD is 1.06 s.

4.2 Descriptive statistics related to energy content

The Peak Ground Velocity (PGV) is computed as the maximum absolute value over all timesteps separately on each component. The PGV is slightly lower on the vertical component than the two horizontal components (Fig. 5). It is very similar between the East-West and North-South components, which is expected since the HEMEW^S-3D database is statistically invariant per

(a) ~~The first wave arrival time at the surface (y-axis) is shown against the mean V_S value (x-axis).~~



(b) ~~The mean PGV over all sensors of a velocity field against the standard deviation of the geological model.~~

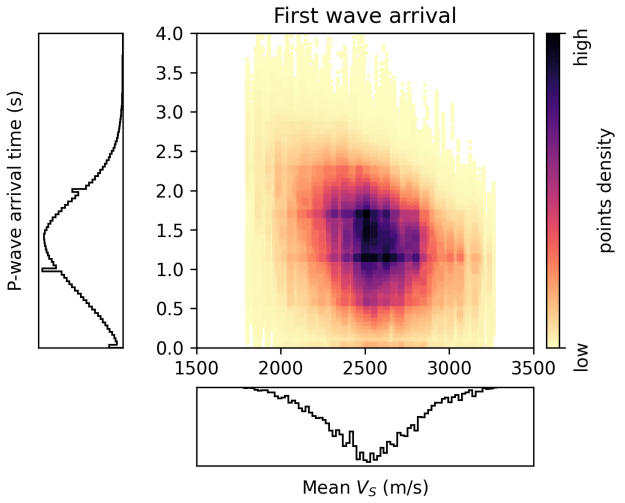


Figure 4. ~~Relationships between velocity fields features~~ For each sample and each sensor, the P-wave arrival time is shown against (~~PGV=Peak Ground Velocity~~a) ~~and geological properties. Each dot corresponds to one sample. Distributions are given~~ the hypocentral distance, (b) the mean S-wave velocity in ~~percentage of~~ the ~~total number of samples~~3D domain.

horizontal rotation. Figure 5 shows that the PGV extends over three orders of magnitude, with the first percentile being equal to $0.89\,\mathrm{cm/s}$ while the 99th percentile equals $129.3\,\mathrm{cm/s}$. The median PGV is $8.9\,\mathrm{cm/s}$. ~~However, it should be noted that, although this is common practice in numerical studies, this does not exactly corresponds to the definition of PGV for recordings.In fact, since synthetic velocity fields contain lower frequencies than recordings, their PGV is also lower than the corresponding recorded PGV .~~

~~Distribution of Peak Ground Velocity (PGV, left) and Relative Significant Duration (RSD, right) for each velocity field and each sensor. The PGV is computed on the arithmetic mean of the three components.~~

~~The sensors' PGV range from $0.0027\,\mathrm{m/s}$ to $1.24\,\mathrm{m/s}$ with a median value of $0.066\,\mathrm{m/s}$ (Figure 5).~~ The 1st- to 99th-percentile interval is in line with ground motion observed within a few kilometers of ~~a~~ moderate-magnitude earthquakes (e.g. Convertito et al. (2022)). ~~When observing the mean PGV over all sensors, Figure **??** shows that geological models with large standard variations tend to produce smaller PGVs since heterogeneitiesinduce~~

When the propagation path is longer, seismic waves encounter more geological heterogeneities. They create a dispersion and diffraction of ~~seismic~~ waves that spread the ~~wavefront along time.~~ energy signal over time. Larger hypocentral distances are associated with longer propagation paths. Figure 6a then shows that the PGV is negatively correlated with the hypocentral distance.

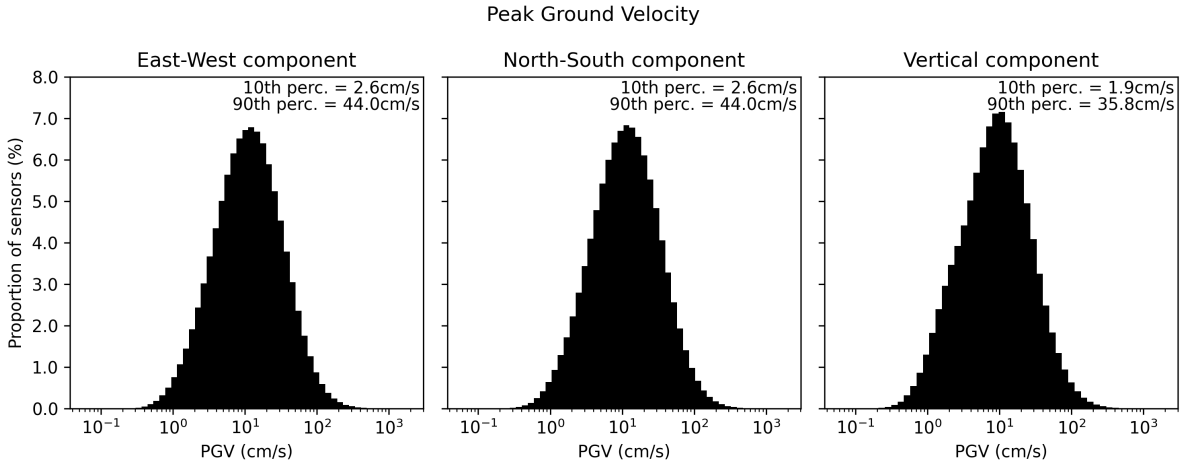


Figure 5. The Peak Ground Velocity (PGV) is computed as the the maximum absolute value over all timesteps separately on each component. There is one value for each of the 32×32 sensors and each of the 30 000 samples.

~~The Relative Significant Duration (RSD) indicates the duration of ground motion. It corresponds to the duration of the signal between 5% and 95% of the Arias intensity (I_A) (Arias, 1970)~~

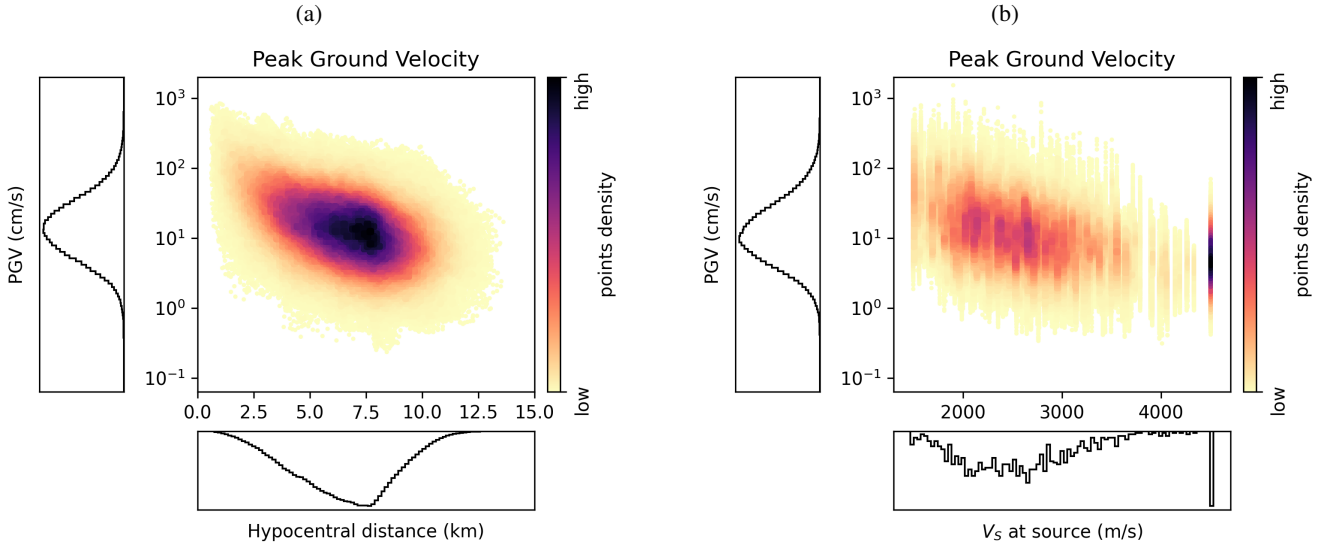


Figure 6. For each sample and each sensor, the PGV is shown against (a) the hypocentral distance, (b) the S-wave velocity at the source location. The PGV is computed on the East-West component, results are very similar for the two other components.

It is also known that the seismic energy E_s generated by a fault rupture is

$$I_A E_s = \frac{\pi}{2g} \int_0^T a^2(t) dt \frac{M_0 \Delta\sigma}{2\mu} \tag{7}$$

where ~~$a(t)$ is the acceleration field and T is the total duration of the signal. The RSD is highly variable between simulations, with a median value of 1.52 s and 90 % of seismic waveforms having a RSD lower than 4.17 s (Figure 3b)~~ M_0 is the seismic moment, $\Delta\sigma$ is the stress drop and μ is the shear modulus at the fault location. Knowing that the shear wave velocity writes $V_S = \sqrt{\mu/\rho}$, equation 7 indicates that the seismic energy is inversely proportional to V_S^2. And Figure 6b confirms that the PGV is negatively correlated with the velocity of S waves at the source location.

