THE SEISMIC WAVEFIELD COMMON TASK FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Seismology faces fundamental challenges in state forecasting and reconstruction (e.g., earthquake early warning and ground motion prediction) and managing the parametric variability of source locations, mechanisms, and Earth models (e.g., subsurface structure and topography effects). Addressing these with simulations is hindered by their massive scale, both in synthetic data volumes and numerical complexity, while real-data efforts are constrained by models that inadequately reflect the Earth's complexity and by sparse sensor measurements from the field. Recent machine learning (ML) efforts offer promise, but progress is obscured by a lack of proper characterization, fair reporting, and rigorous comparisons. To address this, we introduce a Common Task Framework (CTF) for ML for seismic wavefields, demonstrated here on three distinct wavefield datasets. Our CTF features a curated set of datasets at various scales (global, crustal, and local) and task-specific metrics spanning forecasting, reconstruction, and generalization under realistic constraints such as noise and limited data. Inspired by CTFs in fields like natural language processing, this framework provides a structured and rigorous foundation for head-to-head algorithm evaluation. We evaluate various methods for reconstructing seismic wavefields from sparse sensor measurements, with results illustrating the CTF's utility in revealing strengths, limitations, and suitability for specific problem classes. Our vision is to replace ad hoc comparisons with standardized evaluations on hidden test sets, raising the bar for rigor and reproducibility in scientific ML.

1 INTRODUCTION

Earthquake hazards, tragically illustrated by events like the 1971 M6.7 San Fernando, the 1999 M5.9 Athens (143 fatalities), and the 2011 M5.8 Virginia (up to $300M in damage) earthquakes, are among the most challenging domains for prediction. Their underlying physics are inherently multi-physics and multi-scale, yet current geophysics models cannot fully reproduce the vast observational data that are available. Computational simulations of seismic wavefields must accommodate high-dimensional heterogeneous media, with the numerical complexity increasing with the frequencies that need to be resolved for effective hazard mitigation. Consequently, the seismological community has begun to investigate machine learning (ML) and artificial intelligence (AI) techniques to accelerate the accurate reconstruction of wavefields from simulations and data. Early results indicate that AI-accelerated techniques can advance the probabilistic modeling of earthquake ground motions. Recent AI methods for wavefield modeling include Neural Operators (e.g., Yang et al., 2023; Zou et al., 2024; Huang & Alkhalifah, 2025; Kong et al., 2025; Lehmann et al., 2024b), Physics-Informed Neural Networks (Moseley et al., 2023), a combination of both (Huang et al., 2025), and/or reduced-order models that leverage many simulations to discover Proper Orthogonal Decomposition and function bases to reconstruct wavefields at low costs (Rekoske et al., 2023; 2025).

The rapid development and adoption of these methods on seismological data and simulations has outpaced efforts to compare them objectively. In the absence of a common evaluation standard, new methods are not assessed fairly against existing approaches, resulting in weak baselines, reporting bias, and inconsistent evaluations (McGreivy & Hakim, 2024; Wyder et al.). Few scientific and engineering domains have mitigated these problems, instead relying on self-reporting by providing both training and testing datasets to the community. While reducing the evaluation burden on the original authors, self-reporting opens the door to problematic practices such as *p*-hacking and implicit

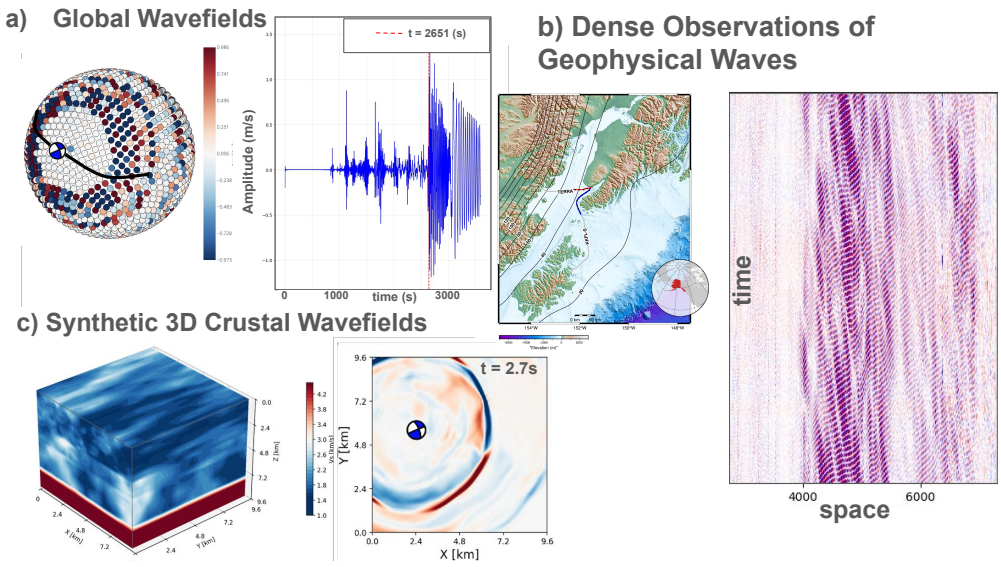

Figure 1: The Seismic Wavefield CTF scores the performance of methods on (a) global wavefields from sparse sensor measurements, (b) dense observations of real geophysical wavefields, and (c) dense simulations of 3D crustal wavefields. These three datasets represent a broad range of challenging datasets encountered frequently by seismologists and present challenges for models in forecasting and state reconstruction.

optimization on the test set. Only with a truly withheld test set is a rigorous and impartial comparison among methods possible.

Common task frameworks (CTFs) have been pivotal in driving and quantifying progress in ML and AI (Donoho, 2017). Landmark CTFs catalyzed key advances: ImageNet (Deng et al., 2009) provided the stage that revitalized Convolutional Neural Networks, with (Krizhevsky et al., 2012) demonstrating their superiority over classical methods in large-scale image recognition. Natural-language challenges, ranging from code generation (Chen et al., 2021) to formal reasoning and mathematics (Hendrycks et al., 2021; Cobbe et al., 2021), have been central to advancing large language models, with recent systems such as DeepSeek-R1-Zero demonstrating competitive performance alongside state-of-the-art models (DeepSeek-AI et al., 2025). Competitive games have served as another CTF: matches in Go and shogi provided the testbed that led to AlphaZero (Silver et al., 2018). Video game environments such as the Arcade Learning Environment (Bellemare et al., 2013), based on Atari 2600 games, enabled the breakthrough results of deep Q-networks (Mnih et al., 2015). Finally, platforms providing models with control inputs, such as the OpenAI Gym (Brockman et al., 2016) and MuJoCo(Todorov et al., 2012), have accelerated the development and testing of reinforcement-learning methods at scale. Mature CTF platforms are therefore critical driving forces of innovation and progress.

