Exploring the scalability and practical usage of non-ergodic ground motion models: A case study of the Los Angeles basin area using a simulated dataset

J. Zhao, G. Lavrentiadis, D. Asimaki,

California Institute of Technology

Abstract

This paper presents a scalable implementation of a non-ergodic ground motion model (NGMM) for the Los Angeles area. The NGMM is trained and validated using simulated ground motion data from a recent SCEC CyberShake study. The NGMM utilizes the ASK14 ergodic ground motion model as its backbone for estimating median intensity measures, complemented by a Gaussian Process (GP) model that accounts for systematic non-ergodic effects. In the simulated dataset, the non-ergodic effects consist of a systematic site effect and a systematic path effect, which are modeled in the GP using Matérn and specialized covariance kernels, respectively. Within-event and between-event kernels are also included in the GP to represent aleatory uncertainties. Implementing the NGMM involves tuning GP hyperparameters and making conditional predictions on large datasets (~1 million datapoints or more). To facilitate hyperparameter tuning, the GPU-accelerated Python package "GPytorch" is employed. For scalable prediction, sparse conditional point selection methods, including k-nearest neighbor and Kullback-Leibler minimization-based approaches, are explored. Preliminary results indicate that the NGMM provides accurate predictions and uncertainty quantification for the simulated data while reducing overall uncertainty in ground motion modeling.

The development of NGMMs will enable more accurate seismic hazard and risk assessments. Future work will focus on exploring and demonstrating the anticipated superior performance of NGMMs in infrastructure risk analysis compared to conventional ergodic ground motion models.

Motivation

Conventional ground motion models (GMMs) use multivariate lognormal (MVN) distributions to represent aleatory uncertainties, i.e., the discrepancy between recorded ground motion intensity measures and GMM predictions. The covariance structure of the MVN distributions is typically expressed through within-event and between-event components, with hyperparameters estimated from empirical data collected across similar tectonic environments. The rationale for combining data in similar tectonic environments is the ergodic assumption, which enriches the data available for developing the GMMs. Although the ergodic assumption provides an accurate prediction of global average ground motions, it also introduces high aleatory uncertainties.

The increasing availability of modern ground motion recordings has motivated the development of non-ergodic ground motion models (NGMMs), which explicitly capture region-specific effects and reduce predictive uncertainty. Previous studies [1, 2] show that NGMM based on Gaussian Process (GP) can successfully reduce predictive uncertainty. The figures below (taken from [2]) demonstrate that such a reduction in predictive uncertainty can decrease the potential for overestimating the probability of unusually large ground motions, resulting in a "steeper" hazard curve. Therefore, NGMMs have the potential to produce more reliable ground motion predictions, as well as avoid overly conservative design.

Despite their improved accuracy, GP-based NGMMs face significant computational challenges when applied to large datasets, which has limited their practical use in seismic hazard and risk analysis. This paper presents a scalable NGMM implementation designed to overcome these challenges.

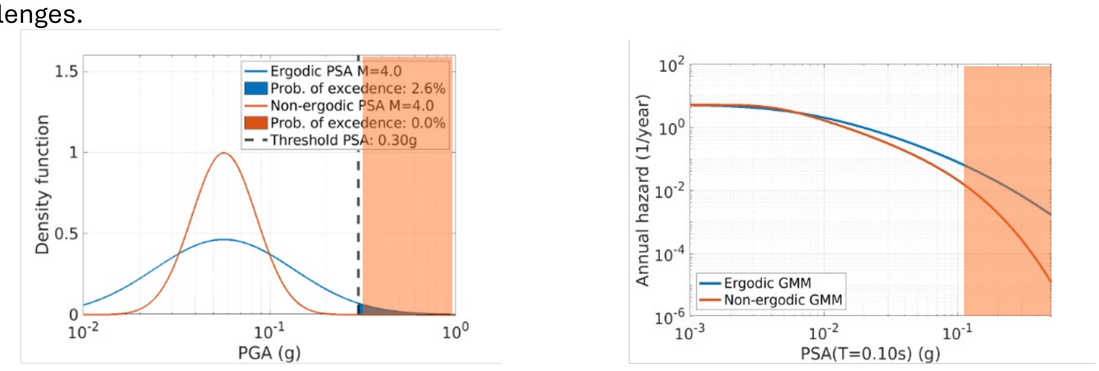


Figure 1. An example of NGMM and GMM predictions.

