https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

Synthetic ground motions in heterogeneous geologies: the HEMEW-3D dataset for scientific machine learning

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

[1]CEA, 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 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. 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.

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

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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 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)). 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; Ren et al., 2022; Song et al., 2023; Wu 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. 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 (HEterogeneous Materials and Elastic Waves 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). Ground motion was synthetized at the surface of the simulation domain for $20\,\mathrm{s}$ on a grid of

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

16×16 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.

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

2 Related Work

Datasets of recorded ground motion are incredibly important for deep learning applications in seismology but their limitations have already been mentioned. In this section, we focus on datasets with 2D or 3D data used in geophysics and seismology.

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 reproduciblity 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 dataset does not provide any ground motion. For 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). The OpenFWI database (Deng et al., 2022) has been recently used with Neural Operators (Li et al., 2022). The above-mentioned wavefields were simulated with the 2D 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). Table 1 summarizes the characteristics of those datasets.

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

Table 1. Summary of datasets providing geological models and seismic wavefields. Dimension of geological models: number of grid points 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)

		Geological models				Seismic wavefields		
Dataset	train/test	Dimensions	Domain	Values V_S	Construction	Dimensions	Comp.	Source
Noddyverse (Jessell et al., 2022)	1M/-	$200 \times 200 \times 200$	$4 \times 4 \times 4\text{km}^3$	categorical	succession of geological events	N/A	N/A	N/A
OpenFWI (Deng et al., 2022)	408k/62k	70×70	$0.7 \times 0.7\text{km}^2$	1500; 4500 m/s	mathematical, from recorded images, and from geological faults	70×1000	1	5 fixed sources at the surface
OpenFWI (Deng et al., 2022) Kimberlina CO2	15k/4k	401×141	$4 \times 1.4\text{km}^2$	1200; 3600 m/s	from real data	101×1251	1	9 fixed sources at the surface
OpenFWI (Deng et al., 2022) Kimberlina 3D	1.6k/163	$400 \times 400 \times 350$	$4 \times 4 \times 3.5\text{km}^3$	?	from real data	$40 \times 40 \times 5001$	1	25 fixed sources at the surface
$\mathbb{E}^{\mathbf{FWI}}$ Feng et al. (2023)	144k/24k	70×70	$0.35 \times 0.35\text{km}^2$	$V_S \in [612; 3000\text{m/s}]$, $V_P \in [1500; 4500\text{m/s}]$	mathematical and from geological faults	70×1000	2	5 fixed sources at the surface
Ovadia et al. (2023)*	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	$10.2 \times 10.2\text{km}^2$	mean V_S=3000 m/s	large-scale random fields for V_P/V_S ratio + small-scale random fields for V_S	$64 \times 64 \times 400$	2	1 source with random location
HEMEW-3D (this study)	27k/3k	$32 \times 32 \times 32$	$9.6 \times 9.6 \times 9.6\text{km}^3$	1071; 4500 m/s	horizontal layers + random fields	$16 \times 16 \times 2000$	3	1 fixed source

*: data are not publicly available, Comp. = Number of components

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

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. 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 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). The source amplitude corresponds to the real seismic moment $M_0 = 2.47 \times 10^{16}$ N m. The source time evolution is given by $t \mapsto 1 - \left(1 + \frac{t}{\tau}\right) e^{-\frac{t}{\tau}}$ with τ = 0.1 s.

3.3 Heterogeneous geological models

The HEMEW-3D database contains pairs $\{V_{S,i}, \dot{\boldsymbol{u}}_i\}_i$ 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.

3.3.1 Homogeneous models

A 1.8 km-thick homogeneous layer is imposed at the bottom of each geological model, with a V_S value of $V_{S,max}$ = 4500 m/s. The minimum S-wave velocity is $V_{S,min}$ = 1071 m/s. Above the bottom layer, the number of horizontal layers and their thickness are randomly chosen for each sample $V_{S,i}$, 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 $\mathcal{U}([\mu_1; \mu_2])$. Values of $\mu_1 = V_{S,min}/0.6$ = 1785 m/s and $\mu_2 = V_{S,max}/1.4$ = 3214 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).