Despite these successes, many domains beyond the major areas of computer vision, natural language processing, and reinforcement learning still suffer from a lack of a community standard for fairly evaluating new methods. Indeed, aside from *critical assessment of protein structure prediction* CASP (predictioncenter.org) (Donoho, 2017), science and engineering have largely ignored CTFs and have instead relied on self-reporting benchmarks. Recent work from Wyder et al. has begun to bridge this gap by providing a CTF for scientific ML models on canonical nonlinear systems. But substantial work remains: discipline-specific CTF suites, standardized evaluation protocols (including withheld test sets), and community-maintained leaderboards are needed to ensure fair comparison, reproducibility, and sustained progress.

1.1 The Seismic Wavefield Common Task Framework

We propose a CTF for seismology that is, in its initial release, primarily focused on evaluating ML and AI algorithms modeling the seismic wavefields shown in Fig. 1. Based on the work from Wyder et al., this CTF provides training datasets with explicit tasks related to forecasting and reconstruction

under various challenges, such as noisy measurements, limited data, and varying system parameters. Participants submit predictions for a withheld test set over specified timesteps. The predictions are evaluated and scored on a diverse set of metrics by an independent referee and posted on a leaderboard.

Scoring is by nature reductive—reducing a method's performance to a single floating point value. We adopt a multi-metric scoring scheme because a single score is often insufficient to characterize suitability for different scientific uses. As a result, we use the carefully designed twelve-score system from Wyder et al. which maps to critical tasks for seismic wavefield data objectively. A summary, or composite score, is also computed that gives the overall score for a given method. Task-specific and overall rankings are highlighted in this paper and displayed on a leaderboard.

For each submission we generate a radar plot that visualizes the twelve task scores (see Fig. 2-e). This profile quickly conveys strengths and weaknesses—e.g., robustness to noise, performance in limited-data regimes, or parametric generalization—so users can choose methods suited to their needs. The composite score is the mean of the task scores; using multiple task-specific metrics avoids a winner-take-all outcome and promotes methods that are fit-for-purpose across diverse seismological applications.

Once the **ctf4seismology** is launched[1], we invite the community to evaluate their methods on the Seismic Wavefield CTF by taking the following steps:

1. Sign-up and Sign-in on Kaggle
2. Train your model with our training data and generate predictions for each task
3. Submit prediction files to the competition platform
4. See your score on the leaderboard

To interact with **ctf4seismology** before the competition launches, visit our GitHub repository, install the **ctf4seismology** package, and evaluate your method on the `seismo` dataset. Our dataset and Python package don't require high-performance hardware and can be run on a laptop.

2 Datasets & Evaluation Metrics

We extend the CTF with three new challenging seismic wavefield datasets and evaluate the first one, the global wavefields dataset, on several commonly used models in scientific machine learning (See Fig. 1). The global wavefields dataset exhibits complex and challenging behavior for the tasks of reconstruction and forecasting under the constraints of noise, limited data, and parametric dependence. While this dataset serves as a starting point, the CTF will evolve to include both more complex data and more challenging tasks. The Seismic Wavefield CTF is a sustainable platform that evolves and grows as the community develops more sophisticated methods and algorithms and faces new challenges.

2.1 Seismic Wavefields as Spatio-Temporal Systems

Seismic wavefields are the response to the elastodynamic wave equation in Earth models of elastic properties that vary in space. When the Earth models are uniform, solutions are simple spatiotemporal fields of ballistic waves: P, S, and, when the Earth's surface is included as a traction-free boundary, surface Rayleigh and Love waves. Spatial and depth variations in Earth properties distort and scatter the wavefields, yielding great complexity in the spatio-temporal pattern of seismic wavefields, with numerical complexity that scales with resolvable seismic frequencies and domain size - the two main bottlenecks in predicting ground motions relevant for seismic hazard analysis.

The first dataset, for which we present results in the leaderboard, is a dataset of globally propagating seismic waves (Fig. 1-a). van Driel et al. (2015) developed an efficient workflow to generate and store Green's functions that can be reused for arbitrary source locations and receivers on the Earth's surface using the AxiSEM numerical solver (Nissen-Meyer et al., 2014). Such a framework has enabled the vast dissemination of complex global wavefields, for Earth and Mars structures, democratizing

[1]We are proposing a launch date of March 1, 2026 on Kaggle

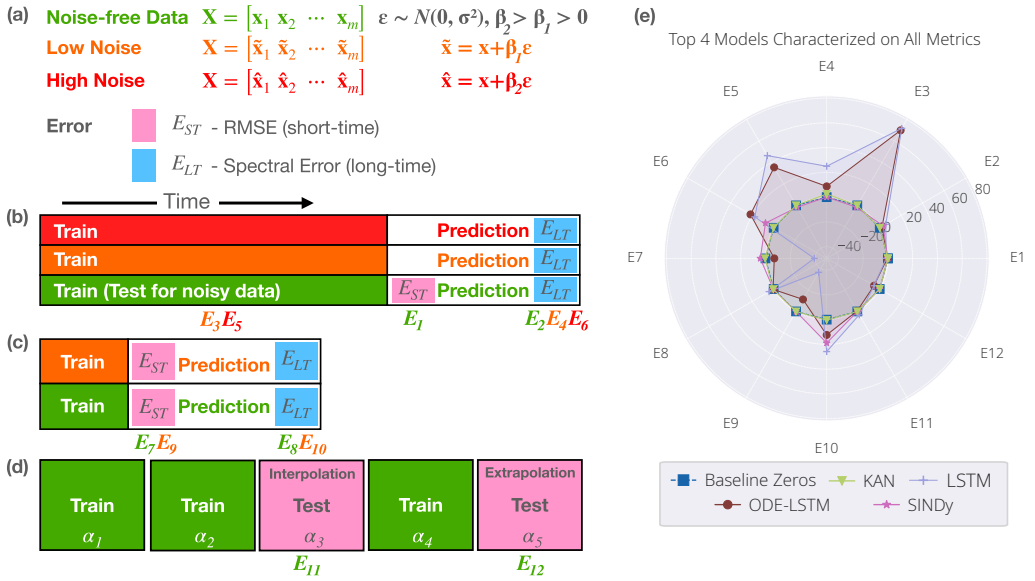

Figure 2: The Seismic Wavefield CTF scores the performance of methods on (e) global wavefields from sparse sensor measurements. (a) Data is collected and organized into matrices, which is then split into testing and training sets. RMSE errors are computed for reconstruction and short-time forecasting, while the spectral error computes the statistics of long-time forecasting (spatial or temporal). (b) Forecasting and reconstruction tasks are evaluated on noise-free, low-noise, and high-noise data. Methods are also evaluated when (c) only limited data is available and (d) for reconstruction of parametrically dependent data.