4.3 Distribution of Pseudo-Spectral Acceleration (PSA)

The Pseudo-Spectral Acceleration (PSA) is a commonly used metric to estimate structural response. It evaluates the maximal acceleration of a one-degree-of-freedom oscillator (with a 5 % damping) with a natural period T. At T=0.2 s, the PSA in the HEMEW^S-3D database is comprised between 2.3×10^{-3} g and 81.2 g. It decreases to the interval 5.8×10^{-4} g ; 6.1 g at T=1.0 s. Figure 7 additionally shows that there exists a negative correlation between the PSA and the hypocentral distance. ~~Finally, Figure **??** indicates that significant ground motion values can be found between 1.6 s and 17 s.~~

(a) ~~Magnitude of the velocity fields. The threshold indicates the detection value for the first wave arrivals.~~

(b)

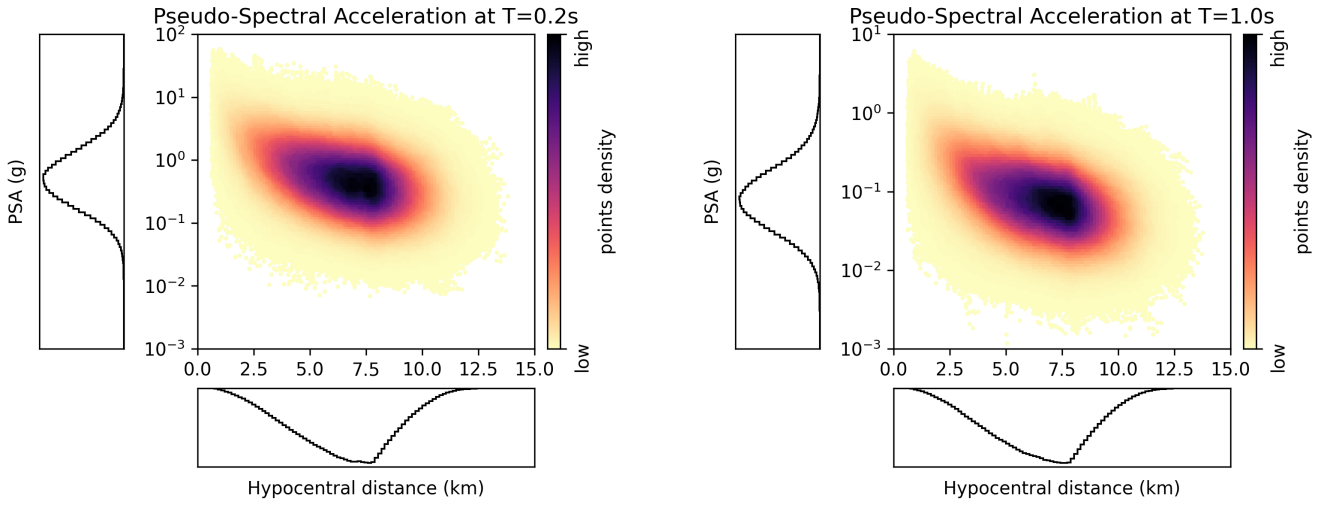


Figure 7. For each sample and each sensor, the Pseudo-Spectral Acceleration is shown against the hypocentral distance at period T=0.2 s (a) and T=1.0 s (b). The PSA is computed on the East-West component, results are very similar for the two other components.

4.4 Dimensionality

In supervised deep learning, it is always challenging to determine whether the size of the database (i.e. the number of samples) is sufficient to represent its variability. This questions relates to the definition of the intrinsic dimension of the dataset, which indicates the number of hidden variables that should be necessary to represent the main features of the samples. In the following, we provide insights on this question with the intrinsic dimension based on the Principal Component Analysis (Section 4.4.1), the correlation dimension (Section 4.4.2), the Maximum Likelihood Estimate (Section 4.4.3), and the Structural Similarity Index (Section 4.5).

4.4.1 Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) ~~is a linear dimensionality reduction method providing~~ decomposes data in principal components that correspond to the directions where data vary the most~~as *principal components*. The principal components form a basis on which data can then be represented by a low number of PCA coefficients. Due to its linearity, the PCArequires a large number of components to accurately represent complex patterns. Table~~ . For different sizes of datasets, we compute the number of principal components required to retain 95 % of variance and define this number as the intrinsic dimension of data. The 3D geological models and the 3D ground motion wavefields are transformed into 1D vectors to perform the PCA. To reduce the memory requirements, ground motions are analyzed only on the East-West component. Geological models are represented by $32 \times 32 \times 32 = 32\,768$ points and ground motions contain $16 \times 16 \times 320 = 81\,920$ points (16 sensors in directions x and y, and 320 time steps between 0 s and 6.4 s). To ease the computation on the large sample covariance matrix, an incremental PCA algorithm was used (Ross et al., 2008).

Table 3 and Figure B1 show that more than 1000 principal components are ~~sufficient~~ needed to reconstruct the geological models with high accuracy ~~(95% of the variance explained), whereas 2500 principal components are needed for the velocity fields~~whereas the intrinsic dimension of ground motion wavefields is around 4900. It is reasonable that the wavefields intrinsic dimension is larger than the geological dimension since wavefields variability is created by geological variations and the source position. However, ~~the PCA~~ due to its linearity, the PCA requires a large number of components to accurately represent complex patterns. Therefore, it may overestimate the intrinsic data dimension.

Table 3. Database intrinsic dimension estimated by PCA, correlation dimension, and MLE for the geological database and the velocity fields database, depending on the number of data samples

	Geological database					Velocity fields database				
Nb. of samples ($\times 10^3$)	2	6	10	20	30	2	6	10	20	30
PCA	491	766	880	~~1005~~ 1006	~~–~~1094	~~933~~ 853	~~1709~~ 2057	~~2086~~ 2853	~~2557~~ 4073	~~–~~4875
Correlation dimension	~~8.2~~ 8.8	~~8.1~~ 8.3	8.4	8.3	8.2	~~8.2~~ 2.2	~~15.7~~ 2.3	~~16.4~~ 2.3	~~16.1~~ 2.3	~~16.2 15.7~~ 2.3
MLE	17.9	23.4	26.7	31.5	33.9	~~29.7~~ 107.2	~~34~~ 129.0	~~34.4~~ 116.6	~~35.7~~ 113.8	~~36.6~~ 110.9

4.4.2 Correlation dimension

An alternative dimensionality measure was introduced by Grassberger and Procaccia (1983) as the correlation dimension, which characterizes the distance between pairs of samples. For a dataset of N samples ~~$\{V_{S,i}\}_{1\leq i\leq N}$~~ $\{V_S^{(i)}\}_{1\leq i\leq N}$ and a given *radius* r, the correlation dimension ($C_N(r)$) is defined as the ratio of sample pairs ~~$(V_{S,i}, V_{S,j})_{i\neq j}$~~ $(V_S^{(i)}, V_S^{(j)})_{i\neq j}$ being at distance less than r

$$C_N(r) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathbb{1}\left(\| V_{S,i}^{(i)} - V_{S,j}^{(j)} \| \leq r \right) \tag{8}$$

~~Table 3 and FigureB3a indicate~~ Figure B2 indicates a correlation dimension of 8 for the geological dataset, which is significantly lower than the PCA dimension. In fact, it is known that the correlation dimension may underestimate the intrinsic dimension, especially "when data are scattered" (Qiu et al., 2023), which is likely to be the case in high-dimensional spaces. ~~The~~ However, the correlation dimension of the ~~velocity fields is around 16 (Tab. 3 and Fig. B3b), which is also lower than the number of PCA components~~ground motion wavefields is debatable as it drops to 2. Figure B3c shows that the log-log representation of $C_N(r)$ does not produce an obvious linear part, which makes it difficult to identify the correlation dimension.

4.4.3 MLE intrinsic dimension

Levina and Bickel (2004) proposed another approach based on the Maximum Likelihood Estimator (MLE) of the distance to the closest neighbours. Figure B4 shows the evolution of the intrinsic dimension as a function of the number of samples for geological models and velocity wavefields. The intrinsic dimension of geological models is 34 while the dimension of velocity wavefields is larger (around 110). Although this method may still underestimate data with high intrinsic dimensionality (Qiu et al., 2023), it provides higher estimates than the correlation dimension~~(Table 3 and Fig. B4). Geological models and surface wavefields have intrinsic dimensions on the same order of magnitude (around 34), with a slightly higher dimension for the wavefields (dimension of 37). This is sound since variability in the wavefields is created entirely from the variability in the geological models, while the source introduces a small complexity reflected by the slightly higher dimension.~~ .

~~Many different methods exist to estimate the data intrinsic dimension and we exemplified the well-known fact that they can lead to different values. Based on the correlation dimension and the MLE, one can argue that the intrinsic dimension of the geological database is around 10 to 30, and that the instrinsic dimension of the surface wavefields ranges around 20-35 instead.~~ It can also be noted that the intrinsic dimension ~~increased~~ increases with the number of samples, as was observed for the PCA~~and the MLE~~. This may reflect a flaw in the intrinsic dimension's definition or it may indicate that despite being already large, our database of 30 000 samples does not capture all the variability. ~~The correlation dimension, however, gives consistent estimates with respect to the number of samples.~~

4.5 Structural similarity

The correlation dimension is computed from the Euclidean distance between pairs of geological models. However, point-wise metrics do not necessarily best represent similarities between geological models, and alternative metrics such as the Structural

Similarity Index Measure (SSIM) have been introduced for this purpose (Wang et al., 2004). This index theoretically ranges from 0 to 1, with 0 indicating no similarity and 1 indicating perfectly similar geological models (although values between -1 and 0 can be obtained numerically from the covariance computation). The SSIM of two geological models A and B is defined as

$$SSIM(A,B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)} \tag{9}$$

where μ_A and μ_B are the means of A and B, σ_A and σ_B are the unbiased estimators of the variance of A and B, σ_{AB} is the unbiased estimator of the covariance of A and B, C_1 and C_2 are constants determined from the range of A and B values.

Figures B5a and B5b illustrates two pairs of geological models with the same SSIM of 0.6, meaning rather high similarity. The first geologies have similar mean values but different heterogeneities, resulting in a low Euclidean distance (Fig. B5a) while the second geologies have different mean values, leading to a higher Euclidean distance (B5b).

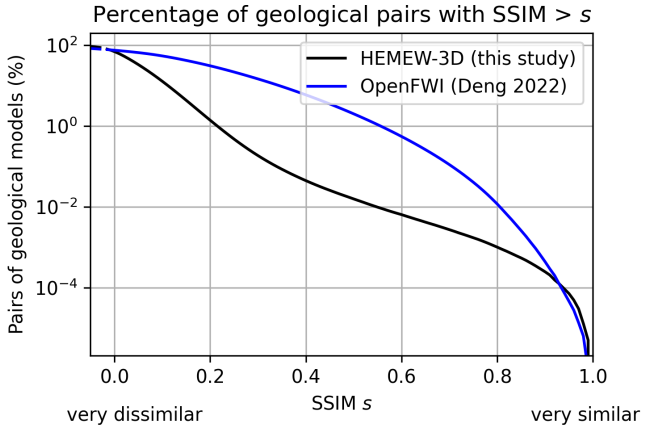


Figure 8. The Structural Similarity Index Measure (SSIM) quantifies the visual ressemblance between images, in a way that should mimic human perception. For each SSIM value s on the x-axis, the percentage of geological pairs being more similar than s is reported on the y-axis.