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

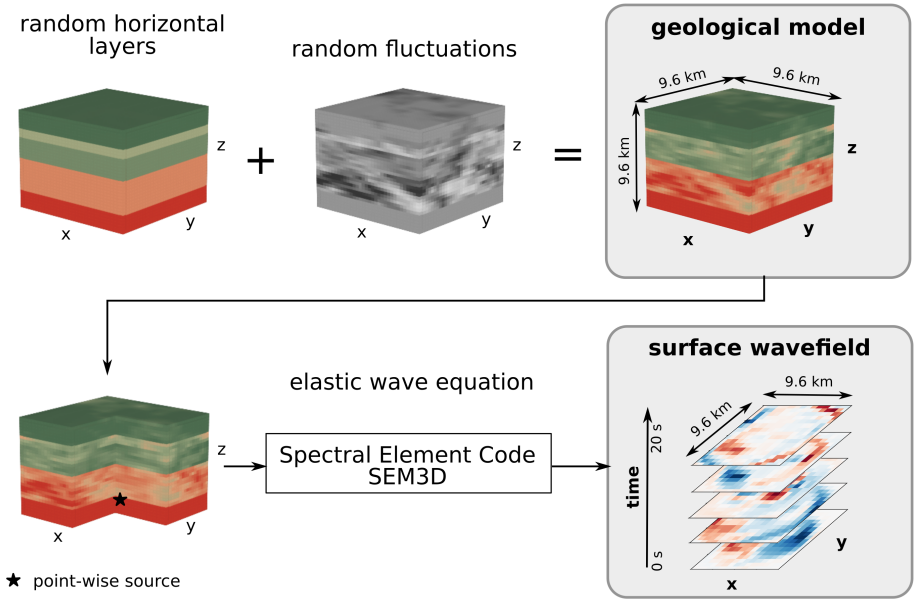


Figure 1. Geological models are built by adding heterogeneities to randomly chosen horizontal layers. Then, elastic waves are propagated from the source to the surface, where velocity wavefields are synthetized.

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.

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)

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

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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

$$Q_P = \max\left(\frac{V_P}{20}, \frac{V_S}{5}\right); Q_S = \frac{V_S}{10} \tag{4}$$

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), we drew random fields with a von Karman correlation kernel and a Hurst exponent of 0.1 (Chernov, 1960).

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 $V_{S,i}(x_B)$. 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 $\mathcal{N}(0.2, 0.1)$. 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).

3.3.3 Representation in the database

Geological realizations $V_{S,i}$ 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 $3.9\,\mathrm{GB}$, split in 15 files of 2000 geological models for easier data management. Additionally, metadata give the minimum, mean, maximum, and standard deviation of the pixel-wise geological values.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

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 $5\,\mathrm{Hz}$, which is relatively high for this type of simulations. Many simulations have been conducted so far with an accuracy up to 1 or $2\,\mathrm{Hz}$ (Rekoske et al., 2023; Rosti et al., 2023), while high-fidelity simulations for local realistic earthquake scenarios extend up to $10\,\mathrm{Hz}$ (Castro-Cruz et al., 2021; De Martin et al., 2021; Heinecke et al., 2014) (and exceptionally up to $18\,\mathrm{Hz}$, 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 sensors was placed between $150\,\mathrm{m}$ and $9450\,\mathrm{m}$ in both horizontal directions. At each monitoring point, the three-component velocity field is synthetized with a $100\,\mathrm{Hz}$ sampling frequency. Figure 2 illustrates velocity waveforms at five virtual sensors.

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 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 $369.9\,\mathrm{GB}$, 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.

4 Dataset analysis

4.1 Descriptive statistics

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 values range from $1756\,\mathrm{m/s}$ to $3145\,\mathrm{m/s}$ (Figure 3a).

Wave arrival times are usually determined from recordings, either manually by experts, or with machine learning methods. However, it is possible to compute approximated 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