access of the scientific community to modeled full wavefields (Krischer et al., 2017), which would be otherwise computationally intractable for most scientists. We leverage this database of Green's function to construct a series of earthquake wavefields recorded at 2048 sensors located at the surface of a sphere. The wavefields are computed in the IASP91 Earth model that was designed to match the arrival time of P and S waves (Kennett & Engdahl, 1991) and has become a standard of radially symmetric Earth models, where seismic properties (P wavespeed, S wavespeed, and Earth material density) increased as a function of depth. Our dataset is built upon these Green's functions, which, when convolved with a source mechanism, deliver a realistic seismogram. We choose 2048 sensors distributed on a Fibonacci sphere of 6371 km radius, a sampling rate of 1 Hz (resampled from the original simulations that resolved as short as 2-second period waves), and time series up to 3600 s. Each dataset file contains NumPy arrays with shape (time_steps, 2048) representing the vertical z-component of velocity seismograms. Given that realistic waveform data span many orders of magnitude, we normalize the dataset to have zero mean and unit variance per earthquake event, ensuring temporal continuity and predictability for each task.

The second dataset (Fig. 1-c) is an extension of the curated datasets from Lehmann et al. (2024a). It comprises synthetic 3D seismic wavefields in a heterogeneous 3D crustal model. Earthquakes were modeled as point sources with a double-couple mechanism represented by 6 parameters; source location and focal mechanism (strike, dip, and rake angle) were drawn at random within the model volume. The modeled spatial and temporal scales are aligned with those used in recent AI models of wavefields (e.g., Kong et al., 2025; Rekoske et al., 2025), relevant for seismic analysis of crustal earthquakes that can pose substantial risk to people and infrastructure when they occur in populated areas. Ten independent simulations were produced. Each simulation yields three-component velocity seismograms on a $32 \times 32 \times 32$ heterogeneous grid. Virtual sensors form a 94×94 grid arranged on the top of the model volume with 100 m spacing. These seismograms are sampled for 6 seconds at a 50 Hz.

The third dataset uses a novel geophysical sensing technology that leverages optical scattering to transform telecommunication fibers into arrays of virtual sensors. Referred to as Distributed Acoustic Sensing (DAS) (Fig. 1-b), this technology is revolutionizing the observations of earthquakes (e.g., Yin et al., 2023), marine mammals (e.g., Wilcock et al., 2023), ocean dynamics (Lindsey et al., 2019), and

structural health of offshore wind turbines. Major challenge for DAS are the massive data volumes (Spica et al., 2023; Ni et al., 2023) and the complexity of the wavefields (Xu et al., 2025), motivating AI-based compression and reconstruction (e.g., Ni et al., 2024). We prepared a dataset of ten 1-minute recordings sampled at 5 Hz and low-pass filtered up to 1 Hz. The data presents interesting overlap between earthquake and oceanographic signals at multiple spatial and temporal scales. The channel spacing is 9.57 m and we trimmed the data files to 3000 channels among the 9000 channels available on that particular fiber. Typical waves that dominate the shallow offshore DAS are the surface swell that follows the dispersion relation $\omega^2 = gk \tanh kh$, where ω is the angular frequency, g is the gravity constant, k is the wavenumber, and h is the water depth. For the test set provided, $h \sim 30m$, making surface gravity waves particularly dispersive in the data.

2.2 METRICS

2.2.1 FORECASTING (2 SCORES)

The first set of tasks, shown in Fig. 2-b, involves the approximation of the future state of the system. Thus, given a data matrix representing the dynamics over $t \in [0, 4T]$ ($\mathbf{X}_1 \in \mathbb{R}^{4m\times n}$), a generated forecast is requested from the model being tested for $t \in [4T, 6T]$ ($\mathbf{X}_{1pred} \in \mathbb{R}^{2m\times n}$) to compare with the ground-truth ($\mathbf{X}_{1test} \in \mathbb{R}^{2m\times n}$), with n being the dimension of the system and m being the number of forecasted time steps. The forecasting score is composed of two scores evaluating both the short-time forecast E_{ST}, which is computed using Root Mean Square Error (RMSE) between the test set and the model's approximation, and the long-time forecast E_{LT}, which is computed using the spectral error based upon the power spectral density, see Fig. 2-a. Short-time forecasting measures trajectory accuracy where deterministic prediction is feasible, while long-time forecasting measures statistical fidelity where only the system's broad statistical properties are recoverable.

For the challenge dynamics of interest, sensitivity of initial conditions is common, making long range forecasting to match the test set an unreasonable task given fundamental mathematical limitations with Lyapunov times. Thus, the long-time error is computed by least-squares fitting of the power spectrum $\mathbf{P}(\mathbf{X}, \mathbf{k}, k) = \ln(|\text{FFT}(\mathbf{X}[-k : -1, \mathbf{k}])|^2)$, where the **fftshift** has been used to model the data in the wavenumber domain and $\mathbf{k} = n/2 - k_{max} : n/2 + (k_{max} + 1)$ with $k_{max} = 100$. This means that we look at the match in the first 100 wavenumbers of the power spectrum over a long time simulation. Let $\hat{\mathbf{X}}$ be the ground-truth matrix, $\tilde{\mathbf{X}}$ be the prediction matrix, and $k \in (0, T)$ an integer specifying how to split the matrices for the short-time and long-time scores. The following two error scores are then computed:

$$S_{\text{ST}}(\tilde{\mathbf{X}}, \hat{\mathbf{X}}) = \frac{\|\hat{\mathbf{X}}[1 : k, :] - \tilde{\mathbf{X}}[1 : k, :]\|}{\|\hat{\mathbf{X}}[1 : k, :]\|}, \tag{1}$$

$$S_{\text{LT}}(\tilde{\mathbf{X}}, \hat{\mathbf{X}}) = \frac{\|\mathbf{P}(\hat{\mathbf{X}}, \mathbf{k}, k) - \mathbf{P}(\tilde{\mathbf{X}}, \mathbf{k}, k)\|}{\|\mathbf{P}(\hat{\mathbf{X}}, \mathbf{k}, k)\|}. \tag{2}$$

It is clear that there are many ways to evaluate the long-range forecasting capabilities. We followed in the footsteps of Wyder et al. and chose a simple and transparent metric fully understanding that more nuanced scoring could be used. To provide a reasonable range we then compute the two scores

$$E_1 = 100(1 - S_{\text{ST}}(\mathbf{X}_{1pred}, \mathbf{X}_{1test})), \quad E_2 = 100(1 - S_{\text{LT}}(\mathbf{X}_{1pred}, \mathbf{X}_{1test})), \tag{3}$$

meaning in each case a score of $E_i = 100$ corresponds to a perfect match. Note that, as a baseline, a solution guess of zeros $\tilde{\mathbf{X}}[1 : k, :] = \mathbf{0}$ (corresponding also to $\mathbf{P}(\tilde{\mathbf{X}}, \mathbf{k}, k) = \mathbf{0}$) gives a score of $E_1 = E_2 = 0$.

