Simulating seismic wavefields using generative artificial intelligen

Rie Nakata[1,2,3], Nori Nakata[1,3,4], Pu Ren[1], Zhengfa Bi[1], Maxime Lacour[5], Benjamin Erichson[1,3], and Michael W. Mahoney[1,3,5]

https://doi.org/10.1190/tle44020123.1

Abstract

Simulating realistic seismic wavefields is crucial for a range of seismic tasks, including acquisition designing, imaging, and inversion. Conventional numerical seismic wave simulators are computationally expensive for large 3D models, and discrepancies between simulated and observed waveforms arise from wave equation selection and input physical parameters such as the subsurface elastic models and the source parameters. To address these challenges, we adopt a data-driven artificial intelligence approach and propose a conditional generative modeling (CGM) framework for seismic wave simulation. The novel CGM framework learns complex 3D wave physics and subsurface heterogeneities from the observed data without relying on explicit physics constraints. As a result, trained CGM-based models act as stochastic wave-propagation operators encoded with a local subsurface model and a local moment tensor solution defined by training data sets. Given these models, we can simulate multicomponent seismic data for arbitrary acquisition settings within the area of the observation, using source and receiver geometries and source parameters as input conditional variables. In this study, we develop four models within the CGM framework — CGM-GM-1D/3D, CGM-Wave, and CGM-FAS — and demonstrate their performance using two seismic data sets: a small low-density data set of natural earthquake waveforms from the San Francisco Bay Area, a region with high seismic risks, and a large high-density data set from induced seismicity records of the Geysers geothermal field. The CGM framework reproduces the waveforms, the spectra, and the kinematic features of the real observations, demonstrating the ability to generate waveforms for arbitrary source locations, receiver locations, and source parameters. We address key challenges, including data density, acquisition geometry, scaling, and generation variability, and we outline future directions for advancing the CGM framework in seismic applications and beyond.

Introduction

Predicting seismic wavefields is a fundamental challenge for a range of seismic tasks in energy exploration and production, including seismic wavefield modeling, data acquisition design, imaging and inversion of subsurface reservoirs, and regularization and interpolation of source and receiver geometries. Seismic wavefields d at a receiver (sensor) x_r generated from a source x_s can be written as

$$d(t,x_s,x_r,m) = S(t,x_s) * R(t,x_r) * G(t,x_s,x_r,m), \quad (1)$$

where $*$ indicates convolution; t denotes the time; S describes source parameters such as source wavelet and mechanisms; R denotes receiver characteristics, coupling, and near-surface static effects; G is the Green's function; and m is the subsurface model.

One popular approach to obtain d is to simulate the wavefields by solving the wave equations numerically, using realistic velocity models m, source and receiver locations x_s, x_r, and source wavelet S as inputs (the term R can be applied postsimulation). This approach, which we refer to as non-machine-learning (non-ML) simulation, may be accurate, but it differs from ground truth observations. This is due to errors in our underlying physics assumptions that determine G (e.g., using acoustic wave equations is the most common approach, but anisotropic viscoelastic behavior is more accurate for the solid earth), the simulation's numerical discretization (e.g., from finite difference methods), and our representation of the subsurface m along estimates in source (e.g., wavelet estimates/assumption) and receiver (e.g., perfect coupling assumption, or no static effects) parameters of S and R. Additionally, even when physically reasonable, the computational cost of the large 3D simulation can be significant.

Machine learning (ML) and artificial intelligence (AI) have emerged as domains with well-developed data-driven methodologies that complement traditional numerical simulation. Many AI/ML approaches can be taxonomized into two streams: discriminative models and generative models. Discriminative models focus on learning labels or targets; for instance, learning an operator f that maps time t to a target wavefield d, as

$$\hat{d} = f(t). \quad (2)$$

For instance, neural operators (NOs) (Sethi et al., 2023; Yang et al., 2023; Zhu et al., 2023) and physics-informed neural networks (PINNs) (Song et al., 2021; Rasht-Behesht et al., 2022; Ren et al. 2024a) have been demonstrated for accelerating the simulation of seismic wave propagation. However, these models still suffer from many of the same limitations as non-ML simulations. For example, PINNs rely on the physics assumptions, and current neural operators use non-ML simulated waveforms that contain subsurface model errors for training. Of course, we note that future neural operators may try to incorporate field data for training to mitigate the issue.

Manuscript received 8 October 2024; revision received 9 December 2024; accepted 7 January 2025.
[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA. E-mail: rnakata@lbl.gov; nnakata@lbl.gov; pren@lbl.gov; zfbi@lbl.gov.
[2]University of Tokyo, Tokyo, Japan.
[3]International Computer Science Institute, Berkeley, California, USA. E-mail: erichson@icsi.berkeley.edu.
[4]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[5]University of California Berkeley, Berkeley, California, USA. E-mail: maxlacour@berkeley.edu; mmahoney@stat.berkeley.edu.

DOI:10.1190/tle44020123.1

In contrast, generative models p aim to learn to approximate the data distribution $p_{data}(d)$ that has generated the observations d. Among other things, this enables the user to generate (new synthetic) waveforms:

$$\hat{d} \sim p(d). \tag{3}$$

Generative models in AI/ML are interesting for modeling seismic waveforms because they aim to directly learn the physics and the underlying subsurface models from simulation/observation data to approximate the true data generating distribution. They have demonstrated great potential in capturing subtle waveform patterns, sophisticated physical laws, and inherent stochasticity that are often challenging to model with existing discriminative methods (Wang et al., 2021; Florez et al., 2022; Lyu et al., 2024; Ren et al., 2024a, 2024b; Shi et al., 2024). They can be taxonomized into unconditional generative models and conditional generative models (CGMs).

Unconditional generative models based on equation 3 provide no mechanism for guidance during the generation process. For example, generating waveforms from $p(d)$ can lead to generally plausible results, but it cannot accurately represent specific data acquisition. On the other hand, conditioning with physical variables v, and considering

$$\hat{d} \sim p(d|v), \tag{4}$$

addresses this limitation by allowing the model to generate wavefields based on additional input physical variables in equation 1, e.g., source-receiver geometries (x_s, x_r). Such CGMs have become a central technique in advancing ML across various fields, including computer vision (Raut and Singh, 2024), natural language processing (Dong et al., 2023), and scientific modeling (Wang et al., 2023).

In this paper, we demonstrate the capability of CGMs in seismic applications and propose a CGM framework for seismic wavefield simulation. By incorporating physical variables in equation 1 as v, the CGM framework can learn underlying physics and subsurface models from data, rather than giving them as an input, as in non-ML simulations, or as constraints, as in PINNs. In the following, we first describe how our CGM framework performs seismic waveform generation, using variational autoencoders (VAEs) as an example. We then demonstrate the powerful waveform generation capabilities of our CGM framework, discussing challenges in generative modeling. We conclude by exploring potential future perspectives.

Conditional generative modeling for seismic wave simulation

If we just use seismic waveforms to train, then (unconditional) generative AI models that perform $\hat{d} \sim p(d)$ generate time series that resemble seismic waveforms. The generated time series exhibit features like P- and S-wave arrival packets and later coda arrivals, but they are not linked to other variables (e.g., source/receiver locations) in equation 1, limiting their usefulness. Time series generated in this manner can be useful for augmenting data for training certain ML models designed for trace-by-trace tasks such as arrival time picking (Wang et al., 2021). However, the generated waveforms tend not to be useful for many other seismic applications that require multiple source-receiver pairs and exploit their properties (i.e., stacking to increase signal to noise or reflection moveout analysis to extract velocity information).