To give insights on the sparsity of the geological database, Figure 8 shows that only $1.4\,\%$ of geological pairs have a SSIM greater than 0.2. This means that geological models are generally very distinct from each other in the ~~HEMEW-3D~~ HEMEW^S-3D database. For comparison, the 2D OpenFWI dataset leads to significantly higher SSIM, with $31\,\%$ of geologies having a SSIM larger than 0.2 (3000 models were chosen from each of the 10 OpenFWI families, (Deng et al., 2022)).

5 ~~Discussion~~Applications

5.1 ~~Applications~~Dimensionality reduction of geological models

~~The geological database was used to study dimensionality reduction methods such as the PCA and a~~ Dimensionality analyses have shown that at least 1000 principal components are necessary to represent geological models with enough accuracy, as measured by the reconstructed variance. This means that the PCA provides a basis of 3D ~~autoencoder (Lehmann et al., 2022) . It was shown that at least~~ models to decompose a wide diversity of geological models. One can consider geological models that are very different from the random fields contained in the HEMEW^S-3D database, for instance embedding a basin shape.

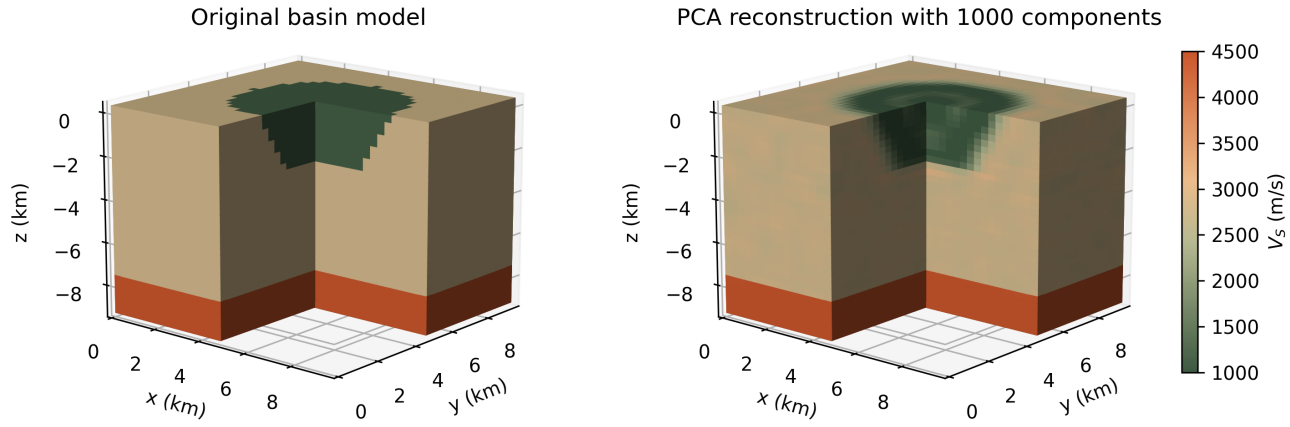


Figure 9. The original geological model (left) contains two homogeneous layers and a circular basin inserted inside the top layer. The PCA reconstruction was obtained with 1000 PCA components (right).

Figure 9 shows that 1000 ~~PCA components should be preserved to reconstruct a geological model whose heterogeneities are not too smoothed. This value matches with the intrinsic dimension estimated with the PCA . More importantly, although the HEMEW-3D database contains only geologies with horizontal heterogeneous layers, the PCA basis allows the reconstruction of specific geological models such as sedimentary basins (Fig. 9~~principal components allow a good reconstruction of the basin shape with correct velocity values inside and outside the basin. Edges are slightly blurred, which is expected since sharp contrasts correspond to high spatial frequencies that require many principal components. This example illustrates the generalization ability of the HEMEW^S-3D database from a geometrical point of view. To match the design of the HEMEW^S-3D database, the velocity values are chosen within the same bounds. In one were to consider real sedimentary basins, rescaling should be applied to target lower velocity values.

The influence of the PCA reconstruction on the generated velocity wavefields was investigated in more details in Lehmann et al. (2022) . It was shown that wavefields created by the propagation of seismic waves inside the reconstructed geological models and the reference model are very similar. When the initial geological model has strong heterogeneities, heterogeneities tend to be blurred in the PCA reconstruction, which reduces the dispersion of seismic waves. As a consequence, velocity wavefields generated inside the reconstructed geological model have slightly larger amplitudes.

5.2 Velocity fields predictions

Since the HEMEW^S-3D database associates geological models and sources with their corresponding velocity wavefields, it can serve to predict the latter from the former. Neural operators are one class of SciML models that have shown great success in the prediction of parametric PDEs. One can mention in particular the Multiple Input Fourier Neural Operator (MIFNO, Lehmann et al. (2024)) that uses the Fast Fourier Transform to learn the frequential representation of the elastic wave equation and a dedicated handling of the source term (Fig. C1).

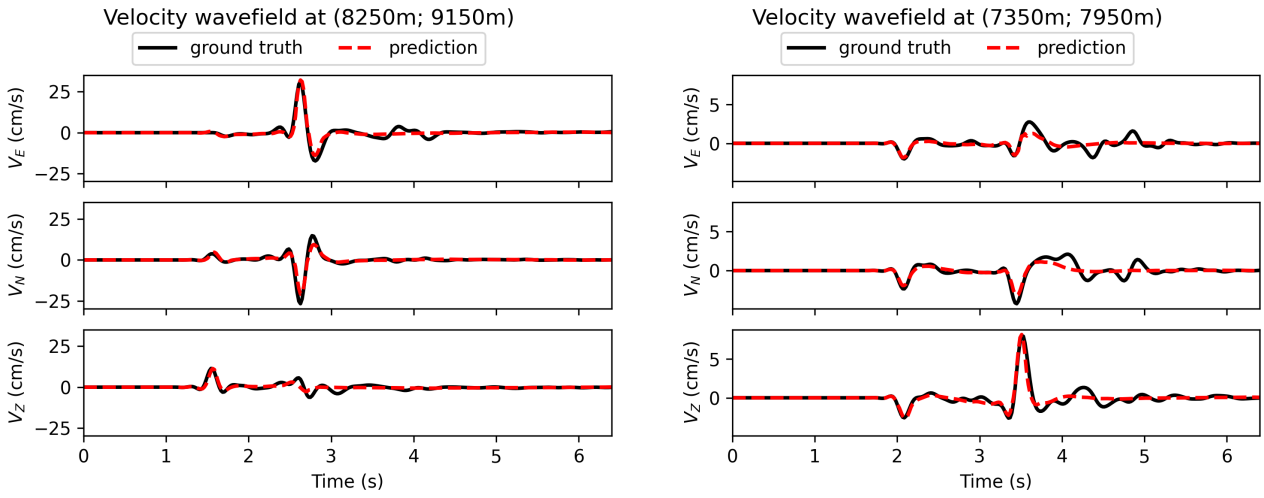


Figure 10. ~~The original geological model (left) contains~~ For two ~~homogeneous layers and a circular basin inserted inside~~ spatial points, the ~~top layer. The PCA reconstruction was obtained~~ velocity field predicted by the MIFNO (red dashed line) is compared with ~~1000 PCA components~~ the reference from the HEMEW^S-3D database (~~right~~black line).

~~Predictions of surface wavefields were also conducted with Fourier Neural Operators based on the HEMEW-3D database(Lehmann et al., . This SciML method takes as inputs 3D geological models and returns 3D velocity fields (functions of two spatial coordinates locating the sensor and a third dimension for time). Furthermore, this model can be specialized for target regions to conduct seismic hazard analyses (Lehmann et al., 2023b)~~For each geological model and source in the HEMEW^S-3D database, the MIFNO predicts the velocity field at each surface point. Figure 10 illustrates that the MIFNO gives accurate predictions for samples with different ground motions. This shows that the variability and size of the HEMEW^S-3D database are appropriate to train complex SciML models.

5.3 Other potential applications

Thanks to the large number of simulations, one can also envision studying the variability of ground motion to capture its statistical distribution~~with the minimal number of simulations~~. In particular, one can investigate the best sampling that minimizes the number of samples while preserving the largest ground motion variability (Tarbali and Bradley, 2015).

Additionally, the ~~HEMEW-3D~~ HEMEW^{S}-3D database could serve to investigate the relationship between geological features and ground motion. Ground motion amplification by geological features close to the surface could especially be explored.

5.4 ~~Limitations and perspectives~~

~~Since the HEMEW-3D~~

6 Limitations and perspectives

Since HEMEW^{S}-3D is the first database providing 3D ground motion, it is constrained by some hypotheses to control its size and allow machine learning applications. ~~Firstly, the earthquake source~~

First, the minimum S-wave velocity of $1071\,\mathrm{m/s}$ is rather high when compared to S-wave velocities in soft sediments (typically of few hundreds of m/s) but coherent for hard sediments (Molinari and Morelli, 2011). V_S values in the HEMEW^{S}-3D database are also in line with national velocity models with a low spatial resolution that display surface values around $2000\,\mathrm{m/s}$. One should also note that the vertical resolution of the geological models is $300\,\mathrm{m}$ while very low V_S values are more commonly encountered in the first tens of meters. These low values would be averaged with higher deeper values in our models. In particular, this means that geological models in the HEMEW^{S}-3D database are not comparable with the common notion of $V_{S,30}$ (average V_S in the first $30\,\mathrm{m}$). Reducing the minimum velocity poses no theoretical limitation but would increase the computational cost of the subsequent numerical simulations since it increases the number of mesh elements.

Second, the maximum S-wave velocity of $4500\,\mathrm{m/s}$ corresponds to existing V_S values at the bottom of the Earth's crust often adopted in velocity models (Duverger et al., 2021; Molinari and Morelli, 2011). In addition, the bottom layer has a fixed ~~location and orientation. In the upcoming months, an updated version will be released with more configurations of the source.~~ thickness and value that originates from earlier works. Therefore, variability is considered only above this constant layer.