Input: $\mathbf{X}_{1train} \in \mathbb{R}^{4m\times n}$; **Output:** $\mathbf{X}_{1pred} \in \mathbb{R}^{2m\times n}$; **Scores:** E_1, E_2.

2.2.2 NOISY DATA (4 SCORES)

The ability to handle noise is critical in all data-driven applications as sensors and measurement technologies are by default embedded with varying levels of noise. Methods that work with numerically accurate data, for example data points that are 10^{-6} accurate, may be useful for model reduction, but

are rarely suitable for discovery and engineering design from real-world data. Both strong and weak noise are considered as these represent realistic challenges to be addressed in practice.

This task is very similar to forecasting described above, but now with noise added to the data. Specifically, the model is provided data matrices $\mathbf{X}_{2train} \in \mathbb{R}^{4m\times n}$ and $\mathbf{X}_{3train} \in \mathbb{R}^{4m\times n}$ representing the evolution with low or high noise respectively. The objective is to first produce a reconstruction of the data itself, i.e. denoise the data to produce an estimate of the true state of the dynamics, $\mathbf{X}_{2pred}, \mathbf{X}_{4pred} \in \mathbb{R}^{4m\times n}$ for $\mathbf{X}_{2train}, \mathbf{X}_{3train}$ respectively, and the second objective is to then forecast the future state, matrices $\mathbf{X}_{3pred}, \mathbf{X}_{5pred} \in \mathbb{R}^{2m\times n}$ for $\mathbf{X}_{2train}, \mathbf{X}_{3train}$ respectively. For the reconstruction objective, a least-square fit is used between the approximation of the denoised data and the truth, and for the forecasting objective, a long-time evaluation is computed, leading to the following scores:

$$E_3 = 100(1 - S_{\text{ST}}(\mathbf{X}_{2pred}, \mathbf{X}_{2test})), \quad E_4 = 100(1 - S_{\text{LT}}(\mathbf{X}_{3pred}, \mathbf{X}_{3test})),$$
$$E_5 = 100(1 - S_{\text{ST}}(\mathbf{X}_{4pred}, \mathbf{X}_{4test})), \quad E_6 = 100(1 - S_{\text{LT}}(\mathbf{X}_{5pred}, \mathbf{X}_{5test})).$$

Input: $\mathbf{X}_{2train}, \mathbf{X}_{3train} \in \mathbb{R}^{4m\times n}$; **Output:** $\mathbf{X}_{2pred}, \mathbf{X}_{4pred} \in \mathbb{R}^{4m\times n}$, $\mathbf{X}_{3pred}, \mathbf{X}_{5pred} \in \mathbb{R}^{2m\times n}$; **Scores:** E_3, E_4, E_5, E_6.

2.2.3 Limited Data (4 scores)

Data limitations are common in real world physical systems and often affect the success of data-driven methods. Thus, testing for model performance on low-data is critically important and provides important insight to potential users.

Figure 2-c demonstrates the nature of the task. In this case only a limited number of snapshots M on numerically accurate data are given $\mathbf{X}_{4train} \in \mathbb{R}^{M\times n}$. From this limited data, a forecast must be made which is evaluated with both error metric equation 1 and equation 2 on the approximated future $\mathbf{X}_{6pred} \in \mathbb{R}^{2m\times n}$. The experiment is repeated with noise on the measurements using the training matrix $\mathbf{X}_{5train} \in \mathbb{R}^{M\times n}$ for which a forecasting prediction matrix is produced $\mathbf{X}_{7pred} \in \mathbb{R}^{2m\times n}$. The performance is evaluated on the following scores representing short and long-time metrics for both noise-free and noisy data respectively.

$$E_7 = 100(1 - S_{\text{ST}}(\mathbf{X}_{6pred}, \mathbf{X}_{6test})), \quad E_8 = 100(1 - S_{\text{LT}}(\mathbf{X}_{6pred}, \mathbf{X}_{6test})),$$
$$E_9 = 100(1 - S_{\text{ST}}(\mathbf{X}_{7pred}, \mathbf{X}_{7test})), \quad E_{10} = 100(1 - S_{\text{LT}}(\mathbf{X}_{7pred}, \mathbf{X}_{7test})).$$

Two error scores (analogous to E_1 and E_2) are produced for the noise-free and noisy limited data. These scores are E_7 (short-time forecast) and E_8 (long-time forecast) for the noise free case and E_9 (short-time forecast) and E_{10} (long-time forecast) for the noisy case.

Input: $\mathbf{X}_{4train}, \mathbf{X}_{5train} \in \mathbb{R}^{M\times n}$; **Output:** $\mathbf{X}_{6pred}, \mathbf{X}_{7pred} \in \mathbb{R}^{2m\times n}$;
Scores: E_7, E_8, E_9, E_{10}.

2.2.4 Parametric Generalization (2 scores)

Finally, the ability of a model to generalize to different parameter values is evaluated. For this case, the model's ability to interpolate and extrapolate to new parameter regimes is considered with noise-free data. The interpolation and extrapolation are each their own score, resulting in two scores that evaluate parametric dependence.

Figure 2-d shows the basic architecture of the task. Three training datasets are provided with three different (unknown) parameter values $\mathbf{X}_{6train}, \mathbf{X}_{7train}, \mathbf{X}_{8train} \in \mathbb{R}^{4m\times n}$. Construction of the dynamics in parametric regimes that are interpolatory $\mathbf{X}_{8pred} \in \mathbb{R}^{2m\times n}$ and extrapolatory $\mathbf{X}_{9pred} \in \mathbb{R}^{2m\times n}$ are required. For both of the tasks, a burn in matrix of size $M \times n$ (where $M < m$) is given ($\mathbf{X}_{9train}$ and $\mathbf{X}_{10train}$ respectively) and the performance is evaluated using the short-time metric equation 1.

$$E_{11} = 100(1 - S_{\text{ST}}(\mathbf{X}_{8pred}, \mathbf{X}_{8test})), \quad E_{12} = 100(1 - S_{\text{ST}}(\mathbf{X}_{9pred}, \mathbf{X}_{9test})).$$

Input: $\mathbf{X}_{6train}, \mathbf{X}_{7train}, \mathbf{X}_{8train} \in \mathbb{R}^{4m\times n}, \mathbf{X}_{9train}, \mathbf{X}_{10train} \in \mathbb{R}^{M\times n}$;

Output: $\mathbf{X}_{8pred}, \mathbf{X}_{9pred} \in \mathbb{R}^{2m\times n}$; **Scores:** E_{11}, E_{12}.