To make the generative AI model more useful, we can incorporate physical variables and terms in equation 1 as conditional variables v in equation 4. The choice of conditional parameters controls what the CGM framework can learn as well as its complexity. When we generate j^{th} component data d_j and use the source-receiver coordinate x_s and x_r along with the source parameters $S(x_s)$ as conditional variables, equation 4 becomes

$$\hat{d}_j \sim p(d_j | x_s, x_r, S). \tag{5}$$

In this case, the CGM framework needs to (implicitly) learn from the data the rest of the terms in equation 1, such as G, m, and R. This means that during training, the CGM framework needs to perform three major seismic tasks: velocity model building to extract subsurface models, m; surface-consistent analysis for near-surface characterization, $R(x_r)$; and wave-equation solver development for modeling. Then, during inference, the CGM framework can be used in a similar manner as conventional non-ML simulations, but it completes them much faster, without needing to specify the physics and the subsurface model. This illustrates the strong potential of the CGM framework in avoiding uncertainties in subsurface models and mitigating errors in physics. At the same time, it also indicates complex and nontrivial aspects of the learning setup. One way to simplify the problem is to use a source-receiver offset (distance) D instead of the source-receiver coordinates, in which case equation 5 becomes

$$\hat{d}_j \sim p(d_j | D, S), \tag{6}$$

and equation 1 becomes

$$d_j(t, D, m_{1D}) = S(t, x_s) * R(t) * G_j(t, D, m_{1D}). \tag{7}$$

In this simplification, the CGM framework needs to learn 1D wave physics and subsurface models along with averaged receiver characteristics. The training of generative models becomes easier, at the cost of accuracy in data representation, as we will see in the following section.

Conditional dynamic variational autoencoder model

A VAE model is designed to learn low-dimensional representations of complex data and generate their variations. The VAE extends the basic concept of an autoencoder, which is a model that compresses input data into a deterministic latent space and then reconstructs an approximation to the data. Different from traditional autoencoders, VAEs introduce a probabilistic framework, where the latent space represents a distribution of possible features rather than a single (deterministic) compressed value. Due to its simplicity and well-established theoretical foundations, the Gaussian distribution, which can be characterized by its mean μ and standard deviation σ, is a common choice for this latent space. As illustrated in Figures 1a and 1b, the VAE framework comprises two key components: an

encoder and a decoder. The encoder first compresses the input data, and then a nonlinear neural network known as a multilayer perceptron (MLP) is used to extract the latent features and output the mean, μ, and standard deviation, σ. The decoder then reconstructs the input variable from the latent variable z sampled from the Gaussian latent space, and the output from the VAE is a probability distribution, instead of a fixed (deterministic) value. With this method, we can generate multiple waveforms per any source-receiver pair.

When modeling time series, temporal dependencies are explicitly incorporated into the probabilistic latent space by using a dynamic VAE architecture (Figure 1c). The latent space variable and statistical parameters are now time-dependent $z_{1:\tau}$, $\mu_{1:\tau}$, and $\sigma_{1:\tau}$, where $1{:}\tau$ indicates discrete time series from time 1 to τ. By regarding the time series as sequence data, the temporal evolution can be modeled by incorporating sequence-to-sequence learning approaches. In our case, a recurrent neural network (RNN) is used due to its lightweight architecture and its demonstrated suitability for time-series modeling (Orvieto et al., 2023). The variable $\tilde{z}_0$ (a latent variable at time 0 as initial condition) is sampled from the Gaussian distribution and then fed into the RNN to obtain the latent variable at a later time. During training, we can then obtain the prior distribution $\tilde{\mu}_{0:\tau}$ and $\tilde{\sigma}_{0:\tau}$ used for generations. Finally, the conditional variables (e.g., x_s, x_r, S) are incorporated through embedding by using MLP to extract their latent information, as shown in Figure 1d.

Training a VAE model involves optimizing both the reconstruction loss and a Kullback-Leibler (KL) divergence that measures the difference between the learned and assumed distributions of the latent space. The KL divergence ensures that the latent space follows a smooth distribution (e.g., a Gaussian distribution). The total loss function to minimize during the training is

$$L = L_{rec} + \alpha \cdot D_{KL}, \tag{8}$$

where L_{rec} represents the reconstruction loss (e.g., the mean squared error [MSE] between the reconstructed and observed data), D_{KL} denotes the KL divergence (that measures the difference between the distribution $[\tilde{\mu}_{1:\tau}, \tilde{\sigma}_{1:\tau}]$ for generation and the distribution $[\mu_{1:\tau}, \sigma_{1:\tau}]$ learned from the data), and α is a tuning coefficient. Note that the KL divergence works as a regularization term, as we will see later. This local gradient-based probabilistic optimization is nontrivial, and it can be made efficient using the reparameterization trick that allows gradients to propagate through the probabilistic latent space during training.

In our implementation, we first transform seismic waveforms from the time domain into the time-frequency domain using the short-time Fourier transform (STFT) (Figure 1a), where the waveform features are decomposed to amplitude and phase spectra. Then we use the dynamic VAE model to learn and generate the amplitude spectra per component. Subsequently, the logarithmic time-frequency amplitudes are normalized to a range of [0,1]. Additionally, the conditional variables are normalized independently within the [0,1] range to ensure consistency across all inputs. Now let us define the number of