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

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

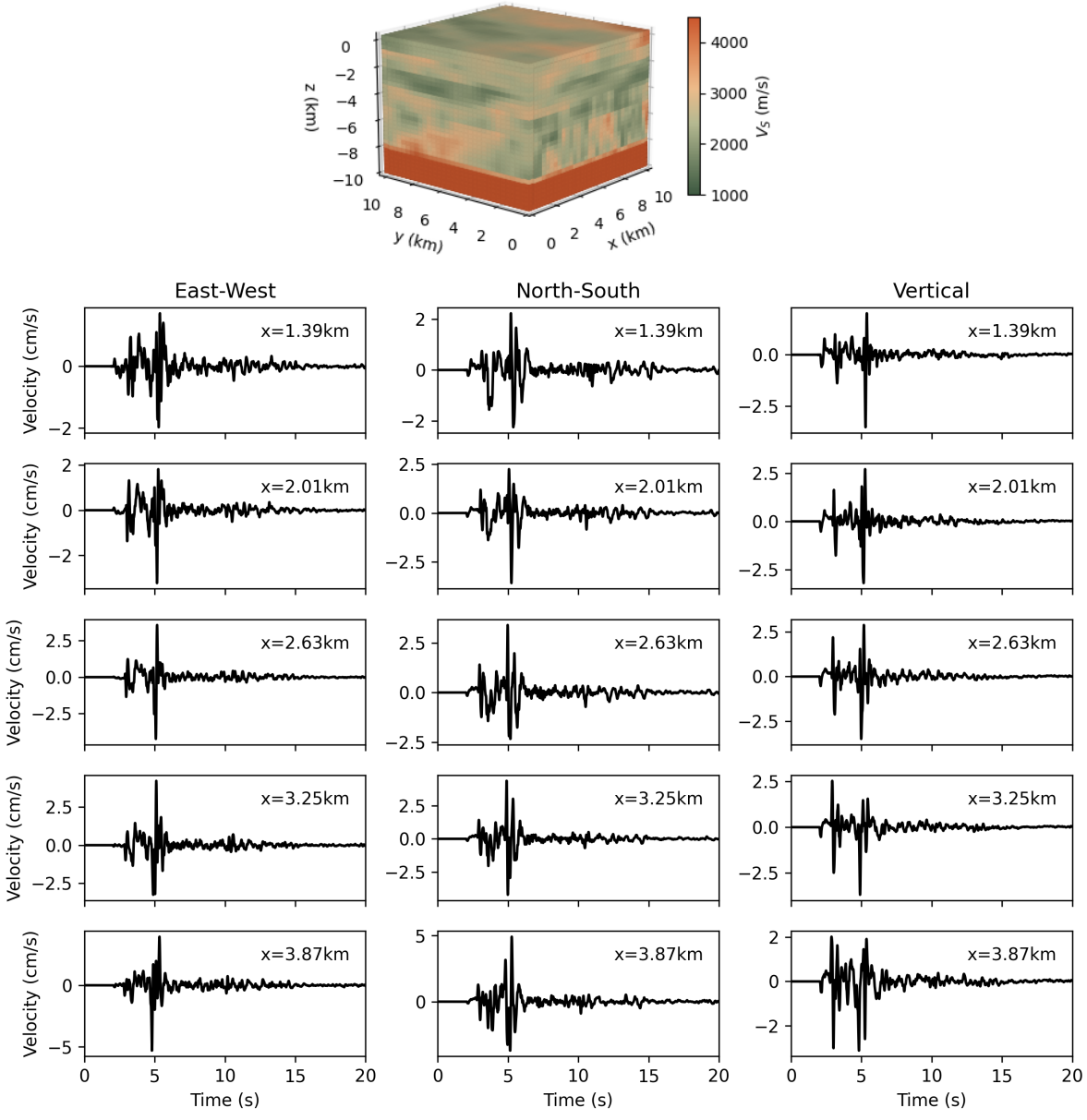
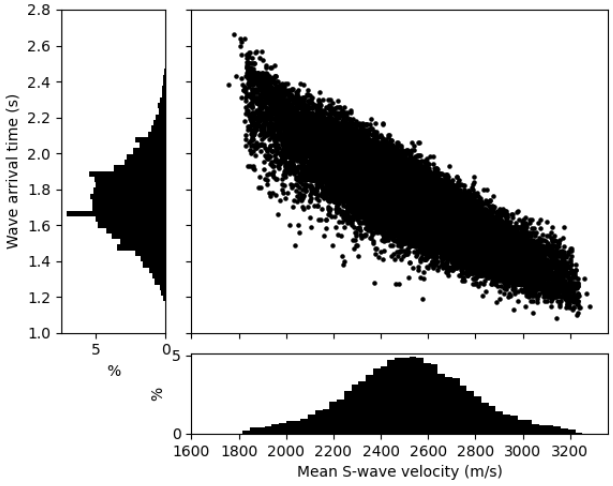


Figure 2. 3-component velocity waveforms syntehtized at 5 virtual sensors with x-coordinate between 1.39 km and 3.87 km, at y-coordinate equal 7.59 km. The corresponding geological model is shown at the top

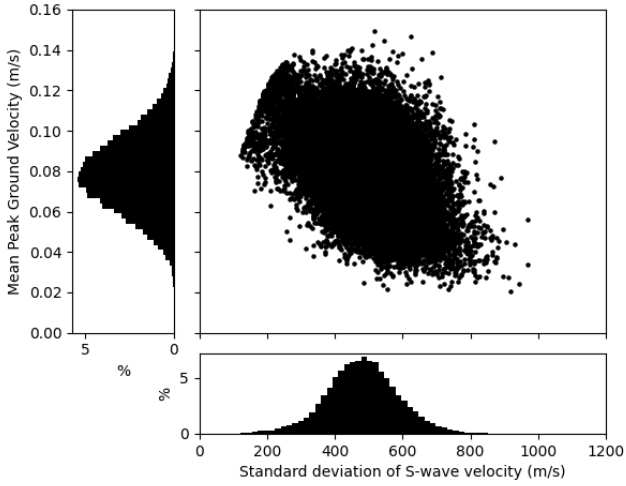
https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

velocity amplitude exceeds 10^{-4}m/s. The 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 first wave arrival time is negatively correlated with the mean V_S value (Figure 3a).