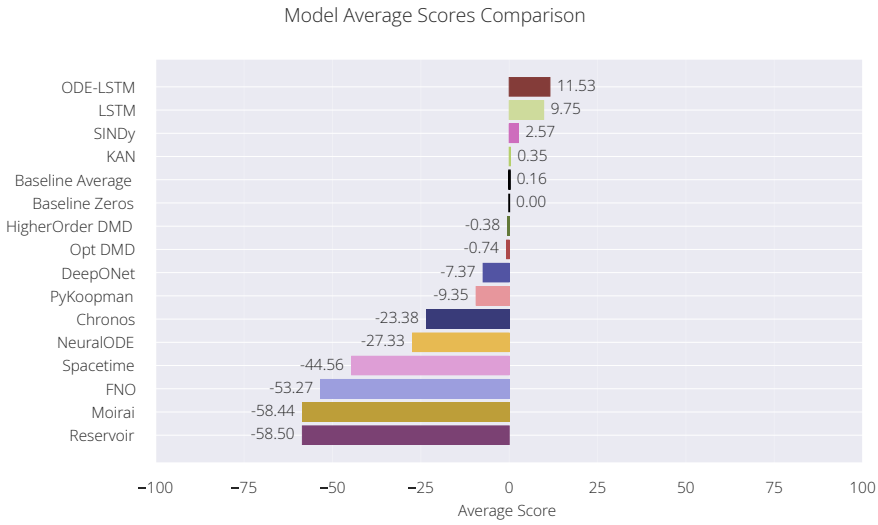

Figure 3: Ranked average scores of each model on the global wavefields dataset.

Table 1: Model Scores on the global wavefields dataset. The evaluated models are: Chronos (Ansari et al., 2024), DeepONet (Lu et al., 2021), FNO (Li et al., 2021), Higher Order DMD (Le Clainche & Vega, 2017), KAN (Liu et al., 2025), LSTM (Hochreiter & Schmidhuber, 1997), Moirai (Liu et al., 2024), NeuralODE (Ruthotto, 2024), ODE-LSTM (Coelho et al., 2024), Opt DMD (Askham & Kutz, 2018), PyKoopman (Brunton et al., 2022; Pan et al., 2024), Reservoir (Jaeger, 2001; Maass & Markram, 2004; Pathak et al., 2018), SINDy (Brunton et al., 2016; Fasel et al., 2022), and Spacetime (Zhang et al., 2023)

Model	Avg Score	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
ODE-LSTM	11.53	-0.57	2.24	69.98	8.71	35.43	21.6	-7.14	0.54	-11.19	12.36	-0.39	-5.33
LSTM	9.75	-1.36	5.63	**71.83**	24.97	**46.41**	17.61	-39.47	4.23	-36.48	25.79	3.16	-5.3
SINDy	2.57	-0.0	3.12	-1.68	0.07	-1.74	7.62	4.18	0.26	0.07	18.89	0.1	0.0
KAN	0.35	-0.0	0.0	0.0	2.55	-0.0	0.13	0.29	0.4	0.08	0.4	-0.0	0.0
Baseline Average	0.16	-0.0	0.0	0.0	0.02	3.59	0.03	-1.73	0.01	-0.02	0.02	-0.03	-0.02
Baseline Zeros	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HigherOrder DMD	-0.38	0.0	-0.0	0.0	0.02	-0.0	0.03	0.01	-1.73	0.02	-0.02	-1.54	-1.34
Opt DMD	-0.74	-4.26	**10.92**	4.39	18.57	0.56	-10.81	-35.65	1.72	-15.6	**27.89**	-3.3	-3.26
DeepONet	-7.37	-0.1	-1.1	6.5	-93.09	2.49	-0.52	-0.66	0.1	-0.66	-1.09	-0.25	-0.05
PyKoopman	-9.35	0.09	0.0	10.54	0.03	0.08	0.06	-3.42	**11.06**	-5.27	-100.0	-25.54	**0.14**
Chronos	-23.38	**17.43**	1.38	-100.0	14.95	-97.0	-59.34	-22.14	0.45	**27.43**	6.93	-35.29	-35.4
NeuralODE	-27.33	12.4	-13.0	-39.3	**30.2**	-100.0	28.1	**8.6**	-100.0	16.8	-60.5	-57.2	-54.1
Spacetime	-44.56	-12.29	-100.0	14.61	-100.0	-8.26	-1.68	-54.8	-100.0	-26.2	-100.0	-34.2	-11.94
FNO	-53.27	-100.0	-100.0	-100.0	-5.34	-100.0	-100.0	-2.56	0.01	-24.02	-100.0	-7.18	-0.14
Moirai	-58.44	-35.4	-50.37	-100.0	-85.68	-82.55	6.1	-35.29	-15.63	-100.0	-94.32	-8.19	-100.0
Reservoir	-58.5	0.76	-100.0	46.5	-100.0	0.0	**50.69**	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0

2.2.5 Composite Score

We compute a composite score ($AvgScore$) per dataset from metrics E_1 through E_{12} by averaging the resulting scores for each method. This score is evaluated per method, not per model. Thus, each method can fit a model for each task and produce the best possible score. All scores are clipped such that $E_i \in [-100, 100]$, thus $AvgScore \in [-100, 100]$. Methods that cannot produce a result for a given task receive the minimum score -100.

3 Methods, Baselines and Results

We characterize twelve highly-cited modeling methods on our initial **ctf4seismology** dataset. Table 1 shows all scored methods and their resulting performance scores. The **ctf4seismology** includes two naive baseline methods: predicting zero and predicting the average. In our evaluations, we use the zero prediction as the reference baseline for the global wavefields dataset. Due to the dataset having zero mean after normalization, the average and zeros baseline report similar scores.

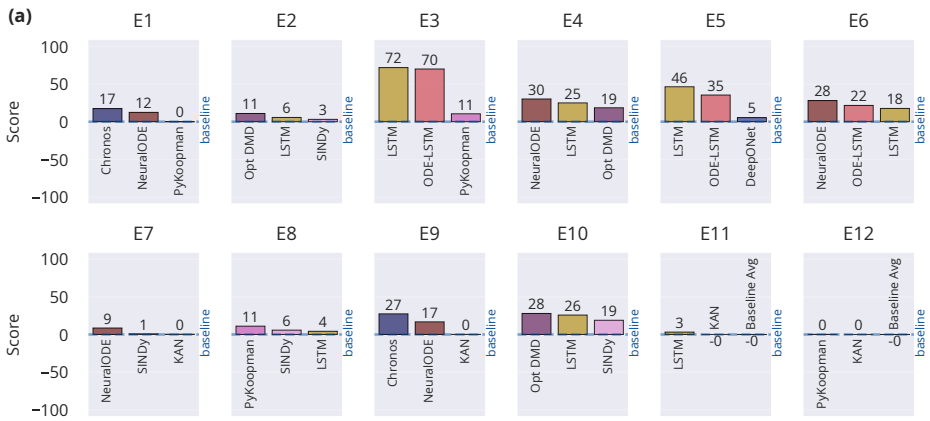

Figure 4: Top three performing models per metric on the global wavefields dataset. Global wavefields uses the constant zero prediction as its baseline.