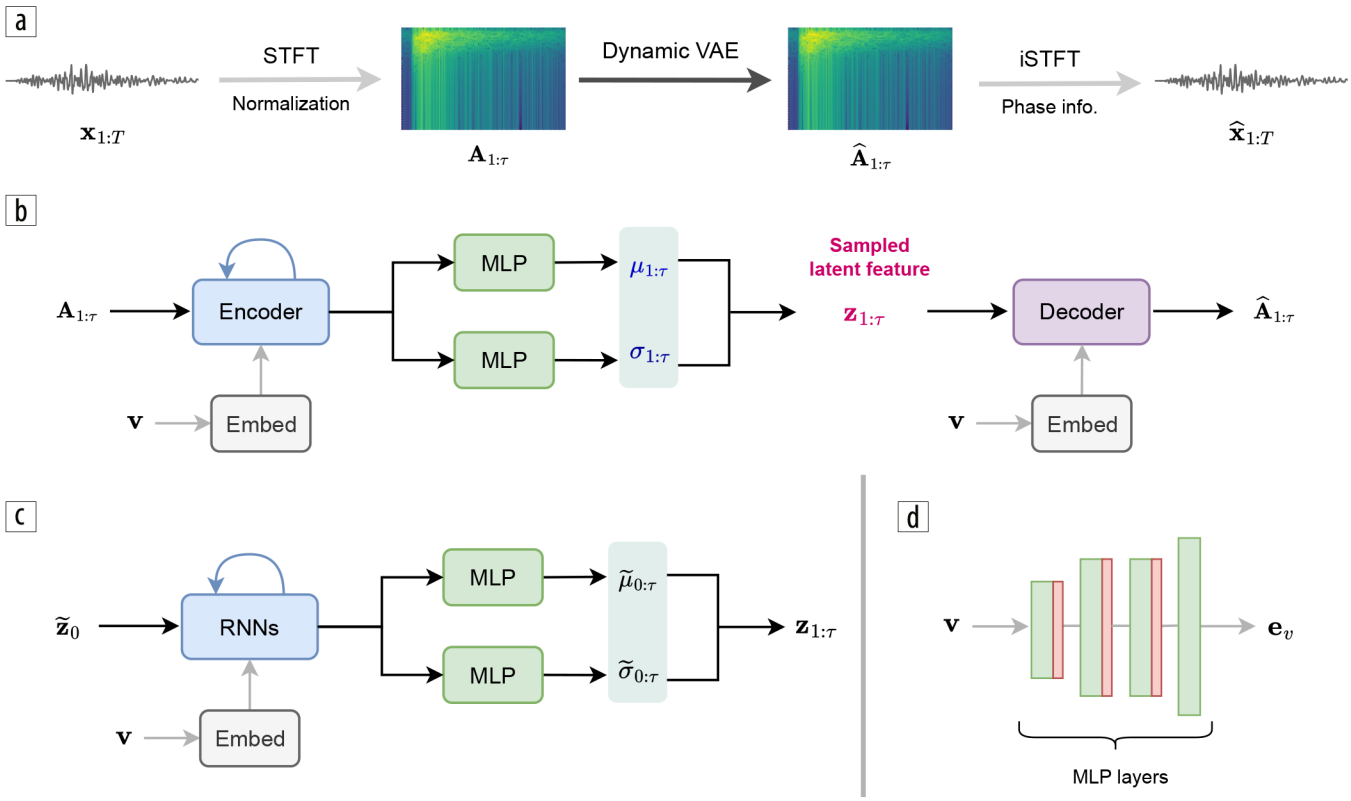

Figure 1. Overview of our CGM framework. Illustrations of (a) the entire process, (b) the core of the dynamic VAE model, which consists of an encoder that compresses the input data into latent features, $z_{1:\tau}$, and a decoder that reconstructs the output from these features, (c) how we use an RNN to model the sequential nature of time-series data, and (d) the embedding process of the conditional variables, v, using multiple layers of neural networks (green) and activation functions (red) to extract latent information.

Table 1. Descriptions of four models used in this study.

Model	Model size	Reconstruction loss	Phase reconstruction	Output	Conditional variables
CGM-GM-1D	0.17 M	Time-domain waveforms, time-frequency-domain amplitude spectra	Griffin-Lim method	Time-domain waveforms	Source location, offset, magnitude
CGM-GM-3D	0.17 M				Source location, receiver locations, source depth, magnitude
CGM-Wave	0.17 M	Time-domain envelope, frequency-domain amplitude spectra	CNN		
CGM-FAS	0.68 M	Frequency-domain amplitude spectra	N/A	Frequency-domain amplitude spectra	

Table 2. Description of data sets before processing used in the study.

Data set	# of earthquakes (per square km)	# of sensors (per square km)	# of samples (per square km)	Area size
SFBA	626,423 (63)	740 (0.07)	626,423 (6)	100 x 100 km
Geysers	30,000 (130)	12,230 (53)	1,000,000 (4,000)	23 x 10 km

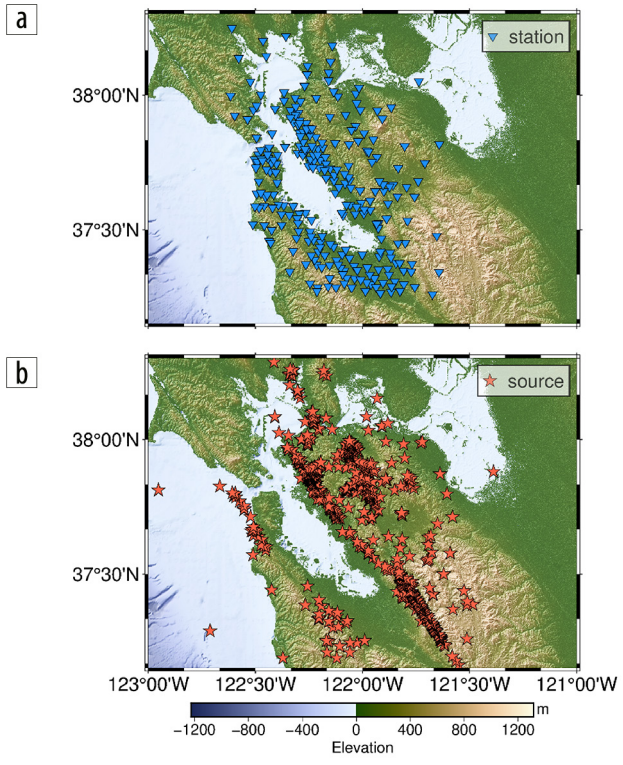

Figure 2. Map of the SFBA along with (a) the sensor distribution and (b) the earthquake source distribution.

batch sizes as b and the frequency and time sequence lengths as $N_f N_f$ and $N_t N_t$, respectively. The training inputs with a tensor shape of $[b, N_t, N_f N_t, N_f]$ are fed into the encoder part, and we obtain a latent feature with a size of $[b, N_z N_z]$, where $N_z N_z$ is the dimension of latent features. The output time-frequency amplitude spectra from the decoder have the same tensor shape as the inputs. We use the true phase information and the inverse short-time Fourier transform to reconstruct time-domain waveforms. Throughout the process, we separate training and test data based on traces, not based on sensor locations.

In the generative process, the inputs of random numbers with a size of $[b, N_z N_z]$ are fed into the decoder along with the conditional variables. As in the training process, the decoder outputs time-frequency amplitude spectra. Then we reconstruct the phase spectra using phase retrieval methods, and we generate time-domain waveforms.

We are developing multiple models to adopt various implementations of the CGM framework. Four are presented in this study (see Table 1 for detailed descriptions). To evaluate their performance, two data examples are used: one from the San Francisco Bay Area (SFBA) and the other from the Geysers geothermal field. These data sets differ significantly in their data characteristics (see Table 2) and thus provide a stress test for the CGM models. Two CGM-GM models (GM stands for ground motion) — the CGM-GM-1D and CGM-GM-3D — are built to generate time-domain waveforms from the natural earthquakes in the SFBA, and we evaluate the effects of the choice of the conditional variables. The CGM-Wave is then developed upon the CGM-GM to improve the performance and to simulate induced-seismicity waveforms from the Geysers geothermal area, and we evaluate data scaling. Finally, the CGM-FAS is used to generate frequency-domain amplitude spectra from the SFBA data, and we evaluate generation variability and data overfit.

San Francisco Bay Area

Located within the San Andreas Fault system, the densely populated SFBA has a history of magnitude 7 or larger earthquakes, and their high seismic risks drive significant interests in predicting seismic waves from potential earthquakes. The large non-ML 3D simulations are computationally expensive; for example, simulating up to 15 Hz takes about 60 hours using approximately 5000 nodes equipped with four NVIDIA A100 GPUs. Our VAE-based CGM framework thus can be an alternative approach for simulating broadband frequency motions.