Training Data: SCEC CyberShake

This study trains the NGMM using data from a recent CyberShake study [3], which simulated 8,358 earthquake ruptures and approximately 420,000 rupture variations with a rupture generation model [3]. The ground motions at 335 sites were computed using a hybrid approach, where low-frequency content (0–1 Hz) was modeled with a finite-difference method, while high-frequency content (1–50 Hz) was generated using stochastic simulations. In this study, half of the earthquake rupture scenarios and 268 out of 335 sites are selected as the training dataset, while the remaining data are used for testing. This study uses one variant to represent each rupture scenario, resulting in nearly 1 million (920,694) observed ground motion data points. Additional rupture variants may be considered in future studies to fully exploit the simulated database.

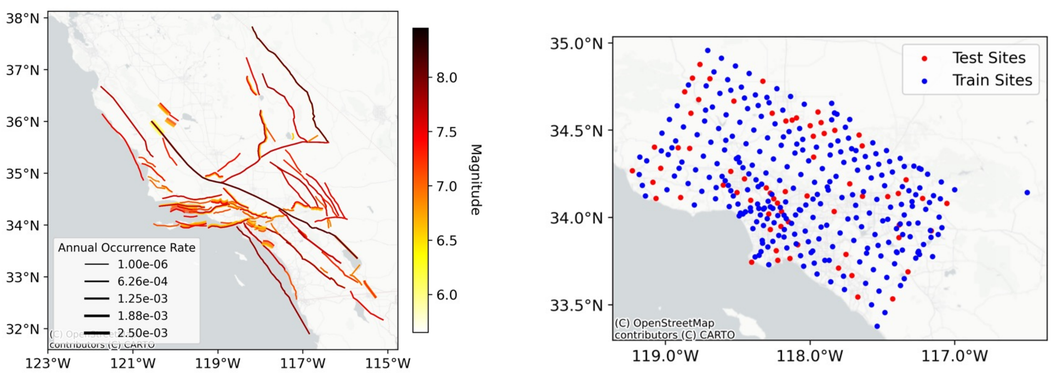


Figure 2. Earthquakes (left) and sites (right) used in NGMM development.

Formulation and scalability of the NGMM

The NGMM adopted in this study has the same formulation as the model in [2], except that the source kernel in [2] is not included. This is because the same rupture generation model is adopted in the CyberShake simulation, and the training data does not capture the non-ergodic source effect.

This study uses the maximum log marginal likelihood method to estimate the GP kernel hyperparameters, which requires repeated evaluation of $\log \mathrm{p}\left(\mathbf{y}_{\mathrm{obs}} \middle| \mathbf{x}_{\mathrm{S,obs}}, \mathbf{x}_{\mathrm{L,obs}}\right)$ and its gradient against the hyperparameters. To enable efficient hyperparameter estimation, the Python package *GPyTorch* is employed. This package enables highly efficient gradient calculation, achieved by interfacing with *PyTorch*, as well as enabling seamless application of GPUs. A mini-batching strategy (similar to stochastic gradient descent) is also adopted to limit GPU memory usage.

Ergodic model

$$\ln(PSa(T)) = \underbrace{f_{GMM}(M, R_{rup}, \dots)}_{\text{Median ground-motion}} + \underbrace{\delta B_e + \delta W_{es}}_{\text{Between- and within-event aleatory}}$$

- $(\delta B_e + \delta W_{es}) \sim \mathcal{N}(0, \kappa_e(h,T))$ where $\kappa_e(h,T)$ is an ergodic covariance (kernel) function that depends on the distance (h) between sites and PSa period (T).
- $\kappa_e(h,T)$ is derived using global data under an ergodic assumption and it describes both true randomness and model simplifications.
- Predictions are made by drawing samples from $\mathcal{N}(0, \kappa_e(h,T))$

Non-ergodic model

$$\delta B_e + \delta W_{es} \approx \underbrace{\delta P2P_{es} + \delta S2S_s}_{\text{Epistemic Terms}} + \underbrace{\delta W_{es}^0 + \delta B_e^0}_{\text{Aleatory terms}}$$

$\delta P2P$: Systematic differences in paths between site-source pair (e.g., repeatable wave propagation effects)
$\delta S2S$: Systematic differences between sites (e.g., repeatable site amplification)

- $(\delta B_e + \delta W_{es}) \sim \mathcal{GP}(0, \kappa_f(\vec{x})) + \delta W_{es}^0 + \delta B_e^0$ where $\kappa_f(\vec{x})$ is a kernel function that depends on a vector of input variables e.g., site coordinates.
- Predictions are made by drawing samples from a conditional distribution of $\boldsymbol{y}_{obs}$ given $\boldsymbol{y}_{prd}$, i.e., $\mathcal{N}(\mathbb{E}[\boldsymbol{y}_{prd}|\boldsymbol{y}_{obs}], \mathrm{COV}[\boldsymbol{y}_{prd}|\boldsymbol{y}_{obs}])$