(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 3. Relationships between velocity fields features (PGV=Peak Ground Velocity) and geological properties. Each dot corresponds to one sample. Distributions are given in percentage of the total number of samples.

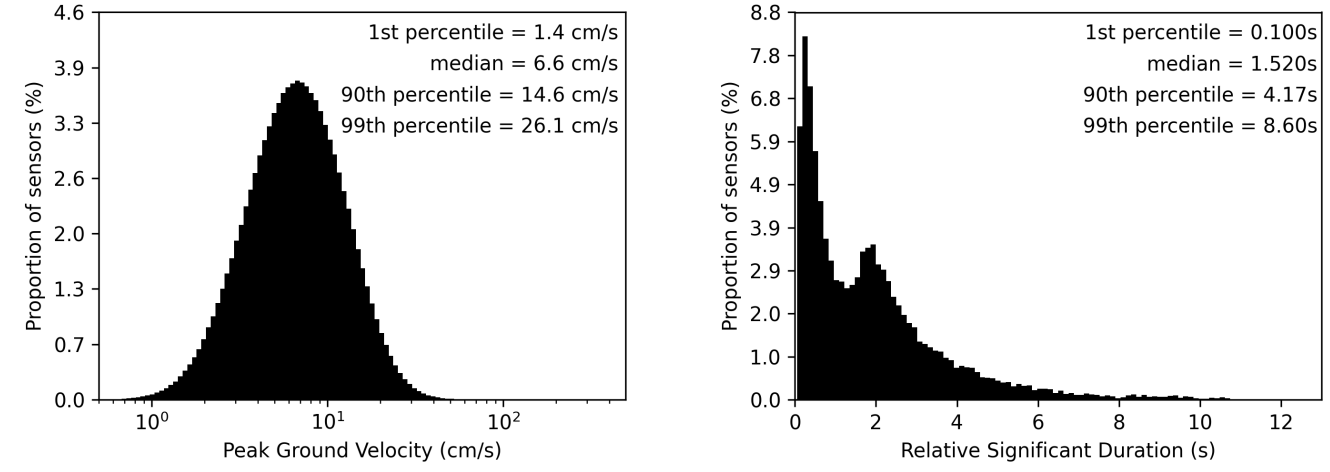
Seismic velocity fields can be characterized by several intensity measures. The Peak Ground Velocity (PGV) is computed as the maximum absolute value over all timesteps. However, it should be noted that, although this is common practice in numerical studies, this does not exactly corresponds to the definition 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.

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 4a). 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 3b shows that geological models with large standard variations tend to produce smaller PGVs since heterogeneities induce dispersion and diffraction of seismic waves that spread the wavefront along time.

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)

$$I_A = \frac{\pi}{2g} \int_0^T a^2(t) dt \tag{5}$$

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024



(a) (b)

Figure 4. 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.

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 $1.52\,\mathrm{s}$ and $90\,\%$ of seismic waveforms having a RSD lower than $4.17\,\mathrm{s}$ (Figure 4b). Finally, Figure 5 indicates that significant ground motion values can be found between $1.6\,\mathrm{s}$ and $17\,\mathrm{s}$.

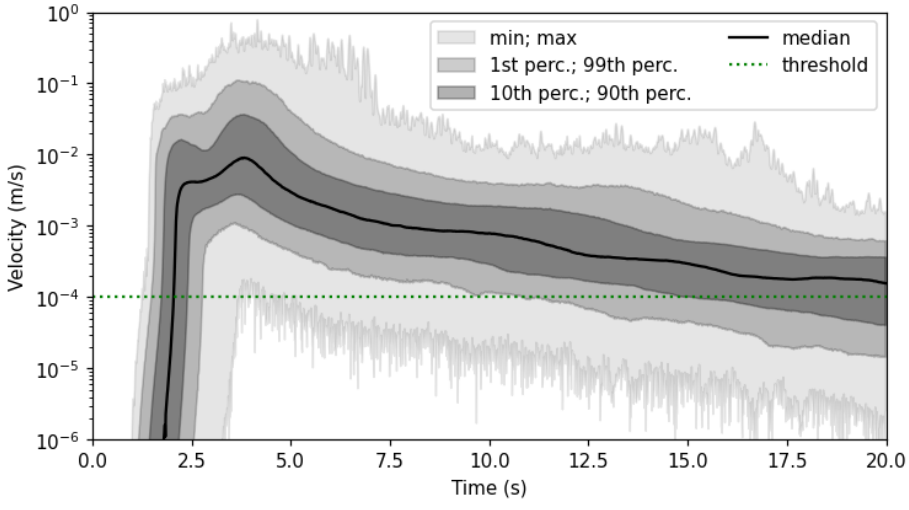


Figure 5. Magnitude of the velocity fields. The threshold indicates the detection value for the first wave arrivals.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

4.2 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 4.2.1), the correlation dimension (Section 4.2.2), the Maximum Likelihood Estimate (Section 4.2.3), and the Structural Similarity Index (Section 4.3).