Our data set consists of two horizontal component waveforms recorded by sensors from small-magnitude (M<4) earthquakes during the period from 1990 to 2022, as shown in Figure 2. We collected a total of 626,423 traces over the 100 × 100 km area, but we discard 98.5% of waveforms that exhibit a signal-to-noise ratio below 3. This results in 10,409 retained traces. This strict data selection is necessary to train the models effectively. Note

that sensor intervals are irregular and sparse for the 15 Hz data, which makes generative modeling further nontrivial.

To evaluate the influence of conditional variables on generating waveforms, we use two CGM-GM models using different conditional variables. We refer to these as CGM-GM-1D and CGM-GM-3D (Table 1). For the CGM-GM-1D, we use source-receiver offset D and source parameter S along with source locations and depths x_s. For the CGM-GM-3D, we include source and receiver coordinates x_s, x_r (both geographical coordinates and depths) along with S. In both cases, S is represented by the earthquake magnitude, as the source wavelet is magnitude dependent. We do not incorporate moment tensor solutions, as many are known to be strike-slip faulting, and the estimates of these small earthquakes either are not available or tend to be unreliable. As we show next, the locally consistent moment tensor solutions are learned and implicitly incorporated into the CGM-GM models. The reconstruction loss is computed using the L2 norm of the amplitude spectra in the time-frequency domain and the waveforms in the time domain. We apply the Griffin-Lim method (Griffin and Lim, 1984), a traditional phase reconstruction method, to reconstruct phase spectra from amplitude spectra. Computational efficiency of our CGM-GM is evident over the physics-based simulations at the same frequency range. Training the CGM-GM-3D takes approximately 2 hours using an A100 GPU, and then inference to generate waveforms takes 0.00178 s per trace.

To demonstrate the differences in the ability of the CGM-GM-1D and the CGM-GM-3D to capture wave propagation physics, we compute amplitude spectra of waveforms generated over a uniform 100 × 100 spatial grid, representing arbitrary sensor locations. The amplitude map for the CGM-GM-1D is shown

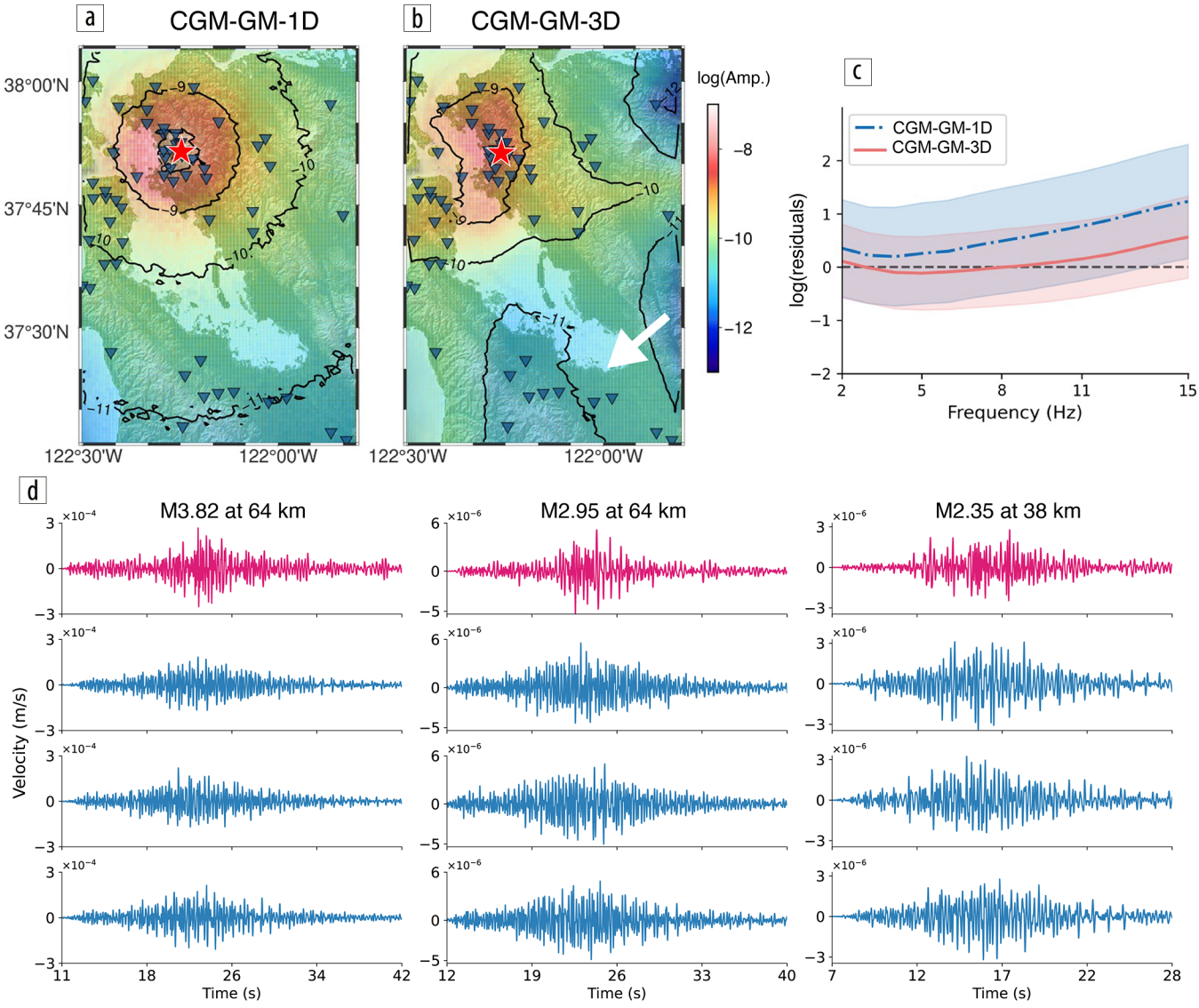

Figure 3. Effects of conditional variables shown for SFBA waveform generation. Spatial variations of the amplitude spectra at 10 Hz using (a) CGM-GM-1D and (b) CGM-GM-3D. The red star and blue triangles denote the earthquake source and receiver locations, respectively. The white arrow in (b) shows the area with large amplitudes. (c) Logarithmic residuals between ground truth (observed) and generated waveforms for all available data. Blue lines represent CGM-GM-1D (as in [a]) and red lines CGM-GM-3D (as in [b]). The solid lines and the shaded area denote the mean curves and the uncertainty region of mean +/- 1 standard deviation. (d) Observed (red) and generated (blue) waveforms for three different source-receiver pairs. In each pair, three realizations of generations are shown.

in Figure 3a for a specific realization. For a 1D medium, we expect that geometrical spreading and attenuation causes amplitudes to decay as a function of the distance, resulting in a symmetric contour map with the source location at its center. Our result aligns with this expectation, showing a radial amplitude decay from the source location. The oscillations in the contours arise from the stochastic nature of generative modeling, where random sampling in the latent space introduces a perturbation in the spatial behavior of waveforms. This result confirms that the CGM-GM-1D captures the 1D earth subsurface averaged across the spatial domain and the 1D wave physics.

In contrast, the amplitude map from the CGM-GM-3D inference, shown in Figure 3b, reveals local features, providing a more nuanced representation of the spatial variability in the waveforms. For example, in the southern SFBA near San Jose (indicated by the white arrow in Figure 3b), the amplitude is large due to the wavefield amplification in the soft Bay Mud layer. These agreements between the generated map and the known geologic and geophysical information strongly support the validity of the spatial recovery achieved by the CGM-GM-3D.