In Fig. 3, we show all evaluated methods per dataset including the naive baselines—constant and average—ranked by their $AvgScore$. While some models score high on specific tasks, no model scores high-across all tasks (see Table 1). Overall, the results demonstrate that the dataset and specific tasks are challenging enough to produce a distribution of scores that characterize the methods. A complete overview of all model's performance metrics on the global wavefields dataset can be found in table 1. The overall score performance for each method in in Fig. 3 while the top three performers in each error category are shown is shown in Fig. 4.

3.1 Observations

The results in Table 1 demonstrate that many complex ML architectures fail to outperform the baseline of predicting zeros on the global wavefields dataset. Our multi-score evaluation scheme provides deeper insight into the difficulty of the dataset as well as the strengths and weaknesses of the evaluated models. Notably, many commonly used ML/AI models perform very poorly on the assigned tasks. Although the current CTF provides limited data, the difficulty most methods have in exceeding the zero baseline indicates that considerable research and development are still needed before ML/AI models make a meaningful impact in the field. The overall best models are the RNN-based architectures: LSTM and ODE-LSTM (Coelho et al., 2024; Hochreiter & Schmidhuber, 1997). They outperform all other models on tasks E_3 and E_5, corresponding to the denoising of low- and high-noise data. Their advantage likely stems from a relatively modest parameter count combined with relatively strong expressive power, allowing them to act as implicit regularizers when training on the limited data available for auto-regressive forecasting. In contrast, statistical approaches such as DMD tend to overfit the noise in these datasets, whereas the RNNs, trained with an MSE loss, are more robust.

This result also demonstrates how the multi-metric scoring in the CTF is better than a single score. While the RNNs achieve the highest average score, they perform exceptionally poorly on tasks E_7 and E_9, corresponding to short-term prediction on limited data (noiseless and noisy respectively). This disparity underscores that no single model dominates across all regimes and that task-specific evaluation is essential. We present these findings to stimulate discussion within the seismology community about the appropriate choice of ML models for different problem settings and to demonstrate the value of a comprehensive CTF framework. As Figure 4 shows, every task on the global seismic wavefields dataset offers room for improvement that can result in real-world benefit. We hope that, as the CTF grows, more models will emerge that can reliably forecast seismic data, perform state reconstruction, and interpolate across diverse parameter regimes.

4 LIMITATIONS & FUTURE WORK

The globally propagating seismic wavefield datasets used here are generated from an axisymmetric Earth model (van Driel et al., 2015), which captures the increase of seismic velocities and densities with depth but omits the heterogeneous geological structures that drive real-world geodynamics. Future datasets, such as those simulated from the REVEAL project (Thrastarson et al., 2024), or others generated for other planets (e.g., Mars - Stähler et al., 2021), could expand the parameter space and enable a more rigorous assessment of model generalization. Furthermore, distributed acoustic sensing (DAS) recordings with earthquake wavefields as more dominant features would add another layer of complexity. Such data contain higher frequency wavefields that exhibit extreme scattering due to real, near-surface structure heterogeneity (e.g., features highlighted in Shi et al., 2025), challenging AI models to capture earthquake ground motion with societal relevance for natural hazard mitigation. Additional datasets may also come from laboratory experiments of earthquake behavior, which exhibit a particularly complex spatio-temporal pattern relevant to dynamical system studies (e.g., Corbi et al., 2022; 2025). Incorporating such experimental data would broaden the scope of our CTF, increasing their relevance for broader tectonic implications and for scientists conducting laboratory experiments.

Additional limitations are inherent to our provided evaluations. While many models were tested in this work, there are many more that should be implemented and scored in the future. We hope that the Kaggle launch will inspire the scientific community to test many more models than what we have here, with the hope that community-driven engagement results in model improvements that vastly outperform our tested models across all tasks. There is also room for improvement in the tasks provided. While we start the CTF with a set of twelve measurements on the global seismic wavefields dataset, more tasks can be provided, yielding a deeper understanding of a model's capabilities. Finally, for the Kaggle competition we will add a larger number of training data sets to potentially unlock the approximation capabilities of the models. With significantly more training data, there is the potential to improve the overall performance of many of the methods and beat the zero baseline. For instance, we will provide $P = 100$ simulations with varying initial data and parametrizations with the goal of predicting new initial data ($Q = 10$) with different parameter settings.

Input: $\mathbf{X}_{Jtrain} \in \mathbb{R}^{4m\times n}$ for $J = 1, 2, \cdots, P$;

Output: $\mathbf{X}_{Jpred} \in \mathbb{R}^{m\times n}$ for $J = 1, 2, \cdots, Q$; **Scores:** $E_{13} - E_{22}$.

5 CONCLUSION

CTFs have been critical to the tremendous advancements in the big three ML fields of computer vision, natural language processing, and reinforcement learning. They have been proven to be catalysts for key model advancements by providing a clear objective measurement in a competitive environment. This work marks the beginning of incubating a similar environment in seismology by providing a challenging dataset and an objective quantification of model performance on tasks that are notoriously challenging. Our aim is to use our platform to evaluate the current state of methods and their usefulness in seismology as well as fostering an environment where novel methods are developed that excel for specific problem classes.

In this work, we introduce the Seismic Wavefield CTF **ctf4seismology** which quantifies the performance of modeling approaches on an amalgam of diverse tasks in seismology. As a first step, we provide the challenging global seismic wavefields dataset and evaluate 14 different models on the data to understand the usefulness of current ML models in forecasting, state reconstruction, and parametric variability. Our work demonstrates that the current state of ML is far behind any meaningful use on challenging seismic datasets. We hope to inspire researchers and engineers to identify and develop models that will advance the current modeling capabilities of seismic wavefields, ultimately leading to more accurate and reliable tools for earthquake science and subsurface characterization. By establishing a CTF, we aim to shift the research focus from incremental improvements on simplified problems to substantive breakthroughs on complex, realistic challenges. The **ctf4seismology** framework is designed to be a living CTF, with plans to expand the datasets and tasks to continuously push the boundaries of what is possible in computational seismology.

Reproducibility Statement

In this work we introduce the Seismic Wavefield CTF. With the goal of evaluating ML and AI algorithms on seismic wavefields, we provide a publicly available GitHub repository containing the training datasets, hyperparameter tuning scripts, visualization notebooks, and all relevant documentation needed to reproduce and extend our results.