Third, we do not constrain the ordering of layer-wise V_S values to provide a large database variability that is essential for machine learning perspectives. This means that some layer arrangements may be unphysical, for instance if the mean values are linearly decreasing with depth. However, it is important to notice that the physics of wave propagation is still satisfied in those situations, which is the main concern of this work. From the metadata provided, users can filter geological models with custom criteria to exclude those *outliers* from their study.

Additionally, more diverse configurations could be designed by relaxing the assumption that all geological parameters depend on a single variable. This would imply~~for instance~~, for instance, varying the V_P/V_S ratio~~but feasibility studies need to be conducted to estimate the size of the database required to cover the added variability. V_S~~. V_S values lower than $1071\,\mathrm{m/s}$ will also be included in future versions of the database. Random anisotropic heterogeneities can also be generated for more diversity (Ta et al., 2010).

~~Also, despite all geological layers being physically plausible, some layer arrangements may be unphysical, for instance if the mean values are linearly decreasing with depth. Those samples can be considered as outliers but they have not been removed from the database since i) velocity inversion is still possible in some layers and ii) the aim is to build a general database from~~

~~which specific configurations can be later studied at a reduced cost. Users can filter geological models with custom criteria to exclude those outliers from their studies.~~

The domain size was limited to $9.6\,\mathrm{km}$ to prove that ~~operator learning~~ SciML was possible with a manageable dataset size. This size ~~already~~ allows reasonable local studies and is ~~in line with existing machine learning models for forward modelling (e.g. Rasht-Behesht et al. (2022); Yang et al. (2023)). To enlarge the domain, one should take advantage of the resolution invariance property of some neural operators (?). In addition, the bottom layer has a fixed thickness and value to guarantee that the energy release is the same in all simulations. Therefore, variability is considered only above this constant layer.~~ already larger than existing 2D databases (Tab. 1). Extending the spatial size is certainly of interest for some seismological applications.

It should also be noted that ~~the~~ numerical simulations are only valid up to a $5\,\mathrm{Hz}$ frequency, due to the mesh design, with numerical pollution for frequencies larger than $5\,\mathrm{Hz}$. We observed that it is crucial to apply a low-pass filter (with a cutoff frequency of $5\,\mathrm{Hz}$) to the velocity fields before using machine learning models, otherwise the model may try to fit numerical noise.

7 Conclusions

We presented the ~~HEMEW-3D~~ HEMEW$^{\mathrm{S}}$-3D database (HEterogeneous Materials and Elastic Waves with Source variability) that contains $30\,000$ geological models, source parameters and the time- and space-dependent surface wavefields generated by the propagation of seismic waves through each geological model. This database was conceived for the forward problem of wave propagation.

Geological models are built from horizontal layers randomly arranged and they correspond the velocity of shear waves ~~V_S~~($\mathrm{V_S}$). They represent a domain of size $9.6\,\mathrm{km} \times 9.6\,\mathrm{km} \times 9.6\,\mathrm{km}$ discretized in $300\,\mathrm{m}$-wide elements. ~~V_S~~ $\mathrm{V_S}$ values are comprised between $1071\,\mathrm{m/s}$ and $4500\,\mathrm{m/s}$. Then, random fields are added independently in each geological layer to create 3D heterogeneities. Their parameters (coefficients of variation and correlation lengths) vary widely to cover diverse geological configurations and are given as metadata. Geological models are provided as cubes with $32 \times 32 \times 32$ voxels.

Seismic waves propagate numerically from the earthquake source to the surface. Point-wise sources have a random position and orientation. They are synthetized at the surface of the propagation domain by a grid of ~~sensors for $20\,\mathrm{s}$~~ 32×32 sensors for $8\,\mathrm{s}$. Simulations are conducted with the High-Performance Computing code SEM3D and amount to a total computational time of ~~1.6 million~~ 9×10^5 equivalent CPU hours. The dataset description shows that the ~~$20\,\mathrm{s}$~~ $8\,\mathrm{s}$ time window covers ~~all~~ most significant ground motion at the surface.

Ground motion characteristics ~~also~~ differ strongly between ~~geological configurations~~samples. They were analyzed in terms of Relative Significant Duration (RSD), P-wave arrival time, Peak Ground Velocity (PGV), and Pseudo-Spectral Acceleration (PSA). In addition to quantifying the distributions of essential intensity measures in seismology, these analyses confirm expected relationships between physical parameters and ground motion characteristics. In particular, hypocentral distance, $\mathrm{V_S}$ at the source location, and mean velocity were investigated.

Due to the size of individual samples in the HEMEW^S-3D database, one may wonder whether data could be represented with less parameters to reduce memory requirements. To this end, we explored different methods to estimate the data intrinsic dimension and we exemplified the well-known fact that they can lead to very different values. Taking the MLE as a lower bound, one can argue that the intrinsic dimension of the geological database is at least 30. In addition, the low values of the SSIM indicate that geologies are sparse and quite distant from each other in the HEMEW^S database.

Concerning the velocity wavefields, the PCA and the MLE confirm the intuition that the intrinsic dimension is larger than the geological dimension since the source adds variability to the time arrival of wavefields as well as their location at the surface. In this situation, it is reasonable to consider that the intrinsic dimension of ground motion is at least on the order of 100. However, if data are decomposed with the PCA, then the number of principal components is a few thousands. The correlation dimension yields questionable estimates of the intrinsic dimension that contradict our intuition and the PCA and MLE outcomes.

By providing a large number of physics-based simulations, the ~~HEMEW-3D~~ HEMEW^S-3D database offers new perspectives to study the relationship between geological properties and surface ground motion. It led to the first neural operator predicting 3D ground motion but many applications, in statistics, scientific machine learning, and deep learning are envisioned. We designed the database to be as generic as possible and we believe that several scientific communities can benefit from it.

8 Code and data availability

The database is referred to as Lehmann (2023) and can be downloaded at https://doi.org/10.57745/LAI6YU. The wave propagation code SEM3D is available at https://github.com/sem3d/SEM. The code used to generate the HEMEW^S-3D database is given at https://github.com/lehmannfa/HEMEW3D.