In Figure 3c, we compare waveform residuals in the frequency domain for real earthquakes at real sensor locations by calculating logarithmic differences between amplitude spectra and averaging over all source-receiver pairs. By increasing the complexity of the conditional variables from the CGM-GM-1D to the CGM-GM-3D, we can improve the waveform reconstruction, reducing the residuals.

The generated waveforms from the CGM-GM-3D shown in Figure 3d accurately capture waveform shapes, frequency contents, amplitudes, and arrival times. For instance, for the first earthquake (M=3.82), the CGM-GM-3D successfully replicates the moderate amplitude P-wave packet around 12 s, followed by the large amplitude S-wave and surface wave packets starting at 18 s. In Figure 4, we display a crossplot of P-wave arrival times from the observed and generated waveforms, confirming the accuracy of P-wave first arrival generation. We observe the same performance for the S-wave arrivals (not shown).

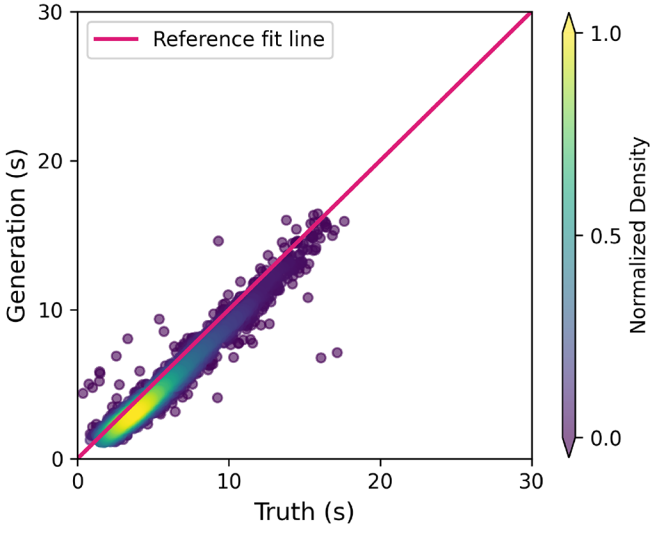

Figure 4. Crossplot of P-wave arrival time between observation (truth) and generation.

Geysers geothermal field

The Geysers geothermal field in Northern California is one of the world's largest and most enduring sources of geothermal energy since the early 1960s. The extensive production has led to significant hydrothermal activity and a high rate of seismic events, with approximately 15,000 annual occurrences based on the United States Geological Survey catalog (see Figure 5a). These induced seismic activities, which concentrate along the Sulfur Creek fault zone and in areas of intense hydrothermal activity, exhibit a positive correlation with steam production and fluid injection processes, thus potentially posing a risk to geothermal operations and nearby infrastructure. The generative modeling approach is particularly appealing given the complex subsurface heterogeneities and dense source and receiver distributions in geothermal regions.

Our CGM-Wave is built to incorporate an ML-based phase-spectra retrieval method and another reconstruction loss. The reconstruction loss of the CGM-Wave is constrained in both the time and frequency domains. Specifically, in the time domain, we compare observed and generated envelopes rather than waveforms, as done in the CGM-GM; and in the frequency domain, the frequency-domain amplitude spectrum is compared with the spectrum of the real data. We employ a convolutional neural network (CNN) to improve the phase spectrum retrieval process and to enhance the overall quality of the generated waveforms, as the Griffin-Lim algorithm often struggles with reconstructing the phase from the amplitude spectrum when the quality of the generation of amplitudes is low.

The CGM-Wave is applied to a data set comprising more than 30,000 seismic events recorded between 2020 and

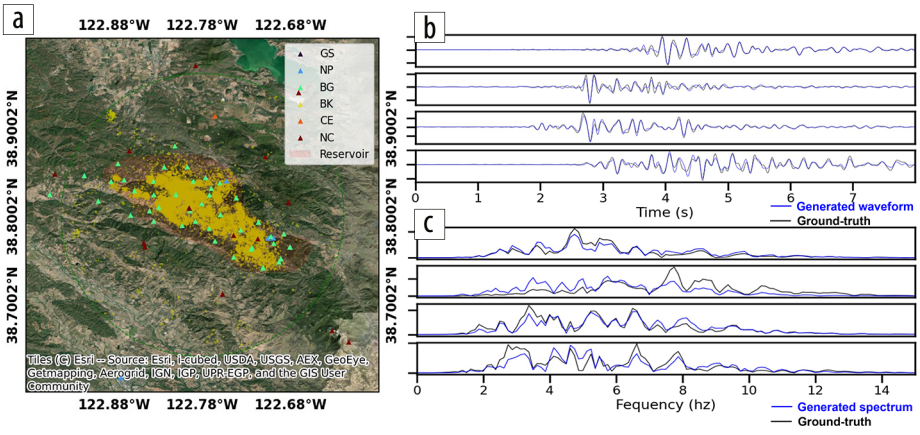

Figure 5. (a) Earthquake and station distribution in the Geysers geothermal field. Triangles represent the locations of seismic stations, while yellow circles indicate the spatial distribution of recorded earthquakes. (b) Comparison of real (black) and synthetic (blue) time-domain waveforms for seismic events excluded from the training data set. (c) Fourier amplitude spectra of observed (black) and synthetic (blue) waveforms.

2023 within the 23 × 10 km area of the Geysers geothermal field (Figure 5a). The number of source-receiver pairs is nearly 1 million.

Figure 5b demonstrates the accuracy of the generated time-domain waveforms by comparing them to ground truth data that were not included in the training set. The close agreement between the synthetic and real waveforms shows that the CGM-Wave successfully captures complex seismic patterns. Figure 5c highlights the alignment between the amplitude spectra of the observed and synthetic data, showing that the model accurately reproduces the spectral characteristics of the seismic events. The generated shot gather of all available receivers associated with a seismic event demonstrates the high fidelity of the CGM-Wave for the entire Geysers field while reproducing the moveout with respect to the distances (Figure 6).

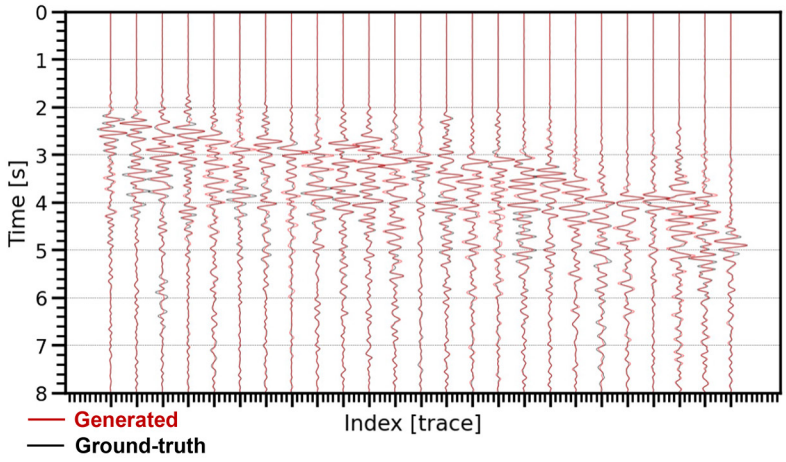

Figure 6. A Geyser geothermal shot gather comparison between generated (red) and real waveforms (black), illustrating a high level of consistency across traces for a single seismic event. Note that the black lines are nearly invisible because of the similarity between the waveforms.