4.2.1 Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) is a linear dimensionality reduction method providing 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 PCA requires a large number of components to accurately represent complex patterns. Table 3 and Figure A1 show that 1000 principal components are sufficient to reconstruct the geological models with high accuracy (95% of the variance explained), whereas 2500 principal components are needed for the velocity fields. However, the PCA 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	-	933	1709	2086	2557	-
Correlation dimension	8.2	8.1	8.3	8.2	8.2	15.7	16.4	16.1	16.2	15.7
MLE	17.9	23.4	26.7	31.5	33.9	29.7	34	34.4	35.7	36.6

4.2.2 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 $\{V_{S,i}\}_{1\leq i\leq N}$ and a given *radius* r, the correlation dimension ($C_N(r)$) is defined as the ratio of sample pairs $(V_{S,i}, V_{S,j})_{i\neq j}$ being at distance less than r

$$C_N(r) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathbb{1}\left(\|V_{S,i} - V_{S,j}\| \leq r\right) \tag{6}$$

Table 3 and Figure A2a indicate 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

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

data are scattered" (Qiu et al., 2023), which is likely to be the case in high-dimensional spaces. The correlation dimension of the velocity fields is around 16 (Tab. 3 and Fig. A2b), which is also lower than the number of PCA components.

4.2.3 MLE intrinsic dimension

Levina and Bickel (2004) proposed another approach based on the Maximum Likelihood Estimator (MLE) of the distance to the closest neighbours. 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. A3). 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 with the number of samples, 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.

4.3 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

$$SSIM(A,B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)} \tag{7}$$

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 A4a and A4b 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. A4a) while the second geologies have different mean values, leading to a higher Euclidean distance (A4b).

To give insights on the sparsity of the geological database, Figure 6 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 database.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

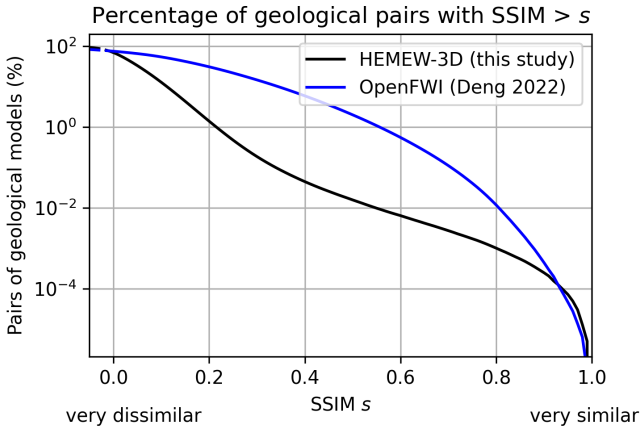


Figure 6. The Structural Similarity Index Measure (SSIM) quantifies the visual ressemblance between images, in a way that should mimic human perception. For each SSIM value s on the x-axis, the percentage of geological pairs being more similar than s is reported on the y-axis.

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

5.1 Applications

The geological database was used to study dimensionality reduction methods such as the PCA and a 3D autoencoder (Lehmann et al., 2022). It was shown that at least 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. 7).

Predictions of surface wavefields were also conducted with Fourier Neural Operators based on the HEMEW-3D database (Lehmann et al., 2023a). 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).

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 (Tarbali and Bradley, 2015).

Additionally, the HEMEW-3D 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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