GP formulation:

$$\begin{bmatrix}\mathbf{y}_{\mathrm{obs}} \\ \mathbf{y}_{\mathrm{prd}}\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I} & \mathbf{k}_{\mathrm{f}} \\ \mathbf{k}'_{\mathrm{f}} & \mathbf{K}_{\mathrm{f}}^*\end{bmatrix}\right)$$, where $\phi_{\mathrm{o}}^2\mathbf{I}$ and $\tau_0^2\mathbf{I}$ are the within- and between-event variability

$$\mathbb{E}[\mathbf{y}_{\mathrm{prd}}|\mathbf{y}_{\mathrm{obs}}] = \mathbf{k}'_{\mathrm{f}}[\mathbf{K}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I}]^{-1}\mathbf{y}_{\mathrm{obs}},\ \mathrm{COV}[\boldsymbol{y}_{prd}|\boldsymbol{y}_{obs}] = \mathbf{K}_{\mathrm{f}}^* - \mathbf{k}'_{\mathrm{f}}[\mathbf{K}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I}]^{-1}\mathbf{k}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I}$$

Log likelihood:

$$\log \mathrm{p}\left(\mathbf{y}_{\mathrm{obs}} \middle| \mathbf{x}_{\mathrm{S,obs}}, \mathbf{x}_{\mathrm{L,obs}}\right) = -\frac{1}{2}\mathbf{y}'_{\mathrm{obs}}[\mathbf{K}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I}]^{-1}\mathbf{y}_{\mathrm{obs}} - \frac{1}{2}\log\left|\mathbf{K}_{\mathrm{f}} + \phi_{\mathrm{o}}^2\mathbf{I} + \tau_0^2\mathbf{I}\right| - \text{constant}$$

Sparse conditional data selection methods, similar to those in [2], are employed to improve the scalability of prediction using NGMMs. These methods select training data that are close to or strongly correlated with prediction data points to calculate $\mathbb{E}[\mathbf{y}_{\mathrm{prd}}|\mathbf{y}_{\mathrm{obs}}]$ and $\mathrm{COV}[\boldsymbol{y}_{prd}|\boldsymbol{y}_{obs}]$, which significantly reduces the size of the matrices being inverted. Methods to select the most informative conditional points are discussed in [4].

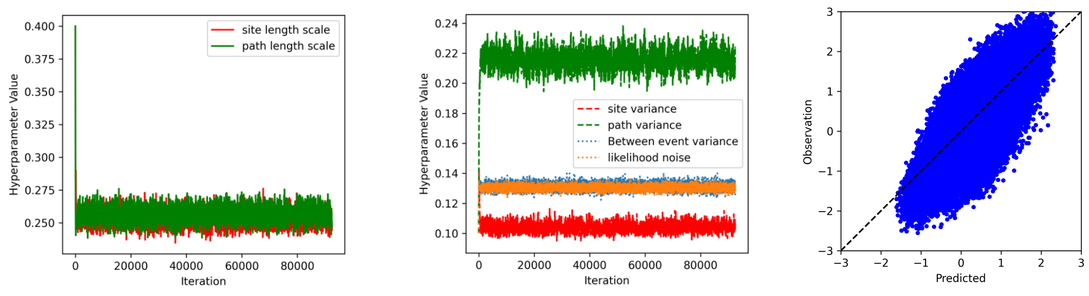


Figure 3. Evolution of the GP hyperparameters during training and a comparison of observations and predictions of the trained model

Preliminary results

Table 1 presents the root mean squared error (RMSE) of NGMM and GMM predictions compared to the CyberShake simulations (used as ground truth). The NGMM developed in this study achieves a reduction of more than 35% in prediction error across all data groups, indicating substantially improved accuracy. Fig. 2 compares hazard curves obtained from the CyberShake and NGMMs trained using 400 and 4,179 earthquakes at an example site. The solid blue and orange lines in Fig. 2 represent hazard curves where exceedance probabilities are computed using the complementary normal CDF, with mean and variance estimated from $\mathbb{E}[\mathbf{y}_{\mathrm{prd}}|\mathbf{y}_{\mathrm{obs}}]$ and $\mathrm{COV}[\boldsymbol{y}_{prd}|\boldsymbol{y}_{obs}]$. Both curves provide reasonable comparison with the CyberShake results, with the NGMM trained on more earthquakes producing a tighter bound than the one trained with fewer earthquakes. Fig. 5 compares the NGMM prediction variance with the GMM mean squared error. These metrics reflect the average predictive uncertainty across all testing scenarios. The results show that the NGMM effectively reduces prediction uncertainty in the majority of testing earthquake scenarios.