Challenges

Our work has shown the strong potential of the CGM approaches. The two data sets employed differ in terms of natural/induced events, the spatial coverage, and the data density, and they are adequate to test the broad applicability of the CGM framework. However, we have identified several important challenges to be considered in order to successfully use these generative approaches more generally.

Phase reconstruction. The current use of the time-frequency domain is appropriate to capture the rich information inherent in seismic wavefields by decomposing temporal scales along frequency and by separating dynamic and kinematic aspects of the wavefields into amplitude and phase components, respectively. Our CGM framework is designed to reconstruct amplitude spectra through a VAE, while phase spectra, required to reconstruct time-domain signals by the inverse STFT, are derived from the amplitude spectra (either using the Griffin-Lim algorithm originally developed for audio data, as in CGM-GM, or by using a CNN as in CGM-Wave). We found that accurate phase reconstruction remains a challenging task, possibly because the phase varies more rapidly than the amplitude. The phase estimation is highly sensitive to the STFT parameters and the quality of the generated amplitude spectra. This sensitivity often leads to inconsistencies in the reconstructed waveforms, which can degrade the quality of synthetic seismic data.

Data volume, sensor and source locations. The size and quality of the data significantly affect the performance of the generative models. When the data size is small compared to the area of coverage, as in the case of the current SFBA data set (0.6 million samples in total; six samples per square kilometer), careful data curation is crucial because lower-quality data degrade model accuracy and overall performance. In contrast, the larger and denser Geysers data set (1 million samples in total; 4000 samples per square kilometer) allows us to reduce the preprocessing efforts,

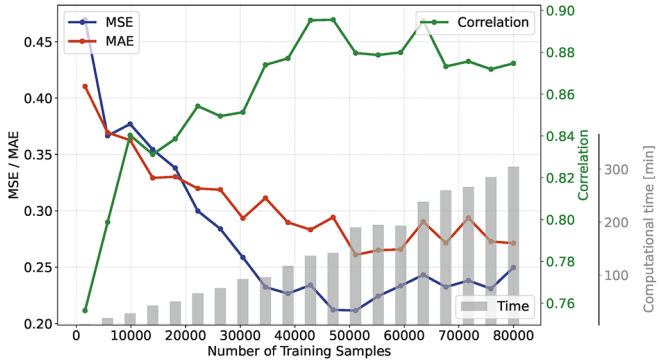

Figure 7. Scaling behavior of the CGM-Wave performance with increasing training data. MSE (red), MAE (blue), the correlation coefficient (green), and computational time (gray bars) are shown.

as the CGM framework can automatically learn to differentiate signal and noise.

The irregular spatial distributions of both sensors and earthquakes create additional challenges. For example, most of the sensor intervals of the SFBA are beyond the Nyquist wavenumber, and thus the waveforms are spatially aliased. Earthquakes are geographically localized along major faults, and wave propagation path coverage is biased. We find that the extrapolation of waveforms is particularly difficult near the boundaries of the sensor network and that the performance degrades in regions with sparse sensor coverage. Despite this, both CGM-GM and CGM-Wave demonstrated strong capability in generating wavefields between sparse sensor locations. The recovered patterns are smooth and reasonable, and we anticipate further improvement in the spatial resolution.

We analyze the scaling behavior of the CGM framework using the CGM-Wave and the Geysers data. We focus on the relationship between training data volume, computational cost, and waveform quality, and we consider three key metrics: MSE, mean absolute error (MAE), and correlation coefficient. These are used to compare observed and simulated time-domain waveforms. Note that we keep the model size consistent across evaluations.

As the training sample size increases, the performance of the CGM-Wave improves, shown by lower MSE and MAE values and a higher correlation coefficient in Figure 7. The MSE, MAE, and correlation coefficient curves show diminishing improvements after approximately 40,000 samples, suggesting that the CGM-Wave reaches a plateau in learning. (Note that adding more data gradually improves the waveform quality; hence, we use 1 million samples in Figure 5.) This steep-to-mild learning curve is typical for small-scale generative models, where the model learns the general trend of data quickly and then hits a gradually diminishing returns regime. Figure 7 also illustrates that computational cost rises linearly, although performance gain with respect to the size follows a nonlinear trend.

This highlights the need for a balanced approach in scaling, ensuring that the improvements in the generated waveform quality are justified against the increased cost in data acquisition and increased computational cost. While larger data sets generally improve the performance of the generative AI, preparing such a data set can require large efforts or may be impractical (i.e., financial costs of acquisition). Efficient computer resource allocation is crucial, especially when working with very large data sets, as training times can become a limiting factor. The findings can guide future efforts in balancing data collection efforts and computational resources to achieve optimal high-quality synthetic seismic waveform generation.

Data fit and generation variability. In traditional inversion/optimization, a typical goal is to fit data as closely as possible, e.g., to reduce the data residuals close to zero. In generative modeling, this is not the case. Errors and incompleteness in acquisition, data preprocessing, and modeling must be accounted for through uncertainties (e.g., standard deviation of the generated data), and variations in the generated data are essential. We apply the CGM-FAS (Fourier amplitude spectra) to the SFBA and show that reducing the residuals excessively limits our capabilities to generate varieties of waveforms. To do so, we use Fourier transform instead of the STFT, omit the RNN part, and construct data loss using amplitude spectra in the frequency domain. Additionally, we focus on wave propagation effects, rewriting equation 1 as

$$d(x_s,x_r,m) = S(x_s) * R(x_r) * (G_o(D,m_{1D}) + \delta G(x_s,x_r,m)), \quad (9)$$

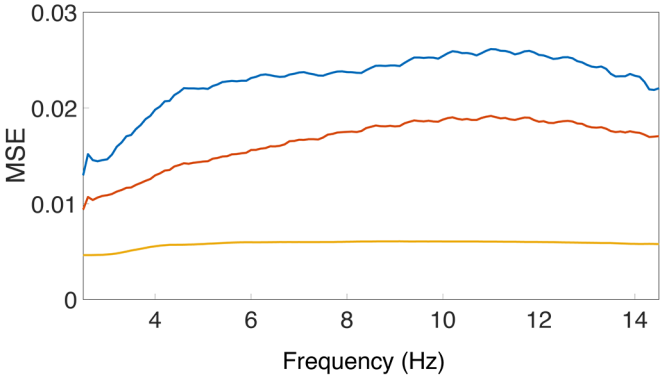

Figure 8. Effects of varying the parameter α on MSEs using CGM-FAS for SFBA data set for (blue) $3e^{-5}$, (orange) $3e^{-4}$, and (yellow) $6e^{-5}$.

where G_o is the background Green's function, and δG is the perturbation in Green's function. We focus on generating δG, and we remove source and receiver effects along with G_o prior to application of the CGM-FAS.