Appendix A: Dataset description

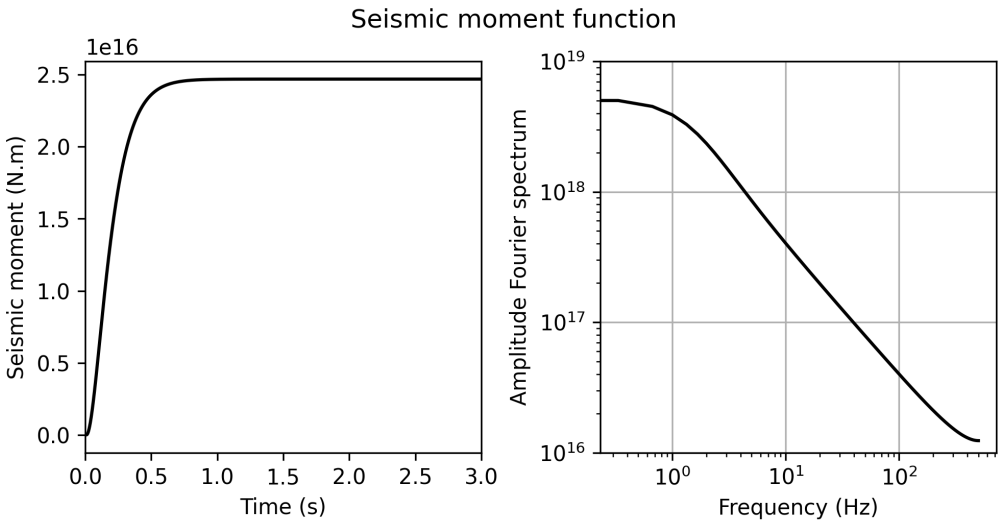


Figure A1. The seismic moment function in the HEMEW^{S}-3D database

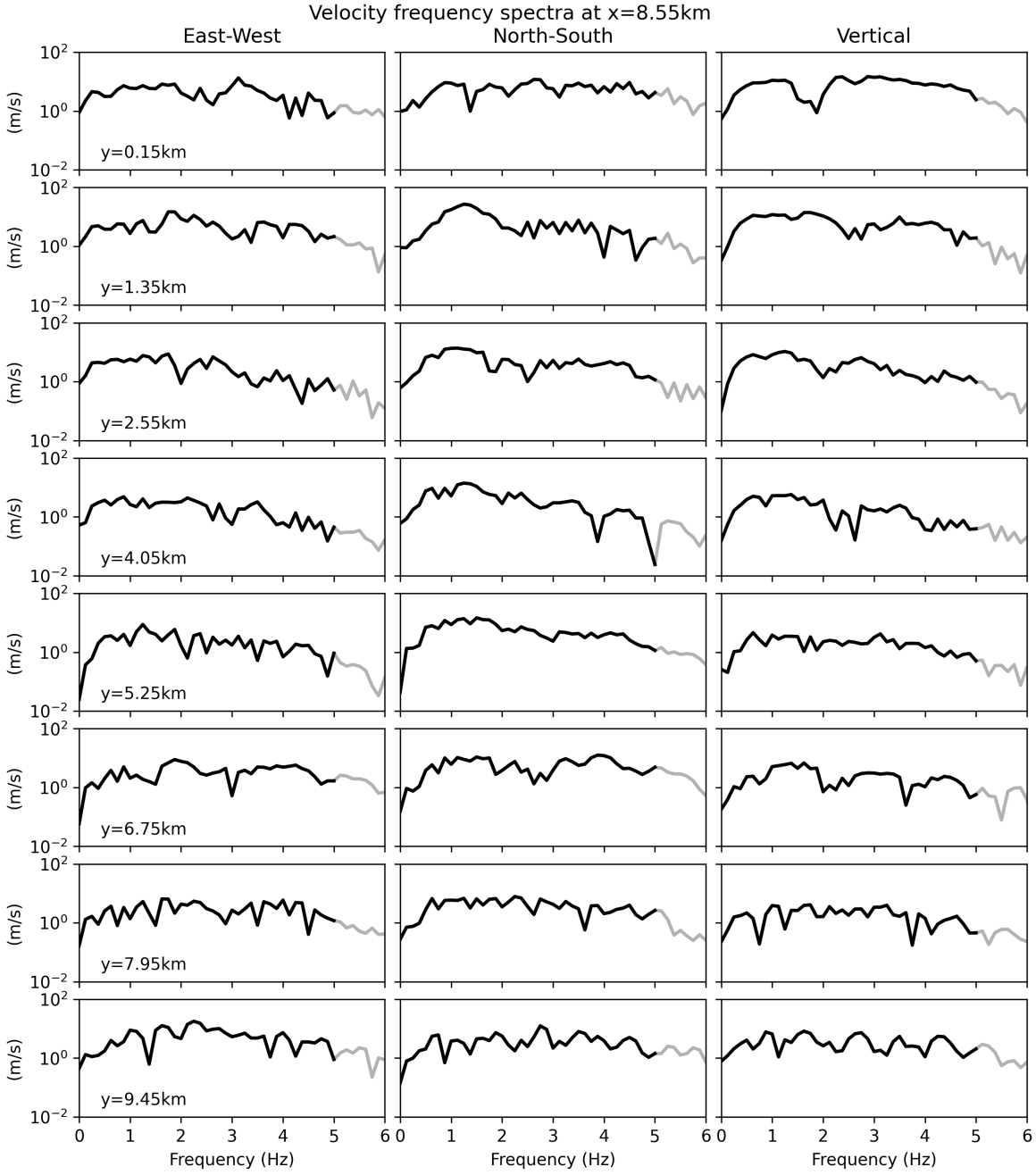


Figure A2. Frequency spectra corresponding to Figure 2. 3-component velocity waveforms synthetized at eight virtual sensors on a line parallel to the y axis at x=8.85 km. The grey line corresponds to frequencies larger than 5 Hz where numerical simulations are not accurate.

Appendix B: Dimensionality of data

B1 Principal Component Analysis

The intrinsic dimension based on the PCA components has been evaluated with the `scikit-dimension` package. Figure B1 illustrates the number of PCA components required to retain 95% of the data variance depending on the number of samples. ~~Due to memory issues, we were unable to estimate the number of components for the full database of 30 000 samples with this method. It can also~~ It can be noted that, for computational reasons, the velocity fields are represented by a single component (the East-West component, parallel to the x axis) for all three methods.

For comparison purposes, the wavefields intrinsic dimension is also computed for a previous version of the database where the source has a fixed position and orientation (HEMEW-3D database[2]). With this database, the wavefields intrinsic dimension was around 3200. It is reasonable that adding degrees of freedom with a random source increases the variability of data.

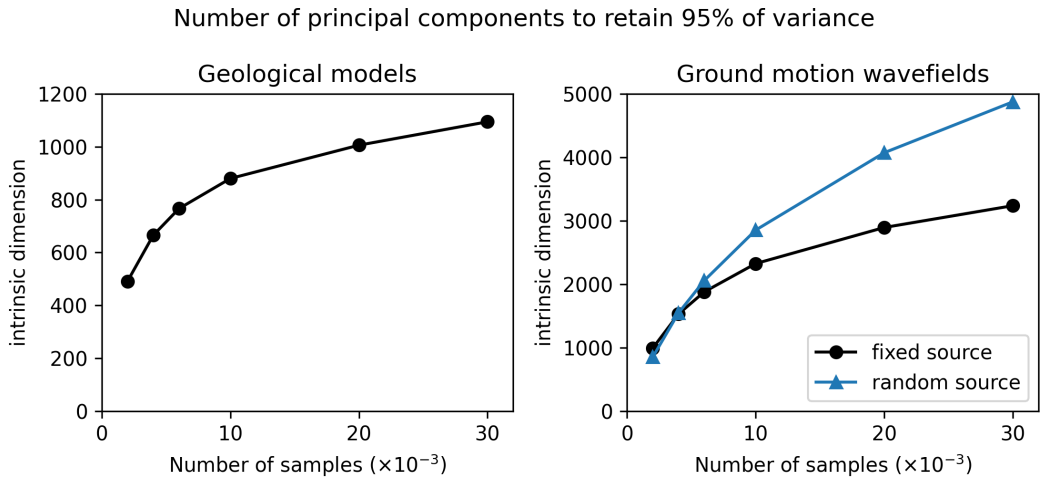


Figure B1. ~~PCA~~ Number of principal components (y-axis) required to represent 95% of the variance in data as a function of the dataset size (x-axis) for geological models (left) and ground motion wavefields (right). ~~Each sample~~ For ground motion, the HEMEW-3D database is ~~described by $32 \times 32 \times 32 = 32\,768$ points~~used for the fixed source (black line) and the HEMEW^S database corresponds to the blue line. ~~PCA of velocity fields. Each sample is described by $16 \times 16 \times 2000 = 512\,000$ points. Number of principal components (y-axis) required to represent 95% of the variance in data as a function of the dataset size (x-axis) for geological models (Figure **??**) and velocity fields (Figure **??**).~~

[2]https://doi.org/10.57745/LAI6YU

B2 Correlation dimension

The correlation dimension is determined as the slope of the linear part in the log-log representation of C_N (Figure ??B3). This definition is subject to some interpretation since one should determine which portion constitutes the linear part. Nevertheless, we found that small variations of the linear part limits had very little influence on the slope estimate (less than one unit).

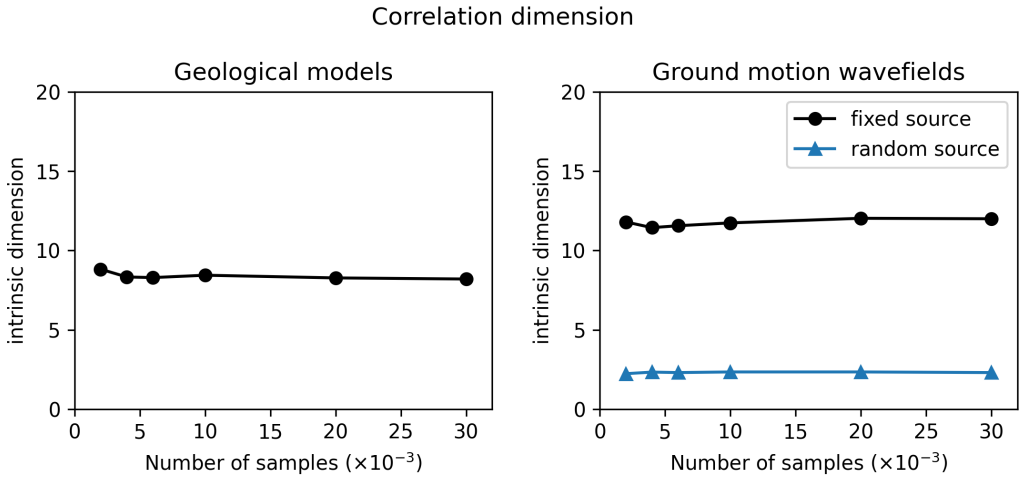
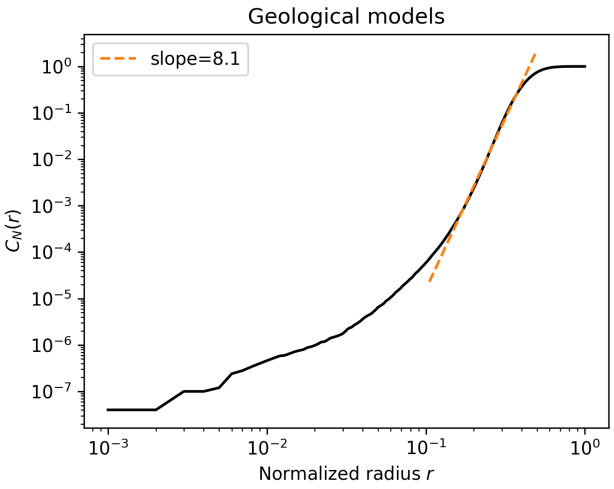
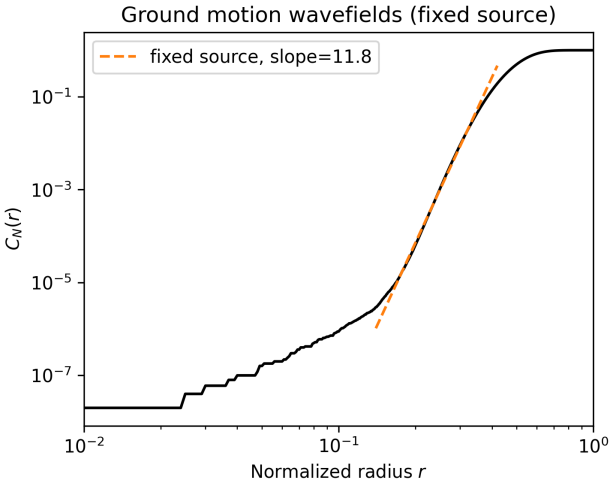


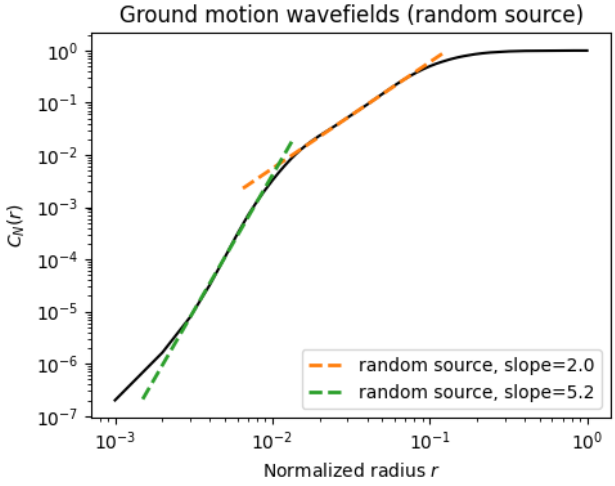
Figure B2. Correlation dimension (y-axis) as a function of the dataset size (x-axis) for geological models (left) and ground motion wavefields (right). For ground motion, the HEMEW-3D database is used for the fixed source (black line) and the HEMEW^S database corresponds to the blue line.