Table 1. Model prediction errors

Data Group	Non-ergodic RMSE	Ergodic RMSE
Training earthquakes, training sites	0.387	0.701
Training earthquakes, testing sites	0.443	0.704
Testing earthquakes, training sites	0.405	0.701
Testing earthquakes, testing sites	0.455	0.703

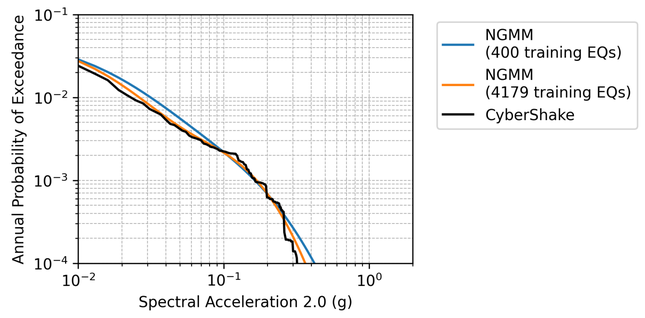


Figure 4. Hazard curves at one studied sites

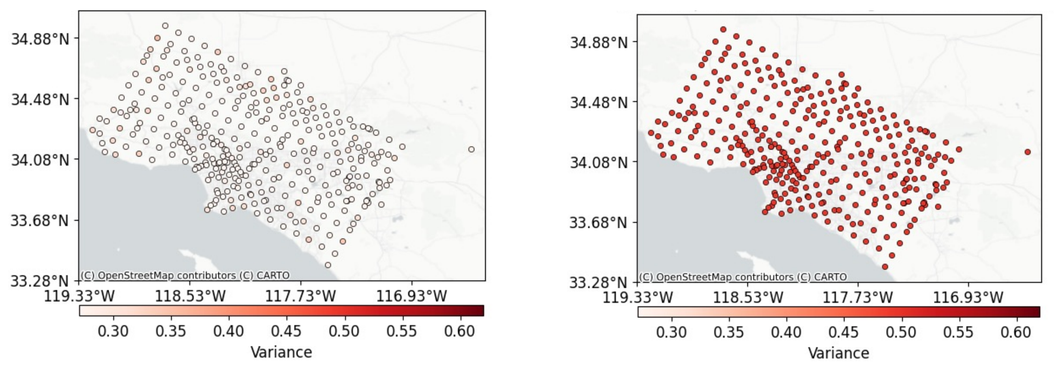
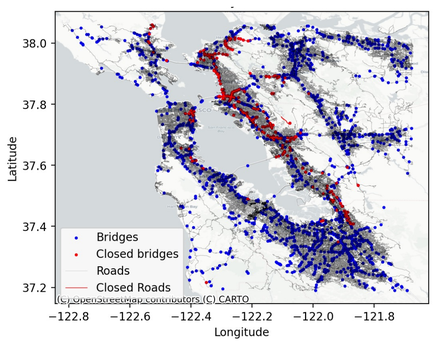


Figure 5. Left: Average prediction variance of the NGMM across testing earthquakes. Right: Mean squared error of the GMM across testing earthquakes.

Ongoing work

Preliminary results show that the developed NGMM provides accurate ground motion prediction, as well as reliable uncertainty quantification. We are currently working on:

- Implement and compare the most effective method for selecting conditional points.
- Demonstrate the NGMM's impact on a regional seismic risk analysis.
- Explore the relationship with underlying geotechnical factors and the NGMM kernels, and use the NGMM to guide the design of experiments and physical model development.



References:

[1] Lavrentiadis, G., Abrahamson, N. A., Nicolas, K. M., Bozorgnia, Y., Goulet, C. A., Babič, A., ... & Walling, M. (2023). Overview and introduction to development of non-ergodic earthquake ground-motion models. *Bulletin of Earthquake Engineering*, *21*(11), 5121-5150.
[2] Lavrentiadis, G., Oral, E., Asimaki, D., Owhadi, H., Susiluoto, J. (In preparation) A non-ergodic ground motion model for the Groninge, Netherlands: merging physics-based simulations and empirical observations.
[3] Graves, R., Jordan, T. H., Callaghan, S., Deelman, E., Field, E., Juve, G., ... & Vahi, K. (2011). CyberShake: A physics-based seismic hazard model for southern California. *Pure and Applied Geophysics*, *168*(3), 367-381.
[4] Huan, S., Guinness, J., Katzfuss, M., Owhadi, H., & Schäofer, F. (2025). Sparse Inverse Cholesky Factorization of Dense Kernel Matrices by Greedy Conditional Selection. *SIAM/ASA Journal on Uncertainty Quantification*, *13*(3)

Acknowledgments

This study was partially supported by the National Science Foundation via the IUCR center Geomechanics and Mitigation of Geohazards and NSF/EAR award #1821853

Grigorios Lavrentidadis was supported by the United States Geological Survey, USGS, under contract G19AC00125.