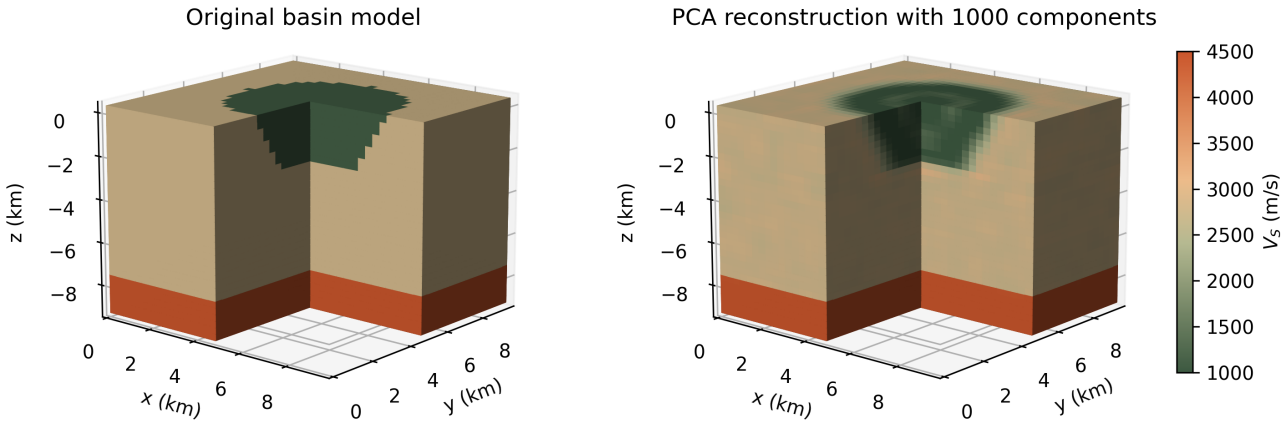


Figure 7. The original geological model (left) contains two homogeneous layers and a circular basin inserted inside the top layer. The PCA reconstruction was obtained with 1000 PCA components (right).

5.2 Limitations and perspectives

Since the HEMEW-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 has a fixed location and orientation. In the upcoming months, an updated version will be released with more configurations of the source.

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 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 values lower than $1071\,\mathrm{m/s}$ will also be included in future versions of the database.

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 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 (Raonic et al., 2023). 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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024


Open Access Earth System Science Data Discussions

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.

6 Conclusions

We presented the HEMEW-3D database (HEterogeneous Materials and Elastic Waves) that contains $30\,000$ geological models 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. 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 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. Geological models are provided as cubes with $32 \times 32 \times 32$ voxels.

Seismic waves propagate numerically from the earthquake source to the surface. They are synthetized at the surface of the propagation domain by a grid of sensors for $20\,\mathrm{s}$. Simulations are conducted with the High-Performance Computing code SEM3D and amount to a total computational time of 1.6 million equivalent CPU hours. The dataset description shows that the $20\,\mathrm{s}$ time window covers all significant ground motion at the surface. Ground motion characteristics also differ strongly between geological configurations.

By providing a large number of physics-based simulations, the HEMEW-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.

7 Code and data availability

The database is referred to as Lehmann (2023) and can be downloaded at https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/LAI6YU. The wave propagation code SEM3D is available at https://github.com/sem3d/SEM. The code used to generate the HEMEW-3D database is given at https://github.com/lehmannfa/HEMEW3D.

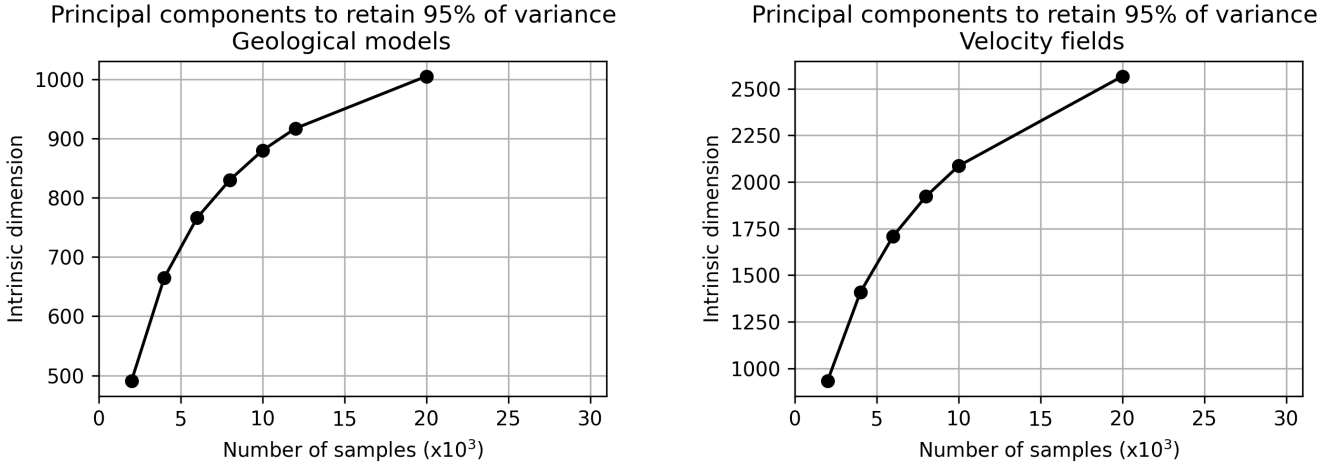
https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

Appendix A: Dimensionality of data

A1 Principal Component Analysis

The intrinsic dimension based on the PCA components has been evaluated with the `scikit-dimension` package. Figure A1 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 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.



(a) PCA of geological models. Each sample is described by $32 \times 32 \times 32 = 32\,768$ points.