(a) Correlation dimension for ~~30,000~~ 30 000 geological models



(b) Correlation dimension for ~~30,000 velocity fields~~30 000 ground motion wavefields with a fixed source (HEMEW-3D database)



(c) ~~$C_N(r)$ is computed from the number of samples being at (Euclidean) distance smaller than r~~ Correlation dimension for ~~different values of r~~ 30 000 ground motion wavefields with a random source (~~Equation 8~~HEMEW^S database)~~. Then, the correlation dimension is obtained as the slope of the linear part in the log-log representation.~~

Figure B3. The correlation dimension $C_N(r)$ is computed from the number of samples being at (Euclidean) distance smaller than r for different values of r (Equation 8). Then, the correlation dimension is obtained as the slope of the linear part in the log-log representation.

B3 MLE based intrinsic dimension

The intrinsic dimension based on the Maximum Likelihood Estimator (MLE) has been computed with the `scikit-dimension` package. Figure B4 shows the evolution of the intrinsic dimension as a function of the number of samples for geological models and velocity fields.

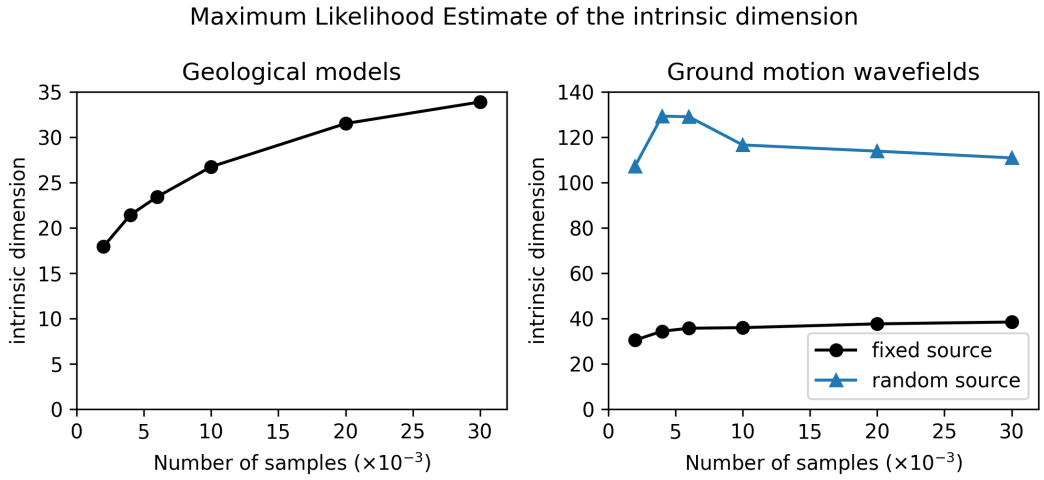
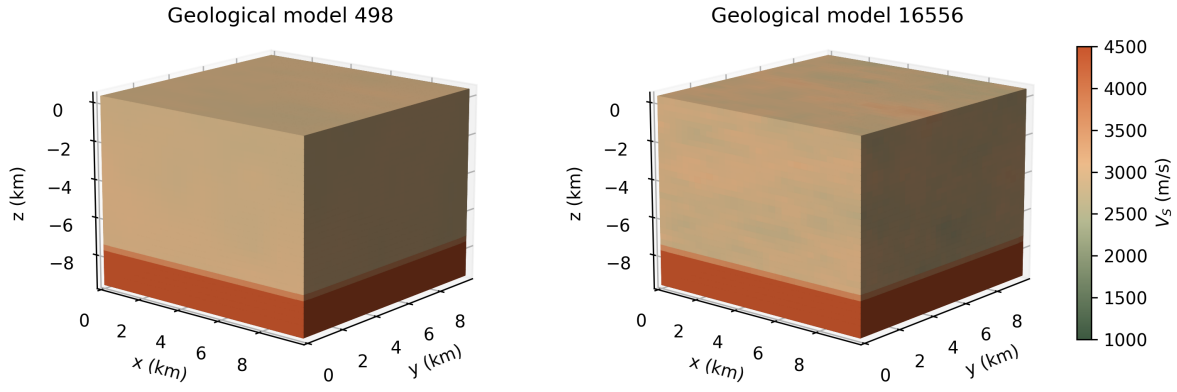


Figure B4. ~~Geological~~ Intrinsic dimension estimated by the MLE (y-axis) as a function of the dataset size (x-axis) for geological models (left) and ground motion wavefields (right). For ground motion, the HEMEW-3D database is used for the fixed source (black line) and the HEMEW^S database corresponds to the blue line.

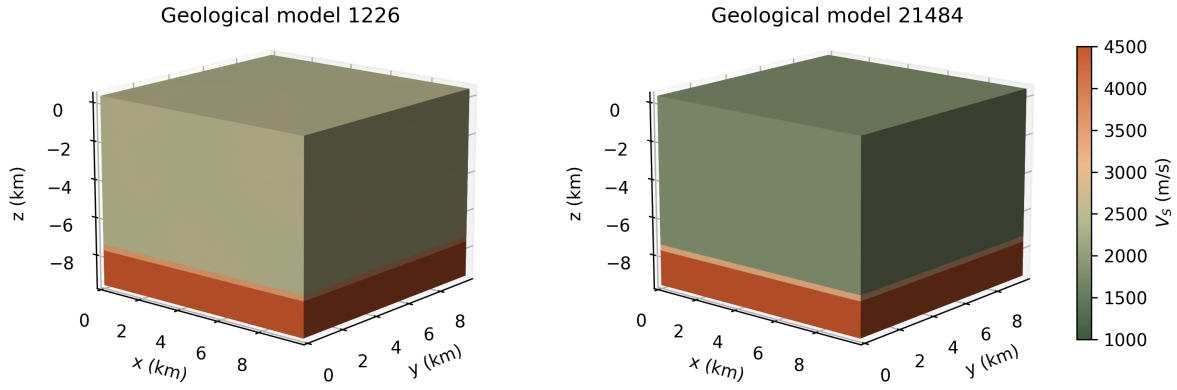
~~Velocity fields The intrinsic dimension determined by the Maximum Likelihood Estimator (y-axis) as a function of the number of samples in the dataset (x-axis), for geological models (Figure ??) and velocity fields (Figure ??).~~

B4 Structural Similarity Index

Figure B5 exemplifies two pairs of geological models with high similarity (SSIM of 0.6) but different properties. The first pair (Fig. B5a) has similar mean values but different heterogeneities while in the second pair, geological models are almost homogeneous but exhibit different mean values (Fig. B5b).



(a) Geological models with SSIM of 0.6 and normalized distance of 0.03



(b) Geological models with SSIM of 0.6 and normalized distance of 0.15

Figure B5. Two pairs of geological models with a high SSIM of 0.6

Appendix C: Multiple Input Fourier Neural Operator (MIFNO)

The MIFNO architecture is shown in Fig. C1.

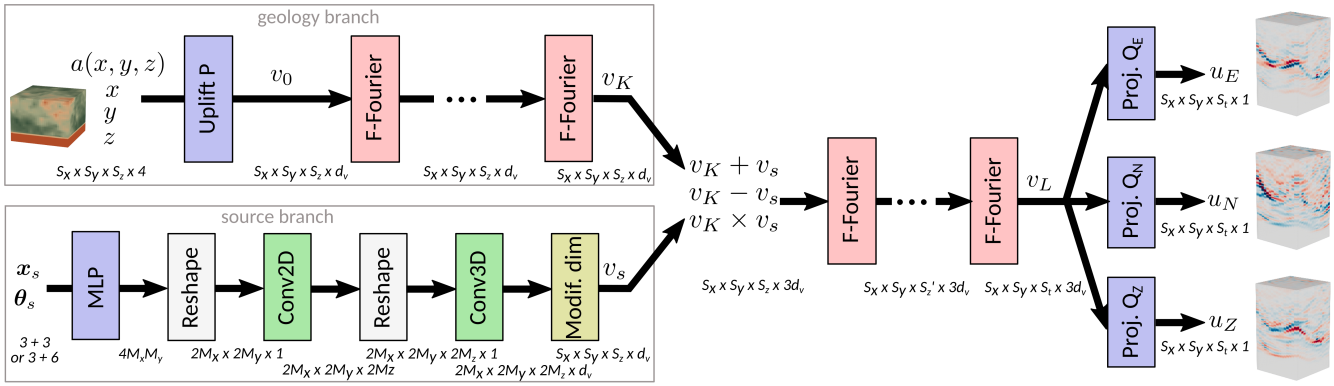


Figure C1. The MIFNO is made of a *geology branch* that encodes the geology with factorized Fourier (F-Fourier) layers, and a *source branch* that transforms the vector of source parameters $(\boldsymbol{x}_s, \boldsymbol{\theta}_s)$ into a 4D variable v_S matching the dimensions of the *geology branch* output v_K. Outputs of each branch are concatenated after elementary mathematical operations and the remaining factorized Fourier layers are applied. Uplift P and projection Q_E, Q_N, Q_Z blocks are the same as in the F-FNO.

Author contributions. F.L., F.G., D.C. designed the study. F.L. conducted the analyses. M.B. and D.C conceived the original idea. F.L. wrote the manuscript with input from all authors.

Competing interests. The authors have no competing interest to declare.

Acknowledgements. The authors are grateful for the resources and human support of the Très Grand Centre de Calcul (TGCC, CCRT, France).

References

Aki, K. and Richards, P. G.: Quantitative Seismology, Freeman, San Francisco, 1980.

Annon: Simulate CO2 Flow with Open Porous Media, https://github.com/microsoft/AzureClusterlessHPC.jl/tree/main/examples/opm, 2022.

Arias, A.: A Measure of Earthquake Intensity, Seismic design for nuclear plants, pp. 438–483, 1970.

Arroucau, P.: A Preliminary Three-Dimensional Seismological Model of the Crust and Uppermost Mantle for Metropolitan Franc, Tech. Rep. SIGMA2-2018-D2014, https://www.sigma-2.net/medias/files/sigma2-2018-d2-014-3d-velocity-model-france-approved-public-.pdf, 2020.

Bahrampouri, M., Rodriguez-Marek, A., Shahi, S., and Dawood, H.: An Updated Database for Ground Motion Parameters for KiK-net Records, Earthquake Spectra, 37, 505–522, https://doi.org/10.1177/8755293020952447, 2021.