One of the key tuning parameters in equation 8 is α, which controls the balance between MSE and KL divergence. By increasing α from $3e^{-5}$ to $6e^{-4}$, we reduce MSE, as shown in Figure 8. However, this reduces the variability in the generated data, as seen in Figures 9a, where the range of 100 generated amplitude spectra narrows. Similar to wavefields, we anticipate spatially varying uncertainties, for example due to the variations in source and receiver coverages. However, a map view of the standard deviation of generated amplitude spectra in Figure 9b suggests that it varies spatially only at $\alpha = 3e^{-4}$, and that it is almost constant across the area for larger or smaller α values. Interestingly, the median amplitude map in Figure 9c exhibits negligible dependence on the values of α. These observations suggest that $\alpha = 3e^{-4}$ is the optimal value among these three, while the smallest MSE with $\alpha = 6e^{-4}$ may be a result of overfitting. This result highlights the importance of inspecting generation variability to avoid overfitting and to achieve the best balance between accuracy and variability.

Perspectives

We have demonstrated that the CGM framework is capable of learning wave physics and underlying earth models by using key physical variables (e.g., source and receiver coordinates) as conditional variables. Once trained, the models — CGM-GM-1D/3D, CGM-Wave, and CGM-FAS — can generate waveforms and/or amplitude spectra for the trained area for arbitrary acquisition settings (i.e., source and sensor locations and parameters), without needing subsurface elastic models. The use of the trained CGM models can be versatile, enabling rapid seismic modeling for assessing many different acquisition design patterns and testing various geologic scenarios for energy exploration. Their ability to simulate plausible waveforms between sensors increases the data density, which provides opportunities for applying high-end high-quality imaging and inversion techniques such as reverse time migration and full-waveform inversion for sparse acquisition data. In a broader context, the CGM framework can be considered as learning partial differential equations governing physics phenomena. One immediate application area will be ground-penetrating radar data, which consist of high-frequency electromagnetic waves that can be treated as acoustic waves. While some verification is needed, we anticipate extensions to other modalities, such as gravity and magnetic survey data.

We have identified several challenges in the current CGM framework application: phase retrieval; data volume/density and quality; sparse data acquisition; efficient and meaningful training strategies; and generation variability. We note that the models based on the CGM framework require retraining from scratch when changing the region of interest. While our CGM framework is designed to be easily portable, as shown in the SFBA and Geysers applications, training for new regions is not trivial, due to data preparation and parameter tuning. We now discuss future potential directions for addressing these challenges, starting with

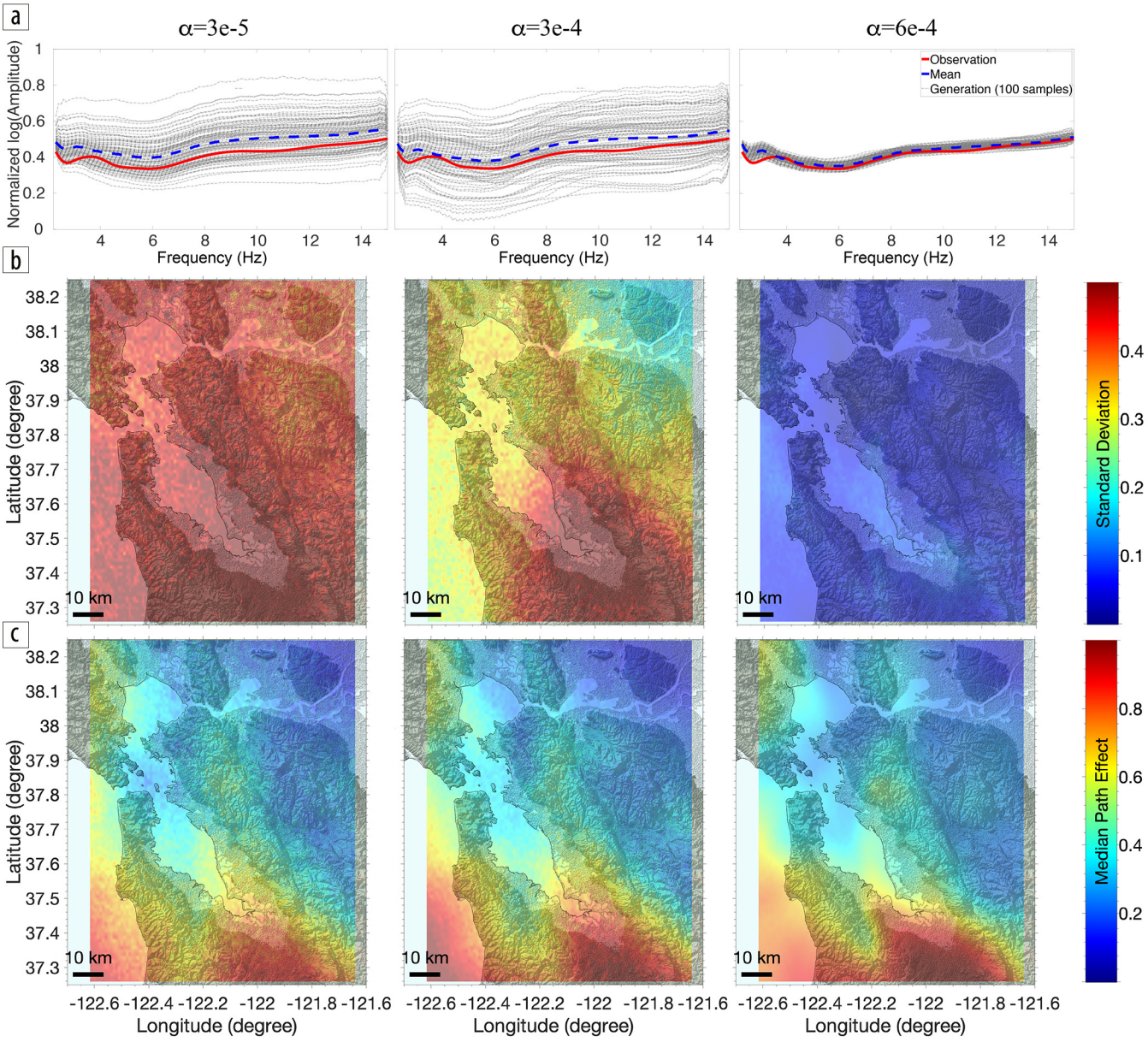

Figure 9. (a) Normalized log amplitude spectra comparisons between observation, generation, and mean of generation for different values of α. (b) Standard deviation and (c) median of generated amplitude spectra data for an earthquake event in SFBA: $\alpha = 3e^{-5}$, $3e^{-4}$, and $6e^{-4}$.

the potential improvements in the current VAE-based CGM framework and then envisioning broader concepts.

Phase retrieval and data acquisition challenges can be mitigated by enhancements to the VAE-based framework. The current CGM framework does not explicitly handle phase information, limiting its capacity to fully capture complex interactions between amplitude and phase in seismic wavefields. Integrating dedicated phase-learning mechanisms or directly building the conditional generative AI model in the time domain could improve performance. Additionally, to better handle sparse sensor data, it may be necessary to incorporate explicit spatial correlations in seismic wavefields, e.g., using the Matérn covariance function (Rasmussen, 2004). Moreover, the CGM framework can benefit from incorporating the correlations of waveforms between different times (i.e., P wave and S wave are not independent).