Ethics Statement

The datasets in this CTF do not contain private or sensitive information. We release this CTF to encourage rigorous, reproducible research and urge the community to use it responsibly. To our best understanding, the datasets and the source code do not provide harm to the scientific and global communities. Throughout this work we adhered to the ICLR Code of Ethics.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Travis Askham and J Nathan Kutz. Variable projection methods for an optimized dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 17(1):380–416, 2018.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. doi: 10.1613/jair.3912. URL http://dx.doi.org/10.1613/jair.3912.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. 2016. URL https://arxiv.org/abs/1606.01540.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1517384113.

Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340, 2022. doi: 10.1137/21M1401243. URL https://doi.org/10.1137/21M1401243.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021. URL https://arxiv.org/abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. 2021. URL https://arxiv.org/abs/2110.14168.

C. Coelho, M. Fernanda P. Costa, and Luis L. Ferrás. Enhancing continuous time series modelling with a latent ode-lstm approach. *Applied Mathematics and Computation*, 475:128727, 2024.

F Corbi, Jonathan Bedford, P Poli, F Funiciello, and Z Deng. Probing the seismic cycle timing with coseismic twisting of subduction margins. *Nature Communications*, 13(1):1911, 2022.

Fabio Corbi, Adriano Gualandi, Giacomo Mastella, and Francesca Funiciello. Scaled seismotectonic models of megathrust seismic cycles through the lens of dynamical system theory. *Seismica*, 4(1), Feb. 2025. doi: 10.26443/seismica.v4i1.1340. URL https://seismica.library.mcgill.ca/article/view/1340.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, 2017.

Urban Fasel, J. Nathan Kutz, Bingni W. Brunton, and Steven L. Brunton. Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2260):20210904, 2022. doi: 10.1098/rspa.2021.0904. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2021.0904.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021. URL https://arxiv.org/abs/2103.03874.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Xinquan Huang and Tariq Alkhalifah. Learned frequency-domain scattered wavefield solutions using neural operators. *Geophysical Journal International*, 241(3):1467–1478, 03 2025. ISSN 1365-246X. doi: 10.1093/gji/ggaf113. URL https://doi.org/10.1093/gji/ggaf113.

Xinquan Huang, Fu Wang, and Tariq Alkhalifah. Physics-informed waveform inversion using pretrained wavefield neural operators. 2025. URL https://arxiv.org/abs/2509.08967.

Herbet Jaeger. "the 'echo state' approach to analyzing and training recurrent neural networks". Technical report, German National Research Center for Information Technology, Technical Report GMD 148, 2001.

B. L. N. Kennett and E. R. Engdahl. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, 105(2):429–465, 05 1991. ISSN 0956-540X. doi: 10.1111/j.1365-246X.1991.tb06724.x. URL https://doi.org/10.1111/j.1365-246X.1991.tb06724.x.

Qingkai Kong, Caifeng Zou, Youngsoo Choi, Eric M. Matzel, Kamyar Azizzadenesheli, Zachary E. Ross, Arthur J. Rodgers, and Robert W. Clayton. Reducing frequency bias of fourier neural operators in 3d seismic wavefield simulations through multistage training. *Seismological Research Letters*, 07 2025. ISSN 0895-0695. doi: 10.1785/0220250085. URL https://doi.org/10.1785/0220250085.

Lion Krischer, Alexander R. Hutko, Martin van Driel, Simon Stähler, Manochehr Bahavar, Chad Trabant, and Tarje Nissen-Meyer. On-demand custom broadband synthetic seismograms. *Seismological Research Letters*, 88(4):1127–1140, 04 2017. ISSN 0895-0695. doi: 10.1785/0220160210. URL https://doi.org/10.1785/0220160210.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Soledad Le Clainche and José M. Vega. Higher order dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 16(2):882–925, 2017. doi: 10.1137/15M1054924. URL https://doi.org/10.1137/15M1054924.

F. Lehmann, F. Gatti, M. Bertin, and D. Clouteau. Synthetic ground motions in heterogeneous geologies from various sources: the hemew^{S}-3d database. *Earth System Science Data*, 16(9):3949–3972, 2024a. doi: 10.5194/essd-16-3949-2024. URL https://essd.copernicus.org/articles/16/3949/2024/.

Fanny Lehmann, Filippo Gatti, Michaël Bertin, and Didier Clouteau. 3d elastic wave propagation with a factorized fourier neural operator (f-fno). *Computer Methods in Applied Mechanics and Engineering*, 420:116718, 2024b. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2023.116718. URL https://www.sciencedirect.com/science/article/pii/S0045782523008411.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Nathaniel J. Lindsey, T. Craig Dawe, and Jonathan B. Ajo-Franklin. Illuminating seafloor faults and ocean dynamics with dark fiber distributed acoustic sensing. *Science*, 366(6469):1103–1107, 2019. doi: 10.1126/science.aay5881. URL https://www.science.org/doi/abs/10.1126/science.aay5881.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks, 2025. URL https://arxiv.org/abs/2404.19756.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Wolfgang Maass and Henry Markram. On the computational power of circuits of spiking neurons. *Journal of Computer and System Sciences*, 69(4):593–616, December 2004. ISSN 0022-0000. doi: 10.1016/j.jcss.2004.04.001. URL https://www.sciencedirect.com/science/article/pii/S0022000004000406.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6 (10):1256–1269, Oct 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00897-5. URL https://doi.org/10.1038/s42256-024-00897-5.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. Finite basis physics-informed neural networks (fbpinns): a scalable domain decomposition approach for solving differential equations. *Advances in Computational Mathematics*, 49(4):62, 2023. doi: https://doi.org/10.1007/s10444-023-10065-9.

Yiyu Ni, Marine A. Denolle, Rob Fatland, Naomi Alterman, Bradley P. Lipovsky, and Friedrich Knuth. An object storage for distributed acoustic sensing. *Seismological Research Letters*, 95(1): 499–511, 10 2023. ISSN 0895-0695. doi: 10.1785/0220230172. URL https://doi.org/10.1785/0220230172.

Yiyu Ni, Marine A. Denolle, Qibin Shi, Bradley P. Lipovsky, Shaowu Pan, and J. Nathan Kutz. Wavefield reconstruction of distributed acoustic sensing: Lossy compression, wavefield separation, and edge computing. *Journal of Geophysical Research: Machine Learning and Computation*, 1(3): e2024JH000247, 2024. doi: https://doi.org/10.1029/2024JH000247. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JH000247. e2024JH000247 2024JH000247.

T. Nissen-Meyer, M. van Driel, S. C. Stähler, K. Hosseini, S. Hempel, L. Auer, A. Colombi, and A. Fournier. Axisem: broadband 3-d seismic wavefields in axisymmetric media. *Solid Earth*, 5 (1):425–445, 2014. doi: 10.5194/se-5-425-2014. URL https://se.copernicus.org/articles/5/425/2014/.