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., and Kashinath, K.: Modelling Atmospheric Dynamics with Spherical Fourier Neural Operators, in: ICLR 2023 Workshop on Tackling Climate Change with Machine Learning, https://www.climatechange.ai/papers/iclr2023/47, 2023.

Castro-Cruz, D., Gatti, F., and Lopez-Caballero, F.: High-Fidelity Broadband Prediction of Regional Seismic Response: A Hybrid Coupling of Physics-Based Synthetic Simulation and Empirical Green Functions, Natural Hazards, 108, 1997–2031, https://doi.org/10.1007/s11069-021-04766-x, 2021.

Chaljub, E., Celorio, M., Cornou, C., Martin, F. D., Haber, E. E., Margerin, L., Marti, J., and Zentner, I.: Numerical Simulation of Wave Propagation in Heterogeneous and Random Media for Site Effects Assessment in the Grenoble Valley, in: He 6th IASPEI/IAEE International Symposium: Effects of Surface Geology on Seismic Motion, 2021.

Chernov, L. A.: Wave Propagation in a Random Medium, Translated by Richard A. Silverman, Mineola, New York, dover publications edn., 1960.

Colvez, M.: Influence of the Earth's Crust Heterogeneities and Complex Fault Structures on the Frequency Content of Seismic Waves, Ph.D. thesis, Université Paris Saclay, Paris-Saclay, 2021.

Convertito, V., De Matteis, R., Amoroso, O., and Capuano, P.: Ground Motion Prediction Equations as a Proxy for Medium Properties Variation Due to Geothermal Resources Exploitation, Scientific Reports, 12, 12 632, https://doi.org/10.1038/s41598-022-16815-x, 2022.

de Carvalho Paludo, L., Bouvier, V., and Cottereau, R.: Scalable Parallel Scheme for Sampling of Gaussian Random Fields over Very Large Domains: Parallel Scheme for Sampling of Random Fields over Very Large Domains, International Journal for Numerical Methods in Engineering, 117, 845–859, https://doi.org/10.1002/nme.5981, 2019.

De Martin, F., Chaljub, E., Thierry, P., Sochala, P., Dupros, F., Maufroy, E., Hadri, B., Benaichouche, A., and Hollender, F.: Influential Parameters on 3-D Synthetic Ground Motions in a Sedimentary Basin Derived from Global Sensitivity Analysis, Geophysical Journal International, 227, 1795–1817, https://doi.org/10.1093/gji/ggab304, 2021.

Delouis, B., Oral, E., Menager, M., Ampuero, J.-P., Guilhem Trilla, A., Régnier, M., and Deschamps, A.: Constraining the Point Source Parameters of the 11 November 2019 Mw 4.9 Le Teil Earthquake Using Multiple Relocation Approaches, First Motion and Full Waveform Inversions, Comptes Rendus. Géoscience, 353, 1–24, https://doi.org/10.5802/crgeos.78, 2021.

Deng, C., Feng, S., Wang, H., Zhang, X., Jin, P., Feng, Y., Zeng, Q., Chen, Y., and Lin, Y.: OpenFWI: Large-scale Multi-Structural Benchmark Datasets for Full Waveform Inversion, in: Advances in Neural Information Processing Systems, edited by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., vol. 35, pp. 6007–6020, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2022/file/27d3ef263c7cb8d542c4f9815a49b69b-Paper-Datasets_and_Benchmarks.pdf, 2022.

Ding, Y., Chen, S., Li, X., Wang, S., Luan, S., and Sun, H.: Self-Adaptive Physics-Driven Deep Learning for Seismic Wave Modeling in Complex Topography, Engineering Applications of Artificial Intelligence, 123, 106 425, https://doi.org/10.1016/j.engappai.2023.106425, 2023.

Duverger, C., Mazet-Roux, G., Bollinger, L., Trilla, A. G., Vallage, A., Hernandez, B., and Cansi, Y.: A Decade of Seismicity in Metropolitan France (2010–2019): The CEA/LDG Methodologies and Observations, p. 25, 2021.

El Haber, E., Cornou, C., Jongmans, D., Lopez-Caballero, F., Youssef Abdelmassih, D., and Al-Bittar, T.: Impact of Spatial Variability of Shear Wave Velocity on the Lagged Coherency of Synthetic Surface Ground Motions, Soil Dynamics and Earthquake Engineering, 145, 106 689, https://doi.org/10.1016/j.soildyn.2021.106689, 2021.

Equinor: Sleipner 2019 Benchmark Model, https://doi.org/10.11582/2020.00004, 2020.

Faccioli, E., Maggio, F., Paolucci, R., and Quarteroni, A.: 2d and 3D Elastic Wave Propagation by a Pseudo-Spectral Domain Decomposition Method, Journal of Seismology, 1, 237–251, https://doi.org/10.1023/A:1009758820546, 1997.

Feng, S., Wang, H., Deng, C., Feng, Y., Liu, Y., Zhu, M., Jin, P., Chen, Y., and Lin, Y.: $\mathbb{E}^{FWI}$: Multi-parameter Benchmark Datasets for Elastic Full Waveform Inversion of Geophysical Properties, http://arxiv.org/abs/2306.12386, 2023.

Fu, H., He, C., Chen, B., Yin, Z., Zhang, Z., Zhang, W., Zhang, T., Xue, W., Liu, W., Yin, W., Yang, G., and Chen, X.: 18.9-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of 18-Hz and 8-Meter Scenarios, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '17, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3126908.3126910, 2017.

Gadylshin, K., Lisitsa, V., Gadylshina, K., Vishnevsky, D., and Novikov, M.: Machine Learning-Based Numerical Dispersion Mitigation in Seismic Modelling, in: Computational Science and Its Applications – ICCSA 2021, edited by Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B. O., Rocha, A. M. A. C., Tarantino, E., and Torre, C. M., pp. 34–47, Springer International Publishing, Cham, 2021.

Gatti, F. and Clouteau, D.: Towards Blending Physics-Based Numerical Simulations and Seismic Databases Using Generative Adversarial Network, Computer Methods in Applied Mechanics and Engineering, 372, 113 421, https://doi.org/10.1016/j.cma.2020.113421, 2020.

Grady, T. J., Khan, R., Louboutin, M., Yin, Z., Witte, P. A., Chandra, R., Hewett, R. J., and Herrmann, F. J.: Model-Parallel Fourier Neural Operators as Learned Surrogates for Large-Scale Parametric PDEs, Computers & Geosciences, 178, 105 402, https://doi.org/10.1016/j.cageo.2023.105402, 2023.

Grassberger, P. and Procaccia, I.: Measuring the Strangeness of Strange Attractors, Physica D: nonlinear phenomena, 9, 189–208, 1983.

Hartzell, S., Harmsen, S., and Frankel, A.: Effects of 3D Random Correlated Velocity Perturbations on Predicted Ground Motions, Bulletin of the Seismological Society of America, 100, 1415–1426, https://doi.org/10.1785/0120090060, 2010.

Heinecke, A., Breuer, A., Rettenberger, S., Bader, M., Gabriel, A. A., Pelties, C., Bode, A., Barth, W., Liao, X. K., Vaidyanathan, K., Smelyanskiy, M., and Dubey, P.: Petascale High Order Dynamic Rupture Earthquake Simulations on Heterogeneous Supercomputers, in: SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 3–14, New Orleans, LA, USA, https://doi.org/10.1109/SC.2014.6, 2014.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Vil-

laume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Imperatori, W. and Mai, P. M.: Broad-Band near-Field Ground Motion Simulations in 3-Dimensional Scattering Media, Geophysical Journal International, 192, 725–744, https://doi.org/10.1093/gji/ggs041, 2013.

Jessell, M., Guo, J., Li, Y., Lindsay, M., Scalzo, R., Giraud, J., Pirot, G., Cripps, E., and Ogarko, V.: Into the Noddyverse: A Massive Data Store of 3D Geological Models for Machine Learning and Inversion Applications, Earth System Science Data, 14, 381–392, https://doi.org/10.5194/essd-14-381-2022, 2022.

Karimpouli, S. and Tahmasebi, P.: Physics Informed Machine Learning: Seismic Wave Equation, Geoscience Frontiers, 11, 1993–2001, https://doi.org/10.1016/j.gsf.2020.07.007, 2020.

Khazaie, S., Cottereau, R., and Clouteau, D.: Influence of the Spatial Correlation Structure of an Elastic Random Medium on Its Scattering Properties, Journal of Sound and Vibration, 370, 132–148, https://doi.org/10.1016/j.jsv.2016.01.012, 2016.

Komatitsch, D. and Tromp, J.: Introduction to the Spectral Element Method for Three-Dimensional Seismic Wave Propagation, Geophysical Journal International, 139, 806–822, https://doi.org/10.1046/j.1365-246x.1999.00967.x, 1999.

Lehmann, F.: Physics-Based Simulations of 3D Wave Propagation with Source Variability: HEMEW^S-3D, https://doi.org/10.57745/LAI6YU, 2023.

Lehmann, F., Gatti, F., Bertin, M., and Clouteau, D.: Machine Learning Opportunities to Conduct High-Fidelity Earthquake Simulations in Multi-Scale Heterogeneous Geology, Frontiers in Earth Science, 10, https://doi.org/10.3389/feart.2022.1029160, 2022.

Lehmann, F., Gatti, F., Bertin, M., and Clouteau, D.: Fourier Neural Operator Surrogate Model to Predict 3D Seismic Waves Propagation, in: 5th ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering, Athens, Greece, https://doi.org/10.7712/120223.10339.20362, 2023a.

Lehmann, F., Gatti, F., Bertin, M., and Clouteau, D.: Seismic Hazard Analysis with a Factorized Fourier Neural Operator (F-FNO) Surrogate Model Enhanced by Transfer Learning, in: NeurIPS 2023 AI for Science Workshop, https://openreview.net/pdf?id=xiNRyrBAjt, 2023b.

Lehmann, F., Gatti, F., and Clouteau, D.: Multiple-Input Fourier Neural Operator (MIFNO) for Source-Dependent 3D Elastodynamics, https://doi.org/10.48550/ARXIV.2404.10115, 2024.