The CGM framework is not limited to VAE models, and alternative architectures may further improve the performance. We are currently testing generative adversarial networks (GANs) and diffusion models. GANs have an "adversarial" or "competitive" component to their training process. Due to this setup, they have shown potential for generating sharper high-frequency components than VAEs, although mode collapse and training instability remain a challenge. We are exploring hybrid VAE-GAN models that combine the strengths of both. Diffusion models, which transform noise into meaningful data by reversing a gradual noise-adding process, are another promising approach. These models have shown success in image and audio generation, and we are exploring their potential for simulating seismic wave behavior. Success with diffusion models will depend on having access to a diverse, high-fidelity seismic data set, making data design a crucial step in their development.

To address the training costs when switching regions, transfer learning offers a powerful practical solution. By using a model pretrained on one region, we can hope to reduce the data requirements and improve the model robustness when applied to a new region (Subramanian et al., 2023). This approach could also improve the robustness and generalizability of the conditional generative AI models across geographic regions, making it an effective strategy for extending the applicability of generative AI to regions with limited data.

Looking ahead, expanding the transfer learning approach could lead to the development of "foundation models" for seismic waves or more broadly for spatiotemporal phenomena (Bommasani et al., 2021; Subramanian et al., 2023). Foundation models are large neural networks trained on a broad range of data sets across multiple scientific domains to capture unified patterns, which can then be fine-tuned for a variety of specific tasks. For seismic waves, this would involve training on vast amounts of seismic waveforms acquired or simulated both by public and private sectors from both active and passive experiments. Our CGM framework could serve as a backbone for such a model, enabling it to learn generalizable features of seismic wave propagation, facilitating more efficient and accurate waveform generation under different scenarios. Such a foundation model could extend beyond seismic data to other scientific tasks, making it a versatile tool for a range of spatiotemporal phenomena. TLE

Acknowledgments

We would like to acknowledge the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory (LBNL) under U.S. Department of Energy contract no. DE-AC0205CH11231. This research used the computational cluster resources, including LBNL's Lawrencium, and NERSC (under contract no. DE-AC02-05CH11231).

Data and materials availability

Waveform data, metadata, or data products for this study were accessed through the Northern California Earthquake Data Center (https://doi.org/10.7932/NCEDC). Processed data of the SFBA earthquakes are available from Lacour et al. (https://doi.org/10.17603/ds2-necm-5q32).

Corresponding author: rnakata@lbl.gov

References

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, et al., 2021, On the opportunities and risks of foundation models: arXiv preprint, https://doi.org/10.48550/arXiv.2108.07258.

Dong, C., Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, 2023, A survey of natural language generation: ACM Computing Surveys, **55**, no. 8, 173, https://doi.org/10.1145/3554727.

Florez, M. A., M. Caporale, P. Buabthong, Z. E. Ross, D. Asimaki, and M.-A. Meier, 2022, Data-driven synthesis of broadband earthquake ground motions using artificial intelligence: Bulletin of the Seismological Society of America, **112**, no. 4, 1979–1996, https://doi.org/10.1785/0120210264.

Griffin, D., and J. Lim, 1984, Signal estimation from modified short-time Fourier transform: IEEE Transactions on Acoustics, Speech, and Signal Processing, **32**, no. 2, 236–243, https://doi.org/10.1109/TASSP.1984.1164317.

Lacour, M., R. Nakata, N. Nakata, and N. Abrahamson, 2024, Earthquake dataset for the study of small magnitude earthquakes in the San Francisco Bay Area: DesignSafe-CL, https://doi.org/10.17603/ds2-necm-5q32.

Lyu, D., R. Nakata, P. Ren, M. W. Mahoney, A. Pitarka, N. Nakata, and N. B. Erichson, 2024, WaveCastNet: An AI-enabled wavefield forecasting framework for earthquake early warning: arXiv preprint, https://doi.org/10.48550/arXiv.2405.20516.

Orvieto, A., S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu and S. De, 2023, Resurrecting recurrent neural networks for long sequences: arXiv preprint, https://doi.org/10.48550/arXiv.2303.06349.

Rasht-Behesht, M., C. Huber, K. Shukla, and G. E. Karniadakis, 2022, Physics-informed neural networks (PINNs) for wave propagation and full waveform inversions: Journal of Geophysical Research: Solid Earth, **127**, no. 5, e2021JB023120, https://doi.org/10.1029/2021JB023120.

Rasmussen, C. E., 2003, Gaussian processes in machine learning, *in* O. Bousquet, U. von Luxburg, and G. Rätsch, eds., Advanced lectures on machine learning: Springer, 63–71, https://doi.org/10.1007/978-3-540-28650-9_4.

Raut, G., and A. Singh, 2024, Generative AI in vision: A survey on models, metrics, and applications: arXiv preprint, https://doi.org/10.48550/arXiv.2402.16369.

Ren, P., C. Rao, S. Chen, J.-X. Wang, H. Sun, and Y. Liu, 2024a, SeismicNet: Physics-informed neural networks for seismic wave modeling in semi-infinite domain: Computer Physics Communications, **295**, 109010, https://doi.org/10.1016/j.cpc.2023.109010.

Ren, P., R. Nakata, M. Lacour, I. Naiman, N. Nakata, J. Song, Z. Bi, et al., 2024b, Learning physics for unveiling hidden earthquake ground motions via conditional generative modeling: arXiv preprint, https://doi.org/10.48550/arXiv.2407.15089.

Sethi, H., D. Pan, P. Dimitrov, J. Shragge, G. Roth, and K. Hester, 2023, Hard enforcement of physics-informed neural network solutions of acoustic wave propagation: Computational Geosciences, **27**, no. 5, 737–751, https://doi.org/10.1007/s10596-023-10232-3.

Subramanian, S., P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. Mahoney, and A. Gholami, 2023, Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior: arXiv preprint, https://doi.org/10.48550/arXiv.2306.00258.

Shi, Y., G. Lavrentiadis, D. Asimaki, Z. E. Ross, and K. Azizzadenesheli, 2024, Broadband ground-motion synthesis via generative adversarial neural operators: Development and validation: Bulletin of the Seismological Society of America, **114**, no. 4, 2151–2171, https://doi.org/10.1785/0120230207.

Song, C., T. Alkhalifah, and U. B. Waheed, 2021, Solving the frequency-domain acoustic VTI wave equation using physics-informed neural networks: Geophysical Journal International, **225**, no. 2, 846–859, https://doi.org/10.1093/gji/ggab010.

Wang, H., T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, et al., 2023, Scientific discovery in the age of artificial intelligence: Nature, **620**, 47–60, https://doi.org/10.1038/s41586-023-06221-2.

Wang, T., D. Trugman, and Y. Lin, 2021, SeismoGen: Seismic waveform synthesis using GAN with application to seismic data augmentation: Journal of Geophysical Research: Solid Earth, **126**, no. 4, e2020JB020077, https://doi.org/10.1029/2020JB020077.

Yang, Y., A. F. Gao, K. Azizzadenesheli, R. W. Clayton, and Z. E. Ross, 2023, Rapid seismic waveform modeling and inversion with neural operators: IEEE Transactions on Geoscience and Remote Sensing, **61**, 5906712, https://doi.org/10.1109/TGRS.2023.3264210.

Zhu, M., S. Feng, Y. Lin, and L. Lu, 2023, Fourier-DeepONet: Fourier-enhanced deep operator networks for full waveform inversion with improved accuracy, generalizability, and robustness: Computer Methods in Applied Mechanics and Engineering, **416**, 116300, https://doi.org/10.1016/j.cma.2023.116300.