(b) PCA of velocity fields. Each sample is described by $16 \times 16 \times 2000 = 512\,000$ points.

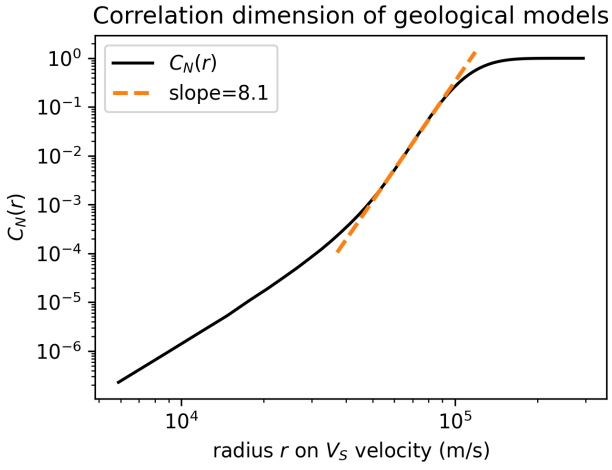
Figure A1. Number of principal components (y-axis) required to represent 95% of the variance in data as a function of the dataset size (x-axis) for geological models (Figure A1a) and velocity fields (Figure A1b).

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

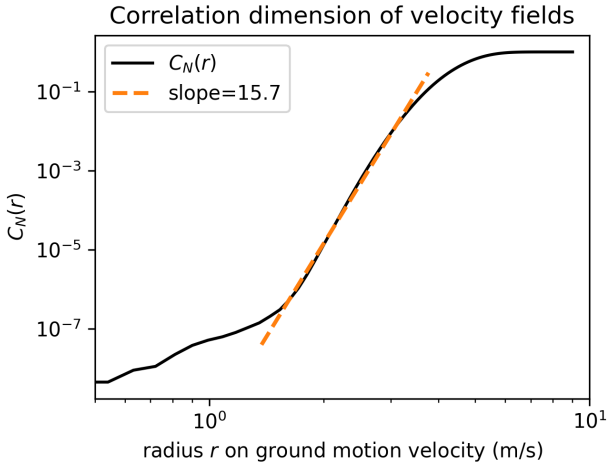
Open Access Earth System Science Data Discussions

A2 Correlation dimension

The correlation dimension is determined as the slope of the linear part in the log-log representation of C_N (Figure A2). 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).



(a) Correlation dimension for 30,000 geological models



(b) Correlation dimension for 30,000 velocity fields

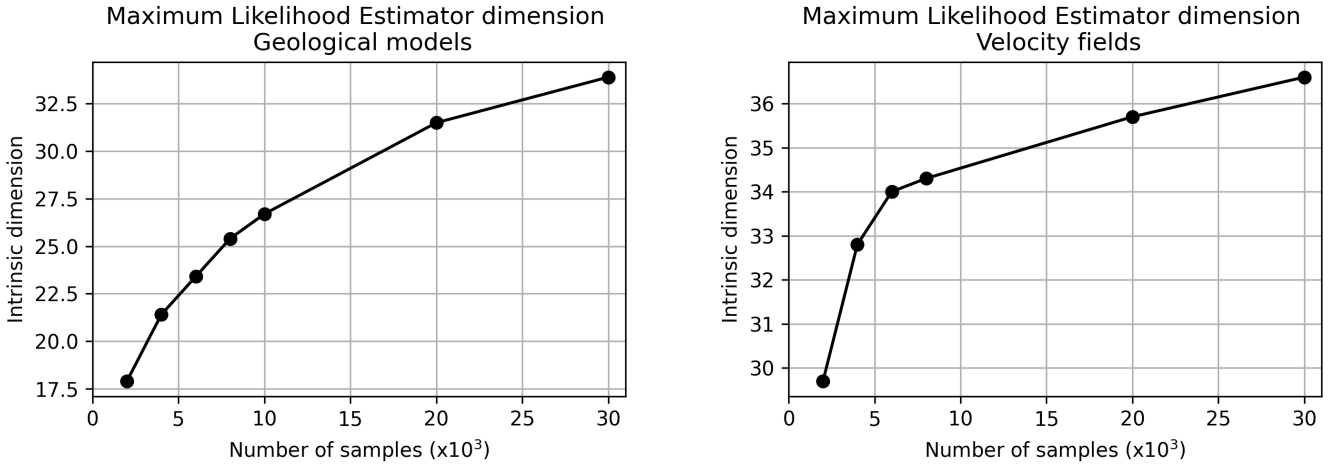
Figure A2. $C_N(r)$ is computed from the number of samples being at (Euclidean) distance smaller than r for different values of r (Equation 6). Then, the correlation dimension is obtained as the slope of the linear part in the log-log representation.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

A3 MLE based intrinsic dimension

320 The intrinsic dimension based on the Maximum Likelihood Estimator (MLE) has been computed with the `scikit-dimension` package. Figure A3 shows the evolution of the intrinsic dimension as a function of the number of samples for geological models and velocity fields.