Levina, E. and Bickel, P.: Maximum Likelihood Estimation of Intrinsic Dimension, in: Advances in Neural Information Processing Systems, edited by Saul, L., Weiss, Y., and Bottou, L., vol. 17, MIT Press, https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf, 2004.

Li, B., Wang, H., Yang, X., and Lin, Y.: Solving Seismic Wave Equations on Variable Velocity Models with Fourier Neural Operator, https://doi.org/10.48550/arXiv.2209.12340, 2022.

Liu, B., Yang, S., Ren, Y., Xu, X., Jiang, P., and Chen, Y.: Deep-Learning Seismic Full-Waveform Inversion for Realistic Structural Models, GEOPHYSICS, 86, R31–R44, https://doi.org/10.1190/geo2019-0435.1, 2021.

Mansoor, K., Buscheck, T., Yang, X., Carroll, S., and Chen, X.: LLNL Kimberlina 1.2 NUFT Simulations June 2018 (V2), https://doi.org/10.18141/1603336, 2020.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V.: INSTANCE – the Italian Seismic Dataset for Machine Learning, Earth System Science Data, 13, 5509–5544, https://doi.org/10.5194/essd-13-5509-2021, 2021.

Moczo, P., Kristek, J., Bard, P.-Y., Stripajová, S., Hollender, F., Chovanová, Z., Kristeková, M., and Sicilia, D.: Key Structural Parameters Affecting Earthquake Ground Motion in 2D and 3D Sedimentary Structures, Bulletin of Earthquake Engineering, 16, 2421–2450, https://doi.org/10.1007/s10518-018-0345-5, 2018.

Molinari, I. and Morelli, A.: EPcrust: A Reference Crustal Model for the European Plate: EPcrust, Geophysical Journal International, 185, 352–364, https://doi.org/10.1111/j.1365-246X.2011.04940.x, 2011.

Moseley, B., Markham, A., and Nissen-Meyer, T.: Solving the Wave Equation with Physics-Informed Deep Learning, https://doi.org/10.48550/arxiv.2006.11894, 2020.

Mousavi, S. M. and Beroza, G. C.: Machine Learning in Earthquake Seismology, Annual Review of Earth and Planetary Sciences, 51, 105–129, https://doi.org/10.1146/annurev-earth-071822-100323, 2023.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, 7, https://doi.org/10.1109/ACCESS.2019.2947848, 2019.

Ovadia, O., Kahana, A., Stinis, P., Turkel, E., and Karniadakis, G. E.: ViTO: Vision Transformer-Operator, http://arxiv.org/abs/2303.08891, 2023.

Paolucci, R., Smerzini, C., and Vanini, M.: BB-SPEEDset: A Validated Dataset of Broadband Near-Source Earthquake Ground Motions from 3D Physics-Based Numerical Simulations, Bulletin of the Seismological Society of America, 111, 2527–2545, https://doi.org/10.1785/0120210089, 2021.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model Using Adaptive Fourier Neural Operators, http://arxiv.org/abs/2202.11214, 2022.

Poursartip, B., Fathi, A., and Tassoulas, J. L.: Large-Scale Simulation of Seismic Wave Motion: A Review, Soil Dynamics and Earthquake Engineering, 129, 105 909, https://doi.org/10.1016/j.soildyn.2019.105909, 2020.

Qiu, H., Yang, Y., and Pan, H.: Underestimation Modification for Intrinsic Dimension Estimation, Pattern Recognition, 140, 109 580, https://doi.org/10.1016/j.patcog.2023.109580, 2023.

Raissi, M., Perdikaris, P., and Karniadakis, G.: Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations, Journal of Computational Physics, 378, 686–707, https://doi.org/10.1016/j.jcp.2018.10.045, 2019.

Rasht-Behesht, M., Huber, C., Shukla, K., and Karniadakis, G. E.: Physics-Informed Neural Networks (PINNs) for Wave Propagation and Full Waveform Inversions, Journal of Geophysical Research: Solid Earth, 127, https://doi.org/10.1029/2021JB023120, 2022.

Rekoske, J. M., Gabriel, A.-A., and May, D. A.: Instantaneous Physics-Based Ground Motion Maps Using Reduced-Order Modeling, Journal of Geophysical Research: Solid Earth, 128, e2023JB026 975, https://doi.org/10.1029/2023JB026975, 2023.

Ren, P., Rao, C., Chen, S., Wang, J.-X., Sun, H., and Liu, Y.: SeismicNet: Physics-informed Neural Networks for Seismic Wave Modeling in Semi-Infinite Domain, Computer Physics Communications, 295, https://doi.org/10.1016/j.cpc.2023.109010, 2024.

Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H.: Incremental Learning for Robust Visual Tracking, International Journal of Computer Vision, 77, 125–141, https://doi.org/10.1007/s11263-007-0075-7, 2008.

Rosti, A., Smerzini, C., Paolucci, R., Penna, A., and Rota, M.: Validation of Physics-Based Ground Shaking Scenarios for Empirical Fragility Studies: The Case of the 2009 L'Aquila Earthquake, Bulletin of Earthquake Engineering, 21, 95–123, https://doi.org/10.1007/s10518-022-01554-1, 2023.

Scalise, M., Pitarka, A., Louie, J. N., and Smith, K. D.: Effect of Random 3D Correlated Velocity Perturbations on Numerical Modeling of Ground Motion from the Source Physics Experiment, Bulletin of the Seismological Society of America, 111, 139–156, https://doi.org/10.1785/0120200160, 2021.

Shinozuka, M. and Deodatis, G.: Simulation of Stochastic Processes by Spectral Representation, Applied Mechanics Reviews, 44, 191–204, https://doi.org/10.1115/1.3119501, 1991.

Smerzini, C., Paolucci, R., and Stupazzini, M.: Comparison of 3D, 2D and 1D Numerical Approaches to Predict Long Period Earthquake Ground Motion in the Gubbio Plain, Central Italy, Bulletin of Earthquake Engineering, 9, 2007–2029, https://doi.org/10.1007/s10518-011-9289-8, 2011.

Song, C., Liu, Y., Zhao, P., Zhao, T., Zou, J., and Liu, C.: Simulating Multi-Component Elastic Seismic Wavefield Using Deep Learning, IEEE Geoscience and Remote Sensing Letters, pp. 1–1, https://doi.org/10.1109/LGRS.2023.3250522, 2023.

Ta, Q.-A., Clouteau, D., and Cottereau, R.: Modeling of Random Anisotropic Elastic Media and Impact on Wave Propagation, European Journal of Computational Mechanics, 19, 241–253, https://doi.org/10.3166/ejcm.19.241-253, 2010.

Tarbali, K. and Bradley, B.: Ground Motion Selection for Scenario Ruptures Using the Generalised Conditional Intensity Measures (GCIM) Method, Earthquake Engineering & Structural Dynamics, https://doi.org/10.1002/eqe.2546, 2015.

Thompson, E. M., Baise, L. G., and Kayen, R. E.: Spatial Correlation of Shear-Wave Velocity in the San Francisco Bay Area Sediments, Soil Dynamics and Earthquake Engineering, 27, 144–152, https://doi.org/10.1016/j.soildyn.2006.05.004, 2007.

Touhami, S., Gatti, F., Lopez-Caballero, F., Cottereau, R., de Abreu Corrêa, L., Aubry, L., and Clouteau, D.: SEM3D: A 3D High-Fidelity Numerical Earthquake Simulator for Broadband (0–10 Hz) Seismic Response Prediction at a Regional Scale, Geosciences, 12, 112, https://doi.org/10.3390/geosciences12030112, 2022.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E.: Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image Processing, 13, 600–612, https://doi.org/10.1109/TIP.2003.819861, 2004.

Wen, G., Li, Z., Long, Q., Azizzadenesheli, K., Anandkumar, A., and Benson, S. M.: Real-Time High-Resolution CO2 Geological Storage Prediction Using Nested Fourier Neural Operators, Energy & Environmental Science, 16, 1732–1741, https://doi.org/10.1039/d2ee04204e, 2023.

Witte, P. A., Konuk, T., Skjetne, E., and Chandra, R.: Fast CO2 Saturation Simulations on Large-Scale Geomodels with Artificial Intelligence-Based Wavelet Neural Operators, International Journal of Greenhouse Gas Control, 126, 103 880, https://doi.org/10.1016/j.ijggc.2023.103880, 2023.

Wu, Y., Aghamiry, H. S., Operto, S., and Ma, J.: Helmholtz Equation Solution in Non-Smooth Media by Physics-Informed Neural Network with Incorporating Quadratic Terms and a Perfectly Matching Layer Condition, GEOPHYSICS, pp. 1–66, https://doi.org/10.1190/geo2022-0479.1, 2023.

Yang, Y., Gao, A. F., Castellanos, J. C., Ross, Z. E., Azizzadenesheli, K., and Clayton, R. W.: Seismic Wave Propagation and Inversion with Neural Operators, The Seismic Record, 1, 126–134, https://doi.org/10.1785/0320210026, 2021.

Yang, Y., Gao, A. F., Azizzadenesheli, K., Clayton, R. W., and Ross, Z. E.: Rapid Seismic Waveform Modeling and Inversion with Neural Operators, IEEE Transactions on Geoscience and Remote Sensing, pp. 1–1, https://doi.org/10.1109/TGRS.2023.3264210, 2023.

Zhang, T., Trad, D., and Innanen, K.: Learning to Solve the Elastic Wave Equation with Fourier Neural Operators, Geophysics, pp. 1–63, https://doi.org/10.1190/geo2022-0268.1, 2023.

Zhu, C., Riga, E., Pitilakis, K., Zhang, J., and Thambiratnam, D.: Seismic Aggravation in Shallow Basins in Addition to One-dimensional Site Amplification, Journal of Earthquake Engineering, 24, 1477–1499, https://doi.org/10.1080/13632469.2018.1472679, 2020.

Zhu, W., Hou, A. B., Yang, R., Datta, A., Mousavi, S. M., Ellsworth, W. L., and Beroza, G. C.: QuakeFlow: A Scalable Machine-Learning-Based Earthquake Monitoring Workflow with Cloud Computing, Geophysical Journal International, 232, 684–693, https://doi.org/10.1093/gji/ggac355, 2022.