Shaowu Pan, Eurika Kaiser, Brian M. de Silva, J. Nathan Kutz, and Steven L. Brunton. PyKoopman: A Python Package for Data-Driven Approximation of the Koopman Operator. *Journal of Open Source Software*, 9(94):5881, 2024. doi: 10.21105/joss.05881. URL https://doi.org/10.21105/joss.05881.

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, 120(2):024102, January 2018. doi: 10.1103/PhysRevLett.120.024102. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.024102. Publisher: American Physical Society.

John M. Rekoske, Alice-Agnes Gabriel, and Dave A. May. Instantaneous physics-based ground motion maps using reduced-order modeling. *Journal of Geophysical Research: Solid Earth*, 128(8): e2023JB026975, 2023. doi: https://doi.org/10.1029/2023JB026975. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023JB026975. e2023JB026975 2023JB026975.

John M Rekoske, Dave A May, and Alice-Agnes Gabriel. Reduced-order modelling for complex three-dimensional seismic wave propagation. *Geophysical Journal International*, 241(1):526–548, 02 2025. ISSN 1365-246X. doi: 10.1093/gji/ggaf049. URL https://doi.org/10.1093/gji/ggaf049.

Lars Ruthotto. Differential equations for continuous-time deep learning. *arXiv preprint arXiv:2401.03965*, 2024.

Qibin Shi, Marine A. Denolle, Yiyu Ni, Ethan F. Williams, and Nan You. Denoising offshore distributed acoustic sensing using masked auto-encoders to enhance earthquake detection. *Journal*

of Geophysical Research: Solid Earth, 130(2):e2024JB029728, 2025. doi: https://doi.org/10.1029/2024JB029728. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JB029728. e2024JB029728 2024JB029728.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Zack J. Spica, Jonathan Ajo-Franklin, Gregory C. Beroza, Biondo Biondi, Feng Cheng, Beatriz Gaite, Bin Luo, Eileen Martin, Junzhu Shen, Clifford Thurber, Loïc Viens, Herbert Wang, Andreas Wuestefeld, Han Xiao, and Tieyuan Zhu. Pubdas: A public distributed acoustic sensing datasets repository for geosciences. *Seismological Research Letters*, 94(2A):983–998, 01 2023. ISSN 0895-0695. doi: 10.1785/0220220279. URL https://doi.org/10.1785/0220220279.

Simon C. Stähler, Amir Khan, W. Bruce Banerdt, Philippe Lognonné, Domenico Giardini, Savas Ceylan, Mélanie Drilleau, A. Cecilia Duran, Raphaël F. Garcia, Quancheng Huang, Doyeon Kim, Vedran Lekic, Henri Samuel, Martin Schimmel, Nicholas Schmerr, David Sollberger, Éléonore Stutzmann, Zongbo Xu, Daniele Antonangeli, Constantinos Charalambous, Paul M. Davis, Jessica C. E. Irving, Taichi Kawamura, Martin Knapmeyer, Ross Maguire, Angela G. Marusiak, Mark P. Panning, Clément Perrin, Ana-Catalina Plesa, Attilio Rivoldini, Cédric Schmelzbach, Géraldine Zenhäusern, Éric Beucler, John Clinton, Nikolaj Dahmen, Martin van Driel, Tamara Gudkova, Anna Horleston, W. Thomas Pike, Matthieu Plasman, and Suzanne E. Smrekar. Seismic detection of the martian core. *Science*, 373(6553):443–448, 2021. doi: 10.1126/science.abi7730. URL https://www.science.org/doi/abs/10.1126/science.abi7730.

Solvi Thrastarson, Dirk-Philip van Herwaarden, Sebastian Noe, Carl Josef Schiller, and Andreas Fichtner. Reveal: A global full-waveform inversion model. *Bulletin of the Seismological Society of America*, 114(3):1392–1406, 04 2024. ISSN 0037-1106. doi: 10.1785/0120230273. URL https://doi.org/10.1785/0120230273.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

M. van Driel, L. Krischer, S. C. Stähler, K. Hosseini, and T. Nissen-Meyer. Instaseis: instant global seismograms based on a broadband waveform database. *Solid Earth*, 6(2):701–717, 2015. doi: 10.5194/se-6-701-2015. URL https://se.copernicus.org/articles/6/701/2015/.

William S. D. Wilcock, Shima Abadi, and Bradley P. Lipovsky. Distributed acoustic sensing recordings of low-frequency whale calls and ship noise offshore central oregon. *JASA Express Letters*, 3(2):026002, 02 2023. ISSN 2691-1191. doi: 10.1121/10.0017104. URL https://doi.org/10.1121/10.0017104.

Philippe Martin Wyder, Judah Goldfeder, Alexey Yermakov, Yue Zhao, Stefano Riva, Jan P Williams, David Zoro, Amy Sara Rude, Matteo Tomasetto, Joe Germany, et al. Common task framework for a critical evaluation of scientific machine learning algorithms. In *Championing Open-source DEvelopment in ML Workshop@ ICML25*.

James T Xu, Linqing Luo, Jaewon Saw, Chien-Chih Wang, Sumeet K Sinha, Ryan Wolfe, Kenichi Soga, Yuxin Wu, and Matthew DeJong. Structural health monitoring of offshore wind turbines using distributed acoustic sensing (das). *Journal of Civil Structural Health Monitoring*, 15(2):445–463, 2025. doi: 10.1007/s13349-024-00883-w.

Yan Yang, Angela F. Gao, Kamyar Azizzadenesheli, Robert W. Clayton, and Zachary E. Ross. Rapid seismic waveform modeling and inversion with neural operators. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. doi: 10.1109/TGRS.2023.3264210.

Jiuxun Yin, Marcelo A. Soto, Jaime Ramírez, Valey Kamalov, Weiqiang Zhu, Allen Husker, and Zhongwen Zhan. Real-data testing of distributed acoustic sensing for offshore earthquake early warning. *The Seismic Record*, 3(4):269–277, 10 2023. ISSN 2694-4006. doi: 10.1785/0320230018. URL https://doi.org/10.1785/0320230018.

Michael Zhang, Khaled K Saab, Michael Poli, Tri Dao, Karan Goel, and Christopher Ré. Effectively modeling time series with simple discrete state spaces. *arXiv preprint arXiv:2303.09489*, 2023.

Caifeng Zou, Kamyar Azizzadenesheli, Zachary E Ross, and Robert W Clayton. Deep neural helmholtz operators for 3-d elastic wave propagation and inversion. *Geophysical Journal International*, 239(3):1469–1484, 09 2024. ISSN 1365-246X. doi: 10.1093/gji/ggae342. URL https://doi.org/10.1093/gji/ggae342.