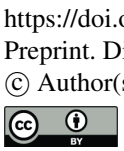


(a) Geological models

(b) Velocity fields

Figure A3. 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 A3a) and velocity fields (Figure A3b).

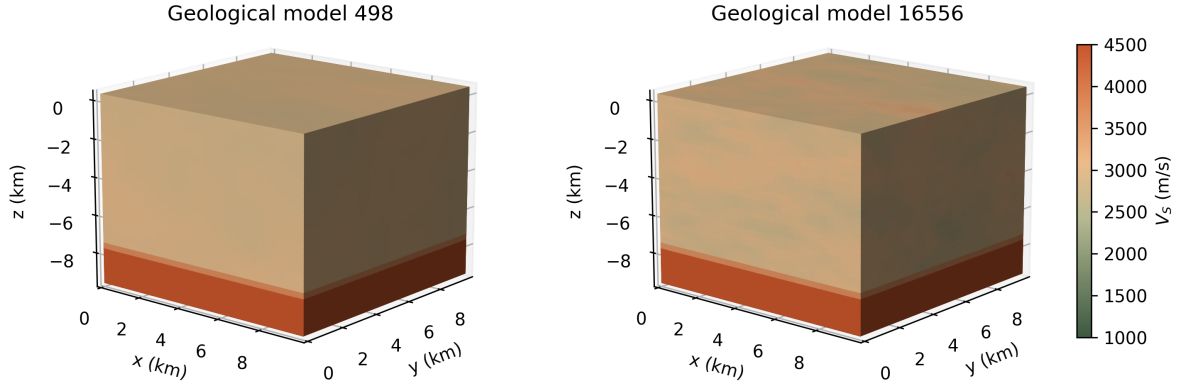
https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024



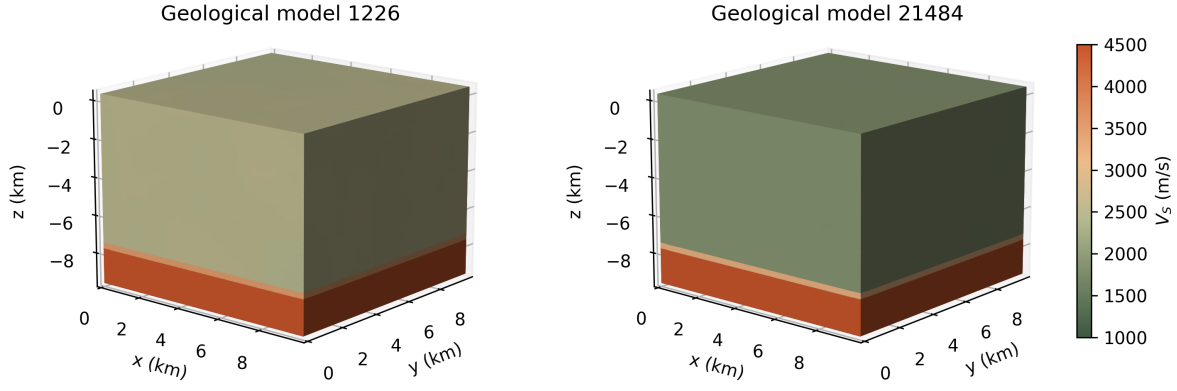
Open Access Earth System Science Data Discussions

A4 Structural Similarity Index

Figure A4 exemplifies two pairs of geological models with high similarity (SSIM of 0.6) but different properties. The first pair (Fig. A4a) has similar mean values but different heterogeneities while in the second pair, geological models are almost homogeneous but exhibit different mean values (Fig. A4b).



(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 A4. Two pairs of geological models with a high SSIM of 0.6

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

Open Access Earth System Science Data Discussions

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).

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

References

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-franceapproved-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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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.

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: Slepiner 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.: $\mathbf{\mathbb{E}\hat{}FWI}$: 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.

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., Villaume, 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.

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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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: A Dataset for Scientific Machine Learning, 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/forum?id=xiNRyrBAjt, 2023b.

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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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.

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.

Raonic, B., Molinaro, R., Rohner, T., Mishra, S., and de Bezenac, E.: Convolutional Neural Operators, in: ICLR 2023 Workshop on Physics for Machine Learning, https://openreview.net/forum?id=GT8p40tiVlB, 2023.

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, http://arxiv.org/abs/2210.14044, 2022.

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.

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.

https://doi.org/10.5194/essd-2023-470
Preprint. Discussion started: 9 January 2024

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.

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.